JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Building a Knowledge-Based Statistical Potential by Capturing High-Order Inter-residue Interactions and its Applications in Protein Secondary Structure Assessment

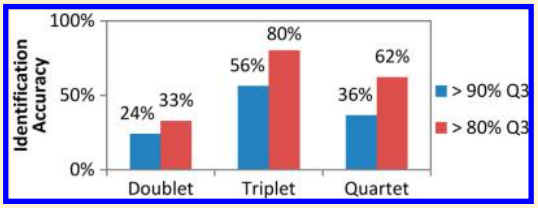
Yaohang Li,*[,†] Hui Liu,[‡] Ionel Rata,[§] and Eric Jakobsson[⊥]

[†]Department of Computer Science, Old Dominion University, Norfolk, Virginia, United States

[‡]Center for Biophysics and Computational Biology and [⊥]Department of Molecular and Integrative Physiology, Beckman Institute, and National Center for Supercomputing Applications, University of Illinois at Urbana−Champaign, Illinois, United States

[§]National Institute for Physics and Nuclear Engineering (IFIN-HH), R-77125, Bucharest-Magurele, Romania

ABSTRACT: The rapidly increasing number of protein crystal structures available in the Protein Data Bank (PDB) has naturally made statistical analyses feasible in studying complex high-order inter-residue correlations. In this paper, we report a context-based secondary structure potential (CSSP) for assessing the quality of predicted protein secondary structures generated by various prediction servers. CSSP is a sequence-position-specific knowledge-based potential generated based on the potentials of mean force approach, where high-order inter-residue interactions are taken into consideration. The CSSP potential is effective in identifying secondary structure predictions with good quality. In 56% of the targets in the CB513 benchmark, the optimal CSSP potential is able to recognize the native secondary structure or a prediction with Q3 accuracy higher than 90% as best scored in the predicted secondary structures generated by 10 popularly used secondary structure prediction servers. In more than 80% of the CB513 targets, the predicted secondary structures with the lowest CSSP potential values yield higher than 80% Q3 accuracy. Similar performance of CSSP is found on the CASP9 targets as well. Moreover, our computational results also show that the CSSP potential using triplets outperforms the CSSP potential using doublets and is currently better than the CSSP potential using quartets.



1. INTRODUCTION

Prediction of protein secondary structure from the primary sequence is an important step toward prediction of tertiary structure. The more accurately the secondary structure can be predicted, the smaller the search space for the tertiary structure prediction. At the core of the secondary structure prediction problem is the derivation of knowledge for secondary structure assignment. The knowledge is contained in the Protein Data Bank (PDB), which includes 83 983 protein structures as of August 21, 2012, specifically in the secondary structure assignment as reported in the PDB. Nevertheless, generation of knowledge for secondary structure assignment is complicated by several sources of inherent error. In the first place, the tertiary structure from which the secondary structure is derived has a resolution ranging from one to a few angstroms, sufficient to alter the local secondary structure assignment. Second, the algorithms that translate the tertiary structure to a secondary structure necessarily have a tolerance for a range of backbone torsion angles that define any of the well-defined secondary structures. These two bases for uncertainty about the precise secondary structure of proteins in PDB contribute to the fact that the maximum meaningful secondary structure prediction accuracy that can ever be obtained, given the noise in the experimental data and its analysis, is significantly less than 100%. It has been estimated at about 88−90%.[1]

Pirovano and Heringa[2] have recently done a critical comparative study of protein secondary structure prediction methods. By the metrics they use in their study, which are generally consistent with other studies and with our group's experience (unpublished), the existing methods provide accuracies near 80%. Wei et al.[3] have utilized linear optimization to provide weighting for a consensus prediction of seven different methods. They report consensus predictions have averagely a couple of percent better than the best single method, suggesting that a consensus method may move the state of the art a significant fraction toward the theoretical maximum, but still far short of the theoretical maximum. As a basis for tertiary structure prediction, moving the percent of inaccuracy from the high teens to 10% 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. Because of a combinatorial expansion of possibilities, such an improvement in secondary structure prediction would reduce the search space for predicting tertiary structure many-fold.

In the present paper, we describe an approach of integrating knowledge for secondary structure assignment into a knowledge-based potential to assess the quality of predicted secondary

Received: April 27, 2012
Published: January 21, 2013

structures. We hypothesize that incorporating higher-order inter-residue correlations into the knowledge-based potential is likely to lead to high accuracy. In particular, we note that it is reasonable to expect correlations of identity for pairs of residues one position removed from each other in turns, two positions removed from each other in β-strands, and three and four positions removed from each other in helices.

When there were relatively few experimental structures available, capturing high-order inter-residue interactions into knowledge-based potentials was difficult due to lack of statistical samples. We note that the sample size for specific doublets in the PDB is 1/20 of that for individual residues (singlets), for specific triplets is 1/20 of that for doublets, and for quartets 1/20 of that for triplets. The fractions are even smaller if rare amino acids are involved. However recently, as an increasing number of high-resolution protein crystal structures are available in the PDB, and powerful computers are available to sort through larger dimension combinatory, it has become feasible to derive knowledge for high-order inter-residue interactions and incorporate it into a knowledge-based potential.

Most of the secondary structure prediction methods[4−13] consider inter-residue correlation implicitly by encoding a window of 15−21 residues in neural networks or other learning machines. Although these methods have achieved certain success, the neural networks or learning machines work like "black boxes," which provide little understandable information in the relation between inter-residue interactions and the secondary structures. Only a few methods have attempted to estimate (high-order) inter-residue correlations explicitly. Miyazawa and Jernigan[14] developed a secondary structure energy using the potentials of mean force method by considering the three-body interactions among three consecutive residues. The GOR4[15] method treats inter-residue interactions as information functions of events and integrates them according to information theory. The original GOR4 program only considers singlets and doublets within a window. The later GOR5 program[16] takes higher-order interactions such as triplets into account but finds that the improvement is only 0.3%. The authors suggest that "better optimization and larger database" are necessary for further accuracy improvement.[16] More recently, Madera et al.[17] proposed a simple k-mer model using a conditional random field to achieve more "realistic" secondary structure predictions.

In this paper, we derive statistics of singlets, doublets, triplets, and quartets of residues with specific relative occurrences at sequence positions and then convert them to inter-residue interaction potentials using the potentials of mean force[18] approach. A context-based secondary structure potential (CSSP) integrating these inter-residue interaction potentials is developed for assessing predicted protein secondary structures. We use the cull data sets (CullPDB) generated by the PISCES server[19] as the training sets for CSSP. We test CSSP by using it to evaluate the predicted secondary structures generated by 10 public secondary structure prediction servers, including GOR4,[15,16] HNN,[4] SAM,[5] Jpred,[6] ProfPHD,[7,8] Psipred,[9] Jufo,[10] Netsurfp,[11] SSPRO4,[12] and Porter,[13] using a commonly used set of sequences known as the CB513 benchmark.[20] In addition to CB513, we also test CSSP on the target sequences of CASP9[21] (critical assessment of protein structure prediction). For the correctness of our computational experiments, chains in the testing sets and their homologues are removed from the CullPDB to ensure the separation of training set and testing set. Accuracy comparisons of potentials with different orders and ranges of inter-residue interactions are also made.

2. METHODS

2.1. Knowledge-Based Statistical Potential for *N*-Residue Fragments with High-Order Inter-Residue Interactions. *2.1.1. Formation of the k-let Potential.* Our formation of the potential is based on the mean-force potential energy according to the Boltzmann formula.[18] We first come up with a statistical potential for a k-let at residue positions $i_1, i_2, ..., i_k$ in a protein sequence. The derivation of a statistical potential $U(R_{i_1}R_{i_2} ... R_{i_k}, C_{i_1}C_{i_2} ... C_{i_k})$ for a sequence-structure correlated k-let starts from the common form of statistical potential calculation using inverse Boltzmann theorem:

$$U(R_{i_1}R_{i_2} ... R_{i_k}, C_{i_1}C_{i_2} ... C_{i_k}) = -RT \ln \frac{P_{obs}(C_{i_1}C_{i_2} ... C_{i_k}|R_{i_1}R_{i_2} ... R_{i_k})}{P_{ref}(C_{i_1}C_{i_2} ... C_{i_k}|R_{i_1}R_{i_2} ... R_{i_k})},$$
$$i_1 \neq i_2 \neq ... \neq i_k$$

where $P_{obs}(C_{i_1}C_{i_2} ... C_{i_k}|R_{i_1}R_{i_2} ... R_{i_k})$ is the observed probability of k-let $R_{i_1}R_{i_2} ... R_{i_k}$ with conformation $C_{i_1}C_{i_2} ... C_{i_k}$, $P_{ref}(C_{i_1}C_{i_2} ... C_{i_k}|R_{i_1}R_{i_2} ... R_{i_k})$ is the probability of the reference state, R is the gas constant, and T is the temperature. Using the frequency values to estimate the probability $P_{obs}(C_{i_1}C_{i_2} ... C_{i_k}|R_{i_1}R_{i_2} ... R_{i_k})$ and applying the conditional probability method described in Samudrala and Moult,[22] $U(R_{i_1}R_{i_2} ... R_{i_k}, C_{i_1}C_{i_2} ... C_{i_k})$ can be written as

$$U(R_{i_1}R_{i_2} ... R_{i_k}, C_{i_1}C_{i_2} ... C_{i_k}) = -RT \ln \frac{\frac{N_{obs}(C_{i_1}C_{i_2} ... C_{i_k}, R_{i_1}R_{i_2} ... R_{i_k})}{N_{obs}(R_{i_1}R_{i_2} ... R_{i_k})}}{\frac{N_{obs}(C_{i_1}C_{i_2} ... C_{i_k})}{N_{total}}}$$

where $N_{obs}(C_{i_1}C_{i_2} ... C_{i_k}, R_{i_1}R_{i_2} ... R_{i_k})$ is the observed number of k-let $R_{i_1}R_{i_2} ... R_{i_k}$ with conformation $C_{i_1}C_{i_2} ... C_{i_k}$ in a protein structure database, $N_{obs}(R_{i_1}R_{i_2} ... R_{i_k})$ is the number of observations of $R_{i_1}R_{i_2} ... R_{i_k}$, $N_{obs}(C_{i_1}C_{i_2} ... C_{i_k})$ is the number of observations of $C_{i_1}C_{i_2} ... C_{i_k}$, and N_{total} is the total number of observations.

Two k-lets are of the same kind if their residue positions $i_1, i_2, ..., i_k$ and $i'_1, i'_2, ..., i'_k$ (in the same or different protein sequences) have the same relative sequence distances: $i_1 - i'_1 = i_2 - i'_2 = \cdots = i_k - i'_k$ Then, $N_{obs}(C_{i_1}C_{i_2} ... C_{i_k}, R_{i_1}R_{i_2} ... R_{i_k})$ can be obtained by counting the total number of occurrences of k-lets having conformation $C_{i_1}C_{i_2} ... C_{i_k}$ at the same relative residue positions as $i_1, i_2, ..., i_k$ in the protein structure database. Similar calculations can be applied to obtain $N_{obs}(R_{i_1}R_{i_2} ... R_{i_k})$, $N_{obs}(C_{i_1}C_{i_2} ... C_{i_k})$, and N_{total}.

For simplicity, we use $U(i_1, i_2, ..., i_k)$ to represent the k-let potential $U(R_{i_1}R_{i_2} ... R_{i_k}, C_{i_1}C_{i_2} ... C_{i_k})$ in the rest of the paper.

2.1.2. Interaction Potential. We denote $INT(i_1, i_2)$ to capture the two-body (doublet) interaction potential energy between residues R_{i_1} and R_{i_2}

$$INT(i_1, i_2) = U(i_1, i_2) - U(i_1) - U(i_2)$$

Similarly, the higher order three-body interactions $INT(i_1, i_2, i_3)$ of triplet residues R_{i_1}, R_{i_2}, and R_{i_3} can be expressed as

$$INT(i_1, i_2, i_3) = U(i_1, i_2, i_3) - U(i_1) - U(i_2) - U(i_3) - INT(i_1, i_2) - INT(i_1, i_3) - INT(i_2, i_3)$$

dx.doi.org/10.1021/ci300207x | J. Chem. Inf. Model. 2013, 53, 500−508

For a k-let, the high order k-body interactions $\mathrm{INT}(i_1, i_2, \ldots i_k)$ of residues $R_{i_1}R_{i_2} \ldots R_{i_k}$ can be generalized as

$$\begin{aligned}\mathrm{INT}(i_1, i_2, \ldots, i_k) &= U(i_1, i_2, \ldots, i_k) - \sum_{j=1}^{k} U(i_j) \\ &- \sum_{j_1=1}^{k}\sum_{j_2=j_1+1}^{k} \mathrm{INT}(i_{j_1}, i_{j_2}) \\ &- \sum_{j_1=1}^{k}\sum_{j_2=j_1+1}^{k}\sum_{j_3=j_2+1}^{k} \mathrm{INT}(i_{j_1}, i_{j_2}, i_{j_3}) \\ &\ldots - \sum_{j_1=1}^{k}\sum_{j_2=j_1+1}^{k} \ldots \sum_{j_{k-1}=j_{k-2}+1}^{k} \mathrm{INT}(i_{j_1}, \ldots, i_{j_{k-1}})\end{aligned}$$

2.1.3. Potential of N-Residue Fragment. By considering up to k-body interactions, we can represent the mean-force potential $U(M+1, M+2, \ldots, M+N)$ of an N-residue fragment $R_{M+1}R_{M+2} \ldots R_{M+N}$ starting at the $(M+1)$th position in a protein sequence as

$$\begin{aligned}&U(M+1, M+2, \ldots, M+N) \\ &= \sum_{j=1}^{N} U(M+j) + \sum_{j_1=1}^{N}\sum_{j_2=j_1+1}^{N} \mathrm{INT}(M+j_1, M+j_2) \\ &+ \sum_{j_1=1}^{N}\sum_{j_2=j_1+1}^{N}\sum_{j_3=j_2+1}^{N} \mathrm{INT}(M+j_1, M+j_2, M+j_3) \\ &\ldots + \sum_{j_1=1}^{N}\sum_{j_2=j_1+1}^{N} \ldots \sum_{j_k=j_{k-1}+1}^{N} \mathrm{INT}(M+j_1, M+j_2, \\ &\ldots, M+j_k)\end{aligned}$$

By substituting the interaction potential with k-let potential and combining the common terms, the potential energy of an N-residue fragment $R_{M+1}R_{M+2} \ldots R_{M+N}$ is simplified as the weighted sum of potentials of singlets, doublets, triplets, ..., and up to k-lets.

$$\begin{aligned}&U(M+1, M+2, \ldots, M+N) \\ &= \underbrace{w_1 \sum_{j=1}^{N} U(j)}_{singlet} + \underbrace{w_2 \sum_{j_1=1}^{N}\sum_{j_2=j_1+1}^{N} U(M+j_1, M+j_2)}_{doublet} \\ &+ \underbrace{w_3 \sum_{j_1=1}^{N}\sum_{j_2=j_1+1}^{N}\sum_{j_3=j_2+1}^{N} U(M+j_1, M+j_2, M+j_3) + \cdots}_{triplet} \\ &+ \underbrace{w_s \sum_{j_1=1}^{N}\sum_{j_2=j_1+1}^{N} \ldots \sum_{j_s=j_{s-1}+1}^{N} U(M+j_1, M+j_2, \ldots, M+j_s) + \cdots}_{s-let} \\ &+ \underbrace{w_k \sum_{j_1=1}^{N}\sum_{j_2=j_1+1}^{N} \ldots \sum_{j_k=j_{k-1}+1}^{N} U(M+j_1, M+j_2, \ldots, M+j_k)}_{k-let}\end{aligned}$$

where the weights w_s are

$$w_s = \sum_{j=s}^{k} (-1)^{j-s} \binom{N-k}{j-s}$$

Using the potential of N-residue fragments and removing the overlapping parts, the overall potential energy of a protein with L residues is

$$\begin{aligned}U_{\mathrm{protein}} &= U(1, \ldots, N) + U(2, \ldots, N+1) - U(2, \ldots, N) \\ &+ U(3, \ldots, N+2) - U(3, \ldots, N+1) + \ldots \\ &= \sum_{j=1}^{L-N+1} U(j, \ldots, j+N-1) \\ &- \sum_{j=2}^{L-N+1} U(j, \ldots, j+N-2)\end{aligned}$$

2.2. Potentials for Secondary Structure Prediction Evaluation. *2.2.1. Data Sets.* We use the CullPDB data sets generated by the PISCES server[19] to collect k-let samples to produce CSSP potentials to evaluate secondary structure predictions. The CullPDB data sets generated on October 21, 2011 with maximum 3.0 Å resolution and maximum 1.0 R-factor are selected. A public benchmark CB513 and targets in CASP9 are used as testing sets to validate our methods. To ensure the correctness of our computational experiments, we enforce the separation of training set and testing set by excluding all sequences with greater than 25% identity to any sequence in the testing set from the CullPDB data sets when the k-let samples are extracted to calculate the statistical potential. Moreover, the k-let samples with missing residues are discarded. Furthermore, due to the fact that PSI-BLAST[23] is usually unable to generate profiles for short sequences, the protein sequences with lengths less than 30 are also removed from the CullPDB data sets.

2.2.2. Estimation of k-let Probability. The weighted frequency value of a k-let of a certain secondary structure appeared in the CullPDB is used to estimate the probability of the k-let sample adopting this secondary structure. The weights of k-let samples are based on the PSSM (position specific score matrix) frequencies at each residue position. PSSM data contains evolutionary information derived from sequence homologues. For a given protein in the CullPDB data sets, PSI-BLAST[23] is used to search against the NR (non-redundant) database with E-value = 0.001 and at most 3 iterations. After the PSSM file is generated, weights are calculated according to the frequency of each residue appearing in a specific position of the sequence. For example, the following figure shows a segment of a PSSM frequency table where a four-residue fragment "ASYK" has secondary structure of "HHHC".Then, in triplet calculation,

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
											...									
1 A	88	0	0	0	0	0	0	0	0	0	0	0	0	0	0	12	0	0	0	0
2 S	24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	29	47	0	0	0
3 Y	0	0	0	0	0	0	0	0	0	0	0	0	0	49	0	0	0	0	51	0
4 K	1	9	6	2	0	13	10	0	2	0	0	34	0	0	1	6	16	0	0	0
											...									

weight of $88/100 \times 24/100 \times 49/100 = 0.103$ is counted toward N_{obs}(HHH, AAF) at triplet position "1_2_3", weight of $88/100 \times 47/100 \times 9/100 = 0.037$ is counted toward N_{obs}(HHH, ATR) at triplet position "1_2_4", weight of $12/100 \times 51/100 \times 34/100 = 0.021$ is counted toward N_{obs}(HHC, SYK) at triplet position "1_3_4", and so on. These combinations give many samples with different weights for calculating the frequency of k-lets at different positions.

3. RESULTS

3.1. CSSP Using Triplets. We first investigate the sensitivity and accuracy of our new knowledge-based statistical potential

dx.doi.org/10.1021/ci300207x | J. Chem. Inf. Model. 2013, 53, 500−508

CSSP by incorporating three-body inter-residue (triplet) interactions using the CB513 benchmark. Since CSSP does not include statistics for unidentified residues, we only consider the 507 out of the 513 targets in CB513 benchmark excluding the six with unidentified residues in the protein sequence. We create a secondary structure set composed of the predicted secondary structures from 10 public prediction servers as well as the native structure and test if the knowledge-based potential can recognize the high quality predictions. The 10 prediction servers we used include GOR4, HNN, SSPRO4, PORTER, NETSURFP, PSIPRED, SAM, PROFPHD, and JUFO. The precisions of the predicted secondary structures are measured by the Q3 accuracy, i.e., the total accuracy of three classes—α-helix, β-strand, and coil. These predicted secondary structure conformations have very different qualities. GOR4 is an early statistical model based on frequencies of amino acid pairs, and HNN is an early method using neural network for classification—both have relatively low accuracy compared to the modern secondary structure prediction servers. On the other hand, SSPRO4 and PORTER take advantage of the homologue structural information for prediction. When structures of homologues with 50% or higher sequence identity are available, SSPRO4 or PORTER can often produce high quality predictions with Q3 accuracy around 80−90% or even 100%.[24] NETSURFP, PSIPRED, SAM, PROFPHD, JUFO, and JPRED are popularly used prediction servers using neural networks,[4,9,12,13] hidden Markov chains,[5] support vector machine (SVM),[8] or consensus[10] methods, typically having Q3 accuracy between 70% and 80%. Table 1 compares the performance of the 10 public servers for secondary structure prediction on CB513.

Table 1. Performance of 10 Secondary Structure Prediction Methods on CB513

methods	no. of targets with Q3 > 90%	no. of targets with Q3 > 80%
GOR4	1 (0.20%)	20 (3.94%)
HNN	8 (1.58%)	40 (7.89%)
NETSURFP	29 (5.72%)	230 (45.36%)
PSIPRED	58 (11.44%)	327 (64.50%)
SAM	25 (4.93%)	226 (44.58%)
SSPRO4	432 (85.21%)	468 (92.31%)
PORTER	453 (89.35%)	492 (97.04%)
PROFPHD	10 (1.97%)	129 (25.44%)
JPRED	32 (6.31%)	269 (53.06%)
JUFO	5 (0.99%)	126 (24.85%)

In this paper, we measure the identification accuracy of the CSSP potentials by the percentage of targets in CB513 in which the predicted structures yielding the lowest potential energy values have Q3 accuracies higher than 80% or 90%. Because the secondary structure assignments based on crystal structure have ∼10% errors themselves[25,26] as inferred from differences between different X-ray structures and NMR models of the same protein and from inconsistency of secondary structure assignments by different methods of different parameters, e.g., DSSP[27] and STRIDE,[28] 90% Q3 prediction accuracy is usually considered as the upper bound of secondary structure prediction. Predictions with 80% Q3 accuracies are also regarded as models with high precision.

A number of tests have been carried out to determine the optimal parameters for CSSP using triplets, including the Cull data sets, the fragment size, and the number of iterations in PSI-BLAST.

Figure 1 compares the identification accuracies when Cull data sets with maximum pairwise mutual sequence identity ranging from 20% to 90% are used to generate the CSSP potentials with fragment size seven in CB513. On one hand, data sets with lower sequence identity have fewer protein sequences and thus fewer triplet samples. On the other hand, samples may bias to certain protein families in data sets with higher sequence identity. Figure 1 shows that the Cull data set with maximum 50% sequence identity have the best compromise of sampling accuracy by showing the highest overall identification percentages. For the Cull data set with maximum 50% sequence identity, in 56.2% and 80.1% of the CB513 targets, CSSP can pick up one from the 10 predicted structures generated by the prediction servers having higher than 90% and 80% Q3 accuracy, respectively.

Figure 2 shows the overall Q3 accuracy in CB513 of varying fragment sizes using CSSP trained by the Cull data set with maximum 50% sequence identity. The CSSP with fragment size of seven yields the best result, with overall Q3 accuracy of 88.2%. The optimum fragment size of seven has certain biological meaning—triplet residues in helix, strand, and coil are strongly correlated at relative positions 1−3−5, 1−4−7, and 1−2−3, respectively. For bigger fragment sizes than seven, the identification accuracies drop gradually, due to the reason that the importance of long distance inter-residue correlation decreases while the statistical sampling noise accumulates.

Since CSSP takes advantage of the evolutionary information to generate the statistics for *k*-lets, the evolutionary distance occupied by a protein and its homologues also affects the accuracy of CSSP. Figure 3 investigates the accuracy of CSSP using weighted frequencies generated from PSSM using 3 and 6 PSI-BLAST iterations. One can find that both CSSPs yield similar performance, but the one based on PSI-BLAST using three iterations is slightly more sensitive. This may be due to the fact that more PSI-BLAST iterations bring in more less-related homologues in the protein family with likely more diverse structures, which reduces the sensitivity of the *k*-let statistics.

In addition to CB513, we also apply the CSSP potential to the nonidentical sequences in recent CASP9, where results are shown in Figure 4. The results on the CASP9 targets are consistent with those on CB513, where in 61.8% and 79.7% targets, the CSSP potential is able to recognize predicted secondary structures with Q3 accuracy higher than 90% and 80% as best scored structures, respectively.

Figure 5 demonstrates the sensitivity of the CSSP potential on 1cdtA in CB513 by comparing the predicted secondary structures by JPred, SAM, and Porter. JPred, SAM, and Porter have Q3 prediction accuracy of 83.3%, 78.6%, and 95.0%, respectively, on 1cdtA. The predicted secondary structure has high Q3 prediction accuracy, which has the similar structure and CSSP potential value as the native. Potential values of each seven-residue fragment in each prediction are displayed in Figure 5. One can notice that mispredicting a β-strand as an α-helix in SAM results in a large spike in the potential values in fragments 35 to 38, indicating that the α-helix is strongly unfavorable. Similarly, the misprediction of a β-strand in JPred leads to significantly higher potential values in fragments from 5 to 21. As a result, Porter's predicted secondary structure has an overall lower potential value (−5.89) than those of JPred (0.28) and SAM (14.64).

In more than 80% of the targets in CB513, our best CSSP potential based on triplets picks the predicted secondary structures generated by SSPRO4 or PORTER. This is due to the fact that both SSPRO4 and PORTER take advantage of the

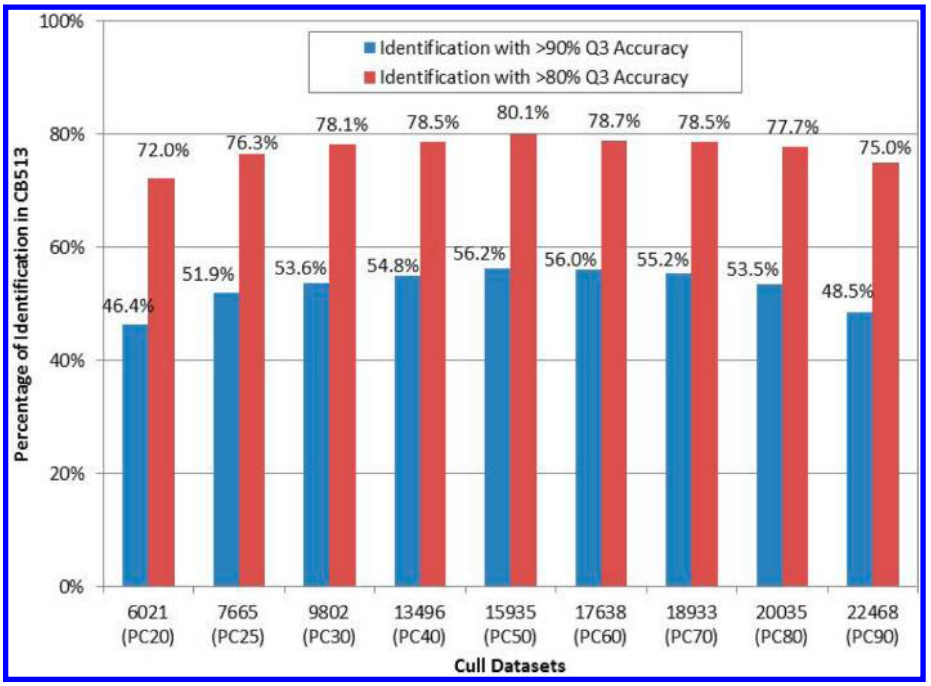


Figure 1. Comparison of identification accuracies of CSSP using different cull data sets with maximum pairwise mutual sequence identity ranging from 20% to 90%. Cull data set with maximum 50% sequence identity yields best identification accuracy.

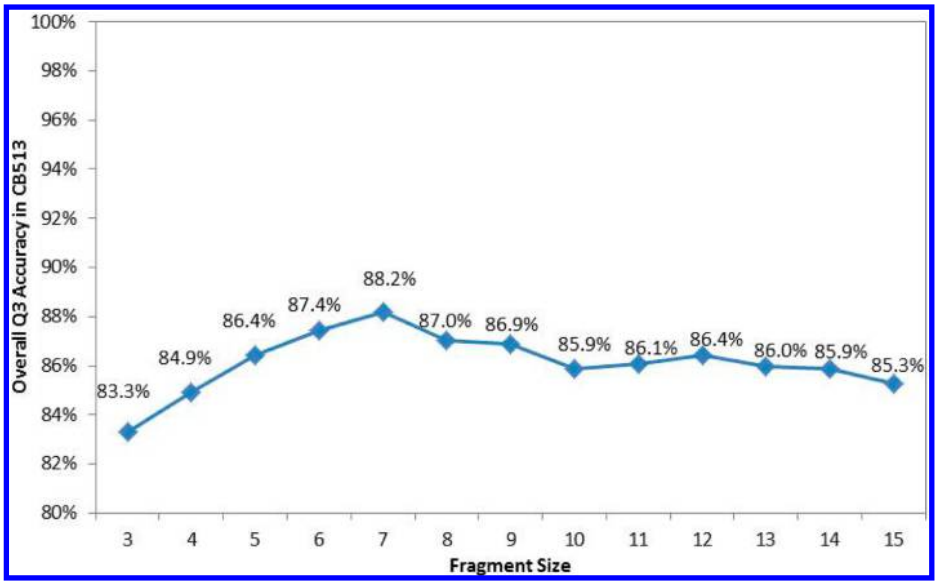


Figure 2. Effect of varying fragment size on the identification accuracy in CSSP using Cull data set with maximum 50% sequence identity. CSSP with fragment size seven has the best performance, yielding 88.2% overall Q3 accuracy in CB513.

structural information of homologues, which is usually helpful to obtain highly accurate prediction. However, when the homologue structures are missing or a wrong homologue template is used, SSPRO4 or PORTER may result in predictions with low accuracy. Figure 6 shows the native secondary structure of 1pyp as well as the predictions of SSPRO4, PORTER, and PSIPRED. Probably due to lack of homologue structural information in PDB, neither SSPRO4 nor PORTER can reach a prediction with more than 80% Q3 accuracy. In this case, CSSP favors the prediction from PSIPRED, which has the lowest potential value (−6.37) and 81% Q3 accuracy.

3.2. CSSP Using Doublets, Triplets, and Quartets. Although theoretically, CSSP can incorporate interactions of *k*-lets for arbitrary *k* value, in practice, the accuracy of *k*-let potential is limited by the number of samples available. Figure 7 compares the identification accuracies of CSSP using doublets, triplets, and quartets on CB513 with fragment size seven and the cull data set with maximum 50% sequence identity. One can find that the identification accuracy of CSSP using triplets is significantly higher than CSSP using doublet by incorporating interactions of three residues. Theoretically, CSSP using quartets should have better precision than the one using triplets since higher order of interactions is taken into account. However, as shown in Figure 7, CSSP using quartets is not as accurate as the one using triplets only. On the basis of the following analysis of sample numbers in doublets, triplets, and quartets, we find that lack of samples in quartets results in significant higher marginal errors in estimating the distribution of secondary structures in quartets than those in

dx.doi.org/10.1021/ci300207x | J. Chem. Inf. Model. 2013, 53, 500−508

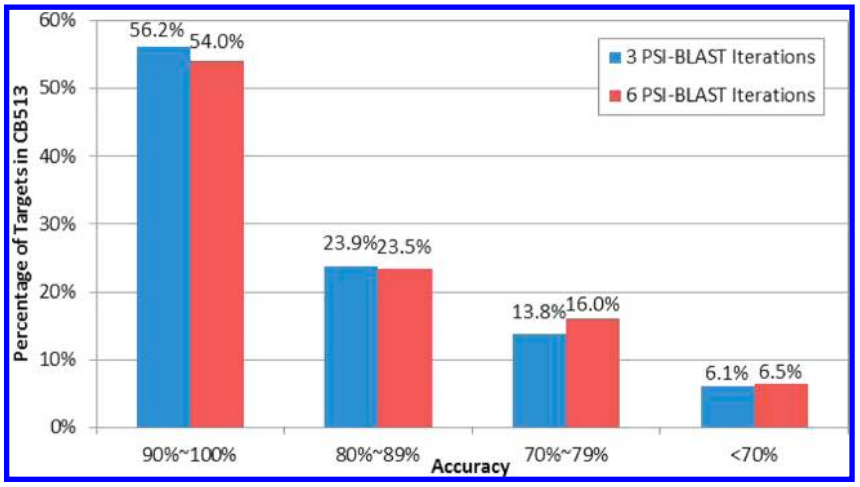


Figure 3. Accuracy comparison of CSSP using weighted frequencies generated from PSSM using three and six PSI-BLAST iterations. CSSP based on PSI-BLAST using three iterations is slightly more sensitive. Cull data set with maximum 50% sequence identity and fragment size of seven is used.

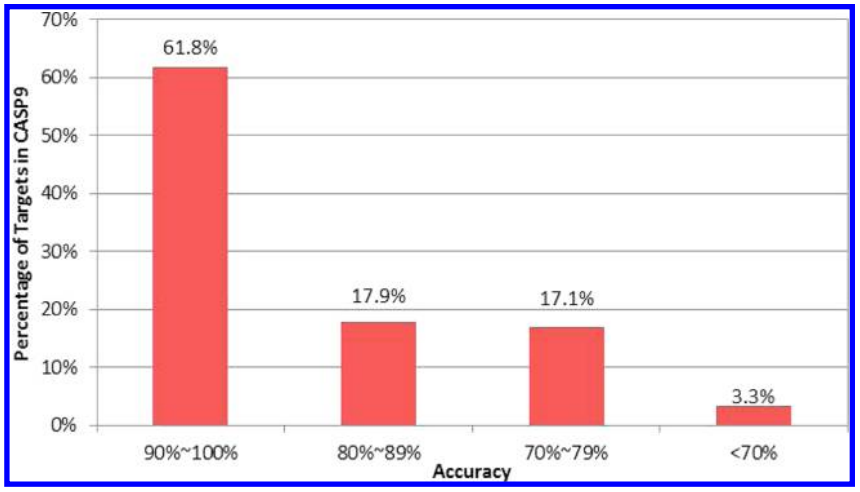


Figure 4. Accuracy of CSSP on nonidentical sequences in CASP9.

triplets. Moreover, CSSP using quartets has almost twice number of terms as CSSP using triplets, which is more prone to suffer from overfitting, particularly when some terms are undersampled.

We use the multinomial distribution to determine the sample size needed to estimate the secondary structure probability of a *k*-let with certain accuracy. Considering statistical samples divided into m mutually exclusive and exhaustive categories and denoting π_i, $i = 1, ..., m$, to be the proportion of the samples in the ith category. The calculation of the sample size n_i for the ith category with precision p_i is

$$n_i = B\pi_i(1 - \pi_i)/(1 - p_i)^2$$

where B is the χ^2 value with $m - 1$ degree of freedom and precision p_i.[29] Let us assume that, in general, the samples are nearly uniformly distributed in the secondary structure categories. The total number of samples needed to estimate the secondary structure distribution of a certain triplet with 99% accuracy (1% marginal error) is

$$n = \frac{45.64\left(\frac{1}{27}\right)\left(1 - \frac{1}{27}\right)}{(1 - 99\%)^2}27 \approx 439496$$

Similarly, the sample size needed to estimate the secondary structure distribution of a certain quartet with 99% accuracy is

$$n = \frac{112.33\left(\frac{1}{81}\right)\left(1 - \frac{1}{81}\right)}{(1 - 99\%)^2}81 \approx 1109432$$

Table 2 displays the most, least, and average number of samples in various doublets, triplets, and quartets at different relative positions when the cull data set with maximum 50% sequence identity is used. Table 2 also shows that 100% and 94.4% of the triplets can achieve 95% and 99% accuracy, respectively. In contrast, only 85.2% of quartets can have 95% accuracy in secondary structure distribution and none of the quartets achieve 99% accuracy. Particularly for the quartets composed of rare amino acids, the estimated secondary structure distribution has low accuracy. For example, the quartet with minimum samples is WWCW at relative position 0_2_3_5, which has only 2172 samples—the accuracy of its secondary structure distribution is approximately 80%. As a result, CSSP using quartet is not as precise and sensitive as the one using triplet only. However, the number of high-resolution, experiment-determined protein structures increases rapidly recently. When the protein data set grows to about 10 times the size of the data set we have now, the average accuracies of secondary

dx.doi.org/10.1021/ci300207x | J. Chem. Inf. Model. 2013, 53, 500−508

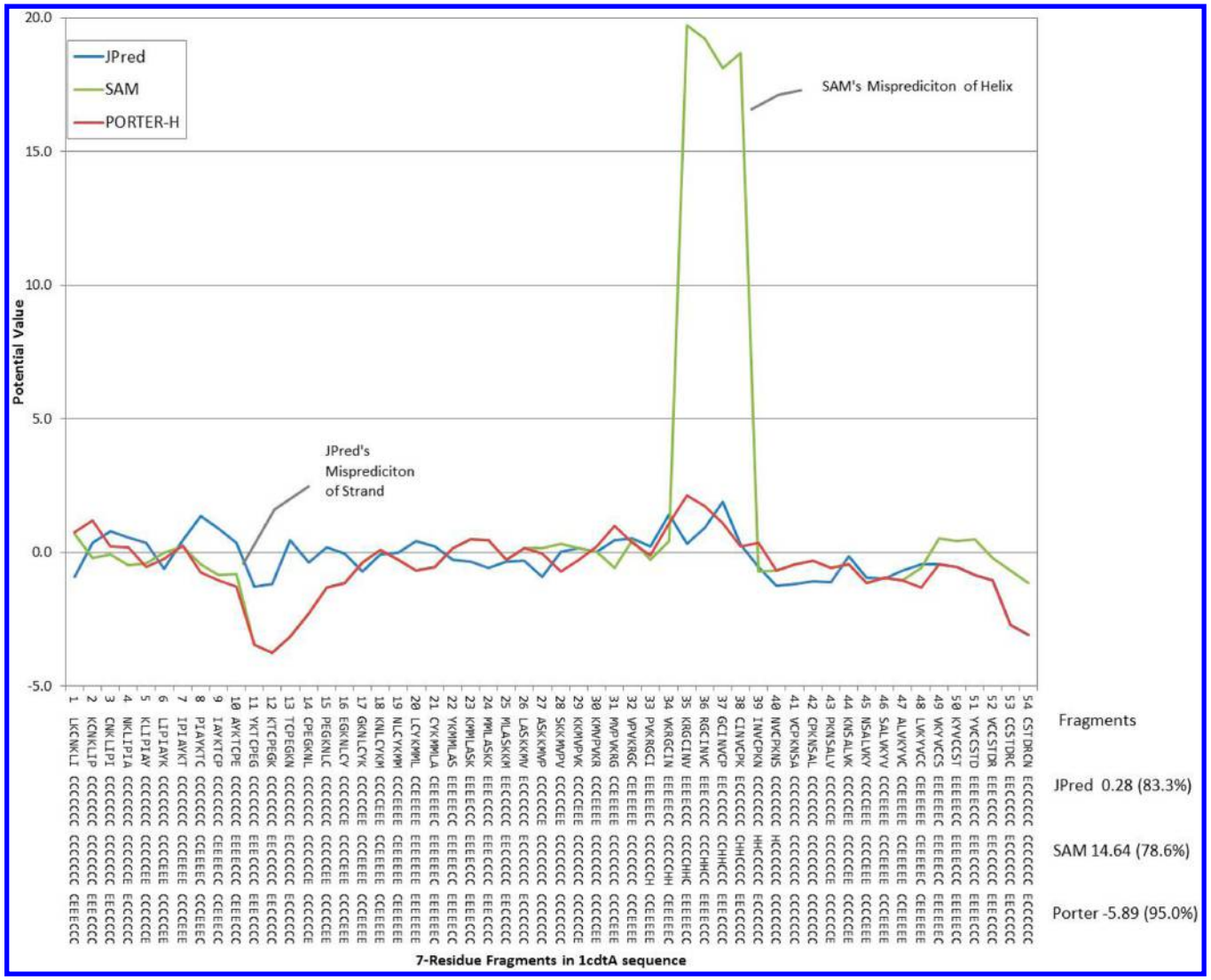


Figure 5. Sensitivity of the knowledge-based potential on 1cdtA. Mispredictions of JPred (fragments 5−21) and SAM (fragments 35−38) lead to higher CSSP potential values than that of Porter's predicted secondary structure.

structure distributions in quartets will reach the current accuracies in triplets and then CSSP using quartets may start to become more effective.

4. DISCUSSION AND SUMMARY

In this paper, we present a context-based secondary structure potential (CSSP) by capturing the high-order inter-residue interactions. The CSSP potential can be effectively used to identify secondary structure predictions with good quality. Moreover, as shown in our computational results and analysis, the CSSP potential using triplets outperforms the CSSP potentials using doublets or quartets. Nevertheless, in the near future when sufficient samples become available, the CSSP potential using quartets may become more effective than the one using triplets.

Although both CSSP and GOR5[16] explicitly consider the high-order inter-residue interactions, the mechanisms of calculating and integrating these interactions are different due to different purposes of CSSP and GOR5. The goal of GOR5 is to predict the secondary structure of each residue. Therefore, the GOR5 scores evaluate how likely a residue adopts a certain secondary structure within its amino acid environment. However, GOR5 is unable to take the influence to a residue from the secondary structures of its neighboring residues into account because they are unknown. In fact, the secondary structures of the neighboring residues play an important role. For example, if the adjacent positions of a residue are not helices, it is impossible for this middle residue to adopt helix as its secondary structure. In contrast, the purpose of CSSP is to assess the qualities of predicted secondary structures, where the favorability of two, three, four, and theoretically up to k residues concurrently adopting certain secondary structures are of interest. Compared to the secondary structure energy by Miyazawa and Jernigan,[14] CSSP considers more general N-body interactions among not necessarily consecutive residues. CSSP is also different from the k-mer model[17] proposed by Madera et al., whose purpose is to refine secondary structure predictions and the k-mer contains the secondary structure information only. In comparison, the k-lets in CSSP measure the high-order correlation between sequence and structure, which include both sequence and structure information.

One of the main disadvantages of the CSSP potential capturing high-order inter-residue interactions is its high computational cost. Considering the CSSP potential with fragment size N and calculating up to k-let inter-residue

dx.doi.org/10.1021/ci300207x | J. Chem. Inf. Model. 2013, 53, 500−508

Figure 6. Sensitivity of the knowledge-based potential 1pyp. SSPRO4 and PORTER have predictions with Q3 accuracies of 75% and 72%, respectively, due to lack of homologue structural information. Our knowledge-based potential favors the prediction by PSIPRED with Q3 accuracy of 81%.

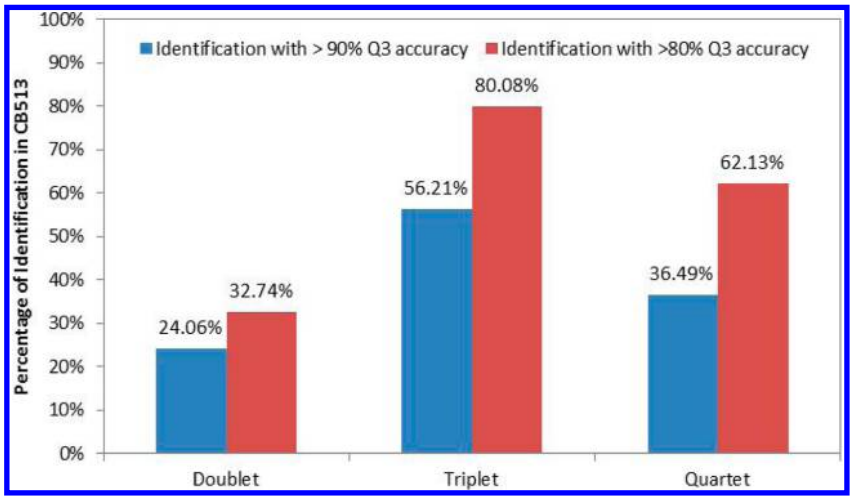


Figure 7. Identification accuracies of CSSP using doublets, triplets, and quartets in CB513.

interactions, the total number of *k*-let calculations for a protein with *P* residues is

$$P/N \sum_{i=1}^{k} \binom{i}{N}$$

In our computational experiments, calculating CSSP potential using triplets for one postulated protein structure takes a few seconds to several minutes on a single processor. CSSP potential using quartets is even more computationally costly. Therefore, we use CSSP to assess predicted secondary structures instead of using CSSP to predict secondary structures. Nevertheless, computing CSSP potential is data-intensive and parallelizable. Taking advantage of the emerging massively parallel computing architectures such as graphics process units (GPU) and data-intensive parallel computing algorithms,[30] one may be able to reduce the computational time of evaluating CSSP significantly and then use CSSP efficiently for secondary structure prediction. Another disadvantage is that the current CSSP is unable to capture global interactions exceeding the fragment size.

Table 2. Most, Least, and Average Numbers of Samples in Doublet, Triplets, and Quartets at Different Relative Positions When Cull Data Set with Maximum 50% Sequence Identity Is Used

	doublets	triplets	quartets
most number of samples	2731381 (AA @ 0_1)	2416997 (AAA @ 0_1_2)	672129 (AAAA @ 0_1_3_4)
least number of samples	270313 (WW @ 0_6)	130213 (WWW @ 0_5_6)	2172 (WWCW @ 0_2_3_5)
average number of samples	1510060	1075624	130565
percentage with 95% accuracy	100%	100%	85.2%
percentage with 99% accuracy	100%	94.4%	0%

AUTHOR INFORMATION

Corresponding Author

*E-mail: yaohang@cs.odu.edu.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

Y.L. acknowledges support from NSF under grant 1066471, ODU 2011 SEECR grant, and ODU 2013 Multidisciplinary Seed grant. I.R. acknowledges support from CNCSIS-UEFISCDI under project number PN-II-PT-PCCA-2011-3.1-1350.

REFERENCES

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

(2) Pirovano, W.; Heringa, J. Protein secondary structure prediction. *Methods Mol. Biol.* **2010**, *609*, 327−348.

(3) Wei, Y.; Thompson, J.; Floudas, C. A. CONCORD: a consensus method for protein secondary structure prediction via mixed integer linear optimization. *Proc. R. Soc. A: Math., Phys. Eng. Sci.* **2012**, *468*, 831−850.

(4) Guermeur, Y. *Combinaison de classifieurs statistiques, application a la prediction de la structure secondaire des proteins*. PhD Thesis, Université Paris 6, 1997.

(5) Karplus, K. SAM-T08, HMM-based protein structure prediction. *Nucleic Acids Res.* **2009**, *37*, W492−W497.

(6) Cole, C.; Barber, J. D.; Barton, G. J. The Jpred 3 secondary structure prediction server. *Nucleic Acids Res.* **2008**, *36*, W197−W201.

(7) Ouali, M.; King, R. D. Cascaded multiple classifiers for secondary structure prediction. *Protein Sci.* **2000**, *9*, 1162−1176.

(8) Rost, B.; Sander, C. Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins* **1994**, *19*, 55−72.

(9) Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **1999**, *292*, 195−202.

(10) Meiler, J.; Baker, D. Coupled prediction of protein secondary and tertiary structure. *Proc. Natl. Acad. Sci.* **2003**, *100*, 12105−12110.

(11) Petersen, B.; Petersen, T.; Andersen, P.; Nielsen, M.; Lundegaard, C. A generic method for assignment of reliability scores applied to solvent accessibility predictions. *BMC Structural Biol.* **2009**, *9*, 51.

(12) Cheng, J.; Randall, A. Z.; Sweredoski, M. J.; Baldi, P. SCRATCH: a protein structure and structural feature prediction server. *Nucleic Acids Res.* **2005**, *33*, W72−W76.

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

(14) Miyazawa, S.; Jernigan, R. L. Evaluation of short-range interactions as secondary structure energies for protein fold and sequence recognition. *Proteins* **1999**, *36*, 347−356.

(15) Garnier, J.; Gibrat, J. F.; Robson, B. GOR method for predicting protein secondary structure from amino acid sequence. *Methods Enzymol.* **1996**, *266*, 540−553.

(16) Kloczkowski, A.; Ting, K. L.; Jernigan, R. L.; Garnier, J. Combining the GOR V algorithm with evolutionary information for protein secondary structure prediction from amino acid sequence. *Proteins* **2002**, *49*, 154−166.

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

(18) Sippl, M. J. Calculation of conformational ensembles from potentials of mena force: an approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol.* **1990**, *213*, 859−883.

(19) Wang, G.; Dunbrack, R. L. PISCES: a protein sequence culling server. *Bioinformatics* **2003**, *19*, 1589−1591.

(20) Cuff, J. A.; Barton, G. J. Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *Proteins* **1999**, *34*, 508−519.

(21) Kinch, L. N.; Shi, S.; Cheng, H.; Cong, Q.; Pei, J.; Mariani, V.; Schwede, T.; Grishin, N. V. CASP9 target classification. *Proteins* **2011**, *79*, 21−36.

(22) Samudrala, R.; Moult, J. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *J. Mol. Biol.* **1998**, *275*, 895−916.

(23) 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.

(24) Mooney, C.; Pollastri, G. Beyond the Twilight Zone: Automated prediction of structural properties of proteins by recursive neural networks and remote homology information. *Proteins* **2009**, *77*, 181−190.

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

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

(27) Kabsch, W.; Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen bonded and geometrical features. *Biopolymers* **1983**, *22*, 2577−2637.

(28) Frishman, D.; Argos, P. Knowledge-based protein secondary structure assignment. *Proteins* **1995**, *23*, 566−579.

(29) Tortora, R. D. A note on sample size estimation for multinomial populations. *Am. Stat.* **1978**, *32*, 100−102.

(30) Yaseen, A.; Li, Y. Accelerating knowledge-based energy evaluation in protein structure modeling with Graphics Processing Units. *J. Parallel Distributed Comput.* **2012**, *72*, 297−307.