Current Protocols in Protein Science
Unit 2.3

Computational Prediction of Protein Secondary Structure from Sequence

Fanchi Meng[1] and Lukasz Kurgan[2*]

[1]Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Canada

[2]Department of Computer Science, Virginia Commonwealth University, Richmond, USA.

* corresponding author
Email: lkurgan@vcu.edu
Phone: 804-827-3986

SIGNIFICANCE STATEMENT

Computational approaches offer a cost- and time-efficient way to predict secondary structure of proteins from protein sequences. The current, third generation of these computational methods provides accurate predictions and is conveniently and freely available as webservers and standalone software. These predictions are widely used across the globe to facilitate prediction of the tertiary protein structure and various functional characteristics of proteins. We provide practical insights on how to perform and interpret the predictions for selected modern methods.

ABSTRACT

Secondary structure of proteins refers to local and repetitive conformations, such as α-helices and β-strands, which occur in protein structures. Computational prediction of secondary structure from protein sequences has long history with three generations of predictive methods. This unit summarizes several recent third-generation predictors. We discuss their inputs and outputs, availability, predictive performance, and explain how to perform and interpret their predictions. We cover methods for the prediction of the 3-class secondary structure states (helix, strand, and coil) as well as the 8-class secondary structure states. Recent empirical assessments and our small-scale analysis reveal that these predictions are characterized by high levels of accuracy between 70 and 80%. We emphasize that modern predictors are available to the end users in the form of convenient to use webservers and standalone software.

Keywords: secondary structure of proteins; helix; strand; coil; prediction; DSSP

INTRODUCTION

Proteins are polymers of 20 types of amino acids. For most proteins, their amino acid chains fold into specific spatial conformations to carry out their biological functions. Thus, it is beneficial to determine these structures to contribute to the understanding of how proteins function at the molecular level. Protein structure is typically categorized into four levels. The primary structure is the linear sequence of amino acid joined by peptide bonds. Secondary structure (SS) refers to local and regularly occurring patterns, such as α-helices and β-strands, which are determined by the dihedral angles and resulting hydrogen bonds between peptide groups. Tertiary structure describes how the protein chains are folded into a three dimensional shape; this corresponds to a specific spatial arrangement of the SSs. Some proteins include multiple polypeptide chains and in these cases the quaternary structure is defined as the spatial arrangements of these chains. The Protein Data Bank (PDB) (Berman et al., 2000) is the worldwide repository of the three-dimensional structural data and the corresponding sequences of large biological molecules, with primary focus on proteins. As of May 2016, there were about 110 thousand protein structures in PDB including about 32 thousand structures of human sequences.

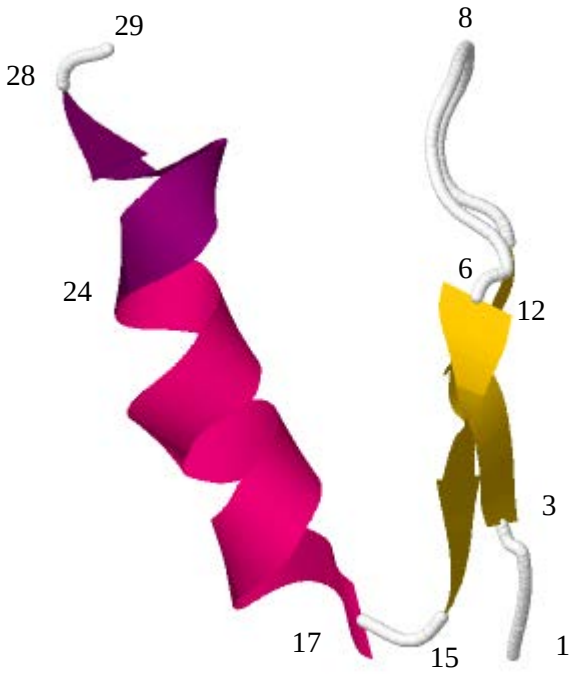
Typically, SS is annotated from the three-dimensional structure. First, the three-dimensional structure is solved and then the SS is computed from the coordinates of the atoms that make up the amino acids that are positioned in the three-dimensional structure. There are two main types of SSs: α-helices and β-strands; they were first postulated by Pauling and his colleagues in 1950s (Pauling et al., 1951). The first method to annotate secondary structure was developed by Michael Levitt (recipient of the 2013 Nobel Prize in Chemistry) and Jonathan Greer in 1976 (Levitt and Greer, 1977). The arguably most widely-used method to assign SSs that is often assumed as the gold standard (Joosten et al., 2011; Kurgan and Disfani, 2011) is the dictionary of proteins SS (DSSP) that was proposed by Wolfgang Kabsch and Christian Sander in 1983 (Kabsch and Sander, 1983). The original article that describes DSSP was cited close to 11 thousand times (source: Google Scholar as of May 2016). The popularity of this method stems from the fact that it is used in the PDB and that it was utilized to evaluate methods for the prediction of SS in two largest community-driven assessments: the Critical Assessment of protein Structure Prediction (CASP) (Moult et al., 1995) and evaluation of automatic protein structure prediction (EVA) (Koh et al., 2003). DSSP assigns one of the following eight SS types for every structured residue (i.e., residue that has three-dimensional coordinates for its atoms):

1. H: α-helix (hydrogen bonds every 4 residues)
2. B: residue in an isolated β-strand
3. E: extended strand that participates in formation of β sheets
4. G: 3_{10} helix (hydrogen bonds every 3 residues)
5. I : π helix (hydrogen bonds every 5 residues)
6. T : hydrogen bonded turn
7. S : bend
8. Blank, –, or C: loop or irregular structure (also referred to as coil or random coil)

The above are known as the DSSP's 8-class classification. These 8 types are often simplified into 3-class classification (3 major types of SSs):

1. H: helix; it encompasses right or left handed cylindrical/helical conformations that include H, G and I types.

2. E: extended strand; it corresponds to pleated sheet structures including E and B types.
3. – or C: other remaining types including blank (– or C), T and S.



A

Residue no	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
Sequence	M	V	Y	V	C	H	F	E	N	C	G	R	S	F	N	D	R	R	K	L	N	R	H	K	K	I	H	T	R
DSSP-8	C	C	E	E	E	C	C	S	S	C	C	E	E	E	S	S	H	H	H	H	H	H	H	G	G	G	G	C	C
SSpro8	C	E	E	E	E	E	E	E	T	T	T	C	C	H	H	H	H	H	H	H	H	H	H	C	E	E	C	C	C
DSSP-3	C	C	E	E	E	C	C	C	C	C	C	E	E	E	C	C	H	H	H	H	H	H	H	H	H	H	H	C	C
PSSpred	C	C	E	E	E	E	E	C	C	C	C	C	C	H	H	H	H	H	H	H	H	H	H	H	H	H	H	C	C
Confidence	9	3	7	8	7	6	1	5	8	8	8	6	3	3	4	5	7	8	8	8	8	7	7	6	5	4	2	4	9
P_C	97	66	16	9	15	18	42	72	86	87	84	74	61	32	27	21	12	9	8	8	8	12	15	17	21	25	35	64	99
P_H	1	1	0	1	1	1	3	5	7	7	9	15	27	58	66	76	87	90	92	92	91	86	82	78	70	64	53	24	0
P_E	2	35	85	92	85	82	56	25	9	5	5	8	8	6	3	2	1	1	1	1	1	2	2	4	5	9	11	6	1

B

Figure 1. Structure of a zinc finger (PDB ID: 2AB7). Panel A shows a cartoon representation of the three-dimensional structure. The N-terminus (C-terminus) of the protein sequence is located in the lower right (upper left) corner. The secondary structures are color-coded as follows: loop and bend (gray), strand (yellow), and helix (red and purple). Numbers show positions of selected residues in the protein sequence. Panel B shows the protein sequence where amino acids are given using 1-letter code together with native and predicted SSs. DSSP-8 and DSSP-3 lines show the native 8- and 3-class SSs, respectively, annotated for each residue using DSSP. The SSpro8 and PSSpred lines give the putative 8- and 3-class SSs predicted with SSpro and PSSpred methods, respectively. The four lines at the bottom provide confidence index and probability (in percent) of predictions for every residue that were generated by the PSSpred method.

Figure 1 shows an example of the three-dimensional structure of a zinc finger together with its corresponding secondary structure annotated with DSSP using the 8 and 3 classes. Following the SS along the protein sequence from the N-terminus we find a segment of two residues that form a loop (gray color), a segments of three residues that forms a strand (yellow), 2-residues long loop, bend (gray color) composed of two residues followed by another 2-residues long loop, second strand that forms a β sheet with the first strand, a short bend, 11-residues long helix that is annotated as α-helix for the first 7 residues (red) and as a 3_{10} helix for the last 4 residues (purple), and a short loop at the C-terminus. We also provide secondary structure predicted from the primary structure (amino acid sequences) using two methods, one for the prediction of 8 classes and another for the prediction of 3 classes of SS. We discuss this structure and predictions at a greater depth later in this unit.

PREDICTION OF SECONDARY STRUCTURE FROM SEQUENCE

Motivation

In spite of the fact that we know structures for over 100 thousand proteins, most protein structures in nature remain unknown. As of May 2016, there were nearly 64 million non-redundant protein sequences in the RefSeq database (Pruitt et al., 2007). Although some of these sequences share similar or even identical structure, arguably many proteins still await structural determination. The large number of unknown structures is one of the barriers that keeps us from learning and studying protein functions. The fast pace at which new protein sequences are accumulated and the lagging number of solved structures motivate development of computational methods for the prediction of protein structures from the sequences. These methods are less expensive to use and substantially more time-efficient compared to the experimental methods. They rely on the observation made by Anfinsen in 1970s that protein sequence uniquely determines the corresponding tertiary structure of proteins (Anfinsen, 1973). This suggests that in principle prediction from the sequence alone could provide correct structure. The fact that the information of how the sequences are folded into SS can then be used to predict the tertiary structure in a stepwise fashion (from the amino acid sequence to SS, and from SS to the tertiary structure) fuels the development of many methods for the prediction of SS.

Brief historical overview

Most of the computational methods predict the 3-class SS, i.e., they predict every residue in the input protein sequence as helix, strand or coil (other type). The first SS predictor was proposed in 1965 (Guzzo, 1965). This method took advantage of correlations between particular amino acid types and SS types. Other early prediction methods that were developed in 1960s and 1970s, i.e., the first generation methods (Rost, 2001), used a similar idea that connects the likelihood of particular amino acid types forming particular SS type (Rost, 2002). The second generation methods which appeared in 1980s and early 1990s (Rost, 2001) extended the single amino acid to a segment of adjacent residues (a sliding window) This approach was motivated by the fact that SSs form segments of consecutive residues in the sequence. However, the accuracy stalled at around 60% due to the fact that these early methods used only local information (a single residue or a segment of adjacent residues in a single input sequence) as their input. This

information was estimated to account for about 65% of the formation of SS (Rost, 2002; Zhang et al., 2011). A major breakthrough has arrived in mid 1990s with the third generation predictors (Rost, 2001). The key advancement was the adoption of evolutionary information generated via multiple sequence alignment, which is used by most of the currently popular methods like PROFsec (Rost and Sander, 1993a; Rost and Sander, 1993b; Rost and Sander, 1994) and PSIPRED (Jones, 1999). The third generation methods have the accuracy for the 3-class prediction at over 70% (Rost, 2002). Recent advances include availability of progressively larger datasets of structurally solved proteins that were used to train predictive models, use of more advanced predictive models, and development of consensus schemes that combine results generated by multiple predictors of SS (Rost and Sander, 2000; Zhang et al., 2011). For instance, authors of PROTEUS (Montgomerie et al., 2006) and RaptorX (Wang et al., 2016) have reported the 3-class accuracy of 81% (Montgomerie et al., 2008) and 84% on their test datasets, respectively. Moreover, in recent years several methods that predict the 8-class SS, such as SSpro8 (Pollastri et al., 2002) and RaptorX (Wang et al., 2016), were released. The prediction of the 8-class secondary structure is arguably more challenging compared to the 3-class prediction.

Computation of predictions

We focus on the sequence-based SS prediction where secondary structure is generated from the input protein sequence. Typically, these predictors accept a single amino acid sequence as the input (in either FASTA format or as a raw sequence). Some methods may also take multiple sequence alignment as the input (to save time required to produce evolutionary profile if it was already computed), and some allow submission of multiple query sequences in a batch. Their output consists of the 3-class or 8-class SS types for every residue in the input sequence. For example, in Figure 1B, the SSpro8 method outputs the predicted 8-class SS types that are based on the DSSP assignment and PSSpred outputs the 3-class types. Some predictors also provide a score that quantifies likelihood that a given prediction is correct and/or multiple scores for each of the possible SS types. PSSpred outputs a confidence index for the predicted SS type (Figure 1B). The confidence is computed as a difference between probability of the predicted type (which is by definition higher than probability of the other types) and the second highest probability. Higher values of this index indicate a higher likelihood that the given prediction is correct. Figure 1B also provides the values of the predicted probabilities for each of the three SS types for every residue that were generated by PSSpred. The predicted SS type is the type that has the highest associated probability.

A list of 13 modern predictors is given in Table 1. These methods were last published no earlier than 2005 and are available as webservers and/or standalone software packages; the latter makes them convenient to use for the end users. They were identified based on a PubMed search and based on the list of methods included in the recent comparative assessment (Zhang et al., 2011) and review (Chen and Kurgan, 2013). We focus on three methods that we recommend as a good starting point when needing to predict SS: (1) SSpro, one of the most recent methods that predicts both 3- and 8-class SSs; (2) PSIPRED, arguably the most popular SS predictor which was reported to receive the highest number of citations per year (Zhang et al., 2011); and (3) PSSpred that is among the most recent method.

Table 1. Summary of recent sequence-based SS prediction methods. The methods are sorted chronologically. WS: Web Server; SP: Standalone Package.

Prediction method	Year last published	Batch submission	3 and/or 8 class	Availability	Reference(s)	URL
SABLE	2005	No	3	WS + SP	(Adamczak et al., 2005)	http://sable.cchmc.org/
YASPIN	2005	No	3	WS	(Lin et al., 2005)	http://www.ibi.vu.nl/programs/yaspinwww/
Porter	2005	Yes	3	WS + SP	(Pollastri and McLysaght, 2005)	http://distillf.ucd.ie/distill/
PROTEUS	2008	Yes	3	WS + SP	(Montgomerie et al., 2008; Montgomerie et al., 2006)	http://www.proteus2.ca/
SPINE-X	2012	No	3	WS + SP	(Faraggi et al., 2012)	http://sparks-lab.org/SPINE-X/
PSIPRED	2013	No	3	WS + SP	(Buchan et al., 2013; Jones, 1999)	http://bioinf.cs.ucl.ac.uk/psipred/
SSPro	2014	No	3 and 8	WS + SP	(Magnan and Baldi, 2014; Pollastri et al., 2002)	http://scratch.proteomics.ics.uci.edu/
SCORPION	2014	No	3 and 8	WS + SP	(Yaseen and Li, 2014)	http://hpcr.cs.odu.edu/c3scorpion/ http://hpcr.cs.odu.edu/c8scorpion
PROFsec	2014	No	3	WS + SP	(Rost and Sander, 1993a; Rost and Sander, 1993b; Rost and Sander, 1994; Yachdav et al., 2014)	https://www.predictprotein.org/
JPred	2015	Yes	3	WS + API	(Drozdetskiy et al., 2015)	http://www.compbio.dundee.ac.uk/jpred/
PSSpred	2015	No	3	WS + SP	(Yang et al., 2015)	http://zhanglab.ccmb.med.umich.edu/PSSpred/
SPIDER^2	2015	No	3	WS + SP	(Heffernan et al., 2015)	http://sparks-lab.org/index.php/Main/Services
RaptorX	2016	Yes	3 and 8	WS + SP	(Källberg et al., 2012; Wang et al., 2016)	http://raptorx.uchicago.edu/

PSIPRED is available as a webserver at http://bioinf.cs.ucl.ac.uk/psipred/ and can be also downloaded as a standalone software for the Linux platform at the same address. This method accepts single amino acid sequence or FASTA-formatted multiple sequence alignment as the input. It outputs predicted SS type for each residue in the input sequence (H: helix, E: strand, C: coil); each prediction comes with a corresponding confidence level that is scaled to the range between 0 (low confidence) and 9 (high confidence).

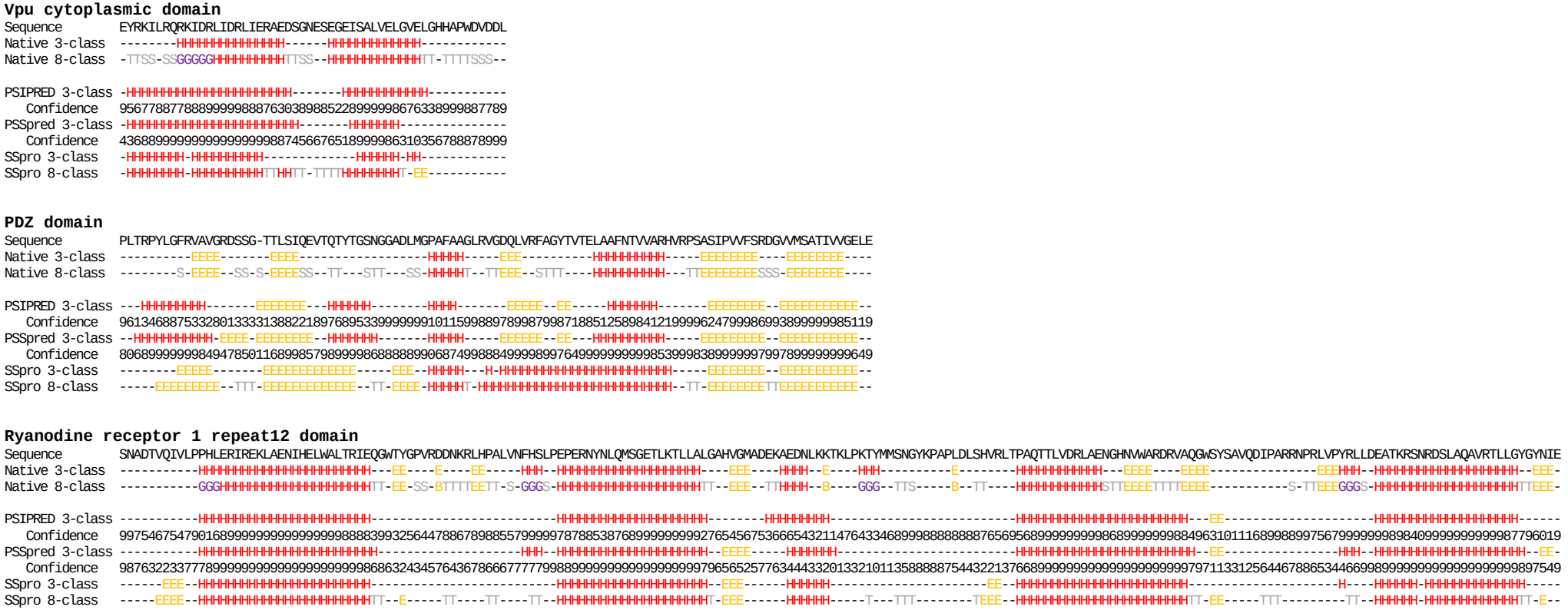
SSpro is available as a webserver as a part of the SCRATCH platform at http://scratch.proteomics.ics.uci.edu/. Its standalone version runs on the Linux platform and can be obtained from the same address. SSpro takes a single raw (unformatted) amino acid sequence up to 1500 residues long as the input. For each residue in the input sequence it generates the 3-class prediction (H: helix, E: strand, C: coil) and the 8-class prediction (H: α helix, G: 3_{10} helix, I: π helix, E: extended strand, B: isolated β-strand, T: turn, S: bend, C: coil).

PSSpred can be used as a webserver and standalone software running on the Linux platform; both versions can be found at http://zhanglab.ccmb.med.umich.edu/PSSpred/. This predictor accepts a single FASTA-formatted amino acid sequence up to 4000 residues long as the input. It outputs the SS type for each residue in the input sequence (H: helix, E: strand, C: coil), the corresponding confidence index that is scaled to the range between 0 (low confidence) and 9 (high confidence), and three probabilities for the three SS types. The predicted secondary structure corresponds to the SS type that secures the highest probability.

Analysis of predictions

We use three protein structures that were recently released in PDB to visualize and compare predictions from PSIPRED, SSpro and PSSpred with each other and with the native structure. These proteins include Vpu cytoplasmic domain (PDB id: 2N29), PDZ domain (PDB id: 2N7P), and Ryanodine receptor 1 repeat12 domain (PDB id: 5C30). We use the recently released structures to minimize predictive bias due to a potential inclusion of these proteins into datasets that were used to build the considered SS predictors. The selected proteins include one short (2N29 with 54 residues), one medium-size (2N7P with 104 residues) and one longer sequence (5C30 with 196 residues). Moreover, the latter two proteins include strands, helices and coils while the shortest chain has helices and coils. This allows us to assess prediction of all major SS types. We note that results on these few proteins should not be assumed to be representative of an overall predictive quality of a given method. The predictions were collected from the corresponding three webservers. Figure 2 that summarizes these predictions reveals that predictions of the three methods generally agree with the native annotations of the secondary structure. Most of the helices and strands were correctly predicted, although their boundaries suffer some errors. The predictors are also relatively consistent with each other. Moreover, we note that the confidence scores offer useful information. For instance, a part of the long helix that was incorrectly predicted at the N-terminus of the Vpu cytoplasmic domain (PDB id: 2N29) by PSIPRED and PSSpred includes residues that have low values of confidence, < 8. Similarly, the helix incorrectly predicted by PSIPRED at the N-terminus of the PDZ domain is also scored with low values of confidence. At the same time, the majority of the correctly predicted helices have the confidence values at 8 and 9. Next, we quantify the predictive quality of the considered methods.

Figure 2. Predictions with PSIPRED, SSpro and PSSpred for Vpu cytoplasmic domain (PDB id: 2N29), PDZ domain (PDB id: 2N7P), and Ryanodine receptor 1 repeat12 domain (PDB id: 5C30). The 3-class SS are annotated as follows: H (helix), E (strand), and – (coil). The 8-class SS are annotated as follows: E (extended strand), B (isolated strand), H (α-helix), G (3_{10} helix) I π helix, T hydrogen bonded turn, S bend, and – (coil). The secondary structures are color-coded as follows: loop and bend (gray), strand (yellow), and helix (red and purple).



Accuracy of a given prediction is evaluated by comparing the predicted SS with the native SS which is typically obtained from an experimentally solved three-dimensional structure. Popular measures that are used to evaluate predictive quality include:

1. $Q_{i\text{pre}}$: the proportion of correctly predicted residues of a given SS type among all residues of the same predicted type, where *i* stands for one of the three or eight types.
2. $Q_{i\text{obs}}$: the proportion of correctly predicted residues of a given SS type among all residues with the same native type, where *i* stands for one of the three or eight types.
3. Q_3 or Q_8 value: the overall rate of correctly predicted residues over all SS types for the 3-class or 8-class predictions.

Overall, a high quality predictor should offer high Q values.

We use the results generated by PSSpred from Figure 1B to demonstrate how these measures are calculated. In the predicted SS sequence, 11 out of 14 predicted helical residues are also native helical residues (as annotated via DSSP-3) and so $Q_{H\text{pred}} = 11/14 \approx 78.6\%$. Similarly, $Q_{E\text{pred}} = 3/5 = 60\%$ and $Q_{C\text{pred}} = 8/10 = 80\%$. In the native SS sequence (DSSP-3 line), 11 out of 11 helix residues are also predicted as helix residues and thus $Q_{H\text{obs}} = 11/11 = 100\%$. Likewise, $Q_{E\text{obs}} = 3/6 = 50\%$ and $Q_{C\text{obs}} = 8/12 \approx 66.7\%$. Overall, 22 out of 29 residues are predicted correctly (i.e., the predicted and native SS type are the same) and so $Q_3 = 22/29 \approx 75.9\%$.

Table 2. Summary of the predictive quality for the predictions with PSIPRED, SSpro and PSSpred for Vpu cytoplasmic domain (PDB id: 2N29), PDZ domain (PDB id: 2N7P), and Ryanodine receptor 1 repeat12 domain (PDB id: 5C30).

Type of SS annotation	Prediction method	Q_3 values (%)			
		2N29	**2N7P**	**5C30**	**Average**
3-class SS types using DSSP	PSIPRED	79.6	66.7	76.6	74.3
	SSpro	70.4	66.7	75.6	70.9
	PSSpred	72.2	65.7	80.6	72.8
8-class SS types using DSSP		Q_8 values (%)			
	SSpro	38.9	55.2	65.2	53.1

The overall accuracies measured with Q_3 (Q_8 for the prediction of the 8-class SS types) for PSIPRED, SSpro and PSSpred for the three sample proteins from Figure 2 are summarized in Table 2. The average (over the three proteins) accuracy ranges between 71 and 74% for the 3-class predictions and equals 53% for the 8-class predictions. The values for the individual proteins vary more widely but overall are correlated between the methods, i.e., all methods predict the Ryanodine receptor (PDB id: 5C30) relatively well and the PDZ domain (PDB id: 2N7P) with the lowest predictive performance. The average values are in agreement with the results from the recent comparative assessment where the Q_3 values ranged between 68% and 82%, depending on the method used (Zhang et al., 2011). We note that these are relatively high values given that the upper limit of predictive quality of the 3-class SS was estimated to be around 90% (Kihara, 2005). This limit was quantified based on differences that were observed between different X-ray structures and NMR models of the same proteins, and inconsistencies in the assignment of SS structures by different annotation protocols.

SUMMARY

Prediction of secondary structure from protein sequences (sequence-based SS prediction) is a mature field of research and these predictors enjoy a widespread use. For instance, many SS predictors including Porter, PROTEUS, RaptorX, PSIPRED, SSpro, and PROFsec were included in comprehensive pipelines for the prediction of protein structure and function. Some methods, like Porter, PROTEUS and SSpro, also incorporate homology search in their predictive models, i.e., they find proteins with sequences that are similar to the sequence of the input protein and use their SSs to perform predictions. Modern sequence-based SS predictors rely on the availability of large databases of proteins that are used to train accurate predictive models and to produce accurate evolutionary information.

Most of the existing SS predictors focus on the 3-class SS (Table 1). Based on the prior comparative reviews and our small-scale evaluation (Table 2), the end users can expect to collect predictions with the average overall accuracy over the three SS types between 70% and 80% (Kurgan and Disfani, 2011; Rost, 2001; Zhang et al., 2011). We also encourage the users to utilize the confidence indices provided by some predictors to judge reliability of predictions for individual residues. This approach was suggested to be “the most successful strategy to find the most reliable predicted regions” by Burkhard Rost, one of the pioneers of the SS prediction (Rost, 2001). We note that recently, between 2014 and 2016, three methods that predict 8-class SS types were released: RaptorX, SSPro, and SCORPION.

Considering usability, most modern prediction methods are provided as webservers (Table 1). This makes it easy and convenient for the end users to submit requests and collect results without the need for dedicated hardware or software. The users just need a modern web browser and an internet-connected computer to obtain the predictions. Some predictors, such as Porter, PROTEUS, RaptorX, and Jpred, accept queries with multiple sequences (batch submissions). This reduces the workload related to predictions on larger datasets of proteins. Many methods (Table 1) also provide standalone packages, which allow the end users to run predictions on their local computers. This is particularly handy when these predictors need to be incorporated into the end user’s computational pipelines. However, computations performed by the modern SS predictors can be relatively time-consuming. The end users should expect that prediction for a short protein (up to 200 amino acids) sequence may take between one and several minutes, and over half an hour for longer proteins (over 500 residues).

LITERATURE CITED

Adamczak, R., Porollo, A., and Meller, J. 2005. Combining prediction of secondary structure and solvent accessibility in proteins. *Proteins: Structure, Function, and Bioinformatics* 59:467-475.

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

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. 2000. The Protein Data Bank. *Nucleic Acids Research* 28:235-242.

Buchan, D.W.A., Minneci, F., Nugent, T.C.O., Bryson, K., and Jones, D.T. 2013. Scalable web services for the PSIPRED Protein Analysis Workbench. *Nucleic Acids Research* 41:W349-W357.

Chen, K. and Kurgan, L. 2013. Computational prediction of secondary and supersecondary structures. *Methods Mol Biol* 932:63-86.

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

Faraggi, E., Zhang, T., Yang, Y., Kurgan, L., and 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* 33:259-267.

Guzzo, A. 1965. The influence of amino-acid sequence on protein structure. *BiophysJ* 5:809-822.

Heffernan, R., Paliwal, K., Lyons, J., Dehzangi, A., Sharma, A., Wang, J., Sattar, A., Yang, Y., and 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.

Jones, D.T. 1999. Protein secondary structure prediction based on position-specific scoring matrices1. *Journal of Molecular Biology* 292:195-202.

Joosten, R.P., te Beek, T.A.H., Krieger, E., Hekkelman, M.L., Hooft, R.W.W., Schneider, R., Sander, C., and Vriend, G. 2011. A series of PDB related databases for everyday needs. *Nucleic Acids Research* 39:D411-D419.

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

Källberg, M., Wang, H., Wang, S., Peng, J., Wang, Z., Lu, H., and Xu, J. 2012. Template-based protein structure modeling using the RaptorX web server. *Nat. Protocols* 7:1511-1522.

Kihara, D. 2005. The effect of long-range interactions on the secondary structure formation of proteins. *Protein Sci* 14:1955-1963.

Koh, I.Y.Y., Eyrich, V.A., Marti-Renom, M.A., Przybylski, D., Madhusudhan, M.S., Eswar, N., Graña, O., Pazos, F., Valencia, A., Sali, A., and Rost, B. 2003. EVA: evaluation of protein structure prediction servers. *Nucleic Acids Research* 31:3311-3315.

Kurgan, L. and Disfani, F.M. 2011. Structural protein descriptors in 1-dimension and their sequence-based predictions. *Curr Protein Pept Sci* 12:470-489.

Levitt, M. and Greer, J. 1977. Automatic identification of secondary structure in globular proteins. *Journal of Molecular Biology* 114:181-239.

Lin, K., Simossis, V.A., Taylor, W.R., and Heringa, J. 2005. A simple and fast secondary structure prediction method using hidden neural networks. *Bioinformatics* 21:152-159.

Magnan, C.N. and 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* 30:2592-2597.

Montgomerie, S., Cruz, J.A., Shrivastava, S., Arndt, D., Berjanskii, M., and Wishart, D.S. 2008. PROTEUS2: a web server for comprehensive protein structure prediction and structure-based annotation. *Nucleic Acids Research* 36:W202-W209.

Montgomerie, S., Sundararaj, S., Gallin, W.J., and Wishart, D.S. 2006. Improving the accuracy of protein secondary structure prediction using structural alignment. *BMC Bioinformatics* 7:301.

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

Pauling, L., Corey, R.B., and Branson, H.R. 1951. The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain. *Proceedings of the National Academy of Sciences* 37:205-211.

Pollastri, G. and McLysaght, A. 2005. Porter: a new, accurate server for protein secondary structure prediction. *Bioinformatics* 21:1719-1720.

Pollastri, G., Przybylski, D., Rost, B., and 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.

Pruitt, K.D., Tatusova, T., and Maglott, D.R. 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research* 35:D61-D65.

Rost, B. 2001. Review: protein secondary structure prediction continues to rise. *J Struct Biol* 134:204-218.

Rost, B. 2002. Prediction In 1D: secondary structure, membrane helices, and accessibility In Structural Bioinformatics. *In* Structural Bioinformatics (P. Bourne and H. Weissig, eds.) pp. 559-588. Wiley New Jersey, USA.

Rost, B. and Sander, C. 1993a. Improved prediction of protein secondary structure by use of sequence profiles and neural networks. *Proc Natl Acad Sci U S A* 90:7558-7562.

Rost, B. and Sander, C. 1993b. Prediction of Protein Secondary Structure at Better than 70% Accuracy. *Journal of Molecular Biology* 232:584-599.

Rost, B. and Sander, C. 1994. Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins* 19:55-72.

Rost, B. and Sander, C. 2000. Third Generation Prediction of Secondary Structures. *In* Protein Structure Prediction, vol. 143 (D. Webster, ed.) pp. 71-95. Humana Press.

Wang, S., Peng, J., Ma, J.Z., and Xu, J.B. 2016. Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields. *Scientific Rep* 6:18962.

Yachdav, G., Kloppmann, E., Kajan, L., Hecht, M., Goldberg, T., Hamp, T., Hönigschmid, P., Schafferhans, A., Roos, M., Bernhofer, M., Richter, L., Ashkenazy, H., Punta, M., Schlessinger, A., Bromberg, Y., Schneider, R., Vriend, G., Sander, C., Ben-Tal, N., and Rost, B. 2014.

PredictProtein—an open resource for online prediction of protein structural and functional features. *Nucleic Acids Research* 42:W337-W343.

Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., and Zhang, Y. 2015. The I-TASSER Suite: protein structure and function prediction. *Nat Meth* 12:7-8.

Yaseen, A. and Li, Y. 2014. Context-Based Features Enhance Protein Secondary Structure Prediction Accuracy. *Journal of Chemical Information and Modeling* 54:992-1002.

Zhang, H., Zhang, T., Chen, K., Kedarisetti, K.D., Mizianty, M.J., Bao, Q., Stach, W., and Kurgan, L. 2011. Critical assessment of high-throughput standalone methods for secondary structure prediction. *Brief Bioinform* 12:672-688.

KEY REFERENCES

Kabsch and Sander, 1983
Describes the most commonly used method for the assignment of secondary structure from the tertiary protein structure

Jones, 1999
A classic reading that describes the most commonly used PSIPRED method for the prediction of the 3-class SS

Zhang et al., 2011
Provides comprehensive empirical assessment of predictive performance of modern methods for the prediction of secondary structure

Chen and Kurgan, 2013
Provides description and detailed discussion of key architectural details of a large number of modern predictors of secondary structure

Magnan and Baldi, 2014
Describes SSpro, one of the most popular and accurate methods for the prediction of the 8-class SS

INTERNET RESOURCES

http://bioinf.cs.ucl.ac.uk/psipred/
PSIPRED's webserver

http://scratch.proteomics.ics.uci.edu/
SSpro's webserver

http://zhanglab.ccmb.med.umich.edu/PSSpred/.
PSSpred's webserver