


2014

# Template-Based C8-Scorpion: A Protein 8 State Secondary Structure Prediction Method Using Structural Information and Context-Based Features

Ashraf Yaseen

Yaohang Li  
*Old Dominion University*

Follow this and additional works at: [https://digitalcommons.odu.edu/computerscience\\_fac\\_pubs](https://digitalcommons.odu.edu/computerscience_fac_pubs)

 Part of the [Biochemistry Commons](#), and the [Computer Sciences Commons](#)

---

## Repository Citation

Yaseen, Ashraf and Li, Yaohang, "Template-Based C8-Scorpion: A Protein 8 State Secondary Structure Prediction Method Using Structural Information and Context-Based Features" (2014). *Computer Science Faculty Publications*. 55.  
[https://digitalcommons.odu.edu/computerscience\\_fac\\_pubs/55](https://digitalcommons.odu.edu/computerscience_fac_pubs/55)

## Original Publication Citation

Yaseen, A., & Li, Y.H. (2014). Template-based c8-scorpion: A protein 8-state secondary structure prediction method using structural information and context-based features. *BMC Bioinformatics*, 15. doi: 10.1186/1471-2105-15-s8-s3

RESEARCH

Open Access

# Template-based C8-SCORPION: a protein 8-state secondary structure prediction method using structural information and context-based features

Ashraf Yaseen, Yaohang Li\*

From Third IEEE International Conference on Computational Advances in Bio and Medical Sciences (ICCABS 2013)

New Orleans, LA, USA. 12-14 June 2013

## Abstract

**Background:** Secondary structures prediction of proteins is important to many protein structure modeling applications. Correct prediction of secondary structures can significantly reduce the degrees of freedom in protein tertiary structure modeling and therefore reduces the difficulty of obtaining high resolution 3D models.

**Methods:** In this work, we investigate a template-based approach to enhance 8-state secondary structure prediction accuracy. We construct structural templates from known protein structures with certain sequence similarity. The structural templates are then incorporated as features with sequence and evolutionary information to train two-stage neural networks. In case of structural templates absence, heuristic structural information is incorporated instead.

**Results:** After applying the template-based 8-state secondary structure prediction method, the 7-fold cross-validated Q8 accuracy is 78.85%. Even templates from structures with only 20%~30% sequence similarity can help improve the 8-state prediction accuracy. More importantly, when good templates are available, the prediction accuracy of less frequent secondary structures, such as 3-10 helices, turns, and bends, are highly improved, which are useful for practical applications.

**Conclusions:** Our computational results show that the templates containing structural information are effective features to enhance 8-state secondary structure predictions. Our prediction algorithm is implemented on a web server named "C8-SCORPION" available at: <http://hpcr.cs.odu.edu/c8scorpion>.

## Background

An important intermediate step in modeling the three-dimensional structure of a protein is to accurately predict its secondary structures [1]. Most often, the secondary structures are classified into three general states, i.e., helices (H), strands (E), and coils (C). Correspondingly, success of secondary structure prediction is typically measured by the Q3 (3-state) accuracy. Many machine learning methods, including statistics analysis, neural networks, hidden Markov chain, support vector machines,

have been developed to predict secondary structures. Correspondingly, there are many secondary structure prediction servers available, including GOR4 [2], PSI-Pred [3], PHD [4], SAM [5], Porter [6], JPred [7], SPINE [8], SSSPRO [9], NETSURF [10], and many others. The modern secondary structure prediction servers can generate prediction results with close to 80% Q3 accuracy.

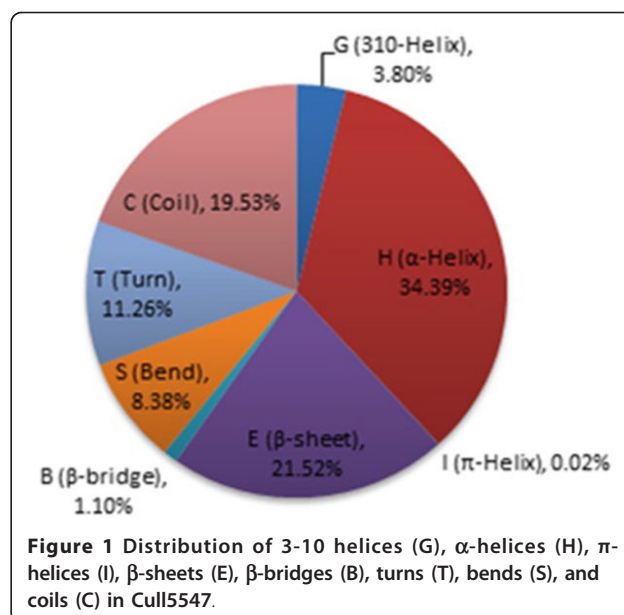
Compared to the general three secondary structure states, the DSSP program [11] has more detailed classifications by assigning secondary structures to eight states, including 3-10 helix (G),  $\alpha$ -helix (H),  $\pi$ -helix (I),  $\beta$ -strand (E), bridge (B), turn (T), bend (S), and others (C). The 8-state secondary structures convey more precise structural information than 3-state, which is particularly

\* Correspondence: [yaohang@cs.odu.edu](mailto:yaohang@cs.odu.edu)  
Department of Computer Science, Old Dominion University, Norfolk, VA 23529, USA

important for a variety of applications. For example, accurate 8-state secondary structures predictions can restrict the variations of backbone dihedral angles within a small range according to the Ramachandran plots [12] and thus reduce the search space in template-free protein tertiary structure modeling. Moreover, differentiations among 3-10 helix,  $\alpha$ -helix, and  $\pi$ -helix in secondary structure prediction aid to assign residues and fit protein structure models in cryo-electron microscopy density maps [13]. Unfortunately, most of the secondary structure prediction software packages or servers only provide 3-state predictions.

To the best of our knowledge, very few methods have been developed for the 8-state secondary structure prediction. Pollastri et al. [9] extended their 3-state prediction method to SSpro8 for 8-state secondary structure prediction. The reported Q8 accuracy of SSpro8 is 62-63% [9]. A more recent prediction method of the 8-state, RaptorXss8, developed by Wang et al [14], has reported 67.9% Q8 accuracy through the use of conditional neural field (CNF) models. Table 1 shows the prediction accuracy of RaptorXss8 on several popularly used secondary structure prediction benchmarks, including CB513, CASP9, Manesh215, and Carugo338. Although nearly 70% Q8 accuracy is achieved, the prediction accuracies of different states vary significantly. In particular, the prediction accuracy of G, I, B, and S are very low, mainly due to the fact of their relatively infrequent appearance in protein data banks (PDB), whose distribution is shown in Figure 1. The low prediction accuracies in these states limit the application of 8-state secondary structure prediction in practice.

Most current secondary structure prediction methods do not rely on similarity to known protein structures; in other words, these methods are *de novo*, where the secondary structure prediction is based on sequence information only. However, we cannot neglect the fact that many protein sequences have some degree of similarity among themselves. Actually, over half of all known protein sequences have some detectable similarity (higher than 25%) to one or more sequences of known structures [15,16]. Around 75% was reported as the percentage of those newly deposited protein structures in the PDB database showing significant similarity to previous deposited structures. Consequently, taking advantage of structural similarity of proteins with sequence similarity may lead to



significant improvement of protein structure prediction. In fact, the latest version of porter [6] has used homology-based templates for 3-state secondary structure prediction [16]. Porter has been reported to achieve prediction accuracy improvement when known structures with >30% sequence similarity are available and even reach theoretical upper bound of secondary structure prediction when such sequence similarity is higher than 50%.

In this paper, we investigate the template-based method for 8-state secondary structure prediction. We extract structural information from known structures of chains with certain sequence similarity to build structural templates. Then, the structural information contained in the templates is incorporated (as features) together with sequence and evolutionary information for neural network training and validation.

In the case where structural information from the structural template is not available for a residue, context-based scores estimating the favorability of that residue adopting a secondary structure conformation in the presence of its neighbors in sequence are used instead. The fundamental idea of the context-based scores is based on the fact that the formation of secondary structure exhibit strong local dependency, particularly, residues in a protein sequence

**Table 1 Prediction Accuracy of RaptorXss8 on Benchmarks of CB513, CASP9, Manesh215, and Carugo338.**

	Q <sub>G</sub>	Q <sub>H</sub>	Q <sub>I</sub>	Q <sub>E</sub>	Q <sub>B</sub>	Q <sub>S</sub>	Q <sub>T</sub>	Q <sub>C</sub>	Q <sub>B</sub>
CB513	17.54	89.96	0.00	77.68	0.09	15.87	48.02	63.29	65.59
CASP9	20.58	92.90	0.00	81.64	0.00	18.11	51.45	59.37	69.31
Manesh215	18.43	90.22	0.00	79.60	0.32	17.80	51.28	63.73	67.69
Carugo338	19.20	89.91	0.00	79.45	0.44	17.14	50.11	63.36	66.64

Prediction accuracies for 3-10 helices (G),  $\pi$ -helices (I),  $\beta$ -bridges (B), and bends (T) are particularly low due to their low appearance frequencies.

are strongly correlated in different sequence positions in coils,  $\beta$ -sheets, 310 helices,  $\alpha$ -helices, and  $\pi$ -helices. We extract statistics to derive context-based scores from a large training data set. These context-based scores are then incorporated as sequence-structure features together with sequence, template, and evolutionary information in neural network training process for 8-state secondary structure prediction.

We test our template-based 8-state prediction method on several popularly used benchmarks including CB513, Manesh215, and Carugo338 as well as the CASP9 targets. The prediction accuracies for the eight states are analyzed.

## Methods

### The protein data sets

We use the protein chain dataset Cull5547 generated by the PISCES server [17] on 10/21/2011 for neural network training and Cull16633 for context-based scores generation. Cull5547 contains 5,547 protein chains with at most 25% sequence identity and 2.0A resolution cutoff, and Cull16633 contains 16,633 protein chains with at most 50% sequence identity and 3.0A resolution cutoff. We eliminate very short chains, whose lengths are less than 40 residues, since the PSI-BLAST program [18] is usually unable to generate profiles for very short sequences, and very large chains whose lengths are greater than 1,000 residues. We also eliminate residue samples with undetermined secondary structures.

Public benchmarks, including CB513 [19], Manesh215 [20], Carugo338 [21], and the recent CASP9 targets [22], that are popularly employed as benchmarks for 3-state secondary structure predictions, are used to benchmark our method in 8-state predictions.

### Template construction

Figure 2 illustrates the procedure of constructing structural templates. First of all, for a given protein sequence target, PSI-BLAST is used to search against the NR (Non-Redundant) database with E-value = 0.001 and at most 3 iterations to generate the PSSM (Position Specific Scoring Matrix) data. Then, the PSSM is used to search against the Protein Data Bank (PDB [23]) for alignments with E-value = 10.0. If known structures are available in PDB, their 8-state assignments are determined by the DSSP program and then a structural template is built for the correspondent residue positions. Among the list of templates constructed, we select the top one that is less than 95% sequence similarity, according to PSI-BLAST ranking.

### Encoding

We use a window size of 15 residues for input encodings. Each residue is represented with 20 values from the PSSM (Position-Specific Scoring Matrix) data, 1 extra input to indicate if the residue window overlaps C- or N-terminal, 1 value for degree of similarity, and 8 values for structural information from template or context-based secondary structure scores [24]. Hence, a total number of 450 values are used to describe each residue

Figure 3 shows an example of encoding residues in a protein sequence. For a residue with available structural information in the template, the corresponding secondary structure state is set to 1 while the other states are set to 0. At the same time, the degree of similarity is set for the sequence similarity. On the other hand, if the structural information for a residue is not available in the template, the degree of similarity is set to zero and the context-based scores are incorporated instead. The context-based

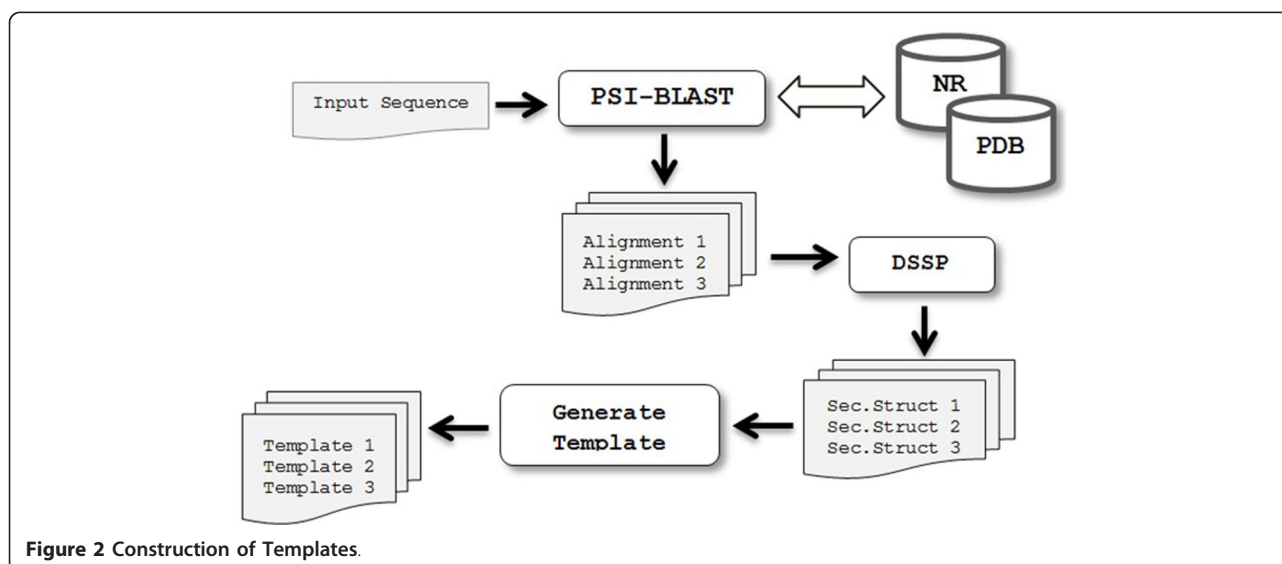
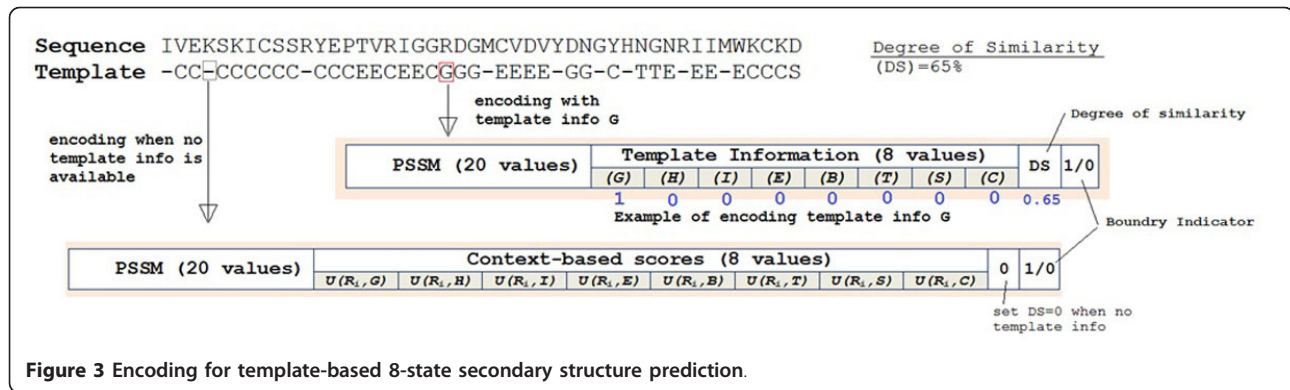


Figure 2 Construction of Templates.



scores are statistics-based pseudo-potentials to specify the favorability of a residue adopting a certain secondary structure in its amino acid context [24].

### Context-based scores

The types and conformations of nearby residues play a critical role in secondary structure conformation that a residue may adopt [24]. In particular, 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,  $\alpha$ -helices, and  $\pi$ -helices, respectively. Residues in contacting parallel or anti-parallel  $\beta$ -sheets are connected by hydrogen bonds in alternative positions. Moreover, the formation of interactions within coils beyond nearest neighbors appears not to contribute with statistical significance in determining coil structure [27]. Hence, correlations among residues provide significant information in predicting secondary structure.

In this method, we will extract statistics of singlets ( $R_i$ ), doublets ( $R_i R_{i+k}$ ), and triplets ( $R_i R_{i+k_1} R_{i+k_2}$ ) residues at different relative positions from protein sequences in Cull16633 dataset. These statistics represent estimations of the probabilities of residues adopting a specific structural state when none, one, or two of their neighbors in context are taken into consideration, respectively.

The observed probabilities of the  $i^{\text{th}}$  residue  $R_i$  in a singlet ( $R_i$ ), doublet ( $R_i R_{i+k}$ ), and triplet ( $R_i R_{i+k_1} R_{i+k_2}$ ) adopting a specific structural state  $C_i$  are respectively estimated by

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

$$P_{obs}(C_i | R_i R_{i+k}) = \frac{N_{obs}(C_i, R_i R_{i+k})}{N_{obs}(R_i R_{i+k})}, \text{ and}$$

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

Here  $N_{obs}(C_i, R_i)$ ,  $N_{obs}(C_i, R_i R_{i+k})$ , and  $N_{obs}(C_i, R_i R_{i+k_1} R_{i+k_2})$  are the weighted observed number of singlet ( $R_i$ ), doublet ( $R_i R_{i+k}$ ), and triplet ( $R_i R_{i+k_1} R_{i+k_2}$ ) with  $R_i$  adopting conformation  $C_i$  in the protein structure database.  $N_{obs}(R_i)$ ,  $N_{obs}(R_i R_{i+k})$ , and  $N_{obs}(R_i R_{i+k_1} R_{i+k_2})$  are the weighted observed number of singlets, doublets, and triplets. The observed numbers will be calculated as

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

$$N_{obs}(R_i R_{i+k}) = \sum_{Protein} \sum_j PSSM_j(R_i) * PSSM_j(R_{i+k}),$$

$$N_{obs}(R_i R_{i+k_1} R_{i+k_2}) = \sum_{Protein} \sum_j PSSM_j(R_i) * PSSM_j(R_{i+k_1}) * PSSM_j(R_{i+k_2}),$$

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

$$N_{obs}(C_i, R_i R_{i+k}) = \sum_{Protein} \sum_{j=C_i} PSSM_j(R_i) * PSSM_j(R_{i+k}), \text{ and}$$

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

where  $PSSM_j(R_i)$  is the PSSM frequency for residue type  $R_i$  at the  $j^{\text{th}}$  position of a protein sequence.

Correspondently, the context-dependent pseudo-potentials are generated using the derived statistics of correlations between each residue and its nearby neighbors based on Sippl's potentials of mean force method [25]. 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 structural state  $C_i$ ,

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



Here  $R$  is gas constant,  $T$  is temperature, and  $P_{ref}(C_i|R_i)$  is the referenced probability. In our method, we will employ the conditional probability approach described in [28] to estimate the referenced probability by

$$P_{ref}(C_i|R_i) = \sum_j^{C_j=C_i} N_{obs}(C_j, R_j) / \sum_j N_{obs}(R_j).$$

Similarly, the mean-force potentials  $U_{doublet}(C_i, R_i R_{i+k})$  and  $U_{triplet}(C_i, R_i R_{i+k_1} R_{i+k_2})$  for residue adopting structural state are

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

and

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

with the corresponding referenced probabilities,

$$P_{ref}(C_i|R_i R_{i+k}) = \sum_j^{C_j=C_i, R_{j+k}=R_{i+k}} N_{obs}(C_j, R_j R_{j+k}) / \sum_j N_{obs}(R_j R_{j+k}),$$

and

$$P_{ref}(C_i|R_i R_{i+k_1} R_{i+k_2}) = \sum_j^{C_j=C_i, R_{j+k_1}=R_{i+k_1}, R_{j+k_2}=R_{i+k_2}} N_{obs}(C_j, R_j R_{j+k_1} R_{j+k_2}) / \sum_j N_{obs}(R_j R_{j+k_1} R_{j+k_2}),$$

respectively.

Then, the context-dependent pseudo-potential for  $R_i$  will be

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

These pseudo-potentials are incorporated as context-based scores representing sequence-structure features in neural network training when structural information from templates is not available.

### Neural network model

We incorporate two phases of standard feed-forward neural network training for the 8-state secondary structure prediction. The first phase is the primary sequence-structure prediction and the second phase is the structure-structure refinement. The numbers of hidden nodes in the first and second networks are 225 and 68, respectively. Figure 4 shows the encoding diagram and the two-phase neural network architecture. Each neural network is trained to predict the secondary structure state of a residue in the middle of the residue window.

### Performance measures

The prediction accuracy is calculated as the average of the seven prediction scores. We use both Q8 and SOV8 (Segment overlap [26]) scores to measure the qualities of our 8-state secondary structure predictions.

### N-fold cross validation

To obtain a reliable estimate of the 8-state secondary structure prediction accuracy, we use 7-fold cross

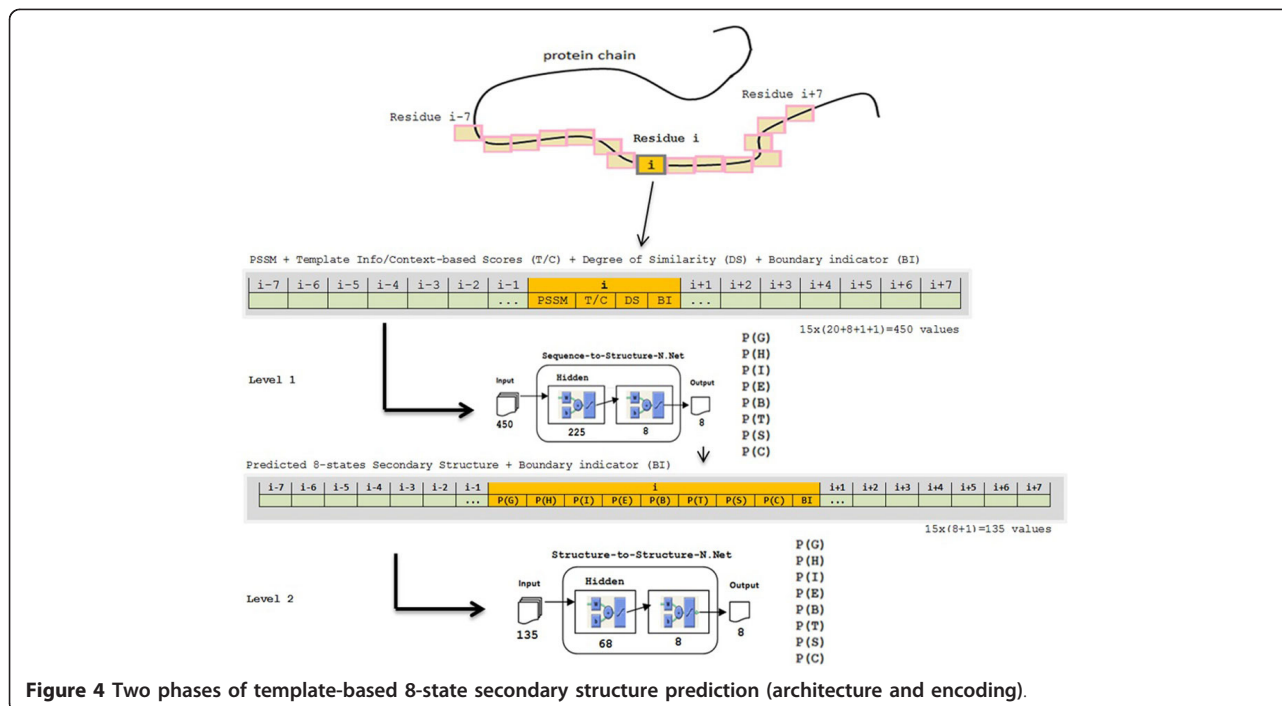


Figure 4 Two phases of template-based 8-state secondary structure prediction (architecture and encoding).

validation on Cull5547. We randomly divide the chains in Cull5547 into 7 subsets with approximately the same size, such that five subsets are used for training, one for testing, and one for validation.

## Results

Upon the selection of the best alignment with similarity less than 95% for all protein chains in the Cull5547 dataset, the final Q8 seven-fold cross validated accuracy after applying the template-based 8-state prediction reaches 78.85%. Table 2 lists the Q8 and SOV8 accuracies of 7-fold cross validation for each state.

Table 3 compares the Q8 and SOV8 accuracy of using predictions with and without templates on benchmarks of CB513, CASP9, Manesh215, and Carugo338. Clearly, when homology structural information is available, the 8-state prediction accuracy is significantly improved. It is also interesting to find that when structural templates are used, the 8-state prediction accuracy improvement in CASP9 is much less than the other benchmark sets. This is due to the fact that in the CASP9 experiment, targets are deliberately selected to have relatively low similarity to sequences with existing structures in PDB.

Figure 5 shows the distribution of the prediction accuracy as a function of sequence similarity in levels in CB513, CASP9, Manesh215, Carugo338 as well as Cull5574 in cross-validation. Without surprise, the better templates with higher sequence similarity level, the more accurate the prediction results are. More importantly, even templates with only 20%~30% sequence similarity can improve the prediction accuracy by near 5% in various benchmark sets compared to predicted results without templates.

Figure 6 uses the A chain of protein 1BTN as an example to demonstrate the effectiveness of template-based 8-state secondary structure prediction. Prediction without template has 73.6% Q8 accuracy. The best template found in PDB has 61% sequence similarity. Under the guidance of the structural template, the mispredicted helix segment and bend segment in template-less prediction (highlighted in Figure 6) are corrected, which leads to overall 89.6% Q8 accuracy.

## Discussion

As shown in Table 1, the prediction accuracies for different states vary largely due to the very unbalanced appearing frequencies of the eight states in protein structures. In this paper, we are particularly interested in

**Table 3 Comparison between 8-state predictions with and without template on CB513, CASP9, Manesh215, and Carugo338.**

	Q <sub>8</sub>		SOV <sub>8</sub>	
	No-Template	With-Template	No-Template	With-Template
CB513	67.22	79.39	67.66	80.64
CASP9	71.54	76.36	73.47	78.15
Manesh215	69.71	81.10	70.79	82.99
Carugo338	68.44	80.39	69.50	81.95

the effectiveness of structural templates in improving the prediction accuracies of those states with low accuracy in prediction without templates. From Cull5547, we create five subsets of chains that have structural templates with similarity level in intervals of (0%, 10%), (10%, 20%), (20%, 40%), (40%, 70%), and (70%, 95%), respectively. Then, 7-fold neural network trainings are carried out for each subset and the average cross validation prediction accuracy for each state is reported in Table 4.

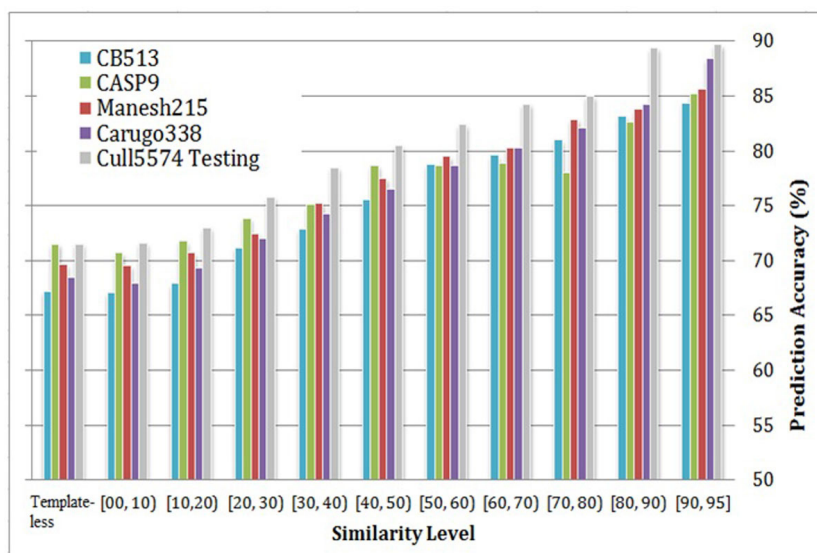
For  $\alpha$ -helices (H), the prediction accuracy using templates with very low sequence similarity (0%, 10%) is already rather high (92.05%), mainly because there are sufficient number of  $\alpha$ -helix samples available and the formation of  $\alpha$ -helix is mainly result from local interactions. Anyway, the structural templates help refine the  $\alpha$ -helix predictions with slight accuracy improvements. When structural templates with 40% or better similarity are available, the prediction accuracy of  $\beta$ -sheets (E) is also improved to above 90%, reaching the theoretical upper bound in secondary structure prediction. 40%+ similarity templates also significantly improve the accuracies of 3-10 helices (G) and bends (S) from 20%+ to 50%+. Similar but not as significant improvements are found in turns (T) and coils (C). However, the prediction results for bridges (B) and  $\pi$ -helices (I) are disappointing. Only when templates with very high similarity (>70%) are available, we can obtain 44% prediction accuracy in bridges (B). The prediction accuracy for  $\pi$ -helices (I) is still 0%. This is mainly due to the facts that  $\pi$ -helices are extremely rare (0.02%) and  $\pi$ -helices (I) are often misclassified into  $\alpha$ -helices (H).

## Conclusions

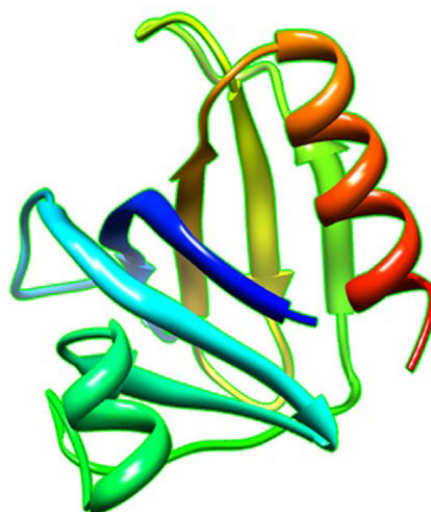
We describe a template-based approach to enhance 8-state secondary structure prediction accuracy in this paper. Our computational results show that the

**Table 2 7-fold cross-validation accuracy in template-based 8-state prediction.**

	G	H	I	E	B	S	T	C	Overall
Q <sub>8</sub>	43.99	92.48	0.00	88.30	27.86	43.46	64.18	75.51	78.85
SOV <sub>8</sub>	47.96	95.19	0.00	92.77	27.57	45.32	66.64	71.45	80.10



**Figure 5** Distribution of 8-state secondary structure prediction accuracy (Q8) as a function of sequence similarity- the first group of bars corresponds to template-less predictions.



MEGFLNRKHEWEAHNKKASSRSWNNVYCVINNQEMGFYKDAKSAASGIPYHSEVPVSLKEAICEVALDYKKKKHVFKLRLSDGNEYLFQAKDDEEMNTWIQAISSA	Sequence
CEEEEEEEEEECSTTCBCSCCCCEEEEEEEEEETEEEEESHHHHTCCSSCCCECTTCEEEECSSCCSSSEEEEECTTSCEEEECSHHHHHHHHHHHHHC	DSSP
EEE-E-EE-E-EE--EEC--CCC-EEEE-----EESSH-H--T-BSS-CCCEC--C---EC-TC-S-SSEEEE-C-SS-EEEECS--HH--H-----H	Template
CEEEEEEEEEETTSCCCSCCEEEEEEEEEETEEEEECCTCCCCSCCECTTEEEECTTCCCTTEEEEEETTSCEEEEECSHHHHHHHHHHHHHC	Template-less pred.
CEEEEEEEEEETTEECTSCCEEEEEEEEEETEEEEESHHHHTCBSSCCCECTTCEEEECTTCCSSSEEEEECTTSCEEEECSHHHHHHHHHHHHHC	Template-based pred.

**Figure 6** Comparison between template-less and template-based predictions on 1BTN chain A.

secondary structure templates, even obtained from sequence with only 20%~30% sequence similarity, can help improve the 8-state prediction accuracy. Overall, 78.85% Q8 accuracy and 80.10% SOV8 accuracy are achieved in 7-fold cross validation. The effectiveness of using structural information in templates has been

demonstrated on popular benchmarks including CB513, CASP9, Manesh215, and Carugo338. More importantly, when good templates are available, the prediction accuracy of less frequent secondary structure states, such as 3-10 helices, turns, and bends, are highly improved, which are suitable for practical use in applications.



**Table 4 Comparison of 7-fold cross validation prediction accuracies in eight states when templates with different sequence similarities are used.**

	(0, 10]	(10, 20]	(20, 40]	(40, 70]	(70, 95]
# of chains	4,426	4,215	3,204	1,437	1,133
Q <sub>H</sub>	92.05	92.70	93.60	94.97	95.94
Q <sub>G</sub>	22.07	23.93	35.09	55.03	69.44
Q <sub>I</sub>	0.00	0.00	0.00	0.00	0.00
Q <sub>E</sub>	83.37	84.53	86.59	90.16	93.61
Q <sub>B</sub>	1.53	3.59	7.24	22.30	44.26
Q <sub>T</sub>	53.35	55.34	60.89	69.66	77.06
Q <sub>S</sub>	22.83	26.41	35.19	54.09	73.40
Q <sub>C</sub>	66.55	67.84	71.81	79.56	86.80
Q <sub>8</sub>	71.33	73.01	76.29	82.11	88.01

A webserver (C8-Scorpion) implementing 8-state secondary structure prediction is currently available at <http://hpcr.cs.odu.edu/c8scorpion>.

#### Competing interests

The authors declare that they have no competing interests.

#### Authors' contributions

YL conceived the context-based scoring method. AY implemented the method and carried out the computation. AY and YL performed the result analysis. Both authors read and approved the final manuscript.

#### Declarations

Publication charges for this work were funded by NSF grant 1066471 to YL. This article has been published as part of *BMC Bioinformatics* Volume 15 Supplement 8, 2014: Selected articles from the Third IEEE International Conference on Computational Advances in Bio and Medical Sciences (ICCABS 2013): Bioinformatics. The full contents of the supplement are available online at <http://www.biomedcentral.com/bmcbioinformatics/supplements/15/S8>.

Published: 14 July 2014

#### References

1. Rost B: Review:Protein secondary structure prediction continues to rise. *J Struct Biol* 2001, **134**(2-3):204-218.
2. Garnier J, Gibrat JF, Robson B: GOR method for predicting protein secondary structure from amino acid sequence. *Methods Enzymol* 1996, **266**:540-553.
3. Jones DT: Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* 1999, **292**(2):195-202.
4. Rost B, Sander C: Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins* 1994, **19**(1):55-72.
5. Karplus K, Barrett C, Cline M, Diekhans M, Grate L, Hughey R: Predicting protein structure using only sequence information. *Proteins-Structure Function and Genetics* 1999, **Suppl** 1:121-125.
6. Pollastri G, McLysaght A: Porter: a new, accurate server for protein secondary structure prediction. *Bioinformatics* 2005, **21**(8):1719-1720.
7. Cole C, Barber JD, Barton GJ: The Jpred 3 secondary structure prediction server. *Nucleic Acids Res* 2008, **36**:W197-W201.
8. Dor O, Zhou YQ: Achieving 80% ten-fold cross-validated accuracy for secondary structure prediction by large-scale training. *Proteins* 2007, **66**(4):838-845.
9. 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-Structure Function and Genetics* 2002, **47**(2):228-235.

10. Petersen B, Petersen TN, 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).
11. Kabsch W, Sander C: Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 1983, **22**(12):2577-2637.
12. Ramachandran GN, Sasisekharan V: Conformation of polypeptides and proteins. *Advances in protein chemistry* 1968, **23**:283-438.
13. Topf M, Baker ML, Marti-Renom MA, Chiu W, Sali A: Refinement of protein structures by iterative comparative modeling and CryoEM density fitting. *J Mol Biol* 2006, **357**(5):1655-1668.
14. Wang ZY, Zhao F, Peng J, Xu JB: Protein 8-class secondary structure prediction using conditional neural fields. *Proteomics* 2011, **11**(19):3786-3792.
15. Montgomerie S, Sundararaj S, Gallin WJ, Wishart DS: Improving the accuracy of protein secondary structure prediction using structural alignment. *Bmc Bioinformatics* 2006, **7**.
16. Pollastri G, Martin AJM, Mooney C, Vullo A: Accurate prediction of protein secondary structure and solvent accessibility by consensus combiners of sequence and structure information. *BMC Bioinformatics* 2007, **8**.
17. Wang GL, Dunbrack RL: PISCES:a protein sequence culling server. *Bioinformatics* 2003, **19**(12):1589-1591.
18. Altschul SF, Madden TL, Schaffer AA, Zhang Z, Miller W, Lipman DJ: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997, **25**(17):3389-3402.
19. Cuff JA, Barton GJ: Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins-Structure Function and Genetics* 2000, **40**(3):502-511.
20. Ahmad S, Gromiha MM, Sarai A: Real value prediction of solvent accessibility from amino acid sequence. *Proteins-Structure Function and Genetics* 2003, **50**(4):629-635.
21. Carugo O: Predicting residue solvent accessibility from protein sequence by considering the sequence environment. *Protein Engineering* 2000, **13**(9):607-609.
22. Kinch LN, Shi S, Cheng H, Cong Q, Pei JM, Mariani V, Schwede T, Grishin NV: CASP9 target classification. *Proteins* 2011, **79**:21-36.
23. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: The Protein Data Bank. *Nucleic Acids Res* 2000, **28**(1):235-242.
24. 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. *Journal of chemical information and modeling* 2013, **53**(2):500-508.
25. Sippl MJ: 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**(4):859-883.
26. Zemla A, Venclovas C, Fidelis K, Rost B: A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment. *Proteins-Structure Function and Genetics* 1999, **34**(2):220-223.
27. Rata I, Li Y, Jakobsson E: Backbone Statistical Potential from Local Sequence-Structure Interactions in Protein Loops. *Journal of Physical Chemistry B* 2010, **114**(5):1859-1869.
28. Samudrala R, Moulton J: An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *Journal of Molecular Biology* 1998, **275**:895-916.

doi:10.1186/1471-2105-15-S8-S3

**Cite this article as:** Yaseen and Li: Template-based C8-SCORPION: a protein 8-state secondary structure prediction method using structural information and context-based features. *BMC Bioinformatics* 2014 **15**(Suppl 8):S3.