Journal of Biomolecular Structure and Dynamics, 2013
http://dx.doi.org/10.1080/07391102.2012.746945

Taylor & Francis
Taylor & Francis Group

Comprehensively designed consensus of standalone secondary structure predictors improves Q_3 by over 3%

Jing Yan, Max Marcus and Lukasz Kurgan*

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

Communicated by Ramaswamy H. Sarma

(*Received 15 August 2012; final version received 2 November 2012*)

Protein fold is defined by a spatial arrangement of three types of secondary structures (SSs) including helices, sheets, and coils/loops. Current methods that predict SS from sequences rely on complex machine learning-derived models andprovide the three-state accuracy (Q_3) at about 82%. Further improvements in predictive quality could be obtained with a consensus-based approach, which so far received limited attention. We perform first-of-its-kind comprehensive design of a SS consensus predictor (SScon), in which we consider 12 modern standalone SS predictors and utilize Support Vector Machine (SVM) to combine their predictions. Using a large benchmark data-set with 10 random training-test splits, we show that a simple, voting-based consensus of carefully selected base methods improves Q_3 by 1.9% when compared to the best single predictor. Use of SVM provides additional 1.4% improvement with the overall Q_3 at 85.6% and segment overlap (SOV_3) at 83.7%, when compared to 82.3 and 80.9%, respectively, obtained by the best individual methods. We also show strong improvements when the consensus is based on ab-initio methods, with Q_3 = 82.3% and SOV_3 = 80.7% that match the results from the best template-based approaches. Our consensus reduces the number of significant errors where helix is confused with a strand, provides particularly good results for short helices and strands, and gives the most accurate estimates of the content of individual SSs in the chain. Case studies are used to visualize the improvements offered by the consensus at the residue level. A web-server and a standalone implementation of SScon are available at http://biomine.ece.ualberta.ca/SSCon/.

Keywords: secondary structure; protein; protein structure; consensus

Introduction

Secondary structure (SS) of proteins is defined as a consecutive fragment of protein sequence that corresponds to a spatially local region in the associated tertiary structure that has distinct geometrical shape. The SS is assigned from protein folds using automated programs (Chen & Kurgan, 2012; Martin et al., 2005). The most popular program, Dictionary of SSs of Proteins (DSSP) (Kabsch & Sander, 1983), assigns each amino acid in a protein sequence to one of eight secondary structure types: α-helix (coded as H), 3_{10}-helix (G), π-helix (I), isolated β-bridge (B), β-sheet (E), hydrogen-bonded turn (T), bend (S) and other structure (‘_’). Typically, these eight types are mapped into three more general states as follows (Eyrich et al., 2003): H, G and I map to helix (H); E and B map to strand (E); and T, S and ‘_’ map to coil (C). The SS can be predicted directly from the protein sequence and most of the predictors consider the abovementioned three states. Knowledge of SS is used to provide structural hierarchy and classification of proteins, such as Structural Classification Of Proteins (SCOP) (Murzin, Brenner, Hubbard, & Chothia, 1995) and protein structure classification at Class, Architecture, Topology and Homologous superfamily levels (CATH) (Orengo et al., 1997). The predicted SS is also adopted for a wide variety of applications including prediction of protein tertiary structure (Hildebrand, Remmert, Biegert, & Söding, 2009; Yang, Faraggi, Zhao, & Zhou, 2011; Wu & Zhang, 2008), solvent accessibility (Garg, Kaur, & Raghava, 2005), folding types (Zhang et al. 2012); identification of post-translational modification (Li, Hu, Niu, Cai, & Chou, 2012) and membrane proteins (Mizianty & Kurgan, 2011); and prediction of protein–protein (Mooney, Pollastri, Shields, & Haslam, 2012) and protein–ligand interactions (Chen, Mizianty, & Kurgan 2012; Lu Wang, Chen, & Zhao, 2012; Zhang et al., 2010), to name a few. These applications are motivated by the fact that a number of well-performing SS predictors were developed over the last few decades. A

*Corresponding author. Email: lkurgan@ece.ualberta.ca recent review discusses 12 SS predictors (Zhang et al., 2011), which include well-known mature methods (Chen & Kurgan, 2012; Pirovano & Heringa, 2010; Rost, 2009) and a selection of new methods that were published in high-impact venues and that are accessible to end users as standalone implementation. The latter allows the users to apply these methods for high throughput and fully automated batch predictions. These predictors can be divided into two classes: (1) methods that utilize homology-based modeling, which include PROTEUS (Montgomerie, Sundararaj, Gallin, & Wishart, 2006; Montgomerie et al., 2008) and SSpro (Cheng, Randall, Sweredoski, & Baldi, 2005; Pollastri, Przybylski, Rost, & Baldi, 2002); and (2) *ab initio* methods such as PHD (Rost, 1996; Rost, Yachdav, & Liu, 2004), PSIPRED (Bryson et al., 2005; Jones, 1999), JNET (Cole, Barber, & Barton, 2008; Cuff & Barton, 2000), SABLE (Adamczak, Porollo, & Meller, 2005), YASPIN (Lin, Simossis, Taylor, & Heringa, 2005), PORTER (Pollastri & McLysaght, 2005), OSSHMM (Martin, Gibrat, & Rodolphe, 2006), SPINE (Dor & Zhou, 2007), P.S.HMM (Won et al., 2007), and SPINEX (Faraggi et al., 2009). The success of the *ab initio* methods primarily comes from the utilization of evolutionary profiles (Rost, 2001), and also from differences in the underlying amino acid composition between different secondary structure states. Figure 1 summarizes these differences; somehow related relation between the composition and protein folding was recently discussed in (Mittal & Jayaram, 2011; Mittal, Jayaram, Shenoy, & Bawa, 2010). According to the comparative survey in Zhang et al. (2011), the best homology-based modeling algorithm is SSpro, which obtains the three-state accuracy (Q_3) at about 82%, and the leading *ab initio* method is SPINEX that has Q_3 around 80.5%. However, there is still room left for further improvements since the theoretical upper limit of the Q_3 accuracy that can be achieved when assigning SS structures from their experimentally determined folds is estimated to be 88% (Rost, 2003). One attractive avenue to improve the predictive quality is to utilize a consensus-based approach that combines results from multiple predictors. This approach received limited attention compared to other related fields, such as prediction of disordered segments (Peng & Kurgan, 2012), protein–ligand interactions (Plewczynski, Łaźniewski, von Grotthuss, Rychlewski, & Ginalski, 2011), and transmembrane topology (Klammer, Messina, Schmitt, & Sonnhammer, 2009), to name a few, where the consensus was shown to provide substantial improvements. The last, related comprehensive study was published in 2003 (Albrecht, Tosatto, Lengauer, & Valle, 2003). The authors considered seven SS predictors and investigated improvements offered by a simple, majority vote-based consensus. This study shows that significant, according to the authors, improvements of about 1.5% in Q_3 can be obtained by implementing the considered consensus of selected three methods. More recent works suffer from a limited scope. The consensus of three SS predictors that are combined with help of a neural network classifier in PROTEUS (Montgomerie et al., 2006, 2008) and ensemble of two predictors in Consensus Data Mining (Cheng, Sen, Jernigan, & Kloczkowski, 2007) were shown to improve predictive quality. In another study, consensus of two SS predictors was found to reduce certain predictive errors (Green, Korenberg, & Aboul-Magd, 2009). These works combine a small set of methods that are selected in an ad hoc fashion, often using in-house predictors.

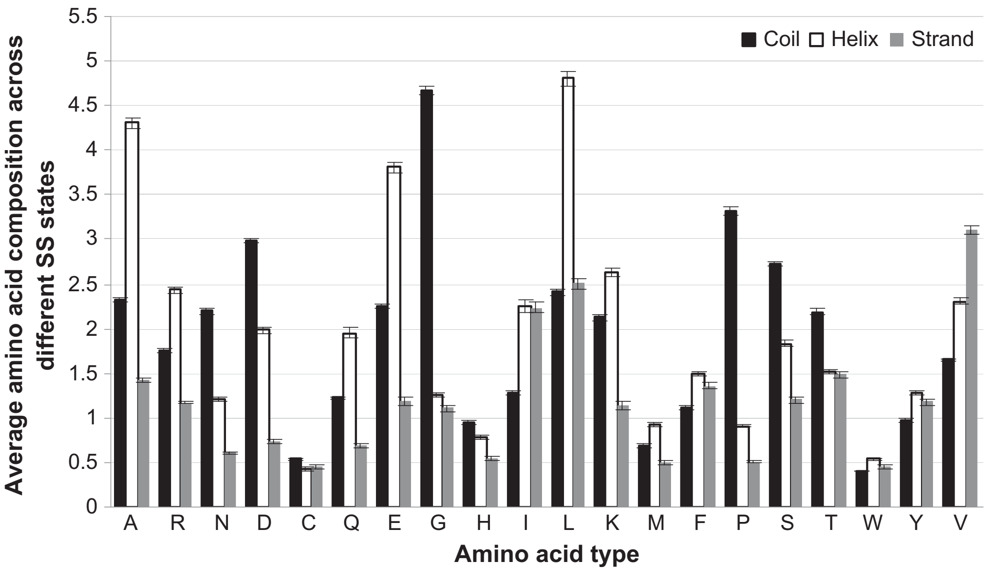

Figure 1. Average amino acid composition across the three SS states computed over 10 different selected test data-sets (see 'Data-sets' section for details). Error bars show the standard deviation of the compositions over the 10 data-sets.

To this end, we design a novel consensus to improve the predictive accuracy. We extend the results by Albrecht et al. (2003) who considered simple majority vote over the seven methods published before 2003. We perform a large-scale investigation of different types of consensuses, utilize a comprehensive list of the above-mentioned 12 modern standalone SS predictors, and take advantage of a more sophisticated model to combine their predictions. Our resulting consensus achieves Q_3 score of 85.6 and 82.3% when using all and only *ab initio* predictors, respectively, resulting in 3.3% and 1.7% improvements. We analyze these improvements at the residue, SS segment, and sequence levels, in the context of the native helix and strand contents, and using two case studies.

Materials and methods

Data-sets

We use a large benchmark data-set with 1975 proteins, which was recently developed in (Zhang et al., 2011), to design and assess the predictive performance of secondary structure predictors, including our consensus approach. These proteins were extracted from PDB (Berman et al., 2000) using PISCES culling server (Wang & Dunbrack, 2003) to include structures solved with high quality and that share low sequence similarity. The pairwise sequence identity in this data-set is below 25%, and each chain has at least 20 residues in length. We randomly split this data-set into two subsets, training set that contains 987 proteins – which are used to design the consensus model, and test set with 988 proteins, which are utilized to perform out-of-sample evaluation and comparison with existing standalone SS predictors. The random split of the data-set into equally sized training/test data-sets is repeated 10 times, and the predictive performance of all considered methods is evaluated using the averaged results obtained from the 10 test sets. As shown in the Results section, results over all splits are similar and thus one particular division was selected to provide detailed insights into the predictive performance of our consensus. The corresponding selected training and test data-sets are available at http://biomine-ws.ece.ualberta.ca/SSCon/training.txt and http://biomine-ws.ece.ualberta.ca/SSCon/test.txt, respectively.

Evaluation protocols

The predictive quality of SS predictors is evaluated using a range of commonly used measures (Zhang et al., 2011), including Q values that were reported in the EVA platform (Eyrich et al., 2001; Koh et al., 2003). They include Q_3, which is defined as a fraction of correct prediction using the three-state SS; Q_{Cobs}, Q_{Hobs}, and Q_{Eobs} that measure the percentage of the correct predictions among all observed native coil, helix, and strand residues, respectively; and Q_{Cpre}, Q_{Hpre}, and Q_{Epre} that quantify the percentage of the correct predictions among all predicted coil, helix and strand residues, respectively. We also compute segment overlap (SOV) values (Zemla, Venclovas, Fidelis, & Rost, 1999), which measure the amount of overlap between the observed and the predicted SS segments. They include SOV_3, which is the average overlap when considering the three SS states, and SOV_C, SOV_H, and SOV_E that estimate overlap for the coil, helix, and strand segments, respectively.

An improvement offered by SScon, when compared against each of the other considered methods, is quantified in terms of Q_3 and SOV_3 scores by calculating average increase in these measures together with the corresponding standard deviations over the 10 random training-test splits. We also evaluate statistical significance of these improvements. The predictive quality measures collected over the 10 runs were tested for normality using Anderson–Darling test (Anderson & Darling, 1952) with .05 significance level; if the measurements follows normal distribution, then we use the paired *t*-test to investigate significance; otherwise the Wilcoxon signed rank test was performed; improvements were assumed significant when $p\text{-value} < .05$. We also perform test of statistical significance of the differences between pairs of predictors for the selected training-test split. We randomly selected 50% proteins from the selected test data-set to calculate the Q_3 and SOV_3 scores for all predictors. This was repeated 10 times and we compared the corresponding 10 paired results using the same procedure as described above.

Simple consensus methods

Motivated by the fact that a simple majority vote consensus of three selected SS predictors was shown to improve predictive quality (Albrecht et al., 2003), we explore a few extended voting strategies by using sequence window, weighting of positions in a window, and ranking. Since, the consensus of three prediction methods was reported to outperform combinations of more than three predictors (Albrecht et al., 2003), we evaluate all triplets of methods selected from among the twelve considered standalone algorithms.

The *Majority Vote* predicts the SS state that corresponds to the most frequent state outputted by three different predictors. In case of a tie (each predictor outputs a different SS state), the state predicted by the algorithm that has the best predictive quality (highest Q_3 value computed on a training data-set) is selected. An extension of the majority voting, named *Window-based Majority Vote* (although in fact it implements plurality vote), uses a local window of size three centered on the predicted residues; use of larger window sizes was found not to improve the predictive quality. This approach is motivated by several studies (Jones, 1999; Madera, Calmus, Thiltgen, Karplus, & Gough, 2010; Montgomerie et al., 2006) that use a window over an initially predicted SS to refine the SS predictions. Residues inside the window are assigned with the Q_3 values of the corresponding input predictors. For each SS state, we sum the Q_3 values assigned to residues in the window of that predicted state, and the consensus predicts the SS state with the highest cumulative Q_3 value. In case of a tie, we choose one of the SS states that are tied that has the highest Q_3 value for the predicted residue (center of the window). We also further extend the window-based voting by weighting positions in the window. The *Weighted Window-based Majority Vote* calculates weighted sum of the Q_3 values where the position in the middle of the window has weight of .5, and the two flaking positions have weights of .25. We also consider two consensuses that utilize ranking instead of voting, based on the Q_3 and SOV_3 values, respectively. The *Ranking on Q_3* method ranks the results from the three input predictors based on Q_C (number of correct predictions for two-class prediction: coil vs. non-coil residues), Q_H (helix vs. non-helix residues), and Q_E (strand vs. non-strand residues) values. The three methods are ranked according to their Q values, computed using a training data-set, for a given SS state. The consensus will output the SS state that corresponds to the predicted SS state that has the best ranking. In case of a two- or three-way tie, we invoke the Majority Vote method. The *Ranking on SOV_3* applies the same procedure but is uses SOV_C, SOV_H and SOV_E instead of the Q values. An example calculation of these five consensuses is shown in Table 1.

We consider two scenarios for each consensus type. The first enumerates all combination of three methods from the considered 12 standalone SS predictors. The second considers all triplets of the 10 *ab initio* methods. The values of the quality measures that are used by these consensuses are estimated based on a training data-set.

Support vector machine (SVM)-based consensus

Our use of SVM is motivated by its recent successful applications in related meta-predictors, including predictors of disordered regions (Mizianty, Stach, Chen, & Kedarisetti, 2010) and transmembrane topology (Klammer et al., 2009). Our SVM-based consensus, named SScon5, uses predictions from five SS predictors including SSpro, PROTEUS, SPINEX, PSIPRED and PORTER, as shown in Figure 2. These five standalone predictors are implemented using different architectures. SSpro uses neural networks and PSI-BLAST-derived profiles (homology analysis). PROTEUS integrates structural alignment with three SS prediction methods (PSIPRED, JNET and TRANSSEC) and applies a jury-of-experts to generate a consensus result. SPINEX is a multistep neural network that couples SS prediction with the prediction of solvent accessibility and backbone torsion angles in an iterative manner. PSIPRED is a two stage neural network that uses PSI-BLAST profiles in the first stage and an initial SS prediction in the second stage. PORTER is based on a two-level ensemble of 45 neural networks. Our SScon5 takes advantage of the architectural differences between these predictors that possibly leads to some complementarity of their predictions, which in turn can be exploited in a consensus. The selection of the five out of the twelve considered predictors is motivated by our results using simple consensus methods (see 'Evaluation of Simple Consensus Methods' Section and Table 3), where we show that these five predictors are consistently selected to provide the best predictive quality. We also designed a second consensus that utilizes three *ab initio* SS predictors: SPINEX, PSIPRED, and PORTER, which is called SScon3.

Our SVM-based consensus predictor consists of three steps, see Figure 2. In the first step, the five (or three) predicted SSs are converted into a vector of numerical features that are fed into the SVM classifier, which in turn predicts one of the three SS states. Three types of features are calculated. They include the three-state SS

Table 1. Example calculation of secondary structure (SS) generated by the considered five consensus-based methods using the three-state SS generated by three input SS predictors. Residues that require a tiebreaker are in bold font. 'X' is used to mark the values of the quality measures (Q_3, Q_C, Q_H, Q_E, SOV_C, SOV_H, and SOV_E) used by a given consensus. The hardcoded values of these quality measures are pre-calculated using a training data-set.

Sequence/consensus type		SS sequence	Q_3	Q_C	Q_H	Q_E	SOV_C	SOV_H	SOV_E
Input predicted SS	Method 1	CCHHHHHCCCCE	78	74	82	73	74	82	78
	Method 2	CCCEHHHHHHEE	79	77	77	75	77	85	75
	Method 3	CCEEEHHCCEEE	80	82	85	78	82	77	73
SS generated by consensus	Majority Vote (MV)	CC**E**EHHHCC**E**EE	X						
	Window-based MV	CCCEHHHHCCEE	X						
	Weighted Window-based MV	CCCEHHH**C**CCEE	X						
	Ranking on Q_3	CCEEEHHCCEEE	X	X	X	X			
	Ranking on SOV_3	CC**EE**HHH**CC**HEE	X				X	X	X

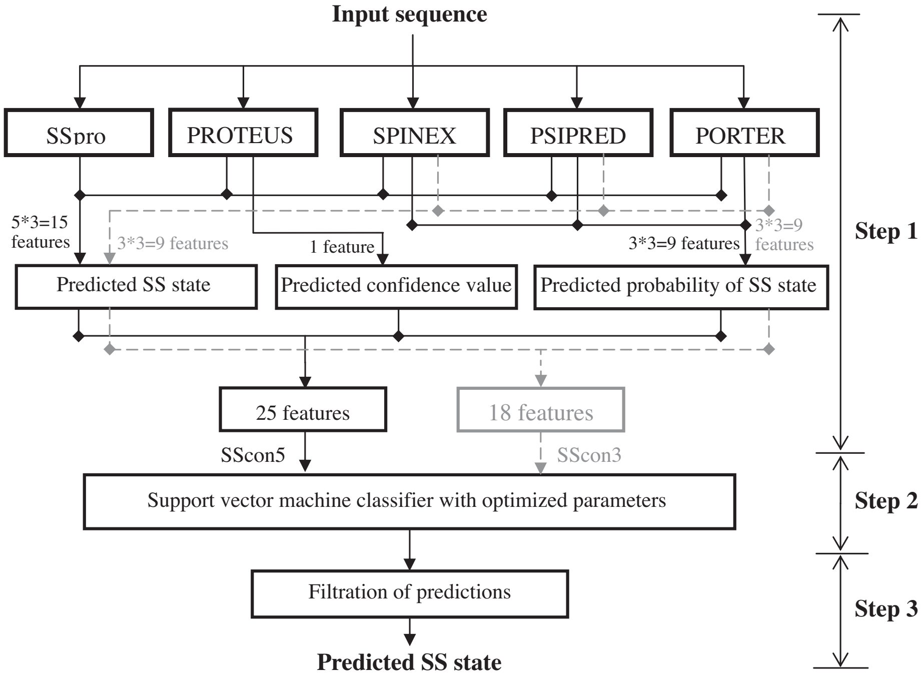

Figure 2. Architecture of our SVM-based consensus predictor. The SScon5 predictor is denoted using solid black lines and SScon3 using dashed gray lines. Boxes with thicker borders describe the standalone predictors utilized in our design.

Table 2. Comparison of predictive quality for default and optimized SVM parameters based on the fivefold cross validation on the selected training data-set.

	SVM parameters		Q_3	
Predictor	Default	Optimized	Default	Optimized
SScon5	$C=1$, $\gamma=.01$	$C=2$, $\gamma=1$	83.9	85.3
SScon3	$C=1$, $\gamma=.01$	$C=1$, $\gamma=4$	80.7	82.2

state predicted by each of the five (three) input predictors, which uses binary encoding, that is, helix (H) is encoded as 001, strand (E) as 010, and coil (C) as 100. We also utilize probability and confidence values that are associated with the predictions; SPINEX, PSIPRED, and PORTER generate probabilities for each of the three SS states, PROTEUS generates a single confidence value, and SSpro generates neither. As a result, SScon5 and SScon3 use 25 and 18 features, respectively. In the second step, we utilize SVM with the Radial Basis Function (RBF) kernel to (re)predict the SS from the 25 (or 18) features. We parameterized the complexity parameter C and the width of the RBF function γ for this classifier using fivefold cross validation on the training data-set. To reduce computational complexity, we randomly selected 20% of residues from the four training folds and we used the entire fifth fold to perform the evaluation. Moreover, we executed parameterization by first fixing C to its default value of 1 and searching for the best performing (that results in highest Q_3) $\gamma=2^n$ where $n=-5$, -4, … ,5. Next, we used the selected to search for the best scoring $C=2^n$ where $n=-5$, -4, … ,5. We run this parameterization for both SScon5 and SScon3. In the third step, we filter the resulting predictions to remove inconsistencies. Specifically, we remove the predicted isolated helical residues, that is, CHC prediction is replaced with CCC, CHE with CCE, EHC with ECC, and EHE with EEE. The SVM model was trained on the whole training data-set using the selected parameters, and this model was tested and compared with other predictors on the corresponding test set.

Our empirical results demonstrate that each of the three steps utilized by SScon5 (SScon3) provides improvements, when tested based on the fivefold cross validation on the selected training data-sets. In the first step, using the 25 (18) features which were fed into the SVM classifier with default parameters, our SScon5 (SScon3) improves Q_3 by 1.6% (.2%) compared with the best-homology based (*ab initio*) method. In the second step, parameterization of the SVM classifier provides

Table 3. Comparison of predictive quality on the selected test data-set for the five considered consensus-based SS predictions designed using the best performing (based on Q_3 on the training data-set) triplets of SS predictors selected from the 12 standalone SS predictors and from the 10 *ab initio* standalone SS predictors. The selected best methods from each group are shown in bold font.

SS predictors	Consensus type	Best triplet of SS predictors	Q_3	Q_{Cobs}	Q_{Hobs}	Q_{Eobs}	Q_{Cpre}	Q_{Hpre}	Q_{Epre}	SOV_3	SOV_C	SOV_H	SOV_E
12 standalone SS predictors	**Majority Vote (MV)**	SSpro, PROTEUS, SPINEX	**84.4**	**83.3**	**88.7**	**78.8**	**79.7**	**89.3**	**84.4**	**82.5**	**78.5**	**83.0**	**78.9**
	Window-based MV	SSpro, PROTEUS, SPINEX	83.7	82.7	87.8	78.6	78.5	89.3	83.9	81.8	77.1	81.5	78.8
	Weighted window-based MV	SSpro, PROTEUS, SPINEX	84.0	83.1	87.9	79.0	78.9	89.6	83.9	82.4	77.9	82.2	79.0
	Ranking on Q_3	SSpro, PROTEUS, SPINEX	83.8	84.7	88.7	74.1	78.1	89.3	85.6	81.5	77.4	83.0	77.5
	Ranking on SOV_3	SSpro, PROTEUS, SPINEX	83.5	83.0	86.7	78.8	78.2	88.6	84.4	80.8	76.8	82.3	78.9
10 *ab initio* standalone SS predictors	**Majority Vote (MV)**	PSIPRED, PORTER, SPINEX	**81.7**	**82.6**	**86.2**	**72.5**	**75.5**	**87.9**	**83.2**	**79.8**	**75.7**	**80.9**	**74.8**
	Window-based MV	PSIPRED, PROTER, SPINEX	81.2	82.3	85.9	71.5	74.8	87.5	83.2	79.1	74.2	79.9	74.4
	Weighted window-based MV	PSIPRED, PORTER, SPINEX	81.5	82.5	86.1	72.1	75.2	87.6	83.3	79.5	74.8	80.3	74.8
	Ranking on Q_3	PSIPRED, PORTER, SPINEX	81.7	82.6	86.2	72.5	75.5	87.9	83.1	79.8	75.7	81.0	74.9
	Ranking on SOV_3	SPINE, PORTER, SPINEX	81.1	82.1	85.8	71.3	74.9	87.3	82.3	78.9	75.1	80.1	74.3

further improvement of 1.4% (total improvement is 1.6% + 1.4% = 3.0%) and 1.5% (total 1.5% + .2% = 1.7%) for the SScon5 and SScon3, respectively; see Table 2. In the last step, the filtration of inconsistent predictions provides additional slight increase in Q_3 by .02 and .01%, respectively.

Results

Evaluation of simple consensus methods

Motivated by results in Albrecht et al. (2003), we select the best triplet of input SS predictors for each of the five considered types of simple voting/ranking-based consensuses using the selected training data-set. The best set of methods is selected based on Q_3 values; we use SOV_3 in case of similar best Q_3 values (within .01). We consider two cases, when using all 12 standalone SS predictors and when selecting from the 10 *ab initio* predictors. We use the quality measures estimated based on the training data-set when implementing the consensuses. The results on the corresponding selected test data-set are summarized in Table 3.

We observe that the simplest majority vote consensus outperforms other voting- and ranking-based consensuses, which means that use of a local window in the predicted SS and ranking does not lead to improvements. This suggests that when the triplet of the input SS predictors is carefully selected, the improvements are due to complementarity between these methods and the predictions at a given residue are sufficient to compute a well-performing consensus. On the other hand, use of ranking biases the consensus toward the best performing method, which potentially limits gains due to the complementarity between methods.

The selected triplets of SS predictors are consistent across different consensus types. The methods selected from among the 12 SS predictors include SSpro, PROTEUS, and SPINEX. These methods are characterized by the highest overall Q_3 values (averaged over 10 runs of random training-test split), which are given in Table 4. Moreover, their predictive performance across the three SS states is complementary, that is, SSpro obtains high Q_{Cobs} and Q_{Epre}, PROTEUS has high Q_{Eobs}, Q_{Cpre}, and Q_{Hpre}, and SPINEX achieves high Q_{Hobs}. When

Table 4. Comparison of average predictive quality on the 10 randomly selected test data-set for three groups of predictors: SVM-based consensuses (SScon5 and SScon3), the best simple majority vote (MV) consensuses using all (MV_{all}) and *ab initio* only ($MV_{abinitio}$) SS predictors, and the 12 standalone algorithms. The best values for each quality index are shown in bold font, and methods are sorted within each group in the descending order by Q_3 values. 'Impr' indicates the average improvements of our SScon5 (SScon3) predictor against other considered methods over the 10 test data-sets, and 'stdv' describes the corresponding standard deviations of these improvements. '+'/ '-'/'=' denote that SScon5 (or SScon3) is statistically significant better/worse/not different with $p < .05$ than another method.

			Improvement in Q_3 by											Improvement in SOV_3 by						
			SScon5		SScon3									SScon5		SScon3				
SS predictor types	Predictors	Q_3	Impr ± stdv	Sig	Impr ± stdv	Sig	Q_{Cobs}	Q_{Hobs}	Q_{Eobs}	Q_{Cpre}	Q_{Hpre}	Q_{Epre}	SOV_3	Impr ± stdv	Sig	Impr ± stdv	Sig	SOV_C	SOV_H	SOV_E
SVM-based consensuses	SScon5	**85.6**			−3.3 ± .1	–	83.7	**90.6**	80.6	**81.2**	90.3	**85.4**	**83.7**			−3.0 ± .3	–	**79.4**	**85.6**	**81.2**
	SScon3	82.3	3.3 ± .1	+			81.6	86.9	76.1	77.2	88.0	82.2	80.7	3.0 ± .3	+			76.3	81.3	77.8
Simple MV consensuses	MV_{all}	84.2	1.4 ± .1	+	−1.9 ± .1	–	83.5	88.5	78.5	79.4	89.3	84.6	82.4	1.3 ± .2	+	−1.7 ± .1	–	78.8	82.7	79.6
	$MV_{abinitio}$	81.7	3.9 ± .2	+	.7 ± .0	+	82.8	86.2	72.4	75.4	87.9	83.5	79.7	4.0 ± .3	+	1.0 ± .1	+	75.9	80.6	75.5
Standalone predictors	SSpro	82.3	3.3 ± .1	+	.0 ± .2	=	**84.5**	86.1	72.5	76.4	87.8	84.7	79.6	4.1 ± .2	+	1.1 ± .3	+	76.4	81.5	77.3
	PROTEUS	81.8	3.8 ± .1	+	.5 ± .2	+	74.3	86.6	**86.4**	79.0	**91.2**	73.1	80.9	2.8 ± .2	+	−.2 ± .3	=	74.1	81.9	**81.2**
	SPINEX	80.6	5.0 ± .1	+	1.7 ± .1	+	79.3	87.4	71.8	75.2	85.8	81.7	77.3	6.4 ± .2	+	3.4 ± .1	+	72.9	80.0	72.4
	PSIPRED	80.1	5.5 ± .2	+	2.2 ± .1	+	79.9	84.9	72.5	75.1	85.4	80.3	78.3	5.4 ± .4	+	2.4 ± .2	+	73.6	80.2	75.1
	SPINE	79.1	6.4 ± .2	+	3.2 ± .1	+	80.2	83.7	69.9	73.8	85.1	79.4	76.6	7.1 ± .3	+	4.1 ± .2	+	73.1	78.0	74.1
	PORTER	78.7	6.9 ± .2	+	3.6 ± .1	+	81.4	83.1	67.1	71.9	86.2	79.8	77.1	6.6 ± .2	+	3.6 ± .1	+	73.5	78.2	71.8
	SABLE	78.0	7.6 ± .2	+	4.3 ± .1	+	82.4	80.7	66.3	70.6	86.5	79.8	75.6	8.1 ± .3	+	5.1 ± .2	+	71.9	76.8	73.0
	YASPIN	76.5	9.1 ± .3	+	5.8 ± .2	+	72.9	81.2	74.6	73.9	83.9	69.5	75.2	8.5 ± .4	+	5.5 ± .3	+	69.8	75.8	75.0
	OSSHMM	74.5	11.1 ± .3	+	7.8 ± .2	+	83.0	78.2	54.4	66.9	82.4	79.5	71.4	12.3 ± .3	+	9.3 ± .3	+	70.0	73.7	66.3
	JNET	74.3	11.3 ± .3	+	8.0 ± .2	+	82.7	74.4	60.0	66.3	84.1	77.5	71.4	12.3 ± .3	+	9.3 ± .2	+	69.1	70.4	68.8
	P.S.HMM	68.5	17.1 ± .2	+	13.9 ± .2	+	74.5	72.6	51.6	65.2	74.0	65.0	65.9	17.8 ± .3	+	14.8 ± .2	+	66.9	67.1	60.4
	PHD	68.0	17.6 ± .2	+	14.4 ± .2	+	73.1	71.7	53.2	64.9	73.6	64.0	66.3	17.4 ± .3	+	14.4 ± .2	+	65.0	67.6	60.6

excluding the homology-based predictors, the selected triplet consists of SPINEX, PSIPRED, and PORTER; these methods are ranked 1st, 2nd, and 4th according to their Q_3. The likely reason why 3rd best SPINE was excluded is that it is similar to the SPINEX. Once again, these methods complement each other, that is, SPINEX outperforms the other methods in Q_{Hobs} and Q_{Epre}, PSIPRED does well in Q_{Eobs}, while PORTER provides high Q_{Cobs} and Q_{Hpre}. This complementarity and consistency motivated our choice of these five methods (note that SPINEX is selected in both cases) to implement the SVM-based consensuses.

The majority vote-based consensuses of these selected triplets of SS predictors provide relatively good improvements over the individual predictors; similar results were also observed in Albrecht et al. (2003). As shown in Table 4, which is based on the results averaged over the 10 randomly selected test data-sets, the combination of SSpro, PROTEUS, and SPINEX improves Q_3 by 1.9% compared to the best Q_3 of the individual predictor SSpro and SOV_3 by 1.5% compared to the best SOV_3 of PROTEUS. Considering the theoretical limit of Q_3 at 88% (Rost, 2003), this improvement corresponds to 100% × (84.2–82.3)/(88–82.3) = 33.3% of the possible improvement in Q_3. Similarly, consensus of the three *ab initio* methods provides 1.1% improvement in Q_3 (that corresponds to 14.9% of the possible improvement) and 1.4% in SOV_3 compared to the best performing SPINEX in Q_3 and PSIPRED in SOV_3, respectively. This shows consistency of the improvements, which hold at both residue and segment levels.

To summarize, the majority vote-based consensuses of three well-performing and complementary SS predictors outperform other considered simple consensus- and ranking-based approaches. Next, we compare these consensuses with a more sophisticated SVM-based solution.

Comparison of consensus and standalone predictors

We randomly split the benchmark data-set 10 times into two equal sized parts, training and test data-sets. The pairwise sequence similarity between these data-sets is below 25%. For each pair of the data-sets, we built our SVM-based consensus on the training data-set and then evaluated it on the corresponding test data-set. Table 4 compares the average (over the 10 test data-sets) prediction quality of the two versions of the SVM-based consensuses, SScon5 and SScon3, the two best simple majority vote (MV) consensuses which are selected from all, named $\mathrm{MV}_{\mathrm{all}}$, and *ab initio* only, named $\mathrm{MV}_{\mathrm{abinitio}}$, SS predictors, and the 12 considered standalone algorithms.

Considering the standalone algorithms, SSpro achieves the highest Q_3 of 82.3% and SPINEX is the best among the *ab initio* algorithms with the Q_3 of 80.6%. The SScon5 and SScon3 substantially improve the predictive quality when compared with the above top single predictors. SScon5 obtains $Q_3 = 85.6\%$ and $\mathrm{SOV}_3 = 83.7\%$, which improve the corresponding values obtained by the best majority vote consensus $\mathrm{MV}_{\mathrm{all}}$ by (average over the 10 test sets ± the corresponding standard deviation) 1.4% ± .1% and by 1.3% ± .2%, respectively. The total improvement over the best standalone predictor is 3.3% ± .1% in Q_3 and 2.8% ± .2% in SOV_3; this translates into 100%(85.6–82.3)/(88–82.3 = 58% of the possible improvement in Q_3. The improvements obtained with SScon5 are statistically significant in both Q_3 and SOV_3 when compared with all other considered approaches over the 10 test data-sets. SScon3 that utilizes only *ab initio* methods achieves $Q_3 = 82.3\%$ and $\mathrm{SOV}_3 = 80.7\%$, which is higher than the corresponding best majority vote consensus $\mathrm{MV}_{\mathrm{abinitio}}$ by .7% ± .05% in Q_3 and by 1.0% ± .1% in SOV_3, respectively. The total improvement in Q_3 over the best *ab initio* method, which is 1.7% ± .1%, equals to 100% × (82.3–80.6)/(88 – 80.6) = 23% of the possible improvement in $Q_{3.}$ SScon3 significantly outperforms all considered *ab initio* solutions in both Q_3 and SOV_3 and provides predictions that are equivalent in the predictive performance to the predictions from the homology modeling-based approaches.

Importantly, the improvements offered by SScon5 are consistent across the three SS states, that is, the $\mathrm{SOV}_{\mathrm{H}}$, $\mathrm{SOV}_{\mathrm{E}}$, and $\mathrm{SOV}_{\mathrm{C}}$ values of SScon5 are higher than the corresponding values of any the other considered methods. SScon5 also shows improvements in terms of Q_{Hobs}, Q_{Cpre} and Q_{Epre}. We emphasize the relatively large increases compared to the standalone methods in $\mathrm{SOV}_{\mathrm{H}}$ by 3.7% and in Q_{Hobs} by 3.2%, which imply that our SScon5 predictor is able to correctly predict more helical segments. Although Q_{Cobs} score of SScon5 is lower by .8% than the highest Q_{Cobs} achieved by SSpro, our consensus obtains Q_{Cpre} that is higher by 4.8%. This suggests that SSpro overpredicts coil residues compared to SScon5. Similarly, SScon5 has a 5.8% deficit in Q_{Eobs} that is coupled with a large 12.3% improvement in Q_{Epre} when compared with PROTEUS; this implies that PROTEUS overpredicts strand residues compared to SScon5.

SScon3 provides improved values of Q_{Eobs}, Q_{Cpre}, Q_{Hpre} and all SOV scores ($\mathrm{SOV}_{\mathrm{C}}$, $\mathrm{SOV}_{\mathrm{H}}$, and $\mathrm{SOV}_{\mathrm{E}}$) compared with the corresponding $\mathrm{MV}_{\mathrm{abinitio}}$ consensus and standalone *ab initio* algorithms. Although SScon3 has Q_{Cobs} and Q_{Epre} lower by 1.2% and 1.3%, respectively, compared to the best $\mathrm{MV}_{\mathrm{abinitio}}$, its Q_{Cpre} and Q_{Eobs} are higher by 1.8% and 3.7%, respectively. Similarly, the drop by .5% in Q_{Hobs} compared to the best SPINEX is matched with the 2.2% increase in Q_{Hpre}. The above shows that SScon3 provides consistent improvements over the three SS states.

Next, we investigate the predictive quality based on a single split into the selected training and test data-sets; see Table 5. We observe that the improvements offered by our consensuses, when compared with the other

Table 5. Comparison of predictive quality on the selected test data-set for three groups of predictors: SVM-based consensuses (SScon5 and SScon3), the best simple majority vote (MV) consensuses using all (MV_{all}) and *ab initio* only ($MV_{abinitio}$) SS predictors, and the 12 standalone algorithms. The best values for each quality index are shown in bold font and methods are sorted within each group in the descending order by Q_3 values. 'Impr' indicates the average improvements of our SScon5 (SScon3) predictor against other considered methods over the 10 randomly selected subsets of 50% of test chains, and 'stdv' describes the corresponding standard deviations of these improvements. '+'/ '-'/'=' denote that SScon5 (or SScon3) is statistically significant better/worse/not different with $p < .05$ than another method.

			Improvement in Q_3 by											Improvement in SOV_3 by						
			SScon5		SScon3									SScon5		SScon3				
SS predictor types	Predictors	Q_3	Impr ± stdv	Sig	Impr ± stdv	Sig	Q_{Cobs}	Q_{Hobs}	Q_{Eobs}	Q_{Cpre}	Q_{Hpre}	Q_{Epre}	SOV_3	Impr ± stdv	Sig	Impr ± stdv	Sig	SOV_C	SOV_H	SOV_E
SVM-based	SScon5	**85.8**			−3.3 ± .2	–	83.3	**91.0**	81.0	**81.9**	90.0	**85.2**	**83.7**			−2.9 ± .2	–	**79.1**	**86.2**	**80.6**
consensuses	SScon3	82.4	3.3 ± .2	+			80.8	87.3	76.9	77.8	87.9	81.3	80.8	2.9 ± .2	+			75.9	81.7	77.7
Simple MV	MV_{all}	84.4	1.3 ± .0	+	−2.0 ± .2	–	83.3	88.7	78.8	79.7	89.3	84.4	82.5	1.3 ± .2	+	−1.6 ± .2	–	78.5	83.0	78.9
consensuses	$MV_{abinitio}$	81.7	4.0 ± .2	+	.7 ± .1	+	82.6	86.2	72.5	75.5	87.9	83.2	79.8	4.0 ± .2	+	1.1 ± .1	+	75.7	80.9	74.8
Standalone predictors	SSpro	82.4	3.2 ± .2	+	−.1 ± .3	=	**84.3**	86.1	72.8	76.7	87.7	84.4	79.2	4.3 ± .3	+	1.4 ± .3	+	75.8	81.4	76.5
	PROTEUS	81.8	4.0 ± .3	+	.6 ± .3	+	74.2	86.8	**86.5**	79.3	**91.2**	72.6	80.7	3.1 ± .2	+	.2 ± .3	=	73.7	82.8	80.3
	SPINEX	80.8	5.0 ± .2	+	1.6 ± .1	+	79.1	87.5	72.3	75.5	85.9	81.4	77.6	6.2 ± .3	+	3.3 ± .2	+	73.0	80.1	71.9
	PSIPRED	80.0	5.8 ± .2	+	2.4 ± .1	+	79.6	84.7	72.4	75.1	85.3	79.8	78.0	5.7 ± .2	+	2.9 ± .2	+	73.1	80.1	74.3
	SPINE	79.0	7.0 ± .2	+	3.7 ± .2	+	79.9	83.4	70.0	73.7	85.1	78.6	76.3	6.7 ± .2	+	3.8 ± .3	+	72.6	78.2	73.5
	PORTER	78.8	6.7 ± .2	+	3.4 ± .1	+	81.3	83.2	67.0	72.1	86.2	79.5	77.0	7.4 ± .3	+	4.5 ± .3	+	73.0	78.4	70.9
	SABLE	77.9	7.8 ± .2	+	4.5 ± .1	+	82.1	80.5	66.4	70.5	86.6	79.1	75.5	8.3 ± .2	+	5.4 ± .3	+	71.5	77.0	72.3
	YASPIN	76.2	9.5 ± .3	+	6.2 ± .2	+	72.5	80.9	74.5	73.9	83.7	68.3	74.7	8.9 ± .3	+	6.1 ± .3	+	69.2	75.8	74.2
	OSSHMM	74.5	11.4 ± .3	+	8.0 ± .2	+	82.5	78.1	54.7	67.1	82.4	78.6	71.3	12.4 ± .3	+	9.6 ± .3	+	69.5	73.9	65.7
	JNET	74.2	11.6 ± .2	+	8.3 ± .2	+	82.5	74.1	60.0	66.3	84.1	76.4	71.3	12.3 ± .2	+	9.4 ± .1	+	68.8	70.4	68.2
	P.S.HMM	68.5	17.3 ± .3	+	14.0 ± .2	+	74.2	72.7	51.6	65.4	74.2	64.1	65.9	17.8 ± .4	+	14.9 ± .3	+	67.0	67.0	60.1
	PHD	68.0	17.8 ± .3	+	14.5 ± .2	+	73.1	71.4	53.3	64.9	74.0	63.3	66.1	17.6 ± .4	+	14.8 ± .3	+	64.7	67.2	59.9

considered predictors, are consistent with the average improvements over the 10 randomly selected test data-sets; see Tables 4 and 5. Hence, we use the results on the selected test data-set in the subsequent sections; this simplifies our analysis without the loss of generality.

Overall, our analysis reveals that the SVM-based consensus substantially improves the prediction quality compared to both standalone methods and simpler consensuses and that these improvements are maintained over the three SS states.

Next, we analyze these results in more detail at the residue, SS segment, and sequence levels using the selected test data-set.

Predictions at residue level

We investigate the predictive quality at the residue level considering the eight SS states that are defined in DSSP. We assume that a given one of the eight SS states is correctly predicted if it matches the corresponding predicted three-state SS, for example, H, G or I states are assumed correct if they are predicted as H. Table 6 shows the Q_{obs} values for three states: α-helix (H), β-bridge (B), and β-sheet (E); the remaining states are either infrequent (G and I) or of potentially lesser interest (coils). SScon5 outperforms the other methods by at least 1.4% in Q_{obs} for α-helices, 1.2% for β-bridges, and 2.1% for β-sheets, except for PROTEUS that has higher values for β-bridge and β-sheet residues. However, PROTEUS in general overpredicts strand residues, that is, it has $Q_{\text{Eobs}} = 86.5$ and $Q_{\text{Epre}} = 72.6$ compared with $Q_{\text{Eobs}} = 81$ and $Q_{\text{Epre}} = 85.2$ for SScon5 (see Table 5), which explains these results. Similarly, SScon3 improves Q_{obs} for β-bridge and β-sheet residues compared to the corresponding $\text{MV}_{\text{abinitio}}$ and standalone *ab initio* methods.

We also analyze the abundance of significant errors where helix is predicted as a strand or vice versa; see the last row in Table 6. We calculate a fraction of these mispredictions among all residues in the selected test data-set. The results reveal that the best standalone method PROTEUS makes these errors for .64% of residues. Although the majority vote-based consensuses and SScon3 fail to improve over this error rate, SScon5 reduces this number to .62% while providing substantially higher overall predictive accuracy (see Table 5). The rate of these ‘strong’ errors obtained by SScon5 is substantially lower, by at least 47%, compared to the other standalone methods.

Predictions at secondary structure segment level

We study the predictive quality at the SS segment level. We assume that a given helix or strand segment (excluding β-bridges) is correctly predicted if at least 50% residues in that segment are predicted correctly. Figure 3 compares eight predictors, including both SVM-based consensuses (SScon5 and SScon3), the two majority vote-based combinations (MV_{all} and $\text{MV}_{\text{abinitio}}$), and the five best standalone SS predictors (SSpro, PROTEUS, SPINEX, PSIPRED, and PORTER), for prediction of short and long SS segments. We define a short helical/strand segment as having 8/6 or less consecutive residues; otherwise the segment is considered to be long. The two cut-offs are selected to best highlight improvements offered by SScon5.

Overall, as it was shown in Zhang et al. (2011), longer helices are easier to predict, see Figure 3(A). SScon5 outperforms the other solutions in the prediction of helical segments. In particular, it provides larger improvements, between 4.3 and 14.5% higher success rates, for the harder to predict short segments. Although PROTEUS achieves the best performance on the strand segments, as mentioned before, it also overpredicts the strand residues. Excluding PROTEUS, our SScon5 provides stronger predictions for the strand segments, again with a larger magnitude for the shorter strands. The improvements range between 7.9 and 14.5%, see Figure 3 (B). SScon3 also provides improvements for both short and long strand segments when compared with the corresponding *ab initio* approaches.

Predictions at sequence level

We analyze the predictive performance at the sequence level by comparing the overall predicted content of each of the three SS states. The content is defined as a fraction of residues in a helix, strand, and coil SS states.

Table 6. Comparison of predictive quality on the selected test data-set at the residue level considering the α-helix, β-bridge, and β-sheet SS states, and the misclassifications between helix and strand residues. The first three rows show the Q_{obs} values for α-helices (% of α-helices of predicted as H), β-bridges (% of β-bridges predicted as E), and β-sheets (% of β-sheets predicted as E). The last row shows the fraction of predictions where helix is classified as strand and vice versa. Best results are shown in bold font.

	SVM-based consensuses		Simple consensuses		Standalone predictors				
	SScon5	SScon3	MV_{all}	$\text{MV}_{\text{abinitio}}$	SSpro	PROTEUS	SPINEX	PSIPRED	PORTER
α-helix (H)	**95.9**	92.5	94.5	91.7	90.1	93.4	93.1	89.4	88.6
β-bridge (B)	**36.2**	23.0	29.8	17.9	35.0	**48.2**	18.3	20.7	18.1
β-sheet (E)	**83.3**	79.6	81.2	75.3	74.7	**88.4**	75.0	75.1	69.5
Mispredictions helix ↔ strand	**.62**	1.21	.95	1.15	1.61	.64	1.18	1.89	1.77

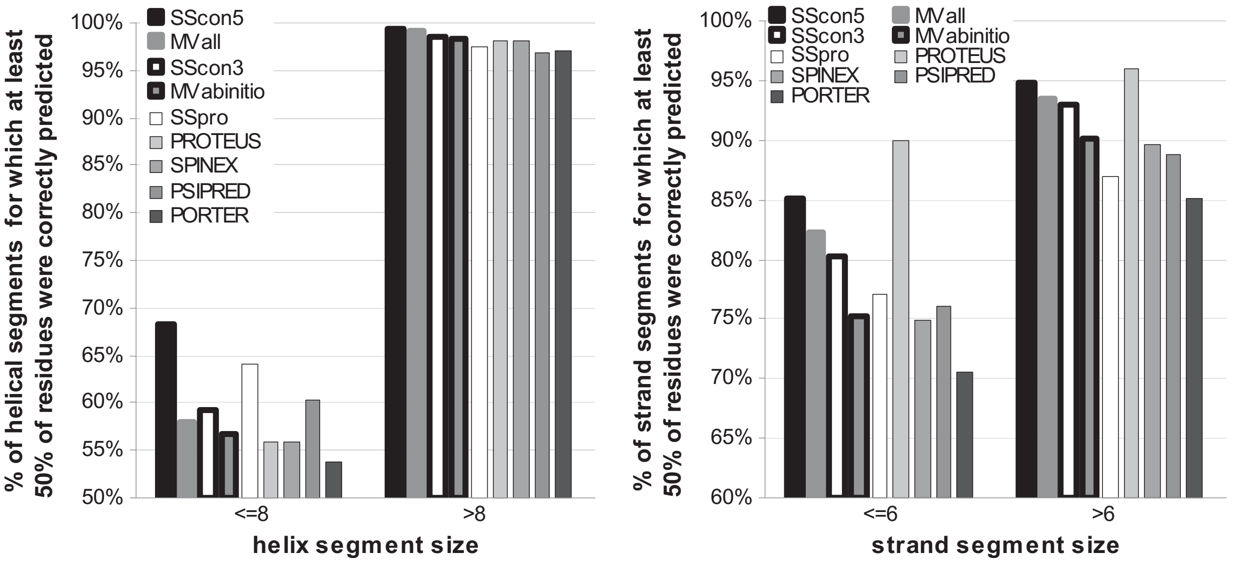

Figure 3. Comparison of predictive quality on the selected test data-set at the segment level. The bars represent fractions of short (segment size ⩽8) and long (>8) helix segments (panel A), and short (⩽6) and long (>6) strand segments (panel B), for which at least 50% residues are correctly predicted.

Table 7. Comparison of the predictive quality on the selected test data-set at the sequence level. The native, defined with DSSP, and predicted SS content and the corresponding average (across the 3 SS states) absolute differences are shown. The best results are shown in bold font.

		SVM-based consensuses		Simple consensuses		Standalone methods				
SS state	Native content	SScon5	SScon3	MV$_{\text{all}}$	MV$_{\text{abinitio}}$	SSpro	PROTEUS	SPINEX	PSIPRED	PORTER
Helix	.386	.390	**.383**	**.383**	.378	.378	.367	.393	**.383**	.372
Strand	.227	**.216**	.214	.211	.198	.196	.270	.201	.206	.191
Coil	.388	**.394**	.403	.405	.424	.426	.362	.406	.411	.437
Average absolute difference	*n*/a	**.007**	.010	.012	.024	.026	.029	.017	.016	.033

We compute the absolute difference of the SS content predicted by the considered standalone and consensus predictors and the native content computed with DSSP. The average of the three absolute differences and the differences for each SS state on the selected test data-set are compared in Table 7. SScon5 provides predictions with the most accurate SS content values for strands and coils and the overall smallest average (across the three SS states) error in content, which is at .7%. This means that the predictions from our consensus properly balance the amount of helix, strand, and coil conformation. To compare, the selected best performing individual SS predictors have the average errors between 1.6 and 3.3%. Similarly, SScon3 has low average error at 1%, while the corresponding *ab initio* solutions have errors at 1.6% or higher. We also note the overprediction of strands by PROTEUS, which we mentioned in previous sections; PROTEUS predicts on average 27% of residues in strand conformation while there are 22.7% based on the native annotations. The standalone predictors, except for PROTEUS, generally overpredict the coil residues. We note that SS content can be computed using specialized predictors, which are faster to compute as they usually do not require the calculation of multiple sequence alignment. However, this comes as a trade-off for a lower predictive quality (Chen, Stach, Homaeian, & Kurgan, 2011; Homaeian, Kurgan, Ruan, Cios, & Chen, 2007).

Mapping of improvements on two-dimensional strand vs. helix content space

We analyze the improvement in Q_3 offered by SScon5 and SScon3 against the corresponding simple majority vote consensuses, MV$_{\text{all}}$ and MV$_{\text{abinitio}}$, respectively, and the best standalone methods, SSpro and SPINEX, respectively. We map these improvements into two-dimensional space defined by the native amount of helix (H$_{\text{content}}$) and strand (E$_{\text{content}}$) residues. The space is binned based on the two content values to have as even as possible number of protein chains; see Table 8. For each bin, we calculate the difference in average Q$_3$ score between a given SVM-based consensus and another method (majority vote consensus or best standalone method). A heat map of

Table 8. Two-dimensional space defined by the binned native amount of helix ($H_{content}$) and strand ($E_{content}$) residues. The values correspond to the percentage of proteins from the selected test data-set in a given bin.

	$E_{content}$				
$H_{content}$	<.2 (%)	.2–.25 (%)	.25–.35 (%)	.35–.45 (%)	>.45 (%)
>.7	11.1	.0	.0	.0	.0
.5–.7	13.5	.0	.0	.0	.0
.375–.5	17.1	7.4	4.3	.0	.0
<37.5	4.4	5.5	14.7	11.8	10.1

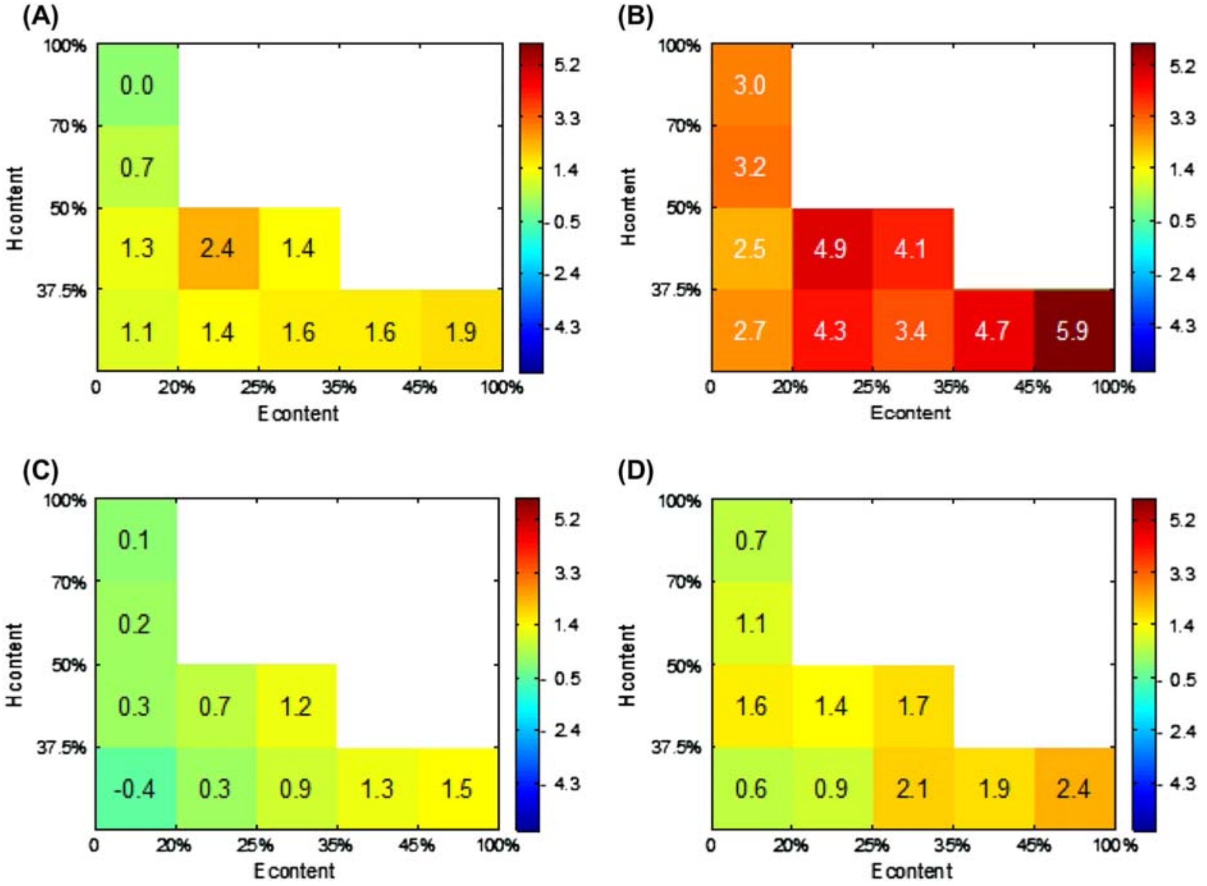

Figure 4. Improvements in Q_3 evaluated on the selected test data-set and mapped into two-dimensional space defined by strand ($E_{content}$) and helix content ($H_{content}$) offered by: SScon5 compared to majority vote-based consensus MV_{all} (panel A); SScon5 compared to the best standalone predictor SSpro (panel B); SScon3 compared to the majority vote-based consensus of *ab initio* methods $MV_{abinitio}$ (panel C); SScon3 compared to the best standalone *ab initio* predictor SPINEX (panel D). White color denotes cells with no proteins.

these differences in the two-dimensional space is given in Figure 4. SScon5 offers improved predictions compared to both the corresponding simple consensus and the standalone method across the entire two-dimensional space. Larger improvements are for proteins with medium to high strand content (over 20% of strand residues) and medium to low helix content (50% of less helix residues); see Figure 4(A) and (B). Similar observations are also true for SScon3. This *ab initio* only consensus provides higher magnitude of improvements for chains with over 25% of strand residues and below 50% helix residues; see Figure 4(C) and (D).

Case studies

We use two case studies selected from the selected test data-set to illustrate mechanisms that lead to the improvements offered by our SScon5 and SScon3

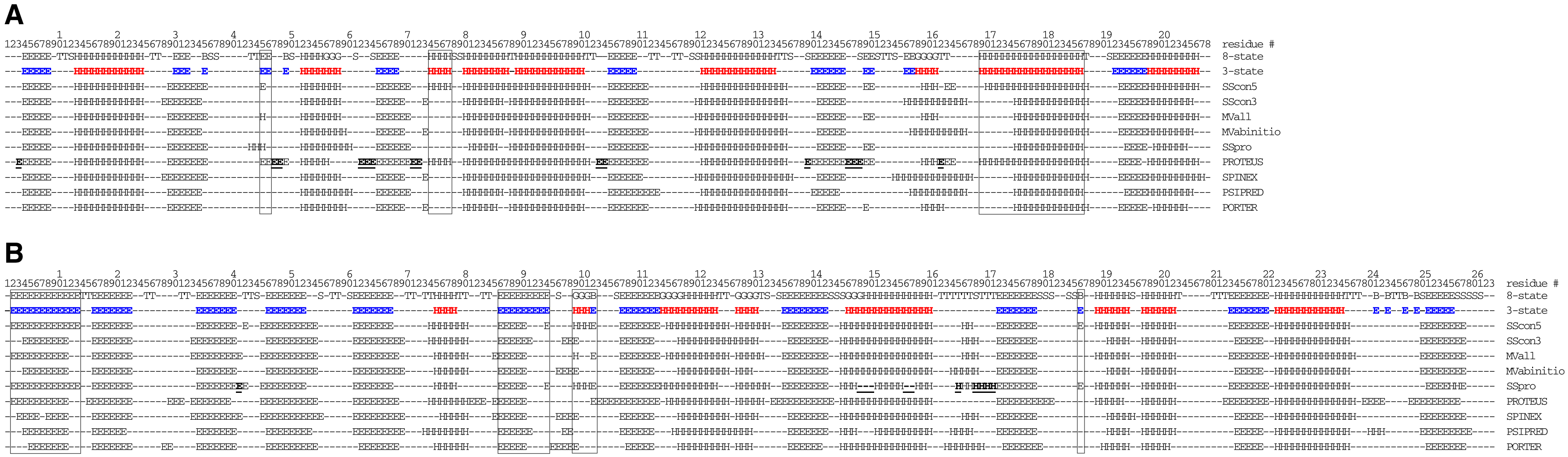

Figure 5. Comparison of native and predicted SS for PA4017 protein (PDBid: 2A35 chain A) (Panel A) and ferredoxin-NADP(H) reductase (PDBid: 2BGI chain A) (Panel B). Consecutive line show the residue number, native eight-state SS assigned with DSSP, three-state native SS assigned with DSSP, predictions from SScon5, SScon3, majority vote-based consensuses MV_{all} and $MV_{abinitio}$, and the top five standalone predictors: SSpro, PROTEUS, SPINEX, PSIPRED, and PORTER. The native three-state SS is color coded where ‘H’ denotes helices, ‘E’ denotes strands and ‘–‘ denotes coils. Boxes highlight segments where SScon provided interesting improvements. Mispredictions that are discussed in text are underlined and shown in bold font.

predictors. We focus on the predictions of strand and helix segments and compare predictions from the two SVM-based consensuses, the two majority vote-based consensuses, and the 5 top-scoring standalone SS predictors.

The first case study is the PA4017 protein (PDBid: 2A35 chain A) that has 208 amino acids; see Figure 5 (A). SScon5 and PROTEUS outperform the other considered predictors; they predict all strand and helix segments. The three boxes regions (residues 45–46, 74–77, and 169–186) show where SScon5 'borrows' from the predictions of PROTEUS, which is one of the inputs to SScon5; this leads to the strong predictive performance of our consensus. At the same time, predictions of PROTEUS that are shown in underline and bold font indicate the strand residues that are overpredicted by this method but correctly predicted by SScon5. This is possible based on the predictions from the other methods used in our consensus. The example shows that the SVM-based meta-predictor, in contrast to the majority vote consensus (MV_{all}) that makes mistakes in the three boxes regions, efficiently combines the input SS predictions without relying on the majority.

The second case study concerns 263 residues long ferredoxin-NADP(H) reductase (PDBid: 2BGI chain A); see Figure 5(B). In this case SScon5 and SSpro outperform the other considered methods, particularly in the four boxes regions (residues 2–13, 86–94, 99–102, and 186). They successfully predict a short 3_{10} helix at positions 99–102 and the β-bridge at position 186. The SScon5 not only successfully transfers these predictions into its outputs, but also removes some of the mistakes generated by SSpro; these errors, which are shown in underline and bold font, were removed using the other four SS predictions that are inputted into SScon5. We also note that the MV_{all} consensus was unable to provide correct predictions in the boxed regions.

Although these case studies should not be considered typical, they highlight how the SVM-based consensus improves over the majority vote and other individual predictors, and what types of improvements are to be expected. For instance, in previous sections we show that improvements are stronger for shorter helix and strand regions and for proteins with larger content of strand residues, which is corroborated in these two case studies.

Conclusions

The availability of SS that is predicted from protein chains fueled numerous investigations related to the prediction and characterization of structural and functional aspects of proteins. We show that SS predictions can be improved based on a well-designed consensus. The novelty of our investigation lies in the comprehensive scope (use of a dozen of modern standalone SS predictors), careful design, and detailed evaluation.

We show that a simple majority vote-based consensus provides improvement if it utilizes a well-selected set of input predictors. These improvements are at about 2% in Q_3 and 1.5% in SOV_3 compared to the best performing individual SS predictors. The other considered versions of the majority vote- and ranking-based consensuses are less effective than the majority vote.

Furthermore, we demonstrate that use of SMV to implement the consensus leads to additional improvements. Our study also reveals that consensus should utilize base/input SS predictors that are characterized by strong predictive performance and complementarity across different SS states.The total improvement offered by the SVM-based consensus over the best standalone predictor is 3.3% in Q_3 and 2.8% in SOV_3; the improvements offered by our SScon5 method are consistent across the three SS states. The SScon5 achieves $SOV_3 = 83.7\%$ and $Q_3 = 85.6\%$, which is close to the theoretical limit of 88%. The SVM-based consensus of *ab initio* methods, SScon3, also provides strong improvements over the corresponding standalone *ab initio* methods with $Q_3 = 82.3\%$ and $SOV_3 = 80.7\%$; these results are on par with the predictive quality of the best template-based approaches.

We also performed in-depth analysis of the improvements generated with the help of the SVM-based consensuses. We show that SScon5 reduces the number of significant errors where helix is confused with a strand, SScon5 and SScon3 perform particularly well for the prediction of short helices and strands, and they provide predictions with the overall amounts of helix, strand, and coil residues that are close to their native ratios. Finally, we demonstrate that SScon5 and SScon3 provide larger improvements for proteins with medium to high strand content and medium to low helix content.

Our future work will consider prediction of richer or different SS alphabets. As one potential extension, instead of using the three state alphabet (helix, strand and coil/loop), we plan to adopt the eight-state alphabet defined by DSSP like it was done in the SSpro8 predictor (Pollastri et al., 2002). Another option is to consider other SS assignments, similarly to SPINEX (Faraggi et al., 2009) where DSSP-based annotation was replaced by a consensus-based SKSP assignment (Zhang, Dunker, & Zhou, 2008), which utilizes four assignment methods.

For user's convenience, a web server and a standalone implementation of SScon5 and SScon3 models are available http://biomine.ece.ualberta.ca/SSCon.

List of abbreviations

C	Coil
DSSP	Dictionary of secondary structures of proteins
E	Strand
H	Helix
MV	Majority vote
RBF	Radial basis function
SOV	Segment overlap
SVM	Support vector machine
SS	Secondary structure
SScon3	Secondary structure prediction based on consensus of 3 *ab initio* predictors
SScon5	Secondary structure prediction based on consensus of 5 predictors

Acknowledgments

The authors thank Mr. Marcin Mizianty for his invaluable help in setting up the web server. J.Y. was supported in part by the University of Alberta Doctoral Recruitment Scholarship.

References

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

Albrecht, M., Tosatto, S. C., Lengauer, T., & Valle, G. (2003). Simple consensus procedures are effective and sufficient in secondary structure prediction. *Protein Engineering, 16*, 459–462.

Anderson, T. W., & Darling, D. A. (1952). Asymptotic theory of certain 'goodness-of-fit' criteria based on stochastic processes. *Annals of Mathematical Statistics, 23*, 193–212.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., … Bourne, P. E. (2000). The protein data bank. *Nucleic Acid Research, 28*(1), 235–242.

Bryson, K., McGuffin, L. J., Marsden, R. L., Ward, J. J., Sodhi, J. S., & Jones, D. T. (2005). Protein structure prediction servers at university college London. *Nucleic Acid Research, 33*, W36–W38.

Chen, K., Kurgan, L. (in press). Computational prediction of secondary and supersecondary structures. *Methods in Molecular Biology.* doi:10.1007/978-1-62703-065-6_5.

Chen, K., Mizianty, M. J., & Kurgan, L. A. (2012). Prediction and analysis of nucleotide binding residues using sequence and sequence-derived structural descriptors. *Bioinformatics, 28*, 331–341.

Chen, K., Stach, W., Homaeian, L., & Kurgan, L. (2011). IFC^2: An integrated web-server for improved prediction of protein structural class, fold type, and secondary structure content. *Amino Acids, 40*, 963–973.

Cheng, J., Randall, A. Z., Sweredoski, M. J., & Baldi, P. (2005). SCRATCH: A protein structure and structural feature prediction server. *Nucleic Acid Research, 33*, W72–W76.

Cheng, H., Sen, T. Z., Jernigan, R. L., & Kloczkowski, A. (2007). Consensus data mining (CDM) protein secondary structure prediction server: Combining GOR V and fragment database mining (FDM). *Bioinformatics, 23*, 2628–2630.

Cole, C., Barber, J. D., & Barton, G. J. (2008). The Jpred 3 secondary structure prediction server. *Nucleic Acid Research, 36*, W197–W201.

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

Dor, O., & Zhou, Y. (2007). Achieving 80% ten-fold cross-validated accuracy for secondary structure prediction by large-scale training. *Proteins, 66*, 838–845.

Eyrich, V. A., Martí-Renom, M. A., Przybylski, D., Madhusudhan, M. S., Fiser, A., Pazos, F., … Rost, B. (2001). EVA: Continuous automatic evaluation of protein structure prediction servers. *Bioinformatics, 17*, 1242–1243.

Eyrich, A. V., Przybylski, D., Koh, I. Y., Grana, O., Pazos, F., Valencia, F., & Rost, B. (2003). CAFASP3 in the spotlight of EVA. *Proteins, 53*, 548–560.

Faraggi, E., Yang, Y., Zhang, S., et al. (2009). Predicting continuous local structure and the effect of its substitution for secondary structure in fragment-free protein structure prediction. *Structure, 17*, 1515–1527.

Garg, A., Kaur, H., & Raghava, G. P. (2005). Real value prediction of solvent accessibility in proteins using multiple sequence alignment and secondary structure. *Proteins, 61*, 318–324.

Green, J. R., Korenberg, M. J., & Aboul-Magd, M. O. (2009). PCI-SS: MISO dynamic nonlinear protein secondary structure prediction. *BMC Bioinformatics, 10*(222).

Hildebrand, A., Remmert, M., Biegert, A., & Söding, J. (2009). Fast and accurate automatic structure prediction with HHpred. *Proteins, 77*, 128–132.

Homaeian, L., Kurgan, L. A., Ruan, J., Cios, K. J., & Chen, K. (2007). Prediction of protein secondary structure content for the twilight zone sequences. *Proteins, 69*, 486–498.

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.

Klammer, M., Messina, D. N., Schmitt, T., & Sonnhammer, E. L. (2009). MetaTM – a consensus method for transmembrane protein topology prediction. *BMC Bioinformatics, 10*(314).

Koh, I. Y., Eyrich, V. A., Marti-Renom, M. A., Przybylski, D., Madhusudhan, M. S., Eswar, N., … Rost, B. (2003). EVA: Evaluation of protein structure prediction servers. *Nucleic Acids Research, 31*, 3311–3315.

Li, B. Q., Hu, L. L., Niu, S., Cai, Y. D., & Chou, K. C. (2012). Predict and analyze S-nitrosylation modification sites with the mRMR and IFS approaches. *Journal of Proteomics, 75*, 1654–1665.

Lin, K., Simossis, V. A., Taylor, W. R., & Heringa, J. (2005). A simple and fast secondary structure prediction algorithm using hidden neural networks. *Bioinformatics, 21*, 152–159.

Lu Y, Wang X, Chen X & Zhao G. (2012). Computational methods for DNA-binding protein and binding residue prediction. *Protein and Peptide Letters*. [Epub ahead of print].

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

Martin, J., Letellier, G., Marin, A., Taly, J. F., de Brevern, A. G., & Gibrat, J. F. (2005 September). Protein secondary structure assignment revisited: A detailed analysis of different assignment methods. *BMC Structural Biology, 5*(17).

Martin, J., Gibrat, J. F., & Rodolphe, F. (2006). Analysis of an optimal hidden Markov model for secondary structure prediction. *BMC Structural Biology, 6*(25).

Mittal, A., & Jayaram, B. (2011). The newest view on protein folding: Stoichiometric and spatial unity in structural and functional diversity. *Journal of Biomolecular Structure & Dynamics, 28*, 669–674.

Mittal, A., Jayaram, B., Shenoy, S., & Bawa, T. S. (2010). A stoichiometry driven universal spatial organization of backbones of folded proteins: Are there Chargaff's rules for protein folding? *Journal of Biomolecular Structure & Dynamics, 28*, 133–142.

Mizianty, M. J., Stach, W., Chen, K., Kedarisetti, K. D., Miri Disfani, F., & Kurgan, L. (2010). Improved sequence-based prediction of disordered regions with multilayer fusion of multiple information sources. *Bioinformatics, 26*, i489–i496.

Mizianty, M. J., & Kurgan, L. A. (2011). Improved identification of outer membrane beta barrel proteins using primary sequence, predicted secondary structure and evolutionary information. *Proteins, 79*(1), 294–303.

Montgomerie, S., Sundararaj, S., Gallin, W. J., & Wishart, D. S. (2006). Improving the accuracy of protein secondary structure prediction using structural alignment. *BMC Bioinformatics, 7*, 301–313.

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

Mooney, C., Pollastri, G., Shields, D. C., & Haslam, N. J. (2012). Prediction of short linear protein binding regions. *Journal of Molecular Biology, 415*(1), 193–204.

Murzin, A. G., Brenner, S. E., Hubbard, T., & Chothia, C. (1995). SCOP: A structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology, 247*, 536–540.

Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B., & Thornton, J. M. (1997). CATH – A hierarchic classification of protein domain structures. *Structure, 5*, 1093–1108.

Peng, Z. L., & Kurgan, L. (2012). Comprehensive comparative assessment of in-silico predictors of disordered regions. *Current Protein and Peptide Science, 13*(1), 6–18.

Pirovano, W., & Heringa, J. (2010). Protein secondary structure prediction. *Methods in Molecular Biology, 609*, 327–348.

Plewczynski, D., Łaźniewski, M., von Grotthuss, M., Rychlewski, L., & Ginalski, K. (2011). VoteDock: Consensus docking method for prediction of protein-ligand interactions. *Journal of Computational Chemistry, 32*, 568–581.

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, 47*, 228–235.

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

Rost, B. (1996). PHD: Predicting one-dimensional protein structure by profile-based neural networks. *Methods Enzymology, 266*, 525–539.

Rost, B. (2001). Review: Protein secondary structure prediction continues to rise. *Journal of Structural Biology, 134*, 204–218.

Rost, B. (2003). Rising accuracy of protein secondary structure prediction. In D. Chasman (Ed.), *Protein structure determination, analysis, and modeling for drug discovery* (pp. 207–249). New York, NY: Dekker.

Rost, B. (2009). Prediction of protein structure in 1D – secondary structure, membrane regions, and solvent accessibility. In P. E. Bourne & H. Weissig (Eds.), *Structural Bioinformatics* (2nd ed., pp. 679–714). Hoboken, NJ: Wiley.

Rost, B., Yachdav, G., & Liu, J. (2004). The predict protein server. *Nucleic Acids Research, 32*, W321–W326.

Wang, G., & Dunbrack, R. L .Jr. (2003). PISCES: A protein sequence culling server. *Bioinformatics, 19*, 1589–1591.

Won, K. J., Hamelryck, T., Prugel-Bennett, A., et al. (2007). An evolutionary method for learning HMM structure: Prediction of protein secondary structure. *BMC Bioinformatics, 8*(357).

Wu, S., & Zhang, Y. (2008). MUSTER: Improving protein sequence profile-profile alignments by using multiple sources of structure information. *Proteins, 72*, 547–556.

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*, 2076–2082.

Zemla, A., Venclovas, C., Fidelis, K., & Rost, B. (1999). A modified definition of SOV, a segment-based measure for protein secondary structure prediction assessment. *Proteins, 34*, 220–223.

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

Zhang, H., Zhang, T., Chen, K., Kedarisetti, K.D., Mizianty, M.J., Bao, Q., … Kurgan, L. (2011). Critical assessment of high-throughput standalone methods for secondary structure prediction. *Briefings in Bioinformatics, 12*(6), 672–688.

Zhang, T., Zhang, H., Chen, K., Ruan, J., Shen, S., & Kurgan, L.A. (2010). Analysis and prediction of RNA-binding residues using sequence, evolutionary conservation, and predicted secondary structure and solvent accessibility. *Current Protein and Peptide Science, 11*(7), 609–628.

Zhang, H., Zhang, T., Gao, J., Ruan, J., Shen, S., & Kurgan, L. A. (2012). Determination of protein folding kinetic types using sequence and predicted secondary structure and solvent accessibility. *Amino Acids, 42*, 271–283.