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