JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

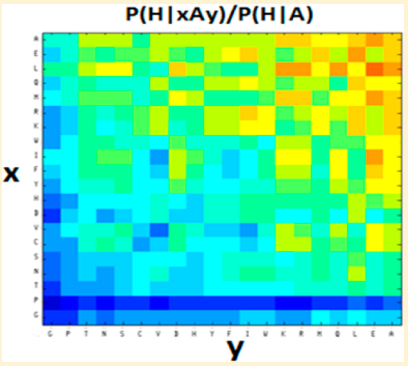
pubs.acs.org/jcim

Context-Based Features Enhance Protein Secondary Structure Prediction Accuracy

Ashraf Yaseen[†] and Yaohang Li*

Department of Computer Science, Old Dominion University, Norfolk, Virginia 23529, United States

ABSTRACT: We report a new approach of using statistical context-based scores as encoded features to train neural networks to achieve secondary structure prediction accuracy improvement. The context-based scores are pseudo-potentials derived by evaluating statistical, high-order inter-residue interactions, which estimate the favorability of a residue adopting certain secondary structure conformation within its amino acid environment. Encoding these context-based scores as important training and prediction features provides a way to address a long-standing difficulty in neural network-based secondary structure predictions of taking interdependency among secondary structures of neighboring residues into account. Our computational results have shown that the context-based scores are effective features to enhance the prediction accuracy of secondary structure predictions. An overall 7-fold cross-validated Q3 accuracy of 82.74% and Segment Overlap Accuracy (SOV) accuracy of 86.25% are achieved on a set of more than 7987 protein chains with, at most, 25% sequence identity. The Q3 prediction accuracy on benchmarks of CB513, Manesh215, Carugo338, as well as CASP9 protein chains is higher than popularly used secondary structure prediction servers, including Psipred, Profphd, Jpred, Porter (ab initio), and Netsurf. More significant improvement is observed in the SOV accuracy, where more than 4% enhancement is observed, compared to the server with the best SOV accuracy. A Q8 accuracy of >70% (71.5%) is also found in eight-state secondary structure prediction. The majority of the Q3 accuracy improvement is contributed from correctly identifying β-sheets and α-helices. When the context-based scores are incorporated, there are 15.5% more residues predicted with >90% confidence. These high-confidence predictions usually have a rather high accuracy (averagely ~95%). The three- and eight-state prediction servers (SCORPION) implementing our methods are available online.



INTRODUCTION

Reliably and accurately predicting three-dimensional structures of proteins from their sequences is one of the grand computational challenges with broad scientific and economic impacts. An important intermediate step toward solving the protein structure modeling problem is to correctly predict the secondary structure. This is due to the fact that proteins form local conformations such as helices and sheets: the correct prediction of protein secondary structures will significantly reduce the degrees of freedom in protein tertiary structure modeling and therefore reduce the difficulty of obtaining high-resolution three-dimensional (3D) models.[1]

Historically, the improvement of secondary structure prediction benefits from the incorporation of new features/information that can separate among secondary structure classes. The early methods of secondary structure prediction[2,3] are based on the statistical analysis of the capability of single amino acids in forming various secondary structure elements, whose Q3 prediction accuracy (percentage of residues predicted correctly in one of the three states: helix, strand, and coil) usually does not exceed 60%.[4] The later generation of secondary structure prediction methods takes advantage of the segment information consisting of 3−51 adjacent residues, which passes 60% Q3 prediction accuracy.[5] The most substantial improvement is due to the use of evolutionary information obtained from the divergence proteins in the same structural family, where a Q3 prediction accuracy of >70% is achieved.[6,7] Nowadays, with the advancement of machine learning algorithms including Bayesian statistics, neural networks, *k*-nearest neighbors, hidden Markov models, support vector machines, and random field, dramatic increase in the number of solved protein structures available in protein data banks, and incorporation of amino acid property features such as physicochemical propensities, solvent accessibilities, etc., in the training and prediction process, quite a few current prediction methods report Q3 accuracies from ~76% to ~80%.[8−16]

Unlike three-state secondary structure prediction, very few methods have been developed for the eight-state prediction, to the best of our knowledge. SSpro8 is an eight-state secondary structure prediction method developed by Pollastri et al.[13] with a Q8 accuracy of 62%−63%. A more recent prediction method (RaptorXss8) developed by Wang et al.[17] reported a Q8 accuracy of 67.9% through the use of the Conditional Neural Field (CNF) model.

Theoretically, a prediction accuracy of 88%−90% is usually considered as the upper bound of secondary structure prediction.[4] This is because the secondary structure assignments based on crystal structure have ~10% errors themselves, as inferred from differences between X-ray structures and nuclear

Received: November 1, 2013
Published: February 26, 2014

   dx.doi.org/10.1021/ci400647u | *J. Chem. Inf. Model.* 2014, 54, 992−1002

magnetic resonance (NMR) models of the same protein and from inconsistency of secondary structure assignments by different methods of different parameters (e.g., DSSP[18] and STRIDE[19]). However, no significant forward steps have been made in the last 10 years to get closer to the upper bound of the prediction. Even obtaining improvements of fractions of a percent has become difficult. Unless homologue templates are available,[20] the Q3 accuracy of secondary structure prediction is stagnated between 76% and 80%. Moreover, many of the reported accuracies from different prediction methods are not cross-validated and/or obtained from several small datasets. On the other hand, a variety of computational applications aimed at modeling protein structures and understanding protein functions strongly rely on the accurate prediction of secondary structures. As a basis for tertiary structure prediction, reducing the percent of inaccuracy would be an enormous improvement in efficiency, because the search space for finding a tertiary structure goes up superlinearly with the fraction of inaccuracy in the secondary structure prediction. Continuous improvement of secondary structure prediction accuracy toward the theoretical upper bound will substantially benefit these applications.

It is well-known that, in machine learning, extracting and selecting "good" features can significantly enhance the prediction performance of a predictor. Probably, the most effective features are the secondary structures of its neighbors to predict the secondary structure of a residue. For example, if both adjacent neighbors are helices, the middle residue is most likely to be a helix. Vice versa, if the adjacent positions of a residue are not helices, it is impossible for this middle residue to adopt a helix as its secondary structure. In fact, our computational results shows that, if the true secondary structures of neighboring residues are encoded, machine learning using a simple feedforward neural network can easily lead to a prediction accuracy of >90%. Unfortunately, using the true secondary structures as features is not feasible, since they cannot be known a priori. However, this inspires us to examine whether the favorability of a residue adopting a certain secondary structure can be also an effective feature. The statistical scores measuring the favorability of a residue adopting a certain secondary structure within its amino acid environment can be obtained from the experimentally determined protein structures in the PDB.[21] Encoding these statistical scores as features provides an approach to address a long-standing difficulty in neural network of taking interdependency among secondary structures of neighboring residues into account.

In this paper, we extract context-based statistical scores to measure favorability of a residue adopting secondary structure from a large training sample set. The fundamental idea is based on the fact that the formation of secondary structure exhibit strong local dependency, particularly, residues in a protein sequence are strongly correlated in different sequence positions in coils, β-sheets, 3_{10}-helices, α-helices, and π-helices. The context-based statistics indicate the favorability of a residue adopting a secondary structure conformation in the presence of its neighbors in sequence. We derive statistics for singlets, doublets, and triplets in a sequence window from experimentally determined structures in the PDB. Scores that measure the pseudo-potentials of a residue adopting a certain secondary structure then are calculated using the potentials of a mean force approach. These scores are incorporated as sequence-structure features together with the Position Specific Scoring Matrix (PSSM) data to train the secondary structure prediction neural networks. Our server implementing this method is called SCORPION (secondary structure prediction). We apply our approach to predict secondary structures in both three-state and eight-state predictions. We test our method on several commonly used benchmarks for secondary structure prediction, including CB513,[22] Manesh215,[23] and Carugo338,[24] as well as the CASP9 targets.[25] We compare our results with a set of popular secondary structure prediction methods including Porter (ab initio),[8] Psipred,[9] PROFphd,[10] Netsurfp,[11] and Jpred[12] for three-state predictions, and with RaptorXss8[17] for eight-state predictions. The prediction accuracy of our method is further analyzed in this paper.

MATERIALS AND METHODS

Datasets. We use the CullPDB data set (Cull16633) generated by the PISCES server[26] on October 21, 2011 to collect the triplet samples to produce the context-based statistics. Cull16633 contains 16 633 chains with a maximum pairwise sequence identity of 50%, a resolution of 3.0 Å, and an R-factor of 1.0. Our recent results have shown that the statistics obtained from the cull library with a maximum sequence identity of 50% yields the optimal accuracy (31).

For neural network training, we use the Cull7987 dataset, which includes 7987 chains with a pairwise sequence identity of, at the most, 25%, a resolution cutoff of 3.0 Å, and an R-factor cutoff of 1.0. We use the PSI-BLAST program[27] to generate the PSSM data. Short chains with less than 40 residues are eliminated, since the PSI-BLAST program is usually unable to generate profiles for very short sequences. Very large chains whose lengths are greater than 1000 residues also are removed. We also exclude residues with undetermined structures from the training set.

Popular public benchmarks for protein secondary structure prediction, including CB513,[22] Manesh215,[23] and Carugo338[24] and the recent CASP9 targets[25] are used to validate our method. We use the Q3 scores (for three-state prediction), Q8 scores (for eight-state prediction), and SOV (segment overlap[28]) scores to measure the accuracy of secondary structure predictions.

The eight-state assignments (G = 3_{10}-helix, H = α-helix, I = π-helix, E = extended strand, B = isolated bridge, S = bend, T = turn, and C = coil) of the Cull16633 and Cull7986 protein datasets, as well as the benchmark protein sets, are determined by the DSSP program. In eight-state prediction, SCORPION directly predicts the DSSP eight states of each residue. In three-state prediction, SCORPION predicts the three dominant secondary structure states by grouping (G, H, I) into helices (H), (E, B) into sheets (E), and (T, S, C) into coils (C), which is consistent with most secondary structure prediction methods.

Context-Based Statistics. The early studies in protein secondary structure show that the types of nearby neighboring residues play a predominant role to the secondary structure conformation that a residue adopts. In particular, the formation of interactions within coils beyond nearest neighbors, in most cases, does not appear to contribute a statistically significant amount in determining coil structure.[29] Residues in contacting parallel or antiparallel β-sheets are connected by hydrogen bonds in alternative positions. The hydrogen bonds between residues at positions i and i+3, i and i+4, and i and i+5 lead to the formation of 3−10 helices, α-helices, and π-helices, respectively. Therefore, the context-based statistics by capturing the correlation between residues provide important information in predicting secondary structure. Actually, the early GOR[30] method employed information theory by taking advantage of the context-based

dx.doi.org/10.1021/ci400647u | J. Chem. Inf. Model. 2014, 54, 992−1002

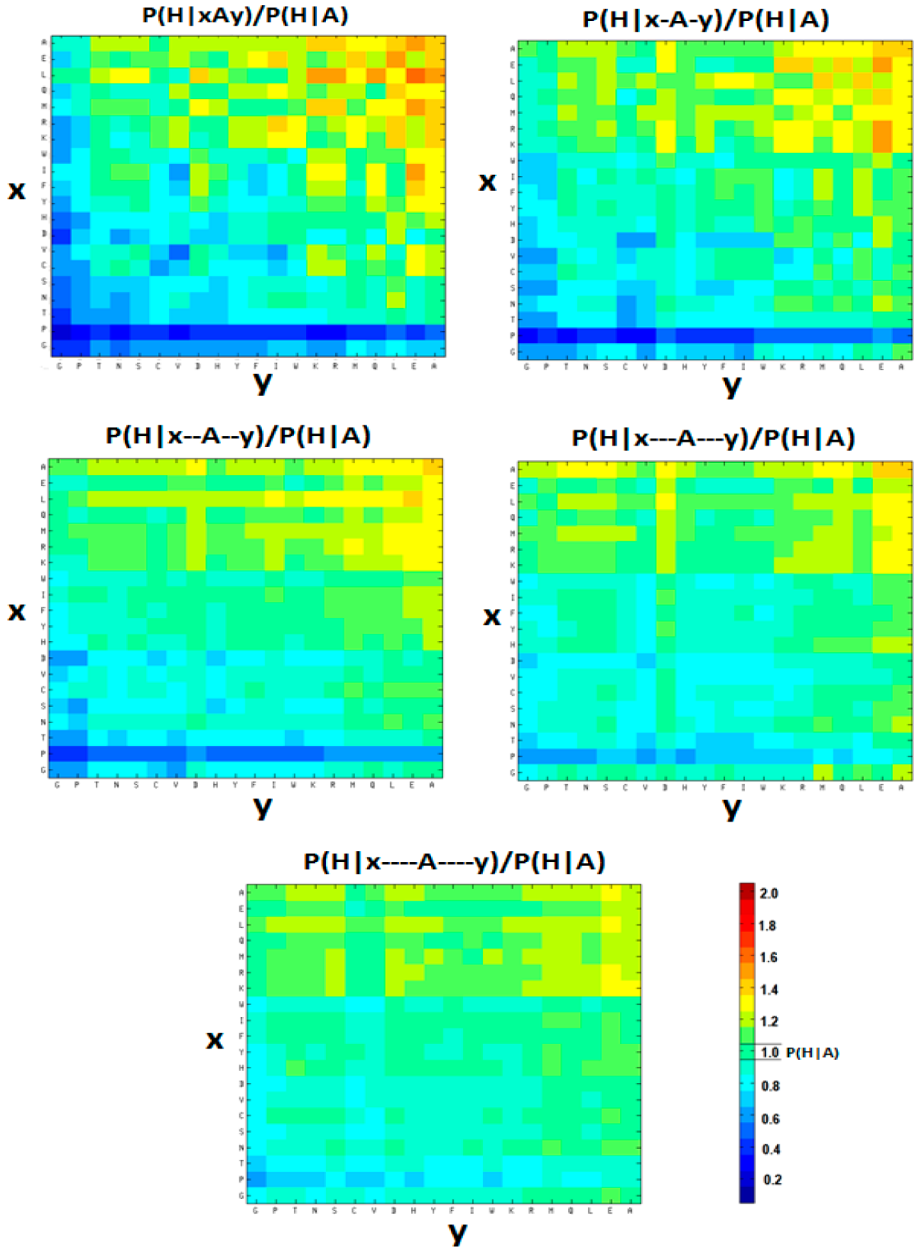


Figure 1. Probability of alanine as the middle residue of a triplet with neighboring residues 1−5 positions away when adopting α-helix as a secondary structure (the symbols "x", "y", and "-" represent the left neighbor, right neighbor, and gap, respectively). The neighboring residues are ordered by their probability of forming α-helix. The nearest neighbors have the strongest influences; however, neighbors 5 positions away still have certain non-negligible influences on the secondary structure conformation of the middle alanine.

pairwise interaction statistics and achieved initial success in secondary structure prediction.

Figure 1 shows the probability of alanine as the middle residue of a triplet with neighboring residues at 1−5 positions away when adopting helix as secondary structure. One can find that, as expected, the nearest neighbors have the strongest influence to the middle alanine and the further the neighbors are away, the weaker the influence. However, even residues five positions away have non-negligible influence on the middle alanine, although much weaker than that of the adjacent residues.

The recent increasing number of determined structures in protein databanks has made the derivation of context-based statistics feasible by characterizing high-order inter-residue interactions. In this work, we extract statistics of singlets (R_i), doublets (R_iR_{i+k}), and triplets ($R_iR_{i+k_1}R_{i+k_2}$) residues at different relative positions in protein sequences, which is further used to generate pseudo-potentials to be incorporated as new features in neural network training and prediction. The statistics of singlets, doublets, and triplets represent estimations of the probabilities of residues adopting a specific secondary structure type when none, one, or two of their neighbors in context are taken into consideration, respectively.

The observed probabilities of the ith residue R_i in a singlet (R_i), doublet (R_iR_{i+k}), and triplet ($R_iR_{i+k_1}R_{i+k_2}$) adopting a specific secondary structure C_i[31] are respectively estimated by

$$P_{\text{obs}}(C_i|R_i) = \frac{N_{\text{obs}}(C_i, R_i)}{N_{\text{obs}}(R_i)}$$

dx.doi.org/10.1021/ci400647u | J. Chem. Inf. Model. 2014, 54, 992−1002

$$P_{obs}(C_i|R_iR_{i+k}) = \frac{N_{obs}(C_i, R_iR_{i+k})}{N_{obs}(R_iR_{i+k})}$$

and

$$P_{obs}(C|R_iR_{i+k_1}R_{i+k_2}) = \frac{N_{obs}(C_i, R_iR_{i+k_1}R_{i+k_2})}{N_{obs}(R_iR_{i+k_1}R_{i+k_2})}$$

Here, $N_{obs}(C_i, R_i)$, $N_{obs}(C_i, R_iR_{i+k})$, and $N_{obs}(C_i, R_iR_{i+k_1}R_{i+k_2})$ are the observed number of singlets (R_i), doublets (R_iR_{i+k}), and triplets ($R_iR_{i+k_1}R_{i+k_2}$) while $N_{obs}(C_i, R_i)$, $N_{obs}(C_i, R_iR_{i+k})$, and $N_{obs}(C_i, R_iR_{i+k_1}R_{i+k_2})$ are those of singlets (R_i), doublets (R_iR_{i+k}), and triplets ($R_iR_{i+k_1}R_{i+k_2}$) with R_i adopting conformation C_i (H, E, C for three-state prediction and G, H, I, E, B, T, S, C for eight-state prediction) in the training set. However, there may not be enough samples for doublets and triplets with less popular amino acids in the training set to obtain statistically meaningful results. We remedy this problem by taking advantage of the PSSM values generated by multiple sequence alignment by summing the frequency weights at each residue position. The frequency weights are derived from the PSSM values at each residue position in a protein sequence, which are generated by PSI-BLAST, using three iterations of searching with an e-value of 0.001 against the NCBI nr database of protein sequences. These observed numbers are then calculated as

$$N_{obs}(R_i) = \sum_{Protein} \sum_{j} PSSM_j(R_i)$$

$$N_{obs}(R_iR_{i+k}) = \sum_{Protein} \sum_{j} PSSM_j(R_i) \times PSSM_j(R_{i+k})$$

$$N_{obs}(R_iR_{i+k_1}R_{i+k_2}) = \sum_{Protein} \sum_{j} PSSM_j(R_i) \times PSSM_j(R_{i+k_1}) \times PSSM_j(R_{i+k_2})$$

$$N_{obs}(C_i, R_i) = \sum_{Protein} \sum_{C_j=C_i}^{j} PSSM_j(R_i)$$

$$N_{obs}(C_i, R_iR_{i+k}) = \sum_{Protein} \sum_{C_j=C_i}^{j} PSSM_j(R_i) \times PSSM_j(R_{i+k})$$

$$N_{obs}(C_i, R_iR_{i+k_1}R_{i+k_2}) = \sum_{Protein} \sum_{C_j=C_i}^{j} PSSM_j(R_i) \times PSSM_j(R_{i+k_1}) \times PSSM_j(R_{i+k_2})$$

where $PSSM_j(R_i)$ is the PSSM frequency for residue type R_i at position j of a protein sequence.

Context-Dependent Pseudo-potentials. The context-dependent pseudo-potentials are generated based on Sippl's potentials of mean force method.[32] According to the inverse-Boltzmann theorem, we calculate the mean-force potential $U_{singlet}(R_i,C_i)$ for a singlet residue R_i adopting a secondary structure C_i:

$$U_{singlet}(C_i, R_i) = -RT \ln\left(\frac{P_{obs}(C_i|R_i)}{P_{ref}(C_i|R_i)}\right)$$

where R is the gas constant, T the temperature, and $P_{ref}(C_i|R_i)$ the referenced probability. In our method, we employ the conditional probability approach described in Samudrala and Moult[33] to estimate the referenced probability by

$$P_{ref}(C_i|R_i) = \frac{\sum_{C_j=C_i}^{j} N_{obs}(C_j, R_j)}{\sum_j N_{obs}(R_j)}$$

Similarly, the mean-force potentials $U_{singlet}(R_i,C_i)$ and $U_{singlet}(R_i,C_i)$ for residues adopting a secondary structure are

$$U_{doublet}(C_i, R_iR_{i+k}) = -RT \ln\left(\frac{P_{obs}(C_i|R_iR_{i+k})P_{ref}(C_i|R_i)}{P_{ref}(C_i|R_iR_{i+k})P_{obs}(C_i|R_i)}\right)$$

and

$$U_{triplet}(C_i, R_iR_{i+k_1}R_{i+k_2}) = -RT \ln\left(\frac{P_{obs}(C_i|R_iR_{i+k_1}R_{i+k_2})P_{ref}(C_i|R_iR_{i+k_2})P_{ref}(C_i|R_iR_{i+k_1})P_{obs}(C_i|R_i)}{P_{ref}(C_i|R_iR_{i+k_1}R_{i+k_2})P_{obs}(C_i|R_iR_{i+k_2})P_{obs}(C_i|R_iR_{i+k_1})P_{ref}(C_i|R_i)}\right)$$

respectively, with corresponding referenced probabilities given as

$$P_{ref}(C_i|R_iR_{i+k}) = \frac{\sum_{\substack{C_j=C_i \\ R_{j+k}=R_{i+k}}}^{j} N_{obs}(C_j, R_jR_{j+k})}{\sum_j N_{obs}(R_jR_{j+k})}$$

and

$$P_{ref}(C_i|R_iR_{i+k_1}R_{i+k_2}) = \frac{\sum_{\substack{C_j=C_i \\ R_{j+k_1}=R_{i+k_1} \\ R_{j+k_2}=R_{i+k_2}}}^{j} N_{obs}(C_j, R_jR_{j+k_1}R_{j+k_2})}{\sum_j N_{obs}(R_jR_{j+k_1}R_{j+k_2})}$$

Then, the context-dependent pseudo-potential for R_i to adopt secondary structure conformation C_i under its amino acid environment is

$$U(C_i, ..., R_{i-1}R_iR_{i+1}, ...) = U_{singlet}(C_i, R_i) + \sum_k U_{doublet}(C_i, R_iR_{i+k}) + \sum_{k_1,k_2} U_{triplet}(C_i, R_iR_{i+k_1}R_{i+k_2})$$

These context-dependent pseudo-potentials are used as context-based scores to be encoded in neural network training and prediction.

Neural Networks. Our SCORPION server incorporates three phases of feed-forward neural network training. The first and second phases are sequence-to-structure and structure-to-

dx.doi.org/10.1021/ci400647u | J. Chem. Inf. Model. 2014, 54, 992−1002

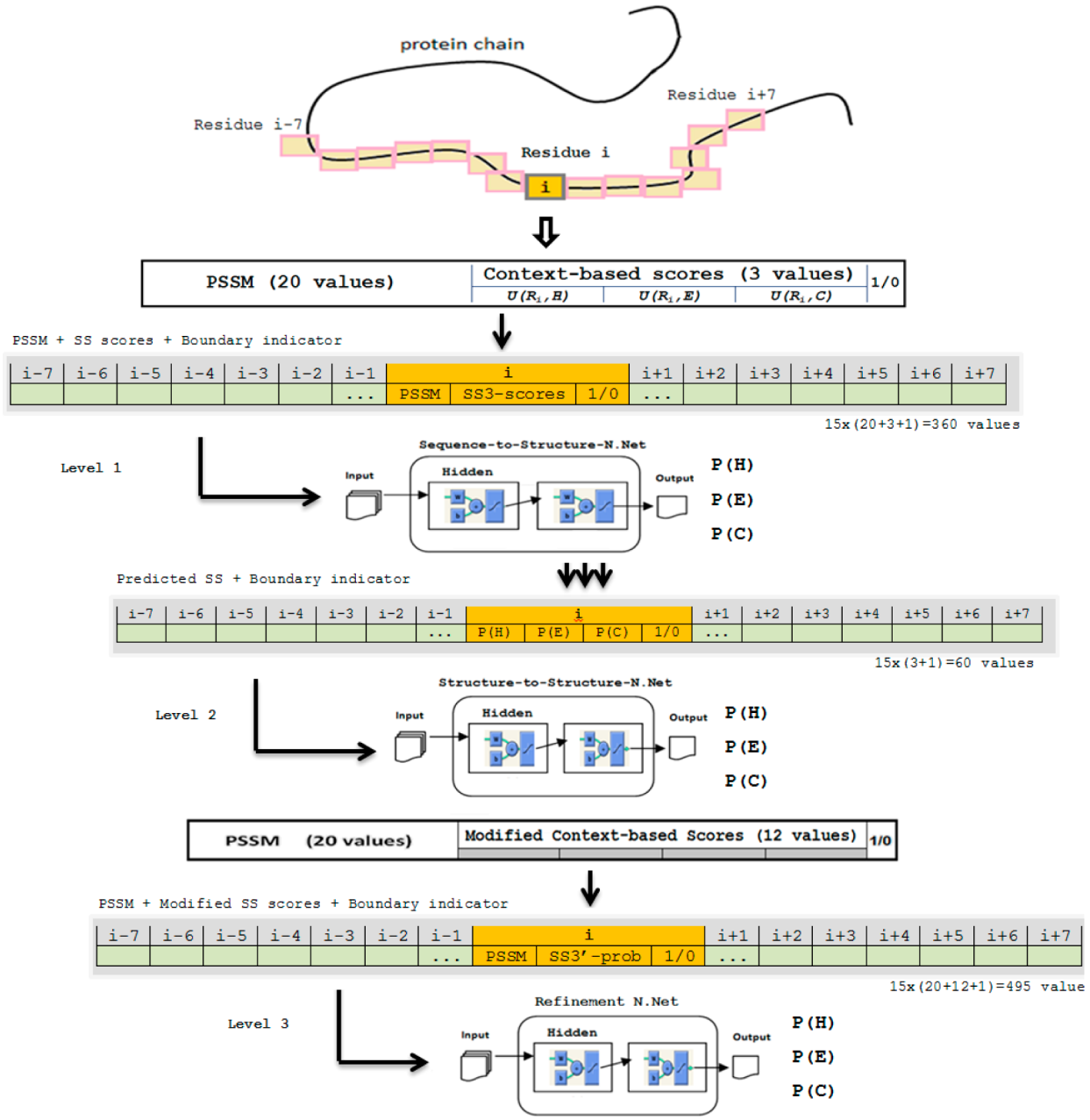


Figure 2. Three phases of neural network (architecture and encoding) for three-state prediction in SCORPION.

structure training, respectively, while the third phase is used to refine the prediction results.

In the sequence-to-structure training, a sliding window of 15 residues is selected, where each neural network is trained to predict the class of that residue in the middle of the window. Each residue is represented by 20 PSSM values and 1 extra value to indicate C- or N-terminals overlap. When the context-based scores are incorporated, additional encoding values (3 for three-state prediction and 8 for eight-state prediction) for the each residue are needed. Overall, 360 and 435 input values are used to encode each residue in three- and eight-state prediction, respectively. After sequence-to-structure training, the next phase is to carry out a structure-to-structure training to eliminate impossible secondary structure predictions. The last phase employs a manner similar to the first one, but setting some context-based scores to "absolute favorable" if the results from structure-to-structure prediction indicates that the probability of a residue adopting a certain secondary structure is >90%. Figure 2 illustrates the neural network encoding and architecture for three phases of training.

7-Fold Cross-Validation. Here, 7-fold cross validation has been performed on Cull7987. We randomly divide the chains in Cull7987 set into seven subsets with approximately the same number of chains. At each fold, five subsets are used for training, one for testing, and one for validation. To ensure complete separation of the training set and testing set for each fold, we generate a set of scores only based on the sequences in the five training subsets and then encode it in training. Hence, totally seven sets of context-based scores are generated for 7-fold cross validation. The overall prediction accuracy is calculated as the average accuracy of the 7-fold predictions.

dx.doi.org/10.1021/ci400647u | J. Chem. Inf. Model. 2014, 54, 992−1002

RESULTS

Three-State Prediction. Table 1 compares the 7-fold cross-validation Q3 and SOV3 accuracies of neural networks for three-state prediction with context-based score encoding (PSSM + context-based score) and without context-based score encoding (PSSM Only). Both neural network trainings go through the same training and cross-validation procedure. When the context-based scores are incorporated, both Q3 and SOV3 accuracy enhancements are observed in all three secondary structure classes. The overall cross-validated Q3 accuracy is 82.74%, which is higher than the reported accuracies (~80%) in the popular secondary structure prediction servers.[8−12] It is important to notice that the most significant improvement in accuracy (4.02%) is found in β-sheets. This is particularly encouraging, because β-sheets are typically harder to predict than helices due to global interactions. 2.55% and 1.47% accuracy improvement are also observed in helices and coils, respectively. Because of the fact that residues in sheets and helices yield stronger correlation to the neighboring residues than those in coils, the context-based scores are more effective on sheets and helices predictions. The overall 7-fold cross-validated SOV3 accuracy reaches 86.25%, which is 2.39% higher than that of using only PSSM encoding.

Table 1. 7-Fold Cross-Validation Q3 and SOV3 Accuracies for Three-State Prediction in SCORPION

	Accuracy (%)	
	PSSM only	PSSM+context-based score
Q_H	84.74	87.29
Q_E	72.72	76.74
Q_C	80.53	82.00
Q_3	80.31	82.74
SOV_H	87.85	90.34
SOV_E	81.87	84.13
SOV_C	81.19	83.31
SOV_3	83.86	86.25

Table 2 shows the Q3 accuracy and the composition frequency of each amino acid type. The prediction accuracy for cysteine is the lowest, mainly due to its lowest composition frequency in protein sequences. Moreover, a cysteine residue may form a disulfide bond with another cysteine residue, which complicates the prediction of their secondary structures.

Tables 3 and 4 compare the Q3 and SOV3 accuracies between our method and the popularly used secondary structure prediction servers, including Porter (ab initio), PsiPred, ProfPhD, Netsurfp, and Jpred on benchmarks of CB513, CASP9, Manesh215, and Carugo338. To enforce fairness comparison, we generate context-based scores by removing all sequences with sequence identities of 25% or higher, relative to the sequences in benchmark from Cull16633, and all homologues with a sequence identity of higher than 25%, relative to the chains presented in these benchmarks are excluded from Cull7987 when training neural networks. It is interesting to notice that our prediction method has significantly higher accuracy in both α-helices and β-sheets than the other servers, with improvements of more than 5% in most cases. However, as a tradeoff, the accuracy of coils is ~5% less compared to PsiPred, ~3.6% less compared to Netsurfp, and ~2% less compared to JPred. After all, compared to PsiPred with the highest Q3 accuracy, our method's improvement on these benchmarks is

Table 2. Q3 Accuracy for Each Amino Acid Type in SCORPION

amino acid, AA	composition frequency (%)	Q3 accuracy (%)
A	8.11	83.59
R	5.16	81.96
N	4.33	81.96
D	5.92	83.67
C	1.26	76.79
Q	3.88	83.27
E	6.88	83.03
G	6.96	83.56
H	2.33	81.20
I	5.85	84.29
L	9.64	83.74
K	5.82	82.07
M	1.62	82.69
F	4.19	80.65
P	4.52	84.31
S	6.03	80.72
T	5.45	80.92
W	1.42	80.02
Y	3.62	79.96
V	7.00	83.70

Table 3. Comparison of Q3 Accuracy between SCORPION and Other Popularly Used Secondary Structure Prediction Methods, Including Porter (Ab Initio), PsiPred, ProfPHD, NetSurfp, and JPred on Benchmarks of CB513, CASP9, Manesh215, and Carugo338

		Q3 Accuracy (%)			
		CB513	CASP9	Manesh215	Carugo338
Porter (ab initio)	Q3	77.53	78.65	77.99	77.5
	QH	81.30	85.45	81.72	80.67
	QE	66.18	67.4	66.73	66.21
	QC	80.49	78.75	80.57	81.1
Psipred	Q3	80.19	81.35	80.67	80.06
	QH	79.28	83.32	78.29	77.09
	QE	68.49	69.68	69.43	67.96
	QC	87.11	85.84	88.63	88.87
Profphd	Q3	76.52	76.91	76.77	76.47
	QH	80.06	84.41	80.02	78.76
	QE	69.18	65.53	68.78	68.82
	QC	77.54	76.48	78.06	78.8
Netsurfp	Q3	77.88	79.35	78.7	78.24
	QH	77.21	82.46	77.39	76.2
	QE	64.36	64.94	66.17	64.65
	QC	85.56	84.27	86.39	87.1
JPred	Q3	78.72	79.24	79.32	78.67
	QH	78.02	79.29	77.72	76.34
	QE	69.04	74.05	71.48	69.37
	QC	84.39	82.09	84.81	85.49
C3-SCORPION	Q3	80.69	83.02	82.66	81.96
	QH	85.27	88.38	86.22	85.51
	QE	72.69	77.66	75.97	74.07
	QC	81.15	81.44	82.95	83.43

dx.doi.org/10.1021/ci400647u | J. Chem. Inf. Model. 2014, 54, 992−1002

Table 4. Comparison of SOV3 Accuracy between SCORPION and Other Popularly Used Secondary Structure Prediction Servers Including Porter (Ab Initio), PsiPred, ProfPHD, NetSurfp, and JPred on Benchmarks of CB513, CASP9, Manesh215, and Carugo338

		SOV3 Accuracy (%)			
		CB513	CASP9	Manesh215	Carugo338
Porter	SOV3	80.21	82.41	80.90	80.03
	SOVH	84.64	88.36	85.07	84.09
	SOVE	76.06	77.24	76.70	76.68
	SOVC	78.76	79.87	79.32	78.61
Psipred	SOV3	78.91	81.24	79.55	77.63
	SOVH	83.61	87.21	84.00	82.40
	SOVE	77.35	78.62	78.56	77.15
	SOVC	75.80	77.19	75.96	74.03
Profphd	SOV3	78.96	79.87	79.62	78.28
	SOVH	83.79	86.42	83.59	82.55
	SOVE	76.19	74.08	76.52	76.21
	SOVC	76.44	77.09	77.64	75.99
Netsurfp	SOV3	77.66	79.74	78.92	77.18
	SOVH	82.21	86.13	82.86	81.62
	SOVE	75.06	75.25	77.02	75.40
	SOVC	75.26	76.38	76.31	74.52
JPred	SOV3	78.82	81.70	79.63	77.98
	SOVH	82.85	83.39	82.88	81.61
	SOVE	77.56	82.52	79.62	78.51
	SOVC	76.11	79.74	76.63	74.71
C3-SCORPION	SOV3	83.98	86.38	85.72	84.45
	SOVH	88.71	89.88	89.52	88.37
	SOVE	80.64	84.57	82.91	82.12
	SOVC	81.84	84.31	83.71	82.57

from 0.5% to 2%. Although an accuracy improvement from 0.5% to 2% over PsiPred does not seem very attractive, it is important to notice an improvement in SOV3 accuracy of more than 5%, compared to PsiPred. Moreover, the SOV3 accuracy of our server is 4.25% higher than Porter (ab initio) and is more than 6% higher compared to Netsurfp, JPred, or PsiPred. This is because the context-based scores incorporating secondary structure information of neighboring residues enhance the coverage of the secondary structure segments.

Eight-State Prediction. The overall Q_8 7-fold cross-validated accuracy on the Cull7987 dataset in SCORPION is 71.5%, where the accuracies of predicting 3_{10}-helix (G), α-helix (H), π-helix (I), extended strand (E), isolated bridge (B), bend (S), turn (T), and coil (C) are 22.7%, 92.4%, 0%, 82.9%, 2.4%, 22.3%, 51.6%, and 66.1%, respectively. The accuracy comparison with prediction without using context-based score encoding is listed in Table 5. The prediction accuracies of the eight different secondary structure states vary significantly. In particular, the prediction accuracy of G, I, B, and S are very low, mainly due to the fact of their infrequent appearances in protein data banks, whose distribution is shown in Figure 3. Hence, the eight-state classification is considered more challenging than the three-state, because of the extremely unbalanced distribution of the eight-starctural states and their composition in native protein structures. As shown in Table 3, when the context-based scores

Table 5. 7-Fold Cross-Validation Accuracy for Eight-State Prediction in SCORPION

	Accuracy (%)	
	PSSM Only	PSSM+Score
Q_G	19.5	22.7
Q_H	91.8	92.4
Q_I	0.0	0.0
Q_E	82.1	82.9
Q_B	2.3	2.4
Q_S	19.4	22.3
Q_T	49.4	51.6
Q_C	65.5	66.1
overall	70.3	71.5

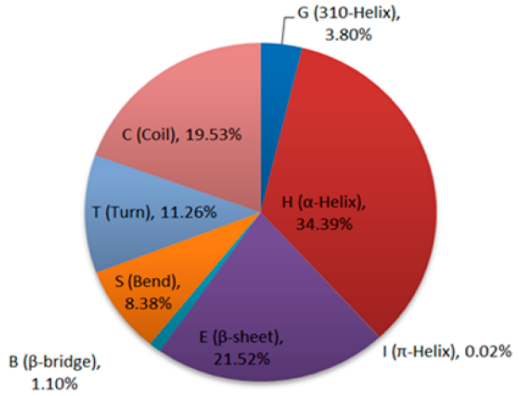


Figure 3. Distribution of eight-state secondary structures (3_{10}-helices (G), α-helices (H), π-helices (I), β-sheets (E), β-bridges (B), turns (T), bends (S), and coils (C)) in Cull16633.

are incorporated, accuracy enhancements are observed in all eight secondary structure classes, except for π-helix remaining at 0%. The very small fraction of residues adopting π-helix (0.02%) structure makes it almost impossible to predict.

To the best of our knowledge, SSpro8 and RaptorXss8 are the only two reported servers for eight-state secondary structure prediction. At the time, when this manuscript is written, SSpro8 is not available online; however, RaptorXss8 has demonstrated higher accuracy than SSpro8 in ref [16]. Tables 6 and 7 respectively compare Q8 and SOV8 accuracies of SCORPION with those of the RaptorXss8 server on CB513, CASP9, Manesh215, and Carugo338. Similar to three-state prediction, in order to guarantee fairness, we generate a new set of context-based scores by removing all sequences with a sequence identity of 25% or higher, relative to the sequences in the benchmarks from Cull16633, and all homologues with a sequence identity of higher than 25%, relative to the chains presented in the benchmarks are excluded from Cull7987 when training eight-state prediction neural networks. SCORPION has a higher accuracy, in seven states except for π-helix, than RaptorXss8, with ~2% improvements in Q8 accuracy and ~3% improvements in SOV8 accuracy.

DISCUSSION

Prediction with High Confidence. The feed-forward neural networks used in SCORPION provide a confidence interval to estimate the uncertainty of the prediction of each residue. When >90% confidence is obtained, the secondary structure prediction of a residue has rather high accuracy (98%

dx.doi.org/10.1021/ci400647u | J. Chem. Inf. Model. 2014, 54, 992−1002

Table 6. Comparison of Q8 Accuracy between SCORPION and RaptorXss8 on Benchmarks of CB513, CASP9, Manesh215, and Carugo338

		Q8 Accuracy (%)			
		CB513	CASP9	Manesh215	Carugo338
RaptorXss8	Q_8	65.59	69.31	67.69	66.64
	Q_G	17.54	20.58	18.43	19.20
	Q_H	89.96	92.90	90.22	89.91
	Q_I	0.00	0.00	0.00	0.00
	Q_E	77.68	81.64	79.60	79.45
	Q_B	0.09	0.00	0.32	0.44
	Q_S	15.87	18.11	17.80	17.14
	Q_T	48.02	51.45	51.28	50.11
	Q_C	63.29	59.37	63.73	63.36
C8-SCORPION	Q_8	67.22	71.54	69.71	68.44
	Q_G	21.81	22.46	23.01	22.42
	Q_H	90.95	93.58	91.42	90.55
	Q_I	00.00	0.00	0.00	0.00
	Q_E	80.31	83.95	82.54	81.44
	Q_B	1.43	1.04	1.79	2.22
	Q_S	19.86	23.41	21.99	21.95
	Q_T	49.44	53.87	53.03	52.65
	Q_C	63.54	62.82	64.93	64.69

Table 7. Comparison of SOV8 Accuracy between SCORPION and RaptorXss8 on Benchmarks of CB513, CASP9, Manesh215, and Carugo338

		SOV8 Accuracy (%)			
		CB513	CASP9	Manesh215	Carugo338
RaptorXss8	SOV_8	64.99	69.84	68.00	66.88
	SOV_G	20.04	21.27	20.81	21.71
	SOV_H	88.95	90.71	89.97	89.24
	SOV_I	0.00	0.00	0.00	0.00
	SOV_E	82.50	84.81	84.15	84.61
	SOV_B	0.09	0.00	0.32	0.44
	SOV_S	17.72	19.74	19.40	18.78
	SOV_T	50.79	53.92	54.73	53.74
	SOV_C	55.13	59.61	58.78	58.01
C8-SCORPION	SOV_8	67.66	73.47	70.79	69.50
	SOV_G	25.39	26.41	27.23	26.01
	SOV_H	92.24	93.66	92.80	91.65
	SOV_I	0.00	0.00	0.00	0.00
	SOV_E	85.25	88.68	87.05	86.57
	SOV_B	1.43	1.04	1.78	2.21
	SOV_S	21.88	25.36	23.70	23.95
	SOV_T	52.98	56.97	56.71	56.89
	SOV_C	56.28	64.28	60.69	60.14

for helices, 94% for sheets, and 90% for coils in three-state prediction and 99% for α-helix and 98% for β-sheet in eight-state prediction). Therefore, if consecutive residues in a helix or sheet segment are predicted with high confidence, misprediction of this helix or sheet segment is very unlikely. This is particularly useful in a variety of applications such as assigning secondary structures to NMR constraints or Cryo-EM density maps as well as limiting backbone torsion angle variations to reduce degree of freedoms in template-free predictions.[34]

Table 8 compares the total number of residues predicted with over 90% confidence in CB513, Manesh215, Carugo338, and CASP in SCORPION with and without context-based score encoding. Overall, there are 153 073 residues in these four benchmarks. For neural networks with PSSM-only encoding, the secondary structures of 60 222 (39.3% of all residues) residues are predicted with >90% confidence. When context-based scores are incorporated, the total number of residue secondary structure predictions with over 90% confidence is enhanced by 15.5%, to 69 537 (45.4% of all residues in benchmarks). Compared to the neural networks using PSSM only encoding, the numbers of residues predicted with over 90% confidence increase by 6.7%, 9.9%, and 42.3% in helices, strands, and coils, respectively.

Table 8. Total Number of Correct Predictions with >90% Confidence on Benchmarks of CB513, CASP9, Manesh215, and Carugo338

	Total Number of Correct Predictions	
	PSSM only	PSSM +score
number of residues predicted as H with 90% confidence	33 292	35 533
number of residues predicted as E with 90% confidence	13 298	14 611
number of residues predicted as C with 90% confidence	13 632	19 393
total number of residues predicted with 90% confidence	60 222	69 537
	(39.3%)	(45.4%)

A Three-State Prediction Example. Figure 4 depicts an example of three-state secondary structure prediction on protein 3NNQ chain A from CASP9 targets. The Q3 accuracy of PSSM-only neural networks is 83.33%. When context-based score encoding is incorporated, the Q3 accuracy is thereby improved to 90.35%. The main prediction difference is on the highlighted α-helix where the PSSM-only neural networks miss. Nevertheless, the context-based scores of the residues in the highlighted helix segment, as shown in Table 9, indicate that secondary structure of helix is highly favorable, which help the neural networks to identify the major part of this helix.

Analysis of Misclassifications. The majority of the prediction errors in the set of benchmarks result from the misclassifications of types E and C (8.34%) and the misclassification of types H and C (7.98%). The misclassification of H and E is much less common (~1.14%).

Table 10 shows the total number of misclassifications on the benchmark set of CB513, CASP9, Manesh215, and Carugo338. A significant reduction of misclassifications is observed upon the incorporation of context-based score encoding in the neural networks in SCORPION. Misclassifying H to E and E to H has been reduced by 14.08%, misclassifying H to C and C to H has been reduced by 8.83%, and the misclassification of E to C and C to E has been reduced by 5.18%.

We further analyze the misclassifications among helices and strands in SCORPION. The benchmark set (CB513, Manesh215, Carugo338, and CASP9) includes 10 011 helices ranging in length from 3 to 18 residues and 13 877 strands ranging in length from 1 to 16 residues. The total number of helices and strands that are correctly predicted as whole structures using our method are 4880 and 5467, respectively (the percentages are 48.75% and 39.40%, respectively), leaving the rest of the predictions with at least one residue misclassification in the

dx.doi.org/10.1021/ci400647u | J. Chem. Inf. Model. 2014, 54, 992−1002

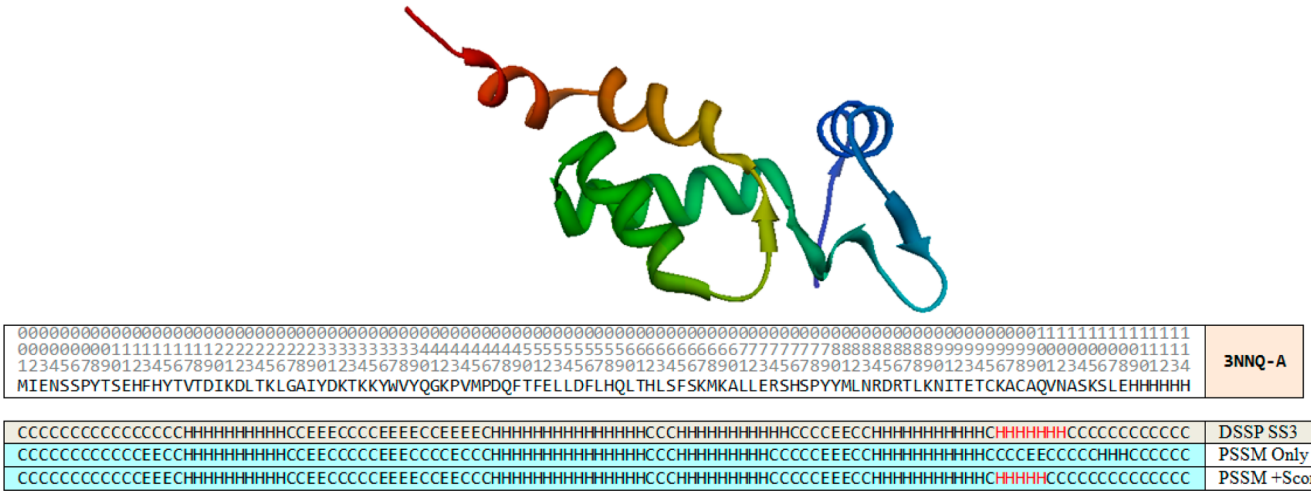


Figure 4. 3-state secondary structure prediction for protein 3NNQ chain A from CASP9 targets.

Table 9. Segment of 3NNQA (82−115)

		Context-Based Scores			Structure Predictions		
	amino acid, AA	U(C, ...)	U(E, ...)	U(H, ...)	PSSM +score	PSSM only	true structure
94	E	1.57	0.73	−1.66	H	H	H
95	T	3.17	0.43	−3.12	H	H	H
96	C	0.12	0.37	−0.08	C	C	C
97	K	5.11	1.01	−5.11	H	C	H
98	A	1.91	0.88	−2.11	H	C	H
99	C	5.5	0.58	−5.31	H	C	H
100	A	0.49	0.5	−0.53	H	E	H
101	Q	2.18	0.22	−2.22	H	E	H
102	V	0.91	−0.1	−0.65	C	C	H
103	N	2.13	0.22	−2.11	C	C	H
104	A	−3.66	0.27	3.68	C	C	C
105	S	−0.51	0.57	0.28	C	C	C

Table 10. Misclassifications of Secondary Structure States on Benchmark Sets

misclassification	PSSM only	PSSM+score
H → E	1864	1717
E → H	2318	1876
H → C	12 997	11 436
C → H	11 550	10 943
E → C	15 125	13 831
C → E	10 383	10 356

structure. Table 11 shows a detailed analysis of the misclassifications in helices and strands in this benchmark set.

Figure 5 shows the length distributions of helices and sheets when the whole structures are missing from prediction in our method. The most misclassified helices are 3_{10}-helices (all three-residue helices and a portion of the longer helices) and the most misclassified sheets are isolated β-bridges (residue size of 1). 3_{10} helices are much rarer in PDB than α-helices, which has a very low prediction accuracy (~14%) in the literature.[35] Also, most of the 3_{10} helices are short with three or four consecutive residues. Similarly, the isolated β-bridges are also rare and short and probably more importantly, the isolated β-bridges are the results of global interactions. Generating statistically meaningful context-based scores to sensitively identify 3_{10} helices and β-bridges is rather difficult. Therefore, the context-based scores help but with limited contribution in correctly identifying 3_{10} helices and β-bridges.

Table 11. Detailed Description of Helix and Strand Misclassifications on the Benchmark Set of CB513, CASP9, Manesh215, and Carugo338

	helix (%)	strand (%)	description
missed at begin (only)	5.36	8.28	only the first residue at the beginning of a structure
missed at middle	15.80	13.89	at least one is misclassified anywhere between the two ends
missed at end (only)	11.59	8.87	only the last residue at the end of a structure
missed at begin and end (only)	1.70	2.08	misclassified at both ends (2 residues)
missed the whole structure	16.80	27.48	the whole structure is misclassified

Sensitivity to Context-Based Scores Accuracy. The context-based scores estimate the favorability of a residue adopting a certain secondary structure conformation within its amino acid environment. The contribution of the context-based scores to secondary structure prediction depends on their accuracy. Figure 6 shows the distribution of the context-based scores accuracy and the corresponding accuracy improvements over the PSSM-only predictions in CB513, CASP9, Manesh215, and Carugo338 benchmarks. When the accuracies are lower than 40%, the context-based scores are close to random and thus their contributions to secondary structure prediction are marginal. The more accurate the scores are, the greater in improvements in accuracy over the PSSM-only predictions are achieved. When the accuracy of the context-based scores exceeds 80%, the average accuracy improvement over PSSM-only predictions reaches 2.5%. Unfortunately, the quality of the context-based scores relies on the statistics of singlets, doublets, and triplets,[31] which are limited by the number of proteins existing in the PDB. Therefore, it is not often that the context-based scores are highly accurate and the average of the context-based scores accuracy in SCORPION training and prediction is ~60%.

Secondary Structure Prediction and Intrinsically Disordered Protein Regions. An important application of secondary structure prediction is to predict intrinsically disordered protein regions. Disordered proteins typically have

dx.doi.org/10.1021/ci400647u | J. Chem. Inf. Model. 2014, 54, 992−1002

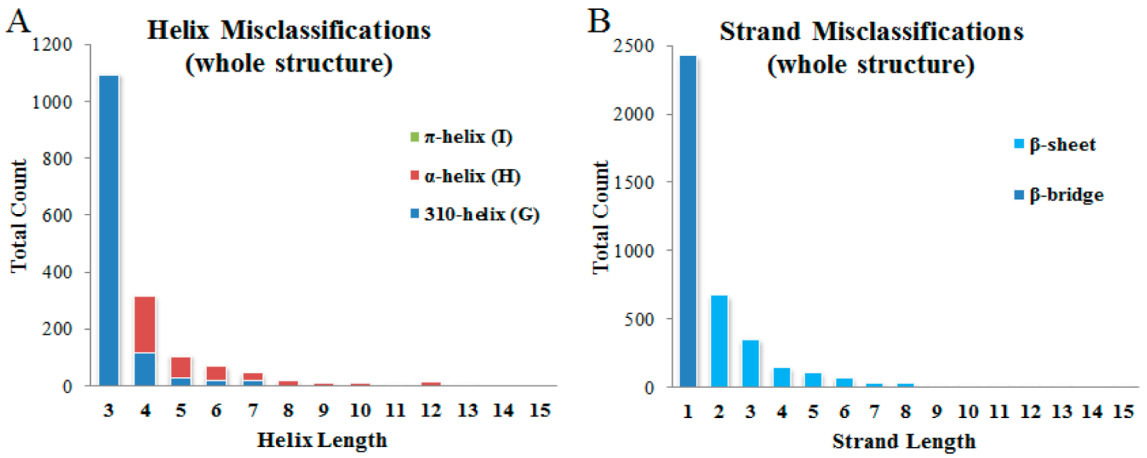


Figure 5. Histograms of the numbers of (A) misclassified helices and (B) strands in SCORPION, based on their lengths.

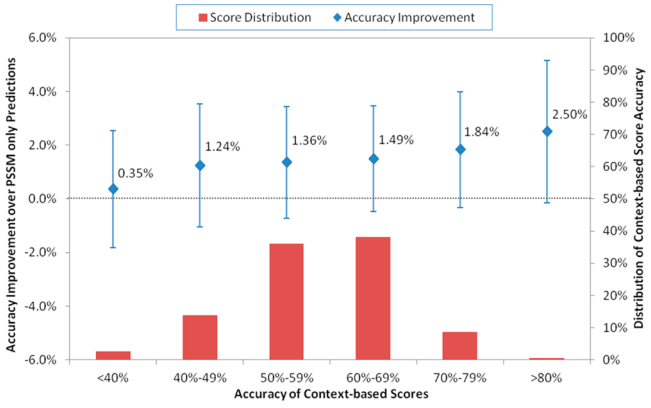


Figure 6. Distribution of context-based scores accuracy and its correlation with accuracy improvement over PSSM only prediction in CB513, CASP9, Manesh215, and Carugo338 benchmarks. The accuracy of context-based scores is measured by calculating the percentage of residues whose lowest secondary structure conformation (H, E, or C) scores agree with the DSSP assignments in a protein sequence.

a low content of secondary structures. In fact, the predicted secondary structures are often incorporated as important information in disorder region predictors such as DISOPRED[36] and SPINE-D.[37] Figure 7 shows an example of the secondary structures predicted by SCORPION on Thylakoid-soluble phosphoprotein TSP9,[38] which is an intrinsic disordered protein that reacts to changes in light conditions from the photosynthetic membrane. The majority of the N-terminal α-helix is predicted correctly with high confidence (8+), except for two residues at the end. For the unstructured coils, 85.3% of the 75 residues are correctly identified while the rest 14.7% are misclassified as helices or strands but with low prediction confidence (6−). Coupled with other residue feature predictors such as solvent accessibility, B-factor, contact, and disulfide bonding state, the accuracy improvement in secondary structure prediction using context-based scores has the potential to enhance determination of intrinsically disordered regions.

Toward the Theoretical Upper Bound. In this paper, we report a 7-fold cross-validation Q3 accuracy of 82.7% in SCORPION. To close the gap in Q3 accuracy to 88%−90%—which is the theoretical upper bound of three-state secondary structure prediction 4—our future efforts in SCORPION will include (1) deriving better context-based scores, particularly for calculating high-order interactions, (2) obtaining more-precise PSSM substitution matrices by better multiple sequence alignment algorithms on increasingly large sequence databases, and (3) developing advanced machine learning algorithms that can capture residue−residue interactions in longer range and handle an increasingly large number of known protein structures.

AUTHOR INFORMATION

Corresponding Author

*E-mail: yaohang@cs.odu.edu (Y. Li).

Present Address

†(A. Yaseen) Department of Mathematics and Computer Science, Central State University, Wilberforce, OH 45384.

Notes

Web services implementing our SCORPION program are available at http://hpcr.cs.odu.edu/c3scorpion for three-state prediction and http://hpcr.cs.odu.edu/c8scorpion for eight-state prediction. A stand-alone SCORPION program is also available for download at the website.

The authors declare no competing financial interest.

```
1       SAAKGTAETK QEKSFVDWLL GKITKEDQFY ETDPILRGGD VKSSGSTSGK   Sequence
        CCCCCCCCCC CCCCHHHHHH HHHCCCCCCC CCCCCCCCCC CCCCCCCCCC   DSSP Assignment
        CCCCCCCCCH HHCCHHHHHH HCCCCCCCCE CCCCCCCCCC CCCCCCCCCC   SCORPION Prediction
        9987887646 6579899997 5667766444 4897667986 5688888888   Prediction Confidence

51      KGGTTSGKKG TVSIPSKKKN GNGGVFGGLF AKKD
        CCCCCCCCCC CCCCCCCCCC CCCCCCCCCC CCCC
        CCCCCCCCCC CECCCCCCCC CCCCEEECEE ECCC
        8898898888 5647887668 9997654445 4789
```

Figure 7. Secondary structures predicted by SCORPION on Thylakoid-soluble phosphoprotein TSP9.

dx.doi.org/10.1021/ci400647u | J. Chem. Inf. Model. 2014, 54, 992−1002

■ ACKNOWLEDGMENTS

Y.L. acknowledges support from NSF (under Grant No. 1066471) and ODU (2013 Multidisciplinary Seed grant).

■ REFERENCES

(1) Lesk, A. M.; Lo Conte, L.; Hubbard, T. J. P. Assessment of novel fold targets in CASP4: Predictions of three-dimensional structures, secondary structures, and interresidue contacts. *Proteins: Struct., Funct., Bioinf.* **2001**, 98−118.
(2) Chou, P. Y.; Fasman, G. D. Prediction of Protein Conformation. *Biochemistry* **1974**, *13*, 222−245.
(3) Garnier, J.; Osguthorpe, D. J.; Robson, B. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.* **1978**, *120*, 97−120.
(4) Rost, B. Review: Protein secondary structure prediction continues to rise. *J. Struct. Biol.* **2001**, *134*, 204−218.
(5) Qian, N.; Sejnowski, T. J. Predicting the Secondary Structure of Globular-Proteins Using Neural Network Models. *J. Mol. Biol.* **1988**, *202*, 865−884.
(6) Rost, B.; Sander, C. Prediction of Protein Secondary Structure at Better Than 70% Accuracy. *J. Mol. Biol.* **1993**, *232*, 584−599.
(7) Rost, B.; Sander, C. Secondary structure prediction of all-helical proteins in two states. *Protein Eng.* **1993**, *6*, 831−837.
(8) Pollastri, G.; McLysaght, A. Porter: A new, accurate server for protein secondary structure prediction. *Bioinformatics* **2005**, *21*, 1719−1720.
(9) Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **1999**, *292*, 195−202.
(10) Rost, B.; Sander, C. Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins: Struct., Funct., Bioinf.* **1994**, *19*, 55−72.
(11) Petersen, B.; Petersen, T. N.; Andersen, P.; Nielsen, M.; Lundegaard, C. A generic method for assignment of reliability scores applied to solvent accessibility predictions. *BMC Struct. Biol.* **2009**, *9*, 51.
(12) Cole, C.; Barber, J. D.; Barton, G. J. The Jpred 3 secondary structure prediction server. *Nucleic Acids Res.* **2008**, *36*, W197−W201.
(13) Pollastri, G.; Przybylski, D.; Rost, B.; Baldi, P. Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins: Struct., Funct., Genet.* **2002**, *47*, 228−235.
(14) Lin, K.; Simossis, V. A.; Taylor, W. R.; Heringa, J. A simple and fast secondary structure prediction method using hidden neural networks. *Bioinformatics* **2005**, *21*, 152−159.
(15) Dor, O.; Zhou, Y. Q. Achieving 80% ten-fold cross-validated accuracy for secondary structure prediction by large-scale training. *Proteins: Struct., Funct., Bioinf.* **2007**, *66*, 838−845.
(16) Ward, J. J.; McGuffin, L. J.; Buxton, B. F.; Jones, D. T. Secondary structure prediction with support vector machines. *Bioinformatics* **2003**, *19*, 1650−1655.
(17) Wang, Z.; Zhao, F.; Peng, J.; Xu, J. Protein 8-class secondary structure prediction using conditional neural fields. *Proteomics* **2011**, *11*, 3786−3878.
(18) Kabsch, W.; Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983**, *22*, 2577−2637.
(19) Heinig, M.; Frishman, D. STRIDE: A web server for secondary structure assignment from known atomic coordinates of proteins. *Nucleic Acids Res.* **2004**, *32*, W500−W502.
(20) Pollastri, G.; Martin, A. J. M.; Mooney, C.; Vullo, A. Accurate prediction of protein secondary structure and solvent accessibility by consensus combiners of sequence and structure information. *BMC Bioinf.* **2007**, *8*, 201.
(21) Sussman, J. L.; Lin, D. W.; Jiang, J. S.; Manning, N. O.; Prilusky, J.; Ritter, O.; Abola, E. E. Protein Data Bank (PDB): Database of three-dimensional structural information of biological macromolecules. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **1998**, *54*, 1078−1084.
(22) Cuff, J. A.; Barton, G. J. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins: Struct., Funct., Genet.* **2000**, *40*, 502−511.
(23) Ahmad, S.; Gromiha, M. M.; Sarai, A. Real value prediction of solvent accessibility from amino acid sequence. *Proteins: Struct., Funct., Genet.* **2003**, *50*, 629−635.
(24) Carugo, O. Predicting residue solvent accessibility from protein sequence by considering the sequence environment. *Protein Eng.* **2000**, *13*, 607−609.
(25) Kinch, L. N.; Shi, S.; Cheng, H.; Cong, Q.; Pei, J. M.; Mariani, V.; Schwede, T.; Grishin, N. V. CASP9 target classification. *Proteins: Struct., Funct., Bioinf.* **2011**, *79*, 21−36.
(26) Wang, G. L.; Dunbrack, R. L. PISCES: A protein sequence culling server. *Bioinformatics* **2003**, *19*, 1589−1591.
(27) Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25*, 3389−3402.
(28) Zemla, A.; Venclovas, C.; Fidelis, K.; Rost, B. A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment. *Proteins: Struct., Funct., Genet.* **1999**, *34*, 220−223.
(29) Rata, I. A.; Li, Y. H.; Jakobsson, E. Backbone Statistical Potential from Local Sequence−Structure Interactions in Protein Loops. *J. Phys. Chem. B.* **2010**, *114*, 1859−1869.
(30) Garnier, J.; Gibrat, J. F.; Robson, B. GOR method for predicting protein secondary structure from amino acid sequence. *Methods Enzymol.* **1996**, *266*, 540−553.
(31) Li, Y.; Liu, H.; Rata, I.; Jakobsson, E. Building a Knowledge-Based Statistical Potential by Capturing High-Order Inter-residue Interactions and Its Applications in Protein Secondary Structure Assessment. *J. Chem. Inf. Model.* **2013**, *53*, 500−508.
(32) Sippl, M. J. Calculation of Conformational Ensembles from Potentials of Mean Force—An Approach to the Knowledge-Based Prediction of Local Structures in Globular-Proteins. *J. Mol. Biol.* **1990**, *213*, 859−883.
(33) Samudrala, R.; Moult, J. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *J. Mol. Biol.* **1998**, *275*, 895−916.
(34) Faraggi, E.; Yang, Y.; Zhang, S.; Zhou, Y. Predicting continuous local structure and the effect of its substitution for secondary structure in fragment-free protein structure prediction. *Structure* **2009**, *17*, 1515−1527.
(35) Pal, L.; Basu, G. Neural network prediction of 3(10)-helices in proteins. *Indian J. Biochem. Biophys.* **2001**, *38*, 107−114.
(36) Ward, J. J.; McGuffin, L. J.; Bryson, K.; Buxton, B. F.; Jones, D. T. The DISOPRED server for the prediction of protein disorder. *Bioinformatics* **2004**, *20*, 2138−2147.
(37) Zhang, T.; Faraggi, E.; Xue, B.; Dunker, A. K.; Uversky, V. N.; Zhou, Y. SPINE-D: Accurate prediction of short and long disordered regions by a single neural-network based method. *J. Biomol. Struct. Dyn.* **2012**, *29*, 799−813.
(38) Song, J.; Lee, M. S.; Carlberg, I.; Vener, A. V.; Markley, J. L. Micelle-induced folding of spinach thylakoid soluble phosphoprotein of 9 kDa and its functional implications. *Biochemistry* **2006**, *45*, 15633−15643.