4OR 4: 91–123 (2006) **Invited survey**

DOI: 10.1007/s10288-006-0089-y

Some operations research methods for analyzing protein sequences and structures*

Jacek Błażewicz, Piotr Łukasiak and Maciej Miłostan

Institute of Computing Science, Poznan University of Technology, Poznan and
Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznan, Poland
(e-mail: Jacek.Blazewicz@cs.put.poznan.pl; Piotr.Lukasiak@cs.put.poznan.pl;
Maciej.Milostan@cs.put.poznan.pl)

Received: February 2006 / Revised version: May 2006

Abstract. Operations Research is probably one of the most successful fields of applied mathematics used in Economics, Physics, Chemistry, almost everywhere one has to analyze huge amounts of data. Lately, these techniques were introduced in biology, especially in the protein analysis area to support biologists. The fast growth of protein data makes operations research an important issue in bioinformatics, a science which lays on the border between computer science and biology. This paper gives a short overview of the operations research techniques currently used to support structural and functional analysis of proteins.

Key words: Protein structure, hidden Markov model, clustering, neural networks, dynamic programming

1 Introduction

Since 1953 (discovery of the double helical structure of DNA) many important findings were obtained in molecular biology. The unraveling of the genetic code was only the beginning. Learning the details of genes and their discontinuous nature in eukaryotic genomes (multicellular organisms) has led to the ability of studying and manipulating the material of that abstract concept of Mendel's, the gene itself. Learning to read the genetic material more and more rapidly has enabled scientists to attempt to decode entire genomes. The rate of innovations in molecular biology is breathtaking. The experimental techniques that must be painstakingly used by one generation of scientists are usually routine for the next generation. The accumulation

* Partially supported by KBN grant No 3T11F00227

of data has necessitated international databases for nucleic acids, for proteins and for individual organisms and even chromosomes. The crudest measure of progress, the size of nucleic databases has an exponential growth rate. Consequently, a new subject, or if that is too grand, a new area of expertise is being created, combining the biological and information sciences. Finding relevant facts and hypotheses in huge databases is becoming essential to biology. Huge amounts of biological data which are available via many databases, need to be analyzed and interpreted, but it is almost impossible to do that by scientists without the help of appropriate computerized techniques.

Probably one of the most important tasks is nowadays to analyze structural and functional features of proteins. Progress in that area can generate profits in medicine, chemistry and ecology, but this progress is impossible without support from computer science. Marriage between biology and computer science can give impressive results, but behind computer science crucial mechanisms are hidden. These mechanisms are needed for data analysis, data interpretation and most of them have roots in operations research.

The aim of this study is to show the usage of selected operations research approaches to solve particular protein analysis problems. Our (subjective) choice includes among others: prediction of protein secondary and tertiary structure, prediction of protein domains and protein fold classification. This study complements earlier survey papers (Błażewicz et al., 1997, 2005) which dealt mainly with DNA sequence analysis.

The organization of the paper is as follows: Section 2 describes basic terms in molecular biology, Section 3 presents basic examples of problems arising in the protein prediction area. Other sections present applications of basic operations research techniques in solving particular problems of protein analysis. Section 4 gives examples of neural networks applied in that context. Section 5 presents dynamic programming approaches to the solution of selected problems of protein analysis. In Section 6 the hidden Markov model approach is presented, while Section 7 gives examples of clustering techniques. Conclusions indicate possible generalizations.

2 Biological primer

DNA or otherwise called deoxyribonucleic acid contains the information living species require to synthesize proteins and their cells to replicate themselves (Waterman, 1995; Setubal and Meidanis, 1997; Pevzner, 2001). In other words, it is the storage repository for the information that is required for any cell to function. Watson and Crick have discovered the structure of DNA in 1953. The famous double-helix structure of DNA has its own significance. There are basically four nucleotide bases, which make up the DNA: Adenine (A), Guanine (G), Thymine (T) and Cytosine (C). A DNA sequence can be presented like a chain, e.g.: "ATTGCTGAAGGTGCGG" where each letter corresponds to one nucleotide base, respectively. DNA is measured in the number of base pairs it consists of, usually in kBp or

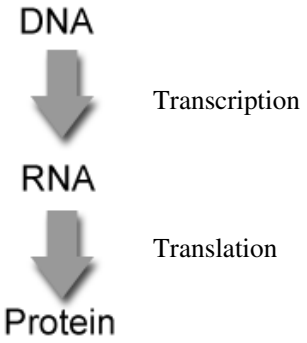

Fig. 1. Central dogma of molecular biology

mBp (kilo/mega base pairs). Each base has its complementary base, which means in the double helical structure of DNA: A will have T as its complementary and similarly G will have C. DNA molecules used to code genetic material of eukaryotic species are incredibly long. If all the DNA bases of the human genome were typed as A, C, T and G, the 3 billion letters would fill 3,000 books of 500 pages each. The DNA is broken down into bits and is tightly cut into coils, which are called chromosomes; human beings having 23 pairs of homolog chromosomes. These chromosomes are further broken down into smaller pieces of code called genes. The 23 pairs of chromosomes consist of about 30,000 genes and every gene has its own function. Finding out the arrangement of the bases in any DNA chain, called DNA sequencing, is usually divided into stages: mapping, assembling and sequencing (see Błażewicz et al., 1997, 2005).

The significance of a DNA is very high. The gene's sequence is like a language that instructs cells to manufacture a particular protein. An intermediate language, encoded in the sequence of ribonucleic acid (RNA), translates a gene's message into a protein's amino acid sequence. It is the protein that determines the trait. This process is called the *central dogma* of molecular biology (Fig. 1).

RNA is somewhat similar to DNA; they both are nucleic acids containing nitric bases joined by a sugar-phosphate backbone, however, structural and functional differences distinguish RNA from DNA. Structurally, RNA is a single strand while DNA is double strand; DNA has Thymine, while RNA has Uracil (U); RNA nucleotides include sugar ribose, rather than the deoxyribose that is a part of DNA. Functionally, DNA maintains the protein-encoding information, whereas RNA uses the information to enable the cell to synthesize the particular protein. RNA takes that information into the cytoplasm, where the cell uses it to construct specific proteins (one may say that the corresponding gene is thus expressed).

Transcription is a process of making an RNA strand from a DNA template, and the obtained RNA molecule is called a transcript. Three types of RNA participate and play different roles in the synthesis of proteins (called *the translation*):

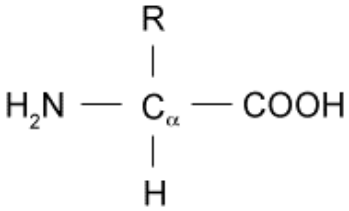

Fig. 2. General structure of α-amino acid

a. *Messenger RNA* (mRNA), which carries the genetic information from DNA and is used as a template for protein synthesis.

b. *Ribosomal RNA* (rRNA), which is a major constituent of the cellular particles called ribosomes on which protein synthesis actually takes place.

c. A set of *transfer RNA* (tRNA) molecules, each of which incorporates a particular amino acid subunit into the constructed protein when it recognizes a specific group of three adjacent bases in the mRNA.

Transcription is highly controlled and complex. In prokaryotes (simple organisms like bacteria), genes are expressed as required, and in eukaryotes (multicellular organisms), specialized cell types express subsets of genes. Transcription factors (*micro RNAs*) recognize sequences near a gene and bind sequentially, creating a binding transcription.

Each three mRNA bases in a row form a *codon* that specifies a particular amino acid.

The mRNA leader sequence binds to rRNA in the small subunit of a ribosome, and the first codon attracts a tRNA 3 letter chain, bearing methionine. Next, as the chain elongates, the large ribosomal subunit attaches and the appropriate aminoacids joined to tRNA molecules (corresponding to consecutive codons), form peptide bonds. At a stop codon, protein synthesis ceases. Protein folding begins as translation proceeds, with enzymes and chaperone proteins assisting the amino acid chain in assuming its final functional form. Translation is efficient and economical, as RNA, ribosomes, enzymes, and key proteins are recycled.

Amino acids are known as α-amino acids because they are built of a primary *amino group* (-NH_2), a carboxylic acid group (-COOH) and a hydrogen atom (H) joined to the *α-carbon atom* (Fig. 2). There are basically 20 standard amino acids having different structures in their side chains (R groups) (see. Figs. 2 and 3). Proline is an exception because it has a secondary amino group (-NH-), but for uniformity it is also treated as α-amino acid.

Peptides (and proteins) are made by joining amino acids together via bonds. Any number of amino acids can be joined together to form peptides of any length.

Small peptides (containing less than a couple of dozen amino acids) are sometimes called oligopeptides. Longer peptides are called polypeptides.

When thinking about peptide (and protein) structure, it is often useful to distinguish between the peptide “backbone” and the side chains. The backbone atoms

Name, Abbreviation (3 letters code, one letter code)	Name, Abbreviation (3 letters code, one letter code)
Alanine, ALA, A	Leucine, LEU, L
Arginine, ARG, R	Lysine, LYS, K
Asparagine, ASN, N	Methionine, MET, M
Aspartic acid, ASP, D	Phenylalanine, PHE, F
Cysteine, CYS, C	Proline, PRO, P
Glutamine, GLN, Q	Serine, SER, S
Glutamic acid, GLU, E	Threonine, THR, T
Glycine, GLY, G	Tryptophan, TRP, W
Histidine, HIS, H	Tyrosine, TYR, Y
Isoleucine, ILE, I	Valine, VAL, V

Fig. 3. List of 20 amino acids – name, three letter code, one letter code

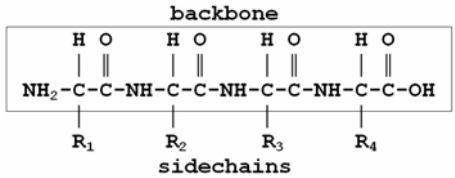

Fig. 4. Backbone – sidechains view of a protein chain

consist of the peptide amino units and the α-carbons; the side chains consist of the remaining atoms in the molecule (i.e. the "R" groups of each amino acid) (Fig.4).

Proteins are not linear molecules as suggested by amino acid sequence –Lys-Ala-Pro-Met-Gly- etc. (or K-A-P-M-G- in one-letter notation), for example. Rather, this "string" folds into an intricate three-dimensional structure that is unique to each protein. It is this three-dimensional structure that allows proteins to function. Thus, in order to understand the details of protein function, one must understand protein structure.

Protein structure is broken down into four levels:

Primary structure refers to the "linear" sequence of amino acids.

Primary structure is sometimes called the "covalent structure" of proteins because, with the exception of disulfide bonds, all of the covalent bonding (standard chemical bonds) within proteins define the primary structure. In contrast, higher orders of protein structures (i.e. secondary, tertiary and quartenary) involve mainly noncovalent interactions.

Secondary structure is "local" ordered structure brought about via hydrogen bonding mainly within the peptide backbone.

The most common secondary structure elements in proteins are the α-helix and the β-strand (sometime called β-pleated strand).

Tertiary structure is the "global" folding of a single polypeptide chain.

A major driving force in determining the tertiary structure of globular proteins is the hydrophobic effect. The polypeptide chain folds such that the side chains of the nonpolar amino acids (i.e. those for which the hydrophobic side chains are chemically unreactive and tend to aggregate rather than to be exposed to the aqueous environment – usually can be found inside the protein molecule) are "hidden" within the structure and the side chains of the polar residues are exposed on the outer surface.

Hydrogen bonding involving groups from both the peptide backbone and the side chains are important in stabilizing tertiary structure. The tertiary structure of some proteins is stabilized by disulfide bonds between cysteine residues.

Quartenary structure involves the association of two or more polypeptide chains into a multi-subunit structure.

Quartenary structure is the stable association of multiple polypeptide chains resulting in an active unit. Not all proteins exhibit quartenary structure. Usually, each polypeptide within a multisubunit protein folds more-or-less independently into a stable tertiary structure and the folded subunits then associate with each other to form the final structure.

Quartenary structures are stabilized mainly by noncovalent interactions; all types of noncolvalent interactions: hydrogen bonding, van der Waals interactions and ionic bonding, are involved in the interactions between subunits. In rare instances, disulfide bonds between cysteine residues in different polypeptide chains are involved in stabilizing quartenary structure.

Most proteins contain multiple *structural domains*. Structural domains are regions that are either compact, globular modules, or are clearly distinguished from flanking regions. Domains can be viewed as semi-independent three-dimensional units in proteins; they may fold independently and may constitute 'units of evolution'. Large scale sequencing efforts have confirmed that eukaryotes differ from all other species in the significantly higher proportion of proteins extending over 1500 residues. These large proteins undoubtedly consist of many structural domains.

3 Selected problems

In the following sections one can find a description of selected bioinformatics problems in the context of protein structure analysis. Methods for solving some of these problems are assessed biennially, since 1994, in Community Wide experiment for Critical Assessment of the Techniques for Protein Structure Prediciton (*CASP*) (`http://predictioncenter.org`).

Fig. 5. The protein tertiary structure with highlighted secondary

3.1 Secondary structure prediction

The three-dimensional, tertiary structure of the protein can be decomposed into local structures created by different fragments of the sequence (Fig. 5). These local fragments are so called protein secondary structures. Among them one can find helices (e.g. α-helix, denoted by H), β-strands (denoted by E), turns (T), loops (S) and other (X).

In the secondary structure prediction problem, one tries to map protein sequence (primary structure) fragments into secondary structures as it is shown in Fig. 6. Most methods applied for these problem are machine learning algorithms, thus, the important part of the prediction scheme is a database of known structures.

3.2 Tertiary structure prediction

Tertiary structure prediction, including determination of protein folding pathways, is one of the most complex computational problem that one can find currently in the analysis of proteins. The protein folding is a process of a creation of a three-dimensional structure from a linear structure of the polypeptide chain. The intellectual puzzle of the structure prediction remains unsolved approximately since the time when Anfinsen's experiments showed reversibility of the folding process (Anfinsen et al., 1961; Anfisen, 1973). Later Levinthal formulated the so called "Levinthal's paradox", concluding that the folding process must be somehow driven through kinetic pathways (Levinthal, 1968). High resolution X-ray structure determination of several hundreds proteins have confirmed that a specific sequence of polypeptide chains has only a single, compact, biologically active fold in the *native state* (Branden and Tooze, 1999). The *native conformation* appears to be the one with significantly lower free energy than others.

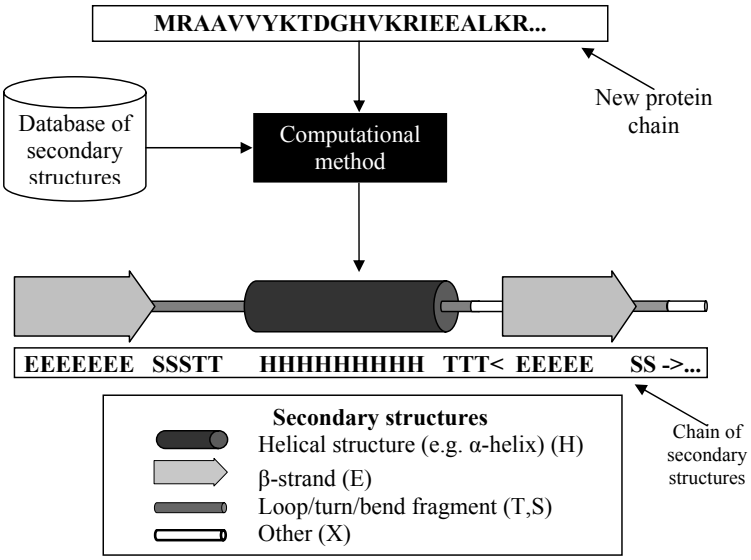

Fig. 6. Visualization of secondary structure prediction problem

3.2.1 Comparative modeling

An increase in number of known protein structures gathered in freely available databases, such as Brookhaven Protein Data Bank (PDB) caused growing popularity of modeling methods based on templates. The template is a protein of a known structure with a sequence similar to the one under investigation. It is possible to obtain a model for such a sequence by copying the backbone elements of the template and adding loops and side chains. The comparative modeling process, known as well as homology modeling, can be divided into four steps: template selection (can be done by BLAST search (Pearson and Lipman 1988; Altschul et al., 1990, 1997)), template and target alignment, backbone building, loops and side chains construction. Protein side chain positioning methods are well described in Chazelle et al. (2003, 2004), Eriksson et al. (2001) and Kingsford et al., (2005).

3.2.2 Fold recognition and new folds

It has to be stressed that different amino acid sequences can create similar tertiary structures that consist of similar motifs. These similar structures are called folds.

The problem of fold recognition is to assign fragments of a primary structure (an amino acid sequence) into the proper tertiary structures (folds) from the database of all known structures (see Fig. 7). The assignment, for example, can be done by sequence similarity analysis that leads to a detection of homologous structures.

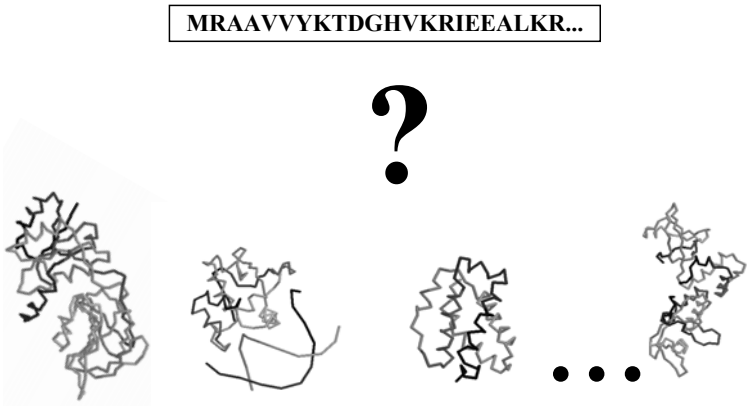

Fig. 7. Fold recognition problem: how to map a sequence into the proper fold if the corresponding structure already exists in database?

If no homolog could be detected, such a sequence possibly creates a new fold and one has to use computational simulation techniques to find a model of such a protein or wait till the structure will be determined experimentally. Simulations are usually done by Monte Carlo method and its variations or other heuristic techniques.

3.2.3 Ab-initio folding and simplified models

The native conformation of the protein is the one with significantly lower free energy than others, thus the protein folding process can be defined as the problem of energy function minimization. The energy function usually takes into account such properties of amino acids like hydrophobicity, electrostatic potential, size, weight, number of chemical bonds (e.g. hydrogen bonds) and others.

Due to the computational complexity of the problem, a protein structure is usually presented in a simplified manner (e.g. using UNRES representation (Nanias et al., 2005)) and placed in a simplified space, based on lattices. The minimization process is very often modeled using Monte Carlo simulations with replica (Kolinski et al., 2004a; Nanias et al., 2005) or other metaheuristic strategies (Błażewicz et al., 2004b, 2005b).

If one obtains the simplified model of the protein, one can switch back to the all atom models and make final refinements of the structure.

Ab-initio prediction methods are useful when comparative modeling approaches fail due to the lack of detectable homologous structures in databases.

Other interested approaches for protein docking and folding energy can be found in Althaus et al. (2002), Doye et al. (2004), Eskow et al. (2004), Koh et al. (2004) and Wagner et al. (2004).

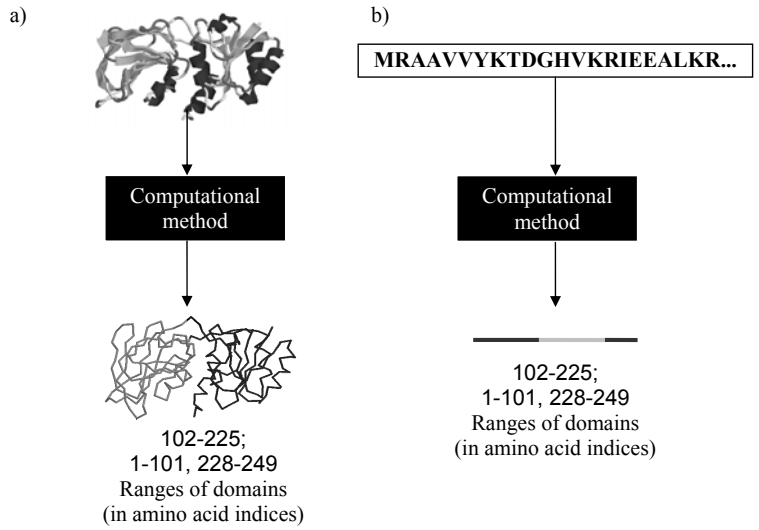

Fig. 8. Domain prediction problem: a) the first version: how to map protein tertiary structure into domains? b) the second version: how to map primary structures into domains?

3.3 Domain prediction

The tertiary structure is formed by packing structural elements (secondary structures) into one or several compact globular units called domains (Branden and Tooze, 1999). Domain prediction can be formulated in two contexts (Fig. 8). The first version of the problem is to find a computational method for proper splitting of the known tertiary structures into an unknown number of domains. These domains should possibly have independent stability and could fold independently.

The second version of the problem is harder than the first question; its aim is to find the correct mapping of the primary structure (amino acid sequence) into independent (possibly discontinued) fragments that probably can fold independently and create stable, functional units.

3.4 Function prediction

A protein in its native state is biologically active and often works as a regulator of metabolic reactions. The regulation is possible because some spots in the surface of protein can react with chemical compounds, so called ligands, and other proteins in the cells of living species. These spots are so called binding sites. Two different proteins with similar binding sites can play a similar role in cell metabolism. Thus, the main aim of computational methods of function prediction is automated detection of binding sites and chemical compounds corresponding to them, based

on geometrical properties and knowledge gathered in databases and gene ontology dictionaries.

3.5 Contact maps

In folded (native) state the protein structure is stabilized by internal contacts (e.g. hydrogen bonds, disulfide bridges) between its building blocks – amino acids. These contacts are important for understanding the protein nature and can be represented as maps or graphs of contacts that indicate strength of interactions. There is a wide range of methods that try to generate such maps from the sequence only. Protein structure comparison methods via contact maps can be found in Barnes et al. (2005), Caprara et al. (2004), Caprara and Lancia (2004), Carr and Lancia (2004); Lancia et al. (2001).

4 Neural networks

Neural networks have been applied to many aspects of predicting protein structure from protein sequence. Initially, methods were designed as a 'quick and dirty' demonstration that artificial intelligence-based approaches could solve real-life problems. At that stage, biologists typically reached higher levels of accuracy (when using their expertise) than computer scientists when using their approaches. However, more thorough investigations enabled the latter researchers to include information used by experts into neural network-based tools. Now, some tools are – on average – as accurate as the best experts; and experts using such tools often arrive at even more accurate predictions. Thus, several neural network-based methods have eventually contributed significantly to advancing the field of bioinformatics, and some are clearly influencing molecular biology. Here, we will not, of course, describe the basics in neural networks design and analysis. The interested reader is referred to (Fiesler and Beale, 1996; Haykin, 1999) for a thorough introduction to the subject.

4.1 Secondary structure prediction

The prediction of a protein's secondary structure – i.e. the formation of regular local structures such as α-helices and β-strands within a single protein sequence – is an essential intermediate step on the way to predicting the full three-dimensional structure of a protein. If the secondary structure of a protein is known, it is possible to derive a comparatively small number of possible tertiary (three-dimensional) structures using knowledge about the ways that secondary structural elements can appear.

Secondary structure prediction methods mainly distinguish between helix (H), strand (E), and non-regular structure (X) (classes T,S usually are also included in

X), as explained in Section 3.1 (see Fig. 5). Some stretches of sequence show a particular preference to be in one of these three states. The task here is to classify a pattern of w adjacent residues as either H, E, or X.

It is known that the effect of long-range interactions among amino acid residues in a protein is very important in the formation of protein secondary structure. Although many machine learning methods could 'learn' to take into account some of the effects of long-range interactions, principles underlying these interactions and their roles in influencing the formation of secondary structures are still difficult to understand. In contrast, if the formation of secondary structures of a protein were dominated by short-range interactions, then all the information for predicting the secondary structure of a residue would be contained in its flanking sequences. In other words, the tertiary structure would have little influence on the formation of secondary structures. If this were true for some residues of a protein, then one should be able to predict their secondary structures at a relatively higher accuracy than others. In recent years, most developments in the secondary structure prediction has been obtained in the area of machine learning techniques applications. Neural networks were first applied to the prediction of secondary structures (Bohr et al., 1988; Qian and Sejnowski, 1988). This stimulated many subsequent studies of neural networks (Holley and Karplus, 1989; Kneller et al., 1990) in the secondary structure prediction. However, a real breakthrough did not come until the work of Rost and Sander (Rost and Sander, 1993a,b, 1994), which was made available as a web server called PHD. The main reason for its success (better than 70% prediction accuracy for the first time), among many others, was the concept of profiles. Thus, instead of presenting a network with a single sequence, many aligned sequences of homologous proteins were presented. The method of profiles continues to improve as more and more sequences are becoming available (Przybylski and Rost, 2002). The reason for the success of the profile based method seems to be that it captures the fact that protein structures are more conserved than sequences. Because only mutations that do not disrupt the three-dimensional structure of a protein will survive the evolution, sequence divergence under structural constraints reflects the interactions between amino acid residues of a protein, where the interactions could be either short range or long range in sequence. Now, most secondary structure prediction methods achieving on average high performance with Q_3[1)] measure reaching 80% (Riis and Krogh, 1996; Baldi et al., 1999; Cuff and Barton, 1999; Jones, 1999; Ouali and King, 2000; Pollastri et al., 2002), make use of PSI-BLAST profiles (Altschul et al., 1997) in combination with an improvement of prediction algorithms. New machine learning methods such as support vector machines (Hua and Sun, 2001) should also benefit from PSI-BLAST profiles.

Let us also mention here a successful attempt of using Logical Analysis of Data to predict secondary structures of protein chains (Błażewicz et al., 2004a, 2005a).

[1] Q_3 measures structure predictability with three structural forms distinguished, and is calculated as follows: $Q_3=(N_{c3}/N)*100$ where N expresses the total number of amino acids in the polypeptide under consideration, N_{c3} expresses the number of correctly predicted amino acids representing c_3 structural form (c_3 expresses one among three structural forms: α-helix, β-structure, random coil).

Its obtained level of accuracy on average varies between 70 and 75% for different classes of secondary structures and oscillates around 95% for membrane proteins (Lukasiak, 2004).

4.2 Functional prediction

Protein function prediction is one of the most important problem in the post-genome era. The classical way for protein function prediction is to find homologies between a protein and other proteins in protein databases using programs such as FASTA (Wilbur and Lipman, 1983; Lipman and Pearson, 1985) and PSI-BLAST (Altschul et al., 1997), and then predict functions based on sequence homologies. Another sequence based approach is called the "Rosetta stone method" where two proteins are inferred to have similar functions if they are together in another genome. By comparing sequenced genomes, the phylogenetic pattern (the presence and absence of the protein in these genomes) of a protein can be determined. It is believed that proteins with similar phylogenetic patterns are likely to share similar functions. Using this idea, the functional links between genes can be predicted based on phylogenetic patterns.

Methods predicting functional similarities between proteins use two types of neural networks: *layered feed-forward networks* usually trained by simple back-propagation (Rumelhart and McClelland, 1986; Arbib, 1995), and *Kohonen maps* (Kohonen, 1982; Arbib, 1995). The first approach is based on multiple networks (Frishman and Argos, 1992) using proteins of similar sequences as input or on single networks using different amino acid features as input (Wu et al., 1996). Some other examples of networks recognizing functional motifs were presented by Hirst and Sternberg (1991, 1992); Ladunga et al. (1991); Schneider and Wrede (1993); Hansen et al. (1998); Nielsen et al. (1997). The second approach is based on using the frequency with which any of the 20*20 possible amino acid pairs occurs in the sequence (Ferran and Pflugfelder, 1993), or on using the information extracted from database annotations (Andrade and Valencia, 1997).

There are two ways to describe the principal difference between these two types of networks. Firstly, the network types can be contrasted in terms of the final results they produce. Kohonen maps provide a more continuous topography of protein function similarity, whereas back-propagation networks differentiate the input into larger categories determined by the number of output units. Secondly, the network types can be contrasted in terms of the way they are trained. Kohonen maps find an unknown classification scheme, whereas back-propagation networks learn from known examples. Consequently, back-propagation networks are useful to learn a classification from known features (e.g. types of secondary structure), while Kohonen maps have been applied to render a general classification scheme of proteins (e.g. A and B, are similar, and A is more similar to C, than B). Such a classification is *a priori* not evident (and it is itself an area of controversial research, e.g., attempting to answer questions like: "Are we more similar to an orangutan than

to a pig?"). One hope guiding such an analysis is to end up with similarities between proteins that might help to learn details in protein functions.

4.3 Multiple sequence alignment

Multiple alignments of protein sequences are important tools in studying proteins. The basic information they provide is the identification of conserved sequence regions. This information is very useful in designing experiments to test and modify the function of specific proteins, in predicting the function and structure of proteins, and in identifying new members of protein families.

Some amino acids can be replaced by others without changing tertiary structure. However, not every amino acid can be replaced by any other. On the contrary, one evolutionary step (exchange of one amino acid) can destabilize a structure. Residue substitution patterns observed in protein families are highly specific for a particular structure, and thus, contain more information about structure than single sequences. These evolutionary patterns were used by experts (Dickerson et al., 1976; Frampton et al., 1989; Benner and Gerlof, 1990; Niermann and Kirschner, 1990). However, this information can also be incorporated into neural networks in the following way. A sequence of unknown structure (U) is aligned against a database of known sequences, and proteins with significant sequence identity to U are extracted. Then, for each sequence position, the profile of residue exchanges in the final multiple alignment is compiled and fed into a network.

Almost any imaginable algorithm has been applied to the secondary structure prediction problem. However, once researchers left the path of trying to optimise black-boxes, it was through neural network applications that many break-throughs were achieved. For example, a neural network system for predicting various aspects of 1D structure based on evolutionary information is by far the most widely used prediction method (Rost et al., 1994). Other network-based methods are unique, or superior in their field (Ferran and Pflugfelder, 1993; Riis and Krogh, 1996; Andrade and Valencia, 1997; Hansen et al., 1997; Nielsen et al., 1997). Furthermore, neural networks revealed data base errors, and principles underlying protein structures (Brunak, 1991; Rost et al., 1994; Tolstrup et al., 1994; Blom et al., 1996). Thus, the neural network approach has matured from loosing the competition against experts to the method used by experts arriving at more reliable predictions than ever before.

4.4 Protein fold classification

In the pioneering works of Ptitsyn and Finkelstein (Ptitsyn and Finkelstein, 1980; Finkelstein and Ptitsyn, 1987) it has been shown that the three-dimensional arrangement of helices and strands in larger proteins can be obtained by the stepwise addition of secondary structure elements (SSEs) to basic structural motifs (Efimov, 1997). Whether this addition of SSEs reflects either a possible folding pathway for

the protein or an evolutionary history is debatable, but irrespective of any of these rationalizations, it provides a valid approach for the classification of protein folds.

Using this approach, the protein folds become organized into a phylogenetic tree (which may include 'missing-links'). Unlike clustering by similarity, the tree can be arbitrarily deep, so relating the most dissimilar folds. A disadvantage, however, is that the construction of the trees is a manual operation that embodies an implicit set of assumptions and rules that are only stated explicitly to varying degrees.

Solving the problems of protein fold classification with domain definition, a series of ideal folds (called forms) were developed and matched to known structures at the level of secondary structures (Taylor, 2000, 2002a). Each successful match simultaneously identifies a fold and defines the domain. The set of ideal folds can be organized in a table that is (not unlike a protein) equivalent to the periodic table of elements (Taylor, 2002b).This analogy is based on the correspondence between layers of a secondary structure with electronic orbitals. Just as the orbitals become filled with electrons, so the layers become filled with secondary structures. In this arrangement,a step in any direction in the table represents the addition or deletion of an SSE in one of the layers.

The organization of known structures, based on their ideal forms,embodies many of the principles discussed above for the alternativeapproaches. For example, the standard feed-forward neural networks combined with an appropriate regularization scheme can classify the fold class of a protein given solely its primary sequence at least as well as other machine learning methods that have been applied to this problem (Edler et al., 2001; Markowetz et al., 2003). They clearly outperformed standard statistical approaches (like the nearest neighbor method etc. (Edler et al., 2001)) and did not perform worse than Support Vector Machines (SVMs).

4.5 Clustering

Clustering of protein sequences from different organisms has been used to identify orthologous and paralogous protein sequences, i.e. to find protein sequences unique to an organism, and to derive the phylogenetic profile for a cluster of protein sequences. These are some of the essential components of a comparative genomics study of protein sequences across several genomes.

With the overwhelming growth of biological sequence databases, handling these amounts of data has increasingly become a problem. Protein sequences constitute one such data type. The number of unique entries in all protein sequence databases together exceeds now about a million. However, biological evolution lets proteins fall into so-called families, thus imposing a natural grouping. A protein family contains sequences that are evolutionally related. Generally, this is reflected by sequence similarity. Therefore, one aims at organizing the set of all protein sequences into clusters based on their sequence similarity.

Clustering a large set of sequences, as opposed to dealing only with the individual sequences, offers several advantages. A frequent problem is the identification

of sequences that are similar to a new query sequence. This task can be executed much quicker when only one comparison to an entire cluster has to be performed rather than one comparison per sequence. Another application lies in the possibility of analyzing evolutionary relationships among the sequences in a cluster and in the species they come from. Moreover, the presence or absence of sequences of a group of species can give useful information about their evolutionary relationship, if their complete set of protein sequences is known.

The aim of clustering protein sequences is to get a biologically meaningful partitioning.

Current systems for the classification of protein structures use different methods. SCOP (Murzin et al., 1995) is principally a manual approach while FSSP (Holm and Sander, 1997) is an automated method. Between these lie CATH (Orengo et al., 1997) and HOMSTRAD (Mizuguchi et al., 1998) which combine automation with manual curation. The basic approach, however, is the same. Proteins are divided into their component domains which are then compared pairwise and clustered into groups of similar structures. This approach works well when there is clear similarity between the structures. For distantly related (or unrelated) structures, however, it becomes difficult to define a rational hierarchy on the relationships. This difficulty is further compounded by the problem of domain definition where differences in the definitions can result in inconsistent fragmentary similarities between structures. Indeed many of the differences between the current classification systems are largely due to different domain definitions (Hadley and Jones, 1999). In contrast, CLUGEN (Ma et al., 2005) is a novel method for the clustering of protein sequences based on a new metric derived from prediction using neural networks and further utilizing the metric to model the transitive sequence homologue to detect the remote homologue. CLUGEN uses fully connected feed-forward back propagation neural network and has one hidden layer with sigmoid activation functions. The output layer of the neural network has one output unit.

4.6 Fold recognition – threading

The threading approach predicts the three-dimensional protein structure by aligning representative template protein structures with an amino acid sequence called the target sequence. The alignment computation and evaluation usually gives a sequence-structure similarity score for each alignment as the result of applying a scoring function. According to the fold recognition protocol, the alignments obtained are then sorted by their score, which yields a ranking list of target-template alignments. The best-scoring alignment should identify the template structure and its corresponding fold class which is the most compatible with the target sequence and thus constitutes a meaningful model for the yet unknown structure of the target sequence.

A number of threading programs have been developed with reduced computational cost. For example, GenTHREADER (Jones, 1999a) uses a classical se-

quence alignment algorithm to generate query-template alignments, and then evaluates the alignments by a threading potential. It provides a confidence measure for each predicted fold recognition using a neural network. The program 3D-PSSM (Kelley et al., 2000) encodes the 1D and 3D profiles, based on multiple sequence alignments among proteins of the same superfamily, into each residue position of each template protein. It finds an optimal alignment between a target sequence or its sequence profile and each template structure by matching their sequence profiles. FUGUE (Shi et al., 2001) represents one of the better performing threading programs currently available. One of its unique feature is that it utilizes environment-specific amino acid substitution tables and structure-dependent gap penalties. Some other groups (Rychlewski et al., 2000; Yona and Levitt, 2002) have applied profile–profile alignment algorithms rather than the traditional sequence–sequence or sequence–profile alignment algorithms. Interested methods in protein folding/threading can be found in Andonow et al. (2004), Balev (2004), Greenberg et al. (2004), Veber et al. (2005), Xu (2003), Xu and Li (2003), Xu et al. (2003) and Xu et al. (2004).

4.7 Domain prediction

Protein structures are usually analyzed at the level of the domain. However, the definition of a domain is not always straightforward. Small structural differences between otherwise similar proteins can have major consequences in the way a protein structure is broken into domains and even within a domain such differences can alter the way in which its fold (or topology) is perceived. Expert judgment can be used to some extent to overcome these problems, but experts do not always agree.

Sequence similarity searching is a crucial step in analyzing newly determined protein sequences. Whereas similarity searching by programs such as BLAST (Pearson and Lipman 1988; Altschul et al., 1990, 1997) or FASTA (Wilbur and Lipman, 1983; Lipman and Pearson, 1985) allows the inference of homology and/or function in many cases, identification of multidomain proteins is often problematic because their similarities point to various unrelated protein families. The best solution to this problem is the use of pattern databases that store the common sequence patterns of domain groups in the form of consensus representations (Attwood, 2000). Various pattern representation methods are in use, including regular expressions (Bairoch and Apweiler, 2000), position-dependent frequency matrices (Gribskov et al., 1987), and hidden Markov models (HMMs) (Sonnhammer et al., 1998). All of these representations are based on multiple sequence alignments. Even though these consensus pattern representations – such as used in PROSITE (Hofmann et al., 1999), PRINTS (Attwood et al., 2000), PFAM (Bateman et al., 2000), PRODOM (Corpet et al., 2000), BLOCKS (Henikoff et al., 2000), PROTFAM (Mewes et al., 2000), INTERPRO (Apweiler et al., 2000), and others (Attwood, 2000) – can be optimized and reached a very high prediction performance.

It is well known that construction of multiple alignments as well as updating them with the stream of new domain sequences requires a substantial human overhead, which is partly due to the high computational complexity of the problem. BLAST or FASTA searches on domain sequence databases (Corpet et al., 2000; Murvai et al., 2000) offer a good alternative, however, the evaluation of the output requires human judgement and/or iterative search strategies, such as those used by PSI-BLAST (Altschul et al., 1997). One of the underlying problem is that the known structural and functional domain groups are quite variable in terms of size, sequence length, as well as similarity between the members, and especially, short and variable domain sequences are sometimes quite hard to detect (Atwood, 2000).

Artificial neural networks (ANNs) have been used very successfully in biological sequence analysis for purposes as diverse as protein secondary structure prediction, recognition of signal peptide cleavage sites, gene recognition, etc. (for review, see Baldi and Brunak, 1998). Representation of sequence data in a form suitable for nonrecursive ANNs in the scope of domain prediction can be quite difficult because of the varying length of the sequences. A common solution to this problem is to use a window sliding over the protein sequence (Jagla and Schuchhardt, 2000). On the other hand, a sequence window encompassing, for example, 19 amino acids can be mapped to a $19 \times 20 = 380$ dimensional vector. Training an ANN recognizer for so many input parameters would require an enormous data set for training (Jagla and Schuchhardt, 2000). Adaptive encoding techniques can be used to find a smaller number of relevant parameters in the course of the training process. However, recognition of a short pattern, such as a signal peptide cleavage site, still required 79 input parameters (Jagla and Schuchhardt, 2000). Recognition of substantially longer protein domains, may thus require a prohibitively large number of parameters.

5 Dynamic programming

Owing to the rapid growth in the number of completely sequencedgenomes, the need for fast, reliable and automated computationaltools to derive structures and functions from protein sequences, is increasing. Recognition of native-like structural folds of an unknown protein from solved protein structures, represents the first step towards understanding its biological functions and serves as the foundation for its detailed tertiary structure prediction by comparative modeling. Dynamic programming method (cf. Bertsekas, 1995; Bertsekas and Tsitsiklis, 1996) for a thorough introduction to dynamic programming principles) is used successfully for this purpose. Before describing this in a greater detail in Section 5.2 we make a comment on the application of this method to the classical sequence alignment (Section 5.1).

5.1 Sequence alignment

The classical sequence alignment is probably the most successful application of dynamic programming in the area of computational biology. Basically, this approach is very similar for both types of sequences: DNA (RNA) and proteins. Its idea comes from Needleman and Wunsch (1970), and Smith and Waterman (1981); see also Setubal and Meidanis (1997) for a discussion of the approach. Since this method was presented in detail in the previous survey (Błażewicz et al., 1997) in the context of DNA chains, we will not discuss this subject here.

5.2 Fold recognition

Methods of protein fold recognition attempt to detect similarities between protein 3D structure that are not accompanied by any significant sequence similarity. There are many approaches, but the unifying theme is a trial to find folds that are comparable with a particular sequence. Unlike sequence-only comparison, these methods take advantage of the extra information made available by 3D structure information. As a result, it turns the protein folding problem on its head: rather than predicting how a sequence will fold, they predict how well a fold will fit a sequence.

Generally, existing fold recognition methods fall into two classes. The first class uses solely sequence information. The hidden Markov model (HMM) methods (Karplus et al., 1999) and PSI-BLAST (Altschul et al., 1997) can be classified into this category. The second class uses structural information in addition to the sequence information, in various ways. In the profile method introduced by Bowie et al. (Bowie et al., 1991), structural information, representing the local environment, is coded into each residue of a structural template. Then, various dynamic programming schemes are used for finding the optimal sequence–profile alignment (Waterman, 1995). In the threading approach, structural information is used more explicitly through evaluating the compatibility between a query sequence and a structural template in terms of residue-residue contacts, hydrophobicity, etc. Threading methods generally require more complicated algorithms to deal with the residue-residue contact term. Previous studies have shown that each approach has its own strength and weakness. For example, a threading-based method such as THREADER (Jones, 1999a) performs worse in homology recognition at the family and superfamily levels than a sequence-based method, while it achieves better performance at the fold level recognition (Lindahl and Elofsson, 2000). Motivated by such observations, researchers have attempted to combine both approaches (Jones, 1999a; Kelley et al., 2000; Panchenko et al., 2000; Shi et al., 2001), although finding an optimal way to take the full advantage of both approaches, turned out to be difficult (Lindahl and Elofsson, 2000).

The biggest disadvantage of threading-based methods is that they are computationally expensive when attempting to solve the sequence-structure alignment problem rigorously. It has been proven that the threading problem is NP-hard (Lathrop,

1994). Hence, most of the existing threading methods employ heuristic approaches to avoid computational difficulty at the expense of performance accuracy. These methods include double dynamic programming (Jones et al., 1992), frozen approximation (Godzik et al., 1992) and the Monte Carlo sampling algorithm (Bryant, 1995).

Another method for fold recognition is added to the general protein structure prediction package PROSPECT II (Kim et al., 2003). This method (PROSPECT II) has four key features. (i) an efficient way to utilize the evolutionary information for evaluating the threading potentials including singleton and pairwise energies. (ii) a two-stage threading strategy: (a) threading using dynamic programming without considering the pairwise energy and (b) fold recognition considering all the energy terms, including the pairwise energy calculated from the dynamic programming threading alignments. (iii) a combined z-score (z-score also referred as z-ratio or z-value is equal to a value of variable X minus the mean of X, divided by the standard deviation) scheme for fold recognition, which takes into consideration z-scores of each energy term. (iv) based on z-scores, a confidence index has been calculated, which measures the reliability of a prediction and a possible structure-function relationship based on a statistical analysis of a large data set consisting of threadings of 600 query proteins against the entire FSSP (Holm and Sander, 1994) templates. Tests on several benchmark sets indicate that the evolutionary information and other features of PROSPECT II greatly improve the alignment accuracy. The performance of PROSPECT II for fold recognition is significantly better (over 10%) than any other method available at all levels of similarity. The improvement in the sensitivity of fold recognition, especially at the superfamily and fold levels, makes PROSPECT II a reliable and fully automated protein structure and function prediction program for genome-scale applications.

6 The hidden Markov models

In nature one can observe a wide range of biological, physical and chemical processes that transform systems from one state into another. In some of the systems, the successive states can be represented as sequences of random variables in which the future variable value is determined by its present value independently of the way in which the present state arose from its predecessors. More precisely, the states observed in the given discrete points of time t_i seem to be determined in a probabilistic manner by the state in the previous point (or k-points) of time t_{i-1} (or $t_{i-k}, t_{i-k+1}, ..., t_{i-1}$).

Assuming that the process runs only from time 0 to time n and that the initial and final states are known, the state sequence is then represented by a finite vector. If such a process has a finite number of states, which are observed in memoryless noise, then it is called a first (or k-) order Markov process. More detailed description of Markov processes can be found in the paper written by Papoulis (Papoulis, 1984).

In 1966 Baum and Petrie (Baum et al., 1966) introduced a hidden Markov model (HMM) as a statistical model in which one assumes that the changes of the system being modeled can be well defined as a probabilistic random process, that is a stochastic Markov process. The challenge is to determine (estimate) the hidden parameters in a precise, well defined manner, from the observable parameters (Rabiner, 1989).

In the following paragraphs one can find a small spectrum of possible applications of the Hidden Markov Models for protein structure analysis.

6.1 Secondary structure prediction

In 1993 Asai and coworkers (Asai et al., 1993) introduced a new method for analyzing the amino acid sequences of proteins using the hidden Markov model (HMM). Secondary structures such as helix, strand and turn (see Lesk, 2001) are learned by HMMs, and after that applied to new sequences without known structures. The output probabilities from the HMMs are used to predict the secondary structures of the sequences. The prediction system was tested on approximately 100 sequences from a public database (Brookhaven PDB) and although the implementation was 'without grammar' (no rule for the appearance patterns of secondary structures), the result was reasonable.

An interesting overview on the biological sequence analysis can be found in the book written by Durbin et al. (Durbin et al., 1998).

In 2005 Li (Li, 2005) proposed a new kind of HMMs, so called Hidden Markov models with states depending on observations (HMMSDO). HMMSDO may have advantages over HMM in some cases such as prediction of protein secondary structures. When using HMM to predict a protein secondary structure, the observations are regarded as amino acid residues, and the states are regarded as tokens of a secondary structure (Asai et al., 1993). According to the basic assumption of biochemistry, i.e. a protein secondary structure depends on its primary structure, it may be theoretically better to use HMMSDO than HMM in this case. The current state in HMMSDO depends both on the immediately preceding state and the immediately preceding observation. Although some experiments show that HMM often outperforms HMMSDO in this application, HMMSDO may perform better than HMM when a large number of training data are used, as can be partly explained by the higher number of parameters used by the former approach (see Lee 2005).

6.2 I-sites and HMMSTR

Proteins have recurrent local sequence patterns that reflect evolutionary selective pressures to fold into stable three-dimensional structures and many of these local patterns correlate with common structural motifs. A general model of a protein sequence that captures these local features could lead to improved methods for gene

finding, protein structure prediction, remote homology detection and other applications relating to the interpretation of genomic sequence information (Bystroff et al., 2000).

Bystroff and coworkers described the development of such a model, based on the so called I-sites library of sequence-structure motifs. The I-sites (invariant or initiation sites) library consists of an extensive set of short (3 to 19 amino acids) sequence motifs obtained by exhaustive clustering of sequence segments from a nonredundant (there is no motif that contains another shorter motif completely) database of known structures (Han et al., 1996; Bystroff and Baker, 1998). However, many of the motifs overlap. The isolated motif model does not capture higher order relationships, such as the distinctly non-random transition frequencies between the different motifs. The redundancy inherent in the I-sites model suggests a better representation that would model both the diversity of the motifs and their higher order relationships. A hidden Markov model is well suited for this purpose. It consists of a set of states, each of which is associated with a probability distribution for generating a symbol, such as an amino acid residue or a secondary structure type, and a set of transition probabilities between the states. Unlike the linear hidden Markov models used to model individual protein families, HMMSTR – a hidden Markov model for local sequence-structure correlations in proteins, has a highly branched topology and captures recurrent local features of protein sequences and structures that transcend protein family boundaries. The model extends the I-sites library by describing the adjacencies of different sequence-structure motifs as observed in the protein database and, by representing overlapping motifs in a much more compact form, achieves a great reduction in parameters. The HMM attributes give a considerably higher probability to the coding sequence than does an equivalent dipeptide model. It predicts a secondary structure with an accuracy of 74,3%, backbone torsion angles better than any previously reported method and the structural context of strands and turns with an accuracy that should be useful for tertiary structure prediction (Bystroff et al., 2000). During the CASP6 (see: `http://predictioncenter.org/`) HMMSTR method was used as one of the components of the method for predicting contact maps (Yuan et al., 2004), so called residue-residue interactions, which gave good average results.

6.3 Contact maps

Pollastri and Baldi (2002) have proposed a set of flexible machine learning architectures for the prediction of contact maps (see Section 3.5). The architectures can be viewed as recurrent neural network implementations of a class of Bayesian networks, so called Generalized Input Output HMMs (GIOHMMs). Contextual information is propagated laterally through four hidden planes, one for each cardinal corner. Experiments showed that these architectures can be trained from examples and yield contact map predictors that outperform previously reported methods. The

method accurately predicted 60.5% of contacts at a distance cutoff below 8 Å and 45% of distant contacts below 10 Å, for proteins of length up to 300.

6.4 Distant homology detection

A common problem in protein structure prediction (PSP), especially in comparative modeling, is the way of finding a correct template structure. A template structure is usually identified on the basis of the sequence similarity (the measure that makes it possible to compare two or more sequences and gives higher values for sequences with a larger number of identical amino acids on the corresponding positions) with the assumption that homologous structures often have a similarity score above average. The problem is how to identify the homologous structures with a low sequence similarity. An interesting solution of this problem has been proposed by Gough et al. (Gough et al., 2001). It is worth noting that the same methods can be used for annotating sequences of unknown structures.

According to Gough et al. (Gough et al., 2001) among the sequence comparison methods, profile-based methods perform with a greater selectivity than those using pairwise comparisons. Of the profile methods, hidden Markov models (HMMs) are apparently the best. In the cited paper (Gough et al., 2001) calculations that improve the performance of HMMs and a good procedure for creating HMMs for sequences of proteins of known structures have been shown. For a family of related proteins, more homologues are detected using multiple models built from diverse single seed sequences than from one model built from a good alignment of those sequences. Some errors arising at the model-building stage of the procedure can be additionally detected and corrected. These two improvements greatly increase selectivity and coverage. Moreover, a library of HMMs, called SUPERFAMILY, has been constructed and it represents essentially all proteins of the known structures. The sequences of the domains in proteins of the known structures, that have identities less than 95% (such threshold decreases a number of redundant structures in the library and increases a diversity of sequences corresponding to the given domain), are used as seeds to build the models. The SUPERFAMILY model library has been used to annotate the sequences of over 50 genomes. The models match twice as many target sequences as are matched by pairwise sequence comparison methods. For each genome, close to half of the sequences are matched in all or in part and, overall, the matches cover 35% of eukaryotic genomes and 45% of bacterial genomes. On average, roughly 15% of genome sequences are labeled as being hypothetical yet homologous to proteins of the known structure. The annotations derived from these matches are available from a public web server at: `http://stash.mrc-lmb.cam.ac.uk/SUPERFAMILY`. This server also enables users to match their own sequences against the SUPERFAMILY model library.

A method similar to the one presented above has been proposed by Tsigelny and coworkers in 2002 (Tsigelny et al., 2002). HMMSPECTR is a tool for finding

putative structural homologs for proteins with known primary sequences. HMMSPECTR contains four major components: a data warehouse with the hidden Markov models (HMM) and alignment libraries; a search program which compares the initial protein sequences with the libraries of HMMs; a secondary structure prediction and comparison program; and a dominant protein selection program that prepares the set of 10–15 "best" proteins from the chosen HMMs. The data warehouse contains four libraries of HMMs. The first two libraries were constructed using different HMM preparation options of the HAMMER program (Eddy, 1998). The third library contains parts ("partial HMM") of initial alignments. The fourth library contains trained HMMs. The program was tested against all of the protein targets proposed in the fourth edition of CASP. The data warehouse included libraries of structural alignments and HMMs constructed on the basis of proteins publicly available in the Protein Data Bank before the CASP4 meeting. The newest fully automated versions of HMMSPECTR 1.02 and 1.02ss produced better results than the best result reported at CASP4 in 64% (HMMSPECTR 1.02) and 79% (HMMSPECTR 1.02ss) of the cases. The improvement is most notable for the difficult fold recognition targets.

7 Clustering

During the analysis of biological data it is often necessary to gather similar objects into some larger collections and group them together. Such grouping or clustering increases efficiency of data analysis and makes it possible to focus only on interesting observations. Moreover, it also allows to deduct some global relationships between biological objects, for example protein structures.

One of the most popular method of clustering is hierarchical clustering proposed by Johnson (Johnson, 1967). The method developed a useful correspondence between any system of clustering and certain kinds of distance measures (Johnson, 1985). Nowadays this kind of methods are commonly used for representing relationships between organisms in a sense of similarities and dissimilarities. Relationships are in most cases based on the genome and protein sequence analysis, sometimes clusters and dendrograms (hierarchy trees) are based on functional and structural properties.

Liu and Rost (2003) presented an interesting overview of the recent manual and automatic methods that attempt to classify proteins. They divided most popular classification strategies into several categories. First of all, they divided the methodologies between human-driven and fully automatic. In the first category they distinguish motif-based classifications, structure-based domain classification, and methods for classifying structural domain-like families. In the second category–fully automatic methods, they described measures for presenting similarities between proteins and their sequences. One important reality of sequence comparisons is that alignment methods optimize the similarity between sequences. 'Less similar' does not imply 'more distant'. To illustrate this point for structural similarity, 90%

of all pairs of proteins that have 15% identical residues over their entire length have different structures; however, 90% of the pairs of proteins with similar structure have less than 15% identical residues (amino acids).

The most widely used and comprehensive databases are SCOP (Murzin et al., 1995), CATH (Orengo et al., 1997), and FSSP (Holm and Sander, 1994), which present three methods of classifying protein structures: purely manual, a combination of manual and automated, and purely automated, respectively. A systematic comparison of these three methods can be found in the paper written by Hadley and Jones (Hadley and Jones, 1999).

The SCOP (Structural Classification of Proteins) database (Murzin et al., 1995; Andreeva et al., 2004) is developed as an evolutionary classification, in which the main focus is to place the proteins in a coherent evolutionary framework, based on their conserved structural features. The database aims at providing a comprehensive and detailed description of the relationships between all proteins whose 3D structures have been determined. A fundamental unit of classification in the SCOP database is the protein domain. A domain is defined as an evolutionary unit observed in nature either in isolation or in more than one context in multidomain proteins. The protein domains are classified hierarchically into families, superfamilies, folds and classes. The seven main classes in the latest release (1.65) contain 40 452 domains organized into 2 327 families, 1 294 superfamilies and 800 folds. These domains correspond to 20 619 entries in the Protein Data Bank (PDB). Statistics of the current and previous releases, summaries and full histories of changes and other information are available from the SCOP website (`http://scop.mrc-lmb.cam.ac.uk/scop/`) together with parsable files encoding all SCOP data (for details see Andreeva et al., 2004).

The CATH database is a hierarchical classification of domains into sequence- and structure-based families and fold groups. In the lowest level, so called S-Level, of the hierarchy, sequences are clustered according to significant sequence similarity (35% identity and above). At higher levels domains are grouped according to whether they share significant sequence, structural and/or functional similarity (homologous superfamilies, or H-level), or just structural similarity (fold or topology group, or T-level). Fold groups sharing similar architectures, i.e. similarities in the arrangements of their secondary structures regardless of connectivity are then merged into the common architectures – this level of the hierarchy is called the A-Level. At the top of the hierarchy, domains are clustered depending on their class, i.e. the percentage of α helices or β-strands (the C-Level) (Pearl et al., 2005).

FSSP is known as Families of Structurally Similar Proteins or Fold classification based on Structure-Structure alignment of Proteins. FSSP is fully automated and does not assign proteins into classes, fold families or superfamilies. Instead proteins of a representative set (sequence similarity between proteins or domains are not greater then 25%) and members of sequence-homologue set (homologues with greater then 25% sequence identity) are structurally compared using the Dali method (Holm and Sander, 1993). A fold tree is constructed using hierarchical clus-

tering methods; an indexing system is also incorporated by dividing the pairwise structural comparisons at z-scores of 2, 3, 4, 5, 10 and 15 (see Hadley and Jones, 1999).

Interesting aggregations of various sequence-motif and sequence-clusters database have been done by Kriventseva et al. (2001). They are collected in InterPro. Because the contributing databases have different clustering principles and scoring sensitivities, the combined assignments complement each other for grouping families and delineating domains.

During the CASP6 (see `http://www.predictioncenter.org`) experiment, Kolinski and Bujnicki used an average linkage hierarchical clustering algorithm, with the distance root-mean-square separation as a measure of structure similarity, as part of a strategy for rebuilding full atom models from the high-resolution reduced lattice CABS model, generated as a result of Replica Exchange Monte Carlo folding simulations (see Kolinski and Bujnicki, 2004).

Eyrich et al. (Eyrich et al., 1999) used a clustering algorithm which sorts the intermediate results of protein folding simulations into geometrically distinct groups, which can then be treated via a higher level methodology. The number of alternative predictions that are passed on to a more accurate (but more expensive) scoring function must be sufficiently small so that those computations are tractable.

Zhang and Skolnick (Zhang and Skolnick, 2004) proposed a simple and efficient strategy to identify near-native folds by clustering protein structures generated during computer simulations. In the method, called SPICKER, the most populated clusters tend to be closer to the native conformation than the lowest energy structures. To assess the generality of the approach, SPICKER was applied to 1489 representative benchmark proteins consisting of 200 residues that cover the PDB at the level of 35% sequence identity; each contains up to 280 000 structure decoys, generated using the TASSER (Threading ASSembly Refinement) algorithm. The best of the top five identified folds has a root-mean-square deviation (RMSD) from the native fold in the top 1.4% of all decoys (structures). For 78% of the proteins, the difference in RMSD from native to the identified models and RMSD from native to the absolutely best individual decoy is below 1. Although native fold identification from divergent decoy structures remains a challenge, these results show significant improvement over previous clustering algorithms.

Tendulkar et al. (2003) presented the geometric invariant-based approach for discovering recurring structural patterns in proteins via clustering. In the method, geometric invariants were used to decide superimposability of structural patterns. As a result the computationally explosive step of pairwise comparison of structures has been eliminated.

Mohseni-Zadeh et al. (2004) proposed an algorithm for the large scale clustering of protein sequences based on the extraction of maximal cliques. The Cluster-C program enables a stand-alone and efficient construction of protein families within whole proteomes (*proteome* is the complete set of proteins present in a cell or in an organism). In the presented analysis the z-value was used as the criterion for

connecting sequences. The clusters built with low threshold were consistent with known protein families.

Clustering large protein databases like the NCBI Non-Redundant database (NR) using even the best currently available clustering algorithms is very time-consuming and only practical at relatively high sequence identity thresholds. In 2001 the program CD-HI written by Li et al. (2002), clustered NR at 90% identity in one hour and at 75% identity in one day on a 1 GHz Linux PC. However, even faster clustering speed is needed because the size of protein databases are rapidly growing and many applications desire lower attainable thresholds. It was shown (Li et al., 2002) that tolerating some redundancy in output database makes far more efficient use of short-word filters and increases the program's speed by 100. Although some redundancy is present after clustering, the new results only differ from the previous results by less than 0.4%. The program and its previous version are available at `http://bioinformatics.burnham.org/cd-hi/`.

8 Summary

Protein analysis plays a central role in understanding the mechanisms of life. With several complete genomes and a reasonably complete set of protein structures, the problem facing bioinformatics shifts from its past challenge of finding weak similarities among sparse data, to one of finding closer similarities in a wealth of data. However, concentrating on protein sequence data, simplifies the data processing problem considerably and the increased computation demands can be met by the equally rapid increase in the power of computer and effective operations research techniques. Nowadays, one can find these techniques in almost each aspect of protein analysis: from secondary and tertiary structure prediction to functional analysis. There is a strong need to decode information obtained from proteins but without operations research techniques any progress seems to be impossible.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ, (1990) Basic local alignment search tool. *Journal of Molecular Biology* 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 Research* 25: 3389–3402

Althaus E, Kohlbacher O, Lenhof H-P, Muller P (2002). A combinatorial approach to protein docking with flexible side-chains. *Journal of Computational Biology*, 9(4):597–612

Andonov R, Balev S, Yanev N (2004) Protein threading: From mathematical models to parallel implementations. *INFORMS Journal on Computing*, 16(4)

Andrade MA, Valencia A (1997) Automatic annotation for biological sequences by extraction of keywords from MEDLINE abstracts. Development of a prototype system. *In Fifth International Conference on Intelligent Systems for Molecular Biology* (Gaasterland, T., Karp, P., Karplus, K., Ouzounis, C., Sander, C. et al., eds.), AAAI Press, Halkidiki, Greece, pp. 25–32

Andreeva A, Howorth D, Brenner SE, Hubbard TJP, Chothia C, Murzin AG (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acid Research* 32: 226–229

Anfinsen CB (1973) Principles that govern the folding of protein chains. *Science* 181: 223–230
Anfinsen CB, Haber E, Sela M, White Jr F (1961) The kynetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *PNAS* 47(9): 1309–1314
Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Bucher P, Codani JJ, Corpet F, Croning MDR, Durbin R (2000) InterPro – An integrated documentation resource for protein families, domains and functional sites. *Bioinformatics,* 16: 1145–1150
Arbib M (1995) *The handbook of brain theory and neural networks.* Bradford Books/The MIT Press, Cambridge, MA
Asai K, Hayamizu S, Handa K (1993) Prediction of protein secondary structure by the hidden Markov model. *Bioinformatics* 9: 141–146
Attwood TK (2000) The quest to deduce protein function from sequence: The role of pattern databases. *Int. J. Biochem. Cell. Biol.,* 32: 139–155
Attwood TK, Croning MD, Flower DR, Lewis AP, Mabey JE, Scordis P, Selley JN, Wright W (2000) PRINTS-S: the database formerly known as prints. *Nucleic Acid Research,* 28: 225–227
Bairoch A, Apweiler R (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Research,* 28: 45–48
Baldi P, Brunak S (1998) *Bioinformatics: The machine learning approach*. MIT Press, Cambridge, MA
Baldi P, Brunak S, Frasconi P, Soda G, Pollastri G (1999) Exploiting the past and the future in protein secondary structure prediction. *Bioinformatics* 15: 937–946
Balev S (2004) Solving the protein threading problem by lagrangian relaxation. *In Proceedings of the Annual Workshop on Algorithms in Bioinformatics (WABI)*, pp 182–193. Springer
Barnes E, Sokol JS, Strickland DM (2005) Optimal protein structure alignment using maximum cliques. *Operations Research*, 53:389–402
Bateman A, Birney E, Durbin R, Eddy SR, Howe KL, Sonnhammer EL (2000) The Pfam protein families database. *Nucleic Acids Research,* 28: 263–266
Baum LE, Petrie T (1966) Statistical inference for probabilistic functions of finite state Markov chains. *Annals of Mathematical Statistics* Vol. 37
Benner SA, Gerloff D (1990) Patterns of divergence in homologous proteins as indicators of secondary and tertiary structure of the catalytic domain of protein kinases. *Adv. Enz. Reg.* 31: 121–181
Bertsekas DP (1995) *Dynamic Programming and Optimal Control.* Vol. 1,2, Athena Scientific, Belmont, MA
Bertsekas DP, Tsitsiklis JN (1996) *Neuro-Dynamic Programming.* Athena Scientific, Belmont, MA
Błażewicz J, Formanowicz P, Kasprzak M (2005) Selected combinatiorial problems of computational biology. *European Journal of Operational Research* 161: 585–597
Błażewicz J, Hammer PL, Lukasiak P (2005a) Predicting Secondary structures of Proteins. *IEEE Engineering in Medicine and Biology* 24(3): 88–94
Błażewicz J, Hammer PL, Lukasiak P (2004a) Logical Analysis of Data as a predictor of protein secondary structures. *In Bioinformatics of Genome Regulations and Structure, Chapter Computational Structural Biology*, Eds. N. Kolchanov & R. Hofestaedt, Kluwer Academic Publisher, Boston, pp 145–154
Błażewicz J, Kasprzak M, Sterna M, Wêglarz J (1997) Selected combinatorial optimization problemsarising in molecular biology. *Ricerca Operativa* 26: 35–63
Błażewicz J, Lukasiak P, Milostan M (2005b) Application of tabu search strategy for finding low energy structure of protein. *Artificial Intelligence in Medicine* 35(1-2): 135–145
Błażewicz J, Dill KA, Lukasiak P, Milostan M (2004b) A Tabu search strategy for finding low energy structures of proteins in HP-model. *Computational Methods in Science and Technology,* 10: 7–19
Blom N, Hansen J, Blaas D, Brunak S (1996) Cleavage site analysis in picornaviral polyproteins: discovering cellular targets by neural networks. *Prot. Sci.* 5: 2203–2216
Bohr H, Bohr J, Brunak S, Cotterill RM, Lautrup B, Norskov L, Olsen OH, Petersen SB (1988) Protein secondary structure and homology by neural networks. The alpha-helices in rhodopsin. *FEBS Lett.* 241: 223–228
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
Branden C, Tooze J (1999) *Introduction to protein structure.* 2^{nd} ed. Garland Science Publishing, pp. 89–120
Bryant SH, Altschul SF (1995) Statistics of sequence-structure threading. *Curr. Opin. Struct. Biol.* 5: 236–244

Brunak S (1991) Non-linearities in training sets identified by inspecting the order in which neural networks learn. *In Neural Networks From Biology to High Energy Physics* (Benhar, O., Bosio, C., Del Giudice, P. & Tabet, E., eds.), Elba, Italy, pp 277–88

Bystroff C, Baker D (1998) Prediction of local structure in proteins using a library of sequence-structure motifs. *Journal of Molecular Biology* 281: 565–577

Bystroff C, Thorsson V, Baker D (2000) HMMSTR: A hidden markov model for local sequence-structure correlations in proteins. *Journal of Molecular Biology* 301: 173–90

Caprara A, Carr B, Istrail S, Lancia G, Walenz B (2004) 1001 optimal pdb structure alignments: Integer programming methods for finding the maximum contact map overlap. *Journal of Computational Biology*, 11(1):27–52

Caprara A, Lancia G (2002) Structural alignment of large-size proteins via lagrangian relaxation. *In Proceedings of the Annual International Conference on Computational Molecular Biology (RECOMB)*, pp 100–108, New York, NY, ACM Press.

Carr RD, Lancia G (2004) Compact optimization can outperform separation: a case study in structural proteomics. *4OR*, 2(3):221–233

Chazelle B, Kingsford C, Singh M (2003) The side-chain positioning problem: A semidefinite programming formulation with new rounding schemes. *In PCK50 - Principles of Computing & Knowledge*, Paris C. Kanellakis Memorial Workshop, pages 86–94. ACM Press

Chazelle B, Kingsford C, Singh M (2004) A semidefinite programming approach to side chain positioning with new rounding strategies. *INFORMS Journal on Computing*, 16(4)

Corpet F, Servant F, Gouzy J, Kahn D (2000) ProDom and ProDom-CG: Tools for protein domain analysis and whole genome comparisons. *Nucleic Acids Research,* 28: 267–269

Cuff JA, Barton GJ (1999) Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *Proteins* 34: 508–519

Dickerson RE, Timkovich R, Almassy RJ (1976) The cytochrome fold and the evolution of bacterial energy metabolism. *Journal of Molecular Biology* 100: 473–491

Doye JPK, Leary RH, Locatelli M, Schoen F (2004) Global optimization of morse clusters by potential energy transformations. *INFORMS Journal on Computing*, 16(4)

Durbin R, Eddy S, Krogh A, Mitchison G (1998) *Biological Sequence Analysis*. Cambridge University Press

Eddy SR (1998) Profile hidden Markov models. *Bioinformatics* 14: 755–763

Edler L, Grassmann J, Suhai S (2001) Role and results of statistical methods in protein fold class prediction. *Mathematical and ComputerModelling* 33: 1401–1417

Efimov AV (1997) Structural trees for protein superfamilies. *Proteins* 28: 241–260

Eriksson O, Zhou Y, Elofsson A (2001) Side chain-positioning as an integer programming problem. In O. Gascuel and B. M. E. Moret, editors, Proceedings of AnnualWorkshop on Algorithms in Bioinformatics (WABI), volume 2149 of *Lecture Notes in Computer Science*, pp 128–141, Springer

Eskow E, Bader B, Byrd R, Crivelli S, Head-Gordon T, Lamberti V, Schnabel R, (2004) An optimization approach to the problem of protein structure prediction. *Mathematical Programming*, 101(3):497–514

Eyrich VA, Standley DM, Friesner RA (1999) Prediction of protein tertiary structure to low resolution: performance for a large and structurally diverse test set. *Journal of Molecular Biology* 288(4): 725–742

Ferrán EA, Pflugfelder B (1993) A hybrid method to cluster protein sequences based on statistics and artificial neural networks. *CABIOS* 9: 671–680

Fiesler E, Beale R (1996) *Handbook of Neural Computation*. Oxford Univ. Press, New York

Finkelstein AV, Ptitsyn OB (1987) Why do globular proteins fit the limited set of folding patterns? *Prog. Biophys. Mol. Biol.* 50: 171–190

Frampton J, Leutz A, Gibson TJ, Graf T (1989) DNA-binding domain ancestry. *Nature* 342: 134–134

Frishman D, Argos P (1992) Recognition of distantly related protein sequences using conserved motifs and neural networks. *Journal of Molecular Biology* 228: 951–962

Godzik A, Skolnick J, Kolinski A (1992) Topology fingerprint approach to the inverse protein folding problem. *Journal of Molecular Biology* 227: 227–238

Gough J, Karplus K, Hughey R, Chothia C (2001) Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *Journal of Molecular Biology,* 313: 903–919

Greenberg H, Hart W, Lancia G (2004) Opportunities for combinatorial optimization in computational biology. *INFORMS Journal on Computing*, 16(3):1–22

Gribskov M, McLachlan AD, Eisenberg D (1987) Profile analysis: Detection of distantly related proteins. *Proc. Natl. Acad. Sci.,* 84: 4355–4358
Hadley C, Jones DT (1999) A systematic comparison of protein structure classifications: SCOP, CATH and FSSP. *Structure* 7: 1099–1112
Han KF, Baker D (1996) Global properties of the mapping between local amino acid sequence and local structure in proteins. *Proc. Natl Acad. Sci.* USA 93: 5814–5818
Hansen JE, Lund O, Tolstrup N, Gooley AA, Williams KL, Brunak S (1998) NetOglyc: Prediction of mucin type O-glycosylation sites based on sequence context and surface accessibility. *Glyconjug. J.* 15: 115–130
Haykin S (1999) *Neural Networks.* 2nd Edition, Prentice Hall
Henikoff JG, Greene EA, Pietrokovski S, Henikoff S (2000) Increased coverage of protein families with the blocks database servers. *Nucleic Acids Research,* 28: 228–230
Hirst JD, Sternberg MJE (1991) Prediction of ATP-binding motifs a comparison of a perceptron-type neural network and a consensus sequence method. *Protein Engineering* 4: 615–623
Hirst JD, Sternberg MJE (1992) Prediction of structural and functional features of protein and nucleic acid sequences by artificial neural networks. *Biochem.* 31: 615–623
Hofmann K, Bucher P, Falquet L, Bairoch A (1999) The PROSITE database, its status in 1999. *Nucleic Acids Researh,* 27: 215–219
Holley H, Karplus M (1989) Protein Secondary Structure Prediction with a Neural Network. *Proc. Natl Acad. Sci. USA* 86: 152–156
Holm L, Sander C (1993) Protein structures comparision by alignment of distance matrices. *Journal of Molecular Biology* 233: 123–138
Holm L, Sander C (1994) The FSSP database of structurally aligned protein fold families. *Nucleic Acids Research* 22: 3600–3609
Holm L, Sander C (1997) Dali/FSSP classification of three-dimensional protein folds. *Nucleic Acids Research* 25: 231–234
Hua S, Sun Z (2001) A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. *Journal of Molecular Biology* 308: 397–407
Jagla B, Schuchhardt J (2000) Adaptive encoding neural networks for the recognition of human signal peptide cleavage sites. *Bioinformatics,* 16: 245–250
Johnson SC (1967) Hierarchical clustering schemes. *Psychometrika* 32: 241–254
Johnson SC (1985) This week's citation classic. *Current Contents* 5: 16
Jones DT (1999) Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology* 292: 195–202
Jones DT (1999a) GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *Journal of Molecular Biology* 287: 797–815
Jones DT, Taylor WR, Thornton JM (1992) A new approach to protein fold recognition. *Nature* 358: 86–89
Karplus K, Barrett C, Cline M, Diekhans M, Grante L, Hughey R (1999) Predicting protein structure using only sequence information. *Proteins* Suppl 3: 121–125
Kelley LA, MacCallum RM, Sternberg MJE (2000) Enhanced genome annotation using structural profiles in the program 3D-PSSM. *Journal of Molecular Biology* 299: 499–520
Kneller D, Cohen F, Langridge R (1990) Improvements in protein secondary structure prediction by an enhanced neural network. *Journal of Molecular Biology* 214: 171–182
Kim D, Xu D, Guo J, Ellrott K, Xu Y (2003) PROSPECT II: protein structure prediction program for genome-scale applications. *Protein Engineering* 16(9): 641–650
Kingsford C, Chazelle B, Singh M (2005) Solving and analyzing side-chain positioning problems using linear and integer programming. *Bioinformatics*, 21(7):1028–1039
Koh SH, Ananthasurehs GK, Croke C (2004) Design of reduced protein models by energy minimization using mathematical programming. *In 10th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference*, pp 1–10
Kohonen T (1982) Self-organized formation of topologically correct feature maps. *Biol. Cybern.* 43: 59–69
Kolinski A, Bujnicki JM (2004) *Combination of Fold–Recognition with De Novo Folding and Evaluation of models.* http://www.forcasp.org/upload/2165.6.pdf
Kolinski A, Skolnick J (2004a) Reduced models of proteins and their applications. *Polymer* 45: 511–524
Kriventseva EV, Biswas M, Apweiler R (2001) Clustering and analysis of protein families *Current opinion in Structural Biolog. 2001* 11: 334–339

Ladunga I, Czakó F, Csabai I, Geszti T (1991) Improving signal peptide prediction accuracy by simulated neural network. *CABIOS* 7: 485–487

Lancia G, Carr R, Walenz B, Istrail S (2001) 101 optimal PDB structure alignments: A branch-and-cut algorithm for the maximum contact map overlap problem. *In Proceedings of the Annual International Conference on Computational Biology (RECOMB)*, pp 193–202, New York, NY, ACM Press

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

Lesk AM (2001) *Introduction to protein architecture*. Oxford University Press

Levinthal C (1968) Are there pathways to protein folding ? *J. Chem. Phys.* 65: 44–45

Lee Y (2005) Hidden Markov models with states depending on observations. *Pattern Recognition Letters,* 26: 977–984

Li W, Jaroszewski L, Godzik A (2002) Tolerating some redundancy significantly speeds up clustering of large protein databases. *Bioinformatics,* 18: 77–82

Liu J, Rost B (2003) Domains, motifs and clusters in protein universe. *Current Opinion in Chemical Biology, 2003* 7: 5–11

Lindahl E, Elofsson A (2000) Identification of related proteins on family, superfamily and fold level. *Journal of Molecular Biology* 295: 613–625

Lipman DJ, Pearson WR (1985) Rapid and sensitive protein similarity searches. *Science,* 227: 1435–1441

Lukasiak P (2004) *Algorithmic aspects of protein secondary structure prediction.* PhD Thesis, Poznan University of Technology

Ma Q, Chirn G-W, Cai R, Szustakowski J, Nirmala NR (2005) Clustering protein sequences with a novel metric transformed from sequence similarity scores and sequence alignments with neural networks. *Bioinformatics* 6: 242

Markowetz F, Edler L, Vingron M (2003) Support vector machines for protein fold class prediction. *Biometrical Journal* 45(3): 377–389

Mewes HW, Frishman D, Gruber C, Geier B, Haase D, Kaps A, Lemcke K, Mannhaupt G, Pfeiffer F, Schuller C (2000) MIPS: Adatabase for genomes and protein sequences. *Nucleic Acids Research,* 28: 37–40

Mizuguchi K, Deane CM, Blundell TL, Overington JP (1998) HOMSTRAD: A database of protein structure alignments for homologous families. *Protein Sci.* 7: 2469–2471

Mohseni-Zadeh S, Brzellec P, Risler J–L (2004) Cluster–C, an algorithm for the large-scale clustering of protein sequences based on the extraction of maximal cliques. *Computational Biology and Chemistry* 28(3): 211–218

Murvai J, Vlahovicek K, Barta E, Cataletto B, Pongor S (2000) The SBASE protein domain library, release 7.0: A collection of annotated protein sequence segments. *Nucleic Acids Research,* 28: 260–262

Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology* 247: 536–540

Nanias M, Chinchio M, O³dziej S, Czaplewski C, Scheraga HA (2005) Protein structure prediction with the UNRES force-field using Replica-Exchange Monte Carlo-with-Minimization; Comparison with MCM, CSA and CFMC. *J. Comput. Chem*., 26: 1472–1486

Needleman S, Wunsch, C (1970) A general method applicable to the search for similarities in the amino acid sequences of two proteins. *Journal of Molecular Biology* 48: 443–453

Nielsen H, Engelbrecht J, Brunak S, von Heijne G (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Engineering* 10: 1–6

Niermann T, Kirschner K (1990) Improving the prediction of secondary structure of 'TIM-barrel' enzymes. *Protein Engineering* 4: 137–147

Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM (1997) CATH-a hierarchic classification of protein domain structures. *Structure* 5: 1093–1108

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

Panchenko AR, Marchler-Bauer A, Bryant SH (2000) In *Quantitative Challenges in the Post-Genome Sequence Era: a Workshop and Symposium*. The La Jolla Interfaces in Science, La Jolla, CA, pp 2

Papoulis A (1984) Brownian Movement and Markoff Processes. *Ch. 15 in Probability, Random Variables, and Stochastic Processes.* 2nd ed. New York: McGraw–Hill, pp 515–553

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

Pearl F, Todd A, Sillitoe I, Dibley M, Redfern O, Lewis T, Bennett C, Marsden R, Grant A, Lee D, Akpor A, Maibaum M, Harrison A, Dallman T, Reeves G, Diboun I, Addou S, Lise S, Johnston C, Sillero A, Thornton J, Orengo C, (2005) The CATH Domain Structure Database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. *Nucleic Acids Research* 33: D247–D251

Pevzner PA (2001) *Computational Molecular Biology An Algorithmic Approach.* MIT Press

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

Pollastri G, Baldi P (2002) Prediction of contact maps by GIOHMMs and recurrent neural networks using lateral propagation from all four cardinal corners. *Bioinformatics* 18(1): S62–S70

Przybylski D, Rost B (2002) Alignments grow, secondary structure prediction improves. *Proteins* 46: 197–205

Ptitsyn OB, Finkelstein AV (1980) Similarities of protein topologies: evolutionary divergence, functional convergence or principles of folding? *Q. Rev. Biophys.* 13: 339–386

Qian N, Sejnowski T (1988) Predicting the secondary structure of globular proteins using neural network models. *Journal of Molecular Biology* 202: 865–884

Rabiner LR (1989) A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE* 77(2): 257–286

Riis SK, Krogh A (1996) Improving prediction of protein secondary structure using structured neural networks and multiple sequence alignments. *J. Comput. Biol.* 3: 163–183

Rost B, Sander C (1993a) Improved Prediction of Protein Secondary Structure by Use of Sequence Profiles and Neural Networks. *Proc. Natl Acad. Sci. USA* 90: 7558–7562

Rost B, Sander C (1993b) Prediction of protein secondary structure at better than 70% accuracy. *Journal of Molecular Biology* 232: 584–599

Rost B, Sander C, Schneider R (1994) PHD - an automatic server for protein secondary structure prediction. *CABIOS* 10: 53–60

Rumelhart DE, McClelland JL (1986) *Parallel distributed processing. Explorations in the microstructure of cognition*. MIT Press, Cambridge, M.A., U.S.A

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

Schneider G, Wrede P (1993) Development of artificial neural filters for pattern recognition in protein sequences. *J. Mol. Evol.* 36: 586–595

Setubal J, Meidanis, J (1997) *Introduction to Computational Biology*. PWS Publishing Company

Shi J, Blundell TL, Mizuguchi K (2001) FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *Journal of Molecular Biology* 310: 243–257

Smith TF, Waterman MS (1981) Identification of common molecular subsequences. *Journal of Molecular Biology* 147: 195–197

Sonnhammer EL, Eddy SR, Birney E, Bateman A, Durbin R (1998) Pfam: Multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Research,* 26: 320–322

Taylor WR (2000) Searching for the ideal forms of proteins. *Biochem. Soc. Trans.* 28: 264–269

Taylor WR (2002a) In Mewes,B. and Weiss,H.S. (eds), *Bioinformatics and Genome Analysis*. Springer-Verlag, Berlin, Ernst Schering Research Foundation Workshop Vol. 38, pp 133–148

Taylor WR (2002b) A 'periodic table' for protein structures. *Nature* 416: 657–660

Tendulkar AV, Wangikar PP, Sohoni MA, Samant VV, Mone ChY (2003) Parameterization and Classification of the Protein Universe via Geometric Techniques. *Journal of Molecular Biology* 334(1): 157–172

Tolstrup N, Toftgård J, Engelbrecht J, Brunak S (1994) Neural network model of the genetic code is strongly correlated to the GES scale of amino acid transfer free energies. *Journal of Molecular Biology* 243: 816–820

Tsigelny I, Sharikov Y, Ten Eyck LF (2002) Hidden Markov Models-based system (HMMSPECTR) for detecting structural homologies on the basis of sequential information, *Protein Engineering* 15(5): 347–352

Veber P, Yanev N, Andonov R, Poirriez V (2005) Optimal protein threading by cost-splitting. *In Proceedings of the Annual Workshop on Algorithms in Bioinformatics (WABI)*, pp 365–375. Springer

Wagner M, Meller J, Elber R (2004) Large-scale linear programming techniques for the design of protein folding potentials. *Mathematical Programming*, 101(2):301–318

Waterman MS (1995) *Introduction to Computational Biology*. Chapman and Hall, London
Wilbur WJ, Lipman DJ (1983) Rapid similarity searches of nucleic acid and protein data banks. *Proc. Natl. Acad. Sci. USA,* 80: 726–730
Wu CH, Zhao S, Chen H-L, Lo C-J, McLarty J (1996) Motif identification neural design for rapid and sensitive protein family search. *CABIOS* 12: 109–118
Xu J (2003) Speedup LP approach to protein threading via graph reduction. *In Proceedings of the Annual Workshop on Algorithms in Bioinformatics (WABI)*, pp 374–388. Springer
Xu J, Li M (2003) Assessment of RAPTOR's linear programming approach in CAFASP3. *Proteins: Structure, Function, and Genetics*, 53(6):579–584
Xu J, Li M, Kim D, Xu Y (2003) RAPTOR: Optimal protein threading by linear programming. *Journal of Bioinformatics and Computational Biology*, 1(1):95–117
Xu J, Li M, Xu Y (2004) Protein threading by linear programming: Theoretical analysis and computational results. *J. of Combinatorial Optimization*, 8(4):403–418
Yona G, Levitt M (2002) Within the twilight zone: a sensitive profile-profile comparison tool based on information theory. *Journal of Molecular Biology* 315: 1257–1275
Yuan X, Hou Y, Huang Y, Shao Y, Bystroff Ch (2004) *Contact Map Prediction Using HMMSTR.* http://www.bioinfo.rpi.edu/ bystrc/pub/casp6abstract.pdf.
Zhang Y, Skolnick J (2004) SPICKER: A clustering approach to identify near-native protein folds. *J. Comput. Chem* 25: 865–71