Secondary Structure Prediction

Jonas Reeb and Burkhard Rost, Technical University of Munich (TUM), Garching/Munich, Germany

Introduction

The study of proteins is fuelled by the desire to understand their function. The molecular understanding of function requires the understanding of structure. However, determining the three-dimensional (3D) structure of proteins is a highly non-trivial process. Despite substantial advances in technologies, the experimental determination of a single protein is still costly and might take years. Conversely, the advent of next generation sequencing techniques has tremendously increased the amount of raw sequence data about which there is very little structural knowledge (Goodwin *et al.*, 2016; Martinez and Nelson, 2010). Due to ever decreasing costs, the speed at which this sequencing data is being generated far outpaces the annotation of the data with structural (or functional) knowledge. This creates a divide that has been referred to as the sequence-structure gap. For instance, the UniProtKB database currently contains almost 60 million proteins which are at most 90% identical (The Uniprot Consortium, 2016; Suzek *et al.*, 2015). In contrast, the Protein Data Bank (PDB), the major database of protein 3D structures, contains just 48,000 protein structures at that similarity cutoff (Berman *et al.*, 2000).

The prediction of the protein secondary structure is one step toward bridging this gap. Due to its simplicity and elementary importance, secondary structure is also one of the earliest features that has been predicted from sequence and has grown over many decades into a mature field where predictions can be performed at relatively little computational cost, yet at a high performance.

If the 3D structure could be predicted directly, then one would also have an implicit prediction of secondary structure. However, despite crucial breakthroughs in the last 5–10 years, the number of proteins for which 3D structure can be predicted from sequence remains limited (Kinch *et al.*, 2016; Hopf *et al.*, 2012; Moult *et al.*, 2016; Marks *et al.*, 2012). Even for most proteins for which comparative modelling can infer 3D structures (Biasini *et al.*, 2014; Yang *et al.*, 2011), dedicated secondary structure prediction methods still tend to perform better (Eyrich *et al.*, 2001). Additionally, the computational costs of more advanced methods are substantial which is in contrast to the decrease in resources needed for generating novel sequence (Muir *et al.*, 2016). Hence, both effective and efficient annotation of this sequencing data is necessary.

On top of the knowledge about the structural elements, secondary structure predictions also serve as input to predict more complex protein features such as disordered regions, interaction sites or aspects of protein function. Solvent accessibility is another feature of protein structure and often predicted by many of the tools for secondary structure prediction.

This contribution succinctly summarizes the underlying biological concepts of protein secondary structure and continues discussing some of the main concepts in the field of secondary structure prediction before presenting a selection of such methods in more detail. Finally, the level at which state-of-the-art methods perform is examined.

Background

Secondary Structure Elements in Protein Folding

Protein 3D structure is encoded by its amino acid sequence which is sometimes misleadingly referred to as the "primary structure". The order in which the 20 biogenic amino acids are arranged uniquely defines the structure and determines its folding in 3D (Anfinsen, 1973). Amino acids are linked by a covalent, so-called peptide bond which forms the backbone of the protein. Given the chemical properties of the amino acids and the physical constraints of this bond, the number of possible protein conformations is immense. The forces that drive the folding process that strives for a stable energetic minimum, are non-covalent interactions such as electrostatic interactions, much weaker but more common van der Waals interactions, or the hydrophobic effect, i.e., non-polar components aggregating in water (Kessel and Ben-Tal, 2011).

For the formation of secondary structure elements, a specific type of electrostatic interactions, the hydrogen bonds, are often claimed as the main stabilizing force. While this is most likely not the case, they still give structure to the elements in question and are important to overall protein stability (Kessel and Ben-Tal, 2011). Hydrogen bonds form between two dipoles which act as donor and acceptor, such as two atoms of amino acid side chains. However, in the case of secondary structure elements, the relevant atoms are part of the peptide bonds. Here, a hydrogen atom that is covalently bound to a nitrogen, gets in proximity of, and is therefore attracted by, an electronegative oxygen (**Fig. 1**). The stable secondary structure elements then further develop interactions between themselves and form a tertiary structure.

Various types of secondary structure are observed repeatedly out of which helices and strands are the most prominent ones. Both are characterized by specific patterns of hydrogen bonds. Due to these patterns, those two types are also often referred to as "regular secondary structure". Most prediction methods target three "states" using those two regular types and merging everything else into the state "other".

doi:10.1016/B978-0-12-809633-8.20267-7

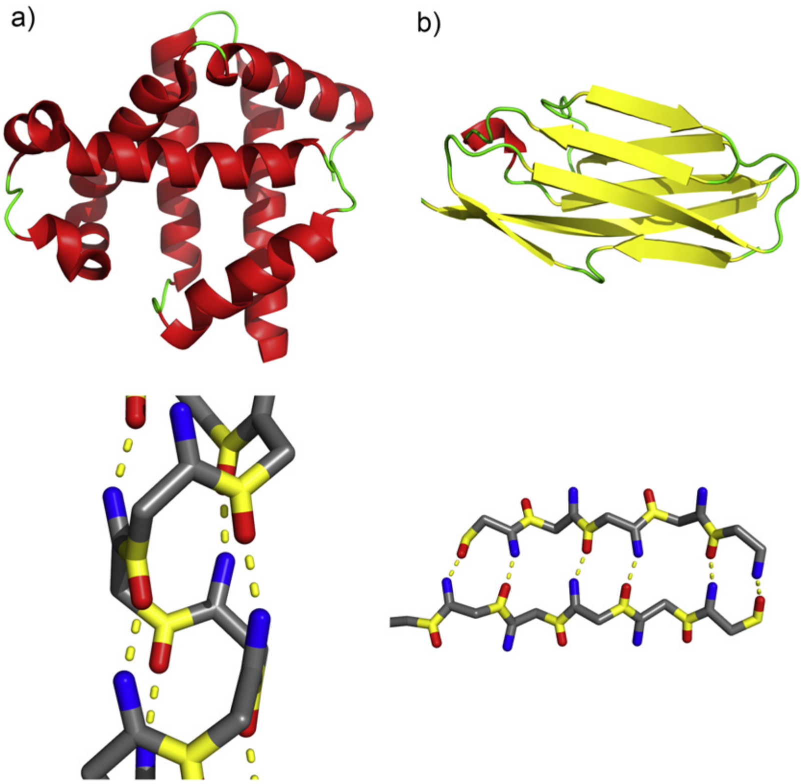

Fig. 1 α-helical and β-sheet secondary structure elements. The two most common secondary structure elements are visualized in PyMol (Schrodinger, 2015). Above, the protein is visualized in a cartoon view that highlights the secondary structure elements and hides details such as the amino acid side chains. Below, the protein backbone, without amino acid sidechains, of parts of the same secondary structures are shown in a sticks view. Here, oxygen atoms are colored blue, nitrogen yellow, hydrogen red and all others grey. Hydrogen bonds between atoms of the protein backbone are visualized as yellow dashed lines. (a) The human protein myoglobin is shown with α-helices highlighted in red and the rest of the protein in green (PDB identifier 3rgk). This is an "all-alpha" protein fold which consists of only α-helices and loops. (b) A part of the human vascular cell adhesion molecule is shown with β-sheets highlighted in yellow and loops in green. A small, single-turn α-helix can be seen in red at the left side.

Helices (α, 3_{10}, π)

As mentioned above, helices are defined by a repeating pattern of hydrogen bonds between residues. In an ideal α-helix, these bonds from between residues i and $i + 4$. This type of helix allows a relatively compact formation where four residues form a single "turn" of a helix. 3_{10} helices are more compact with just three residues (or ten atoms) forming a single "turn". π-helices on the other hand are wider and complete one "turn" every five residues. Among these three types, α-helices are energetically most favorable and thus most often observed. In literature and the output of prediction methods, helices are typically abbreviated by the letter H.

Strands and sheets (β-strands, parallel and antiparallel β-sheets)

β-sheets are a more spacious type of secondary structure formed from β-strands. Strands consist of the protein backbone "zig-zagging", typically for four to ten residues. Single β-strands are not energetically favorable. However, they can form β-sheets which are characterized by a pattern of hydrogen bonds between the residues on two different β-strands. Residue i in strand 1 forms a hydrogen bond with residue j in strand 2. The next bond can be formed in two different ways: either residue $i + 2$ in strand 1 binds to $j + 2$ in strand 2 (referred to as a parallel β-sheet), or residue $i + 2$ in strand 1 binds to $j - 2$ in strand 2 (referred to as anti-parallel β-sheet). In contrast to the situation for helices, the hydrogen bonding residues in β-sheets might be far away in sequence. Most β-sheets are formed from more than two strands. Typically, β-strands and -sheets are abbreviated by the letter E.

Other/loops

All residues that do not participate in helices or strands are often joined in one category as "other". Historically, this is also referred to as "loop" or, very misleadingly, as "(random) coil". Although the residues in question do not form hydrogen-bonded regular secondary structure, they are often constrained by bonds to other residues and are certainly not all random. Some methods distinguish different types within the class "other" such as "(reverse) turns" or "bends" where the residues in question change the direction of the amino acid chain. This non-regular class is typically abbreviated by the letter L.

Inference of Secondary Structure Elements From 3D Protein Structures

DSSP

Define Secondary Structure of Proteins (DSSP) is the standard tool for the annotation of secondary structure elements from protein structures (Kabsch and Sander, 1983; Touw *et al.*, 2015). Based primarily on hydrogen bonding patterns and some geometric constraints, it assigns every residue to one of eight possible states. States corresponding to helical structures are α-, 3_{10}- and π-helices. β-bridges are short fragments that show β-sheet like binding patterns. Multiple consecutive β-bridges form the state referred to as β-ladder. Multiple ladders can then form the above-mentioned β-sheets. However, DSSP does not have a separate state for β-sheet residues and the one for β-ladders is used. The remaining three states, called turn, bend and other, describe loop structures. Typically, secondary structure prediction methods simplify these eight states into just three: α-helix, β-sheet and loop.

STRIDE

Given a high-resolution 3D structure, annotation of secondary structure elements remains a matter of definition to some degree. STRIDE represents an alternative approach to DSSP. It aims to provide secondary structure assignments that are more consistent with the assignments performed by experimentalists who determined the protein structure (Frishman and Argos, 1995; Heinig and Frishman, 2004). Next to hydrogen bonds it includes other aspects such as the backbone geometry in the form of dihedral angle propensities. Another difference to DSSP is that some of its decision thresholds have been optimized on data from the PDB which makes it a knowledge-based approach. Nonetheless, the agreement between DSSP and STRIDE annotations is very high at 95% (Martin *et al.*, 2005).

Assessing Secondary Structure Predictions

Assessing secondary structure predictions requires all the care generally exercised when evaluating the performance of machine learning models. One aspect is the careful choice of a meaningful scoring model.

The simplest and most commonly used performance measure is the three-state per-residue accuracy referred to as Q_3. It represents the fraction of all residues that have been predicted accurately in the three states helix, strand, other. Despite its popularity, this score has several shortcomings. Q_3 ignores the uneven distribution in the three states: most residues are in the state other, i.e., methods that poorly predict the more important regular states (in particular the least common state of strand) might still reach high performance. Additional scores to measure these aspects include Q_H and Q_E, the percentages of correctly predicted helix and strand residues. Many other scores have been proposed and are useful in different contexts (Rost and Sander, 1993b). All these per-residue ignore the fact that secondary structure segments span over several residues. This shortcoming is corrected by per-segment scores (Rost *et al.*, 1994). The simplest such score compares the average length of predicted and observed segments. More advanced is a composite score referred to as segment overlap (SOV) (Fidelis *et al.*, 1999). Instead of measuring single-residue accuracy, SOV scores the correct placement of secondary structure segments. Such scores tend to give leeway with respect to predicting the precise begin and end positions of segments. Predictions that roughly predict the location of most segments score high.

Given performance measures, one also needs a test dataset to evaluate predictions. Authors exercise varying degrees of care in trying to report unbiased performance estimates. Best would be to assess all relevant methods based on proteins that differ substantially in sequence from those used for development of any of the methods.

One effort for such an assessment was EVA, a server that collected newly released protein structures as an unknown test set and then automatically queried secondary structure prediction methods for their predictions of those proteins' secondary structure elements (Eyrich *et al.*, 2001). This process was performed on a weekly basis and the results published online. A similar concept was followed by LiveBench (Rychlewski and Fischer, 2005). Unfortunately, both servers have now been offline for more than 10 years. CAMEO is a current effort in automated protein structure prediction evaluation and its operators have stated that it may include an assessment of secondary structure prediction in the future if there is a community interest (Haas *et al.*, 2013).

Approaches

Amino Acid Propensities

Since the very beginning, the prediction of secondary structure elements was based on the fact that amino acid occurrences are not evenly distributed between those elements but show preferences (Kessel and Ben-Tal, 2011). For example, due to their compactness helices often contain amino acids with small, or more precisely linear, side chains such as alanine, glutamic acid or methionine. On the other hand, proline is rarely found in helices. Its pyrrolidine ring and bound backbone amide group disturb the geometry and hydrogen bonding pattern of the helix. This leads to a kink in the helix that is energetically unfavorable and gives proline the nickname "helix breaker". In the same way β-sheets have preferences for certain amino acids, in particular at their termini. Although weak, these propensities suffice to predict secondary structure better than random. In one way or another even the most recent methods developed today still exploit such features. However, finding the right relationships is far from trivial as the formation of secondary structure depends on more than just the local amino acid content. The most extreme evidence for this are so-called chameleon sequences. These sequence fragments can fold into both helix and strand conformations depending on context (Jacoboni *et al.*, 2000; Guo *et al.*, 2007; Li *et al.*, 2015).

Basic Concepts of Secondary Structure Prediction

Even before high-resolution protein structures became available, the earliest methods used exactly these amino acid propensities mentioned in the previous section together with empirically defined rules to gain insights into the secondary structure content of proteins of interest. For example, Szent-Györgyi and Cohen described the relationship between helices and proline as outlined above (Szent-Györgyi and Cohen, 1957). Chou and Fasman determined a simple rule as to which a clustering of certain helix or strand "former" residues leads to formation of the respective secondary structure element (Chou and Fasman, 1974). One could refer to those early approaches as "first generation" methods because they ultimately scanned for features of single residues.

Stepping up, the "second generation" methods improved performance by including more information. The basic idea was to consider the sequence neighborhood of a residue, i.e., for a residue at position i those before it $(i-1, i-2, \ldots)$ and those directly after it $(i+1, i+2, \ldots)$. At the end of the 80s such statistical approaches were complemented by the first machine learning solutions. While any of the usual machine learning models is potentially applicable, (artificial) neural networks (NNs) have been particularly popular in the field. Early machine learning-based methods used a limited set of input features which may be as simple as the amino acid sequence itself. Typically, the sequence would be parsed in a sliding window of uneven size X, i.e.,. to predict the secondary structure of the central amino acid residue, all $\lfloor X/2 \rfloor$ neighboring residues are considered. With increasingly complex models and more computational power becoming available, a multitude of other features was added to increase the chances of the model finding the intricate correlations between amino acids which lead to a respective secondary structure. This allowed to predict strands and helices at equal levels of accuracy (Rost and Sander, 1993b). However, all those tools seemingly hit a performance limit since the information available within the window was too small and increasing the window size introduced more noise than signal.

This led to the leap into the "third generation" methods that combined machine learning with evolutionary information (Rost and Sander, 1993a). To use evolutionary information, one first needs to create a multiple sequence alignment (MSA). Toward this end, one needs to find all proteins with sufficiently high sequence similarity to a query which guarantees that those proteins have similar secondary structure. Most important here is not the number of related proteins, but the amount of diversity, i.e., the sequences more distantly related are more important to improve the prediction than those that are very similar. The MSA is then converted into a position-specific scoring matrix (PSSM), often also referred to as "profile". This profile can be used by the prediction method to replace a much simpler 20-dimensional binary vector with 1 for the amino acid at that position and 0 for all 19 amino acids not at that position. An example of a simple neural network with homology information as input is shown in Fig. 2.

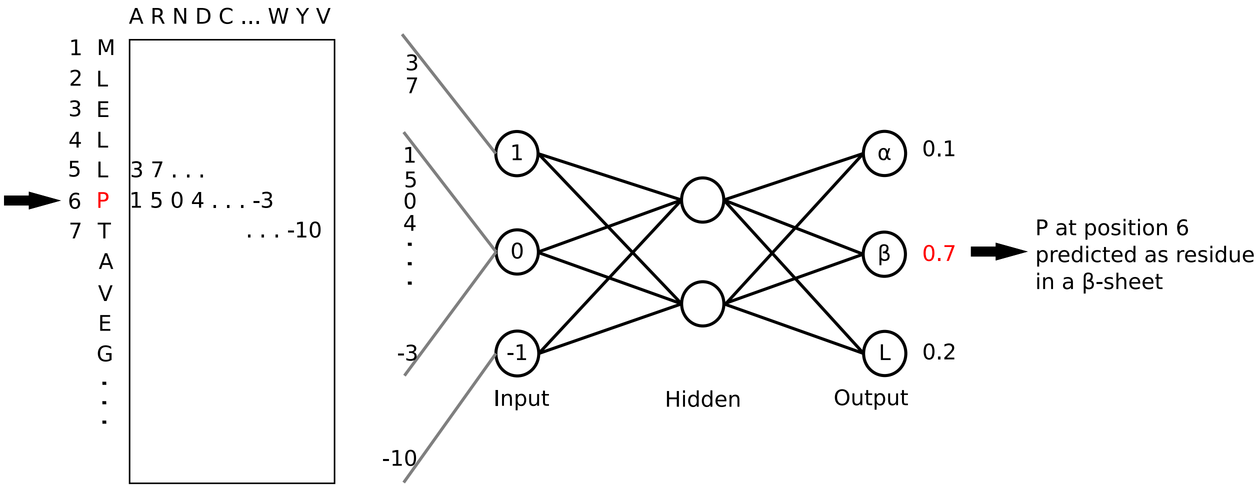

Fig. 2 A simple feed-forward neural network using homology information for the prediction of secondary structure elements. On the left, the input sequence for which the secondary structure is to be predicted runs from top to bottom. Right of it, the position specific scoring matrix (PSSM) for this input sequence is shown. Such a PSSM is created for example by performing a PSI-BLAST query with the input sequence against a large sequence database. The matrix contains information about which amino acids are commonly found in homologous sequences and which substitutions are rare. Typically, these values are used as log-odds such that negative values determine an unlikely substitutions and positives ones a likely one. These values can then be used as input to the neural network which leads to 20 inputs for every residue. To allow for a better visualization, the network shown here uses a very small sliding window of just three residues. Therefore, when predicting the secondary structure state at position 6, the PSSM values of positions 5, 6 and 7 are used as input to the network. The three input nodes labelled -1, 0 and 1 are in fact 20 nodes each, leading to a total of 60 inputs. These values are then fed through a hidden layer to the output layer which predicts one of the three secondary structure states. Weights on the connections between these layers are optimized during the training of the network. A typical neural network for the prediction of secondary elements may use a sliding window of size 13 and have hundreds of nodes in the hidden layer. In deep learning, more than one hidden layer is used and in recurrent neural networks nodes from one layer may be connected to those of a previous layer.

Over the last few years, increasingly cheap, yet powerful hardware capable of highly parallelized calculations has given a new rise to deep learning approaches that has also reached the field of secondary structure prediction. These approaches are also based on NNs. However, they have more than one hidden layer and their architecture often differs significantly from that of traditional feed-forward NNs.

Results and Discussions

Secondary Structure Prediction Methods

In the following, a few secondary structure prediction methods are described in more detail starting with some of the oldest approaches to the current state-of-the-art (**Table 1**). The methods were chosen to represent some of the historically most valuable publications, to give an overview of how the field developed and highlight some of the currently most promising methods.

Prediction based on amino acid composition statistics

GOR, named by the initials of its authors is an information theoretical approach first published in 1978 (Garnier *et al*., 1978, 1996). Using a sliding window of 17 residues the method calculates the most likely secondary structure element based on amino acid propensities which were extracted from a database of proteins with known secondary structure. The method was later extended by updating the underlying database with more recent data in GOR II. In addition to the previous single-residue statistics, GOR III included correlations between pairs of residues, i.e., between the central amino acid and amino acids at all other positions in the window. GOR IV further increased the number of residue pairs considered. Instead of only pairs with the central residue, all possible residue pairs in the window were now included in the calculation. In 2002 the final version of the method, GOR V, further extended the approach by adding statistics of amino acid triplets, as well as by including homology information, choosing the windows size depending on sequence length and again updating the underlying database (Kloczkowski *et al*., 2002).

Prediction based on two levels of neural networks

PHDsec is one of the earliest methods to apply machine learning to the task of secondary structure prediction and the first to harness homology information to further improve it (Rost and Sander, 1993a). The method is based on NNs and uses only sequence information as input. For a given input sequence an MSA with homologous sequences is constructed from the HSSP database (Sander and Schneider, 1991). Then the amino acid occurrences at every position in the sequence are calculated. The method uses a sliding window of 13 residues and every position translates to 21 input units which are fed with the amino acid occurrences from the MSA at that position. 21 units are required (as opposed to just 20) to model cases near the N- and C-terminus of the input sequence where parts of the sliding window are “empty”. The total 273 input units are connected to a hidden layer and an output layer which contains three nodes corresponding to an α-helix, β-sheet or loop prediction for the central residue in the sliding window. The output of this first NN is used as input to another NN with a 17-residue sliding window. This second layer aims to smooth the output such that segments are predicted more continuously. Later extensions to the method added more input features derived from the MSA as well as global amino acid composition information (Rost and Sander, 1994). The method has also been renamed several times, with the most recent version, ReProf, being available as part of the PredictProtein webserver

Table 1 A selection of secondary structure prediction methods discussed in this article

Method	*Learning model*	*Other predicted features*	*Primary reference*	*Webserver address*
GOR I-V	Bayes rules on amino acid propensities	None	Kloczkowski *et al.* (2002)	
PhDsec/ReProf	Neural Network	Solvent accessibility	Rost and Sander (1994)	https://predictprotein.org/
PSIPRED	Neural network	None	Jones (1999)	http://bioinf.cs.ucl.ac.uk/psipred/
JPred4	Neural network	None	Drozdetskiy *et al.* (2015)	http://www.compbio.dundee.ac.uk/jpred/
Frag1D	Database Lookup (Fragment matching)	None	Zhou *et al.* (2010)	http://frag1d.bioshu.se/
SSpro5	Database lookup + Deep Neural Network	None	Magnan and Baldi (2014)	http://scratch.proteomics.ics.uci.edu/
S2D	Neural Network	Disorder	Sormanni *et al.* (2015)	http://www-mvsoftware.ch.cam.ac.uk/
SPIDER2	Deep Neural Network	Solvent accessibility, torsion angles	Heffernan *et al.* (2015)	http://sparks-lab.org/server/SPIDER2/
Raptor X Property	Deep Neural Network	Solvent accessibility, disorder	Wang *et al.* (2016b)	http://raptorx.uchicago.edu/StructurePropertyPred/predict/

(Yachdav *et al.*, 2014). PSIPRED is another common NN based prediction method with an approach similar to that of the PHDsec family (Jones, 1999). It uses sequence profiles generated by PSI-BLAST (Altschul, 1997) as input to a NN with a sliding window of size 15, followed by a second smoothing NN with the same window size. The method has been continually updated since its first publication and the network architecture was recently improved (Buchan *et al.*, 2013). Finally, JPred4, a webserver built on the underlying Jnet also employs the two-level neural network methodology and has been updated many times since its first publication in 2000 (Drozdetskiy *et al.*, 2015; Cole *et al.*, 2008; Cuff and Barton, 2000). The method builds sequence profiles using PSI-BLAST as well as HMMER3, a hidden Markov model-based sequence search tool (Eddy, 2011), and trains two separate NN-pairs on the respective data. If these two models do not agree on a residue's prediction a third "jury" NN is used on the output of the previous two.

Prediction based on deep learning

SSpro5 is the most recent iteration of a secondary structure predictor which first performs a lookup of already known structures (Magnan and Baldi, 2014). The input sequence is compared to structures in the PDB using BLAST (Camacho *et al.*, 2009). If the resulting (local) alignments satisfy a set of similarity criteria the most common secondary structure class annotated in DSSP is used as the prediction for the respective residue. If no similar positions exist or the DSSP annotations do not form a majority, a neural network predictor is invoked to fill in predictions for the missing residues. This model consists of 100 bi-directional recurrent NNs. In contrast to the window-based feed-forward NNs covered so far, these networks can parse information of an arbitrarily large input sequence at once. In addition to the middle segment with the residue of interest and its neighbors, two other components of the network parse the rest of the sequence (Pollastri *et al.*, 2002). This could allow the recognition of global contacts which go beyond the correlations captured in the local window. The number of 100 networks result from the double cross-validation procedure which first splits the training set into 10 folds and then performs 10-fold cross-validation on each of them.

SPIDER2 is a deep-learning approach that combines the prediction of secondary structure elements with the prediction of solvent accessibility, backbone torsion angles and two more angles around the backbone C_α atoms (Heffernan *et al.*, 2015). The underlying idea is that these properties all correlate with each other to some degree. This makes the machine learning model aware of all classes at the same time which might improve prediction performance. SPIDER2 achieves this with an iterative approach: Given a PSI-BLAST PSSM of the input sequence and seven physicochemical indices, three NNs are separately trained for the prediction of secondary structure, solvent accessibility and the backbone angles. In the two following iterations, the input to each respective network is extended by the input from the other two networks. For example, predicted angles and solvent accessibility are used as input to the secondary structure predictor. In sum, this leads to three deep NN predictors with three hidden layers each. An improvement of the method using long short-term memory neural networks is currently in development.

Similar to SPIDER2, RaptorX Property is a webserver predicting several sequence features, including secondary structure and solvent accessibility (Wang *et al.*, 2016a,b). However, the prediction models are treated separately and do not interfere with each other. The authors use a deep convolutional neural field, which is the combination of a deep convolutional NN (DCNN) with seven hidden layers and a final output layer which together with the last hidden layer forms a conditional neural field (CNF). The rationale behind this design is that the DCNN can capture global relationships that go beyond the local window size of 11 residues while the final CNF learns the correlations between the secondary structure of adjacent residues. The input to the network is either just the sequence or a PSI-BLAST profile of that sequence.

Prediction based on other approaches

Frag1D performs secondary structure prediction based on a fragment matching approach (Zhou *et al.*, 2010). Given a query sequence, a sliding window of nine residues is used to find the 100 best-scoring fragments in a database of proteins with known structure. These 100 fragments are selected by comparing PSI-BLAST profiles and structural profiles between the input and database sequences. The structural profiles were created from so-called ShapeStrings which are defined based on the torsion angle pairs of the peptide bond. Same as for the direct sequence comparison, the database was searched for similar nine-residue stretches of ShapeStrings to build the structural profile which represents the conservation pattern in similar local structures. Further weighting the 100 fragments results in a preliminary secondary structure prediction. Given this prediction, the structural profile of the query sequence can now be computed and the whole process is repeated once to yield the final prediction.

s2D combines secondary structure prediction with the identification of disordered regions in proteins (Sormanni *et al.*, 2015). Such regions do not adapt a well-defined structure but instead fluctuate between different conformations (Habchi *et al.*, 2014). Typically, other methods extract secondary structure elements mostly from X-Ray crystallography structures through tools such as DSSP. In contrast, the s2D training data consists of annotations extracted from the chemical shifts of structures determined by nuclear magnetic resonance. In this way, all residues that show neither α-helical nor β-sheet properties can be considered as some form of disordered segments. Given these annotations, two NNs are trained with window sizes 11 and 15 but otherwise equal architecture. PSI-BLAST profiles are used as input to these networks. Next, a global network is trained with the same profiles and the output of the previous two networks resulting in a prediction of the mean secondary structure content in the whole sequence. Finally, the output of that network is combined with the initial inputs in a NN that smooths the prediction with a sliding window of five residues.

Estimates of Prediction Performance

As mentioned above in Section Assessing Secondary Structure Predictions two previous efforts for an automated independent assessment of secondary structure prediction are not maintained any longer. A related effort is CASP, a biannual challenge in which organizers collect target proteins with known but unpublished structures (Moult *et al.*, 1995). The target sequences are given to the community who can submit their predictions for the proteins' structures. After the structures are publicly disclosed, the assessors evaluate the performance of the prediction methods on these previously unseen proteins. This enables an independent evaluation where neither the competition organizers nor the participating groups know the answer at the time of submission. CASP has a focus on protein 3D structure prediction but in the past also included an evaluation of secondary structure prediction methods. In CASP5 assessors found that helices are predicted better than sheets and that sequences with no homology to a solved structure were harder to predict (Aloy *et al.*, 2003).

Overall, methods seem to have approached a saturation point with only small improvements over the previous years. Therefore, public assessments were stopped. This means that current estimates of prediction performance are only regularly given by the authors of newly developed methods. Although these may have the best intentions in providing a fair comparison to previous approaches they typically focus only on the flawed Q_3 score and their reported numbers may still be biased to favor their own method – for example, by the selection of a specific dataset. Traditionally, reported performance values have often proven to be overestimated. Apart from over-training, over-estimates can also be caused by discovering completely new types of structures. Ultimately, one therefore has two ways to evaluate performance by either using old results from the "last public" assessments, thereby excluding all new methods, or using method-skewed values from publications, likely over-estimating performance for the methods reported. In this chapter, the second solution was favored. However, readers must be aware that the published evaluation numbers need to be viewed with substantial skepticism.

Keeping that in mind, the most recent methods such as s2D, JPred4, RaptorX Property and SSpro5, all report Q_3 values around 80–85%. These values seem realistic, given the last independently determined performance estimates along with the increase in database size. However, choosing the best method among them is impossible given their heterogeneous evaluation schemes.

Future Directions

Without any independent evaluations, it is hard to say whether the top has already been reached in secondary structure prediction. Reinstating secondary structure prediction evaluation through a CAMEO revival of approaches such as EVA and LiveBench may help greatly. Certainly, creating such an independent evaluation entity is far from trivial and great care will be necessary when choosing the underlying datasets and scoring procedure. Clearly the limit is not given by a Q_3 of 100%. Due to errors in experimental structure determination and since proteins are in motion, secondary structure assignments will always contain some amount of error (Kihara, 2005). For the time being, the main problem with the assessment of secondary structure prediction is that although the tools are essential as input to many subsequent "higher level" prediction methods, there are very few incentives to invest substantial efforts to doing evaluations right. At the same time, it has become so difficult to publish new prediction methods in this field that almost nothing less than the claim that the ultimate has been reached will even appear on the screen of experts. There are many good reasons for these two opposing realities but they clearly make it difficult to assess today's state-of-the-art. It also remains debatable whether there is a large value in another method that shows increasingly smaller improvements over the current state-of-the-art. With the recent advances in the prediction of 3D protein structures from sequence, the interest in dedicated secondary structure prediction methods may continue to fade. However, at this point they cannot be substituted yet, since the difference in runtime between the two fields reaches several orders of magnitude and secondary structure prediction methods currently still perform better than inferring helices and sheets from predicted 3D structures (Faraggi *et al.*, 2012).

Closing Remarks

Predicting protein secondary structure elements from sequence is among the oldest prediction approaches in the field. Benefitting from decades of work, predictions are now both fast and highly accurate although determining just how accurate exactly is proving hard.

Recently, interest in these predictions has decreased, partly owing to a focus shift towards higher goals such as the prediction of protein 3D structures. Nonetheless, the respective methods are still relevant and a crucial help in making sense of the current sequence data deluge.

References

Aloy, P., Stark, A., Hadley, C., Russell, R.B., 2003. Predictions without templates: New folds, secondary structure, and contacts in CASP5. Proteins: Structure, Function and Genetics 53, 436–456.

Altschul, S., 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Research 25, 3389–3402.

Anfinsen, C.B., 1973. Principles that govern the folding of protein chains. Science 181, 223–230.

Berman, H.M., Westbrook, J., Feng, Z., *et al.*, 2000. The protein data bank. Nucleic Acids Research 28, 235–242.

Biasini, M., Bienert, S., Waterhouse, A., *et al.*, 2014. SWISS-MODEL: Modelling protein tertiary and quaternary structure using evolutionary information. Nucleic Acids Research 42, W252–W258.

Buchan, D.W.A., Minneci, F., Nugent, T.C.O., Bryson, K., Jones, D.T., 2013. Scalable web services for the PSIPRED protein analysis workbench. Nucleic Acids Research 41, 349–357.

Camacho, C., Coulouris, G., Avagyan, V., *et al.*, 2009. BLAST+: Architecture and applications. BMC Bioinformatics 10, 421.

Chou, P.Y., Fasman, G.D., 1974. Prediction of protein conformation. Biochemistry 13 (2), 222–245.

Cole, C., Barber, J.D., Barton, G.J., 2008. The Jpred 3 secondary structure prediction server. Nucleic Acids Research 36, 197–201.

Cuff, J.A., Barton, G.J., 2000. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. Proteins 40, 502–511.

Drozdetskiy, A., Cole, C., Procter, J., Barton, G.J., 2015. JPred4: A protein secondary structure prediction server. Nucleic Acids Research 43, W389–W394.

Eddy, S.R., 2011. Accelerated profile HMM searches. PLOS Computational Biology 7. doi:10.1371/journal.pcbi.1002195.

Eyrich, V., Martí-Renom, M.A., Przybylski, D., *et al.*, 2001. EVA: Continuous automatic evaluation of protein structure prediction servers. Bioinformatics 17, 1242–1243.

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. Journal of Computational Chemistry 33, 259–267.

Fidelis, K., Rost, B., Zemla, A., 1999. A modified definition of sov, a segment-based measure for protein secondary structure prediction assessment. Proteins 223, 220–223.

Frishman, D., Argos, P., 1995. Knowledge-based protein secondary structure assignment. Proteins-Structure Function and Genetics 23, 566–579.

Garnier, J., Gibrat, J.-F., Robson, B., 1996. GOR method for predicting protein secondary structure from amino acid sequence. Methods in Enzymology 266, 540–553.

Garnier, J., Osguthorpe, D.J., Robson, B., 1978. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. Journal of Molecular Biology 120, 97–120.

Goodwin, S., Mcpherson, J.D., Mccombie, W.R., 2016. Coming of age: Ten years of next-generation sequencing technologies. Nature Reviews Genetics 17, 333–351.

Guo, J.-T., Jaromczyk, J.W., Xu, Y., 2007. Analysis of chameleon sequences and their implications in biological processes. Proteins 67, 548–558.

Haas, J., Roth, S., Arnold, K., *et al.*, 2013. The protein model portal – A comprehensive resource for protein structure and model information. Database 2013, 1–8.

Habchi, J., Tompa, P., Longhi, S., Uversky, V.N., 2014. Introducing protein intrinsic disorder. Chemical Reviews 114, 6561–6588.

Heffernan, R., Paliwal, K., Lyons, J., *et al.*, 2015. Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning. Scientific Reports 5, 11476. doi:10.1038/srep11476.

Heinig, M., Frishman, D., 2004. STRIDE: A web server for secondary structure assignment from known atomic coordinates of proteins. Nucleic Acids Research 32, 500–502.

Hopf, T.A., Colwell, L.J., Sheridan, R., *et al.*, 2012. Three-dimensional structures of membrane proteins from genomic sequencing. Cell 149, 1607–1621.

Jacoboni, I., Martelli, P.L., Fariselli, P., Compiani, M., Casadio, R., 2000. Predictions of protein segments with the same aminoacid sequence and different secondary structure: A benchmark for predictive methods. Proteins 41, 535–544.

Jones, D.T., 1999. Protein secondary structure prediction based on position-specific scoring matrices. Journal of Molecular Biology 292, 195–202.

Kabsch, W., Sander, C., 1983. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22, 2577–2637.

Kessel, A., Ben-Tal, N., 2011. Protein structure. In: Introduction to Proteins. Boca Raton, FL: CRC Press.

Kihara, D., 2005. The effect of long-range interactions on the secondary structure formation of proteins. Protein Science: A Publication of the Protein Society 14, 1955–1963.

Kinch, L.N., Li, W., Monastyrskyy, B., Kryshtafovych, A., Grishin, N.V., 2016. Evaluation of free modeling targets in CASP11 and ROLL. Proteins: Structure, Function and Bioinformatics. doi:10.1002/prot.24973.

Kloczkowski, A., Ting, K.L., Jernigan, R.L., Garnier, J., 2002. Combining the GOR V algorithm with evolutionary information for protein secondary structure prediction from amino acid sequence. Proteins: Structure, Function and Genetics 49, 154–166.

Li, W., Kinch, L.N., Karplus, P.A., Grishin, N.V., 2015. ChSeq: A database of chameleon sequences. Protein Science 24, 1075–1086.

Magnan, C.N., Baldi, P., 2014. SSpro/ACCpro 5: Almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity. Bioinformatics (Oxford, England) 30, 2592–2597.

Marks, D.S., Hopf, T.A., Chris, S., Sander, C., 2012. Protein structure prediction from sequence variation. Nature Biotechnology 30, 1072–1080.

Martinez, D.A., Nelson, M.A., 2010. The next generation becomes the now generation. PLOS Genetics 6, e1000906.

Martin, J., Letellier, G., Marin, A., *et al.*, 2005. Protein secondary structure assignment revisited: A detailed analysis of different assignment methods. BMC Structural Biology 5, 17.

Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T., Tramontano, A., 2016. Critical assessment of methods of protein structure prediction: Progress and new directions in round XI. Proteins 84 (Suppl. 1), 4–14.

Moult, J., Pedersen, J.T., Judson, R., Fidelis, K., 1995. A large-scale experiment to assess protein structure prediction methods. Proteins: Structure, Function, and Genetics 23, ii–iv.

Muir, P., Li, S., Lou, S., *et al.*, 2016. The real cost of sequencing: Scaling computation to keep pace with data generation. Genome Biology 17, 53.

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: Structure, Function, and Bioinformatics 47, 228–235.

Rost, B., Sander, C., 1993a. Improved prediction of protein secondary structure by use of sequence profiles and neural networks. Proceedings of the National Academy of Sciences 90, 7558–7562.

Rost, B., Sander, C., 1993b. Prediciton of protein secondary structure at better than 70% accuracy. Journal of Molecular Biology 232 (2), 584–599.

Rost, B., Sander, C., 1994. Combining evolutionary information and neural networks to predict protein secondary structure. Proteins: Structure, Function, and Bioinformatics 19, 55–72.

Rost, B., Sander, C., Schneider, R., 1994. Redefining the goals of protein secondary structure prediction. Journal of Molecular Biology 235, 13–26.

Rychlewski, L., Fischer, D., 2005. LiveBench-8: The large-scale, continuous assessment of automated protein structure prediction. Protein Science 14, 240–245.

Sander, C., Schneider, R., 1991. Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins: Structure, Function, and Bioinformatics 9, 56–68.

Schrodinger, L.L.C., 2015. The PyMOL molecular graphics system, Version 1.8.

Sormanni, P., Camilloni, C., Fariselli, P., Vendruscolo, M., 2015. The s2D method: Simultaneous sequence-based prediction of the statistical populations of ordered and disordered regions in proteins. Journal of Molecular Biology 427, 982–996.

Suzek, B.E., Wang, Y., Huang, H., Mcgarvey, P.B., Wu, C.H., 2015. UniRef clusters: A comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics 31, 926–932.

Szent-Györgyi, A.G., Cohen, C., 1957. Role of proline in polypeptide chain configuration of proteins. Science 126, 697.

The Uniprot Consortium, 2016. UniProt: The universal protein knowledgebase. Nucleic Acids Research 45, 1–12.

Touw, W.G., Baakman, C., Black, J., *et al.*, 2015. A series of PDB-related databanks for everyday needs. Nucleic Acids Research 43, D364–D368.

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

Wang, S., Peng, J., Ma, J., Xu, J., 2016b. Protein secondary structure prediction using deep convolutional neural fields. Scientific Reports 6, 18962.

Yachdav, G., Kloppmann, E., Kajan, L., *et al.*, 2014. PredictProtein – An open resource for online prediction of protein structural and functional features. Nucleic Acids Research 42, W337–W343.

Yang, Z., Lasker, K., Schneidman-Duhovny, D., *et al.*, 2011. UCSF Chimera, MODELLER, and IMP: An integrated modeling system. Journal of Structural Biology 179 (3), 269–278.

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

Relevant Websites

https://www.cameo3d.org
CAMEO.

https://www.wwpdb.org/
The Protein Data Bank.