SPANDIDOS PUBLICATIONS INTERNATIONAL JOURNAL OF MOLECULAR MEDICINE 28: 295-310, 2011

State-of-the-art bioinformatics protein structure prediction tools (Review)

ATHANASIA PAVLOPOULOU[1] and IOANNIS MICHALOPOULOS[2]

[1]Department of Pharmacy, School of Health Sciences, University of Patras, Rion-Patras; [2]Cryobiology of Stem Cells, Centre of Immunology and Transplantation, Biomedical Research Foundation, Academy of Athens, Athens, Greece

Received April 18, 2011; Accepted May 9, 2011

DOI: 10.3892/ijmm.2011.705

Abstract. Knowledge of the native structure of a protein could provide an understanding of the molecular basis of its function. However, in the postgenomics era, there is a growing gap between proteins with experimentally determined structures and proteins without known structures. To deal with the overwhelming data, a collection of automated methods as bioinformatics tools which determine the structure of a protein from its amino acid sequence have emerged. The aim of this paper is to provide the experimental biologists with a set of cutting-edge, carefully evaluated, user-friendly computational tools for protein structure prediction that would be helpful for the interpretation of their results and the rational design of new experiments.

Contents

1. Introduction

Experimental determination of protein structure and function is becoming increasingly important, as proteins have attracted interest as drug targets. Although the large-scale sequencing projects have generated an abundance of protein sequence data, the experimental determination of a protein structure and/or function is labour intensive, time consuming and expensive. As a result, the 'sequence-structure gap', the gap between the number of protein sequences and the number of proteins with experimentally determined structure and function, is growing rapidly. Therefore, the use of computational tools for assigning structure to a novel protein represents the most efficient alternative to experimental methods (1). To overcome this problem, a plethora of automated methods to predict protein structure have evolved as computational tools over the past decade (2).

In the present review, we present a set of state-of-the-art bioinformatics tools that cover most aspects of protein structure prediction, including automated methods for primary, secondary and tertiary structure prediction from the amino acid sequence of the query protein alone. Each of these methods has its strengths and limitations. The scope of this paper is to provide an overview of the tools available for protein structure prediction and offer suggestions on how to use these tools more efficiently.

2. Phylogenetic analysis

Sequences which have diverged from a single common ancestor (homologs) tend to have similar structure and subsequently function. Homologous sequences that have risen after speciation are called orthologs, and they tend to have similar functions; sequences rising from gene duplication are called paralogs, and they tend to have different function (3). Therefore, the first step in inferring the structure and/or function of a novel protein is to compare this protein with that of an evolutionarily related protein of known structure. The shared evolutionary origins of sequences are assessed by phylogenetic analysis.

Biological sequence databases. The main repositories of biological sequences are the publicly available sequence databases. Data mining and analysis tools are provided in these databases.

The nucleotide databases independently store data derived from sequencing projects. The primary public and comprehensive repositories of nucleotide sequence entries are: GenBank (4), the European Molecular Biology Laboratory Nucleotide Sequence Database (EMBL) (5) and the DNA DataBank of Japan (DDBJ) (6). These are members of the

Correspondence to: Dr Ioannis Michalopoulos, Cryobiology of Stem Cells, Centre of Immunology and Transplantation, Biomedical Research Foundation, Academy of Athens, Soranou Efessiou 4, 11527 Athens, Greece
E-mail: imichalop@bioacademy.gr

Key words: protein structure prediction, multiple sequence alignment, pairwise sequence similarity search, phylogenetic analysis

Table I. Major sequence databases.

Database	Comments	Web link
Nucleotide sequence databases		
DDBJ	Primary sequence repository in Japan	http://www.ddbj.nig.ac.jp/
EMBL	Primary sequence repository in Europe	http://www.ebi.ac.uk/Databases/
GenBank	Primary sequence repository in the USA	http://www.ncbi.nlm.nih.gov/Genbank/
Genome sequence databases		
ENSEMBL	Analysis and annotation of metazoan genomes	http://www.ensembl.org/
Entrez genome	Analysis and annotation of genomes from plasmids, viruses, archaea, bacteria and eukaryotes	http://www.ncbi.nlm.nih.gov/sites/entrez?db=genome
Protein sequence databases		
GenPept	Translations of GenBank coding nucleotide entries	http://www.ncbi.nlm.nih.gov/Genbank/
PIR	International protein database	http://pir.georgetown.edu/
RefSeq	Curated, non-redundant with expert annotation	http://www.ncbi.nlm.nih.gov/RefSeq/
UniProt/SwissProt	Reviewed, manually annotated entries	http://www.uniprot.org/help/uniprotkb
UniProt/TrEMBL	Automatically classified and annotated entries	http://www.uniprot.org/help/uniprotkb
Specialized databases		
FlyBase	Integrated genetic and genomic data on *Drosophila*	http://flybase.org/
JGI	Analysis and annnotation of all publicly available microbial genomes from eukaryotes, bacteria and archaea	http://img.jgi.doe.gov/cgi-bin/pub/main.cgi
Protein classification databases		
CATH	Proteins classified based on class, architecture, topology and homology	http://www.cathdb.info/
SCOP	Structural classification of proteins	http://scop.mrc-lmb.cam.ac.uk/scop
ProtClustDB	Proteins classified based on sequence similarity	http://www.ncbi.nlm.nih.gov/proteinclusters
Protein structure databases		
PDB	Resolved 3D biomolecular structures	http://www.rcsb.org/pdb

International Nucleotide Sequence Database Collaboration (INSDC) and they are cross-referenced against each other on a daily basis.

The protein databases contain amino acid sequences derived from translations of the sequences stored in the nucleotide databases or resolved protein structures. The major protein sequence databases are GenPept (4), RefSeq (7), the Protein Information Resource (PIR) (8), the UniProt Knowledgebase (UniProtKB) (9), which consists of the non-redundant, manually curated UniProtKB/Swiss-Prot and its computer-annotated supplement, UniProtKB/TrEMBL, which contains protein sequences translated from the EMBL nucleotide sequence database.

The specialized genome databases contain partial or complete genomes of different organisms. Examples of specialized genome databases are FlyBase (10) and JCI (11). Examples of genome databases are ENSEMBL (12) and Entrez Genome (13).

The Protein Data Bank (PDB) (14) is the universal repository for the three-dimensional structural data of biological macromolecules (proteins and nucleic acids), typically obtained by X-ray crystallography and NMR spectroscopy, cryoelectrical microscopy and theoretical modelling. The data deposited in PDB contain three-dimensional coordinates of the deposited protein structures, information regarding the method used for the structure determination and general information such as the names of all components in the deposited structure, full sequence of all macromolecular components, literature citations, chemical structures of cofactors and prosthetic groups. The main structure protein classification schemes are the automated CATH (15) and manually curated SCOP (16); they are both hierarchical and despite their differences are in agreement

Sequences producing High-scoring Segment Pairs:	High Score	Smallest Sum Probability P(N)	N
ENSXETP00000006440 pep:known scaffold:JGI4.1:scaffold_698...	1974	7.8e-192	1
ENSXETP00000006442 pep:known scaffold:JGI4.1:scaffold_698...	1966	4.9e-191	1
ENSXETP00000006282 pep:known scaffold:JGI4.1:scaffold_153...	1799	1.8e-174	1
ENSXETP00000029682 pep:novel scaffold:JGI4.1:scaffold_216...	1497	1.7e-144	1
ENSXETP00000029684 pep:known scaffold:JGI4.1:scaffold_216...	1322	3.8e-127	1

Figure 1. Sample BLAST output. In this case, the protein sequence hKLK1 was used as query to search the ENSEMBL database. Lower probability values (P) indicate higher probabilities that the results are accurate.

about which proteins should be assigned to the same group. Furthermore, the Protein Clusters Database (ProtClustDB) (17) contains clusters of related proteins encoded by complete eukaryotic and prokaryotic genomes derived from RefSeq (Table I).

Amino acid substitution models. During the course of evolution, protein sequences are prone to changes in the form of substitutions between amino acid residues with similar physicochemical properties (e.g. aromatic, polar, acidic). These changes are not detrimental to the overall structure and function of a protein. The amino acid substitution models estimate the replacement rate of an amino acid residue by another, and they are essential in the different steps of phylogenetic analysis (see below). The amino acid substitution models used to estimate the substitution probabilities are essential in phylogenetic analysis. The Point Accepted Mutation (PAM) matrix (18) is calculated based on the alignments of closely related amino acid sequences. The PAM matrix estimates the rate of substitution that would be expected if 1% of the amino acid sequence had changed. The BLOck SUbstitution Matrix (BLOSUM) (19) series of matrices are calculated based on conserved, functionally important, blocks found in aligned distantly-related amino acid sequences. For example, the BLOSUM62 matrix is calculated from observed substitutions between proteins that share 62% sequence identity. Thereby, the higher numbered BLOSUM matrices (e.g. BLOSUM80) are used for aligning closely related amino acid sequences and, conversely, the lower numbered (e.g. BLOSUM45) for distantly related sequences. The model JTT (20) estimates amino acid substitution rates from a set of closely related pairs of amino acid sequences (85% identity). The observed amino acid changes for each sequence pair are counted and processed. The results of all sequence pairs are averaged.

Pairwise sequence similarity searches. Pairwise sequence similarity search methods are employed in order to search the sequence databases for protein sequences (templates) similar to the query (target) protein; the target sequence is aligned with each of the template sequences in a database. There are two main methods for pairwise sequence alignment: local and global. The former method identifies local regions of similarity within sequences that are overall dissimilar to the query sequence. The latter method aligns the entire length of the query sequence to the database sequences.

The computational tools for local alignment include BLAST (21), and SSEARCH, which has implemented the Smith-Waterman algorithm (22). The filter parameter in BLAST is employed to mask off regions of the query sequence that have low compositional complexity. Moreover, in BLAST searches, the amino acid substitution matrices BLOSUM45 and PAM70 could be chosen to detect sequences distantly related to the query; matrices BLOSUM80 and PAM30 would be recommended for the identification of highly conserved regions. There are different variants of BLAST, including the position-specific iterated BLAST (PSI-BLAST) (23) which uses position-specific scores derived for the multiple alignment of homologous sequences to search for related sequences. However, the potential problem with PSI-BLAST is contamination with sequences unrelated to the query sequence. An example of the use of BLAST is shown in Fig. 1. The filter parameter was chosen. The human kallikrein 1 (KLK1) protein sequence was used as a query to search the genome of *Xenopus tropicalis* (frog) in the ENSEMBL database. The hit with the highest score was the sequence ENSXETP00000006440 which, in a previous study (24), was shown to be a *bona fide* KLK1.

The tools for global alignment include FASTA (25) and GGSEARCH which is based on the Needleman-Wunsch algorithm (26) (Table II). Reciprocal searches are employed to identify true orthologs. In a reciprocal search, a query sequence from database A is searched against database B. The highest-scoring sequence from B is then searched against database A. If this returns the sequence originally used as the highest scorer, then the two sequences are considered true orthologs.

Multiple sequence alignment. Accurate multiple sequence alignment (MSA) is a critical step in phylogenetic reconstruction (27). The most widely used method for aligning multiple sequences is the 'progressive sequence alignment' method (28). This process involves the construction of a crude 'guide tree' which determines the order in which the sequences are added to the alignment, starting with the most closely related sequences and progressively adding the more distant. This method suffers from the drawback that misalignments made early in the process cannot be rectified ('once a gap always a gap'). CLUSTALW (29) is the most popular implementation of this method. ProbCons (30), which is based on probabilistic consistency, also uses this method. T-Coffee (31) rectifies the mistakes made in the progressive alignment: a library of pairwise alignments is constructed which is represented as a list of aligned residue pairs and the information in this library is used to perform progressive alignment. Programs like MUSCLE (32) and MAFFT 5.3 (33,34) apply extensive iterative refinement to improve the classical progressive alignment

Table II. Sequence similarity search web-tools.

Program	Comments	Web link
BLAST	Basic local alignment search tool	http://blast.ncbi.nlm.nih.gov/Blast.cgi
FASTA	Global alignment search tool; recommended for distant homologs	http://www.ebi.ac.uk/Tools/fasta33/
GGSEARCH	Global alignment search tool	http://www.ebi.ac.uk/Tools/fasta33/index.html?program=GGSEARCH
SSEARCH-Protein	Local alignment search tool against proteins	http://www.ebi.ac.uk/Tools/fasta33/index.html?program=SSEARCH
SSEARCH-Proteomes	Local alignment search tool against proteomes	http://www.ebi.ac.uk/Tools/fasta33/proteomes.html?program=SSEARCH

	Query	Subject	Comments
BLAST variants			
BLASTN	Nucleotide	Nucleotide	Identification of the most similar DNA sequences
BLASTP	Protein	Protein	Identification of the most similar protein sequences
TBLASTN	Protein	Nucleotide[a]	Identification of non-annotated coding DNA sequences
BLASTX	Nucleotide[a]	Protein	Identification of novel DNA sequences and ESTs
TBLASTX	Nucleotide[a]	Nucleotide[a]	EST identification
PSI-BLAST	Protein	Protein	Identification of distant homologs in a protein family

[a]Six-frame conceptual translation. EST, expressed sequence tag.

Table III. Multiple sequence alignment web-tools.

Program	Comments	Web link
CLUSTALW	Progressive alignment; widely-used	http://www.ebi.ac.uk/tools/clustalw2
MAFFT	Progressive alignment; accurate	http://www.ebi.ac.uk/Tools/mafft/
MUMMALS	Use of HMM and local structural information; improved alignment	http://prodata.swmed.edu/mummals/mummals.php
MUSCLE	Progressive alignment; accurate	http://www.ebi.ac.uk/Tools/muscle/index.html
ProbCons	Probabilistic consistency	http://probcons.stanford.edu/
PROMALS	Use of database searches and structure information; distant homologs alignment	http://prodata.swmed.edu/promals/promals.php
PROMALS3D	Use of 3D structural information; improved alignment	http://prodata.swmed.edu/promals3d/promals3d.php
T-Coffee	Progressive alignment; accurate, sequences limit, high CPU time	http://tcoffee.vital-it.ch/cgi-bin/Tcoffee/tcoffee_cgi/index.cgi?stage1=1&daction=TCOFFEE::Regular
3D-Coffee	Use of 3D structural information; improved alignment	http://www.phylogeny.fr/version2_cgi/one_task.cgi?task_type=expresso

method. Given that homologous proteins sharing a small degree of sequence identity tend to share a similar structure, several MSA programs incorporate structural information in order to generate improved alignments (especially of distantly related sequences). For example, MUMMALS (35) improves alignment quality by exploring the use of Hidden Markov models (described in detail later in the text) that describe local structural information. The program PROMALS (36) explores the use of PSI-BLAST sequence database searches and secondary structure information for accurately aligning distantly related protein sequences. PROMALS3D (37) is a derivation of PROMALS and uses three-dimensional information in order to improve alignment accuracy. Likewise, a T-Coffee variant, 3D-Coffee (38), combines sequence alignment and uses three-dimensional structure-sequence information for improved alignment (Table III).

Table IV. Phylogenetic reconstruction programs.

Program	Methods	Comments	Web link
MEGA	Distance, maximum parsimony, maximum composite likelihood	Molecular evolutionary genetics analysis software	http://www.megasoftware.net
MrBayes	Bayesian	Fast; estimates phylogeny from large datasets	http://mrbayes.csit.fsu.edu/
PAUP*	Maximum parsimony, distance, maximum likelihood	Phylogenetic analysis using parsimony (*and other methods)	http://paup.csit.fsu.edu/
PHYLIP	Distance, maximum parsimony, maximum likelihood	PHYLogenetic inference package	http://evolution.genetics.washington.edu/phylip.html
PhyML	Maximum likelihood	Fast and accurate method	http://www.atgc-montpellier.fr/phyml/
RAxML	Maximum likelihood	Estimates phylogeny from large datasets	http://phylobench.vital-it.ch/raxml-bb/
SplitsTree	Split decomposition	Well-resolved branches	http://www.splitstree.org/

	Comments	Web link
Tree visualization programs		
Dendroscope	Interactive visualization and editing of large phylogenetic trees	http://www-ab.informatik.uni-tuebingen.de/software/dendroscope
iTOL	Visualization, manipulation and annotation of phylogenetic trees; interactive pruning and collapsing	http://itol.embl.de
NJPlot	Visualization and manipulation of phylogenetic trees	http://pbil.univ-lyon1.fr/software/njplot.html
TreeView	Visualization and manipulation of phylogenetic trees	http://taxonomy.zoology.gla.ac.uk/rod/treeview.html

Construction of phylogenetic trees. Proteins identified by sequence similarity searches using a query protein, can be used to perform phylogenetic analyses in order to resolve their evolutionary relationships (i.e. if they are orthologs or paralogs).

Phylogenetic trees are used to depict evolutionary relationships. The methods for re-constructing phylogenetic trees are divided into two major categories: distance-matrix and tree-searching methods. The former (e.g. Neighbor Joining, UPGMA etc.) infer a phylogenetic tree by calculating the distance (defined as the percentage difference) of all combinations of sequence pairs. The latter searches for the tree that best fits the information present in each column of the multiple sequence alignment. Examples of tree-searching methods are the Maximum Parsimony which searches for the tree with the minimum total length and the Maximum Likelihood, which searches for the tree with the greatest probability or likelihood of observing method which is based on the posterior probability principle: the probability that is estimated based on some prior expectations which may simply be the expectation of any tree from all the possible trees that could be obtained from the given dataset.

The most comprehensive phylogenetic software packages are PHYLIP (39), PAUP* (40) and the user-friendly MEGA (41,42). PhyML (43,44) and RAxML (45) are based on the maximum likelihood method. MrBayes (46) performs Bayesian inference of phylogeny. The SplitsTree (47) uses the split decomposition method to infer phylogeny (48). The interpretation of the inferred phylogenetic tree is facilitated by its graphical representation. Tree visualization programs such as TreeView (49), NJPlot (50), Dendroscope (51) and the web-based iTOL (52) allow the display of various tree shapes, as well as editing and manipulation of trees (Table IV).

Bootstrapping. The reliability of the inferred tree is tested using the bootstrapping process, in which random subsamples are taken from the original dataset; individual trees are built from each of these which are scored from the frequency of node appearance (53,54). Jack-knifing (55) is a similar technique, according to which 50% of the original dataset is re-sampled.

3. Protein primary structure prediction

The basic information about the structure of a protein comes from its primary sequence. The first step in the analysis of the protein primary sequence is to divide it into its constituent parts (domains) and handle each one of them separately (56). The domains are often defined as compact, spatially distinct

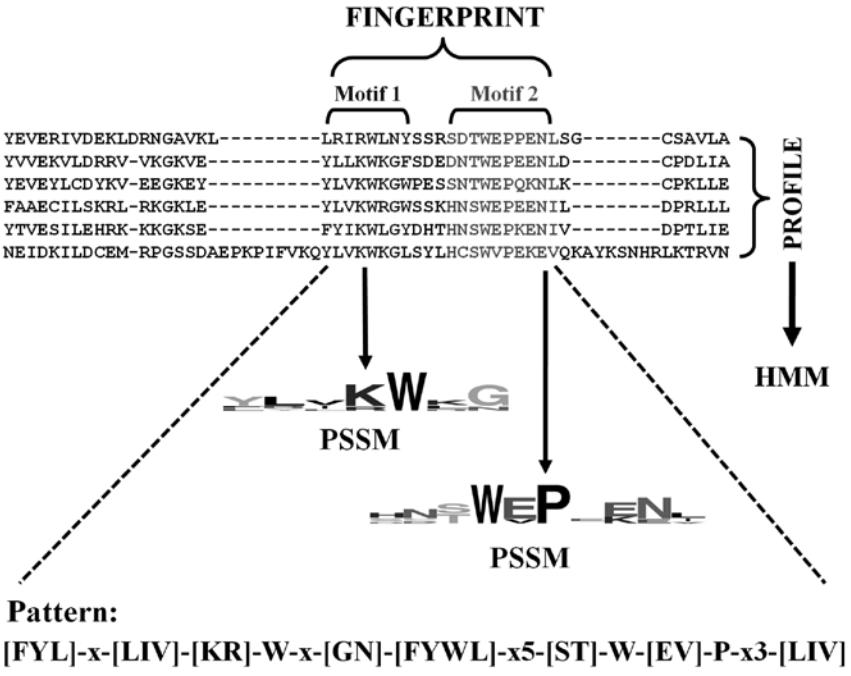

Figure 2. Schematic representation of the different types of signatures describing the chromo domain family. The motif is a single conserved region. A fingerprint is used to describe a group of motifs. PSSM are generated by adding a scoring to motifs. Each PSSM column corresponds to a motif position and contains values based on the amino acid residue frequencies at each position. Weblogo was employed for the generation of the PSSM sequence logos; the height of each amino acid residue depicts the frequency of the corresponding residue, and the letters are ordered so that the most frequent one is on the top. The patterns represent the core functional and structural features of the sequence. The square brackets and x(n) indicate alternative amino acids at each position in the pattern. The residue W is a key catalytic residue of the chromo domain family and therefore there are no alternative residues at this position. Profiles are the complete conserved regions including gaps. The Hidden Markov models are statistical models derived from the profiles.

functional units, which appear in a variety of otherwise unrelated proteins. In biochemistry, the domains are defined as protein regions with determined experimental functions. Many proteins with a complex function consist of a combination of interacting and cooperating domains (57,58).

Protein domains or protein family databases are useful for the assignment of function to uncharacterized proteins. These databases are often called 'signature databases' because they contain collections of 'signatures' which are consensus representations of different domain types or protein families deduced from multiple sequence alignments (59,60). Signatures may be diagnostic of structure or function, and they are derived using a number of different methods which are briefly discussed (Fig. 2). Motifs (or blocks) are ungapped multiple sequence alignments, typically 10-20 amino acids in length. A set of motifs representing a protein domain family is called a 'fingerprint'. The principle advantage of motifs/fingerprints is that they can detect distant sequence relationships (61). The position-specific score matrix (PSSM) is a common representation of motifs. A PSSM calculates scores at each position in the motif independently from the other positions (62). Any information from a single conserved region reduced into a consensus sequence (or regular expression) results in the so-called sequence patterns. Due to their shortness, patterns are restricted to detect the most conserved protein regions (63). The sequence profiles (also referred to as gapped weight matrices) describe larger conserved sequence fragments that include variable regions which may contain useful information (64). They are used for sensitive detection of larger domains (64). In the Hidden Markov Models (HMMs), which are related to sequence profiles, each position of a sequence can be described as a match, insert or delete state; this allows the query sequence to be aligned by assigning to its amino acids the most probable state transition (65).

The methods above have given rise to a number of signature databases (Table V). Each of these signature databases has different diagnostic strengths and it is cross-referenced to other databases in order to provide complementary information (61). The databases BLOCKS (66) and PRINTS (67) are based on motifs and fingerprints, respectively. PSSM models are used by the Conserved Domains Database (CDD) (68). PROSITE (69) is based on both sequence patterns and sequence profiles. HMMs have been adopted by the CDD (68), Gene3D (70), PANTHER (71) Pfam (72), PIRSF (73), SMART 6 (Simple Modular Architecture Research Tool release 6) (74), SUPERFAMILY (75) and the TIGRFAMs (76) signature databases. There are databases which identify protein domain families using sequence clustering. For instance, ProDom (77) is created automatically from databases of known protein domain family sequences using BLAST followed by clustering together of similar sequence fragments from different proteins; the resulting protein domain families are aligned using MultAlin (78), a program which aligns very large sequence families. InterPro is a meta-site which combines several major signature databases (79). There are types of domains for which a signature is not easy to determine, due to the weak similarity between their sequences. SBASE 12.0, a collection of protein sequence fragments with known structure and/or function, cross-references to all major sequence databases and several signature databases, and overcomes this

Table V. Main databases of protein signatures.

Database	Signature type	External source	Web link
BLOCKS	Blocks		http://blocks.fhcrc.org/blocks/
CDD	HMM, MSA[a]	Pfam, SMART, COGs, ProtClustDB	http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml
Gene3D	HMM	CATH	http://gene3d.biochem.ucl.ac.uk/Gene3D/
InterPro	Integrated signature types of its member databases	Gene3D, PANTHER, Pfam, PIRSF, PRINTS, ProDom, PROSITE, SMART, SUPERFAMLY, TIGRFAMs	http://www.ebi.ac.uk/interpro/
Pfam	HMM, MSA[a]	UniProtKB, GenPept, metagenomics datasets	http://pfam.sanger.ac.uk/
PRINTS	Fringerprints[b]		http://www.bioinf.manchester.ac.uk/dbbrowser/PRINTS/index.php
ProDom		UniProtKB, SCOP	http://prodom.prabi.fr/prodom/current/html/home.php
PROSITE	Patterns, profiles	UniProtKB/SWISS-PROT	http://au.expasy.org/prosite/
SBASE		BLOCKS, Pfam, PRINTS, ProDom, PROSITE	http://hydra.icgeb.trieste.it/sbase/
SMART	HMM[b]		http://smart.embl.de/
SUPERFAMILY	HMM	SCOP	http://supfam.org/SUPERFAMILY/

	Features/supplements	Web link
Pfam	Pfam A: manually curated PfamB: automatic generation of HMMs from ProDom	http://pfam.sanger.ac.uk/
PRINTS	prePRINTS: automatic supplement	http://www.bioinf.manchester.ac.uk/dbbrowser/prePRINTS/index.php
ProDom	ProDom-SG: selects candidate proteins for structural genomics	http://prodom.prabi.fr/prodom/current/html/formSG.php

[a]Multiple sequence alignment. [b]Manually-curated.

problem by performing BLAST searches and incorporating biological information derived from known protein domain groups (80).

Signature databases can be queried with novel protein sequences via search engines available in these databases (Table VI). Various search algorithms are utilized by these tools. For example, SBASE 12.0 (80) uses pairwise sequence similarity methods to identify close homologues in the signature databases. On the contrary, FingerPRINTScan (81) approach exploits the contextual information contained in the multiple motifs within a fingerprint to identify distantly-related homologs in the databases. The InterProScan search tool (82) allows a simultaneous search of its member databases. However, the user is advised to refer to the original databases to obtain richer information. For instance, BLOCKS utilizes three different search engines, namely the 'traditional' Block Searcher and two of its variants: the Block Searcher IMPALA and the Block Searcher RPS-BLAST. The former variant utilizes Integrating Matrix Profile and Local Alignment (IMPALA) algorithm (83) to compare a query protein against the PSSM models which represent the blocks in the BLOCKS database. The latter variant is using Reverse-Position-Specific BLAST (RPS-BLAST), a PSI-BLAST variant algorithm (23), to compare a protein query sequence against the signature databases. Furthermore, ScanProsite (84) is supplemented by a context-dependent annotation transfer system, called ProRule (85) in order to detect intra-domain features, such as active sites, substrate binding sites and disulfide-bridges. Several databases also provide rich graphical outputs, such as: highlighting the intra-domain conserved elements, distribution of domain families across the major taxonomic groups, interactive networks where the interaction patterns of the particular domain are displayed etc.

A consensus sequence represents the frequency of a residue in a particular position in a multiple sequence alignment. The sequence logos are used for the graphical representation of these frequencies. The web-based application Weblogo (86) is used for the generation of sequence logos (Fig. 2).

Table VI. Web-based tools for searching against signature databases.

Tool	Database	Search algorithm	Output	Web link
Block searcher	BLOCKS	Block PSSM search	Top ranking signature display	http://blocks.fhcrc.org/blocks/blocks_search.html
Block searcher IMPALA	BLOCKS	IMPALA	Top ranking signature display	http://blocks.fhcrc.org/blocks/impala.html
Block searcher RPS-BLAST	BLOCKS and PRINTS	RPS-BLAST	Top ranking signature display	http://blocks.fhcrc.org/blocks/rpsblast.html
CD-Search	CDD	RPS-BLAST	Domain boundaries display; rich graphical output	http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi
FingerPRINTScan	PRINTS	FingerPRINTScan	Top ranking signature display	http://www.bioinf.manchester.ac.uk//cgi-bin/dbbrowser/fingerPRINTScan/muppet/FPScan.cgi
InterProScan	InterPro member databases	Integrated search algorithm of its member databases	Predictions from InterPro member databases combined	http://www.ebi.ac.uk/Tools/InterProScan/
Pfam	Pfam	HMM	Domain boundaries display; rich graphical output	http://pfam.sanger.ac.uk/
ProDom	ProDom	MultAlin	Display of all matching domain arrangements	http://prodom.prabi.fr/prodom/current/html/form.php
SBASE	SBASE	BLAST	Domain boundaries display	http://hydra.icgeb.trieste.it/servers/protein/sbase/
ScanProsite	PROSITE	ProRule	Domain boundaries display; rich graphical output	http://www.expasy.org/tools/scanprosite/
SMART	SMART	HMM	Domain boundaries display; rich graphical output	http://smart.embl.de/

B

FingerPrint Name	Motif Number	IdScore	PfScore	Pval	Sequence	Length	low	Pos	high
CHYMOTRYPSIN	1 of 3	63.74	566	1.50e-09	GGVLINKWWVLTAAHC	16	0	35	0
	2 of 3	43.64	290	3.86e-07	TYNNDIMLLKLASQA	15	0	89	0
	3 of 3	68.09	495	2.94e-08	EDTCVGDSGGPLV	13	0	180	0

Figure 3. Primary structure analysis of ENSXETP00000006440. The output of (A) ScanProsite and (B) FingerPRINTScan predictors. The predicted domain and its boundaries are shown in (A). The three residues (H49, D93, S187) are indicated by squares (A). The vertical grey lines connected by horizontal lines represent the four putative disulfide bridges. The query sequence's matching regions against the PRINTS motifs are shown (B).

To identify the domain organization of the protein sequence ENSXETP00000006440, its amino acid sequence (downloaded from ENSEMBL) was searched against the protein signature databases using the ScanProsite and the FingerPRINTScan search engines. Default parameters were chosen. These tools were also used to detect other biological features in the query protein sequence such as active sites, disulfide bonds etc. A single known domain, namely the (chymo)trypsin-like serine protease protein domain was identified in the query by comparing the outputs of ScanProsite. A significant match to all three motifs held in PRINTS for this protein domain family (accession code: PR00722) was found (Fig. 3).

4. Protein secondary structure prediction

The secondary structure (SS) of a protein is defined by the patterns of hydrogen bonds between the backbone amide and the carboxyl groups. They have a regular geometry, restrained to allowed values of the dihedral angles ψ and ϕ on the Ramachandran plot (87). The SS is often defined in three

Table VII. Web-based tools for protein secondary structure prediction.

Tool	Comments	Web link
CDM	FDM + GOR	http://gor.bb.iastate.edu/cdm/
FDM	PDB mining for structural fragments	http://gor.bb.iastate.edu/cdm/
GOR	Information theory, Bayesian statistics, PSSM profiles	http://gor.bb.iastate.edu/cdm/
Jpred	HMM and PSSM profiles; NNs; RSA	http://www.compbio.dundee.ac.uk/www-jpred/
PHD	Multiple sequence alignments; NNs	http://www.predictprotein.org/
PORTER	PSSM profiles; NNs	http://distill.ucd.ie/porter/
PSIPRED	PSSM profiles; NNs	http://bioinf.cs.ucl.ac.uk/psipred/
SABLE	PSSM profiles; NNs; RSA	http://sable.cchmc.org/
SSpro	PSSM profiles; NNs and SVMs; RSA; 8-state prediction	http://www.ics.uci.edu/~baldig/scratch/

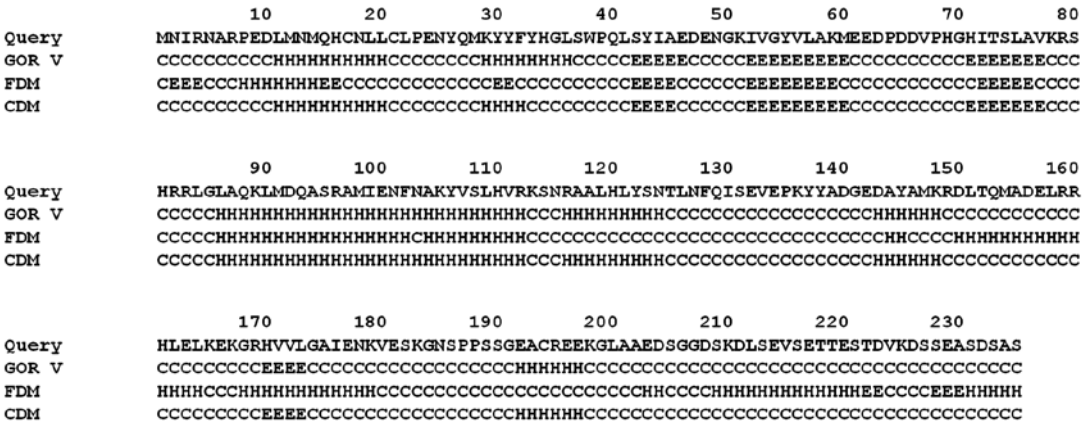

Figure 4. Secondary structure prediction of the protein N-acetyltransferase using CDM. H, α-helix; E, β-strand; C, coil.

conformational states, helix (H), β-strand (E) or coil (C). To extract as much information available in the atomic structure of a protein, the dictionary of protein secondary structure (DPSS) (88) defines eight states of SS: H (α-helix), G (3_{10}-helix) and I (π-helix) for helices; E (extended strand in parallel and/or anti-parallel β-strand conformation) and B (β-bridge) for β-strands; S (bend), T (turn) and C (coil) for coils (88). The protein SS prediction is proposed to be an intermediate step in the tertiary structure prediction when known or homologous three-dimensional structures are not available in the PDB.

The first step in the protein SS prediction is to search the PDB for proteins of experimentally determined tertiary structure which are homologous to the query sequence. For example, the Fragment Database Mining (FDM) (89) method performs a BLAST query for a given sequence against the PDB and collects structural fragments with sequence similarity to the query (Table VII).

Earlier methods for the prediction of protein SS were based on single amino acid propensities, i.e. certain amino acid residues have a higher probability to be in a particular SS state than other residues. For example, leucine, isoleucine and valine are usually found in β-strands (90). The Garnier-Osguthorpe-Robson (GOR) (91) method is based on this approach. GOR has been improved by including Bayesian statistics and considering pairwise interactions of the target amino acid and its flanking residues (92). The Consensus Data Mining (CDM) combines the advantages of the FDM and GOR methods (93). FDM is recommended for SS prediction when the template fragments available in the PDB are highly similar to the target sequence, whereas GOR is successfully used when the sequence similarity is low (Table VII).

Given that proteins with >30% sequence identity adopt similar structures (94), many methods have significantly improved the overall SS prediction by incorporating evolutionary information in the form of multiple sequence alignments, such as PHD (95-97), or sequence profiles (HMM and PSSM profiles) (91,98-102). The prototypic method that implements profiles is PSIPRED (102), which pioneered the use of PSI-BLAST output profiles (PSSM). In order to avoid contamination of the profile with unrelated proteins, the database search is first filtered (102). Notably, evolutionary information resulting from larger training sets and better search strategies increased the prediction accuracy (103) (Table VII).

Moreover, many modern methods for SS prediction are based on machine learning techniques such as Support Vector Machines (SVMs) and Neural Networks (NNs) trained with sequence profiles or SS information of resolved structures deposited in the PDB in order to achieve higher prediction accuracy (95-102) (Table VII).

Table VIII. Web-based tools for transmembrane protein prediction.

Tool	Method	Predicts	Web link
DAS-TMfilter	DAS	AHTM	http://mendel.imp.ac.at/sat/DAS/DAS.html
MINNOU	RSA/SS	AHTM and TMB	http://minnou.cchmc.org/
PRED-TMMB	HMM	TMB	http://bioinformatics.biol.uoa.gr/PRED-TMBB/input.jsp
PRED-TMR	Hydrophobicity profile	AHTM and TMB	http://athina.biol.uoa.gr/PRED-TMR/input.html
SOSUI	Hydropathy scale	AHTM	http://bp.nuap.nagoya-u.ac.jp/sosui/sosui_submit.html
TMBETA-NET	Amino acid composition; NNs	TMB	http://psfs.cbrc.jp/tmbeta-net/
TMB-Hunt	*k*-NN algorithm	TMB	http://bmbpcu36.leeds.ac.uk/~andy/betaBarrel/AACompPred/aaTMB_Hunt.cgi
TMMOD	HMM profile	AHTM	http://liao.cis.udel.edu/website/servers/TMMOD/scripts/frame.php?p=submit
TSEG	Tandem clusters of membrane proteins	AHTM and TMB	http://www.genome.ad.jp/SIT/tsegdir/tseg_exe.html

Several SS prediction methods also consider the relative solvent accessibility (RSA) of the amino acid residues in proteins (101) (Table VII); RSA measures the degree to which a residue is accessible to the solvent. Chan and Dill (104) demonstrated that the burial of core residues is a strong driving force in protein folding. Therefore, differentiating between exposed and buried residues further improves SS prediction. The results of the prediction of the SS of the human protein N-acetyltransferase (GenBank Accession number: NP_003482) are shown in Fig. 4.

Transmembrane protein prediction. Transmembrane (TM) proteins span the entire lipid membrane (105,106). TM proteins are implicated in various important biological functions, such as transmembrane transport, cell signaling and energy production. TM proteins are divided into α-helical TM (AHTM) and TM β-barrel (TMB) proteins. The AHTM proteins are located in the inner membranes of bacterial cells and the plasma membranes of eukaryotes. Their membrane spanning segments are formed by α-helices connected by polar loops (106). The TMB proteins are the least characterized and, at present, they have been found in the outer membrane of Gram-negative bacteria and presumably in the outer membrane of mitochondria and chloroplasts. Their membrane-spanning segments are antiparallel β-strands which form a barrel-like channel (107).

However, it is difficult to determine the three-dimensional structure of TM proteins by applying experimental methods, such as X-ray crystallography and NMR, and, thus, only a limited number of TM proteins have their 3D structure resolved compared to the number of globular proteins. Therefore, a variety of computational methods have been developed as an alternative to predict the topology of TM proteins. Most of these methods identify the membrane-spanning segments of a protein based solely on its amino acid sequence.

The AHTM segments can be predicted by a continuous run of 15-30 predominantly hydrophobic residues (106). Many prediction methods also assess the orientation of α-helices with respect to the membrane based on the 'positive inside' rule (108). According to this, the proportion of positively charged residues is higher in the loops in the cytoplasmic side of the membrane, compatible with the hydrophobic environment in lipid membranes (108). To identify AHTM segments, the SOSUI (109) method relies on hydropathy scales (i.e. observed preferences of amino acid residues for TM proteins). The DAS-TMfilter (110) uses the Dense Alignment Surface (DAS) approach to distinguish between AHTM and non-AHTM at a reasonably low rate of false positive predictions. TMMOD (111) is based on HMM profiles for AHTM prediction (Table VIII).

In contrast to AHTM, the TMB proteins cannot be easily distinguished because the segments embedded in the membrane are only seven residues long and numerous residues at the barrel inside are nonpolar (107). For this reason, computational techniques applicable only to TMB have been developed. TMBarrel-Hunt (112) classifies protein sequences as TMB or non-TMB based on a modified *k*-nearest neighbor (*k*-NN) based algorithm. TMB-Hunt (112) uses whole sequence amino acid composition profiles. TMBETA-NET (113) is based on the amino acid composition for discriminating TM proteins. Moreover, this program uses an algorithm based on NNs trained with amino acid sequences alone for identifying TMB segments (113). PRED-TMMB (114) relies on HMMs to predict TMB (114) (Table VIII).

Several methods predict both AHTM and TMB segments. Membrane protein IdeNtificatioN withOUt explicit use of hydropathy profiles and evolutionary profiles (MINNOU) (115) predicts AHTM and TMB segments. MINNOU relies on the PSSM profile, based on the representation of amino acid residues at a given position in a protein family, which consists of predicted RSA and SS of each amino acid (115).

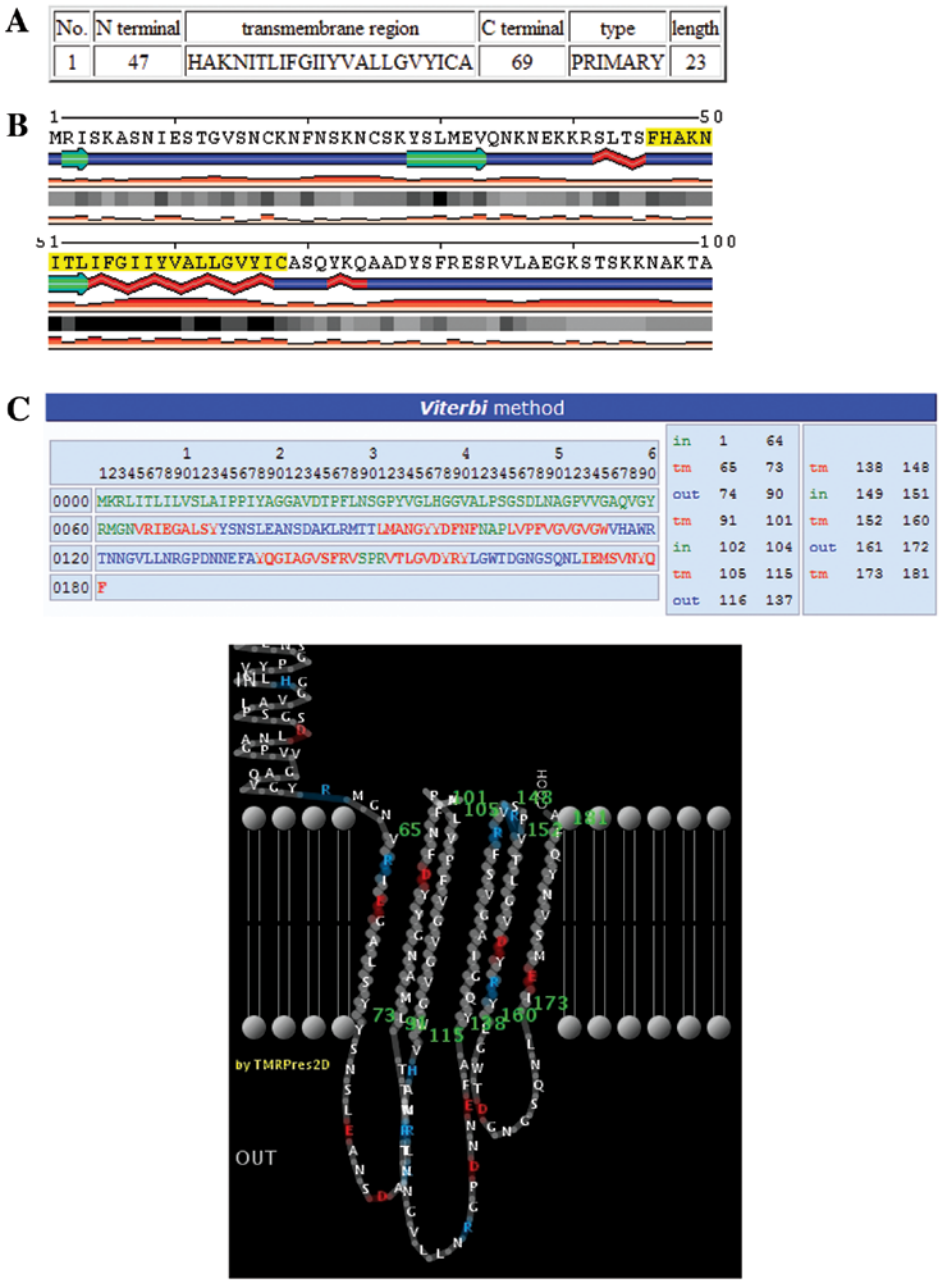

Figure 5. Secondary structure analysis of the AHTM protein Gbph using the methods (A), SOSUI and (B), MINNOU and the TMB protein porin using (C), PRED-TMMB. The membrane-spanning segment is highlighted in (B). In (C), tm stands for transmembrane. The two-dimensional representation of the TMB protein is shown below.

Transmembrane SEGments (TSEG) (116) predictor of TM segments is based on the analysis of tandem clusters of TM protein-coding genes and sequence similarities in the complete genome sequences. This method predicts TM segments after the removal of amino-terminal signal peptides (116). PREDiction of TM Regions (PRED-TMR) (117) is based on a standard hydrophobicity analysis to detect potential termini (starts and ends) of AHTM and TMB domains. Thereby, it predicts TM proteins discarding any highly hydrophobic stretches of residues without clear termini (117) (Table VIII).

Fig. 5 shows the results of the prediction of the AHTM protein Gbph (GenBank accession code: XP_001348185) from the parasite *Plasmodium falciparum*, which belongs in the glycophorin binding protein family, and the TMB protein porin (GenBank accession code: NP_820583) from the bacterium *Coxiella burnetii*.

5. Protein tertiary structure prediction

The protein tertiary structure is the full three-dimensional atomic structure of a single amino acid sequence (118,119). The biological function of a protein is highly correlated with its tertiary structure. Therefore, knowledge of the structure is critical for the functional annotation of uncharacterized proteins. However, due to the 'sequence-structure gap', the use of computational tools to assign a three-dimensional structure to a protein represents the most efficient alternative to experimental methods. The protein tertiary structure prediction

Table IX. Web-based tools for homology modelling.

Tool	Comments	Web link
CPHmodels	PSSM profile-based search for templates	http://www.cbs.dtu.dk/services/CPHmodels/
Domain Fishing	Domain split	http://www.bmm.icnet.uk/servers/3djigsaw/dom_fish
ESyPred3D	Target-template alignment generated by different programs	http://www.fundp.ac.be/sciences/biologie/urbm/bioinfo/esypred/
Geno3D	Target-template alignment using PSI-BLAST	http://geno3d-pbil.ibcp.fr
SWISS-MODEL	Integrated service	http://swissmodel.expasy.org/workspace/
TASSER-Lite	Iterative threading of the PDB for template selection; structure assembly	http://cssb.biology.gatech.edu/skolnick/webservice/tasserlite/index.html

is based on the observation that the three-dimensional structure of a protein tends to be better conserved than its amino acid sequence. There are knowledge-based and *ab initio* prediction methods. Knowledge-based methods depend on information extracted from databases of known structures to model the query proteins and they include homology modelling and fold recognition. *Ab initio* methods are based on physicochemical principles to determine a protein structure. The best strategy for protein tertiary structure prediction first involves homology modelling followed by fold recognition, and if not successful, *ab initio* prediction. The protein tertiary structure prediction tools mentioned below have been evaluated by LiveBench (120), CASP (121), CAFASP (122) and EVA (123).

Homology modelling. The most reliable method for the protein three-dimensional structure prediction is homology modelling, also known as comparative modelling. This method refers to constructing a full three-dimensional atomic model of an unknown (or 'target') protein from its amino acid sequence by using the solved three-dimensional structure of an evolutionarily-related (homologous) protein (or 'template'). Homology modelling is based on the principle that proteins sharing significant sequence identity adopt the same fold. The procedure for homology modelling involves three steps: (i) template selection, (ii) target-template alignment where residues in the target sequence are superimposed to residues in the template sequences and (iii) three-dimensional model construction. The accuracy of homology modelling is largely dependent on the accuracy of the target-template alignment, particularly when the aligned sequences share less than 30% identity. The accuracy of the generated model is assessed by the root mean square deviation (RMSD) of the distance between the α-carbon atoms of the aligned residues; lower RMSD values represent better alignments.

Some homology modelling methods focus on certain steps of the homology modelling process. For example, methods such as CPHmodels (124) and Domain Fishing (125) focus on the first step of the process, the template selection. CPHmodels (124) iteratively searches a sequence database to build a PSSM profile, which is then used to search a database of proteins with solved structures. In Domain Fishing (125), the query protein sequence is split into single domains to optimize the search for candidate structural templates. For each domain, a list of templates is generated, extracted from PDB, Pfam and SCOP. Finally, SS matching is used to remove any false templates. ESyPred3D (126) and Geno3D (127) methods focus on the second step, the target-template alignment, which is the most critical step of homology modelling. ESyPred3D (126) is an automated program which handles the target-template alignment problem by combining the results of multiple alignment programs and subsequently weighing and screening these results to filter out the false matches. In Geno3D (127), the target and template protein sequences are aligned using PSI-BLAST, and a protein homology model is constructed by extracting spatial restraints (dihedral angles and distances) from the pairwise alignment (Table IX).

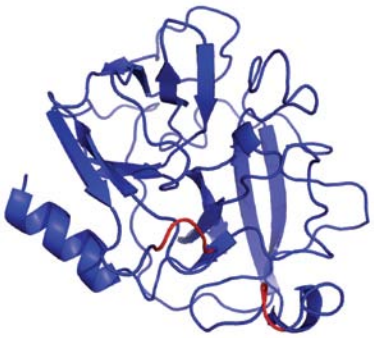

Figure 6. The predictor's SWISS-MODEL output. The prostate specific antigen (kallikrein 3) from stallion seminal plasma (PDB ID: 1GVZ) was the user specified template used for modelling the *X. tropicalis* KLK1-like protein.

SWISS-MODEL (128) is a fully automated homology modelling integrated service which allows the user to construct protein homology models by manually modifying and validating the different steps of modelling. TASSER-Lite (Threading/ASSEmbly/Refinement-Lite) uses the threading program PROSPECTOR_3 (129) to iteratively search the PDB to identify the template for modelling the query sequence (Table IX).

In the example in Fig. 6, the result of the homology modelling predictor SWISS-MODEL is shown. The *Xenopus tropicalis* KLK1-like protein (mentioned previously) was used as the input to the SWISS-MODEL.

Fold recognition. The homology modelling method, however, rapidly looses accuracy in the 'twilight zone', where the target and template sequences share <30% sequence identity. The

Table X. Fold recognition computational tools.

Tool	Comments	Web link
FFAS	Profile-profile comparison	http://ffas.burnham.org/ffas-cgi/cgi/ffas.pl
FUGUE	Environment-specific substitution tables; structure-dependent gap penalties	http://tardis.nibio.go.jp/fugue/prfsearch.html
M-TASSER	Multimeric threading; multimer model assembly; refinement	http://cssb.biology.gatech.edu/skolnick/webservice/MetaTASSER/index.html
pGenTHREADER	Profile-profile comparison	http://bioinf.cs.ucl.ac.uk/downloads/pGenTHREADER
PHYRE	Profile-profile comparison	http://www.sbg.bio.ic.ac.uk/~phyre/
SP[5]	Torsion angle profiles; profile-based gap penalty	http://sparks.informatics.iupui.edu/SP5/

Table XI. Web-based tools for *ab initio* protein structure prediction.

Tool	Comments	Web link
3Dpro	Fragment libraries; energy functions	http://www.ics.uci.edu/~baldig/scratch/
Bhageerath	Limits the search space of small globular proteins	http://www.scfbio-iitd.res.in/bhageerath/index.jsp
PROTINFO	Simulated annealing method	http://protinfo.compbio.washington.edu/protinfo_abcmfr/
ROSETTA	Models structurally variable regions	http://boinc.bakerlab.org/rosetta/

more efficient alternative method is the fold recognition or threading. This method is based on the observation that two proteins may adopt a similar fold even if they are evolutionarily distantly-related. The goal of fold recognition is to identify a known fold for a query sequence even if it does not share a significant degree of sequence identity to any of the proteins of known structure. The fold recognition approaches can be broadly divided into sequence-based and structure-based. The former approaches are based on multiple sequence alignments to construct profiles. The latter approaches attempt to optimally align a target sequence to the three-dimensional backbone of a template protein and assess the compatibility of a target sequence with each known structure by using knowledge-based structural profiles. Several fold recognition methods, also, combine the advantages of both approaches in order to produce better results.

The profile-profile comparison methods are sequence-based methods. The tools that implement this method accept a query protein sequence and automatically construct a sequence profile, which subsequently is compared with several sets of sequence profiles of proteins of known structure. pGenTHREADER (parametricGenTHREADER) (130) constructs PSSM profiles and incorporates structural information and solvation potentials in order to improve fold recognition. In a similar manner, FFAS03 (Fold and Function Assignment System 03) (131) and PHYRE (132) automatically generate a PSSM profile from the user-supplied protein sequence, which is then compared against the sequence profiles of proteins of known structure (Table X).

The structure-based methods for fold recognition include M-TASSER (Meta-Threading/ASSEmbly/Refinement) (133,134), a hierarchical method for protein fold recognition which employs multimeric threading for template identification, followed by multimer model assembly and refinement. Another method, SP[5] (135), improves protein fold recognition by using dihedral torsion angles (ϕ and ψ) profile along with a profile-based gap penalty model and real-value RSA profiles. Several fold recognition methods, also, combine the advantages of both sequence-based and structure-based approaches in order to produce better results. FUGUE (136), which belongs in this category, has implemented environment-specific substitution tables and environment-dependent gap penalties in order to increase the target-template alignment and homology recognition performance (Table X).

Ab initio protein structure prediction. The *ab initio* protein structure prediction methods are used for the prediction of native protein structures in the absence of reliable template structures to construct accurate models of the target sequence. The *ab initio* methods are based on fundamental physico-chemical principles, and they search for protein conformations which are thermodynamically favorable and stereochemically allowed.

ROSETTA (137) is the most popular *ab initio* protein structure prediction method. It models long structurally variable regions (SVR) based on a databank of known structures; the non-local interactions are approximated with a scoring function and Monte Carlo minimization. 3Dpro (138) utilizes fragment libraries built from the PDB and energy functions for protein three-dimensional structure prediction. In PROTINFO (139), tertiary protein structures are generated using a simulated annealing method which minimizes a target scoring function. Bhageerath (140) is an energy based protein tertiary structure prediction method which narrows down the

search space of small globular proteins to generate probable native-like structures (Table XI).

6. Conclusions

In this review, we presented a bioinformatics 'toolkit' particularly useful for bench biologists. We suggest a hierarchical approach to protein structure prediction that would consist of a BLAST search in the databases to retrieve sequences similar to the query. Phylogenetic analysis is suggested in order to assess the evolutionary relationships of the retrieved sequences. The sequence of interest should be queried against the signature databases for domain identification. The next step would be the prediction of the SS of the query protein. Homology modelling would be the next step in the prediction of the three-dimensional protein structure of the query sequence. In case the sequence identity between the query protein sequence and the template protein with solved structure is in the 'twilight zone', fold recognition would be recommended. If the sequence identity drops below 10% ('midnight zone') (141), then *ab initio* structure prediction should be employed. Finally, we believe that the predictions at the individual steps of this process are improved when the results produced by different methods are combined.

References

1. Neerincx PB and Leunissen JA: Evolution of web services in bioinformatics. Brief Bioinform 6: 178-188, 2005.
2. Fischer D: Servers for protein structure prediction. Curr Opin Struct Biol 16: 178-182, 2006.
3. Fitch WM: Homology a personal view on some of the problems. Trends Genet 16: 227-231, 2000.
4. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J and Sayers EW: GenBank. Nucleic Acids Res 37: D26-D31, 2009.
5. Kulikova T, Akhtar R, Aldebert P, *et al*: EMBL nucleotide sequence database in 2006. Nucleic Acids Res 35: D16-D20, 2007.
6. Sugawara H, Ogasawara O, Okubo K, Gojobori T and Tateno Y: DDBJ with new system and face. Nucleic Acids Res 36: D22-D24, 2008.
7. Pruitt KD, Tatusova T and Maglott DR: NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 35: D61-D65, 2007.
8. Wu CH, Huang H, Arminski L, *et al*: The protein information resource: an integrated public resource of functional annotation of proteins. Nucleic Acids Res 30: 35-37, 2002.
9. Bairoch A, Apweiler R, Wu CH, *et al*: The universal protein resource (UniProt). Nucleic Acids Res 33: D154-D159, 2005.
10. Drysdale R: FlyBase: a database for the *Drosophila* research community. Methods Mol Biol 420: 45-59, 2008.
11. Markowitz VM, Szeto E, Palaniappan K, *et al*: The integrated microbial genomes (IMG) system in 2007: data content and analysis tool extensions. Nucleic Acids Res 36: D528-D533, 2008.
12. Hubbard TJ, Aken BL, Ayling S, *et al*: Ensembl 2009. Nucleic Acids Res 37: D690-D697, 2009.
13. Wheeler DL, Barrett T, Benson DA, *et al*: Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 36: D13-D21, 2008.
14. Berman HM, Westbrook J, Feng Z, *et al*: The Protein Data Bank. Nucleic Acids Res 28: 235-242, 2000.
15. Pearl F, Todd A, Sillitoe I, *et al*: The CATH domain structure database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. Nucleic Acids Res 33: D247-D251, 2005.
16. Andreeva A, Howorth D, Brenner SE, Hubbard TJ, Chothia C and Murzin AG: SCOP database in 2004: refinements integrate structure and sequence family data. Nucleic Acids Res 32: D226-D229, 2004.
17. Klimke W, Agarwala R, Badretdin A, *et al*: The National Center for Biotechnology Information's protein clusters database. Nucleic Acids Res 37: D216-D223, 2009.
18. Dayhoff MO, Schwartz RM and Orcutt BC: A Model of Evolutionary Change in Proteins. In: Atlas of Protein Sequence and Structure. Vol. 5. 3rd edition. Dayhoff MO (ed). National Biomedical Research Foundation, Washington DC, pp345-352, 1978.
19. Henikoff S and Henikoff JG: Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 89: 10915-10919, 1992.
20. Jones DT, Taylor WR and Thornton JM: The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 8: 275-282, 1992.
21. Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ: Basic local alignment search tool. J Mol Biol 215: 403-410, 1990.
22. Smith TF and Waterman MS: Identification of common molecular subsequences. J Mol Biol 147: 195-197, 1981.
23. Altschul SF, Madden TL, Schaffer AA, *et al*: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389-3402, 1997.
24. Pavlopoulou A, Pampalakis G, Michalopoulos I and Sotiropoulou G: Evolutionary history of tissue kallikreins. PLoS One 5: e13781, 2010.
25. Pearson WR: Rapid and sensitive sequence comparison with FASTP and FASTA. Methods Enzymol 183: 63-98, 1990.
26. Needleman SB and Wunsch CD: A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48: 443-453, 1970.
27. Pei J: Multiple protein sequence alignment. Curr Opin Struct Biol 18: 382-386, 2008.
28. Feng DF and Doolittle RF: Progressive sequence alignment as a prerequisite to correct phylogenetic trees. J Mol Evol 25: 351-360, 1987.
29. Larkin MA, Blackshields G, Brown NP, *et al*: Clustal W and Clustal X version 2.0. Bioinformatics 23: 2947-2948, 2007.
30. Do CB, Mahabhashyam MS, Brudno M and Batzoglou S: ProbCons: Probabilistic consistency-based multiple sequence alignment. Genome Res 15: 330-340, 2005.
31. Notredame C, Higgins DG and Heringa J: T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol 302: 205-217, 2000.
32. Edgar RC: MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792-1797, 2004.
33. Katoh K, Kuma K, Toh H and Miyata T: MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res 33: 511-518, 2005.
34. Katoh K and Toh H: Recent developments in the MAFFT multiple sequence alignment program. Brief Bioinform 9: 286-298, 2008.
35. Pei J and Grishin NV: MUMMALS: multiple sequence alignment improved by using hidden Markov models with local structural information. Nucleic Acids Res 34: 4364-4374, 2006.
36. Pei J and Grishin NV: PROMALS: towards accurate multiple sequence alignments of distantly related proteins. Bioinformatics 23: 802-808, 2007.
37. Pei J, Kim BH and Grishin NV: PROMALS3D: a tool for multiple protein sequence and structure alignments. Nucleic Acids Res 36: 2295-2300, 2008.
38. O'Sullivan O, Suhre K, Abergel C, Higgins DG and Notredame C: 3DCoffee: combining protein sequences and structures within multiple sequence alignments. J Mol Biol 340: 385-395, 2004.
39. Felsenstein J: PHYLIP – Phylogeny Inference Package (Version 3.2). Cladistics 5: 164-166, 1989.
40. Swofford DL: PAUP*, Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Inc. Publishers, Sunderland, MA, 2000.
41. Kumar S, Nei M, Dudley J and Tamura K: MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences. Brief Bioinform 9: 299-306, 2008.
42. Tamura K, Dudley J, Nei M and Kumar S: MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol 24: 1596-1599, 2007.
43. Guindon S and Gascuel O: A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52: 696-704, 2003.
44. Guindon S, Lethiec F, Duroux P and Gascuel O: PHYML Online–a web server for fast maximum likelihood-based phylogenetic inference. Nucleic Acids Res 33: W557-W559, 2005.
45. Stamatakis A: RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22: 2688-2690, 2006.

46. Ronquist F and Huelsenbeck JP: MrBayes 3: bayesian phylogenetic inference under mixed models. Bioinformatics 19: 1572-1574, 2003.
47. Huson DH: SplitsTree: analyzing and visualizing evolutionary data. Bioinformatics 14: 68-73, 1998.
48. Bandelt HJ and Dress AW: Split decomposition: a new and useful approach to phylogenetic analysis of distance data. Mol Phylogenet Evol 1: 242-252, 1992.
49. Page RD: TreeView: an application to display phylogenetic trees on personal computers. Comput Appl Biosci 12: 357-358, 1996.
50. Perriere G and Gouy M: WWW-query: an on-line retrieval system for biological sequence banks. Biochimie 78: 364-369, 1996.
51. Huson DH, Richter DC, Rausch C, Dezulian T, Franz M and Rupp R: Dendroscope: an interactive viewer for large phylogenetic trees. BMC Bioinformatics 8: 460, 2007.
52. Letunic I and Bork P: Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics 23: 127-128, 2007.
53. Felsenstein J: Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39: 783-791, 1985.
54. Hillis DM and Bull JJ: An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Syst Biol 42: 182-192, 1993.
55. Efron B: Bootstrap methods: another look at the jackknife. Ann Stat 7: 1-26, 1979.
56. Paliakasis CD, Michalopoulos I and Kossida S: Web-based tools for protein classification. Methods Mol Biol 428: 349-367, 2008.
57. Ponting CP and Russell RR: The natural history of protein domains. Annu Rev Biophys Biomol Struct 31: 45-71, 2002.
58. Holland TA, Veretnik S, Shindyalov IN and Bourne PE: Partitioning protein structures into domains: why is it so difficult? J Mol Biol 361: 562-590, 2006.
59. Mulder NJ and Apweiler R: Tools and resources for identifying protein families, domains and motifs. Genome Biol 3: Reviews2001, 2002.
60. Wu CH, Huang H, Yeh LS and Barker WC: Protein family classification and functional annotation. Comput Biol Chem 27: 37-47, 2003.
61. Attwood TK: The role of pattern databases in sequence analysis. Brief Bioinform 1: 45-59, 2000.
62. Henikoff S and Henikoff JG: Embedding strategies for effective use of information from multiple sequence alignments. Protein Sci 6: 698-705, 1997.
63. Hofmann K: Sensitive protein comparisons with profiles and hidden Markov models. Brief Bioinform 1: 167-178, 2000.
64. Gribskov M, McLachlan AD and Eisenberg D: Profile analysis: detection of distantly related proteins. Proc Natl Acad Sci USA 84: 4355-4358, 1987.
65. Eddy SR: Profile hidden Markov models. Bioinformatics 14: 755-763, 1998.
66. Pietrokovski S, Henikoff JG and Henikoff S: The Blocks database–a system for protein classification. Nucleic Acids Res 24: 197-200, 1996.
67. Attwood TK, Bradley P, Flower DR, *et al*: PRINTS and its automatic supplement, prePRINTS. Nucleic Acids Res 31: 400-402, 2003.
68. Marchler-Bauer A, Anderson JB, Chitsaz F, *et al*: CDD: specific functional annotation with the Conserved Domain Database. Nucleic Acids Res 37: D205-D210, 2009.
69. Hulo N, Bairoch A, Bulliard V, *et al*: The 20 years of PROSITE. Nucleic Acids Res 36: D245-D249, 2008.
70. Lees J, Yeats C, Redfern O, Clegg A and Orengo C: Gene3D: merging structure and function for a thousand genomes. Nucleic Acids Res 38: D296-D300, 2010.
71. Mi H, Dong Q, Muruganujan A, Gaudet P, Lewis S and Thomas PD: PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium. Nucleic Acids Res 38: D204-D210, 2010.
72. Finn RD, Tate J, Mistry J, *et al*: The Pfam protein families database. Nucleic Acids Res 36: D281-D288, 2008.
73. Nikolskaya AN, Arighi CN, Huang H, Barker WC and Wu CH: PIRSF family classification system for protein functional and evolutionary analysis. Evol Bioinform Online 2: 197-209, 2006.
74. Letunic I, Doerks T and Bork P: SMART 6: recent updates and new developments. Nucleic Acids Res 37: D229-D232, 2009.
75. Wilson D, Pethica R, Zhou Y, *et al*: SUPERFAMILY–sophisticated comparative genomics, data mining, visualization and phylogeny. Nucleic Acids Res 37: D380-D386, 2009.
76. Haft DH, Selengut JD and White O: The TIGRFAMs database of protein families. Nucleic Acids Res 31: 371-373, 2003.
77. Bru C, Courcelle E, Carrere S, Beausse Y, Dalmar S and Kahn D: The ProDom database of protein domain families: more emphasis on 3D. Nucleic Acids Res 33: D212-D215, 2005.
78. Corpet F: Multiple sequence alignment with hierarchical clustering. Nucleic Acids Res 16: 10881-10890, 1988.
79. Hunter S, Apweiler R, Attwood TK, *et al*: InterPro: the integrative protein signature database. Nucleic Acids Res 37: D211-D215, 2009.
80. Vlahovicek K, Kajan L, Agoston V and Pongor S: The SBASE domain sequence resource, release 12: prediction of protein domain-architecture using support vector machines. Nucleic Acids Res 33: D223-D225, 2005.
81. Scordis P, Flower DR and Attwood TK: FingerPRINTScan: intelligent searching of the PRINTS motif database. Bioinformatics 15: 799-806, 1999.
82. Mulder N and Apweiler R: InterPro and InterProScan: tools for protein sequence classification and comparison. Methods Mol Biol 396: 59-70, 2007.
83. Schaffer AA, Wolf YI, Ponting CP, Koonin EV, Aravind L and Altschul SF: IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. Bioinformatics 15: 1000-1011, 1999.
84. de Castro E, Sigrist CJ, Gattiker A, *et al*: ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. Nucleic Acids Res 34: W362-W365, 2006.
85. Sigrist CJ, De Castro E, Langendijk-Genevaux PS, Le Saux V, Bairoch A and Hulo N: ProRule: a new database containing functional and structural information on PROSITE profiles. Bioinformatics 21: 4060-4066, 2005.
86. Crooks GE, Hon G, Chandonia JM and Brenner SE: WebLogo: a sequence logo generator. Genome Res 14: 1188-1190, 2004.
87. Pauling L, Corey RB and Branson HR: The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. Proc Natl Acad Sci USA 37: 205-211, 1951.
88. Kabsch W and Sander C: Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22: 2577-2637, 1983.
89. Cheng H, Sen TZ, Kloczkowski A, Margaritis D and Jernigan RL: Prediction of protein secondary structure by mining structural fragment database. Polymer (Guildf) 46: 4314-4321, 2005.
90. Chou PY and Fasman GD: Prediction of protein conformation. Biochemistry 13: 222-245, 1974.
91. Garnier J, Osguthorpe DJ and Robson B: Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. J Mol Biol 120: 97-120, 1978.
92. Kloczkowski A, Ting KL, Jernigan RL and Garnier J: Combining the GOR V algorithm with evolutionary information for protein secondary structure prediction from amino acid sequence. Proteins 49: 154-166, 2002.
93. Cheng H, Sen TZ, Jernigan RL and Kloczkowski A: Consensus data mining (CDM) protein secondary structure prediction server: combining GOR V and fragment database mining (FDM). Bioinformatics 23: 2628-2630, 2007.
94. Rost B: Twilight zone of protein sequence alignments. Protein Eng 12: 85-94, 1999.
95. Rost B, Sander C and Schneider R: PHD–an automatic mail server for protein secondary structure prediction. Comput Appl Biosci 10: 53-60, 1994.
96. Rost B and Sander C: Prediction of protein secondary structure at better than 70% accuracy. J Mol Biol 232: 584-599, 1993.
97. Rost B and Sander C: Improved prediction of protein secondary structure by use of sequence profiles and neural networks. Proc Natl Acad Sci USA 90: 7558-7562, 1993.
98. Cole C, Barber JD and Barton GJ: The Jpred 3 secondary structure prediction server. Nucleic Acids Res 36: W197-W201, 2008.
99. Pollastri G, Przybylski D, Rost B and Baldi P: Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. Proteins 47: 228-235, 2002.
100. Pollastri G and McLysaght A: Porter: a new, accurate server for protein secondary structure prediction. Bioinformatics 21: 1719-1720, 2005.
101. Adamczak R, Porollo A and Meller J: Combining prediction of secondary structure and solvent accessibility in proteins. Proteins 59: 467-475, 2005.
102. McGuffin LJ, Bryson K and Jones DT: The PSIPRED protein structure prediction server. Bioinformatics 16: 404-405, 2000.

103. Rost B: Review: protein secondary structure prediction continues to rise. J Struct Biol 134: 204-218, 2001.
104. Chan HS and Dill KA: Origins of structure in globular proteins. Proc Natl Acad Sci USA 87: 6388-6392, 1990.
105. von Heijne G: Principles of membrane protein assembly and structure. Prog Biophys Mol Biol 66: 113-139, 1996.
106. Schulz GE: The structure of bacterial outer membrane proteins. Biochim Biophys Acta 1565: 308-317, 2002.
107. Schulz GE: beta-Barrel membrane proteins. Curr Opin Struct Biol 10: 443-447, 2000.
108. von Heijne G: Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. J Mol Biol 225: 487-494, 1992.
109. Hirokawa T, Boon-Chieng S and Mitaku S: SOSUI: classification and secondary structure prediction system for membrane proteins. Bioinformatics 14: 378-379, 1998.
110. Cserzo M, Eisenhaber F, Eisenhaber B and Simon I: TM or not TM: transmembrane protein prediction with low false positive rate using DAS-TMfilter. Bioinformatics 20: 136-137, 2004.
111. Kahsay RY, Gao G and Liao L: An improved hidden Markov model for transmembrane protein detection and topology prediction and its applications to complete genomes. Bioinformatics 21: 1853-1858, 2005.
112. Garrow AG, Agnew A and Westhead DR: TMB-Hunt: a web server to screen sequence sets for transmembrane beta-barrel proteins. Nucleic Acids Res 33: W188-W192, 2005.
113. Gromiha MM, Ahmad S and Suwa M: TMBETA-NET: discrimination and prediction of membrane spanning beta-strands in outer membrane proteins. Nucleic Acids Res 33: W164-W167, 2005.
114. Bagos PG, Liakopoulos TD, Spyropoulos IC and Hamodrakas SJ: PRED-TMBB: a web server for predicting the topology of beta-barrel outer membrane proteins. Nucleic Acids Res 32: W400-W404, 2004.
115. Cao B, Porollo A, Adamczak R, Jarrell M and Meller J: Enhanced recognition of protein transmembrane domains with prediction-based structural profiles. Bioinformatics 22: 303-309, 2006.
116. Kihara D and Kanehisa M: Tandem clusters of membrane proteins in complete genome sequences. Genome Res 10: 731-743, 2000.
117. Pasquier C, Promponas VJ, Palaios GA, Hamodrakas JS and Hamodrakas SJ: 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, 1999.
118. Bowie JU, Luthy R and Eisenberg D: A method to identify protein sequences that fold into a known three-dimensional structure. Science 253: 164-170, 1991.
119. Luthy R, Bowie JU and Eisenberg D: Assessment of protein models with three-dimensional profiles. Nature 356: 83-85, 1992.
120. Rychlewski L and Fischer D: LiveBench-8: the large-scale, continuous assessment of automated protein structure prediction. Protein Sci 14: 240-245, 2005.
121. Moult J, Fidelis K, Zemla A and Hubbard T: Critical assessment of methods of protein structure prediction (CASP): round IV. Proteins (Suppl 5): S2-S7, 2001.
122. Fischer D, Rychlewski L, Dunbrack RL Jr, Ortiz AR and Elofsson A: CAFASP3: the third critical assessment of fully automated structure prediction methods. Proteins 53 (Suppl 6): S503-S516, 2003.
123. Eyrich VA, Marti-Renom MA, Przybylski D, *et al*: EVA: continuous automatic evaluation of protein structure prediction servers. Bioinformatics 17: 1242-1243, 2001.
124. Nielsen M, Lundegaard C, Lund O and Petersen TN: CPHmodels-3.0–remote homology modeling using structure-guided sequence profiles. Nucleic Acids Res 38: W576-W581, 2010.
125. Contreras-Moreira B and Bates PA: Domain fishing: a first step in protein comparative modelling. Bioinformatics 18: 1141-1142, 2002.
126. Lambert C, Leonard N, De Bolle X and Depiereux E: ESyPred3D: Prediction of proteins 3D structures. Bioinformatics 18: 1250-1256, 2002.
127. Combet C, Jambon M, Deleage G and Geourjon C: Geno3D: automatic comparative molecular modelling of protein. Bioinformatics 18: 213-214, 2002.
128. Kiefer F, Arnold K, Kunzli M, Bordoli L and Schwede T: The SWISS-MODEL Repository and associated resources. Nucleic Acids Res 37: D387-D392, 2009.
129. Skolnick J, Kihara D and Zhang Y: Development and large scale benchmark testing of the PROSPECTOR_3 threading algorithm. Proteins 56: 502-518, 2004.
130. Lobley A, Sadowski MI and Jones DT: pGenTHREADER and pDomTHREADER: new methods for improved protein fold recognition and superfamily discrimination. Bioinformatics 25: 1761-1767, 2009.
131. Jaroszewski L, Rychlewski L, Li Z, Li W and Godzik A: FFAS03: a server for profile–profile sequence alignments. Nucleic Acids Res 33: W284-W288, 2005.
132. Kelley LA and Sternberg MJ: Protein structure prediction on the Web: a case study using the Phyre server. Nat Protoc 4: 363-371, 2009.
133. Zhou H, Pandit SB and Skolnick J: Performance of the Pro-sp3-TASSER server in CASP8. Proteins 77 (Suppl 9): S123-S127, 2009.
134. Chen H and Skolnick J: M-TASSER: an algorithm for protein quaternary structure prediction. Biophys J 94: 918-928, 2008.
135. Zhang W, Liu S and Zhou Y: SP5: improving protein fold recognition by using torsion angle profiles and profile-based gap penalty model. PLoS One 3: e2325, 2008.
136. Shi J, Blundell TL and Mizuguchi K: FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. J Mol Biol 310: 243-257, 2001.
137. Rohl CA, Strauss CE, Misura KM and Baker D: Protein structure prediction using Rosetta. Methods Enzymol 383: 66-93, 2004.
138. Cheng J, Randall AZ, Sweredoski MJ and Baldi P: SCRATCH: a protein structure and structural feature prediction server. Nucleic Acids Res 33: W72-W76, 2005.
139. Hung LH, Ngan SC, Liu T and Samudrala R: PROTINFO: new algorithms for enhanced protein structure predictions. Nucleic Acids Res 33: W77-W80, 2005.
140. Jayaram B, Bhushan K, Shenoy SR, *et al*: Bhageerath: an energy based web enabled computer software suite for limiting the search space of tertiary structures of small globular proteins. Nucleic Acids Res 34: 6195-6204, 2006.
141. Rost B: Protein structures sustain evolutionary drift. Fold Des 2: S19-S24, 1997.