# 2

# First Steps of Protein Structure Prediction

*Karolina Majorek, Łukasz Kozłowski, Marcin Jąkalski and Janusz M. Bujnicki*

## 2.1 Introduction

The famous hypothesis formulated by the Nobel Prize laureate Christian Anfinsen states that

> *The three-dimensional structure of a native protein [. . .] is determined by the totality of interatomic interactions and hence by the amino acid sequence, in a given environment. In terms of natural selection through the 'design' of macromolecules during evolution, this idea emphasized the fact that a protein molecule only makes stable, structural sense when it exists under conditions similar to those for which it was selected – the so-called physiological state.*[1,2]

On the one hand, essentially all globular domains in proteins studied so far appear to conform to this rule. Most proteins (or at least their major fragments) have been found to fold into unique, well-defined, stable three-dimensional structures under very broadly defined 'physiological conditions' that also include 'laboratory conditions' under which protein samples are prepared for biophysical and biochemical characterization. In agreement with Anfinsen's hypothesis, variations in conditions (e.g. change of pH or addition of a ligand) or changes in a sequence (e.g. due to proteolytic cleavage and removal of a sequence fragment) may result in structural changes that are often functionally relevant, e.g. if a protein's function requires opening and closing of a cavity that is used for binding of another molecule. On the other hand, a growing number of protein sequences (or sequence fragments) have been found to be mostly unstructured (review:[3]). These 'intrinsically disordered proteins' (IDPs) may assume a defined structure only under very specific

*Prediction of Protein Structures, Functions, and Interactions* Edited by Janusz Bujnicki

conditions, e.g. in the presence of another molecule (e.g. upon binding to another protein or a ligand). In the absence of a stabilizing factor, these sequences exist as an ensemble of rapidly interconverting different conformations. Thus, for many IDPs, the Anfinsen's 'stable physiological state for which the protein was selected' is significantly different from 'standard' conditions assumed for other proteins.

Anfinsen's hypothesis implies that the knowledge of amino acid sequence and of a given environment should be sufficient to infer the native structure of a protein (or to predict the lack of a stable structure). However, despite seemingly solid theoretical foundations, accurate prediction of protein structure from the primary chemical structure of the polypeptide chain remains one of the greatest challenges in biology. Thus far, there have been two major approaches to predict the unknown protein structure from known amino acid sequence: one relies on our knowledge of 'first principles', i.e. the laws of physics, while the other is based on rules inferred from comparative analysis of experimentally solved protein structures. Both approaches had some successes, but neither of them has achieved the final goal.

The 'physical' approach has been successfully applied already in 1951 by Linus Pauling and Robert Corey, who predicted the existence of two periodic structural motifs formally defined by the pattern of hydrogen bonds that may be formed by the protein backbone: the spiral $\alpha$-helix with 3.6 amino acid residues per turn[4] and the flat $\beta$-sheet comprising two or more $\beta$-strands having an extended zigzag conformation.[5] These two secondary structure elements (SSEs) now are known to be major features of protein architecture, with $>50\%$ of residues of an average protein assuming either helical or extended conformation. Helices and strands provide a natural frame for insightful protein structure visualization (with a helix often represented as a tube or a spiral and strand as an arrow), and are widely used to describe protein three-dimensional folds. They are also used by many programs that use simplified protein structure representation (e.g. SSEs instead of individual amino acids) to speed up calculations, e.g. for superposition of protein structures.

Secondary structure is much more conserved in the evolution than amino acid sequence; therefore accurate prediction of SSEs from sequence would be of great benefit in structural bioinformatics. For instance the knowledge of SSEs can help to guide sequence alignment or improve existing sequence alignment of remotely related sequences with low sequence similarity (see Chapter 1 by Kaminska *et al*. in this volume). Secondary structure prediction is also a good starting point toward elucidating the three-dimensional structure – it serves as an intermediate step in the protein fold-recognition procedure, i.e. identification of templates for comparative modeling (see Chapter 4 by Kosinski *et al*.) and may provide useful restraints both in comparative modeling and in *de novo* modeling (see Chapter 5 by Gront *et al*.). However, it has been found that it is quite difficult to predict accurately, which type of secondary structure is assumed by each amino acid residue of a protein.[6] Computational simulations of peptide and protein folding based on the 'physics-based' approach have been carried out in attempt to predict both local structure and global conformation (review:[7]). The first applications of force field methods to study peptide conformations date back to calculations performed by Nemethy and Scheraga.[8] However, due to an extremely large number of degrees of freedom and very complex calculations of energies, such simulations require extremely large computer resources, such as supercomputers or massively parallel distributed computing.[9] Alas, despite recent advancement in computer hardware and software, physics-based simulation techniques remain incapable of confidently predicting structures of even moderately sized proteins ($>100$ amino acid residues).

On the other hand, as soon as the first protein structures solved by X-ray crystallography have been determined, it has been observed that secondary structures tend to exhibit regular arrangements of amino acid residues of certain type. The regularities are due to the local periodicity of helical and extended conformations (3.6 and 2 residues per repeated segment, respectively), and the global tendency of a protein to form a well-packed hydrophobic core and to place hydrophilic residues at the surface (at least in water-soluble proteins). For example, SSEs buried in the protein core are composed mainly of hydrophobic residues, while SSEs at the protein surface tend to be amphipathic and show an alternating pattern of hydrophobic and hydrophilic residues (usually 010101 for strands and 0011011 or 0010011 for helices), allowing the respective side chains to be buried in the protein core or exposed to the solvent. Thus, as soon as a sufficient number of protein structures have been determined to make useful statistics, 'knowledge-based' methods have been developed, aiming at predicting SSEs based on calculated tendencies of different residues or peptide fragments to assume particular conformations. The first attempt of predicting secondary structure of polypeptides using the 'knowledge-based' approach dates back to the same time as the afore-mentioned physics-based analyses. It was performed in 1965 by Guzzo,[10] who inferred simple rules for preferences of different amino acid residues to form helical and non-helical regions. With the growing number of available protein structures and sequences the statistics have quickly improved, and new algorithms have been developed, yielding methods that are far from perfect, but achieve useful accuracy of about 80%. Currently the 'knowledge-based' approach is used by essentially all methods for secondary structure prediction, as well as methods for order/disorder prediction and for inference of other simple structural features from the primary sequence. This chapter aims at providing a comprehensive overview of these methods. The 'physical' approach to protein structure prediction is beyond the scope of this chapter and will not be reviewed here – instead, it will be referred to in other chapters, in particular Chapter 5 by Gront *et al*. Because of significant differences between proteins that function as water-soluble and those that are embedded in biological membranes, secondary structure prediction methods for each of two types are different and will be discussed separately. We will also discuss prediction of higher order motifs formed by certain types of SSEs, the so called 'supersecondary' structures (e.g. coiled coils or $\beta$ hairpins), and prediction of contacts between residues that are remote in primary structure.

## 2.2 Definition of Secondary Structure and Its Assignment for Known Protein Structures

One caveat of the knowledge-based structure prediction is the requirement of unambiguous definition of secondary structure elements or ordered vs. disordered regions. Statistical methods and machine learning methods require the input data to be appropriately classified in order to make meaningful predictions. Unfortunately, defining the boundaries between disordered and ordered regions or between helix, sheet, and coil structures is arbitrary, and commonly accepted standard assignments do not exist. Therefore, various researchers employed different criteria that in some cases have lead to considerably different assignments.

Secondary structure is formally defined by the hydrogen bonds, but the hydrogen bonding is correlated with other features, such as dihedral angels generally adopted by particular types of secondary structure, which has given rise to less formal definitions of SSEs. To standardize secondary structure assignment, the Dictionary of Secondary Structure in Proteins (DSSP) was designed.[11] It was the first method for protein secondary structure assignment available as a computer program, and has remained most popular until now. DSSP classifies each amino acid residue in a protein with known 3D structure into one of 8 types of secondary structure, based on recognition of hydrogen bonding patterns and geometrical features defined in terms of the concepts torsion and curvature of differential geometry. The 7 types of SSEs include the $\alpha$-helix (H), the $\beta$-strand (E, for 'extended'), and less frequent types of secondary structure namely non-$\alpha$ helices: 3/10 (G) and $\pi$ (I), isolated $\beta$-bridge (B), highly curved bend (S) and hydrogen-bonded turn (T), while the remaining residues are classified as outside of SSEs.[11]

Another quite popular method for secondary structure assignment named STRIDE was proposed later by Frishman and Argos.[12] STRIDE is based on the combined use of hydrogen bond energy and statistically derived backbone torsional angle information, with parameters of the pattern recognition procedure optimized to improve (compared to DSSP) agreement with manual designations provided by the crystallographers as a standard-of-truth. More recently developed methods for secondary structure assignment include P-SEA,[13] SECSTR,[14] KAKSI,[15] SEGNO,[16] and PALSSE.[17] These methods are based on either the hydrogen-bond pattern, geometric features, expert knowledge or their combinations. However, they often disagree on their assignments, up to 25%. The discrepancy among different methods is caused by nonideal configurations of helices and sheets in experimentally solved structures and by different definitions of helices and strands. Of particular interest is PALSSE, which identifies only two types of SSEs that can be approximated by vectors: helix and strand. In contrast to other algorithms, which identify a secondary structure state for every residue in a protein chain, PALSSE attributes residues to SSEs in such a way that consecutive elements may overlap, thus allowing residues located at the overlapping region to have more than one secondary structure type. This method is robust to coordinate errors and can be used to define SSEs even in poorly refined and low-resolution structures (e.g. if only C-$\alpha$ atoms are available, thus if no hydrogen-bonds are present). PALSSE usually assigns a larger fraction of residues to SSEs as compared to other methods, e.g. 80% vs. 53% in the case of DSSP.[17]

Discrepancies with structural assignment concern not only algorithms. Protein structures are dynamic objects with some regions more mobile than others. Local conformations near the ends of secondary structures vary under native conditions, but may be forced to assume a single conformation in crystals due to packing constraints, hence secondary structure assignments differ by about 5–15 percentage points between different X-ray versions or different NMR models for the same protein[6] (Figure 2.1). This inherent protein flexibility is the main reason why the theoretical upper limit of secondary structure prediction accuracy is about 90%, for a particular SSE assignment method. Recently it has been proposed that instead of relying on single structures, structure assignment methods should be assessed based on the similarity of the secondary structures assigned to established pairwise sequence-alignment benchmarks, where these benchmarks are determined by prior structural alignments of the protein pairs. The use of this criterion has led to identification of STRIDE and KAKSI as the most robust methods (PALSSE was not included in that

***Figure 2.1*** *Illustration of secondary structure, conformational variability, and order vs disorder. Three-dimensional structure of small protein SUMO-1 solved by NMR (1a5r in Protein Data Bank), 10 alternative models shown in the 'cartoon' representation. Black spirals indicate α-helices, dark grey arrows indicate β-strands, white coils indicate loops. While the central part shows relatively ordered structure (with only some fluctuations at one end of the helix), the N- and C-terminal regions (left and right, respectively) show 'intrinsic disorder'. Interestingly, a short helical region persists in the disordered N-terminal tail, demonstrating the presence of secondary structure despite the absence of stable tertiary structure*

comparison), and to development of a consensus of STRIDE, KAKSI, SECSTR, and P-SEA, called SKSP, which is 2–3% higher in agreement with structurally aligned residues than DSSP for three established alignment benchmarks.[18]

Summarizing, assignments of secondary structure for one particular protein may vary, depending on the method used. Thus, in theoretical protein structure prediction it is important to select, which type of assignment is going to be predicted. Thus far, DSSP has been used as the 'golden standard' because of its popularity among crystallographers, but it is likely that methods for secondary structure assignment that are more consistent with the 3D structure alignments (e.g. SKSP) may lead to improved secondary structure prediction. Another caveat of secondary structure prediction methods is that they are aimed at predicting only the three basic classes of local structure: (H)elix, (E)xtended, and (C)oil; thus, e.g. an 8-letter 'structural alphabet' used in the DSSP notation is reduced to a 3-letter alphabet. In different algorithms it is done according to different conversion rules, which may yield an apparent increase of accuracy, but cause errors when the predicted secondary structure is used to predict 3D structure.[6] Karplus and coworkers found that the replacement of simple alphabets of secondary structure with highly informative, detailed alphabets can improve detection and alignment of structurally similar, but remotely related proteins.[19] Examples of such alphabets include STR, an enhanced version of DSSP that subdivides DSSP letter E (strand) into six letters, according to properties of a residue's relationship to its strand partners (number of partners and their parallel/antiparallel character) or Protein Blocks, a set of overlapping protein backbone fragments of length 5 amino acid residues.[20] HMMSTR[21] implements yet another solution to this problem: in this method protein structure is represented by short structural fragments taken from the database of known structures; for secondary structure it uses an alphabet of 11 conformation states, 10 corresponding to Φ-Ψ angle regions and one for cis-peptide bonds. A recently developed method Real-SPINE[22,23] predicts real values of torsion angles from a given sequence.

## 2.3 Prediction of Secondary Structure and Solvent Accessibility for Water-soluble Proteins

The first empirical prediction system aiming at predicting SSEs from protein sequence, based on statistics calculated from structures solved by X-ray crystallography, was developed by Fasman and Chou.[24,25] This very simple method is based on analysis of the relative frequency of each amino acid in helices, strands and coils and on assumption that single residue individually influences secondary structure. Subsequently, a more sophisticated GOR method has been developed, which is based on information theory and Bayesian statistics, and takes into account not only one residue but also adjacent positions in the sequence. These methods have been built-in into commercial software packages for protein sequence analysis and structure modeling and have become very popular among biologists despite their accuracy was only slightly better than random. The main limitation of these early methods was a small amount of known 3D structures, from which parameters could be derived. Besides, these methods did not utilize any evolutionary information and were applicable to single sequences rather than for multiple sequence alignments (MSAs) of homologous sequences.[26] Although over the years, both the Chou-Fasman[27] and the GOR methods[28] have been improved, the level of their accuracy is inferior to the best modern methods.

A significant improvement in prediction accuracy (>70%) has been achieved by 'second generation' methods such as PHD,[29] SAM-T98,[30] and PSIPRED,[31] which utilized MSA-derived information concerning sequence conservation, often combined with machine learning techniques such as artificial neural networks (ANNs), nearest-neighbor search (NNS) methods, and support vector machines (SVMs), or advanced statistical methods such as Hidden Markov Models (HMM) (review:[32]). These methods were also made available as web servers (Table 2.1). MSA, provided by the user or generated by an internal routine of an algorithm, is usually based on identification of homologs by searches of protein sequence databases (see Chapter 1 by Kaminska *et al*. in this volume). It is important to note that PSIPRED was the first method, in which iterative PSI-BLAST sequence searches have been introduced, compared to single-pass searches in earlier methods. Currently, iterative database searches to obtain the input MSA for prediction methods are considered a standard. Typically, patterns in sequence variability observed in MSA provide information on conservation of core elements (hydrophobic core and regions important for protein function), while the location of insertions and deletions (indels) hints at a position of surface-exposed loops. Incorporated machine learning techniques allow training the methods on known structures to learn characteristic sequence-structure patterns and then use those patterns to predict the secondary structure of the query protein. Most of the SSE prediction methods of the above-mentioned generation, or their derivatives developed later, have been associated with predictors of solvent accessibility used to identify residues that are buried to different extents in the hydrophobic core (Table 2.2).

In addition to methods for predicting the three main types of SEEs, there are several methods based on sequence profile analysis for predicting certain types of local structure, including various hairpin structures,[52] or specialized in $\alpha$-turns,[71,72] $\beta$-turns,[51,73] $\gamma$-turns,[50] and $\pi$-turns.[74] There are also methods to predict conformation of individual residues, e.g. the trans/cis state of Pro.[75,76] To our knowledge, these types of methods have not yet been

***Table 2.1*** *Software for secondary structure prediction*

| Program | URL (http://) |
|---|---|
| Three-state (Helix/Extended/Coil) prediction | |
| IPSSP[33] (for single sequences) | exon.gatech.edu/genemark/ipssp/webIPSSP.cgi |
| PSIPRED[31] | bioinf.cs.ucl.ac.uk/psipred/ |
| SSPRO[34] | scratch.proteomics.ics.uci.edu/ |
| PHD[29] | www.predictprotein.org/ |
| PROFsec[35] | www.predictprotein.org/ |
| PRED2ARY[36] | alexander.ucsf.edu/∼jmc/pred2ary/ |
| APSSP2[37] | www.imtech.res.in/raghava/apssp2/ |
| PREDATOR[38] | ftp://ftp.ebi.ac.uk/pub/software/unix/predator/ |
| HMMSTR[21] | www.bioinfo.rpi.edu/∼bystrc/hmmstr/ |
| NPREDICT[39] | www.cmpharm.ucsf.edu/∼nomi/nnpredict.html |
| PORTER[40] | distill.ucd.ie/porter/ |
| HYPROSPII[41] | bioinformatics.iis.sinica.edu.tw/HYPROSPII/ |
| SAM-T06[42] | www.soe.ucsc.edu/compbio/SAM_T06/T06-query.html |
| JNET[43] | www.compbio.dundee.ac.uk/Software/JNet/jnet.html |
| SABLE[44] | sable.cchmc.org/ |
| YASSPP[45] | glaros.dtc.umn.edu/yasspp/ |
| YASPIN[46] | ibivu.cs.vu.nl/programs/yaspinwww/ |
| CRNPred[47] | ftp.bioinformatics.org/pub/crnpred/ |
| JUFO3D[48] | www.meilerlab.org/index.php |
| SPINE[22] | sparks.informatics.iupui.edu/SPINE/spine.html |
| Other types of secondary and supersecondary structure, and other types of local conformation | |
| TURNS ($\alpha$, $\beta$, $\gamma$)[49,50] | imtech.res.in/raghava/ |
| $\beta$-Turn[51] | serine.umdnj.edu/∼zhangq3/betaturn/prediction.htm |
| TURNPRED[52] | www.meilerlab.org/index.php |
| COILS[53] | www.ch.embnet.org/software/COILS_form.html |
| MARCOIL[54] | www.isrec.isb-sib.ch/webmarcoil/webmarcoilC1.html |
| PCOILS[55] | toolkit.tuebingen.mpg.de/pcoils |
| PairCoil2[56] | groups.csail.mit.edu/cb/paircoil2/paircoil2.html |
| MultiCoil[57] | groups.csail.mit.edu/cb/multicoil/cgi-bin/multicoil.cgi |
| LearnCoil[58] | groups.csail.mit.edu/cb/learncoil-vmf/cgi-bin/vmf.cgi |
| 'Meta-servers' for secondary structure prediction | |
| JPRED[59] | www.compbio.dundee.ac.uk/∼www-jpred/ |
| NPS@[60] | npsa-pbil.ibcp.fr |
| META-PP[61] | www.predictprotein.org/meta.php |
| PROTEUS[62] | wishart.biology.ualberta.ca/proteus |
| DISTILL[63] | distill.ucd.ie |
| GeneSilico[64] | genesilico.pl/meta2/ |

integrated into metaservers for secondary or tertiary structure prediction and their practical utility for protein modeling and function prediction remains to be established.

Currently the recommended approach to secondary structure prediction involves combining the results of different methods; it may involve advanced machine learning approaches, such as voting, linear discrimination, neural networks or decision trees[77] or even simple

***Table 2.2*** *Software for solvent accessibility prediction*

| Program | URL (http://) |
|---|---|
| Jnet[43] | www.compbio.dundee.ac.uk/Software/JNet/jnet.html |
| PHDacc[35] | www.predictprotein.org/ |
| PROFacc[35] | www.predictprotein.org/ |
| SABLE[65] | sable.cchmc.org |
| NETASA[66] | www.netasa.org |
| MLRprdsec[67] | spg.biosci.tsinghua.edu.cn/ |
| WESA[68] | pipe.scs.fsu.edu/wesa.html |
| ACCpro[69] | scratch.proteomics.ics.uci.edu/ |
| SARpred[70] | www.imtech.res.in/raghava/sarpred/ |
| SPINE[22] | sparks.informatics.iupui.edu/SPINE/spine.html |
| PaleAle[63] | distill.ucd.ie/paleale/ |

consensus.[78] The idea of combining different prediction methods was first implemented in JPRED,[59] a consensus meta-server that standardizes input and output requirements of a range of secondary structure prediction algorithms, each representing a different prediction strategy, and computes a consensus of PHD, NNSSP, DSC, and PREDATOR secondary structure predictions. In addition, the output of the JPRED server includes predictions of solvent accessibility by the JNET method,[43] as well as predictions of coiled-coil regions and transmembrane helices (see below), which however are not directly incorporated in the calculation of the secondary structure.

The most recent class of meta-approaches, exemplified by PROTEUS[62] exploits the observation that if an experimentally determined three-dimensional structure of a closely related protein is known, then copying the secondary structure assignment from the known structure provides a better result than by predicting it *de novo*. PROTEUS initially carries out a sequence similarity search against the PDB database in order to determine if the whole or a part of the query sequence is significantly similar to a known structure, and if such a template structure is found, secondary structure mapping is carried out from the template to the query based on a sequence alignment. For the sequence segments that are not covered by template structures, *de novo* secondary structure prediction is carried out with three different, high quality neural network approaches (PSIPRED, JNET and TRANSSEC), whose results are combined into a consensus prediction by the fourth neural network. Merging template-based predictions and *de novo* predictions allows PROTEUS to yield a full sequence prediction, regardless of the extent of sequence overlap to a PDB hit (when complete 3D-to-2D mapping is achieved, when only partial coverage is provided and when no homologue with known structure can be found), and to achieve high average accuracy of $>80\%$ per residue. A similar approach of merging template-based and *de novo* predictions of secondary structure and solvent accessibility has been implemented in the DISTILL suite.[63]

While the early methods of secondary structure prediction were about 60–65% accurate, with accuracy for $\beta$-strands only slightly better than random,[6] the best modern methods reach about 80% accuracy per residue,[22] with $\sim$10% lower accuracy for $\beta$-strands. The difference between theoretical upper limit of prediction accuracy and actual secondary

structure prediction accuracy, and between level of prediction accuracy of $\alpha$-helices and $\beta$-strands, is mainly due to difficult to detect long-range interactions that may influence secondary structure formation. It has been shown that the same amino acid sequence of substantial length may fold as $\alpha$-helix when in one position in primary protein sequence but as $\beta$-sheet when in another sequence context.[79] Besides, during the folding process, a certain fragment of a protein might first adopt a secondary structure preferred by the local sequence and later, because of non-local interactions, be transformed to another secondary structure. The latter concern has been addressed in the method 3D-JUFO,[48] which combines iterative *de novo* secondary structure prediction using an approach similar to PSIPRED with tertiary structure prediction with the ROSETTA method (see Chapter 5 by Gront *et al*.), followed by re-prediction of SSEs based on local environment of particular residues observed in models of tertiary structure. 3D-JUFO achieves remarkable accuracy of 80%, with notable improvement of accuracy for $\beta$-strand prediction (76%) over sequence-only methods. Another interesting recently developed method that brings the accuracy of secondary structure prediction close to the theoretical limit combines bioinformatics methodology with experimental techniques of circular dichroism (CD) and Fourier transform infrared (FTIR) spectroscopy for assessing the overall secondary structure content.[80]

## 2.4 Prediction of Secondary Structure for Transmembrane Proteins

Membrane proteins are different from water-soluble proteins in that a large fraction of their surface is hydrophobic to enable stability in the environment of a lipid bilayer. They constitute about 20–30% of all proteins in the fully sequenced genomes, and are typically involved in cell signaling, molecular pumping and energy transduction. Integral membrane proteins consist of one or more transmembrane (TM) segments and can be divided into two structural classes: the $\alpha$-helical TM proteins and the $\beta$-barrel TM proteins, varying in structure, localization and physicochemical features. Typical TM proteins of the more abundant $\alpha$-helical class are present in all types of biological membranes including outer membranes. They comprise one or more hydrophobic $\alpha$-helical membrane spanning regions separated by hydrophilic loops that are exposed into the solvent (review:[81]). TM $\beta$-barrel proteins are found only in outer membranes of Gram-negative bacteria, cell wall of Gram-positive bacteria, and outer membranes of mitochondria and chloroplasts. They consist of different number of antiparallel, membrane spanning $\beta$-strands with a simple up-and-down topology.[82]

TM proteins aggregate and precipitate in water and require detergents or nonpolar solvents for extraction, therefore they are much more difficult to analyze experimentally than their soluble counterparts, in relation to all steps from overexpression to high-resolution structure determination. Although TM proteins represent the most important drug targets, their structure determination has lagged behind that for soluble proteins; currently they represent less than 1% of available crystal structures.[83] On the one hand, this situation generates a great deal of pressure to develop effective methods for predicting the structure of TM proteins. On the other hand, the paucity of structural data hampers the development of knowledge-based approaches. Nonetheless, for both types of TM proteins specialized structure predictors have been designed, but due to the relatively easily detectable patterns

of hydrophobic residues forming $\alpha$-helical TM segments and much smaller amount of known $\beta$-barrel TM proteins structures, the majority of them was focused on the $\alpha$-helical TM proteins until quite recently (review:[84]).

Prediction of TM helices should be intuitively easy due to their hydrophobic nature. However, predictions based solely on hydrophobicity profiles have high error rates. Besides, hydrophobic signal peptides may be confused with TM helices. It is also a consecutive challenge to predict TM proteins topology. Prediction of the way in which TM segments cross the membrane (inside-out or outside-in) is done mainly by considering the different charge distribution between the inside (cytoplasmic) and outside (extracellular) regions connecting TM segments, and by application of the so-called 'positive-inside rule'[85] based on the observation that there is an overrepresentation of positively charged residues in the intracellular loops of TM proteins. Contemporary approaches usually predict both localization of TM segments and their orientation (topology). The best methods such as PHOBIUS[86] or MEMSAT3[87] utilize evolutionary information as well as discriminate against signal peptides. Prediction of TM segments for $\beta$-barrel proteins is more difficult, because the strands are amphipathic. They contain 10–22 residues with alternating hydrophobic side chains facing the lipid bilayer and hydrophilic side chains facing the internal pore. To predict the $\beta$-barrel type of TM proteins a small number of specialized algorithms have been developed based on standard statistical and machine learning techniques including HMMs, ANNs, or SVMs (Table 2.3).

As in the case of secondary structure prediction for globular soluble proteins, consensus methods perform much better compared to each individual prediction method separately and the recommended strategy for identification membrane spanning segments and their orientation in membranes is to use many different methods and combine results into a consensus prediction. Examples of 'metaservers' that combine the results of several individual methods, providing a more accurate consensus prediction, include BPROMPT,[100] ConPredII,[110] and PONGO[111] for $\alpha$-helical TM proteins, and ConBBPRED[114] for $\beta$-barrel TM proteins. The newest trends in TM structure prediction include meta-predictions that utilize predictions of solvent accessibility and secondary structure propensity typical for globular proteins in the form of 'structural profiles'.[102] There have also been attempts to make concurrent prediction of secondary and tertiary structure by simulating folding in lipid membranes, e.g. with modified versions of *de novo* structure prediction methods FRAGFOLD[115] and ROSETTA.[116,117]

In addition to predictors specific for TM proteins, a new method MeTaDoR has been recently proposed that predicts membrane-binding peripheral proteins that do not form an integral part of the membrane, but bind to it mostly in a reversible manner and thereby function in various important processes, including cell signaling and membrane trafficking.[113]

## 2.5 Prediction of Supersecondary Structure

Individual SSEs may be arranged in simple geometrical shapes forming recurring supersecondary structures. There is a number of well-defined $\alpha$–$\alpha$, $\beta$–$\beta$, $\alpha$–$\beta$ and $\beta$–$\alpha$ structural motifs that serve as 'building blocks' of tertiary structure. Prediction of supersecondary structures can be an important step toward building a tertiary structure from the specified secondary structure elements.[118] The $\beta$–hairpin, comprising two adjacent antiparallel

***Table 2.3*** *Software for prediction of TM regions in proteins*

| Program | URL (http://) |
|---|---|
| | $\alpha$-TM proteins |
| HMMTOP[88] | www.enzim.hu/hmmtop/ |
| DAS[89] | www.sbc.su.se/∼miklos/DAS/ |
| PHDhtm[90] | www.predictprotein.org/ |
| TMAP[91] | bioinfo4.limbo.ifm.liu.se/tmap/index.html |
| TMHMM[92] | www.cbs.dtu.dk/services/TMHMM/ |
| Tmpred[93] | www.ch.embnet.org/software/TMPRED_form.html |
| MEMSAT3[87] | bioinf.cs.ucl.ac.uk/memsat |
| TopPred2[94] | bioweb.pasteur.fr/seqanal/interfaces/toppred.html |
| WHAT[95] | saier-144-37.ucsd.edu/what.html |
| THUMBUP[96] | sparks.informatics.iupui.edu/Softwares-Services_files/thumbup.htm |
| UMDHMM[96] | sparks.informatics.iupui.edu/Softwares-Services_files/umdhmm.htm |
| PRED-TMR[97] | athina.biol.uoa.gr/PRED-TMR/ |
| HMM-TM[98] | biophysics.biol.uoa.gr/HMM-TM/ |
| ORIENTM[99] | athina.biol.uoa.gr/orienTM/ |
| BROMPT[100] | www.jenner.ac.uk/BPROMPT |
| LOCALIZOME[101] | localizome.org |
| PHOBIUS[86] | phobius.sbc.su.se/ |
| MINNOU[102] | minnou.cchmc.org |
| | $\beta$-Transmembrane proteins |
| BBF*[103] | www-biology.ucsd.edu/∼msaier/transport/software/bbfsource.tar.gz |
| HMM-B2TMR**[104] | gpcr.biocomp.unibo.it/biodec/ |
| MINNOU[102] | minnou.cchmc.org |
| B2TMPRED[105] | gpcr.biocomp.unibo.it/cgi/predictors/outer/pred_outercgi.cgi |
| PRED-TMBB[106] | bioinformatics.biol.uoa.gr/PRED-TMBB/ |
| PROFtmb[107] | cubic.bioc.columbia.edu/services/proftmb/ |
| TMBETA-NET[108] | psfs.cbrc.jp/tmbeta-net/ |
| BOMP[109] | www.bioinfo.no/tools/bomp |
| | Metaservers |
| BPROMPT ($\alpha$)[100] | www.jenner.ac.uk/BPROMPT |
| ConPredII ($\alpha$)[110] | bioinfo.si.hirosaki-u.ac.jp/∼ConPred2/ |
| PONGO ($\alpha$)[111] | pongo.biocomp.unibo.it/pongo/ |
| TUPS ($\alpha$)[112] | sparks.informatics.iupui.edu/Softwares-Services_files/tups.htm |
| ConBBPRED ($\beta$)[106] | bioinformatics.biol.uoa.gr/ConBBPRED/ |
| | Membrane-binding peripheral proteins |
| MeTaDo[113] | proteomics.bioengr.uic.edu/metador |

*The BBF program is freely available to academic users upon request to the corresponding author.
**HMM-B2TMR is a commercial program, demo version is available.

hydrogen bonded $\beta$-strands, is an example of the frequently occurring motif for which predictors have been developed. BhairPred[119] is an example of a method for discriminating hairpins from non-hairpins; obviously it achieves high accuracy only if the prediction of secondary structure is correct. Coiled coils are another type of super-secondary structure characterized by a bundle of two or more $\alpha$-helices wrapping around each other. Coiled coil structures have been implicated in inter- and intraprotein interactions, and may be

formed by helices formed by segments of sequence distant in the primary structure or even contributed by different proteins. Thus, coiled coils allow monomeric building blocks to form complex assemblages that can serve as molecular motors and springs (review: ref.[120]). The helices forming coiled coils have a unique pattern of hydrophobicity, which repeats every seven residues (five hydrophobic and two hydrophilic). This sequence periodicity has prompted the development of special algorithms to predict the location of $\alpha$-helices that form coiled coils. According to the recent benchmark, the two best computational methods are a HMM-based MARCOIL[54] and PCOILS,[55] followed by PairCoil2[56] (Table 2.2).

## 2.6 Disorder Prediction

During the past decade, the literature has exploded with reports on intrinsically unstructured proteins (IDPs). Currently it is estimated that 30–60% of proteins are predicted to contain long stretches of disordered residues. Many of the disordered regions have been confirmed experimentally; they have been often found to be essential for protein function. Interestingly, intrinsic disorder appears to be significantly correlated with certain terms from functional ontologies and with specific functional motifs.[121–124] In particular, linear motifs[125] that harbor sites of posttranslational modification, such as phosphorylation, or sites of protein–protein interactions, often fall into regions that are locally disordered or undergo order-disorder transition in different, biologically relevant situations.[126,127] (see Chapter 1 by Kaminska *et al*. in this volume). With respect to molecular/biochemical function, IDPs have been frequently implicated in protein-nucleic acid interactions as transcription factors or in protein–protein interactions as e.g. regulators of enzyme activity. With respect to cellular roles, they have been implicated in regulatory processes, in particular in regulation of gene expression on the level of transcription and RNA processing, and in cellular signaling. On a more general level, IDPs are crucial for cell survival, proliferation, differentiation and apoptosis. Dysfunctions of IDPs may therefore lead to cancer, which makes them particularly important from a biomedical point of view. On the other hand, disordered regions often prevent crystallization of proteins, or the generation of interpretable NMR data, and in protein bioinformatics – they introduce compositional biases that hamper comparison of sequences of ordered regions. Recognition of disordered regions in a protein is therefore important for delineating boundaries of stably folded protein domains for structural and functional studies and for reducing bias in sequence similarity analyses by avoiding alignment of disordered regions against ordered ones (reviews:[128,129]). Detection of disordered regions may also facilitate identification of domains (see Chapter 1 by Kaminska *et al*.). A very important resource for disorder is the DISPROT database (http://www.disprot.org).[130] It links structure and function information for proteins that contains at least one experimentally determined disordered region.

The relatively frequent occurrence of IDPs and their importance in understanding protein structure-function relationships and cellular processes make it worthwhile to develop predictors of protein disordered regions. Since the SEG algorithm for identification of low-complexity regions that are typically associated with molecular disorder was developed in 1994,[131] an increasing number of groups have been developing such methods. However, as with secondary structure, it is not immediately clear how to unambiguously define

'disorder'. The lack of stable structure and conformational heterogeneity can manifest itself either at the secondary or tertiary level, and may include sites with varying extent of residual secondary structure and conformational mobility: molten globules, pre-molten globules, liquid-like collapsed-disordered state, or gas-like extended-disordered state.[132] Various researchers employed different criteria for defining disorder, resulting in numerous predictors that attempt to identify different features. Thus, depending on a research question being asked, using a single disorder predictor may be insufficient to achieve a meaningful prediction. Ferron *et al.*[128] presented an informative review of a number of methods published until 2006, highlighting their advantages and drawbacks. Table 2.4 presents succinct descriptions of disorder predictors, taking into account also the most recently published methods.

According to our own benchmark focused at accuracy of predicting regions of short disorder (using the criterion employed in CASP-7, i.e. the absence of resolved coordinates in crystal structures[153]), the best methods include POODLE,[147] DisPSSMP,[138] and iPDA.[142] We have developed a meta-predictor that reports the results of two primary coiled-coil predictors (COILS and Marcoil; see above and Table 2.2), and 10 primary disorder predictors (DISOPRED2, GlobPlot, Spritz, DISPROT (VSL2), IUPred, POODLE-L, POODLE-S, iPDA, PrDOS, and DisPSSMP, see Table 2.4). It also calculates a consensus prediction. The disorder meta-predictor is available via the gateway of the GeneSilico metaserver[64] at http://genesilico.pl/meta2/.

## 2.7 Prediction of Long-range Contacts between Amino Acid Residues

In addition to predicting local structure, a number of methods have been developed to predict contacts between residues that are remote in primary structure. This type of information is of particular interest, because it has been shown that it is possible to directly infer three-dimensional protein structures, if a sufficiently large number of contacts are known with sufficient accuracy. It has been estimated that as few as one contact on average per seven residues may be sufficient.[154] Various measures of distance and various thresholds may be used to define a contact between two residues (see e.g. ref.[155]), however the most common definition of contact used in prediction methods is as a C$\beta$-C$\beta$ pair (C$\alpha$ in the case of Gly residue) less than or equal to 8 Å apart.[156] According to the recent benchmark within the framework of the CASP7 experiment, the best contact predictor is an ANN associated with the SAM-T06 structure prediction server.[157] Other well-performing programs (according to the CASP7 benchmark or to other tests published by their authors) that are available as web-servers have been summarized in Table 2.5.

Special kinds of methods for long-range contact prediction are those for identification of Cys residues involved in disulfide bond formation (review: ref.[158]). Disulfide bonds are primary covalent cross-links between two Cys residues in proteins that play critical roles in stabilizing the protein structures. They can impose a substantial distance and angular constraint on the backbone of protein, thus making a large contribution to the stabilization of protein tertiary structures. A number of proposed algorithms for prediction of disulfide bonding states of Cys (involved in disulfide formation or not), as well as prediction of disulfide connectivity patterns (with the prior knowledge of disulfide bonding states) have been implemented as freely available web servers (Table 2.5). Most of these methods

***Table 2.4*** *Software for disorder prediction*

| Program | URL (http://) | Short description |
|---|---|---|
| DisEMBL™[133] | Dis.embl.de/ | ANN trained to predict classic loops (DSSP), flexible loops with high B-factors, missing coordinates in X-ray structures, regions of low-complexity and prone to aggregation. |
| DISOPRED2[134] | bioinf.cs.ucl.ac.uk/disopred/disopred.html | SVM trained to predict residues with missing coordinates. Standalone version available. |
| DISpro[135] | www.ics.uci.edu/∼baldig/dispro.html | Recursive neural networks (RNNs) trained to predict missing coordinates. |
| DISPROT[136,137] | www.ist.temple.edu/disprot/predictor.php | VL2 (least-squares regression) and VL3 (ANN) predict long disorder, VSL2 predicts both short and long disorder. Standalone version available. |
| DisPSSMP[138] | biominer.bime.ntu.edu.tw/dispssmp/ | Radial Basis Function Network (RBFN) trained to predict missing coordinates. |
| DRIP-PRED[139] | sbcweb.pdc.kth.se/cgi-bin/maccallr/disorder/submit.pl | Self-organizing maps (SOMs) trained to predict missing coordinates. |
| FoldIndex©[140] | Bip.weizmann.ac.il/fldbin/findex | Simple method to predict whether a given protein will fold or not, based on average hydrophobicity and net charge. |
| FoldUnfold[141] | skuld.protres.ru/∼mlobanov/ogu/ogu.cgi | A statistical method to predict regions of weak packing density (less than 8 Å between heavy atoms of non-adjacent residues). |
| GlobPlot2[133] | globplot.embl.de/ | A simple method based on several hydrophobicity scales to predict regions of missing coordinates and loops with high B-factors. |
| iPDA[142] | biominer.cse.yzu.edu.tw/ipda | A successor of DisPSSMP. Incorporates information about sequence conservation, predicted secondary structure, sequence complexity and hydrophobic clusters. |
| IUPred[143] | iupred.enzim.hu/ | Estimates pairwise interaction energies using a statistical potential. Disordered regions tend to exhibit poor inter-residue contact capacity. |
| NORSp[144] | www.rostlab.org/services/NORSp/ | Predicts long regions exposed to the solvent, with no regular secondary structure. |
| PONDR®[145] | www.pondr.com/ | A commercial package containing several predictors based on FFNNs. |

***Table 2.4*** *(continued)*

| Program | URL (http://) | Short description |
|---|---|---|
| POODLE (S,L,W)[146,147] | mbs.cbrc.jp/poodle/poodle.html | L predicts long disorder using an SVM. S adds analysis of PSSMs generated by PSI-BLAST to detect short disorder. W uses Joachims' spectral graph transducer (SGT) to classify entire proteins as either disordered or ordered. |
| PrDOS[148] | prdos.hgc.jp | Predicts missing coordinates using SVM and PSSMs from PSI-BLAST. |
| PreLink[149] | genomics.eu.org/prelink/ | Identifies regions with biased composition and poor in hydrophobic clusters to predict regions with missing coordinates. |
| RONN[150] | www.strubi.ox.ac.uk/RONN | Predicts missing coordinates using an ANN. |
| SEG[131] | mendel.imp.ac.at/METHODS/seg.server.html | Ancient precursor of modern disorder predictors, identifies regions of low sequence complexity. |
| SPRITZ[151] | distill.ucd.ie/spritz/ | Predicts long and short disorder (missing coordinates) using two separate SVMs. Utilizes secondary structure predicted by PORTER. |
| Grishin Lab Disorder Predictor[152] | prodata.swmed.edu/disorder/disorder_prediction/predict.cgi | Predicts missing coordinates based on a PSSM and optimized propensities of amino acid residues toward disorder. |
| GeneSilico[64] | genesilico.pl/meta2/ | A metaserver that predicts different types of disorder using weighted consensus of several methods. |

employ combinations of various machine-learning techniques and utilize information from multiple sequence alignments (e.g. to identify correlated Cys pairs) and predicted secondary structure. Their reported prediction accuracy reaches 80%, but different methods have not been compared directly with each other on the same test set.

## 2.8 Summary

To obtain better quality of secondary structure prediction, when no related structures are known, it is advisable to follow some general rules:

First, it is important to use multiple sequence information, but if target sequence shows high similarity to none or to only a few other proteins it is worth trying to search different databases (e.g. not only the non-redundant database at the NCBI, but also protein sequences deduced from unfinished genomes and environmental sequencing projects) to find moderately divergent sequences that can be used to build MSA (see also Chapter 1

***Table 2.5*** *Software for prediction of long-range contacts and disulfide bonds*

| Program | URL (http://) |
|---|---|
| SAM-T06[157] | www.soe.ucsc.edu/research/compbio/SAM_T06/T06-query.html |
| GPCPred[159] | sbcweb.pdc.kth.se/cgi-bin/maccallr/gpcpred/submit.pl |
| PROFcon[160] | www.predictprotein.org/submit_profcon.html |
| CONpro[161] | www.ics.uci.edu/∼baldig/scratch/ |
| SVMcon[162] | www.bioinfotool.org/svmcon.html |
| CRNPRED[47] | ftp.bioinformatics.org/pub/crnpred/ |
| CMAPpro[69] | scratch.proteomics.ics.uci.edu/ |
| Distill[151] | distill.ucd.ie/distill/ |
| PoCM[163] | foo.maths.uq.edu.au/∼nick/Protein/contact.html |
| CMA[164] | ligin.weizmann.ac.il/cma/ |
| HMMSTR-CM[165] | www.bioinfo.rpi.edu/∼bystrc/hmmstr/server.php |
| BETApro ($\beta$)[166] | www.ics.uci.edu/∼baldig/betasheet.html |
| DiANNA (C-C)[167] | clavius.bc.edu/∼clotelab/DiANNA/ |
| DIpro 2.0 (C-C)[168] | contact.ics.uci.edu/bridge.html |
| DISULFIND (C-C)[169] | cassandra.dsi.unifi.it/disulfind/index.php |
| GDAP (C-C)[170] | www.doe-mbi.ucla.edu/∼boconnor/GDAP/ |
| DCON (C-C)[171] | gpcr.biocomp.unibo.it/cgi/predictors/cys-cys/pred_dconcgi.cgi |
| DISULFIDE (C-C)[172] | foo.maths.uq.edu.au/∼huber/disulfide/ |

by Kaminska *et al*.). If the sequence is a true 'ORFan' with no homologs, a specialized method IPSSP[33] (Table 2.1) may be used. Alignments that include remotely related sequences should be inspected in the most divergent regions, and sequences that cannot be aligned with confidence should be removed. In case of secondary structure prediction algorithms that do not accept an MSA as an input (but e.g. construct one from scratch by themselves), secondary structure may be predicted independently for a few homologous sequences and checked for mutual consistency. Correctly aligned positions should display similar structure; therefore regions of low sequence similarity with different predictions should be checked for possible errors in MSA.

Second, we recommend using meta-servers for disorder and secondary structure prediction, because combining results of several best prediction methods into a consensus prediction is more reliable than relying on any individual method alone. Agreement between methods usually indicates confident prediction, while disagreement may indicate various things: different peculiarities of methods used, poor alignment in the input data, and/or non-standard type of secondary structure, such as surface-exposed $\beta$-strands with bulges that are often mispredicted as helices due to their irregular pattern of hydrophobic and hydrophilic residues. It is also important to remember that specialized methods are usually better for predicting particular types of structure than general-purpose methods for secondary structure prediction. Therefore, it may be useful to use methods for prediction of TM regions to pre-screen sequence for non-globular elements and 'mask' them before considering regular secondary structure prediction.

Third, when selecting 'best' methods for consensus prediction it is important to remember that many authors use different benchmarks to assess their methods and that many

published accuracies have been shown to be overestimated, when these methods were assessed in rigorous blind tests on standard benchmarks, such as within CASP[173] or EVA.[174] Although secondary structure predictions are no longer assessed in CASP, the EVA website (http://cubic.bioc.columbia.edu/eva/) is updated automatically each week, to cope with the large number of existing prediction servers and the constant changes in the prediction methods. EVA currently assesses servers for secondary structure prediction, contact prediction, comparative protein structure modeling and threading/fold recognition. The identity of the test set assures that the competition is fair, while a large sample of targets assures that methods are compared reliably.

Fourth, we recommend making simultaneous predictions of secondary structure, solvent accessibility, and disorder, as they usually reinforce each other (e.g. regions of disorder usually exhibit little tendency to form secondary structure and their residues are predicted to be largely solvent-exposed). However, discrepancies in this regard (i.e. presence of confidently predicted secondary structure and/or buried residues within regions of disorder) may indicate interesting structural and functional elements, such as partially folded molten globule-like structures or candidates for linear motifs (see Chapter 1 by Kaminska *et al*. in this volume). Thus, again we recommend using meta-servers for making predictions on the level of primary and secondary structure, in particular if they are going to be used as restraints for modeling of protein tertiary structure (see articles by Kosinski *et al*. and Gront *et al*. in this volume).

## Acknowledgements

We thank present and former members of the Bujnicki lab in IIMCB and at the UAM for stimulating discussions and collaboration in development of some of the methods mentioned in this article. The authors acknowledge the support from past and current grants for the development of structure prediction methods from Polish Ministry of Science, NIH, Framework Programme of the EU, EMBO, and HHMI.

## References

1. C.J. Epstein, R.F. Goldberger, and C.B. Anfinsen, The genetic control of tertiary protein structure. Model systems. *Cold Spring Harb Symp Quant Biol*, **28**, 439–449 (1963).
2. C.B. Anfinsen, Studies on the principles that govern the folding of protein chains. Nobel lecture, December 11, 1972, in *Nobel Lectures, Chemistry 1971–1980*, T. Frängsmyr (Ed), World Scientific Publishing Co., Singapore, 1993.
3. H.J. Dyson, and P.E. Wright, Intrinsically unstructured proteins and their functions, *Nat Rev Mol Cell Biol*, **6**, 197–208 (2005).
4. L. Pauling, R.B. Corey, and H.R. Branson, The structure of proteins; Two hydrogen-bonded helical configurations of the polypeptide chain, *Proc Natl Acad Sci U S A*, **37**, 205–211 (1951).
5. L. Pauling, and R.B. Corey, The pleated sheet, a new layer configuration of polypeptide chains, *Proc Natl Acad Sci U S A*, **37**, 251–256 (1951).
6. B. Rost, Review: Protein secondary structure prediction continues to rise, *J Struct Biol*, **134**, 204–218 (2001).
7. G.A. Chasse, A.M. Rodriguez, M.L. Mak, *et al*., Peptide and protein folding, *J Mol Struct THEOCHEM*, **537**, 319–361 (2001).

8. G. Nemethy, and H.A. Scheraga, Theoretical determination of sterically allowed conformations of a polypeptide chain by a computer method, *Biopolymers*, **3**, 155–181 (1965).
9. H.A. Scheraga, M. Khalili, and A. Liwo, Protein-folding dynamics: Overview of molecular simulation techniques, *Annu Rev Phys Chem*, **58**, 57–83 (2007).
10. A.V. Guzzo, The influence of amino-acid sequence on protein structure, *Biophys J*, **5**, 809–822 (1965).
11. W. Kabsch, and C. Sander, Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features, *Biopolymers*, **22**, 2577–2637 (1983).
12. D. Frishman, and P. Argos, Knowledge-based protein secondary structure assignment, *Proteins*, **23**, 566–579 (1995).
13. G. Labesse, N. Colloc'h, J. Pothier, and J.P. Mornon, P-Sea: A new efficient assignment of secondary structure from C alpha trace of proteins, *Comput Appl Biosci*, **13**, 291–295 (1997).
14. M.N. Fodje, and S. Al-Karadaghi, Occurrence, conformational features and amino acid propensities for the Pi-helix, *Protein Eng*, **15**, 353–358 (2002).
15. J. Martin, G. Letellier, A. Marin, J.F. Taly, A.G. de Brevern, and J.F. Gibrat, Protein secondary structure assignment revisited: A detailed analysis of different assignment methods, *BMC Struct Biol*, **5**, 17 (2005).
16. M.V. Cubellis, F. Cailliez, and S.C. Lovell, Secondary structure assignment that accurately reflects physical and evolutionary characteristics, *BMC Bioinformatics*, **6 Suppl 4**, S8 (2005).
17. I. Majumdar, S.S. Krishna, and N.V. Grishin, Palsse: A program to delineate linear secondary structural elements from protein structures, *BMC Bioinformatics*, **6**, 202 (2005).
18. W. Zhang, A.K. Dunker, and Y. Zhou, Assessing secondary structure assignment of protein structures by using pairwise sequence-alignment benchmarks, *Proteins*, (2007).
19. R. Karchin, M. Cline, Y. Mandel-Gutfreund, and K. Karplus, Hidden Markov models that use predicted local structure for fold recognition: Alphabets of backbone geometry, *Proteins*, **51**, 504–514 (2003).
20. C. Etchebest, C. Benros, S. Hazout, and A.G. de Brevern, A structural alphabet for local protein structures: Improved prediction methods, *Proteins*, (2005).
21. C. Bystroff, V. Thorsson, and D. Baker, Hmmstr: A hidden Markov model for local sequence-structure correlations in proteins, *J Mol Biol*, **301**, 173–190 (2000).
22. O. Dor, and Y. Zhou, Achieving 80% ten-fold cross-validated accuracy for secondary structure prediction by large-scale training, *Proteins*, **66**, 838–845 (2007).
23. B. Xue, O. Dor, E. Faraggi, and Y. Zhou, Real-value prediction of backbone torsion angles, *Proteins*, (2008).
24. P.Y. Chou, and G.D. Fasman, Prediction of protein conformation, *Biochemistry*, **13**, 222–245 (1974).
25. P.Y. Chou, and G.D. Fasman, Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins, *Biochemistry*, **13**, 211–222 (1974).
26. J.M. Levin, S. Pascarella, P. Argos, and J. Garnier, Quantification of secondary structure prediction improvement using multiple alignments, *Protein Eng*, **6**, 849–854 (1993).
27. H. Chen, F. Gu, and Z. Huang, Improved Chou-Fasman method for protein secondary structure prediction, *BMC Bioinformatics*, **7 Suppl 4**, S14 (2006).
28. A. Kloczkowski, K.L. Ting, R.L. Jernigan, and J. Garnier, Combining the Gor V algorithm with evolutionary information for protein secondary structure prediction from amino acid sequence, *Proteins*, **49**, 154–166 (2002).
29. B. Rost, C. Sander, and R. Schneider, Phd: An automatic mail server for protein secondary structure prediction, *Comput Appl Biosci*, **10**, 53–60 (1994).
30. J. Park, K. Karplus, C. Barrett, *et al*., Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods, *J Mol Biol*, **284**, 1201–1210 (1998).
31. D.T. Jones, Protein secondary structure prediction based on position-specific scoring matrices, *J Mol Biol*, **292**, 195–202 (1999).
32. J. Heringa, Computational methods for protein secondary structure prediction using multiple sequence alignments, *Curr Protein Pept Sci*, **1**, 273–301 (2000).

33. Z. Aydin, Y. Altunbasak, and M. Borodovsky, Protein secondary structure prediction for a single-sequence using hidden semi-Markov models, *BMC Bioinformatics*, **7**, 178 (2006).
34. G. Pollastri, D. Przybylski, B. Rost, and P. Baldi, Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles, *Proteins*, **47**, 228–235 (2002).
35. B. Rost, G. Yachdav, and J. Liu, The Predictprotein Server, *Nucleic Acids Res*, **32**, W321–326 (2004).
36. J.M. Chandonia, and M. Karplus, Neural networks for secondary structure and structural class predictions, *Protein Sci*, **4**, 275–285 (1995).
37. G.P.S. Raghava. Apssp2: A combination method for protein secondary structure prediction based on neural network and example based learning. Casp5 Online Abstract. A-132, 2002.
38. D. Frishman, and P. Argos, Seventy-five percent accuracy in protein secondary structure prediction, *Proteins*, **27**, 329–335 (1997).
39. D.G. Kneller, F.E. Cohen, and R. Langridge, Improvements in protein secondary structure prediction by an enhanced neural network, *J Mol Biol*, **214**, 171–182 (1990).
40. G. Pollastri, and A. McLysaght, Porter: A new, accurate server for protein secondary structure prediction, *Bioinformatics*, **21**, 1719–1720 (2005).
41. H.N. Lin, J.M. Chang, K.P. Wu, T.Y. Sung, and W.L. Hsu, Hyprosp Ii-a knowledge-based hybrid method for protein secondary structure prediction based on local prediction confidence, *Bioinformatics*, **21**, 3227–3233 (2005).
42. K. Karplus, S. Katzman, G. Shackleford, *et al*., Sam-T04: What is new in protein-structure prediction for Casp6, *Proteins*, **61 Suppl 7**, 135–142 (2005).
43. J.A. Cuff, and G.J. Barton, Application of multiple sequence alignment profiles to improve protein secondary structure prediction, *Proteins*, **40**, 502–511 (2000).
44. R. Adamczak, A. Porollo, and J. Meller, Combining prediction of secondary structure and solvent accessibility in proteins, *Proteins*, **59**, 467–475 (2005).
45. G. Karypis, Yasspp: Better kernels and coding schemes lead to improvements in protein secondary structure prediction, *Proteins*, **64**, 575–586 (2006).
46. K. Lin, V.A. Simossis, W.R. Taylor, and J. Heringa, A simple and fast secondary structure prediction method using hidden neural networks, *Bioinformatics*, **21**, 152–159 (2005).
47. A.R. Kinjo, and K. Nishikawa, Crnpred: Highly accurate prediction of one-dimensional protein structures by large-scale critical random networks, *BMC Bioinformatics*, **7**, 401 (2006).
48. J. Meiler, and D. Baker, Coupled prediction of protein secondary and tertiary structure, *Proc Natl Acad Sci U S A*, **100**, 12105–12110 (2003).
49. H. Kaur, and G.P. Raghava, Prediction of beta-turns in proteins from multiple alignment using neural network, *Protein Sci*, **12**, 627–634 (2003).
50. H. Kaur, and G.P. Raghava, A neural-network based method for prediction of gamma-turns in proteins from multiple sequence alignment, *Protein Sci*, **12**, 923–929 (2003).
51. Q. Zhang, S. Yoon, and W.J. Welsh, Improved method for predicting {beta}-turn using support vector machine, *Bioinformatics*, (2005).
52. M. Kuhn, J. Meiler, and D. Baker, Strand-loop-strand motifs: prediction of hairpins and diverging turns in proteins, *Proteins*, **54**, 282–288 (2004).
53. A. Lupas, M. Van Dyke, and J. Stock, Predicting coiled coils from protein sequences, *Science*, **252**, 1162–1164 (1991).
54. M. Delorenzi, and T. Speed, An Hmm model for coiled-coil domains and a comparison with Pssm-based predictions, *Bioinformatics*, **18**, 617–625 (2002).
55. M. Gruber, J. Soding, and A.N. Lupas, Comparative analysis of coiled-coil prediction methods, *J Struct Biol*, **155**, 140–145 (2006).
56. A.V. McDonnell, T. Jiang, A.E. Keating, and B. Berger, Paircoil2: Improved prediction of coiled coils from sequence, *Bioinformatics*, **22**, 356–358 (2006).
57. E. Wolf, P.S. Kim, and B. Berger, Multicoil: A program for predicting two- and three-stranded coiled coils, *Protein Sci*, **6**, 1179–1189 (1997).
58. M. Singh, B. Berger, and P.S. Kim, Learncoil-Vmf: Computational evidence for coiled-coil-like motifs in many viral membrane-fusion proteins, *J Mol Biol*, **290**, 1031–1041 (1999).

59. J.A. Cuff, M.E. Clamp, A.S. Siddiqui, M. Finlay, and G.J. Barton, Jpred: A consensus secondary structure prediction server, *Bioinformatics*, **14**, 892–893 (1998).
60. Y. Guermeur, C. Geourjon, P. Gallinari, and G. Deleage, Improved performance in protein secondary structure prediction by inhomogeneous score combination, *Bioinformatics*, **15**, 413–421 (1999).
61. V.A. Eyrich, and B. Rost, Meta-Pp: Single interface to crucial prediction servers, *Nucleic Acids Res*, **31**, 3308–3310 (2003).
62. S. Montgomerie, S. Sundararaj, W.J. Gallin, and D.S. Wishart, Improving the accuracy of protein secondary structure prediction using structural alignment, *BMC Bioinformatics*, **7**, 301 (2006).
63. G. Pollastri, A.J. Martin, C. Mooney, and A. Vullo, Accurate prediction of protein secondary structure and solvent accessibility by consensus combiners of sequence and structure information, *BMC Bioinformatics*, **8**, 201 (2007).
64. M.A. Kurowski, and J.M. Bujnicki, Genesilico protein structure prediction meta-server, *Nucleic Acids Res*, **31**, 3305–3307 (2003).
65. R. Adamczak, A. Porollo, and J. Meller, Accurate prediction of solvent accessibility using neural networks-based regression, *Proteins*, **56**, 753–767 (2004).
66. S. Ahmad, and M.M. Gromiha, Netasa: Neural network based prediction of solvent accessibility, *Bioinformatics*, **18**, 819–824 (2002).
67. S. Qin, Y. He, and X.M. Pan, Predicting protein secondary structure and solvent accessibility with an improved multiple linear regression method, *Proteins*, **61**, 473–480 (2005).
68. H. Chen, and H.X. Zhou, Prediction of solvent accessibility and sites of deleterious mutations from protein sequence, *Nucleic Acids Res*, **33**, 3193–3199 (2005).
69. J. Cheng, A.Z. Randall, M.J. Sweredoski, and P. Baldi, Scratch: A protein structure and structural feature prediction server, *Nucleic Acids Res*, **33**, W72–76 (2005).
70. A. Garg, H. Kaur, and G.P. Raghava, Real value prediction of solvent accessibility in proteins using multiple sequence alignment and secondary structure, *Proteins*, **61**, 318–324 (2005).
71. H. Kaur, and G.P. Raghava, Prediction of alpha-turns in proteins using Psi-blast profiles and secondary structure information, *Proteins*, **55**, 83–90 (2004).
72. Y. Wang, Z. Xue, and J. Xu, Better prediction of the location of alpha-turns in proteins with support vector machine, *Proteins*, **65**, 49–54 (2006).
73. H. Kaur, and G.P. Raghava, A neural network method for prediction of beta-turn types in proteins using evolutionary information, *Bioinformatics*, **20**, 2751–2758 (2004).
74. Y. Wang, Z.D. Xue, X.H. Shi, and J. Xu, Prediction of Pi-turns in proteins using Psi-blast profiles and secondary structure information, *Biochem Biophys Res Commun*, **347**, 574–580 (2006).
75. M.L. Wang, W.J. Li, M.L. Wang, and W.B. Xu, Support vector machines for prediction of peptidyl prolyl Cis/trans isomerization, *J Pept Res*, **63**, 23–28 (2004).
76. J. Song, K. Burrage, Z. Yuan, and T. Huber, Prediction of Cis/trans isomerization in proteins using Psi-blast profiles and secondary structure information, *BMC Bioinformatics*, **7**, 124 (2006).
77. R.D. King, M. Ouali, A.T. Strong, *et al*., Is it better to combine predictions?, *Protein Eng*, **13**, 15–19 (2000).
78. M. Albrecht, S.C. Tosatto, T. Lengauer, and G. Valle, Simple consensus procedures are effective and sufficient in secondary structure prediction, *Protein Eng*, **16**, 459-462 (2003).
79. D.L. Minor, Jr., and P.S. Kim, Context-dependent secondary structure formation of a designed protein sequence, *Nature*, **380**, 730–734 (1996).
80. J.G. Lees, and R.W. Janes, Combining sequence-based prediction methods and circular dichroism and infrared spectroscopic data to improve protein secondary structure determinations, *BMC Bioinformatics*, **9**, 24 (2008).
81. J.L. Popot, and D.M. Engelman, Helical membrane protein folding, stability, and evolution, *Annu Rev Biochem*, **69**, 881–922 (2000).
82. G.E. Schulz, Transmembrane beta-barrel proteins, *Adv Protein Chem*, **63**, 47–70 (2003).
83. K. Lundstrom, Structural genomics for membrane proteins, *Cell Mol Life Sci*, **63**, 2597–2607 (2006).

84. A. Elofsson, and G. von Heijne, Membrane protein structure: Prediction versus reality, *Annu Rev Biochem*, **76**, 125–140 (2007).
85. L. Sipos, and G. von Heijne, Predicting the topology of eukaryotic membrane proteins, *Eur J Biochem*, **213**, 1333–1340 (1993).
86. L. Kall, A. Krogh, and E.L. Sonnhammer, A combined transmembrane topology and signal peptide prediction method, *J Mol Biol*, **338**, 1027–1036 (2004).
87. D.T. Jones, Improving the accuracy of transmembrane protein topology prediction using evolutionary information, *Bioinformatics*, **23**, 538–544 (2007).
88. G.E. Tusnady, and I. Simon, The Hmmtop transmembrane topology prediction server, *Bioinformatics*, **17**, 849–850 (2001).
89. M. Cserzo, E. Wallin, I. Simon, G. von Heijne, and A. Elofsson, Prediction of transmembrane alpha-helices in prokaryotic membrane proteins: The dense alignment surface method, *Protein Eng*, **10**, 673–676 (1997).
90. B. Rost, P. Fariselli, and R. Casadio, Topology prediction for helical transmembrane proteins at 86% accuracy, *Protein Sci*., **5**, 1704–1718 (1996).
91. F. Milpetz, P. Argos, and B. Persson, Tmap: A new email and WWW service for membrane-protein structural predictions, *Trends Biochem Sci*, **20**, 204–205 (1995).
92. E.L. Sonnhammer, G. von Heijne, and A. Krogh, A hidden Markov model for predicting transmembrane helices in protein sequences, *Proc Int Conf Intell Syst Mol Biol*, **6**, 175–182 (1998).
93. K. Hofmann, and W. Stoffel, Tmbase – a database of membrane spanning proteins segments, *Biol Chem Hoppe Seyler*, **347**, 166 (1993).
94. G. von Heijne, Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule, *J Mol Biol*, **225**, 487–494 (1992).
95. Y. Zhai, and M.H. Saier, Jr., A web-based program (What) for the simultaneous prediction of hydropathy, amphipathicity, secondary structure and transmembrane topology for a single protein sequence, *J Mol Microbiol Biotechnol*, **3**, 501–502 (2001).
96. H. Zhou, and Y. Zhou, Predicting the topology of transmembrane helical proteins using mean burial propensity and a hidden-Markov-model-based method, *Protein Sci*, **12**, 1547–1555 (2003).
97. C. Pasquier, and S.J. Hamodrakas, An hierarchical artificial neural network system for the classification of transmembrane proteins, *Protein Eng*, **12**, 631–634 (1999).
98. P.G. Bagos, T.D. Liakopoulos, and S.J. Hamodrakas, Algorithms for incorporating prior topological information in Hmms: Application to transmembrane proteins, *BMC Bioinformatics*, **7**, 189 (2006).
99. T.D. Liakopoulos, C. Pasquier, and S.J. Hamodrakas, A novel tool for the prediction of transmembrane protein topology based on a statistical analysis of the Swissprot database: The Orientm algorithm, *Protein Eng*, **14**, 387–390 (2001).
100. P.D. Taylor, T.K. Attwood, and D.R. Flower, Bprompt: A consensus server for membrane protein prediction, *Nucleic Acids Res*, **31**, 3698–3700 (2003).
101. S. Lee, B. Lee, I. Jang, S. Kim, and J. Bhak, Localizome: A server for identifying transmembrane topologies and Tm helices of eukaryotic proteins utilizing domain information, *Nucleic Acids Res*, **34**, W99–103 (2006).
102. B. Cao, A. Porollo, R. Adamczak, M. Jarrell, and J. Meller, Enhanced recognition of protein transmembrane domains with prediction-based structural profiles, *Bioinformatics*, **22**, 303–309 (2006).
103. Y. Zhai, and M.H. Saier, Jr., The Beta-Barrel Finder (Bbf) Program, Allowing identification of outer membrane beta-barrel proteins encoded within prokaryotic genomes, *Protein Sci*, **11**, 2196–2207 (2002).
104. P.L. Martelli, P. Fariselli, A. Krogh, and R. Casadio, A sequence-profile-based Hmm for predicting and discriminating beta barrel membrane proteins, *Bioinformatics*, **18 Suppl 1**, S46–53 (2002).
105. I. Jacoboni, P.L. Martelli, P. Fariselli, V. De Pinto, and R. Casadio, Prediction of the transmembrane regions of beta-barrel membrane proteins with a neural network-based predictor, *Protein Sci*, **10**, 779–787 (2001).

106. P.G. Bagos, T.D. Liakopoulos, I.C. Spyropoulos, and S.J. Hamodrakas, Pred-Tmbb: A web server for predicting the topology of beta-barrel outer membrane proteins, *Nucleic Acids Res*, **32**, W400–404 (2004).
107. H.R. Bigelow, D.S. Petrey, J. Liu, D. Przybylski, and B. Rost, Predicting transmembrane beta-barrels in proteomes, *Nucleic Acids Res*, **32**, 2566–2577 (2004).
108. M.M. Gromiha, S. Ahmad, and M. Suwa, Neural network-based prediction of transmembrane beta-strand segments in outer membrane proteins, *J Comput Chem*, **25**, 762–767 (2004).
109. F.S. Berven, K. Flikka, H.B. Jensen, and I. Eidhammer, Bomp: A program to predict integral beta-barrel outer membrane proteins encoded within genomes of gram-negative bacteria, *Nucleic Acids Res*, **32**, W394–399 (2004).
110. M. Arai, H. Mitsuke, M. Ikeda, *et al*., Conpred Ii: A consensus prediction method for obtaining transmembrane topology models with high reliability, *Nucleic Acids Res*, **32**, W390–393 (2004).
111. M. Amico, M. Finelli, I. Rossi, *et al*., Pongo: A web server for multiple predictions of all-alpha transmembrane proteins, *Nucleic Acids Res*, **34**, W169–172 (2006).
112. H. Zhou, C. Zhang, S. Liu, and Y. Zhou, Web-based toolkits for topology prediction of transmembrane helical proteins, fold recognition, structure and binding scoring, folding-kinetics analysis and comparative analysis of domain combinations, *Nucleic Acids Res*, **33**, W193–197 (2005).
113. N. Bhardwaj, R.V. Stahelin, R.E. Langlois, W. Cho, and H. Lu, Structural bioinformatics prediction of membrane-binding proteins, *J Mol Biol*, **359**, 486–495 (2006).
114. P.G. Bagos, T.D. Liakopoulos, and S.J. Hamodrakas, Evaluation of methods for predicting the topology of beta-barrel outer membrane proteins and a consensus prediction method, *BMC Bioinformatics*, **6**, 7 (2005).
115. M. Pellegrini-Calace, A. Carotti, and D.T. Jones, Folding in lipid membranes (film): A novel method for the prediction of small membrane protein 3D structures, *Proteins*, **50**, 537–545 (2003).
116. V. Yarov-Yarovoy, J. Schonbrun, and D. Baker, Multipass membrane protein structure prediction using Rosetta, *Proteins*, **62**, 1010–1025 (2006).
117. P. Barth, J. Schonbrun, and D. Baker, Toward high-resolution prediction and design of transmembrane helical protein structures, *Proc Natl Acad Sci U S A*, **104**, 15682–15687 (2007).
118. X. de la Cruz, E.G. Hutchinson, A. Shepherd, and J.M. Thornton, Toward predicting protein topology: An approach to identifying beta hairpins, *Proc Natl Acad Sci U S A*, **99**, 11157–11162 (2002).
119. M. Kumar, M. Bhasin, N.K. Natt, and G.P. Raghava, Bhairpred: Prediction of beta-hairpins in a protein from multiple alignment information using Ann and Svm techniques, *Nucleic Acids Res*, **33**, W154–159 (2005).
120. A.N. Lupas, and M. Gruber, The structure of alpha-helical coiled coils, *Adv Protein Chem*, **70**, 37–78 (2005).
121. A. Lobley, M.B. Swindells, C.A. Orengo, and D.T. Jones, Inferring function using patterns of native disorder in proteins, *PLoS Comput Biol*, **3**, e162 (2007).
122. H. Xie, S. Vucetic, L.M. Iakoucheva, *et al*., Functional anthology of intrinsic disorder. 3. Ligands, post-translational modifications, and diseases associated with intrinsically disordered proteins, *J Proteome Res*, **6**, 1917–1932 (2007).
123. S. Vucetic, H. Xie, L.M. Iakoucheva, *et al*., Functional anthology of intrinsic disorder. 2. Cellular components, domains, technical terms, developmental processes, and coding sequence diversities correlated with long disordered regions, *J Proteome Res*, **6**, 1899–1916 (2007).
124. H. Xie, S. Vucetic, L.M. Iakoucheva, *et al*., Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions, *J Proteome Res*, **6**, 1882–1898 (2007).
125. V. Neduva, and R.B. Russell, Linear motifs: Evolutionary interaction switches, *FEBS Lett*, **579**, 3342–3345 (2005).
126. M. Fuxreiter, P. Tompa, and I. Simon, Local structural disorder imparts plasticity on linear motifs, *Bioinformatics*, **23**, 950–956 (2007).
127. Y. Zhang, B. Stec, and A. Godzik, Between order and disorder in protein structures: Analysis of 'dual personality' fragments in proteins, *Structure*, **15**, 1141–1147 (2007).

128. F. Ferron, S. Longhi, B. Canard, and D. Karlin, A practical overview of protein disorder prediction methods, *Proteins*, **65**, 1–14 (2006).
129. Z. Dosztanyi, M. Sandor, P. Tompa, and I. Simon, Prediction of protein disorder at the domain level, *Curr Protein Pept Sci*, **8**, 161–171 (2007).
130. M. Sickmeier, J.A. Hamilton, T. LeGall, *et al*., Disprot: The database of disordered proteins, *Nucleic Acids Res*, **35**, D786–793 (2007).
131. J.C. Wootton, Non-globular domains in protein sequences: Automated segmentation using complexity measures, *Comput.Chem*., **18**, 269–285 (1994).
132. V.N. Uversky, Natively unfolded proteins: A point where biology waits for physics, *Protein Sci*, **11**, 739–756 (2002).
133. R. Linding, L.J. Jensen, F. Diella, P. Bork, T.J. Gibson, and R.B. Russell, Protein disorder prediction: Implications for structural proteomics, *Structure*, **11**, 1453–1459 (2003).
134. J.J. Ward, J.S. Sodhi, L.J. McGuffin, B.F. Buxton, and D.T. Jones, Prediction and functional analysis of native disorder in proteins from the three kingdoms of life, *J Mol Biol*, **337**, 635–645 (2004).
135. J. Cheng, M. Sweredoski, and P. Baldi, Accurate prediction of protein disordered regions by mining protein structure data, *Data Mining and Knowledge Discovery*, **11**, 213–222 (2005).
136. S. Vucetic, C.J. Brown, A.K. Dunker, and Z. Obradovic, Flavors of protein disorder, *Proteins*, **52**, 573–584 (2003).
137. Z. Obradovic, K. Peng, S. Vucetic, P. Radivojac, and A.K. Dunker, Exploiting heterogeneous sequence properties improves prediction of protein disorder, *Proteins*, **61 Suppl 7**, 176–182 (2005).
138. C.T. Su, C.Y. Chen, and Y.Y. Ou, Protein disorder prediction by condensed Pssm considering propensity for order or disorder, *BMC Bioinformatics*, **7**, 319 (2006).
139. R.M. MacCallum. Order/disorder prediction with self organising maps. CASP 6 meeting, Online paper. http://www.forcasp.org/paper2127.html
140. J. Prilusky, C.E. Felder, T. Zeev-Ben-Mordehai, *et al*., Foldindex: A simple tool to predict whether a given protein sequence is intrinsically unfolded, *Bioinformatics*, **21**, 3435–3438 (2005).
141. O.V. Galzitskaya, S.O. Garbuzynskiy, and M.Y. Lobanov, Foldunfold: Web server for the prediction of disordered regions in protein chain, *Bioinformatics*, **22**, 2948–2949 (2006).
142. C.T. Su, C.Y. Chen, and C.M. Hsu, Ipda: Integrated protein disorder analyzer, *Nucleic Acids Res*, **35**, W465–472 (2007).
143. Z. Dosztanyi, V. Csizmok, P. Tompa, and I. Simon, Iupred: Web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content, *Bioinformatics*, **21**, 3433–3434 (2005).
144. J. Liu, and B. Rost, Norsp: Predictions of long regions without regular secondary structure, *Nucleic Acids Res*, **31**, 3833–3835 (2003).
145. P. Romero, Z. Obradovic, X. Li, E.C. Garner, C.J. Brown, and A.K. Dunker, Sequence complexity of disordered protein, *Proteins*, **42**, 38–48 (2001).
146. K. Shimizu, S. Hirose, and T. Noguchi, Poodle-S: Web application for predicting protein disorder by using physicochemical features and reduced amino acid set of a position-specific scoring matrix, *Bioinformatics*, **23**, 2337–2338 (2007).
147. S. Hirose, K. Shimizu, S. Kanai, Y. Kuroda, and T. Noguchi, Poodle-L: A two-level Svm prediction system for reliably predicting long disordered regions, *Bioinformatics*, **23**, 2046–2053 (2007).
148. T. Ishida, and K. Kinoshita, Prdos: Prediction of disordered protein regions from amino acid sequence, *Nucleic Acids Res*, **35**, W460–464 (2007).
149. K. Coeytaux, and A. Poupon, Prediction of unfolded segments in a protein sequence based on amino acid composition, *Bioinformatics*, **21**, 1891–1900 (2005).
150. Z.R. Yang, R. Thomson, P. McNeil, and R.M. Esnouf, Ronn: The bio-basis function neural network technique applied to the detection of natively disordered regions in proteins, *Bioinformatics*, **21**, 3369–3376 (2005).

151. A. Vullo, O. Bortolami, G. Pollastri, and S.C. Tosatto, Spritz: A server for the prediction of intrinsically disordered regions in protein sequences using kernel machines, *Nucleic Acids Res*, **34**, W164–168 (2006).
152. N.B. Holladay, L.N. Kinch, and N.V. Grishin, Optimization of linear disorder predictors yields tight association between crystallographic disorder and hydrophobicity, *Protein Sci*, **16**, 2140–2152 (2007).
153. L. Bordoli, F. Kiefer, and T. Schwede, Assessment of disorder predictions in Casp7, *Proteins*, **69 Suppl 8**, 129–136 (2007).
154. J. Skolnick, A. Kolinski, and A.R. Ortiz, Monsster: A method for folding globular proteins with a small number of distance restraints, *J Mol Biol*, **265**, 217–241 (1997).
155. M.J. Pietal, I. Tuszynska, and J.M. Bujnicki, Protmap2d: Visualization, comparison, and analysis of 2D maps of protein structure, *Bioinformatics*, (2007).
156. J.M. Izarzugaza, O. Grana, M.L. Tress, A. Valencia, and N.D. Clarke, Assessment of intramolecular contact predictions for Casp7, *Proteins*, **69 Suppl 8**, 152–158 (2007).
157. G. Shackelford, and K. Karplus, Contact prediction using mutual information and neural nets, *Proteins*, **69 Suppl 8**, 159–164 (2007).
158. C.H. Tsai, C.H. Chan, B.J. Chen, C.Y. Kao, H.L. Liu, and J.P. Hsu, Bioinformatics approaches for disulfide connectivity prediction, *Curr Protein Pept Sci*, **8**, 243–260 (2007).
159. R.M. MacCallum, Striped Sheets and Protein Contact Prediction, *Bioinformatics*, **20 Suppl 1**, i224–231 (2004).
160. M. Punta, and B. Rost, Profcon: Novel prediction of long-range contacts, *Bioinformatics*, **21**, 2960–2968 (2005).
161. G. Pollastri, P. Baldi, P. Fariselli, and R. Casadio, Prediction of coordination number and relative solvent accessibility in proteins, *Proteins*, **47**, 142–153 (2002).
162. J. Cheng, and P. Baldi, Improved residue contact prediction using support vector machines and a large feature set, *BMC Bioinformatics*, **8**, 113 (2007).
163. N. Hamilton, K. Burrage, M.A. Ragan, and T. Huber, Protein contact prediction using patterns of correlation, *Proteins*, **56**, 679–684 (2004).
164. V. Sobolev, E. Eyal, S. Gerzon, *et al*. Space: A suite of tools for protein structure prediction and analysis based on complementarity and environment, *Nucleic Acids Res*, **33**, W39–43 (2005).
165. Y. Shao, and C. Bystroff, Predicting interresidue contacts using templates and pathways, *Proteins*, **53 Suppl 6**, 497–502 (2003).
166. J. Cheng, and P. Baldi, Three-stage prediction of protein beta-sheets by neural networks, alignments and graph algorithms, *Bioinformatics*, **21 Suppl 1**, i75–84 (2005).
167. F. Ferre, and P. Clote, Dianna 1.1: An extension of the Dianna web server for ternary cysteine classification, *Nucleic Acids Res*, **34**, W182–185 (2006).
168. J. Cheng, H. Saigo, and P. Baldi, Large-scale prediction of disulphide bridges using kernel methods, two-dimensional recursive neural networks, and weighted graph matching, *Proteins*, **62**, 617–629 (2006).
169. A. Ceroni, A. Passerini, A. Vullo, and P. Frasconi, Disulfind: A disulfide bonding state and cysteine connectivity prediction server, *Nucleic Acids Res*, **34**, W177–181 (2006).
170. B.D. O'Connor, and T.O. Yeates, Gdap: A web tool for genome-wide protein disulfide bond prediction, *Nucleic Acids Res*, **32**, W360–364 (2004).
171. P. Fariselli, P. Martelli, and R. Casadio, A neural network based method for predicting the disulfide connectivity in proteins, in *Knowledge Based Intelligent Information Engineering Systems and Allied Technologies*, E. Damiani (ed.), International Conference on Knowledge-Based Intelligent Engineering (Kes 2002), IOS Press, Amsterdam, 2002.
172. J. Song, Z. Yuan, H. Tan, T. Huber, and K. Burrage, Predicting disulfide connectivity from protein sequence using multiple sequence feature vectors and secondary structure, *Bioinformatics*, **23**, 3147–3154 (2007).
173. P. Aloy, A. Stark, C. Hadley, and R.B. Russell, Predictions without templates: New folds, secondary structure, and contacts in Casp5, *Proteins*, **53 Suppl 6**, 436–456 (2003).
174. I.Y. Koh, V.A. Eyrich, M.A. Marti-Renom, *et al*. Eva: Evaluation of protein structure prediction servers, *Nucleic Acids Res*, **31**, 3311–3315 (2003).