

Computational prediction of secondary and supersecondary structures

Ke Chen and Lukasz Kurgan

Department of Electrical and Computer Engineering, University of Alberta, Edmonton, CANADA

Summary

The sequence-based prediction of the secondary and supersecondary structures enjoys strong interest and finds applications in numerous areas related to the characterization and prediction of protein structure and function. Substantial efforts in these areas over the last three decades resulted in the development of accurate predictors, which take advantage of modern machine learning models and availability of evolutionary information extracted from multiple sequence alignment. In this chapter, we first introduce and motivate both prediction areas and introduce basic concepts related to the annotation and prediction of the secondary and supersecondary structures, focusing on the β hairpin, coiled coil, and α -turn- α motifs. Next, we overview state-of-the-art prediction methods, and we provide details for twelve modern secondary structure predictors and four representative supersecondary structure predictors. Finally, we provide several practical notes for the users of these prediction tools.

Key words: secondary structure prediction; supersecondary structure prediction; beta-hairpins; coiled coils; helix-turn-helix; Greek key; multiple sequence alignment.

1. Introduction

Protein structure is defined at three levels: *primary structure* which is the sequence of amino acids joined by peptide bonds; *secondary structure* that concerns regular local sub-structures including α -helices and β -strands, which were first postulated by Pauling and coworkers (1, 2); and *tertiary structure* which is the three-dimensional structure of a protein molecule. The supersecondary structure (SSS) bridges the two latter levels and concerns specific combinations / geometric arrangements of a few secondary structure elements. Common supersecondary structures include α -helix hairpins, β hairpins, coiled coils, Greek key, and β - α - β , α -turn- α , α -loop- α , and Rossmann motifs. The secondary and SSS elements are combined together, with help of various types of coils, to form the tertiary structure. An example that displays the secondary structures and the β hairpin supersecondary structure is given in Figure 1.

In early 1970s Anfinsen demonstrated that the native tertiary structure is encoded in the primary structure (3) and this observation fueled the development of methods that predict the structure from the sequence. The need for these predictors is motivated by the fact that the tertiary structure is known for a relatively small number of proteins, i.e., as of mid 2011 about 70 thousand protein structures are deposited in the Protein Data Bank (PDB) (4) when

compared with 12.5 million nonredundant protein sequences in the RefSeq database (5), and the fact that experimental determination of protein structure is relatively expensive and time-consuming and cannot keep up with the rapid accumulation of the sequence data (6-9). One of the successful ways to predict the tertiary structure is to proceed in a step-wise fashion. First, we predict how the sequence folds into the secondary structure, then how these secondary structure elements come together to form SSSs, and finally the information about the secondary and supersecondary structures is used to help in computational determination of the full three-dimensional molecule (10-15).

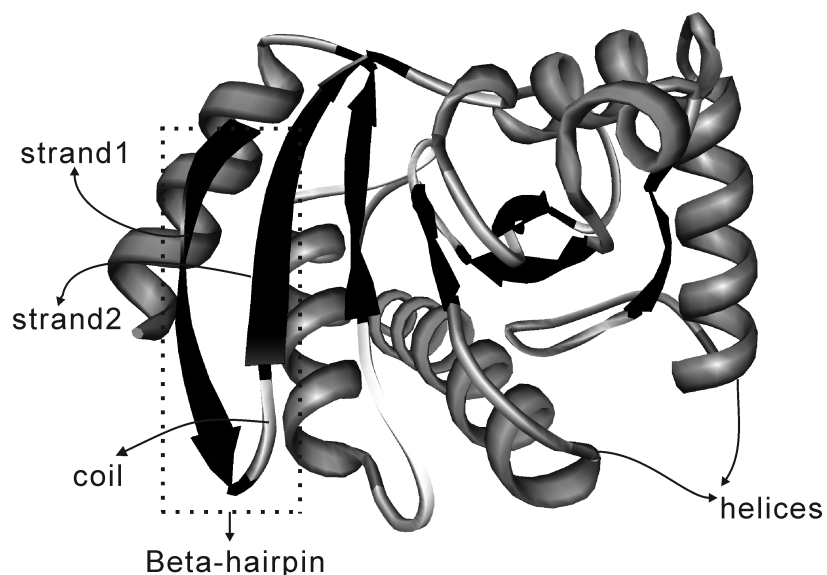


Figure 1. Cartoon representation of the tertiary structure of chain A of AF1521 protein (PDB code: 2BFR). The α -helices are shown in dark gray, β -strands in black, and coils in light gray. The β hairpin supersecondary structure motif, which consists of strand1, strand2, and the coil between the two strands, is denoted using the dotted rectangle.

Last three decades observed strong progress in the development of accurate predictors of the secondary structure, which currently provide predictions with about 82% accuracy (16). Besides being useful for the prediction of the tertiary structure, the secondary structure predicted from the sequence is widely applied for the analysis and prediction of numerous structural and functional characteristics of proteins. These characteristics include multiple alignment (17), prediction of protein-ligand interactions (18-20), prediction of residue depth (21, 22), structural classes and folds (23-25), residue contacts (26, 27), disorder (28-30), folding rates and types (31-33), and target selection for structural genomics (34, 35), to name just a few. The secondary structure predictors enjoy strong interest, which could be quantified by the massive workloads that they handle. For instance, the web server of the one of the most popular methods, PSIPRED, was reported in 2005 to receive over 15,000 requests per month (36). Another indicator is the fact that many of these methods receive high citations counts. A recent review (37) reported that seven methods were cited over 100 times and two of them, PSIPRED (36, 38, 39) and PHD (40, 41) were cited over 1300 times.

The prediction of the SSS includes methods specialized for specific types of these structures, including β hairpins, coiled coils, and helix–turn–helix motifs. The first methods were developed in 1980s and to date about twenty predictors were developed. Similarly as the secondary structure predictors, the predictors of SSS found applications in numerous areas including analysis of amyloids (42, 43), microbial pathogens (44), and synthases (45), simulation of protein folding (46), analysis of relation between coiled coils and disorder (47), genome-wide studies of protein structure (48, 49), and prediction of protein domains (50). One interesting aspect is that the prediction of the secondary structure should provide useful information for the prediction of SSS. Two examples that exploit this relation are a prediction method by the Thornton's group (51) and the BhairPred method (52), both of which predict the β hairpins.

The secondary structure prediction field was reviewed a number of times. The earlier reviews summarized the most important advancements in this field, which were related to the use of sliding window, evolutionary information extracted from multiple sequence alignment, and machine-learning classifiers (53-55), and more recently due to the utilization of consensus-based approaches (56). More recent reviews concentrate on the evaluations and applications of the secondary structure predictors and provide practical advice for the users, such as the information concerning availability (16, 57, 58). The SSS prediction area was reviewed less extensively. The β hairpin and coiled coil predictors, as well as the secondary structure predictors were overviewed in 2006 (59) and a comparative analysis of the coiled coil predictors was presented in the same year (60). In this chapter, we summarize a more comprehensive set of recent secondary structure and SSS predictors. We also demonstrate how the prediction of the secondary structure is used to implement a SSS predictor and provide several practical notes for the users.

2. Materials

2.1. Assignment of secondary structure

The secondary structure, which is assigned from the tertiary structure, is used for a variety of applications, including visualization (61-63) and classification of the protein folds (64-67), and as a ground truth to develop and evaluate the secondary and SSS predictors. Several annotation protocols were developed over the last few decades. The first implementation was done in late 1970s by Levitt and Greer (68). This was followed by Kabsch and Sander who developed a method called Dictionary of Protein Secondary Structure (DSSP) (69), which is based on the detection of hydrogen bonds defined by an electrostatic criterion. Other, more recent, assignment methods include DEFINE (70), P-CURVE (71), STRIDE (72), P-SEA (73), XTLSSTR (74), SECSTR (75), KAKSI (76), Segno (77), PALSSE (78), SKSP (79), PROSIGN (80), and SABA (81). Moreover, the 2Struc web server provides an integrated access to multiple annotation methods, which enables convenient comparison between different assignment protocols (82).

The DSSP remains to be the most widely-used protocol (76), which is likely due to the fact that it is used to annotate depositions in the PDB and since it was used to evaluate secondary structure predictions in the two largest community based assessments: the Critical Assessment of techniques for protein Structure Prediction (CASP) (83) and the Evaluation of Automatic protein structure prediction (EVA) continuous benchmarking project (84). DSSP determines the secondary structures based on the patterns of hydrogen bonds, which are categorized into three major states: helices, sheets, and regions with irregular secondary structure. This method assigns one of the following eight secondary structure states for each of the structured residues (residues that have three-dimensional coordinates) in the protein sequence:

- G: (3-turn) 3_{10} helix, where the carboxyl group of a given amino acid forms a hydrogen bond with amid group of the residue three positions down in the sequence forming a tight, right-handed helical structure with 3 residues per turn.
- H: (4-turn) α -helix, which is similar to the 3-turn helix, except that the hydrogen bonds are formed between consecutive residues that are 4 positions away.
- I: (5-turn) π -helix, where the hydrogen bonding occurs between residues spaced 5 positions away. Most of the π -helices are right-handed.
- E: extended strand, where 2 or more strands are connected laterally by at least two hydrogen bonds forming a pleated sheet.
- B: an isolated beta-bridge, which is a single residue pair sheet formed based on the hydrogen bond.
- T: hydrogen bonded turn, which is a turn where a single hydrogen bond is formed between residues spaced 3, 4, or 5 positions away in the protein chain.
- S: bend, which corresponds to a fragment of protein sequence where the angle between the vector from C^{α}_i to C^{α}_{i+2} (C^{α} atoms at the i^{th} and $i+2^{\text{th}}$ positions in the chain) and the vector from C^{α}_{i-2} to C^{α}_i is below 70° . The bend is the only non-hydrogen bond-based regular secondary structure type.
- –: irregular secondary structure (also referred to as loop and random coil), which includes the remaining conformations.

These eight secondary structure states are often mapped into the following three states (see Figure 1):

- H: α -helix, which corresponds to the right or left handed cylindrical/helical conformations that include G, H, and I states.
- E: β -strand, which corresponds to pleated sheet structures that encompass E and B states.
- C: coil, which covers the remaining S, T, and – states.

The DSSP program is freely available from <http://swift.cmbi.ru.nl/gv/dssp/>.

2.2. Assignment of supersecondary structures

The SSS is composed of several adjacent secondary structure elements. Therefore, the assignment of the SSS relies on the assignment of the secondary structure. Among more than a dozen types of the SSSs, the β hairpins, coiled coils, and α -turn- α motifs received more attention due to the fact that they are present in a large number of protein structures and they have pivotal roles in the biological functions of proteins. The β hairpin motif comprises the second largest group of protein domain structures and is found in diverse protein families,

including enzymes, transporter proteins, antibodies, and in viral coats (52). The coiled coil motifs mediate the oligomerization of a large number of proteins and are involved in regulation of gene expression, e.g., transcription factors (85). The α -turn- α (helix-turn-helix) motif is instrumental for DNA binding, i.e., majority of the DNA-binding proteins interact with DNA through this motif (86). The β hairpins, coiled coils, and α -turn- α motifs are defined as follows:

- β hairpin motif contains two strands that are adjacent in the primary structure, oriented in an antiparallel arrangement, and linked by a short loop;
- coiled coil is build by two or more α -helices that wind around each other to form a supercoil.
- α -turn- α motif is composed of two α -helices joined by a short turn structure.

The β hairpin motifs are commonly annotated by PROMOTOF program (87), which also assigns several other SSS types, e.g., psi-loop and β - α - β motifs. Similar to DSSP, the PROMOTOF program assigns SSS based on the distances and hydrogen bonding between the residues. The coiled coils are usually assigned with the SOCKET program (88), which locates/annotates coiled-coil interactions based on the distances between multiple helical chains. The DNA-binding α -turn- α motifs are usually manually extracted from the DNA-binding proteins, since these motifs that do not interact with DNA are of lesser interest.

For users convenience, certain supersecondary structures, such as the coiled coils and β - α - β motifs, can be accessed, analyzed, and visualized using specialized repositories such as CCPLUS (89) and TOPS (90). CCPLUS archives coiled coil structures identified by SOCKET for all structures in PDB. The TOPS database stores topological descriptions of protein structures, including the secondary structure and the chiralities of selected SSSs, e.g., β hairpins and β - α - β motifs.

2.3. Multiple sequence alignment

Multiple sequence alignment profile was introduced into the pipelines for the prediction of the secondary structure in early 1990s (91). Using the multiple sequence alignment profile rather than the primary sequence has led to a large improvement by 10% accuracy in the secondary structure prediction (91). The alignment profile is also often used in the prediction of the supersecondary structure (52, 59, 60). The multiple sequence alignment profile is generated from a given protein sequence in two steps. In the first step, sequences that are similar to the given input sequence are identified from a large sequence database, such as the *nr* (non-redundant) database provided by the National Center for Biotechnology Information (NCBI). In the second step, multiple sequence alignment is performed between the input sequence and its similar sequences and the profile is generated. An example of the multiple sequence alignment is given in Figure 2 where eight similar sequences are identified for the input protein (we use the protein from Figure 1). Each position of the input (query) sequence is represented by the frequencies of amino acid derived from the multiple sequence alignment to derive the profile. For instance, for the boxed position in Figure 2, the counts of amino acids Tyr (Y), Ala (A) and Gly (G) are 5, 2, and 2, respectively. Therefore, this position is represented by a 20-dimensional vector (2/9, 0, 0, 0, 0, 2/9, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0).

5/9), where each value indicates the fraction of the corresponding amino acid type (amino acids are sorted in alphabetical order) in multiple sequence alignment at this position. The profile is composed of these 20-dimensional vectors for each position in the input protein chain.

Query protein	...	K R L E H G G G V A	Y	A I A K A C A G D A G L	...
YP_002995377	...	K Y L E H G G G V A	Y	A I A K A A S G D V R E	...
YP_002958591	...	K Y L E H G G G V A	Y	A I A K A A A G N V A E	...
YP_003418650	...	S Y L Q H G G G V A	Y	A I V K K G G - - - - -	...
YP_002828572	...	S Y L Q H G G G V A	Y	A I V K K G G - - - - -	...
ZP_04861702	...	G M L K H V G G V A	A	A I V K K G G - - - - -	...
ZP_05391340	...	G A L K H G G G A A	A	A I V K A G G - - - - -	...
YP_003345806	...	E Y L K H G G G V A	G	A I V R A G G - - - - -	...
YP_003496764	...	S H L K M G G G V A	G	A I R R A G G - - - - -	...

Figure 2. Multiple sequence alignment between the input (query) sequence, which is a fragment of chain A of the AF1521 protein shown in Figure 1, and similar sequences identified in the *nr* database. The first row shows the query chain and the subsequent rows show the eight aligned proteins. Each row contains the protein sequence ID (the first column) and the corresponding amino acid sequence (the third and subsequent columns), where “...” denotes continuation of the chain and “-” denotes a gap, which means that this part of the sequence could not be aligned. The boxed column is used as an example to discuss generation of the multiple sequence alignment profile in section 2.3.

The PSI-BLAST (Position-Specific Iterated BLAST) (92) algorithm was developed for the identification of distant similarity to a given input sequence. First, a list of closely related protein sequences is identified from a sequence database, such as the *nr* database. These sequences are combined into a general "profile", which summarizes significant features present in these sequences. Another query against the sequence database is run using this "profile", and a larger group of sequences is found. This larger group of sequences is used to construct another "profile", and the process is repeated. PSI-BLAST is more sensitive in picking up distant evolutionary relationships than a standard protein-protein BLAST that does not perform iterative repetitions. Since late 1990s, the PSI-BLAST is commonly used for the generation of multiple sequence alignment profile, which is named position-specific scoring matrix (PSSM) and which is often utilized in the prediction of secondary and supersecondary structures. An example PSSM profile is given in Figure 3. The BLAST and PSI-BLAST programs are available at <http://blast.ncbi.nlm.nih.gov/>.

		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	...
1	K	-1	2	0	-1	-3	1	1	-2	-1	-3	-3	5	-2	-3	-1	0	-1	-3	-2	-3	...
2	R	-2	2	-2	-3	-3	-1	-2	-3	1	-2	-1	0	-1	2	-3	-2	-2	2	7	-2	...
3	L	-2	-2	-4	-4	-1	-2	-3	-4	-3	2	4	-3	2	0	-3	-3	-1	-2	-1	1	...
4	E	-1	0	0	2	-4	2	5	-2	0	-4	-3	1	-2	-4	-1	0	-1	-3	-2	-3	...
5	H	-2	0	1	-1	-3	0	0	-2	8	-4	-3	-1	-2	-1	-2	-1	-2	-3	2	-3	...
6	G	0	-3	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-3	-3	-4	...
7	G	0	-3	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-3	-3	-4	...
8	G	0	-3	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-3	-3	-4	...
9	V	0	-3	-3	-4	-1	-2	-3	-4	-3	3	1	-3	1	-1	-3	-2	0	-3	-1	4	...
10	A	4	-2	-2	-2	0	-1	-1	0	-2	-1	-2	-1	-1	-2	-1	1	0	-3	-2	0	...
11	Y	-2	-2	-3	-4	-2	-2	-2	-4	1	0	1	-2	0	3	-3	-2	-2	2	6	-1	...
12	A	4	-2	-2	-2	0	-1	-1	0	-2	-1	-2	-1	-1	-2	-1	1	0	-3	-2	0	...
13	I	-1	-3	-4	-3	-1	-3	-4	-4	-4	5	2	-3	1	0	-3	-3	-1	-3	-1	3	...
14	A	4	-2	-2	-2	0	-1	-1	0	-2	-1	-2	-1	-1	-2	-1	1	0	-3	-2	0	...
15	K	-1	2	0	-1	-3	1	1	-2	-1	-3	-3	5	-2	-3	-1	0	-1	-3	-2	-3	...
16	A	4	-2	-2	-2	0	-1	-1	0	-2	-1	-2	-1	-1	-2	-1	1	0	-3	-2	0	...
17	C	2	-3	-3	-3	9	-2	-3	-2	-3	-1	-1	-2	-1	-3	-2	0	-1	-3	-2	-1	...
18	A	4	-1	-1	-1	-1	-1	-1	0	-2	-2	-2	-1	-1	-3	-1	2	0	-3	-2	-1	...
19	G	0	-3	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-3	-3	-4	...
20	D	-2	-2	1	6	-4	0	2	-1	-1	-3	-4	-1	-3	-4	-2	0	-1	-5	-3	-4	...
21	A	3	-2	-2	-2	-1	-1	-1	-1	-2	0	-1	-1	-1	-2	4	0	0	-3	-2	1	...
22	G	0	3	0	-1	-3	3	0	2	-1	-3	-3	1	-2	-3	-2	1	0	-3	-2	-3	...
23	L	-1	0	-1	0	-3	1	4	-2	-1	-1	1	2	0	-2	-2	-1	-1	-3	-2	-1	...

Figure 3. Position-specific scoring matrix generated by PSI-BLAST for the input (query) sequence, which is a fragment of chain A of the AF1521 protein shown in Figure 1. The first and second columns are the residue number and type, respectively, in the input protein chain. The subsequent columns provide values of the multiple sequence alignment profile for a substitution to an amino acid type indicated in the first row. Initially, a matrix $\{p_{ij}\}$, where p_{ij} indicates the probability that the j^{th} amino acid type (in columns) occurs at i^{th} position in the input chain (in rows), is generated. The position-specific scoring matrix $\{m_{ij}\}$ is defined as $m_{ij} = \log(p_{ij} / b_j)$, where b_j is the background frequency of the j^{th} amino acid type.

3. Methods

3.1. Current secondary structure prediction methods

The prediction of the secondary structure is defined as mapping of each amino acid in the primary structure to one of the three (or eight) secondary structure states, most often as defined by the DSSP. Virtually all recent secondary structure predictors use a sliding window approach in which a local stretch of residues around a central position in the window is

utilized to predict the secondary structure state at the central position. Moreover, as one of the first steps in the prediction protocol, the state-of-the-art methods use PSI-BLAST to generate multiple alignment and/or PSSM that, with the help of the sliding window, are used to encode the input sequence. The early predictors were implemented based on a relatively simple statistical analysis of composition of the input sequence. The modern methods adopt sophisticated machine learning-based classifiers to represent the relation between the input sequence (or more precisely between the evolutionary information generated with PSI-BLAST) and the secondary structure states. In majority of cases, the classifiers are implemented using neural networks. However, different predictors use different numbers of networks (between one and hundreds), different types of networks (e.g., feed-forward and recurrent), and different sizes of the sliding window. These prediction methods are provided to the end users as standalone applications and/or as web servers. The standalone programs are suitable for higher volume (for a large number of proteins) predictions and they can be incorporated in other predictive pipelines, but they require installation by the user on a local computer. The web servers are more convenient since they can be run using a web browser and without the need for the local installation, but they are more difficult to use when applied to predict a large set of chains, i.e., some servers allow submission of one chain at the time and may have long wait times due to limited computational resources and a long queue of requests from other users. Moreover, recent comparative survey (16) shows that the differences in the predictive quality for a given predictor between its standalone and web server versions depend on the frequency with which the underlying databases, which are used to calculate the evolutionary information and to perform homology modeling, are updated. Sometimes these updates are more frequent for the web server, and in other cases for the standalone package.

Table 1. Summary of the recent sequence-based predictors of secondary structure. The “year last published” column provides the year of the publication of the most recent version of a given method. The “availability” column identifies whether a standalone program (SP) and/or a web server (WS) is available. The methods are sorted by the year of their last publication in the descending order.

Name	Year last published	Prediction model	Availability
PSIPRED	2010	Neural network	WS+SP
SPINE	2009	Neural network	WS+SP
Frag1D	2009	Scoring function	SP
DISSPred	2009	Support vector machine + clustering	WS
SAM-T	2009	Neural network	WS+SP
PROTEUS	2008	Neural network	WS+SP
Jpred	2008	Neural network	WS
P.S.HMM	2007	Neural network + hidden Markov model	WS
Porter	2007	Neural network	WS+SP
OSS-HMM	2006	Hidden Markov model	SP
YASSPP	2006	Support vector machine	WS
YASPIN	2005	Neural network + hidden Markov model	WS
SABLE	2005	Neural network	WS+SP
SSpro	2005	Neural network	WS+SP

Table 1 summarizes 15 methods, including PSIPRED (36, 38, 39), SPINE (93, 94), Frag1D (95), DISSPred (96), SAM-T (97-101), PROTEUS (102, 103), Jpred (104-106), P.S.HMM (107), Porter (108, 109), OS-HMM (110), YASSPP (111), YASPIN (112), SABLE (113), and SSpro (114, 115), that predict the 3-state secondary structure and which were published since 2005 inclusive. Older methods were reviewed in (53-55). We note that only a few methods, including SSpro8 (115) and SAM-T08 (101), predict the 8-state secondary structure. Following, we discuss in greater detail the methods that offer web servers, as arguably these are used by a larger number of users. We summarize their architecture, provide location of their implementation, and briefly discuss their predictive performance. We note that the predictive quality should be considered with a grain of salt since different methods were evaluated on different datasets and using different test protocols (see Note 1). However, we primarily utilize fairly consistent results that were published in two recent comparative studies (see Note 3) (16, 37). Moreover, recent research shows that improved predictive performance could be obtained by post-processing of the secondary structure predictions (see Note 4) (116).

3.2. PSIPRED

PSIPRED is one of the most popular prediction methods (see Note 2); e.g., it received the largest number of citations as shown in (16, 37). This method was developed in late 1990s by Jones group at the University College London (38), and later improved and updated, with the most recent version 3.0 (39). PSIPRED is characterized by a relatively simple design which utilizes just two neural networks. This method was ranked as top predictor in the CASP3 and CASP4 competitions, and was recently evaluated to provide 3-state secondary structure predictions with 81% accuracy (16, 39). The current version bundles the secondary structure predictions with the prediction of transmembrane topology and fold recognition.

Inputs: PSSM generated from the input protein sequence using PSI-BLAST

Architecture: ensemble of two neural networks

Availability: <http://bioinf.cs.ucl.ac.uk/psipred/>

3.3. Jpred

Jpred was developed in late 1990s by Barton group at the University of Dundee (105). This method was updated a few times, with the most recent version Jpred 3 (104, 106). Similarly as PSIPRED, Jpred was demonstrated to provide about 81% accuracy for the 3-state secondary structure prediction (104). The web server implementation of Jpred couples the secondary structure predictions with the prediction of solvent accessibility and prediction of coiled coils using COILS algorithm (117).

Inputs: hidden Markov model profiles and PSSM generated from the input protein sequence using HMMer (118) and PSI-BLAST, respectively

Architecture: ensemble of neural networks

Availability: <http://www.compbio.dundee.ac.uk/www-jpred/>

3.4. SSpro

SSpro was introduced in early 2000 by the Baldi group at the University of California, Irvine (115). The latest version 4.5 (114) utilizes homology modeling, which is based on alignment to known tertiary structures from PDB, and achieves over 82% accuracy (16). The SSpro 4.0 was also ranked as one of the top secondary structure prediction servers in the EVA benchmark (119). SSpro is part of a comprehensive prediction center called SCRATCH, which also includes predictions of secondary structure in 8-states using SSpro8 (115), and prediction of solvent accessibility, disorder, contact numbers and contact maps, domains, disulfide bonds, B-cell epitopes, solubility upon overexpression, antigenicity, and tertiary structure.

Inputs: sequence profiles generated from the input protein sequence using PSI-BLAST

Architecture: ensemble of recurrent neural networks

Availability: <http://scratch.proteomics.ics.uci.edu/>

3.5. SAM-T

SAM-T is a family of methods which are under development since late 1990s by Karplus lab at the University of California at Santa Cruz. They include SAM-T98 (97), SAM-T99 (98), SAM-T02 (99), SAM-T04 (100), and SAM-T08 (101). The server outputs secondary structure prediction using multiple annotation protocols, including the 3- and 8-state DSSP. It also offers a number of other predictions (the predicted secondary structure is used as an input to calculate some of these predictions) including the tertiary structure, solvent accessibility, residue-residue contacts, multiple sequence alignments of putative homologs, and lists and alignment to potential templates with known structure.

Inputs: multiple alignment generated from the input protein sequence using PSI-BLAST

Architecture: neural network

Availability: http://compbio.soe.ucsc.edu/SAM_T08/T08-query.html

3.6. SABLE

The SABLE predictor was developed by Meller group at the University of Cincinnati (113). The web server that implements this method was used close to 200,000 times since it became operational in 2003. Two recent comparative studies (16, 39) and prior evaluations within the framework of the EVA initiative show that SABLE achieves accuracy of about 78%. The web server of the current version 2 also includes prediction of solvent accessibility and transmembrane domains.

Inputs: PSSM generated from the input protein sequence using PSI-BLAST

Architecture: ensemble of recurrent neural networks

Availability: <http://sable.cchmc.org/>

3.7. YASPIN

The YASPIN method was developed by Heringa lab at the Vrije Universiteit in 2004 (112). This is a hybrid method that utilizes a neural network and a hidden Markov model. One of the key characteristics of this method is that, as shown by the authors, it provides accurate predictions of β -strands (112). The predictive performance of YASPIN was evaluated using EVA benchmark and more recently in two comparative assessments (16, 39), which show that this method provides predictions with accuracy in the 76 to 79% range.

Inputs: PSSM generated from the input protein sequence using PSI-BLAST

Architecture: Two-level hybrid design with neural network in the 1st level and hidden Markov model in the 2nd level

Availability: <http://www.ibi.vu.nl/programs/yaspinwww/>

3.8. PORTER

This predictor was developed by Pollastri group at the University College Dublin (109). The web server that implements PORTER was utilized over 170,000 times since 2004 when it was released. This predictor was upgraded in 2007 to include homology modeling (108). The original and the homology-enhanced versions were recently shown to provide 79% (16) and 83% accuracy (37), respectively. PORTER is a part of a comprehensive predictive platform called DISTILL (120), which also incorporates predictors of relative solvent accessibility, residue-residue contact density, contacts maps, subcellular localization, and tertiary structure.

Inputs: PSSM generated from the input protein sequence using PSI-BLAST

Architecture: ensemble of recurrent neural networks

Availability: <http://distill.ucd.ie/porter/>

3.9. YASSPP

YASSPP was designed by Karypis lab at the University of Minnesota in 2005 (111). This is one of the few modern predictors that do not utilize neural network classifiers, but instead it uses multiple support vector machine learners. This method was shown to provide similar predictive quality to PSIPRED (111). The YASSPP predictor is bundled with several other predictors for transmembrane helices, disorder, solvent accessibility, contact order, and DNA-binding and ligand-binding residues in the MONSTER server at <http://bio.dtc.umn.edu/monster/>.

Inputs: PSSM generated from the input protein sequence using PSI-BLAST

Architecture: ensemble of six support vector machines

Availability: <http://glaros.dtc.umn.edu/yasspp/>

3.10. PROTEUS

This secondary structure prediction approach was developed by Wishart group at the University of Alberta around 2005 (103). PROTEUS is a consensus-based method in which

outputs of three secondary structure predictors, namely PSIPRED (37), Jnet (106), and an in-house TRANSSEC (103), are fed into a neural network. The predictions from the neural network are combined with the results based on homology modeling to generate the final output. PROTEUS is characterized by accuracy of about 81%, which was shown both the authors (102) and in a recent comparative survey (16). This predictor was incorporated into an integrated system called PROTEUS2, which additionally offers prediction of signal peptides, transmembrane helices and strands, and tertiary structure (102).

Inputs: multiple alignment generated from the input protein sequence using PSI-BLAST

Architecture: neural network that utilizes consensus of three secondary structure predictors

Availability: <http://wks16338.biology.ualberta.ca/proteus2/>

3.11. SPINE

The SPINE method originated at the Zhou group at the Indiana University–Purdue University in mid 2000s. The initial implementation (94), which was completed at the SUNY at Buffalo, was recently upgraded to create SPINE X (93). This predictor is characterized by relatively strong predictive performance with accuracy at about 81% (16). An important feature of this method is that it also provides predictions of backbone torsion angles, which give more detailed insights into the conformation of the backbone when compared with the secondary structure. The web server of SPINE X also provides predictions of solvent accessibility (93) and fluctuations of the torsion angles (121).

Inputs: PSSM generated from the input protein sequence using PSI-BLAST

Architecture: ensemble of neural networks

Availability: <http://sparks.informatics.iupui.edu/SPINE-X/>

3.12. P.S.HMM

The P.S.HMM predictor was developed at the University of Copenhagen and University of Southampton (107). Similar to YASPIN, this is a hybrid of a neural network and a hidden Markov model. The P.S.HMM method uses the hidden Markov model to produce initial predictions that are refined with help of a small neural network, while YASPIN performs predictions in the reverse order. The unique characteristic of this method is the fact that the hidden Markov model was designed utilizing genetic algorithms. This predictor provides outputs with 69% accuracy, as recently evaluated in (16), which is consistent with results presented by the authors (107).

Inputs: sequence profiles generated from the input protein sequence using PSI-BLAST

Architecture: Two-level hybrid design with hidden Markov model in the 1st level and neural network in the 2nd level

Availability: <http://wonk.med.upenn.edu/>

3.13. DISSPred

The DISSPred approach was recently introduced by Hirst group at the University of

Nottingham (96). Similar to SPINE, this method predicts both the 3-state secondary structure and the backbone torsion angles. The unique characteristic of DISSPred is that the predictions are cross-linked as inputs, i.e., predicted secondary structure is used to predict torsion angles and vice versa. The author estimated the accuracy of this method to be at 80% (96).

Inputs: PSSM generated from the input protein sequence using PSI-BLAST

Architecture: ensemble of support vector machines and clustering

Availability: <http://comp.chem.nottingham.ac.uk/disspred/>

3.14. Supersecondary structure prediction methods

Since SSS predictors are designed for a specific type of the supersecondary structures, e.g., SpiriCoil only predicts the coiled coils (48), the prediction of the SSS is defined as the assignment of each residue in the primary structure to two states: a state indicating the formation of a certain SSS type and another state indicating any other conformation. Similar to the prediction of the secondary structure, majority of the recent SSS predictors use a sliding window approach in which a local stretch of residues around a central position in the window is utilized to predict the SSS state at the central position. However, the architectures of the methods that were proposed for the prediction of different types of SSSs vary more substantially when compared with the fairly uniform architectures of the modern secondary structure predictors.

One of the early attempts for the prediction of β hairpin utilized the predicted secondary structure and similarity score between the predicted sequence and a library of β hairpin structures (51). More recent β hairpin predictors use the predicted secondary structure and some sequence-based descriptors to represent the predicted sequence (52, 122-126). Moreover, several types of prediction algorithms, including neural networks, support vector machines, quadratic discriminants, and random forests, were used for the prediction of β hairpin motifs.

The first attempt to predict coiled coils was based on scoring the propensity for formation of coiled coils in the predicted (input) sequence by calculating similarity to a position-specific scoring matrix derived from a statistical analysis of a coiled coil database (67). More recent studies utilize the hidden Markov models and the PSSM profile to represent the input sequence (48, 127-131).

The initial study on the prediction of α -turn- α motif was also based on scoring similarity between the predicted sequence and the α -turn- α structure library (132). Subsequently, a statistical method that utilizes a pattern dictionary of the primary sequences was developed (133). A more recent predictor exploits the potential for using structural knowledge to improve the detection of the helix-turn-helix motifs (134). This method uses a linear predictor that takes similarity scores between the input protein structure and a template library of α -turn- α structures as its inputs.

Table 2 summarizes 16 supersecondary structure prediction methods, including 6 β hairpin predictors: method by de la Cruz *et al.* (51), BhairPred (52), and methods by Hu *et al.* (125),

Zou *et al.* (124), Xia *et al.* (123), and Jia *et al.* (122); 7 recent coiled coil predictors: MultiCoil (135), MARCOIL (131), PCOILS (130), bCIPA (129), Paircoil2 (128), CCHMM_PROF (127), and SpiriCoil (48); and 3 α -turn- α predictors: method by Dodd and Egan (132), GYM (133), and HTHquery (see Note 6) (134). The older coiled coil predictors were reviewed in (60).

We note that some of the methods for the prediction of β hairpin and α -turn- α structures do not offer any implementation, i.e., neither a standalone program nor a web server, which substantially limits their utility. Following, we discuss in greater detail the representative predictors for each type of the SSSs, with particular emphasis on the β hairpin predictors that utilize the predicted secondary structure.

Table 2. Summary of the recent sequence-based predictors of supersecondary structure. The “year last published” column provides the year of the publication of the most recent version of a given method. The “availability” column identifies whether a standalone program (SP), and/or a web server (WS), or neither (NA) is available. The methods are sorted by the year of their last publication in the descending order for a given type of the supersecondary structures.

Supersecondary structure type	Name (<i>authors</i>)	Year last published	Prediction model	Availability
β hairpin	<i>Jia et al.</i>	2011	Random forest	NA
	<i>Xia et al.</i>	2010	Support vector machine	NA
	<i>Zou et al.</i>	2009	Increment of diversity + quadratic discriminant analysis	NA
	<i>Hu et al.</i>	2008	Support vector machine	NA
	BhairPred	2005	Support vector machine	WS
	<i>de la Cruz et al.</i>	2002	Neural network	NA
Coiled coil	SpiriCoil	2010	Hidden Markov model	WS
	CCHMM_PROF	2009	Hidden Markov model	WS
	Paircoil2	2006	Pairwise residue probabilities	WS+SP
	bCIPA	2006	no model	WS
	PCOILS	2005	Residue probabilities	WS
	MARCOIL	2002	Hidden Markov model	SP
	MultiCoil	1997	Pairwise residue probabilities	WS+SP
α -turn- α	HTHquery	2005	Linear predictor	WS
	GYM	2002	Statistical method	WS
	<i>Dodd et al.</i>	1990	Similarity scoring	NA

3.15. BhairPred

The BhairPred predictor was developed by Raghava group at the Institute of Microbial Technology, India in 2005 (52). The predictions are performed using a support vector machine-based model, which is shown by the authors to outperform a neural network-based

predictor. Each residue is encoded using its PSSM profile, secondary structure predicted with PSPRED, and solvent accessibility predicted with the NETASA method (136). BhairPred was shown to provide predictions with accuracy in the 71 to 78% range on two independent test sets (52).

Inputs: PSSM generated from the input protein sequence using PSI-BLAST, 3-state secondary structure predicted using PSIPRED, and solvent accessibility predicted with NETASA

Architecture: support vector machine

Availability: <http://www.imtech.res.in/raghava/bhairpred/>

3.16. CCHMM_PROF

The CCHMM_PROF predictor was developed by Fariselli group at the University of Bologna in 2009 (127). CCHMM_PROF is the first hidden Markov model-based predictor of coiled-coils that exploits the PSSM profile to encoding the input sequence. The major difference between CCHMM_PROF and other hidden Markov models is that the states of CCHMM_PROF produce vectors instead of symbols. The CCHMM_PROF achieved accuracy of 97% when discriminating between sequence that do and do not contain coiled coils (127). This predictor finds the location of the coiled coil segments with 80% success rate and was shown to outperform older solutions (127).

Inputs: PSSM generated from the input protein sequence using PSI-BLAST

Architecture: hidden Markov model

Availability: http://gpcr.biocomp.unibo.it/cgi/predictors/cchmmprof/pred_cchmmprof.cgi

3.17. HTHquery

This method was developed by Thornton group at the European Bioinformatics Institute in 2005 (134). HTHquery takes a protein structure as input and tests whether this structure has a helix–turn–helix motif which could bind to DNA. The input protein is compared with a set of structural templates and putative α –turn– α regions with the smallest RMSD to each template in a template library are determined using Kabsch algorithm (137). The accessible surface area and the electrostatic motif score are computed for each of these putative regions using NACCESS (<http://www.bioinf.manchester.ac.uk/naccess/>) and the methods described in (138), respectively. Next, these inputs, i.e., the minimum RMSD, the accessible surface area, and the electrostatic motif score, are inputted into a linear predictor. HTHquery provides predictions with a true positive rate of 83.5% and a false positive rate of 0.8% (134).

Inputs: protein structure

Architecture: linear predictor

Availability: <http://www.ebi.ac.uk/thornton-srv/databases/HTHquery>

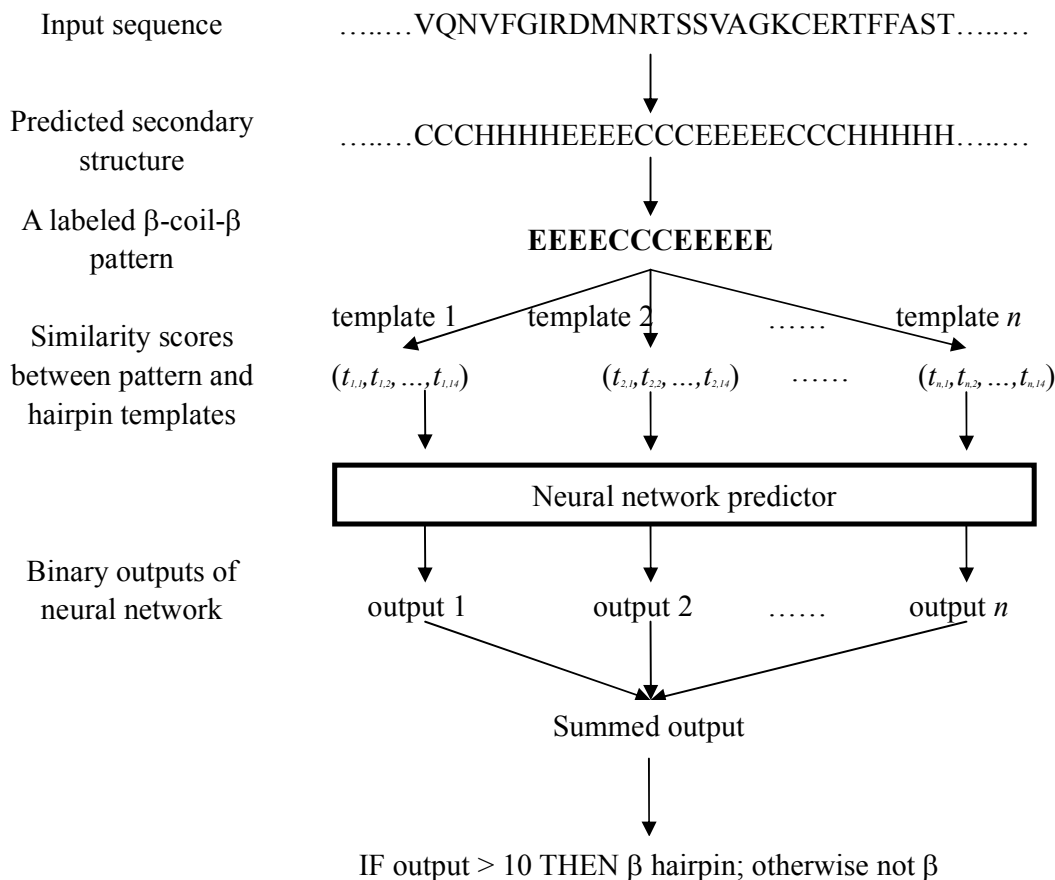


Figure 4. The architecture of the β hairpin predictor proposed by the Thornton group (51).

3.18. Supersecondary structure prediction by using predicted secondary structure

Since supersecondary structure is composed of several adjacent secondary structure elements, the prediction of the secondary structures should be a useful input to predict SSS (see Note 5). Two SSS predictors, BhairPred (52) and the method developed by Thornton group (51), have utilized the predicted secondary structure for the identification of β hairpins. Following, we discuss latter method to demonstrate how the predicted secondary structure is used for the prediction of the SSS. The Thornton *et al.* method consists of 5 steps:

Step 1. Predict the secondary structure for a given input sequence using PHD method (40).

Step 2. Label all β -coil- β patterns in the predicted secondary structure.

Step 3. Score similarity between each labeled pattern and each hairpin structure in a template library. The similarity vector between a β -coil- β pattern and a hairpin structure consists of 14 values, including 6 values that measure similarity of the secondary structures, 1 value that measures similarity of the solvent accessibility, 1 value that indicates the presence of turns, 2 values that describe specific pair interactions and nonspecific distance-based contacts, and 4 values that represent the secondary

structure patterns related to residue length.

Step 4. The 14 similarity scores are processed by a neural network that produces a discrete output, 0 or 1, indicating that the strand-coil-strand pattern is unlikely or likely, respectively, to form a β hairpin.

Step 5. For a given labeled β -coil- β pattern, a set of similarity scores is generated for each template hairpin, and therefore the neural network generates an output for each template hairpin. The labeled β -coil- β pattern is predicted as β hairpin if the outputs are set to 1 for more than 10 template hairpins.

The working of the Thornton *et al.* method is visualized in Figure 4.

4. Notes

1. The predictive quality of the secondary structure predictors was empirically compared in several large-scale, world-wide initiatives including CASP (83), Critical Assessment of Fully Automated Structure Prediction (CAFASP) (139), and EVA (84, 119). Only the early CASP and CAFASP meetings, including CASP3 in 1998, CASP4 and CAFASP2 in 2000, and CASP5 and CAFASP3 in 2002, included the evaluation of the secondary structure predictions. Later on, the evaluations were carried out within the EVA platform. Its most recent release monitored thirteen predictors. However, EVA was last updated in mid 2008.
2. The arguably most popular secondary structure predictor is PSIPRED. This method is implemented as both a standalone application (version 2.6) and a web server (version 3.0). PSIPRED is continuously improved, usually with a major upgrade every year and with weekly updates of the databases. The current (as of June 2011) count of citations in the ISI Web of Knowledge to the paper that describes the original PSIPRED algorithm (38) is close to 1700, which demonstrates the high utility of this method.
3. A recent large-scale comparative analysis (16) has revealed a number of interesting and practical observations concerning state-of-the-art in the secondary structure prediction. The accuracy of the 3-state prediction based on the DSSP assignment is currently at 82%, and the use of a simple consensus-based prediction improves the accuracy by additional 2%. The homology modeling-based methods, such as SSpro and PROTEUS, are shown to be better by 1.5% accuracy than the ab-initio approaches. The neural network-based methods are demonstrated to outperform the hidden Markov model-based solutions.
4. As shown in (16), the current secondary structure predictors are characterized by several drawbacks, which motivate further research in this area. They confuse 1-6% of strand residues with helical residues and vice versa (these are significant mistakes) and they perform poorly when predicting residues in the beta-bridge and 3_{10} helix conformations.
5. The major obstacle to utilize the predicted secondary structure in the prediction of the supersecondary structures, which was observed in mid 2000s, was (is) the inadequate quality of the predicted secondary structure. For instance, only about half of the native β hairpins were predicted with the strand-coil-strand secondary structure pattern (51). The use the native rather than the predicted secondary structure was shown to lead to a significant improvement in the prediction of the supersecondary structures (52).
6. Prediction of the supersecondary structures could be potentially improved by utilizing a consensus of different approaches. As shown in a relatively recent comparative analysis of

coiled coil predictors (60), the best-performing Marcoil has generated many false positives for highly charged fragments, while the runner-up PCOILS provided better predictions for these fragments. This suggests that the results generated by different coiled coil predictors could be complementary.

Acknowledgment

This work was supported by the Alberta Ingenuity and Alberta Innovates Graduate Student Scholarship to KC and the NSERC Discovery grant to LK.

References

1. Pauling L., Corey R.B., Branson H.R. (1951) The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. *Proc Natl Acad Sci USA*. 37, 205-11.
2. Pauling L., Corey R.B. (1951) The pleated sheet, a new layer configuration of polypeptide chains. *Proc Natl Acad Sci USA*. 37, 251-6.
3. Anfinsen C.B. (1973) Principles that govern the folding of protein chains. *Science*, 181, 223–230.
4. Berman H.M., Westbrook J., Feng Z., et al. (2000) The Protein Data Bank. *Nucleic Acids Res*. 28, 235-242.
5. Pruitt K.D., Tatusova T., Klimke W., et al. (2009) NCBI Reference Sequences: current status, policy, and new initiatives. *Nucleic Acids Res*. 37(Database issue), D32-6
6. Gronwald W., Kalbitzer H.R. (2010) Automated protein NMR structure determination in solution. *Methods Mol Biol*. 673, 95-127.
7. Chayen N.E. (2009) High-throughput protein crystallization. *Adv Protein Chem Struct Biol*. 77, 1-22.
8. Zhang Y. (2009) Protein structure prediction: when is it useful? *Curr Opin Struct Biol*. 19, 145-55.
9. Ginalski K. (2006) Comparative modeling for protein structure prediction. *Curr Opin Struct Biol*. 16, 172-7.
10. Yang Y., Faraggi E., Zhao H., et al. (2011) Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of the query and corresponding native properties of templates. *Bioinformatics*. *Bioinformatics* 27(15):2076-2082
11. Roy A., Kucukural A., Zhang Y. (2010) I-TASSER: a unified platform for automated protein structure and function prediction. *Nat Protoc*. 5, 725-38
12. Faraggi E., Yang Y., Zhang S., et al. (2009) Predicting continuous local structure and the effect of its substitution for secondary structure in fragment-free protein structure prediction. *Structure*. 17, 1515-27.
13. Wu S., Zhang Y. (2008) MUSTER: Improving protein sequence profile-profile alignments by using multiple sources of structure information. *Proteins*. 72, 547-56.
14. Zhou H., Skolnick J. (2007) Ab initio protein structure prediction using chunk-TASSER. *Biophys J*. 93, 1510-8.

15. Skolnick J. (2006) In quest of an empirical potential for protein structure prediction. *Curr Opin Struct Biol.* 16, 166-71.
16. Zhang H., Zhang T., Chen K., et al. (2011) Critical assessment of high-throughput standalone methods for secondary structure prediction. *Briefings in Bioinformatics.* 12(6):672-688
17. Pei J., Grishin N.V. (2007) PROMALS: towards accurate multiple sequence alignments of distantly related proteins. *Bioinformatics.* 23, 802-8.
18. Zhang T., Zhang H., Chen K., et al. (2010) Analysis and prediction of RNA-binding residues using sequence, evolutionary conservation, and predicted secondary structure and solvent accessibility. *Curr Protein Pept Sci.* 11, 609-28.
19. Pulim V., Bienkowska J., Berger B. (2008) LTHREADER: prediction of extracellular ligand-receptor interactions in cytokines using localized threading. *Protein Sci.* 17, 279-92.
20. Fischer J.D., Mayer C.E., Söding J. (2008) Prediction of protein functional residues from sequence by probability density estimation. *Bioinformatics.* 24, 613-20.
21. Song J., Tan H., Mahmood K., et al. (2009) Prodepth: predict residue depth by support vector regression approach from protein sequences only. *PLoS One.* 4, e7072.
22. Zhang H., Zhang T., Chen K., et al. (2008) Sequence based residue depth prediction using evolutionary information and predicted secondary structure. *BMC Bioinformatics.* 9, 388.
23. Mizianty M.J., Kurgan L. (2009) Modular prediction of protein structural classes from sequences of twilight-zone identity with predicting sequences. *BMC Bioinformatics.* 10, 414.
24. Kurgan L., Cios K., Chen K. (2008) SCPRED: accurate prediction of protein structural class for sequences of twilight-zone similarity with predicting sequences. *BMC Bioinformatics.* 9:226.
25. Chen K., Kurgan L. (2007) PFRES: protein fold classification by using evolutionary information and predicted secondary structure. *Bioinformatics.* 23, 2843-50.
26. Xue B., Faraggi E., Zhou Y. (2009) Predicting residue-residue contact maps by a two-layer, integrated neural-network method. *Proteins.* 76, 176-83.
27. Cheng J., Baldi P. (2007) Improved residue contact prediction using support vector machines and a large feature set. *BMC Bioinformatics.* 8,113.
28. Mizianty M.J., Stach W., Chen K., et al. (2010) Improved sequence-based prediction of disordered regions with multilayer fusion of multiple information sources. *Bioinformatics.* 26, i489-96.
29. Mizianty M.J., Zhang T., Xue B., et al. (2011) In-silico prediction of disorder content using hybrid sequence representation. *BMC Bioinformatics.* 12, 245.
30. Schlessinger A., Punta M., Yachdav G., et al. (2009) Improved disorder prediction by combination of orthogonal approaches. *PLoS One.* 4, e4433.
31. Zhang H., Zhang T., Gao J., et al. (2012) Determination of protein folding kinetic types using sequence and predicted secondary structure and solvent accessibility. *Amino Acids.* 42(1):271-283.
32. Gao J., Zhang T., Zhang H., et al. (2010) Accurate prediction of protein folding rates from sequence and sequence-derived residue flexibility and solvent accessibility.

- Proteins. 78, 2114-30.
33. Jiang Y., Iglinski P., Kurgan L. (2009) Prediction of protein folding rates from primary sequences using hybrid sequence representation. *J Comput Chem.* 30, 772-83.
 34. Mizianty M., Kurgan L. (2011) Sequence-based prediction of protein crystallization, purification, and production propensity. *Bioinformatics.* 27, i24-i33
 35. Slabinski L., Jaroszewski L., Rychlewski L., et al. (2007) XtalPred: a web server for prediction of protein crystallizability. *Bioinformatics,* 23, 3403-3405.
 36. Bryson K., McGuffin L.J., Marsden R.L., et al. (2005) Protein structure prediction servers at University College London. *Nucleic Acids Res.* 33, W36-8.
 37. Kurgan L., Miri Disfani F. (2011) Structural protein descriptors in 1-dimension and their sequence-based predictions. *Current Protein and Peptide Science.* 12(6):470-489
 38. Jones D. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* 292, 195-202.
 39. Buchan D.W., Ward S.M., Lobley A.E., et al. (2010) Protein annotation and modelling servers at University College London. *Nucleic Acids Res.* 38, W563-8.
 40. Rost B. (1996) PHD: predicting one-dimensional protein structure by profile-based neural networks. *Methods Enzymol.* 266, 525-539.
 41. Rost B., Yachdav G, Liu J. (2004) The PredictProtein Server. *Nucleic Acids Res.* 32(Web Server issue), W321-W326.
 42. O'Donnell C.W., Waldspühl J., Lis M., et al. (2011) A method for probing the mutational landscape of amyloid structure. *Bioinformatics.* 27, i34-i42.
 43. Bryan A.Jr., Menke M., Cowen L.J., et al. (2009) BETASCAN: probable beta-amyloids identified by pairwise probabilistic analysis. *PLoS Comput Biol.* 5, e1000333.
 44. Bradley P., Cowen L., Menke M., et al. (2001) BETAWRAP: successful prediction of parallel beta -helices from primary sequence reveals an association with many microbial pathogens. *Proc Natl Acad Sci U S A.* 98, 14819-24.
 45. Hornung T., Volkov O.A., Zaida T.M., et al. (2008) Structure of the cytosolic part of the subunit b-dimer of Escherichia coli F0F1-ATP synthase. *Biophys J.* 94, 5053-64.
 46. Sun Z.R., Cui Y., Ling L.J., et al. (1998) Molecular dynamics simulation of protein folding with supersecondary structure constraints. *J Protein Chem.* 17, 765-9.
 47. Szappanos B., Süveges D., Nyitray L., et al. (2010) Folded-unfolded cross-predictions and protein evolution: the case study of coiled-coils. *FEBS Lett.* 584, 1623-7.
 48. Rackham O.J., Madera M., Armstrong C.T., et al. (2010) The evolution and structure prediction of coiled coils across all genomes. *J Mol Biol.* 403, 480-93.
 49. Gerstein M., Hegyi H. (1998) Comparing genomes in terms of protein structure: surveys of a finite parts list. *FEMS Microbiol Rev.* 22, 277-304.
 50. Reddy C.C., Shameer K., Offmann B.O., et al. (2008) PURE: a webserver for the prediction of domains in unassigned regions in proteins. *BMC Bioinformatics.* 9, 281.
 51. de la Cruz X., Hutchinson E.G., Shepherd A., et al. (2002) Toward predicting protein topology: an approach to identifying beta hairpins. *Proc Natl Acad Sci USA.* 99, 11157-62.
 52. Kumar M., Bhasin M., Natt N.K., et al. (2005) BhairPred: prediction of beta-hairpins in a protein from multiple alignment information using ANN and SVM techniques. *Nucleic Acids Res.* 33(Web Server issue), W154-9.

53. Barton G.J. (1995) Protein secondary structure prediction. *Curr Opin Struct Biol.* 5, 372-6.
54. Heringa J. (2000) Computational methods for protein secondary structure prediction using multiple sequence alignments. *Curr Protein Pept Sci.* 1, 273-301.
55. Rost B. (2001) Protein secondary structure prediction continues to rise. *J Struct Biol.* 134, 204-18.
56. Albrecht M., Tosatto S.C., Lengauer T., et al. (2003) Simple consensus procedures are effective and sufficient in secondary structure prediction. *Protein Eng.* 16,459-62.
57. Rost B. (2009) Prediction of protein structure in 1D – secondary structure, membrane regions, and solvent accessibility. In: Bourne PE, Weissig H (eds). *Structural Bioinformatics*, 2nd Edition. Wiley, 2009, 679-714.
58. Pirovano W., Heringa J. (2010) Protein secondary structure prediction. *Methods Mol. Biol.* 609, 327-48.
59. Singh M. Predicting Protein Secondary and Supersecondary Structure. In: *Handbook of Computational Molecular Biology*. Aluru S. (Ed). Chapman and Hall / CRC Press, 2006, pp.29.1-29.29
60. Gruber M., Söding J., Lupas A.N. (2006) Comparative analysis of coiled-coil prediction methods. *J Struct Biol.* 155, 140-5.
61. Kolodny R., Honig B. (2006) VISTAL-a new 2D visualization tool of protein 3D structural alignments. *Bioinformatics.* 22, 2166-7.
62. Moreland J.L., Gramada A., Buzko O.V., et al. (2005) The Molecular Biology Toolkit (MBT): a modular platform for developing molecular visualization applications. *BMC Bioinformatics.* 6, 21.
63. Porollo A.A., Adamczak R., Meller J. (2004) POLYVIEW: a flexible visualization tool for structural and functional annotations of proteins. *Bioinformatics.* 20, 2460-2.
64. Murzin A.G, Brenner S. E., Hubbard T., et al. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247, 536-540.
65. Orengo C.A., Michie A.D., Jones S., et al. (1997) CATH--a hierarchic classification of protein domain structures. *Structure.* 5, 1093-108.
66. Andreeva A., Howorth D., Chandonia J.M., et al. (2008) Data growth and its impact on the SCOP database: new developments. *Nucl. Acids Res.* 36, D419-D425
67. Cuff A.L., Sillitoe I., Lewis T., et al. (2011) Extending CATH: increasing coverage of the protein structure universe and linking structure with function. *Nucleic Acids Res.* 39(Database issue):D420-6.
68. Levitt M., Greer J. (1997) Automatic identification of secondary structure in globular proteins. *J. Mol. Biol.*, 114, 181-239.
69. Kabsch W., Sander C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22, 2577-2637.
70. Richards F., Kundrot C.E. (1988) Identification of structural motifs from protein coordinate data: secondary structure and first-level supersecondary structure. *Proteins*, 3, 71-84.
71. Sklenar H., Etchebest C., Lavery R. (1989) Describing protein structure: a general algorithm yielding complete helicoidal parameters and a unique overall axis. *Proteins.* 6,

- 46-60.
72. Frishman D., Argos P. (1995) Knowledge-based protein secondary structure assignment. *Proteins*. 23, 566-579.
 73. Labesse G., Colloc'h N., Pothier J., et al. (1997) P-SEA: a new efficient assignment of secondary structure from C alpha trace of proteins. *Comput. Appl. Biosci.* 13, 291-295.
 74. King S., Johnson W.C. (1999) Assigning secondary structure from protein coordinate data. *Proteins*, 3, 313-320.
 75. Fodje M., Al-Karadaghi S. (2002) Occurrence, conformational features and amino acid propensities for the pi-helix. *Protein Eng.* 15, 353-358.
 76. Martin J., Letellier G., Marin A., et al. (2005) Protein secondary structure assignment revisited: a detailed analysis of different assignment methods. *BMC Struct. Biol.* 5, 17.
 77. Cubellis M.V., Cailliez F., Lovell S.C. (2005) Secondary structure assignment that accurately reflects physical and evolutionary characteristics. *BMC Bioinformatics*. 6 Suppl 4, S8.
 78. Majumdar I., Krishna S.S., Grishin N.V. (2005) PALSSE: a program to delineate linear secondary structural elements from protein structures. *BMC Bioinformatics*. 6, 202.
 79. Zhang W., Dunker A.K., Zhou Y. (2008) Assessing secondary structure assignment of protein structures by using pairwise sequence-alignment benchmarks. *Proteins*. 71, 61-7.
 80. Hosseini S.R., Sadeghi M., Pezeshk H., et al. (2008) PROSIGN: a method for protein secondary structure assignment based on three-dimensional coordinates of consecutive C(alpha) atoms. *Comput Biol Chem.* 32, 406-11.
 81. Park S.Y., Yoo M.J., Shin J., et al. (2011) SABA (secondary structure assignment program based on only alpha carbons): a novel pseudo center geometrical criterion for accurate assignment of protein secondary structures. *BMB Rep.* 44, 118-22.
 82. Klose D.P., Wallace B.A., Janes R.W. (2010) 2Struc: the secondary structure server. *Bioinformatics*. 26, 2624-5.
 83. Moulton J., Pedersen J.T., Judson R., et al. (1995) A large-scale experiment to assess protein structure prediction methods. *Proteins*. 23, ii-v.
 84. Koh I.Y., Eyrich V.A., Marti-Renom M.A., et al. (2003) EVA: Evaluation of protein structure prediction servers. *Nucleic Acids Res.* 31, 3311-15.
 85. Parry D.A. (2008) Fifty years of coiled-coils and alpha-helical bundles: a close relationship between sequence and structure. *J. Struct. Biol.* 163, 258-269.
 86. Pellegrini-Calace M., Thornton J.M. (2005) Detecting DNA-binding helix-turn-helix structural motifs using sequence and structure information. *Nucleic Acids Res.* 33, 2129-40.
 87. Hutchinson E.G., Thornton J.M. (1996) PROMOTIF—a program to identify and analyze structural motifs in proteins. *Protein Sci.* 5, 212–220.
 88. Walshaw J., Woolfson D.N. (2001) Socket: a program for identifying and analysing coiled-coil motifs within protein structures. *J Mol Biol.* 307, 1427-50.
 89. Testa O.D., Moutevelis E., Woolfson D.N. (2009) CC+: a relational database of coiled-coil structures. *Nucleic Acids Res.* 37(Database issue), D315-22.
 90. Michalopoulos I., Torrance G.M., Gilbert D.R., et al. (2004) TOPS: an enhanced database of protein structural topology. *Nucleic Acids Res.* 32(Database issue), D251-4.
 91. Rost B., Sander C. (1993) Improved prediction of protein secondary structure by use of

- sequence profiles and neural networks. *Proc Natl Acad Sci U S A.* 90, 7558-62.
92. Altschul S.F., Madden T.L., Schäffer A.A., et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, 25, 3389-402.
 93. Faraggi E., Xue B., Zhou Y. (2009) Improving the prediction accuracy of residue solvent accessibility and real-value backbone torsion angles of proteins by guided-learning through a two-layer neural network. *Proteins.* 74, 847-56.
 94. Dor O., Zhou Y. (2007) Achieving 80% ten-fold cross-validated accuracy for secondary structure prediction by large-scale training. *Proteins*, 66, 838-845.
 95. Zhou, T., Shu N., Hovmöller S. (2010) A novel method for accurate one-dimensional protein structure prediction based on fragment matching. *Bioinformatics*, 26, 470-477.
 96. Kountouris P., Hirst J.D. (2009) Prediction of backbone dihedral angles and protein secondary structure using support vector machines. *BMC Bioinformatics*, 10, 437.
 97. Karplus K., Barrett C., Hughey R. (1998) Hidden Markov models for detecting remote protein homologies. *Bioinformatics.* 14, 846-56.
 98. Karplus K., Karchin R., Barrett C., et al. (2001) What is the value added by human intervention in protein structure prediction? *Proteins. Suppl 5*, 86-91.
 99. Karplus K., Karchin R., Draper J., et al. (2003) Combining local-structure, fold-recognition, and new fold methods for protein structure prediction. *Proteins*, 53, 491-496.
 100. Karplus K., Katzman S., Shackelford G., et al. (2005) SAM-T04: what is new in protein-structure prediction for CASP6. *Proteins*, 61(Suppl 7), 135-142.
 101. Karplus K. SAM-T08, HMM-based protein structure prediction. (2009) *Nucleic Acids Res.* 37(Web Server issue), W492-W497.
 102. Montgomerie S., Cruz J.A., Shrivastava S., et al. (2008) PROTEUS2: a web server for comprehensive protein structure prediction and structure-based annotation. *Nucleic Acids Res.* 36(Web Server issue), W202-W209.
 103. Montgomerie S., Sundararaj S., Gallin W.J., et al. (2006) Improving the accuracy of protein secondary structure prediction using structural alignment. *BMC Bioinformatics.* 7, 301.
 104. Cole C., Barber J.D., Barton G.J. (2008) The Jpred 3 secondary structure prediction server. *Nucleic Acids Res.* 2008, 36, W197-W201.
 105. Cuff J.A., Clamp M.E., Siddiqui A.S., et al. (1998) JPred: a consensus secondary structure prediction server. *Bioinformatics.* 14, 892-893.
 106. Cuff J., Barton G.J. (2000) Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins*, 40, 502-511.
 107. Won K., Hamelryck T., Prügél-Bennett A., et al. (2007) An evolutionary method for learning HMM structure: prediction of protein secondary structure. *BMC Bioinformatics.* 8, 357.
 108. Pollastri G, Martin A.J.M., Mooney C., et al. (2007) Accurate prediction of protein secondary structure and solvent accessibility by consensus combiners of sequence and structure information. *BMC Bioinformatics.* 8, 201
 109. Pollastri G, McLysaght A. (2005) Porter: a new, accurate server for protein secondary structure prediction. *Bioinformatics*, 21, 1719-1720.

110. Martin J., Gibrat J.F., Rodolphe F. (2006) Analysis of an optimal hidden Markov model for secondary structure prediction. *BMC Struct. Biol.* 6, 25.
111. Karypis G. (2006) YASSPP: better kernels and coding schemes lead to improvements in protein secondary structure prediction. *Proteins*, 64, 575-586.
112. Lin K., Simossis V.A., Taylor W.R., et al. (2005) A simple and fast secondary structure prediction algorithm using hidden neural networks. *Bioinformatics*, 21, 152-159.
113. Adamczak R., Porollo A., Meller J. (2005) Combining prediction of secondary structure and solvent accessibility in proteins. *Proteins*. 59, 467-475.
114. Cheng J., Randall A.Z., Sweredoski M.J., et al. (2005) SCRATCH: a protein structure and structural feature prediction server. *Nucleic Acids Res.* 33, W72-W76.
115. Pollastri G., Przybylski D., Rost B., et al. (2002) Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins*. 47, 228-235.
116. Madera M., Calmus R., Thiltgen G., et al. (2010) Improving protein secondary structure prediction using a simple k-mer model. *Bioinformatics*. 26, 596-602.
117. Lupas A., Van Dyke M., Stock J. (1991) Predicting coiled coils from protein sequences. *Science*, 252, 1162-1164.
118. Eddy S.R. (1998) Profile hidden Markov models. *Bioinformatics*. 14, 755-63.
119. Eyrich V.A., Martí-Renom M.A., Przybylski D., et al. (2001) EVA: continuous automatic evaluation of protein structure prediction servers. *Bioinformatics*. 17, 1242-3.
120. Bau D., Martin A.J., Mooney C., et al. (2006) Distill: a suite of web servers for the prediction of one-, two- and three-dimensional structural features of proteins. *BMC Bioinformatics*. 7, 402.
121. Zhang T., Faraggi E., Zhou Y. (2010) Fluctuations of backbone torsion angles obtained from NMR-determined structures and their prediction. *Proteins*. 78, 3353-62.
122. Jia S.C., Hu X.Z. (2011) Using Random Forest Algorithm to Predict β -Hairpin Motifs. *Protein Pept Lett.* 18, 609-17.
123. Xia J.F., Wu M., You Z.H., et al. (2010) Prediction of beta-hairpins in proteins using physicochemical properties and structure information. *Protein Pept Lett.* 17, 1123-8.
124. Zou D., He Z., He J. (2009) Beta-hairpin prediction with quadratic discriminant analysis using diversity measure. *J Comput Chem.* 30, 2277-84.
125. Hu X.Z., Li Q.Z. (2008) Prediction of the beta-hairpins in proteins using support vector machine. *Protein J.* 27, 115-22.
126. Kuhn M., Meiler J., Baker D. (2004) Strand-loop-strand motifs: prediction of hairpins and diverging turns in proteins. *Proteins*. 54, 282-8.
127. Bartoli L., Fariselli P., Krogh A., et al. (2009) CCHMM_PROF: a HMM-based coiled-coil predictor with evolutionary information. *Bioinformatics*. 25, 2757-63.
128. McDonnell A.V., Jiang T., Keating A.E., et al. (2006) Paircoil2: improved prediction of coiled coils from sequence. *Bioinformatics*. 2006, 22, 356-8.
129. Mason J.M., Schmitz M.A., Müller K.M., et al. (2006) Semirational design of Jun-Fos coiled coils with increased affinity: Universal implications for leucine zipper prediction and design. *Proc Natl Acad Sci U S A.* 103, 8989-94.
130. Gruber M., Söding J., Lupas A.N. (2005) REPPER--repeats and their periodicities in fibrous proteins. *Nucleic Acids Res.* 33(Web Server issue), W239-43.

131. Delorenzi M., Speed T. (2002) An HMM model for coiled-coil domains and a comparison with PSSM-based predictions. *Bioinformatics*. 18, 617-25.
132. Dodd I.B., Egan J.B. (1990) Improved detection of helix-turn-helix DNA-binding motifs in protein sequences. *Nucleic Acids Res.* 18, 5019-5026.
133. Narasimhan G., Bu C., Gao Y., et al. (2002) Mining protein sequences for motifs. *J Comput Biol.* 9, 707-20.
134. Ferrer-Costa C., Shanahan H.P., Jones S., et al. (2005) HTHquery: a method for detecting DNA-binding proteins with a helix-turn-helix structural motif. *Bioinformatics*. 21, 3679-80.
135. Wolf E., Kim P.S., Berger B. (1997) MultiCoil: a program for predicting two- and three-stranded coiled coils. *Protein Sci.* 6(6), 1179-89.
136. Ahmad S., Gromiha M.M. (2002) NETASA: neural network based prediction of solvent accessibility. *Bioinformatics* 18(6), 819-24.
137. Kabsch W. (1976) A solution for the best rotation to relate two sets of vectors. *Acta Cryst.* A32, 922-923.
138. Shanahan H., Garcia M., Jones S., et al. (2004) Identifying DNA binding proteins using structural motifs and the electrostatic potential. *Nucleic Acids Res.* 32, 4732-41
139. Fischer D., Barret C., Bryson K., et al. (1999) CAFASP-1: critical assessment of fully automated structure prediction methods. *Proteins Suppl* 3, 209-17.