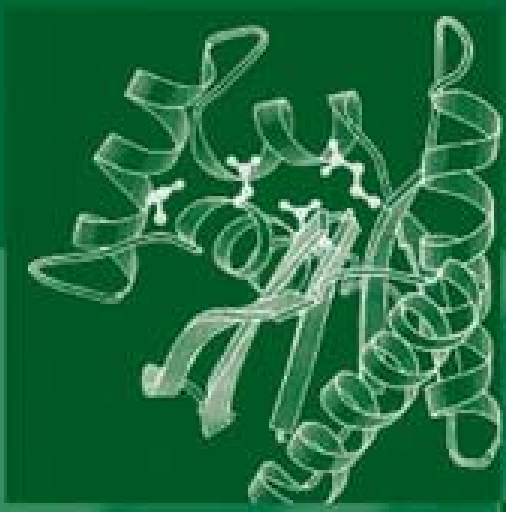
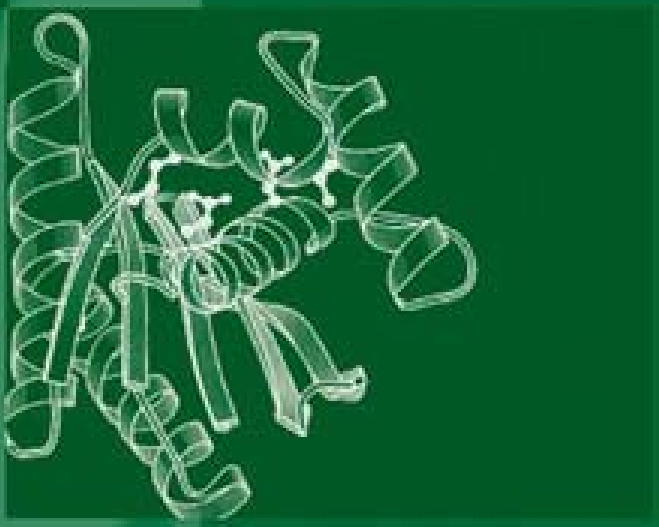
15 Nucleic Acids and Molecular Biology
Hans Joachim Gross (Ed.)

Practical Bioinformatics

Janusz M. Bujnicki (Ed.)





Springer

Nucleic Acids and Molecular Biology

15

Series Editor
H. J. Gross

Janusz M. Bujnicki (Ed.)

Practical Bioinformatics

With 53 Figures, 10 of Them in Color, and 14 Tables

Springer

Dr. Janusz M. Bujnicki
Bioinformatics Laboratory
International Institute of
Molecular and Cell Biology
Trojdena 4
02-109 Warsaw
Poland

ISSN 0933-1891
ISBN 978-3-540-74267-8 e-ISBN 978-3-540-74268-5

Library of Congress Control Number: 2007934269

Springer-Verlag is a part of Springer Science+Business Media

springer.com

5 4 3 2 1 0 – Printed on acid free paper

Preface

The past decade has witnessed not only a flood of protein sequence and structure data generated by large-scale genomic sequencing and structural genomics projects, but also an ensuing growth of size and number of databases and computer programs designed to manage and process these data. The multitude of bioinformatic tools available to molecular biologists offers multiple solutions to various steps of process sequence–structure–function analyses. Often the choice of which tool to use depends more on its popularity among relatively naïve *users*, sometimes stemming from the availability of an intuitive web-server interface, rather than on an understanding of the underlying principles or on the user's ability to utilize all the information returned by the program, including the assessment of confidence of the results. Being educated and trained in molecular biology and biochemistry and self-taught in bioinformatics, I am interested in both the development of computational tools and their optimal application in the realm of experimental biology, especially in the studies of protein–nucleic acid interactions. Despite the abundance of literature on bioinformatics and on molecular biology of proteins that interact with nucleic acids, there are few (if any) timely volumes dedicated to the synthesis of these two research areas. Hence, I was delighted to accept the invitation to act as an editor of a "*Practical Bioinformatics*" volume of *Nucleic Acids and Molecular Biology* and to consolidate key bioinformatic methods for studying protein sequence–structure-function relationships into a convenient source.

This volume is mainly for the biochemist or molecular biologist who wants to analyze, search or manipulate protein structure or sequence data and to integrate these analyses with their experimental investigations to interpret the obtained results or to plan further studies better. Thus, the first part of the volume comprises reviews of methodology solicited from developers of bioinformatic software (with the emphasis on methods that explicitly utilize experimental information and/or are designed to guide experimental research), while the second part comprises useful strategies for studying protein function with the aid of bioinformatics, described in the form of "case

studies" by at-the-bench scientists. Methods and strategies range from protein structure prediction by template-dependent (comparative modeling, fold-recognition) and template-independent (ab initio) approaches, to prediction of protein–protein and protein–nucleic acid interactions, to identification of proteins exerting a defined function or prediction of the function for newly identified proteins. In the spirit of this series, all case studies involve analyses of proteins involved in interactions with nucleic acids – from ribosome assembly and structure, to posttranscriptional RNA modification, to DNA restriction and repair.

The bioinformatics field is a very fast-moving one, and every effort was made to produce this volume as rapidly as possible so the methods would be timely. In this regard, I am grateful to all the authors for taking their time to contribute and for adhering to a set of rigid deadlines; without their participation this volume would not have been possible. I hope that *Practical Bioinformatics* will serve as a useful compendium of methods both to newcomers in the field of bioinformatics-aided experimental molecular biology and biochemistry as well as to scientists actively engaged in research in this area.

Warsaw, July 2003 *Janusz M. Bujnicki*

Contents

Contributors

Alber, Frank

Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry, and California Institute for Quantitative Biomedical Research, Mission Bay Genentech Hall, Suite N472D, 600 16th Street, University of California at San Francisco, San Francisco, California 94143–2240, USA

Bujnicki, Janusz M. (e-mail: iamb@genesilico.pl)

Bioinformatics Laboratory, International Institute of Molecular and Cell Biology in Warsaw, Trojdena 4, 02-109 Warsaw, Poland

Bystroff, Christopher (e-mail: bystrc@rpi.edu)

Department of Biology, Rensselaer Polytechnic Institute, Troy, New York 12180, USA

Catherinot, Vincent

Centre de Biochimie Structurale, INSERM U554 - CNRS UMR5048 - Université Montpellier I. 15, Ave. Charles Flahault, 34060 Montpellier Cedex, France

Cohen-Gonsaud, Martin

Centre de Biochimie Structurale, INSERM U554 - CNRS UMR5048 - Université Montpellier I. 15, Ave. Charles Flahault, 34060 Montpellier Cedex, France

Crécy-Lagard, Valérie de (e-mail: vcrecy@scripps.edu)

Molecular Biology Department, The Scripps Research Institute, BCC-379, 10550 N. Torrey Pines Road, La Jolla, California 92037, USA

Cymerman, Iwona A.

Bioinformatics Laboratory, International Institute of Molecular and Cell Biology in Warsaw, Trojdena 4, 02-109 Warsaw, Poland

Douguet, Dominique

Centre de Biochimie Structurale, INSERM U554 - CNRS UMR5048 - Université Montpellier I. 15, Ave. Charles Flahault, 34060 Montpellier Cedex, France

Droogmans, Louis

Laboratoire de Microbiologie, Université Libre de Bruxelles, 1 av. E. Gryson, B-1070 Bruxelles, Belgium, and Laboratoire de Génétique des Procaryotes, Université Libre de Bruxelles, 12 rue des Professeurs Jeener et Brachet, 6041 Gosselies, Belgium

Eswar, Narayanan

Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry, and California Institute for Quantitative Biomedical Research, Mission Bay Genentech Hall, Suite N472D, 600 16th Street, University of California at San Francisco, San Francisco, California 94143–2240, USA

Feder, Marcin

Bioinformatics Laboratory, International Institute of Molecular and Cell Biology in Warsaw, Trojdena 4, 02-109 Warsaw, Poland

Fischer, Daniel

Bioinformatics, Dept. Computer Science, Ben Gurion University, Beer-Sheva 84015, Israel

Friedhoff, Peter (e-mail: Friedhoff@chemie.bio.uni-giessen.de)

Institut für Biochemie, FB 08, Justus-Liebig Universität Giessen, Heinrich-Buff-Ring 58, 35395 Giessen, Germany

Grosjean, Henri

Laboratory of Structural Enzymology and Biochemistry, CNRS, 1 av. De la Terrasse, 91198 Gif-sur-Yvette, France

Kurowski, Michał A.

Bioinformatics Laboratory, International Institute of Molecular and Cell Biology in Warsaw, Trojdena 4, 02-109 Warsaw, Poland

Labesse, Gilles (e-mail: Labesse@cbs.cnrs.fr)

Centre de Biochimie Structurale, INSERM U554 - CNRS UMR5048 - Université Montpellier I. 15, Ave. Charles Flahault, 34060 Montpellier Cedex, France

Lapeyre, Bruno

Centre de Recherche de Biochimie Macromoléculaire du CNRS, 1919 Route de Mende 34293, Montpellier, France

Linge, Jens P.

Unité de Bio-Informatique Structurale, Institut Pasteur, 25–28 rue du docteur Roux, 75015 Paris, France

Mushegian, Arcady (e-mail: arm@stowers-institute.org)

Stowers Institute for Medical Research, 1000 E 50th St., Kansas City, Missouri 64110, USA

Nilges, Michael (e-mail: nilges@pasteur.fr)

Unité de Bio-Informatique Structurale, Institut Pasteur, 25–28 rue du docteur Roux, 75015 Paris, France

Pawłowski, Marcin

Bioinformatics Laboratory, International Institute of Molecular and Cell Biology in Warsaw, Trojdena 4, 02-109 Warsaw, Poland

Purushothaman, Suresh K.

Centre de Recherche de Biochimie Macromoléculaire du CNRS, 1919 Route de Mende 34293 Montpellier, France

Sali, Andrej (e-mail:sali@salilab.org)

Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry, and California Institute for Quantitative Biomedical Research, Mission Bay Genentech Hall, Suite N472D, 600 16th Street, University of California at San Francisco, San Francisco, California 94143–2240, USA

Shao, Yu

Department of Biology, Rensselaer Polytechnic Institute, Troy, New York 12180, USA

Zharkov, Dmitry O. (e-mail: dzharkov@niboch.nsc.ru)

Institute of Chemical Biology and Fundamental Medicine, Siberian Division of Russian Academy of Sciences, Novosibirsk 630090, Russia

Computational Methods for Protein Structure Prediction and Fold Recognition

I.A. Cymerman, M. Feder, M. Pawłowski, M.A. Kurowski, J.M. Bujnicki

1 Primary Structure Analysis

Amino acid sequence analysis provides important insight into the structure of proteins, which in turn greatly facilitates the understanding of its biochemical and cellular function. Efforts to use computational methods in predicting protein structure based only on sequence information started 30 years ago (Nagano 1973; Chou and Fasman 1974). However, only during the last decade, has the introduction of new computational techniques such as protein fold recognition and the growth of sequence and structure databases due to modern high-throughput technologies led to an increase in the success rate of prediction methods, so that they can be used by the molecular biologist or biochemist as an aid in the experimental investigations.

1.1 Database Searches

Sequence similarity searching is a crucial step in analyzing newly determined (hereafter called "target") protein sequences. Typically, large sequence databases such as the non-redundant (nr) database at the NCBI (synthesis of GenBank, EMBL and DDBJ databases) or genome sequences are scanned for DNA or amino acid sequences that are similar to a target sequence. Alignments of the target sequence are constructed for each database entry, typically using dynamic programming algorithms (Needleman and Wunsch 1970; Smith and Waterman 1981), scores derived from these alignments are used to identify statistically significant matches. Matches which have a low probability of occurrence by chance are interpreted as likely to indicate homology, i.e. that

I. Cymerman, M. Feder, M. Pawłowski, M.A. Kurowski, J.M. Bujnicki
Bioinformatics Laboratory, International Institute of Molecular and Cell Biology in Warsaw, Trojdena 4, 02-109 Warsaw, Poland

Nucleic Acids and Molecular Biology, Vol. 15
Janusz M. Bujnicki (Ed.)
Practical Bioinformatics

the target protein and the matched protein share a common ancestor and their sequences have diverged by accumulating a number of substitutions. However, pairwise similarities (especially if confined to very short regions) can also reflect convergent evolution or simply coincidental resemblance. Hence, percent identity or percent similarity should not be used as a primary criterion for homology. Modern methods for database searches usually employ extreme value distributions to estimate the distribution of the scores between the target and the database entries and a probability of a random match (Pearson 1998; Pagni and Jongeneel 2001) For the search for homologues to be effective and the score to be accurately estimated, the database must contain many unrelated sequences.

Traditionally, searches were carried out using programs for pairwise sequence comparisons like FASTA (Pearson and Lipman 1988) or BLAST (Altschul et al. 1990). However, sequences of homologous proteins can diverge beyond the point where their relationship can be recognized by pairwise sequence comparisons. The most sensitive methods available today use the initial search for homologues to construct a multiple sequence alignment (MSA), which provide insight into the positional constraints of the amino acid composition, and allow the identification of conserved and variable regions in the family, comprising the target and its presumed homologues. The MSA is then converted to a position-specific score matrix (PSSM) and used as a target to search the database for more distant homologues that share similarity not only with the initial target, but with the whole family of related sequences in the MSA. The MSA can be updated with new sequences and searches can be carried out in an iterative fashion until no new sequences are reported with the score above the threshold of statistical significance; PSI-BLAST (Altschul et al. 1997; Aravind and Koonin 1999; Schaffer et al. 2001) is well-optimized and currently the most popular tool in which the PSSM-based search strategy has been implemented. Alternatively to PSSMs, the MSA can be used to create a Hidden Markov Model (HMM), which also can be iteratively compared with the database to identify new statistically significant matches (Karplus et al. 1998).

A related "intermediate sequence search" (ISS) strategy (Park et al. 1997, 1998) employs a series of database scans initiated with the target and then continued with its homologues. Saturated BLAST is a freely available software package that performs ISS with BLAST in an automated manner (Li et al. 2000). This strategy is computationally more demanding than iterative MSA-based searches (all homologues should be used as search targets), but it can sometimes identify links to remotely related outliers, which may be missed by PSI-BLAST or HMM, which preferentially detect sequences most similar to the *average* of the family. However, MSA-based searches can be used to search for new sequences that are compatible with very subtle trends of sequence conservation in the target family, which may be undetectable in any pairwise comparisons. Recently, it was suggested that an increased number of target

homologues can be found by a combination of various pairwise alignment methods for database searches (Webber and Barton 2003). The recommended strategy in database searches (as well as in other bioinformatic tasks) is to use multiple methods and take the agreement between methods as confirmation.

1.2 Protein Domain Identification

Most proteins are composed from a finite number of evolutionarily conserved modules or domains. Protein domains are distinct units of three-dimensional protein structures, which often carry a discrete molecular function, such as the binding of a specific type of molecule or catalysis (reviews: Thornton et al. 1999; Aravind et al. 2002). Proteins can be composed of single or multiple domains. If this information is available, it can be used to make a detailed prediction about the protein function (for instance a protein composed of a phosphodiesterase domain and a DNA-binding domain can be speculated to be a deoxyribonuclease), but if the domain structure is obscure, it can lead to erroneous conclusions about the output of software for sequence analysis.

A common problem in sequence searches is homology of various parts of the target to different protein families, which is often the case in multidomain proteins. Naïve exhaustive ISS searches that detect and use multidomain proteins can result in an erroneous inference of homology between unrelated proteins, which happen to be related to different domains fused together in one of the sequences extracted from a database. Hence, domain identification should be an essential step in analyzing protein sequences, preferably preceding or concurrent to sequence database searches.

A few thousand conserved domains, which cover more than two thirds of known protein sequences have been identified and described in literature. Several searchable databases have been created, which store annotated MSAs (sometimes in the form of PSSMs or HMMs) of protein domains, which can be used to identify conserved modules in the target sequence (Table 1). PFAM and SMART databases are the largest collections of the manually curetted protein domains of information. Each deposited domain family is extensively annotated in the form of textual descriptions, as well as cross-links to other resources and literature references. Both resources contain friendly but powerful web-based interfaces, which provide several types of database search and exploration. The database can be queried using a protein sequence or an accession number to examine its domain organization. Alternatively, the domains can be searched by keywords or browsed via an alphabetical index. Apart from PFAM and SMART there are a number of other databases that classify the domains according to their mutual similarity or inferred evolutionary relationships (Table 1). They differ from each other either through the technical aspects or by concentrating on a specific group of domains. The MSA deposited in these databases as well as their annotations (e.g. in the form

Table 1. Searchable databases of protein domains

Program	Reference	URL (http://)
PFAM	Bateman et al. (2002)	sanger.ac.uk/Software/Pfam/
SMART	Letunic et al. (2002)	smart.embl-heidelberg.de/
TIGRFAMs	Haft et al. (2003)	www.tigr.org/TIGRFAMs/
PRODOME	Servant et al. (2002)	prodes.toulouse.inra.fr/prodom/2002.1/html/home.php
PROSITE	Sigrist et al. (2002)	us.expasy.org/prosite/
SBASE	Vlahovicek et al. (2003)	hydra.icgeb.trieste.it/~kristian/SBASE/
BLOCKS	Henikoff et al. (2000)	bioinfo.weizmann.ac.il/blocks/
COGs	Tatusov et al. (2001)	www.ncbi.nlm.nih.gov/COG/
CDD	Marchler-Bauer et al. (2003)	www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml
INTERPRO	Mulder et al. (2003)	www.ebi.ac.uk/interpro/

of keywords or links to literature and/or other databases) can be generated completely automatically or manually and corrected by experts. The usefulness of each database varies, depending on which problem needs to be solved, so it is reasonable to use more than one method and infer domain boundaries from judicious analysis of all results. In order to facilitate such analyses, the InterPro (Mulder et al. 2003) and Conserved Domain Database (CDD; Marchler-Bauer et al. 2003) have integrated the information from several resources and allow simultaneous searches of multiple domain databases. InterPro and CDD are also used for the primary structural and functional annotation of sequence databases, SWISS-PROT and RefSeq, respectively.

The Clusters of Orthologous Groups (COG) database is one of the most useful resources included in CDD, which may be used to predict protein function or conserved sequences modules. COGs comprise only proteins from fully sequences genomes. COG entries consists of individual orthologous proteins or orthologous sets of paralogs from at least three lineages. Orthologs typically have the same function, so functional information from one member is automatically transferred to an entire COG. The COGnitor tool (http://www.ncbi.nlm.nih.gov/COG/cognitor.html) allows for the comparison of the target protein with the COG database and infers the location of the individual domains, as well as a study of their genomic context, such as the frequency of occurrence of particular genomic neighbors.

1.3 Prediction of Disordered Regions

Recently, it has been suggested that the classical protein structure-function paradigm should be extended to proteins and protein fragments whose native and functional state is unstructured or disordered (Wright and Dyson 1999). Many protein domains, especially in eukaryotic proteins appear to lack a folded structure and display a random coil-like conformation under physiological conditions (reviews: Liu et al. 2002; Tompa 2002). A significant fraction of the intrinsically unstructured sequences exhibits low complexity, i.e. a non-random compositional bias (Wootton 1994).

On the one hand, low-complexity sequences create a serious problem for database searches, as they are not encompassed by the random model used by these methods to evaluate alignment statistics. For instance running a database search with a target sequence including a compositionally biased fragment may lead to erroneous identification of a large number of matches with spuriously high similarity scores. Algorithms such as SEG (Wootton and Federhen 1996) may be used to *mask* the low-complexity segments for database searches.

On the other hand, identification of disordered, non-globular regions may help to delineate domains. Independently folded globular structures can be separated from each other if a flexible linker that connects them is identified. Alternatively, if a protein with many low-complexity regions is known to comprise only a single domain, its rigid core can be identified by *masking off* flexible insertions. The latter case is typical for many proteins from human pathogens such as Plasmodium or Trypanosomes, which use the large flexible loops as hypervariable immunodominant epitopes that contribute to a smoke-screen strategy enacted by the parasite against the host immunogenic response (Pizzi and Frontali 2001). In any case, dissection of the target sequence into a set of relatively rigid, independently folded domains may greatly facilitate tertiary structure prediction, especially by fold-recognition methods (see below). The freely available on-line servers for prediction of disordered *loopy* regions in proteins are: NORSP (http://cubic.bioc.columbia.edu/services/NORSp/), DISOPRED (http://bioinf.cs.ucl.ac.uk/disopred/), DISEMBL (http://dis.embl.de/), and GLOBPLOT (http://globplot.embl.de/). The state-of-the art commercial program PONDR is available from Molecular Kinetics (http://www.pondr.com/); at the time of writing the company promised to introduce a free academic license in the near future.

2 Secondary Structure Prediction

2.1 Helices and Strands and Otherwise

Globular protein domains are typically composed of the two basic secondary structure types, the α-helix and the β-strand, which are easily distinguishable

because of their regular (periodic) character. Other types of secondary structures such as different turns, bends, bridges, and non-α helices (such as 3/10 and π) are less frequent and more difficult to observe and classify for a non-expert. The non-α, non-β structures are often referred to as coil or loop and the majority of secondary structure prediction methods are aimed at predicting only these three classes of local structure. Given the observed distribution of the three states in globular proteins (about 30 % α-helix, 20 % β-strand and 50 % coil), random prediction should yield about 40 % accuracy per residue. The accuracy of the secondary structure prediction methods devised earlier, such as Chou-Fasman (1974) or GOR (Garnier et al. 1978) is in the range of 50–55 %. The best modern secondary structure prediction methods (Table 2) have reached a sustained level of 76 % accuracy for the last 2 years, with α-helices predicted with ca. 10 % higher accuracy than β-strands (Koh et al. 2003). Hence, it is quite surprising that the early mediocre methods are still used in good faith by many researchers; maybe even more surprising that they are sometimes recommended in contemporary reviews of bioinformatic software or built in as a default method in new versions of commercial software packages for protein sequence analysis and structure modeling.

Modern secondary structure prediction methods typically perform analyses not for the single target sequences, but rather utilize the evolutionary information derived from MSA provided by the user or generated by an internal routine for database searches and alignment (Levin et al. 1993). The information from the MSA provides a better insight into the positional conservation of physico-chemical features such as hydrophobicity and hints at a position of loops in the regions of insertions and deletions (indels) corresponding to *gaps* in the alignment. It is also recommended to combine different methods for secondary structure prediction; the ways of combing predictions may include the calculation of a simple consensus or more advanced approaches, including machine learning, such as voting, linear discrimination, neural networks and decision trees (King et al. 2000). JPRED (Cuff et al. 1998) is an example of a consensus *meta-server* that returns predictions from several secondary structure prediction methods (mostly third-party algorithms) and infers a consensus using a neural network, thereby improving the average accuracy of prediction. In addition, JPRED predicts the relative solvent accessibility of each residue in the target sequence, which is very useful for identification of solvent-exposed and buried faces of amphipathic helices.

In general, the most effective secondary structure prediction strategies follow these rules: (1) if an experimentally determined three-dimensional structure of a closely related protein is known, copy the secondary structure assignment from the known structure rather than attempt to predict it de novo. (2) If no related structures are known, use multiple sequence information. If your target sequence shows similarity to only a few (or none) other proteins with sequence identity <90 %, try different databases (for example preliminary data from unfinished genomes) to build an MSA comprising a

Table 2. Software for secondary structure prediction

Program	Reference	URL (http://)
Three-state (α/β/coil) prediction		
PSIPRED	Jones (1999b)	bioinf.cs.ucl.ac.uk/psipred/
SSPRO	Pollastri et al. (2002)	www.igb.uci.edu/tools/scratch/
PHD	Rost et al. (1994)	cubic.bioc.columbia.edu/predictprotein/
PROF	Ouali and King (2000)	www.aber.ac.uk/~phiwww/prof/
PRED2ARY	Chandonia and Karplus (1995)	www.cmpharm.ucsf.edu/~jmc/pred2ary/
APSSP2	G.P. Raghava (unpubl.)	www.imtech.res.in/raghava/apssp2/
PREDATOR	Frishman and Argos (1997)	ftp://ftp.ebi.ac.uk/pub/software/unix/predator/
NNSSP	Salamov and Solovyev (1995)	bioweb.pasteur.fr/seqanal/interfaces/nnssp-simple.html
HMMSTR	Bystroff et al. (2000)	www.bioinfo.rpi.edu/~bystrc/hmmstr/
NPREDICT	Kneller et al. (1990)	www.cmpharm.ucsf.edu/~nomi/nnpredict.html
Other types of secondary structure		
TURNS	Kaur and Raghava (2003a, b)	imtech.res.in/raghava/
COILS	Lupas et al. (1991)	www.ch.embnet.org/software/COILS_form.html
"Meta-servers" for secondary structure prediction (gateways to several different methods)		
JPRED	Cuff et al. (1998)	www.compbio.dundee.ac.uk/~www-jpred/
NPS@	Combet et al. (2000)	npsa-pbil.ibcp.fr
META-PP	Eyrich and Rost (2003)	cubic.bioc.columbia.edu/meta/

number of moderately diverged sequences. Discard too strongly diverged sequences, which cannot be aligned with confidence and carefully refine the MSA in the most diverged regions. (3) If the particular algorithm does not accept MSA as an input, try to predict the secondary structure for the target and a few of its distant homologues and use the consensus pattern of secondary structures as an additional indicator of reliability of the prediction. (4) Run as many good methods as possible and use the agreement between their results to infer a consensus prediction. (5) If for a given region only a few methods predicted a β-strand and most coil or an α-helix, the β-strand prediction should be considered as a plausible alternative, as this type of secondary structure is predicted with lower accuracy by virtually all available

methods. (6) Reconfirm the prediction of loops by correlating their presence with regions of indels in the MSA.

In our own hands, the application of these rules in a semi-automated manner (i.e. human post-processing of prediction generated by various individual methods) led to a very high accuracy of 83 % per residue (better than any single server or any other human predictor) according to the recent evaluation within the CASP-5 experiment (http://predictioncenter.llnl.gov/casp5/).

2.2 Transmembrane Helices

Membrane proteins are an abundant and functionally relevant subset of proteins predicted to include up to 30 % of proteins in the fully sequenced genomes. Membrane proteins are associated with the cell membrane and comprise one or more transmembrane segments. Because of the hydrophobic environment within the cell membrane, the transmembrane segments are generally hydrophobic too. On the one hand, typical cytoplasmic membrane proteins comprise hydrophobic α-helical regions separated by hydrophilic loops. On the other hand, bacterial and organellar outer membrane proteins exhibit a characteristic β-barrel structure comprising different even numbers of β-strands. Specialized structure predictors have been designed for both types of membrane proteins. Because both sides of the lipid bilayer are non-equivalent, structure prediction methods for transmembrane proteins often attempt to identify not only the secondary structure elements (α-helices or β-strands), but also the topology of the protein, i.e. the orientation of the elements with respect to both surfaces (which side of transmembrane protein is intra- or extracellular). For instance, the "positive inside rule" (von Heijne 1986, 1992) indicates that the positively charged residues have a preference for the inside of internal membrane proteins.

As with *orthodox* secondary structure prediction methods, the recommended strategy for identification of transmembrane segments and prediction of their distribution and topology in protein sequences is to use many different methods and refer to the consensus as the most robust structural model (Ikeda et al. 2002). Table 3 lists available programs for prediction of transmembrane segments and topology. A meta-server BPROMPT for prediction of transmembrane helices has been recently developed that combines the results of other prediction methods, providing a more accurate consensus prediction (Taylor et al. 2003).

3 Protein Fold-Recognition

The success of the prediction of protein tertiary (three-dimensional) structure from its amino acid sequence is limited by deficiencies in the conforma-

Table 3. Software for prediction of transmembrane regions in proteins

Program	Reference	URL (http://)
α-Transmembrane proteins		
HMMTOP	Tusnady and Simon(2001)	www.enzim.hu/hmmtop/
DAS	Cserzo et al. (1997)	www.sbc.su.se/~miklos/DAS/
PHDhtmn	Rost et al. (1996)	cubic.bioc.columbia.edu/predictprotein/
SOSUI	Hirokawa et al. (1998)v	sosui.proteome.bio.tuat.ac.jp/sosuiframe0.html
TMAP	Milpetz et al. (1995)	www.mbb.ki.se/tmap/
TMHMM	Sonnhammer et al. (1998)	www.cbs.dtu.dk/services/TMHMM-2.0/
TMpred	Hofmann and Stoffel (1993)	www.ch.embnet.org/software/TMPRED_form.html
MEMSAT	Jones et al. (1994)	bioinf.cs.ucl.ac.uk/psipred/
TopPred2	von Heijne (1992)	www.sbc.su.se/~erikw/toppred2/
WHAT	Zhai and Saier (2001)	saier-144–37.ucsd.edu/what.html
UMDHMM	Zhou and Zhou (2003)	phyyz4.med.buffalo.edu/Softwares-Services_files/umdhmm.htm
PRED-TMR2	Pasquier et al. (1999)	biophysics.biol.uoa.gr/PREDTMR2/input.html
ORIENTM	Liakopoulos et al. (2001)	biophysics.biol.uoa.gr/OrienTM/submit.html
BPROMPT	Taylor et al. (2003)	www.jenner.ac.uk/BPROMPT
β-Transmembrane proteins		
BBF	Zhai and Saier (2002)	www-biology.ucsd.edu/~msaier/transport/software/bbfsource.tar.gz
HMM	Martelli et al. (2002)	www.biocomp.unibo.it

tional search procedures aimed at finding the global energy minimum and in the effective potentials used to evaluate the free energies of possible structures. However, despite the number of possible conformations is practically unlimited, the universe of protein folds (i.e. spatial arrangement of secondary structure elements) is not only finite, but the total number of folds is estimated to be relatively small, in the range of a few thousand (Chothia 1992; Gerstein and Levitt 1997; Zhang and DeLisi 1998; Wolf et al. 2000; Koonin et al. 2002). The notion that proteins can share a similar fold (even in the absence of significant sequence similarity) prompted the development of structure prediction methods that limit the search of the vast conformational space to known protein three-dimensional structures.

The protein fold-recognition approach to structure prediction aims to identify the known structural framework (i.e. the backbone of an experimentally determined protein structure) that accommodates the target protein sequence in the best way. Typically, a fold-recognition program comprises four components: (1) the representation of the template structures (usually corresponding to proteins from the Protein Data Bank database), (2) the evaluation of the compatibility between the target sequence and a template fold, (3) the algorithm to compute the optimal alignment between the target sequence and the template structure, and (4) the way the ranking is computed and the statistical significance is estimated (Fischer et al. 1996).

Two main types of fold-recognition algorithms may be defined: those that detect sequence similarity (without utilizing structural information from the template) and those that detect structure similarity (Table 4).

Sequence-based fold recognition methods do not utilize explicitly the structural information from the templates. The simplest sequence-only fold-recognition operation is to use BLAST or PSI-BLAST to search the Protein Data Bank for structurally characterized proteins that exhibit significant sequence similarity to the target protein. However, the principal task of protein fold-recognition methods is to identify sequence similarities that most biologists wouldn't easily call evident and that cannot be identified in trivial database searches. The evolutionary information used to detect remote relationships is usually compiled in the form of a profile, or a HMM. However, the most sensitive sequence-based fold-recognition methods available today are more advanced than sequence-profile comparisons implemented in methods such as PSI-BLAST, IMPALA or HMMs and utilize the evolutionary information available both for the target and the template by performing profile-profile alignment and the evaluation of the likelihood that two protein families are related to each other; examples include FFAS (Rychlewski et al. 2000) and the prof_sim algorithm (Yona and Levitt 2002). A recently developed method ORFeus uses sequence profiles and disregards the experimental structural information from the template, and attempts to predict the structure de novo both for the target and the template families (Ginalski et al. 2003b).

Structure-based fold-recognition, often referred to as *threading*, utilizes the experimentally determined structural information from the template. The target sequence can be enhanced by including sequence-derived (predicted) structural features of the target. The two typically used structural features are the patterns of secondary structure elements and local environment classes (combination of solvent accessibility, polarity of the side chain environment and local backbone conformation). The target-template compatibility functions of the early threading methods were based mainly on physicochemical properties and evaluation of pseudo-energy of interactions and utilized either distance-based (Godzik et al. 1992; Jones et al. 1992; Sippl and Weitckus 1992; Bryant and Lawrence 1993) or profile-based scoring-functions (Bowie et al. 1991; Ouzounis et al. 1993). The compatibility score is computed by

Table 4. Fold-recognition servers

Program	Reference	URL (http://)
Sequence-based fold-recognition		
FFAS	Rychlewski et al. (2000)	ffas.ljcrf.edu
SAM-T99	Karplus et al. (1998)	www.cse.ucsc.edu/research/compbio/HMM-apps/T99-query.html
ESyPred3D	Lambert et al. (2002)	www.fundp.ac.be/urbm/bioinfo/esypred/
ORFEUS	Ginalski et al. (2003b)	grdb.bioinfo.pl/
Structure-based fold recognition ("threading")		
3DPSSM	Kelley et al. (2000)	www.sbg.bio.ic.ac.uk/~3dpssm/
FUGUE	Shi et al. (2001)	www-cryst.bioc.cam.ac.uk/~fugue/
GENThreader	Jones (1999a)	bioinf.cs.ucl.ac.uk/psipred/
INBGU	Fischer (2000)	www.cs.bgu.ac.il/~bioinbgu/form.html
PROTINFO	Samudrala and Levitt (2002)	protinfo.compbio.washington.edu/
RPFOLD	G.P. Raghava (unpubl.)	imtech.res.in/raghava/rpfold/
RAPTOR	Xu et al. (2003)	www.cs.uwaterloo.ca/~j3xu/RAPTOR_form.htm
PROSPECT	Xu and Xu (2000)	compbio.ornl.gov/PROSPECT/
LOOPP	Elber and Meller (unpubl.)	ser-loopp.tc.cornell.edu/cbsu/loopp.htm
SAM-T02	Karplus et al. (2001)	www.soe.ucsc.edu/research/compbio/HMM-apps/T02-query.html
Selected fold-recognition "meta-servers" (gateways to several different methods)		
BIOINFO	Bujnicki et al. (2001 c)	bioinfo.pl/meta/
GENESILICO	Kurowski and Bujnicki (2003)	genesilico.pl/meta/
@TOME	Douguet and Labesse (2001)	bioserv.cbs.cnrs.fr/HTML_BIO/frame_meta.html

adding up the compatibility scores of each residue and subtracting a penalty for any gaps in the target-template alignment. Computing an optimal alignment with a distance-based multipositional compatibility function that takes into account residues adjacent in space but not necessarily in the primary sequence, is an NP-complete problem (Lathrop 1994). In practice it means that the time required to find the best alignment grows exponentially with the length of the protein. Thus, many methods implemented various approximations to encode all structural properties into a one-dimensional string of symbols, thereby allowing target-template matching using conventional dynamic programming algorithms (Needleman and Wunsch 1970; Smith and

Waterman 1981), as in sequence-based methods. The early threaders were quite successful in identification of the correct fold, however the quality of the reported target-template alignments was often poor. Apparently, correct fold-recognition could be achieved, despite poor alignment quality, by a generally unspecific maximization of the hydrophobic interactions, and a reasonably good prediction of the local secondary structure (Lemer et al. 1995).

Modern fold-recognition methods utilize both the structural information (experimentally determined for the potential templates and predicted for the target) and the evolutionary information inferred from the MSA available for the target and the templates. According to the recent evaluations (Bujnicki et al. 2001a, b), best fold-recognition algorithms are able to make up to 40 % of correct structural predictions for targets, which exhibit no significant similarity to any of the potential templates (i.e. similarities that cannot be detected by BLAST or PSI-BLAST searches run with default parameters). One of the most significant unsolved problems is the lack of an accurate scoring function for discrimination between correct and incorrect fold-recognition alignments. It is quite often the case that the correct template is reported among the best ten results returned by a fold-recognition server, but its score is very similar to scores for nine *false positives* or it is below the threshold of statistical significance. In other words, the sensitivity and specificity of fold-recognition methods are insufficient to confidently identify the correct template, if it exists in the Protein Data Bank. Recently, consensus meta-servers have been developed which greatly increase the sensitivity and specificity of fold-recognition (Douguet and Labesse 2001; Bujnicki et al. 2001 c; Lundstrom et al. 2001; Kurowski and Bujnicki 2003; Ginalski et al. 2003a). Most of them combine not only fold-recognition methods, but integrate many different kinds of protein structure prediction methods described in this article, from identification of domains, to secondary structure prediction, to modeling of the target based on the best-scoring template structures (for detailed description of two examples see the following section and a separate review by Cohen et al. (this Vol.); a separate discussion on various aspects of *meta* prediction is provided in a review by Bujnicki and Fischer).

4 Predicting All-in-One-Go

The GeneSilico meta-server (http://genesilico.pl/meta/; Kurowski and Bujnicki 2003) will serve here as an example of a freely available on-line service for integrated prediction of different aspects of protein structure. As mentioned earlier, the recommended strategy is to predict the target protein structure using not only the single sequence information, but to enhance it with aligned homologous sequences. The GeneSilico meta-server allows submission of single sequences or user-defined multiple alignments (MSA). A single sequence is processed further by individual methods, which often gen-

erate their own alignments, typically using PSI-BLAST (Altschul et al. 1997) with different parameters. Automatically generated sequence alignments are usually sufficient, but sometimes the target sequence has an unusual amino acid composition or atypical insertions, which may cause the default iterated database search to produce erroneous alignments that will degrade the evolutionary signal instead of enhancing it. Moreover, some sequences have only a few homologues in the traditionally used databases such as NRDB or Swiss-Prot and in order to build a useful alignment, additional searches of other databases are necessary. Therefore, it is strongly recommended for experienced predictors to submit their own MSA, in addition to the single-sequence queries. The GeneSilico meta-server will forward the MSA to those servers that allow such input, while for the others, which accept only single-sequence queries, a single consensus sequence will be calculated from the MSA using one of many different options selected by the user (from majority-rule to scoring derived from different substitution matrices). Furthermore, the user will have an option to delete or retain loopy regions corresponding to gaps in the sequence alignment – this option causes a limitation on the fold-recognition analysis to regions most likely to correspond to the true globular core of the target protein.

As mentioned earlier, the crucial step in protein structure prediction is to identify protein domains in the target sequence. This task is accomplished by the HMMPFAM tool, which scans the PFAM database of known protein domains (Bateman et al. 2002) with the HMMER method (Eddy 1996). If the results obtained from the HMMPFAM search suggest the presence of more than one domain in the target sequence, it is strongly recommended to split the target into the respective fragments (possibly retaining some regions of overlap, 10–50 aa, depending on the confidence of the domain prediction) and resubmit the individual domains as separate prediction queries.

Secondary structure is predicted in three states (α, β, and coil) by PSI-PRED (Jones 1999b), PROF (Ouali and King 2000), and SAM-T02 (Karplus et al. 2001). Identification of potential transmembrane helices is attempted using TMPRED (Hofmann and Stoffel 1993) MEMSAT (Jones et al. 1994), and TMHMM (Sonnhammer et al. 1998). If all methods predict a transmembrane segment or a long region with no α or β structure in the target sequence, it is again strongly recommended to remove such regions, as they are unlikely to form any globular domain identifiable by fold-recognition methods, and to resubmit the remaining part of the target as a new prediction query.

The GeneSilico metaserver serves as a gateway for a number of third-party fold-recognition methods, both sequence-dependent, and structure-dependent, including FUGUE (Shi et al. 2001), 3DPSSM (Kelley et al. 2000), SAM-T02 (Karplus et al. 2001), GENTHREADER (Jones 1999a), FFAS (Rychlewski et al. 2000), INBGU (Fischer 2000), and RAPTOR (Xu et al. 2003). However, before the extensive fold-recognition calculations are carried out, the PDB database is searched with the PSI-BLAST method to identify trivial similarities of the

target to proteins of known structure (three iterations against the NRDB database are carried out with the target sequence to generate a MSA, which is subsequently used to search the PDB database for significant similarities). If the target exhibits significant similarity to a known structure, the fold-recognition analysis is halted and the user is notified; otherwise (or if the user decides to resume the analysis) the query (i.e. the single sequence or the MSA) is sent to the above-mentioned fold-recognition servers. Typically, the collection of results from all servers (up to ten target-template alignments per server) requires about 24 h, however some sequence-based servers return their predictions within a few minutes. The meta-server presents all target-template alignments and the corresponding confidence scores assigned by the individual methods according to their internal criteria. These scores are mutually incompatible and further analysis is required to provide a common ranking of results returned by different fold-recognition servers. Hence, when all results are available, they are further processed by the consensus server PCONS (two different versions, 2 and 5; Lundstrom et al. 2001; Wallner and Elofsson 2003), which does not produce any new predictions, but selects the ten potentially best target-template alignments from those reported by the original methods and assigns its own confidentiality scores. It has been shown that PCONS is more sensitive (i.e. able to identify correct templates) and specific (i.e. able to generate significant scores) than any individual method incorporated as a *slave* in the prediction pipeline.

Finally, the user of the GeneSilico server has an opportunity to generate preliminary three-dimensional models of the target structure based on the alignments proposed by all servers. These models may be incomplete and contain significant errors even if they are based on correct templates, but usually serve as a useful starting point for further refinement. The preliminary evaluation is carried out using the VERIFY3D method, whose score tells how much the characteristics of the model resemble the features of high-resolution crystal structures i.e. how much the theoretical model is protein-like or protein-unlike, compared to the known structures.

5 Pitfalls of Fold Recognition

As soon as the sequence of the target protein is optimally mounted on the presumably best template structure, the corresponding sequence-structure alignment can be used to initiate reconstruction of a complete full-atom model of the target protein by various comparative modeling techniques (reviewed by Cohen et al. in this volume; see also the following references: (Sanchez and Sali 2000; Krieger et al. 2003)). The comparative modeling approach assumes that the target and the template share the polypeptide backbone and the differences are limited to the solvent-exposed loops and the conformation of the side chains, according to the notion that protein spatial

structures are more conserved in evolution than amino acid sequences (Chothia and Lesk 1986). This assumption is certainly valid in many cases, especially if the sequence identity between the target and the template is very high (>50 %). However, the recent sequence and structure analyses led to the accumulation of examples of homologous proteins with globally distinct structures. It has been found that even in proteins with significant sequence similarity, insertions, deletions and mutual conversions of α-helices and β-strands can occur both at the periphery and in the core of the fold; moreover, the global topology of the fold can be changed by circular permutations, and rearrangements in the order of strands in β-sheets (reviews: Murzin 1998; Grishin 2001a). Such structural changes are usually undetectable by computational methods that operate on the level of protein sequence similarities and even for structure-based threading methods it is extremely hard to predict differences between the three-dimensional folds of the target and the template other than the deletion or insertion of secondary structure elements.

It also becomes clear that domains are not the only units of homology. Some protein superfamilies have been reported to contain segments of homology often limited to a few elements of secondary structure unable to fold independently, such as the ββα-Me finger in many nucleases, embedded into non-homologous regions acquired independently between proteins (Kuhlmann et al. 1999; Grishin 2001b). In contrast, unrelated segments acquired independently could be embedded into the regions of homology. In such cases, detection of a strong local homology by fold-recognition programs can be erroneously extended to the entire length of the target and the template. Currently, no fully automated methods exist for prediction of fold irregularities. However, recent progress in the ab initio protein structure prediction field, especially the development of methods that use confident predictions of the protein core made by fold-recognition methods to initiate extensive folding simulation to assemble the peripheral elements (Simons et al. 1997; Kihara et al. 2001) suggest that in the near future these limitations of the current fold-recognition methods may be overcome.

Presently, the best strategy, however, is to validate the computational prediction of the protein fold by experimental analyses which on their own would not be sufficient to *solve* protein structure, but when combined with bioinformatics, may serve to identify one reasonable structural model and then guide its refinement. Such experimental investigations may include generation of both specific and non-specific distance restraints by intramolecular cross-linking, chemical modification, or simple NMR analyses, identification of solvent-exposed loops by proteolysis, identification of important residues by mutagenesis etc. Several examples of combination of computational and experimental analyses are discussed elsewhere in this volume (see chapters by Linge and Nilges; Alber et al; and Friedhoff). Clearly, the development of a convenient computational method for automated combination of heterologous experimental data and low-resolution structure prediction by

fold-recognition and ab initio bioinformatic methods would greatly facilitate structural analyses of proteins and bring protein modeling closer to the workbench of a biochemist or a molecular biologist.

Acknowledgements. The authors' research on various aspects of combination of computational and experimental methods for protein structure analysis is supported by KBN (grants 6P04B00519, 3P04A01124, and 3P05A02024). J.M.B. is an EMBO and Howard Hughes Medical Institute Young Investigator and a Fellow of the Foundation for Polish Science.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Aravind L, Koonin EV (1999) Gleaning non-trivial structural, functional and evolutionary information about proteins by iterative database searches. J Mol Biol 287:1023–1040

Aravind L, Mazumder R, Vasudevan S, Koonin EV (2002) Trends in protein evolution inferred from sequence and structure analysis. Curr Opin Struct Biol 12:392–399

Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL (2002) The Pfam protein families database. Nucleic Acids Res 30:276–280

Bowie JU, Luthy R, Eisenberg D (1991) A method to identify protein sequences that fold into a known three-dimensional structure. Science 253:164–170

Bryant SH, Lawrence CE (1993) An empirical energy function for threading protein sequence through the folding motif. Proteins 16:92–112

Bujnicki JM, Elofsson A, Fischer D, Rychlewski L (2001a) LiveBench-1: continuous benchmarking of protein structure prediction servers. Protein Sci 10:352–361

Bujnicki JM, Elofsson A, Fischer D, Rychlewski L (2001b) LiveBench-2: Large-scale automated evaluation of protein structure prediction servers. Proteins 45:184–191

Bujnicki JM, Elofsson A, Fischer D, Rychlewski L (2001 c) Structure prediction Meta Server. Bioinformatics 17:750–751

Bystroff C, Thorsson V, Baker D (2000) HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins J Mol Biol 301:173–190

Chandonia JM, Karplus M (1995) Neural networks for secondary structure and structural class predictions. Protein Sci 4:275–285

Chothia C (1992) Proteins. One thousand families for the molecular biologist. Nature 357:543–544

Chothia C, Lesk AM (1986) The relation between the divergence of sequence and structure in proteins. EMBO J 5:823–826

Chou PY, Fasman GD (1974) Prediction of protein conformation. Biochemistry 13:222–245

Combet C, Blanchet C, Geourjon C, Deleage G (2000) NPS@: network protein sequence analysis. Trends Biochem Sci 25:147–150

Cserzo M, Wallin ESimon Ivon Heijne G, Elofsson A (1997) Prediction of transmembrane alpha-helices in prokaryotic membrane proteins: the dense alignment surface method. Protein Eng 10:673–676

Cuff JA, Clamp ME, Siddiqui AS, Finlay M, Barton GJ (1998) JPred: a consensus secondary structure prediction server. Bioinformatics 14:892–893

Douguet D, Labesse G (2001) Easier threading through web-based comparisons and cross-validations. Bioinformatics 17:752–753

Eddy SR (1996) Hidden Markov models. Curr Opin Struct Biol 6:361–365

Eyrich VA, Rost B (2003) META-PP: single interface to crucial prediction servers. Nucleic Acids Res 31:3308–3310

Fischer D (2000) Hybrid fold recognition: combining sequence derived properties with evolutionary information. Pac Symp Biocomput , pp 119–130

Fischer D, Elofsson A, Rice D, Eisenberg D (1996) Assessing the performance of fold recognition methods by means of a comprehensive benchmark. Pac Symp Biocomput , pp 300–318

Frishman D, Argos P (1997) Seventy-five percent accuracy in protein secondary structure prediction. Proteins 27:329–335

Garnier J, Osguthorpe DJ, Robson B (1978) Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. J Mol Biol 120:97–120

Gerstein M, Levitt M (1997) A structural census of the current population of protein sequences. Proc Natl Acad Sci USA 94:11911–11916

Ginalski K, Elofsson A, Fischer D, Rychlewski L (2003a) 3D-Jury: a simple approach to improve protein structure predictions. Bioinformatics 19:1015–1018

Ginalski K, Pas J, Wyrwicz LS, von Grotthuss M, Bujnicki JM, Rychlewski L (2003b) ORFeus: detection of distant homology using sequence profiles and predicted secondary structure. Nucleic Acids Res 31:3804–3807Godzik A, Kolinski A, Skolnick J (1992) Topology fingerprint approach to the inverse protein folding problem. J Mol Biol 227:227–238

Grishin NV (2001a) Fold change in evolution of protein structures. J Struct Biol 134:167–185

Grishin NV (2001b) Treble clef finger–a functionally diverse zinc-binding structural motif. Nucleic Acids Res 29:1703–1714

Haft DH, Selengut JD, White O (2003) The TIGRFAMs database of protein families. Nucleic Acids Res 31: 371–373

Henikoff JG, Greene EA, Pietrokovski S, Henikoff S (2000) Increased coverage of protein families with the blocks database servers. Nucleic Acids Res 28:228–230

Hirokawa T, Boon-Chieng S, Mitaku S(1998) SOSUI: classification and secondary structure prediction system for membrane proteins. Bioinformatics 14:378–379

Hofmann K, Stoffel W (1993) TMbase – a database of membrane spanning proteins segments. Biol Chem 374:166

Ikeda M, Arai M, Lao DM, Shimizu T (2002) Transmembrane topology prediction methods: a re-assessment and improvement by a consensus method using a dataset of experimentally-characterized transmembrane topologies. In Silico Biol 2:19–33

Jones DT (1999a) GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. J Mol Biol 287:797–815

Jones DT (1999b) Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 292:195–202

Jones DT, Taylor WR, Thornton JM (1992) A new approach to protein fold recognition. Nature 358:86–89

Jones DT, Taylor WR, Thornton JM (1994) A model recognition approach to the prediction of all-helical membrane protein structure and topology. Biochemistry 33:3038–3049

Karplus K, Barrett C, Hughey R (1998) Hidden Markov models for detecting remote protein homologies. Bioinformatics 14:846–856

Karplus K, Karchin R, Barrett C, Tu S, Cline M, Diekhans M, Grate L, Casper J, Hughey R (2001) What is the value added by human intervention in protein structure prediction? Proteins 45(Suppl 5):86–91

Kaur H, Raghava GP (2003a) A neural-network based method for prediction of gamma-turns in proteins from multiple sequence alignment. Protein Sci 12:923–929

Kaur H, Raghava GP (2003b) Prediction of beta-turns in proteins from multiple alignment using neural network. Protein Sci 12:627–634

Kelley LA, McCallum CM, Sternberg MJ (2000) Enhanced genome annotation using structural profiles in the program 3D-PSSM. J Mol Biol 299:501–522

Kihara D, Lu H, Kolinski A, Skolnick J (2001) TOUCHSTONE: an ab initio protein structure prediction method that uses threading-based tertiary restraints. Proc Natl Acad Sci USA 98:10125–10130

King RD, Ouali M, Strong AT, Aly A, Elmaghraby A, Kantardzic M, Page D (2000) Is it better to combine predictions? Protein Eng 13:15–19

Kneller DG, Cohen FE, Langridge R (1990) Improvements in protein secondary structure prediction by an enhanced neural network. J Mol Biol 214:171–182

Koh IY, Eyrich VA, Marti-Renom MA, Przybylski D, Madhusudhan MS, Eswar N, Grana O, Pazos F, Valencia A, Sali A, Rost B (2003) EVA: evaluation of protein structure prediction servers. Nucleic Acids Res 31:3311–3315

Koonin EV, Wolf YI, Karev GP (2002) The structure of the protein universe and genome evolution. Nature 420:218–223

Krieger E, Nabuurs SB, Vriend G (2003) Homology modeling. Methods Biochem Anal 44:509–523

Kuhlmann UC, Moore GR, James R, Kleanthous C, Hemmings AM (1999) Structural parsimony in endonuclease active sites: should the number of homing endonuclease families be redefined? FEBS Lett 463:1–2

Kurowski MA, Bujnicki JM (2003) GeneSilico protein structure prediction meta-server. Nucleic Acids Res 31:3305–3307

Lambert C, Leonard N, De B, X, Depiereux E (2002) ESyPred3D: Prediction of proteins 3D structures. Bioinformatics 18:1250–1256

Lathrop RH (1994) The protein threading problem with sequence amino acid interaction preferences is NP-complete. Protein Eng 7:1059–1068

Lemer CM, Rooman MJ, Wodak SJ (1995). Protein structure prediction by threading methods: evaluation of current techniques. Proteins 23:337–355

Letunic I, Goodstadt L, Dickens NJ, Doerks T, Schultz J, Mott R, Ciccarelli F, Copley RR, Ponting CP, Bork P (2002) Recent improvements to the SMART domain-based sequence annotation resource. Nucleic Acids Res 30:242–244

Levin JM, Pascarella S, Argos P, Garnier J (1993) Quantification of secondary structure prediction improvement using multiple alignments. Protein Eng 6:849–854

Li W, Pio F, Pawlowski K, Godzik A (2000) Saturated BLAST: an automated multiple intermediate sequence search used to detect distant homology. Bioinformatics 16:1105–1110

Liakopoulos TD, Pasquier C, Hamodrakas SJ (2001) A novel tool for the prediction of transmembrane protein topology based on a statistical analysis of the SwissProt database: the OrienTM algorithm. Protein Eng 14:387–390

Liu J, Tan H, Rost B (2002) Loopy proteins appear conserved in evolution. J Mol Biol 322:53–64

Lundstrom J, Rychlewski L, Bujnicki JM, Elofsson A (2001) Pcons: a neural-network-based consensus predictor that improves fold recognition. Protein Sci 10:2354–2362

Lupas A, Van Dyke M, Stock J (1991) Predicting coiled coils from protein sequences. Science 252:1162–1164

Marchler-Bauer A, Anderson JB, DeWeese-Scott C, Fedorova ND, Geer LY, He S, Hurwitz DI., Jackson JD, Jacobs AR, Lanczycki CJ, Liebert CA, Liu C, Madej T, Marchler GH, Mazumder R, Nikolskaya AN, Panchenko AR, Rao BS, Shoemaker BA, Simonyan V, Song JS, Thiessen PA, Vasudevan S, Wang Y, Yamashita RA, Yin JJ, Bryant SH (2003) CDD: a curated Entrez database of conserved domain alignments. Nucleic Acids Res 31:383–387

Martelli PL, Fariselli P, Krogh A, Casadio R (2002) A sequence-profile-based HMM for predicting and discriminating beta barrel membrane proteins. Bioinformatics 18(Suppl 1):S46-S53

Milpetz F, Argos P, Persson B (1995) TMAP: a new email and WWW service for membrane-protein structural predictions. Trends Biochem Sci 20:204–205

Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Barrell D, Bateman A, Binns D, Biswas M, Bradley P, Bork P, Bucher P, Copley RR, Courcelle E, Das U, Durbin R, Falquet L, Fleischmann W, Griffiths-Jones S, Haft D, Harte N, Hulo N, Kahn D, Kanapin A, Krestyaninova M, Lopez R, Letunic I, Lonsdale D, Silventoinen V, Orchard SE, Pagni M, Peyruc D, Ponting CP, Selengut JD, Servant F, Sigrist CJ, Vaughan R, Zdobnov EM (2003) The InterPro Database, 2003 brings increased coverage and new features. Nucleic Acids Res 31:315–318

Murzin AG (1998) How far divergent evolution goes in proteins. Curr Opin Struct Biol 8 380–387

Nagano K (1973) Logical analysis of the mechanism of protein folding. I. Predictions of helices, loops and beta-structures from primary structure. J Mol Biol 75:401–420

Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48:443–453

Ouali M, King RD (2000) Cascaded multiple classifiers for secondary structure prediction. Protein Sci 9:1162–1176

Ouzounis C, Sander C, Scharf M, Schneider R (1993) Prediction of protein structure by evaluation of sequence-structure fitness. Aligning sequences to contact profiles derived from three- dimensional structures. J Mol Biol 232:805–825

Pagni M, Jongeneel CV (2001) Making sense of score statistics for sequence alignments. Brief Bioinform 2:51–67

Park J, Karplus K, Barrett C, Hughey R, Haussler D, Hubbard T, Chothia C (1998) Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. J Mol Biol 284:1201–1210

Park J, Teichmann SA, Hubbard T, Chothia C (1997). Intermediate sequences increase the detection of homology between sequences. J Mol Biol 273:349–354

Pasquier C, Promponas VJ, Palaios GA, Hamodrakas JS, Hamodrakas SJ (1999) A novel method for predicting transmembrane segments in proteins based on a statistical analysis of the SwissProt database: the PRED-TMR algorithm. Protein Eng 12:381–385

Pearson WR (1998) Empirical statistical estimates for sequence similarity searches. J Mol. Biol 276:71–84

Pearson WR, Lipman DJ (1988) Improved tools for biological sequence comparison. Proc Natl Acad Sci U. S. A. 85:2444–2448

Pizzi E, Frontali C.(2001) Low-complexity regions in *Plasmodium falciparum* proteins. Genome Res 11:218–229

Pollastri G, Przybylski D, Rost B, Baldi P (2002) Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. Proteins 47:228–235

Rost B, Fariselli P, and Casadio R (1996) Topology prediction for helical transmembrane proteins at 86 % accuracy. Protein Sci 5:1704–1718

Rost B, Sander C, Schneider R (1994) PHD–an automatic mail server for protein secondary structure prediction. Comput Appl Biosci 10:53–60

Rychlewski L, Jaroszewski L, Li W, Godzik A (2000) Comparison of sequence profiles. Strategies for structural predictions using sequence information. Protein Sci 9:232–241

Salamov AA, Solovyev VV (1995) Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignments. J Mol Biol 247: 11–15

Samudrala R, Levitt M (2002) A comprehensive analysis of 40 blind protein structure predictions. BMC Struct Biol 2:3

Sanchez R, Sali A (2000) Comparative protein structure modeling. Introduction and practical examples with modeller. Methods Mol Biol 143:97–129

Schaffer AA, Aravind L, Madden TL, Shavirin S, Spouge JL, Wolf YI, Koonin EV, Altschul SF (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res 29:2994–3005

Servant F, Bru C, Carrere S, Courcelle E, Gouzy J, Peyruc D, Kahn D (2002) ProDom: automated clustering of homologous domains. Brief Bioinform 3(3):246–251

Shi J, Blundell TL, Mizuguchi K (2001) Fugue: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. J Mol Biol 310:243–257

Sigrist CJ, Cerutti L, Hulo N, Gattiker A, Falquet L, Pagni M, Bairoch A, Bucher P (2002) PROSITE: a documented database using patterns and profiles as motif descriptors. Brief Bioinform 3:265–274

Simons KT, Kooperberg C, Huang E, Baker D (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol 268:209–225

Sippl MJ, Weitckus S (1992) Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a data base of known protein conformations. Proteins 13:258–271

Smith TF, Waterman MS (1981) Identification of common molecular subsequences. J Mol Biol 147:195–197

Sonnhammer EL, von Heijne G, Krogh A (1998) A hidden Markov model for predicting transmembrane helices in protein sequences. Proc Int Conf Intell Syst Mol Biol 6:175–182

Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res 29:22–28

Taylor PD, Attwood TK, Flower DR (2003) BPROMPT: a consensus server for membrane protein prediction. Nucleic Acids Res 31:3698–3700

Thornton JM, Orengo CA, Todd AE, Pearl FM (1999) Protein folds, functions and evolution. J Mol Biol 293:333–342

Tompa P (2002) Intrinsically unstructured proteins. Trends Biochem Sci 27:527–533

Tusnady GE, Simon I (2001) The HMMTOP transmembrane topology prediction server. Bioinformatics 17:849–850

Vlahovicek K, Kajan L, Murvai J, Hegedus Z, Pongor S (2003) The SBASE domain sequence library, release 10: domain architecture prediction. Nucleic Acids Res 31:403–405

von Heijne G (1986) The distribution of positively charged residues in bacterial inner membrane proteins correlates with the trans-membrane topology. EMBO J 5:3021–3027

von Heijne G (1992) Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. J Mol. Biol 225:487–494

Wallner B, Elofsson A (2003) Can correct protein models be identified? Protein Sci 12:1073–1086

Webber C, Barton GJ (2003) Increased coverage obtained by combination of methods for protein sequence database searching. Bioinformatics 19:1397–1403

Wolf YI, Grishin NV, Koonin EV (2000) Estimating the number of protein folds and families from complete genome data. J Mol Biol 299:897–905

Wootton JC (1994) Sequences with "unusual" amino acid composition. Curr Opin Struct Biol 4:413–421

Wootton JC, Federhen S (1996) Analysis of compositionally biased regions in sequence databases. Methods Enzymol 266:554–571

Wright PE, Dyson HJ (1999) Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. J Mol Biol 293:321–331

Xu J, Li M, Lin G, Kim D, Xu Y (2003) Protein structure prediction by linear programming. Pac Symp Biocomput 264:75

Xu Y, Xu D (2000) Protein threading using PROSPECT: design and evaluation. Proteins 40 (3):343–354

Yona G, Levitt M (2002) Within the twilight zone: a sensitive profile-profile comparison tool based on information theory. J Mol Biol 315:1257–1275

Zhai Y, Saier MH Jr (2001) A web-based program (WHAT) for the simultaneous prediction of hydropathy, amphipathicity, secondary structure and transmembrane topology for a single protein sequence. J Mol Microbiol Biotechnol 3:501–502

Zhai Y, Saier MH Jr (2002) The beta-barrel finder (BBF) program, allowing identification of outer membrane beta-barrel proteins encoded within prokaryotic genomes. Protein Sci 11:2196–2207

Zhang C, DeLisi C (1998) Estimating the number of protein folds. J Mol. Biol 284:1301–1305

Zhou H, Zhou Y (2003) Predicting the topology of transmembrane helical proteins using mean burial propensity and a hidden-Markov-model-based method. Protein Sci 12:1547–1555

'Meta' Approaches to Protein Structure Prediction

J.M. Bujnicki, D. Fischer

1 Introduction

The computational assignment of three-dimensional structures to newly determined protein sequences is becoming an increasingly important element in experimental structure determination and in structural genomics (Fischer et al. 2001a). In particular, fold-recognition methods aim to predict approximate three-dimensional (3D) models for proteins bearing no evident sequence similarity to any protein of known structure (see the review by Cymerman et al., this Vol.). The assignment is carried out by searching a library of known structures (usually obtained from the Protein Data Bank) for a compatible fold. A variety of fold-recognition methods has been published, both *structure-dependent* (i.e.*threading*) (Sippl and Weitckus 1992; Godzik et al. 1992; Jones et al. 1992; Ouzounis et al. 1993; Bryant and Lawrence 1993; Rost 1995; Alexandrov et al. 1996; Di Francesco et al. 1997; Fischer 2000; Kelley et al. 2000; Shi et al. 2001) and *sequence-only dependent* (Karplus et al. 1998; Rychlewski et al. 2000). The state-of-the-art in the field of fold recognition is currently to combine the evolutionary information available from multiple sequence alignments for the target and the template (to detect remote homology between protein families) and the structural information from the template (to detect similarities of folds of compared proteins regardless of their evolutionary relationship, i.e. analogs and homologues as well).

J.M. Bujnicki
Bioinformatics Laboratory, International Institute of Molecular and Cell Biology in Warsaw, Trojdena 4, 02-109 Warsaw, Poland
D. Fischer
Current address: Bioinformatics, Dept. Computer Science, Ben Gurion University, Beer-Sheva 84015, Israel

Nucleic Acids and Molecular Biology, Vol. 15
Janusz M. Bujnicki (Ed.)
Practical Bioinformatics

2 The Utility of Servers as Standard Tools for Protein Structure Prediction

Automatic structure prediction has witnessed significant progress during the last few years. A large number of fully automated servers, covering various aspects of structure prediction, are currently available to the scientific community. In addition to the biannual Critical Assessment of Structure Prediction (CASP) experiment, which evaluates the state-of-the-art in the methodology and the skills of modeling teams and individual modelers (Moult et al. 1995, 1997, 1999, 2001), a number of evaluation experiments exist that are aimed at assessing the capabilities and limitations of the servers. These experiments assess the reliability of the programs when applied to specific prediction targets and provide predictors with valuable information that can help them in choosing which programs to use and thereby make best use of the automated tools. One of these experiments is CAFASP (Fischer et al. 1999, 2001b), where the evaluation is carried out over the set of the CASP prediction targets by fully automatic web servers that submit the predictions without any human-expert intervention. CAFASP servers cover various aspects of protein structure prediction, such as secondary structure, inter-residue contacts, and tertiary structure. Another experiment is LiveBench, which differs from CAFASP in that it is run continuously and on a much larger set of targets. The targets are selected from protein structures newly submitted to the Protein Data Bank, if their sequences show no trivial similarity to any of the previously available structures (Bujnicki et al. 2001a, b).

However, despite significant progress, protein structure prediction methods still have a number of limitations. Fully automated fold-recognition methods can currently produce reliable sequence-structure assignments for only a fraction of target sequences with no significant sequence similarity to proteins of known structure (Bujnicki et al. 2001b). In the case of remote structural similarities, the sequence alignments between the target and the template reported by fold recognition often contain large errors (shifts). Needless to say, fold-recognition methods perform poorly when the target protein exhibits only partial structural similarity (i.e. not the same, but a related fold) to proteins in the database or when the sought fold is completely novel and cannot be recognized among the known structures. Another limitation of fold-recognition methods is the uncertainty as to the identity of the best model among the *top candidates*. Quite often, the correct fold is reported within the best ten predictions, but with a non-significant confidence score, buried among *false positives*.

2.1 Consensus 'Meta-Predictors': Is the Whole Greater Than the Sum of the Parts?

The use of a number of models and methods to produce better predictions has already proven useful in a number of areas, including artificial intelligence and computer vision (Marr 1982). Not surprisingly, this approach works well also in protein structure prediction. It has been observed in protein secondary structure prediction (consensus of various methods (Cuff et al. 1998; Selbig et al. 1999; Cuff and Barton 2000)), in homology modeling (multiple-parent structures; (Marti-Renom et al. 2000)) and in ab initio protein folding methods (clustering models and deriving recurring constraints from various models (Simons et al. 1999; Kihara et al. 2001; Kolinski et al. 2001).

The most vigorous development of *meta* approaches has been recently in the field of protein fold recognition. From the series of CASP experiments, it has become clear that often a correct protein fold prediction can be obtained by one server but not by the others. It has also been observed that no server can reliably distinguish between weak hits (predictions with below-threshold scores) and wrong hits, and that often a correct model is found among the top hits of the server, but scoring below a number of incorrect models. From such, and other, observations, many human expert predictors have realized that in order to produce better predictions, the results from a number of independent methods need to be analyzed.

CASP has shown that the combined use of human expertise and automated methods can often result in successful predictions. This, however, requires extensive human intervention, because a human predictor has to improve the model manually, has to determine whether the rank-1 model obtained is correct, whether there is a lower ranking model that corresponds to a correct prediction, or whether the results of the method indicate that no prediction at all can be obtained. To this end, human expert predictors have developed a number of semi-automated strategies. One such strategy has been the application of a number of independent methods to extract a prediction from the top ranking predictions. This has proven useful because for some prediction targets, one method may succeed in producing a correct prediction while others fail, yet for other targets, this same method may fail while the others succeed. Because it is impossible to determine a priori for which targets a given method will succeed, human expert predictors attempt to extract any useful information from results obtained with different methods.

To study whether it was possible to obtain a better prediction using a very simple consensus method that utilized the information from several servers, in CASP4, a group of four human predictors, Leszek Rychlewski, Arne Elofsson and both authors of this chapter, pioneered the consensus idea by submitting to CASP manually selected *consensus* predictions under the groupname CAFASP-CONSENSUS. The consensus predictions were obtained by

analyzing the predictions of the fold-recognition servers that participated in the parallel CAFASP2 experiment. This group performed better than any of the CAFASP servers and ranked seventh among all other human predictors of CASP (Fischer et al. 2001b). This finding illustrated the utility of the servers' results when taken as a whole. Since then, meta-prediction has become the most successful approach, and has been applied by a large number of human predictors, including some of the best CASP5 performers.

For example, in the comparative modeling section of CASP5, three groups excelled (Tramontano 2003), including the GeneSilico group (Janusz Bujnicki and colleagues). This group applied a new semi-automated multi-step meta-protocol named *Frankenstein's Monster*, which uses the results of diverse fold-recognition methods to generate initial target-template alignments (Kosinski et al. 2003; Kurowski and Bujnicki 2003). Full-atom models were built by a series of steps aimed at assembling hybrid models using the most conserved and most reliable fragments from the various models. Because this procedure required extensive human intervention (over 24 h/model), it is clear that *human-meta-predicting* is a difficult task requiring extensive expertise, and that automated procedures are sorely needed.

2.2 Automated Meta-Predictors

Following the proven success of manual meta-predictors, several groups have already implemented fully-automated versions of the meta-approach (Table 1). Automated meta-predictors can be divided into two types: (1) selectors, which simply select models from the input and (2) added-value meta-predictors, which use the input models to generate new models.

One of the earliest meta-predictors was developed by Arne Elofsson by implementing the CAFASP-CONSENSUS ideas from CASP4 into the automated program Pcons (Lundstrom et al. 2001). Pcons receives, as input the top models produced by different fold-recognition servers and selects the models that are evaluated to be more likely to be correct, based on the structural similarities among the input models. That is, it does not produce any new models, only re-ranks the existing ones, based on their mutual similarity and the original scores assigned by the individual servers. Pcons corroborated the strength of the consensus idea in the subsequent LiveBench experiments (Bujnicki et al. 2001b). It was demonstrated that PCONS2 (version trained specifically for a few *original*, i.e. non-meta servers) combined the sensitivity of the most sensitive original method (3D-PSSM; Kelley et al. 2000) with a very high specificity (higher than any individual server). The most important feature contributing to the improved performance of an early version of PCONS was its scoring system, which allowed to confidently identify the correct models, although it was not always able to identify the absolutely best model among similar *top solutions*. The newest version of PCONS, reinforced

Table 1. Meta-servers for protein structure prediction. Most fold-recognition (FR) servers utilize PSI-BLAST and SS methods and therefore are considered "Meta^0". Here, only those that explicitly utilize more than one FR method are included

Method	Input	Reference (http:// URL)	Comments
Meta^1			
BIOINBGU selector	GONP, GONPM, SEQPPRF, SEQPMPRF, PRFSEQ	www.cs.bgu.ac.il/~bioinbgu/	Consensus prediction obtained from the results of five different methods run locally
LIBELULLA selector	SAM-T99, 3DPSSM	www.pdg.cnb.uam.es/servers/libellula.html	
Meta^2, utilize Meta^1			
PCONS/PROQ selector		Available only via 3D-JURY and GeneSilico servers	Predicts the quality of models generated by various FR methods by all-against-all comparison and evaluation with the ProQ method
@TOME selector	PDB-BLAST, 3D-PSSM GENTHR., FUGUE, SAM-T99	bioserv.cbs.cnrs.fr	Predicts the quality of models generated by FR methods using the TITO method. Top models ranked by Verify3D and PROSAII
SHGU/SHGUM added-value	BIOINBGU	www.cs.bgu.ac.il/~bioinbgu/	Utilizes the 3D-SHOTGUN approach to create hybrid models. SHGU produces raw C-α models, SHGUM produces full-atom models using MODELLER
3DS3/3DS5* added-value	BIOINBGU, FFAS, 3DPSSM, FUGUE*, GENTHR.*	www.cs.bgu.ac.il/~bioinbgu/	Utilize the 3D-SHOTGUN approach to create hybrid models (C-α). Assemble hybrid models from fragments of models generated by original and consensus methods

Table 1. (*Continued*)

Method	Input	Reference (http:// URL)	Comments
Meta[3], utilize Meta[2] ROBETTA added-value	PDB-BLAST PCONS, ROSETTA	Not available publicly	Takes the Pcons results to initiate the folding simulation using the Rosetta method
GENESILICO selector	A variety of primary FR methods + PCONS	genesilico.pl/meta/	Common input and output. Predicts SS with several methods, runs several FR methods and re-ranks the results using Pcons. All models additionally ranked by Verify3D
Meta[4], utilize Meta[3] 3D-JURY selector	A variety of primary FR methods + SHOTGUN, PCONS	bioinfo.pl/meta/	Common input and output. Predicts SS with several methods, runs a variety of FR methods and consensus methods. All results (including the original models and consensus models) are re-ranked by the 3D-JURY system
Meta[5], utilize Meta[4] PRCM added-value	3D-JURY	protinfo.compbio.washington.edu	Builds, minimizes and ranks the full-atom models starting from the crude FR models selected by 3D-JURY
Meta[6], utilize Meta[5] ALEPH0-JURY selector	ROBETTA, PRCM, SHGUM	Fischer (unpubl.)	Selects the best full-atom model

by the PROQ method for protein model evaluation (Wallner and Elofsson 2003), exhibits even higher specificity; moreover, it is able to use the set of any external (original or meta) methods as model generators.

LIBELULLA (Juan et al. 2003) is a system of neural networks trained to select correct folds from among the results of two *primary* fold-recognition methods implemented as web servers, SAM-T99 (Karplus et al. 1999) and 3DPSSM (Kelley et al. 2000). It uses a set of associated characteristics such as the quality of the sequence-structure alignment, distribution of sequence features (sequence-conserved positions and apolar residues), and compactness of the resulting models.

Another fully-automated meta-predictor that simply selects models from those produced by other servers is 3D-JURY (Ginalski et al. 2003). It takes as input any set of models, structurally compares all against all using MaxSub (Siew et al. 2000), and selects one that appears to contain the largest recurrent subset of common coordinates. It does not use any special characteristics of the models or of the servers. 3D-JURY is coupled to the BioInfo.PL Meta-server and, thus, can use any model including selection of the *most common* model from a user-defined subset.

2.3 Hybrid Methods: Going Beyond the "Simple Selection" of Models

Some automated meta-predictors go beyond the simple selection. PMOD uses MODELLER (Sali and Blundell 1993) to generate full-atom models based on the selection of fold-recognition results reported by PCONS, amended by secondary structure predicted by PSI-PRED (Jones 1999). These models are evaluated using the PROQ method (Wallner and Elofsson 2003). ROBETTA (D. Baker, unpubl.) builds full-atom models using the ROSETTA fragment insertion method (Simons et al. 1997), starting from structures detected by PDB-BLAST or PCONS and aligned by the K*SYNC alignment method. PRCM takes as input the top models selected by 3D-JURY and builds full-atom models, which are minimized and evaluated using energy functions. ALEPH0-JURY (D. Fischer, unpubl.) selects a model from those of ROBETTA, PRCM, and SHGUM using a combination of the 3D-SHOTGUN technology (see below) and evaluation using knowledge-based potentials.

Another successful practice observed in previous CASPs was to build hybrid models from fragments (e.g. Bujnicki and GeneSilico; see above). Automated meta-predictors using this approach have also been developed. Conceptually, the first method to use the *fragment-splicing* approach (which nevertheless should not be considered a meta-server) was David Baker's ROSETTA protein folding simulation algorithm that uses the fragment insertion Monte Carlo approach (Simons et al. 1997). The general premise of this method is that the protein conformation is reasonably well approximated by the distribution of local structures adopted by known, not necessarily homol-

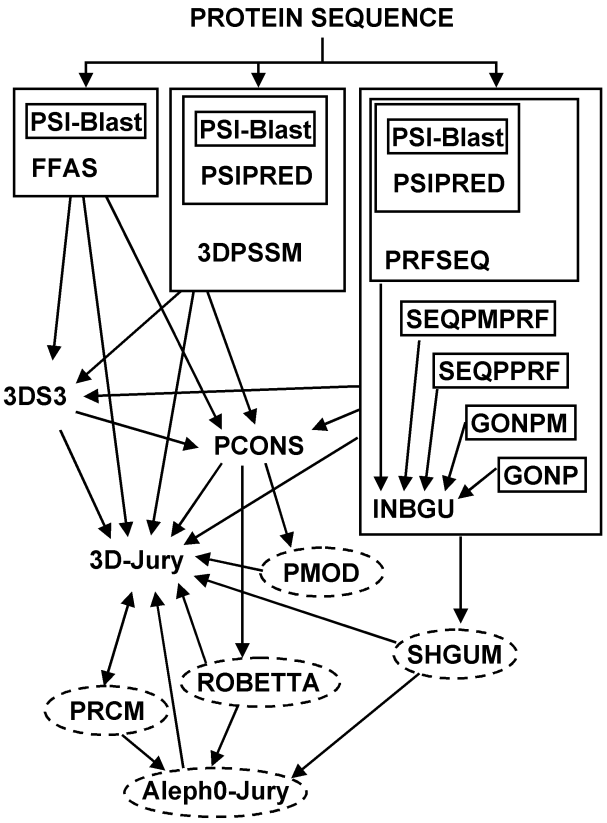


Fig. 1. Mutual interdependencies of meta-servers and their reliance on the original methods. Meta-servers *encircled by broken lines* produce refined (energy-minimized) full-atom models, other meta-servers produce crude C-α models

ogous, protein structures. Protein structure fragments are obtained from the protein structure database (Simons et al. 1997). The original version of ROSETTA utilized the I-sites (invariant or initiation sites) library developed by Chris Bystroff (reviewed elsewhere in this volume), which consists of a set of short motifs, lengths 3 to 19, obtained by a clustering of sequence segments from the Protein Data Bank (Bystroff and Baker 1998). ROSETTA has been notoriously succesful in CASP3, CASP4, and CASP5, demonstrating that protein structure modeling by recombination of fragments derived from experimentally solved structures is a powerful approach.

3D-SHOTGUN (Fischer 2003) is the first fully automated meta-predictor, which assembles hybrid C-α models by combining the structures of individual models, independently obtained from different fold-recognition methods. The 3D-SHOTGUN approach is superior to "pure" selection, as the resulting hybrid models are on the average more complete and more accurate than the input models. There are three versions of 3D-SHOTGUN: (1) an independent version named SHGU using as input models generated by the BIOINBGU server (Fischer 2000); (2) 3DS3 and (3) 3DS5, which uses as input the models

from three or five different independent fold-recognition servers, respectively. A new automated version of the SHOTGUN series, which was very successful in CASP is SHGUM, which generates full-atom, refined models, without the collisions and gaps seen in some of the *raw* spliced models. SHGUM is an independent server using the same input as SHGU (i.e. the results of BIOINBGU).

Figure 1 shows the diagram of mutual interdependencies of "meta^{N}-servers" and their reliance on the input from original servers and other "meta^{N-1}-servers".

3 Future Prospects

The recent CASP5/CAFASP3 evaluation has clearly shown that the meta-servers, on average, perform much better than these primary servers and the higher the *N* in the meta^{N}, the more the meta-server is likely to succeed. This works because no program is suitable for all cases, and each program has its strengths and weaknesses, and with each layer of "meta"- analysis the strengths can be amplified. The idea of meta-servers, or more precisely – the post-CASP5 proliferation of meta-servers – has met, however, with ambiguous reactions of the community of developers of bioinformatic methods. On the one hand, it is much easier to develop meta-servers (especially a relatively simple selector) than to develop a new, original fold-recognition method. On the other hand, because of the *out-sourcing*, the existing meta-servers are very slow: always slower than the slowest of the external servers used to generate primary predictions.

The idea of "meta-prediction" is well known in areas such as artificial intelligence or the stock market, where independent agents are used to obtain a consensus prediction that will be on average more accurate than any of the individual agents. Initially, you consult with various external brokers, but if you have the money, you just hire them to sit at your location. Thus, the richest will be the winner. Likewise, in order to obtain a fast protein fold-recognition meta-server-type method applicable at genomic scales, meta-predictors will run each of the "lower-rank" components locally, without the dependence on other external servers. Many of the existing fold-recognition methods have implemented a local version of PSI-BLAST (Altschul et al. 1997) for database searches and generation of multiple sequence alignment, and PSI-PRED (Jones 1999) (which itself utilizes PSI-BLAST) for prediction of secondary structure. It has been envisaged that the future meta-servers would utilize local implementations of original fold-recognition methods and lower-rank meta-protocols. Hence, some of the criticism attributed to the first generation of meta-servers may not be justified and will certainly fade away in the future, when fast, powerful independent meta-predictors will challenge the best human predictors. Whether this will happen at CASP6 or later, remains to be seen.

References

Alexandrov NN, Nussinov R, Zimmer RM (1996) Fast protein fold recognition via sequence to structure alignment and contact capacity potentials. Pac Symp Biocomput, pp 53–72

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Bryant SH, Lawrence CE (1993) An empirical energy function for threading protein sequence through the folding motif. Proteins 16:92–112

Bujnicki JM, Elofsson A, Fischer D, Rychlewski L (2001a) LiveBench-1: continuous benchmarking of protein structure prediction servers. Protein Sci 10:352–361

Bujnicki JM, Elofsson A, Fischer D, Rychlewski L (2001b) LiveBench-2: Large-scale automated evaluation of protein structure prediction servers. Proteins 45:184–191

Bystroff C, Baker D (1998) Prediction of local structure in proteins using a library of sequence- structure motifs. J Mol Biol 281:565–577

Cuff JA, Barton GJ (2000) Application of multiple sequence alignment profiles to improve protein secondary structure prediction. Proteins 40:502–511

Cuff JA, Clamp ME, Siddiqui AS, Finlay M, Barton GJ (1998) JPred: a consensus secondary structure prediction server. Bioinformatics 14:892–893

Di Francesco V, Geetha V, Garnier J, Munson PJ (1997) Fold recognition using predicted secondary structure sequences and hidden Markov models of protein folds. Proteins (Suppl 1):123–128

Fischer D (2000) Hybrid fold recognition: combining sequence derived properties with evolutionary information. Pac Symp Biocomput, pp 119–130

Fischer D (2003) 3D-SHOTGUN: a novel, cooperative, fold-recognition meta-predictor. Proteins (in press)

Fischer D, Baker D, Moult J (2001a) We need both computer models and experiments. Nature 409:558

Fischer D, Barrett C, Bryson K, Elofsson A, Godzik A, Jones D, Karplus KJ, Kelley LA, MacCallum RM, Pawlowski K, Rost B, Rychlewski L, Sternberg M (1999) CAFASP-1: critical assessment of fully automated structure prediction methods. Proteins (Suppl 3):209–217

Fischer D, Elofsson A, Rychlewski L, Pazos F, Valencia A, Rost B, Ortiz AR, Dunbrack RL Jr (2001b) CAFASP2: The second critical assessment of fully automated structure prediction methods. Proteins 45 (Suppl 5):171–183

Ginalski K, Elofsson A, Fischer D, Rychlewski L (2003) 3D-Jury: a simple approach to improve protein structure predictions. Bioinformatics 19:1015–1018

Godzik A, Kolinski A, Skolnick J (1992) Topology fingerprint approach to the inverse protein folding problem. J Mol Biol 227:227–238

Jones DT (1999) Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 292:195–202

Jones DT, Taylor WR, Thornton JM (1992) A new approach to protein fold recognition. Nature 358:86–89

Juan D, Grana O, Pazos F, Fariselli P, Casadio R, Valencia A (2003) A neural network approach to evaluate fold recognition results. Proteins 50:600–608

Karplus K, Barrett C, Cline M, Diekhans M, Grate L, Hughey R (1999) Predicting protein structure using only sequence information. Proteins (Suppl 3):121–125

Karplus K, Barrett C, Hughey R (1998) Hidden Markov models for detecting remote protein homologies. Bioinformatics 14:846–856

Kelley LA, McCallum CM, Sternberg MJ (2000) Enhanced genome annotation using structural profiles in the program 3D-PSSM. J Mol Biol 299:501–522

Kihara D, Lu H, Kolinski A, Skolnick J (2001) TOUCHSTONE: an ab initio protein structure prediction method that uses threading-based tertiary restraints. Proc Natl Acad Sci USA 98:10125–10130

Kolinski A, Betancourt MR, Kihara D, Rotkiewicz P, Skolnick J (2001) Generalized comparative modeling (GENECOMP): a combination of sequence comparison, threading, and lattice modeling for protein structure prediction and refinement. Proteins 44:133–149

Kosinski J, Cymerman IA, Feder M, Kurowski MA, Sasin JM, Bujnicki JM (2003) A 'Frankenstein's monster' approach to comparative modeling: merging the finest fragments of fold-recognition models and iterative model refinement aided by 3D structure evaluation. Proteins (in press)

Kurowski MA, Bujnicki JM (2003) GeneSilico protein structure prediction meta-server. Nucleic Acids Res 31:3305–3307

Lundstrom J, Rychlewski L, Bujnicki JM, Elofsson A (2001) Pcons: A neural-network-based consensus predictor that improves fold recognition. Protein Sci 10:2354–2362

Marr D (1982) Vision. Freeman, New York

Marti-Renom MA, Stuart AC, Fiser A, Sanchez R, Melo F, Sali A (2000) Comparative protein structure modeling of genes and genomes. Annu Rev Biophys Biomol Struct 29:291–325

Moult J, Fidelis K, Zemla A, Hubbard T (2001) Critical assessment of methods of protein structure prediction (CASP): round IV. Proteins (Suppl 5):2–7

Moult J, Hubbard T, Bryant SH, Fidelis K, Pedersen JT (1997) Critical assessment of methods of protein structure prediction (CASP): round II. Proteins (Suppl 1):2–6

Moult J, Hubbard T, Fidelis K, Pedersen JT (1999) Critical assessment of methods of protein structure prediction (CASP): round III. Proteins (Suppl 3):2–6

Moult J, Pedersen JT, Judson R, Fidelis K (1995) A large-scale experiment to assess protein structure prediction methods. Proteins 23:ii–iv

Ouzounis C, Sander C, Scharf M, Schneider R (1993) Prediction of protein structure by evaluation of sequence-structure fitness. Aligning sequences to contact profiles derived from three-dimensional structures. J Mol Biol 232:805–825

Rost B(1995) TOPITS: threading one-dimensional predictions into three-dimensional structures. ISMB 3:314–321

Rychlewski L, Jaroszewski L, Li W, Godzik A (2000) Comparison of sequence profiles. Strategies for structural predictions using sequence information. Protein Sci 9:232–241

Sali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 234:779–815

Selbig J, Mevissen T, Lengauer T (1999) Decision tree-based formation of consensus protein secondary structure prediction. Bioinformatics 15:1039–1046

Shi J, Blundell TL, Mizuguchi K (2001) Fugue: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. J Mol Biol 310:243–257

Siew N, Elofsson A, Rychlewski L, Fischer D (2000) MaxSub: an automated measure for the assessment of protein structure prediction quality. Bioinformatics 16:776–785

Simons KT, Kooperberg C, Huang E, Baker D (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol 268:209–225

Simons KT, Ruczinski I, Kooperberg C, Fox BA, Bystroff C, Baker D.(1999) Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. Proteins 34:82–95

Sippl MJ, Weitckus S (1992) Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a data base of known protein conformations. Proteins 13:258–271

Tramontano A (2003) Of men and machines. Nat Struct Biol 10:87–90
Wallner B, Elofsson A (2003) Can correct protein models be identified? Protein Sci 12:1073–1086

From Molecular Modeling to Drug Design

M. Cohen-Gonsaud, V. Catherinot, G. Labesse, D. Douguet

1 Introduction

1.1 General Context

Today, the pace of genome sequencing rapidly increases the number of protein sequences. This may lead to a description of living organisms at an unprecedented level of both detail and completeness. It will require the characterization of the biophysical properties and of the biological role of each macromolecular assembly. The growing number of known protein sequences largely exceeds the number of protein structures determined experimentally by NMR and X-ray crystallography (Baker and Sali 2001). However, at the same time, new folds are now rarely discovered despite significant efforts to determine structures of unrelated proteins (see CASP5 results). Meanwhile, a huge number of small molecules can now be easily synthesized and tested experimentally thanks to robotics. Libraries of chemical compounds are rapidly growing while the structural, thermodynamic and dynamic characterization of ligand-macromolecule complexes is still tedious and difficult. These observations suggest that new *in silico* methods (taking advantage of the increasing power of computers) need to be developed in the field of pharmacogenomics.

Since the first modeled protein structure (Browne et al. 1969), numerous modeling studies have been published. Among them, several have highlighted new needs or new strategies pushing forward the field (Crawford et al. 1987). Sequence comparisons allow biologists to identify protein homologies and to routinely derive functional and/or structural information (Rost and Sander 1996). In absence of any significant similarities and in some particular cases (mainly small proteins), ab initio methods may suggest a potential fold but,

M. Cohen-Gonsaud, V. Catherinot, G. Labesse, D. Douguet
Centre de Biochimie Structurale, INSERM U554–CNRS UMR5048, Université Montpellier I, 15, Ave. Charles Flahault 34060 Montpellier Cedex, France

Nucleic Acids and Molecular Biology, Vol. 15
Janusz M. Bujnicki (Ed.)
Practical Bioinformatics

currently, at a resolution too low for ligand docking (Baker and Sali 2001). We will not discuss further the use of ab initio methods except in the particular case of modeling insertions/deletions (hereafter "indels"; see Sect. 2.3.3).

Computational methods for ligand docking into macromolecular structures are more recent (early 1980s) but currently represent a very active field, and are essential at several steps of drug design strategies. However, the combination of the two major *in silico* methodologies, comparative modeling and virtual screening, remains largely unexplored despite tremendous potential applications. In this chapter we shall describe first the modeling of protein structures and the manner by which the resulting theoretical models may be evaluated and used in the context of drug design.

1.2 Comparative Modeling

The sequence similarities a protein shares with proteins already characterized at the functional and/or structural level(s) are widely used to overcome the low output and the cost of experimental biochemical characterization (Orengo et al. 1999; Saveanu et al. 2002). First, one must search for such sequence similarities (e.g. using PSI-BLAST; Altschul et al. 1997) or so-called sequence-structure compatibilities (e.g. using fold recognition). Herein, we shall briefly describe our server, @TOME (http://bioserv.cbs.cnrs.fr/), dedicated to threading (Douguet and Labesse 2001) and molecular modeling (Douguet et al. unpubl.), in order to highlight some specific features in relation to the characterization of ligand binding sites and drug design. The next step is to evaluate the quality of the structural alignment and analyze the derived partial structure or "common core", which corresponds to the aligned residues. This step is not yet successfully automated and represents a major

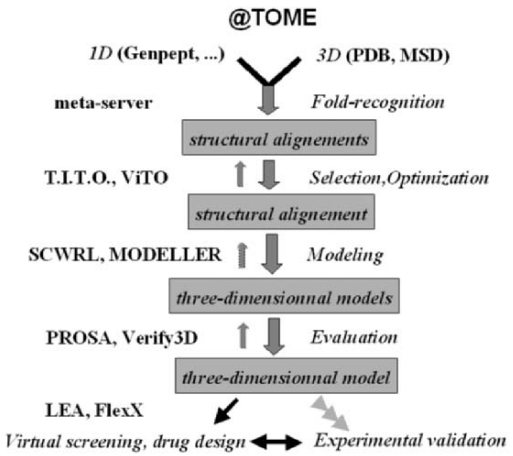


Fig. 1. Flowchart of the pipeline @TOME for macromolecular modeling and drug design. Both fully automated or semi-automated (with user intervention) use of the pipeline is possible allowing error corrections at the various steps connecting genomics to pharmacogenomics

bottleneck for macromolecular modeling. While model building can be performed by fully automatic methods, it is not currently able to recover from a wrong template choice or an incorrect alignment. Several steps are required for complete model building and we will describe them below. However, we do encourage the reader to refer to dedicated reviews of this field for further details (Forster 2002; Marti-Renom et al. 2000). The last (but not the least) step in molecular modeling is the evaluation of the predicted structure. Different analyses are needed depending on the target-template similarities but also the expected use of the resulting models. Docking of other macromolecules or of small ligands will require evaluations at distinct locations and at different resolutions.

1.3 Drug Design and Screening

In silico docking of small molecules (ligands) into large macromolecules (also named "target" or "receptor") has been developed in order to probe their potential interactions. It may be used to explain the mode of action of drugs as well as defining the way to improve them (e.g. derivatization to optimize their specificity and/or affinity). This subtopic in structure-based drug design requires structural data for both the putative ligands and the targeted receptor. Different aspects of virtual docking and drug design have been reviewed recently in more detail elsewhere. The speed of the current software allows the search for good ligands of a given macromolecule in large chemical databases (thousands of molecules a day per workstation). In recent years, several successful structure-based virtual screening studies have been reported (Boehm et al. 2000; Doman et al. 2002; Perola et al. 2000; Gruneberg et al. 2001). Virtual screens, and in particular receptor-based virtual screens, have emerged as a reliable, inexpensive method for identifying "leads" (compounds used as a starting point for drug design). Advances in computational techniques have enabled virtual screening to have a positive impact on the discovery process (Lyne 2002). *In silico* docking becomes a complementary approach to experimental high-throughput screening in the lead identification stage (Jenkins et al. 2003). However, the lack of an experimentally determined structure of the targeted protein frequently limits the application of structure-based drug design methods. Efforts have been made to overcome some limitations and examples of model-based drug design have emerged (see Chap. 3). Some applications have been initiated for several important protein families such as GPCRs (which are targeted by one third of commercial drugs and represent 3 % of the human genome; Klabunde and Hessler 2002) or drug metabolizing enzymes (e.g. P450; Zamora et al. 2003) just to mention a few highly challenging examples. This situation calls for the development of efficient computational methods for structure modeling and ligand screening as well as a global effort to evaluate their limitations.

We will discuss the use of ligand database screening and docking on a small scale to evaluate and/or to refine modeled protein structures. On a larger scale, *in silico* docking will require high quality models. One might envision as well the use of molecular models of various but related proteins to evaluate the specificity of a set of ligands in order to predict potential side effects (Rockey and Elcock 2002). The usefulness of these approaches in the context of genomic biology will be discussed.

2 Comparative Modeling

2.1 Sequence Gathering and Alignment

Before, comparative molecular modeling, i.e. three-dimensional structure building, can be initiated, sequence alignment of the target and (at least) one template is necessary. However, the lower the sequence identity, the harder it is to detect similarity and to align sequences. While obvious at high sequence identity (above 30 %), the detection might not be straightforward at lower sequence identity. A prerequisite is generally to find and align close homologues of the target.

2.1.1 Sequence Database Searches

Sequence database searches were efficiently automated one decade ago through the development of BLAST and its derivatives (Altschul et al. 1990, 1997). Most recent methods, such as fold recognition (see Sect. 2.2.1), include such searches prior to sequence-structure comparison and their efficiency heavily relies on the search output. The use of the template's homologues is also helpful, especially through profile-based methods (Rychlewski et al. 2000). Checking for the availability of a sufficient number of homologues in the sequence databases may be necessary to ascertain the quality of the outputs (alignment, fold recognition, secondary structure prediction). In some cases, this verification is highly recommended, especially, for eukaryotic sequences belonging to small families with no prokaryotic equivalent (Ganem et al. 2003) or particular proteins specific to a phylogenic "niche" (Carret et al. 1999). The number of fully sequenced genomes of prokaryotes usually warrants the construction of reasonable multiple sequence alignments for most proteins of bacterial or archaeal origin. However, some sequence subfamilies might lead to the convergence of PSI-BLAST searches, which is too rapid in the absence of "joining" intermediates between too distantly related subfamilies (Labesse et al. 2001). At the same time, the efficiency of the sequencing projects makes PSI-BLAST searches more and more successful. It may detect true sequence similarity even at a very low level of sequence identity (~15 % over 60–90 % of the protein length; see CASP5

results). In these cases, a reliable alignment is more likely to be achieved using sequence-structure comparison methods and/or the manual edition of sequence-structure alignment (hereafter, named structural alignment) by experts.

2.1.2 Multiple Sequence Alignments

Once similar sequences have been gathered, various sequence alignment methods are available (e.g. CLUSTALW, DIAILIGN, etc.) and can be directly connected to molecular modeling (Lambert et al. 2002). PSI-BLAST itself provides multiple sequence alignments. However, the latter correspond to similarity matches and do not always cover the full-length hit sequences. Compared to pairwise alignment, multiple alignments may reveal more meaningful sequence conservation (Labesse 1996). Computer programs such as MEME (Bailey et al 1997) are available to pick up among aligned sequences, common motifs that usually correspond to functionally or structurally important regions. However, fine functional assignment may require tracing subtle changes aside from common motifs that may not be automatically detected (Labesse et al. 1994; Reid et al. 2003).

The overall quality of the alignment depends mainly on the mean pairwise sequence identity. The statistical significance of a multiple alignment can now be estimated (Pei et al. 2003). At a low level of sequence identity (below 25 %), structural information will be needed to improve the alignment quality (e.g. avoiding insertion or deletion inside secondary structure elements; Gracy et al. 1993).

2.2 Structural Alignments

We wish to put, herein, strong emphasis on the essential step of sequence-structure alignment also called, fold recognition. This requirement is reinforced by the growing use of sequence-structure comparison methods to derive alignments in the so-called twilight and midnight zones (for sequence identity levels between 15–25 and 0–15 %, respectively). We shall illustrate here, with several examples, the need for careful refinement of structural alignment as well as the usefulness of the crude models one can derive from these alignments. Fold recognition is usually performed to search structure databases using "frozen approximation" for speed. It allows rapid similarity detection. In contrast, true three-dimensional threading evaluates pairwise contacts (in between amino acids or atoms) instead of profile-profile matches. The enhanced sensitivity of pairwise contacts suggests that it should be used after profile-profile comparison. This strategy has been implemented in PROSPECT (Xu et al. 2000) or PROSPECTOR (Skolnick and Kihara 2001) and is also made available on the server @TOME. Various factors may interfere

with the achievement of a correct sequence-structure alignment and their identification may require going through all the following steps: alignment refinement (Sect. 2.2.3), model building (Sect. 2.3) and model evaluation (Sect. 2.4).

2.2.1 Fold Recognition

Fold-recognition programs usually produce sequence alignments that are generally more reliable than those derived from purely sequence-based methods. Furthermore, they can detect distant homologues with sequence identity as low as 10 % (Kinch and Grishin 2002). However, the current rate of the success of individual threaders reaches at best 40 % for distantly related structures (Bujnicki et al. 2001). This can be partially overcome by using consensus scoring schemes such as those provided by several web servers (http://BioInfo.PL/meta/meta.html: Bujnicki et al. 2001; http://GeneSilico.pl/meta/: Kurowski and Bujnicki 2003; @TOME). On the server @TOME, structural alignments are further evaluated through a common threading tool (T.I.T.O.; Labesse and Mornon 1998) using a potential of mean force, PKB (Bryant and Lawrence 1993). The use of a common scoring scheme helps to choose a better template and/or a better structural alignment. When distinct folds are proposed to be compatible for the same region of the query sequence, the proposed similarity is doubtful and extra care must be taken before going through the following steps of structure modeling.

Usually, different threaders will find similar compatible folds but their sequence-structure alignments may differ locally. In case of high sequence similarities (above 25 %, over more than 100 residues), discrepancies occur mainly in the vicinity of indels. A few amino acids on each side might be improperly aligned usually due to spurious sequence identity instead of the geometrical likelihood of the indels. Under the level of 25 % sequence identity or in the case of small proteins the significance of the alignment might be questioned (Sander and Schneider 1991). Below 10 % sequence identity, it might be considered that a correct alignment cannot be achieved (except by chance). Difficulties in alignment refinement may arise from sequence divergence but also from structure changes and function variations.

2.2.2 Structural Alignment Refinement

Currently, few tools tackle the problem of automatic refinement of sequence alignments but promising approaches have been described recently (Deane et al. 2001; Pei et al. 2003). However, various internal controls may be used for the selection and refinement of structural alignments using available techniques including three-dimensional structure visualization.

One may evaluate the "stability" of a given alignment by adding new sequences significantly similar either to the template or to the target as well as

experimentally solved structures that superpose well onto the template. Checking the agreement of secondary structure predictions (for the query sequence) with secondary structure assignment (for the template) is important for distantly related proteins (Errami et al. 2003; Callebaut et al. 1997). Other criteria may be taken into account (particular phi/psi angles, burial, hydrogen bonding capabilities, helix capping, etc.) and may be visualized on the structural alignment using the JOY format (Mizuguchi et al. 1998). However, at low levels of sequence conservation, structural alignment should also be evaluated more precisely at the three-dimensional level.

One may build (or rather extract) rapidly a "crude model" (e.g. using the program T.I.T.O. (Labesse and Mornon 1998)). Such a partial structure includes only strictly conserved residues (including both backbone and side-chain atoms) and the backbone of distinct but aligned residues. Neither optimization nor loop building at the indels are performed, adding no error due to the more complex model building methods, that could mask alignment errors. Clusters of strictly conserved residues (e.g. catalytic triad) and/or conservation of topohydrophobic residues (Poupon and Mornon 1998) would suggest functional conservation (e.g. catalytic mechanism) and/or indicate a lower global structure divergence, respectively. A related approach was implemented in THREADLIZE (Pazos et al. 1999). Visual evaluation of a structural alignment quality often suggests numerous local changes in the sequence alignment. These changes may be transposed into a new "crude model". A new round of alignment edition, common-core extraction and assessment is necessary for this trial-and-error optimization. Until recently, the various steps involved in this tedious and time-consuming process, have been performed by several programs, e.g. a multiple-alignment editor such as SEAVIEW (Galtier et al. 1996), T.I.T.O. (Labesse and Mornon 1998) and a macromolecular structure visualization tool such as XmMol (Tuffery 1995), Swiss-PDB viewer (Guex and Peitsch 1997) or Rasmol (Sayle and Milner-White 1995). Two programs gathering most of the previous properties (i.e. editing and visualization) are now available to help this task (Modview: Ilyin et al. 2002; ViTO: Catherinot and Labesse, unpubl.).

2.2.3 Active Site Recognition

Determination of the active site location and prediction of the protein function are essential steps in the "post-genomic era". This may become automated soon based on both modeled structures and sequence conservation using "evolutionary traces" (Lichtarge et al. 1996; Aloy et al. 2001; Yao et al. 2003). Another methodology, based on sequence conservation and active site geometry analysis (Fetrow and Skolnick 1998) has been recently developed for comparative searches. The methods for recognition of active sites may also show loss-of-function evolution (Kniazeff et al. 2002). The significance of the conservation of a cluster of amino acids can also be used to identify subfami-

lies of related proteins. This can be performed using statistical tools such as PATTINPROT (Combet et al. 2000) or PHI-BLAST (Zhang et al. 1998) even at a low level of sequence conservation (~15 %) to confirm fold recognition (Labesse et al. 2001) or to characterize the catalytic mechanism and/or ligand specificity (Carret et al. 1999; Ganem et al. 2003; Reid et al. 2003). Identification of the amino acids involved in the protein activity may also be useful at the model completion step by providing additional restraints (see Sect. 2.3).

2.2.4 A Biological Application

As an example, we have described the study of the human copper transporter Hah1, the crystal structure of which has been solved (Wernimont et al. 2000). Correct identification of the compatible folds may now be obtained using any sequence-structure comparison tools even at a very low sequence identity (e.g. 12 %). A similar approach was previously applied to correctly model this protein at 20 % sequence identity (Hung et al. 1998). Perfect alignment could be achieved by restraining, as much as possible, the deletions to lie in between positions close in space to each other (measured as Cαi-Cαi+1 distances in

```
Hah1(1FE4)    --MPKHEFSVD-MTCGGCAEAVSRVLNKLGGV-KYDIDL
1AFJ          -ATQTVTLAVPGMTCAACPITVKKALSKVEGVSKVDVGF
hah1_TITO     --MPKHEFSV-DMTCGGCAEAVSRVLNKLG-GVKYDIDL
1AFJ_TITO     -ATQTVTLAVPGMTCAACPITVKKALSKVEGVSKVDVGF
hah1_mGT      --MPKHEFSV-DMTCGGCAEAVSRVLNKLGGVK-YDIDL
1AFJ_mGT      -ATQTVTLAVPGMTCAACPITVKKALSKVEGVSKVDVGF
hah1_3DP      -MPKHE-FSV-DMTCGGCAEAVSRVLNKLGGV-KYDIDL
1AFJ_3DP      -ATQTVTLAVPGMTCAACPITVKKALSKVEGVSKVDVGF
hah1_T99      --MPKHEFSV-DMTCGGCAEAVSRVLNKLGGVK-YDIDL
1AFJ_T99      AT-QTVTLAVPGMTCAACPITVKKALSKVEGVSKVDVGF
              ***   *   **                    ****

hah1(1FE4)    PNKKVCIESE---HSMDTLLATLKKTGKTVSYLGLE---
1AFJ          EKREAVVTFDDTKASVQKLTKATADAGYPSSVKQ-----
hah1_TITO     PNKKVCIESE---HSMDTLLATLKKTGKTVSYLGLE---
1AFJ_TITO     EKREAVVTFDDTKASVQKLTKATADAGYPSSVKQ-----
hah1_mGT      PNKKVCIESE---HSMDTLLATLKKTGKTVSYLGLE---
1AFJ_mGT      KREAVVTFDDTKASVQKLTKATADAGYPSSVKQ-----
hah1_3DP      PNKKVCIESEH---SMDTLLATLKKTGKTVSYLGLE---
1AFJ_3DP      EAVVTFDDTKASVQKLTKATADAGYPSSVKQ-----
hah1_T99      PNKKVCIESE---HSMDTLLATLKKTGKTVSYLGLE---
1AFJ_T99      EKREAVVTFDDTKASVQKLTKATADAGYP-----SSVKQ
                        ****               ******
```

Fig. 2. Sequence-structure alignments of Hah1 (PDB1FE4) and PDB1AFJ. Sequence-structure alignment produced by optimal superposition or as published before the determination of the crystal structure PDB1FE4 or as computed by the programs mGenThreader (Jones 1999), 3D-PSMM (Kelley et al 2000) or SAM-T99 (Karplus et al. 1998). Discrepancies among alignments are indicated by the *asterisk* (*)

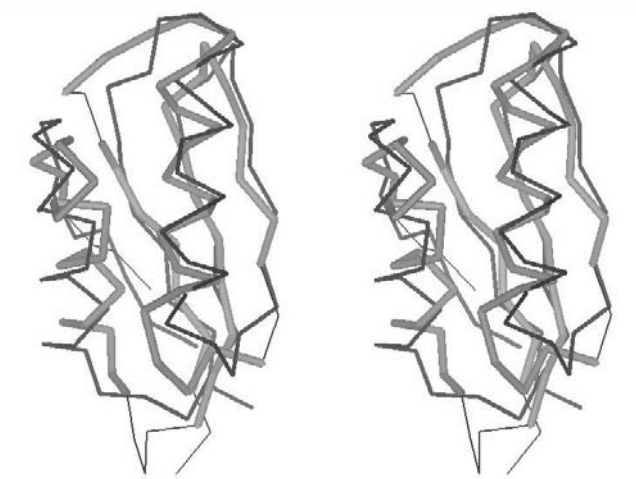


Fig. 3. Stereographic view of the superposition of Ca traces of PDB1FE4 and PDB1AFJ according to the sequence-structure alignment produced by the program SAM-T99. Crystal structure of PDB1FE4 (Wernimont et al. 2000) is drawn in *thin* and *black lines*. PDB1AFJ (Steele and Opella, 1997) is in *grey* and *thick* (*aligned*) or *thin* (*"indels"*) *lines*

the resulting "crude model") and outside of secondary structure elements. The completion of the structure model highlighted additional features such as putative salt-bridges. Model-guided experiments (directed mutagenesis, DTNB labeling or UV-visible spectroscopy of the cobalt-Hah1 complex) quickly validated the proposed alignment (Hung et al. 1998).

2.3 Complete Model Achievement

The frequent need for manual refinement of sequence-structure alignments at a low level of sequence identity (<25 %; see Sect. 2.2), would suggest that no automatic modeling should currently be directly connected to sequence similarity searches. However, subsequent completion of the three-dimensional structure modeling may sometimes result in good models implying that, in this case, a correct structural alignment was achieved. Automatic modeling using several unrefined structural alignments may be performed in parallel using a pipeline dedicated to protein structure modeling such as @TOME. Otherwise, alternative alignments (e.g. suboptimal alignments according to the scoring schemes of automatic procedures) are to be generated and tested. Recognizing the correct model out of numerous incorrect ones will, then, be the next important step (see Sect. 2.4) before one might consider that the resulting macromolecular models are relevant for drug design (see Chap. 3).

2.3.1 Global Structure Modeling

Once a structural alignment is available, a common core is deduced (corresponding to aligned residues; see Sect. 2.2.2) and amino acid changes and indels are delineated. A complete structure may be built from this starting point using various approaches. Completion of the model implies either adding missing parts, or fragments, onto the common core or building and folding the whole structure at once. These methodologies were inspired by the manner in which structures are modeled by X-ray crystallographers or the approach to folding structures using NMR constraints. In between these two approaches, a hybrid methodology is based on databases of protein structure fragments which are used to build missing parts and also to rebuild (or optimize) any parts including the common core. At CASP5, difficult targets (e.g. T0130) were modeled in a better way by mixing large fragments from different but related three-dimensional structures. Such chimeric structures might appear also at a finer level as illustrated by the mycobacterial TMP kinase (Munier-Lehmann et al 2001). The extension of this technique is already available through the use of several templates by more popular modeling programs such as MODELER (Sali and Blundell 1993) and COMPOSER (Srinivasan and Blundell 1993). Other programs and web servers are also available (e.g. SWISS-MODEL: Gueix and Peitsch 1997; Geno3D: Combet et al. 2003). The speed and efficiency of the current modeling software allow the building of models to improve gene detection in genomes (Gopal et al. 2001) or to set up databases such as ModBase (Sánchez et al. 1999) covering, so far, roughly 25 % of protein sequences.Twofold higher coverage can be obtained, but, at the expense of significantly lower structure alignment and structure model quality.

At high levels of sequence identity (above 25 %) little difference in the quality of the modeled structures is observed regardless of which software is used. However, more precise or particular modeling studies will require taking advantage of some specific features of these tools (additional restraints in MODELER such as inter-atomic distance or secondary structure predictions). Otherwise, dedicated programs may be required for specific tasks such as side-chain conformational searches (Sect. 2.3.2), indel building (Sect. 2.3.3) and/or energy minimization (Sect. 2.3.5). We emphasize, here, their general use, their complementarity as well as their potential use for ligand docking and drug design.

2.3.2 Optimization of Side-Chains Conformation

Several tools such as SMD (Tuffery et al. 1997) or SCWRL (Dunbrack and Karplus 1993) are available to build side-chains onto a fixed backbone. They use dedicated rotamer libraries and optimized space search procedures. SCWRL is one of the most popular and is currently made available on the

Fig. 4. View of the active site of the protein kinase AKT. The active site structure was modeled using as a template PKA (Engh et al. 1996). The Ca trace (*thin*) and the ligand H8 (*thick)* are indicated by *grey lines*. Side chains of three residues threonine T141, aspartate D142 and a methionine M131 (T183, D184 and L173 in PKA: PDB1YDS), are indicated by *black* and *thick lines*. Their orientations were computed by SCWRL (Dunbrack and Karplus 1993) using, in absence of H8, either no restraint or constraining the strictly conserved side-chains (e.g. T141 and D142). For clarity, the ligand H8 is shown in its position in PKA

server @TOME. Predicted orientations of side-chains are up to 80 % correct (percent of dihedral angle chi1 within 40Åã of the actual value) for models built by homology. Current improvement now comes from the use of a huge number of conformers for each amino acid, to overcome potentially misleading small van der Waals clashes (but at the expense of the CPU time required). Optimized scoring functions are another way of improvement (Liang and Grishin 2002).

At a low level of sequence identity, active site residues (even those strictly conserved) are usually not properly optimized (generally due to a particular environment and specific conformational constraints). In our experience constraining the original side-chain orientations (to those observed in the template) is often more accurate. This approximation is valid only when similar ligands are expected to bind and/or similar conditions are modeled (e.g. similar allosteric conformations). The use of constraints on the strictly conserved residues has yet to be carefully evaluated (on a larger scale and ahead of ligand docking experiments). Similarly, maintaining a bound ligand while optimizing side-chain conformations may be important prior to virtual screening or docking of ligand analogs. This is illustrated by the catalytic aspartate in protein kinases (D184 in PKA) whose orientation is dramatically changed in the presence of the inhibitor H8 compared to other ligands (Engh et al. 1996). The stabilization of this particular conformation comes from a neighboring threonine (T183 in PKA) hydrogen bonded to the side chain of aspartate D184. Similarly, to maintain the active site pocket "open" enough to allow ligand docking, one may favor modeling a complex with a ligand kept bound.

Template choice (when possible) and specific constraints will depend on the conformation to target and/or the type of ligands to search. Setting up constraints should be carefully revised when significant structural rearrangements are expected in the vicinity (e.g. due to indels).

2.3.3 Insertions/Deletions Building

Different techniques are required according to the length of the "indels", which are generally considered to correspond to loop segments. However, this is no longer true at low levels of sequence identity (below 25 %) as secondary structure elements may vary in length and number among related structures. Modeling of substantial indels, taking into account local secondary structure predictions, is still in its infancy and mainly carried out manually (Aloy et al. 2000). Short indels (usually between three and eight amino acids in length) are modeled more accurately than longer ones. Modeling of indels may be based solely on their own sequences (Deane and Blundell 2001) or it may take into account the potential influence of the surrounding environment (Burke et al. 2001).

Short loops are mainly modeled by taking into account the flanking elements and the sequence of the loop itself. Families of short-loop structures have been defined showing some clear clusters (Kwasigroch et al. 1997; Wojcik et al. 1999) despite the known flexibility of these protein regions. This kind of loop is efficiently modeled using fragments sharing similar sequences and/or compatible geometries (fitting to flanking elements). The fragment-based approaches rely on protein structure databases that should be optimally set up due to high redundancies in the PDB (http://www.rcsb.org/pdb/). Criteria that are too stringent will remove closely related fragments from such preprocessed databases preventing a fine-grained search while ensuring higher speed.

For longer loops (above 12 amino acids in length), additional restraints are necessary to achieve convergence. Their construction may better rely on ab initio modeling (Bystroff and Baker 1998; De Pristo et al. 2003) rather than on comparative modeling despite the need to take into account the surrounding structural elements and the anchoring points. In some cases, very long indels correspond to subdomains that can be modeled independently and are fused later on (see CASP5 results).

The most promising improvement comes from conformation optimization using a specific force field including terms from a potential of mean force at the atomic level. This force field is too CPU-intensive to be used on the global structure. This new loop building approach significantly improved the likelihood of the conformation and it was shown to lower the RMSD (down to 2 Å) of most modeled loops (Fiser et al. 2000). Further improvements (Fiser et al. 2002; de Bakker et al. 2003) come from the use of Generalized-Born solvation approximation to select and/or optimize loop conformations.

2.3.4 Modeling Protein Quaternary Structures

Protein-protein associations play a major role in biology, notably in signaling cascades in eukaryotes or in complex biosynthetic pathways (ribosome, photosynthesis, etc.) and may represent therapeutic targets. The huge number of possible complexes, especially in eukaryotic cells, due to the large protein families involved (e.g. more than 200 human SH3 domains) calls also for the analysis of their specificity through quaternary structure modeling. Furthermore, active sites might be formed or stabilized through macromolecular interactions (e.g. the dimer of the target T0132 at CASP5). Predictions of the quaternary structures have long been too demanding in CPU time and are also dependent on the experimental determination of complexes. However, potentially rapid experimental evaluation of the quaternary structure (or interactions) makes these predictions more attractive. Such predictions may also be performed in conjunction with low-resolution structure determination (Beckmann et al. 2001). Furthermore, the recently developed macromolecular structure database (PQS; http://pqs.ebi.ac.uk/) facilitates the retrieval of most likely quaternary structures from crystal structures. Our server @TOME provides an easy way towards the modeling of the quaternary structure, using MODELER, when structural data are available in the PQS.

In some particular cases, analysis of the putative quaternary structure may confirm putative similarity. For example, modeling of a trimeric structure of the major porin from *Campylobacter jejuni* has confirmed weak sequence similarities (~15% over 400 residues) with better-known enterobacterial malto- and sucroporins (Labesse et al. 2001). The best conserved sequence motifs in these bacterial porins lay at the monomer-monomer interface especially on the trimer axis. In contrast, the external loops as well as the strands facing the lipid membrane show little or no sequence conservation. Furthermore, a putative di-cation binding site at the interface in the model (each monomer providing an aspartate) could then be predicted (Labesse et al. 2001). MultiPROSPECTOR (Lu et al. 2002) represents an automation of this approach by taking advantage of the potential conservation of the quaternary structure to refine threading searches.

Modeling indels and positioning of side chains may be improved if performed in the correct macromolecular context. Furthermore, theoretical evaluation of a modeled structure (see Sect. 2.4) in an incorrect environment (exposing residues normally buried at the interface) might be misleading. The example of the CDK/cyclin complex (Davies et al. 2001) shows that the binding of a macromolecular partner can favorably influence the active site geometry.

All this would prompt us to predict and to build correctly the actual quaternary structure. At a high level of sequence identity, quaternary structure is likely conserved. It will be easily modeled using methods developed for

monomeric structures. At lower sequence identity its conservation may be more questionable and model building will require additional skills.

Evolutionary traces (see Sect. 2.2.3) for large protein families is a convenient tool to predict common interfaces based on structural alignments. Servers are now available to perform rapidly such analysis (Armon et al. 2001). A posteriori analysis might also be convenient to identify a potential interface. One way is to evaluate each monomer first separately and then embedded in the putative complex using tools for model quality evaluation such as Verify3D (see Sect. 2.4.1), which is made available on the server @TOME.

Another way to model quaternary structure is to build partners independently and then try to bring them in contact. This field has been reviewed recently (Smith and Sternberg 2002) and several docking programs are available (Katchalski-Katzir et al. 1992; Smith and Sternberg 2003; Nussinov and Wolfson 1999; Goodsell et al. 1996; Lorber et al. 2002). The use of different methods in parallel and consensus scoring are convenient ways to improve current performance. Low-resolution protein–protein docking (Vakser 1996) is a convenient tool for docking modeled structures (screening out small discrepancies in the monomeric models; Tovchigrechko et al. 2002). Some applications have been recently published such as the modeling of vitronectin, a multi-domain protein, using threading, modeling and docking (Xu at al. 2001). However, the results of the experiment CAPRI (http://capri.ebi.ac.uk) suggest that more developments are necessary before protein–protein docking can be used in routine (Janin et al. 2003).

2.3.5 Energy Minimization and Molecular dynamics

Additional steps may be required to regularize the geometry of the modeled structure, especially in the vicinity of indels (see Sect. 2.3.3). Energy minimization may improve bond length and valence angle values as well as eliminate severe van der Waals clashes. It will not bring atoms closer to their actual position. Due to the roughness of the energy landscape, energy minimizations are easily trapped in local minima. These limitations explain why energy-minimized structures, generally, show slightly increased global deviation (as measured by atomic root-mean-square deviation versus the actual structure) compared to the un-minimized models (or the starting template).

Besides energy minimization, trajectory simulation (molecular dynamics) may be also performed with similar master equations. Molecular dynamics may be used to explore the conformational space. Snapshots in the trajectory may result in models as good as the starting ones (according to various structural criteria; Flohil et al. 2002). This may be used to show the precision (or error) of the models. In MODELER (Sali and Blundell 1993), energy minimization and molecular dynamics are used to optimize and generate distinct models of the same query sequence. Largely deviating regions generally cor-

respond to long indels and may be considered to be incorrectly modeled. Further improvements in available CPU and forcefields may lead, in the near future, to more suitable energy simulation for models optimization.

2.4 Model Validation

2.4.1 Theoretical Model Validation

Several tools are now available to validate three-dimensional structures at different levels of accuracy. At a very high level of sequence identity (above 50 %), small deviations from actual coordinates may be achieved and programs dedicated to experimental structure evaluation are suitable (e.g.: WHAT-CHECK; Hooft et al. 1996). At lower sequence identity (25–50 %), deviation from standard stereochemistry may not correlate with the overall quality of the model (especially after energy minimization; see Sect. 2.3.5). Non-bonding interatomic interactions may be more suitable using atomic statistical potentials such as ERRAT (Colovos and Yeates 1993), ANOLEA (Melo and Feytmans 1998) or SOESA (Wall et al. 1999). Below 25 % sequence identity, model evaluation should rather be performed at the residue level. PROSA II (Sippl 1993) and Verify3D (Eisenberg et al. 1997) are used to assess automatic modeling by MODELER on the server @TOME. In our experience, mainly at low levels of sequence identity (15–25 %), good models have a mean score between 0.3 and 0.4 using Verify3D and between –0.7 and –1.0 in PROSA.

Precise and local analysis may be required in particular cases. Simultaneous visualization of the score and the three-dimensional structure may be done using visualization programs (using the B-factor values to input scores). Specific features remain to be implemented to handle original configurations, which are mostly observed in the active sites (or binding sites). Residues contacting ions (especially, those involved in metal coordination) and/or deeply buried ligands (especially co-factors) have a non-classical environment resulting in disturbed evaluation. Interactions with charged compounds may imply clustering of similarly charged residues (e.g. lysines and arginines for phosphate binding). Similarly, particularities may be observed in thermostable proteins, which may be stabilized by buried salt bridges (or even a buried ion binding site such as -amylases). When buried in the modeled structure, charged or highly hydrophilic residues are often considered to be incorrectly modeled. Attention must be paid to the conservation of these polar and buried residues and/or looking at counterbalancing residues (especially correlated substitution) or chemical groups (backbone atoms, and substrate or co-factor). When such particular features are observed, evaluation of the model quality requires the assessment of the template structure as well.

When a protein structure has been determined under various conditions and shows some rearrangements, models of homologues built using the vari-

ous known forms might indicate some preferred conformations. To what extent this technique can be generalized remains an open question. However, application of this strategy to the eukaryotic cyclin-dependent kinase CDK7 suggested that it might not require cyclin binding for full activity due to subtle amino acid changes in the vicinity of the activation loop. Among these changes, one is a tyrosine to phenylalanine substitution (tyrosine Y15 in CDK2) in the glycine-rich loop and other changes occurred at the N-terminus of the activation loop. The predicted higher stability of the active form due to these correlated changes is in agreement with the observed behavior of this CDK (Martinez et al. 1997).

2.4.2 Ligand-Based Model Selection

Methods testing the complementarity with known ligands may better rank protein models than general structural criteria (e.g. sequence identity, intermolecular energies, etc.). This has been applied recently by Johnson et al. (2003) in the case of the anti-*Shigella flexneri* Y monoclonal antibody complexes. Virtual docking methods (described in Sect. 3.2) may be used on a limited set of experimentally characterized binders (or derived obviously from clear protein homology).

The docking of a common substrate (e.g. TMP) in three TMP kinases (from *Haemophilus influenzae*, *Yersinia pestis*, *Bacillus subtilis*, respectively) modeled using the related TMP kinase from *Escherichia coli*, (75, 75 and 30 % identical, respectively) was used to check the quality of the modeled active site structure (Pochet et al. 2002). Correct docking scores and position were obtained for the enterobacteria while a poor docking score was obtained for the enzyme from *B. subtilis*. This discrepancy is due to a van der Waals clash with a buried proline not present in the template structure as shown by docking on a modeled mutant form (P104A) of the same TMP kinase. This suggested some difficulties in taking into account structural constraints due to the substitution toward a proline in a buried helix. Remodeling this TMP kinase locally would be necessary prior to further ligand screening at high resolution.

2.4.3 Experimental Evaluation of Models

Several biochemical and biophysical characterizations of proteins structures are likely to provide restraints to evaluate a theoretical model at a very low cost in time and in material. However, one should make sure to use methods eliminating alternate models (Hurle et al. 1987). As an example limited proteolysis can be extremely powerful, especially when the cleavage site lies in the protein active site (Bucurenci et al. 1996) or one particular face of the protein (Labesse et al. 2001). Similarly, tryptophan fluorescence may help to monitor substrate orientation and/or a putative induced-fit in the active site (Mar-

rakchi et al. 2002, Cohen-Gonsaud et al. 2002). Mass spectrometry is currently the method of choice in conjunction with other techniques including specific labeling, cross-linking (Young et al. 2000), endo- and exo-proteolysis or, in the case of small proteins, oxidation/reduction (Hung et al. 1998). When quaternary structures are predicted, model evaluation might be easily performed using cross-linking or gel permeation. This, in turn, may highlight some instability or the importance of some conformational change (Marrakchi et al. 2002; Cohen-Gonsaud et al. 2002). Directed mutagenesis is an alternative way to check the functional role of particular residues (Labesse et al. 2001; Kniazeff et al. 2002; Ganem et al. 2003) but it is usually more demanding while at risk of pleiotropic effects making the results difficult to analyze. Chimera of closely related proteins with distinct ligand specificities are an elegant means of building new targets to assess precisely predicted modes of binding (Malherbe et al. 2003). The most precise and most useful validation may be functional assessment through enzymology or affinity measurements especially prior to drug design (Carret et al. 1998; Ganem et al. 2003). With a significantly larger amount of sample (~10 mg), SAXS and ultracentrifugation might be used to assess the overall structure of oligomers as well as the structure of monomers (Bada et al. 2000). Solving experimentally the protein structure, at atomic resolution, will correspond to a final assessment. Only good models are currently suitable to speed up X-ray crystallography using molecular replacement (Jones 2001). Models may potentially also facilitate NMR spectroscopy, in the near future. Experimental structures are usually more suitable for drug design and virtual screening (see Chap. 3) but are determined, currently, at a low output. Prior macromolecular modeling in connection with tuned ligand docking may lead to easier and faster experimental structure determination (e.g. by identifying or by providing a stabilizing ligand) which, in turn, will help further ligand optimization.

2.5 Current Limitations

The methods described above may not be well suited to predict and model specific structural rearrangements such as inter-domain swapping or for particular protein subtypes (membrane or cytoskeletal proteins). Modeling and evaluation tools have to be redesigned or used with extra care in order to tackle these special tasks.

Domain swapping and strand exchanges, as exemplified by the bacterial YajQ (CASP5: T0148), are currently very hard to predict (Saveanu et al. 2002). Contact map predictions might help in some cases but their low overall accuracy is a severe limitation for their general use.

Similarly, membrane protein structures are still difficult to align, to model and to assess. Several reasons explain this situation: few structures have been solved experimentally compared to their soluble counterparts, and moreover,

specific rules apply in the context of protein-lipid interactions. Some particularities can be used to help sequence alignment and model refinement in the case of membrane embedded proteins. Observed in both all-alpha and all-beta membrane proteins, two crowns of mainly aromatic amino acids are underlining the upper and lower limit of the membrane spanning segments. This might correlate with the particular biophysical behavior of these amino acids (solvation energy). Dedicated computer programs have been developed (Diederichs et al. 1998) and may predict structural characteristics (in absence of any detected sequence similarity) more precisely than those usually proposed for soluble and globular proteins (e.g. topology prediction versus simple secondary structure predictions). This kind of prediction is especially useful because experimental tools can assess the topology efficiently (reporter fusions). Large families of membrane proteins have been defined based on a conserved topology, such as GPCRs (Bockaert and Pin 1999), despite sequence identity level below 15 % (Bhave et al. 2003).

3 Model-Based Drug Design

Several examples of protein structure models used for ligand search and/or optimization have already been published (Ring et al. 1993; Munier-Lehmann et al. 2001). However, the accuracy of modeled structures is often thought to be unsuitable for drug design due to too high deviation from the actual structure (Baker and Sali 2001). Nonetheless, structure-structure comparisons regularly show conservation of both fold and ligand binding mode in distantly related proteins (roughly 20 % identical). Analysis of the relationship between sequence and functional descriptors has defined an empirical limit for automatic pairwise-based functional annotations of two of the four EC digits (classification according to the IUBMB Enzyme Nomenclature Committee) at 15 % identity (Devos and Valencia 2000). This suggests that at least the mechanism of action (catalysis for enzyme) is roughly conserved. It may be sufficient to predict potential inhibitors that would mimic the reaction intermediate (Meinhart et al. 2003; Ganem et al. 2003). At higher sequence identity (>30 %) the global shape and nature of the ligand is usually similar. At the same time, small discrepancies in the active sites may appear, suggesting that while the general mode of binding is conserved, specific substitutions may be built (Munier-Lehmann et al 2001). Such correlations may be detected by comparative modeling and used as a starting point for further more-elaborated drug design.

Structure-based drug design or structure-based virtual screening usually involves explicit molecular docking of molecules (mostly small compounds) into the binding site of targets (or receptors). It predicts a binding mode of the compounds and measures, or rather “scores”, the quality of the intermolecular interactions. There are a large number of classical docking programs

available for virtual screening. They differ in the sampling algorithms used, the handling of ligand and protein flexibility, the scoring functions they employ, and the CPU time required to dock a molecule to a given target. Taylor et al. (2002) and Wong and McCammon (2003) have recently described the current state-of-the-art of such methods. The various docking techniques require different types of model qualities and were mainly derived for crystal structures. In order to circumvent discrepancies in the measured and calculated affinities, alternatives schemes of docking and scoring have been developed. Among them, the development of potentials of mean force (Sect. 3.2.1), the representation of protein flexibility (Sect. 3.2.4) and the fragment-based approaches (Sect. 3.2.5) are promising methodologies in connection with macromolecular modeling. Extending comparative modeling of protein structure to ligand recognition may represent a convenient use of sequence similarities that we shall call hereafter: comparative drug-design.

3.1 Comparative Drug Design

Fold-recognition techniques are able to find structures compatible with sequences sharing as little as 10 % identity at the primary level. Among those sequences some are true homologues (or even orthologues) sharing the same function as well as true analogs. The latter may possess neither the same ligand specificity nor the same mechanism of action (esterase and dehalogenases in the alpha/beta hydrolase fold superfamily). However, most of the time, the ligand is bound in a related position relative to the protein fold (especially in alpha/beta folds). Recognition of a potential active site (or ligand binding site) is an important step prior to drug design but also for validation of sequence or structural alignments. It may, also, be used to add new compatible templates sharing a common active site but more distant in the sequence space. This is especially fruitful when these new structures have been solved in complex with small molecules. The latter may be used as lead compounds in drug design. The position of the ligand may also serve to point out functionally important residues (amino acids to be found conserved in the alignment of true homologues). Identification of the likely active site already allows analysis of the local conservation compared to the global one. It may indicate either reminiscence of partial homology or, on the contrary, a dramatic rearrangement of the active site. In the latter case little may be said about the potential ligand and the model may be very approximate and doubtful (e.g. UMP binding in UMP kinase which is 18 % identical to carbamate kinase while UMP and carbamate are dissimilar; Labesse et al. 2002). Evidence of homology would suggest that known ligands (and related compounds) of the characterized homologous proteins could be tested. Identification of a first inhibitor is especially helpful for further function testing by enzymology and derivatization of a known ligand

may serve to validate or to refine locally the model (or select one among several models). It may also indicate the global shape of the binding site and suggest how suitable it is for the design of a specific ligand. In that case, it may serve as a starting point for oriented design of subtle chemical substitutions and/or synthesis of focused compound libraries (Pochet et al. 2002). Identification of a largely open-binding site such as that of ATP in UMPK (Labesse et al. 2002) would suggest little chance of designing a good ligand using a low-resolution model.

A striking example is leukotrien A4 hydrolase/aminopeptidase (LTA4H), a bifunctional enzyme. Both reactions are catalyzed in the same Zn-containing active site (Thunnissen et al. 2001). The presence of some short conserved sequence residues was sufficient to prompt investigations of the relationship of this enzyme to M1 metalopeptidase. It suggested that peptidase inhibitors might be investigated as potential ligands despite the distinct biochemical function and substrate structure. Indeed, the aminopeptidase inhibitor bestatin appeared to inhibit LTB4 biosynthesis (Orning et al. 1991a). Furthermore, another metalloprotease inhibitor, captopril which also inhibits LTA4H has been derived in new compounds with a nanomolar range inhibition (Orning et al. 1991b).

Structural comparisons to search protein structures experimentally determined in the presence of a ligand may indicate new compounds to test prior to an unoriented drug screening. Such a comparative search is especially suitable for a low-resolution model as the ligand docking may be performed exploiting protein structure similarity. Once structures are superimposed, ligands are brought into equivalent regions. The ligand-structure compatibility can be evaluated using various scores described below and used in classical docking methods (see Sect. 3.2). Some databases facilitate comparative searches of putative ligands. Among them, LigBase (Stuart et al. 2002; http://alto.rockefeller.edu/ligbase/) provides structural alignments produced by global superposition using the program CE (Shindyalov and Bourne 1998) and a link to homologous models that were made using any of the aligned templates (gathered in ModBase; Sanchez and Sali 1999). However, distantly related proteins may suffer global rearrangements (particularly domain reorientation) puzzling the programs for global structure superposition. Alternatively, a search by compound similarity scale or by superposition restrained to the binding site may overcome previous difficulties. The Relibase+ database provides convenient search engines in these cases (http://relibase.rutgers.edu; Hendlich 1998). This database contains more than 11,938 protein entries and 43,741 ligand-binding sites for a total of 3,509 unique ligands (May 2000). Once a potential ligand is clearly identified, refinement of the model by inclusion of the chemical compound during the modeling steps may be useful (see Sect. 2.3). One drawback of this approach might be the extraction of mainly biological ligand i.e. natural products. However, close to half of the best-selling pharmaceuticals are either natural prod-

ucts or derivatives thereof (Cragg et al. 1997). Indeed, it has been observed that the hit rates in high-throughput screens determined for natural products collections are often dramatically higher than the rates found for large classical libraries (Breinbauer et al. 2002).

One should remember that the low resolution models derived from macromolecular comparative modeling at low levels of sequence identity (15–25 %) are probably not suitable for large scale drug design and virtual screening. However, such models may indicate a few primary "lead" compounds as starting points for further model and drug design refinement. Such a compound may also serve to stabilize the targeted macromolecule and facilitate its experimental structural characterization.

3.2 Docking Methodologies

Two classes of docking methods will be described below: faster but empirical evaluations (Sects. 3.3.1 and 3.3.2) or more expensive free energy calculations (Sect. 3.3.3). Further information can be found in the review by Gohlke and Klebe (2001). Independent comparative studies have been recently published (Charifson et al. 1999; Bissantz et al. 2000). Consensus scoring approaches suggest that, at present no individual scoring function adequately treats all of the effects important for protein-ligand binding.

Most docking softwares take into account several conformations of a potential ligand. In contrast, the majority of docking tools currently make the assumption that the protein target is held fixed in one given conformation. This approximation is generally necessary in the interest of speed and simplicity, avoiding the computational cost required to accurately sample the flexibility of the binding site. However, some efforts have been made to incorporate protein flexibility. This new strategy may help overcome some inaccuracies in receptor models (see Sect. 3.2.4). This may be useful for the few substituted side-chains one has to model in the active site. In absence of primary tests (already known binders) to discriminate among the different modeled conformations, one will have to consider them equally. Alternatively, one may consider the receptor model to be globally incorrect but locally very accurate. Fragments of putative binders will either fit very well or not at all. Starting from a few docked fragments, new and more complex compounds may be designed to obtained a ligand with increased affinity (see Sect. 3.2.5).

3.2.1 Knowledge-Based Potentials

In this approach, one analyzes the increasing number of experimentally determined protein-ligand complexes by statistical means to extract rules on preferred interaction geometries (frequencies of interatomic contacts). Compared with force-field potentials, knowledge-based potentials tend to be

softer, allowing better handling of the uncertainties and deficiencies of computed interaction geometries (e.g. in modeled structures). Furthermore, such a statistical approach implicitly incorporates physical effects not yet fully described from a theoretical point of view (see Sect. 3.3.3). Examples of knowledge-based scoring functions include PMF (Muegge 2000), Bleep (Mitchell et al. 1999), SmoG (Ishchenko and Shakhnovich 2002) and Drugscore (Gohlke et al. 2000). Their accuracy is comparable with that of empirical-based methods, and they are fast to compute. However, they do require structural data and, at present, they are limited by a paucity of suitable information. Nevertheless, picomolar ligands have been designed by *in silico* screening onto experimental structures using such potentials in combination with a fragment-based approach (Grzybowski et al. 2002).

3.2.2 Regression-Based (or Empirical) Methods

These empirical scoring functions estimate the binding affinity of protein-ligand complexes by adding up weighted interaction terms (hydrogen bonding, hydrophobic interactions, etc.). The weights are assigned by regression methods; fitting predicted and experimentally determined affinities to a given set of training complexes. FlexX (Rarey et al. 1996), SCORE (Bohm 1994), ChemScore (Eldridge et al. 1997), LUDI (Bohm 1992) or PLP (Gehlhaar et al. 1995) use such additive approximations to estimate the binding free energy. These empirical-based scoring functions are fast and therefore are employed often by most docking algorithms. However, the definition of the training set is a major step and may be focused to a too small number of protein or ligand types.

3.2.3 Physics-Based Methods

Physics-based forcefields, may be employed using free energy perturbation (FEP) or thermodynamic integration (TI) methods, to estimate binding free energies (Kollman 1993). They are the best choice for accurately assessing fine chemical modifications that can be made to existing inhibitors to improve their binding affinity. For example, Kuhn and Kollman (2000) were able to predict a derivative that binds stronger than biotin to avidin by changing different C-H groups of biotin into C-F groups. The protein and ligand flexibility are inherent to the method for evaluation of the binding affinities at the expense of CPU time. Furthermore, evaluation of the solvent contributions still represents a major challenge in view of the computational demands and accuracy. Related methods use approximations to the binding free energy of protein-ligand complexes by adding up the individual contributions of different types of interactions. These terms are derived from physico-chemical theory and are not determined by fitting to experimental affinities. In most cases, gas-phase molecular mechanical energies are combined with implicit solvent

models, such as MM/PBSA (molecular mechanics/Poisson-Boltzmann surface area; Kuhn and Kollman 2000) or the Generalized-Born model (Dominy and Brooks 1999). Nevertheless, it is still difficult to examine the binding of a large number of compounds to a receptor with these highly CPU-consuming methods. They may rather be used, at an early stage, to improve the modeled active site or in the latter stage to help the "lead-compound" optimization process.

For docking and virtual screening purposes, physics-based scoring functions employ a reduced force field. Among them the most commonly used are: DOCK (Ewing et al. 2001), AutoDock (Goodsell and Olson 1990), QXP (McMartin and Bohacek 1997), ICM (Internal Coordinates Mechanics) (Abagyan et al. 1994) and Prodock (Trosset and Scheraga 1999). ICM has been tested in the CASP-2 experiments (Totrov and Abagyan 1997) to predict eight complexes (with resulting RMSD values varying between 1.8 and 10.6 A).

3.2.4 Flexible Models

Using only one rigid protein structure of aldose reductase for virtual screening, one would have missed potential inhibitors whereas the latter can be docked, taking into account conformational changes (Claussen et al. 2001). Furthermore, considering the flexibility of the protein in the case of a protein model is also important in order to resolve some inaccuracies in atom position. Modeling protein flexibility during docking of each ligand may be still too CPU-demanding for general purposes but represents an interesting development in the future in connection with precise analysis of model local accuracies. The easiest way to take into account macromolecule flexibility is currently to build an ensemble of static models. They may be generated, for example, by randomization and/or from various templates or by molecular dynamic simulation (Kollman 1996; Brooks et al. 1983). As an example, the docking program FlexE works on an ensemble of structures (Claussen et al. 2001). The FlexE approach is based on a united protein description generated from the superimposed structures of the ensemble. For varying parts of the protein, discrete alternative conformations are explicitly taken into account, which can be combinatorially joined to create new valid protein structures.

A new method called "relaxed complex methods" has been described by Lin et al. (2002). It allows an induced fit of the targeted protein. First, several target conformations should be generated as above. In a second phase, a simple, coarse-grained scoring algorithm is used to allow fast docking of a small set of molecules. The last step corresponds to a more accurate positioning and evaluation of the free energies of binding of the best complexes. The program Slide also enables the motion and relaxation of binding-site side chains in response to the presence of a docked ligand (the so-called induced fit) (Schnecke and Kuhn 2000).

Precise evaluations of these various approaches in connection with macromolecular modeling remain to be performed on distinct protein and ligand types.

3.2.5 Fragment-Based Drug Design

Fragment-based methods determine energetically favorable binding site positions for various functional chemical group types (fragments) or small chemical compounds (methane, methanol, etc.). It would represent the first step to de novo drug-design while the second step would correspond to the assembly of multiple fragments into a chemical compound. GRID (Goodford 1985) and MCSS (Miranker and Karplus 1991) are examples of software using the fragment positioning approach. Another alternative may be the development of a dynamic pharmacophore model based on a number of snapshots from molecular dynamics simulations. For each snapshot, Carlson et al. (2000) determined components of a pharmacophore model by identifying favorable binding sites of chemical functional groups using MUSIC program. This program identifies favorable binding sites of a large number of small probe molecules. Strong binding sites tend to cluster many probe molecules in well-defined orientations and locations. The deduced pharmacophore can be used in identifying potent inhibitors from a database of molecules by chemical similarity. This approach is derived from two successful experimental methods namely "SAR by NMR" (Shuker et al. 1996) and the "tether method" (Erlanson et al. 2000).

3.3 Virtual Screening Using Models

While comparative docking takes advantage of structure similarities with previously determined ligand-receptor structures, classical docking (see Sect. 3.2) may be used to search chemical compound databases to highlight potential binders with new structures and new modes of binding. While low-resolution models are no longer suitable, the following applications show that significant similarities are already sufficient for successful search of micromolar "lead" compounds.

3.3.1 Docking onto Medium Resolution Models

The development of antiparasitic agents by Ring et al. (1993) using model-based virtual screenings is among the first and few such docking studies published so far. Two protease structures have been modeled and used to search a potential binder using ligand docking with program DOCK3.0. Among the 55,313 compounds, 52 and 31 compounds where selected for the two proteases. Because of the uncertainties in the models built, the authors have cho-

sen chemically diverse compounds that were predicted to interact in different ways. After experimental testing, three inhibitors displayed activity against the enzymes at micromolar concentrations.

Combinations of model-based docking with ligand data, used by and derived from 3D quantitative structure-activity relationships (QSAR) may lead to improved results and may overcome model discrepancies (Schafferhans and Klebe 2001).

3.3.2 Docking onto High-Resolution Models

In the case of high-resolution models, the active site is likely to be very well modeled. The description of the active site may be finely prepared, for example, by the addition of hydrogens required for finer docking and energy computation. The appropriate protonation states of ionizable residues need to be determined and the correct tautomer for histidines should be assigned, as well. Sometimes, the positions of the hydrogens are relaxed by energy minimization to avoid any steric clashes. At last, in some instances, tightly bound water(s), ions and/or cofactor(s) might need to be maintained for the docking stage.

In high-throughput virtual screening, the source of the ligands typically corresponds to a corporate collection of physically available compounds, or a database of compounds available externally from chemical vendors (e.g. the MDL® Available Chemicals Directory (ACD) from MDL Information Systems; http://mdli.com). An additional source that is sometimes considered for virtual screening is an *in silico* virtual library corresponding to compounds constructed from a list of reagents and a database of known chemistries. These compounds may be easily purchased or synthesized in order to evaluate their true affinities.

As an example, Wang and coworkers used a homology model for Bcl-2, derived from the solution structure of Bcl-x_L (47.2 % i.d.) (Wang et al. 2000). A total of 193,833 compounds were screened using the program DOCK 3.5 to score shape complementarity for each virtual compound bound to the Bcl-2 model in a variety of conformations. Among a total of 28 compounds available commercially, one proved to be a ligand for Bcl-2 with a IC_{50} of ~9 μM.

To illustrate the need for water molecules and ions in the representation of the active site we have studied docking of cAMP onto the human phosphodiesterase PDE4. To date, there is only one crystal structure of this enzyme, which does not contain any ligand but crystal water molecules and ions (Mg and Zn metals). Metals ions are essential for the hydrolysis by PDE4 of cAMP and/or cGMP in AMP and GMP products, respectively (Liu et al. 2001). In the absence of metals neither cAMP nor the sugar moiety (i.e. cyclic-monophosphoryl ribose) could be docked into the active site using the program FlexX (Rarey et al. 1996). Including ions and a few water molecules coordinated to the metals allowed for better docking results. Indeed, the phosphoryl oxygens

appeared to coordinate Mg and Zn ions in agreement with the hypothesis of Liu et al. (2001). Nevertheless, no satisfactory docking could be achieved with the substrate cAMP, suggesting that protein flexibility should be taken into account.

3.4 Pharmacogenomic Applications

3.4.1 A Challenging Application: the GPCRs

With the single exception of bovine rhodopsin, there are no experimental 3D structures available for G-protein-coupled receptors (GPCRs). GPCRs are membrane proteins hardly overexpressed or purified while their pharmacology is better characterized (Bockaert and Pin 1999). This situation has encouraged theoretical modeling of GPCRs and model evaluation using docking of well-characterized ligands and/or directed mutagenesis (Gershengorn and Osman 2001; Klabunde and Hessler 2002).

Bissantz et al. (2003) have investigated whether comparative models of GPCRs are reliable enough to be used for virtual screening of chemical databases. They first constructed "antagonist-bound" molecular models of three human GPCRs (dopamine D3 receptor, acetylcholine muscarinic M1 receptor, vasopressin V1a receptor). The sequence identity between the template ("antagonist-bound" rhodopsin) and the sequence to model is between 21 and 29 %. Preliminary attempts to dock known ligands into the starting models usually failed regardless of which docking tool was used. Energy minimization of the putative complex ligand/protein was necessary. Then random compounds (990) and known antagonists (10) have been virtually screened against these models. The results show that these models were suitable to retrieve known antagonists of different structural classes from a database of structurally different molecules (hit rates are 20- to 40-fold higher than what can be obtained by random screening). Nevertheless, this strategy could not be applied to derive a model of the agonist-state of such receptor (dopamine D3 receptor, beta2-adrenergic receptor and mu-opioid receptor). Docking efficiency is limited in this case by the capability to model conformational change in protein structure. In contrast, the quality of the antagonist-state models is validated for numerous proteins at once.

3.4.2 Family-Wide Docking

Docking on models of related proteins may be necessary in order to characterize the specificity as well as the mode of binding of a set of substrate analogs as exemplified, here, with various homologous TMP kinases. These enzymes are essential for cell proliferation and have been studied intensively over the last few years. Two crystal structures (TMP kinases from *Escherichia*

coli, *Mycobacterium tuberculosis*) and three modeled structures (sequences from *Haemophilus influenzae*, *Yersinia pestis*, *Bacillus subtilis*) have been used for focused docking (Pochet et al. 2002). While the catalytic residues appeared strictly conserved, several substitutions were observed in the vicinity of the bound nucleotide (dTMP). No nucleoside has currently been co-crystallized with a TMP kinase. We predicted a mode of binding of dT that suggested a reorientation of the 5'-hydroxyl group (instead of a phosphoryl group in dTMP) to form a specific hydrogen bond. Sequence variations among the TMP kinases in the vicinity of the reoriented chemical group were used to test the hypothesis (presence of an asparagine in *B. subtilis* instead of an aspartate in other TMP kinases). Replacement of the 5′-hydroxyl group by an amino group only slightly affects the affinity for the mycobacterial enzyme but dramatically decreases the binding to that from *B. subtilis*.

3.4.3 Side Effect Predictions

A small molecule may bind not only to one unique receptor but also potentially to various protein-receptors. An essential issue in virtual screening is target-selectivity i.e. the capacity to predict the range of related proteins one drug-candidate will actually bind to. Genome-sequencing projects provide us with the complete set of proteins. However, one needs the protein structure to apply efficient docking strategies. Combined use of comparative modeling and selective docking has been recently described (Rockey and Elcock 2002). Being able to rationally tune the target-range of a chemical compound would limit the potential side effects that currently represent a major bottleneck in drug design and development.

3.4.4 Drug Metabolism Predictions

In addition to the receptor-ligand affinity, another important aspect of drug design is also the behavior (named ADMET for absorption, distribution, metabolism, excretion and toxicity) of drug-candidates in the targeted organisms (hosts and parasites). Empiric rules are now available to predict the behavior of drug candidates. Characterizations of the proteins involved in the various processes (transport, metabolism, detoxification, etc.) may rationalize the former approach. Several specialized databases have been developed for ADMET-associated proteins: transporter (http://lab.digibench.net/transporter/), cytochromes P450 (http://medicine.iupui.edu/flockhart/ and http://p450.abc.hu/) and ADME-AP (http://xin.cz3.nus.edu.sg/group/admet/admet.asp). Obtaining the structure of the corresponding proteins is a major challenge of "integrated pharmacogenomics" as illustrated by the study of cytochromes P450.

The cytochromes P450 constitute a huge superfamily of heme-thiolate enzymes involved in the metabolism of a large number and structural diver-

sity of substrates. P450s from pathogens (including *M. tuberculosis* which possesses more than 20 P450) represent important drug targets while human P450s present in liver (CYP1, CYP2 and CYP3 families) are associated with the oxydative metabolism of the majority of drugs in current clinical use (limiting their half-time). Only a few experimental structures of P450 s have been solved to date (including CYP102 (Ravichandran et al. 1993) and CYP2C5 (Williams et al. 2000)). Comparative modeling of human P450 and ligand docking results are largely consistent with currently available experimental information from site-directed mutagenesis and substrate metabolism studies (Lewis 2002). In the near future, P450 models should allow for the screening of drug candidates in order to better define their potential efficiency (Zamora et al. 2003).

4 Conclusions

Depending on both the sequence and the functional conservation among proteins will provide structural models of different quality. Comparative macromolecular modeling may already provide some functional clues even at a low level of sequence identity (at the resolution of fold-recognition techniques). At higher sequence identity (above 25 %), clear homology may give rise to medium resolution models of good quality in the active site (usually better conserved). At even higher sequence identity (>45 %) theoretical models are equivalent to low-resolution experimental structures and may provide very good templates for large-scale virtual screening and fine drug design. The different level of accuracies will become an important issue for large-scale pharmacogenomics, especially in order to predict potential side effects, a major difficulty in current drug development. Adapting the drug design strategy to the likelihood of the modeled active site will be an important step to develop further comparative drug design and model-based virtual screening. Developing new bioinformatic tools (software and databases) will be necessary for a rapidly increasing number of biological applications. A critical assessment of the combination of modeling and docking techniques might require a community-wide extension of those set up for structure prediction (CASP: Venclovas et al. 2001; EVA: Eyrich et al. 2001; LiveBench: Bujnicki et al. 2001b) on one side and those for protein-ligand interaction prediction on the other side (CASP2, CATFEE; http://uqbar.ncifcrf.gov/~catfee).

Acknowledgements. The authors wish to thank Hélène Munier-Lehmann, Jean-Philippe Pin, Sylvie Pochet Jean-Luc Pons, and Annaik Quémard for helpful discussions and Laetitia Martin, Laurent Chiche, Stefan Arold and Cathy Royer for critical reading of the manuscript.

References

Abagyan R, Totrov M, Kuznetsov D (1994) ICM – a new method for protein modeling and design: applications to docking and structure prediction from the distorted native conformation. J Comput Chem 15:488–506

Aloy P, Querol E, Aviles FX, Sternberg MJ (2001) Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. J Mol Biol 311:395–408

Aloy P, Mas JM, Marti-Renom MA, Querol E, Aviles FX, Oliva B (2000) Refinement of modeled structures by knowledge-based energy profiles and secondary structure prediction:application to the human procarboxypeptidase A2. J Comput Aided Mol Des 14:83–92

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Altschul SF, Madden TL, Schaffer A, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST : a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Armon A, Graur D, Ben-Tal N (2001) ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. J Mol Biol 307:447–463

Bada M, Walther D, Arcangioli B, Doniach S, Delarue M (2000) Solution structural studies and low-resolution model of the *Schizosaccharomyces pombe* sap1 protein. J Mol Biol 300:563–574

Bailey TL, Baker ME, Elkan CP. (1997) An artificial intelligence approach to motif discovery in protein sequences: application to steriod dehydrogenases. J Steroid Biochem Mol Biol. 62:29–44

Baker D, Sali A (2001) Protein structure prediction and structural genomics. Science 294:93–96

Beckmann R, Spahn CM, Eswar N, Helmers J, Penczek PA, Sali A, Frank J, Blobel G (2001) Architecture of the protein-conducting channel associated with the translating 80S ribosome. Cell 107:361–372

Bhave G, Nadin BM, Brasier DJ, Glauner KS, Shah RD, Heinemann SF, Karim F, Gereau IV RW (2003) Membrane topology of a metabotropic glutamate receptor. J Biol Chem 278:30294–30301

Bockaert J, Pin JP (1999) Molecular tinkering of G protein-coupled receptors: an evolutionary success. EMBO J 18:1723–9

Bissantz C, Bernard P, Hibert M, Rognan D (2003) Protein-based virtual screening of chemical databases. II. Are homology models of G-Protein Coupled Receptors suitable targets? Proteins 50:5–25

Bissantz C, Folkers G, Rognan D (2000) Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. J Med Chem 43:4759–4767

Bohm HJ (1992) The computer program LUDI: a new method for the de novo design of enzyme inhibitors. J Comput Aided Mol Des 6:61–78

Bohm HJ (1994) The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. J Comput Aided Mol Des 8:243–256

Boehm HJ, Boehringer M, Bur D, Gmuender H, Huber W, Klaus W, Kostrewa D, Kuehne H, Luebbers T, Meunier-Keller N, Mueller F (2000) Novel inhibitors of DNA gyrase: 3D structure based needle screening, hit validation by biophysical methods, and 3D

guided optimization. A promising alternative to random screening. J Med Chem 43:2664–2674
Breinbauer R, Vetter IR, Waldmann H (2002) From protein domains to drug candidates-natural products as guiding principles in the design and synthesis of compound libraries. Angew Chem Int Ed Engl 41:2879–2890
Brooijmans N, Kuntz ID (2003) Molecular Recognition and Docking Algorithms. Annu Rev Biophys Biomol Struct 32:335–373
Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M (1983) CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. J Comput Chem 4:187–217
Browne WJ, North AC, Phillips DC, Brew K, Vanaman TC, Hill RL (1969) A possible three-dimensional structure of bovine alpha-lactalbumin based on that of hen's egg-white lysozyme. J Mol Biol. 42:65–86
Bryant SH, Lawrence CE (1993) An empirical energy function for threading protein sequence through the folding motif Proteins 16:92–112
Bucurenci N, Sakamoto H, Briozzo P, Palibroda N, Serina L, Sarfati RS, Labesse G, Briand G, Danchin A, Barzu O, Gilles AM (1996) CMP kinase from *Escherichia coli* is structurally related to other nucleoside monophosphate kinases. J Biol Chem 271:2856–2862
Bujnicki JM, Elofsson A, Fischer D, Rychlewski L (2001) Structure prediction meta server. Bioinformatics 17:750–751
Bujnicki JM, Elofsson A, Fischer D, Rychlewski L (2001) LiveBench-1: Continuous Benchmarking of structure prediction server. Protein Sci 10:352–361
Burke DF, Deane CM (2001) CODA: a combined algorithm for predicting the structurally variable regions of protein models. Protein Sci 10:599–612
Bystroff C, Baker D (1998) Prediction of local structure in proteins using a library of sequence-structure motifs. J Mol Biol 281:565–77
Callebaut I, Labesse G, Durand P, Poupon A, Canard L, Chomilier J, Henrissat B, Mornon JP (1997) Deciphering protein sequence information through Hydrophobic Cluster Analysis (HCA) : current status and perspectives. Cell Mol Life Sci 53:621–645
Carlson HA, Masukawa KM, Rubins K, Bushman FD, Jorgensen WL (2000) Developing a dynamic pharmacophore model for HIV-1 integrase. J Med Chem 43:2100–2114
Carret C, Delbecq S, Labesse G, Carcy B, Precigout E, Moubri K, Schetters TPM, Gorenflot A (1999) Characterization and molecular cloning of an adenosine kinase from *Babesia canis rossi*. Eur J Biochem 265:1015–1021
Charifson PS, Corkery JJ, Murcko MA, Walters WP (1999) Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. J Med Chem 42:5100–5109
Claussen H, Buning C, Rarey M, Lengauer T (2001) FlexE: efficient molecular docking considering protein structure variations. J Mol Biol 308:377–395
Cohen-Gonsaud M, Ducasse S, Hoh F, Zerbib D, Labesse G, Quémard A (2002) Crystal structure of MabA from *Mycobacterium tuberculosis*, a reductase involved in long-chain fatty acid biosynthesis. J Mol Biol 320:249–261
Colovos C, Yeates TO (1993) Verification of protein structures: patterns of nonbonded atomic interactions. Protein Sci. 2:1511–1519
Combet C, Blanchet C, Geourjon C, Deleage G (2000) NPS@: network protein sequence analysis. Trends Biochem Sci 25:147–150
Combet C, Jambon M, Deleage G, Geourjon C (2002) Geno3D: automatic comparative molecular modeling of protein. Bioinformatics 18:213–214
Cragg CM, Newman DJ, Snader KM (1997) For an outstanding analysis of the role of natural products in pharmaceuticals. J Nat Prod 60:52–60

Crawford IP, Niermann T, Kirschner K (1987) Prediction of secondary structure by evolutionary comparison: application to the alpha subunit of tryptophan synthase. Proteins 2:118–129

Davies TG, Tunnah P, Meijer L, Marko D, Eisenbrand G, Endicott JA, Noble ME (2001) Inhibitor binding to active and inactive CDK2: the crystal structure of CDK2-cyclin A/indirubin-5-sulphonate. Structure 9:389–397

De Bakker PI, De Pristo MA, Burke DF, Blundell TL (2003) Ab initio construction of polypeptide fragments: Accuracy of loop decoy discrimination by an all-atom statistical potential and the AMBER force field with the Generalized Born solvation model. Proteins 51:21–40

Deane CM, Blundell TL (2001) Improved protein loop prediction from sequence alone. Protein Eng 14:473–478

Deane CM, Kaas Q, Blundell TL (2001) SCORE: predicting the core of protein models. Bioinformatics 17:541–550

DePristo MA, De Bakker PI, Lovell SC, Blundell TL (2003) Ab initio construction of polypeptide fragments: efficient generation of accurate, representative ensembles. Proteins 51:41–55

Devos D, Valencia A (2000) Practical limits of function prediction. Proteins 41:98–107

Diederichs K, Freigang J, Umhau S, Zeth K, Breed J (1998) Prediction by a neural network of outer membrane beta-strand protein topology. Protein Sci 7:2413–2420

Doman TN, McGovern SL, Witherbee BJ, Kasten TP, Kurumbail R, Stallings WC, Connolly DT, Shoichet BK (2002) Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1B. J Med Chem 45:2213–2221

Dominy BN, Brooks CL 3rd (1999) Methodology for protein-ligand binding studies: application to a model for drug resistance, the HIV/FIV protease system. Proteins 36:318–331

Douguet D, Labesse G (2001) Easier threading through Web-based comparisons and cross-validations. Bioinformatics 17:752–753

Dunbrack RL Jr, Karplus M (1993) Backbone-dependent rotamer library for proteins. Application to side-chain prediction. J Mol Biol 230:543–74

Eisenberg D, Luthy R, Bowie JU (1997) VERIFY3D: assessment of protein models with three-dimensional profiles. Methods Enzymol 277:396–404.

Eldridge MD, Murray CW, Auton TR, Paolini GV, Mee RP (1997) Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. J Comput Aided Mol Des 11:425–445

Engh RA, Girod A, Kinzel V, Huber R, Bossemeyer D (1996) Crystal structures of catalytic subunit of cAMP-dependent protein kinase in complex with isoquinolinesulfonyl protein kinase inhibitors H7, H8, and H89. Structural implications for selectivity. J Biol Chem 271:26157–26164

Erlanson DA, Braisted AC, Raphael DR, Randal M, Stroud RM, Gordon EM, Wells JA (2000) Site-directed ligand discovery. Proc Natl Acad Sci USA 97:9367–9372

Errami M, Geourjon C, Deleage G (2003) Detection of unrelated proteins in sequences multiple alignments by using predicted secondary structures. Bioinformatics 19:506–512

Fiser A, Do RK, Sali A (2000) Modeling of loops in protein structures. Protein Sci 9:1753–1773

Ewing TJ, Makino S, Skillman AG, Kuntz ID (2001) DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. J Comput Aided Mol Des 15:411–428

Eyrich VA, Marti-Renom MA, Przybylski D, Madhusudhan MS, Fiser A, Pazos F, Valencia A, Sali A, Rost B (2001) EVA: continuous automatic evaluation of structure prediction servers. Bioinformatics 17:1242–1243

Fetrow JS, Skolnick JA (1998) Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T1 ribonucleases. J Mol Biol 281:949–68.
Feig M, BrooksIII CL, Sali A (2002) Evolution and physics in comparative protein structure modeling. Acc Chem Res 35:413–421
Flohil JA, Vriend G, Berendsen HJC (2002) Completion and refinement of 3-D homology models with restricted molecular dynamics: Application to targets 47, 58, and 111 in the CASP modeling competition and posterior analysis. Proteins 48:593–604
Forster MJ (2002) Molecular modeling in structural biology. Micron 33:365–384
Galtier N, Gouy M, Gautier C (1996) SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. Comput Appl Biosci 12:543–548
Ganem C, Devaux F, Torchet C, Jacq C, Quevillon-Cheruel S, Labesse G, Facca C, Faye G (2003) Ssu72 is a phosphatase essential for transcription termination of snoRNAs and specific mRNAs in yeast. EMBO J 22:1588–1598
Gehlhaar DK, Moerder KE, Zichi D, Sherman CJ, Ogden RC, et al. (1995) De novo design of enzyme inhibitors by Monte Carlo ligand generation. J Med Chem 38:466–472
Gershengorn MC, Osman R (2001) Minireview: insights into G protein-coupled receptor function using molecular models. Endocrinology 142:2–10
Goodford PJ (1985) A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. J Med Chem 28:849–857
Goodsell DS, Morris GM, Olson AJ (1996) Automated docking of flexible ligands: applications of AutoDock. J Mol Recogn 9:1–5
Goodsell DS, Olson AJ (1990) Automated docking of substrates to proteins by simulated annealing. Proteins 8:195–202
Gohlke H, Hendlich M, Klebe G (2000) Knowledge-based scoring function to predict protein-ligand interactions. J Mol Biol 295:337–356
Gohlke H, Klebe G (2001) Statistical potentials and scoring functions applied to protein-ligand binding. Curr Opin Struct Biol 11:231–235
Gopal S, Schroeder M, Pieper U, Sczyrba A, Aytekin-Kurban G, Bekiranov S, Fajardo EJ, Eswar N, Sánchez N, Sali A, Gaasterland T (2001) Homology-based annotation yields 1,042 new candidate genes in the *Drosophila melanogaster* genome. Nat Genet 27:337–340
Gracy J, Chiche L, Sallantin J (1993) Improved alignment of weakly homologous protein sequences using structural information. Protein Eng 6:821–829
Gruneberg S, Wendt B, Klebe G (2001) Subnanomolar Inhibitors from Computer Screening: A Model Study Using Human Carbonic Anhydrase II. Angew Chem Int Ed Engl 40:389–393
Grzybowski BA, Ishchenko AV, Kim CY, Topalov G, Chapman R, Christianson DW, Whitesides GM, Shakhnovich EI (2002) Combinatorial computational method gives new picomolar ligands for a known enzyme. Proc Natl Acad Sci USA 99:1270–1276
Guex N, Peitsch MC (1997) SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling. Electrophoresis 18:2714–2723
Hendlich M (1998) Databases for protein-ligand complexes. Acta Crystallogr D Biol Crystallogr 54:1178–1182
Hooft RW, Vriend G, Sander C, Abola EE (1996) Errors in protein structures. Nature 381:272
Hung IH, Casareno RL, Labesse G, Mathews FS, Gitlin JD (1998) HAH1 is a copper-binding protein with distinct amino acid residues mediating copper homeostasis and antioxidant defense. J Biol Chem 273:1749–1754
Hurle MR, Matthews CR, Cohen FE, Kuntz ID, Toumadje A, Johnson WC Jr (1987) Prediction of the tertiary structure of the alpha-subunit of tryptophan synthase. Proteins 2:210–224

Ilyin VA, Pieper U, Stuart AC, Martí-Renom MA, McMahan L, Sali A (2002) ModView, visualization of multiple protein sequences and structures. Bioinformatics 19:165–166

Ishchenko AV, Shakhnovich EI (2002) Small molecule growth 2001 (SMoG2001): an improved knowledge-based scoring function for protein-ligand interactions. J Med Chem 45:2770–2780

Janin J, Henrick K, Moult J, Eyck LT, Sternberg MJ, Vajda S, Vakser I, Wodak SJ (2003) CAPRI: A critical assessment of predicted interactions. Proteins 52:2–9

Jenkins JL, Kao RY, Shapiro R (2003) Virtual screening to enrich hit lists from high-throughput screening: a case study on small-molecule inhibitors of angiogenin. Proteins 50:81–93

Johnson MA, Hoog C, Pinto BM (2003) A novel modeling protocol for protein receptors guided by bound-ligand conformation. Biochemistry 42:1842–1853

Jones DT (1999) GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. J Mol Biol 287:797–815

Jones DT (2001) Evaluating the potential of using fold-recognition models for molecular replacement. Acta Crystallogr D Biol Crystallogr 57:1428–1434

Karplus K, Barrett C, Hughey R (1998) Hidden Markov models for detecting remote protein homologies. Bioinformatics 14:846–856

Katchalski-Katzir E, Shariv I, Eisenstein M, Friesem AA, Aflalo C, Vakser IA (1992) Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. Proc Natl Acad Sci USA 89:2195–2199

Kelley LA, MacCallum RM, Sternberg MJE (2000) Enhanced Genome Annotation using Structural Profiles in the Program 3D-PSSM. J Mol Biol 299:501–522

Kinch LN, Grishin NV (2002) Expanding the nitrogen regulatory protein superfamily: Homology detection at below random sequence identity. Proteins 48:75–84

Klabunde T, Hessler G (2002) Drug design strategies for targeting G-protein-coupled receptors. Chembiochem. 3:928–944.

Kniazeff J, Galvez T, Labesse G, Pin JP (2002) Ligand binding in the GB2 subunit of the GABAB heteromeric receptor is not required for receptor activation and allosteric interaction between the subunits. J Neurosci 22:7352–7361

Kollman PA (1993) Free energy calculations: application to chemical and biochemical phenomena. Chem Rev 93:2395–2417

Kollman PA (1996) AMBER 5.0. University of California, San Francisco

Kuhn B, Kollman PA (2000) Binding of a diverse set of ligands to avidin and streptavidin: an accurate quantitative prediction of their relative affinities by a combination of molecular mechanics and continuum solvent models. J Med Chem 43:3786–3791

Kuhn B, Kollman PA (2000) A ligand that is predicted to bind better to avidin than biotin: insights from computational fluorine scanning. J Am Chem Soc 122:3909–3916

Kurowski MA, Bujnicki JM (2003) GeneSilico protein structure prediction meta-server Nucleic Acids Res 31:3305–3307

Kwasigroch JM, Chomilier J, Mornon JP (1996) A global taxonomy of loops in globular proteins. J Mol Biol 259:855–772

Labesse G, Vidal-Cros A, Chomilier J, Gaudry M, Mornon JP (1994) Structural comparisons lead to the definition of a new superfamily of NAD(P)(H)-oxydoreductases: the single domain reductases/epimerases/dehydrogenases, the 'RED' family. Biochemical J 304:95–99

Labesse G (1996) MulBlast 1.0: a multiple alignment of BLAST output to boost protein sequence similarity analysis. CABIOS 12:463–467

Labesse G, Mornon JP (1998) Incremental threading optimization (TITO) to help alignment and modeling of remote homologues. Bioinformatics 14:206–211

Labesse G, Garnotel E, Bonnel S, Dumas C, Pagés JM, Bolla JM (2001) MOMP a divergent porin from *Campylobacter jejuni* cloning and primary structural characterization. BBRC 280:380–387

Labesse G, Bucurenci N, Douguet D, Sakamoto H, Landais S, Gagyi C, Gilles AM, Bârzu O (2002) Comparative modeling and immunochemical studies of *Escherichia coli* UMP kinase. BBRC 294:173–179

Lambert C, Leonard N, De Bolle X, Depiereux E (2002) ESyPred3D: prediction of proteins 3D structures. Bioinformatics 18:1250–1256

Lewis DFV (2002) modeling human cytochromes P450 involved in drug metabolism from the CYP2C5 crystallographic template. J Inorg BioChem 91:502–514

Liang S, Grishin NV (2002) Side-chain modeling with an optimized scoring function. Protein Sci 11:322–331

Lichtarge O, Bourne HR, Cohen FE (1996) An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol 257:342–358

Lin JH, Perryman AL, Schames JR, McCammon JA (2002) Computational drug design accommodating receptor flexibility: the relaxed complex scheme. J Am Chem Soc 124:5632–5633

Liu S, Laliberte F, Bobechko B, Bartlett A, Lario P, Gorseth E, Van Hamme J, Gresser MJ, Huang Z (2001) Dissecting the cofactor-dependent and independent bindings of PDE4 inhibitors. Biochemistry 40:10179–10186

Lorber DM, Udo MK, Shoichet BK (2002) Protein–protein docking with multiple residue conformations and residue substitutions. Protein Sci 11:1393–1408

Lu L, Lu H, Skolnick J (2002) MULTIPROSPECTOR: An algorithm for the prediction of protein-protein interactions by multimeric threading Proteins 49:350–364

Lyne PD (2002) Structure-based virtual screening: an overview. Drug Discov Today 7:1047–1055

Malherbe P, Kratochwil N, Knoflach F, Zenner MT, Kew JN, Kratzeisen C, Maerki HP, Adam G, Mutel V (2003) Mutational analysis and molecular modeling of the allosteric binding site of a novel, selective, noncompetitive antagonist of the metabotropic glutamate 1 receptor. J Biol Chem 278:8340–8347

Marrakchi H, Ducasse S, Labesse G, Margeat E, Montrozier H, Emorine L, Charpentier X, Lanéelle G, Quémard A (2002) Biochemical and structural studies of the *Mycobacterium tuberculosis* MabA (FabG1) protein involved in the fatty acid elongation system FAS-II. Microbiology 148:951–960

Marti-Renom MA, Stuart AC, Fiser A, Sanchez R, Melo F, Sali A (2000) Comparative protein structure modeling of genes and genomes. Annu Rev Biophys Biomol Struct 29:291–325

Martinez AM, Afshar M, Martin F, Cavadore JC, Labbe JC, Doree M (1997) Dual phosphorylation of the T-loop in cdk7: its role in controlling cyclin H binding and CAK activity. EMBO J 16:343–354

McMartin C, Bohacek RS (1997) QXP: powerful, rapid computer algorithms for structure-based drug design. J Comput Aided Mol Des 11:333–344

Meinhart A, Silberzahn T, Cramer P (2003) The mRNA transcription/processing factor ssu72 is a potential tyrosine phosphatase. J Biol Chem 278:15917–15921

Melo F, Sánchez R, Sali A (2002) Statistical potentials for fold assessment. Protein Science 11: 430–448

Melo F, Feytmans E (1998) Assessing protein structures with a non-local atomic interaction energy. J Mol Biol 277:1141–1152

Miranker A, Karplus M (1991) Functionality maps of binding sites: a multiple copy simultaneous search method. Proteins 11:29–34

Mitchell JBO, Laskowski RA, Alex A, Thornton JM (1999) BLEEP-potential of mean force describing protein-ligand interactions. I. Generating potential. J Comput Chem 20:1165–1176

Mizuguchi K, Deane CM, Blundell TL, Johnson MS, Overington JP (1998) JOY: protein sequence-structure representation and analysis. Bioinformatics. 14:617–623

Muegge I (2000) A knowledge-based scoring function for protein-ligand interactions: probing the reference state. Perspect Drug Des Discov 20:99–114

Munier-Lehmann H, Chaffotte A, Pochet S, Labesse G (2001) Thymidylate Kinase of *Mycobacterium tuberculosis*: A chimera sharing properties common to Eucaryotic and bacterial enzymes. Prot Science 10:1195–1205

Nussinov R, Wolfson HJ (1999) Efficient computational algorithms for docking and for generating and matching a library of functional epitopes I. Rigid and flexible hinge-bending docking algorithms. Comb Chem High Throughput Screen 2:249–259

Orengo CA, Todd AE, Thornton JM (1999) From protein structure to function. Curr Opin Struct Biol 9:374–382

Orning L, Krivi G, Bild G, Gierse J, Aykent S, Fitzpatrick FA (1991a) Inhibition of leukotriene A4 hydrolase/aminopeptidase by captopril. J Biol Chem 266:16507–16511

Orning L, Krivi G, Fitzpatrick FA (1991b) Leukotriene A4 hydrolase. Inhibition by bestatin and intrinsic aminopeptidase activity establish its functional resemblance to metallohydrolase enzymes. J Biol Chem 266:1375–1378

Pazos F, Rost B, Valencia A (1999) A platform for integrating threading results with protein family analyses. Bioinformatics 15:1062–1063

Pei J, Sadreyev R, Grishin NV (2003) PCMA: fast and accurate multiple sequence alignment based on profile consistency. Bioinformatics 19:427–428

Perola E, Xu K, Kollmeyer TM, Kaufmann SH, Prendergast FG, Pang YP (2000) Successful virtual screening of a chemical database for farnesyltransferase inhibitor leads. J Med Chem 43:401–408

Pochet S, Dugue L, Douguet D, Labesse G, Munier-Lehmann H (2002) Nucleoside analogs as inhibitors of Thymidylate Kinases: possible therapeutical applications. ChemBioChem 3:108–110

Poupon A, Mornon JP (1998) Populations of hydrophobic amino acids within protein globular domains: identification of conserved "topohydrophobic" positions. Proteins 33:329–342

Rarey M, Wefing S, Lengauer T (1996) Placement of medium-sized molecular fragments into active sites of proteins. J Comput Aided Mol Des 10:41–54

Ravichandran KG, Boddupalli SS, Hasermann CA, Peterson JA, Deisenhofer J (1993) Crystal structure of hemoprotein domain of P450BM-3, a prototype for microsomal P450's. Science 261:731–736

Reid R, Piagentini M, Rodriguez E, Ashley G, Viswanathan N, Carney J, Santi DV, Hutchinson CR, McDaniel R (2003) A model of structure and catalysis for ketoreductase domains in modular polyketide synthases. Biochemistry 42:72–79

Ring CS, Sun E, McKerrow JH, Lee GK, Rosenthal PJ, Kuntz ID, Cohen FE (1993) Structure-based inhibitor design by using protein models for the development of antiparasitic agents. Proc Natl Acad Sci USA 90:3583–3587

Rockey WM, Elcock AH (2002) Progress toward virtual screening for drug side effects. Proteins 48:664–671

Rost B, Sander C (1996) Bridging the protein sequence-structure gap by structure prediction. Annu Rev Biophys Biomol Struct 25:113–136

Rychlewski L, Jaroszewski L, Li W, Godzik A (2000) Comparison of sequence profiles. strategies for structural predictions using sequence information. Protein Sci 9:232–241

Sali A, Blundell TL (1993) Comparative protein modeling by satisfaction of spatial restraints. J Mol Biol 234:779–815.

Sanchez R, Sali A (1999) ModBase: a database of comparative protein structure models. Bioinformatics 15:1060–1061

Sander C, Schneider R (1991) Database of homology derived protein structures and the structural meaning of sequence alignment. Proteins 9:56–68

Saveanu C, Miron S, Borza T, Craescu CT, Labesse G, Gagyi C, Popescu A, Schaeffer F, Namane A, Laurent-Winter C, Bârzu O, Gilles AM (2002) Structural and ligand binding properties of YajQ and YnaF, two proteins from *Escherichia coli* with unknown biological function. Protein Sci 11:2551–2560

Sayle RA, Milner-White EJ (1995) RASMOL: biomolecular graphics for all. Trends Biochem Sci 20:374

Schafferhans A, Klebe G (2001) Docking ligands onto binding site representations derived from proteins built by homology modeling. J Mol Biol 307:407–427

Schnecke V, Kuhn LA (2000) Virtual screening with solvation and ligand-induced complementarity. Perspect Drug Discov 20:171–190

Shindyalov IN, Bourne PE (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng 11:739–747

Shuker SB, Hajduk PJ, Meadows RP, Fesik SW (1996) Discovering high-affinity ligands for proteins: SAR by NMR. Science 274:1531–1534

Sippl MJ (1993) Recognition of errors in three-dimensional structures of proteins. Proteins 17:355–362

Skolnick J, Kihara D (2001) Defrosting the frozen approximation: PROSPECTOR–a new approach to threading. Proteins 42:319–331

Smith GR, Sternberg MJ (2002) Prediction of protein-protein interactions by docking methods. Curr Opin Struct Biol 12:28–35

Smith GR, Sternberg MJ (2003) Evaluation of the 3D-Dock protein docking suite in rounds 1 and 2 of the CAPRI blind trial. Proteins 52:74–79

Srinivasan N, Blundell TL (1993) An evaluation of the performance of an automated procedure for comparative modeling of protein tertiary structure. Protein Eng 6:501–512

Steele RA, Opella SJ (1997) Structures of the reduced and mercury-bound forms of MerP, the periplasmic protein from the bacterial mercury detoxification system. Biochemistry 36:6885–6895

Stuart AC, Ilyin VA, Sali A (2002) LigBase: a database of families of aligned ligand binding sites in known protein sequences and structures. Bioinformatics 18:200–201

Taylor RD, Jewsbury PJ, Essex JW (2002) A review of protein-small molecule docking methods. J Comput Aided Mol Des 16:151–166

Thunnissen MM, Nordlund P, Haeggstrom JZ (2001) Crystal structure of human leukotriene A(4) hydrolase, a bifunctional enzyme in inflammation. Nat Struct Biol 8:131–135

Totrov M, Abagyan R (1997) Flexible protein-ligand docking by global energy optimization in internal coordinates. Proteins Suppl 1:215–220

Tovchigrechko A, Wells CA, Vakser IA (2002) Docking of protein models. Protein Sci 11:1888–1896

Trosset JY, Scheraga HA (1999) PRODOCK: software package for protein modeling and docking. J Comput Chem 20:412–427

Tuffery P, Etchebest C, Hazout S (1995) Prediction of protein side chain conformations: a study on the influence of backbone accuracy on conformation stability in the rotamer space. Protein Eng 10:361–372

Vakser IA (1996) Low-resolution docking: prediction of complexes for underdetermined structures. Biopolymers. 39:455–464

Venclovas C, Zemla A, Fidelis K, Moult J (2001) Comparison of performance in successive CASP experiments. Proteins 5:163–170

Wall ME, Subramaniam S, Phillips GN Jr (1999) Protein structure determination using a database of interatomic distance probabilities. Protein Sci 8:2720–2727

Wang JL, Liu D, Zhang ZJ, Shan S, Han X, Srinivasula SM, Croce CM, Alnemri ES, Huang Z (2000) Structure-based discovery of an organic compound that binds Bcl-2 protein and induces apoptosis of tumor cells. Proc Natl Acad Sci USA 97:7124–7129

Wernimont AK, Huffman DL, Lamb AL, O'Halloran TV, Rosenzweig AC (2000) Structural basis for copper transfer by the metallochaperone for the Menkes/Wilson disease proteins. Nat Struct Biol 7:766–771

Wojcik J, Mornon JP, Chomilier J (1999) New efficient statistical sequence-dependent structure prediction of short to medium-sized protein loops based on an exhaustive loop classification. J Mol Biol 289:1469–1490

Williams PA, Cosme J, Sridhar V, Johnson EF, McRee DE (2000) Mammalian microsomal cytochrome P450 monooxygenase: structural adaptations for membrane binding and functional diversity. Mol Cell 5:121–131

Wong CF, McCammon JA (2003) Protein flexibility and computer-aided drug design. Annu Rev Pharmacol Toxicol 43:31–45

Xu D, Baburaj K, Peterson CB, Xu Y (2001) Model for the three-dimensional structure of vitronectin: predictions for the multi-domain protein from threading and docking. Proteins 44:312–320

Yao H, Kristensen DM, Mihalek I, Sowa ME, Shaw C, Kimmel M, Kavraki L, Lichtarge O (2003) An accurate, sensitive, and scalable method to identify functional sites in protein structures. J Mol Biol 326:255–261

Young MM, Tang N, Hempel JC, Oshiro CM, Taylor EW, Kuntz ID, Gibson BW, Dollinger G (2000) High throughput protein fold identification by using experimental constraints derived from intermolecular cross-links and mass spectrometry. Proc Natl Acad Sci USA 97:5802–5806

Zamora I, Afzelius L, Cruciani G (2003) Predicting drug metabolism: a site of metabolosim prediction tool applied to the cytochrome P450 2C9. J Med Chem 46:2313–2324

Zhang Z, Schaffer AA, Miller W, Madden TL, Lipman DJ, Koonin EV, Altschul SF (1998) Protein sequence similarity searches using patterns as seeds. Nucleic Acids Res 26:3986–3990

Structure Determination of Macromolecular Complexes by Experiment and Computation

F. Alber, N. Eswar, A. Sali

1 Introduction

The function of a protein is defined by its interactions with other molecules in its environment. The interactions can be either transient, such as protein–protein interactions involved in intracellular signaling, or relatively stable, such as the protein–protein and protein–RNA interactions in ribosomes. A structural description of these interactions is an important step toward understanding the mechanisms of biochemical, cellular, and higher order biological processes. There is a need to integrate structural information gathered at multiple levels of the biological hierarchy – from atoms to cells – into a common framework. Recent developments in several experimental and computational techniques allow structural biology to shift its focus from the structures of the individual proteins to larger assemblies (Sali et al. 2003; Baumeister 2002).

Macromolecular assemblies vary widely in their functions and sizes (Alberts 1998; Goto et al. 2002; Grakoui et al. 1999; Courey 2001; Noji and Yoshida 2001). They play crucial roles in most cellular processes, and are often depicted as molecular machines (Alberts 1998). This metaphor accurately captures many of their characteristic features, such as modularity, complexity, cyclic functions, and energy consumption (Nogales and Grigorieff 2001). For instance, the nuclear pore complex, a 50–100 MDa protein assembly, regulates and controls the traffic of macromolecules through the nuclear envelope (Rout et al. 2000); the ribosome is responsible for protein biosynthesis; the RNA polymerase catalyzes the formation of RNA (Murakami and Darst 2003); and the ATP synthase catalyzes the formation of ATP (Noji and Yoshida 2001).

F. Alber, N. Eswar, A. Sali (✉)
Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry, and California Institute for Quantitative Biomedical Research, Mission Bay Genentech Hall, Suite N472D, 600 16th Street, University of California at San Francisco, San Francisco, California 94143–2240, USA

Nucleic Acids and Molecular Biology, Vol. 15
Janusz M. Bujnicki (Ed.)
Practical Bioinformatics

Macromolecular assemblies are also involved in transcription control (i.e., IFNβ enhanceosome) (Courey 2001; Nogales 2000), regulation of cellular transport (i.e., microtubulines in complex with molecular motors myosin or kinesin) (Vale 2003; Goldstein and Yang 2000; Vale and Milligan 2000), and are crucial components in neuronal signaling (e.g., the postsynaptic density complexes) (Gomperts 1996).

The estimation of the total number of macromolecular complexes in a proteome is not a trivial task. This difficulty can be partly ascribed to the multitude of component types (e.g., proteins, nucleic acids, nucleotides, metal ions), and the varying lifespan of the complexes (e.g., transient complexes such as those involved in signaling and stable complexes such as the ribosome).

The Protein Quaternary Structure Database (PQSD; Nov 2002) contains ~10,000 structurally defined protein assemblies of presumed biological significance, derived from a variety of organisms (http://pqs.ebi.ac.uk/pqs-doc.shtml). Each assembly consists of at least two protein chains. These assemblies can be organized into ~3,000 groups that contain chains with more than 30 % sequence identity to at least one other member of the group (Fig. 1; Sali et al. 2003).

The most comprehensive information about protein-protein interactions is available for the yeast proteome consisting of ~6,200 proteins. The lower bound on binary protein-protein interactions and functional links in yeast has been estimated to be in the range of ~30,000 (Kumar and Snyder 2002; von Mering et al. 2002); this number corresponds to ~9 protein partners per protein, though not necessarily all at the same time. The human proteome may have an order of magnitude of more complexes than the yeast cell; and the number of different complexes across all relevant genomes may be several times larger still. Therefore, there may be thousands of biologically relevant macromolecular complexes whose structures are yet to be characterized (Abbott 2002).

In contrast to structure determination of the individual macromolecules, structural characterization of macromolecular assemblies usually poses a more difficult challenge. A comprehensive description of large complexes generally requires the use of several experimental methods, underpinned by a variety of theoretical approaches to maximize efficiency, completeness, accuracy, and resolution of the determination of assembly composition and structure.

X-ray crystallography has been the most prolific technique for the structural analysis of proteins and protein complexes, and is still the 'gold standard' in terms of accuracy. Structures of several macromolecular assemblies have recently been solved: the RNA polymerase (Cramer et al. 2001), the ribosomal subunits (Ban et al. 2000; Harms et al. 2001; Wimberly et al. 2000); the complete ribosome and its functional complexes (Yusupov et al. 2001); the proteasome (Lowe et al. 1995); GroEl (Braig et al. 1994); the cellular transport

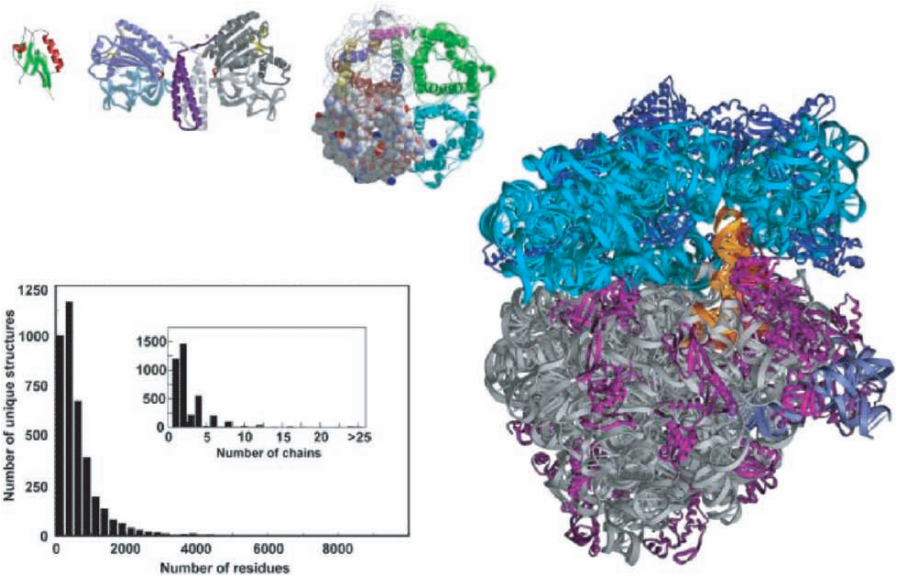


Fig. 1. Illustration of the size range of biomolecular structures solved by X-ray crystallography and the size distribution of structures contained in the Protein Quaternary Structure database. Structures of (*top left* to *right*) the PDZ domain, a molecular recognition domain that leads to protein–protein interactions; CheA, a dimeric multidomain bacterial signaling molecule; aquaporin, which serves as a transmembrane water channel; and 70S ribosome, which is the molecular machine for protein biosynthesis. The histogram shows the distribution of the size of the entries in the Protein Quaternary Structure (PQS) database (http://pqs.ebi.ac.uk). The 15,190 entries with at least one protein chain of at least 30 residues, when compared with each other, produced 3,876 clusters with more than 30 % sequence identity and less than 30 residue length differences among the members within the same cluster. The distributions of the numbers of residues and chains (*inset*) in the representative structures for each group are shown. As expected, the structures of large complexes are under-represented, given an estimated average size of a yeast complex of 7.5 proteins (see text)

machinery (Goldstein and Yang 2000; Vale 2003), and various viral capsid and virion structures (Grimes et al. 1995; Oda et al. 2000). However, the number of structures of macromolecular assemblies solved by X-ray crystallography is still quite small compared to that of individual proteins (Fig. 1). This discrepancy is due mainly to the difficult production of sufficient quantities of the sample and its crystallization.

There are several variants of electron microscopy, including single-particle electron microscopy (EM; Frank 1996), electron tomography (Baumeister 2002), and electron crystallography of regular two-dimensional arrays of the sample (Nogales et al. 1998). For large particles with molecular weights larger than 250 to 500 kDa, single particle cryo-EM can reveal the shape and symmetry of an assembly at resolutions of 1–2 nm. Although the electron micro-

scope produces images that represent only 2D projections of the specimen, the full 3D structure of the object can be reconstructed from many such projections, each showing the object from a different angle (Frank 1996). More importantly, imaging by cryo-EM at these resolutions requires neither large quantities of the sample nor the sample in a crystalline form.

In the absence of high-resolution assembly crystal structures, approximate atomic models of assemblies can still be derived by combining low-resolution cryo-EM data of whole protein assemblies with computational docking of atomic resolution structures of their subunits (Nogales et al. 1998; Volkmann et al. 2000; Spahn et al. 2001; Beckmann et al. 2001; Chiu et al. 2002; Chacon and Wriggers 2002). Recent developments in the methods for interpretation of low-resolution cryo-EM maps have suggested that docking and fitting of atomic resolution subunit structures can enhance the structural information content of the maps to a large extent. It has been estimated that using fitting techniques improves the accuracy up to one tenth the resolution of the original EM reconstruction (Volkmann and Hanein 1999; Roseman 2000; Wriggers et al. 2000; Rossmann et al. 2001; Wriggers and Birmanns 2001).

Unfortunately, atomic resolution crystal structures of the isolated subunits are frequently not available. Alternatively, the induced fit may severely limit their utility in the reconstruction of the whole assembly. In such cases, it might frequently be possible to get useful comparative protein structure models of the subunits (Blundell et al. 1987; Greer 1990; Sali and Blundell 1993; Marti-Renom et al. 2000; Sauder and Dunbrack Jr. 2000; Murzin and Bateman 2001). This approach is increasingly more applicable because of the structural genomics initiative. One of the main goals of structural genomics is to determine a sufficient number of appropriately selected structures from each domain family, so that all sequences are within modeling distance of at least one known protein structure (Baker and Sali 2001). It has also been shown that the number of models that can be constructed with useful accuracy is already two orders of magnitude higher than the number of available experimental structures (Pieper et al. 2002).

We begin by introducing the need for a multi-scale description of macromolecular assemblies that integrates information derived from multiple sources and variable resolution into a common computational framework (Sect. 2). Next, we review the role comparative modeling may play in the determination of atomic structures by EM (Sect. 3). In particular, we introduce automated comparative protein structure modeling (Sect. 3.1), its errors (Sect. 3.2), ways to predict errors (Sect. 3.3), and utility of comparative models in docking of assembly subunits into EM maps. Finally, we illustrate combined comparative modeling and map fitting with two applications, the determination of partial atomic models of the 80S ribosome from *Saccharomyces cerevisiae* and the 70S ribosome of *Escherichia coli* (Sects. 3.5 and 3.6).

2 Hybrid Approaches to Determination of Assembly Structures

Although X-ray crystallography and EM in combination with atomic structure docking have been successfully employed to solve structures of protein assemblies, they are not capable of efficiently characterizing the myriad of complexes that exist in a cell. For example, most of the transient complexes cannot be addressed with these approaches. Therefore, there is a great need for hybrid methods where accuracy, high throughput, and/or high resolution are improved by integrating information from all available sources (Fig. 2) (Malhotra et al. 1990; Aloy et al. 2002). Information about the structure of an assembly can be provided by a number of experimental and theoretical methods (Fig. 2). For instance, the shape, density and symmetry of a complex or its subunits may be derived from X-ray crystallography (Ban et al.

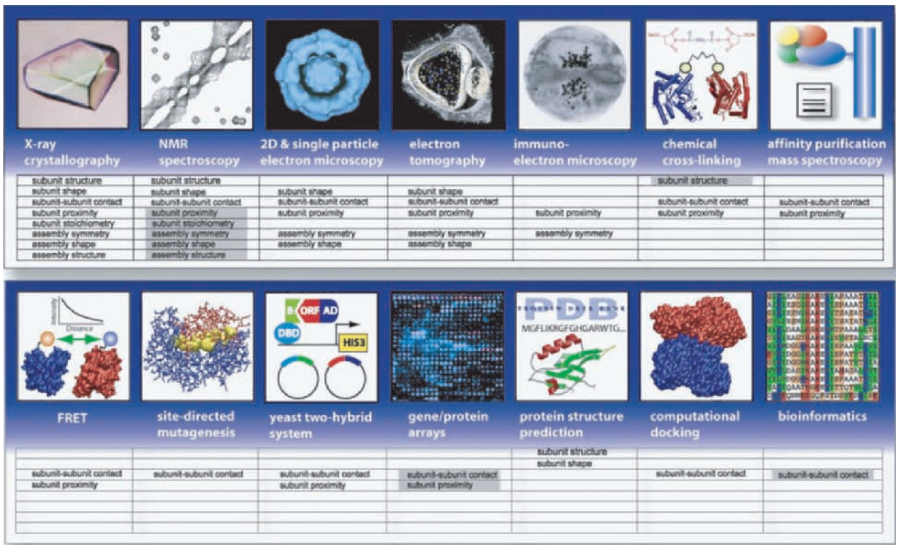


Fig. 2. Experimental and theoretical methods that can provide information about a macromolecular assembly structure. The annotations below each of the panels list the aspects of an assembly that might be obtained by the corresponding method. Subunit and assembly structure indicate an atomic or near atomic resolution at 3 Å or better. Subunit and assembly shape indicate the density or surface envelope at a low-resolution of worse than 3 Å. Subunit–subunit contact indicates knowledge about protein pairs that are in contact with each other, and in some cases about the face that is involved in the contact. Subunit proximity indicates whether two proteins are close to each other relative to the size of the assembly, but not necessarily in direct contact. Subunit stoichiometry indicates the number of subunits of a given type that occur in the assembly. Assembly symmetry indicates the symmetry of the arrangement of the subunits in the assembly. *Gray boxes* indicate extreme difficulty in obtaining the corresponding information by a given method. (Sali et al. 2003)

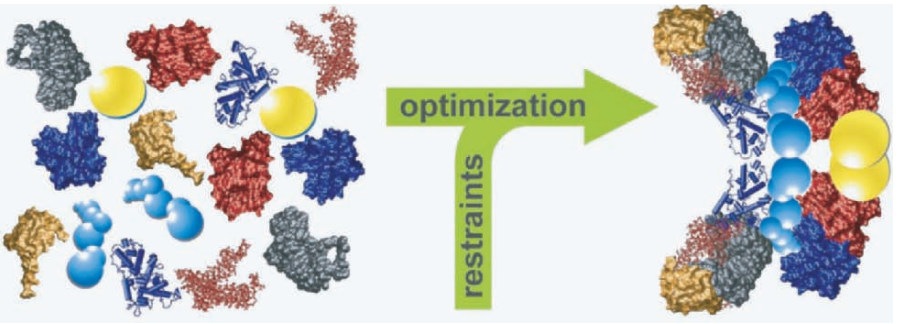


Fig. 3. The scheme that illustrates how the subunits of a hypothetical complex (*left*) may be assembled through optimization with respect to restraints from a variety of methods to obtain the final assembly model (*right*). (Sali et al. 2003)

2000; Zhang et al. 1999) and electron microscopy (Frank 2002). Upper- distance bounds on residues from different proteins may be obtained from NMR spectroscopy (Fiaux et al. 2002) and chemical cross-linking (Rappsilber et al. 2000; Young et al. 2000); information that two proteins bind to each other may be discovered by yeast two-hybrid (Phizicky et al. 2003; Uetz et al. 2000) or micro-calorimetry (Lakey and Raggett 1998) experiments; two proteins can be assigned to be close to each other (relative to the size of the assembly) if they are part of an isolated sub-complex, characterized, for example, by an immuno-purification experiment (Rout et al. 2000; Aebersold and Mann 2003; Phizicky et al. 2003).

To develop a framework for computing the 3D models of a given protein assembly that are consistent with all available information about its composition and structure, we express structure determination of assemblies as an optimization problem (Fig.3). This approach consists of three components (Fig. 4): (1) a representation of the modeled assembly (Fig. 4a): (2) a scoring function consisting of the individual spatial restraints (Fig. 4b): and (3) optimization of the scoring function to obtain the models (Fig. 4 c). The most important aspect of this approach is to accurately capture all available information about the structure of the complex, whether it is high- or low-resolution, experimental, or theoretical. The method should also be capable of calculating all the models that satisfy the input spatial restraints. We illustrate this method by a description of its application to the low-resolution modeling of the configuration of proteins in a given assembly.

2.1 Modeling the Low-Resolution Structures of Assemblies

Some large assemblies, such as the nuclear pore complex, consist predominantly of subunits whose structures have not yet been defined. If comparative

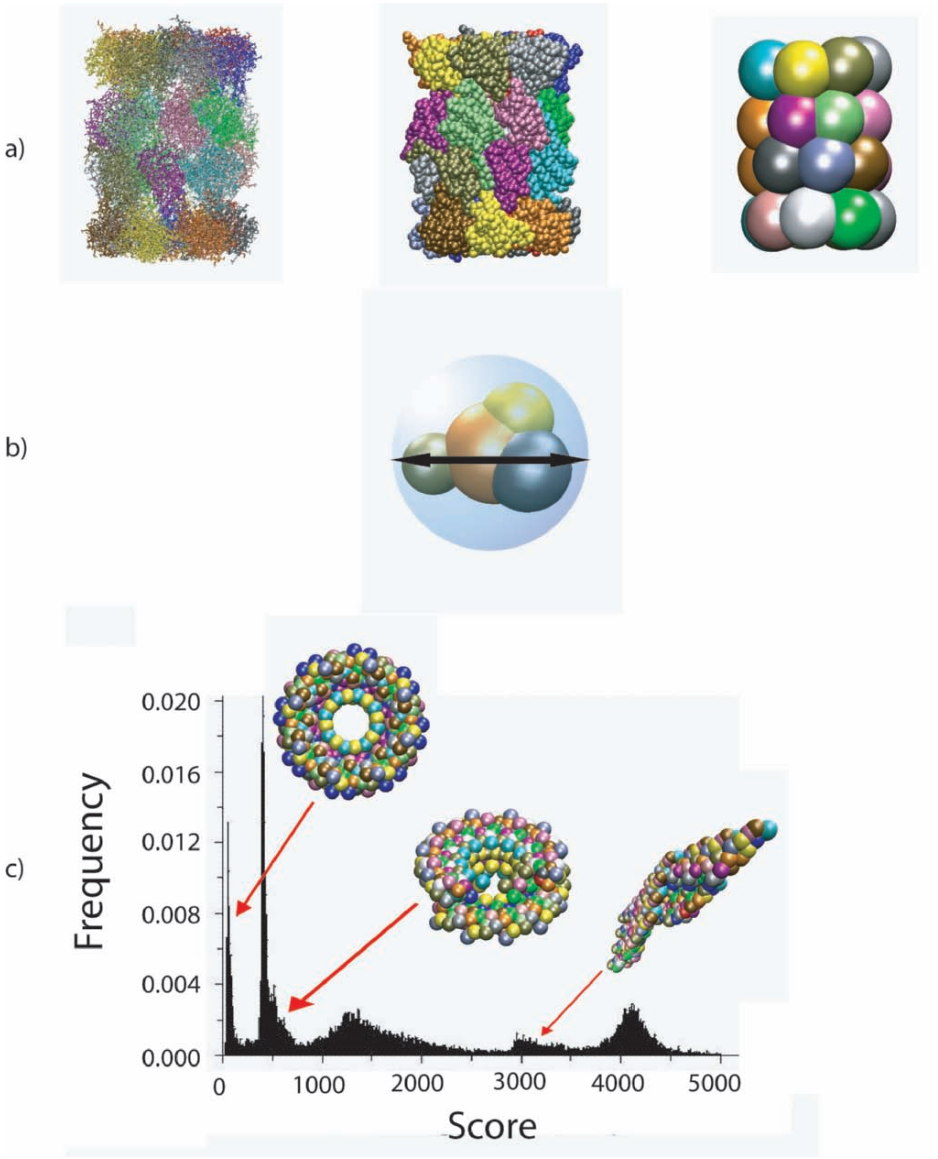


Fig. 4. Modeling of the configuration of proteins in an assembly by satisfaction of spatial restraints. **a** From *left* to *right*, representations of the proteasome assembly of 28 proteins with points per atom, residue and protein, respectively. **b** Derivation of upper distance bounds on all pairs of proteins that have been shown to be a part of the same subcomplex by an affinity chromatography experiment. An estimate for the diameter of the whole subcomplex is needed and can be obtained, for example, from the measured Stokes radius or the total number of residues in the subcomplex. **c** The distribution of an objective function score for many optimized configurations. A desired ring structure is indicated on the *left*, but stochastic optimization that starts from random configurations also results in a variety of other distorted solutions that do not satisfy input restraints

modeling attempts cannot provide atomic structures, such assemblies may be characterized only by low-resolution information about their overall shape and protein–protein proximity. In other words, we can expect to be able to model only the configuration of the proteins in the assembly, not their individual conformations. The following sections outline the three essential aspects of modeling by satisfaction of spatial restraints, introduced above. It has been applied to the low-resolution modeling of the configuration of proteins in the yeast nuclear complex (Alber et al. 2004, in prep.).

2.1.1 Representation of Molecular Assemblies

The system is represented by points that are restrained by spatial restraints. In the absence of any atomic structures, we need to represent each of the assembly proteins as a point. A slightly higher resolution may be achieved by parsing the protein into individual domains, using either bioinformatics tools or biochemical experiments, such as limited proteolysis followed by mass spectroscopy.

2.1.2 Scoring Function Consisting of Individual Spatial Restraints

The most important aspect of low-resolution modeling is to accurately capture all of the experimental and theoretical information about the structure of the modeled assembly. This aim may be achieved by defining the scoring function as a sum of individual spatial restraints.

The restrained spatial features may include distances, angles, and dihedral angles defined by points and gravity centers of sets of points, as well as symmetry between sets of points. The distance restraints are defined based on the available information about the modeled complex. Typical examples are given below.

Excluded Volume Restraints. Lower bounds on protein–protein distances are the sum of the corresponding estimated protein radii (Russel et al. 1997). The radius can be estimated from the number of amino acid residues or from the experimentally determined Stokes radius (Harding and Colfen 1995).

Symmetry Restraints. If EM images and stoichiometry considerations indicate symmetry (Yang et al. 1998), the appropriate result can be achieved by imposing a distance root-mean-square term on the parts of the model that need to have similar conformation or configuration.

Protein Localization Restraints. Immunolabeling experiments (Rout et al. 2000) can be readily expressed as distance restraints on the labeled protein, relative to a reference point such as another labeled protein or the gravity center of the complex. This data can be arrived from superimposition of the individual electron microscopy or tomography images containing the labeled proteins.

Protein Proximity Restraints. "Pullout" experiments (Rout et al. 2000; Aebersold and Mann 2003; Phizicky et al. 2003), chemical cross-linking (Rappsilber et al. 2000; Young et al. 2000), foot-printing (Kiselar et al. 2002), or yeast two-hybrid system assays (Uetz et al. 2000) can be translated into weak upper bounds on the protein-protein distances. Such restraints may also be inferred from a bioinformatics analysis of protein sequences (e.g. an analysis of correlated mutations (Pazos and Valencia 2002)).

Shape Restraints. EM (Frank 1996) and tomography images (Baumeister 2002) may allow defining the volume density map for the complex. The configuration of the proteins in the complex can then be restrained by maximizing the correlation coefficient between the EM map and that implied by a model, similarly to the fitting of higher-resolution atomic models into the EM maps (Roseman 2000; Wriggers et al. 2000; Rossmann et al. 2001; Wriggers and Birmanns 2001).

2.1.3 Optimization of the Scoring Function

An "ensemble" of models that minimize violations of the input restraints may be obtained by optimization of the scoring function. For example, it is possible to start with a random configuration of the proteins, and then apply a combination of the conjugate gradients minimization and simulated annealing with molecular dynamics to the Cartesian coordinates of the points representing the system. Since the optimization is stochastic, a large number of models are generally calculated by starting from a large number of independently generated random configurations (e.g., 100,000). The aim of this sampling is to find all possible models that satisfy the input restraints.

2.1.4 Analysis of the Models

Depending on the resolution of the modeling, a variety of geometrical criteria for comparing two given configurations of points can be used. Examples include the distance root-mean-square deviation that focuses on the protein-protein contacts and a root-mean-square deviation that focuses on the positions of the individual proteins.

Assessing the accuracy of the results is an important and highly non-trivial part of the modeling. There are three conceivable ways of estimating the accuracy of the models, in the absence of a directly determined structure.

1. Similarity among the well scoring models is a necessary, but not sufficient condition for their accuracy. If the well scoring models are not similar to each other, there is not sufficient information in the input restraints to define the configuration of the whole complex.
2. The consistency between the model and the data not used in the model calculation also measures the accuracy of the model. For example, a criterion

similar to the crystallographic free R-factor could be used to assess both the model accuracy and the harmony among the input restraints.
3. The number and properties of the restraints can be correlated with the expected accuracy of the resulting models. Such correlations can be estimated by the use of "toy" models where the native structure of an assembly is known, the restraints are simulated, and their information content is estimated by exhaustive simulation.

3 Comparative Modeling for Structure Determination of Macromolecular Complexes

Comparative modeling can play an important role in the structure determination of large protein assemblies. Due to the progress in structural biology and structural genomics, the structures of the individual subunits of larger assemblies are frequently already known. Additionally, the structures of large assemblies and their constituent parts also tend to be conserved in evolution. Therefore, it is possible to calculate relatively accurate comparative models of the individual subunits that have no available experimental structure. While only ~2 % of known protein sequences have had their structures determined by experiment, comparative modeling can currently be used to predict at least the folds for approximately 30 % of all domains in the known sequences. This indicates that there is a growing need to improve the use of homologous subunit structures in the modeling of protein assemblies. We will now review the comparative modeling method and its limitations, and then continue with its application to the docking of subunit structures into EM maps.

3.1 Automated Comparative Protein Structure Modeling

Comparative modeling consists of four main steps (Marti-Renom et al. 2000): (1) fold the assignment that identifies the similarity between the target sequence of interest and at least one known protein structure (the template); (2) the alignment of the target sequence and the template(s); (3) building a model based on the chosen template(s); and (4) assessing the model for its accuracy. These steps were assembled into a completely automated pipeline (Sanchez and Sali 1998; Eswar et al. 2003). Manual intervention is usually required only in difficult cases. Automation of the procedure makes comparative modeling accessible to both experts and the non-specialists alike and enables the calculation of models for more sequences than is practical by hand. There are a number of servers for automated comparative modeling (http://salilab.org/bioinformatics_resources.shtml). Many of these servers are tested at the bi-annual CAFASP meetings (Fischer et al. 2001) and continually by the LiveBench (Bujnicki et al. 2001) and EVA (Eyrich et al. 2001; Koh et al.

2003) web servers for assessment of automated protein structure prediction methods. We will now describe ModPipe, which is our version of an automated scheme for large-scale comparative modeling (Sanchez and Sali 1998; Eswar et al. 2003).

ModPipe is an automated software pipeline for comparative protein structure modeling that can calculate comparative models for a large number of protein sequences, using many different template structures and sequence-structure alignments (Fig. 5; Sanchez and Sali 1998; Marti-Renom et al. 2000; Pieper et al. 2002; Eswar et al. 2003). Sequence-structure matches are established by aligning the PSI-BLAST sequence profile (Altschul et al. 1997) of the target sequence against each of the template sequences extracted from the Protein Data Bank (PDB) (Berman et al. 2002), as well as by scanning the target sequence against a database of the template profiles (Schaffer et al. 1999). Significant alignments covering distinct regions of the target sequence are chosen for modeling. Models are calculated for each of the sequence-struc-

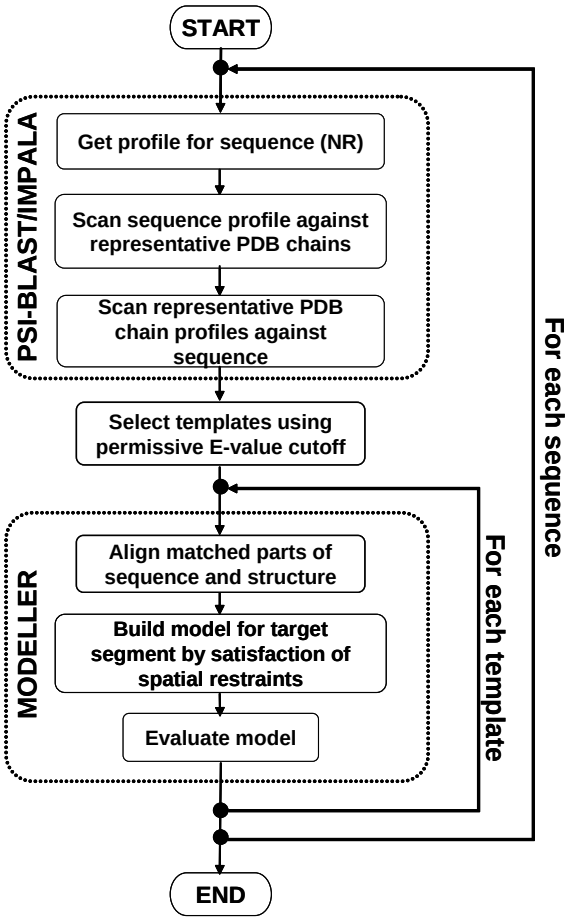


Fig. 5. Flowchart of ModPipe, a large-scale protein structure modeling pipeline (Eswar et al. 2003). See text for details

ture matches using MODELLER, which implements comparative protein structure modeling by satisfaction of spatial restraints (Sali and Blundell 1993). The resulting models are then evaluated by a composite model quality criterion that depends on the compactness of a model, the sequence identity of the sequence-structure match, and statistical energy Z-scores (Melo et al. 2002).

The thoroughness of a search for the best model is modulated by a number of parameters, including the E-value thresholds for identifying useful sequence-structure relationships and the degree of conformational sampling given a sequence-structure alignment. The validity of sequence-structure relationships is not pre-judged at the detection of the fold, but is obtained after the construction of the model and its subsequent evaluation. This approach enables a thorough exploration of fold assignments, sequence-structure alignments, and conformations, with the aim of finding the model with the best model quality score.

ModPipe has been used to calculate models for all sequences in the SwissProt database (Boeckmann et al. 2003) with detectable similarity to a known protein structure. The results are available through ModBase, a relational database that allows flexible and efficient querying of its contents (http://salilab.org/modbase) (Pieper et al. 2002). Currently, ModBase contains models for domains in 415,937 out of 733,239 (~57 %) unique protein sequences found in SwissProt (March 2002). Most of the models are based on less than 30 % sequence identity to the closest structure and cover only a single domain in the protein sequence, corresponding on average to one third of the whole protein. The automation and archival of such comparative models reflect the ultimate goal of the structural genomics initiative (Sali 1998; Sanchez et al. 2000; Vitkup et al. 2001; Burley and Bonanno 2002).

3.2 Accuracy of Comparative Models

The accuracy of comparative models is most easily quantified by the extent of sequence similarity between the sequence and the known structure (Chothia and Lesk 1986; Sanchez and Sali 1998; Marti-Renom et al. 2000; Baker and Sali 2001). Accuracy of a model tends to increase with the target-template sequence identity (Fig. 6). In general, models based on alignments with more than 40 % sequence identity frequently tend to have close to 80 % of their backbone atoms superposable with their actual structures with an RMS error less than 3.5 Å (Sanchez and Sali 1998).

High accuracy comparative models are based on more than 50 % sequence identity to their templates (Marti-Renom et al. 2000; Fiser and Sali 2001). They tend to have approximately 1 Å RMS error for the main-chain atoms, which is comparable to the accuracy of a medium resolution nuclear magnetic resonance (NMR) spectroscopy structure or a low-resolution X-ray structure. The

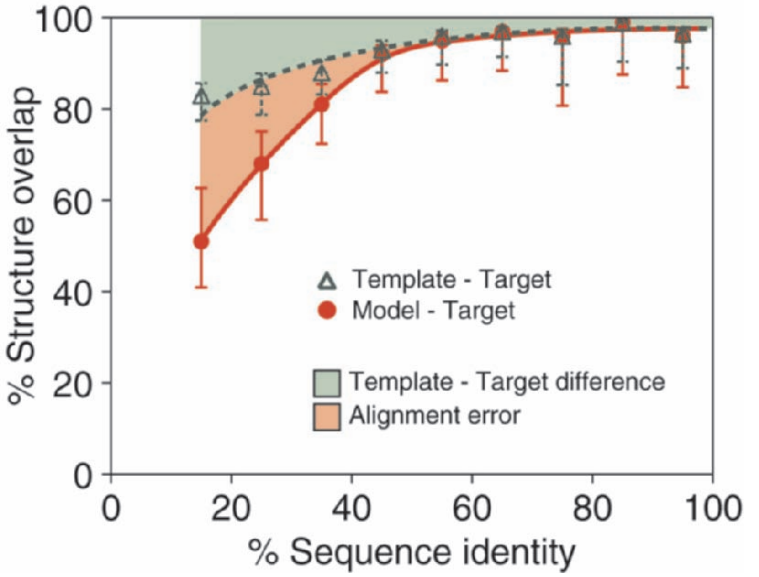


Fig. 6. The relationship between the accuracy of a reliable model and the percentage sequence identity to the template. The overlaps of an experimentally determined protein structure with its model (*red continuous line*) and with a template on which the model was based (*green dashed line*) are shown as a function of the target-template sequence identity. The structure overlap is defined as the fraction of the equivalent C^{α} atoms. For comparison of the model with the actual structure (*filled circles*), two C^{α} atoms were considered equivalent if they were within 3.5 Å of each other and belonged to the same residue. For comparison of the template structure with the actual target structure (*open triangles*), two C^{α} atoms were considered equivalent if they were within 3.5 Å after alignment and rigid-body superposition. The *points* correspond to the median values, and the *error bars* in the positive and negative directions correspond to the average positive and negative differences from the median, respectively. (Sanchez and Sali 1998)

errors are mostly mistakes in side-chain packing, small shifts or distortions of the core main-chain regions, and occasionally larger errors in loops. Medium accuracy comparative models are usually based on 30–50 % sequence identity. They tend to have approximately 90 % of the main-chain modeled with 1.5 Å RMS error. There are more frequent side-chain packing, core distortion, and loop modeling errors, and there are occasional alignment mistakes. And finally, low accuracy comparative models are generally based on less than 30 % sequence identity. The alignment errors increase rapidly below 30 % sequence identity and become the most significant origin of errors in comparative models. In addition, when a model is based on an almost insignificant alignment to a known structure, it may also have an entirely incorrect fold.

3.3 Prediction of Model Accuracy

The folds of the comparative models in ModPipe are evaluated by a composite scoring function (Melo et al. 2002; John and Sali 2003):

$$\text{GA341} = 1 - \left[\cos\left(\text{sequence\_identity}\right)\right]^{(\text{compactness}+\text{sequence\_identity})/\exp(\text{z-score})}$$

Sequence identity is the fraction of positions with identical residues in the target-template alignment. Structural compactness is the ratio between the sum of the standard volumes of the amino acid residues in the protein and the volume of the sphere with the diameter equal to the largest dimension of the model. The Z-score is calculated for the combined statistical potential energy of a model, using the mean and standard deviation of the 200 random sequences with the same composition and structure as the model (Melo et al. 2002). The combined statistical potential energy of a model is the sum of the solvent accessibility terms for all C^{β} atoms and distance-dependent terms for all pairs of C^{α} and C^{β} atoms. The solvent accessibility term for a C^{β} atom depends on its residue type and the number of other C^{β} atoms within 10 Å; the non-bonded terms depend on the atom and residue types spanning the distance, the distance itself, and the number of residues separating the distance-spanning atoms in sequence. These potential terms reflect the statistical preferences observed in 760 non-redundant proteins of known structure. The GA341 scoring function was evolved by a genetic algorithm that explored many combinations of a variety of mathematical functions and model features, to optimize the discrimination between good and bad models in a training set of models. The GA341 score ranges from 0 for models that tend to have an incorrect fold to 1 for models that tend to be comparable to at least low-resolution X-ray structures. GA341 scores greater than 0.7 indicate a correct fold with more than 35 % of the backbone atoms superposable to those better than 3.5 Å.

3.4 Docking of Comparative Models into Low-Resolution Cryo-EM Maps

The usefulness of comparative models is limited by their accuracy and the resolution of the density map; similar limitations may also apply to the experimentally determined subunit structures, due to the induced fit. It is usually possible to generate a set of comparative models that are based on alternate alignments, templates, and domain orientations; some of these models may be more accurate than others. The best subunit models and their positions in the complex may then be identified by manual or automated docking of the alternate models into the electron density data from electron microscopy or low-resolution X-ray crystallography. Ultimately, the best protein assembly model may be obtained by satisfying simultaneously the homology-derived

restraints on the individual subunits and shape restraints on the whole complex.

The useful accuracy of comparative models for docking into the EM density map varies with the resolution of the map (Fig. 7). At resolutions worse than 10 Å, only the shape and size of a subunit can be identified and models based on different but related template structures could be chosen for the docking without loss of accuracy. The different template structures could account for variable conformations of the subunit (e.g., open/closed forms) or different orientations of the constituent rigid bodies. At medium resolutions, between 5 and 10 Å, it is usually possible to discern the positions of secondary structural elements and the domain structure of the components. In these cases, models calculated with one or more templates but with several variations in the alignments to reposition secondary structures and loops could be useful for identifying the optimal fit of the structure in the density map. Additionally, loop regions can be independently optimized to account for differences in conformations between the model and the observed density. The backbone trace as well as the positions and boundaries of the secondary structure elements can be identified more accurately at even higher resolutions (~5 Å). Models of at least medium accuracy (Sect. 3.2) are required for docking into maps at this resolution. In addition to the use of multiple templates, multiple models could also be sampled by an optimization scheme that

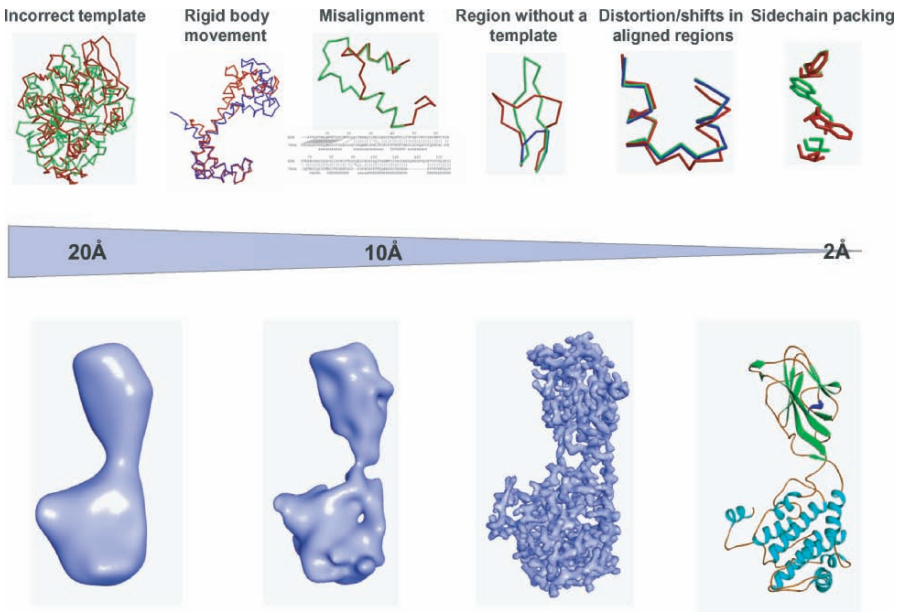


Fig. 7. Usefulness of comparative models for docking into EM electron density (the maps are courtesy of Dr. Wah Chiu). Examples of errors in comparative models that can be identified at various resolutions of the density maps are indicated. See text for details

explores the conformational degrees of freedom for the backbone and side-chains based on a single target-template alignment.

3.5 Example 1: A Partial Molecular Model of the 80S Ribosome from Saccharomyces cerevisiae

As an illustration of the integrated strategies introduced earlier, we now describe the fitting of comparative protein structure models into electron density maps of the whole yeast (Spahn et al. 2001) and *E. coli* ribosomes (Spahn et al. 2001; Gao et al. 2003). Partial or complete molecular models of the ribosomes are obtained by the use of information from two sources, experimental low-resolution (10 Å) cryo-EM maps and all-atom comparative models for the individual RNA and protein components of the ribosomes.

Ribosomes are macromolecular machines responsible for protein biosynthesis in the cell and consist of ribosomal RNA (rRNA) molecules and 50–80 ribosomal proteins. They are made up of two subunits, a small subunit responsible for decoding in protein translation (i.e., selection of cognate tRNA) (Carter et al. 2000) and a large subunit, primarily responsible for the catalytic activity (i.e., peptidyl transferase) (Nissen et al. 2000). Atomic resolution X-ray structures are available for the small 30S subunit from the thermophile bacterium *Thermus thermophilus* (Schluenzen et al. 2000; Wimberly et al. 2000) as well as the large 50S subunit from the halophile archaebacterium *Haloarcula marismortui* (Ban et al. 2000) and mesophilic eubacterium *Deinococus radiodurans* (Harms et al. 2001). While a relatively large amount of high-resolution structural information is available for prokaryotic ribosomes or their individual subunits, there is only sparse data for their eukaryotic counterparts. Fortunately, the eukaryotic ribosomal RNA and proteins are evolutionarily related to their prokaryotic homologues. Despite the different sizes of the rRNA, additional proteins, and more complex functions of the eukaryotic ribosome, it can be anticipated that the overall spatial arrangement of the subunits and the fundamental process of protein biosynthesis are similar to those in the prokaryotes.

To gain structural insights into the machinery of eukaryotic ribosomes, we combined a low-resolution cryo-EM map (~15 Å) of the *Saccharomyces cerevisiae* ribosome with comparative modeling and docking (Spahn et al. 2001). The yeast ribosomal complex is made up of a 40S small subunit, composed of a 1798 nucleotide (nt) long 18S rRNA and 32 ribosomal proteins, and a large 60S subunit composed of a 25S rRNA (3392 nt), 5.8S rRNA (158 nt), 5S rRNA (121 nt), and 45 ribosomal proteins (Spahn et al. 2001). To facilitate the docking, the map of the 80S ribosome was computationally separated into the protein and RNA parts, using a method that takes into account the differences in the density distribution of RNA and proteins, as well as the molecular masses and contiguity constraints (Spahn et al. 2000).

rRNA models from the crystal structures of the 30S subunit from *T. thermophilus* (Wimberly et al. 2000) and the 50S subunit from *H. marismortui* (Ban et al. 2000) were fitted into the resulting maps for the small subunit rRNA and large subunit rRNA of yeast, respectively. Where necessary, the X-ray models were modified by moving the non-fitting parts (e.g. helices) as rigid bodies relative to the rest of the model.

Comparative models for the yeast ribosomal proteins were constructed using ModPipe (Sanchez and Sali 1998; Spahn et al. 2001) and are available through ModBase (http://salilab.org/modbase). Structural templates used to calculate the models consisted of all the individual chains from structures in PDB (as of September 2000), clustered so that the sequences of no two chains

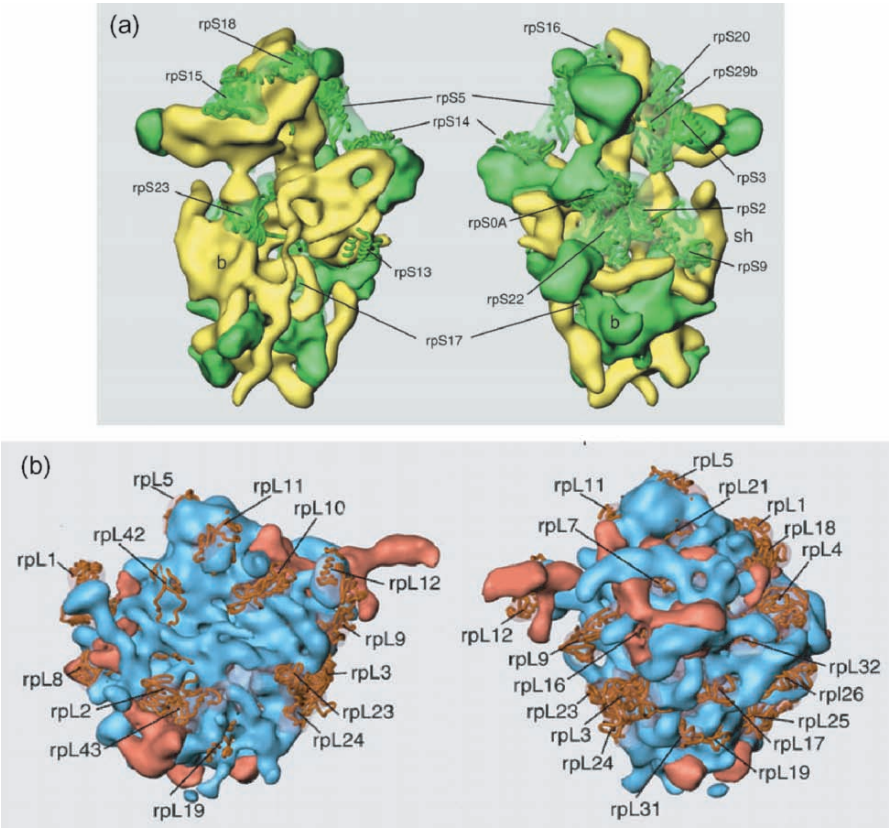


Fig. 8. Structures of the **a** 40S small subunit and **b** 60S large subunit of the yeast ribosome. The RNA and protein partitions are shown in *yellow* and *turquoise* respectively for the small subunit; they are depicted as *blue* and *orange* respectively for the large subunit. Wherever comparative models could be docked into the map, the protein partition is shown transparently. Therefore, solid parts of the protein partition predict the position of additional proteins with no homologous counterparts in prokaryotes. (Spahn et al. 2000)

from any two clusters were more than 95 % identical. In addition, the structures of the small subunit from *T. thermophilus* (PDB code: 1FJF) and the large subunit from *H. morismortui* (PDB code: 1FKF) were considered as separate sets of templates. In total, comparative models were obtained for 43 yeast ribosomal proteins; 15 for the 40S subunit (Fig. 8a) and 28 for the 60S subunit (Fig. 8b). The models were derived from alignments with sequence identities in the range of 20–56 % (with an average of 32 %) and E-values better than 0.0001. The coverage of the models (fraction of the yeast ribosomal sequence modeled) ranges between 34–99 % (with an average of 75 %). Docking of atomic models into the cryo-EM density map was done manually using program O (Jones et al. 1991).

The composite map, consisting of docked RNA and comparative models of proteins into the 15.4 Å cryo-EM map, provides for the structural interpretation of the eukaryotic ribosome complex. The common core of the eukaryotic ribosome was found to agree well with X-ray structures of the bacterial and archaebacterial subunits. It reinforces the notion that the fundamental mechanism of protein synthesis is highly conserved throughout all kingdoms. The differences in the structures of the prokaryotic and eukaryotic ribosomes could be localized to regions in the density map corresponding to either yeast proteins without homologous counterparts or those with additional domains. These differences occur mainly on the solvent exposed faces of the subunits, conserving the core of the ribosome. It was also found that the inter-subunit interactions, important for communication between the subunits, and the ribosome-tRNA interactions were largely conserved. Additionally, the structure enabled the identification of four new protein-protein contacts. For more information, see references (Spahn et al. 2001; Beckmann et al. 2001).

3.6 Example 2: A Molecular Model of the E. coli 70S Ribosome

The aim of this study was to capture the dynamic features of the ribosome, the 'ratchet-like' inter-subunit motion, by trapping functionally meaningful states by cryo-EM (Gao et al. 2003). The limited resolution of the cryo-EM maps was overcome by docking comparative models of rRNA and proteins into the maps of the different states of the ribosome: (1) a 11.5-Å map (Gabashvili et al. 2000) of the control, an initiation-like complex with fMet-tRNA$_f^{Met}$ at the P site (Malhotra et al. 1998); and (2) a 12.3-Å map of the EF-G·GTP-bound complex (i.e. a ribosome complex with EF-G in the presence of a non-hydrolysable GTP analog) (Frank and Agrawal 2000). The *E. coli* 70S ribosome consists of two subunits: the 30S subunit, comprising 16S rRNA (1542 nt) and 21 proteins, and the 50S subunit, comprising 23S rRNA (2904 nt), 5S rRNA (120 nt), and 36 proteins. The models of *E. coli* 23S rRNA and 5S rRNA were generated from the crystal structure of *H. marismortui* (PDB code 1FFK) (Ban et al, 2000), while the model of *E. coli* 16S rRNA was generated from the crystal

structure of *T. thermophilus* (PDB code 1IBL) (Ogle et al. 2001) using the molecular modeling package Insight II (Accelrys Inc. Insight II 2003).

Models for the *E. coli* ribosomal proteins were calculated by ModPipe as described earlier. The crystal structures of the proteins from the small subunit of *T. thermophilus* (PDB code 1FJG) (Carter et al. 2000) were chosen as the structural templates to model 19 proteins of the 30S small subunit (S2-S20) of *E. coli*. For proteins of the 50S subunit, 29 out of the 36 *E. coli* proteins were modeled based on the crystal structures of *H. marismortui* (PDB code 1JJ2) (Klein et al. 2001), *D. radiodurans* (1LNR) (Harms et al. 2001), and *T. thermophilus* (1GIY; L9, L25) (Yusupov et al. 2001).

The starting models of the whole ribosome were built by manually docking the individual rRNA and protein models as rigid bodies into the cryo-EM density maps using the interactive program O (Jones et al. 1991). The initial positions of each of the rRNA structures and those of the proteins were taken from the corresponding positions of the template crystal structures. The program RSRef (Chapman 1995), a real-space refinement module for the TNT program (Tronrud 1997), was then employed for automatically and simultaneously refining both the stereochemistry and the fit of the atomic structures to the density map. Since the resolutions of the experimental density maps are not suitable for refinement of independent atoms, a multi-rigid-body refinement was employed.

A comparison of the two resulting atomic models revealed that the ribosome changes from a compact structure in the initiation-like form to a looser one in the EF-G bound form. This change is coupled with the rearrangement of many of the proteins. Furthermore, it could be seen that in contrast to the unchanged inter-subunit bridges formed wholly by RNA, the bridges involving ribosomal proteins undergo large conformational changes following the ratchet-like motion. Observations suggested an important role of ribosomal proteins in facilitating the dynamics of translation.

4 Conclusions

We are now poised to integrate structural information gathered at multiple levels of the biological hierarchy – from atoms to cells – into a common framework. The goal is a comprehensive description of the multitude of interactions between molecular entities, which in turn is a prerequisite for the discovery of general structural principles that underlie all cellular processes. In contrast to structure determination of individual proteins, structural characterization of macromolecular assemblies usually requires diverse sources of information (Sali et al. 2003). This information may vary greatly in terms of its accuracy and resolution, and includes data from both experimental and computational methods, such as X-ray crystallography, NMR spectroscopy, electron microscopy, chemical cross-linking, affinity purification, yeast two-

hybrid system experiments, calorimetry, computational docking, and bioinformatics analysis of protein sequences and structures. Structural genomics will bring us closer to a comprehensive dictionary of proteins in the foreseeable future, while electron microscopy techniques and other approaches will allow us to assemble proteins into complexes. A comprehensive description of large complexes will generally require the use of a number of experimental methods, underpinned by a variety of theoretical approaches to maximize efficiency, completeness, accuracy, and resolution of the experimental determination of assembly composition and structure. In conjunction with the non-invasive 3D imaging of whole cells, these approaches might ultimately enable us to read the molecular book of the cell.

Acknowledgements. We are grateful to Fred Davies, Damien Devos, Dimitry Korkin, and Maya Topf for discussions about the modeling of assembly structures. This review is based on the following publications: Sali et al. (2003), Spahn et al. (2001), Gao et al. (2003), Eswar et al. (2003), and Marti-Renom et al. (2000).

References

AbbottA (2002) The society of proteins. Nature 417:894–896

AebersoldR, MannM (2003) Mass spectrometry-based proteomics. Nature 422:198–207

Alberts B (1998). The cell as a collection of protein machines: preparing the next generation of molecular biologists. Cell 92:291–294

Aloy P, Ciccarelli FD, Leutwein C, Gavin AC, Superti-Furga G, Bork P, Bottcher B, Russell RB (2002) A complex prediction: three-dimensional model of the yeast exosome. EMBO Rep 3:628–635

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Baker D, Sali A (2001) Protein structure prediction and structural genomics. Science 294:93–96

Ban N, Nissen P, Hansen J, Moore PB, Steitz TA (2000) The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. Science 289:905–920

Baumeister W (2002) Electron tomography: towards visualizing the molecular organization of the cytoplasm. Curr Opin Struct Biol 12:679–684

Beckmann R, Spahn CM, Eswar N, Helmers J, Penczek PA, Sali A, Frank J, Blobel G (2001) Architecture of the protein-conducting channel associated with the translating 80S ribosome. Cell 107:361–372

Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, Burkhardt K, Feng Z, Gilliland GL, Iype L, Jain S., Fagan P, Marvin J, Padilla D, Ravichandran V, Schneider B, Thanki N, Weissig H, Westbrook JD, Zardecki C (2002) The protein data bank. Acta Crystallogr D Biol Crystallogr 58:899–907

Blundell TL, Sibanda BL, Sternberg MJ, Thornton JM (1987) Knowledge-based prediction of protein structures and the design of novel molecules. Nature 326: 347–352

Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M (2003). The SWISS-PROT

protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res 31: 365–370

Braig K, Otwinowski Z, Hegde R, Boisvert DC, Joachimiak A, Horwich AL, Sigler PB (1994). The crystal structure of the bacterial chaperonin GroEL at 2.8 Å. Nature 371: 578–586

Bujnicki JM, Elofsson A, Fischer D, Rychlewski L (2001). LiveBench-2: large-scale automated evaluation of protein structure prediction servers. Proteins (Suppl 5):184–191

Burley SK, Bonanno JB (2002) Structural genomics of proteins from conserved biochemical pathways and processes. Curr Opin Struct Biol 12: 383–391

Carter AP, Clemons WM, Brodersen DE, Morgan-Warren RJ, Wimberly BT, Ramakrishnan V (2000). Functional insights from the structure of the 30S ribosomal subunit and its interactions with antibiotics. Nature 407: 340–348

Chacon P, Wriggers W (2002). Multi-resolution contour-based fitting of macromolecular structures. J Mol Biol 317:375–384

Chapman MS (1995) Restrained real-space macromolecular atomic refinement using a new resolution-dependent electron density function. Acta Crystallogr A A51:69–80

Chiu W, Baker ML, Jiang W, Zhou ZH (2002) Deriving folds of macromolecular complexes through electron cryomicroscopy and bioinformatics approaches. Curr Opin Struct Biol 12:263–269

Chothia C, Lesk AM (1986) The relation between the divergence of sequence and structure in proteins. EMBO J 5:823–826

Courey AJ (2001) Cooperativity in transcriptional control. Curr Biol 11:R250-R252

Cramer P, Bushnell DA, Kornberg RD (2001) Structural basis of transcription: RNA polymerase II at 2.8 angstrom resolution. Science 292:1863–1876

Eswar N, John B, Mirkovic N, Fiser A, Ilyin VA, Pieper U, Stuart AC, Marti-Renom MA, Madhusudhan MS, Yerkovich B, Sali A (2003) Tools for comparative protein structure modeling and analysis. Nucleic Acids Res 31:3375–3380

Eyrich VA, Marti-Renom MA, Przybylski D, Madhusudhan MS, Fiser A, Pazos F, Valencia A, Sali A, Rost B (2001) EVA: continuous automatic evaluation of protein structure prediction servers. Bioinformatics 17:1242–1243

Fiaux J, Bertelsen EB, Horwich AL, Wuthrich K (2002) NMR analysis of a 900 K GroEL GroES complex. Nature 418:207–211

Fischer D, Elofsson A, Rychlewski L, Pazos F, Valencia A, Rost B, Ortiz AR, Dunbrack RL Jr (2001) CAFASP2: the second critical assessment of fully automated structure prediction methods. Proteins (Suppl 5):171–183

Fiser A, Sali A (2001) MODELLER: generation and refinement of homology models. In: Carter CW, Sweet RM (eds) Methods in enzymology. Academic Press, New York

Frank J (1996) Three-dimensional electon microscopy of macromolecular assemblies. Academic Press, London

Frank J (2002) Single-particle imaging of macromolecules by cryo-electron microscopy. Annu Rev Biophys Biomol Struct 31:303–319

Frank J,Agrawal RK (2000) A ratchet-like inter-subunit reorganization of the ribosome during translocation. Nature 406:318–322

Gabashvili IS, Agrawal RK, Spahn CM, Grassucci RA, Svergun DI, Frank J, Penczek P (2000) Solution structure of the E. coli 70S ribosome at 11.5 A resolution. Cell 100:537–549

Gao H, Sengupta J, Valle M, Korostelev A, Eswar N, Stagg SM, Van Roey P, Agrawal RK, Harvey SC, Sali A, Chapman MS, and Frank J (2003) Study of the structural dynamics of the E coli 70S ribosome using real-space refinement. Cell 113:789–801

Goldstein LS, Yang Z (2000) Microtubule-based transport systems in neurons: the roles of kinesins and dyneins. Annu Rev Neurosci 23:39–71

Gomperts SN (1996) Clustering membrane proteins: It's all coming together with the PSD-95/SAP90 protein family. Cell 84:659–662

Goto NK, Zor T, Martinez-Yamout M, Dyson HJ, Wright PE (2002) Cooperativity in transcription factor binding to the coactivator CREB-binding protein (CBP) The mixed lineage leukemia protein (MLL) activation domain binds to an allosteric site on the KIX domain. J Biol Chem 277: 43168–43174

Grakoui A, Bromley SK, Sumen C, Davis MM, Shaw AS, Allen PM, Dustin ML (1999) The immunological synapse: a molecular machine controlling T cell activation. Science 285:221–227

Greer J (1990) Comparative modeling methods: application to the family of the mammalian serine proteases. Proteins 7:317–334

Grimes J, Basak AK, Roy P, Stuart D (1995) The crystal structure of bluetongue virus VP7. Nature 373:167–170

Harding SE, Colfen H (1995) Inversion formulae for ellipsoid of revolution macromolecular shape functions. Anal Biochem 228:131–142

Harms J, Schluenzen F, Zarivach R, Bashan A, Gat S, Agmon I, Bartels H, Franceschi F, Yonath A (2001) High resolution structure of the large ribosomal subunit from a mesophilic eubacterium. Cell 107:679–688

John B, Sali A (2003). Comparative protein structure modeling by iterative alignment, model building and model assessment. Nucleic Acids Res 31:3982–3992

Jones TA, Zou JY, Cowan SW, Kjeldgaard (1991) Improved methods for building protein models in electron density maps and the location of errors in these models. Acta Crystallogr A 47 (Pt 2):110–119

Kiselar JG, Maleknia SD, Sullivan M, Downard KM, Chance MR (2002) Hydroxyl radical probe of protein surfaces using synchrotron X-ray radiolysis and mass spectrometry. Int J Radiat Biol 78:101–114

Klein DJ, Schmeing TM, Moore PB, Steitz TA (2001) The kink-turn: a new RNA secondary structure motif. EMBO J 20:4214–4221

Koh IYY, Eyrich VA, Marti-Renom MA, Przybylski D, Madhusudhan MS, Eswar N, Grana O, Pazos ,F, Valencia A, Sali A, Rost B (2003) EVA: evaluation of protein structure prediction servers. Nucleic Acids Res (in press)

Kumar A, Snyder M (2002). Protein complexes take the bait. Nature 415:123–124

Lakey JH, Raggett EM (1998). Measuring protein-protein interactions. Curr Opin Struct Biol 8:119–123

Lowe J, Stock D, Jap B, Zwickl P, Baumeister W, Huber R (1995). Crystal structure of the 20S proteasome from the archaeon T. acidophilum at 3.4 A resolution. Science 268:533–539

Malhotra A, Penczek P, Agrawal RK, Gabashvili IS, Grassucci RA, Junemann R, Burkhardt N, Nierhaus KH, Frank J (1998) Escherichia coli 70 S ribosome at 15 A resolution by cryo-electron microscopy: localization of fMet-tRNAfMet and fitting of L1 protein. J Mol Biol 280:103–116

Malhotra A, Tan RK, Harvey SC (1990) Prediction of the three-dimensional structure of *Escherichia coli* 30S ribosomal subunit: a molecular mechanics approach. Proc Natl Acad Sci USA 87:1950–1954

Marti-Renom MA, Stuart AC, Fiser A, Sanchez R, Melo F, Sali A (2000) Comparative protein structure modeling of genes and genomes. Annu Rev Biophys Biomol Struct 29:291–325

Melo F, Sanchez R, Sali A (2002) Statistical potentials for fold assessment. Protein Sci 11:430–448

Murakami KS, Darst SA (2003) Bacterial RNA polymerases: the wholo story. Curr Opin Struct Biol 13:31–39

Murzin AG, Bateman A (2001) CASP2 knowledge-based approach to distant homology recognition and fold prediction in CASP4. Proteins (Suppl 5):76–85
Nissen P, Hansen J, Ban N, Moore PB, Steitz TA (2000) The structural basis of ribosome activity in peptide bond synthesis. Science 289:920–930
Nogales E (2000) Recent structural insights into transcription preinitiation complexes. J Cell Sci 113 (Pt 24):4391–4397
Nogales E, Grigorieff N (2001) Molecular Machines: putting the pieces together. J Cell Biol 152:F1–10
Nogales E, Wolf SG, Downing KH (1998) Structure of the alpha beta tubulin dimer by electron crystallography. Nature 391:199–203
Noji H, Yoshida M (2001) The rotary machine in the cell, ATP synthase. J Biol Chem 276: 1665–1668
Oda Y, Saeki K, Takahashi Y, Maeda T, Naitow H, Tsukihara T, Fukuyama K (2000) Crystal structure of tobacco necrosis virus at 2.25 2.25 Å resolution. J Mol Biol 300:153–169
Ogle JM, Brodersen DE, Clemons WM Jr, Tarry MJ, Carter AP, Ramakrishnan V (2001) Recognition of cognate transfer RNA by the 30S ribosomal subunit. Science 292:897–902
Pazos F, Valencia A (2002) In silico two-hybrid system for the selection of physically interacting protein pairs. Proteins 47:219–227
Phizicky E, Bastiaens PI, Zhu H, Snyder M, Fields S (2003) Protein analysis on a proteomic scale. Nature 422:208–215
Pieper U, Eswar N, Stuart AC, Ilyin VA, Sali A (2002) MODBASE, a database of annotated comparative protein structure models. Nucleic Acids Res 30:255–259
Rappsilber J, Siniossoglou S, Hurt EC, Mann M (2000) A generic strategy to analyze the spatial organization of multi-protein complexes by cross-linking and mass spectrometry. Anal Chem72:267–275
Roseman AM (2000) Docking structures of domains into maps from cryo-electron microscopy using local correlation. Acta Crystallogr D Biol Crystallogr 56 (Pt 10): 1332–1340
Rossmann MG, Bernal R, Pletnev SV (2001) Combining electron microscopic with x-ray crystallographic structures. J Struct Biol136:190–200
Rout MP, Aitchison JD, Suprapto A, Hjertaas K, Zhao Y, Chait BT (2000) The yeast nuclear pore complex: composition, architecture, and transport mechanism. J Cell Biol 148:635–651
Russel M, Linderoth NA, Sali A (1997) Filamentous phage assembly: variation on a protein export theme. Gene 192:23–32
Sali A (1998) 100,000 protein structures for the biologist. Nat Struct Biol 5:1029–1032
Sali A, Blundell TL (1993)Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 234:779–815
Sali A, Glaeser R, Earnest T, Baumeister W (2003) From words to literature in structural proteomics. Nature 422:216–225
Sanchez R, Pieper U, Melo F, Eswar N, Marti-Renom MA, Madhusudhan MS, Mirkovic N, Sali A (2000) Protein structure modeling for structural genomics. Nat Struct Biol 7 (Suppl):986–990
Sanchez R, Sali A (1998) Large-scale protein structure modeling of the *Saccharomyces cerevisiae* genome. Proc Natl Acad Sci USA 95:13597–13602
Sauder JM, Dunbrack RL Jr (2000) Genomic fold assignment and rational modeling of proteins of biological interest. Proc Int Conf Intell Syst Mol Biol 8:296–306
Schaffer AA, Wolf YI, Ponting CP, Koonin EV, Aravind L, Altschul SF (1999) IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. Bioinformatics 15:1000–1011

Schluenzen F, Tocilj A, Zarivach R, Harms J, Gluehmann M, Janell D, Bashan A, Bartels H, Agmon I, Franceschi F, Yonath A (2000) Structure of functionally activated small ribosomal subunit at 3.3 angstroms resolution. Cell 102:615–623

Spahn CM, Beckmann R, Eswar N, Penczek PA, Sali A, Blobel G, Frank J (2001). Structure of the 80S ribosome from Saccharomyces cerevisiae–tRNA-ribosome and subunit-subunit interactions. Cell 107:373–386

Spahn CM, Penczek PA, Leith A, Frank J (2000) A method for differentiating proteins from nucleic acids in intermediate-resolution density maps: cryo-electron microscopy defines the quaternary structure of the *Escherichia coli* 70S ribosome. Structure Fold Des 8:937–948

Tronrud DE (1997) TNT refinement package. Methods Enzymol. 277:306–319

Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, Qureshi-Emili A, Li Y, Godwin B, Conover D, Kalbfleisch T, Vijayadamodar G, Yang M, Johnston M, Fields S, Rothberg JM (2000) A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. Nature 403:623–627

Vale RD (2003) The molecular motor toolbox for intracellular transport. Cell 112:467–480

Vale RD,Milligan RA (2000) The way things move: looking under the hood of molecular motor proteins. Science 288:88–95

Vitkup D, Melamud E, Moult J, Sander C (2001) Completeness in structural genomics. Nat Struct Biol 8:559–566

Volkmann N, Hanein D (1999). Quantitative fitting of atomic models into observed densities derived by electron microscopy. J Struct Biol 125:176–184

Volkmann N, Hanein D, Ouyang G, Trybus KM, DeRosier DJ, Lowey S (2000) Evidence for cleft closure in actomyosin upon ADP release. Nat Struct Biol 7: 1147–1155

von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P (2002) Comparative assessment of large-scale data sets of protein-protein interactions. Nature 417: 399–403

Wimberly BT, Brodersen DE, Clemons WM Jr, Morgan-Warren RJ, Carter AP, Vonrhein C, Hartsch T, Ramakrishnan V (2000) Structure of the 30S ribosomal subunit. Nature 407:327–339

Wriggers W, Agrawal RK, Drew DL, McCammon A, Frank J (2000) Domain motions of EF-G bound to the 70S ribosome: insights from a hand-shaking between multi-resolution structures. Biophys J 79:1670–1678

Wriggers W, Birmanns S (2001) Using situs for flexible and rigid-body fitting of multiresolution single-molecule data. J Struct Biol 133:193–202

Yang Q, Rout MP, Akey CW (1998) Three-dimensional architecture of the isolated yeast nuclear pore complex: functional and evolutionary implications. Mol Cell 1:223–234

Young MM, Tang N, Hempel JC, Oshiro CM, Taylor EW, Kuntz ID, Gibson BW, Dollinger G (2000) High throughput protein fold identification by using experimental constraints derived from intramolecular cross-links and mass spectrometry. Proc Natl Acad Sci USA 97:5802–5806

Yusupov MM, Yusupova GZ, Baucom A, Lieberman K, Earnest TN, Cate JH, Noller HF (2001). Crystal structure of the ribosome at 5.5 Å resolution. Science 292: 883–896

Zhang G, Campbell EA, Minakhin L, Richter C, Severinov K, Darst SA (1999) Crystal structure of *Thermus aquaticus* core RNA polymerase at 3.3 Å resolution. Cell 98:811–824

Modeling Protein Folding Pathways

C. Bystroff, Y. Shao

1 Introduction: Darwin Versus Boltzmann

All computational models that predict something have certain underlying assumptions that constitute the physical basis for the model. In protein structure prediction, there are two physical/biological processes that can be modeled: the process of evolution, or the process of folding. We name these two paradigms, Darwin and Boltzmann, after the scientists who defined the fundamental principles of evolutionary biology and statistical thermodynamics, respectively.

Most of the work in protein structure prediction is Darwin-based, using the well-known premise that sequences that have a common ancestor have similar folds, and they strive to extrapolate this principle to increasingly distant sequence relationships. Methods that use multiple sequence alignment, structural alignment, or "threading potentials" are implicitly searching for a common ancestor. Despite the often-used "energy-like" scoring functions, these methods do not address the physical process of folding. Evolution happens on the time scale of millions of years, folding on the time scale of fractions of a second.

Protein structure prediction of the Boltzmann kind is perceived to be a very difficult problem. Many have tried their hand at it over the last thirty years, and an equal number have failed to improve upon Darwin-based methods. The problem of predicting folding pathways may be perceived to be even harder, since it *should* depend on first solving the protein folding problem. However, this is not true, as we shall see. Prediction of the protein folding pathway may be evaluated by looking at the success in predicting sub-segments or substructures of proteins. If the computational model has the right underlying assumptions about what comes first in the pathway, and what comes next, and so on, then blind predictions, such as those done as part of

C. Bystroff, Y. Shao
Department of Biology, Rensselaer Polytechnic Institute, Troy, New York, USA

Nucleic Acids and Molecular Biology, Vol. 15
Janusz M. Bujnicki (Ed.)
Practical Bioinformatics

CASP, the Critical Assessment of Protein Structure Prediction bi-annual worldwide experiment (Moult et al. 2001), may validate that model. And the pathway model that eventually arises from this process will tell us more than just a final answer.

In this chapter, we present a series of bioinformatics and simulation experiments related to predicting protein structure by modeling the folding pathway. We will conclude that ab initio predictions can be done either by simulations or by a rule-based fragment assembly method, and that it is possible to find folds that are not present in the database of structures. We will discuss issues of accuracy and resolution and present some possible directions for the future.

1.1 Protein Folding Pathway History

The early work of Levinthal and Anfinsen established that a protein chain folds spontaneously and reproducibly to a unique three dimensional structure when placed in aqueous solution. Levinthal proved that the folding process cannot occur by random diffusion. Anfinsen proposed that proteins must form intermediate structures in a time-ordered sequence of events, or "pathway" (Anfinsen and Scheraga 1975). The nature of the pathways, whether they are restricted to partially native states or whether they might include non-specific interactions, such as an early collapse driven by the hydrophobic effect, was left unanswered.

Over the years, the theoretical models for folding have converged somewhat (Baldwin 1995; Colon and Roder 1996; Oliveberg et al. 1998; Pande et al. 1998), in part due to a better understanding of the structure of the "unfolded state" (Dyson and Wright 1996; Gillespie and Shortle 1997; Mok et al. 1999) and to a more detailed description of kinetic and equilibrium folding intermediates (Eaton et al. 1996; Gulotta et al. 2001; Houry et al. 1996). An image of the transition state of folding can now be mapped out by point mutations, or "phi-value analysis" (Fersht et al. 1992; Grantcharova et al. 2000; Heidary and Jennings 2002; Mateu et al. 1999; Nolting et al. 1997). The "folding funnel" model (Chan et al. 1995; Onuchic et al. 1997) has reconciled hydrophobic collapse with the alternative nucleation-condensation model (Nolting and Andert 2000) by envisioning a distorted, funicular energy landscape (Laurents and Baldwin 1998) and a "minimally frustrated" pathway (Nymeyer et al. 2000; Shoemaker and Wolynes 1999) through this landscape. The view remains of a channeled, counter-entropic search for the hole in the funnel as the predominant barrier to folding (Zwanzig 1997).

Simulations using various simplified representations of the protein chain, including lattice models, have clarified the basic nature of folding pathways (Kolinski and Skolnick 1997; Mirny and Shakhnovich 2001; Shakhnovich 1998; Thirumalai and Klimov 1998). The topology of the fold plays a domi-

nant role in defining the critical positions that effect the folding rate (Ortiz and Skolnick 2000; Shea and Brooks 2001). Models that represent the chain in atomistic detail show that minimally frustrated, low-energy pathways may involve the propagation of structure along the chain like a zipper (Alm and Baker 1999; Munoz et al. 1998). All-atom, explicit solvent molecular dynamics simulations have reproduced the experimentally determined conformations for short peptides (Cavalli et al. 2002; Duan and Kollman 1998; Garcia and Sanbonmatsu 2001; Krueger and Kollman 2001; Bystroff and Garde 2003). This large body of work is still inconclusive, but clearly folding is best represented by an ensemble rather than a single pathway.

2 Knowledge-Based Models for Folding Pathways

The approach that began with I-sites is an attempt to build a hierarchical series of models mirroring the hierarchy of folding events, from initiation to nucleation to propagation and condensation. The hierarchy can be roughly described as "local to global." Each model builds on the model before it. At each point the results are an ensemble of conformational states.

"Local structure" is a generic term for the conformations of short pieces of the protein chain, usually 3–20 residue pieces. Local structure motifs include the two common forms (alpha helix and beta strand) along with a few dozen turns, half-turns, caps, bulges and coils. The role of local structure motifs with regards to the initiation of folding has been discussed by Baldwin, Rooman and others (Baldwin and Rose 1999; Efimov 1993; Rooman et al. 1990).

2.1 I-sites: A Library of Folding Initiation Site Motifs

I-sites is a library of 262 sequence patterns that map to local structures. A sequence pattern is expressed as a position-specific scoring matrix (PSSM). Recurrent sequence patterns had been previously used for prediction of structural motifs, including the Schellman motif (Schellman 1980), the hydrophobic staple (Munoz et al. 1995), and various types of coiled coil (Woolfson and Alber 1995). Recurrent sequence patterns of various lengths were found by exhaustively clustering short segments of sequence profiles for proteins in a non-redundant database of known structures (Bystroff et al. 1996; Han and Baker 1996, 1995, Han et al. 1997). Bystroff and Baker mapped recurrent sequence patterns to their predominant structural motifs and used reinforcement learning to optimize the sequence-structure correlation (Bystroff and Baker 1998). The resulting I-sites Library (Fig. 1) has been used in various prediction experiments (Bystroff and Baker 1997; Bystroff and Shao 2002) and has inspired numerous experimental studies since its publication (Jacchieri 2000; Mendes et al. 2002; Northey et al. 2002b; Skolnick and

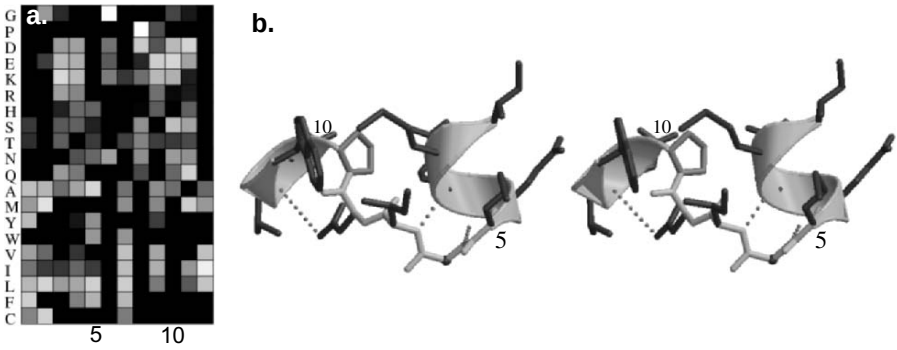


Fig. 1. **a** I-sites profile for alpha–alpha corner motif. *Boxes* are *shaded lighter* in proportion to the log-likelihood ratio of each amino acid at each position relative to the start of the motif. **b** Stereo image of the alpha–alpha corner motif

Kolinski 2002; Steward and Thornton 2002). I-sites motifs have been linked to local structure stability in both NMR studies (Blanco et al. 1994; Munoz et al. 1995; Viguera and Serrano 1995; Yi et al. 1998) and molecular dynamics simulations (Bystroff and Garde 2003; Gnanakaran and Garcia 2002; Krueger and Kollman 2001). Mutations in high-confidence I-sites motif regions are found to have dramatic effects on folding (Mok et al. 2001; Northey et al. 2002a). About one third of all residues in all proteins are found in high-confidence (>70 %) I-sites motif regions and these sites are predicted to be conformationally stable and early folding.

2.2 HMMSTR: A Hidden Markov Model for Grammatical Structure

The I-sites library was condensed to a single, non-linear hidden Markov model (HMM), called HMMSTR ("hamster"). This model, trained on a large database of protein structures and multiple sequence alignments, removes the fragment length dependence of I-sites motif predictions, models the adjacencies of motifs in proteins, and puts all of the motifs on the same probability scale (Fig. 2). Unlike profile HMMs (Eddy 1996; Gough and Chothia 2002; Karplus et al. 1998), HMMSTR has a highly branched and cyclic connectivity, containing for example a seven-residue cycle of helix states representing the amphipathic helix heptad repeat motif. By modeling the adjacencies of motifs, HMMSTR is a model for the ways that local structure can be arranged along the sequence, similar to the ways that words can be arranged in a sentence. This is, in a simple way, a model for the grammatical structure of protein sequences, from words to phrases.

The result of an HMMSTR prediction is like that of any HMM, an ensemble of Markov state strings. Each string is a state, one state for each position in the

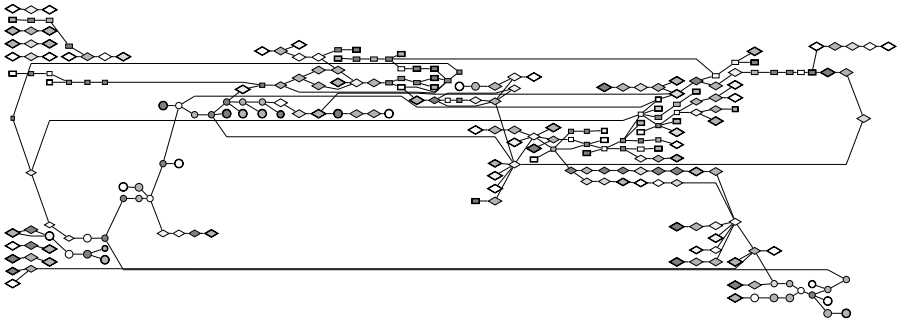


Fig. 2. HMMSTR represented as a direct graph. The symbol shape represents the secondary structure type; *circles* helix; *rectangles* beta sheet; *diamonds* other motifs. *Shading* represents the amino acid preference; *dark gray* non-polar; *gray* polar; *light-gray* proline; *lightest gray* glycine; *white* no preference. Only high-probability transitions are shown

sequence, represents a probable arrangement of mutually-compatible local structure motifs. A single prediction may be obtained from the ensemble by either selecting the most probable state string, or better, by a voting procedure over the whole ensemble (Bystroff et al. 2000). HMMSTR improved the overall accuracy in local structure prediction over the I-sites method from 43– 60 % for eight residue fragments with RMSD <1.4 Å (Bystroff et al. 2000). HMMSTR has been used for local and secondary structure prediction (Bystroff et al. 2000; Rost 2001), inter-residue contact prediction (Zaki et al. 2000), and as the source of a fragment library for Rosetta simulations (Bystroff and Shao 2002). Previous HMMs have modeled proteins globally, not as fragments (Eddy 1996; Gough and Chothia 2002; Karplus et al. 1998).

3 ROSETTA: Folding Simulations Using a Fragment Library

The ROSETTA folding simulation algorithm uses Monte Carlo Fragment Insertion (MCFI) to predict the 3D structures of small proteins or protein fragments without the use of structural templates (Bonneau and Baker 2001; Bonneau et al. 2001; Simons et al. 1997, 1999a, b). MCFI is a mostly downhill search in a knowledge-based energy landscape. Each MCFI move consists of replacing the backbone angles of segments of the chain with fragments in a library. ROSETTA has been successful in prediction experiments (CASP (Moult et al. 2001)) either using fragments from the database, from HMMSTR, or from the I-sites motif library.

In the version of ROSETTA that runs as a public server (www.bioinfo.rpi.edu), the fragment library is derived from I-sites fragment predictions, and the highest confidence I-sites were restrained to their predicted backbone

angles to increase efficiency. Fragment insertion was allowed in the restrained regions, but moves were constrained to deviate by more then 60° from the I-sites prediction. Also, long sequences were simulated as overlapping short fragments of approximately 50 residues each, again for efficiency. The resulting predictions are spliced together at the end, using a genetic algorithm in conjunction with the ROSETTA knowledge-based energy function. Detailed descriptions of each of the algorithms have been previously published (Bystroff and Shao 2002; Simons et al. 1997, 1999b).

3.1 Results of Fully Automated I-SITES/ROSETTA Simulations

3.1.1 Summary

A web server was used to predict 31 protein structures in the CASP4 experiment (2000) and 44 in the CASP5 experiment (2002). The successes and failures of the server may be summarized in a few broad statements. The statistics and conclusions presented here refer to bona fide blind predictions sent automatically to the CAFASP site as part of their "Fully-Automated" satellite experiment (Fischer et al. 2001). A more detailed analysis of this and other methods can be obtained from the associated publications (Bystroff and Shao 2002; Shao and Bystroff 2003).

Over the 75 targets, 64 % of the residues were found in "topologically correct" large fragments, defined as fragments of 30 residues or more with RMSD <6 Å. At 6 Å RMSD, the correct overall chain trace has been reproduced, but not the finer details of structure. Occasionally, beta strand may be out of order in a sheet, and strands may be substituted for helices.

A smaller percentage of all 30-residue fragments, 44 %, were predicted with a 5 Å RMSD. At 5 Å precision, secondary structure is occasionally mispredicted, loop structures may be wrong in detail, and axial rotations of secondary structure units are possible. However, much or most of the non-local packing interactions are faithfully though roughly reproduced at this level of accuracy, and strand mispairing is not observed.

In practice, the details of the local structure are often correctly predicted when a fragment was globally correct, but the RMSD measure is insensitive to this. Therefore, another measure is used to evaluate the local accuracy of the predictions. The maximum deviation in backbone angles (*mda*) over a window of eight residues is usually ~180° or smaller, and serves as a strictly local measure of correctness. Eight-residue peptides that have *mda* <90° and obey all of the stereochemical constraints of a polypeptide, have an RMSD of 1.4 Å at most (Bystroff and Baker 1998). Unfortunately, when *mda* is plotted alongside RMSD, it is immediately obvious that the good local structure predictions do not always coincide with the good, large-fragment predictions.

3.1.2 Topologically Correct Large Fragment Predictions are Found

Figure 3 shows a 97-residue fragment prediction with 5.9 Å RMSD. At this level of precision, the residues found in the core are correct and their 3D arrangement is roughly correct. In fragments that contained helices, the N and C capping residues were usually but not always correctly located, and the direction of the chain coming off of the helix was generally correct. The orientation of parallel sheets to helices was reproduced to within about 60°, and the axial orientation of the helices with respect to strands was almost always correct, even though rolling the helix would not greatly affect the RMSD value.

Some characteristics of even the "correct" fragment predictions suggested ways in which the algorithm could be improved. The most obvious of these is the distortion of alpha helices. True native helices retain very straight helix axes despite variability in the backbone angles. Helices in the predictions, however, were often distorted, sometimes bending the axis by 90° over its length. A combination of factors produces these errors. ROSETTA has no energy penalty for helix distortion, while it gives a large energetic bonus for packing hydrophobic residues into the core and for maintaining a low radius of gyration. Bent helices are found to replace helix kinks and alpha-alpha corners. Adding a penalty for helix distortion might fix this problem.

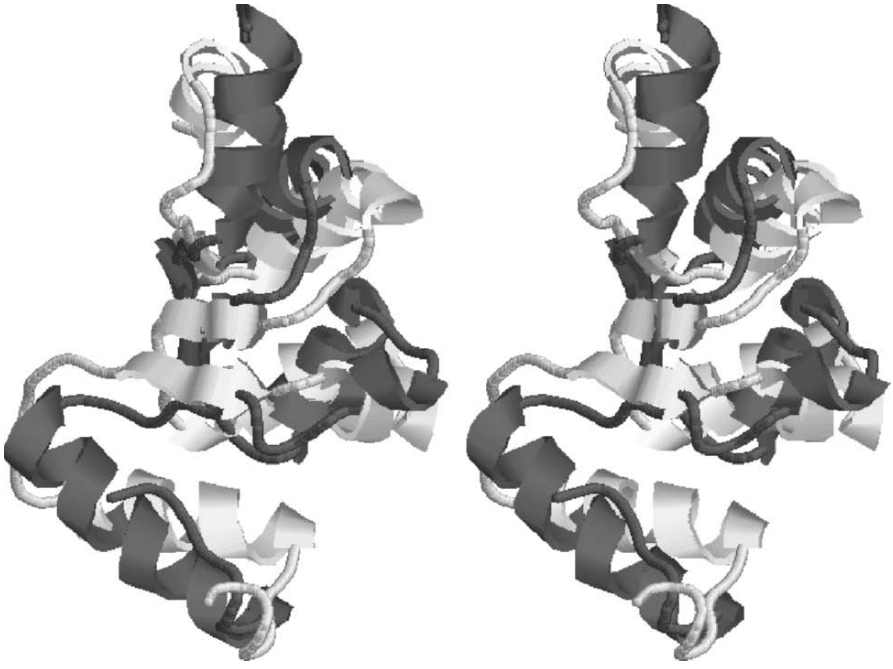


Fig. 3 ROSETTA-predicted (*dark gray*) and true (*light gray*) structure of tryptophan synthase alpha subunit from *P. furiosus* (PDB code 1GEQ) residues 57–153

Topological correctness is a weak criterion for usefulness, since it means that only the handedness of the chain reversals and most of the secondary structure are right. However, these fragmentary predictions may narrow the search space for a structural analog or remote homologue, and may therefore be useful in combination with other methods. The I-sites Server correctly identified the overall anti-parallel β topology of one of the CASP5 targets, the F-actin capping protein (PDB code 1IZN), a new fold at the time.

3.1.3 Good Local Structure Correlates Weakly with Good Tertiary Structure

If the ROSETTA simulations followed a "local structure first" pathway, then we would expect to see good super-secondary structure predictions coinciding with good local structure predictions. However, this is not always the case. Frequently, the topologically correct large fragments have the wrong local structure. This is true despite the fact that at least 90 % of the target sequences are covered by at least one fragment with the correct local structure in the fragment library.

Three-state secondary structure (SS) predictions were made using a version of HMMSTR that was trained on a large dataset of proteins of known structure with SS states assigned using DSSP (Kabsch and Sander 1983). The accuracy of these predictions over the 31 targets was 73.3 %, only slightly lower than the state of the art in SS prediction (Jones 1998). SS predictions based on tertiary structure (TS) predictions from ROSETTA had the potential of benefiting from the added TS information, however this did not improve the prediction accuracy.

Using SS assignments derived from the TS predictions using DSSP or STRIDE (Frishman and Argos 1995), the prediction accuracy was low (50–60 % Q3) because these programs depend on precise positioning of the hydrogen-bonding residues in assigning the strand state (E). Instead, the SS predictions were derived from the fragments in the fragment library, using SS assignments from their native proteins. Using this method, the overall Q3 score improved to 72.4 %, but this is still no better than the SS predictions that use sequence alone without running a simulation.

If the simulation were reproducing the folding process, one might expect that the correctly-predicted tertiary interactions would add information to the secondary structure prediction. One explanation for the lack of improvement in secondary structure, despite some success in tertiary packing, is that topologically correct tertiary structures are possible even when the wrong local structure is used to build it.

3.1.4 Average Contact Order is too Low

Relative contact order (Plaxco et al. 1998) is calculated from the coordinates as follows:

$$CO = \frac{1}{L \bullet N} \sum^{N} \Delta S_{ij} \tag{1}$$

where ΔS_{ij} is the sequence separation $|i\text{-}j| \geqq 5$, for residues, *ij*, that are in contact (Cα-Cα distance <8 Å). *N* is the number of contacts, and L is the length of the sequence. The overall average *CO* in the targets was 0.252, while the *CO* for the 32 predictions was 0.119. The lower *CO* is mostly the result of an increased number of beta hairpins. Contacts that are local, such as those in beta hairpins, are easier to find in a conformational search, and thus may represent kinetic intermediates, trapped at the end of the simulation. Kinetic trapping may be exacerbated by the more computationally efficient server protocol. A possible solution is to do more replicates and rely on cluster analysis to identify the global energy minimum. Practical limitations currently stand in the way of implementing this.

Alternatively, the predominance of beta hairpins may reflect an error in the energy function with regard to the backbone angles. Positive ϕ angles, favored only in glycine residues and usually required for turns, are found in the same proportion in the targets (8 %) and in the predictions (7 %), but in the targets, 44 % of these turn residues are glycines, while in the prediction only 16 % are glycines. This suggests that a larger energetic penalty for positive ϕ angles in non-glycine residues might correct the overabundance of hairpin turns.

3.1.5 How Could Automated ROSETTA Be Improved?

Our results suggest that a combination of improvements in efficiency may increase the potential of the ROSETTA algorithm as a high-throughput engine for tertiary structure prediction at the 30–100 residues length scale. We suggest that a combination of structure comparison metrics be used for the evaluation of correctness; a low RMSD in the context of low backbone angle deviations is shown to identify predictions that were "correct for the right reasons".

Secondary structure assignments were not improved by the use of tertiary structure predictions, partly because it was possible to obtain a globally correct tertiary structure prediction by inserting fragments of the wrong local structure.

An overall low contact order was observed in the predictions relative to the true structures. This is at least partly due to the absence of an energetic penalty for unfavorable backbone torsion angles. These may also represent kinetically trapped intermediate structures from a simulation that was too short.

4 HMMSTR-CM: Folding Pathways Using Contact Maps

HMMSTR-CM is a pathway-based method for predicting protein structure using contact maps. Contact maps are square symmetrical Boolean matrices that represent protein tertiary structures in a two-dimensional format. The 2D format has simplified the process of developing a rule-based algorithm for folding pathways. Contact maps may be projected into three-dimensions using existing methods (Aszodi et al. 1997; Brunger et al. 1986; Crippen 1988; Vendruscolo et al. 1997).

Two-dimensional flat images are more readily discernable to the eye and more memorable than complex, rotating three-dimensional images. With only a little training, a student can learn to quickly distinguish a contact map for an α/β barrel from a 3-layer α/β fold, different topologies which are very similar in their secondary structures. Similarities between distant homologues or analogs of α/β and all β folds can be seen easily in contact maps, even when the 3D structures superimpose poorly. It makes sense that if our eyes can recognize protein folds from 2D patterns, we should be able to program a computer to do so, and thereby create a new tool for learning the rules of folding.

Previous contact map prediction methods have used neural nets (Fariselli and Casadio 1999; Pollastri and Baldi 2002), correlated mutations (Olmea and Valencia 1997; Ortiz et al. 1998; Singer et al. 2002), and association rules (Hu et al. 2002; Zaki et al. 2000). Neural net-based predictions had an average accuracy of about 21 % overall (Fariselli et al. 2001), while higher accuracies were reported for local contacts (Pollastri and Baldi 2002), but the accuracy is lower for all-α proteins.

Our earlier work (Zaki et al. 2000) led us to believe that two important factors were missing in contact map predictions. First, typical predicted contact maps were structurally ambiguous or physically impossible, representing either multiple or zero possible folds when projected into three dimensions. Secondly, the order of appearance of contacts (i.e., the pathway) was not considered, even though much is known about the general character of folding pathways (Baldwin 1995; Fersht 1995; Galzitskaya et al. 2001; Nolting and Andert 2000). In the new approach we tried to enforce "physicality" and protein-like characteristics by using protein templates and simple rules. The rules consist of common sense facts for the packing of secondary structures (Table 1). Rules for the order of appearance were derived from the general assumptions of a nucleation/propagation pathway (Nolting and Andert 2000).

4.1 A Knowledge-Based Potential for Motif–Motif Interactions

The first step in predicting a contact map is to assign an energy to each potential contact. The energy in this case is the database-derived likelihood of con-

Table 1. Physicality and propagation rules

1. Maximum neighbor rule: One residue can have at the most 12 contacts.
2. Maximum mutual contact rule: If residue *i* and *j* are in contact, there are at the most six residues in contact with both *i* and *j*.
3. Beta pairing rule: A beta strand can be in contact with at the most two other beta strands.
4. Beta sheet rule: Any two pairing strands are either parallel or antiparallel.
5. Helix mutual contact rule: No residue can be in contact at the same time with the residues on the opposite sides of a helix.
6. Helix rule: Only the contacts between residues *i* and *i*+4 are allowed in a helix.
7. Beta rule: No contacts ($|j-i|>3$) are allowed within any strand.
8. Right-hand crossover rule: Crossovers between parallel strands of the same sheet (paired or not) are right-handed, especially if the crossover contains a helix.
9. Helix crowding rule: If a helix can go to either side of a sheet, it picks the side with fewer crossovers.
10. Strand burial rule: If a strand can pair with either of two other strands, it chooses the one that is more non-polar.
11. Propagation rule: A contact cannot be assigned between *i* and *j* if there are more than eight residues in the intervening sequence that have no assigned contacts.

tact between any two local structure motifs. This implies that local structure forms first, then these sub-structures condense to form larger units, subject to a free energy of interaction, similar to a binding energy. But like its predecessors I-sites and HMMSTR, HMMSTR-CM is a Bayesian ensemble approach; each residue is represented as a probability distribution of motifs, rather than as a single motif. Thus, each contact potential models a pair of flickering local structures, interacting in proportion to their structural content.

The energetic interaction potential of two motifs is modeled as the statistical interaction potential between two corresponding Markov states of the HMMSTR model. Knowledge-based Markov state "pair potentials" were summed from the CATH database of protein domains. Each domain was first preprocessed into Markov state probability distributions using the Forward/Backward algorithm (Rabiner 1989) to get the position-dependent Markov state probability distribution γ (Eq. 2).

$$\gamma(i,q) = P(q \mid i) \tag{2}$$

The pairwise contact potential between any two HMMSTR states *p* and *q* (*G(p,q,s)*) was calculated as the log of the mutual probability of these two states in contacting residues (Cα -Cα distance <8 Å), for proteins in the PDB-select database (Hobohm and Sander 1994) (Eq. 3).

$$G(p,q,s) = -\log \frac{\sum_{PDBSelect} \sum_{i \exists D_{i,i+s}<8\text{Å}} \gamma(i,p) * \gamma(i+s,q)}{\sum_{PDBSelect} \sum_{i} \gamma(i,p) * \gamma(i+s,q)} \quad (3)$$

The sensitivity of discriminating contacts from non-contacts improved greatly by calculating G as a function of the sequence separation $s=|j-i|$ ($4 \leqq s \leqq 20$. For sequence separations greater than 20, s=20 was used.) The total number of potential functions G was 1037153, one for every pair of 247 Markov states in HMMSTR and every separation distance from 4 to 20. G may be viewed as the knowledge-based energy of contacts between local structure motifs.

The target contact potential map E (Eq. 4) is the matrix of contact potentials between every two residues in the target sequence. The contact potential between residues i and j ($E(i,j)$) in the target was calculated as the probability-weighted sum of the pairwise potential functions G.

$$E(i,j) = \sum_{p} \sum_{q} \gamma(i,p) * \gamma(j,q) * G(p,q,s) \quad (4)$$

where $s=|i-j|$. In general, the contact potential map readily identifies possible contacts between β strands, and also finds super-secondary structure motifs such as the right-handed parallel βαβ motif and the αα corner.

4.2 Fold Recognition Using Contact Potential Maps

The flowchart in Fig. 4 summarizes the steps in a contact map prediction using HMMSTR-CM. Target sequences were aligned to database sequences using PSI-BLAST (Altschul et al. 1997). The resulting multiple sequence alignment was converted to an amino acid probability distribution or sequence profile, as described previously (Bystroff and Baker 1998). The target sequence profile and 1239 template profiles from the PDBselect database (Hobohm and Sander 1994) were converted to HMMSTR γ-matrices (Eq. 2), and γ^{target} was aligned against each γ^{template} using Bayesian adaptive alignment (Zhu et al. 1998). The alignment matrix in this case was the sum over all joint probabilities of Markov states (Eq. 5). The alignments were evaluated using contact potential maps to choose the best template.

$$\mathrm{A_{ij}} = \sum_{q} \gamma_{iq}^{target} \gamma_{jq}^{template} \quad (5)$$

Candidate target contact maps were generated for each alignment, and each was evaluated by the contact free energy (CFE), as described below, and

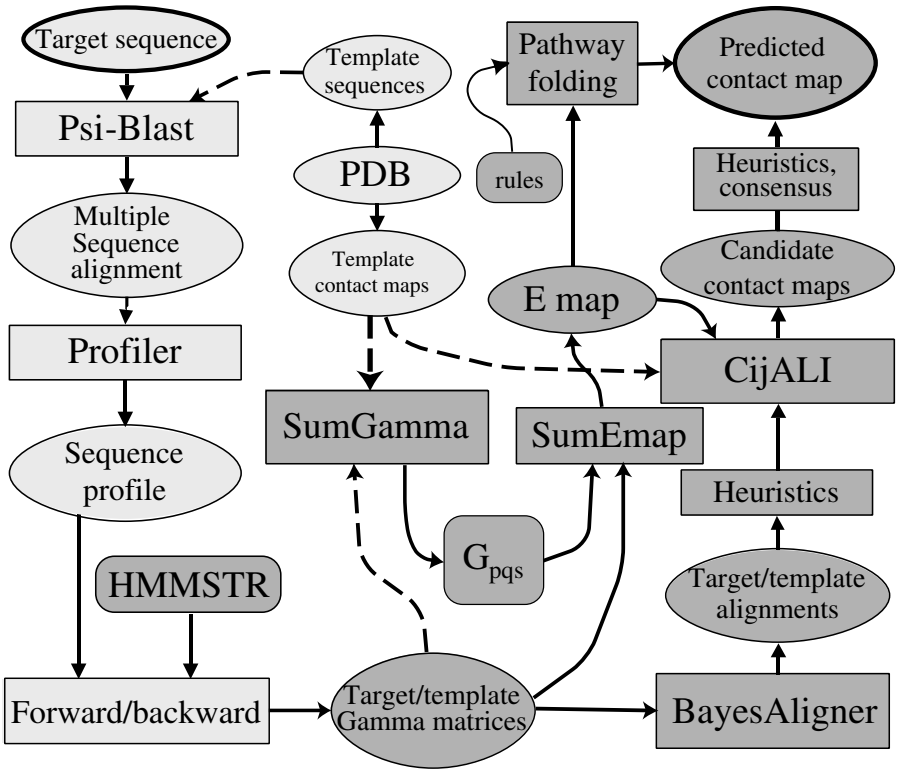


Fig. 4. Flowchart for HMMST-CM contact map prediction. *Rectangles* represent algorithms, *ovals* are data, and *rounded rectangles* are models. *Dashed lines* apply to training set data (templates) and *solid lines* apply to both templates and targets. *Light gray* items are described in referenced material. *Dark gray* items are described in this text as follows: HMMSTR, Section 2.2; gamma matrices, Eq. (2); SumGamma, Gpqs, Eq. (3); SumEmap, E map, Eq. (4): Rules, Pathpath foling, Section 4.4, Table 1; BayesAligner, Target/template alignments, Section 4.2, Eq. (5), Fig. 5a; Heuristics, Eq. (6); CijALI, Section 4.2, Eq. (7); Heuristics, consensus, Section 4.3, Fig. 6

other measures. The BayesAligner produced a single score and any number of alignments. Templates with low alignment scores were rejected. Otherwise, 100 alignments were selected at random for further evaluation.

BayesAligner produces a probability distribution over all possible alignments with no more than k gaps (k depends on the sequence lengths). The quality of the alignment distribution (see Fig. 5a) was a strong indicator of the quality of the template. Templates and/or alignments were removed from this set if they were highly fragmented. This was assessed using a "compactness score" which is simply the length of the longest contiguously aligned region, ignoring small gaps (three residues). The template distance at the ends of the aligned blocks was enforced to be physically possible values (Eq. 6) by trimming the aligned blocks if necessary.

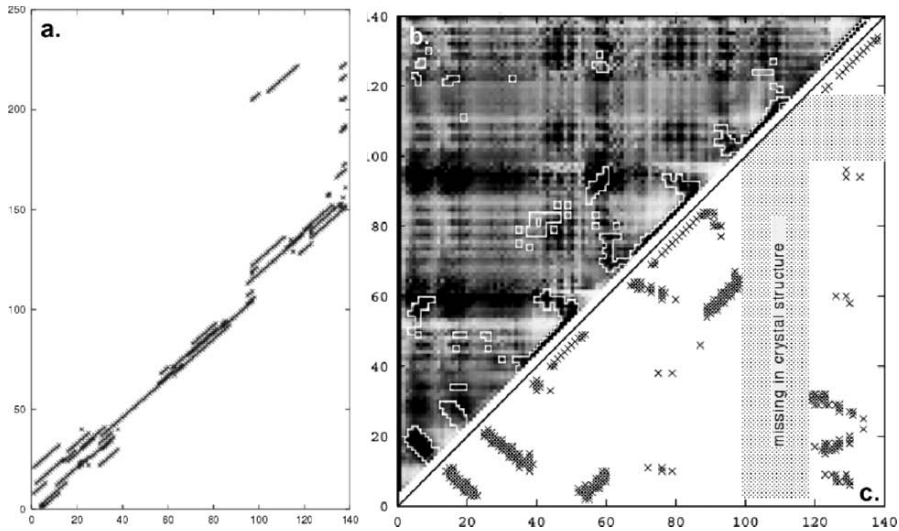


Fig. 5. **a** BayesAligner summary of the most probable alignments between YqgF (*x*-axis) and 1HJR (*y*-axis). **b** Contact potential map for YqgF; *darker* is lower energy $E(i,j)$. Predicted contacts are outlined in *white*. **c** Contact map from crystal structure of YqgF, hypothetical protein from *E. coli*

$$D_{i'j'} \leq 3.8\text{Å} \times \left| i - j \right| \tag{6}$$

Candidate contact maps (C) were generated using the alignments and the contact maps of each of the templates that had the top ten compactness scores, scored using the "contact free energy" (*CFE*, Eq. 7). CFE was calculated by summing the relative contact potential E over all contacts, C. Contacts with sequence separations $|j–i|$ of less than 4 were ignored.

$$CFE = \sum_{i,j \exists C_{ij}=1 \cap (j>(i+3))} E(i,j) - \left\langle E \right\rangle \tag{7}$$

where $<E>$is the mean contact potential for the target. For each template, we calculated the *CFE* for all contact map candidates and chose the one with the best energy as the best alignment to that template.

After we carried out the above procedure for every template in our dataset, we usually accumulated several hundred target contact map predictions. How to evaluate them and choose one as the final prediction became a problem itself. The decision was made using the following four parameters: CFE, the BayesAligner score, the compactness score and the similarity between sequence lengths of the target and the template. The primary parameter was the CFE, since it represented the free energy of the sequence when folded to the template structure. However, we observed that better alignments and similar lengths improved the perceived prediction quality.

The automated selection of templates was sometimes overridden by our ab initio analysis, described below. If the propagation rules favored one topology over another and a template of the favored topology was present in our list of top scorers, we would select that template over a higher scoring one.

4.3 Consensus and Composite Contact Map Predictions

Often several of the top-scoring templates contained the same fold or substructure. Consensus was considered a strong indicator, especially if the fold was uncommon. Multiple candidates were sometimes used to construct a single composite map. In practice, consensus similarity between many structures is difficult to see in a 3D multiple superposition, but is easy to see in superimposed contact maps.

This prediction can be done in different ways when the top scoring templates share a similar fold. When they disagree on some contacts, the consensus contacts (not necessarily those from the best scoring template) are used; when some templates aligned well in one region and other templates aligned well in another region, the predictions from these templates were spliced to maximize the coverage. For some recurrent contact patterns, e.g., the parallel $\beta\alpha\beta$ motif, the parallel β contacts or the helix contacts were sometimes incomplete because of misalignment of the template. By combining the top scoring predictions, we could "grow" the incomplete pattern into a complete one.

Simply combining the contact maps introduces "noise" – contacts that make the prediction non-physical. (A "non-physical" contact map cannot be projected into three-dimensions.) Manual post-processing, including pathway-based editing (discussed next) was needed to enforce the physicality of the final contact map.

4.4 Ab Initio Rule-Based Pathway Predictions

The fold-recognition methods described above have their roots in evolution, but contact maps as a representation of protein structures were chosen not with the intention of building a Darwin-based prediction strategy, but with the intention of modeling the folding pathway. Contact maps simplify the conformational search. However, as we have pointed out, not all contact maps represent physically-possible three-dimensional objects. Therefore, external information about proteins must be included. A set of aligned templates is one source of external information. Here we present a set of fundamental rules (Table 1) and energies (Eq. 4) that serve the same purpose – to restrict the conformational search to contact maps that are physically possible and protein-like.

A rule-based structure propagation model was used either in conjunction with templates or ab initio (without templates). In CASP5, ab initio predictions were sometimes done on targets found later to be remote homologues by CASP5 assessors, but because our alignment method was not always able to recognize remote homology, we treated them as potential new folds. The procedure is as follows.

Starting from a contact potential map, E, we kept the contacts that were better than a cut-off value. The cut-off value was chosen so that blocks of contacts were found between most secondary structural units, especially between β strands. As a result, the initial contact map was often characterized by dense blocks of contacts between β strands and sparse contacts to helices and between helices.

If we kept all of these contacts, clearly the map would be physically impossible. For example, a β strand element cannot be paired with more than two other β strands. A set of common-sense rules were compiled to weed out the possible contacts from the impossible or unlikely, and to enforce protein-like characteristics, such as right-handed crossovers and exposed reverse turns (Table 1). These rules were enforced as the prediction was propagated.

The folding pathway consisted of "assigning" or "erasing" contacts. Contacts were assigned if the energy $E(i,j)$ passed a threshold and the corresponding contact $C_i = 1$ did not violate any of the rules, otherwise they were erased. Blocks of potential contacts were considered together, and the order in which blocks were considered depended on their proximity to previously assigned blocks of contacts (Table 1, rule 11), following the principles of the nucleation/condensation folding mechanism.

To start the folding pathway, we selected one or more local regions with high confidence contacts as the "nucleation site(s)". We then propagated the prediction in both directions by assigning or erasing blocks of contacts around and between the nucleation site(s), subject to our set of rules. TOPS diagrams (Sternberg and Thornton 1976) were drawn for the growing structure as a visual aid, since some rules applied to the topology. The pathway, and the prediction, was complete when all of the remaining contacts were rejected. The method is best described using examples, as in the next section.

4.5 Selected Results of HMMSTR-CM Blind Structure Predictions

HMMSTR-CM was used to predict contact maps as part of the CASP5 experiment. Targets in the FR (fold recognition) and NF (new fold) categories were predicted using the three methods described above: threading, consensus and ab initio, collectively called HMMSTR-CM. In all three of these methods, the overall accuracy of the contact map prediction depends on the accuracy of the secondary structure prediction, which was done using HMMSTR.

4.5.1 A Prediction Using Templates and a Pathway

YqgF, a hypothetical protein from *E. coli*, was successfully predicted using the template-based approach in conjunction with a pathway prediction. All visible secondary structure units are correctly predicted (note that the 17 residues from 102 to 118 are missing in the crystal structure), and all of the true contacts have a higher-than-average *E(i,j)* score. After aligning the contact potential matrix, *E*, to each of the 1258 templates, a consensus contact map was plotted using the top-scoring six templates. This map was used to construct a folding pathway. Nucleating the pathway at $\beta_4\alpha_2\beta_5$ and propagating produced a TOPS diagram that agreed with only one of the templates, 1HJR, this template was therefore chosen to construct the consensus contact map. 1HJR had the third highest CFE score. In the prediction based on 1HJR, the N-terminal three-strand β meander is slightly under-predicted, and a contact between helices 1 and 2 is slightly over-predicted. Nevertheless, the topology is correct throughout (Fig. 5b). The two higher-scoring templates that were not chosen had very different, and incorrect, topologies.

4.5.2 A Prediction Using Several Templates

Ycdx, another hypothetical protein from *E. coli*, was successfully predicted using multiple templates. The threading approach found four templates that had high CFE scores and also shared common structural components. Three of those templates were eight-stranded α/β barrels and the other consisted of two parallel α/β domains. Ycdx turned out to be an αβ barrel with seven parallel β strands (PDB code 1M65). Templates with good CFE scores existed but none of them predicted all of the first five helices and the parallel β strand contacts correctly. However, by combining the results from the top scoring templates, we made a consensus prediction that was better than any of the contact maps made from the single templates. In particular, we correctly found parallel contacts between the first six β strands (Fig. 6).

The sixth helix and the contacts between the sixth and the seventh strands were predicted but misaligned. Our method mispredicted the C-terminus to be a parallel βαβ motif, as in a standard eight-stranded TIM barrel, but the true structure is three short helices connected by loops. Visual inspection of the templates confirmed that they share the same topology, and a consensus fold prediction would have been obvious given this result. However, finding structural similarity and combining structures is more easily automated in the 2D contact map format than in 3D coordinate space. Consensus in contact maps provides a way to merge and "grow" the incomplete contact maps of different targets into a more complete contact map.

Ycdx also revealed a weakness of the method. HMMSTR, which is trained to recognize recurrent super-secondary motifs, does not recognize the unusual substructure at the C-terminus of this protein, three short helices

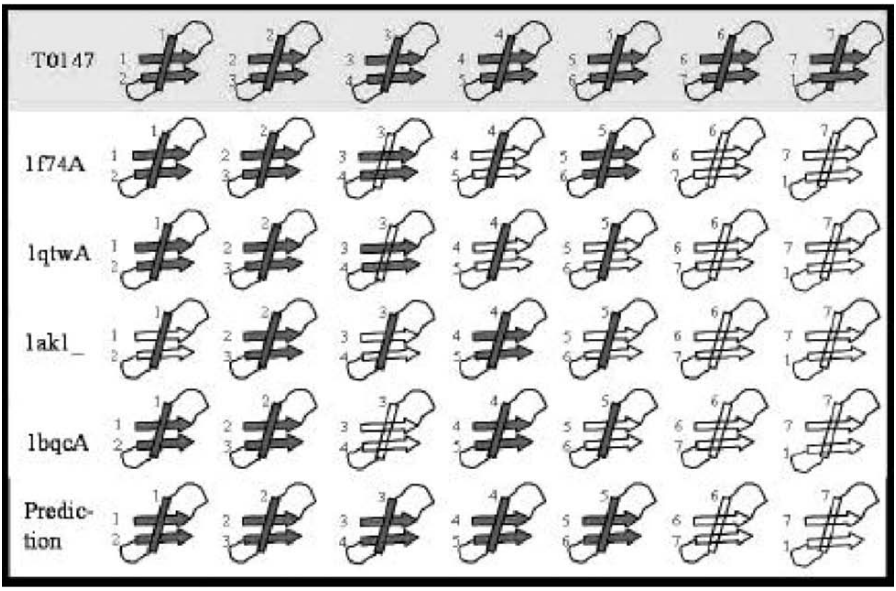


Fig. 6. Summary of strand-strand (*arrows*) contacts and helix predictions for four templates aligned to Ycdx (T0147). *Shaded symbols* represent contacts that were correctly predicted using the template specified in the margin. The last line shows contacts that were correctly predicted after combining the four templates and using the consensus set

instead of the usual βαβ motif. The consensus method, as we have defined it, tends to bias the prediction toward the more common folds. In fact, this is a problem with any template-based method.

4.5.3 Correct Prediction Using Only the Folding Pathway

Hypothetical protein HI0073 from *H. influenzae* is an example of a successful ab initio prediction. It has 116 residues arranged in a three-layer all-parallel α/β sandwich. The contact potential map (Fig. 7a) shows that most of the true contacts are assigned favorable (darker) contact potentials. However, many other favorable regions are also correctly predicted as non-contacts. Depending on the choice of nucleation sites, there was more than one way to derive a physically possible and high-scoring topology. In this case, the nucleation site was selected to be $\beta_2\alpha_2\beta_3$. Contacts were assigned or erased in four steps, as follows:

1. Parallel β contacts were assigned between β_2 and β_3.
2. Anti-parallel β contacts were assigned to β_1 and β_2. All other β contacts to β_2 were erased.
3. There were two ways to make a right-handed crossover from β_3 to β_4, as shown in Fig. 3 c, d. Since β_1 was more hydrophobic and β_3 more polar, we

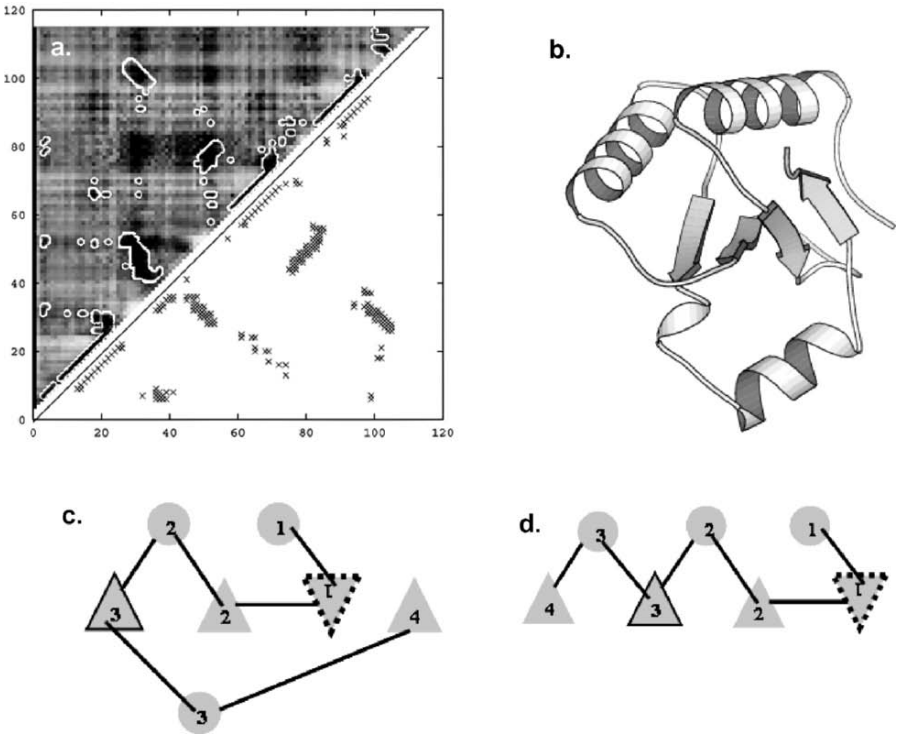


Fig. 7. **a** *Upper triangle* Contact potential map for HI0073 showing predicted contacts as *white outlines*. *Darker* means lower energy, $E(i,j)$. *Lower triangle* True contacts. **b** Molscript drawing of the crystal structure of HI0073, a hypothetical protein from *Haemophilus influenzae*. **c** Correct TOPS diagram showing non-polar strand (*dashed*) buried. **d** Incorrect TOPS diagram, consistent with all rules except strand burial rule

paired β_1 and β_4. All other β contacts to β_1 and contacts between α_2 and α_3 were erased.

4. α_1 must be on the opposite side of the sheet from α_3, since α_3 extends across the sheet. Therefore, contacts were assigned between α_1 and α_2 and erased between α_1 and α_3.

The completed TOPS diagram and contact map accurately match the true structure (Fig. 7b). The contact map prediction has 42 % contact coverage and 29 % accuracy. However, accuracy and coverage are not good measures of the quality of a contact map prediction, since near-contacts and gross errors are counted equally. Most of the false positive contacts in the HI0073 prediction are adjacent to true contacts. If we count near misses (±1 residue), then the coverage is 75 % and the accuracy is 57 %. Note that the long-range contacts between the β_1 and β_4 were correctly predicted, which speaks to the power of rule-based methods over raw statistics.

Identification of the folding nucleation site is the critical step in this approach. Once the nucleation site is chosen, the subsequent contact assignments are often unambiguous. After assigning secondary structures and choosing $\beta_2\alpha_2\beta_3$ as the nucleation site, only one folding pathway was possible, and it leads to the correct structure (Fig. 7 c). It is interesting to note that this pathway also predicts a possible misfolded state (Fig. 7d). At step 3 in the pathway, a critical decision is made that depends on the sequences of strands one and three. If strand one was more polar and strand three more hydrophobic, then the alternative structure would be predicted. A simple mutation experiment might tell us whether our model is on the right track.

The choice of the nucleation site in HI0073 was relatively easy. Only one of the three potential βαβ units had a high score. The hairpin between β_1 and β_2 would also be a correct choice, but the selection of $\beta_2\alpha_1\beta_3$ eliminated more of the potential incorrect folding pathways.

4.5.4 False Prediction Using the Folding Pathway. What Went Wrong?

The KaiA N-terminal domain from *S. elongatus* (PDB code 1M2E) is an example where we chose the wrong nucleation site. KaiA is 135 residues long and has five β and five α units. From its contact potential, two possible nucleation sites could be identified, $\beta_2\alpha_2\beta_3$, or $\beta_3\alpha_3\beta_4$. We chose $\beta_2\alpha_2\beta_3$ as the nucleation site instead of the correct, and higher scoring, $\beta_3\alpha_3\beta_4$ unit in order to favor a region of non-local high confidence contacts between β_1 and β_3 and between β_1 and β_4. Our mistake was in assigning non-local contacts before assigning local ones. If we had chosen the correct nucleation site, $\beta_3\alpha_3\beta_4$, there would be an unambiguous choice of the N-terminal βαβαβ segment. This sequence of five secondary structures is most commonly found in a three-stranded parallel sheet, and since in this case β_2 is polar and β_3 already pairs with another strand, only β_1 could be placed in the middle of the sheet. This would have given the correct 2134 strand order (Fig. 8a), and the helices would have been

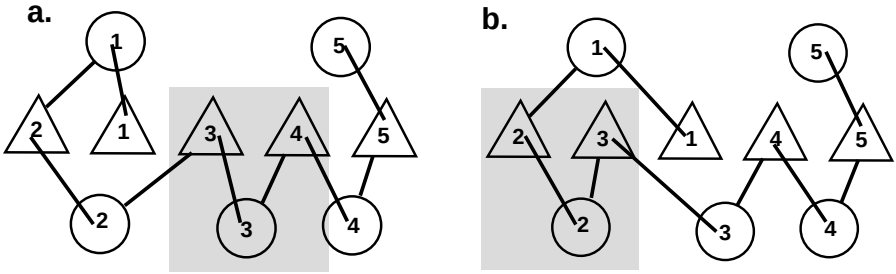


Fig. 8. **a** Correct TOPS diagram for KaiA, generated using the pathway described in the text using the *shaded* βαβ unit as the nucleation site. **b** Incorrect TOPS diagram, similar to the actual prediction, generated using a similar pathway but starting with the wrong nucleation site (*shaded*)

correctly placed according to our propagation rules (particularly the right-handed crossover rule). Our erroneous choice of the nucleation site led to the incorrect strand order 2314 (Fig. 8b), instead of 2134. For the record, here is the correct pathway for KaiA using HMMSTR-CM:

1. Nucleation site at $\beta_3\alpha_3\beta_4$.
2. The N-terminal parallel $\beta\alpha\beta\alpha\beta$ unit must have β_1 in the middle, since β_2 is polar and β_3 cannot be in the middle. To satisfy the right-handed crossover rule, α_2 must be on the same side of the sheet as α_3.
3. β_5 must pair with β_4 since it cannot pair with β_2, due to crossovers on both sides of the sheet.
4. α_5 must go on the same side of the sheet as α_1, due to helix crowding on the other side.

For other targets, pathway construction and CFE score alignment methods failed if the secondary structure prediction was inaccurate. In several targets, including HIP1R N-terminal domain from rat, an all-helix protein, secondary structure prediction by HMMSTR significantly under-predicted the helices. The wrong secondary structure pattern led to the wrong assignment of contact potentials, and therefore the wrong assumption of possible topologies. Under-prediction of helices was identified as a problem in HMMSTR.

4.6 Future Directions for HMMTR-CM

By gaining insight about how different parts of the protein pack together, we can improve the accuracy of the ab initio method. This will be necessary to make the whole prediction process automatic. The rule-based pathway approach depends on the correct assignment of the fold class of the target (all-α, α/β, $\alpha+\beta$ or all β (Zhou 1998)), since the rules of propagation depend on choices of the final topology. Generally this assignment is not difficult. So far, it has been applied only to the α/β class, but a different set of rules may be envisioned for the packing of helices and all β proteins.

The difficulty of choosing the correct nucleation site increases with protein size, since there are more to choose from. For larger proteins, more than one correct choice may be required. One possible approach could be a recursive algorithm to exhaust all the possible topologies by starting with each potential nucleation site, and then evaluate the topologies using the contact potential.

Another improvement might be to attempt to make the contact map prediction more protein- like. Our predictions have many false contacts adjacent to true contacts, e.g. a "fat" β-hairpin prediction – even though it is predicted at the right position. Rules to prune this type of false contacts – in other words, to beautify the predicted contact blocks – would increase the accuracy of our prediction. This will require better secondary structure predictions.

5 Conclusions

We have developed methods for calculating an inter-residue contact potential map for a protein sequence, for aligning that map to templates, and for pruning that map using a folding pathway model. Results on CASP5 targets reveal that the folding pathways for some α/β proteins are unambiguous given the correct choice of the folding nucleation site. Pathway predictions improved the selection of a remote homologue for one threading target. Consensus contact maps are more complete than maps from single templates. The contact map representation of a protein structure is a useful intermediate-level detail that facilitates rule-based algorithm development.

References

Alm E, Baker D (1999) Prediction of protein-folding mechanisms from free-energy landscapes derived from native structures. Proc Natl Acad Sci USA 96:11305–11310

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Anfinsen CB, Scheraga HA (1975) Experimental and theoretical aspects of protein folding. Adv Protein Chem 29:205–300

Aszodi A, Munro RE, Taylor WR (1997) Distance geometry based comparative modelling. Fold Des 2:S3–S6

Baldwin RL (1995) The nature of protein folding pathways: the classical versus the view. J Biomol NMR 5:103–109

Baldwin RL, Rose GD (1999) Is protein folding hierarchic? I. Local structure and peptide folding. Trends Biochem Sci 24:26–33

Blanco FJ Rivas G, Serrano L (1994) A short linear peptide that folds into a native stable beta-hairpin in aqueous solution. Nat Struct Biol 1:584–590

Bonneau R, Baker D (2001) Ab initio protein structure prediction: progress and prospects. Annu Rev Biophys Biomol Struct 30:173–189

Bonneau R, Strauss CE, Baker D (2001) Improving the performance of Rosetta using multiple sequence alignment information and global measures of hydrophobic core formation. Proteins 43:1–11

Brunger AT, Clore GM, Gronenborn AM, Karplus M (1986) Three-dimensional structure of proteins determined by molecular dynamics with interproton distance restraints: application to crambin. Proc Natl Acad Sci USA 83:3801–3805

Bystroff C, Baker D (1997) Blind predictions of local protein structure in CASP2 targets using the I-sites library. Proteins (Suppl 1):167–171

Bystroff C, Baker D (1998) Prediction of local structure in proteins using a library of sequence-structure motifs. J Mol Biol 281:565–577

Bystroff C, Garde S (2003) Helix propensities of short peptides: Molecular dynamics versus bioinformatics. Proteins 50:552–562

Bystroff C, Shao Y (2002) Fully automated ab initio protein structure prediction using I-SITES, HMMSTR and ROSETTA. Bioinformatics 18 (Suppl 1):S54–S61

Bystroff C, Simons KT, Han KF, Baker D (1996) Local sequence-structure correlations in proteins. Curr Opin Biotechnol 7:417–421

Bystroff C, Thorsson V, Baker D (2000). HMMSTR: A hidden markov model for local sequence-structure correlations in proteins. J Mol Biol 301:173–190

Cavalli A, Ferrara P, Caflisch A (2002) Weak temperature dependence of the free energy surface and folding pathways of structured peptides. Proteins 47:305–314
Chan HS, Bromberg S, Dill KA (1995) Models of cooperativity in protein folding. Philos Trans R Soc Lond B Biol Sci 348:61–70
Colon W, Roder H (1996) Kinetic intermediates in the formation of the cytochrome c molten globule. Nat Struct Biol 3:1019–1025
Crippen GM, Havel TF (1988) Distance geometry and molecular conformation. Chemometrics Series, 15. Wiley, New York
Duan Y, Kollman PA (1998) Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution [see comments]. Science 282:740–744
Dyson HJ, Wright PE (1996) Insights into protein folding from NMR. Annu Rev Phys Chem 47:369–395
Eaton WA, Thompson PA, Chan CK, Hage SJ, Hofrichter J (1996) Fast events in protein folding. Structure 4:1133–1139
Eddy SR (1996) Hidden Markov models. Curr Opin Struct Biol 6:361–365
Efimov AV (1993) Standard structures in proteins. Prog Biophys Mol Biol 60:201–39
Fariselli P, Casadio R (1999) A neural network based predictor of residue contacts in proteins. Protein Eng 12:15–21
Fariselli P, Olmea O, Valencia A, Casadio R (2001) Progress in predicting inter-residue contacts of proteins with neural networks and correlated mutations. Proteins (Suppl 5):157–162
Fersht AR (1995) Optimization of rates of protein folding: the nucleation-condensation mechanism and its implications. Proc Natl Acad Sci USA 92:10869–10873
Fersht AR, Matouschek A, Serrano L (1992) The folding of an enzyme. I. Theory of protein engineering analysis of stability and pathway of protein folding. J Mol Biol 224:771–782
Fischer D, Elofsson A, Rychlewski L, Pazos F, Valencia A, Rost B, Ortiz AR, Dunbrack RL Jr (2001) CAFASP2: the second critical assessment of fully automated structure prediction methods. Proteins (Suppl 5):171–183
Frishman D, Argos P (1995) Knowledge-based protein secondary structure assignment. Proteins 23:566–579
Galzitskaya OV, Ivankov DN, Finkelstein AV (2001) Folding nuclei in proteins. FEBS Lett 489:113–118
Garcia AE, Sanbonmatsu KY (2001) Exploring the energy landscape of a beta hairpin in explicit solvent. Proteins 42:345–354
Gillespie JR, Shortle D (1997) Characterization of long-range structure in the denatured state of staphylococcal nuclease. II. Distance restraints from paramagnetic relaxation and calculation of an ensemble of structures. J Mol Biol 268:170–184
Gnanakaran S, Garcia AE (2002) folding of a highly conserved diverging turn motif from the SH3 domain. Biophys J
Gough J, Chothia C (2002) SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. Nucleic Acids Res 30:268–272
Grantcharova VP, Riddle DS, Baker D (2000) Long-range order in the src SH3 folding transition state. Proc Natl Acad Sci USA 97:7084–7089
Gulotta M, Gilmanshin R, Buscher TC, Callender RH, Dyer RB (2001) Core formation in apomyoglobin: probing the upper reaches of the folding energy landscape. Biochemistry 40:5137–5143
Han KF, Baker D (1996) Global properties of the mapping between local amino acid sequence and local structure in proteins. Proc Natl Acad Sci USA 93:5814–5818
Han KF, Baker D (1995) Recurring local sequence motifs in proteins. J Mol Biol 251:176–187

Han KF, Bystroff C, Baker D (1997) Three-dimensional structures and contexts associated with recurrent amino acid sequence patterns. Protein Sci 6:1587–1590
Heidary DK, Jennings PA (2002) Three topologically equivalent core residues affect the transition state ensemble in a protein folding reaction. J Mol Biol 316:789–798
Hobohm U, Sander C (1994) Enlarged representative set of protein structures. Protein Sci 3:522–524
Houry WA, Rothwarf DM, Scheraga HA (1996) Circular dichroism evidence for the presence of burst-phase intermediates on the conformational folding pathway of ribonuclease A. Biochemistry 35:10125–10133
Hu J, Shen X, Shao Y, Bystroff C, Zaki MJ (2002) BIOKDD 2002, Edmonton, Canada
Jacchieri SG (2000) Mining combinatorial data in protein sequences and structures. Mol Divers 5:145–152
Jones DT (1998) Critical assessment of protein structure prediction 3, Asilomar, California
Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637
Karplus K, Barrett C, Hughey R (1998) Hidden Markov models for detecting remote protein homologies. Bioinformatics 14:846–856
Kolinski A, Skolnick J (1997) High coordination lattice models of protein structure, dynamics and thermodynamics. Acta Biochim Pol 44:389–422
Krueger BP, Kollman PA (2001) Molecular dynamics simulations of a highly charged peptide from an SH3 domain: possible sequence-function relationship. Proteins 45:4–15
Laurents DV, Baldwin RL (1998) Protein folding: matching theory and experiment. Biophys J 75:428–434
Mateu MG, Sanchez Del Pino MM, Fersht AR (1999) Mechanism of folding and assembly of a small tetrameric protein domain from tumor suppressor p53. Nat Struct Biol 6:191–198
Mendes J, Guerois R, Serrano L (2002) Energy estimation in protein design. Curr Opin Struct Biol 12:441–446
Mirny L, Shakhnovich E (2001) Protein folding theory: from lattice to all-atom models. Annu Rev Biophys Biomol Struct 30:361–396
Mok YK, Elisseeva EL, Davidson AR, Forman-Kay JD (2001) Dramatic stabilization of an SH3 domain by a single substitution: roles of the folded and unfolded states. J Mol Biol 307:913–928
Mok YK, Kay CM, Kay LE, Forman-Kay J (1999) NOE data demonstrating a compact unfolded state for an SH3 domain under non-denaturing conditions. J Mol Biol 289:619–638
Moult J, Fidelis K, Zemla A, Hubbard T (2001) Critical assessment of methods of protein structure prediction (CASP): round IV. Proteins (Suppl 5):2–7
Munoz V, Blanco FJ, Serrano L (1995) The hydrophobic-staple motif and a role for loop-residues in alpha- helix stability and protein folding. Nat Struct Biol 2:380–385
Munoz V, Henry ER, Hofrichter J, Eaton WA (1998) A statistical mechanical model for beta-hairpin kinetics. Proc Natl Acad Sci USA 95:5872–5879
Nolting B, Andert K (2000) Mechanism of protein folding. Proteins 41:288–298
Nolting B, Golbik R, Neira JL, Soler-Gonzalez AS, Schreiber G, Fersht AR (1997) The folding pathway of a protein at high resolution from microseconds to seconds. Proc Natl Acad Sci USA 94:826–830
Northey JG, Di Nardo AA, Davidson AR (2002a) Hydrophobic core packing in the SH3 domain folding transition state. Nat Struct Biol 9:126–130
Northey JGB, Maxwell KL, Davidson AR (2002b) Protein folding kinetics beyond the Phi value: Using multiple amino acid substitutions to investigate the structure of the SH3 domain folding transition state. J Mol Biol 320:389–402

Nymeyer H, Socci ND, Onuchic JN (2000) Landscape approaches for determining the ensemble of folding transition states: success and failure hinge on the degree of frustration. Proc Natl Acad Sci USA 97:634–639
Oliveberg M, Tan YJ, Silow M, Fersht AR (1998) The changing nature of the protein folding transition state: implications for the shape of the free-energy profile for folding. J Mol Biol 277:933–943
Olmea O, Valencia A (1997) Improving contact predictions by the combination of correlated mutations and other sources of sequence information. Fold Des 2:S25–S32
Onuchic JN, Luthey-Schulten Z, Wolynes PG (1997) Theory of protein folding: the energy landscape perspective. Annu Rev Phys Chem 48:545–600
Ortiz AR, Kolinski A, Skolnick J (1998) Fold assembly of small proteins using monte carlo simulations driven by restraints derived from multiple sequence alignments. J Mol Biol 277:419–448
Ortiz AR, Skolnick J (2000) Sequence evolution and the mechanism of protein folding. Biophys J 79:1787–1799
Pande VS, Grosberg A, Tanaka T, Rokhsar DS (1998) Pathways for protein folding: is a new view needed? Curr Opin Struct Biol 8:68–79
Plaxco KW, Simons KT, Baker D (1998) Contact order, transition state placement and the refolding rates of single domain proteins. J Mol Biol 277:985–994
Pollastri G, Baldi P (2002) Prediction of contact maps by GIOHMMs and recurrent neural networks using lateral propagation from all four cardinal corners. Bioinformatics 18 (Suppl 1):S62–S70
Rabiner LR (1989) A tutorial on hidden Markov models and selected applications in speech recognition. Proc IEEE 77:257–286
Rooman MJ, Rodriguez J, Wodak SJ (1990) Automatic definition of recurrent local structure motifs in proteins. J Mol Biol 213:327–36
Rost B (2001) Review: protein secondary structure prediction continues to rise. J Struct Biol 134:204–218
Schellman C (1980) Protein folding: the proceedings of the 28th Conference of the German Biochemical Society, University of Regensburg, Sept 10–12, 1979, Regensburg
Shakhnovich EI. (1998). Folding nucleus: specific or multiple? Insights from lattice models and experiments. Fold Des 3:R108–R111 (discussion R107)
Shao Y, Bystroff C (2003) Predicting inter-residue contacts using templates and pathways. Proteins, structure, function and genetics 53 Suppl 6:497–502
Shea JE, Brooks CL 3rd (2001) From folding theories to folding proteins: a review and assessment of simulation studies of protein folding and unfolding. Annu Rev Phys Chem 52:499–535
Shoemaker BA, Wolynes PG (1999) Exploring structures in protein folding funnels with free energy functionals: the denatured ensemble. J Mol Biol 287:657–674
Simons KT, Bonneau R, Ruczinski I, Baker D (1999a) Ab initio protein structure prediction of CASP III targets using ROSETTA. Proteins (Suppl 3):171–176
Simons KT, Kooperberg C, Huang E, Baker D (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol 268:209–225
Simons KT, Ruczinski I, Kooperberg C, Fox BA, Bystroff C, Baker D (1999b) Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. Proteins 34:82–95
Singer MS, Vriend G, Bywater RP (2002) Prediction of protein residue contacts with a PDB-derived likelihood matrix. Protein Eng 15:721–725
Skolnick J, Kolinski A (2002) A unified approach to the prediction of protein structure and function. In: Computational methods for protein folding, vol 120. pp 131–192
Sternberg MJ, Thornton JM (1976) On the conformation of proteins: the handedness of the beta-strand-alpha-helix-beta-strand unit. J Mol Biol 105:367–382

Steward RE, Thornton JM (2002) Prediction of strand pairing in antiparallel and parallel beta- sheets using information theory. Proteins-Structure Function and Genetics 48:178–191

Thirumalai D, Klimov DK (1998) Fishing for folding nuclei in lattice models and proteins. Fold Des 3:R112–R118 (discussion R107)

Vendruscolo M, Kussell E, Domany E (1997) Recovery of protein structure from contact maps. Fold Des 2:295–306

Viguera AR, Serrano L (1995) Experimental analysis of the Schellman motif. J Mol Biol 251:150–160

Woolfson DN, Alber T (1995) Predicting oligomerization states of coiled coils. Protein Sci 4:1596–1607

Yi Q, Bystroff C, Rajagopal P, Klevit RE, Baker D (1998) Prediction and structural characterization of an independently folding substructure in the src SH3 domain. J Mol Biol 283:293–300

Zaki MJ, Shan J, Bystroff C (2000) Proceedings IEEE International Symposium on Bio-Informatics and Biomedical Engineering, Arlington, Virginia

Zhou GP (1998) An intriguing controversy over protein structural class prediction. J Protein Chem 17:729–738

Zhu J, Liu JS, Lawrence CE (1998) Bayesian adaptive sequence alignment algorithms. Bioinformatics 14:25–39

Zwanzig R (1997) Two-state models of protein folding kinetics. Proc Natl Acad Sci USA 94:148–150

Structural Bioinformatics and NMR Structure Determination

J.P. Linge, M. Nilges

1 Introduction: NMR and Structural Bioinformatics

It has become common ground to start a bioinformatics article by mentioning the flood of data overwhelming the research community. Indeed, the large amount of data has led to the insight that even the average wet lab needs several computers, e.g. to manage micro array results or to run BLAST searches via the internet. However, it is more than that: bioinformatics has matured to a research discipline in its own right. The main reason is that bioinformatics not only allows for the solving of problems that are tedious in a traditional approach (e.g. protein function can often be inferred from homologous proteins in other species), but contributes to a new way of looking at biological systems: from a reductionist approach, to a systemic view of biology (Noble 2002). With the large amounts of data on whole systems, the focus in biomedical research is increasing steadily: from a single protein to complexes, from an enzyme-catalyzed reaction to metabolic networks. Even virtual cells or tissues are no longer science fiction.

Recently, the term bioinformatics has been used more and more to describe research related to databases (integration of resources, web access, etc.) and sequence analysis (homology searches, multiple sequence alignments, phylogenetic trees). Computational biology refers to the simulation of complex networks, e.g. metabolic or signalling pathways, cell and tissue simulations. In recent years, the field has seen an -omics explosion (e.g. genomics, proteomics, transcriptomics, metabolomics), five of which have created several research communities eager to integrate their data.

Structural bioinformatics focuses on the relationship between sequence, three-dimensional structure, and the function of proteins and other biologi-

J.P. Linge, M. Nilges
Unité de Bio-Informatique Structurale, Institut Pasteur, 25–28 rue du docteur Roux, 75015 Paris, France

Nucleic Acids and Molecular Biology, Vol. 15
Janusz M. Bujnicki (Ed.)
Practical Bioinformatics

cal macromolecules, using, among others, modelling techniques and molecular dynamics (MD) simulations.

2 Algorithms for NMR Structure Calculation

Molecular modelling has a central position in the derivation of NMR solution structures. Experimental data are sparse and measurable for only a fraction of the atoms (mostly the hydrogens). Most of the data describe relative positions of atoms, and do not directly correspond to the global conformation of the molecule. It is therefore necessary to use additional data (prior information) valid for all molecules of the same type. This information can be derived from molecular dynamics (MD) or molecular mechanics force fields.

Algorithm development for structure calculation is very important since models are not manually built, but automatically calculated. The methods used for NMR structure calculations originally came from other fields of structural bioinformatics or computational biology. Distance geometry methods have been developed with the aim of predicting protein 3D structure. Nonlinear optimisation usually employs MD algorithms.

Since several good reviews exist on the different methods, for example for distance geometry (Braun 1987; Havel et al. 1983; Havel 1991), for simulated annealing in Cartesian coordinates (Brünger and Nilges 1993; Brünger et al. 1997; Nilges and O'Donoghue 1998), and for torsion angle dynamics (Güntert 1998; Brünger and Adams 2002), we discuss these approaches only very briefly.

2.1 Distance Geometry and Data Consistency

Distance geometry was a natural choice for structure calculations, since the principal data are distances derived from NOE measurements. Most of the prior knowledge can also be expressed in form of distances (van der Waals radii, bond lengths, etc.), with the exception of chiral information. All data points are only known with some uncertainty; there are for example experimental errors, and the measured quantities (e.g., the NOEs) are converted to structural parameters (inter-proton distances) by an approximate theory (the isolated spin pair approximation). The uncertainties for all data points are expressed in the form of lower and upper distance bounds.

Distance geometry provides a framework to analyse the distance bounds in terms of their geometric consistency and checks the bounds systematically by applying the triangle inequalities, before the structures calculation itself. The triangle inequalities are a fundamental property of distances, and the bounds need to be set generously enough to ensure that they can be satisfied. The structure calculation itself is initiated by constructing a complete distance

matrix by selecting distances from within the bounds (Kuszewski et al. 1992; Havel 1991; Hodsdon et al. 1997). Structures are then calculated with tools from multivariate data analysis (principal coordinate analysis). The resulting approximate structure is the starting point for further refinement by non-linear optimisation techniques.

2.2 Nonlinear Optimisation

With the arrival of powerful minimisation strategies based on the idea of simulated annealing, it became less important to obtain an approximate starting structure. The distance geometry concept of lower and upper bounds, in particular for distances derived from NOEs, is generally maintained in optimisation. It is incorporated into the optimisation via an appropriate pseudo-potential. The total energy that is minimised combines a physical energy term (similar to an MD force field) and pseudo-potentials penalising deviations from experimental data. If the bounds have been set sufficiently wide, the exact values of the relative weights in the experimental terms in the energy function are not very important.

The difficulties of applying minimisation strategies directly to the problem of calculating NMR structures are due to the complicated energy landscape produced by the physical force field and the experimental terms, having deep local energy minima separated by high barriers. Methods inspired by simulated annealing (Kirkpatrick et al. 1983) can overcome energy barriers. Because of the high degree of correlation between the principal degrees of freedom (the torsion angles), Monte Carlo based methods do not converge very well for biological macromolecules, and the simulated annealing algorithms in NMR structure determination and X-ray crystal structure refinement are usually based on molecular dynamics. The key to using MD as a minimisation tool is temperature control (e.g. Berendsen et al. 1984). Minimisation schemes usually employ additional methods to overcome energy barriers, for example the scaling of the non-bonded energy (Nilges and O'Donoghue 1998).

In Cartesian coordinates, simulated annealing with MD consists of numerically solving Newton's equations of motion with forces derived from the physical energy term and the pseudo-potential incorporating the experimental data. The resulting trajectory obviously does not reflect the dynamics of the molecule, but is only a means to minimisation. Using torsion angles as only degrees of freedom has decisive advantages, since the covalent geometry of the molecule is automatically maintained, permitting the use of longer time steps. The nature of the MD equations becomes much more complicated, with an explicit dependency of the angular acceleration on angular velocities, and a time-dependent non-diagonal mass matrix. Torsion angle dynamics shows, in general, better convergence than Cartesian dynamics.

2.3 Sampling Conformational Space

All structure calculation methods have a random element. For the simulated annealing methods, the initial coordinates and the initial velocities are set randomly. In distance geometry using the metric matrix approach, distances are randomly chosen within their bounds, and the distances can be chosen in random sequence (Kuszewski et al. 1992; Havel 1991; Hodsdon et al. 1997). The standard procedure is then to use the calculation with identical data set at several times, varying only the random number seed for initial coordinate or distance generation. In this way, the conformational space consistent with the data is sampled, to test if the data determines the structure. The result is a more or less distributed structure ensemble.

This distribution depends on many factors, for example the choice of distance bounds (Chalaoux et al. 1999), the shape of the restraint potential, the calculation method (metrisation in distance geometry, simulated annealing schedule, etc.). The limitations of this empirical procedure are well recognised, and re-sampling strategies have been suggested (Spronk et al. 2003).

2.4 Modelling Structures with Limited Data Sets

There is great interest in methods to reduce the amount of data necessary to obtain an NMR structure, in order to extend the NMR methodology to larger proteins (Mueller et al. 2000) and to speed up the structure determination process for its use in structural genomic efforts (Prestegard et al. 2001). Even with automated data analysis (seeSect. 5 below), the time required for structure determination by traditional NMR methods is still too long.

NMR data such as residual dipolar couplings (Prestegard et al. 2000; Bax et al. 2001) have an important impact on speeding up structure determination. In fortunate cases the residual dipolar couplings alone may be sufficient to determine the fold (Hus et al. 2001). Searching databases with threading techniques is a very promising method (Andrec et al. 2001, 2002) in reducing the requirement of the completeness of the data. Also, the chemical shifts of the NMR-active nuclei may be sufficient to predict the fold of a structure by threading the secondary structure derived from the chemical shifts against a structural data base (Ayers et al. 1999). Residual dipolar couplings also offer an alternate approach to simultaneous resonance assignment and structure determination of protein backbones (Tian et al. 2001).

Further development of methods combining database searches, molecular modelling, and NMR data will lead to increasingly reliable NMR structures from minimal data sets. The use of sparse NMR data in combination with ab initio protein 3D structure prediction algorithms can significantly reduce the amount of necessary data (Bowers et al. 2000; Rohl and Baker 2002). The future will show if structures based on very limited data sets can be

made accurate enough to allow for detailed structural analysis beyond fold assignment.

3 Internal Dynamics and NMR Structure Determination

3.1 Calculating NMR Parameters from Molecular Dynamics Simulations

NMR is a rich source of structural data (inter-atomic distances, angles, and orientations), However, the sensitivity of all structural data obtainable from NMR experiments to internal dynamics makes them particularly difficult to interpret in structural terms. The measured quantities are averages over time and a large ensemble of structures, while in a standard structure calculation the lower and upper bounds refer to instantaneous distances. The calculation of NMR parameters was one of the first applications of MD simulations, and the simulations often play a central role in the analysis of biomolecular NMR data (see recent reviews by Case 2002; Brüschweiler 2003).

MD simulations on peptides are of great interest, since one can perform fully solvated simulations in the hundreds of nanoseconds range, and the peptides may show very complicated dynamics, including reversible folding (Daura et al. 1999; Peter et al. 2001). In these cases, the interpretation of NMR relaxation data is particularly difficult, and the use of standard methods for structure determination is bound to fail. However, the direct back calculation of the complete spectra is possible from long MD simulations, therefore, there is no need to separate internal and rotational dynamics. The NMR parameters predicted from the simulation can be directly compared to the experiments.

3.2 Inferring Dynamics from NMR Data

In terms of structure refinement, this is somewhat unsatisfactory, since one has to rely entirely on the accuracy of the MD simulations in atomic detail, and the experimental data do not enter directly. Simulations of sufficient length are still impossible for larger biological macromolecules. Inferring dynamics from the (sparse) experimental data during a structure calculation is difficult, since we require additional data to characterise the molecular motions in addition to the structure. The structure ensembles generated by standard structure calculation methods (see the section on sampling conformational space above) are sometimes taken as a reflection of the dynamics the molecule shows in solution. There are indeed similarities to independently performed dynamics simulations (Abseher et al. 1998) and experimental relaxation parameters (Redfield et al. 1992). Simple contact models suffice to predict relaxation parameters (Zhang and Brüschweiler 2002; Haliloglu

and Bahar 1999) with similar quality as detailed MD simulations (Philippopoulos et al. 1997). This suggests that this resemblance is caused mostly by non-specific interactions: On the surface of the molecule, atoms have more conformational freedom since there are fewer experimental restraints in the NMR refinement, and also fewer non-bonded contacts.

Refinements with an ensemble of structures (ensemble refinement) or a trajectory (time-averaged refinement) attempt to account for the conformational averaging (reviewed by Bonvin et al. 1993a). Care has to be taken to avoid over-fitting of the data, for example by cross-validation (Bonvin and Brünger 1995). Clearly, the precise effect of local dynamics on the NMR data cannot be determined from the data, and MD simulations have shown that simple conformational averages are an oversimplification (Brüschweiler et al. 1992; Schneider et al. 1999; Peter et al. 2001). It is however difficult to integrate this knowledge with the experimental data into one consistent picture of a dynamic structure. One can use correction factors derived from MD simulations (Bonvin et al. 1993b) and other theoretical calculations, such as normal modes (Brüschweiler and Case 1994). A general framework for the interpretation of relaxation data from nonfolded and folded proteins has been developed, using structures generated from MD simulations and principal component analysis (Prompers and Brüschweiler 2002).

The major problem with NOE data is that structural and dynamic effects cannot be separated. Residual dipolar couplings, in contrast, offer new ways to analyse the internal dynamics of macromolecules. Here, the effects of structure and dynamics can be separated to first order, and thus a simultaneous extraction of structural and motional parameters from residual dipolar coupling data becomes possible (Tolman et al. 2001). Alternatively, a simultaneous analysis of results from many alignment media can yield dynamic properties directly from the data (Meiler et al. 2003).

4 Structure Validation

The increased speed of structure determination necessary for the structural genomics projects makes an independent validation of the structures (by comparison to expected properties) particularly important (reviewed by Laskowski et al. 1998). Structure validation helps to correct obvious errors (e.g. in the covalent structure) and leads to a more standardised representation of structural data, e.g. by agreeing on a common atom-name nomenclature. The knowledge of the structure quality is a prerequisite for further use of the structure, e.g. in molecular modelling or drug design.

The quality of structures is largely influenced by the quality of the data (Doreleijers et al. 1998) and the energy parameters used in the refinement (Linge et al. 2003 c). Validation programmes check the agreement of the three-dimensional structure with the experimental data; with the a priori informa-

tion used in the refinement (e.g. nonbonded contacts, covalent interactions); and evaluate structural properties that depend directly neither on the data nor the energy parameters, by comparing the structures to statistics derived from a database of solved protein structures.

5 Structural Genomics by NMR

While X-ray crystallographers can resolve a protein structure only hours after data collection at the synchrotron, high-throughput NMR still faces several technical problems. Data analysis and even storage of all the parameters involved in NMR structure determination are cumbersome. Despite recent efforts, chemical-shift assignment and NOE assignment are not yet fully automated.

5.1 Automated Assignment and Data Analysis

Most approaches to automated NMR structure determination require an independent chemical shift assignment as a first step. Several approaches exist to assign at least the backbone resonances automatically (cf. review by Moseley et al. 1999).

The major bottleneck in the analysis of NMR data and structure calculation is the assignment of NOESY cross-peaks. The large number of possible assignments for each peak, overlap and artefacts in the spectra render manual NOE assignment tedious. An important part of the automatic structure calculation is data analysis, since incorrect peaks have to be automatically recognised.

Several programmess for automated NOE assignment exist: CLOUDS (Grishaev and Llinas 2002); CANDID (Herrmann et al. 2002); NOAH (Mumenthaler 1997); AUTOSTRUCTURE (Montelione et al. 2000); and ARIA (Nilges and O'Donoghue 1998; Linge et al. 2001, 2003a). CLOUDS does not require independent chemical-shift assignment and is akin in spirit to direct methods in X-ray crystal structure determination.

Using the concept of ambiguous distance restraints, our own development, ARIA, automatically assigns NOEs in an iterative manner (see Fig. 1 for an overview). ARIA attempts to obtain optimal distance estimates in an efficient way, by employing a fast spin diffusion correction (Linge et al. 2003 c). This spin diffusion correction permits the use of tighter bonds for the distance restraints, facilitating the discrimination between signal and noise. ARIA 2.0 (Habeck et al. 2003) also provides for efficient communication with databases and supports the collaborative computing project for NMR (CCPN).

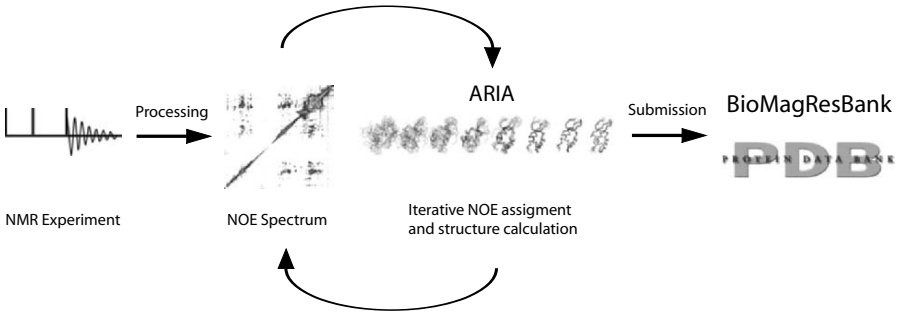


Fig. 1. Data flow in a typical NMR structure determination project using ARIA (reprinted with permission from Linge et al. 2003a). The most time-consuming step is the cycle of iterative NOE assignment and structure calculation

5.2 Collaborative Computing Project for NMR (CCPN)

The CCPN (Fogh et al. 2002) aims to provide services for NMR spectroscopists analogous to the highly successful CCP4 project for the X-ray community. CCPN develops a data model that covers all key areas of macromolecular NMR from the initial experimental data to the validation of the final structures. A data model is a description of the organisation of the data without reference to a particular format. The description of the data and their relationships are implemented in the UML language (see Fig. 2 for an example). CCPN provides software to automatically generate Application Programme Interfaces (APIs) for Python and C starting from the UML description. Programmes can directly store, access and share NMR data via the APIs. This facilitates data inter-change between NMR software, storage, and submission of NMR data to the PDB and BMRB databases, including 'data harvesting': all known information about a particular structure determination is carried forward from one program to another, through all the stages to the final deposition in the database. Eventually, programs will exchange data between each other via CCPN. Thus, the user himself does not need to write any scripts to convert input and output formats.

A first version of the data model is available. The CCPN software suite provides tools to process raw experimental data, analyse and assign spectra, and read and write input files for the most common programmes for spectra handling, assignment and structure calculation. The FormatConverter offers a GUI to manage several files (e.g. for chemical shift lists or cross peak assignments) at the same time.

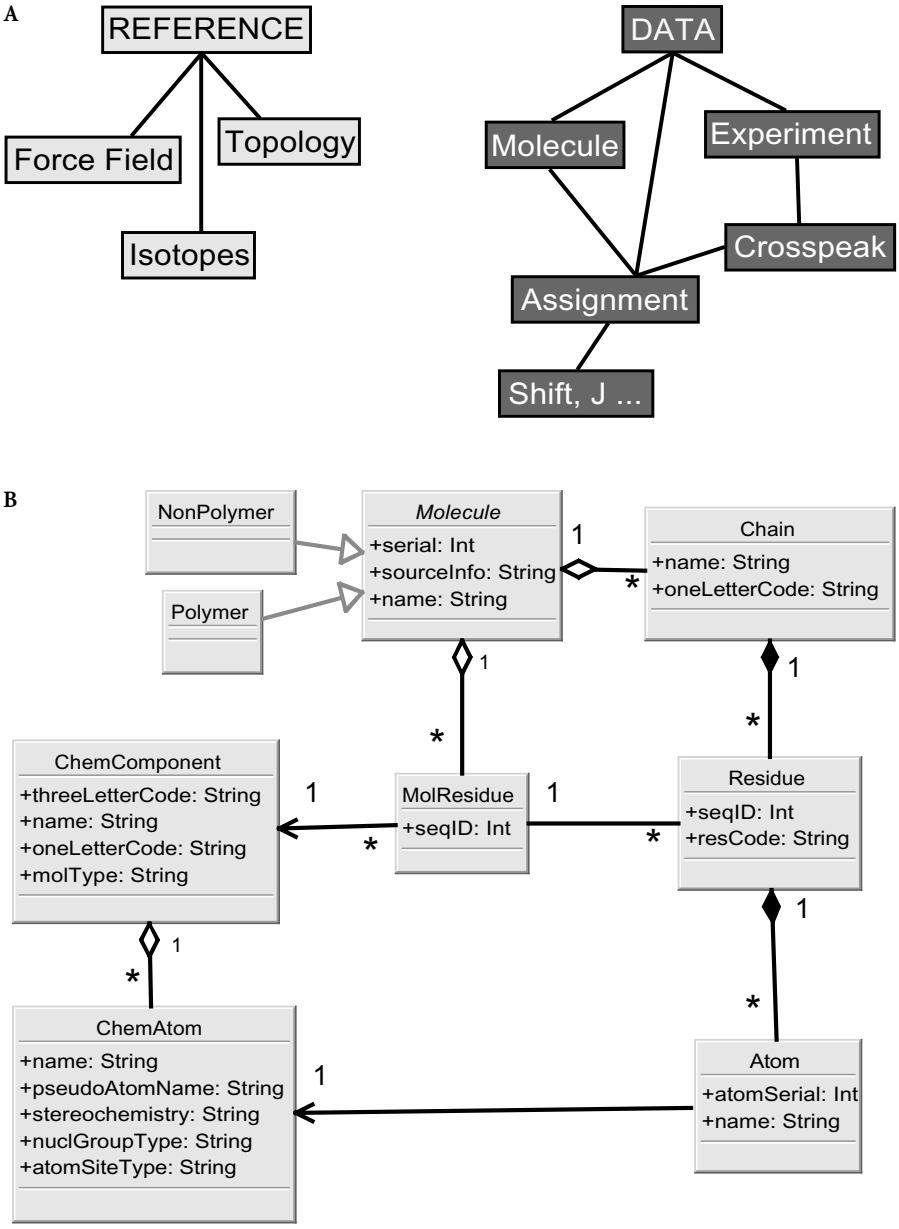


Fig. 2. **a** Overview of the CCPN data model. **b** Representation of a molecule within the CCPN data model. (Reprinted with permission from Fogh et al. 2002)

5.3 SPINS

SPINS (standardised protein NMR storage) (Baran et al. 2002) is a relational database system for NMR data organisation and archiving. It allows for the storing of all information from the raw NMR data to protein structures and offers tools to read and write BMRB compliant NMRStar files. SPINS is already being used in structural genomics projects.

6 Databanks and Databases

Databanks store biological raw data as repositories whereas databases provide additional annotation and functionality. Examples of databanks are GenBank (Benson et al. 2003) and EMBL (Stoesser et al. 2003) for primary DNA sequences and PDB (Berman et al. 2000) for protein structures. SwissProt (Boeckmann et al. 2003) and FlyBase (FlyBase Consortium 2003) are well-known databases which provide genetic and functional annotation. Examples of higher-level databases are PFAM (Bateman et al. 2002), SCOP (Murzin et al. 1995) and KEGG (Kanehisa et al. 2002). All of them have one feature in common: their fast growth. The aforementioned explosion of data can be quantified: DNA databases are currently doubling every 9 months. The PDB is expected to grow faster due to structural genomics efforts (3298 structures were deposited in 2001, 3381 structures in 2002 with a total of 19,623 entries at the end of 2002).

An unsolved problem is the integration of biological databases. Since each database only contains a subset of biological knowledge, databases have to be combined to gather all of the available information. Several methods to integrate biological databases exist, but technical challenges are enormous (cf. review by Stein 2003). Link integration is the most common integration method so far, as employed in the sequence retrieval system (SRS) (Zdobnov et al. 2002) and Entrez (Schuler et al. 1996). Severe problems are naming clashes (e.g. genes and gene products using the same name) and stale hyperlinks to outdated database entries. When trying to combine information from several resources, scientists have to access several web sites (often using "copy & paste" within different browser windows). Obviously, this approach is tedious and cannot be scaled up.

The underlying data models of the databases are changing quickly in order to account for new technological developments and to describe the data in more detail. Unfortunately, this creates additional problems when accessing their content (software has to be rewritten, etc.). Furthermore, each database uses its own vocabulary to describe molecular function or cellular localisation. Even the meaning of attributes such as protein function may be different, e.g. one database may annotate the protein function of the human Titin protein as muscle protein, whereas another database may describe its function as kinase.

Ontologies give hope in overcoming these problems. In information science, an ontology is an explicit formal specification of how to represent objects, concepts, entities that are assumed to exist in some area of interest, and the relationships among them. Ontologies provide sophisticated vocabulary to describe the key concepts. They do not integrate databases themselves, but serve as a basis to help in the merging of several databases.

A major problem is error propagation in databanks and databases. DNA sequences may contain frame shifts, deletions, contaminations from cloning vectors, etc., functional annotations may be unverified or outdated. PDB structures often use non-standard atom names. NMR restraint files often show a different atom-name nomenclature than their PDB structure counterparts. This compromises the overall quality and usefulness of the stored data. Without expert knowledge, a lot of time and money could be wasted.

6.1 BioMagResBank and PDB/RCSB

For NMR, the principal databases for storage of NMR experimental data and solved structures are the BioMagResBank, and the Protein Data Bank (PDB) curated by the Research Collaboratory for Structural Bioinformatics (RCSB). The BMRB stores all non-coordinate biomolecular NMR data (Doreleijers et al. 2003): chemical shifts, NOEs, coupling constants, residual dipolar couplings (RDCs), hydrogen exchange rates and protection factors, order parameters, atomic relaxation parameters, and molecular correlation times. The PDB is the central repository for all coordinates and also manages restraint files used for NMR structure calculation (Berman et al. 2000). Most journals require structures and NMR data to be published in PDB and BMRB.

Exploiting the databases, several methods for the prediction of chemical shifts, dihedral angles, secondary and tertiary structure have been developed. A well-known example is the TALOS programme (Cornilescu et al. 1999) for the empirical prediction of phi and psi backbone torsion angles. The method exploits a subset of high-resolution X-ray PDB structures for which accurate NMR chemical-shift data are available. Since the difference between chemical shifts and their corresponding random coil values is often correlated with protein secondary structure, TALOS is able to make quantitative predictions for phi and psi, using only secondary shift and sequence information.

7 Conclusions

NMR is unique in its ability to measure experimental data on both the structure and dynamics of biological macromolecules in solution at atomic resolution. NMR therefore provides valuable input for the functional characterisation of biological macromolecules. On the other hand, the interpretation of

the data benefits from bioinformatics infrastructures and tools. Databases relating structures and dynamics to NMR parameters are useful for interpreting new experimental data. They may reduce the time and the amount of data necessary for determining a structure or interpreting dynamics data. Methodological advances from the fields of protein 3D structure prediction and MD simulations have been essential for the development of structural biology by NMR in the last few decades.

Acknowledgements. J. P. L. thanks the Pasteur Institute and the SPINE network (EU 5th Framework programme, contract number QLG2-CT-2002–00988) for financial support. We thank Michael Habeck and Wolfgang Rieping for comments on the manuscript.

References

Abseher R, Horstink L, Hilbers CW, Nilges M (1998) Essential spaces defined by NMR structure ensembles and molecular dynamics simulation show significant overlap. Proteins 31:370–382

Andrec M, Du P, Levy RM (2001) Protein backbone structure determination using only residual dipolar couplings from one ordering medium. J Biomol NMR 21:335–347

Andrec M, Harano Y, Jacobson MP, Friesner RA, Levy RM (2002) Complete protein structure determination using backbone residual dipolar couplings and sidechain rotamer prediction. J Struct Funct Genomics 2:103–111

Ayers DJ, Gooley PR, Widmer-Cooper A, Torda AE (1999) Enhanced protein fold recognition using secondary structure information from NMR. Protein Sci 8:1127–1133

Baran MC, Moseley HN, Sahota G, Montelione GT (2002) SPINS: Standardized protein NMR storage. A data dictionary and object-oriented relational database for archiving protein NMR spectra. J Biomol NMR 24:113–121

Bateman A, Birney AE, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL (2002) The pfam protein families database. Nucleic Acids Res 30:276–280

Bax A, Kontaxis G, Tjandra N (2001) Dipolar couplings in macromolecular structure determination. Methods Enzymol 339:127–174

Benson KA, Karsch-Mizrachi I, Ostell,LDJJ, Wheeler DL (2003) Genbank. Nucleic Acids Res 31:23–27

Berendsen HJC, Postma JPM, van Gunsteren WF, DiNola A, Haak JR (1984) Molecular dynamics with coupling to an external bath. J Chem Phys 81:3684–3690

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The protein data bank. Nucleic Acids Res 28:235–242

Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M (2003) The SWISS-PROT protein knowledge base and its supplement TrEMBL in 2003. Nucleic Acids Res 31:365–370

Bonvin AMJJ, Brünger AT (1995) Do NOE distances contain enough distance information to access the relative populations of multiconformer structures? J Biomol NMR 5:72–76

Bonvin AMJJ, Boelens R, Kaptein R (1993a) Determination of biomolecular structures by NMR: Use of relaxation matrix calculations. In: van Gunsteren WF, Weiner PK,

Wilkinson J (eds) Computer simulation of biomolecular systems: theoretical and experimental applications, vol 2. Escom, Leiden, pp 407–440

Bonvin A, Rullmann AMJJ, Lamerichs JA, Boelens RM, Kaptein R (1993b) "Ensemble" iterative relaxation matrix approach: a new NMR refinement protocol applied to the solution structure of crambin. Proteins Struct Funct Genet 15:385–400

Bowers PM, Strauss CE, Baker D (2000) De novo protein structure determination using sparse NMR data. J Biomol NMR 18:311–318

Braun W (1987) Distance geometry and related methods for protein structure determination from NMR data. Q Rev Biophys 19:115–157

Brünger AT, Adams PD (2002) Molecular dynamics applied to X-ray structure refinement. Acc Chem Res 35:404–412

Brünger AT, Nilges M (1993) Computational challenges for macromolecular structure determination by X-ray crystallography and solution NMR-spectroscopy. Q Rev Biophys 26:49–125

Brünger AT, Adams PDL, Rice LM (1997) New applications of simulated annealing in X-ray crystallography and solution NMR. Struct 5:325–336

Brüschweiler R (2003) New approaches to the dynamic interpretation and prediction of NMR relaxation data from proteins. Curr Opin Struct Biol 13:175–183

Brüschweiler R Case DA (1994) Characterization of biomolecular structure and dynamics by NMR cross relaxation. Prog NMR Spec 26:27–58

Brüschweiler R, Roux B, Blackledge M, Griesinger C, Karplus M, Ernst RR (1992) Influence of rapid intramolecular motion of NMR cross-relaxation rates: a molecular dynamics study of antanamide in solution. J Am Chem Soc 114:2289–2302

Case DA (2002) Molecular dynamics and NMR spin relaxation in proteins. Acc Chem Res 35:325–331

Chalaoux FR, O'Donoghue SI, Nilges M (1999) Molecular dynamics and accuracy of NMR structures: effects of error bounds and data removal. Proteins Struct Funct Genet 34:453–463

Cornilescu G, Delaglio F, Bax A (1999) Protein backbone angle restraints from searching a database for chemical shift and sequence homology. J Biomol NMR 13:289–302

Daura X, Antes I, van Gunsteren WF, Thiel W, Mark AE (1999) The effect of motional averaging on the calculation of NMR derived structural properties. Proteins Struct Funct Genet 36:542–555

Doreleijers JF, Mading S, Maziuk D, Sojourner K, Yin L, Zhu J, Markley JL, Ulrich EL (2003) BioMagResBank database with sets of experimental NMR constraints corresponding to the structures of over 1400 biomolecules deposited in the Protein Data Bank. J Biomol NMR 26:139–146

Doreleijers JF, Rullmann JA, Kaptein R (1998) Quality assessment of NMR structures: a statistical survey. J Mol Biol 281:149–164

FlyBase Consortium (2003) The FlyBase database of the Drosophila genome projects and community literature. Nucleic Acids Res 31:172–175

Fogh RH, Ionides J, Ulrich E, Boucher W, Vranken W, Linge JP, Habeck M, Rieping W,Bhat TN, Westbrook J, Henrick K, Gilliland G, Berman H, Thornton J, Nilges M, Markley J, Laue E (2002). The CCPN project: an interim report on a data model for the NMR community. Nature Struct Biol 9:416–418

Grishaev A, Llinas M (2002) CLOUDS, a protocol for deriving a molecular proton density via NMR. Proc Natl Acad Sci USA 99:6707–6712

Güntert P (1998) Structure calculation of biological macromolecules from NMR data. Q Rev Biophys 31:145–237

Habeck M, Rieping W, Linge JP, Michael M (2003) NOE assignment with ARIA 2.0 - the nuts and bolts. In: Downing K (ed) Protein NMR techniques. Humana Press (in press)

Haliloglu T, Bahar I (1999) Structure-based analysis of protein dynamics: comparison of theoretical results for hen lysozyme with X-ray diffraction and NMR relaxation data. Proteins Struct Funct Genet 37:654–667

Havel TF (1991) An evaluation of computational strategies for use in the determination of protein structure from distance constraints obtained by nuclear magnetic resonance. Prog Biophys Mol Biol 56:43–78

Havel TF, Kuntz ID, Crippen GM (1983) The combinatorial distance geometry method for the calculation of molecular conformation. I. A new approach to an old problem. J Theor Biol 104:359–381

Herrmann T, Güntert P, Wüthrich K (2002) Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA. J Mol Biol 319:209–227

Hodsdon ME, Ponder JW, Cistola DP (1997) The NMR solution structure of intestinal fatty acid-binding protein complexed with palmitate: application of a novel distance geometry algorithm. J Mol Biol 264:585–602

Hus JC, Marion D, Blackledge M (2001) Determination of protein backbone structure using only residual dipolar couplings. J Am Chem Soc 123:1541–1542

Kanehisa M, Goto S, Kawashima S, Nakaya A (2002) The KEGG databases at Genomenet. Nucleic Acids Res 30:42–46

Kirkpatrick S, Gelatt CDJ, Vecchi MP (1983) Optimization by simulated annealing. Science 220:377–385

Kuszewski J, Nilges M, Brünger AT (1992) Sampling and efficiency of metric matrix distance geometry: a novel partial metrization algorithm. J Biomol NMR 2:33–56

Laskowski RA, MacArthur MW, Thornton JM (1998) Validation of protein models derived from experiment. Curr Opin Struct Biol 8:631–639

Linge JP, O'Donoghue SI, Nilges M (2001) Automated assignment of ambiguous nuclear Overhauser effects with ARIA. Methods Enzymol 339:71–90

Linge JP, Habeck M, Rieping W, Nilges M (2003a) ARIA: automated NOE assignment and NMR structure calculation. Bioinformatics 19:315–316

Linge JP, Williams MA, Spronk CA, Bonvin AMJJ, Nilges M (2003b) Refinement of protein structures in explicit solvent. Proteins Struct Funct Genet 20:496–506

Linge JP, Habeck M, Rieping |W, Nilges M (2003 c) Correction of spin diffusion during iterative automated NOE assignment. J Magn Reson (submitted)

Meiler J, Peti W, Griesinger C (2003) Dipolar couplings in multiple alignments suggest alpha helical motion in ubiquitin. J Am Chem Soc 125:8072–8073

Montelione GT, Zheng D, Huang YJ, Gunsalus KC, Szyperski T (2000) Protein NMR spectroscopy in structural genomics. Nat Struct Biol 7:982–985

Moseley HN, Montelione G T (1999) Automated analysis of NMR assignments and structures for proteins. Curr Opin Struct Biol 9:635–642

Mueller GA, Choy WY, Yang D, Forman-Kay JD, Venters RA, Kay LE (2000) Global folds of proteins with low densities of NOEs using residual dipolar couplings: application to the 370-residue maltodextrin-binding protein. J Mol Biol 300:197–212

Mumenthaler C,Güntert P, Braun W, Wüthrich K (1997) Automated combined assignment of NOESY spectra and three-dimensional protein structure determination. J Biomol NMR 10:351–362

Murzin AG, Brenner E, Hubbard T, Chothia C, (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247:536–540

Nilges M, Macias MJ, O'Donoghue SI, Oschkinat H (1997) Automated NOESY interpretation with ambiguous distance restraints: the refined NMR solution structure of the pleckstrin homology domain from β-spectrin. J Mol Biol 269:408–422

Nilges M, O'Donoghue SI (1998) Ambiguous NOEs and automated NOESY assignment. Prog NMR Spec 32:107–139

Noble D (2002) The rise of computational biology. Nat Rev Mol Cell Biol 3:460–463

Peter C, Daura X, van Gunsteren WF (2001). Calculation of NMR-relaxation parameters for flexible molecules from molecular dynamics simulations. J Biomol NMR 20:297–310

Philippopoulos M, Mandel AM, Palmer AG, Lim C (1997) Accuracy and precision of NMR relaxation experiments and MD simulations for characterizing protein dynamics. Proteins Struct Funct Genet 28:481–493

Prestegard JH, al-Hashimi HM, Tolman JR (2000) NMR structures of biomolecules using field oriented media and residual dipolar couplings. Q Rev Biophys 33:371–424

Prestegard JH, Valafar H, Glushka J, Tian F (2001) Nuclear magnetic resonance in the era of structural genomics. Biochemistry 40:8677–8685

Prompers JJ, Brüschweiler R (2002) General framework for studying the dynamics of folded and nonfolded proteins by NMR relaxation spectroscopy and MD simulation. J Am Chem Soc 124:4522–4534

Redfield C, Boyd J, Smith LJ, Smith RAG, Dobson CM (1992) Loop mobility in a four helix bundle protein: 15 N NMR relaxation measurements on human Interleukin-4. Biochemistry 31:10431–10437

Rohl CA, Baker D (2002) De novo determination of protein backbone. J Am Chem Soc 124:2723–2729

Schneider T, Brünger AT, Nilges M (1999) Influence of internal dynamics on accuracy of protein NMR structures: derivation of realistic model distance data from a long molecular dynamics trajectory. J Mol Biol 285:727–740

Schuler GD, Epstein JA, Ohkawa H, Kans JA (1996) Entrez: molecular biology database and retrieval system. Methods Enzymol 266:141–162

Spronk CA, B Nabuurs SB, Bonvin AMJJ, Krieger E, Vuister GV, Vriend G (2003) The precision of NMR structure ensembles revisited. J Biomol NMR 25:225–234

Stein LD (2003) Integrating biological databases. Nature Reviews Genet 4:337–345

Stoesser G, Baker W, van den Broek A, Garcia-Pastor M, Kanz C, Kulikova T, Leinonen R, Lin Q, Lombard V, Lopez R, Mancuso R, Nardone F, Stoehr P, Tuli MA, Tzouvara K, Vaughan R (2003) The EMBL nucleotide sequence database: major new developments. Nucleic Acids Res 31:17–22

Tian F, Valafar H, Prestegard JH (2001) A dipolar coupling based strategy for simultaneous resonance assignment and structure determination of protein backbones. J Am Chem Soc 123:11791–11796

Tolman JR, Al-Hashimi HM, Kay LE, Prestegard JH (2001) Structural and dynamic analysis of residual dipolar coupling data for proteins. J Am Chem Soc 123:1416–1424

Zdobnov EM, Lopez R, Apweiler R, Etzold T (2002) The EBI SRS server – new features. Bioinformatics 18:1149–1150

Zhang F, Brüschweiler R (2002) Contact model for the prediction of NMR N-H order parameters in globular proteins. J Am Chem Soc 124:12654–12655

Bioinformatics-Guided Identification and Experimental Characterization of Novel RNA Methyltransferases

J.M. Bujnicki, L. Droogmans, H. Grosjean, S.K. Purushothaman, B. Lapeyre

1 Introduction

1.1 Diversity of Methylated Nucleosides in RNA

Naturally occurring RNAs contain numerous chemically altered nucleosides. They are formed by enzymatic modification of the primary transcripts during the complex RNA maturation process. To date, a total of 96 structurally distinguishable modified nucleosides originating from different types of RNAs from many diverse organisms of the three major phylogenetic domains of life have been reported (Rozenski et al. 1999); http://medstat.med.utah.edu/RNAmods; and references therein). The pattern of modifications (type and location) depends on the RNA molecule considered, as well as, on the organism or the organelle they originate from. However, the largest number of modified nucleosides with the greatest structural diversity (a total of 81) is found in transfer RNAs, especially in tRNAs from higher organisms (Sprinzl et al. 1998; http://www.uni-bayreuth.de/departments/biochemie/trna). Other types of RNA (snRNA, snoRNA, rRNA, mRNA) also contain modified nucleosides (see http://rna.wustl.edu/snoRNAdb), however, their occurrence and

J.M. Bujnicki
Bioinformatics Laboratory, International Institute of Molecular and Cell Biology in Warsaw, Trojdena 4, 02–109 Warsaw, Poland
L. Droogmans
Laboratoire de Microbiologie, Université Libre de Bruxelles, 1 av. E. Gryson, 1070 Bruxelles, Belgium, and Laboratoire de Génétique des Procaryotes, Université Libre de Bruxelles, 12 rue des Professeurs Jeener et Brachet, 6041 Gosselies, Belgium
H. Grosjean
Laboratory of Structural Enzymology and Biochemistry, CNRS, 1 av. De la Terrasse, 91198 Gif-sur-Yvette, France
S.K. Purushothaman, B. Lapeyre
Centre de Recherche de Biochimie Macromoléculaire du CNRS, 1919 Route de Mende, 34293 Montpellier, France

Nucleic Acids and Molecular Biology, Vol. 15
Janusz M. Bujnicki (Ed.)
Practical Bioinformatics

particularly their diversity are lower than in tRNAs (see, for example, Limbach et al. 1995; Motorin and Grosjean 1998).

Among the naturally occurring nucleoside modifications, base and/or ribose methylations are by far the most frequently encountered and diverse (Fig. 1). They arise by single or multiple methylation(s) of an endocyclic carbon (like in m^5U, m^5C, or m^2A), an endocyclic nitrogen (like in m^3C, m^3U, $m^1\Psi$, m^1A, m^1G, m^7G, m^1I, mimG) or an exocyclic amino group (a nitrogen like in m^4C, m^6A, $m^6{}_2A$, m^2G, $m^2{}_2G$, mnm^5U; an oxygen as in mcm^5U, $mchm^5U$, mo^5U, $mcmo^5U$, or a sulfur atom like in ms^2A). Methyl groups are also present in the structure of the so-called hypermodified nucleosides which result from the attachment of a more complex side chain to one atom of the canonical base (like ms^2t^6A, m^6t^6A, $m^1acp^3\Psi$, wybutosine). It can also be bound to the exocyclic 2′ oxygen of ribose (Cm, Um, Gm, Am, Im, Ψm). In few cases, the methyl group has been found on both the base and the ribose, leading to hypermethylated nucleosides, as found in tRNAs from hyperthermophilic Archaea. A few methyl-derivatives like m^1G, m^1A, Cm, Gm, Um, m^5U are found in RNAs from all three major biological domains, while all others are clearly domain-specific. This suggests that only some of the corresponding modification enzymes may have a common evolutionary origin, while the majority of the other ones have evolved after the emergence of the three domains (Cermakian and Cedergren 1998).

Despite an impressive amount of research, the function of methylated nucleosides is poorly understood. Generally speaking, the presence of methyl

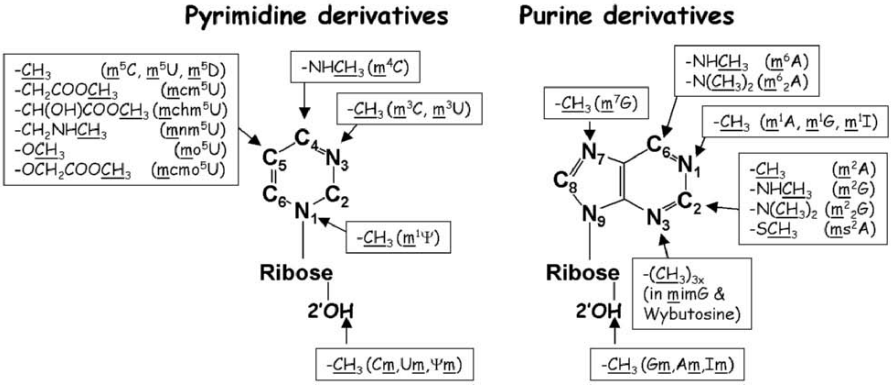


Fig. 1. Type of post-transcriptional methylated derivatives in cellular RNAs. Conventional numbering of each atom in the pyrimidine and purine *rings* are shown. Functional groups that are characteristic of each type of modified nucleoside derivative are *boxed*. Only the methyl groups that arise from an identified or still putative MTases are *underlined*. The conventional symbols of each modified nucleoside are in *brackets*. Further information on the structure, occurrence and location of each methylated nucleosides in RNAs and corresponding literature citations can be found in Limbach et al. (1994)

groups in the RNA molecules changes a local chemical microenvironment by increasing hydrophobicity, but also by increasing polarity as in positively charged m^1A and m^7G. These discrete chemical changes may help the RNA to fine tune its folding into a functional 3D architecture, and avoid misfolding or improving its recognition by proteins (maturation enzymes, structural proteins, aminoacyl-tRNA synthetases, initiation or elongation factors, etc.). Therefore, depending on their locations in the RNA molecules, one can expect the attachment of a methyl group to have more influence on one or the other of these properties. In the case of 2′-*O*-methyl ribose derivatives, this can also protect against nucleolytic degradation, while in eukaryotic cells, methylation could promote efficient transport of pre-rRNAs to the cytoplasm (for reviews see: Agris 1996; Bjork 1995; Curran 1998; Davis 1998).

1.2 RNA Methyltransferases

RNA methylation is carried out by a diverse group of RNA methyltransferases (MTases). S-adenosyl-L-methionine (AdoMet) is the most common methyl donor of a majority of RNA MTases identified so far. However, in some instances, the methyl groups have been shown to originate from folic acid (5,10-methylenetetrahydrofolate coupled with FADH2), as for the enzymatic formation of m^5U in tRNA from *B. subtilis* or *S. faecalis* (Romeo et al. 1974).

Hitherto, the RNA MTase activity has always been found associated with a protein enzyme (a single molecule or a protein complex). With respect to the target specificity (i.e., recognition of the nucleoside(s) to be methylated), there are, however, two major distinct mechanisms for RNA methylation. In the first mechanism, an "all-protein enzyme" can be site-specific, region-specific or multisite-specific, depending on the complexity and occurrence of the target-structural motif within the various RNA molecules (see Pintard et al. 2002b for a few examples). In the second mechanism, the modification is performed by a ribonucleoprotein (RNP), in which a protein carries out the catalytic activity, but the recognition is ensured by a guide RNA and few accessory proteins bound to the guide RNA. The advantage of this second system is that a single MTase bound to different guide-RNAs can catalyze methylation at various different positions in a given RNA and also in different RNAs. To date, RNA-guided modification has been identified only for ribose methylation and pseudouridine formation (review: Terns and Terns 2002).

For a mechanistic reason, enzymes that methylate different types of atoms of a particular nucleoside may possess different active sites and use different reaction chemistry. The same methylated nucleoside in different RNAs or in different positions of a given RNA are often catalyzed by distinct type- and/or site-specific enzymes. Also, certain phylogenetically conserved methylated residues in a given position of an RNA molecule but of different organisms can be generated by distinct mechanisms. For example, 2′-*O*-methylation of

ribose at position 34 in the majority of archaeal pre-tRNA^{Trp} are catalyzed by a special MTase within a multiprotein complex that uses the internal complementary sequence of the intron-containing pre-tRNA as a guide to target the methylation reactions (Clouet d'Orval et al. 2001). The same 2′-*O*-methyl ribose in *S. cerevisiae* is generated by a single protein (Trm7p, see below), which carries both the RNA recognition capacity and the catalytic activity (Pintard et al. 2000). Moreover, some RNA MTases fulfill other functions rather than just the methylation of specific atoms within RNA. In few cases it was demonstrated that the catalytic activity of a given MTase can be abolished by site-directed mutagenesis without affecting the growth rate of the mutated cells, while the disruption or the deletion of the corresponding ORF within the genome lead to severe slow growth rate or even to lethality probably due to a defect in the "RNA quality-control" function of some MTases (Lafontaine et al. 1998; Johansson and Bystrom 2002).

1.3 Structural Biology of RNA MTases and Their Relatives

Most of the known MTases, whose structures were solved by X-ray crystallography or NMR (currently over 30 structures in the Protein Data Bank) belong to a large superfamily related to Rossmann-fold proteins (denoted as RFP; Bujnicki 1999; Fauman et al. 1999). Compared to RFP, Rossmann-fold MTases (RFMs) exhibit a characteristic insertion of the C-terminal 7th β-strand into the common central β-sheet: 6-7-5-4-1-2-3. The "classical" RFPs, which bind NAD(P), and the RFMs, which bind AdoMet, use structurally equivalent and evolutionarily conserved cofactor-binding sites (between strands 1, 2 and 3) and they interact with the adenosine and ribose moieties of their ligands in a very similar manner. Typical RFM, acting on RNA, comprises the common catalytic domain and an auxiliary domain, often involved in substrate/target recognition (hence dubbed TRD, for target-recognition domain) or sometimes in oligomerization. TRDs of different RFMs are usually unrelated to each other and to any known domains in the database. Some RFMs lack auxiliary domains and use protuberances of the catalytic domain to recognize and bind their substrates (Fauman et al. 1999).

According to crystallographic and/or bioinformatic analyses, most of the experimentally studied RNA MTases belong to the RFM superfamily and include enzymes that generate a wide variety of methylated bases and ribose-2′-*O*-methylated nucleosides (Table 1). RFM enzymes are typically monomeric, although di-, tri-, or tetrameric structures have been reported; among the crystal structures of RNA MTases, Mj0882 forms a homodimer and Rv2118 c forms a homotetramer.

There are several superfamilies of AdoMet-dependent MTases, which neither share the RFM/RFP fold nor are structurally or evolutionarily related to one another (review: Schubert et al. 2003). The activation domain of methio-

Table 1. Structurally characterized RNA MTases

Protein	Organism	Specificity	PDB	Reference
RFM superfamily				
ErmAM	*Streptococcus pneumoniae*	23S rRNA:m^6_2A2058	1yub	Yu et al. (1997)
ErmC′	*Bacillus subtilis*	23S rRNA:m^6_2A2058	1qam	Schluckebier et al. (1999)
mtTFB	*Saccharomyces cerevisiae*	[16S rRNA:m^6_2A1518,1519][a]	1i4w	Schubot et al. (2001)
AviRa	*Streptomyces viridochromogenes*	23S rRNA:m^2G2535	1o9 h	Mosbacher et al. (2003)
Mj0882	*Methanococcus jannaschii*	[predicted m^2G]	1dus	Bujnicki and Rychlewski (2002b)
Rv2118 c	*Mycobacterium tuberculosis*	[predicted tRNA:m^1A58]	1i9 g	Gupta et al. (2001)
RrmJ	*Escherichia coli*	23S rRNA:Um2552[b]	1eiz	Bugl et al. (2000)
Mj0697	*Methanococcus jannaschii*	2′-O-ribose (guided)[c]	1fbn	Wang et al. (2000)
MT-1	*Human reovirus*	mRNA:cap 0 (m^7G)[d]	1ej6	Reinisch et al. (2000)
MT-2	*Human reovirus*	mRNA:cap 1 (2′-O-)[d]	1ej6	Reinisch et al. (2000)
vp39	*Vaccinia virus*	mRNA:cap 1 (2′-O-)	1av6	Hodel et al. (1998)
NS5 MT	*Dengue virus*	mRNA:cap 1 (2′-O-)	1l9 k	Egloff et al. (2002)
SPOUT superfamily				
RlmB	*Escherichia coli*	23S rRNA:Gm2251	1gz0	Michel et al. (2002)
TrmD	*Haemophilus influenzae*	tRNA:m^1G37	1uam	Ahn et al. (2003)
MT1	*Meth. thermooautotrophicum*	Unknown	1 k3r	Zarembinski et al. (2002)
RrmA	*Thermus thermophilus*	Unknown	1ipa	Nureki et al. (2002)
YibK	*Haemophilus influenzae*	Unknown	1 mxi	Lim et al. (2003)

The most representative entry from the Protein Data Bank (PDB) (http://www.rcsb.org) has been chosen for each enzyme, with the preference for protein–ligand complexes and structures solved at possibly highest resolution. Ss, *Sulfolobus solfataricus*; Pf, *Pyrococcus furiosus*; Sc, *Saccharomyces cerevisiae*; Mj, *Metanococcus jannaschii*; Af, *Archaeoglobus fulgidus*. [ ] Function predicted (not determined experimentally).

[a] Function of a human ortholog of Sc mtTFB was determined by Seidel-Rogol et al. (2003).

[b] Function determined by Caldas et al. (2000).

[c] Function determined by Omer et al. (2002).

[d] Function postulated by Bujnicki and Rychlewski (2001).

nine synthase (MetH) (Drennan et al. 1994) and the B12 biosynthetic enzyme CbiF (Schubert et al. 1998) are single examples of structurally characterized representatives of superfamilies with alternative folds that can support AdoMet-dependent methyltransfer reactions (reviews: Dixon et al. 1999). Several members of the SET superfamily (all protein-lysine MTases) have been recently characterized structurally and functionally (review: Marmorstein 2003). It remains to be shown whether there are any RNA MTases that belong to these superfamilies; to date, no such relationship has been reported.

Recently, another superfamily of AdoMet-dependent MTases has been defined based on bioinformatics analyses and dubbed SPOUT from the two major lineages: SpoU (catalyzing 2′-*O*-ribose methylation in tRNA) and TrmD (catalyzing the formation of m^1G in tRNA)(Anantharaman et al. 2002b). SPOUT MTases comprise two domains, which exhibit different spatial arrangements. The large, catalytic domain common to all these enzymes (the actual SPOUT domain) exhibits a novel and unusual α/β fold with a deep knot (topology of the central β-sheet: 6-4-5-1-2-3; AdoMet is bound between the strands 4 and 5). The smaller domain is not conserved and exhibits structural similarity to various unrelated RNA-binding proteins; it can be found either fused N-terminally to the catalytic domain, or inserted into it. YibK is a "minimal" member of the SPOUT superfamily, which comprises only the catalytic domain. Figure 2 shows the RFM and SPOUT folds of catalytic domains of RNA MTases. All structurally characterized SPOUT MTases form homodimers. Five structures of SPOUT-superfamily members have been reported

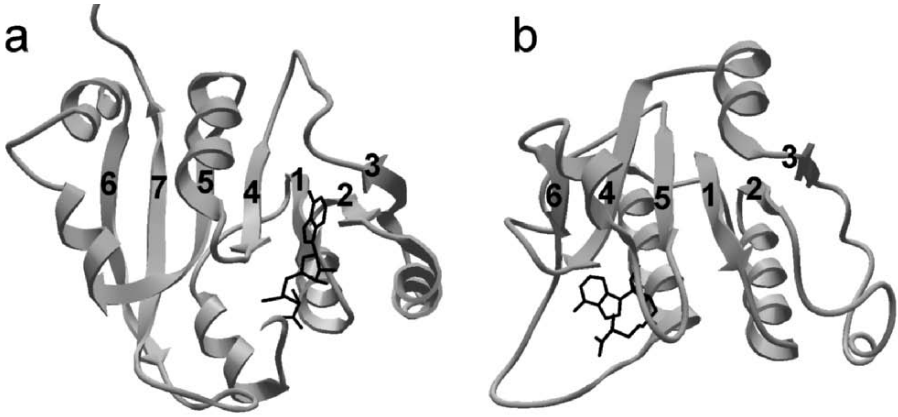


Fig. 2. Comparison of the RFM and SPOUT folds of catalytic domains of RNA MTases. **a** The catalytic domain of rRNA:m^6A MTase ErmC′ (Schluckebier et al. 1999), **b** putative MTase YibK (Lim et al. 2003). The position of the cofactor is indicated. The consecutive b-strands are *numbered* from the N-terminus to the C-terminus, revealing different topologies of these two MTase folds

to date (Table 1), however, only two of them have been characterized biochemically (including determination of the substrate specificity), while the function of the others remains putative.

2 Traditional and Novel Approaches to Identification of New RNA-Modification Enzymes

The first RNA modification enzyme that was identified in the 1960s by Borek and coworkers was a tRNA MTase. Using RCRel mutants of *E. coli* which accumulate undermodified RNAs during methionine starvation, Borek's group showed for the first time that an AdoMet-dependent enzyme can mediate methylation in vitro of specific bases in macromolecular precursors of RNA, i.e., posttranscriptionally after polynucleotide synthesis, which at that time was not evident at all (Fleissner and Borek 1962). The tRNA:m^5U MTase (RUMT) was renamed as TrmA after the corresponding gene in the *E.coli* genome was identified (Persson et al. 1992). It catalyzes the formation of the almost universal m^5U at position 54. Later, using the sequence information of the bacterial enzyme, the orthologous gene (and the corresponding enzyme) was identified in yeast (Nordlund et al. 2000).

For almost three decades, identification and purification of an RNA modifying enzyme, as well as, the assignment of its corresponding structural genes within a genome were daunting tasks. The main reason was that these enzymes are usually present in the cell at low level and are therefore difficult to purify to homogeneity. Also, tools to generate transcripts of synthetic genes allowing easy detection of the enzymatic activity in vitro were lacking (review: Grosjean et al. 1998). Nevertheless, a few RNA modifying enzymes were identified and purified from cellular extracts by standard chromatographic techniques (Garcia and Goodenough-Lashua 1998). Also, classical genetic approaches, coupled with screening depending on translational suppression or resistance to selected antibiotics, allowed for identification of a few genes corresponding to RNA modifying enzymes, including MTases, mostly in *E.coli* and *Salmonella thyphimurium* (review: Winkler 1998).

Nowadays, such "classical" biochemical and genetic methods for identification of new RNA MTases (and other proteins) can be efficiently supplemented by "reverse genetics" and large-scale "-omics" approaches, including biochemical genomics (Martzen et al. 1999, reviewed in Hopper and Phizicky 2003) and bioinformatics/phylogenomics (Eisen 1998). These two approaches have proven to be very efficient and complementary in the search for novel RNA MTases.

In the biochemical genomics approach, a set of clones that express a representative of each protein of a proteome is generated (e.g., all ORFs from a given genome are cloned) and the biochemical function of the corresponding proteins is analyzed on a genome-wide basis (Martzen et al. 1999; Phizicky et

al. 2002). For example, Phizicky and coworkers generated an array of 6144 individual yeast strains, each containing a different yeast ORF, N-terminally fused to glutathione S-transferase (GST) that facilitates the isolation of the corresponding protein. For the identification of ORF-associated activities, strains were grown in defined pools, and GST-ORFs were purified. Then, pools were assayed for activities, and active pools were deconvoluted to identify the source strains. This method has led to isolation of several novel RNA modification enzymes such as tRNA:D17 dihydrouridine synthase (Xing et al. 2002), tRNA:m^7G46 MTase Trm8p (Alexandrov et al. 2002), and tRNA: m^1G9 MTase Trm10p (Jackman et al. 2003). The biochemical genomics has proven to be exceptionally powerful in isolating "non-conventional" RNA modification enzymes that had been predicted by none of the bioinformatics analyses.

In the biochemical genomics approach, the tagged proteins are tested in relatively small pools (using 96 wells boxes), hence a theoretical limitation could be that the complexes formed by the association of several protein subunits and/or other macromolecules like RNA as in RNP would be missed. This has proven not to be the case experimentally, since Trm8p, which was isolated by this method, is part of an heterodimer with Trm82p. Apparently, there was enough Trm82p contaminating the preparation for Trm8p to be active (Alexandrov et al. 2002). A serious limitation of the biochemical genomics approach is that the tagged enzymes have to be active in vitro on the substrate that is provided for the test, and this is not always the case. For instance when tagged on its N-terminus, the tRNA MTase Trm7p (identified by bioinformatics; see below) has been found to be inactive (L. Pintard, F. Lecointe, H. G. and B.L., unpubl. data).

The bioinformatics/phylogenomics approach, which relies on the computational (comparative) analysis that combines genome sequence information and phylogenetic studies, will be reviewed in detail in this chapter. It has advantages over the biochemical genomics approach, since the initial screening procedure does not require manipulation of the gene or handling of the recombinant protein, which is sometimes an endless source of difficulty. However, it is also very limited by the experimental data available at a certain time that will drive the search for new candidates, and by the database, in which the new candidates are sought. In practice, two types of computational analyses have been used to predict biological function for uncharacterized ORFs: the "homology" and "non-homology" methods are summarized below.

3 Bioinformatics: Terminology, Methodology, and Applications to RNA MTases

Homology is defined as a synonym of common evolutionary origin, i.e., all genes that have arisen from a common ancestor are homologous. Genes in different species that originate from a single gene in the last common ances-

tor of these species are termed *orthologs*. Orthologous genes have often very similar biological roles in the present-day organisms. Homologues generated by gene duplication are termed *paralogs* – they often share the same generic function, such as the type of the reaction catalyzed, but may differ in certain details, such as specificity towards different substrates. Typically, homologous genes can be grouped into *families* and *superfamilies*, in which the majority of members share the same function (e.g., pseudouridine synthases, adenosine deaminases, dihydrouridine synthases, tRNA-splicing nucleases – to name but a few families of enzymes involved in RNA modification and processing). In the homology-based approach, the inference of common evolutionary origin is used to hypothesize a common function, which can be transferred between the experimentally characterized member of the family and other members, for which only sequence information is available.

Sometimes homologous gene products have strong sequence similarities, so that the inference of homology is straightforward. This is especially the case when orthologs from closely related species are compared. A BLAST (Altschul et al. 1990) or FASTA (Pearson and Lipman 1988) search of a sequence database can reveal potential homologues of the "query" sequence together with pairwise alignments and an estimation of the statistical significance of similarities. However, accumulation of multiple substitutions in the course of the divergent evolution can make two homologous sequences as dissimilar as any two proteins chosen randomly from the database. Several bioinformatics approaches have been developed to identify remote homology in the absence of pairwise sequence similarity; one of the most popular methods is *protein fold recognition* (FR; reviewed in another chapter of this volume). Briefly, FR detects homology based on a combination of evolutionary criteria (conservation of key residues in multiple sequence alignments) and structural considerations (similar linear patterns of secondary structure elements or estimation of the physico-chemical compatibility of one protein's sequence with another protein's three-dimensional structure). FR methods can be used virtually in the same manner as traditional methods for sequence database searches, with the key difference: the database to be searched by FR comprises only proteins with experimentally determined structures rather than all known protein sequences. Hence, the availability of a related structure in the Protein Data Bank is an essential (but not sufficient) precondition for the success of FR-based identification of homology.

However, homology is defined on the basis of evolution, rather than function. On the one hand, homologues can fulfill different functions and share only very general similarities; even orthologs may fulfill non-identical roles (reviews: Todd et al. 2002; Rost 2002). On the other hand, numerous cases of very remote homologues or even non-homologues, which developed the same function "by convergence" have been reported (Koonin et al. 1996; Galperin et al. 1998). Moreover, orthology is not necessarily a one-to-one relationship because a single gene in one genome may correspond to a whole family of

paralogs in another genome (which may be functionally diversified; see examples below). Hence, there is a pitfall of over-prediction (i.e., too specific functional assignment) to be avoided when annotating ORFs' function by homology, using either simple or sophisticated bioinformatics tools.

The so-called non-homology methods rely on properties shared by functionally-related proteins other than the hallmarks of homology, i.e., sequence or structural similarity. Instead, conserved gene position (similar genomic neighbors), fusions with domains of similar function, correlation in gene occurrence (shared "phyletic patterns"), co-evolution, or co-expression is sought. These methods can predict functions for ORFs that are without characterized homologues (reviews: Marcotte 2000; Galperin and Koonin 2000). However, the types of functional predictions also differ from what might be learned from detection of homology.

The domain fusion method finds functional relationships of ORFs found separately in one genome by identification of their co-occurrence as a single ORF (i.e., fused protein) in another genome (Marcotte et al. 1999). Similarly, the products of ORFs can be predicted to interact (physically and/or functionally) if they are repeatedly found as neighbors in different genomes (Overbeek et al. 1999). Such co-occurring ORFs typically encode subunits of a multi-protein complex or components of an enzymatic pathway. However, it has been suggested that in prokaryotic genomes, some genes are maintained within operons because of the advantage of expression at a level that is typical of the given neighborhood rather than because of functional association (Rogozin et al. 2002). Co-operating or interacting proteins be also be identified by detection of families with similar phylogenetic profiles i.e., correlated patterns of inheritance (presence or absence) in known genomes (Pellegrini et al. 1999). Likewise, proteins involved in some specific biological process can be identified by studying the correlation of their phylogenetic profile with the presence or absence of a particular phenotype (Huynen et al. 1998).

Typically, the non-homology methods offer only very general functional predictions in terms of metabolic pathways or multi-protein complexes, rather than inference of a specific biochemical activity. However, if the analyzed ORF or some of its identified interaction partners have a known function (or if their biochemical function can be predicted based on homology), the prediction of specific biochemical functions can be extended to other putative components of a complex or a pathway.

Two general approaches (termed top-down, and bottom-up) have been developed for identification of proteins with a desired function ; these approaches combine various homology and non-homology methods (Fig. 3). These approaches can be applied to guide experimental characterization of virtually any protein superfamily; here, we describe their application to identify new candidates for RNA MTases and predict their function as specifically as possible.

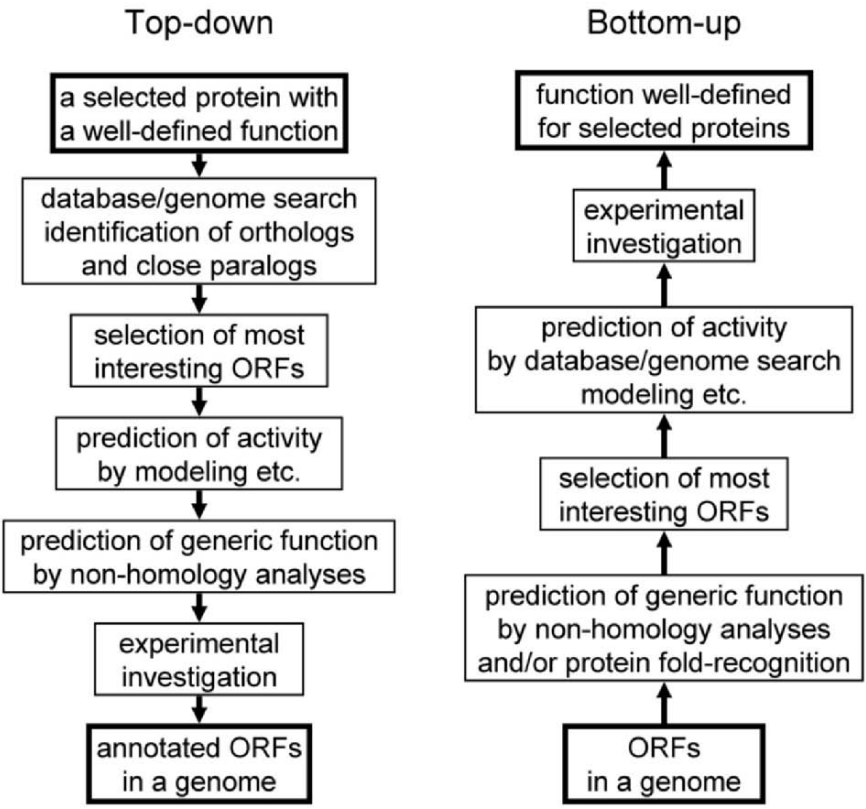


Fig. 3. Steps involved in the *top-down* and *bottom-up* approaches to gene identification by a combination of bioinformatics and experiment

3.1 The Top-Down Approach

The *top-down* approach has been traditionally used to identify relatives of a newly identified and functionally characterized protein. It involves database searches with the functionally characterized protein's sequence as a "query", with the aim of identification of its orthologs in different organisms (likely to share the same function) and paralogs (likely to share the general function, but for instance exhibit different specificity). The specific aim of the top-down analysis is typically to identify close relatives of a known (often, newly characterized) protein, which may exhibit a slightly different function, for instance, the same activity (to be verified with a similar assay as used for the "founding member" of the family), but different substrate specificity. Often, proteins with a specific function, closely related to the function of another, known protein, are sought. The precondition for this type of analysis is the knowledge of the molecular mechanism of the "query" protein and the ability

to classify its homologues into proteins with (potentially) similar mechanisms of activity and proteins with different mechanisms. For instance, the knowledge of residues forming the active site may be essential to recognize enzymes that catalyze a given type of RNA methylation and distinguish them from MTases that catalyze different types of reaction. The "non-homology" analysis in the top-down approach is usually limited to help with the prediction of the subcellular localization of individual members of the family or their putative association with some known metabolic pathway or a protein complex.

All early bioinformatic analyses of RNA MTases were carried out according to the top-down approach and involved iterative and/or transitive searches of genome sequences. Typically, the BLAST program (or its iterative version, PSI-BLAST) (Altschul et al. 1997) was used to identify homologues of a query with a known, recently identified function, and some of these homologues were further used as new, additional queries. Using this approach, Koonin and Rudd have identified a family of homologues of SpoU and predicted that they all share a function of rRNA 2′-*O*-MTase determined for one functionally characterized member of a family, Tsr from *Streptomyces azureus* (Koonin and Rudd 1993). Subsequently, it has been shown that SpoU (renamed as TrmH) is in fact a tRNA:Gm18 2′-*O*-MTase (Persson et al. 1997). Subsequent "top-down" analysis carried out for a larger database by Bachellerie and coworkers, revealed additional putative RNA 2′-*O*-MTases from Bacteria, Archaea, and Eukaryota (Cavaille et al. 1999). Finally, the aforementioned analysis by Koonin and coworkers (Anantharaman et al. 2002b) allowed for linkage of the SpoU family of 2′-*O*-MTases with the TrmD family of tRNA:m^1G MTases and with a few other families of uncharacterized proteins. The conservation of the putative cofactor-binding site in the newly defined SPOUT superfamily implied the common MTase function of all its members; nonetheless, the lack of conservation in the active site suggested that their specificities may be different (and not predictable from homology alone).

The top-down approach has also been used by Santi and coworkers to identify new candidates for RNA:m^5U MTase, and SpoU-related RNA:2′-*O*-MTases (Gustafsson et al. 1996) and RNA:m^5C MTases (Reid et al. 1999), by Phizicky and coworkers to identify orthologs of their biochemically-discovered fungal tRNA:m^7G46 MTase (Alexandrov et al. 2002) and new paralogs of tRNA:m^1G9 MTase (Jackman et al. 2003), and by Bujnicki and coworkers to identify new putative base-MTases with the following potential specificities: RNA:m^7G (Bujnicki et al. 2001), RNA:m^2G (Bujnicki 2000; Bujnicki and Rychlewski 2002b), tRNA:m^1A (Bujnicki 2001a), RNA:$m^2{}_2$G (Bujnicki et al. 2002c), 23S rRNA:m^1G (Bujnicki et al. 2002a), RNA:m^6A (Bujnicki et al. 2002b), and various 2′-*O*-MTases that belong to the RFM superfamily rather than the SPOUT superfamily (Bujnicki and Rychlewski 2000, 2002a; Pintard et al. 2002a, b; Feder et al. 2003). The analysis of m^5C, m^1A and 2′-*O*-MTases combined with

experimental studies has prompted and guided identification and characterizations of several novel MTases, which will be described in more detail below, as "case studies".

3.1.1 Top-Down Search for Novel RNA:m^{5}C MTases in Yeast

In all eubacterial organisms, 5-methylribocytosine (m^{5}C) has been found only in ribosomal RNAs. For example, two m^{5}C residues were located at positions 967 and 1407 in *E. coli* 16S rRNA and one at position 1962 in 23S rRNA (Smith et al. 1992), whereas none of the *E.coli* tRNAs sequenced so far contain m^{5}C (Sprinzl et al. 1998). In contrast, in Eukaryota and Archaea, m^{5}C is found in both tRNAs and rRNAs. In particular, in yeast tRNAs, m^{5}C is found at four positions (34, 40, 48 and 49), but the most frequently occurring cluster of m^{5}C residues is located at positions 48 and 49 at the border of the variable loop and the T-stem (review: Auffinger and Westhof 1998).

The first member of the RNA methyltransferases family catalyzing such a reaction in *E.coli* rRNA was found independently by two groups using two different approaches. Koonin (1994) predicted that a family comprised of the human proliferation-associated nucleolar protein P120 and the product of the bacterial *fmu/fmv/SUN* gene is made up of RNA MTases. Guided by this prediction, Santi and his group cloned, expressed the *E. coli* Fmu protein and tested its activity in vitro using [^{3}H]AdoMet and various RNA transcripts as substrates. The identity of the methylated residue was rigorously established by chromatographic analysis allowing for the claim that Fmu (now renamed RsmB) is indeed an m^{5}C-MTase acting exclusively at position 967 of *E.coli* 16S rRNA (Gu et al. 1999). Independently, Ofengand and coworkers used a "classical" biochemical approach. They purified a protein from the *E.coli* extract that catalyzes the formation of m^{5}C at position 946 in 16S rRNA. After verification of the specificity of the newly purified enzyme using in vitro produced 16S rRNA transcript as a substrate, the N-terminus of the protein was micro-sequenced. The resulting peptide sequence was then used as a query to identify the corresponding gene (Fmu) in the *E.coli* genome (Tscherne et al. 1999).

Subsequently, the sequence of RsmB was used to search for homologous proteins (potential RNA:m^{5}C MTases) in the *S. cerevisiae* genome. This top-down approach allowed for the detection of three yeast proteins with obvious sequence similarity to RsmB: (1) YNL061w/Nop2p, a nucleolar protein implicated in the rRNA maturation of 26S ribosomal RNA (Hong et al. 1997); (2) an uncharacterized ORF YNL022 c; and (3) YBL024w/Ncl1p, a non-essential nuclear protein (Wu et al. 1998). It was found that the Ncl1p protein catalyzes the formation of m^{5}C at four different positions (34, 40, 48 and 49) in tRNAs and hence it was renamed Trm4p (Motorin and Grosjean 1999). That only Trm4p is responsible for the formation of all the m^{5}Cs in the various *S. cerevisiae* tRNAs, was confirmed by the deletion experiment. Thus, based on a

bioinformatic prediction of a general function (RNA MTase) for Fmu and experimental determination of its specificity, a family of eukaryotic m^5C MTases was identified.

3.1.2 Top-Down Search for Bacterial and Archaeal m^1A MTases

The methylated nucleoside 1-methyladenosine (m^1A) is found in the T-loop of tRNAs from many organisms belonging to the three domains of life (Bacteria, Eukarya and Archaea). In eukaryotic and bacterial tRNAs, m^1A is present at position 58, whereas in archaeal tRNAs it is present at position(s) 58 and/or 57. Archaeal m^1A57 is an obligatory intermediate in the biosynthesis of 1-methylinosine (m^1I57). In contrast to the biosynthesis of m^1I37 in the anticodon loop of *S. cerevisiae* tRNAs, which proceeds by a deamination of A to I, followed by a methylation step, the biosynthesis of m^1I57 in archaeal tRNAs proceeds by the methylation of A57 into m^1A57, followed by a deamination leading to m^1I57 (reviewed in Grosjean et al. 1996). The enzyme responsible for the m^1A modification has been studied for a long time using cell extracts or (semi-)purified enzymes from a variety of organisms: mammals, *Tetrahymena pyriformis*, *Dictyostelium discoideum*, *Thermus flavus* and *Thermus thermophilus* (reviewed in Garcia and Goodenough-Lashua 1998)). However, the genes encoding these tRNA:m^1A MTases remained unidentified.

A major breakthrough in the identification of the genes encoding tRNA:m^1A MTases was made by Anderson et al. (1998), who were characterizing mutations affecting the regulation of the *S. cerevisiae GCN4* gene encoding a transcription factor acting as a general regulator of amino acids biosynthesis. The expression of *GCN4* itself is a highly regulated process that involves translational control. The *GCN4* messenger RNA contains four short ORFs upstream of the main *GCN4* ORF, and translation of these short ORFs controls the level of *GCN4* translation. A series of *trans* acting mutations were obtained (called *gcd* mutations) leading to the translational derepression of *GCN4* translation. Among the different *gcd* mutations were *gcd10* and *gcd14* which were found to affect the maturation of the initiator tRNA. A more detailed analysis showed that the formation of m^1A in tRNAs is affected in the *gcd10* and *gcd14* mutants (Anderson et al. 1998). The Gcd10p and Gcd14p proteins form a nuclear complex with tRNA:m^1A MTase activity, in which Gcd14p (now renamed Trm6a) is responsible for transferring the methyl group from AdoMet to tRNA whereas Gcd10p (now renamed Trm6b) is required for tRNA binding (Anderson et al. 2000). These results confirmed a previous computational sequence analysis suggesting that only Trm6a, and not Trm6b possesses conserved motifs typical of the RFM family of MTases (Calvo et al. 1999).

A top-down approach allowed for the identification of Trm6a orthologs in a variety of organisms belonging to the three domains of life. In contrast, Trm6b orthologs were found only in Eukaryota (Bujnicki 2001a). Moreover,

protein fold-recognition analysis revealed that the Trm6a and Trm6b families are evolutionarily related and probably evolved from a common ancestor (after a gene duplication in the ancestor of extant Eukaryota). Therefore, the archaeal and prokaryotic tRNA:m^1A MTases were postulated to be homomultimers of a Trm6a-like polypeptide (Bujnicki 2001a). This hypothesis was reinforced after the resolution of the crystal structure of a bacterial Trm6a homologue, the Rv2118 c protein from *Mycobacterium tuberculosis* (Gupta et al. 2001). The crystal structure revealed that Rv2118 c exhibits an RFM fold, that it binds AdoMet and forms a homotetramer corresponding to a weak dimer of strong dimers. Nonetheless, the MTase function of Rv2118 c has not yet been demonstrated.

Guided by bioinformatics, bacterial and archaeal homologues of Trm6a have been cloned and characterized functionally. As expected, the bacterial protein cloned from *Thermus thermophilus* genomic DNA (termed TrmI) turned out to be a homotetrameric, site-specific AdoMet-dependent MTase, able to methylate m^1A58 in tRNA in the absence of any other protein (Droogmans et al. 2003). Surprisingly, the archaeal ortholog of TrmI (cloned from *Pyrococcus abyssi*) was found to exhibit not only the m^1A58, but also the m^1A57 specificity (Droogmans and coworkers, unpublished data). The latter finding was quite surprising: a typical non-homology approach aimed at the identification of an Archaea-specific RNA modification enzyme would suggest searches for an Archaea-specific gene. This study suggested that a function specific to a given phylogenetic lineage could be in fact conferred by a "moonlighting" protein (Jeffery 1999), which developed a novel activity, while maintaining the original one.

3.1.3 Top-Down Search for Novel Yeast 2'-O-MTases

In yeast, ribose methylation is guided mainly by the numerous snoRNA that base-pair with their cognate targets onto the pre-rRNA precursor. Several lines of evidence suggest that Nop1p is the major snoRNA-dependent rRNA MTase (Tollervey et al. 1993; Wang et al. 2000). However, there are still a few ribose methylations for which no guide RNA has yet been identified (Lowe and Eddy 1999). One of these orphan nucleosides is located within the catalytic site of the large rRNA molecule, at the peptidyl-transferase center. Not only was this structure highly conserved throughout the evolution, but the very same nucleoside – a uridine at the 5′ end of the loop – is always 2′-*O*-methylated (Fig. 4). Interestingly, in *E. coli*, the homologous position (Um_{2552}) of the 23S rRNA has been shown to be specifically methylated by RrmJ, an RFM-superfamily enzyme that belongs to an heat-shock operon (Caldas et al. 2000).

Top-down sequence searches, initiated with RrmJ revealed a large family of proteins from Bacteria, Archaea, Eukaryota and various viruses, characterized by a common K-D-K-E tetrad of residues separated in a primary sequence,

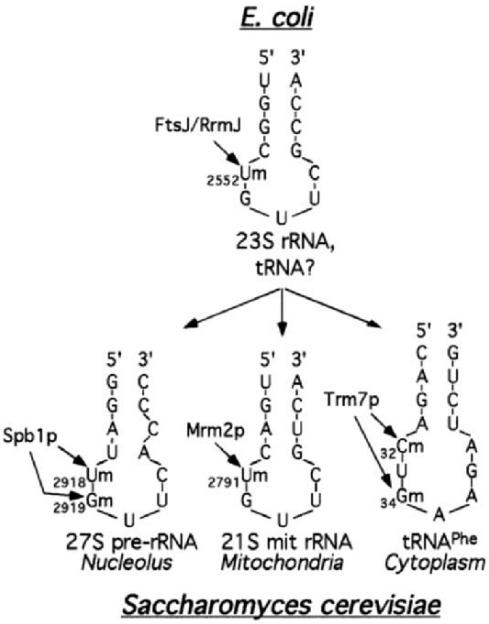


Fig. 4. Three yeast orthologs of one bacterial MTase: Duplication and horizontal gene transfer lead the way to subfunctionalization

but adjacent in space in the RrmJ crystal structure (Bugl et al. 2000; Feder et al. 2003). Phylogenetic studies revealed that RrmJ is the closest prokaryotic homologue of Spb1p, a nucleolar protein involved in 25S rRNA maturation in yeast (Kressler et al. 1999; Pintard et al. 2000). Interestingly, the top-down search revealed that the yeast genome encodes two additional proteins with striking similarities to RrmJ/Spb1p, namely Mrm2p and Trm7p. Mrm2p has recently been shown to be a mitochondrial ortholog of RrmJ and to methylate position U_{2791} of the peptidyl-transferase center that corresponds to U_{2552} in *E. coli* (Pintard et al. 2002a). Figure 4 shows a secondary structure representation of the hairpin loops (in the *E. coli* 23S rRNA, the *S. cerevisiae* 28S rRNA and the yeast mitochondrial 21S rRNA) that contain the methylated nucleoside always located at the same position. Taken together, these results strongly suggest that the position of the peptidyltransferase center of the large rRNA is modified by site-specific enzymes rather than by the snoRNA-guided mechanism, perhaps since this well conserved position plays a key role, different from all other 2′-*O*-methylations.

The functions of Spb1p (a nucleolar protein involved in 25S rRNA maturation), and Mrm2p (a mitochondrial protein involved in 21S rRNA) were quite obvious (methylation of the same position in orthologous rRNA molecules), the function of Trm7p was quite puzzling. A key experiment – the demonstration that Trm7p is mostly cytoplasmic – has limited the search for its substrates to the sole RNA known to be modified within the cytoplasm (Pintard et al. 2002b). Obvious candidates were tRNAs, for which certain modifications

occur after their export from the nucleus. There are striking similarities between the anticodon loop of certain tRNA that are 2′-*O*-methylated and the peptidyl transferase center of the rRNA recognized by FtsJ, Mrm2p and possibly Spb1p. One major difference is the length of the loop that is seven nucleotides long in tRNA as compared to 5 nucleotides in rRNA. However, once methylated, nucleotide at position 32 can base pair with the nucleotide at position 38 therefore reducing the length of the loop to five nucleotides, rendering it more like the rRNA loop (Auffinger and Westhof 2001). It turned out that Trm7p is required for the formation of both Cm32 and Gm34 in $tRNA^{Phe,\ Tyr\ and\ Leu}$ (Pintard et al. 2002b). Interestingly, the kinetics of the two reactions are different, Cm32 being made rapidly and without delay, while Gm34 is made more slowly and after a long delay. This observation suggested that the reaction could be sequential, Cm32 being first catalyzed, then this modification would be followed by a structural rearrangement of tRNA that would expose G34 to the action of the enzyme. A homology model of Trm7p was built, to which the $tRNA^{Phe}$ structure was docked. Strikingly, only the mature form of the tRNA fits well with the modeled structure, exposing the 2′-OH group of the ribose to the methyl group of AdoMet bound to the enzyme. In sharp contrast, the 2′-OH group of ribose at position 32 is not accessible to the enzyme when the tRNA has already adopted its mature 3D structure. This suggested that the tRNA is being modified at position 32 prior to the adoption of its mature 3D-structure, when its structure is still flexible enough to expose its 2′-OH group to the action of the enzyme. Then, after methylation of position 32, base pairing between nucleotides 32 and 38 would take place and the tRNA structure would flip and adopt its more rigid mature structure. Only then would modification of position 34 take place, therefore explaining the different kinetic reactions. This view is supported by the observation that formation of Cm32 is not very sensitive to mutations affecting tRNA secondary structure, while formation of Gm34 is strongly dependent on the rigid structure adopted by the mature tRNA (F. Lecointe and H.G., unpubl. results). It is a noteworthy point that the protein produced in *E. coli* had no activity in vitro, suggesting that certain modifications achieved in eukaryotic cells are essential for the enzyme to become active. Alternatively, it is conceivable that Trm7p requires some other cellular components as is the case for the aforementioned heterodimeric tRNA MTases Trm6a/Trm6b (alias Gcd10p/Gcd14p) (Anderson et al. 2000) or Trm8/Trm82 (Alexandrov et al. 2002).

3.2 The Bottom-Up Approach

The *bottom-up* approach aims to identify as many members of a protein superfamily in a given genome (or set of genomes) as possible. Typically, the generic function is predicted first (by detection of distant homology) and the functional details (up to the level of specificity) are predicted by combination

of homology and non-homology methods. This type of analyses is typically intended to provide a large number of potential candidates with only "crude" functional prediction, for experimental testing and determination of biochemical function.

Here, we will focus on the bottom-up prediction of RNA MTases in the yeast genome. The earliest genome-wide analysis of potential MTases among yeast ORFs did not specifically focus on RNA methylation and involved simple identification of proteins that possess motifs conserved in the RFM superfamily (four of the most conserved motifs were selected out of a total of nine) (Niewmierzycka and Clarke 1999). The search resulted in 33 candidate ORFs with identifiable motifs. Seven of these ORFs turned out to be known MTases (the authors note that they failed to detect motifs in several genuine MTases), while the other 26 so-called good and marginal matches were put forward for experimental analysis. Disruptions were made in seven of the corresponding genes, revealing one lethal and two slow growth phenotypes. One of the mutants showed a methylation defect of a novel type of arginine derivative, leading to the prediction of the specificity of the corresponding enzyme, ultimately confirmed by its biochemical characterization (Niewmierzycka and Clarke 1999).

Recently, a large-scale analysis of candidate proteins involved in RNA metabolism has been presented by Anantharaman et al. (2002a). Their analysis involved different types of enzymes and non-enzymes and all prokaryotic and eukaryotic genomes available, naturally including also RNA MTases and the yeast genome. They used exhaustive iterative searches of sequence databases queried by representatives of all known families of proteins implicated in RNA metabolism. All sequences retrieved from the searches were pooled together and potential orthologous sets were delineated by clustering (according to BLAST scores, as implemented in BLASTCLUST). The initial groups of orthologs and paralogs were corrected and optimized by multiple sequence alignment analysis and phylogenetic tree reconstruction. The domain architecture of each individual protein was then determined by comparison with multiple sequence alignments of all known protein domains. Finally, the conservation of functional complexes and pathways was assessed by combining the results of protein domain analysis with the experimental evidence extracted from the literature. Detection of homologues of proteins involved in RNA metabolism required corrections to exclude those domains and proteins that were known to be primarily involved in DNA metabolism. A distinction between RNA and DNA MTases, whose catalytic domains are often very similar, was made based on the non-homology criteria: RNA MTases are typically highly conserved (both with respect to the protein sequence and phylogenetic distribution) and are often associated with known RNA-binding domains, such as PUA (Aravind and Koonin 1999), S4 (Staker et al. 2000), THUMP (Aravind and Koonin 2001), or TRAM (Anantharaman et al. 2001). On the other hand, DNA MTases are typically found associated with

restriction endonucleases in restriction-modification systems (review: Bujnicki 2001b), exhibit sporadic phylogenetic distribution due to intense horizontal gene transfer (Jeltsch and Pingoud 1996) and have never been found to contain RNA-binding domains. For a few putative MTase families, Anantharaman et al. predicted the RNA MTase function based on the identification of fusions with known RNA MTases, other RNA modification enzymes, or with known RNA-binding domains (Anantharaman et al. 2002a). Curiously, the predicted rRNA MTases from the HemK/YfcB family turned out to be active as protein MTases, which catalyzes the methylation of polypeptide chain release factors such as RF1 and RF2 (Nakahigashi et al. 2002; Heurgue-Hamard et al. 2002). This would suggest that the function of these proteins was significantly overpredicted (the prediction should be as general as "methylation implicated in the translation process" rather than very specific "rRNA methylation"). Interestingly, the RF2 exhibits "molecular mimicry" and its 3D structure mimics that of tRNA (Vestergaard et al. 2001). It is therefore tempting to speculate that HemK and its homologues could have diverged from an ancestral RNA MTases that were "cheated out of" by the RNA-like structure of the protein substrate; no wonder that humans were misled too.

3.2.1 Bottom-Up Search for New Yeast RNA MTases

Isolation of yeast MTases Mrm2p and Trm7p (see above) represented a perfect example of the characterization of new RNA MTase enzymes guided by a top-down bioinformatic analyses. We have extended the search for novel RNA MTases to the whole yeast genome, using the bottom-up approach (Fig. 3.). Firstly, identification of all putative yeast AdoMet-dependent MTases was attempted; secondly, these MTase candidates (MTCs) were ranked according to their predicted potential to act on RNA; thirdly, prediction of the substrate specificity was attempted; fourthly, the functional predictions were experimentally tested by cloning and in vivo and in vitro characterization of the respective MTCs.

In the first step, all MTases identified previously (regardless of their substrate specificity) were included in the "MTCs" database. The yeast proteome was "purged" from all functionally characterized (i.e., annotated) proteins, which were unlikely to encode an MTase function. All functionally uncharacterized yeast ORFs were subjected to the bioinformatic analysis, aimed at the identification of novel homologues of MTases with known structures. In the first step, the IMPALA (Schaffer et al. 1999) and PDB-BLAST (Li et al. 2002) methods were used to identify trivial similarities to characterized proteins or domains. Both methods utilize the PSI-BLAST (Altschul et al. 1997) algorithm to construct position-specific score matrices (PSSMs) from sequence profiles and conduct sequence-profile matching. The key difference is in the "side" on which PSI-BLAST is used to collect homologues: IMPALA compares the query

sequence to a set of pre-computed PSSMs corresponding to protein domains, while PDB-BLAST computes a PSSM for the query and compares it with single sequences from the PDB. Sequences with significant similarities to known MTases were putatively annotated as MTCs and retained for further analysis, while those significantly similar to other proteins were excluded from the query database. The remaining sequences (including all MTCs) were analyzed using the fold-recognition methods, which utilize both sequence and structure information to identify similarities between the query protein and proteins with known structure (reviewed elsewhere in this volume).

The fold-recognition results (J.M.B., unpubl. data) were used to predict further MTase homologues (added to the MTCs database), to identify auxiliary domains in the MTC sequences, and to rank the MTCs according to their relative similarity to any of the known nucleic acid MTase structures (Table 1) versus similarity to other MTases with demonstrated non-RNA MTase activity. Since *S. cerevisiae* does not encodes any DNA MTases, all predicted nucleic acid MTases are obvious candidates for RNA MTases. MTCs with obvious non-RNA MTase specificities (high similarity to non-RNA MTases, genomic context strongly suggesting non-RNA MTase function, etc.) were down-ranked. The remaining yeast MTCs were putatively labeled as primary RNA MTase candidates if they (or their close homologues from different genomes): (1) contained known or predicted nucleic acid-binding domain(s); (2) exhibited strong similarity to known RNA MTases (either in simple sequence searches or advanced fold-recognition analyses); (3) exhibited conservation of predicted catalytic residues characteristic for nucleic acid MTases. MTCs were putatively labeled as secondary RNA MTase candidates if they (or their close homologues from different genomes): (1) exhibited genomic association with proteins involved in RNA metabolism; (2) exhibited phylogenetic correlation with the occurrence of a particular methylated nucleoside. All remaining MTCs were considered as possible RNA MTase candidates (especially if they exhibited a sequence conservation characteristic of many known RNA MTases), but their experimental characterization was given low priority.

The bioinformatic search led to the identification of 20 MTCs, some of which had been previously identified as putative AdoMet-binding proteins (Niewmierzycka and Clarke 1999), others as putative RNA MTases (Anantharaman et al. 2002a). The corresponding genes have been deleted and tRNA modification defects have been studied in vivo and in vitro. Wild-type and mutant cells are labeled with [^{32}P]orthophosphate, then the total tRNA is extracted, digested with various enzymes and the nucleosides are separated on 2D-TLC plates. The analysis is complex, since the same modification can occur at different positions in the same or in different tRNAs. For instance, m^{1}G can be found at position 9 (catalyzed by Trm10p) and at position 37 (catalyzed by Trm5p). Therefore, in vivo, in a strain deleted for Trm5p m^{1}G would still be detected due to the activity of Trm10p, and vice-versa. In some cases,

Table 2. Known and predicted RNA MTases in the yeast *S. cerevisiae*

Protein/ORF	Superfamily	Function/specificity	Identification[a]
Bona fide RNA MTases			
Trm1p	RFM	tRNA:$m^2{}_2G26$	GEN
Trm2p	RFM	tRNA:m^5U54	GEN, BIO
Trm3p	SPOUT	tRNA:Gm18	BIO
Trm4p	RFM	tRNA:$m^5C34,40,48,49$	BIO
Trm5p	RFM	tRNA:m^1G37	BIO
Trm6a&b	RFM	tRNA:m^1A58	GEN
Trm7p	RFM	tRNA:Cm32,Gm34	BIO
Trm8p (MTC1) &82p	RFM	tRNA:m^7G46	BGE, BIO
Trm9p (MTC2)	RFM	tRNA:mcm^5U34/mcm^5s^2U34	GEN, BIO
Trm10p	?	tRNA:m^1G9	BGE
Trm11p (MTC12)	RFM	tRNA:m^2G10	BIO
Nop1p	RFM	Nm (snoRNA-guided)	GEN
Nop2p	RFM	25S rRNA:m^5C	BIO
Dim1p	RFM	18S rRNA:$m^6{}_2A1779,1780$	GEN
Pet56p	SPOUT	mt 21S rRNA:Gm2270	GEN
Mrm2p	RFM	mt 21S rRNA:Um2791	BIO
Tsg1p (MTC20)	RFM	sn(o)RNA:$m^2{}_2{}^7G$	GEN, BIO
Abd1p	RFM	mRNA:m^7G	GEN
Ime4p (MTC17)	RFM	mRNA:m^6A	GEN, BIO
Predicted RNA MTases			
Spb1p	RFM	[25S rRNA:Um2918,Gm2919][b]	GEN, BIO
sc-mtTFB	RFM	[mt 15S rRNA:$m^6{}_2A$[c]][b]	GEN
Rrp8p	RFM	(Nucleolar localization)[c]	GEN
YNL024 c (MTC3)	RFM	[c,d]	BIO
YLR137w (MTC4)	RFM	[c,d]	BIO
YBR271w (MTC5)	RFM	[c,d]	BIO
YDR140 W (MTC6)	RFM	[c]	BIO
YIL110w (MTC7)	RFM	[c,d]	BIO
YJR129 c (MTC8)	RFM	[c,d]	BIO
YLR285w (MTC9)	RFM	[c,d]	BIO
YML005w (MTC10)	RFM	[c]	BIO
YNL063w (MTC11)	RFM	[Putative protein MTase[c]][b]	BIO
KAR4 (MTC13)	RFM	[Inactivated][b]; cofactor of Ime4p[c]	BIO
YNL092 W (MTC14)	RFM	[c]	BIO
YMR209C (MTC15)	RFM	[c]	BIO
YNL022 c (MTC16)	RFM	[RNA:m^5C[c]][b]	BIO
YGR283C (MTC18)	SPOUT	[c]	BIO
YMR310C (MTC19)	SPOUT	[c]	BIO

[a] BIO, bioinformatics; GEN, genetics; BGE, biochemical genomics.
[b] Square brackets indicate function predicted with relatively high confidence.
[c] Function unknown.
[d] A family of paralogous putative MTases.

it has been necessary to purify by hybrid-selection the tRNA labeled in vivo in order to analyze only a certain tRNA. Alternatively, it was also possible to use a double-deleted strain, as it had been the case for Trm11p (see below).

Meanwhile, many of the MTCs from the "top 20" list were demonstrated to be true RNA MTases. *MTC1* was found to correspond to Trm8p, a catalytic subunit of a heterodimeric MTase required for the formation of m^7G46 in yeast tRNA (Alexandrov et al., 2002). *MTC2* was shown by us and by others to be required for the formation of a complex modification at position 34 of the anticodon loop in yeast tRNA (the corresponding MTase has been named Trm9p–S. Clarke, Saccharomyces Genome Database; S.K.P., J.M.B, H.G. and B.L., unpubl. observ.). *MTC19* (Ime4p), which was originally described as controlling meiosis, catalyzes the formation of m^6A in mRNA (Clancy et al. 2002). *MTC20* was found to encode the enzyme (Tgs1p) that catalyzes the trimethylation of the cap of snRNAs and some snoRNAs (Mouaikel et al. 2002). Recently, we have found that *MTC12* (Trm11p) is required for the formation of m^2G10 in yeast tRNA (S.K.P., J.M.B, H.G. and B.L., unpubl. observ.). In yeast tRNA there is also some m^2G, along with $m^2{}_2G$ made at position 26 by Trm1p. A double mutant strain *trm1–0*, *trm11–0* was constructed, in which neither m^2G nor $m^2{}_2G$ were detected at all (S.K.P., J.M.B., H.G. and B.L., unpubl. observation). In our hands, other MTCs exhibited no detectable tRNA MTase activity; they remain plausible candidates for novel MTases acting on other RNAs, and interesting objects for experimental characterization.

4 Conclusions

Although the first RNA MTase was discovered about 40 years ago, progress in the study of this and related enzymes has been very slow until it was possible to identify and clone their genes, and produce recombinant proteins. Currently, the growth in sequence data through large-scale genome sequencing projects has made the identification of novel proteins possible based on comparative analyses. However, in order to understand the detailed biochemical function of enzymes encoded in the genomes, the knowledge of linear protein sequences must be interpreted in the context of their three-dimensional structures. This can be achieved by a combination of structural genomics (providing the template structures) and bioinformatics (providing the links between the templates and protein sequences). The protein function can be inferred by interpretation of the sequence/structure data in the evolutionary context – thereby conserved sites implicated in substrate binding and/or catalysis can be identified. However, the most common way is to annotate new genes and proteins by transferring the function "by homology" without detailed considerations. The weakness of such approach lies in our insufficient understanding of how sequence similarity translates to functional similarity. It may be useful to provide useful hints, but to date this has resulted in

too many overpredictions that are too specific of function for uncharacterized homologues. It is now known that paralogous proteins may exhibit quite distinct new functions since the divergence from the common ancestor they shared with a well-characterized protein used as a reference in the annotation process. Needless to say that experiments in vivo and in vitro are essential to validate the predicted functions based on homology.

Despite the progress in identification of new RNA modification enzymes by the "-omics" and "-matics" approaches, our knowledge of the details of the enzymology of RNA modification remains limited relative to the number and variety of modified nucleotides that have been identified so far in the various RNAs from the three biological domains. The availability of complete genome sequences of many organisms from distinct branches of the Tree of Life creates the opportunity to explore the functional content of the genomes and evolutionary relationships between them at a new qualitative level. The analysis of the conserved genome neighborhood and phyletic profiling allows for the prediction of new functions without referring to homology or protein structure and to detect the cases of functional convergence in evolution. This methodology, however, is dependent on our knowledge of biochemical reactions and metabolic pathways leading to a generation of modified nucleosides in RNA and on the experimental data concerning the presence or absence of particular modifications in organisms with fully sequenced genomes.

The development of non-identical functions by orthologs and functional convergence of unrelated enzymes is poorly understood. RNA modification (and especially RNA methylation) is a perfect object for the analysis of these processes, as many cases of such functionally and structurally diversified RNA MTases have been reported. The study of enzymology of RNA modification by a combination of theoretical and experimental approaches may provide the key to understanding the basic evolutionary processes and determining the relationships between Eubacteria, Archaea, and Eukaryota, and help to reconstruct the true Tree of Life.

Acknowledgements. J.M.B.'s research on RNA MTases is supported by EMBO and Howard Hughes Medical Institute (Young Investigator Programme award) and by the Fellowship for Young Scientists from the Foundation for Polish Science. L.D. is a Research Associate of the FNRS (Fonds National de la Recherche Scientifique) and is supported by grants from the FRFC (Fonds pour la Recherche Fondamentale Collective), from the French Community of Belgium (Actions de Recherches Concertées) and from the Université Libre de Bruxelles (Fonds E. Defay). H.G. is supported by the CNRS (Programme Interdépartemental de Géomicrobiologie des Environnements Extrèmes, Geomex 2002–2003). S.K.P. has a fellowship from the Ministère des Affaires Étrangères. B.L. is a recipient of grant No. 5914 from the "Association pour la Recherche sur le Cancer" and is supported by the "Fondation pour la Recherche Médicale" and the "Centre National de la Recherche Scientifique".

References

Agris PF (1996) The importance of being modified: roles of modified nucleosides and Mg^{2+} in RNA structure and function. Prog Nucleic Acid Res Mol Biol 53:79–129

Ahn HJ, Kim HW, Yoon HJ, Lee BI, Suh SW, Yang JK (2003) Crystal structure of tRNA(m'G37)methyltransferase: insights into tRNA recognition. EMBO J 22:2593–2603

Alexandrov A, Martzen MR, Phizicky EM (2002) Two proteins that form a complex are required for 7-methylguanosine modification of yeast tRNA. RNA 8:1253–1266

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Anantharaman V, Koonin EV, Aravind L (2001) TRAM, a predicted RNA-binding domain, common to tRNA uracil methylation and adenine thiolation enzymes. FEMS Microbiol Lett 197:215–221

Anantharaman V, Koonin EV, Aravind L (2002a) Comparative genomics and evolution of proteins involved in RNA metabolism. Nucleic Acids Res 30:1427–1464

Anantharaman V, Koonin EV, Aravind L (2002b) SPOUT: a class of methyltransferases that includes SpoU and TrmD RNA methylase superfamilies, and novel superfamilies of predicted prokaryotic RNA methylases. J Mol Microbiol Biotechnol 4:71–75

Anderson J, Phan L, Cuesta R, Carlson BA, Pak M, Asano K, Bjork GR, Tamame M, Hinnebusch AG (1998) The essential Gcd10p-Gcd14p nuclear complex is required for 1-methyladenosine modification and maturation of initiator methionyl-tRNA. Genes Dev 12:3650–3662

Anderson J, Phan L, Hinnebusch AG (2000) The Gcd10p/Gcd14p complex is the essential two-subunit tRNA(1-methyladenosine) methyltransferase of *Saccharomyces cerevisiae*. Proc Natl Acad Sci USA 97:5173–5178

Aravind L, Koonin EV (1999) Novel predicted RNA-binding domains associated with the translation machinery. J Mol Evol 48:291–302

Aravind L, Koonin EV (2001) THUMP – a predicted RNA-binding domain shared by 4-thiouridine, pseudouridine synthases and RNA methylases. Trends Biochem Sci 26: 215–217.

Auffinger P, Westhof E (1998) Location and distribution of modified nucleosides in tRNA. In: Grosjean H, Benne R (eds) Modification and editing of RNA. ASM Press, Washington, pp 569–576

Auffinger P, Westhof E (2001) An extended structural signature for the tRNA anticodon loop. RNA 7:334–341

Bjork GR (1995) Genetic dissection of synthesis and function of modified nucleosides in bacterial transfer RNA. Prog Nucleic Acid Res Mol Biol 50:263–338

Bugl H, Fauman EB, Staker BL, Zheng F, Kushner SR, Saper MA, Bardwell JC, Jakob U (2000) RNA methylation under heat shock control. Mol Cell 6:349–360

Bujnicki JM (1999) Comparison of protein structures reveals monophyletic origin of the AdoMet-dependent methyltransferase family and mechanistic convergence rather than recent differentiation of N^4-cytosine and N^6-adenine DNA methylation. In Silico Biol 1:1–8 (http://www.bioinfo.de/isb/1999–01/0016/)

Bujnicki JM (2000) Phylogenomic analysis of 16S rRNA:(guanine-N^2) methyltransferases suggests new family members and reveals highly conserved motifs and a domain structure similar to other nucleic acid amino-methyltransferases. FASEB J 14:2365–2368

Bujnicki JM (2001a) In silico analysis of the tRNA:m^{1}A58 methyltransferase family: homology-based fold prediction and identification of new members from Eubacteria and Archaea. FEBS Lett 507:123–127

Bujnicki JM (2001b) Understanding the evolution of restriction-modification systems: clues from sequence and structure comparisons. Acta Biochim Pol 48:1–33

Bujnicki JM, Feder M, Radlinska M, Rychlewski L (2001) mRNA:guanine-N^7 cap methyltransferases: identification of novel members of the family, evolutionary analysis, homology modeling, and analysis of sequence-structure-function relationships. BMC Bioinformatics 2:2

Bujnicki JM, Blumenthal RM, Rychlewski L (2002a) Sequence analysis and structure prediction of 23S rRNA:m^{1}G methyltransferases reveals a conserved core augmented with a putative Zn-binding domain in the N-terminus and family-specific elaborations in the C-terminus. J Mol Microbiol Biotechnol 4:93–99

Bujnicki JM, Feder M, Radlinska M, Blumenthal RM (2002b) Structure prediction and phylogenetic analysis of a functionally diverse family of proteins homologous to the MT-A70 subunit of the human mRNA:m^{6}A methyltransferase. J Mol Evol 55:431–444

Bujnicki JM, Leach RA, Debski J, Rychlewski L (2002 c) Bioinformatic analyses of the tRNA:(guanine 26, *N*^{2},*N*^{2})-dimethyltransferase (Trm1) family. J Mol Microbiol Biotechnol 4:405–415

Bujnicki JM, Rychlewski L (2000) Prediction of a novel RNA 2'-O-ribose methyltransferase subfamily encoded by the *Escherichia coli* YgdE open reading frame and its orthologs. Acta Microbiol Pol 49:253–260

Bujnicki JM, Rychlewski L (2001) Reassignment of specificities of two cap methyltransferase domains in the reovirus λ2 protein. Genome Biol 2:38

Bujnicki JM, Rychlewski L (2002a) In silico identification, structure prediction and phylogenetic analysis of the 2′-O-ribose (cap 1) methyltransferase domain in the large structural protein of ssRNA negative-strand viruses. Protein Eng 15:101–108

Bujnicki JM, Rychlewski L (2002b) RNA:(guanine-N^2) methyltransferases RsmC/RsmD and their homologs revisited – bioinformatic analysis and prediction of the active site based on the uncharacterized Mj0882 protein structure. BMC Bioinformatics 3:10

Caldas T, Binet E, Bouloc P, Costa A, Desgres J, Richarme G (2000) The FtsJ/RrmJ heat shock protein of *Escherichia coli* is a 23 S ribosomal RNA methyltransferase. J Biol Chem 275:16414–16419

Calvo O, Cuesta R, Anderson J, Gutierrez N, Garcia-Barrio MT, Hinnebusch AG, Tamame M (1999) Gcd14p, a repressor of GCN4 translation, cooperates with Gcd10p and Lhp1p in the maturation of initiator methionyl-tRNA in *Saccharomyces cerevisiae*. Mol Cell Biol 19:4167–4181

Cavaille J, Chetouani F, Bachellerie JP (1999) The yeast *Saccharomyces cerevisiae* YDL112w ORF encodes the putative 2′-O-ribose methyltransferase catalyzing the formation of Gm18 in tRNAs. RNA 5:66–81

Cermakian N, Cedergren R (1998) Modified nucleosides always were: an evolutionary model. In: Grosjean H, Benne R (eds) Modification and editing of RNA. ASM Press, Washington, pp 535–542

Clancy MJ, Shambaugh ME, Timpte CS, Bokar JA (2002) Induction of sporulation in *Saccharomyces cerevisiae* leads to the formation of N^6-methyladenosine in mRNA: a potential mechanism for the activity of the IME4 gene. Nucleic Acids Res 30:4509–4518

Clouet d'Orval B, Bortolin ML, Gaspin C, Bachellerie JP (2001) Box C/D RNA guides for the ribose methylation of archaeal tRNAs. The tRNA^{Trp} intron guides the formation of two ribose-methylated nucleosides in the mature tRNA^{Trp}. Nucleic Acids Res 29:4518–4529

Curran JF (1998) Modified nucleosides in translation. In: Grosjean H, Benne R(eds), Modification and editing of RNA. ASM Press, Washington, pp 493–516

Davis DR (1998) Properties of modified nucleosides, In: Grosjean H, Benne R (eds) Modification and editing of RNA. ASM Press, Washington, pp 85–102

Dixon M, Fauman EB, Ludwig ML (1999) The black sheep of the family: AdoMet-dependent methyltransferases that do not fit the consensus structural fold. . In: Cheng X, Blumenthal RM (eds) S-Adenosylmethionine-dependent methyltransferases: structures and functions. World Scientific Inc, Singapore, pp 39–54

Drennan CL, Huang S, Drummond JT, Matthews RG, Lidwig ML (1994) How a protein binds B12: A 3.0 A X-ray structure of B12-binding domains of methionine synthase. Science 266:1669–1674

Droogmans L, Roovers M, Bujnicki JM, Tricot C, Hartsch T, Stalon V, Grosjean H (2003) Cloning and characterization of tRNA (m^1 A58) methyltransferase (TrmI) from *Thermus thermophilus* HB27, a protein required for cell growth at extreme temperatures. Nucleic Acids Res 31:2148–2156

Egloff MP, Benarroch D, Selisko B, Romette JL, Canard B (2002) An RNA cap (nucleoside-2′-O-)-methyltransferase in the flavivirus RNA polymerase NS5: crystal structure and functional characterization. EMBO J 21:2757–2768

Eisen JA (1998) Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. Genome Res 8:163–167

Fauman EB, Blumenthal RM, Cheng X. (1999) Structure and evolution of AdoMet-dependent MTases. In: Cheng X, Blumenthal RM (eds) S-Adenosylmethionine-dependent methyltransferases: structures and functions, World Scientific Inc, Singapore, pp 1–38

Feder M, Pas J, Wyrwicz LS, Bujnicki JM (2003) Molecular phylogenetics of the RrmJ/fibrillarin superfamily of ribose 2′-O-methyltransferases. Gene 302:129–138

Fleissner E, Borek E (1962) A new enzyme of RNA synthesis: RNA methylase. Proc Natl Acad Sci USA. 48:1199–1203

Galperin MY, Koonin EV (2000) Who's your neighbor? New computational approaches for functional genomics. Nat Biotechnol 18:609–613

Galperin MY, Walker DR, Koonin EV (1998) Analogous enzymes: independent inventions in enzyme evolution. Genome Res 8:779–790

Garcia GA, Goodenough-Lashua DM (1998) Modifying and editing enzyme mechanisms. In: Grosjean H, Benne R (eds) Modification and editing of RNA. ASM Press, Washington, pp135–168

Grosjean H, Auxilien S, Constantinesco F, Simon C, Corda Y, Becker HF, Foiret D, Morin A, Jin YX, Fournier M, Fourrey JL (1996) Enzymatic conversion of adenosine to inosine and to N^1-methylinosine in transfer RNAs: a review. Biochimie 78:488–501

Grosjean H, Motorin Y, Morin A (1998) RNA-modifying and RNA-editing enzymes: methods for their identification. In: Grosjean H, .Benne R(eds) Modification and editing of RNA. ASM Press, Washington, pp.21–46

Gu XR, Gustafsson C, Ku J, Yu M, Santi DV (1999) Identification of the 16S rRNA m^5C967 methyltransferase from *Escherichia coli*. Biochemistry 38:4053–4057

Gupta A, Kumar PH, Dineshkumar TK, Varshney U, Subramanya HS (2001) Crystal structure of Rv2118 c: An AdoMet-dependent methyltransferase from *Mycobacterium tuberculosis* H37Rv. J Mol Biol 312:381–391

Gustafsson C, Reid R, Greene PJ, Santi DV (1996) Identification of new RNA modifying enzymes by iterative genome search using known modifying enzymes as probes. Nucleic Acids Res 24:3756–3762

Heurgue-Hamard V, Champ S, Engstrom A, Ehrenberg M, Buckingham RH (2002) The hemK gene in *Escherichia coli* encodes the N^5-glutamine methyltransferase that modifies peptide release factors. EMBO J 21:769–778

Hodel AE, Gershon PD, Quiocho FA (1998) Structural basis for sequence-nonspecific recognition of 5′-capped mRNA by a cap-modifying enzyme. Mol Cell 1:443–447
Hong B, Brockenbrough JS, Wu P, Aris JP (1997) Nop2p is required for pre-rRNA processing and 60S ribosome subunit synthesis in yeast. Mol Cell Biol 17:378–388
Hopper AK, Phizicky EM (2003) tRNA transfers to the limelight. Genes Dev 17:162–180
Huynen M, Dandekar T, Bork P (1998) Differential genome analysis applied to the species-specific features of Helicobacter pylori. FEBS Lett 426:1–5
Jackman JE, Montange RK, Malik HS, Phizicky EM (2003) Identification of the yeast gene encoding the tRNA m^1G methyltransferase responsible for modification at position 9. RNA 9:574–585
Jeffery CJ (1999) Moonlighting proteins. Trends Biochem Sci 24:8–11
Jeltsch A, Pingoud A (1996) Horizontal gene transfer contributes to the wide distribution and evolution of type II restriction-modification systems. J Mol Evol 42:91–96
Johansson MJ, Bystrom AS (2002) Dual function of the tRNA(m^5U54)methyltransferase in tRNA maturation. RNA 8:324–335
Koonin EV (1994) Prediction of an rRNA methyltransferase domain in human tumor-specific nucleolar protein P120. Nucleic Acids Res 22:2476–2478
Koonin EV, Mushegian AR, Bork P (1996) Non-orthologous gene displacement. Trends Genet 12:334–336
Koonin EV, Rudd KE (1993) SpoU protein of *Escherichia coli* belongs to a new family of putative rRNA methylases. Nucleic Acids Res 21:5519
Kressler D, Rojo M, Linder P, Cruz J (1999) Spb1p is a putative methyltransferase required for 60S ribosomal subunit biogenesis in *Saccharomyces cerevisiae*. Nucleic Acids Res 27:4598–4608
Lafontaine DL, Preiss T, Tollervey D (1998) Yeast 18S rRNA dimethylase Dim1p: a quality control mechanism in ribosome synthesis? Mol Cell Biol 18:2360–2370
Li W, Jaroszewski L, Godzik A (2002) Sequence clustering strategies improve remote homology recognitions while reducing search times. Protein Eng 15:643–649
Lim K, Zhang H, Tempczyk A, Krajewski W, Bonander N, Toedt J, Howard A, Eisenstein E, Herzberg O (2003) Structure of the YibK methyltransferase from *Haemophilus influenzae* (HI0766): a cofactor bound at a site formed by a knot. Proteins 51:56–67
Limbach PA, Crain PF, McCloskey JA (1994) Summary: the modified nucleosides of RNA. Nucleic Acids Res 22:2183–2196
Limbach PA, Crain PF, McCloskey JA (1995) Characterization of oligonucleotides and nucleic acids by mass spectrometry. Curr Opin Biotechnol 6:96–102
Lowe TM, Eddy SR (1999) A computational screen for methylation guide snoRNAs in yeast. Science 283:1168–1171
Marcotte EM (2000) Computational genetics: finding protein function by nonhomology methods. Curr Opin Struct Biol 10:359–365
Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D (1999) Detecting protein function and protein-protein interactions from genome sequences. Science 285:751–753
Marmorstein R (2003) Structure of SET domain proteins: a new twist on histone methylation. Trends Biochem. Sci. 28:59–62
Martzen MR, McCraith SM, Spinelli SL, Torres FM, Fields S, Grayhack EJ, Phizicky EM (1999) A biochemical genomics approach for identifying genes by the activity of their products. Science 286:1153–1155
Michel G, Sauve V, Larocque R, Li Y, Matte A, Cygler M (2002) The structure of the RlmB 23S rRNA methyltransferase reveals a new methyltransferase fold with a unique knot. Structure 10:1303–1315
Mosbacher TG, Bechthold A, Schulz GE (2003) Crystal structure of the avilamycin resistance-conferring methyltransferase AviRa from *Streptomyces viridochromogenes*. J Mol Biol 329:147–157

MotorinY, Grosjean H (1998) Chemical structures and classification of posttranscriptionally modified nucleosides in RNA. In: Grosjean H,Benne R (eds) Modification and editing of RNA. ASM Press, Washington, pp 543–549

Motorin Y, Grosjean H (1999). Multisite-specific tRNA:m^5C-methyltransferase (Trm4) in yeast *Saccharomyces cerevisiae*: identification of the gene and substrate specificity of the enzyme. RNA 5:1105–1118

Mouaikel J, Verheggen C, Bertrand E, Tazi J, Bordonne R (2002) Hypermethylation of the cap structure of both yeast snRNAs and snoRNAs requires a conserved methyltransferase that is localized to the nucleolus. Mol Cell 9:891–901

Nakahigashi K, Kubo N, Narita SS, Shimaoka T, Goto S, Oshima T, Mori H, Maeda M, Wada C, Inokuchi H (2002) HemK, a class of protein methyl transferase with similarity to DNA methyl transferases, methylates polypeptide chain release factors, and hemK knockout induces defects in translational termination. Proc Natl Acad Sci USA 99:1473–1478

Niewmierzycka A, Clarke S (1999) S-Adenosylmethionine-dependent methylation in *Saccharomyces cerevisiae*. Identification of a novel protein arginine methyltransferase. J Biol Chem 274: 814–824

Nordlund ME, Johansson JO, Pawel-Rammingen U, Bystrom AS (2000) Identification of the TRM2 gene encoding the tRNA(m^5U54) methyltransferase of *Saccharomyces cerevisiae*. RNA 6:844–860

Nureki O, Shirouzu M, Hashimoto K, Ishitani R, Terada T, Tamakoshi M, Oshima T, Chijimatsu M, Takio K, Vassylyev DG, Shibata T, Inoue Y, Kuramitsu S, Yokoyama S (2002) An enzyme with a deep trefoil knot for the active-site architecture. Acta Crystallogr D Biol Crystallogr 58:1129–1137

Omer AD, Ziesche S, Ebhardt H, Dennis PP (2002) In vitro reconstitution and activity of a C/D box methylation guide ribonucleoprotein complex. Proc Natl Acad Sci USA 99:5289–5294

Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N (1999) The use of gene clusters to infer functional coupling. Proc Natl Acad Sci USA 96:2896–2901

Pearson WR, Lipman DJ (1988) Improved tools for biological sequence comparison. Proc Natl Acad Sci USA 85:2444–2448

Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci USA 96:4285–4288

Persson BC, Gustafsson C, Berg DE, Bjork GR (1992) The gene for a tRNA modifying enzyme, m^5U54-methyltransferase, is essential for viability in *Escherichia coli*. Proc Natl Acad Sci USA 89:3995–3998

Persson BC, Jager G, Gustafsson C (1997) The spoU gene of *Escherichia coli*, the fourth gene of the spoT operon, is essential for tRNA (Gm18) 2′-O-methyltransferase activity. Nucleic Acids Res 25:4093–4097

Phizicky EM, Martzen MR, McCraith SM, Spinelli SL, Xing F, Shull NP, Van Slyke C, Montagne RK, Torres FM, Fields S, Grayhack EJ (2002) Biochemical genomics approach to map activities to genes. Methods Enzymol 350:546–559

Pintard L, Bujnicki JM, Lapeyre B, Bonnerot C (2002a) MRM2 encodes a novel yeast mitochondrial 21S rRNA methyltransferase. EMBO J 21:1139–1147

Pintard L, Kressler D, Lapeyre B (2000) Spb1p is a yeast nucleolar protein associated with Nop1p and Nop58p that is able to bind S-adenosyl-L-methionine in vitro. Mol Cell Biol 20:1370–1381

Pintard L, Lecointe F, Bujnicki JM, Bonnerot C, Grosjean H, Lapeyre B (2002b) Trm7p catalyses the formation of two 2′-O-methylriboses in yeast tRNA anticodon loop. EMBO J 21:1811–1820

Reid R, Greene P, Santi DV (1999) Exposition of a family of RNA m^5C methyltransferases from searching genomic and proteomic sequences. Nucleic Acids Res 27:3138–3145

Reinisch KM, Nibert ML, Harrison SC (2000) Structure of the reovirus core at 3.6 A resolution. Nature 404:960–967

Rogozin IB, Makarova KS, Murvai J, Czabarka E, Wolf YI, Tatusov RL, Szekely LA, Koonin EV (2002) Connected gene neighborhoods in prokaryotic genomes. Nucleic Acids Res 30:2212–2223

Romeo JM, Delk AS, Rabinowitz JC (1974) The occurrence of a transmethylation reaction not involving S-adenosylmethionine in the formation of ribothymidine in *Bacillus subtilis* transfer-RNA. Biochem Biophys Res Commun 61:1256–1261

Rost B (2002) Enzyme function less conserved than anticipated. J Mol Biol 318:595–608

Rozenski J, Crain PF, McCloskey JA (1999) The RNA Modification Database: 1999 update. Nucleic Acids Res 27:196–197

Schaffer AA, Wolf YI, Ponting CP, Koonin EV, Aravind L, Altschul SF (1999) IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. Bioinformatics 15:1000–1011

Schluckebier G, Zhong P, Stewart KD, Kavanaugh TJ, Abad-Zapatero C (1999) The 2.2 A structure of the rRNA methyltransferase ErmC′ and its complexes with cofactor and cofactor analogs: implications for the reaction mechanism. J Mol Biol 289:277–291

Schubert HL, Blumenthal RM, Cheng X (2003) Many paths to methyltransfer: a chronicle of convergence. Trends Biochem Sci 28:329–335

Schubert HL, Wilson KS, Raux E, Woodcock SC, Warren MJ (1998) The X-ray structure of a cobalamin biosynthetic enzyme, cobalt- precorrin-4 methyltransferase. Nat Struct Biol 5:585–592

Schubot FD, Chen CJ, Rose JP, Dailey TA, Dailey HA, Wang BC (2001) Crystal structure of the transcription factor sc-mtTFB offers insights into mitochondrial transcription. Protein Sci 10: 1980–1988

Seidel-Rogol B.L., McCulloch V, Shadel GS (2003) Human mitochondrial transcription factor B1 methylates ribosomal RNA at a conserved stem-loop. Nat Genet 33:23–24

Smith JE, Cooperman BS, Mitchell P (1992) Methylation sites in *Escherichia coli* ribosomal RNA: localization and identification of four new sites of methylation in 23S rRNA. Biochemistry 31:10825–10834

Sprinzl M, Horn C, Brown M, Ioudovitch A, Steinberg S (1998) Compilation of tRNA sequences and sequences of tRNA genes. Nucleic Acids Res 26:148–153

Staker BL, Korber P, Bardwell JC, Saper MA (2000) Structure of Hsp15 reveals a novel RNA-binding motif. EMBO J 19:749–757

Terns MP, Terns RM (2002) Small nucleolar RNAs: versatile trans-acting molecules of ancient evolutionary origin. Gene Exp 10:17–39

Todd AE, Orengo CA, Thornton JM (2002) Sequence and structural differences between enzyme and nonenzyme homologs. Structure (Camb) 10:1435–1451

Tollervey D, Lehtonen H, Jansen R, Kern H, Hurt EC (1993) Temperature-sensitive mutations demonstrate roles for yeast fibrillarin in pre-rRNA processing, pre-rRNA methylation, and ribosome assembly. Cell 72:443–457

Tscherne JS, Nurse K, Popienick P, Michel H, Sochacki M, Ofengand J (1999). Purification, cloning, and characterization of the 16S RNA m^5C967 methyltransferase from *Escherichia coli*. Biochemistry 38:1884–1892

Vestergaard B, Van LB, Andersen GR, Nyborg J, Buckingham RH, Kjeldgaard M (2001) Bacterial polypeptide release factor RF2 is structurally distinct from eukaryotic eRF1. Mol Cell 8:1375–1382

Wang H, Boisvert D, Kim KK, Kim R, Kim SH (2000) Crystal structure of a fibrillarin homologue from *Methanococcus jannaschii*, a hyperthermophile, at 1.6 A resolution. EMBO J 19:317–323

Winkler ME (1998) Genetics and regulation of base modification in the tRNA and rRNA of Prokaryotes and Eukaryotes, In:Grosjean H, Benne R (eds) Modification and editing of RNA. ASM Press, Washington, pp 441–470

Wu P, Brockenbrough JS, Paddy MR, Aris JP (1998) NCL1, a novel gene for a non-essential nuclear protein in *Saccharomyces cerevisiae*. Gene 220:109–117

Xing F, Martzen MR, Phizicky EM (2002) A conserved family of *Saccharomyces cerevisiae* synthases effects dihydrouridine modification of tRNA. RNA 8: 370–381

Yu L, Petros AM, Schnuchel A, Zhong P, Severin JM, Walter K, Holzman TF, Fesik SW (1997) Solution structure of an rRNA methyltransferase (ErmAM) that confers macrolide-lincosamide-streptogramin antibiotic resistance. Nat Struct Biol 4:483–489

Zarembinski T, Kim Y, Peterson K, Christendat D, Dharamsi A, Arrowsmith CH, Edwards AM, Joachimiak A (2002) Deep trefoil knot implicated in RNA binding found in an archaebacterial protein. Proteins 50:177–183

Finding Missing tRNA Modification Genes: a Comparative Genomics Goldmine

V. de Crécy-Lagard

1 Missing tRNA Modification Genes

1.1 tRNA Modifications

As the adapters between mRNAs and the elongating peptide chain, transfer RNAs (tRNA) are at the nexus of the genetic code and of the translation apparatus. Prior to their participation in translation, tRNAs must undergo extensive processing of the nascent transcript. The post-transcriptional processing of tRNAs involves a number of functionally distinct events essential for tRNA maturation (Altman et al. 1995; Björk 1995; Deutscher 1995; Westaway and Abelson 1995). The phenomenon of nucleoside modification is perhaps the most remarkable of these events, and results in a wealth of structural changes to the canonical nucleosides (Björk 1995). Although other RNA species also exhibit varying degrees of nucleoside modification, it is only in the tRNA that a rich structural diversity is realized.

Nucleoside modification typically occurs to ~10 % of the nucleosides in a particular tRNA, but can involve as many as 25 % of the nucleosides (Björk 1995). Over 80 modified nucleosides have been characterized (Björk 1995), many of which are conserved across broad phylogenetic boundaries. The nature of nucleoside modification varies from simple methylation of the base or ribose ring to extensive "hypermodification" of the canonical bases, the latter of which can result in radical structural changes and involve multiple enzymatic steps to complete. The lack of mutant phenotypes for some modification enzymes was initially interpreted as precluding an important physiological role for tRNA modification. However, with the realization that modified nucleosides are conserved in phylogenetically diverse organisms, and

V. de Crécy-Lagard
Molecular Biology Department, The Scripps Research Institute, BCC-379, 10550 N. Torrey Pines Road, La Jolla, California 92037, USA

Nucleic Acids and Molecular Biology, Vol. 15
Janusz M. Bujnicki (Ed.)
Practical Bioinformatics

that an impressive amount of genetic information codes for tRNA-modifying enzymes [an estimated 1 % of the total genome in the bacterium *Salmonella typhimurium* (Björk and Kohli 1990) and the eukaryote *Saccharomyces cerevisiae* (Hopper and Phizicky 2003)], an appreciation for the importance of modified nucleosides to the basic physiology of the cell is emerging. It is now recognized that modified nucleosides are integral to tRNA function at many levels, influencing translation (Björk 1992; Muramatsu et al. 1988; Yokoyama and Nishimura 1995), tRNA structure and stability (Björk 1995; Derrick and Horowitz 1993; Horie et al. 1985; Kowalak et al. 1994; Perret et al. 1990), and regulatory events (Persson 1993). In spite of the importance of modified nucleosides to tRNA function, the contributions that specific modifications make to tRNA are well established in only a few cases (Björk 1995; Björk and Kohli 1990), and our understanding of the biosynthesis of the various modified nucleosides is mainly rudimentary.

1.2 Compilation of the Missing tRNA Modification Genes

The lack of fundamental knowledge about the biosynthetic pathways involved in nucleoside modification is due to their resistance to traditional biochemical and genetic characterization. Identification and purification of relevant enzyme activities from crude cell-free extracts are complicated by the difficulty of obtaining appropriate tRNA substrates, the presence of endogenous RNases that degrade the RNA substrates and products, a lack of appropriate assays, and the typically low abundance of the enzymes involved in RNA modification. Traditional genetic approaches are hindered by the lack of clearly defined phenotypes in mutants, and the fact that unambiguous identification of a gene involved in RNA modification is ultimately dependent on determining the presence or absence of the specific modified nucleoside in tRNA, a laborious and technically challenging process when working with large libraries of mutants. As a consequence, it is estimated that approximately 30–50 % of the tRNA modification genes remain uncharacterized (Eastwood-Leung et al. 1998). *Escherichia coli* and *Saccharomyces cerevisiae* are the best-characterized organisms with respect to tRNA modifications genes. A compilation of the *E. coli* tRNA modification genes is given in Table 1; approximately 25 % has not been clearly identified. In *S. cerevisiae* 28 genes have been identified and it has been estimated that 15–20 are missing (H. Grosjean, pers. comm.). The status of the current literature is summarized in Fig. 1.

Out of the 81 modifications found in tRNA (Motorin and Grosjean 2001), the synthesis of 20 has been fully genetically characterized, the pathways for 33 are only partially elucidated and, for the remaining 27, no genetic information is known. The number of missing tRNA modification genes can only be estimated as in some cases (e.g., wyeosine) the number of steps in the pathways are unknown and could implicate between four and eight genes (Droog-

Table 1. Known *E. coli* tRNA modification genes

Modification[a]	Locus	Swiss-Prot	Reference
Ψ_{13}	*truD=ygbO*	O57261	Kaya and Ofengand (2003)
Ψ_{32}	*rluA=yabO*	P39219	Raychaudhuri et al. (1999)
Ψ_{38-40}	*truA= hisT*	P07649	Kammen et al. (1988)
Ψ_{55}	*truB=yhbA*	P09171	Gutgsell et al. (2000; Nurse et al. (1995)
Ψ_{65}	*truC=yqcB*	Q46918	Del Campo et al. (2001)
$D_{16,17,20,20a}$	*dusA=yjbN*	P32695	Bishop et al. (2002)
	dusB=yhdG	P25717	
	dusC=yohI	P33371	
I_{34}	*tadA=yfhC*	P30134	Wolf et al. (2002)
m^2A_{37}	*trmG=yfiF?*[b]	P33635	Gustafsson et al. (1996, this work)
m^6A_{37}	?[c]		
Cm_{32}	?		
Um_{32}	?		
m^5U_{54}	*trmA*	P23003	Ny and Björk (1980); Persson et al. (1992)
m^1G_{37}	*trmD*	P07020	Bystrom and Björk (1982)
Gm_{18}	*trmH= spoU*	P19396	Persson et al. (1997)
m^7G_{46}	*trmB=yggH*	P32049	De Bie et al. (2003)
s^2C_{32}	*stcA=?*[d]		
s^4U_8	*thiI =nuvA*	P77718	Mueller et al. (1998)
	icsS=nuvC	P39171	Lauhon and Kambampati (2000)
i^6A_{37}	*miaA*	P16384	Caillet and Droogmans (1988)
$ms^2i^6A_{37}$	*miaB= yleA*	P77645	Esberg et al. (1999)
	icsS		Lauhon (2002); Nilsson et al. (2002)
s^2U_{34}	*mnmA=trmU=asuE*		P25745 Green et al. (1996); Kambampati and Lauhon (2003)
	icsS=nuvC	P39171	
$cmnm^5s^2U_{34}$	*mnmE=trmE*	P25522	Cabedo et al. (1999)
$cmnm^5U_{34}$	*mnmG=gidA*	P17112	Bregeon et al. (2001)
	gidB?		Kambampati and Lauhon (2003)
$mnm^5s^2U_{34}$	*mnmC=yfcK?*[b]	P77182	Björk and Kjellin-Straby (1978, this work)
mnm^5U_{34}			
$mnm^5se^2U_{34}$	*selD*	P16456	Leinfelder et al. (1990)
	icsS		Mihara et al. (2002)
	?		
mnm^5Um_{34}	?		
Q_{34}	*queA*	P21516	Reuter et al. (1991)
	tgt	P19675	Frey et al. (1988)
	queB=?		
preQ0, preQ1	*queC=ybaX*	P77756	Reader et al. (in press)
	queD=ygcM	Q46903	
	queE=ygcF	P55139	
	queF=yqcD	Q46920	
mo^5U_{34}	*aroABCDE*		Björk (1980); Hagervall et al. (1990)
cmo^5U_{34}	?		
$mcmo^5U_{34}$			
ac^4A_{34}	?		

Table 1. (*Continued*)

Modification[a]	Locus	Swiss-Prot	Reference
s^2T	?		
t^6A_{37}	?		
mt^6A_{37}	*mtaA*[d]		
k^2C_{34}	*tilS*[d]		Soma et al. (2003)
acp^3U_{47}	?		
Predicted missing		12	
Total known	30		

[a] The list of the modified bases found in *E. coli* was taken from Björk (1996). The abbreviations are taken from Motorin and Grosjean (2001).
[b] Question marks denote missing pathways or steps.
[c] Prediction from genomics, no experimental proof.
[d] The gene has been identified, but the accession number is not yet available.

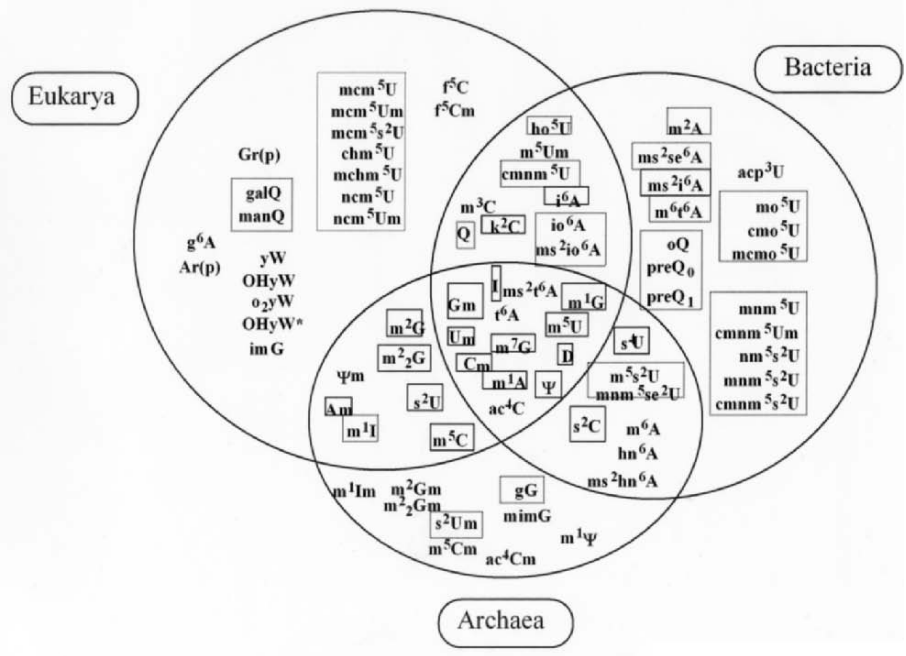


Fig. 1. Distribution of modified bases found in tRNA among kingdoms. The bases for which all the pathway genes have been identified are *boxed in full*, those for which the pathway genes are partially identified are *boxed with dash lines*

mans and Grosjean 1987; Munch and Thiebe 1975; Smith et al. 1985). The analysis is further complicated by the fact that the same modification might be synthesized by different enzymes or pathways in different organisms. For example, the mechanisms of formation of m^1I are different in Archaea and in Eukaryots. In the latter the deamination occurs before the methylation whereas the reverse happens in Archaea (Grosjean et al. 1995, 1996).

It is clear from this analysis, however, that at least 30 % of the modification genes are still missing. The availability of hundreds of whole genome sequences allows the use of radically new approaches to identify them.

2 Comparative Genomics: an Emerging Tool to Identify Missing Genes

The data generated from genome sequencing programs (16 Archaea, 106 Bacteria and 18 Eukarya fully sequenced and published to date) (http://wit.integratedgenomics.com/GOLD/), has revealed how much we have yet to learn before understanding the roles of all the proteins in a cell. Even in the best genetically characterized organisms, a third of the genes have no assigned function (Blattner et al. 1997; Kunst et al. 1997). Systematic approaches such as structural genomics initiatives or systematic interaction mapping can lead to elucidation of some functions (Huynen et al. 2003; Mittl and Grutter 2001). However, there remains a plethora of enzymatic activities or pathways for which the genes remain unknown (Cordwell 1999), and "comparative genomics" is emerging as a powerful approach for identifying these "missing" genes (Osterman and Overbeek 2003; Fig. 2).

These methods integrate several types of genomic data to make predictions that can then be tested experimentally. The kind of information that can be derived from whole genome datasets include:

1. Clustering data: Genes of a given pathway have a higher probability of being physically linked on the chromosome (Overbeek et al. 1999).
2. Protein fusion events: Genes of the same pathway can be fused to encode multi-domain proteins in some organisms (Enright et al. 1999).
3. Phylogenetic occurrence profiles or signatures: phylogenetic profiles can be generated from the profile of one known gene in the pathway under study or from information about the presence or absence of a given pathway among sequenced organisms (Pellegrini et al. 1999).
4. Shared regulatory sites: pathway genes are often regulated by a common protein recognizing a specific DNA sequence (Gelfand et al. 2000).
5. Co-expression: now that expression array data is available, particularly for *S. cerevisiae* (http://db.yeastgenome.org/cgi-bin/SGD/expression/expressionConnection.pl) and *E. coli* (http://www.genome.wisc.edu/functional/microarray.htm), co-expression correlations can allow for identification of genes that are in the same metabolic pathway.

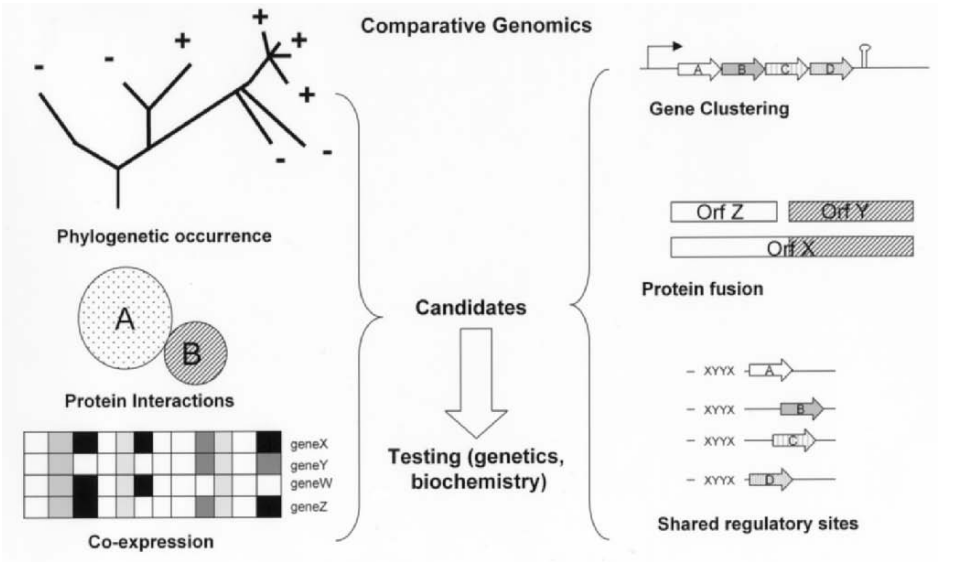


Fig. 2. Summary of comparative genomics approaches

6. Protein interaction networks: using two-hybrid screens protein interaction networks have been established for several organisms and can be used to make predictions (Legrain et al. 2001).

This information can be gathered and organized using Web-based tools such as the freely accessible Cluster of Orthologous Groups database (Tatusov et al. 2001) and the proprietary ERGO database (Overbeek 2003). Though the comparative genomics field is still young, these tools have allowed the genetic characterization of a number of critical metabolic pathways that had eluded scientific inquiry for decades (Osterman and Overbeek 2003). For example, predictions based exclusively on occurrence profiling resulted in the identification of the last steps of the non-mevalonate isoprenoid pathway (Smit and Mushegian 2000). Protein fusion analysis allowed the identification of missing coenzyme A biosynthesis genes in *Homo sapiens* (Daugherty et al. 2002). Chromosome clustering analysis revealed a missing fatty acid synthesis gene (a target of antibacterial compounds) in *Streptococcus pneumoniae* (Heath and Rock 2000). A search for regulator sites allowed the identification of many missing thiamine biosynthesis genes (Rodionov et al. 2002). These methods can be combined as described below to find both genes encoding simple tRNA modification enzymes or whole new pathways involved in the synthesis of the more complex modifications.

3 Finding Genes for Simple tRNA Modifications

3.1 Paralog- and Ortholog-Based Identifications

Methylation and pseudourylation are the most common and abundant modifications in tRNAs. The first genes were discovered more than 10 years ago (Nurse et al. 1995; Ny and Björk 1980) and most of the work of the last decade has been sorting out the exact catalytic functions of the paralogs of these genes identified by blast searches. These sequence homology based searches were very successful and identified most of the methylases or pseudouridine synthases involved in tRNA modification (Del Campo et al. 2001; Gustafsson et al. 1996; Motorin and Grosjean 1999). In several cases, however, the family has diverged too much to be identified by these methods. Ofengand and colleagues recently identified the TruD family that modifies position 13 on $tRNA_{Asp}$ using a traditional enzyme purification approach (Kaya and Ofengand 2003). In a similar fashion, Phizicky and colleagues identified the methylase that modifies m1G9 in *S. cerevisiae* by a "biochemical genomic approach", (Jackman et al. 2003). In both cases, these new families could not have been identified by homology searches.

The identification of tRNA methylases presents a second difficulty: the Cluster of Orthologous Group analysis (Tatusov et al. 2001) is not sensitive enough to differentiate orthologs and paralogs and cannot be efficiently used to predict functions. More than 16 genes are in the same methylase cluster COG0500 in *E.coli,* and these genes encode RNA, DNA and protein methylases. To differentiate between the different methylase subclasses more sensitive methods are needed, such as structure-based protein alignments (Anantharaman et al. 2002a, b) which discriminate between subfamilies**.** Another way to circumvent the problem of insensitivity in the COG analysis is to combine the COG analysis with genetic mapping information when available (bearing in mind that genetic mapping information can be erroneous). For example, the gene *trmG* involved in m^2A_{37} formation has been mapped between 56–61 min on the *E. coli* chromosome (Björk 1996). Orf *yfiF*, found at 58 min and annotated as an rRNA methylase (GOG0566) is an obvious candidate for *trmG*. In a similar fashion, the last methylase steps involved in the formation of $cmnm^5s^2U$, encoded by *trmC*, had been mapped to the 50-min region in *E. coli* (Hagervall and Björk 1984). Analysis of the region allowed for the identification of the *ycfK* gene, at position 52.59 that encodes a bifunctional protein combining a SAM-dependent methylase domain (COG0500) and another domain (COG0665) identified as Glycine/D-amino acid oxidase (deaminating) domain that could be the *mnmC* gene. This prediction has recently been confirmed experimentally (L. Droogmans, pers. comm.).

3.2 Comparative Genomics-Based Identifications

In cases where no genetic or biochemical information is available comparative genomics methods are valuable to identify the missing genes, as we have recently shown with the identification of the tRNA 5,6-dihydrouridine synthase (Dus) family (Bishop et al. 2002). 5,6-Dihydrouridine (D) is one of the most common and abundant modifications of tRNA (Sprinzl et al. 1998), and is also present in some 23S RNA (Kowalak et al. 1995), but both the genes and enzymes were unknown. As detailed in Fig. 3, by combining occurrence pro-

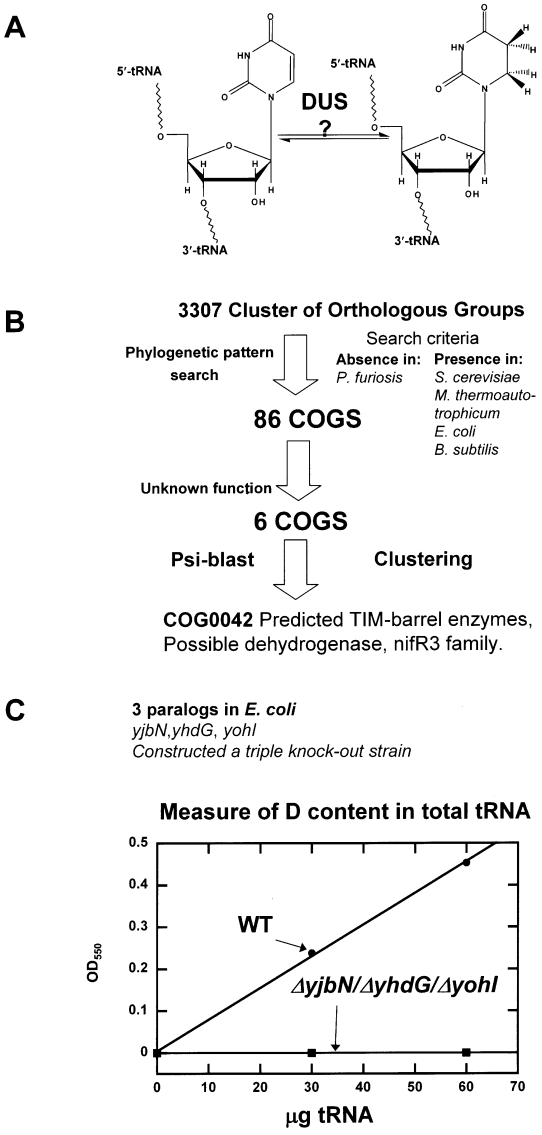


Fig. 3. Identification of the Dus family by comparative genomics. **A** tRNA modification catalyzed by the missing Dus enzyme. **B** The strategy followed to identify the candidates. Experimental validation was obtained by constructing an *E. coli* mutant deleted in all the genes of this family, and demonstrating that tRNA purified from this strain lacked any detectable D as shown in **C**

filing, chromosome clustering, and homology searches, the *dus* family of genes that contains orthologs in most sequenced species was identified.

McCloskey and coworkers observed an inverse correlation between the D content and the growth temperature of a given organism (Dalluge et al. 1997). Generally, thermophiles have little or no D and psychrophiles contain high amounts of D. These authors also showed that short oligonucleotides containing D favor the C2′-endo ribose conformation (compared with the equivalent U-containing oligonucleotide), while the C3′-endo conformation is necessary for base stacked RNA (Dalluge et al. 1996). Thus it was proposed that D confers local flexibility to tRNAs that is required at lower temperatures and detrimental at higher temperatures. However, there is no direct evidence for this theory. Available genomic information allowed us to plot the number of DUS paralogs in a given organism against its optimum growth temperature (Fig. 4). It is clear that while mesophilic organisms can have one to three genes, all thermophiles and hyperthermophiles have one or less. Access to the whole genome sequences of psychrophilic organisms such as *Methanogenium frigidum* and *Methanococcoides burtonii* (Saunders et al. 2003) will soon reveal if this trend is confirmed.

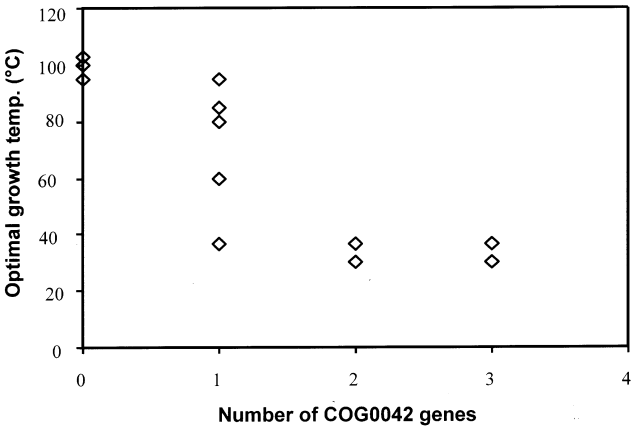


Fig. 4. Inverse correlation between the number of *dus* genes in a given genome and the growth temperature of the organism

4 Finding Complex Modification Pathway Genes

4.1 Finding Missing Steps in Known Pathways

For many complex tRNA modifications such as queuosine, ms^2i^6A or Wyeosine, several biosynthetic steps are needed. Few pathways have yet been totally characterized and fully reconstituted in vitro. In the case of the synthesis of $mnm^5s^2U_{34}$ (Fig. 5A), five enzymes have been characterized but at least one enzyme is still missing (Kambampati and Lauhon 2003). Gene clustering around the *gidA*, *mnmE*, and *mnmA* genes was analyzed. One candidate, *gidB,* clearly stood out: it is linked to *gidA* and *mnmE* in many genomes (Fig. 5B), and the structure of GidB has recently been determined and shown to have a methyltransferase fold (Romanowski et al. 2002). GidB is clearly a candidate for the missing step in the ms^2i^6A modification pathway and should be investigated further.

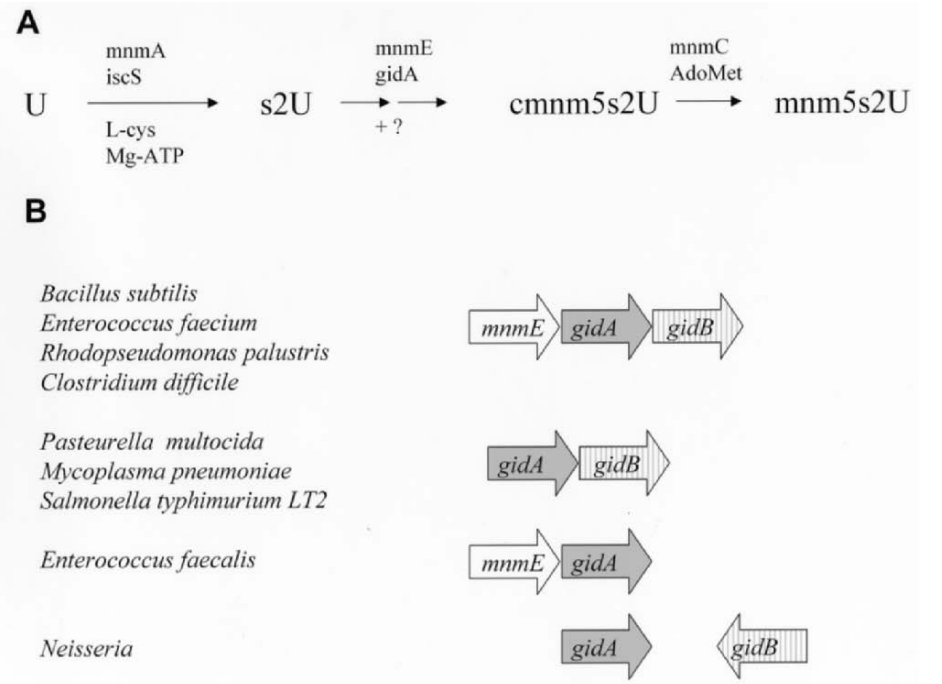


Fig. 5. **A** Biosynthesis pathway for $mnm^5s^2U_{34}$ (adapted with permission from Kambampati and Lauhon 2003). **B** Clustering examples of the *mnmE*, *gidA* and *gidB* genes

4.2 Finding Uncharacterized Pathway Genes

4.2.1 Identification of the PreQ Biosynthesis Pathway Genes

Some of the most complex modifications known to occur in tRNA are the 7-deazaguanosine nucleosides queuosine (Q) and archaeosine (gG), and the tricyclic wyosine (Yt) family of nucleosides (Fig. 6). Both queuosine and archaeosine share the unusual 7-deazaguanosine core, but differ in the extent of further elaboration of this core structure; queuosine is characterized by a cyclopentenediol ring appended to (7-aminomethyl)-7-deazaguanosine (Kasai et al. 1975a; Ohgi et al. 1979), which in some mammalian tRNAs is glycosylated with galactose or mannose at the C5″ hydroxyl (Okada and Nishimura 1977), while archaeosine possesses an amidine functional group at the 7-position (Gregson et al. 1993).

Queuosine and its derivatives occur exclusively at position 34 (the wobble position) in the anticodons of tRNAs coding for the amino acids asparagine, aspartic acid, histidine, and tyrosine (Frey et al. 1988). These tRNAs share two common nucleosides in their anticodon sequence GUN (positions 34–36), where N defines the identity of the codon and can be any nucleoside. Queuosine is ubiquitous throughout studied eukaryotic and bacterial phyla (with the exception of the tRNA of yeast and *Mycoplasma*), but is absent from the tRNA of the Archaea. In marked contrast, archaeosine is present only in the Archaea, where it is found in the majority of tRNA species, specifically at position 15 in the dihydrouridine loop (D-loop)(Sprinzl et al. 1989), a site not modified in any tRNA outside of the archaeal domain.

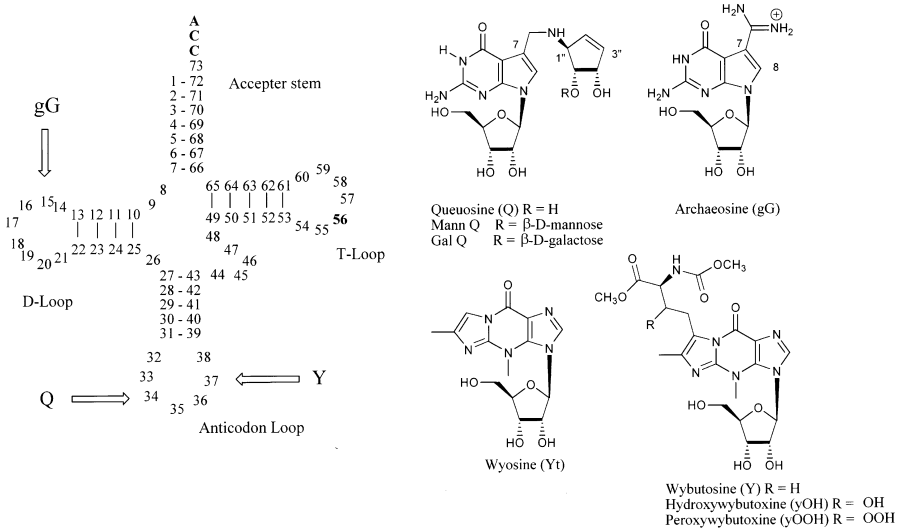


Fig. 6. Structures of hypermodified guanosines and positions modified in tRNA

The biosynthetic pathways of queuosine and archaeosine are partially characterized and summarized in Fig. 7. GTP is known to be the precursor in queuosine biosynthesis, and the first established intermediate in the pathway is 7-cyano-7-deazaguanine ($preQ_0$) (Okada et al. 1978), which presumably then undergoes reduction to 7-aminomethyl-7-deazaguanine ($preQ_1$) by an as yet uncharacterized dehydrogenase. $PreQ_1$ is subsequently inserted into the tRNA by the enzyme tRNA-guanine transglycosylase (TGT), a reaction in which the genetically encoded base (guanine) is eliminated (Okada et al. 1979; Okada et al. 1978). The remainder of queuosine biosynthesis occurs at the level of the tRNA, and involves the unprecedented utilization of *S*-adenosylmethionine (AdoMet) in the construction of an epoxycyclopentandiol ring (Kinzie et al. 2000; Slany et al. 1993; Slany et al. 1994) to give epoxyqueuosine (oQ), followed by an apparent B_{12}-dependent step in which the epoxide in oQ is reduced to give queuosine (Frey et al. 1988).

Although queuosine is ubiquitous in both the Eukarya and Bacteria, only Bacteria are capable of de novo queuosine biosynthesis. Eukaryotes acquire queuosine as a nutrient factor from the intestinal flora (Frey et al. 1988), and insert queuine, the free base of queuosine, directly into the appropriate tRNAs (Shindo-Okada et al. 1980) by a eukaryotic TGT.

The presence of a 7-substituted 7-deazaguanine core structure in both queuosine and archaeosine, along with the structural similarity of $preQ_0$ to archaeosine base, is consistent with identical biosynthetic pathways in Archaea and Bacteria for the formation of $preQ_0$. These pathways presumably diverge at $preQ_0$, with $preQ_0$ serving as the substrate for an archaeal TGT in the key base substitution reaction. Evidence in support of this scenario came with the isolation of both $preQ_0$ and an archaeal TGT from *Haloferax volcanii* (Watanabe et al. 1997), followed by the identification and cloning of a putative *tgt* gene from *M. jannaschii* (Bai et al. 2000), and the biochemical characterization of the recombinant enzyme as a TGT (Bai et al.

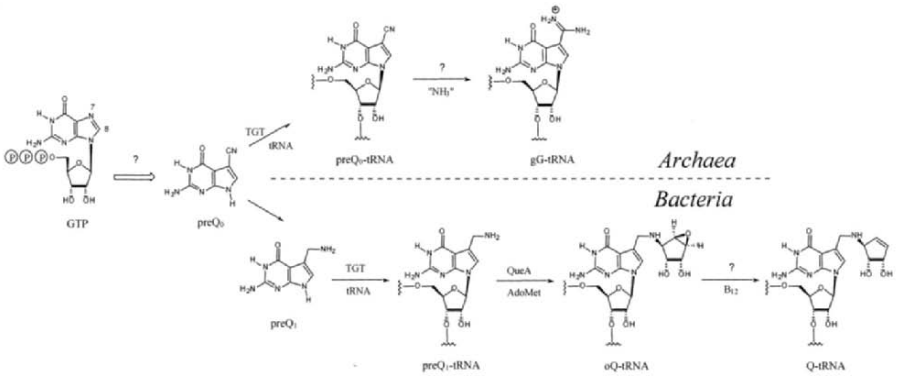


Fig. 7. The de novo biosynthesis of queuosine and archaeosine

2000). The formation of archaeosine can then in principle occur through the formal addition of ammonia to the nitrile of $preQ_0$ after incorporation into the polynucleotide.

As summarized in Fig. 7, at least three steps are missing in Q and gG biosynthetic pathways; the respective last steps and the steps leading from GTP to $preQ_0$. To identify the missing genes several types of genomic information were combined as shown in Fig. 8 and detailed below.

1. Biochemical information: because GTP is the precursor in queuosine biosynthesis, several authors have proposed that an uncharacterized GTP cyclohydrolase-like enzyme catalyzes the first step of the biosynthesis (Morris and Elliott 2001). A search of the COG database (Tatusov et al. 2001) using the "GTP cyclohydrolase" keywords identified the two known GTP cyclohydrolase families (FolE and RibA), but also identified the COG0780 family, annotated as "enzymes related to GTP cyclohydrolase I"(Fig. 8A).
2. Clustering data: when analyzing the neighboring regions of COG0780 family members in many organisms we found that, in *B. subtilis*, the COG0780 member *ykvM* was the last gene of the *ykvJKLM* operon (Fig. 8B). In 80 % of the totally sequenced organisms, different combinations containing two or three of these four genes were found in operonic structures.
3. Phylogenetic distribution: an occurrence profile was generated using the presence or absence of *tgt* as a marker for the occurrence of the Q biosyn-

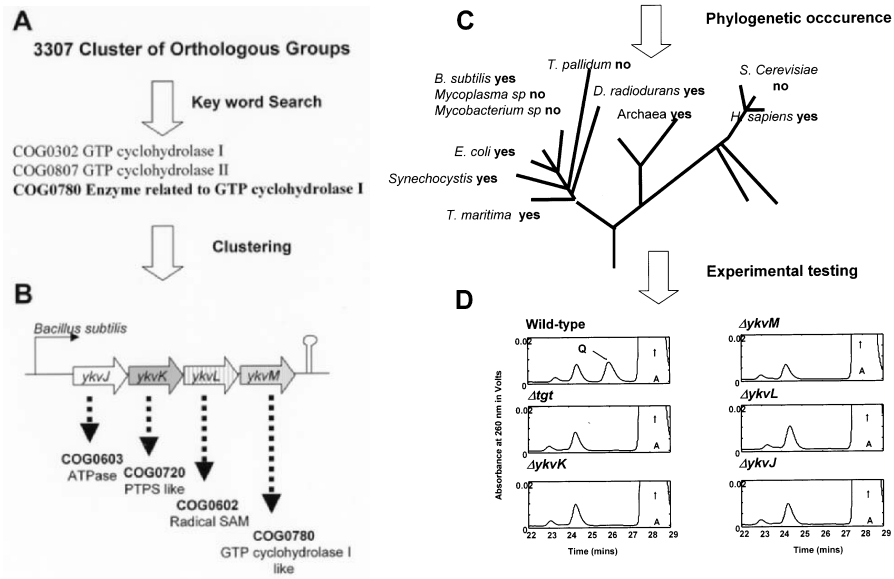


Fig. 8A–D. Identification of four new queuine biosynthesis genes by comparative genomics

thetic pathway in a given organism (Fig. 8C): *tgt* homologues are absent from *S. cerevisiae* and *Mycoplasma* as predicted from the literature (Andachi et al. 1989; Kasai et al. 1975b; Katze et al. 1982). Unexpectedly, *tgt* homologues are also absent from most mycobacterial sp. and *Treponema pallidum*, suggesting that Q is absent from these species as well. The four genes *ykvJKLM* are members of COGs that all followed the required occurrence criteria (absence in *S. cerevisiae*, *Mycoplasma* and *Mycobacterium*).

The combination of phylogenetic occurrence, clustering and biochemical data led to the hypothesis that the four enzyme families encoded by the *ykvJKLM* operon in *B. subtilis* are involved in Q synthesis. The hypothesis was tested by constructing four Acinetobacter ADP1 mutants deleted in the corresponding genes. In all cases, HPLC analysis of digests of bulk tRNA prepared from these strains shows the disappearance of the Q peak present in the WT strain (Fig. 8D).

By combining genomics approaches with genetics, we were able to identify four new queuosine genes in a short space of time (Reader et al., in press). We are currently testing the hypothesis that these genes encode the biosynthetic enzymes for $preQ_0$ /$preQ_1$.

4.2.2 Hunting for the Wyeosine Biosynthesis Genes

Little is known about the biosynthesis of the wyosine family (Fig. 9). Most of the relevant studies have been performed in *S. cerevisiae*, where it has been demonstrated that wyosine originates from the genetically encoded guanine (Blobstein et al. 1973), and the first step is N1 methylation by the m^1G methylase (Droogmans and Grosjean 1987), an AdoMet-dependent methylase encoded by the gene YHR070w (Trm5). Notably this methylase is non-specific, being responsible for N1-methylation at G37 in tRNAs that code for Leu, His, Asp, Trp, and Pro (Björk et al. 2001).

Two strategies can be followed to identify genes in the wyosine pathway. The first is clustering analysis using the methylase as the starting gene. Clustering is not as strong in Eukarya as in Bacteria or Archaea, but as the first genes of the pathway should be present in Archaea (involved in m^1G and mimG formation), this is still a useful approach. The second strategy is to use the observation that *Drosophila melanogaster* $tRNA^{Phe}$ does not have Y but m^1G at position 37 (Sprinzl et al. 1998).

Using an ERGO macro (Overbeek 2003), a phylogenetic occurrence query was performed to identify genes that are present in *Homo sapiens*, *S. cerevisiae, S. pombe* and *M. janaschii*, but absent in *Drosophila melanogaster*, *E. coli* and *B.subtilis.* Only one family followed this distribution (COG0731). Members of this family are found in all sequenced Archaea and in all sequenced Eukarya except *D. melanogaster* and *Anopheles gambiae*. It has been annotated as a Fe-S oxidoreductase and is a member of the SAM radical

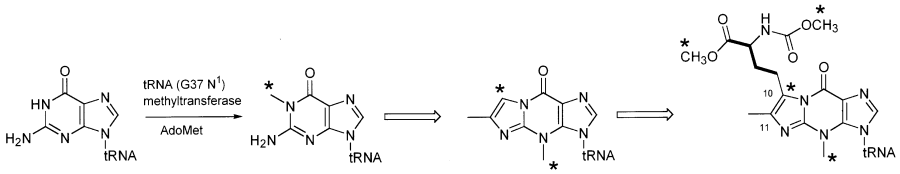


Fig. 9. The biosynthesis of wybutosine in *S. cerevisiae*. *Asterisks* denote methyl groups from AdoMet, and *heavy bonds* denote the origin of these carbons from the 3-amino-3-carboxylpropyl group of methionine

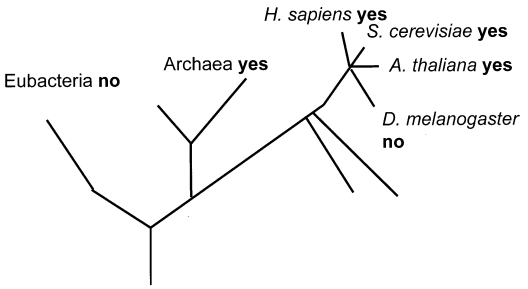


Fig. 10. Phylogenetic occurrence profile for the wyosine family of modified nucleosides

superfamily (Sofia et al. 2001) and is a plausible candidate for an enzyme involved in the formation of the tricyclic ring.

In the case of *S. cerevisiae*, a number of post- genomic tools such as protein interaction data, mRNA expression data or gene deletion data are available in integrated databases such as SGD (http://www.yeastgenome.org/). Analysis of the data available on the yeast COG0731 member YPL207w, revealed that a deletion mutant has been constructed and is viable (Giaever et al. 2002). Analysis of the genes that are co-expressed with YPL207w during the cell cycle (Cho et al. 1998) or in response to DNA damaging agents (Gasch et al. 2001) shows strong linkage (P>10–5) with ribosome biogenesis and RNA processing/ RNA metabolism genes. Experiments are underway to test if indeed, if YPL207w is a wyeosine biosynthesis gene (Fig. 10).

5 Conclusions

As we have shown in several cases comparative genomics approaches are well suited in identification of the missing tRNA modification genes. Compared with other biosynthesis clusters in which major pathway genes were missing such as the coenzyme (NAD, CoA, FAD), the use of the comparative genomics approach has given us a closing of the number of unknowns in a very short time (Gerdes et al. 2002). We can anticipate that before this review goes to

press it will already be obsolete as more genes are found. However, the type of approaches that were described in this review can be applied to any pathway and should be an integral part of the experimental biologist tools when tackling a biological problem.

Acknowledgements. I am grateful to Henri Grosjean for his constant encouragement and help in getting started in the tRNA modification field and for his careful revision of this manuscript. I thank Dirk Iwata-Reuyl for significant contributions to the manuscript and for providing several figures. I thank Glenn Björk, Louis Droogmans and Tsutomu Suzuki for unpublished information and Charles Lauhon for suggestions. This work would not have been possible without the support of Paul Schimmel.

References

Altman S, Kirsebom L, Talbot S (1995) Recent studies of RNaseP. In: RajBhandary UL (ed) tRNA: structure, biosynthesis, and function. ASM Press, Washington,DC, pp 67–78

Anantharaman V, Koonin EV, Aravind L (2002a) Comparative genomics and evolution of proteins involved in RNA metabolism. Nucleic Acids Res 30:1427–1464

Anantharaman V, Koonin EV, Aravind L (2002b) SPOUT: a class of methyltransferases that includes *spoU* and *trmD* RNA methylase superfamilies, and novel superfamilies of predicted prokaryotic RNA methylases. J Mol Microbiol Biotechnol 4:71–75

Andachi Y, Yamao F, Muto A, Osawa S (1989) Codon recognition patterns as deduced from sequences of the complete set of transfer RNA species in *Mycoplasma capricolum*. Resemblance to mitochondria. J Mol Biol 209:37–54

Bai Y, Fox DT, Lacy JA, Van Lanen SG, Iwata-Reuyl D (2000) Hypermodification of tRNA in *Thermophilic archaea*. Cloning, overexpression, and characterization of tRNA-guanine transglycosylase from *Methanococcus jannaschii*. J Biol Chem 275:28731–28738

Bishop AC, Xu J, Johnson RC, Schimmel P, de Crécy-Lagard V (2002) Identification of the tRNA-dihydrouridine synthase family. J Biol Chem 277:25090–25095

Björk G R (1980) A novel link between the biosynthesis of aromatic amino acids and transfer RNA modification in *Escherichia coli*. J Mol Biol 140:391–410

Björk GR (1992) The role of modified nucleosides in tRNA interactions. In: Pirtle RM (ed) Transfer RNA in protein ynthesis. CRC Press, Boca Raton, pp 23–85

Björk GR (1995) Biosynthesis and function of modified nucleosides. In: RajBhandary UL(ed) tRNA: structure, biosynthesis, and function. ASM Press, Washington, DC, pp 165–206

Björk GR (1996) Stable RNA modification. In: Neidhart FC(ed) *Escherichia coli and Salmonella*. cellular and molecular biology. ASM Press, Washington, DC,pp 861–886

Björk GR, Jacobsson K, Nilsson K, Johansson MJ, Bystrom AS, Persson OP (2001) A primordial tRNA modification required for the evolution of life? Embo J 20:231–239

Björk GR, Kjellin-Straby K (1978) *Escherichia coli* mutants with defects in the biosynthesis of 5-methylaminomethyl-2-thio-uridine or 1-methylguanosine in their tRNA. J Bacteriol 133:508–517

Björk GR, Kohli J (1990) Synthesis and function of modified nucleosides in tRNA. In: Kuo K(ed) Chromatography and modification of nucleosides. Part b biological roles and function of modification. Elsevier, Amsterdam, pp B13-B67

Blattner FR, Plunkett G 3rd, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, et al. (1997) The complete genome sequence of *Escherichia coli* K-12. Science 277:1453–1474

Blobstein SH, Grunberger D, Weinstein IB, Nakanishi K (1973) Isolation and structure determination of the fluorescent base from bovine liver phenylalanine transfer ribonucleic acid. Biochemistry 12:188–193

Bregeon D, Colot V, Radman M, Taddei F (2001) Translational misreading: a tRNA modification counteracts a +2 ribosomal frameshift. Genes Dev 15:2295–2306

Bystrom AS, Björk GR (1982) Chromosomal location and cloning of the gene (*trmD*) responsible for the synthesis of tRNA (m1G) methyltransferase in *Escherichia coli* K-12. Mol Gen Genet 188:440–446

Cabedo H, Macian F, Villarroya M, Escudero JC, Martinez-Vicente M, Knecht E, Armengod ME (1999) The *Escherichia coli trmE* (*mnmE*) gene, involved in tRNA modification, codes for an evolutionarily conserved GTPase with unusual biochemical properties. EMBO J 18:7063–7076

Caillet J, Droogmans L (1988) Molecular cloning of the *Escherichia coli miaA* gene involved in the formation of delta 2-isopentenyl adenosine in tRNA. J Bacteriol 170:4147–4152

Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, Wodicka L, Wolfsberg TG, Gabrielian AE, Landsman D, Lockhart DJ, Davis RW (1998) A genome-wide transcriptional analysis of the mitotic cell cycle. Mol Cell 2:65–73

Cordwell SJ (1999) Microbial genomes and "missing" enzymes: redefining biochemical pathways. Arch Microbiol 172:269–279

Dalluge JJ, Hamamoto T, Horikoshi K, Morita RY, Stetter KO, McCloskey JA (1997) Post-transcriptional modification of tRNA in psychrophilic bacteria. J Bacteriol 179:1918–1923

Dalluge JJ, Hashizume T, McCloskey JA (1996) Quantitative measurement of dihydrouridine in RNA using isotope dilution liquid chromatography-mass spectrometry (LC/MS). Nucleic Acids Res 24:3242–3245

Daugherty M, Polanuyer B, Farrell M, Scholle M, Lykidis A, de Crécy-Lagard V, Osterman A (2002) Complete reconstitution of the human coenzyme A biosynthetic pathway via comparative genomics. J Biol Chem 277:21431–21439

De Bie LG, Roovers M, Oudjama Y, Wattiez R, Tricot C, Stalon V, Droogmans L, Bujnicki JM (2003) The *yggH* gene of *Escherichia coli* encodes a tRNA (m7G46) methyltransferase. J Bacteriol 185:3238–3243

Del Campo M, Kaya Y, Ofengand J (2001) Identification and site of action of the remaining four putative pseudouridine synthases in *Escherichia coli*. Rna 7:1603–1615

Derrick WB, Horowitz J (1993) Probing structural differences between native and in vitro transcribed *Escherichia coli* valine transfer RNA: evidence for stable base modification-dependent conformers. Nucleic Acids Res 21:4948–4953

Deutscher MP (1995) tRNA processing nucleases. In: RajBhandary UL (ed) tRNA: structure, biosynthesis, and function. ASM Press, Washington, DC, pp 51–66

Droogmans L, Grosjean H (1987) Enzymatic conversion of guanosine 3¢ adjacent to the anticodon of yeast tRNAPhe to N1-methylguanosine and the wye nucleoside: dependence on the anticodon sequence. EMBO J 6:477–483

Eastwood Leung H-C, G H T, Björk G R, Winkler M E (1998). Genetic locations and database accession numbers of RNA-modifying and -editing enzymes. In: Benne R(ed) modification and editing of RNA. ASM Press, Washington, DC, pp 561–568

Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA (1999) Protein interaction maps for complete genomes based on gene fusion events. Nature 402:86–90

Esberg B, Leung HC, Tsui HC, Björk GR, Winkler M E (1999) Identification of the *miaB* gene, involved in methylthiolation of isopentenylated A37 derivatives in the tRNA of *Salmonella typhimurium* and *Escherichia coli*. J Bacteriol 181:7256–7265

Frey B, McCloskey JA, Kersten W, Kersten H (1988) New function of vitamin B_{12}: cobamide-dependent reduction of Epoxyqueuosine to Queuosine in tRNAs of *Escherichia coli* and *Salmonella typhimurium*. J Bacteriol 170:2078–2082

Gasch AP, Huang M, Metzner S, Botstein D, Elledge SJ, Brown PO (2001) Genomic expression responses to DNA-damaging agents and the regulatory role of the yeast ATR homolog Mec1p. Mol Biol Cell 12:2987–3003

Gelfand MS, Novichkov PS, Novichkova ES, Mironov AA (2000) Comparative analysis of regulatory patterns in bacterial genomes. Brief Bioinform 1:357–371

Gerdes SY, Scholle MD, D'Souza M, Bernal A, Baev MV, Farrell M, Kurnasov OV, Daugherty MD, Mseeh F, Polanuyer, BM et al. (2002) From genetic footprinting to antimicrobial drug targets: examples in cofactor biosynthetic pathways. J Bacteriol 184:4555–4572

Giaever G, Chu AM, Ni L, Connelly C, Riles L, Veronneau S, Dow S, Lucau-Danila A, Anderson K, Andre B et al. (2002) Functional profiling of the *Saccharomyces cerevisiae* genome. Nature 418:387–391

Green SM, Malik T, Giles IG, Drabble WT (1996) The *purB* gene of *Escherichia coli* K-12 is located in an operon. Microbiology 142:3219–3230

Gregson JM, Crain PF, Edmonds CG, Gupta R, Hashizume T, Phillipson DW, McCloskey JA (1993) Structure of Archaeal transfer RNA nucleoside GG-15 (2-Amino-4,7-dihydro-4-oxo-7-β-D-ribofuranosyl-1*H*-pyrrolo[2,3-*d*]pyrimidine-5-carboximidamide (Archaeosine)). J Biol Chem 268:10076–10086

Grosjean H, Auxilien S, Constantinesco F, Simon C, Corda Y, Becker H F, Foiret D, Morin A, Jin YX, Fournier M, Fourrey JL (1996) Enzymatic conversion of adenosine to inosine and to N1-methylinosine in transfer RNAs: a review. Biochimie 78:488–501

Grosjean H, Constantinesco F, Foiret D, Benachenhou N (1995) A novel enzymatic pathway leading to 1-methylinosine modification in *Haloferax volcanii* tRNA. Nucleic Acids Res 23:4312–4319

Gustafsson C, Reid R, Greene PJ, Santi DV (1996) Identification of new RNA modifying enzymes by iterative genome search using known modifying enzymes as probes. Nucleic Acids Res 24:3756–3762

Gutgsell N, Englund N, Niu L, Kaya Y, Lane BG, Ofengand J (2000) Deletion of the *Escherichia coli* pseudouridine synthase gene *truB* blocks formation of pseudouridine 55 in tRNA in vivo, does not affect exponential growth, but confers a strong selective disadvantage in competition with wild-type cells. RNA 6:1870–1881

Hagervall TG, Björk GR (1984) Genetic mapping and cloning of the gene (*trmC*) responsible for the synthesis of tRNA (mnm5s2U)methyltransferase in *Escherichia coli* K12. Mol Gen Genet 196:201–207

Hagervall TG, Jonsson YH, Edmonds CG, McCloskey JA, Björk GR (1990) Chorismic acid, a key metabolite in modification of tRNA. J Bacteriol 172:252–259

Heath RJ, Rock CO (2000) A triclosan-resistant bacterial enzyme. Nature 406:145–146

Hopper AK, Phizicky EM (2003) tRNA transfers to the limelight. Genes Dev 17:162–180

Horie N, Hara-Yokoyama M, Yokoyama S, Watanabe K, Kuchino Y, Nishimura S, Miyazawa T (1985) Two tRNAIle1 species from an extreme thermophile, *Thermus thermophilus* HB8: effect of 2-thiolation of ribothymidine on the thermostability of tRNA. Biochemistry 24:5711–5715

Huynen MA, Snel B, Mering C, Bork P (2003) Function prediction and protein networks. Curr Opin Cell Biol 15:191–198

Jackman JE, Montange RK, Malik HS, Phizicky EM (2003) Identification of the yeast gene encoding the tRNA m1G methyltransferase responsible for modification at position 9. RNA 9:574–585

Kambampati R, Lauhon C T (2003) MnmA and IscS are required for in vitro 2-thiouridine biosynthesis in *Escherichia coli*. Biochemistry 42:1109–1117

Kammen, HO, Marvel CC, Hardy L, Penhoet EE (1988) Purification, structure, and properties of *Escherichia coli* tRNA pseudouridine synthase I. J Biol Chem 263:2255–2263

Kasai H, Kuchino Y, Nihei K, Nishimura S (1975a) Distribution of the modified nucleoside Q and its derivatives in animal and plant transfer RNA's. Nucleic Acids Res 2:1931–1939

Kasai H, Ohashi Z, Harada F, Nishimura S, Oppenheimer NJ, Crain, PF, Liehr JG, von Minden DL, McCloskey JA (1975b) Structure of the modified nucleoside Q isolated from *Escherichia coli* transfer ribonucleic acid. 7-(4,5-*cis*-dihydroxy-1-cyclopenten-3-ylaminomethyl)-7-deazaguanosine. Biochemistry 14:4198–4208

Katze JR, Basile B, McCloskey JA (1982) Queuine, a modified base incorporated post-transcriptionally into eukaryotic transfer RNA: wide distribution in nature. Science 216:55–56

Kaya Y, Ofengand J (2003) A novel unanticipated type of pseudouridine synthase with homologs in Bacteria, Archaea, and Eukarya. RNA 9:711–721

Kinzie SD, Thern B, Iwata-Reuyl D (2000) Mechanistic studies of the tRNA-modifying enzyme QueA: a chemical imperative for the use of AdoMet as a "ribosyl" donor. Org Lett 2:1307–1310

Kowalak JA, Dalluge JJ, McCloskey JA, Stetter KO (1994) The role of posttranscriptional modification in stabilization of transfer RNA from hyperthermophiles. Biochemistry 33:7869–7876

Kowalak JA, Bruenger E, McCloskey JA (1995) Posttranscriptional modification of the central loop of domain V in *Escherichia coli* 23 S ribosomal RNA. J Biol Chem 270:17758–17764

Kunst F, Ogasawara N, Moszer I, Albertini AM, Alloni G, Azevedo V, Bertero MG, Bessieres P, Bolotin A, Borchert S, et al. (1997) The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. Nature 390:249–256

Lauhon C T (2002) Requirement for *IscS* in biosynthesis of all thionucleosides in *Escherichia coli*. J Bacteriol 184:6820–6829

Lauhon CT, Kambampati R (2000) The *iscS* gene in *Escherichia coli* is required for the biosynthesis of 4-thiouridine, thiamin, and NAD. J Biol Chem 275:20096–20103

Legrain P, Wojcik J, Gauthier J M (2001) Protein–protein interaction maps: a lead towards cellular functions. Trends Genet 17:346–352

Leinfelder W, Forchhammer K, Veprek B, Zehelein E, Bock A (1990) In vitro synthesis of selenocysteinyl-tRNA(UCA) from seryl-tRNA(UCA): involvement and characterization of the *selD* gene product. Proc Natl Acad Sci USA 87:543–547

Mihara H, Kato S, Lacourciere GM, Stadtman TC, Kennedy RA, Kurihara T, Tokumoto U, Takahashi Y, Esaki N (2002) The *iscS* gene is essential for the biosynthesis of 2-selenouridine in tRNA and the selenocysteine-containing formate dehydrogenase H. Proc Natl Acad Sci USA 99:6679–6683

Mittl PR, Grutter MG (2001) Structural genomics: opportunities and challenges. Curr Opin Chem Biol 5:402–408

Morris RC, Elliott MS (2001) Queuosine modification of tRNA: a case for convergent evolution. Mol Genet Metab 74:147–159

Motorin Y, Grosjean H (1999) Multisite-specific tRNA:m5C-methyltransferase (Trm4) in yeast *Saccharomyces cerevisiae*: identification of the gene and substrate specificity of the enzyme. RNA 5:1105-1118

Motorin Y, Grosjean H (2001) tRNA modification. Encyclopedia of life sciences. Nature publishing group/www.els.net

Mueller EG, Buck CJ, Palenchar PM, Barnhart LE, Paulson JL (1998) Identification of a gene involved in the generation of 4-thiouridine in tRNA. Nucleic Acids Res 26:2606–2610

Munch HJ, Thiebe R (1975) Biosynthesis of the nucleoside Y in yeast tRNAPhe: incorporation of the 3-amino-3-carboxypropyl-group from methionine. FEBS Lett 51:257–258

Muramatsu T, Nishikawa K, Nemoto F, Kuchino Y, Nishimura S, Miyazawa T, Yokoyama S (1988) Codon and amino-acid specificities of a transfer RNA are both converted by a single post-transcriptional modification. Nature 336:179–181

Nilsson K, Lundgren HK, Hagervall TG, Björk G R (2002) The cysteine desulfurase IscS is required for synthesis of all five thiolated nucleosides present in tRNA from *Salmonella enterica* serovar *typhimurium*. J Bacteriol 184:6830–6835

Nurse K, Wrzesinski J, Bakin A, Lane BG, Ofengand J (1995) Purification, cloning, and properties of the tRNA psi 55 synthase from *Escherichia coli*. RNA 1:102–112

Ny T, Björk GR (1980) Cloning and restriction mapping of the *trmA* gene coding for transfer ribonucleic acid (5-methyluridine)-methyltransferase in *Escherichia coli* K-12. J Bacteriol 142:371–379

Ohgi T, Kondo T, Goto T (1979) Total Synthesis of Optically Pure Nucleoside Q. Determination of Absolute Configuration of natural Nucleoside Q. J Am Chem Soc 101:3629–3633

Okada N, Nishimura S (1977) Enzymatic Synthesis of Q* Nucleoside Containing Mannose in the Anticodon of tRNA: Isolation of a Novel Mannosyltransferase from a Cell-Free Extract of Rat Liver. Nucleic Acids Res 4:2931–2937

Okada N, Noguchi S, Nishimura S, Ohgi T, Goto T, Crain PF, McCloskey JA (1978) Structure determination of a nucleoside Q precursor isolated from *E. coli* tRNA: 7-(aminomethyl)-7-deazaguanosine. Nucleic Acids Res 5:2289–2296

Okada N, Noguchi S, Kasai H, Shindo-Okada N, Ohgi T, Goto T, Nishimura S (1979) Novel Mechanism of Post-transcriptional Modification of tRNA. J Biol Chem 254:3067–3073

Osterman A, Overbeek R (2003) Missing genes in metabolic pathways: a comparative genomics approach. Curr Opin Chem Biol 7:238–251

Overbeek R, et al. (2003) The ERGO Genome Analysis and Discovery System. Nucleic Acids Res 31:1–8

Overbeek R, Fonstein M, D'Souza M, Pusch G D, Maltsev N (1999) The use of gene clusters to infer functional coupling. Proc Natl Acad Sci USA 96:2896–2901

Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci USA 96:4285–4288

Perret V, Garcia A, Puglisi J, Grosjean H, Ebel JP, Florentz C, Giege R (1990) Conformation in solution of yeast tRNA(Asp) transcripts deprived of modified nucleotides. Biochimie 72:735–743

Persson BC (1993) Modification of tRNA as a regulatory device. Mol Microbiol 8:1011–1016

Persson BC, Gustafsson C, Berg DE, Björk GR (1992) The gene for a tRNA modifying enzyme, m5U54-methyltransferase, is essential for viability in *Escherichia coli*. Proc Natl Acad Sci USA 89:3995–3998

Persson BC, Jager G, Gustafsson C (1997) The *spoU* gene of *Escherichia coli*, the fourth gene of the *spoT* operon, is essential for tRNA (Gm18) 2¢-O-methyltransferase activity. Nucleic Acids Res 25:4093–4097

Raychaudhuri S, Niu L, Conrad J, Lane BG, Ofengand J (1999) Functional effect of deletion and mutation of the *Escherichia coli* ribosomal RNA and tRNA pseudouridine synthase RluA. J Biol Chem 274:18880–18886

Reader JS, Metzgar D, Schimmel P, de Crézy-Lagard V (2004) Identification of four genes necessary for biosynthesis of the modified nucleoside queuosine. J Biol Chem (in press)

Reuter K, Slany R, Ullrich F, Kersten H (1991) Structure and organization of *E. coli* genes involved in biosynthesis of the Deazaguanine Derivative Queuine, a nutrient factor for Eukaryotes. J Bacteriol 173:2256–2264

Rodionov DA, Vitreschak AG, Mironov AA, Gelfand M S (2002) Comparative genomics of thiamin biosynthesis in procaryotes: new genes and regulatory mechanisms. J Biol Chem 277:48949–48959

Romanowski MJ, Bonanno JB, Burley SK (2002) Crystal structure of the *Escherichia coli* glucose-inhibited division protein B (GidB) reveals a methyltransferase fold. Proteins 47:563–567

Saunders NF, Thomas T, Curmi PM, Mattick JS, Kuczek E, Slade R, Davis J, Franzmann PD, Boone D, Rusterholtz K et al. (2003) Mechanisms of thermal adaptation revealed from the genomes of the Antarctic Archaea *Methanogenium frigidum* and *Methanococcoides burtonii*. Genome Res 12:12

Shindo-Okada N, Okada N, Ohgi T, Goto T, Nishimura S (1980) Transfer Ribonucleic Acid Guanine Transglycosylase isolated from rat liver. Biochemistry 19:395–400

Slany RK, Bosl M, Crain PF, Kersten H (1993) A new function of S-Adenosylmethionine: The ribosyl moiety of AdoMet is the precursor of the Cyclopentenediol moiety of the tRNA Wobble Base Queuine. Biochemistry 32:7811–7817

Slany RK, Bosl M, Kersten H (1994) Transfer and isomerization of the ribose moiety of AdoMet during the biosynthesis of queuosine tRNAs, a new unique reaction catalyzed by the QueA protein from *Escherichia coli*. Biochimie 76:389–393

Smit A, Mushegian A (2000) Biosynthesis of isoprenoids via mevalonate in Archaea: the lost pathway. Genome Res 10:1468–1484

Smith C, Schmidt PG, Petsch J, Agris PF (1985) Nuclear magnetic resonance signal assignments of purified [13C]methyl-enriched yeast phenylalanine transfer ribonucleic acid. Biochemistry 24:1434–1440

Soma A, Ikeuchi Y, Kanemasa S, Koyabashi K, Ogawasara N, Ote T, Kato J, Watanabe K, Sekine Y, Suzuki T (2003) An RNA-modifying enzyme that governs both the codon and amino acid specificities of isoleucine tRNA. Mol Cell 12:689–98

Sofia HJ, Chen G, Hetzler BG, Reyes-Spindola JF, Miller NE (2001) Radical SAM, a novel protein superfamily linking unresolved steps in familiar biosynthetic pathways with radical mechanisms: functional characterization using new analysis and information visualization methods. Nucleic Acids Res 29:1097–1106

Sprinzl M, Hartmann T, Weber J, Blank J, Zeidler R (1989) Compilation of tRNA sequences and sequences of tRNA genes. Nucleic AcidsRes 17:r1-r67

Sprinzl M, Horn C, Brown M, Ioudovitch A, Steinberg S (1998) Compilation of tRNA sequences and sequences of tRNA genes. Nucleic Aids Res 26:148–153

Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND Koonin E V (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res 29:22–28

Watanabe M, Matsuo M, Tanaka S, Akimoto H, Asahi S, Nishimura S, Katz JR, Hashizume T, Crain PF, McCloskey JA, Okada N (1997) Biosynthesis of Archaeosine, a Novel Derivative of 7-Deazaguanosine Specific to Archaeal tRNA, Proceeds via a Pathway Involving Base Replacement of the tRNA Polynucleotide Chain. J Biol Chem 272: 20146–20151

Westaway SK, Abelson J (1995) Splicing of tRNA Precursors. In: RajBhandary UL (ed) tRNA: structure, biosynthesis, and function. ASM Press, Washington, DC, pp79–92

Wolf J, Gerber AP, Keller W (2002) *tadA*, an essential tRNA-specific adenosine deaminase from *Escherichia coli*. EMBO J 21:3841–3851

Yokoyama S, Nishimura S (1995) Modified nucleosides and codon recognition. In: RajBhandary UL, Söll D (eds) tRNA: Structure, biosynthesis, and function. ASM Press, Washington, DC, pp 207–233

Evolution and Function of Processosome, the Complex that Assembles Ribosomes in Eukaryotes: Clues from Comparative Sequence Analysis

A. Mushegian

1 Introduction

An assembly of functioning ribosomes starts with the biosynthesis of ribosomal RNAs and proteins. In all living species, polycistronic pre-ribosomal RNA (pre-rRNA) is processed to mature rRNAs and is covalently modified at multiple positions, with the aid of specific protein enzymes and small nucleolar guide RNAs (snoRNAs). There are more than 40 known types of covalent rRNA modifications, the two most common ones being pseudouridylation and methylation (Crain et al. 2003). Ribosomal proteins, some of which are also covalently modified, are then assembled into mature ribosome subunits with rRNA. In eukaryotic cells, cytoplasmically synthesized ribosomal proteins have to be imported into the nucleus, and the assembled ribosomes are exported from the nucleus back into the cytoplasm.

Until very recently, the molecular components of the apparatus for ribosome assembly in eukaryotes were not well uncharacterized. Lately, the high-throughput proteomic analysis of fractionated yeast cells resulted in the determination of the protein composition of several complexes involved in various stages of ribosome assembly (Dragon et al. 2002; Grandi et al. 2002; Nissan et al. 2002; Schafer et al. 2003; reviewed in: Fatica and Tollervey 2002). The multistage process of ribosome maturation and export is associated with the modification of these complexes through the addition and removal of specific proteins; altogether, there are more than 80 such protein components in yeast, specific to at least some stages of the ribosome maturation and export pathway. This protein set is collectively referred to as processosome.

A. Mushegian
Stowers Institute for Medical Research, 1000 E 50th St., Kansas City, Missouri 64110, USA

Nucleic Acids and Molecular Biology, Vol. 15
Janusz M. Bujnicki (Ed.)
Practical Bioinformatics

The knowledge about the composition of the processosome complex is now guiding the molecular dissection of protein–protein and protein-RNA interactions within the complex (Fatica et al. 2002, 2003; Oeffinger et al. 2002; Wehner et al. 2002; Gadal et al. 2002; Granneman et al. 2003). In addition, sequence database searches detected some sequences highly similar to the processosome components, predicting specific biochemical activities for many of them (e.g., Fatica et al. 2003). One goal of this chapter is to take the next step in that analysis, namely, to detect additional sequence signals that may be indicative of protein function. I will focus on the newly discovered (putative) components of processosome, as they are defined in Fatica and Tollervey (2002; see their Fig. 2 for the details), leaving out the analysis of better-studied and typically better-conserved RNA processing enzymes themselves.

Processosome composition is also interesting for evolutionary reasons. Ribosomes are essential organelles shared by all three domains of life, Bacteria, Archaea, and Eukarya. Despite the prokaryotic cellular organization of both Bacteria and Archaea, the ribosomal proteins and other factors involved in mRNA translation set Bacteria and Archaea apart, and unite Archaea with Eukarya, in the two following senses: first, whenever the orthologous proteins exist in all three domains, the evolutionary distance between archaeal and eukaryal proteins is much closer than between archaeal and bacterial orthologs (Koonin et al. 1997); secondly, when orthologous proteins are found only in two domains out of three, they are typically Archaea and Eukarya, to the exclusion of bacteria (Anantharaman et al. 2002). Similar trends are observed with the proteins involved in many other aspects of RNA metabolism (Koonin et al. 2001; Anantharaman et al. 2002). In the case of the processosome components, one could therefore expect substantial similarity between yeasts (and other eukaryotes) and archaea. Yet, the complex organization of eukaryotic karyoplasm and nucleocytoplasmic transport requires eukaryote-specific adaptations. Separation of ancestral from eukaryote-specific components of processosome would therefore provide some clues to the evolution of multiprotein complexes associated with the emergence of cellular nucleus.

2 Sequence Analysis of the Processosome Components

I extracted sequences of yeast processome components listed in Fatica and Tollervey (2002) from GenBank. For comparative purposes, five other protein sets were prepared: (1) complete yeast proteome, as distributed by NCBI (http://www.ncbi.nlm.nih.gov). (2) Proteins localized throughout the nucleus, as determined by the high-throughput tagging project at Yale University (http://ygac.med.yale.edu/triples/basic_search.asp). (3) Proteins with nuclear localization, determined by various computational and in vivo approaches,

from the MIPS Yeast database (http://mips.gsf.de/proj/yeast/CYGD/db/index.html). (4) Structural components of nuclear pore, as annotated in the first dataset. (5) Yeast cytoplasmic ribosomal proteins, as annotated in the first dataset. Several properties of these datasets are summarized in Table 1.

Most of the analysis described in this chapter was performed using the programs from NCBI Toolkit (http://www.ncbi.nlm.nih.gov), in particular PSI-BLAST and RPS-BLAST (Altschul et al. 1997; Schaffer et al. 1999, 2001). Intrinsic features were predicted using the SEG program for detection of low-complexity and non-globular sequences (Wootton and Federhen 1996) and the Coils2 program for prediction of left-handed coiled coils (Lupas 1996a). The SEALS suite (Walker and Koonin 1997) was used to manage the pipelines for analysis of sequence batches, and the tax_collector program from that package was used to automatically detect homologues of yeast proteins from different clades. Secondary structures and three-dimensional folds were predicted using the meta-server (Ginalski et al. 2003). Orthologous and paralogous relationships of the homologues were determined using the described criteria (Tatusov et al. 1997; Sonnhammer and Koonin 2002). Automatic assignment to NCBI COG database (Tatusov et al. 2001; http://www.ncbi.nlm.nih.gov/COG/new/) was performed using the modified cognitor program (Tatusov et al. 1997; M. Coleman and ARM, unpubl.).

2.1 Intrinsic Features

Properties of a biopolymer that can be computed without sequence database searches are sometimes called "intrinsic" features. In the case of proteins, these features include, for example, a fraction of amino acids with certain properties; the random vs. biased amino acid composition of whole proteins or certain regions within these proteins; or short strings of letters that can serve as biologically relevant "tags", for instance, intracellular sorting signals. Some of the intrinsic properties of the processosome components are summarized in Table 1. Proteins that participate in ribosome assembly appear, on average, to be similar in most respects to larger sets of nuclear proteins, although quite different from the known structural components of the nuclear pore.

A relatively high proportion of negatively-charged amino acids are observed in processosome components. Interestingly, they tend to occur in short (typically 3–7 amino acids) homopolymeric or mixed poly-aspartic/ poly-glutamic acid clusters (data not shown). Preliminary analysis indicates that this clustering may not be merely an artifact of a high proportion of charged residues in these proteins, but may be required for some aspects of processosome function, such as, facilitating interactions with and compensating the positive net charge of ribosomal proteins.

The runs of negatively charged amino acids are an extreme example of low compositional complexity of a protein sequence. The low-complexity regions

Table 1. Intrinsic features of the processosome proteins compared to other categories of yeast proteins

	Processosome	Nuclear proteins – Yale tagging project	Nuclear proteins – MIPS annotations	Nuclear pore components	Complete yeast proteome	Cytoplasmic ribosomal proteins
Number of proteins	85	122	636	16	6298	137
Average length	636.4	594.6	574.1	771.1	472.2	157.6
Percentage of negatively charged residues D+E	15	14.2	13.5	10.2	11.2	9.4
Percentage of negatively charged residues and amines, D+E+N+Q	24	25	24.7	23.9	21	17.2
Percentage of positively charged residues K+R	14.6	12.2	12.4	9.9	12	19.5
Percentage of residues that belong to predicted non-globular regions	26.7	33.1	26.5	50.9	22.7	21.3
Percentage of residues that belong to coiled coils	15.6	12.3	12.1	6.1	9.1	10.3

also include longer stretches enriched in one or only a few amino acids; if such a segment is on the order of 20 amino acids or longer, it typically adopts non-compact, elongated or flexible (non-globular) conformation (Wootton and Federhen 1996; Wan et al. 2003).

Perhaps the most distinctive feature of the processosome proteins is a high proportion of amino acids that are predicted to belong to left-handed coiled coils (Lupas 1996a, b). A coiled coil is a ubiquitous protein motif consisting of several (commonly two or three) alpha helices wound around each other, similar to the threads of a rope. Most coiled-coil sequences are based on heptad repeats, the seven-residue patterns in which the first and fourth residues, called core positions, are hydrophobic (typically leucine). As there are 3.6 residues in each turn of the alpha helix, these residues form a hydrophobic seam that slowly moves around the helix. The coiled-coil motifs in the same or several different molecules can join each other to bury their hydrophobic seams. The intercalation of side chains between neighboring helices ("knobs-into-holes" arrangement) stabilizes the coiled coils. Most coiled-coil prediction algorithms are based on detecting this heptad periodicity (Lupas 1996a, b).

The low-complexity sequences and coiled coils are two examples of regions with biased sequence composition, as compared to a more random distribution of individual residues in sequence databases as a whole (Altschul et al. 1994). Both of these types of regions tend to be non-globular (fibrillar or elongated), and frequently serve as hinges between globular domains in multidomain proteins, or as interfaces for protein oligomerization (Lupas 1996b; Wootton and Federhen 1996). A substantial fraction of amino acid residues in processosome components belongs to non-globular and coiled-coil regions (Table 1), and the majority of proteins have both types of these regions (67 out of 85 proteins contain non-globular segments and 59 contain coiled coils).

Thus, analysis of intrinsic sequence features in processosome components reveals many potential interfaces for protein-protein interactions. As the information on pairwise interactions between individual processosome proteins accumulates, one can use the prediction of these features in planning experiments on more precise mapping of these interfaces. Another practical application of the simple sequence regions is that they often help to demarcate the borders of globular domains in large multidomain proteins (Mushegian et al. 1997; see Fig. 1 for an example).

2.2 Evolutionarily Conserved Sequence Domains

I compared sequences of processosome components to the databases of protein sequences and of conserved sequence families and domains at NCBI (http://www.ncbi.nlm.nih.gov/BLAST). The summary of these results is shown in Table 2. Perhaps the main general conclusion of this analysis is that

Table 2. Phyletic patterns and functional annotations of the processosome components. Boldface indicates annotations from the NCBI COG database

Gene name	GenBank ID	Phyletic pattern[a]	Functional annotation[b]	R/D	PP	E
Imp3	6321942	MPFAb	Part of small (ribosomal) subunit (SSU) processosome (contains U3 snoRNA), interacts with Mpp10, Imp3p is a specific component of the U3 snoRNP and is required for pre-18S rRNA processing. It is not required for U3 snoRNA stability. Ribosomal Protein 4-like RNA-binding domain **COG0522 ribosomal protein S4 and related protein**	●		
Imp4	1730744	MPFa	Conserved Imp4/Brx1/Ssf1/Peter Pan family, consists of two domains; N-terminal domain is conserved in Archaea, the C-terminal region appears to be eukaryote-specific **COG2136 predicted exosome subunit/U3 small nucleolar ribonucleoprotein component, contains IMP4 domain**	●		
Lcp5	6320974	MPF– –	Lethal with conditional pap1 allele, Lcp5p			
Mpp10	6322461	MPF– –	Part of small (ribosomal) subunit (SSU) processosome (contains U3 snoRNA), Mpp10p			
Nop1	6320190	MPFAb	**COG1889 fibrillarin-like rRNA methylase**			●
Nop56	2833223	MPFA-	**COG1498 protein implicated in ribosomal biogenesis, Nop56p homologue**			
Nop58	6324886	MPFA-	**COG1498 protein implicated in ribosomal biogenesis, Nop56p homologue**			
Snu13	6320809	MPFAb	**COG1358 ribosomal protein HS6-type (S12/L30/L7a)**	●		
Sof1	6323018	MPFab	Part of small (ribosomal) subunit (SSU) processosome (contains U3 snoRNA), 56 kDa nucleolar snRNP protein that shows homology to beta subunits of G-proteins and the splicing factor Prp4, Sof1p **COG2319 WD40 repeat**		●	
Bms1	6325039	MPFab	BMH1-sensitive, Bms1p **COG0532 translation initiation factor 2 (IF-2) COG5192 GTP-binding protein required for 40S ribosome biogenesis**			●

Gene name	GenBank ID	Phyletic pattern[a]	Functional annotation[b]	R/D	PP	E
Dhr1	6323776	MPFab	Part of small (ribosomal) subunit (SSU) processosome (contains U3 snoRNA), ExtraCellular Mutant DEAH-box protein involved in ribosome synthesis, Ecm16p **COG1643 HrpA-like helicases**			●
Dim1	6324989	MPFAB	Dimethyladenosine transferase, (rRNA(adenine-N6,N6-)-dimethyltransferase), responsible for m6[2]Am6[2]A dimethylation in 3′-terminal loop of 18S rRNA, Dim1p **COG0030 dimethyladenosine transferase (rRNA methylation)**			●
Dip2	6323158	MPFab	**COG2319 WD40 repeat**		●	
Emg1	6323215	MPFA-	Essential for mitotic growth, Emg1p **COG1756 uncharacterized conserved protein**			
Enp1	6319724	MPF– –	Essential nuclear protein, Enp1p – homologue of eukaryotic bystin which has cytoplasmic/adhesion-related function; Enp1p also has glycosylation-deficient phenotype			
Kre31	6320926	MPFab	**COG2319 WD40 repeat**		●	
Kre33	6324197	MPFAB	Killer toxin resistant, Kre33p ; **COG1444 predicted P-loop ATPase fused to an acetyltransferase COG0454 histone acetyltransferase HPA2 and related acetyltransferases** (see text)	●		●
Krr1	6319791	MPFab	Involved in cell division and spore germination, Krr1p **COG1094 predicted RNA-binding protein, KH domains**	●		
Ltv1	6322706	MPF– –	Protein required for viability at low temperature, Ltv1p			
Nan1	6325131	MPFab	Part of small (ribosomal) subunit (SSU) processosome (contains U3 snoRNA), Net1-associated nucleolar protein 1, Nan1p **COG2319 WD40 repeat**		●	
Nop14	6320053	MPF– –	Part of small (ribosomal) subunit (SSU) processosome (contains U3 snoRNA), nucleolar protein 14, Nop14p			

Table 2. (*Continued*)

Gene name	GenBank ID	Phyletic pattern[a]	Functional annotation[b]	R/D	PP	E
Pwp2	6319903	MPF– –	Part of small (ribosomal) subunit (SSU) processosome (contains U3 snoRNA), eight WD-repeats with homology with G protein beta subunits flanked by nonhomologous N-terminal and C-terminal extensions, Pwp2p **COG2319 WD40 repeat**		●	
Rcl1	2500651	MPFAB	RNA cyclase **COG0430 RNA 3′-terminal phosphate cyclase**	●		●
Rok1	6321267	MPFab	High-copy suppressor of kem1 null mutant, Rok1p **COG0513 superfamily II DNA and RNA helicases**	●		●
Rrp5	6323885	MPFab	Part of small ribosomal subunit (SSU) processosome (contains U3 snoRNA). Rrp5p is the only ribosomal RNA processing trans-acting factor that is required for the synthesis of both 18S and 5.8S rRNAs. Rrp5p **COG0539 ribosomal protein S1 COG0457 TPR repeat**	●	●	
Rrp7	6319818	MPFab	Involved in rRNA processing, Rrp7p RRM-type RNA-binding domain			
Rrp9	6325394	MPFab	Part of small (ribosomal) subunit (SSU) processosome (contains U3 snoRNA) **COG2319 WD40 repeat**		●	
Rrp12	6325245	MPF– –	Required for normal pre-rRNA Processing. Member of a group of seven genes whose expression is repressed during growth on glucose before and during the diauxic shift, Rrp12p			
Utp20	14270688	MPF– –	Hypothetical 287.5 kDa protein in PDR3-HTA2 intergenic region			
Utp4	6320531	MPFab	Part of small (ribosomal) subunit (SSU) processosome (contains U3 snoRNA), Utp4p **COG2319 WD40 repeat**		●	
Utp6	6320657	MPFab	Part of small (ribosomal) subunit (SSU) processosome (contains U3 snoRNA), Utp6p **COG0457 TPR repeat**		●	
Ygr081	6321518	–––––	Hypothetical ORF, Ygr081 cp			

Gene name	GenBank ID	Phyletic pattern[a]	Functional annotation[b]	R/D	PP	E
Ygr090	6321527	MPF– –	U3 protein, Ygr090wp			
Utp8	6321567	– – – – –	Part of small (ribosomal) subunit (SSU) processosome (contains U3 snoRNA), Utp8p			
Ygr145	6321584	MPFab	Similar to hypothetical protein FLJ14075 [Homo sapiens] **COG2319 WD40 repeat**		●	
Utp9	6321990	mpfab	Part of small (ribosomal) subunit (SSU) processosome (contains U3 snoRNA), Utp9p. Contains repeats that may be distantly related to WD40 repeats			
Utp10	6322352	MPF– –	Part of small (ribosomal) subunit (SSU) processosome (contains U3 snoRNA), Utp10p			
Utp11	6322750	MPF– –	Part of small (ribosomal) subunit (SSU) processosome (contains U3 snoRNA), Utp11p **COG5223 uncharacterized conserved protein**			
Ykr060	6322913	MPFAb	Ykr060wp Domain related to ribosomal protein L1 **COG0081 ribosomal protein L1**	●		
Utp13	6323251	MPFab	**COG2319 WD40 repeat**		●	
Ylr409	6323441	MPFab	Protein required for cell viability, Ylr409 cp **COG2319 WD40 repeat**		●	
Utp15	6323740	MPFab	part of small (ribosomal) subunit (SSU) processosome (contains U3 snoRNA), Utp15p **COG2319 WD40 repeat**			
Nob1	6324630	MPFAb	Nin1 (one) binding protein, Nob1p **COG1439 Predicted nucleic acid-binding protein, consists of a PIN domain and a Zn-ribbon module.** Some PIN domains appear to have nuclease activity	●		●
Ypr144	6325402	MPF– –	U3 protein, localized to the nucleolus, Ypr144 cp			
Rio2	6324122	MPFAb	Protein required for cell viability, Rio2p **COG0478 RIO-like serine/threonine protein kinase fused to N-terminal HTH domain**	●	●	●
Pno1	6324720	MPFAb	Partner of Nob1, Pno1p **COG1094 predicted RNA-binding protein (contains KH domains)**			

Table 2. (*Continued*)

Gene name	GenBank ID	Phyletic pattern[a]	Functional annotation[b]	R/D	PP	E
Ssf1	6321857	MPFa-	Putative involvement in mating, Ssf1p Conserved Imp4/Brx1/Ssf1/Peter Pan domain or possibly two domains; N-terminal and middle portions are also conserved in Archaea, but the C-terminal region appears to be eukaryote-specific COG2136 Predicted exosome subunit/U3 small nucleolar ribonucleoprotein (snoRNP) component, contains IMP4 domain COG5154 RNA-binding protein required for 60S ribosomal subunit biogenesis	●		
Brx1	6324496	MPFa-	Conserved Imp4/Brx1/Ssf1/Peter Pan family, consists of two domains; N-terminal domain is conserved in Archaea, the C-terminal region appears to be eukaryote-specific **COG2136 predicted exosome subunit/U3 small nucleolar ribonucleoprotein component, contains IMP4 domain COG5154 RNA-binding protein required for 60S ribosomal subunit biogenesis**	●		
Rpf1	6321880	MPFa-	Conserved Imp4/Brx1/Ssf1/Peter Pan family, consists of two domains; N-terminal domain is conserved in archaea, the C-terminal region appears to be eukaryote-specific **COG2136 predicted exosome subunit/U3 small nucleolar ribonucleoprotein component, contains IMP4 domain**	●		
Rpf2	6322934	MPFa-	Conserved Imp4/Brx1/Ssf1/Peter Pan family, consists of two domains; N-terminal domain is conserved in archaea, the C-terminal region appears to be eukaryote-specific **COG2136 predicted exosome subunit/U3 small nucleolar ribonucleoprotein component, contains IMP4 domain COG5154 RNA-binding protein required for 60S ribosomal subunit biogenesis**	●		
Nop7	6321540	MPFab	Pescadillo homologue 1, containing BRCT domain **COG5163 Protein required for biogenesis of the 60S ribosomal subunit**		●	
Dbp9	6323306	MPFAB	**COG0513 Superfamily II DNA and RNA helicases**	●		●
Drs1	6323021	MPFAB	**COG0513 Superfamily II DNA and RNA helicases**	●		●

Gene name	GenBank ID	Phyletic pattern[a]	Functional annotation[b]	R/D	PP	E
Ebp2	6322676	MPF– –	EBNA1-binding protein homologue, Ebp2p			
Erb1	6323693	MPFab	Eukaryotic ribosome biogenesis, Erb1p **COG2319 WD40 repeat**		●	
Has1	6323947	MPFab	Helicase Associated with SET1, Has1p **COG0513 superfamily II DNA and RNA helicases**	●		●
Mak16	6319294	MPF– –	Putative nuclear protein, Mak16p **COG5129 nuclear protein with HMG-like acidic region**	●		
Nip7	6325045	MPFAb	**COG1374 Protein involved in ribosomal biogenesis, contains PUA domain**	●		
Nop2	6324268	MPFAB	May participate in nucleolar function during the transition from stationary phase to rapid growth, Nop2p **COG0144 tRNA and rRNA cytosine-C5-methylases**	●		●
Nop16	6320838	MPF– –	Nucleolar protein 16, Nop16p		●	
Nsa1	6321327	MPFab	Nop seven associated, Nsa1p Modified WD40-like domains (mostly FD)			
Nog1	6325164	MPFAb	**COG0536 Predicted GTPase COG1084 Predicted GTPase COG0486 predicted GTPase**			●
Nop4	6325213	MPFab	NOP77p **COG0724 RNA-binding proteins (RRM domain)**	●		
Rlp7	6324326	MPFA-	**COG1841 ribosomal protein L30/L7E**			
Rrp1	6320292	MPFAb	Involved in processing rRNA precursor species to mature rRNAs			
Tif6	6325273	MPFAb	**COG1976 translation initiation factor 6 (eIF-6)** Pentein-like structure, 5 copies of an alpha-beta repeat which may have a remote relationship to the S1 domain	●		
Ykl082	6322768	MPF– –	Required for normal pre-rRNA Processing, Ykl082 cp homologous to Surf6 protein in higher eukaryotes			
Ytm1	6324846	MPFab	Microtubule-associated protein, Ytm1p **COG2319 WD40 repeat**		●	
Loc1	14318523	– – – – –	Localization of mRNA			
Mrt4	6322843	MPFAB	60S acidic ribosomal protein P0 (L10E) **COG0244 ribosomal protein L10**	●		

Table 2. (*Continued*)

Gene name	GenBank ID	Phyletic pattern[a]	Functional annotation[b]	R/D	PP	E
Nop15	6324219	MPFAb	Nucleolar protein 15, Nop15p **COG0724 RNA-binding proteins (RRM domain)**	●		
Nsa2	6320973	MPFA-	**COG2007 [J] ribosomal protein S8E** family member	●		
Nsa3	6321843	MPFAB	Core interacting component 1, Cic1p **COG0081 ribosomal protein L1** family member	●		
Rlp24	6323037	MPFA-	**COG2075 ribosomal protein L24E**	●		
Spb1	6319796	MPFAB	Suppressor of PaB1 mutant, involved in 60S ribosomal subunit biogenesis, Spb1p **COG0293 23S rRNA methylase COG3269 predicted RNA-binding protein, contains TRAM domain**	●		
Nug1	6320842	MPFab	Nuclear GTPase, Nug1p **COG1161 predicted GTPases**			●
Nug2	6324381	MPFab	Nuclear/nucleolar GTP-binding protein 2, Nog2p **COG1160 predicted GTPases COG1161 predicted GTPases**			●
Dbp10	6320173	MPFab	Dead box protein 10, Dbp10p **COG0513 superfamily II DNA and RNA helicases**			●
Noc2	6324780	MPF– –	Nucleolar complex 2, involved in nuclear export of pre-ribosomes, Noc2p CBF/Mak21/Noc2–3 family has DNA-binding homologues, perhaps binds to RNA **COG5117 protein involved in the nuclear export of pre-ribosomes COG5604 uncharacterized conserved protein**	●		
Noc3	6323030	MPF– –	Nucleolar complex 2, involved in the nuclear export of pre-ribosomes, Noc3p CBF/Mak21/Noc2–3 family has DNA-binding homologues, perhaps binds to RNA **COG5117 protein involved in the nuclear export of pre-ribosomes**			
Rix1	6321991	MPF– –	Involved in processing ITS2, Yhr197wp			
Sda1	6321684	MPF– –	Severe depolymerization of actin			
Ycr072	10383804	MPFab	Protein required for cell viability, Ycr072 cp **COG2319 WD40 repeat**		●	

Gene name	GenBank ID	Phyletic pattern[a]	Functional annotation[b]	R/D	PP	E
Mdn1	6323135	MPFab	A large protein with a conserved N-terminal domain, a central AAA ATPase domain (with similarity to dynein) composed of 6 tandem AAA protomers, and a C-terminal M-domain containing MIDAS (Metal Ion Dependent Adhesion Site) sequence motifs, Mdn1p **COG0714 MoxR-like ATPases COG5271 AAA ATPase containing von Willebrand factor type A (vWA) domain**			●
Lpi2	6325111	MPF– –	Protein required for cell viability, Ypl146 cp			

[a] M, Mammals; P, plants; F, fungi other than *S. cerevisiae*; A, Archaea; B, bacteria. *Uppercase* indicates ortholog(s), *lowercase* indicates only paralogs in a given lineage.

[b] Functional annotation is derived from the definition lines for yeast proteins in GenBank and for their orthologs in other species, as well as from the COG database at NCBI (shown in bold) and from this study.

[c] In the last three columns, *R/D* indicates RNA-binding or DNA-binding activity, *PP* indicates protein–protein interaction domains, and *E* indicates enzymatic activity. *Black circles* indicate that respective activity has been observed or predicted.

```
>gi|6324197|ref|NP_014267.1| Killer toxin REsistant; Kre33p [Saccharomyces
cerevisiae]
1-189 First predicted globular domain (SEG program), corresponds almost
precisely to the first distinct homology region with predicted alpha/beta fold
(PSI-BLAST and fold recognition data), except that the latter extends to amino
acid 211 (Fig. 2A)
MAKKAIDSRIPSLIRNGVQTKQRSIFVIVGDRARNQLPNLHYLMMSADLKMNKSVLWAYKKKLLGFTSHRKKRENKIKK
EIKRGTREVNEMDPFESFISNQNIRYVYYKESEKILGNTYGMCILQDFEALTPNLLARTIETVEGGGIVVILLKSMSSL
KQLYTMTMDVHARYRTEAHGDVVARFNERFI
190-255 First predicted non-globular domain (SEG)
LSLGSNPNCLVVDDELNVLPLSGAKNVKPLPPKEDDELPPKQLELQELKESLEDVQPAGSLVSLSK
256-679 ATPase/helicase homology domain (PSI-BLAST and fold recognition)
TVNQAHAILSFIDAISEKTLNFTVALTAGRGRGKSAALGISIAAAVSHGYSNIFVTSPSPENLKTLFEFIFKGFDALGY
QEHIDYDIIQSTNPDFNKAIVRVDIKRDHRQTIQYIVPQDHQVLGQAELVVIDEAAAIPLPIVKNLLGPYLVFMASTIN
GYEGTGRSLSLKLIQQLRNQNNTSGRESTQTAVVSRDNKEKDSHLHSQSRQLREISLDEPIRYAPGDPIEKWLNKLLCL
DVTLIKNPRFATRGTPHPSQCNLFVVNRDTLFSYHPVSENFLEKMMALYVSSHYKNSPNDLQLMSDAPAHKLFVLLPPI
DPKDGGRIPDP
583-679 Acetyltransferase homology domain, N-terminal part (Fig 2B; PSI-BLAST
and fold recognition)
LCVIQIALEGEISKESVRNSLSRGQRAGGDLIPWLISQQFQDEEFASLSGARIVRIATNPEYASMGYGSRAIELLRDYF
EGKFTDMSEDVRPKDYSI
680-714 Predicted non-globular region connecting two halves of the
acetyltransferase homology region
KRVSDKELAKTNLLKDDVKLRDAKTLPPLLLKLSE
715-807 Acetyltransferase homology domain, C-terminal part (Fig 2B; PSI-BLAST
and fold recognition
QPPHYLHYLGVSYGLTQSLHKFWKNNSFVPVYLRQTANDLTGEHTCVMLNVLEGRESNWLVEFAKDFRKRFLSLLSYDF
HKFTAVQALSVIES
808-828 Short coiled-coil region (Coils2)
SKKAQDL
SDDEKHDNKELTRT
829-925 C-terminal mostly helical domain, possibly nucleic acid-binding (PSI-
PRED and PSI-BLAST)
HLDDIFSPFDLKRLDSYSNNLLDYHVIGDMIPMLALLYFGDKMGDSVKLSSVQSAILLAIGLQRKNIDTIAKELNLPSN
QTIAMFAKIMRKMSQYFR
926-1056 C-terminal non-globular domain (SEG)
QLLSQSIEETLPNIKDDAIAEMDGEEIKNYNAAEALDQMEEDLEEAGSEAVQAMREKQKELINSLNLDKYAINDNSEEW
AESQKSLEIAAKAKGVVSLKTGKKRTTEKAEDIYRQEMKAMKKPRKSKKAAN
```

Fig. 1. Segmentation of the Kre33p sequence into predicted structural and functional domains. The programs used to predict each particular feature are indicated; see text for the references to each program

for about one half of the proteins, some prediction of their molecular functions, interactions, or three-dimensional folds can be made on the basis of sequence similarities (note that this does not include annotations of knockout phenotypes and roles in ribosome synthesis available for the yeast proteins used as queries). Thus, the knowledge base about protein families, domains, their structure and function enables predictions even when experimental information is insufficient. Below, some of these predictions are described in more detail.

2.2.1 Kre33p, or Possibly AtAc: Protein with Multiple Predicted Activities

Kre33p (YNL132w) is associated with the 90S pre-ribosomal complex, thought to contain the 35S pre-rRNA and the U3 snoRNA. It is not detected in the pre-60S complex, suggesting a function either within the 90S complex

itself, or perhaps in the less-studied early pre-40S particle. The phenotypes of Kre33p knockouts in yeast and nematode (lethal in homozygote, haploinsufficient and K1 killer toxin resistant in yeast; slow-growing and locomotion-impaired in worm) suggest no clue to its biochemical activity.

Searches of the sequence databases and libraries of conserved domains, however, detect orthologs of Kre33 in all completely sequenced eukaryotic genomes, in most Archaea, and in many Proteobacteria from the gamma subdivision. By definition of an ortholog, a similar domain architecture provides for the alignment extended along the whole lengths of these proteins. Genes for Kre33p-like proteins are generally found in a single copy per genome, and their evolutionary tree is consistent with the generally accepted species tree (data not shown).

What might be the (shared) function of Kre33p proteins? Analysis of matches to the libraries of conserved sequence domains clearly predicts two distinct enzymatic activities of the Kre33p family. The second globular region (Fig. 1) in yeast Kre33p aligns with a number of predicted ATP-hydrolyzing enzymes with a Walker-type gamma phosphate-binding loop (Leipe et al. 2002). The downstream DEAA motif flanked by the predicted N-terminal beta strand and C-terminal alpha helix is almost certainly the Mg^{2+}-binding site, similarly to the well-known DEAD/H box in DNA and RNA helicases (data not shown). The set of other conserved motifs, partly overlapping with the helicase-specific signatures, suggests that Kre33p-like ATPases interact with DNA or RNA, possibly mediating the assembly/disassembly of RNA-protein complexes within the 90S pre-ribosome.

The region located to the C-terminus of the predicted ATPase domain (Fig. 2B) aligns to the large family of GNAT-like acetyltransferases, a group of enzymes that includes members with a wide range of substrate specificities, from small ligands to macromolecules. The three-dimensional structures of many GNAT-like enzymes, in wild type and mutated forms and in complexes with various adducts, have pinpointed the residues important for interaction with the donor of acetyl group, Acetyl-CoA (Angus-Hill et al. 1999). Most of these residues are well conserved in Kre33p-like sequences, indicating that they must be active enzymes.

Acetyltransferases are involved in a wide range of biological processes, from inactivation of antibiotics in bacteria to chromatin remodeling via acetylating histones and thereby changing their ability to interact with DNA. Recently, GNAT-like acetyltransferases have been shown to play a role in the acetylation of many other protein substrates, such as nuclear import proteins (Bannister et al. 2000) and multiple components of proteasome (Kimura et al. 2003). A hint at another possible role for Kre33p in ribosome assembly is given by the mass-spectrometry data on mammalian 40S ribosomal subunit and yeast 60S subunit (Louie et al. 1996; Lee et al. 2002), indicating that at least 13 ribosomal proteins are acetylated. Most of the activities responsible for these modifications are unknown, and Kre33p-like proteins are good candi-

```
secondary       hh sssssss          hhhhh    ssssss      hhhhhhhhhhh    sssss  hhhhhhhhhhh      hhhhhhhh       ssss     sss
At_21363086 18 LRHRSMFVIIGDKA 73 YKDSERILGNTFGICILQDFEALTPNLLARTIETVEGGGLVVLL13 MDVHDRFRTESHS  1 ASGRFNERF  8 ACVVMDDELNLLPL
Hs_21362641 19 ERQRSLFVVVGDRG 73 YNETHKILGNTFGMCVLQDFEALTPNLLARTVETVEGGGLVVIL13 MDVHSRYRTEAHQ  1 VVGRFNERF  8 KCLVIDDQLNILPL
Ce_17507621 18 SGHRSMFAVVGDKA 73 YNETEKILGNTFGVLVLQDFEAMTPNLLARTIETIEGGGMVILL13 MDVHNRYRTEAHN  1 ITARFNERF  8 SVLVLDDQLRVLPM
Dm_21363085 19 LGHRTMFIVIGDKA 70 YSETHAVLGRTYGVCVLQDFEALTPNLLARTVETVEGGGLIILL13 MDVHKRFRTEAHQ  1 VTCRFNERL  9 CLVVNDDLTVLPLL
Sp_19114238 19 EKQRSFFVVVGDRA 74 YKESEKILGQTYGMLVLQDFEALTPNLLARTIETVEGGGIVVLL13 MDIHSRYRTEAHS  1 VTARFNERF  8 NCLVIDDELNVLPL
Sc_6324197  19 TKQRSIFVIVGDRA 74 YKESEKILGNTYGMCILQDFEALTPNLLARTIETVEGGGIVVIL13 MDVHARYRTEAHG  1 VVARFNERF  8 NCLVVDDELNVLPL
Ap_14602036 38 KRHRVMVVISGSDP 64 YEESDRYLGTTFQALVLDLTNDLKPNDVGRLVEVVEGGGLIILQ 2 WERWDTALTLFKK  3 VPGHEEPRH 15 GIIVYDADSRKILA
Mk_20093586 36 HRHRRMIVFQGDEG 67 FSGSKRALGTTWDLMVIDLSYDLDPDAIGRLVETVRGGGLVIFQ 2 FDRWRNMWTAFHK 11 VGKRFNRRF  8 GVWIVDTDEWTAET
Af_11499117 19 NRHRFMVILCSKEL 46 YKDSPKVLGQTYSSLLIDFTEGFHPNDLGIIVETIKEGGIIIAF 5 WENLKSRWHEELV  8 VTGRFFARF  8 GIIIFDADRRKIIS
Pa_14521475 37 NFHRRMIILQGDTL 74 FKYSRDILGRTFDILVLDLSYDYSPNDLGRIIETVRGGGLIFIL 2 FEKWKDMWTGFHK 11 VKKRFNRRL  8 GIYIVDADKKKIET
Pm_15602908  4 NKQRKVQCWVGENA 32 FKQAKNMLGQEFALIVYDARGGLNLDALAIASGTLQQDGVLLIL 9 QDPDSVRWSGEND  2 LTPHFLHYF  0 QQKVRDYGFPVYHL
St_16765805 16 EGIRRLLVISGDAA 31 PQALQTLLGREFRHAIFDAWQGFDAAAFAALSGTLQAGSWLLLL 9 PDTDSLRWSDCAQ  2 PTPQFAQHL  1 RTLSRDPQTLLWRM
Ec_15832590 16 EGIRRLLVLSGEEG 31 PSALQTLLGREFRHAVFDARHGFDAAAFAALSGTLKAGSWLVLL 9 PDADSLRWSDCPD  2 ATPHFVQHF  1 RVLTADNDAIFWRL
Yp_16123233 15 QGIRRLLILSGQAD 32 PTAARTLLGQENLHGVFDATDGLNIEALAIVAGTLRAGSWLLLL 9 PDKDSLRWSEQPA  2 VTANFIRHL  1 RQFLADPDVVLWQV
consensus       xxxRxUUUUxO=xO   UxxxxxUUGxxUxxxUUDxxxxUxxxxUOxxxxTxxOOOUUUUU  xDxxxxxxxxxxx   UxxxUxxxU    xUUUUDxxxxxUxU
                          *                  *                                                          *
```

A

```
secondary        sssss   sssssss        sssssss  hhhhhhhhhhhhhh      sssssss    hhhhhhh  sssss        sssssssss
secondary 1KUY   sssss   sssssss     sssssssssss       hhhhhhhh     sssssss      hhhhh   ssssss          sssssss
11387097_1KUY 80 SLGWFVEGRLVAFII 19 RGHSAHLHALAVHRSFRQQGKGSVLLWRY 10 RRAVLMCEDA 0 LVPFYQRFGFHPAGPCAIVVGSLTFTEMHCSL
At_21363086  595 VIQVCLEGQISRKSA 29 KLSGARIVRIAVHPNAMKMGYGSAAVELL 59 RLHYIGVSFG 4 LFRFWRKHKFAPFYISQI--PSAVTGEHTCML
Hs_21362641  575 VIQVCLEGEISRQSI 29 GLSGGRVVRIAVHPDYQGMGYGSRALQLL 55 RLDYLGVSYG 4 LLKFWKRAGFVPVYLRQT--PNDLTGEHSCIM
Ce_17507621  568 VVQVCLEGRLDSDNI 29 TLCGGRIVRVAVHPDYQSMGYGGRAVQLI 60 RLDYLGVSFG 4 LLKFWKRNEFVPVYIRQN--SNDITGEHTCII
Dm_21363085  550 VIQVALEGQISAQSI 29 KLCGVRIVRVATHPNYQRMGYGKRAIQLL 58 HVDYIGTSYG 4 LLKFWKNAGFVPVYLSQK--SNELTAEHSCIM
Sp_19114238  575 VIQLALEGSISRESI 29 ALGGARIVRIAVSPEHVKMGYGTRAMQLL 60 PLHYVGVSYG 4 LQKFWKREGYCPLYLRQT--ANDLTGEHTCVM
Sc_6324197   587 VIQIALEGEISKESV 29 SLSGARIVRIATNPEYASMGYGSRAIELL 60 YLHYLGVSYG 4 LHKFWKNNSFVPVYLRQT--ANDLTGEHTCVM
Ap_14602036  538 AAQVAEEGGLGEELI 27 RMVGWRIVRIATHPEVQGRGIGSWLLERL  7 GYDWLGSGFG 4 LLRFWLRNGFLVLHLSPD--RNPVSGEYTSLV
Mk_20093586  540 ALQVAREGTIPRDVI 28 RMKGLRIVRIATHPDIMRHGLGSRALKEL  7 DYDWIGTGFG 4 LTRFWLRNGFVPVHISPN--RNPVSGEYSVAV
Af_11499117  502 SLHIAVEGNLDDETI 28 KLVGARVVRIATHPSVMDMGIGSFALAKI  7 DMDWVGSGFG 4 LLRFWAKNHFTPVHITPQ--RNEVSGEHTLIV
Pa_14521475  536 AIQIAKEGGIPKAVI 28 KLRGYRIVRIATHPDAMDLGLGSKALELL  7 GLDWVGSGFG 4 LIRFWVRNGFAVVHLSPT--RNPVSGEYTAIV
Pm_15602908  406 AAWLVQEGSLQSEQL 29 RLSSLRISRIAVQPNWQQQGIGQQLVKNI  4 QVDFLSVSFG 4 LAAFWQKCGFLLVHLGEQ--KEASSGCYSVIG
St_16765805  407 ALWLVEEGGLSAELS 29 TLVGRRVSRIAVHPARQREGIGQQLIACA  5 QCDYLSVSFG 4 LWRFWQRCGFVLVRMGNH--REASSGCYTAMA
Ec_15832590  407 ALWLVDEGGLSQELS 29 TLRGRRVSRIAVHPARQREGVGQQIIASA  5 GLDYLSVSFG 4 LWRFWQRCGFVLVRMGNH--REASSGCYTAMA
Yp_16123233  417 ALWLVDEGGLSEALA 29 TLLSRRISRVAVTAAWRQQGIARRMIAAE  7 QCDFLSVSFG 4 LAHFWHRCGFRLVRIGSH--KEASSGCYTAMA
consensus        OuxUxxEGxUxxxxU   xUxOxxUxxUxUxxxxxxxxGxOOxxUxxx   xxxUUxxxxO   UxxUUxxxxGUxPxxxxxxxxxxxxxxxEUxCxU
CoA interaction                          * *       *  *
```

B

Fig. 2. Conserved sequence motifs in two of the four predicted domains in the Kre33p protein family. Species abbreviations, sequences identifier in GenBank, and where applicable, in the PDB database of three-dimensional structures (http://www.rcsb.org), and the distances from the N-terminus of the protein are shown before each sequence. Conserved sequence residues are shown in *bold face*. In the consensus line, + stands for a positively charged residue (K or R), = stands for negatively charged residue (D or E), *U* stands for a bulky hydrophobic residue (I, L, V, M, F, Y, or W), *O* stands for a residue with a small side chain (A, G, or S), and *x* stands for any residue. In the secondary structure line, assignment to an *alpha helix (h)* or *beta strand (s)* is taken from the PDB entry where indicated, or predicted using the PSI-PRED program (McGuffin et al. 2000). **A** N-terminal putative enzymatic domain with alpha-beta fold. *Asterisks* indicate conserved acidic residues that follow *beta strands* (see text). **B** GNAT acetyltransferase-like domain. *Asterisks* indicate the residues making contacts with the substrate acetyl-CoA in serotonin N-acetyltransferase (pdb structure 1KUY). Species are abbreviated as follows: *At*, *Arabidopsis thaliana*; *Hs*, *Homo sapiens*; *Ce*, *Caenorhabditis elegans*; *Dm*, *Drosophila melanogaster*; *Sp*, *Schizosaccharomyces pombe*; *Sc*, *Saccharomyces cerevisiae*; *Ap*, *Aeropyrum pernix*; *Mk*, *Methanopyrus kandleri AV19*; *Af*, *Archaeoglobus fulgidus*; *Pa*, *Pyrococcus abyssi*; *Pm*, *Pasteurella multocida*; *St*, *Salmonella typhimurium LT2*; *Ec*, *Escherichia coli O157:H7*; *Yp*, *Yersinia pestis*

dates for that role. It would be also of interest to know whether rRNA and any of snoRNAs are acetylated in vivo and what the function of such a modification might be. Given the presence of *AT*Pase and *Ac*etyltransferase domains in Kre33p and its orthologs, I provisionally rename this family AtAc (pronounced *attack*), even though this name is already used as a synonym to lymphotactin (a cytokine; e.g. Dorner et al. 2003).

The name AtAc may be short-lived, though, because two more conserved regions can be defined in the family. The most C-terminal region is predicted to consist mostly of alpha helices, and displays marginally significant yet potentially interesting similarity, in PSI-BLAST searches, to C-terminal regions of bacterial sigma-70 transcription factors. The sigma-70 sequences most similar to the AtAc C-terminal domain were the sporulation-specific factors from Gram-positive bacteria (sigma-G factor from *Bacillus cereus*, GI 30021992), matched the C-terminus of AtAc from *Acidianus sp*., GI 14279357, at thefifth PSI-BLAST iteration with the probability of random match $E=10^{-3}$, but the domain in question is found in a wide variety of sigma factors (data not shown). Known under the name of sigma-4, it is one of the most conserved and functionally important regions in the sigma factors, consisting of three tightly packed helices (Campbell et al. 2002). The mode of interaction of sigma-4 with DNA (in bacteria, the prime target is the –35 promoter region) seems to be exclusively via the major groove of the double-stranded form, so the immediate implications to the AtAc activity are not clear.

Recently, structural and functional similarity has been described between essentially the same helix-turn-helix motif in sigma-4 and a region in a different family of proteins also involved in the processosome function, namely the Imp4 family (Wehner and Baserga 2002). I will discuss the significance of these observations in the next section (see Sect. 2.2.2).

The N-terminal region of the AtAc proteins, located upstream of the ATPase-like domain and separated from it by a putative low-complexity sequence linker, is not similar to any other sequences in the database. However, sensitive methods of secondary structure prediction (McGuffin et al. 2000) and fold recognition based on probabilistic modeling and threading (Ginalski et al. 2003) indicate that this domain belongs to the alpha/beta class, and may be a representative of a Rossmann-like fold within that class. Indeed, weak sequence and structure similarity has been observed between AtAc-N domain and several S-adenosyl-methionine dependent methyltransferases that have a Rossmann-like structure (Fig. 2A and data not shown). Conserved acidic residues, located close to the C-termini of the predicted beta-strands, are characteristic of many enzymes that belong in this fold (Lesk 1995) and can be observed in AtAc-N. Thus, AtAc-N is likely to be an active enzyme, though the type of reaction it catalyzes is not possible to predict with certainty. I speculate that there might be yet another transferase involved in the modification of ribosomal proteins or RNA. Alternatively, this or other enzy-

matic domains in AtAc's may be involved in regulating, by covalent modification, the activities of other components of processosome.

2.2.2 Imp4/Ssf1/Rpf1/Brx1/Peter Pan Family of Proteins

Five putative processosome components, namely Imp4p (YNL075w), Ssf1p (YHR066w), Brx1p (YOL077c), Rpf1p (YHR088w), and Rpf2p (YKR081c), share significant sequence similarity. All these proteins, as well as their other recognizable homologues in yeast, Ssf2p (YDR312w), are involved in the formation and activity of RNP complexes in the nucleus (Wehner et al. 2002; Wehner and Baserga 2002). The known roles of Rpf1 and Rpf2 are mainly in ribosome maturation, whereas Ssf1 appears to be additionally, and Ssf2 exclusively, involved in splicing. Multiple homologues of these proteins are found in all eukaryotes, including the fruit fly protein, Peter Pan, required for larval growth. One homologue of the Imp4/Peter Pan gene per genome is detected in some (but not all) Archaea, which are usually found next to the genes encoding the exosome components, i.e. the RNA processing nucleases whose homologues in yeast are involved in pre-rRNA processing (Koonin et al. 2001).

Based on this evidence, the RNA-binding roles have been suggested to the members of the Imp4/Peter Pan family, and experiments have confirmed that Imp4p, Rpf1p, and Rpf2p bind single-stranded (ss) RNA, displaying somewhat different nucleotide preferences, and do not bind double-stranded RNA or single-stranded DNA (Wehner and Baserga 2002). The substantial part of the ssRNA-binding activity resided in the C-terminal domains of Rpf1p and Rpf2p; corresponding segments of these proteins conferred RNA-binding ability to an unrelated protein (luciferase).

I aligned eukaryotic and archaeal Peter Pan-like proteins and attempted to predict their secondary structure and tertiary fold. The family alignment indicates that similarity between Eukarya and Archaea is distributed over most of the length of archaeal proteins, but is confined to the middle portion of the larger eukaryotic homologues (Fig. 3). Thus, eukaryotic proteins appear to consist of two conserved sequence domains, one shared with Archaea and the other unique (and an additional, non-globular and less well-conserved, N-terminal regions – data not shown). No statistically significant sequence similarity was observed between either of the two domains and any proteins from other families.

Wehner and Baserga (2002) used a search with the eMOTIF program and library (Huang and Brutlag 2001) to match the C-terminal RNA-binding segment of eukaryotic Peter Pan-like proteins to the sigma-70 helix-turn helix motifs in bacterial transcription factors. Unlike the well-studied statistics of database searches used in the BLAST suite of programs (Altschul et al. 2001; Schaffer et al. 2001), the statistical properties of short motifs is not well understood, and I was not able to quantify the specific sequence affinity between Peter Pan-C domain and sigma-70. Moreover, secondary structure prediction

```
            Secondary          eeee    hhhhhhhhhhhhhh      hhhhhhh  hhhhhhh      eeeeeee      eeeee
Archaea     Ma_20090593    2 ITTSRKPSAKTRTLCKLLSRFITGR-CITRGKMGMQELIGFAEGGP  0 LIVVGEYHGNPGELGFYDDA
            Ss_15897633    9 ITSSRDPSIRTRNFLNVLTFVLPDSVKITRGKKSKIEIFERAINLG  3 LLFVLAKNGNPLRIIVYDLE
            Pa_18313175    7 ITTSREPSKKTLELVNDLVNSLPGTSKIVRGKKSFITLLEEAVACG  3 IAFIWERRGMPFALLFYDVI
            Ap_14601409   17 VTTSRRPSPRIRSFVKDLSATIPGAFRFTRGHYSMEELAREAIIRG  3 IVVVGERRGNPGIIRVYAVE
            Mk_20093815    5 ITTSQRPARRTRSLCRDLECALPDATYVLRGTKNLRDTVLEALESG  3 LFYVTEAKGNPARLHVIDLG
Eukarya     Dm_24649922   74 LTTSHNPSSRLKMFMKELRLIIPNAQQMNRGNYQLTTLMHACRANN 16 LVVCHLPYGPTAFFNISDVV
            At_18411672  130 LTTCRFNSTRGPALISELLSVIPNSHYQKRGTYDLKKIVEYATKKD 16 LLIIGLPNGPTAHFKLSNLV
            Sp_19114697   87 VTTSREPSSRLAQFAKEVRLLIPNSYRLNRGNIVVGSLVEAARAND 16 LVISHLPYGPTLSFSLHNVV
            Ce_17544642   81 ITTSRDPSSRLKMFAKEMKLIFPNAQRINRGHYDVKQVVQASKAQD 16 MLVCHLPFGPTAFFSMANVV
            Ce_17534031  180 ITMTPKAKITTFKFCFELQKCIPNSEIFTRKNVLLKTIIEQAKERE 16 IIFCHLPEGPTAYFKINSLT
            Ce_17554174   66 VLCSRGADVRTRYLMKDIKDLLPHAKGDSKLDQQKSLNVLNEIAEM 18 LWMSNVEKGPSIKFLVHNVH
  Imp4p_Sc_6324254   89 VTTSRDPSTRLSQFAKEIKLLFPNAVRLNRGNYVMPNLVDACKKSG 16 LTISHFPHGPTAQFSLHNVV
            Nc_11595521   87 VTTSRDPSSRLGQFSKEIRLLLPTSVRLNRGNLVLEDLVGAAKAQN 16 MTISHFPHGPTLMVSLHNVV
            At_15222534   76 LTTSRNPSAPLIRFTKELKFVFPNSQRINRGSQVISEIIETARSHD 16 LIISHLPFGPTAYFGLLNVV
            Tc_4262378    86 VTTSREPSQKLLEFAKEIRLVIPSAVRMNRGNLSVRQLMDAARRGQ 16 LTVSHLPLGPTVVFTIHNLV
            Hs_13376605  145 ITTSDRPHGRTVRLCEQLSTVIPNSHVYYRRGLALKKIIPQCIARD 16 LILSHLPNGPTAHFKMSSVR
            Sp_19111899   46 VLSSRGVTYRQRHLLNDLVSMMPHSKKDSKLDSKDRLYQLNELAEL 18 LHIARAPNGPTVKFHVENLH
            Nc_18376164   39 ILSSRGVTYRHRHLLNDLAAMLPHGRKDVKFDSKSKLYQLNELAEL 18 MWFSKVPNGPTVKFHAQNLH
            consensus        UTTS+xPOx+xxxUxxxUxxUUPxxxxxx+GxxxUxxUUxxOxOxx     UUUxxxxxGxxxxUxUxxxx

secondary       hhhhhhhhh        eeee          hhhhhhhh      eeeee     eeeeeee      eeeeee      hhhhhhhhhhhhh
Dm_24649922 166 MRHDIPDIGHMS 2 KPHLIFNNFKTPIGLR--TVKILKHLFPVPK  4 RVMSFLNHNDSIIFRHH  4 VNKELELTEVGPRFSLKLYQIKLGTL
At_18411672 213 LRKDIKNHGNPT 2 QPELVLNNFTTRLGNR--VGRFFQSLFPPDP  5 RVVTFHNQRDFIFFRHH 16 ETIKPRLQECGPRFTLKLVTLQHGTF
Ce_19114697 169 LRHDIPNTGTMS 2 YPHLIFENLTSKLGKR--VKTALSALFPPDP  5 RVVTFANTDDYISFRHH  5 GPKQIILSEAGPRFEMKLFEITLGTV
Ce_17544642 163 MRHDIPNCGTMS 2 YPHLIFDNLNSKLGHR--FTTILKHLFPVPK  4 RIITFSNSEDYISFRHH  5 NDGEVELTEAGPRFELKPYQIKLGTL
Ce_17534031 216 FTQDLKKFGEST 2 FPEVILNNFNTRLGHN--IARMLACLFPHDP  5 RVVTFHNQRDYIFFRHH  6 EGSKAALLELGPRFTLRLKWLQKGTF
Ce_17554174 150 TMKELKMSGNCL 3 RPVLSFDDAFDKKPQLKLIKAVLMQTLGTPH  9 DHVFNFSVGEGDKIWFR  4 VDESLQLQEVGPRFVLEMVRLFAGSF
Sc_6324254  171 MRHDIINAGNQS 2 NPHLIFDNFTTALGKR--VVCILKHLFNAGP  5 RVITFANRGDFISVRQH  4 TREGVEIAEVGPRFEMRLFELRLGTL
Nc_11595521 169 LRADIPKSIKGT 4 YPRLIFEGFSTKLGER--VVKILKHLFPPRE  9 RVITFVNNDDTIEVRHH  5 SYDSVELSEVGPRFTMKPFKITMGTL
At_15222534 168 TRHDISDKKSIG 5 YPHLIFNNFTTQMGQR--VGNILKHIFPAPK  4 RIVTFSNQSDYISFRNH  7 GPKSIELKEIGPRFELRLYQVKLGTV
Tc_4262378  168 TRHDIQDVGTMS 2 HPHLIFENFTTRLGRR--VRDVLKFLFPVPK  4 RVLTFDNQNDFVSFRHH  5 KGREVQLTEVGPRMDVAPYRITLGTL
Hs_13376605 226 LRKEIKRRGKDP 3 IPEIILNNFTTRLGHS--IGRMFASLFPHNP  5 QVATFHNQRDYIFFRFH  5 SEKKVGIQELGPRFTLKLRSLQKGTF
Sp_19111899 130 TMDELNMTGNAL 3 RPILSFDKTFDTAPHLKVVKELLQQTFGIPK 10 RVCTLTIADGKIWFRNY  8 SKDPVTLIEIGPRFVMTIINILEGSF
Nc_18376164 124 TMEELHFTGNCL 3 RPLLSFDLAFEQEPHYRVIKEVFLQMFGVPQ 10 HVMSFSIADGKIWVRNY 39 KDTDVSLVEIGPRFVLTPIIIQEGAF
consensus       xxx=UxxxGxxx   xPxUUUxxUxxxxGxxxxUxxUUxxUUxxxx     xUUxUxxxx=UuxU+xx     xxxxUxUxEUGPRUxUxUUxUxxGxU
```

Fig. 3. Sequence conservation in Peter Pan family. *Above* Domain shared by archaeal and eukaryotic proteins. *Below* Eukaryote-specific, C-terminal domain. Species abbreviations: *Nc*, *Neurospora crassa*; *Tc*, *Trypanosoma cruzi*. Other designations are as in Fig. 2

suggests an alpha-beta structure for both domains in Peter Pan-like proteins (Fig. 3), and specifically, indicates that the area to the N-terminus of the turn with the characteristic ExG signature is more likely to be a beta-strand than an alpha helix. Determination of the structure of the C-terminal, RNA-binding region of eukaryotic Peter Pan-like proteins will resolve the question of similarity between these domains and sigma-70, and of the mode of RNA binding by Peter Pan-C. The apparently independent existence of Peter Pan-N domain in archaea suggests a distinct, second molecular function.

The dedicated involvement of the Peter Pan/Imp4 family in nucleolar RNA metabolism, and the addition of eukaryote-specific domain to an archaea-specific module are reminiscent of another protein superfamily involved in an RNA-directed process. The post-transcriptional gene silencing (PTGS) pathway occurs in cytoplasm, is mediated by specific small RNAs, and requires several specific protein components. A protein family that appears to be dedicated to the PTGS is the eukaryotic Piwi/Argonaute/Zwille family that consists of two distinct protein domains. The centrally located PAZ domain is eukaryote-specific, and is also found in another component of the PTGS pathway, the helicase/exonuclease Dicer, and is thought to mediate heterodimerization of Argonaute and Dicer (Anantharaman et al. 2002). The C-terminal Piwi domain is found in eukaryotes and, as a stand-alone protein, in Archaea. In a purely superficial analogy with this system, it could be argued that the Peter Pan-N domain performs a function shared by Archaea and eukaryotes, and the Peter Pan-C domain may be involved in the maturation of eukaryote-specific supramolecular complexes.

2.2.4 Diverse RNA-Binding Domains and Limited Repertoire of Globular Protein Interaction Modules

The substrates of processosome are pre-rRNA, its various processed forms and, finally, ribosome subunits that, even in mature form, interface with the environment mostly via RNA (Moore and Steitz 2002). Thus, it is not unexpected to detect RNA-binding domains in many processosome components. Indeed, a comparison of the databases of sequences and conserved sequence domains, as well as biochemical evidence, predicts an RNA-binding function and, typically, the existence of known RNA-binding domains, for at least 30 of the processosome proteins. What may be less expected is the amazing variety in sequences and structures of these domains. I counted at least 17 distinct types of known or purported RNA-binding domains (Table 2).

Some of the processosome components are RNA-modifying enzymes that employ discrete domains to interact with their substrates. This is the case for helicases (C-terminal regions, including, probably, distinct HELICc domain) and methyltransferases (PUA domain). Other proteins seem to consist mostly of RNA-binding domains, such as KH, RRM, or TRAM (Letunic et al. 2002; http://smart.embl-heidelberg.de/). Furthermore, a number of processosome

components appear to be the paralogs of ribosomal proteins, many of which contain RNA interaction domains named after them (e.g. various forms of S1 domain and S4 versions of the PUA domain). Finally, several domains are found in both DNA-binding and RNA-binding proteins, such as versions of Helix-Turn-Helix domain in AtAc and in Rio2p protein kinase, of HELICc domain in helicases, and of a yet uncharacterized domain in CBF/Mak21/Noc2/Noc3 a family related to CCAAT-binding transcriptional coactivator subunits (Table 2).

Not only do the sequences of the nucleic acid-binding modules belong to a broad variety of sequence families, but the spatial structures of these domains are also extremely diverse. Five high-level structural classes are commonly described (http://scop.mrc-lmb.cam.ac.uk/scop/data/scop.b.html), and representatives of most of them are predicted among the processosome. For example, the helical domains mentioned in previous sections belong to all-alpha folds, L1 insertion domain (YKR060w) to alpha-beta, ribosomal protein S4 –like fold (Imp3) is segregated alpha+beta, and the C-terminal domain of Nob1p is predicted to form a zinc ribbon from the class of small ligand-stabilized proteins.

A significant fraction of processosome proteins (at least 10) also contains specific conserved domains that are likely to be involved in protein-protein interactions. In contrast to the RNA-binding domains, there are only a few distinct classes of such domains in processosome. With a single exception of one BRCT domain in Nop7p, specific protein interaction domains in processosome are either of WD40, or of TPR/halfTPR type.

Although many proteins contain one interaction domain, having the same type of protein interaction module does not imply that they all share one and the same interaction partner. Despite the ease with which WD40 and TPR domains can be recognized in sequence similarity searches, their sequences are highly diverged in the processosome proteins. Moreover, mutagenesis experiments have shown how a change of even a few amino acids in these domains may result in dramatically different interactions (see domain annotations in http://smart.embl-heidelberg.de/ for the details). Apparently, larger protein-protein interfaces, with a potential of chemically diverse interactions, afford substantial specialization of specific processosome components, even though the interacting modules are broadly of the same type. A different principle seems to operate in protein-RNA recognition, where interfaces may be smaller, and the sequence of one interaction partner (ribosomal RNA) is evolutionarily constrained.

3 Phyletic Patterns

In an attempt to reconstruct at least some steps in the evolutionary history of processosome, I gathered information on the evolutionary lineages in which

the homologues of the processosome proteins are found. The presence and absence of orthologous genes in different lineages can be presented as a binary vector (set of 0′s and 1′s) or coded with letters (the third column in Table 2), and is called a phyletic pattern. I traced the phyletic patterns of processosome proteins in three eukaryotic lineages, metazoans, plants, and fungi, as well as in Archaea and Bacteria.

Whenever the orthologs of a given gene can be found in all three major domains, i.e., Bacteria, Archaea, and Eukarya, this suggests that an orthologous gene has been present in the common ancestor of three domains. If orthologs are found only in Archaea and Eukarya, this indicates the emergence of the gene later in evolution. These patterns can be adulterated by horizontal gene transfer, which is apparently a widespread process, as extensively documented, for example, in Mirkin et al. (2003); however, it will not be considered here. At a higher resolution, certain genes may be present in some but not all lineages within a domain, suggesting lineage-specific gene losses and possibly displacements of these functions by different, non-orthologous enzymes (Koonin et al. 1996). Taken as a whole, the distribution of phyletic patterns helps one to understand the rate and order of the evolutionary accrual of different domains with their specific functions.

A special, and important, case of sequence similarity is paralogous similarity, i.e. the relationship resulting from domain duplication and rearrangement. If a gene from an evolutionarily more recent lineage lacks orthologs, but has paralogs, in a more ancient lineage, the emergence of a present-day function can be explained by cooptation of a copy of an ancestral gene for a new function. Often, more complex functional systems in eukaryotic cells can be parsed into components, most of which are traceable to simpler prokaryotic systems.

All told, seven phyletic patterns are represented in processosome components. At one extreme, three proteins, Ygr081, Utp8p, and Loc1p, are missing even in fungi other than *Saccharomyces cerevisiae*. Assuming that their association with processosome is functionally relevant, these proteins have either evolved beyond recognition in other lineages, or have been displaced by unrelated proteins with the same function. At the other extreme, there are nine proteins represented by orthologs everywhere. Seven of them are enzymes with known biochemical activities, dimethyl adenosine transferase Dim1p, RNA methylases Nop2p and Spb1p, the multidomain enzyme Kre33p/AtAc described above, RNA 3'-teminal phosphate cyclase Rcl1p (one copy in bacteria and archaea, lineage-specific duplication in eukaryotes – Bill et al. 2000), helicases Dbp9p and Drs1p, and two proteins are related to distinct classes of ribosomal proteins, Mrt4p and Nsa3p.

There are 20 proteins with the MPF– – pattern (orthologs in all lineages of eukaryotes, no orthologs or paralogs in prokaryotes). None of such proteins are predicted to have an enzymatic activity; Mac16p, Noc2p and Noc3p contain RNA interaction domains. There are five proteins with pattern MPFa–

(orthologs in eukaryotes, paralogs in Archaea) – all from the Peter Pan family – and six MPFA– proteins, again with no known enzymatic activities, but with three ribosomal protein-related domains. The largest class has the pattern MPFab, with 29 proteins, 7 of which are Walker-type ATP/GTP-binding enzymes with ancient, ubiquitous core domain (Leipe et al. 2002), and the others are non-enzymatic proteins dominated by protein–protein interaction domains. Finally, 12 MPFAb proteins are the most diverse ones, including both interaction adaptors and several enzymes with diverse specificities.

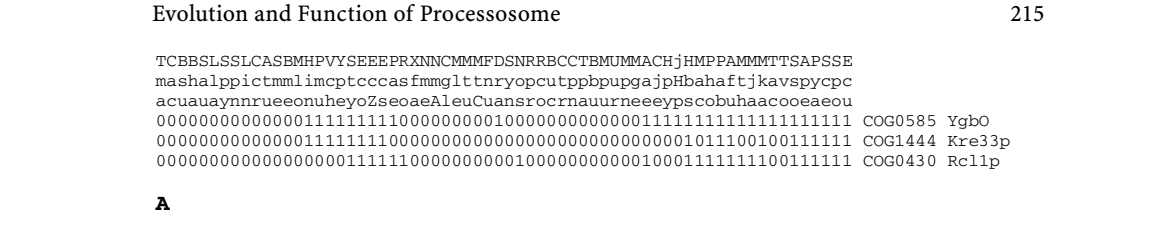
From this brief outline, the following trends emerge. Essentially all the enzymes involved in ribosome assembly have been present early in the evolution of Life, perhaps in the last common ancestor of Bacteria, Archaea and Eukarya. Many types of modules for protein–protein and protein-RNA interaction have also been present at that stage. Later in the evolution of eukaryotes, copies of some enzymes were recruited to perform additional functions, and interaction domains greatly proliferated in numbers and diverged in sequences. Archaea-specific "inventions" in processosome are mostly limited to the Peter Pan-N domain. About one-fourth of processosome components are eukaryote-specific, but very few clues to their function or origins currently exist.

A more detailed evolutionary analysis has to take into account not just the deep branches of the Tree of Life, but examine in more detail the gene content of the recent lineages. For example, although both Rcl1p and Kre33p are found in Eukarya, Archaea, and Bacteria, they are actually missing from a few Archaea and are found in only a few bacteria, mostly from the gamma subdivision of proteobacteria (Fig. 4A).

While the reconstruction of the most probable path of evolution leading to such "patchy" patterns is possible yet technically challenging (Mirkin et al. 2003), the pattern itself can be used for the reconstruction of functional pathways. The basic idea is that genes whose phyletic patterns are similar to each other are more likely to be functionally linked (Pellegrini et al. 1999).

Recently, we have developed the statistical framework for a comparison of binary patterns using a variety of distance measures and clustering techniques (G.V.Glazko and ARM, in prep.). A fragment of a neighbor-joining clustering based on a distance derived from Pearson correlation measure, is shown in Fig. 4A. The patterns of Rcl1p and Kre33p appear to be highly correlated with each other and with the pattern for an uncharacterized family of proteins (NCBI COG0585), represented in *E. coli* by the YgbO gene product and in yeast by YOR243 cp.

Initial sequence analysis did not provide convincing clues to YgbO function. However, I noticed statistically insignificant PSI-BLAST matches between the N-terminal regions of the YgbO protein from *E. coli* and different pseudouridine synthases of the TruB family. As pseudouridine synthesis is one of the essential modifications of rRNAs, tRNAs, and snoRNAs, and because the matches contained a conserved aspartic acid residue thought to



A

```
Secondary - YgbO          sssss    ssss            sss     sssssssss      hh
Secondary - 1K8W            ssssss     hhhhhhh       sss         sssssssshhhhhh
YgbO/TruD_Ec_26249151  17 GLLKANPEDFVVVEDLGF  37 REVSFAGQKDKHAVTEQWLCARVPGKEMP 55
YgbO_Cj_15792774       22 AYFSKNSDDFVVRERPLY  37 ADFGYAGLKDKQGSTFQYLSMPKKFESFL 55
YgbO_Mt_15679525       15 GRIRVHNRDFEVEEIPLT  37 GRMGFAGMKDKRAVTRQWICVSNTAPSEV 58
Vng0243c_Hsp_25409280 166 GRLRDSPADFRVRELEAF  43 GRVRWAGTKDRHAVTTQLFAVRDLDAAQV 56
Yor243c_Sc_6324817     57 GQIKQRYTDFLVNEIDQE 173 GVIRYAGTKDRRAVTCQRVSISKIGLDRL 62
YgbO-like_Os_20160919  64 GALKQRYSDFVVHEVALD 107 GSFGFAGTKDKRAVTTQQVTVFKVQASRL 57
1K8W_Ec_18158779       26 VLLLDKPQGMSSNDALQK   9 VRAGHTGALDPLATGMLPICLGEATKFSQ 43
TruB_At_15241405      322 VLLVNKPKGWTSFTVCGK   9 VKVGHAGTLDPMATGLLIVCVGKATKVVD 44
Minifly_Dm_4406198     88 FINLDKPSNPSSHEVVAW   9 FKTGHSGTLDPKVTGCLIVCIDRATRLVK 27
TRUB_BORBU_3915166      4 FLLINKEQGKTSFETLFP   9 FRVGHAGTLDKFASGILVCLVGKYTKLSG 43
TRUB_BACSU_3183559      4 VLLLHKPVGMTSHDCVMK   9 VKVGHTGTLDPEVSGVLPICVGRATKIVE 45
TRUB_DEIRA_13959603     2 VIAADKPLHLTSHDVVNR   9 VRVGHTGTLDPLATGVVVLCVDDSTKLVQ 44
TruB_Ma_20089974       43 VVNIDKPSGPTSHEVAAW   9 VTAGHAGSLDPKVTGLLPTLLGKATKAVP 26
TruB_Sp_19113192       51 LIAINKPSGRTSAQCLNE  38 LKIGHGGTLDPLASGVLVVGLGTGTKQLS 43
PUS4_YEAST_1353109      3 IFAIEKPSGITSNQFMLK  46 IKMGHGGTLDPLASGVLVIGIGAGTKKLA 45
TRUB_SYNY3_ 2498045     3 FLNLHKPLHLTSHDCVAK   9 FRVGHGGTLDPLAEGVLPLAVGSATRLLP 43
Consensus                 xUUxxxPxxUxxx=Uuxx     x+UOxOOxxDxxOxxxxUUxUxxxxxxxx
Active center                                        ***

Secondary - YgbO                        hhhhhhhhhh    hhhhhhh          ss   hhhhhhhhhhhh
Secondary - 1K8W       hhh   ssss           ss     sshhhhhhh          sssssss     hhhhhh
YgbO/TruD_Ec_26249151 GVPNYFGAQR 30  LSAARSALFNQ 2 AERLKKADVNQ 103 FWLPAGSFATSVVRELI
YgbO_Cj_15792774      GFANYFGYQR 29  ISAFQSELFNR 2 SKRVELSHFAN 127 FFLQKGSYATVVLEEIL
YgbO_Mt_15679525      GVPNYYGWQR 96  VHAYQSYLFNR 2 SERAALGINTH 103 FSIPRGCYATSVLREIM
Vng0243c_Hsp_25409280 GTPNYFGQQR 89  VHAAQSYVFNR 2 SERMARGLPFG 114 FALPSGSYATVVLREFL
Yor243c_Sc_6324817    GFINYFGMQR 98  VHAYQSYVWNS 2 SKRIELHGLKL 158 FQLGTSAYATMALRELM
YgbO-like_Os_20160919 GFINYYGLQL 25  VHSYQSYLWNH 2 SMRVQKYGISR 160 FTLPASSYATMAIRELM
1K8W_Ec_18158779      ALDTFRGDIE  1  IPSMYSALKYQ 0 GKKLYEYARQG  22 NELEIHCSKGTYIRTII
TruB_At_15241405      ALTSFLGEIW  1  VPPMFSAIKVG 0 GEKMYEKARRG  26 LIFRVICSKGTYIRSLC
Minifly_Dm_4406198    GLEKLRGALF  1  RPPLISAVKRQ 7 DSKLLDYDETR   2 GVFWVSCEAGSYIRTMC
TRUB_BORBU_3915166    KLKDFVGEIY  1  SPPRFSSVHID 0 GSRAYKLALNG  26 LSLKISCSKGTYIRSIA
TRUB_BACSU_3183559    VLNSLKGKQE  1  IPPMYSAVKVN 0 GKKLYEYARAG  29 FRFTVTCSKGTYVRTLA
TRUB_DEIRA_13959603   VLKGFLGPQQ  1  IPPQYSAIQIG 0 GQRAYAVARAG  56 LLLRAHVGSGTYLRSLA
TruB_Ma_20089974      VCEEFTGPIY  1  MPPIKSAVKRV 7 YIEVLEIEGMS   0 VLFRVGCEAGTYIRKLC
TruB_Sp_19113192      GLDAFRGDIS  1  LPPLYSALHIQ 0 GKRLYEYAREG 120 AVLDMTVSSGFYVRSLI
PUS4_YEAST_1353109    VEEKFVGQLK  1  TPPIYAALKMD 0 GKPLHEYAREG 100 LHFKANVSSGTYIRSLV
TRUB_SYNY3_ 2498045   ILPTFLGEIE  1  IPPQYSAIQVN 0 GKRLYELARAG  27 LDLQVTCGEGTYIRALA
Consensus             xUxxUUGxxx     UxxUxSxUxxx   Ox+Uxxxxxxx     UxUxxxxxOxxUURxUU
Active center                                                      ****
```

B

Fig. 4. YgbO/Pus7 is a pseudouridine synthase related to TruB and functionally linked to Rcl1p and Kre33p/AtAc. **A** Phyletic patterns of Rcl1p, Kre33p, and YgbO families are similar. *Three-letter species abbreviations* are according to the COG database (http://www.ncbi.nlm.nih.gov/COG/new/). **B** Alignment of YgbO and TruB families. *Asterisks* indicate the residues lining the active-site cavity of *E.coli* TruB (pdb 1K8 W), including the principal catalytic aspartate and conserved arginine thought to form a salt bridge with each other (*bold asterisks*). Species abbreviations: *Cj*, *Campylobacter jejeuni*; *Mt*, *Methanothermobacter thermautotrophicus*; *Hsp*, *Halobacterium sp. NRC-1*; Os, *Oryza sativa*; *Ma*, *Methanosarcina acetivorans str. C2A*. Other designations are as in Fig. 2

be involved in catalysis in TruB (see below), I analyzed multiple alignments of the YgbO and TruB families, in an attempt to define all conserved sequence motifs shared by both families. As shown in Fig. 4B, the YgbO family contains all the main motifs recognized in a TruB family (some of which are also shared with other families of pseudouridine synthases; Koonin 1996), and the sequence elements conserved between YgbO and TruB appear to correspond to the set of beta-strands which form the core of the N-terminal catalytic domain in the known three-dimensional structure of TruB (Hoang and Ferre-D'Amare 2001). The C-terminal (presumably RNA-binding) PUA-like domain of TruB enzymes (Anantharaman et al. 2002) is missing from YgbO, but the latter family apparently contains long sequence inserts, some of which may fold independently and play an equivalent role in YgbO interaction with the substrate. Notably, all residues that make contact with the substrate analog 5-fluorouracyl in the TruB co-crystal (Hoang and Ferre-D'Amare 2001) are well conserved in the YgbO family (Fig. 4B). Thus, YgbO enzymes are likely to be pseudouridine synthases distantly related to the TruB family.

When this manuscript was under final revision, two reports showed the pseudouridylate synthase activity of YgbO and Yor243p. Kaya and Ofengand (2003) reported purification of a pseudoridylate synthase responsible for the modification of nucleotide 13 in $tRNA^{Glu}$ in *E.coli*, by a direct biochemical approach unaided by sequence comparison. In another work, Ma et al. (2003) selected the enzyme that modifies position 35 in yeast U2 snoRNA, based on its biochemical activity, from the library of expressed GST fusions of all yeast proteins (Martzen et al. 1999). Neither group reported the similarity to TruB, except for noticing the most conserved acidic residue in the GxKD motif and showing that it is essential for catalysis (Kaya and Ofengand 2003).

Analysis of phyletic patterns and sequence conservation thus indicates that Rcl1p, AtAc/Kre33p, and pseudouridylate synthase now renamed TruD/Pus7, may act in concert, modifying snoRNAs and/or rRNA. The nature of these modifications and their role in ribosome assembly remains to be investigated.

4 Concluding remarks

In this chapter, I surveyed the sequences of yeast proteins that are found in specific complexes associated with ribosome assembly and nuclear export. Several computational methods, i.e., examination of intrinsic sequence features, database searching and analysis of homologous domains, and quantification of similarities between phyletic patterns, can be used together to predict novel molecular functions for proteins that have not been sufficiently studied. While "molecular function" is not the same as the detailed understanding of the biological role of a protein, the list of "what is possible" for a given protein can be dramatically shortened, and further experimentation will be increasingly guided, by the information inferred from computational approaches.

References

Altschul SF, Bundschuh R, Olsen R, Hwa T (2001) The estimation of statistical parameters for local alignment score distributions. Nucleic Acids Res 29:351–361

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Altschul SF, Boguski MS, Gish W, Wootton JC (1994) Issues in searching molecular sequence databases. Nat Genet 6:119–129

Anantharaman V, Koonin EV, Aravind L (2002) Comparative genomics and evolution of proteins involved in RNA metabolism. Nucleic Acids Res 30:1427–1464

Angus-Hill ML, Dutnall RN, Tafrov ST, Sternglanz R, Ramakrishnan V (1999) Crystal structure of the histone acetyltransferase Hpa2: A tetrameric member of the Gcn5-related N-acetyltransferase superfamily. J Mol Biol 294:1311–1325

Bannister AJ, Miska EA, Gorlich D, Kouzarides T (2000) Acetylation of importin-alpha nuclear import factors by CBP/p300. Curr Biol 10:467–470

Bill E, Wegierski T, Nasr F, Filipowicz W (2000) Rcl1p, the yeast protein similar to the RNA 3′-phosphate cyclase, associates with U3 snoRNP and is required for 18S rRNA biogenesis. EMBO J 19:2115–2126

Campbell EA, Muzzin O, Chlenov M, Sun JL, Olson CA, Weinman O, Trester-Zedlitz ML, Darst SA (2002) Structure of the bacterial RNA polymerase promoter specificity σ subunit. Mol Cell 9:527–539

Crain PF, Rozenski J, McCloskey JA (2003) The RNA Modification Database. http://medlib.med.utah.edu/RNAmods/

Dorner BG, Steinbach S, Huser MB, Kroczek RA, Scheffold A. Single-cell analysis of the murine chemokines MIP-1alpha, MIP-1beta, RANTES and ATAC/lymphotactin by flow cytometry (2003) J Immunol Methods 274:83–91

Dragon F, Gallagher JE, Compagnone-Post PA, Mitchell BM, Porwancher KA, Wehner KA, Wormsley S, Settlage RE, Shabanowitz J, Osheim Y, Beyer AL, Hunt DF, Baserga SJ (2002) A large nucleolar U3 ribonucleoprotein required for 18S ribosomal RNA biogenesis. Nature 417:967–970

Fatica A, Dlakic M, Tollervey D (2002) Naf1 p is a box H/ACA snoRNP assembly factor. RNA 8:1502–1514

Fatica A, Tollervey D (2002) Making ribosomes. Curr Opin Cell Biol 14:313–318

Fatica A, Oeffinger M, Dlakic M, Tollervey D (2003) Nob1p is required for cleavage of the 3¢ end of 18S rRNA. Mol Cell Biol 23:1798–1807

Gadal O, Strauss D, Petfalski E, Gleizes PE, Gas N, Tollervey D, Hurt E (2002) Rlp7p is associated with 60S preribosomes, restricted to the granular component of the nucleolus, and required for pre-rRNA processing. J Cell Biol 157:941–951

Ginalski K, Elofsson A, Fischer D, Rychlewski L (2003) 3D-Jury: a simple approach to improve protein structure predictions. Bioinformatics 19:1015–1018

Grandi P, Rybin V, Bassler J, Petfalski E, Strauss D, Marzioch M, Schafer T, Kuster B, Tschochner H, Tollervey D, Gavin AC, Hurt E (2002) 90S pre-ribosomes include the 35S pre-RNA, the U3 snoRNP, and 40S subunit processing factors but predominantly lack 60S synthesis factors. Mol Cell 10:105–115

Granneman S, Gallagher JE, Vogelzangs J, Horstman W, van Venrooij WJ, Baserga SJ, Pruijn GJ (2003) The human Imp3 and Imp4 proteins form a ternary complex with hMpp10, which only interacts with the U3 snoRNA in 60–80S ribonucleoprotein complexes. Nucleic Acids Res 31:1877–18787

Hoang C, Ferre-D'Amare AR (2001) Cocrystal structure of a tRNA Ψ55 pseudouridine synthase: nucleotide flipping by an RNA-modifying enzyme. Cell 107:929–939

Huang JY, Brutlag DL (2001) The EMOTIF database. Nucleic Acids Res 29:202–204

Kaya Y, Ofengand J (2003) A novel unanticipated type of pseudouridine synthase with homologs in Bacteria, Archaea, and Eukarya. RNA 9:711–721

Kimura Y, Saeki Y, Yokosawa H, Polevoda B, Sherman F, Hirano H (2003) N-Terminal modifications of the 19S regulatory particle subunits of the yeast proteasome. Arch Biochem Biophys 409:341–348

Koonin EV, Wolf YI, Aravind L (2001) Prediction of the archaeal exosome and its connections with the proteasome and the translation and transcription machineries by a comparative-genomic approach. Genome Res 11:240–252

Koonin EV, Mushegian AR, Galperin MY, Walker DR (1997) Comparison of archaeal and bacterial genomes: computer analysis of protein sequences predicts novel functions and suggests a chimeric origin for the archaea. Mol Microbiol 25:619–637

Koonin EV, Mushegian AR, Bork P (1996) Non-orthologous gene displacement. Trends Genet 12:334–336

Koonin EV (1996) Pseudouridine synthases: four families of enzymes containing a putative uridine-binding motif also conserved in dUTPases and dCTP deaminases. Nucleic Acids Res 24:2411–2415

Lee SW, Berger SJ, Martinovic S, Pasa-Tolic L, Anderson GA, Shen Y, Zhao R, Smith RD (2002) Direct mass spectrometric analysis of intact proteins of the yeast large ribosomal subunit using capillary LC/FTICR. Proc Natl Acad Sci USA 99:5942–5947

Leipe DD, Wolf YI, Koonin EV, Aravind L (2002) Classification and evolution of P-loop GTPases and related ATPases. J Mol Biol 317:41–72

Lesk AM (1995) NAD-binding domains of dehydrogenases. Curr Opin Struct Biol 5:775–783

Letunic I, Goodstadt L, Dickens NJ, Doerks T, Schultz J, Mott R, Ciccarelli F, Copley RR, Ponting CP, Bork P (2002) Recent improvements to the SMART domain-based sequence annotation resource. Nucleic Acids Res 30:242–244

Louie DF, Resing KA, Lewis TS, Ahn NG (1996) Mass spectrometric analysis of 40S ribosomal proteins from Rat-1 fibroblasts. J Biol Chem 271:28189–28198

Lupas A (1996a) Prediction and analysis of coiled-coil structures. Methods Enzymol 266:513–525

Lupas AN (1996b) Coiled coils: new structures and new functions. Trends Biochem Sci 21:375–382

Ma X, Zhao X, Yu YT (2003) Pseudouridylation (Ψ)of U2 snRNA in S.cerevisiae is catalyzed by an RNA-independent mechanism. EMBO J 22:1889–1897

Martzen MR, McCraith SM, Spinelli SL, Torres FM, Fields S, Grayhack EJ, Phizicky EM (1999) A biochemical genomics approach for identifying genes by the activity of their products. Science 286:1153–1155

McGuffin LJ, Bryson K, Jones DT (2000) The PSIPRED protein structure prediction server. Bioinformatics 16:404–405

Mirkin BG, Fenner TI, Galperin MY, Koonin EV (2003) Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. BMC Evol Biol 3:2 (Epub 2003 Jan 06)

Moore PB, Steitz TA (2002) The involvement of RNA in ribosome function. Nature 418:229–235

Mushegian AR, Bassett DE Jr, Boguski MS, Bork P, Koonin EV (1997) Positionally cloned human disease genes: patterns of evolutionary conservation and functional motifs. Proc Natl Acad Sci USA 94:5831–5836

Nissan TA, Bassler J, Petfalski E, Tollervey D, Hurt E (2002) 60S pre-ribosome formation viewed from assembly in the nucleolus until export to the cytoplasm. EMBO J 21:5539–5547

Oeffinger M, Leung A, Lamond A, Tollervey D, Lueng A. (2002) Yeast Pescadillo is required for multiple activities during 60S ribosomal subunit synthesis. RNA 8:626–636

Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci USA 96:4285–4288

Schafer T, Strauss D, Petfalski E, Tollervey D, Hurt E (2003) The path from nucleolar 90S to cytoplasmic 40S pre-ribosomes. EMBO J 22:1370–1380

Schaffer AA, Aravind L, Madden TL, Shavirin S, Spouge JL, Wolf YI, Koonin EV, Altschul SF (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res 29:2994–3005

Schaffer AA, Wolf YI, Ponting CP, Koonin EV, Aravind L, Altschul SF (1999) IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. Bioinformatics. 15:1000–1011

Sonnhammer EL, Koonin EV (2002) Orthology, paralogy and proposed classification for paralog subtypes. Trends Genet 18:619–620

Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res 2:22–28

Tatusov RL, Koonin EV, Lipman DJ (1997) A genomic perspective on protein families. Science 278:631–637

Walker DR, Koonin EV (1997) SEALS: a system for easy analysis of lots of sequences. Proc Int Conf Intell Syst Mol Biol 5:333–339

Wan H, Li L, Federhen S, Wootton JC (2003) Discovering simple regions in biological sequences associated with scoring schemes. J Comput Biol 10:171–185

Wehner KA, Baserga SJ (2002) The σ^{70}-like motif: a eukaryotic RNA binding domain unique to a superfamily of proteins required for ribosome biogenesis. Mol Cell 9:329–339

Wehner KA, Gallagher JE, Baserga SJ (2002) Components of an interdependent unit within the SSU processome regulate and mediate its activity. Mol Cell Biol 22:7258–7267

Wootton JC, Federhen S (1996) Analysis of compositionally biased regions in sequence databases. Methods Enzymol 266:554–571

Bioinformatics-Guided Experimental Characterization of Mismatch-Repair Enzymes and Their Relatives

P. Friedhoff

1 Introduction

The increasing information in databases of protein sequences and structures together with the development of bioinformatic tools has helped the biochemists to identify and validate the function of many different proteins. In this chapter, we will show the successful application of two methods, protein-fold recognition (FR) and evolutionary-trace (ET) analysis to learn about the function of a group of proteins which belong to the class restriction endonucleases, namely the type II restriction endonuclease (REase) Sau3AI and the mismatch repair (MMR) protein MutH.

The first method, fold recognition, makes use of sequence information to predict the secondary structure, the topology, and finally the tertiary structure of a protein. These methods are of great value for many purposes in modern biology since the available sequence information has by far exceeded the available structural information and the knowledge about the fold of a given protein is an important step towards the understanding of its function. Here, we applied several fold-recognition programs and the consensus server Pcons available via "metaservers" (Bujnicki et al. 2001; Kurowski and Bujnicki 2003) to predict the structure of the C-terminal domain (CTD) of Sau3AI.

The second method, the evolutionary-trace analysis, makes use of phylogenetic and structural information of a protein family in order to identify functional sites in proteins. This method has been successfully used to predict the functional sites in a variety of different proteins (review: Lichtarge et al. 2002). Here, we use the evolutionary-trace method to identify amino acid residues in MutH, which are involved in sensing the methylation status of its recognition sequence.

P. Friedhoff
Institut für Biochemie, FB 08, Justus-Liebig Universität Giessen, Heinrich-Buff-Ring 58, 39392 Giessen, Germany

Nucleic Acids and Molecular Biology, Vol. 15
Janusz M. Bujnicki (Ed.)
Practical Bioinformatics

In both cases, the bioinformatic analysis guided the biophysical and biochemical analysis of the proteins, whose results in turn confirmed that the predictions were highly significant.

1.1 Sau3AI and Related Restriction Endonucleases

Sau3AI belongs to the family of type II restriction endonucleases. In the presence of Mg^{2+}, Sau3AI cleaves double-stranded DNA (recognition sequence: /GATC), producing sticky ends with four nucleotide 5′-overhangs. Its partner enzyme, the Sau3AI DNA methyltransferase (MTase) protects the hosts DNA via methylation of the recognition sequence to give C5-methylation (m^5C) at position 4, thereby preventing cleavage by Sau3AI (Seeber et al. 1990). Moreover, Sau3AI is also inhibited by N4-methylcytosine (m^4C) at position 4 but not by N6-methyladenine (m^6A) at position 2.

"Orthodox" type II REases recognize short palindromic sequences, 4 to 8 base pairs in length, and cut both strands and have recently been classified into several subgroups (reviews: Pingoud and Jeltsch 2001; Roberts et al. 2003). Most of these enzymes, now classified as type IIP enzymes, have a symmetric recognition and cleavage site. They often act as homodimers, thereby forming one DNA binding site that contains two catalytic centers allowing the simultaneous cleavage of both strands, e.g. EcoRI or EcoRV. One variation to this scheme is the type IIF REases, which tetramerize, thereby forming two DNA binding sites. In order to be fully active both DNA binding sites most be occupied (Embleton et al. 2001). On the other hand, types IIE REases are two-domain proteins, which upon dimerization will form two DNA binding sites. Similar to type IIF REases, both DNA binding sites of type IIE enzymes have to be occupied to achieve maximum activity. However, in contrast to type IIF enzymes, only one DNA binding site of a IIE REase contains a catalytically active site while the other is an activator site, e.g., in NaeI or EcoRII (Huai et al. 2001; Mucke et al. 2002).

While Sau3AI has been used widely as a tool by molecular biologists, rather little information regarding the biochemistry of this enzyme was known. Sau3AI contains 449 amino acid residues and, therefore, has twice the size of the subunit of a typical type IIP REase. Therefore, the enzyme might belong to one of the various subtypes of the type II class of restriction endonucleases. Special interest on Sau3AI was raised when the structure of the DNA mismatch repair endonuclease MutH was solved and the sequence similarity between the N-terminal domain of Sau3AI and MutH became apparent (Ban and Yang 1998).

1.2 DNA Mismatch Repair

DNA mismatch repair is one of several repair pathways to ensure the stability of the genome of most organisms. The difference between mismatch repair and most other repair mechanisms is that the mismatch repair machinery corrects sequence information errors, i.e., base–base mismatches or insertion/deletions, while the other repair mechanism corrects changes in the structure of the DNA, e.g. strand breaks or alkylation of bases. The main but not exclusive function of DNA mismatch repair is the correction of errors of the replicative DNA polymerases, which escaped their built-in proofreading function. Therefore, the mismatch repair system has two tasks to fulfill: (1) recognition of the replication error and (2) identification of the erroneous strand, which by definition is the newly synthesized daughter strand (Modrich and Lahue 1996).

The paradigm for DNA mismatch repair is the methylation-directed MutHLS system of *Escherichia coli* and related proteobacteria. The system has been extensively studied both in vivo and in vitro and has been reconstituted in vitro from purified compounds by Modrich and coworkers (Lahue et al. 1989). Moreover, the crystal structures for the key players, MutS, MutL and MutH, have been solved recently (Ban and Yang 1998; Ban et al. 1999; Lamers et al. 2000).

Mismatch repair is initiated by recognition of the mismatch by MutS. MutL then acts as a "molecular matchmaker" and helps to recruit other components involved in this process. The first protein recruited is the sequence-specific endonuclease MutH, which after activation by MutS and MutL nicks the unmethylated erroneous DNA strand at hemimethylated *dam* sites (GATC), which can be up to 1000 base pairs away from the mismatch (Modrich and Lahue 1996). Thereby, strand discrimination is achieved and repair is directed to the newly synthesised daughter strand. In vitro reconstitution experiments have shown that the excision re-synthesis steps can occur bidirectionally by involving the action of DNA helicase II, exonucleases (ExoI or ExoX for 3′-5′-removal; ExoVII or RecJ for 5′-3′-removal, single-strand DNA binding protein, DNA polymerase III holoenzyme and DNA ligase (Modrich 1991; Au et al. 1992; Cooper et al. 1993).

Several mechanisms for the communication between the mismatch recognition and the strand discrimination step have been discussed in the literature. These models all have one thing in common, a physical interaction between MutS and MutL on the one hand, and MutL and MutH, on the other. The three most frequently discussed models include an ATP-dependent DNA translocation model (Allen et al. 1997); the molecular switch or sliding clamp model (Gradia et al. 1997) and the DNA looping model (Junop et al. 2001).

Although the structures of the MutHLS proteins are available, it is not known how MutH is activated by MutL, or how MutH discriminates between the unmethylated and methylated DNA strands, since the structure of MutH was solved in the absence of its cognate DNA substrate.

1.3 Nicking Endonuclease MutH

E. coli MutH is a 229 amino acid monomeric endonuclease, which in the presence of Mg^{2+} nicks DNA strands (recognition sequence: /GATC;Welsh et al. 1987). MutH only nicks at GATC when the adenine is not methylated. The natural and preferred substrates of MutH are hemimethylated GATC sites with m^6A on the (parental) strand, which is not cleaved by MutH. DNA fully methylated (m^6A) at GATC) is not cleaved at all. However, the endonuclease activity of MutH is low (turnover number $<1\ h^{-1}$) but gets stimulated 20- to 50-fold in a mismatch-independent manner by MutL or in a mismatch-dependent manner by MutS and MutL (Ban et al. 1999; Hall and Matson 1999). The mechanism of strand discrimination by MutH seems to be limited to a subset of ɣ-proteobacteria, since close homologues of MutH have not been found outside these taxa (Eisen and Hanawalt 1999). Hence, the mechanisms for strand discrimination must be different in most other Bacteria, Archaea and Eukarya.

Sequence comparisons revealed that MutH shows sequence similarity to Sau3AI (Ban and Yang 1998). In addition to the sequence similarity, MutH and Sau3AI have the cleavage of the same recognition sequence at the same position, i.e. /GATC, in a Mg^{2+}-dependent manner in common (Ban and Yang 1998). Both enzymes are inhibited by m4C and m5C methylation at position 4 (Friedhoff, unpubl. results). However, whereas MutH only nicks DNA in unmethylated (m^6A) GATC sites, Sau3AI makes a double-strand break regardless of the methylation status at the adenine at position 2. Moreover, Sau3AI is almost twice the size of MutH and does not require activation by additional factors. The question, therefore, was raised, whether these additional residues have a function in controlling the activity of Sau3AI similar to how MutL activates MutH (Ban and Yang 1998).

The structure of the *E. coli* MutH protein was the first structure of the MutHLS system solved (Ban and Yang 1998) and revealed the structural similarity to restriction endonucleases. As might be expected from the function, the crystallographic analysis suggested a monomeric structure for MutH. Moreover, three structural variations of MutH have been observed (pdb codes: 1azo, 2azo_A and 2azo_B). One of these (2azo_B) is believed to be the catalytic competent form of MutH, where the active site has a similar geometry as the closest structural relative PvuII. The three MutH structures differ in terms of the relative orientation of the two subdomains (Yang 2000) and the order of several loop residues. At least six highly conserved residues Lys48, Glu56, Asp70, Glu77, Lys79 and Lys116 have been reported to be crucial for enzymatic activity, since the individual alanine mutants are almost devoid of catalytic activity while still being able to bind to the substrate (Yang 2000; Loh et al. 2001; Friedhoff et al. 2002). By structural similarity the DNA binding cleft could be assigned but no details on how the monomeric MutH recognizes the DNA and discriminates unmethylated DNA from methylated one were reported.

A mechanism for activation might be the stabilization of the active form of MutH upon binding to activator protein MutL. The protruding C-terminal α-helix F and the following hydrophobic have been suggested to act as a molecular lever for MutS and MutL to activate MutH (Ban and Yang 1998). Moreover, the binding site for MutL was mapped experimentally to a region around the end of α-helix E (Toedt et al. 2003).

2 Sau3AI – Similar Folds for N- and C-Terminal Domains

In order to get an insight into the function of the C-terminal domain (CTD) of Sau3AI we made use of fold-recognition programs by using the power of metaservers, which combine and judge the results of several fold-recognition programs (Bujnicki et al. 2001; Ginalski et al. 2003; Kurowski and Bujnicki 2003).

2.1 Fold Recognition for the C-Terminal of Sau3AI

MutH-related sequences were identified using a variety of BLAST and PSI-BLAST searches (Altschul et al. 1997) of a non-redundant (nr) database and publicly available nucleotide sequences from both complete and unfinished genome projects at NCBI (http://www. ncbi.nlm.nih.gov). BLAST searches of individual genome sequences were performed using the GOLD Genomes OnLine Database using sequences of *E. coli* MutH and Sau3AI as queries. Multiple alignments were extracted from the BLAST output with the BIBVIEW software (http://bioinfo.pl/bibview.pl) and were corrected manually, taking into account preservation of the continuity of the observed secondary structural elements.

In a search for proteins with a similar fold of the C-terminal domain of Sau3AI, we submitted the individual sequences for the seven C-terminal domains to the Structure Prediction Meta Server (http://bioinfo.pl/meta/; (Bujnicki et al. 2001; Ginalski et al. 2003) or the alignment of the C-terminal domains of the seven REases to the Fold Prediction Metaserver at Genesilico (http://genesilico.pl/meta) (Kurowski and Bujnicki 2003). Both metaservers use several fold-recognition programs (links to the individual structure prediction servers are provided on the above-mentioned websites).

The best hits by far, from the metaserver for all seven REases was the structure of MutH (pdb-code 1azo or 2azo), suggesting that both the N-terminal and C-terminal domain of the REases adopt a similar fold as MutH (Table 1). Moreover, when the sequence alignment of all seven C-terminal domains was submitted, which is now possible with the Fold Prediction Metaserver at GeneSilico, the PDB-BLAST had already resulted in a significant hit, namely 1azo with a score of 3e-86.

Table 1. Results of the Structure Prediction Meta Server

Protein	Pcons2	PDB code	3D Jury	PDB code
Sau3AI	2.12	1azo	64.67	1azo
LlaKR2I	3.07	1azo	79.00	2azo_A
Sth368I	3.88	1azo	84.25	2azo_A
RE_Spn	4.17	1azo	93.00	1azo
RE_Blo	3.12	1azo	81.40	1azo
RE_Cpe	4.48	1azo	95.00	2azo_A
RE_Lmo	3.15	1azo	85.86	1azo

Scores of >1.5 or 50 for Pcons2 or 3D Jury, respectively, are regarded as significant.

These results suggest that both the N- and the C-terminal domains of Sau3AI adopt a MutH-like fold, which has several implications for the quaternary structure of the enzyme. For instance, this result may suggest that Sau3AI is a pseudodimer, i.e., the DNA binding site is formed by the two subdomains from a single polypeptide chain similar to the one which has been observed for other DNA cleaving enzymes, e.g., the homing endonucleases PI-SceI (Christ et al. 1999). Another possibility is that the protein forms a dimer with two distinct DNA binding sites formed by the N and the C-terminal domain, respectively. This has been shown for the type IIE REases NaeI and EcoRII (Huai et al. 2001; Mucke et al. 2002; Zhou et al. 2003), though the DNA binding domains of these enzymes have different folds.

To learn more about the function of the C-terminal domain of Sau3AI we performed a protein sequence alignment based on the fold-recognition predictions with the N-and C-terminal domains of the REases and MutH proteins. MutH protein sequences and the N-terminal domains of the REases were aligned using the ClustalX program (Thompson et al. 1997). The C-terminal domains of the REases were aligned separately and thereafter aligned to the former alignment guided by the results of the prediction servers and manually refined using the program BioEdit (Hall 1999,Fig. 1). Moreover, a phylogenetic and molecular evolutionary analyses conducted using MEGA version 2.1 (Kumar et al. 2001) indicated that the C-terminal domains of the REases are more distantly related to the MutH sequences than to the N-terminal domains (Fig. 2). A similar result of a phylogenetic study was reported by Bujnicki (2001).

One of the important questions was, whether the C-terminal domain will also be able to bind and cleave the DNA. Catalytic important residues in MutH (E56, D70, E77, K79 and K116) have been identified by structural similarity (Ban and Yang 1998) and verified by mutational analyses (Friedhoff et al. 2002; Wu et al. 2002; Junop et al. 2003). The sequence analysis showed that the C-terminal domains lack the characteristic active site residues of the PD-(D/E)XK motif (Fig. 1). Moreover, by mapping the sequence conservation

onto the structure of MutH we noticed that most conserved residues were located in the protein core rather than on the protein surface, with the exception of a few residues probably involved in DNA binding and recognition. These residues included K48 and E91 of MutH, which are involved in DNA binding and recognition as revealed by mutational analyses (Friedhoff et al. 2002; Wu et al. 2002).

To validate our bioinformatic analysis we analyzed the biochemical and biophysical properties of Sau3AI in more detail.

2.2 Biochemical and Biophysical Analysis – Evidence for a Pseuotetramer That Induces DNA Looping

Since efforts to clone the entire R-M system of either Sau3AI or LlaKR2I in a variety of *E. coli* strains failed (Seeber et al. 1990; Twomey et al. 1998), we were not able to perform a mutational analysis on Sau3AI. However, we could show by gel filtration and sedimentation analysis that Sau3AI in the absence of DNA is a monomer (Friedhoff et al. 2001). This might suggest that Sau3AI forms a pseudodimer, however, it does not rule out that Sau3AI could dimerize upon binding to DNA. Similar results have been observed for type IIS REase that presumably are monomeric in the absence of DNA but dimerize in the presence of DNA, as shown for FokI (Bitinaite et al. 1998; Wah et al. 1998).

Next, we addressed the question whether Sau3AI contains a single DNA binding site with two catalytic centers – as is the case for most type II restriction endonucleases due to their homodimerization – competent to perform a double-strand break in one binding event. The results of a DNA cleavage analysis in which the conversion of plasmid DNA from the supercoiled form to either the open circular form (by nicking) or to the linear form (by double-strand breakage) was monitored, and showed no indication for an accumulation of the open circular form and thus no indication for DNA nicking (Friedhoff et al. 2001). Hence, we concluded that Sau3AI contains one DNA binding site with two catalytic centers.

However, these results did not answer the question regarding the function of the additional C-terminal domain of Sau3AI. It was known from type IIE REases that an additional DNA binding domain without catalytic activity can act as an activator thereby regulating the activity of the catalytic domain, as shown for example for NaeI, EcoRII or FokI (Bitinaite et al. 1998; Embleton et al. 2001; Mucke et al. 2002). Consequently, we addressed this issue by analyzing the cleavage of DNA substrates containing either one or two GATC sites. The outcome of this analysis revealed that a substrate with two recognition sites was cleaved significantly faster than the substrate with a single site (Friedhoff et al. 2001). Moreover, the two sites were cleaved one at a time similar to that observed for type IIE REase, e.g., NaeI or EcoRII, which contains two DNA binding sites, one of which is the catalytic site and one of which is the activator site. In addition it was shown

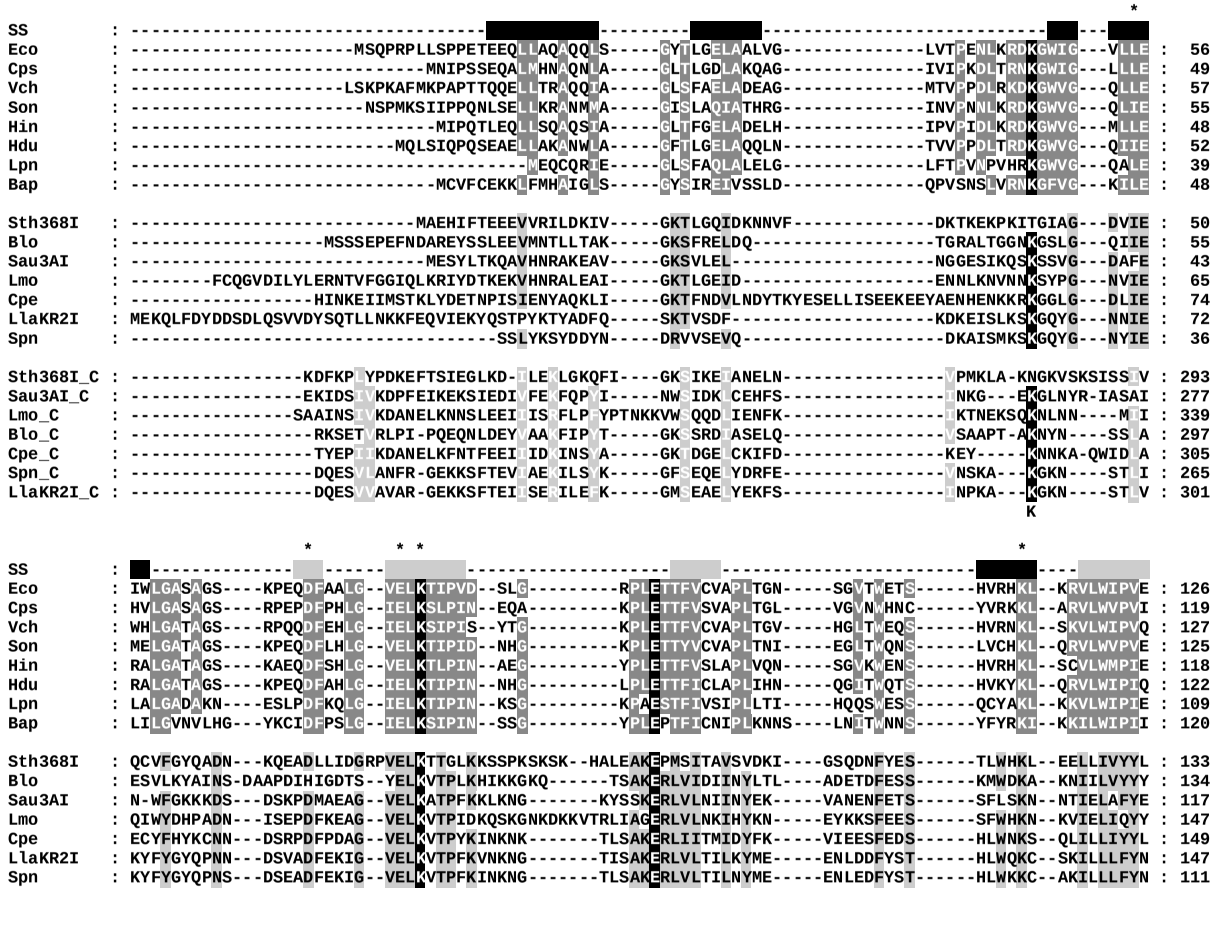
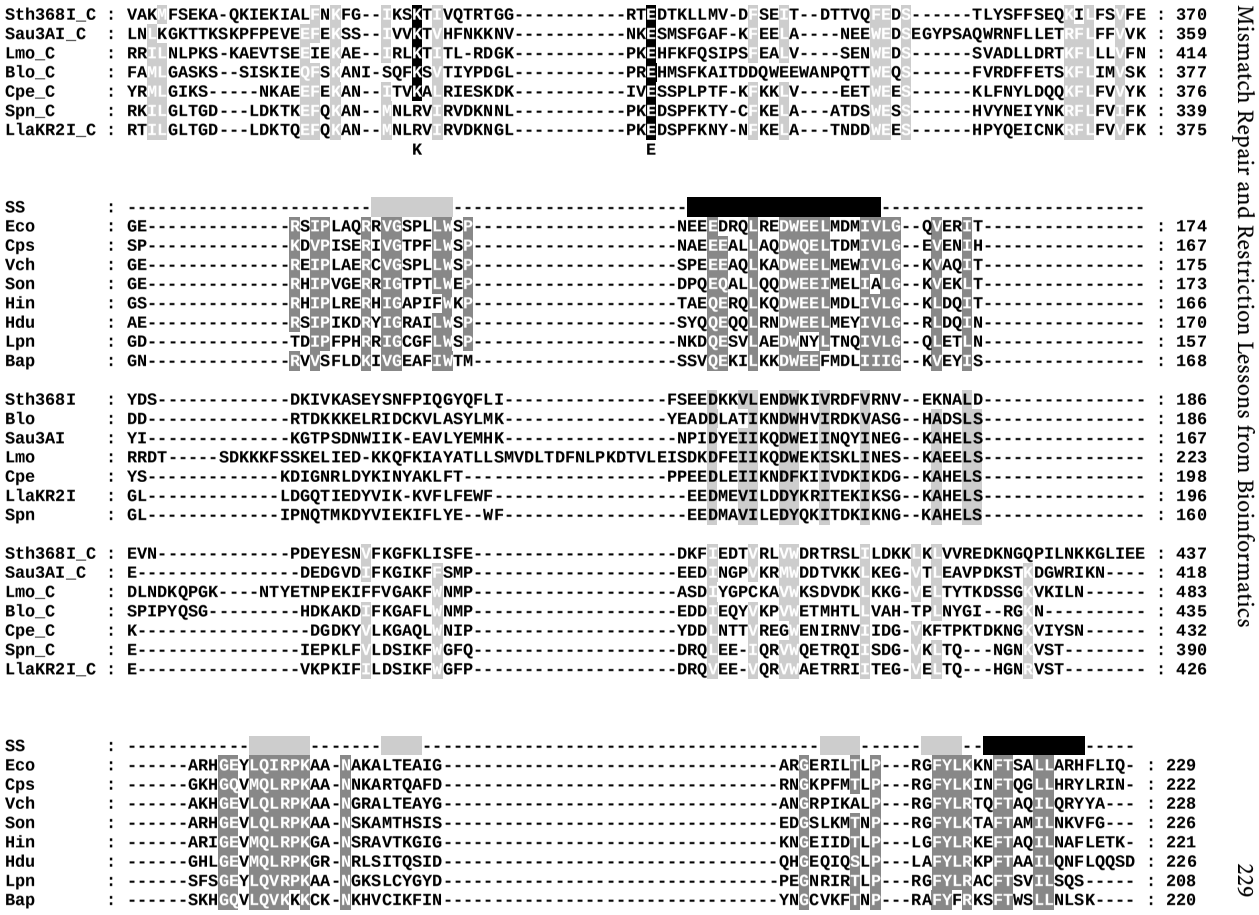


Fig. 1. Sequence alignment of MutH and related REases. Protein sequence alignment of the MutH proteins, the N-terminals domains and the C-terminal domains, respectively, of REases related to Sau3AI. The secondary structure elements of *E. coli* MutH are indicated also. α-Helices and β-strands are indicated by *black* and *gray bars*, respectively. Residues conserved in all groups are shown in *black* with *white lettering*. Residues conserved within each group are shown in *gray* with *white lettering*, *dark gray* or *light gray*, respectively. The catalytic important residues are marked by *asterisks*. The following abbreviations were used: MutH proteins: *Bpa*, *Buchnera aphidi-*

colaK; *Cps*, *Colwellia psychroerythraea* 34H; *Eco*, *Escherichia coli* K12- MG1655; *Hdu*, *Haemophilus ducreyi* 35000HP; *Hin*, *Haemophilus influenzae* KW20; *Lpn*, *Legionella pneumophila* Philadelphia-1; *Son*, *Shewanella oneidensis* (putrefaciens) MR-1 ATCC700550; *Vch*, *Vibrio cholerae* serotype O1, biotype El Tor, strain N16961; REases: *Blo*, *Bifidobacterium longum* NCC2705; *Cpe*, *Clostridium perfringens* ATCC 13124; *Lmo*, *Listeria monocytogenes* 4b



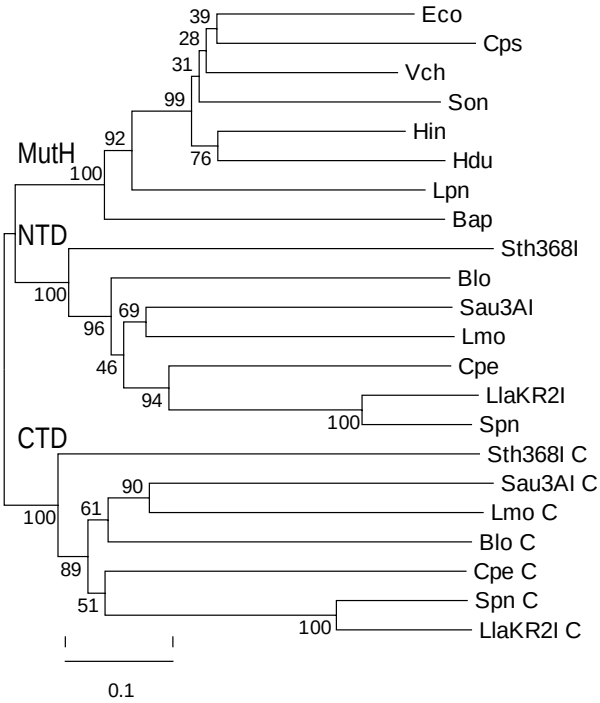


Fig. 2. Phylogenetic tree of MutH and REases (N- and C-terminal domains). The phylogenetic tree of MutH, the N-terminal domain (NTD) and C-terminal domain (CTD) of REases was constructed with the neighbor-joining method (Saitou and Nei 1987) using the alignment shown in Fig. 1 after removing some regions with large gaps. The *numbers* at the *nodes* correspond to the statistical support of the branching order by the bootstrap criterion. The *bar below* the phylogram indicates the evolutionary distance to which the branch lengths are scaled based on the estimated divergences

for Sau3AI that the cleavage of long DNA substrates containing a single GATC site is stimulated rather than inhibited by the addition of small oligonucleotides containing a GATC site (Hermann and Jeltsch 2003).

The enzymological analysis of Sau3AI was corroborated by an electron microscopy study using the same DNA substrates as was used for the cleavage analysis. This analysis clearly demonstrated the ability of Sau3AI to introduce DNA loops on substrates containing two GATC sites (Fig. 3), indicating the presence of two DNA binding sites in the enzyme.

These results taken together suggest that Sau3AI forms a dimer upon binding to DNA. This dimer contains a DNA binding site formed from the N-terminal domain each of which contains a catalytically competent active site that allows the enzyme to perform a double-strand break in a single DNA binding event. The second DNA binding site is formed from two C-terminal domains that lack a catalytically active site and thereby do not cleave the DNA upon binding. Based on these conclusions a model of quaternary structure of Sau3AI complexed with the target DNA was generated by Bujnicki (2001; Fig. 4). These results have implications for the evolution, structure and function of bacterial DNA repair of enzymes and restriction endonucleases.

The exact function of the additional C-terminal DNA binding domains remains elusive although it has been speculated for other type II REase, e.g.,

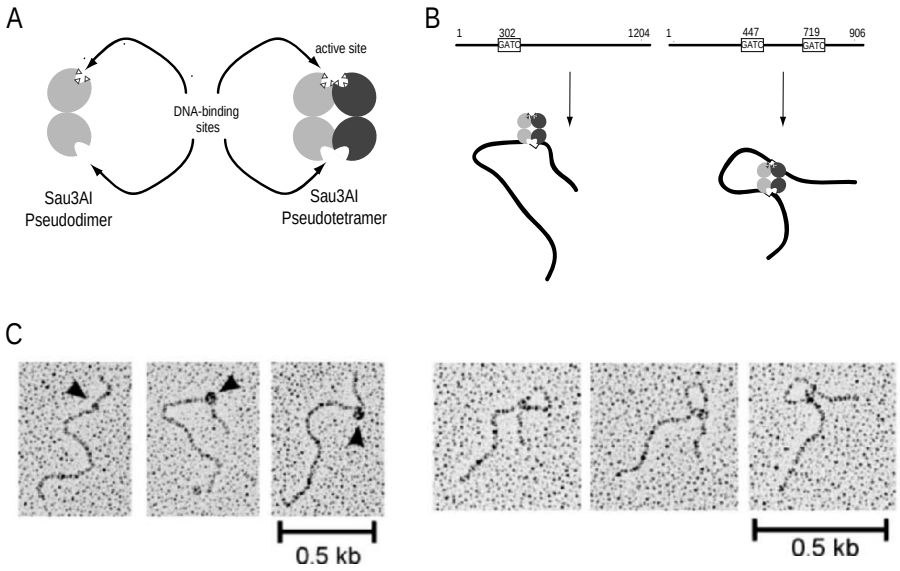


Fig. 3. Looping of DNA with two GATC sites by Sau3AI. **A** Schematic diagram of the pseudodimer of Sau3AI (consisting of two similar domains) showing the two DNA binding half sites one of which contains the catalytic amino acid active residues (indicated by *three triangles*). Upon dimerization, the pseuotetramer is formed containing two DNA binding sites. The active site is formed by the two N-terminals domains while the second DNA binding site is formed by the C-terminal domains. **B** Schematic representation of Sau3AI binding to DNA containing a single or two GATC sites. When two binding sites are present, DNA looping is possible. **C** Electron microscope analysis of Sau3AI binding to two different DNAs as shown in **B** revealing the formation of DNA loops in the substrate with two DNA binding sites

the type IIE and typeIIF enzymes that the second DNA binding site could increase the accuracy of these enzymes (Embleton et al. 2001). The function of the C-terminal domain of Sau3AI is reminiscent of the function of MutL, which also can bind to DNA (Bende and Grafstrom 1991; Ban et al. 1999). By having an additional DNA binding site, the affinity of Sau3AI for DNA has probably increased.

One of the functions of MutL could be the formation of a MutL-MutH complex that has higher DNA binding affinity compared to MutH, which in turn would result in stimulating the activity of MutH. Indeed, this has been observed although the stimulation of the catalytic activity was shown to be higher compared to the stimulation of DNA binding affinity (Loh et al. 2001). Another possibility is the stabilization of the catalytic competent form of MutH by MutL upon complex formation. In case of Sau3AI, binding of DNA to the C-terminal domain might also lead to a stabilization of the catalytic active form of the N-terminal domain.

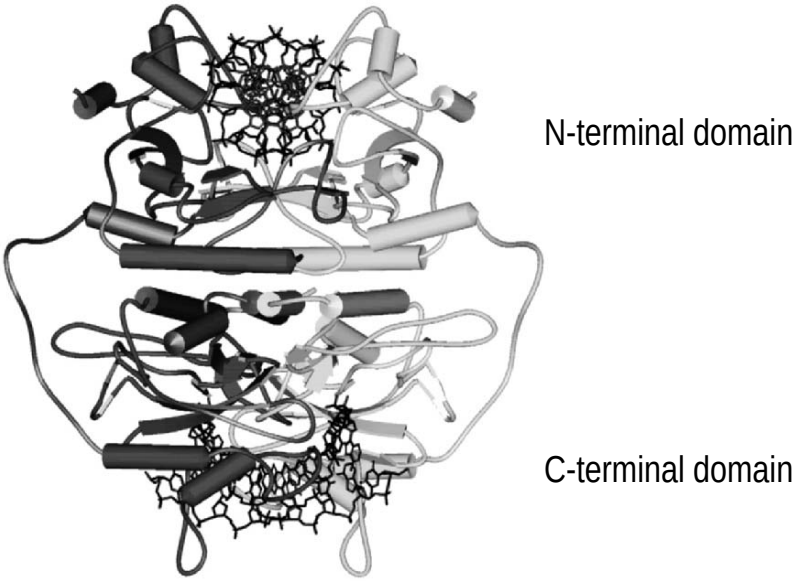


Fig. 4. Model of the Sau3AI pseudotetramer. Homology modeling of the Sau3AI structure was performed by J.M. Bujnicki 2001). The two subunits of Sau3AI are shown in a schematic drawing in *dark* and *light gray*. The two DNAs bound to the binding sites formed by the N- and C-terminal domains, respectively, are shown as *sticks*

In summary, our analysis showed that Sau3AI and its relatives belong to the type IIE REases adding one more facet to this class by having two subdomains with a similar but not identical fold.

3 Identification of the Methylation Sensor of MutH

One of the important differences between MutH and Sau3AI is their different sensitivity towards m^6A methylation of the GATC recognition sequence. While Sau3AI cleaves DNA substrates regardless of the methylation status at the adenine, MutH only cleaves unmethylated DNA strands with a preference for the unmethylated DNA strand in a hemimethylated (its natural) substrate over fully unmethylated DNA. Using the above mentioned sequence comparison between the family of MutH proteins and the family of Sau3AI related REases (Fig. 1) we wondered whether the biochemical difference between these two families would also be reflected as differences in the sequence. This would allow for making a prediction regarding the function of several amino acid residues in MutH. Since the preferential cleavage of unmethylated GATC sites in hemimethylated DNA is crucial for the in vivo function of MutH in mismatch repair, elucidation of the mechanism underlying the strand discrimination is an important issue.

3.1 Evolutionary Trace Analysis

Functional important residues can be predicted by the evolutionary trace (ET) analysis (Lichtarge et al. 1996) or related methods such as ConSurf (Armon et al. 2001; Pupko et al. 2002; Glaser et al. 2003). These analyses rely on the presence of a family of proteins having sequence similarity and fall into distinct classes. The evolutionary trace analysis is a method of identifying functional residues in a protein sequence by looking for conserved residues in the branches of an evolutionary tree (Lichtarge and Sowa 2002).

It has been shown for several proteins that functional differences could be correlated with differences in the sequence by the ET method. Moreover, in combination with protein structures this evolutionary information can be mapped onto the structure, thereby increasing the likelihood of identifying a functional epitope. We therefore started such an analysis that became possible when more sequence homologues of both MutH and Sau3AI became available. However, an evolutionary trace analysis using the Evolutionary Trace Server (TraceSuite II) (Lichtarge et al. 1996; Innis et al. 2000) did not gave us a significant result.

Hence, we decided to perform a modification of the evolutionary trace analysis using the program GeneDoc and its built-in function for defining groups of sequences (Nicholas et al. 1997). We included only MutH sequences in this analysis with a maximum pairwise sequence identity of 60 %. Moreover, we removed the Sth368I sequence that was the most divergent sequence among the Sau3AI family members (Fig. 2). Thus, we ended up with a set of eight MutH and five Sau3AI-related protein sequences. In contrast to the original ET method, we constructed the consensus sequence at 80 % sequence identity. Residues were then regarded as conserved when they were identical in both consensus sequences, as class specific when they were different in both consensus sequences or as neutral, when they were only conserved in one consensus sequence. Thus we identified 16 conserved and 24 class specific residues, i.e., residues conserved in both groups but not identical between the groups, (Fig. 5). Finally, we mapped this evolutionary trace to the structure of MutH. The result of this analysis is given in Fig. 6.

The alignment of the amino acid sequence of these proteins revealed that these nucleases share only a limited number of conserved amino acid residues, which presumably are involved in common functions, viz. DNA binding, recognition and cleavage or folding. Some of the class specific trace residues are right next to the active site residues and, therefore, are likely candidates for being involved in DNA recognition and sensing the methylation status of the GATC recognition sequence. The most prominent residues were Phe94, Arg184 and Tyr212 (Fig. 6).

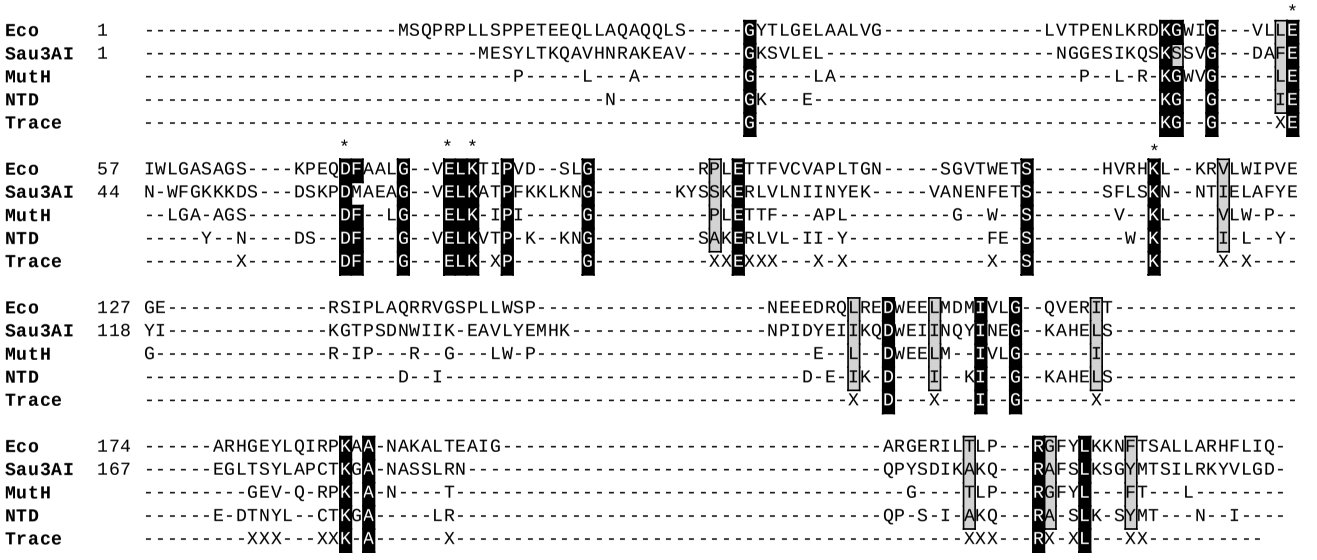


Fig. 5. Evolutionary trace analysis of MutH and REases. The sequences of *E. coli* MutH (Eco) and Sau3AI aligned as in Fig. 1 are shown. Catalytically important residues have been marked with an *asterisk* above the sequences. The consensus sequences for the MutH family (*MutH*) and N-terminal domains of the REases (*NTD*) were obtained at 65% identity of the families (see text for details). The evolutionary trace (*Trace*) is shown in the *bottom line*. Conserved residues are indicated explicitly and *shaded* in *black* with *white lettering*, whereas class specific residues are indicated by an *X*, and *shaded* in *gray* whenever in both consensus sequences different but similar conserved residues were observed

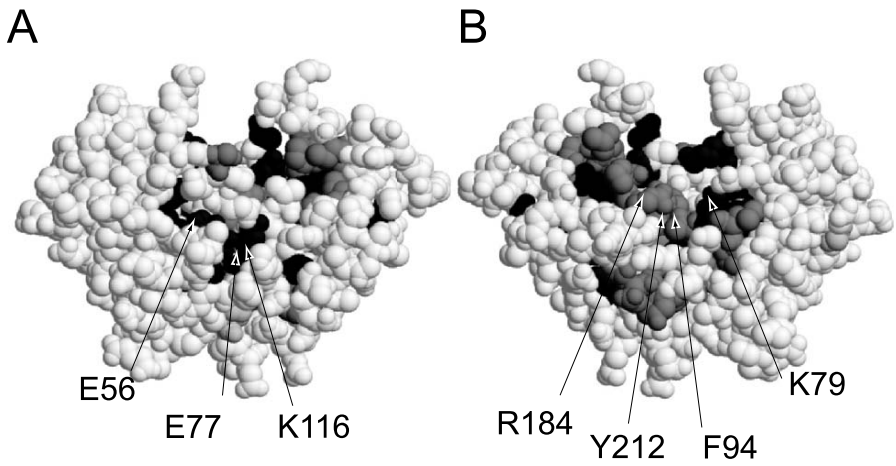


Fig. 6A, B. Mapping the evolutionary trace on the structure of MutH. Space-fill display of the MutH structure (pdb 2azo_B): the coloring is according to the evolutionary trace analysis between the MutH and REase (Fig. 5). Conserved residues are shown in *dark gray*, while class specific residues are shown in *light gray*. The position of the catalytically important active site residues *E56*, *D70*, *E77*, *K79* and *K116* are indicated. The position of three class-specific residues located in the DNA binding site (F94, R184 and Y212) is indicated in *gray*

3.2 Superposition of MutH with REases in Complexes with DNA

To find out which of the class-specific amino acid residues are most likely to be located in the protein DNA interface, we superimposed the structure of MutH with 11 structures of restriction enzyme-DNA complexes (Friedhoff et al. 2003), using the residues of the catalytic center, i.e., D70, E77 and K79 in case of MutH, as a seed. After superposition we searched for amino acid residues in MutH that are at a distance of 0.5 nm to the bases of each superimposed DNA molecule equivalent to the adenines of the GATC recognition sequence of MutH (Friedhoff et al. 2003). Residues that were conserved in the MutH group were then regarded as candidates involved in recognizing the methylation status of the adenines.

In the superimposed structures, the following residues in MutH turned out to be close to the nucleobases of the DNA, corresponding to the two adenine residues in the double stranded DNA sequence: Lys48 facing the minor groove and Phe94, Arg184 and Tyr212 facing the major groove (Fig. 7). As the N6 position of the adenine residues is located in the major groove, only Phe94, Arg184 and Tyr212 are good candidates to sense the methylation of N6 in one strand and the absence of methylation in the other strand. Moreover, Lys48 is also conserved in the Sau3AI family and might therefore have a general function in DNA binding and recognition. Tyr212 seemed to be of particular

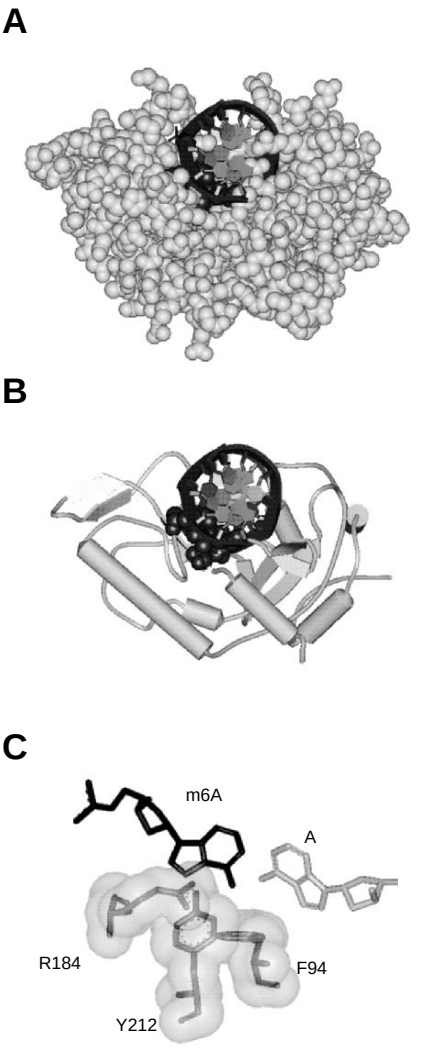


Fig. 7A–C. Superposition of MutH with MunI. **A** Result of the superposition of the active site of MutH with that of MunI (pdb code 1d02) in similar orientation as in Fig. 6B. **A** Space-fill representation of MutH and the DNA of the MunI-DNA complex as a *cartoon*. **B** Same as in **A** but with MutH in a schematic drawing showing *strand* and *helices* as *arrows* and *tubes*, respectively. The atoms of *F94*, *R184* and *Y212* are shown in space-fill representations. **C** Blow up of the three amino acid residues as well as the two adenine bases of the GATC sequence (in the MunI-DNA structure: AATT). The adenine on the strands to be cleaved must be unmethylated (A), while the adenine on the opposite strand (m^6A) can be methylated

interest, as the superposition suggests that it is located close to the adenine residues in both strands.

3.3 Mutational Analysis of MutH

Consequently, we performed a mutational analysis of the three amino acids identified by the bioinformatic analysis and generated the MutH variants F94A, R184A and Y212S. These variants were tested in vivo and in vitro for the activity in DNA mismatch repair. The in vivo analysis suggested that the MutH variants R184A and Y212S were severely impaired in their function in DNA mismatch repair (Table 2). Since this could be due to several factors

Table 2. Mutational analysis of MutH

Protein	In vivo activity (%)	DNA binding (%)	DNA cleavage (%)	Strand discrimination
Wild type	100	100	100	>100
F94A	50	10	290	78
R184A	0.3	5	3	>100
Y212S	0.5	7	86	2.5

Data are modified from Friedhoff et al. (2003). In vivo activity of MutH was monitored in a *mutH*-deficient strain as the ability of a plasmid-encoded MutH (wild type or variants) to reduce the mutation frequency. DNA binding was monitored as binding of MutH to a 19-mer oligonucleotides in a electrophoretic mobility shift assay. DNA cleavage was monitored as described in Fig. 8. Strand discrimination was calculated as the ratio of the cleavage rates for the unmethylated and methylated DNA strands in a hemimethylated DNA substrate.

(improper folding, reduced DNA binding/cleavage, or change in specificity) we purified the proteins to homogeneity and analyzed the cleavage of a hemimethylated DNA substrate that carried two different fluorophores on the unmethylated and the methylated strand. The analysis of the cleavage products obtained by incubation of these substrates with MutH was carried out by capillary electrophoresis with laser induced fluorescence detection using denaturing polyacrylamide gels. The results were very clear: The wild type MutH protein and the MutH variants F94A and R184A were only able to cleave the unmethylated DNA strand, though the R184A variant showed a reduced catalytic activity due to decreased DNA binding affinity (Table 2). On the other hand, the variant Y212S has lost its ability to discriminate between the unmethylated and the methylated strand being able to cleave both strands with a similar rate (Fig. 8). Hence, this variant has lost its activity by a change in specificity rather than activity and, therefore, cannot function as a strand discrimination factor in DNA mismatch repair in vivo.

Taken together the evolutionary trace analysis correctly predicted a functional site in the MutH protein family. One of these residues, Tyr212, turned out to be responsible for sensing the methylation status of the GATC site. The function of the Arg184 is mainly in DNA binding/recognition while the role of Phe84 remains to be solved by a more detailed analysis.

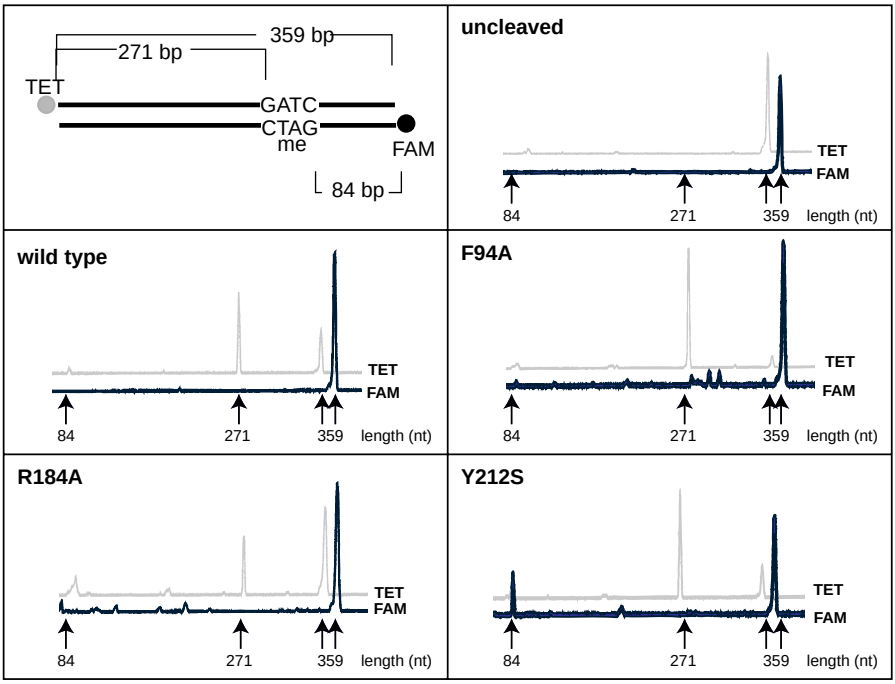


Fig. 8. Mutational analysis of MutH – identification of a methyl group sensor. Analysis of DNA cleavage by wild-type MutH and variants of MutH with hemimethylated DNA substrates. The DNA is labeled with the fluorophores FAM in the methylated strand and TET in the unmethylated strand. Cleavage in the unmethylated strand will lead to a 271-nt-long fragment labeled with TET while cleavage of the methylated strand will lead to an 84-nt-long DNA fragment labeled with FAM. Note that only the MutH variant Y212S has lost its ability to discriminate between the methylated and the unmethylated DNA strand

4 Conclusions

The results presented above describe examples of how the combination of bioinformatic predictions can guide biochemical and biophysical experiments to elucidate the functions of proteins. When done thoroughly the bioinformatic analysis leads to a testable hypothesis that can be specifically addressed by the biochemists. To this end, it will be very interesting to conduct similar analysis for other components of the DNA mismatch repair machinery. One issue will regard the yet unknown structure of the C-terminal domain of the evolutionary conserved protein MutL, which is important for the formation of homodimers and heterodimers in prokaryotes and eukaryotes, respectively. Another unresolved issue is the topology and structure of the mismatch repair complex. The identification of the protein–protein interfaces will be, therefore, an important step towards the understanding of this important biochemical pathway.

Acknowledgements. The expert technical assistance of Ina Steindorf is gratefully acknowledged. I thank Tomek Jurkowski for help with the phylogenetic analysis and Prof. A. Pingoud for critical reading of the manuscript. This work was supported by the Deutsche Forschungsgemeinschaft (Pi-122/12–4 and Pi-122/13–2) and the Dr Herbert Stolzenberg Stiftung.

References

Allen DJ, Makhov A, Grilley M, Taylor J, Thresher R, Modrich P, Griffith JD (1997) MutS mediates heteroduplex loop formation by a translocation mechanism. EMBO J 16: 4467–4476

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Armon A, Graur D, Ben-Tal N (2001) ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. J Mol Biol 307:447–463

Au KG, Welsh K, Modrich P (1992) Initiation of methyl-directed mismatch repair. J Biol Chem 267:12142–12148

Ban C, Yang W (1998) Structural basis for MutH activation in *E. coli* mismatch repair and relationship of MutH to restriction endonucleases. EMBO J 17:1526–1534

Ban C, Junop M, Yang W (1999) Transformation of MutL by ATP binding and hydrolysis: a switch in DNA mismatch repair. Cell 97:85–97

Bende SM, Grafstrom RH (1991) The DNA binding properties of the MutL protein isolated from Escherichia coli. Nucleic Acids Res 19:1549–1555

Bitinaite J, Wah DA, Aggarwal AK, Schildkraut I (1998) FokI dimerization is required for DNA cleavage. Proc Natl Acad Sci USA 95:10570–10575

Bujnicki JM (2001) A model of structure and action of Sau3AI restriction endonuclease that comprises two MutH-like endonuclease domains within a single polypeptide. Acta Microbiol Pol 50:219–231

Bujnicki JM, Elofsson A, Fischer D, Rychlewski L (2001) Structure Prediction Meta Server. Bioinformatics 17:750–751

Christ F, Schoettler S, Wende W, Steuer S, Pingoud A, Pingoud V (1999) The monomeric homing endonuclease PI-SceI has two catalytic centres for cleavage of the two strands of its DNA substrate. EMBO J 18:6908–6916

Cooper DL, Lahue RS, Modrich P (1993) Methyl-directed mismatch repair is bidirectional. J Biol Chem 268:11823–11829

Eisen JA, Hanawalt PC (1999) A phylogenomic study of DNA repair genes, proteins, and processes. Mutat Res 435:171–213

Embleton ML, Siksnys V, Halford SE (2001) DNA cleavage reactions by type II restriction enzymes that require two copies of their recognition sites. J Mol Biol 311:503–514

Friedhoff P, Lurz R, Luder G, Pingoud A (2001) *Sau*3AI, a monomeric type II restriction endonuclease that dimerizes on the DNA and thereby induces DNA loops. J Biol Chem 276:23581–23588

Friedhoff P, Sheybani B, Thomas E, Merz C, Pingoud A (2002) *Haemophilus influenzae* and *Vibrio cholerae* genes for *mutH* are able to fully complement a *mutH* defect in *Escherichia coli*. FEMS Microbiol Lett 208:121–126

Friedhoff P, Thomas E, Pingoud A (2003) Tyr-212: A key residue involved in strand discrimination by the DNA mismatch repair *endonuclease MutH*. J Mol Biol 325:285–297

Ginalski K, Elofsson A, Fischer D, Rychlewski L (2003) 3D-Jury: a simple approach to improve protein structure predictions. Bioinformatics 19:1015–1018

Glaser F, Pupko T, Paz I, Bell RE, Bechor-Shental D, Martz E, Ben-Tal N (2003) ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. Bioinformatics 19:163–164

Gradia S, Acharya S, Fishel R (1997) The human mismatch recognition complex hMSH2-hMSH6 functions as a novel molecular switch. Cell 91:995–1005

Hall MC, Matson SW (1999) The *Escherichia coli MutL* protein physically interacts with MutH and stimulates the MutH-associated endonuclease activity. J Biol Chem 274:1306–1312

Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp Ser 41:95–98

Hermann A, Jeltsch A (2003) Methylation sensitivity of restriction enzymes interacting with GATC sites. BioTechniques 34:924–926, 928, 930

Huai Q, Colandene JD, Topal MD, Ke H (2001) Structure of NaeI-DNA complex reveals dual-mode DNA recognition and complete dimer rearrangement. Nat Struct Biol 8:665–669

Innis CA, Shi J, Blundell TL (2000) Evolutionary trace analysis of TGF-beta and related growth factors: implications for site-directed mutagenesis. Protein Eng 13:839–847

Junop MS, Obmolova G, Rausch K, Hsieh P, Yang W (2001) Composite active site of an ABC ATPase: MutS uses ATP to verify mismatch recognition and authorize DNA repair. Mol Cell 7:1–12

Junop MS, Yang W, Funchain P, Clendenin W, Miller JH (2003) In vitro and in vivo studies of MutS, MutL and MutH mutants: correlation of mismatch repair and DNA recombination. DNA Repair (Amst) 2:387–405

Kumar S, Tamura K, Jakobsen IB, Nei M (2001) MEGA2: molecular evolutionary genetics analysis software. Bioinformatics 17:1244–1245

Kurowski MA, Bujnicki JM (2003) GeneSilico protein structure prediction meta-server. Nucleic Acids Res 31:3305–3307

Lahue RS, Au KG, Modrich P (1989) DNA mismatch correction in a defined system. Science 245:160–164

Lamers MH, Perrakis A, Enzlin JH, Winterwerp HH, de Wind N, Sixma TK (2000) The crystal structure of DNA mismatch repair protein MutS binding to a G x T mismatch. Nature 407:711–717

Lichtarge O, Bourne HR, Cohen FE (1996) An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol 257:342–358

Lichtarge O, Sowa ME (2002) Evolutionary predictions of binding surfaces and interactions. Curr Opin Struct Biol 12:21–27

Lichtarge O, Sowa ME, Philippi A (2002) Evolutionary traces of functional surfaces along G protein signaling pathway. Methods Enzymol 344:536–556

Loh T, Murphy KC, Marinus MG (2001) Mutational analysis of the MutH protein from *Escherichia coli*. J Biol Chem 276:12113–12119

Modrich P (1991) Mechanisms and biological effects of mismatch repair. Annu Rev Genet 25:229–253

Modrich P, Lahue R (1996) Mismatch repair in replication fidelity, genetic recombination, and cancer biology. Annu Rev Biochem 65:101–133

Mucke M, Grelle G, Behlke J, Kraft R, Kruger DH, Reuter M (2002) EcoRII: a restriction enzyme evolving recombination functions? EMBO J 21:5262–5268

Nicholas KB, Nicholas HBJ, Deerfield DWI (1997) GeneDoc: Analysis and Visualization of Genetic Variation. EMBnet News 4:14

Pingoud A, Jeltsch A (2001) Structure and function of type II restriction endonucleases. Nucleic Acids Res 29:3705–3727

Pupko T, Bell RE, Mayrose I, Glaser F, Ben-Tal N (2002) Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. Bioinformatics 18 (Suppl 1):S71–S77

Roberts RJ, Belfort M, Bestor T, Bhagwat AS, Bickle TA, Bitinaite J, Blumenthal RM et al. (2003) A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes. Nucleic Acids Res 31:1805–1812

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425

Seeber S, Kessler C, Gotz F (1990) Cloning, expression and characterization of the Sau3AI restriction and modification genes in *Staphylococcus carnosus* TM300. Gene 94:37–43

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 24:4876–4882

Toedt G, Krishnan R, Friedhoff P (2003) Site-specific protein modification to identify the MutL interface of MutH. Nucleic Acids Res 31:819–825

Twomey DP, McKay LL, O'Sullivan DJ (1998) Molecular characterization of the *Lactococcus lactis* LlaKR2I restriction-modification system and effect of an IS982 element positioned between the restriction and modification genes. J Bacteriol 180:5844–5854

Wah DA, Bitinaite J, Schildkraut I, Aggarwal AK (1998) Structure of FokI has implications for DNA cleavage. Proc Natl Acad Sci USA 95:10564–10569

Welsh KM, Lu AL, Clark S, Modrich P (1987) Isolation and characterization of the *Escherichia coli* mutH gene product. J Biol Chem 262:15624–15629

Wu TH, Loh T, Marinus MG (2002) The function of Asp70, Glu77 and Lys79 in the *Escherichia coli* MutH protein. Nucleic Acids Res 30:818–822

Yang W (2000) Structure and function of mismatch repair proteins. Mutat Res 460:245–256

Zhou XE, Wang Y, Reuter M, Mackeldanz P, Kruger DH, Meehan EJ, Chen L (2003) A single mutation of restriction endonuclease EcoRII led to a new crystal form that diffracts to 2.1 A resolution. Acta Crystallogr D Biol Crystallogr 59:910–912

Predicting Functional Residues in DNA Glycosylases by Analysis of Structure and Conservation

D.O. Zharkov

1 Introduction

Almost every biochemist and molecular biologist with an interest in protein research confronts the question of the role played by individual amino acid residues in a specific polypeptide. The wide variety of experimental techniques available to address this question can be categorized into two general approaches: functional and structural. In the former case, the residue in question is chemically modified or mutated; in the latter, the relationships with neighboring residues are defined and biological function is inferred. Each approach has its advantages and limitations and the most accurate information is provided when both are used together.

Elsewhere (Zharkov and Grollman 2002), I have outlined theoretical and practical methods for using information derived from protein structure and the conservation of amino acid residues to predict the biochemical function(s) of these residues. Such residues then become candidates for functional testing by site-directed mutagenesis or chemical modification. In that work, the principles of this bioinformatics approach were illustrated by analyzing two families of DNA glycosylases. These enzymes, which initiate the repair of damaged DNA, are members of the same structural family but exhibit sharply different substrate specificities. Penicillin-binding proteins (Goffin and Ghuysen 1998), transferrins (Gu 1999), Myc proteins (Gu 1999), cyclooxygenases (Gu 2001) and caspases (Wang and Gu 2001) have been analyzed by similar methods, but without the inclusion of structural information. The general methodology of our analysis will be reviewed here only briefly; readers interested in detailed information should consult the original paper (Zharkov and Grollman 2002). In this review, I will focus on recent methodological develop-

D.O. Zharkov
Institute of Chemical Biology and Fundamental Medicine, Siberian Division of Russian Academy of Sciences, Novosibirsk 630090, Russia

Nucleic Acids and Molecular Biology, Vol. 15
Janusz M. Bujnicki (Ed.)
Practical Bioinformatics

ments, placing emphasis on their practical application. I will demonstrate their use in refining predictions and present examples of hypotheses generated by this type of analysis.

2 Generating Predictions: Sequence Selection and Analysis

The basis for structure-conservation analysis is intuitively clear (Zharkov and Grollman 2002). Consider the situation in which a group of sequence-related proteins (orthologous, paralogous, or both) is divided into two or more functional subgroups. Then, residues conserved across all subgroups should be important for functions common to all proteins, whereas residues conserved in some subgroups, but not others, would be important only for functions of subgroups in which they are conserved. For example, among DNA glycosylases, the endonuclease III family may be divided into the subgroups Nth and MutY, which are similar structurally, but differ in the types of damage they recognize and process (Aravind et al. 1999; Eisen and Hanawalt 1999). All positions in the sequence alignment may be classified as *similar* (residues conserved in all subgroups), *dissimilar* (residues conserved in some but not all subgroups), or *unconserved* (Fig. 1). The availability of a three-dimensional structure for a representative member of the subgroup provides an outline of the structurally conserved core or active center for the entire subgroup. Mapping dissimilar residues on such a structure reveals positions of residues responsible for substrate specificity or other subgroup-specific functions.

Application of this strategy requires careful consideration of several issues. Which sequences should be included and which should be intentionally excluded in the comparison (sequence *selection* and *qualification*)? Which are the best subgroup divisions (sequence *classification*)? How can conservation be quantified? The answer to any one question may depend upon the answer to another.

The most important parameter in this analysis is the algorithm used to quantify conservation. The two main approaches use publicly available software. One, exemplified by the AMAS (Analysis of Multiply Aligned Sequences) algorithm (Livingstone and Barton 1993, 1996), takes into account the physicochemical properties of individual amino acid residues. For each position of the sequence alignment, AMAS calculates a "conservation number", C_n, representing the number of borders in the Venn diagram that must be crossed to include all amino acids at this position. The Venn diagram of physicochemical properties may be custom defined, but a standard Taylor set (Taylor 1986) is often a good starting point. To be considered conserved, the value for C_n must exceed a certain threshold value, after which a simple set of rules (Fig. 1) may be followed to classify every position (Zharkov and Grollman 2002). AMAS is available as a Web-based server (barton.ebi.ac.uk/servers/amas_server.html). An alternative approach, developed by Gu (1999,

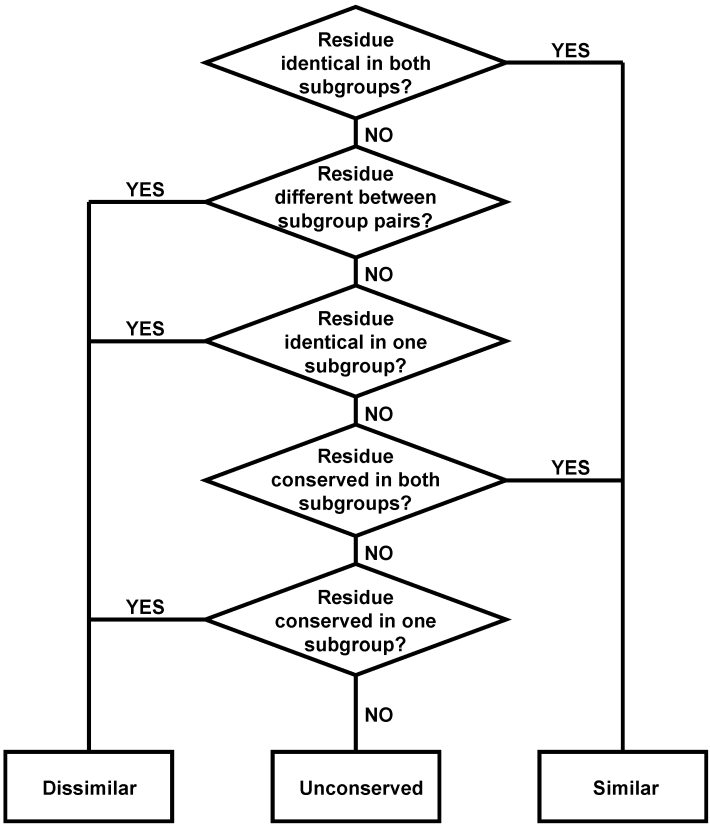


Fig. 1. Flow chart for the similarity classification algorithm. The chart illustrates the procedure for assigning residues to a subgroup. Output of the AMAS algorithm is used as the input for this classification

2001; Wang and Gu 2001) and later implemented as a Gu99 algorithm in the DIVERGENCE software (xgu1.zool.iastate.edu/software.html), essentially disregards the physicochemical properties of individual residues. This algorithm, provided with two clusters in a tree of related sequences, uses statistical methods to estimate the so-called posterior probability that divergence at every position will contribute to the total value for divergence between the clusters. With some reservations, this value is presumed to reflect selective pressure at this particular position after gene duplication and functional divergence. Thus, if a certain residue is important for a subgroup-specific function, it will have a higher posterior probability of being related to divergence between the subgroups.

In AMAS, the selection of sequences for analysis is of utmost importance as the algorithm does not distinguish between conservation by descent and conservation by function. Inclusion of many closely related sequences will influ-

ence C_n if a fraction of atypical residues is ignored, as often is the case when minimizing noise created by a chance inclusion of a wrongly classified or unqualified sequence. In contrast, Gu99 is not adversely affected by sequence relationships, as it is already incorporated into the procedure. Still, practical limitations of computing power favor pre-selection of sequences. In earlier work (Zharkov and Grollman 2002), we used the National Center for Biotechnology Information's Clusters of Orthologous Groups database (Tatusov et al. 1997, 2001), which offers a selection of sequences from a broad set of phylogenetic lineages. Alternative options are, however, available. The greatest number of relevant sequences related to a given sequence is obtained by using the latter as a query for a nonrestricted BLAST search (Altschul et al. 1990, 1997) in a nonredundant sequence database. If the search is limited to certain genomes, these should be chosen to represent a large number of phylogenetic lineages. Alternatively, if a nonrestricted search is used, a subset of identified sequences may be selected for subsequent additional analysis. In practice, we found the latter approach to be the most useful, especially when analyzing proteins that occur only in a limited number of lineages. The optimal level of selection depends on the total number of sequences; for example, we selected one sequence per phylogenetic order to analyze the Nth family of DNA glycosylases (see below).

Problems of qualification and classification are important for technical reasons primarily, as functional information is generally available for only a small number of proteins under analysis (Zharkov and Grollman 2002). Obviously, if residues with crucial functions have already been determined biochemically for the entire family or subgroup(s), then the sequences to be analyzed should be checked for conservation of such elements to avoid an artificial decrease in similarity. Classification of sequences into subgroups is best done by constructing the phylogenetic tree and considering the relationship of the retrieved sequences to prototype subgroup members although, in some cases, corrections may be made on the basis of well-established functionally important subgroup-specific residues. For example, biochemical experiments have shown that the *E. coli* MutY protein is similar to *E. coli* Nth but possesses an additional C-terminal domain that regulates its substrate specificity. Therefore, the presence of this domain designates a sequence as MutY even if it clusters with Nth in a tree of Nth family members.

Given the conceptual difference in the approaches characterizing the AMAS and Gu99 algorithms, it is of interest to compare results obtained by both algorithms in a defined protein family. Endonuclease III (Nth) from *E. coli* is the prototype of the largest superfamily of DNA repair glycosylases (Asahara et al. 1989; Thayer et al. 1995), enzymes that excise damaged bases from DNA, thereby helping to maintain genomic stability (Friedberg et al. 1995). This superfamily is characterized by the presence of a helix-hairpin-helix motif and a conserved loop ending in an aspartic acid residue (Thayer et al. 1995; Nash et al. 1996). DNA glycosylases belonging to this superfamily

are collectively capable of repairing almost the full repertoire of base lesions. A subset of the superfamily, termed the Nth family, combines enzymes that share the two signature motifs and an overall three-dimensional organization with *E. coli* Nth (Eco-Nth).

Nth family proteins are bilobal, with one lobe comprising a six-helix barrel and the other containing a $[4Fe\text{-}4S]^{2+}$ iron-sulfur cluster. Despite close sequence and structural similarity, these close relatives of Eco-Nth are surprisingly diverse with respect to the spectrum of lesions removed from DNA. The entire Nth family may be divided into four subgroups by substrate specificity: (1) the Nth subgroup, specific for oxidized or reduced pyrimidines, such as thymine glycols and 5,6-dihydropyrimidines, where Eco-Nth is a representative member (Asahara et al. 1989); (2) the Pdg subgroup, specific for pyrimidine dimer UV photoproducts, where *Micrococcus luteus* pyrimidine dimer glycosylase (Mlu-Pdg) is a representative member (Piersen et al. 1995); (3) MutY subgroup, specific for adenine mismatched either with guanine or 8-oxoguanine; *E. coli* MutY protein (Eco-MutY) is a representative member (Michaels et al. 1990); and (4) Tdg subgroup, specific for thymine or uracil mismatched with guanine; with *Methanothermobacter thermautotrophicus* thymine-DNA glycosylase, Mth-Tdg, being a representative member (Begley and Cunningham 1999).

Nth and MutY are ubiquitous, whereas Pdg and Tdg are restricted in their appearance in the tree of life. MutY differs from Nth in two respects. First, Nth possesses concomitant AP lyase activity, for which Lys-120 (in Eco-Nth) is absolutely required. Eco-MutY, however, has a serine residue in this position and possesses glycosylase, but not AP lyase, activity. Secondly, as noted above, MutY contains an additional C-terminal domain important for recognition of a mismatched base opposite adenine.

To perform the comparative analysis of predictions generated by AMAS and Gu99, we searched the nonredundant National Center for Biotechnology Information's protein sequence database with BLASTP (Altschul et al. 1997), using sequences of Eco-Nth and Eco-MutY as queries. For every taxonomic order, as defined in the NCBI Taxonomy Browser, a single best hit was selected for further analysis (Table 1). After qualification, based on the presence of an intact iron-sulfur cluster, sequences were classified as Nth if they possessed lysine at the position corresponding to K120 in *E. coli* Nth and had no C-terminal domain; alternatively, sequences were classified as MutY if they had the C-terminal domain and a residue other than lysine at the K120-related position. Sequences were aligned using ClustalW (Thompson et al. 1994), and a tree was constructed by the neighbor-joining method (Saitou and Nei, 1987), producing well-defined clusters for Nth and MutY. Positions were then either classified as similar or dissimilar at C_n=8 by AMAS (Fig. 2A), and posterior probability was calculated for each cluster by Gu99 (Fig. 2B).

As shown in Fig. 2, the two algorithms did not produce identical results. The cluster of dissimilar residues on the upper "lip" of the DNA-binding

Table 1. Sequence selection for Nth and MutY

Phylum	Class	Order	Representative species
Bacteria			
Actinobacteria	Actinobacteria	Actinomycetales	*Thermobifida fusca* (Nth, MutY)
		Bifidobacteriales	*Bifidobacterium longum* DJO10A (Nth, MutY)
Aquificae	Aquificae	Aquificales	*Aquifex aeolicus* (Nth)
Bacteroidetes	Bacteroides	Bacteroidales	*Bacteroides thetaiotaomicron* VPI-5482 (Nth, MutY)
	Sphingobacteria	Sphingobacteriales	*Cytophaga hutchinsonii* (Nth, MutY)
Chlorobi	Chlorobia	Chlorobiales	*Chlorobium tepidum* TLS (Nth)
Chlamydiae	Chlamydiae	Chlamydiales	*Chlamydia trachomatis* (MutY)
Chloroflexi	Chloroflexi	Chloroflexales	*Chloroflexus aurantiacus* (Nth)
Cyanobacteria		Chroococcales	*Thermosynechococcus elongatus* BP-1 (MutY)
		Nostocales	*Nostoc punctiforme* (Nth)
		Oscillatoriales	*Trichodesmium erythraeum* IMS101 (Nth)
		Prochlorophytes	*Prochlorococcus marinus* str. MIT 9313 (MutY)
Deinococcus-Thermus	Deinococci	Deinococcales	*Deinococcus radiodurans* (Nth, MutY)
Firmicutes	Bacilli	Bacillales	*Bacillus halodurans* (Nth)
			Oceanobacillus iheyensis HTE831 (MutY)
		Lactobacillales	*Enterococcus faecalis* V583 (Nth)
			Lactococcus lactis subsp. *lactis* (MutY)
	Clostridia	Clostridiales	*Heliobacillus mobilis* (Nth); *Desulfitobacterium hafniense* (MutY)
		Thermoanaerobacteriales	*Thermoanaerobacter tengcongensis* (Nth)
Fusobacteria	Fusobacteria	Fusobacterales	*Fusobacterium nucleatum* subsp. *nucleatum* ATCC 25586 (Nth)
Planctomycetes	Planctomycetacia	Planctomycetales	*Pirellula* sp. (MutY)
Proteobacteria	Alpha-proteobacteria	Caulobacterales	*Caulobacter crescentus* CB15 (Nth, MutY)
		Rhizobiales	*Bradyrhizobium japonicum* (Nth); *Agrobacterium tumefaciens* (MutY)
		Rhodobacterales	*Rhodobacter sphaeroides* (Nth, MutY)
		Rhodospirillales	*Magnetospirillum magnetotacticum* (Nth, MutY)
		Rickettsiales	*Rickettsia conorii* (Nth)
		Sphingomonadales	*Novosphingobium aromaticivorans* (Nth, MutY)

Table 1. (*Continued*)

Phylum	Class	Order	Representative species
	Beta-proteobacteria	Burkholderiales	*Burkholderia fungorum* (Nth, MutY)
		Neisseriales	*Neisseria meningitidis* MC58 (Nth); *Neisseria meningitidis* Z2491 (MutY)
		Nitrosomonadales	*Nitrosomonas europaea* ATCC 19718 (Nth, MutY)
	Gamma proteobacteria	Alteromonadales	*Shewanella oneidensis* MR-1 (Nth, MutY)
		Enterobacteriales	*Escherichia coli* K12 (Nth, MutY)
		Legionellales	*Coxiella burnetii* RSA 493 (Nth, MutY)
		Pasteurellales	*Haemophilus influenzae* Rd (Nth, MutY)
		Pseudomonadales	*Pseudomonas fluorescens* PfO-1 (Nth) *Pseudomonas putida* KT2440 (MutY)
		Vibrionales	*Vibrio vulnificus* CMCP6 (Nth); *Vibrio cholerae* (MutY)
		Xanthomonadales	*Xanthomonas campestris* pv. *campestris* str. ATCC 33913 (Nth, MutY)
	Delta-proteobacteria	Desulfovibrionales	*Desulfovibrio desulfuricans* G20 (Nth, MutY)
	Epsilon-proteobacteria	Campylobacterales	*Campylobacter jejuni* (Nth); *Helicobacter hepaticus* ATCC 51449 (MutY)
	Magnetotactic cocci		*Magnetococcus* sp. MC-1 (Nth, MutY)
Spirochaetes	Spirochaetes	Spirochaetales	*Treponema pallidum* (Nth) *Leptospira interrogans* serovar lai str. 56601 (MutY)
Thermotogae	Thermotogae	Thermotogales	*Thermotoga maritima* (Nth)
Archaea			
Crenarchaeota	Thermoprotei	Desulfurococcales	*Aeropyrum pernix* (Nth)
		Sulfolobales	*Sulfolobus solfataricus* (Nth)
		Thermoproteales	*Pyrobaculum aerophilum* (Nth)
Euryarchaeota	Archaeoglobi	Archaeoglobales	*Archaeoglobus fulgidus* DSM 4304 (Nth)
	Halobacteria	Halobacteriales	*Halobacterium* sp. NRC-1 (Nth)
	Methanococci	Methanococcales	*Methanocaldococcus jannaschii* (Nth)
	Methanopyri	Methanopyrales	*Methanopyrus kandleri* AV19 (Nth)
	Thermococci	Thermococcales	*Pyrococcus abyssi* (Nth)

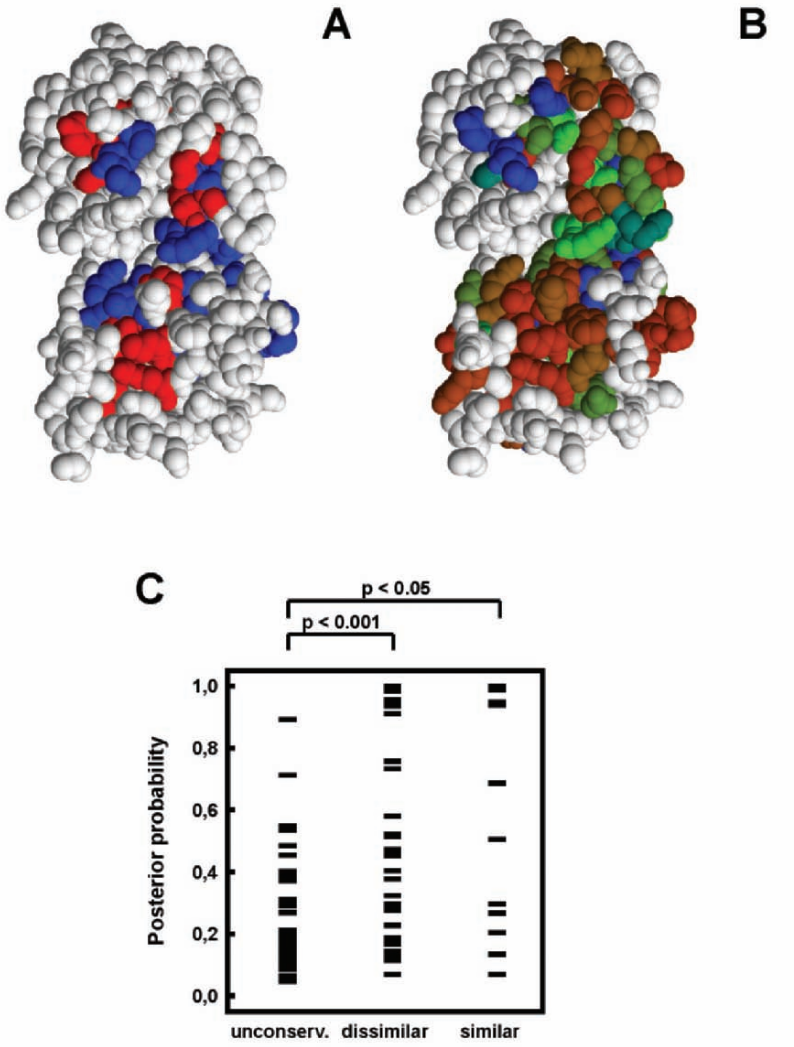


Fig. 2. Correlation between predictions of AMAS and Gu99 algorithms. The classification of positions into unconserved, dissimilar and similar was performed using AMAS on the nonrestricted order-level nonredundant Nth/MutY dataset at C_n=8 (see text for details). The sequences possessing an intact iron-sulfur cluster were deemed qualified; classification was based on the presence of (1) the C-terminal domain (excluded from the alignment and tree building) and (2) position 120 in *E. coli* Nth/MutY. **A** Similar (*red*), dissimilar (*blue*) and unconserved (*gray*) residues mapped on the structure of *E. coli* Nth. **B** The same structure with residues colored according to their posterior probability (*P*=0, *red*, *P*=0.5, *green*, *P*=1, *blue*, with *halftones* in between; residues for which the DIVERGENCE v1.04 software produced no *P* value are *shaded gray*). **C** Posterior probabilities of the site being related to functional divergence plotted for 99 sites. Correlation between groups of data was estimated from the biserial correlation coefficient; the significance is given by Student's t-test for the biserial correlation coefficient

groove was identified only partly by posterior probability, as were the dissimilar residues in the active site pocket at the center of the protein globule. In particular, Lys-120, which is well established biochemically to be essential for Nth but not MutY activity, and which served for sequence classification, had P=0.47, a low value for a residue of crucial functional difference. Residues of the dissimilarity cluster at the lower lip of the DNA-binding groove also displayed low posterior probability. Indeed, although there is a statistically significant difference in posterior probability among the unconserved, similar and dissimilar residues, the predictive power of correlation is not high (biserial correlation coefficient r_{bs}=0.45 for "dissimilar" vs. "unconserved" group; r_{bs}=0.25 for "similar" vs. "unconserved" group; Fig. 2C).

The reason for this discrepancy may lie in the heterogeneity of the "dissimilar" group, as defined by AMAS. Consider the general situation of two subgroups. If a dissimilar position is conserved in one group, it may be conserved in another ("asymmetric dissimilarity") or conserved as a residue with different physicochemical properties ("symmetric dissimilarity"). Symmetrically dissimilar positions are associated with anomalously low P values in Gu99, because they do not accumulate many changes even when the physicochemical properties of the respective residues differ tremendously, as is the case for Lys-120 in Nth and Ser-120 in MutY. In fact, if symmetrically dissimilar positions are excluded from the dissimilar group, then r_{bs} for dissimilar vs. unconserved increases from 0.45 to 0.53, and the symmetrically dissimilar group becomes significantly different from the asymmetrically dissimilar (r_{bs}=0.38, P<0.05) but not from unconserved or similar categories. Similarly, Gu99 does not distinguish efficiently between similar and unconserved positions, as neither is likely to contribute to functional divergence.

Ideally, both approaches should be used to identify primary candidates for site-directed mutagenesis. Below, we discuss examples of mutations at positions identified by both procedures. To date, AMAS appears to have better predicting power if the sequences are chosen with care. Ideally, an algorithm for functional conservation prediction would take both phylogeny and physicochemical properties into account; to our knowledge, no such algorithm has been developed.

3 Testing the Predictions: Mutational Analysis of Residues Defining Substrate Specificity in Formamidopyrimidine-DNA Glycosylase

Formamidopyrimidine-DNA glycosylase (Fpg or MutM) and endonuclease VIII (Nei) from *E. coli* are structurally related DNA glycosylases that excise oxidized bases from DNA (Tchou et al. 1991; Melamede et al. 1994). Although these enzymes reportedly act on many different substrates, Fpg primarily excises redox-modified purines (8-oxoguanine and formamidopyrimidines), while

Nei is most active on redox-modified pyrimidines (David and Williams 1998). Eco-Fpg and Eco-Nei display high sequence similarity (Jiang et al. 1997a) and serve as protypes for the Fpg superfamily of DNA repair glycosylases (Zharkov et al. 2003), including more than 100 bacterial, plant and vertebrate proteins. The biochemistry of Fpg and Nei is well-established, including functional requirements for the N-terminal PE catalytic dyad (Tchou and Grollman 1995; Zharkov et al. 1997; Lavrukhin and Lloyd 2000) and the C-terminal Cys_4 zinc finger (O'Connor et al. 1993; Tchou et al. 1993). Crystal structures were determined recently for several members of the Fpg family (Fromme and Verdine 2002; Gilboa et al. 2002; Serre et al. 2002; Zharkov et al. 2002).

To identify candidate subgroup-specific residues for Fpg or Nei, we performed a BLAST search in the NCBI nonredundant protein sequence database using Eco-Fpg and Eco-Nei as queries. Qualification of sequences was based on the presence of the N-terminal PE and C-terminal Cys_4 zinc finger motifs. Following alignment and tree construction, the sequences most distant from the query sequences were used as queries in the second round of the search. We found 147 homologues of Fpg and Nei, showing an even distribution of Fpg (but not Nei) in bacteria, although no Fpg homologues were found in Archaea. The sequences identified were classified as belonging, or not belonging, to the Fpg subgroup, based on the presence of the N-terminal motif PE(L/I/M)PE (124 sequences); the remaining 23 sequences carrying the N-terminal signature motif PEG were considered to be Nei. This highly skewed distribution toward Fpg made it difficult to reliably predict Nei-specific residues due to high noise levels; thus, we restricted the analysis to Fpg. However, although Nei sequences formed a cluster clearly separated from Fpg in the tree, the low conservation of many elements functionally important for Eco-Nei suggest that Nei as defined here may be an artificial subgroup and perhaps should be further classified into narrower subgroups. Nevertheless, as presently constructed, this subgroup functions as a "non-Fpg" category.

To analyze the conservation of physicochemical properties of Fpg residues, subsets of the sequences were randomly selected (with the exception that *E. coli* K12 sequences were always present) to represent no more than one per order and were then re-aligned. Conservation in the aligned sequences were analyzed by AMAS with the threshold C_n=9. The structure of *E. coli* Fpg covalently complexed with DNA (1K82 in the Protein Data Bank) (Gilboa et al. 2002) was used for mapping. The alignment also was analyzed by Gu99, in which case the Nei subgroup was extended to one sequence per genus to provide a cluster of a workable size.

Two residues identified by this similarity analysis (Arg-108 and His-89) proved to be of particular interest when mapped on the three-dimensional structure. Arg-108 (together with Met-73 and Phe-110) is part of a void-filling triad inserted into the DNA helix, compensating for the void produced when the damaged base flips out of the helix into the active site pocket (Gilboa et al. 2002). Arg-108 forms two hydrogen bonds with the orphaned cytosine, con-

tributing to the opposite-base specificity of the enzyme (Tchou et al. 1994). In Eco-Nei, this function is performed by Gln-69, which forms a void-filling triad with Leu-70 and Tyr-71 (Zharkov et al. 2002); surprisingly, this element is not conserved in the Nei subgroup. Unlike Eco-Fpg, Eco-Nei does not discriminate among opposite bases (Jiang et al. 1997b). In contrast, Arg-108 is highly conserved among bacterial Fpg homologues but not in Nei proteins. A supplementary analysis by Gu99 indicated an extremely high posterior probability for this residue (P=0.9999) to be involved in the functional divergence between Fpg and Nei subgroups. His-89 (P=0.995) also forms two hydrogen bonds with DNA but makes these contacts with the phosphates of the strand opposite the lesion, possibly contributing to early steps of lesion recognition by indirect readout (Zharkov et al. 2004).

Mutations of Arg-108 and His-89 produced marked overall effects on enzyme activity, but the mechanisms involved were different, as expected from their different predicted functions (Fig. 3; Zaika et al. 2004). For exam-

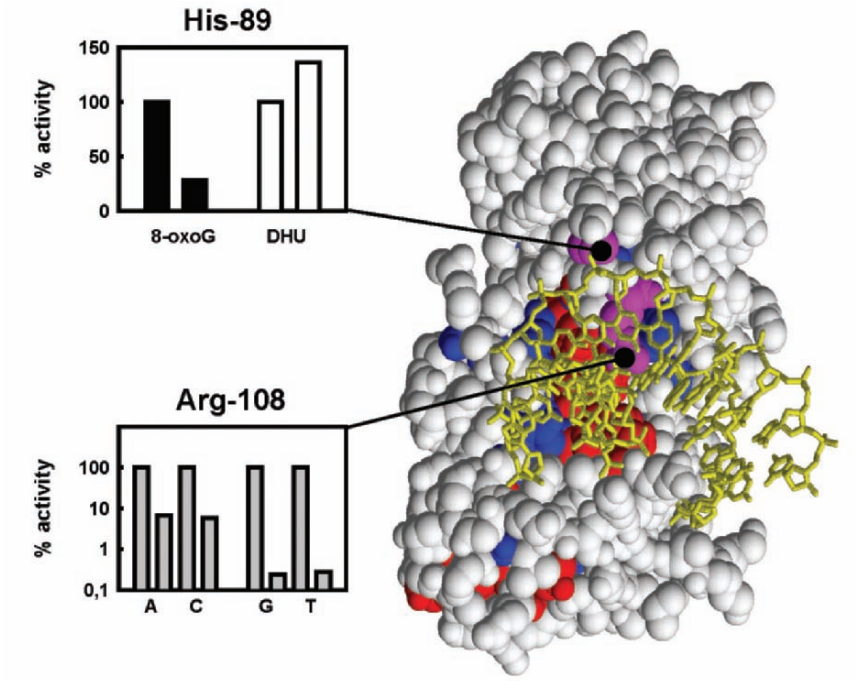


Fig. 3. Effects of mutations of dissimilar residues in Fpg. Similar (*red*) and dissimilar (*blue*) residues are defined and mapped as described in the text. His-89 and Arg-108 are colored *magenta*. For the H89A mutation, the specific activity (k_{cat}/K_M) of the wild-type enzyme toward 8-oxoG:C and DHU:C is plotted as 100 % in the *left bar*, and that of the mutant is shown in the *right bar* of a respective group. For the R108A mutation, the specific activity against 8-oxoG:A, 8-oxoG:C, 8-oxoG:G, and 8-oxoG:T for the wild-type enzyme is plotted in the same way (note the log scale of the ordinate)

ple, the H89A mutation significantly decreased activity of the enzyme towards DNA containing 8-oxoguanine (8-oxoG), but not dihydrouracil (DHU), the latter being an unspecific substrate for this enzyme, presumably recognized via a different mechanism than 8-oxoG (Karakaya et al. 1997). The R108A mutation influenced the activity toward 8-oxoG in an opposite base-specific manner. For example, substrates containing C and A opposite 8-oxoG were affected far less than those with G and T opposite the lesion (Fig. 3). Our site-directed mutagenesis experiments, therefore, confirmed the predictions made by structural conservation analysis (Zaika et al. 2004).

4 Refining the Predictions: Analysis of Substrate Specificity in the Endonuclease III Family

The Nth family of DNA glycosylases may be divided into four subgroups by substrate specificity: Nth, Pdg, MutY and Tdg (see above). Although the overall structures of these enzymes are similar, their substrate specificities are quite different. From two subgroups previously considered for the Nth family, namely Nth and MutY (Zharkov and Grollman 2002), I shall now extend this analysis to all subgroups of the Nth family.

The following 11 archaeal and 44 bacterial genomes were searched by BLAST in the NCBI microbial genome database: *Aeropyrum pernix, Sulfolobus solfataricus, Pyrobaculum aerophilum, Archaeoglobus fulgidus, Halobacterium* sp. NRC-1, *Methanothermobacter thermautotrophicus, Methanocaldococcus jannaschii, Methanopyrus kandleri* AV19, *Methanosarcina mazei* Goe1, *Pyrococcus furiosus* DSM 3638, *Thermoplasma volcanium, Mycobacterium tuberculosis* H37Rv, *Streptomyces coelicolor* A3(2), *Aquifex aeolicus, Chlorobium tepidum* TLS, *Chlamydia trachomatis, Chlamydophila pneumoniae* CWL029, *Nostoc* sp. PCC 7120, *Synechocystis* sp. PCC 6803, *Bacillus subtilis, Clostridium perfringens, Enterococcus faecium, Mycoplasma pneumoniae, Ureaplasma urealyticum, Lactococcus lactis* subsp. *lactis*, *Listeria innocua, Thermoanaerobacter tengcongensis, Staphylococcus aureus* subsp. *aureus* N315, *Streptococcus pyogenes* M1 GAS, *Fusobacterium nucleatum* subsp. *nucleatum* ATCC 25586, *Magnetococcus* sp. MC-1, *Caulobacter crescentus* CB15, *Agrobacterium tumefaciens* str. C58 (U. Washington), *Mesorhizobium loti, Rhodobacter sphaeroides, Rickettsia prowazekii, Ralstonia solanacearum, Neisseria meningitidis* Z2491, *Nitrosomonas europaea, Campylobacter jejuni, Helicobacter pylori* 26695, *Escherichia coli* K12, *Yersinia pestis, Buchnera aphidicola* str. Sg, *Vibrio cholerae, Xanthomonas campestris* pv. *campestris* str. ATCC 33913, *Xylella fastidiosa* 9a5 c, *Haemophilus influenzae* Rd, *Pasteurella multocida, Pseudomonas aeruginosa, Salmonella typhimurium* LT2, *Borrelia burgdorferi, Treponema pallidum, Thermotoga maritima, Deinococcus radiodurans*. These genomes represent 46 phylogenetic groups, with no more than two genomes per group; four genomes were unfinished at the time of our analysis. Sequences of Eco-

Nth, Mlu-Pdg, Eco-MutY, and Mth-Tdg were used as queries and 100 top-scoring sequences for each query were pooled. All sequences were aligned and the tree was constructed by the neighbor-joining method. Sequences outside the root common for Eco-Nth, Mlu-Pdg, Eco-MutY, and Mth-Tdg were discarded. Remaining sequences were re-aligned and classified into one of the four subgroups according to rooting with the closest query sequence. The manual sequence qualification step was omitted. Physicochemical properties of residues in the aligned sequences were analyzed by AMAS (C_n=7, 10 % atypical residues allowed, no gaps ignored, cysteines considered reduced). X-ray crystallographic structures of Nth (2ABK; Thayer et al. 1995), MutY (1MUY; Guan et al. 1998), and Tdg (1KEA; Mol et al. 2002) were used for mapping.

The BLAST search in 55 microbial genomes recovered a total of 103 sequences similar to the four query sequences, including the Mlu-Pdg and Mth-Tdg sequences, although genomes of the respective species were not searched. As 100 top-scoring sequences were taken from each search, the small number of sequences in the pool reveals that the Nth family is well conserved and shares little similarity with other sequences. Each query produced many sequences from other subgroups, e.g., the Eco-Nth query identified the Eco-MutY sequence as a homologue. Following classification and qualification steps, 80 sequences belonging to 38 phylogenetic lineages remained in the analysis. The Nth subgroup included 33 sequences of 29 phylogenetic lineages; the Pdg subgroup, 8 sequences of 6 lineages; the MutY subgroup, 34 sequences of 28 lineages; and the Tdg subgroup, 5 sequences of 4 lineages. Visual inspection of the alignment for sequence length and subgroup-specific conserved motifs confirmed the correctness of the group composition. Interestingly, the Tdg subgroup included archaean sequences only.

In general, Nth proteins grouped with Pdg proteins and MutY proteins with Tdg proteins formed two subgroup pairs. Many positions are conserved in Nth and Pdg and in MutY and Tdg, but not between these two pairs. Positions conserved in one protein and in either member of the other pair are rare. As Nth and Pdg participate in the repair of damaged bases, while MutY and Tdg repair mismatched bases, such groupings may reflect either functional differences between these enzymes, or their evolutionary relationships. The latter possibility appears less likely because of the exclusively archaean origin of Tdg.

Representative enzymes from three of the four subgroups (excluding Pdg) have been crystallized and their three-dimensional structures determined by X-ray diffraction methods (Kuo et al. 1992; Thayer et al. 1995; Guan et al. 1998; Mol et al. 2002). Enzymes of the Nth family bound to their cognate DNA have not yet been structurally analyzed[1], and their DNA-binding site and active site

[1] After this manuscript was completed, a structure of Nth covalently bound to DNA was published (Fromme and Verdine 2003); since the structures of DNA-bound MutY and Tdg are not available, the structure of free Nth was nevertheless used for illustrative purposes here.

are inferred from biochemical and mutagenesis evidence. Analysis of residues conserved across all subgroups of the Nth family were conducted previously (Zharkov and Grollman 2002) and the results did not differ significantly when the present data set was included. Residues that appeared to be specific for the Nth, MutY, or Tdg subgroup were mapped on the appropriate structure. Mapping serves as a useful visualization tool and helps in the understanding of the possible roles of the subgroup-specific amino acids (Zharkov and Grollman 2002). Residues were considered specific for a subgroup if (1) they were conserved ($C_n \geq 7$) within the subgroup and (2) they were not conserved in the other subgroup of the same pair. Most residues fulfilling these two criteria were not conserved in any of the three remaining subgroups.

Proteins in the Eco-Nth, Eco-MutY, and Mth-Tdg subgroups contain two lobes separated by a positively charged interdomain cleft, where DNA is presumably bound. A deep pocket opens into the bottom of this cleft, containing residues important for the enzyme's catalytic activity. The groove usually has well-defined rims or "lips." The inferred mechanism of action for enzymes in the Nth family postulates that damaged DNA is bound into the enzyme's cleft and kinked at the site of the lesion. The base to be excised is then extruded (flipped out) of the double helix and inserted into the enzyme's active site pocket, where a series of chemical reactions take place (McCullough et al. 1999).

Mapping of subgroup-specific residues reveals them to be slightly more scattered across the enzyme globule as compared with the previous analysis of Nth and MutY (Zharkov and Grollman 2002). This is especially evident in Tdg, likely due to the small sample size. Nevertheless, many subgroup-specific residues clearly cluster on the lips of the interdomain groove and in the active-site pocket.

In the Eco-Nth (Fig. 4A), the highly conserved E23 and Y185 residues close the far-left part of the groove and form a hydrogen bond between the Y185 hydroxyl and one of the E23 Oε atoms (Kuo et al. 1992). (Orientation here and elsewhere is given with the six-barrel domain pointing upward). H176 and H177 form the bottom of the active-site pocket.

In Eco-MutY (Fig. 4B), R19 closes the far-left part of the groove. G139 and A189 are located on the lower lip and Y82 is on the upper lip of the groove. Activity of the Y82C Eco-MutY mutant is severely compromised, and a mutation converting the corresponding tyrosine of a human MutY homologue into a cysteine is associated with familial adenomatous polyposis (Al-Tassan et al. 2002). The entrance into the active-site pocket is occupied by Q41, a residue that likely interacts with the base opposite A, thus directly contributing to MutY specificity (Guan et al. 1998). Deeper in the pocket, one finds E37 and A124. An E37S mutation completely inactivates the enzyme, and it has been proposed that this residue forms hydrogen bonds with the N7 and N^6 of the adenine to be excised (Guan et al. 1998).

Finally, Mth-Tdg (Fig. 4C) does not contain subgroup-specific residues in the postulated active-site pocket. However, R46 R47, and L87 cluster on the

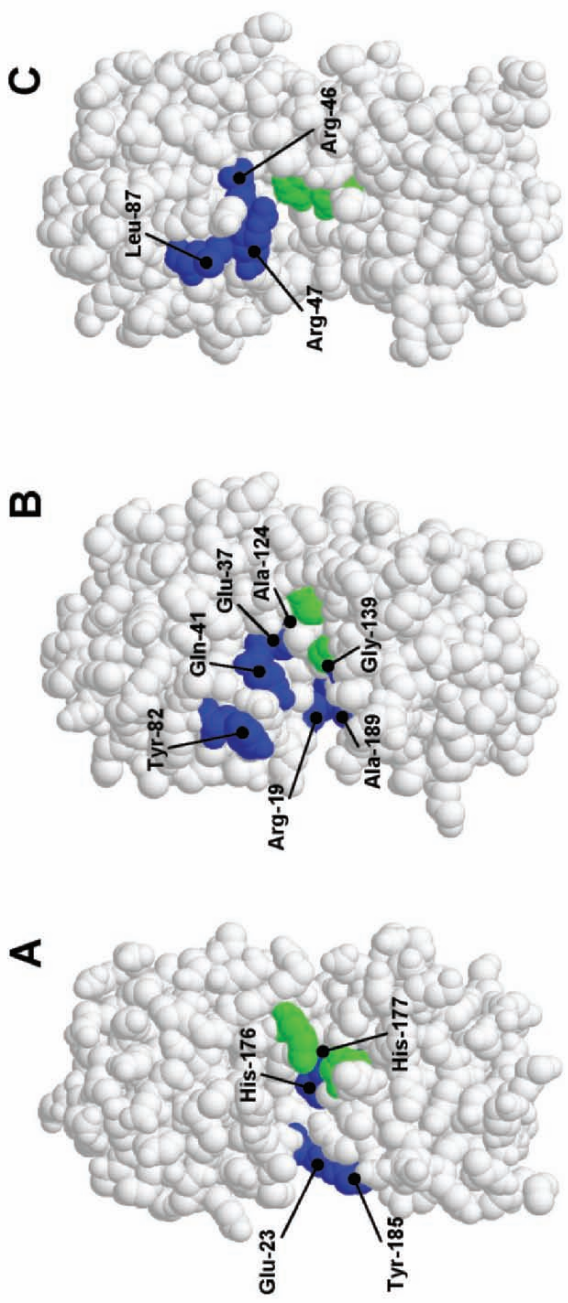


Fig. 4A–C. Structures of Eco-Nth, Eco-MutY, and Mth-Tdg with the subgroup-specific residues mapped on their surface. **A** Eco-Nth; **B** Eco-MutY; **C** Mth-Tdg. The protein molecules are oriented so that the six-helix barrel domain points upward, and the DNA-binding groove faces the reader. Selected subgroup-specific residues discussed in the text are highlighted in *blue* and *labeled*. The catalytic dyad 120/138 is shown in *green* for easier orientation

upper lip of the DNA-binding groove. It is suggested that R47 is inserted into the DNA double helix and assists in base flipping (Mol et al. 2002); mutation of this residue to alanine reduces enzymatic activity 20-fold.

In some cases, residues identified as being specific for Nth and MutY subgroups differ from the Nth- or MutY-specific residues identified earlier (Zharkov and Grollman 2002). This situation results from improvements in separation of enzyme groups with different substrate specificities and in sequence selection based on microbial genome search rather than on predefined clusters of orthologous groups (Tatusov et al. 1997). For example, Pdg enzymes are often annotated as Nth. Consequently, Pdg is one of three "endonuclease III proteins" found in the *D. radiodurans* genome and included as such in the Nth COG of the Clusters of Orthologous Groups Database (Tatusov et al. 2001). The two others relate less to Nth or Pdg than Nth and Pdg relate to each other. This ambiguous annotation leads to an artificial decrease in conservation of positions that are truly specific for Nth and thus hinders their identification. The present approach allows for better resolution of residues important for the specific function of each subgroup.

The currently accepted mechanism by which DNA glycosylases search for and "recognize" cognate lesions includes several steps where an enzyme could exert substrate specificity. In the initial encounter, the enzyme binds nonspecifically to DNA and moves along one or the other groove by facilitated one-dimensional diffusion (von Hippel and Berg 1989) until the lesion is encountered. Recognition of the lesion is accomplished through the action of a "reading head," a part of the enzyme directly involved in scanning DNA. The damaged base is then everted from the helix and stabilized through interactions in the active-site pocket. Interactions with the reading head and the binding pocket are likely to be different as, in some DNA glycosylases, canonical bases may not fit the pocket (Kavli et al. 1996). Residues located at the edges of the DNA-binding groove of Nth, MutY, or Tdg are good candidates for reading-head groups. These residues are often bulky and capable of intercalation between base pairs in the DNA duplex, as in the prototypical tyrosine/arginine reading head of uracil-DNA glycosylase (Parikh et al. 1998). Alternatively, residues positioned on the edge may detect atypical patterns of hydrogen bond donors and acceptors exposed in the major or minor groove of DNA, as proposed for *E. coli* Fpg protein (Grollman et al. 1994). Residues located within the active-site pocket likely stabilize the everted base through formation of specific hydrogen bonds. Interestingly, some glycosylases, such as Tdg or alkylpurine-DNA glycosylase AlkA (Zharkov and Grollman 2002), contain no specific amino acids in the active-site pocket. These enzymes may rely on nonspecific van der Waals contacts to stabilize the everted base (Labahn et al. 1996). Alternatively, the enzymes may form hydrogen bonds with amino acids that are not unique to the subgroup (Mol et al. 2002), in which case substrate specificity would most likely occur during the scanning step. Residues from both classes, identified by a combined structural and

bioinformatics approach, are primary candidates for site-directed mutagenesis studies designed to clarify their roles in determining substrate specificity.

Acknowledgements. I am grateful to Dr. Arthur Grollman for numerous discussions regarding the mechanism of DNA glycosylases, and to Annette Oestreicher for help in editing the manuscript. Analyses conducted while the author was in the Laboratory of Chemical Biology of the State University of New York, Stony Brook, New York, were supported by grant 47995 from the National Cancer Institute. He is currently supported by grants from the Wellcome Trust (UK), the Russian Foundation for Basic Research (02–04–49605) and the Russian Ministry of Education (PD02–1.4–469).

References

Al-Tassan N, Chmiel NH, Maynard J, Fleming N, Livingston AL, Williams GT, Hodges AK, Davies DR, David SS, Sampson JR et al (2002) Inherited variants of MYH associated with somatic G:C→T:A mutations in colorectal tumors. Nat Genet 30:227–232

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Aravind L, Walker DR, Koonin EV (1999) Conserved domains in DNA repair proteins and evolution of repair systems. Nucleic Acids Res 27:1223–1242

Asahara H, Wistort PM, Bank JF, Bakerian RH, Cunningham RP (1989) Purification and characterization of *Escherichia coli* endonuclease III from the cloned *nth* gene. Biochemistry 28:4444–4449

Begley TJ, Cunningham RP (1999) *Methanobacterium thermoformicicum* thymine DNA mismatch glycosylase: conversion of an N-glycosylase to an AP lyase. Protein Eng 12:333–340

David SS, Williams SD (1998) Chemistry of glycosylases and endonucleases involved in base-excision repair. Chem Rev 98:1221–1261

Eisen JA, Hanawalt PC (1999) A phylogenomic study of DNA repair genes, proteins, and processes. Mutat Res 435:171–213

Friedberg EC, Walker GC, Siede W (1995). DNA repair and mutagenesis. ASM Press, Washington, DC

Fromme JC, Verdine GL (2002) Structural insights into lesion recognition and repair by the bacterial 8-oxoguanine DNA glycosylase MutM. Nat Struct Biol 9:544–52

Fromme JC, Verdine GL (2003) Structure of a trapped endonuclease III–DNA covalent intermediate. EMBO J 22:3461–3471

Gilboa R, Zharkov DO, Golan G, Fernandes AS, Gerchman SE, Matz E, Kycia JH, Grollman AP, Shoham G (2002) Structure of formamidopyrimidine-DNA glycosylase covalently complexed to DNA. J Biol Chem 277:19811–19816

Goffin C, Ghuysen J-M (1998) Multimodular penicillin-binding proteins: an enigmatic family of orthologs and paralogs. Microbiol Mol Biol Rev 62:1079–1093

Grollman AP, Johnson F, Tchou J, Eisenberg, M (1994) Recognition and repair of 8-oxoguanine and formamidopyrimidine lesions in DNA. Ann N Y Acad Sci 726:208–214

Gu X (1999) Statistical methods for testing functional divergence after gene duplication. Mol Biol Evol 16:1664–1674

Gu X (2001) Maximum-likelihood approach for gene family evolution under functional divergence. Mol Biol Evol 18:453–464

Guan Y, Manuel RC, Arvai AS, Parikh SS, Mol CD, Miller JH, Lloyd RS, Tainer JA (1998) MutY catalytic core, mutant and bound adenine structures define specificity for DNA repair enzyme superfamily. Nat Struct Biol 5:1058–1064

Jiang D, Hatahet Z, Blaisdell JO, Melamede RJ, Wallace SS (1997a) *Escherichia coli* endonuclease VIII: cloning, sequencing and overexpression of the *nei* structural gene and characterization of *nei* and *nei nth* mutants. J Bacteriol 179:3773–3782

Jiang D, Hatahet Z, Melamede RJ, Kow YW, Wallace SS (1997b) Characterization of *Escherichia coli* endonuclease VIII. J Biol Chem 272:32230–32239

Karakaya A, Jaruga P, Bohr VA, Grollman AP, Dizdaroglu, M (1997) Kinetics of excision of purine lesions from DNA by *Escherichia coli* Fpg protein. Nucleic Acids Res 25:474–479

Kavli B, Slupphaug G, Mol CD, Arvai AS, Peterson SB, Tainer JA, Krokan H E (1996) Excision of cytosine and thymine from DNA by mutants of human uracil- DNA glycosylase. EMBO J 15:3442–3447

Kuo C-F, McRee DE, Fisher CL, O'Handley SF, Cunningham RP, Tainer JA (1992) Atomic structure of the DNA repair [4Fe-4S] enzyme endonuclease III. Science 258:434–440

Labahn J, Schärer OD, Long A, Ezaz-Nikpay K, Verdine GL, Ellenberger TE (1996) Structural basis for the excision repair of alkylation-damaged DNA. Cell 86:321–329

Lavrukhin OV, Lloyd RS (2000) Involvement of phylogenetically conserved acidic amino acid residues in catalysis by an oxidative DNA damage enzyme formamidopyrimidine glycosylase. Biochemistry 39:15266–15271

Livingstone CD, Barton GJ (1993) Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. Comput Appl Biosci 9:745–756

Livingstone CD, Barton GJ (1996) Identification of functional residues and secondary structure from protein multiple sequence alignment. Methods Enzymol 266:497–512

McCullough AK, Dodson ML, Lloyd RS (1999) Initiation of base excision repair: glycosylase mechanisms and structures. Annu Rev Biochem 68:255–285

Melamede RJ, Hatahet Z, Kow YW, Ide H, Wallace SS (1994) Isolation and characterization of endonuclease VIII from *Escherichia coli*. Biochemistry 33:1255–1264

Michaels ML, Pham L, Nghiem Y, Cruz C, Miller JH (1990) MutY, an adenine glycosylase active on G-A mispairs, has homology to endonuclease III. Nucleic Acids Res 18:3841–3845

Mol CD, Arvai AS, Begley TJ, Cunningham RP, Tainer JA (2002) Structure and activity of a thermostable thymine-DNA glycosylase: evidence for base twisting to remove mismatched normal DNA bases. J Mol Biol 315:373–384

Nash HM, Bruner SD, Shärer OD, Kawate T, Addona TA, Spooner E, Lane WS, Verdine GL (1996) Cloning of a yeast 8-oxoguanine DNA glycosylase reveals the existence of a base-excision DNA-repair protein superfamily. Curr Biol 6:968–980

O'Connor TR, Graves RJ, de Murcia G, Castaing B, Laval J (1993) Fpg protein of *Escherichia coli* is a zinc finger protein whose cysteine residues have a structural and/or functional role. J Biol Chem 268:9063–9070

Parikh SS, Mol CD, Slupphaug G, Bharati S, Krokan HE, Tainer JA (1998) Base excision repair initiation revealed by crystal structures and binding kinetics of human uracil-DNA glycosylase with DNA. EMBO J 17:5214–5226

Piersen CE, Prince MA, Augustine ML, Dodson ML, Lloyd RS (1995) Purification and cloning of *Micrococcus luteus* ultraviolet endonuclease, an *N*-glycosylase/abasic lyase that proceeds via an imino enzyme-DNA intermediate. J Biol Chem 270:23475–23484

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425

Serre L, Pereira de Jesus K, Boiteux S, Zelwer C, Castaing B (2002) Crystal structure of the *Lactococcus lactis* formamidopyrimidine-DNA glycosylase bound to an abasic site analogue-containing DNA. EMBO J 21:2854–2865

Tatusov RL, Koonin EV, Lipman DJ (1997) A genomic perspective on protein families. Science 278:631–637

Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res 29:22–28

Taylor WR (1986) Classification of amino acid conservation. J Theor Biol 119:205–218

Tchou J, Bodepudi V, Shibutani S, Antoshechkin I, Miller J, Grollman AP, Johnson F (1994) Substrate specificity of Fpg protein. Recognition and cleavage of oxidatively damaged DNA. J Biol Chem 269:15318–15324

Tchou J, Grollman AP (1995) The catalytic mechanism of Fpg protein. Evidence for a Schiff base intermediate and amino terminus localization of the catalytic site. J Biol Chem 270:11671–11677

Tchou J, Kasai H, Shibutani S, Chung M-H, Laval J, Grollman AP, Nishimura S (1991) 8-oxoguanine (8-hydroxyguanine) DNA glycosylase and its substrate specificity. Proc Natl Acad Sci USA 88:4690–4694

Tchou J, Michaels ML, Miller JH, Grollman AP (1993) Function of the zinc finger in *Escherichia coli* Fpg protein. J Biol Chem 268:26738–26744

Thayer MM, Ahern H, Xing D, Cunningham RP, Tainer JA (1995) Novel DNA binding motifs in the DNA repair enzyme endonuclease III crystal structure. EMBO J 14:4108–4120

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680

von Hippel PH, Berg OG (1989) Facilitated target location in biological systems. J Biol Chem 264:675–678

Wang Y, Gu X (2001) Functional divergence in the caspase gene family and altered functional constraints: statistical analysis and prediction. Genetics 158:1311–1320

Zaika EI, Perlow RA, Matz E, Broyde S, Gilboa R, Grollman AP, Zharkov DO (2004) Mutational analysis of substrate discrimination by formamidopyrimidine-DNA glycosylase. J Biol Chem(in press)

Zharkov DO, Golan G, Gilboa R, Fernandes AS, Gerchman SE, Kycia JH, Rieger RA, Grollman AP, Shoham G (2002) Structural analysis of an *Escherichia coli* endonuclease VIII covalent reaction intermediate. EMBO J 21:789–800

Zharkov DO, Grollman AP (2002) Combining structural and bioinformatics methods for the analysis of functionally important residues in DNA glycosylases. Free Radic Biol Med 32:1254–1263

Zharkov DO, Rieger RA, Iden CR, Grollman AP (1997) NH_2-terminal proline acts as a nucleophile in the glycosylase/AP-lyase reaction catalyzed by *Escherichia coli* formamidopyrimidine-DNA glycosylase (Fpg) protein. J Biol Chem 272:5335–5341

Zharkov DO, Shoham G, Grollman AP (2003) Structural characterization of the Fpg family of DNA glycosylases. DNA Repair 2:839–862

Subject Index