Chapter 4

Computational Prediction of Secondary and Supersecondary Structures from Protein Sequences

Christopher J. Oldfield, Ke Chen, and Lukasz Kurgan

Abstract

Many new methods for the sequence-based prediction of the secondary and supersecondary structures have been developed over the last several years. These and older sequence-based predictors are widely applied for the characterization and prediction of protein structure and function. These efforts have produced countless accurate predictors, many of which rely on state-of-the-art machine learning models and evolutionary information generated from multiple sequence alignments. We describe and motivate both types of predictions. We introduce concepts related to the annotation and computational prediction of the three-state and eight-state secondary structure as well as several types of supersecondary structures, such as β hairpins, coiled coils, and α-turn-α motifs. We review 34 predictors focusing on recent tools and provide detailed information for a selected set of 14 secondary structure and 3 supersecondary structure predictors. We conclude with several practical notes for the end users of these predictive methods.

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*, which concerns regular local substructures including α-helices and β-strands that were first postulated by Pauling and coworkers [1, 2]; and *tertiary structure*, which is the three-dimensional structure of a protein molecule. Supersecondary structure (SSS) bridges the two latter levels and concerns specific combinations/geometric arrangements of just 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 the help of various types of coils, to form the tertiary structure. An example that displays the

Alexander E. Kister (ed.), *Protein Supersecondary Structures: Methods and Protocols*, Methods in Molecular Biology, vol. 1958, https://doi.org/10.1007/978-1-4939-9161-7_4, © Springer Science+Business Media, LLC, part of Springer Nature 2019

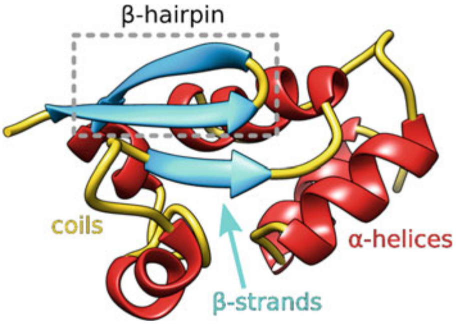

Fig. 1 Cartoon representation of the tertiary structure of the T1 domain of human renal potassium channel Kv1.3 (PDB code: 4BGC). Secondary structures are color-coded: α-helices (red), β-strands (blue), and coils (yellow). The β hairpin supersecondary structure motif, which consists of two β-strands and the coil between them, is denoted using the dotted rectangle

secondary structures and the β hairpin supersecondary structure is given in Fig. 1.

In the 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 May 2018, about 140,000 structures for 44,000 distinct protein sequences are included in the Protein Data Bank (PDB) [4, 5] when compared with 110.3 million nonredundant protein sequences in the RefSeq database [6, 7]. Moreover, the experimental determination of protein structure is relatively expensive and time-consuming and cannot keep up with the rapid accumulation of the sequence data [8–14]. One successful way to predict the tertiary structure is to proceed in a stepwise fashion. First, we predict how the sequence folds into the secondary structure and 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 [15–22].

The last three decades observed strong progress in the development of accurate predictors of the secondary structure, with predictions with about 82% accuracy [23]. This number has climbed in recent years to 84%, which is approaching the estimated accuracy limit of approximately 88% [24]. Besides being useful for the prediction of the tertiary structure, the secondary structure predicted from the sequence is widely adopted for analysis and prediction of numerous structural and functional characteristics of proteins. These applications include computation of multiple alignment [25], target selection for structural genomics [26–28], and prediction of protein-nucleic acids interactions [29–32], protein-ligand interactions [33–35], residue depth [36, 37], beta-turns [38], structural classes and folds [39–43], residue contacts [44, 45], disordered regions [46–51], disordered linker regions [52], disordered protein-binding regions [53, 54], and folding rates and types [55–57], to name selected few. Secondary structure predictors enjoy strong interest, which could be quantified by the massive workloads that they handle. For instance, the web server of the arguably one of the most popular methods, PSIPRED, was reported already in 2005 to receive over 15,000 requests per month [58]. Another indicator is the fact that many of these methods receive high citations counts. A review [59] reported that 7 methods were cited over 100 times, and two of them, PSIPRED [58, 60, 61] and PHD [62, 63], 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 the 1980s, and to date about 20 predictors were developed. Similarly as the secondary structure predictors, the predictors of SSS found applications in numerous areas including analysis of amyloids [64, 65], microbial pathogens [66], and synthases [67], simulation of protein folding [68], analysis of relation between coiled coils and disorder [69], genome-wide studies of protein structure [70, 71], and prediction of protein domains [72]. 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 group [73] and the BhairPred method [74], 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 [75–77] and the utilization of consensus-based approaches [78, 79]. Several 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 [23, 80, 81]. Most recent reviews cover many of the current state-of-the-art secondary structure prediction methods, but they lack the coverage of the supersecondary structure predictors [24, 82]. The SSS prediction area has been reviewed less extensively. The β hairpin and coiled coil predictors, as well as the secondary structure predictors, were overviewed in 2006 [83], and a comparative analysis of the coiled coil predictors was presented in the same year [84]. A recent review provides an in depth guide to the prediction of coiled coils [85]. To the best of our knowledge, there were only two surveys that covered both secondary and supersecondary structure predictors [86, 87]. This chapter extends our review from 2013 on the predictors of secondary and supersecondary structures [87] by including the most recent advancements and methods in these active areas of research. We summarize a comprehensive set of 34 recent secondary structure and SSS predictors, with 17 methods for each type of predictions. We also demonstrate how the prediction of the secondary structure is used to implement a SSS predictor and provide several practical notes for the end users.

2 Materials

2.1 Assignment of Secondary Structure

Secondary structure, which is assigned from experimentally determined protein structure, is used for a variety of applications, including visualization [88–90] and classification of protein folds [91–96], 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 the late 1970s by Levitt and Greer [97]. This was followed by Kabsch and Sander who developed a method called Dictionary of Protein Secondary Structure (DSSP) [98], which is based on the detection of hydrogen bonds defined by an electrostatic criterion. Many other secondary structure assignment methods have been developed, including (in chronological order): DEFINE [99], P-CURVE [100], STRIDE [101], P-SEA [102], XTLSSTR [103], SECSTR [104], KAKSI [105], Segno [106], PALSSE [107], SKSP [108], PROSIGN [109], SABA [110], PSSC [111], PCASSO [112], and SACF [113]. Moreover, the 2Struc web server provides an integrated access to multiple annotation methods and enables convenient comparison between different assignment protocols [114].

DSSP remains the most widely used protocol [105], 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) [115] and the EValuation of Automatic protein structure prediction (EVA) continuous benchmarking project [116]. DSSP determines secondary structures based on 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 amide group of the residue three positions down in the sequence forming a tight, right-handed helical structure with three 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 four positions away.
- I: (5-turn) π-helix, where the hydrogen bonding occurs between residues spaced five positions away. Most of the π-helices are right-handed.
- E: extended strand, where two 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_i^α to C_{i+2}^α (C^α atoms at the *i*th and $i + 2$th positions in the chain) and the vector from C_{i-2}^α to C_i^α 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* Fig. 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 https://swift.cmbi.umcn.nl/gv/dssp/.

2.2 Assignment of Supersecondary Structures

SSS is composed of several adjacent secondary structure elements. Therefore, the assignment of SSS relies on the assignment of the secondary structure. Among more than a dozen types of SSSs, 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 [74]. The coiled coil motif mediates the oligomerization of a large number of proteins, are involved in regulation of gene expression, and serve as molecular spacers [117, 118]. The α-turn-α (helix-turn-helix) motif is instrumental for DNA binding and transcription regulation [119, 120]. The β hairpin, coiled coil, and α-turn-α motifs are defined as follows:

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

β hairpins are commonly annotated by PROMOTIF program [121], which also assigns several other SSS types, e.g., psi-loop and β-α-β motifs. Similar to DSSP, the PROMOTIF program assigns SSS based on the distances and hydrogen bonding between the residues. Coiled coils are usually assigned with the SOCKET program [122], which locates/annotates coiled coil interactions based on the distances between multiple helical chains. 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 user convenience, certain supersecondary structures, such as the coiled coils and β-α-β motifs, can be accessed, analyzed, and visualized using specialized databases like CCPLUS [123] and TOPS [124]. 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 chirality of selected SSSs, e.g., β hairpins and β-α-β motifs.

2.3 Multiple Sequence Alignment

Multiple sequence alignments were introduced to prediction of secondary structure in the early 1990s [125]. Using multiple sequence alignment information rather than only protein sequence has led to a 10% accuracy improvement in secondary structure prediction [125]. Multiple sequence alignments are also often used in the prediction of SSS [74, 83, 84]. The strength of including multiple sequence alignment information in prediction is the evolutionary information they contain, which is much richer (or accessible) than a single sequence. Multiple sequence alignments can be obtained 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* (nonredundant) 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

```
Query protein  ... E R V V I N I S G L R F E T Q L K T L - Q F P E ...
Q61923         ... E R L V I N I S G L R F E T Q L R T L S L F P D ...
Q8I4B0         ... Q I V T I N V S G M R F Q T F E S T L S R Y P N ...
P17972         ... N R V V L N V G G I R H E T Y K A T L K K I P A ...
Q63881         ... A L I V L N V S G T R F Q T W Q D T L E R Y P D ...
Q01956         ... G K I V I N V G G V R H E T Y R S T L R T L P G ...
P97557         ... D C L T V N V G G S R F V L S Q Q A L S C F P H ...
Q0P583         ... D S F T V N V G G S R F V L S Q Q A L S C F P H ...
O18868         ... R R V R L N V G G L A H E V L W R T L D R L P R ...
```

Fig. 2 Multiple sequence alignment between the input (query) sequence, which is a fragment of the T1 domain of human renal potassium channel Kv1.3 shown in Fig. 1, and similar sequences. 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 Subheading 2.3

the multiple sequence alignment is given in Fig. 2 where eight similar sequences are identified for the input protein (we use the protein from Fig. 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 Fig. 2, the counts of amino acids glutamic acid (E), glutamine (Q), and valine (V) are 5, 2, and 2, respectively. Therefore, this position can be represented by a 20-dimensional vector $(0, 0, 0, 5/9, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2/9, 0, 0, 0, 2/9, 0, 0)$, 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. A multiple sequence alignment profile is composed of these 20-dimensional vectors for each position in the input protein chain and is a common representation of a multiple sequence alignment.

The PSI-BLAST (Position-Specific Iterated BLAST) [126] 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 position-specific scoring matrix (PSSM), which is similar to a profile discussed above with the exception that values are the log-odds of observing a given residue. Another query against the sequence database is run using this first PSSM, and a larger group of sequences is found. This larger group of sequences is used to construct another PSSM, 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 the late 1990s, PSI-BLAST is commonly used for the generation of PSSMs that are often used directly in the prediction of secondary and

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

Fig. 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 Fig. 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, where $p_{i,j}$ indicates the probability that the jth amino acid type (in columns) occurs at ith position in the input chain (in rows), is generated. The position-specific scoring matrix M is defined as $m_{i,j} = \log(p_{i,j}/b_j)$, where b_j is the background frequency of the jth amino acid type

supersecondary structures. An example PSSM profile is given in Fig. 3. The BLAST and PSI-BLAST programs are available at http://blast.ncbi.nlm.nih.gov/.

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. Many secondary structure predictors use a sliding window approach in which a local stretch of residues around a central position in the window is used to predict the secondary structure state at the central position. Moreover, as one of the first steps in the prediction protocol, many 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. Early predictors

Table 1
Summary of the recent sequence-based predictors of secondary structure

Name	Year last published	Prediction model	States	Availability
MUFOLD-SS	2018	Deep neural network	8	SP
SPIDER3	2017	Bidirectional recurrent neural network	3	WS + SP
RaptorX	2016	Deep conditional neural fields	8	WS
Jpred	2015	Neural network	3	WS + API
SCORPION	2014	Neural network	8	WS
PSIPRED	2013	Neural network	3	WS + SP + API
PORTER	2013	Bidirectional recurrent neural network	3	WS + SP
SPARROW	2012	Quadratic model + neural network	3	SP
Frag1D	2010	Scoring function	3	WS + SP
DISSPred	2009	Support vector machine + clustering	3	WS
PCI-SS	2009	Parallel cascade identification	3	WS + API
PROTEUS	2008	Neural network	3	WS + SP
OSS-HMM	2006	Hidden Markov model	3	SP
YASSPP	2006	Support vector machine	3	WS
YASPIN	2005	Neural network + hidden Markov model	3	WS
SABLE	2005	Neural network	3	WS + SP
SSpro	2005	Neural network	8	WS + SP

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), a web server (WS), and/or an application programmer's interface (API) is available. The methods are sorted by the year of their last publication in the descending order

were implemented based on a relatively simple statistical analysis of composition of the input sequence. Modern methods adopt sophisticated machine learning-based classifiers to represent the relation between the input sequence (or more precisely between evolutionary information encoded in its PSSM) and the secondary structure states. In the 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), different scales of the networks (e.g., regular and deep), and different sizes of the sliding windows.

Prediction methods are provided to the end users as standalone applications and/or as web servers. 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 [23] 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 summarizes 17 methods in the reverse chronologic order: MUFOLD-SS [127], SPIDER3 [128], RaptorX [129, 130], Jpred [131–134], SCORPION [135], PSIPRED [58, 60, 61, 136], PORTER [137–139], SPARROW [140], Frag1D [141], DISSPred [142], PCI-SS [143], PROTEUS [131, 144, 145], OS-HMM [146], YASSPP [147], YASPIN [148], SABLE [149], and SSpro [150, 151]. This list is limited to predictors published since 2005 and that have standalone programs or websites available at the time of this writing. Older methods were comprehensively reviewed in [75–77]. Note that only 4 of the 17 methods (MUFOLD-SS, RaptorX, SCORPION, and SSpro) predict the 8-state secondary structure, and these methods also provide the 3-state predictions.

Below, we discuss in detail 14 methods that are listed in reverse chronologic order. These methods 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 observe 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 comparative studies (*see* **Note 2**) [23, 59]. Moreover, research shows that improved predictive performance could be obtained by post-processing of the secondary structure predictions (*see* **Note 3**) [152].

3.1.1 SPIDER3

The SPIDER series of predictors have been developed by the Zhou group at the Griffith University. In particular, SPIDER3 [128] is inspired by its predecessors that have come from the same lab: SPIDER2 [153, 154] and SPINE X [155]. SPIDER3 is designed to consider long-range sequence information to improve secondary structure prediction accuracy. Common neural network architectures require a fixed size input window, and the network is applied to the sequence repeatedly over a sliding window. In contrast, bidirectional recurrent neural networks (BRNN) used in SPIDER3, in a sense, consider the entire sequence simultaneously, with network state being shared along the sequence [156]. Moreover, use of specialized long short-term memory (LSTM) network nodes improves the flow of information between distant sequence positions [157]. The authors demonstrate that the LSTM-BRNN improves prediction accuracy, particularly for residues with many long-range contacts. SPIDER3 achieves a Q3 score of nearly 84% on an independent test dataset.

Inputs: hidden Markov model profiles and PSSM generated from the input protein sequence using HHBlits [158] and PSI-BLAST, respectively.

Architecture: LSTM-BRNN.

Availability: http://sparks-lab.org/server/SPIDER3/.

3.1.2 RaptorX

RaptorX [129, 130] was developed by the Xu group at the University of Chicago. Like SPIDER3, this predictor also considers long-range sequence information but through the use of deep convolutional neural networks (DeepCNF). For each layer of the network, inputs are taken from the previous layer from neighboring positions. This architecture considers information from distant sequence positions, where the depth of the network determines the maximum distance considered. RaptorX uses a window size of 11 residues and 5 layers, resulting in an effective window size of 51 residues. On an independent test set, the authors find RaptorX to outperform all other tested methods.

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

Architecture: deep convolutional neural network.

Availability: http://raptorx2.uchicago.edu/StructurePropertyPred/predict/.

3.1.3 Jpred

Jpred was originally developed in the late 1990s by Barton group at the University of Dundee [132]. This method was updated a few times, with the most recent version Jpred 4 [134]. Similar to PSIPRED, Jpred was demonstrated to provide about 82% accuracy for the three-state secondary structure prediction [134]. 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 [159].

Inputs: hidden Markov model profiles and PSSM generated from the input protein sequence using HMMer [160] and PSI-BLAST, respectively.

Architecture: ensemble of neural networks.

Availability: http://www.compbio.dundee.ac.uk/jpred/.

3.1.4 SCORPION

The Li group of Old Dominion University developed the SCORPION method to take advantage of local sequence features for the prediction of secondary structure [135]. This is accomplished by using statistics derived from residue pairs and triplets at defined sequence distances within a short local window. The other key aspect of this method is its stacked architecture. Stacking is an approach where the output of the first predictor is used for the input to the second predictor, second to third, etc. SCORPION uses a series of three stacked neural networks to refine secondary structure predictions. By the authors' assessment, SCORPION shows superior accuracy to other prediction methods, in both Q3 (three-state accuracy) and Q8 (eight-state accuracy) measures. Though improvement in Q3 accuracy over PSIPRED is small, SCORPION shows a large improvement in segment-based accuracy, indicating a better match to secondary structure elements.

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

Architecture: three stacked neural networks.

Availability: three-state prediction, http://hpcr.cs.odu.edu/c3scorpion/; eight-state prediction, http://hpcr.cs.odu.edu/c8scorpion/.

3.1.5 PSIPRED

PSIPRED is one of the most popular prediction methods (*see* **Note 4**); for example, it received the largest number of citations as shown in [23, 59]. This method was developed in the late 1990s by Jones group at the University College London [60] and was later improved and updated in 2020 and 2013 [61, 136]. 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 three-state secondary structure predictions with 81% accuracy [23, 61]. 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/.

3.1.6 PORTER

This predictor was developed by Pollastri group at the University College Dublin [138]. 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 [137]. The original and the homology-enhanced versions were recently shown to provide 79% [23] and 83% accuracy [59], respectively. PORTER is a part of a comprehensive predictive platform called DISTILL [161], which also incorporates predictors of relative solvent accessibility, residue-residue contact density, contacts maps, subcellular localization, and tertiary structure. The most recent upgrades to PORTER in 2009 [162] and 2013 [139] expanded the training set and architecture, resulting in a significate increase in performance.

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

Architecture: ensemble of recurrent neural networks.

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

3.1.7 Frag1D

The idea behind Frag1D, created by the Hovmöller group at Stockholm University, is that similar sequence fragments will have similar structures [141]. A database of fragments from known structures is compared to fragments of the query protein, where the most similar segments are used to predict secondary structure. Fragment comparison is scored using profile-profile comparison, augmenting sequence-derived profiles with structure-derived profiles. For the query sequence, structure profiles are unknown and are approximated through an iterative procedure of fragment matching. By the authors' evaluation, this method has comparable performance to PSIPRED.

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

Architecture: scoring function.

Availability: http://frag1d.bioshu.se.

3.1.8 DISSPred

The DISSPred approach was recently introduced by Hirst group at the University of Nottingham [142]. Similar to SPIDER3 and its predecessors, this method predicts both the three-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% [142].

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/.

3.1.9 PCI-SS

PCI-SS [143] is a unique approach to secondary structure prediction, developed by the Green group at the Carleton University. The approach, known as parallel cascade identification (PCI), progressively refines predictions using a series of linear/nonlinear function layers. The parameter space of PCI contains discrete components, intractable to conventional training methods, so the authors employ genetic algorithm for model training. The authors find that this method performs comparably to other, more conventional, methods.

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

Architecture: parallel cascade identification.

Availability: http://bioinf.sce.carleton.ca/PCISS.

3.1.10 PROTEUS

This secondary structure prediction approach was developed by Wishart group at the University of Alberta [145]. PROTEUS is a consensus-based method, in which outputs of three secondary structure predictors, namely PSIPRED, Jnet [163], and an in-house TRANSSEC [145], 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 by both the authors [144] and in a comparative survey [23]. This predictor was incorporated into an integrated system called PROTEUS2, which additionally offers prediction of signal peptides, transmembrane helices and strands, and tertiary structure [144].

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/.

3.1.11 YASSPP

YASSPP was designed by Karypis lab at the University of Minnesota in 2005 [147]. Rather than typically used neural network classifiers, YASSPP instead utilizes multiple support vector machine learners. This method was shown to provide similar predictive quality to PSIPRED [147].

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/.

3.1.12 YASPIN

The YASPIN method was developed by Heringa lab at the Vrije Universiteit in 2004 [148]. 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 [148]. The predictive performance of YASPIN was evaluated using EVA benchmark and two comparative assessments [23, 61], which show that this method provides predictions with accuracy in the 76–79% range.

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

Architecture: Two-level hybrid design with neural network in the first level and hidden Markov model in the second level.

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

3.1.13 SABLE

The SABLE predictor was developed by Meller group at the University of Cincinnati [149]. The web server that implements this method was used close to 200,000 times since it became operational in 2003. Two recent comparative studies [23, 61] 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/.

3.1.14 SSpro

SSpro was introduced in early 2000 by the Baldi group at the University of California, Irvine [151]. Its version 4.5 [150] utilizes homology modeling, which is based on alignment to known tertiary structures from PDB, and achieves over 82% accuracy [23]. The SSpro 4.0 was also ranked as one of the top secondary structure prediction servers in the EVA benchmark [164]. SSpro's most recent version 5.2 is part of a comprehensive prediction center called SCRATCH, which also includes predictions of secondary structure in eight states using SSpro8 [151] and prediction of solvent accessibility, intrinsic disorder, contact numbers and contact maps, domains, disulfide bonds, B-cell epitopes, solubility upon overexpression, antigenicity, viral capsid and tail proteins, 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/.

3.2 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 [70], the prediction of 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. The architectures of the methods for the prediction of different types of SSSs vary more substantially when compared with the fairly uniform architectures of the modern secondary structure predictors that primarily rely on the neural networks.

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 [73]. More recent β hairpin predictors use the predicted secondary structure and some sequence-based descriptors to represent the predicted sequence [74, 165–169]. Moreover, several types of prediction algorithms, such as neural networks, support vector machines, quadratic discriminant functions, and random forests, were used for the prediction of the β hairpin motifs. The most recent predictor, STARPDB-beta hairpin [170], is based on a simple alignment into structurally annotated proteins collected from PDB.

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 [94]. More recent studies utilize hidden Markov models and a PSSM profile to represent the input sequence [70, 171–175]. Predictive quality of these tools was empirically evaluated in a recent comparative review [85] (*see* **Note 5**).

The initial study on the prediction of the α-turn-α motif was also based on scoring similarity between the predicted sequence and the α-turn-α structure library [176]. Subsequently, the method uses a pattern dictionary developed from known α-turn-α structures [177, 178], where predictions are made directly from pattern similarity [177] or using a classifier over pattern occurrences [178].

Table 2 summarizes 17 supersecondary structure prediction methods, including 7 β hairpin predictors (in chronological order) (method by de la Cruz et al. [73], BhairPred [74], methods by Hu et al. [168], Zou et al. [167], Xia et al. [166], and Jia et al. [165], and the STARPDB-beta hairpin method [170]); 7 coiled coil predictors (MultiCoil2 [179, 180], MARCOIL [175], PCOILS [174], bCIPA [173], Paircoil2 [172], CCHMM_PROF [171], and SpiriCoil [70]); and 3 α-turn-α predictors (method by Dodd and Egan [176], GYM [177], and method by Xiong et al. [178]). Older coiled coil predictors were reviewed in [84].

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.

3.2.1 BhairPred

The BhairPred predictor was developed by Raghava group at the Institute of Microbial Technology, India, in 2005 [74]. 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 PSPPRED, and solvent accessibility predicted with the NETASA method [181]. BhairPred was shown to provide predictions with accuracy in the 71–78% range on two independent test sets [74].

Table 2
Summary of the recent sequence-based predictors of supersecondary structure

Supersecondary structure type	Name (*authors*)	Year last published	Prediction model	Availability
β hairpin	STARPDB-beta hairpin	2016	Sequence similarity	WS
	Jia et al.	2011	Random forest	NA
	Xia et al.	2010	Support vector machine	NA
	Zou et al.	2009	Increment of diversity + quadratic discriminant analysis	NA
	Hu et al.	2008	Support vector machine	NA
	BhairPred	2005	Support vector machine	WS
	de la Cruz et al.	2002	Neural network	NA
Coiled coil	MultiCoil2	2011	Markov random field	WS + SP
	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
α-turn-α	*Xiong et al.*	2009	Support vector machine	NA
	GYM	2002	Statistical method	WS
	Dodd et al.	1990	Similarity scoring	NA

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. NA denotes that neither SP nor WS 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

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

Architecture: support vector machine.

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

3.2.2 MultiCoil2

The MultiCoil2 prediction method [179] extends the MultiCoil method [180], which in turn is an extension of the Paircoil method [182]. The Paircoil method is based on the pairwise residue statistics of positions within the heptapeptide coiled coil repeat, calculated from a set of known coiled coils. This idea is based on the structural constraints of coiled coil packing and the compensatory nature of neighboring positions with the α-helix. MultiCoil extends this idea to simultaneously predicting two- or three-way coiled coils by employing a separate set of statistics for each. MultiCoil2 improves this approach by replacing the simple scoring scheme of previous methods with a Markov random field. This improvement yields a substantial improvement over both Paircoil and Multicoil when trained on the same dataset [179]; MultiCoil2 correctly identifies nearly 92% of coiled coil residues with only 0.3% of non-coiled coils incorrectly identified, according to the authors' assessment [179] (*see* **Note 6**).

Inputs: protein sequence.

Architecture: Markov random field.

Availability: http://cb.csail.mit.edu/cb/multicoil2/cgi-bin/multicoil2.cgi.

3.2.3 GYM

The GYM prediction method is based on mining patterns from known helix-turn-helix examples and matching those patterns from novel sequences [177]. Patterns are defined as two or more residues occurring at the same positions within different helix-turn-helix examples. These patterns are discovered with a novel algorithm which ensures that they are maximal (i.e., not a portion of another pattern) and occur with a minimum defined frequency. When matching patterns to novel sequences, the GYM2 method uses the BLOSUM62 matrix to weight patterns by similarity, rather than strict matching.

Inputs: protein sequence.

Architecture: motif scoring.

Availability: http://users.cis.fiu.edu/~giri/bioinf/GYM2/prog.html.

3.3 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 7**). Two SSS predictors, BhairPred [74] and the method developed by Thornton group [73], have utilized the predicted secondary structure for the identification of β hairpins. Following, we discuss the latter method to demonstrate how the predicted secondary structure is used for the prediction of the SSS. This method consists of five steps:

Step 1. Predict secondary structure for a given input sequence using the PHD method [62].

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.

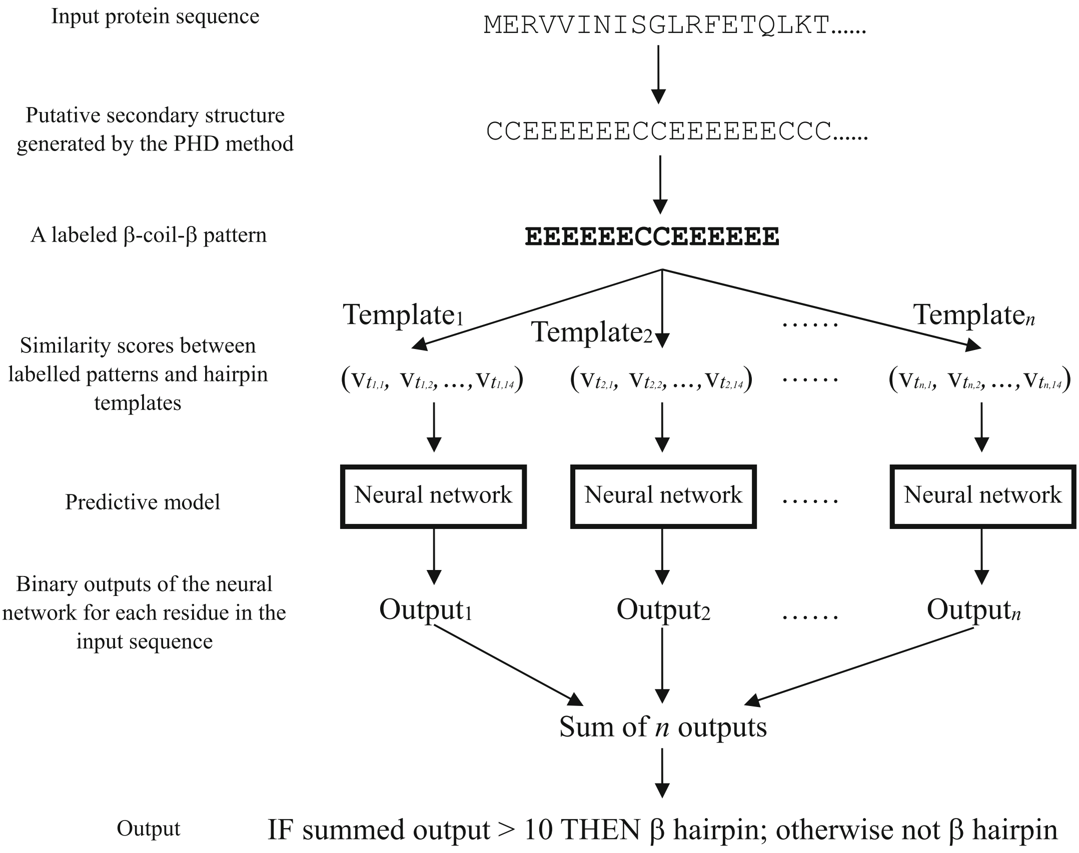
Input protein sequence
MERVVINISGLRFETQLKT......
Putative secondary structure generated by the PHD method
CCEEEEEEECCEEEEEEECCC......
A labeled β-coil-β pattern
EEEEEECCEEEEEE
Template1
Template2
......
Templaten
Similarity scores between labelled patterns and hairpin templates
$(v_{t_{1,1}}, v_{t_{1,2}}, \ldots, v_{t_{1,14}})$
$(v_{t_{2,1}}, v_{t_{2,2}}, \ldots, v_{t_{2,14}})$
......
$(v_{t_{n,1}}, v_{t_{n,2}}, \ldots, v_{t_{n,14}})$
Predictive model
Neural network
Neural network
......
Neural network
Binary outputs of the neural network for each residue in the input sequence
Output1
Output2
......
Outputn
Sum of n outputs
Output
IF summed output > 10 THEN β hairpin; otherwise not β hairpin

Fig. 4 The architecture of the β hairpin predictor proposed by the Thornton group. The prediction concerns the input (query) sequence, which is a fragment of chain A of the AF1521 protein shown in Fig. 1

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 one for more than ten template hairpins.

The working of the de la Cruz et al. method developed by the Thornton group [73] is illustrated in Fig. 4.

4 Notes

1. The predictive quality of the secondary structure predictors was empirically compared in several large-scale, worldwide initiatives including CASP [115], Critical Assessment of Fully

Automated Structure Prediction (CAFASP) [183], and EVA [116, 164]. 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 13 predictors. However, EVA was last updated in 2008.

2. A large-scale comparative analysis [23] has revealed a number of interesting and practical observations concerning structure prediction. The accuracy of the three-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. A recent comparative analysis [24] finds that accuracy has climbed to about 84%. Further, they find that errors most commonly confuse helices and coils and strands and coils. Prediction errors for helices and strands are most frequent at the ends of elements, whereas coils show little location bias in errors. Errors are also elevated for residues with many long-range contacts, relative to residues with few long-range contacts.
3. As shown in [23], the current secondary structure predictors are characterized by several drawbacks, which motivate further research in this area. Depending on the predictor, they confuse between 1 and 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.
4. 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 May 2018) count of citations in the ISI Web of Knowledge to the paper that describes the original PSIPRED algorithm [60] is close to 3187, which demonstrates the broad usage of this method.
5. A recent assessment evaluated all coiled coil predictors listed in Table 2 [85]. In general, MultiCoil2 and CCHMM_PROF were found to have the highest prediction performance. Due to its architecture, MultiCoil2 cannot detect coiled coils shorter than 21 amino acids, and on a length restricted set,

MultiCoil2 has the best performance. However, on an unrestricted set, CCHMM_PROF has the best performance.

6. Prediction of the supersecondary structures could be potentially improved by utilizing a consensus of different approaches. As shown in a comparative analysis of coiled coil predictors [84], 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. However, another comparative study highlights some problem cases for consensus prediction and instead calls for further predictor development [85].
7. The major obstacle to utilize the predicted secondary structure in the prediction of the supersecondary structures, which was observed in the 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 [73]. The use of the native rather than the predicted secondary structure was shown to lead to a significant improvement in the prediction of the supersecondary structures [74].

Acknowledgments

This work was supported by the Qimonda Endowment funds to L.K.

References

1. Pauling L, Corey RB (1951) The pleated sheet, a new layer configuration of polypeptide chains. Proc Natl Acad Sci 37 (5):251–256
2. Pauling L, Corey RB, Branson HR (1951) The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. Proc Natl Acad Sci 37(4):205–211
3. Anfinsen CB (1973) Principles that govern the folding of protein chains. Science 181 (4096):223–230
4. Berman HM (2000) The protein data bank. Nucleic Acids Res 28(1):235–242
5. Burley SK, Berman HM, Kleywegt GJ, Markley JL, Nakamura H, Velankar S (2017) Protein data bank (PDB): the single global macromolecular structure archive. Methods Mol Biol 1607:627–641
6. Pruitt KD, Tatusova T, Klimke W, Maglott DR (2009) NCBI Reference Sequences: current status, policy and new initiatives. Nucleic Acids Res 37(Database):D32–D36
7. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH, Rausch D,

Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, Tatusova T, DiCuccio M, Kitts P, Murphy TD, Pruitt KD (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 44(D1):D733–D745

8. Gronwald W, Kalbitzer HR (2010) Automated protein NMR structure determination in solution, Methods in molecular biology. Humana Press, Totowa
9. Chayen NE (2009) High-throughput protein crystallization. Adv Protein Chem Struct Biol 77:1–22
10. Zhang Y (2009) Protein structure prediction: when is it useful? Curr Opin Struct Biol 19 (2):145–155
11. Ginalski K (2006) Comparative modeling for protein structure prediction. Curr Opin Struct Biol 16(2):172–177
12. Mizianty MJ, Fan X, Yan J, Chalmers E, Woloschuk C, Joachimiak A, Kurgan L (2014) Covering complete proteomes with X-ray structures: a current snapshot. Acta Crystallogr D Biol Crystallogr 70 (Pt 11):2781–2793
13. Gao J, Wu Z, Hu G, Wang K, Song J, Joachimiak A, Kurgan L (2018) Survey of predictors of propensity for protein production and crystallization with application to predict resolution of crystal structures. Curr Protein Pept Sci 19(2):200–210
14. Grabowski M, Niedzialkowska E, Zimmerman MD, Minor W (2016) The impact of structural genomics: the first quindecennial. J Struct Funct Genom 17(1):1–16
15. Yang Y, Faraggi E, Zhao H, Zhou Y (2011) Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates. Bioinformatics 27 (15):2076–2082
16. Roy A, Kucukural A, Zhang Y (2010) I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc 5(4):725–738
17. Faraggi E, Yang Y, Zhang S, Zhou Y (2009) Predicting continuous local structure and the effect of its substitution for secondary structure in fragment-free protein structure prediction. Structure 17(11):1515–1527
18. Wu S, Zhang Y (2008) MUSTER: improving protein sequence profile-profile alignments by using multiple sources of structure information. Proteins 72(2):547–556
19. Zhou H, Skolnick J (2007) Ab initio protein structure prediction using chunk-TASSER. Biophys J 93(5):1510–1518
20. Skolnick J (2006) In quest of an empirical potential for protein structure prediction. Curr Opin Struct Biol 16(2):166–171
21. Zhang W, Yang J, He B, Walker SE, Zhang H, Govindarajoo B, Virtanen J, Xue Z, Shen HB, Zhang Y (2016) Integration of QUARK and I-TASSER for ab initio protein structure prediction in CASP11. Proteins 84(Suppl 1):76–86
22. Czaplewski C, Karczynska A, Sieradzan AK, Liwo A (2018) UNRES server for physics-based coarse-grained simulations and prediction of protein structure, dynamics and thermodynamics. Nucleic Acids Res 46(W1): W304–W309
23. Zhang H, Zhang T, Chen K, Kedarisetti KD, Mizianty MJ, Bao Q, Stach W, Kurgan L (2011) Critical assessment of high-throughput standalone methods for secondary structure prediction. Brief Bioinform 12 (6):672–688
24. Yang Y, Gao J, Wang J, Heffernan R, Hanson J, Paliwal K, Zhou Y (2018) Sixty-five years of the long march in protein secondary structure prediction: the final stretch? Brief Bioinform 19(3):482–494
25. Pei J, Grishin NV (2007) PROMALS: towards accurate multiple sequence alignments of distantly related proteins. Bioinformatics 23(7):802–808
26. Mizianty MJ, Kurgan L (2011) Sequence-based prediction of protein crystallization, purification and production propensity. Bioinformatics 27(13):i24–i33
27. Slabinski L, Jaroszewski L, Rychlewski L, Wilson IA, Lesley SA, Godzik A (2007) XtalPred: a web server for prediction of protein crystallizability. Bioinformatics 23(24):3403–3405
28. Wang H, Feng L, Webb GI, Kurgan L, Song J, Lin D (2017) Critical evaluation of bioinformatics tools for the prediction of protein crystallization propensity. Brief Bioinform. https://doi.org/10.1093/bib/bbx1018
29. Zhang T, Zhang H, Chen K, Ruan J, Shen S, Kurgan L (2010) Analysis and prediction of RNA-binding residues using sequence, evolutionary conservation, and predicted secondary structure and solvent accessibility. Curr Protein Pept Sci 11(7):609–628
30. Yan J, Kurgan L (2017) DRNApred, fast sequence-based method that accurately

predicts and discriminates DNA- and RNA-binding residues. Nucleic Acids Res 45 (10):e84

31. Yan J, Friedrich S, Kurgan L (2016) A comprehensive comparative review of sequence-based predictors of DNA- and RNA-binding residues. Brief Bioinform 17(1):88–105
32. Peng Z, Kurgan L (2015) High-throughput prediction of RNA, DNA and protein binding regions mediated by intrinsic disorder. Nucleic Acids Res 43(18):e121
33. Pulim V, Bienkowska J, Berger B (2008) LTHREADER: prediction of extracellular ligand-receptor interactions in cytokines using localized threading. Protein Sci 17 (2):279–292
34. Fischer JD, Mayer CE, Söding J (2008) Prediction of protein functional residues from sequence by probability density estimation. Bioinformatics 24(5):613–620
35. Chen K, Mizianty MJ, Kurgan L (2012) Prediction and analysis of nucleotide-binding residues using sequence and sequence-derived structural descriptors. Bioinformatics 28 (3):331–341
36. Song J, Tan H, Mahmood K, Law RHP, Buckle AM, Webb GI, Akutsu T, Whisstock JC (2009) Prodepth: predict residue depth by support vector regression approach from protein sequences only. PLoS One 4(9):e7072
37. Zhang H, Zhang T, Chen K, Shen S, Ruan J, Kurgan L (2008) Sequence based residue depth prediction using evolutionary information and predicted secondary structure. BMC Bioinformatics 9(1):388
38. Zheng C, Kurgan L (2008) Prediction of beta-turns at over 80% accuracy based on an ensemble of predicted secondary structures and multiple alignments. BMC Bioinformatics 9:430
39. Mizianty MJ, Kurgan L (2009) Modular prediction of protein structural classes from sequences of twilight-zone identity with predicting sequences. BMC Bioinformatics 10 (1):414
40. 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 (1):226
41. Chen K, Kurgan L (2007) PFRES: protein fold classification by using evolutionary information and predicted secondary structure. Bioinformatics 23(21):2843–2850
42. Kong L, Zhang L (2014) Novel structure-driven features for accurate prediction of protein structural class. Genomics 103 (4):292–297
43. Kurgan LA, Zhang T, Zhang H, Shen S, Ruan J (2008) Secondary structure-based assignment of the protein structural classes. Amino Acids 35(3):551–564
44. Xue B, Faraggi E, Zhou Y (2009) Predicting residue-residue contact maps by a two-layer, integrated neural-network method. Proteins 76(1):176–183
45. Cheng J, Baldi P (2007) Improved residue contact prediction using support vector machines and a large feature set. BMC Bioinformatics 8(1):113
46. Mizianty MJ, Stach W, Chen K, Kedarisetti KD, Disfani FM, Kurgan L (2010) Improved sequence-based prediction of disordered regions with multilayer fusion of multiple information sources. Bioinformatics 26(18): i489–i496
47. Mizianty MJ, Zhang T, Xue B, Zhou Y, Dunker A, Uversky VN, Kurgan L (2011) In-silico prediction of disorder content using hybrid sequence representation. BMC Bioinformatics 12(1):245
48. Schlessinger A, Punta M, Yachdav G, Kajan L, Rost B (2009) Improved disorder prediction by combination of orthogonal approaches. PLoS One 4(2):e4433
49. Mizianty MJ, Peng ZL, Kurgan L (2013) MFDp2: Accurate predictor of disorder in proteins by fusion of disorder probabilities, content and profiles. Intrinsically Disord Proteins 1(1):e24428
50. Mizianty MJ, Uversky V, Kurgan L (2014) Prediction of intrinsic disorder in proteins using MFDp2. Methods Mol Biol 1137:147–162
51. Walsh I, Martin AJ, Di Domenico T, Vullo A, Pollastri G, Tosatto SC (2011) CSpritz: accurate prediction of protein disorder segments with annotation for homology, secondary structure and linear motifs. Nucleic Acids Res 39(Web Server issue):W190–W196
52. Meng F, Kurgan L (2016) DFLpred: high-throughput prediction of disordered flexible linker regions in protein sequences. Bioinformatics 32(12):i341–i350
53. Yan J, Dunker AK, Uversky VN, Kurgan L (2016) Molecular recognition features (MoRFs) in three domains of life. Mol BioSyst 12(3):697–710
54. Sharma R, Raicar G, Tsunoda T, Patil A, Sharma A (2018) OPAL: prediction of MoRF regions in intrinsically disordered protein sequences. Bioinformatics 34 (11):1850–1858

55. Zhang H, Zhang T, Gao J, Ruan J, Shen S, Kurgan L (2010) Determination of protein folding kinetic types using sequence and predicted secondary structure and solvent accessibility. Amino Acids 42(1):271–283
56. Gao J, Zhang T, Zhang H, Shen S, Ruan J, Kurgan L (2010) Accurate prediction of protein folding rates from sequence and sequence-derived residue flexibility and solvent accessibility. Proteins 78(9):2114–2130
57. Jiang Y, Iglinski P, Kurgan L (2009) Prediction of protein folding rates from primary sequences using hybrid sequence representation. J Comput Chem 30(5):772–783
58. Bryson K, McGuffin LJ, Marsden RL, Ward JJ, Sodhi JS, Jones DT (2005) Protein structure prediction servers at University College London. Nucleic Acids Res 33(Web Server): W36–W38
59. Kurgan L, Miri Disfani F (2011) Structural protein descriptors in 1-dimension and their sequence-based predictions. Curr Protein Pept Sci 12(6):470–489
60. Jones DT (1999) Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 292(2):195–202
61. Buchan DWA, Ward SM, Lobley AE, Nugent TCO, Bryson K, Jones DT (2010) Protein annotation and modelling servers at University College London. Nucleic Acids Res 38 (Web Server):W563–W568
62. Rost B (1996) PHD: predicting one-dimensional protein structure by profile-based neural networks. Methods Enzymol 266:525–539
63. Rost B, Yachdav G, Liu J (2004) The PredictProtein server. Nucleic Acids Res 32(Web Server):W321–W326
64. O'Donnell CW, Waldispühl J, Lis M, Halfmann R, Devadas S, Lindquist S, Berger B (2011) A method for probing the mutational landscape of amyloid structure. Bioinformatics 27(13):i34–i42
65. Bryan AW, Menke M, Cowen LJ, Lindquist SL, Berger B (2009) BETASCAN: probable β-amyloids identified by pairwise probabilistic analysis. PLoS Comput Biol 5(3):e1000333
66. Bradley P, Cowen L, Menke M, King J, Berger B (2001) BETAWRAP: successful prediction of parallel β-helices from primary sequence reveals an association with many microbial pathogens. Proc Natl Acad Sci 98 (26):14819–14824
67. Hornung T, Volkov OA, Zaida TMA, Delannoy S, Wise JG, Vogel PD (2008) Structure of the cytosolic part of the subunit b-dimer of Escherichia coli F0F1-ATP synthase. Biophys J 94(12):5053–5064
68. Sun ZR, Cui Y, Ling LJ, Guo Q, Chen RS (1998) Molecular dynamics simulation of protein folding with supersecondary structure constraints. J Protein Chem 17(8):765–769
69. Szappanos B, Süveges D, Nyitray L, Perczel A, Gáspári Z (2010) Folded-unfolded cross-predictions and protein evolution: the case study of coiled-coils. FEBS Lett 584 (8):1623–1627
70. Rackham OJL, Madera M, Armstrong CT, Vincent TL, Woolfson DN, Gough J (2010) The evolution and structure prediction of coiled coils across all genomes. J Mol Biol 403(3):480–493
71. Gerstein M, Hegyi H (1998) Comparing genomes in terms of protein structure: surveys of a finite parts list. FEMS Microbiol Rev 22 (4):277–304
72. Reddy CCS, Shameer K, Offmann BO, Sowdhamini R (2008) PURE: a webserver for the prediction of domains in unassigned regions in proteins. BMC Bioinformatics 9(1):281
73. de la Cruz X, Hutchinson EG, Shepherd A, Thornton JM (2002) Toward predicting protein topology: an approach to identifying β hairpins. Proc Natl Acad Sci 99 (17):11157–11162
74. Kumar M, Bhasin M, Natt NK, Raghava GPS (2005) BhairPred: prediction of β-hairpins in a protein from multiple alignment information using ANN and SVM techniques. Nucleic Acids Res 33(Web Server): W154–W159
75. Barton GJ (1995) Protein secondary structure prediction. Curr Opin Struct Biol 5 (3):372–376
76. Heringa J (2000) Computational methods for protein secondary structure prediction using multiple sequence alignments. Curr Protein Pept Sci 1(3):273–301
77. Rost B (2001) Protein secondary structure prediction continues to rise. J Struct Biol 134(2–3):204–218
78. Albrecht M, Tosatto SCE, Lengauer T, Valle G (2003) Simple consensus procedures are effective and sufficient in secondary structure prediction. Protein Eng Des Sel 16 (7):459–462
79. Yan J, Marcus M, Kurgan L (2014) Comprehensively designed consensus of standalone secondary structure predictors improves Q3 by over 3%. J Biomol Struct Dyn 32(1):36–51
80. Rost B (2009) Prediction of protein structure in 1D—secondary structure, membrane

regions, and solvent accessibility. Structural bioinformatics, 2nd edn. Wiley, New York

81. Pirovano W, Heringa J (2010) Protein secondary structure prediction. Methods Mol Biol 609:327–348
82. Meng F, Kurgan L (2016) Computational prediction of protein secondary structure from sequence. Curr Protoc Protein Sci 86:2.3.1–2.3.10
83. Singh M (2006) Predicting protein secondary and supersecondary structure, Chapman & Hall/CRC Computer & Information Science Series. Chapman and Hall/CRC, New York
84. Gruber M, Söding J, Lupas AN (2006) Comparative analysis of coiled-coil prediction methods. J Struct Biol 155(2):140–145
85. Li C, Ching Han Chang C, Nagel J, Porebski BT, Hayashida M, Akutsu T, Song J, Buckle AM (2016) Critical evaluation of in silico methods for prediction of coiled-coil domains in proteins. Brief Bioinform 17(2):270–282
86. Ho HK, Zhang L, Ramamohanarao K, Martin S (2013) A survey of machine learning methods for secondary and supersecondary protein structure prediction. Methods Mol Biol 932:87–106
87. Chen K, Kurgan L (2013) Computational prediction of secondary and supersecondary structures. Methods Mol Biol 932:63–86
88. Kolodny R, Honig B (2006) VISTAL—a new 2D visualization tool of protein 3D structural alignments. Bioinformatics 22 (17):2166–2167
89. Moreland JL, Gramada A, Buzko OV, Zhang Q, Bourne PE (2005) BMC Bioinformatics 6(1):21
90. Porollo AA, Adamczak R, Meller J (2004) POLYVIEW: a flexible visualization tool for structural and functional annotations of proteins. Bioinformatics 20(15):2460–2462
91. Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: A structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247(4):536–540
92. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM (1997) CATH—a hierarchic classification of protein domain structures. Structure 5 (8):1093–1109
93. Andreeva A, Howorth D, Chandonia JM, Brenner SE, Hubbard TJP, Chothia C, Murzin AG (2007) Data growth and its impact on the SCOP database: new developments. Nucleic Acids Res 36(Database):D419–D425
94. Cuff AL, Sillitoe I, Lewis T, Clegg AB, Rentzsch R, Furnham N, Pellegrini-Calace M, Jones D, Thornton J, Orengo CA (2011) Extending CATH: increasing coverage of the protein structure universe and linking structure with function. Nucleic Acids Res 39 (Database):D420–D426
95. Sillitoe I, Dawson N, Thornton J, Orengo C (2015) The history of the CATH structural classification of protein domains. Biochimie 119:209–217
96. Andreeva A, Howorth D, Chandonia JM, Brenner SE, Hubbard TJ, Chothia C, Murzin AG (2008) Data growth and its impact on the SCOP database: new developments. Nucleic Acids Res 36(Database issue):D419–D425
97. Levitt M, Greer J (1977) Automatic identification of secondary structure in globular proteins. J Mol Biol 114(2):181–239
98. Kabsch W, Sander C (1983) Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22(12):2577–2637
99. Richards FM, Kundrot CE (1988) Identification of structural motifs from protein coordinate data: secondary structure and first-level supersecondary structure. Proteins Struct Funct Genet 3(2):71–84
100. Sklenar H, Etchebest C, Lavery R (1989) Describing protein structure: a general algorithm yielding complete helicoidal parameters and a unique overall axis. Proteins Struct Funct Genet 6(1):46–60
101. Frishman D, Argos P (1995) Knowledge-based protein secondary structure assignment. Proteins Struct Funct Genet 23 (4):566–579
102. Labesse G, Colloc'h N, Pothier J, Mornon JP (1997) P-SEA: a new efficient assignment of secondary structure from Cα trace of proteins. Bioinformatics 13(3):291–295
103. King SM, Johnson WC (1999) Assigning secondary structure from protein coordinate data. Proteins Struct Funct Genet 35 (3):313–320
104. Fodje MN, Al-Karadaghi S (2002) Occurrence, conformational features and amino acid propensities for the π-helix. Protein Eng Des Sel 15(5):353–358
105. Martin J, Letellier G, Marin A, Taly J-F, de Brevern AG, Gibrat J-F (2005) BMC Struct Biol 5(1):17
106. Cubellis M, Cailliez F, Lovell SC (2005) Secondary structure assignment that accurately reflects physical and evolutionary characteristics. BMC Bioinformatics 6(Suppl 4):S8
107. Majumdar I, Krishna SS, Grishin NV (2005) PALSSE: a program to delineate linear

secondary structural elements from protein structures. BMC Bioinformatics 6(1):202

108. Zhang W, Dunker AK, Zhou Y (2008) Assessing secondary structure assignment of protein structures by using pairwise sequence-alignment benchmarks. Proteins 71(1):61–67

109. Hosseini S-R, Sadeghi M, Pezeshk H, Eslahchi C, Habibi M (2008) PROSIGN: A method for protein secondary structure assignment based on three-dimensional coordinates of consecutive C^{α} atoms. Comput Biol Chem 32(6):406–411

110. Park S-Y, Yoo M-J, Shin J-M, Cho K-H (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(2):118–122

111. Zacharias J, Knapp EW (2014) Protein secondary structure classification revisited: processing DSSP information with PSSC. J Chem Inf Model 54(7):2166–2179

112. Law SM, Frank AT, Brooks CL 3rd (2014) PCASSO: a fast and efficient Calpha-based method for accurately assigning protein secondary structure elements. J Comput Chem 35(24):1757–1761

113. Cao C, Wang GS, Liu A, Xu ST, Wang LC, Zou SX (2016) A new secondary structure assignment algorithm using C-alpha backbone fragments. Int J Mol Sci 17(3):333

114. Klose DP, Wallace BA, Janes RW (2010) 2Struc: the secondary structure server. Bioinformatics 26(20):2624–2625

115. Moult J, Pedersen JT, Judson R, Fidelis K (1995) A large-scale experiment to assess protein structure prediction methods. Proteins Struct Funct Genet 23(3):ii–iv

116. Koh IYY (2003) EVA: evaluation of protein structure prediction servers. Nucleic Acids Res 31(13):3311–3315

117. Parry DAD, Fraser RDB, Squire JM (2008) Fifty years of coiled-coils and α-helical bundles: a close relationship between sequence and structure. J Struct Biol 163(3):258–269

118. Truebestein L, Leonard TA (2016) Coiled-coils: the long and short of it. BioEssays 38 (9):903–916

119. Pellegrini-Calace M (2005) Detecting DNA-binding helix-turn-helix structural motifs using sequence and structure information. Nucleic Acids Res 33(7):2129–2140

120. Aravind L, Anantharaman V, Balaji S, Babu MM, Iyer LM (2005) The many faces of the helix-turn-helix domain: transcription regulation and beyond. FEMS Microbiol Rev 29 (2):231–262

121. Hutchinson EG, Thornton JM (1996) PROMOTIF-A program to identify and analyze structural motifs in proteins. Protein Sci 5(2):212–220

122. Walshaw J, Woolfson DN (2001) SOCKET: a program for identifying and analysing coiled-coil motifs within protein structures. J Mol Biol 307(5):1427–1450

123. Testa OD, Moutevelis E, Woolfson DN (2009) CC+: a relational database of coiled-coil structures. Nucleic Acids Res 37(Database):D315–D322

124. Michalopoulos I (2004) TOPS: an enhanced database of protein structural topology. Nucleic Acids Res 32(90001):D251–D254

125. Rost B, Sander C (1993) Improved prediction of protein secondary structure by use of sequence profiles and neural networks. Proc Natl Acad Sci 90(16):7558–7562

126. Altschul S (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25(17):3389–3402

127. Fang C, Shang Y, Xu D (2018) MUFOLD-SS: new deep inception-inside-inception networks for protein secondary structure prediction. Proteins 86(5):592–598

128. Heffernan R, Yang Y, Paliwal K, Zhou Y (2017) Capturing non-local interactions by long short-term memory bidirectional recurrent neural networks for improving prediction of protein secondary structure, backbone angles, contact numbers and solvent accessibility. Bioinformatics 33(18):2842–2849

129. Wang S, Li W, Liu S, Xu J (2016) RaptorX-Property: a web server for protein structure property prediction. Nucleic Acids Res 44 (W1):W430–W435

130. Wang S, Peng J, Ma J, Xu J (2016) Protein secondary structure prediction using deep convolutional neural fields. Sci Rep 6:18962

131. Cole C, Barber JD, Barton GJ (2008) The Jpred 3 secondary structure prediction server. Nucleic Acids Res 36(Web Server): W197–W201

132. Cuff JA, Clamp ME, Siddiqui AS, Finlay M, Barton GJ (1998) JPred: a consensus secondary structure prediction server. Bioinformatics 14(10):892–893

133. Cuff JA, Barton GJ (2000) Application of multiple sequence alignment profiles to improve protein secondary structure prediction. Proteins Struct Funct Genet 40 (3):502–511

134. Drozdetskiy A, Cole C, Procter J, Barton GJ (2015) JPred4: a protein secondary structure

prediction server. Nucleic Acids Res 43(W1): W389–W394

135. Yaseen A, Li Y (2014) Context-based features enhance protein secondary structure prediction accuracy. J Chem Inf Model 54 (3):992–1002

136. Buchan DW, Minneci F, Nugent TC, Bryson K, Jones DT (2013) Scalable web services for the PSIPRED Protein Analysis Workbench. Nucleic Acids Res 41(Web Server issue):W349–W357

137. Pollastri G, Martin AJM, Mooney C, Vullo A (2007) Accurate prediction of protein secondary structure and solvent accessibility by consensus combiners of sequence and structure information. BMC Bioinformatics 8 (1):201

138. Pollastri G, McLysaght A (2005) Porter: a new, accurate server for protein secondary structure prediction. Bioinformatics 21 (8):1719–1720

139. Mirabello C, Pollastri G (2013) Porter, PaleAle 4.0: high-accuracy prediction of protein secondary structure and relative solvent accessibility. Bioinformatics 29 (16):2056–2058

140. Bettella F, Rasinski D, Knapp EW (2012) Protein secondary structure prediction with SPARROW. J Chem Inf Model 52 (2):545–556

141. Zhou T, Shu N, Hovmöller S (2010) A novel method for accurate one-dimensional protein structure prediction based on fragment matching. Bioinformatics 26(4):470–477

142. Kountouris P, Hirst JD (2009) Prediction of backbone dihedral angles and protein secondary structure using support vector machines. BMC Bioinformatics 10(1):437

143. Green JR, Korenberg MJ, Aboul-Magd MO (2009) PCI-SS: MISO dynamic nonlinear protein secondary structure prediction. BMC Bioinformatics 10:222–222

144. Montgomerie S, Cruz JA, Shrivastava S, Arndt D, Berjanskii M, Wishart DS (2008) PROTEUS2: a web server for comprehensive protein structure prediction and structure-based annotation. Nucleic Acids Res 36(Web Server):W202–W209

145. Montgomerie S, Sundararaj S, Gallin WJ, Wishart DS (2006) Improving the accuracy of protein secondary structure prediction using structural alignment. BMC Bioinformatics 7:301

146. Martin J, Gibrat JF, Rodolphe F (2006) Analysis of an optimal hidden Markov model for secondary structure prediction. BMC Struct Biol 6:25

147. Karypis G (2006) YASSPP: better kernels and coding schemes lead to improvements in protein secondary structure prediction. Proteins 64(3):575–586

148. Lin K, Simossis VA, Taylor WR, Heringa J (2005) A simple and fast secondary structure prediction method using hidden neural networks. Bioinformatics 21(2):152–159

149. Adamczak R, Porollo A, Meller J (2005) Combining prediction of secondary structure and solvent accessibility in proteins. Proteins 59(3):467–475

150. Cheng J, Randall AZ, Sweredoski MJ, Baldi P (2005) SCRATCH: a protein structure and structural feature prediction server. Nucleic Acids Res 33(Web Server):W72–W76

151. Pollastri G, Przybylski D, Rost B, Baldi P (2002) Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. Proteins Struct Funct Genet 47(2):228–235

152. Madera M, Calmus R, Thiltgen G, Karplus K, Gough J (2010) Improving protein secondary structure prediction using a simple k-mer model. Bioinformatics 26(5):596–602

153. Yang Y, Heffernan R, Paliwal K, Lyons J, Dehzangi A, Sharma A, Wang J, Sattar A, Zhou Y (2017) SPIDER2: a package to predict secondary structure, accessible surface area, and main-chain torsional angles by deep neural networks. Methods Mol Biol 1484:55–63

154. Heffernan R, Paliwal K, Lyons J, Dehzangi A, Sharma A, Wang J, Sattar A, Yang Y, Zhou Y (2015) Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning. Sci Rep 5:11476

155. Faraggi E, Zhang T, Yang Y, Kurgan L, Zhou Y (2012) SPINE X: improving protein secondary structure prediction by multistep learning coupled with prediction of solvent accessible surface area and backbone torsion angles. J Comput Chem 33(3):259–267

156. Schuster M, Paliwal KK (1997) Bidirectional recurrent neural networks. IEEE Trans Signal Process 45(11):2673–2681

157. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9 (8):1735–1780

158. Remmert M, Biegert A, Hauser A, Soding J (2011) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat Methods 9(2):173–175

159. Lupas A, Van Dyke M, Stock J (1991) Predicting coiled coils from protein sequences. Science 252(5009):1162–1164

160. Eddy SR (1998) Profile hidden Markov models. Bioinformatics 14(9):755–763

161. Baú D, Martin AJM, Mooney C, Vullo A, Walsh I, Pollastri G (2006) Distill: a suite of web servers for the prediction of one-, two- and three-dimensional structural features of proteins. BMC Bioinformatics 7(1):402

162. Mooney C, Pollastri G (2009) Beyond the Twilight Zone: automated prediction of structural properties of proteins by recursive neural networks and remote homology information. Proteins 77(1):181–190

163. Cuff JA, Barton GJ (2000) Application of multiple sequence alignment profiles to improve protein secondary structure prediction. Proteins 40(3):502–511

164. Eyrich VA, Marti-Renom MA, Przybylski D, Madhusudhan MS, Fiser A, Pazos F, Valencia A, Sali A, Rost B (2001) EVA: continuous automatic evaluation of protein structure prediction servers. Bioinformatics 17 (12):1242–1243

165. Jia S-C, Hu X-Z (2011) Using random forest algorithm to predict β-hairpin motifs. Protein Pept Lett 18(6):609–617

166. Xia J-F, Wu M, You Z-H, Zhao X-M, Li X-L (2010) Prediction of β-hairpins in proteins using physicochemical properties and structure information. Protein Pept Lett 17 (9):1123–1128

167. Zou D, He Z, He J (2009) β-Hairpin prediction with quadratic discriminant analysis using diversity measure. J Comput Chem 30 (14):2277–2284

168. Hu XZ, Li QZ (2008) Prediction of the β-hairpins in proteins using support vector machine. Protein J 27(2):115–122

169. Kuhn M, Meiler J, Baker D (2004) Strand-loop-strand motifs: Prediction of hairpins and diverging turns in proteins. Proteins 54 (2):282–288

170. Singh H, Raghava GPS (2016) BLAST-based structural annotation of protein residues using Protein Data Bank. Biol Direct 11:4

171. Bartoli L, Fariselli P, Krogh A, Casadio R (2009) CCHMM_PROF: a HMM-based coiled-coil predictor with evolutionary information. Bioinformatics 25(21):2757–2763

172. McDonnell AV, Jiang T, Keating AE, Berger B (2006) Paircoil2: improved prediction of coiled coils from sequence. Bioinformatics 22(3):356–358

173. Mason JM, Schmitz MA, Muller KM, Arndt KM (2006) Semirational design of Jun-Fos coiled coils with increased affinity: universal implications for leucine zipper prediction and design. Proc Natl Acad Sci 103 (24):8989–8994

174. Gruber M, Soding J, Lupas AN (2005) REPPER—repeats and their periodicities in fibrous proteins. Nucleic Acids Res 33(Web Server):W239–W243

175. Delorenzi M, Speed T (2002) An HMM model for coiled-coil domains and a comparison with PSSM-based predictions. Bioinformatics 18(4):617–625

176. Dodd IB, Egan JB (1990) Improved detection of helix-turn-helix DNA-binding motifs in protein sequences. Nucleic Acids Res 18 (17):5019–5026

177. Narasimhan G, Bu C, Gao Y, Wang X, Xu N, Mathee K (2002) Mining protein sequences for motifs. J Comput Biol 9(5):707–720

178. Xiong W, Li T, Chen K, Tang K (2009) Local combinational variables: an approach used in DNA-binding helix-turn-helix motif prediction with sequence information. Nucleic Acids Res 37(17):5632–5640

179. Trigg J, Gutwin K, Keating AE, Berger B (2011) Multicoil2: predicting coiled coils and their oligomerization states from sequence in the twilight zone. PLoS One 6 (8):e23519

180. Wolf E, Kim PS, Berger B (1997) MultiCoil: a program for predicting two-and three-stranded coiled coils. Protein Sci 6 (6):1179–1189

181. Ahmad S, Gromiha MM (2002) NETASA: neural network based prediction of solvent accessibility. Bioinformatics 18(6):819–824

182. Berger B, Wilson DB, Wolf E, Tonchev T, Milla M, Kim PS (1995) Predicting coiled coils by use of pairwise residue correlations. Proc Natl Acad Sci U S A 92(18):8259–8263

183. Fischer D, Barret C, Bryson K, Elofsson A, Godzik A, Jones D, Karplus KJ, Kelley LA, MacCallum RM, Pawowski K, Rost B, Rychlewski L, Sternberg M (1999) CAFASP-1: critical assessment of fully automated structure prediction methods. Proteins Suppl 3:209–217