Computational Methods for Protein Structure Prediction and Fold Recognition

I. 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/) 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