W298–W302 *Nucleic Acids Research, 2012, Vol. 40, Web Server issue* *Published online 2 May 2012*
doi:10.1093/nar/gks361

DSP: a protein shape string and its profile prediction server

Jiangming Sun[1], Shengnan Tang[1], Wenwei Xiong[1,2], Peisheng Cong[1] and Tonghua Li[1,*]

[1]Department of Chemistry, Tongji University, 1239 Siping Road, Shanghai 200092, China and [2]Department of Biology and Molecular Biology, Montclair State University, Montclair, NJ 07043, USA

Received January 30, 2012; Revised March 31, 2012; Accepted April 10, 2012

ABSTRACT

Many studies have demonstrated that shape string is an extremely important structure representation, since it is more complete than the classical secondary structure. The shape string provides detailed information also in the regions denoted random coil. But few services are provided for systematic analysis of protein shape string. To fill this gap, we have developed an accurate shape string predictor based on two innovative technologies: a knowledge-driven sequence alignment and a sequence shape string profile method. The performance on blind test data demonstrates that the proposed method can be used for accurate prediction of protein shape string. The DSP server provides both predicted shape string and sequence shape string profile for each query sequence. Using this information, the users can compare protein structure or display protein evolution in shape string space. The DSP server is available at both http://cheminfo.tongji.edu.cn/dsp/ and its main mirror http://chemcenter.tongji.edu.cn/dsp/.

INTRODUCTION

Fast and accurate structure comparison are fundamental in structural and evolutionary biology. Such a task depends on choosing an appropriate representation for a protein 3D structure. The obvious representation of a protein is atom coordinates, which is commonly used in structural alignment or structural superposition methods (1–3). Many of the very recent structural alignment approaches reduce the protein to a coarse metric, such as structural fragments (4–6) or secondary structure elements (7–9), which can also produce sensible alignments.

A Ramachandran plot (10,11) is a plot of dihedral torsion angles *phi* versus *psi* angles. It maps the entire conformational space of a polypeptide and illuminates the allowed and disallowed conformations. Since the allowed combinations of *phi*/*psi* angles in the Ramachandran plot are highly clustered, Ison *et al.* (12) clustered phi/psi torsion angle pairs of backbone protein structure into eight distinct regions, and assigned these clusters as eight symbols to describe the backbone protein structure [see Figure 1 (12) for detail]. Thus a sequence of such shape symbols, one per residue, called shape string, is a 1D structural alphabet representation for protein tertiary structure.

Many studies have demonstrated that shape string is an appropriate structure representation, and indicate precisely the backbone conformation of protein structure, which can carry more structural information than classical secondary structure representation (12,13). Our previous work (14,15) has demonstrated that shape string is extremely important at identifying tight turns as well. In our further study, we find that both shape string and its profile play an important role in DNA-binding residues prediction and protein post-translational modification prediction.

Here we present DSP server, based on two innovative technologies: a knowledge-driven sequence alignment and a sequence structure profile method (16). When tested on a benchmark set, DSP produces superior segment overlap (SOV) (17) measure (82.0%), overall accuracy for three-state shape stings (S3, 83.6%) and eight-state shape stings (S8, 74.4%) values, outperforming Frag1D (13) by 4.7% in SOV, 6.9% in S3 and 6.8% in S8. To assess the DSP method on newly measured proteins, we construct a non-redundant independent test data set (25% sequence identity, 916 entries). DSP achieves an S3 of 84.6% and an S8 of 75.3%.

METHODS

The flowchart of the shape string prediction is shown in Figure 1A. The PSI-BLAST (18) algorithm was initially employed to match a query sequence against a protein database constructed by a non-redundant PDB chain set

*To whom correspondence should be addressed. Tel: +86 21 65983987; Fax: +86 21 65983987; Email: lith@tongji.edu.cn

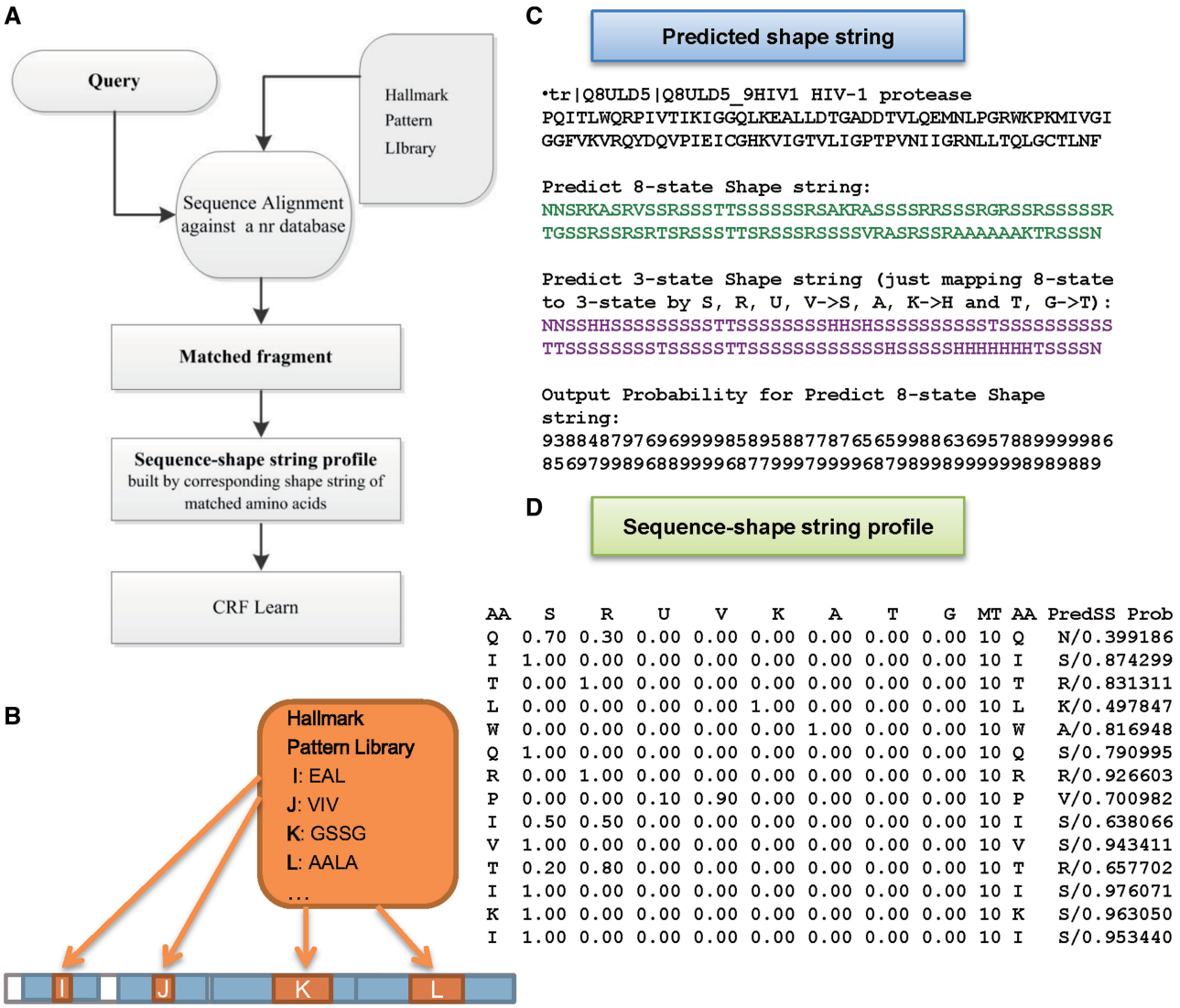

Figure 1. (**A**) The flowchart of the prediction of shape string and (**B**) sequence alignment with hallmark patterns as seeds. An example of (**C**) the predicted shape string and (**D**) the output sequence shape string profile. AA, amino acid; MT, match times; PredSS, predicted shape string; Prob, output probability.

(nr3PDB, NCBI MMDB 2009 December, 40 849 entries in total), resulting in two parts: matched fragments and unmatched fragments. Then, we utilized the hallmark patterns in the hallmark pattern library (HPL) (see below) to hit the unmatched fragments and obtain the hit segments (Figure 1B). These hit segments and their flanking amino acids (+n and −n, default is 5 in this study) were aligned together against nr3PDB using PHI-BLAST (19), which found more matched shorter sequences. The matched fragments obtained by the first alignment and the shorter sequences obtained by the subsequent alignments were encoded based on corresponding shape string element profiles. The shape string element profile was composed of eight elements (*S, R, U, V, K, A, T* and *G*), which was employed as feature for predicting the shape string of the query. Lastly, conditional random field (CRF) was performed for modeling and prediction.

Hallmark pattern generated

One innovative character of our approach was a knowledge-driven sequence alignment guided by seeds in a constructed HPL, which was instrumental in searching structural similarities among highly divergent proteins. Initially, we began a traversal search for consecutive sequence patterns with sufficient frequency in a representative non-redundant PDB chain set (nr0PDB, NCBI MMDB 2009 Dec, 7775 entries, 0-level non-redundancy, two sequences are considered similar if they have a BLAST *E*-value of 10^{-7} or less). In our previous study (20), we introduced an algorithm that could extract local combinational variables with fixed locations from equal length sequences. Here, the algorithm was developed to extract candidate patterns from unequal length sequences without sequence alignment (Figure 2). These short patterns were merged with every other single fragment

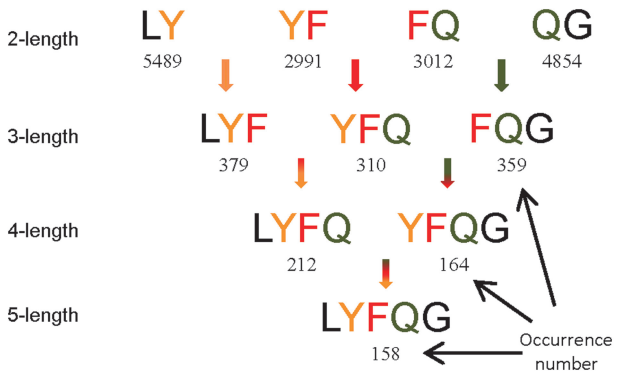

Figure 2. An illustration of consecutive sequence pattern mining.

that contained the same residue as the former fragment in order to form potentially longer sequences while maintaining the frequency criterion. We set the frequency criterion to 100 and a total of 5667 consecutive sequence patterns were obtained. The entire pattern extraction process progressed as the fragment grew longer, a process known as the bottom-up method.

Second, hallmark patterns were defined as conservative both in sequence patterns and shape string structures. For each position of a consecutive sequence pattern, the *P*-value of the corresponding shape string of the amino acid at this position was calculated according to a binomially distributed model [see Equation (1) below],

$$\mathrm{P}(N,m_j,q_j) = \sum_{i=m}^{N} \binom{N}{i} q_j^i (1-q_j)^{N-i} \quad (1)$$

where N denoted the occurrence number of the pattern; m_j denoted the count of maximum occurrence shape string at position j in the pattern; and q_j denoted the corresponding shape string background probability of residue at position j. If one of the *P*-values of a pattern was $<10^{-6}$, the consecutive sequence pattern was identified as a significant hallmark pattern.

Thirdly, based on the *P*-values, we selected 2761 hallmark patterns with lengths ranging between 2 and 4 residues that typically exhibited conserved structures to construct the library. The HPL represented remote homology in the sequences and shape strings and was an indispensable tool in our approach.

Sequence shape string profile

The sequence shape string profile was another innovative character of our approach, which was generated as follows: In the first step, the query sequence was aligned against the nr3PDB (NCBI MMDB 2009 December, 3-level non-redundancy, 40 849 entries in total) resulting in the top n ($n = 10$ in this work) subjects. Then, the shape strings of the n subjects were retrieved. Finally, the shape string elements of every amino acid were counted and stored in eight boxes. These boxes constituted a vector that represents the sequence shape string profile for each residue and was considered to include the structural evolutionary information [more details about sequence structure profile can be found in our previous study (16)].

RESULTS

Characteristics of the data set

The data set we used is nr0PDB with X-ray resolution better or equal to 3.0 Å. There are eight elements of shape string (expressed by italic in this study), which are shape *S* (β-sheets), *R* (polyproline type α structure), *U*, *V* (bridging regions), *K* (3_{10} helices), *A* (α-helices), *T* (Turns) and *G* (almost entirely glycine). Besides we define *N* for missing shape strings where *phi* or *psi* angle is undefined, or gaps in a PDB entry where no atom coordinates exists for parts of the structure. The actual corresponding shape strings are retrieved from http://www.fos.su.se/~pdbdna/ according to their tertiary structures. The shape string composition for each amino acid is presented in Table 1. It can be seen that glycine, proline and valine are significantly exhibited as shape *T*, *R* and *S*, respectively, and all other amino acids prefer to be shape *A*.

Performance evaluation

The training data set, containing 4234 chains, was derived from the PDB (21) released before 2010 and was determined by X-ray diffraction with a resolution of ≤2.0 Å, an *R*-factor of ≤0.25 and was cutoff at 25% sequence identity using PISCES (22).

We used the overall accuracy for eight-state shape stings (S8) to evaluate the performances of a 5-fold cross-validation on train set. To calculate overall accuracy for three-state shape strings (S3), we mapped eight-state shape string to the three state by $[S,R,U,V] \rightarrow S, [A,K] \rightarrow H$ and $[T,G] \rightarrow T$ as defined by Zhou *et al.* (13). The DSP achieved an overall per-residue accuracy for the three-state shape stings and eight-state shape strings of 88.7% and 80.9%, respectively, and an SOV of 86.4%, which was very close to the theoretical upper limit of accuracy of the secondary structure prediction (23).

To assess DSP and the effect of the hallmark patterns, we used the latest EVA set (24) as an independent test set, which contained 79 proteins (1 was abolished out of 80 entries in the EVA set). The detailed results are listed in Figure 3. The prediction by our method produced superior SOV (82.0%) and S3 (83.6%) values, outperforming an existing state-of-the-art method, Frag1D, by at least 6.9% in S3 and 4.7% in SOV. The more difficult S8 measure showed a remarkable improvement in performance (S8 74.4%, outperforming Frag1D by 6.8%) as well. The same trend occurred when the hallmark patterns were employed (outperforming when the hallmark patterns are not used by 9.2% in S3 and 6.2% in S8).

To assess the DSP method on newly measured proteins, we constructed independent test data by retrieving protein data released in the year 2010 from PDB, which were determined by X-ray diffraction with a resolution of ≤2.0 Å, an *R*-factor of ≤0.25, culled at 25% sequence identity and contained 916 chains. Our method achieved an S3 of 84.6% and an S8 of 75.3%. The accuracy of prediction on three-state shape strings *S*, *H* and *T* were

Table 1. The shape string composition for 20 amino acids

	S	*R*	*U*	*V*	*K*	*A*	*T*	*G*
A	0.156	0.157	0.012	0.021	0.048	0.590	0.013	0.003
R	0.220	0.134	0.017	0.013	0.060	0.526	0.026	0.003
N	0.223	0.115	0.039	0.023	0.159	0.324	0.112	0.004
D	0.211	0.156	0.026	0.016	0.122	0.413	0.052	0.004
C	0.308	0.192	0.039	0.010	0.064	0.361	0.023	0.003
Q	0.172	0.146	0.017	0.013	0.074	0.547	0.029	0.002
E	0.164	0.132	0.011	0.013	0.057	0.601	0.020	0.002
G	0.158	0.122	0.005	0.006	0.042	0.199	0.334	0.134
H	0.275	0.144	0.030	0.020	0.110	0.375	0.043	0.003
I	0.416	0.113	0.003	0.005	0.039	0.422	0.002	0.001
L	0.250	0.125	0.006	0.011	0.042	0.557	0.008	0.001
K	0.206	0.142	0.010	0.013	0.052	0.544	0.031	0.003
M	0.242	0.146	0.012	0.013	0.049	0.521	0.016	0.002
F	0.332	0.139	0.023	0.011	0.073	0.404	0.017	0.001
P	0.005	0.559	0.006	0.018	0.026	0.386	0.001	0.000
S	0.261	0.203	0.013	0.008	0.052	0.439	0.020	0.004
T	0.338	0.177	0.009	0.004	0.073	0.391	0.006	0.002
W	0.266	0.183	0.016	0.010	0.049	0.462	0.012	0.002
Y	0.332	0.137	0.025	0.010	0.095	0.383	0.017	0.002
V	0.455	0.130	0.005	0.003	0.041	0.362	0.003	0.001

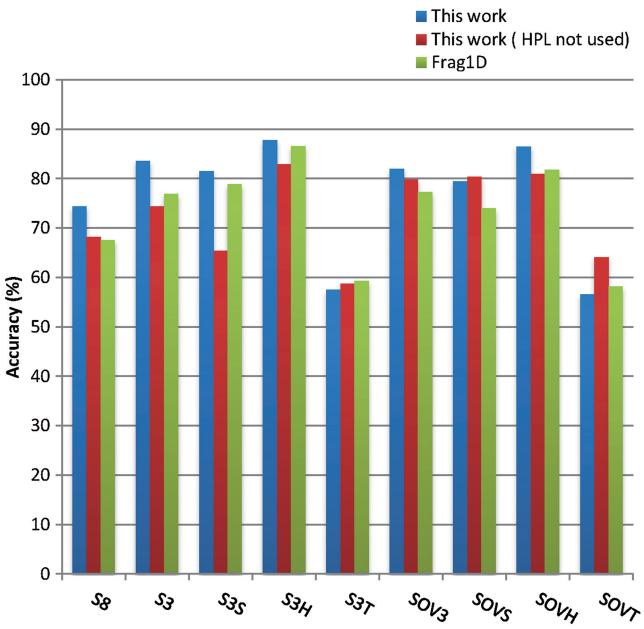

Figure 3. Performance comparison on EVA benchmark set.

82.0, 88.4 and 69.5%, respectively. The performance on newly measured PDB data demonstrated that the proposed method can be used for accurate prediction of protein shape string.

WEBSERVER

DSP provides an interactive web-based platform for predicting protein shape string and its profile. The input sequences (FASTA format, unique id required) can be entered either directly or by uploading a sequence file. An example input is provided by the server, which can be easily loaded to try out the DSP. There is also an option whether HPL is used or not. The default is false for most cases. If selected, only 20 entries are allowed, and it would produce more accurate predicted results and may take a longer time for predicting sequence which shares low homology to the structures of known proteins.

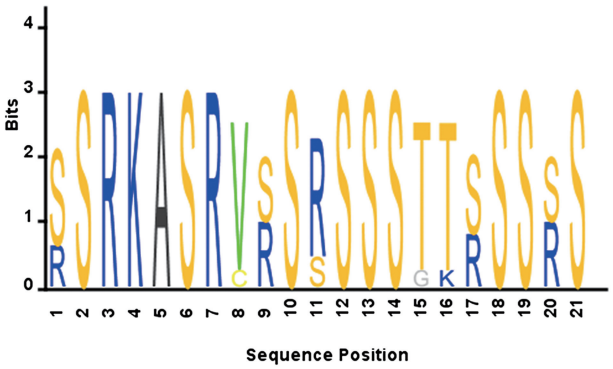

Figure 4. A sequence logo created by sequence-shape string profile. (It should be noted that letter C is denoted as shape *U*).

Output description

There are two main outputs: predicted shape string (Figure 1C) and sequence shape string profile (Figure 1D) presented in the result page. Upon submitting a sequence set, a progress window will be launched with a link to the result page and additional information about the job being performed. The page will refresh every 20 s until the output is ready. Once the job is completed, the predicted results will automatically appear on the web browser. The results contain query sequence, predicted eight-state shape string with output probability and predicted three-state shape string (just map eight state to three state by $[S, R, U, V] \rightarrow S, [A, K] \rightarrow H$ and $[T, G] \rightarrow T$). An email notification containing a link to the final results will be sent, if any email provided. User can download all predicted shape string in FASTA format and its profile with output probability in the result page as well.

The output sequence shape string profile (Figure 1D) is displayed as: residue, profile for shape *S*, *R*, *U*, *V*, *K*, *A*, *T* and *G*, match times, residue, predicted shape string and output probability each row. The profile can be used as features for modeling or protein structure comparison. Figure 4 is a sequence logo (25) constructed by the sequence shape string profile, which displays protein evolution information in shape string space.

Software

We also provide source code and binary executable program, which enable users performing the prediction on local machine. The software is developed in c# 4.0, which enables it run on multi-platforms, such as Linux, OS X and Microsoft Windows. The program can be freely downloaded via (http://cheminfo.tongji.edu.cn/dsp/Home/Downloads).

CONCLUSION

Despite the developments in discovering motifs, little research has focused on the relationship between sequence patterns and their corresponding structures. The hallmark pattern that we propose in this study is a union of sequences and shape strings. We use these conformational restricted hallmark patterns as seeds to guide sequence alignment and the sequence shape string profile as features, and show that protein shape string can be accurately predicted. The DSP server provides services for predicting and analysing protein backbone conformation in shape string space. User can obtain the shape string quickly and accurately or download program to run on local machine for great calculation demand. In the near future, we will add structure comparison by protein shape string services.

ACKNOWLEDGEMENTS

The authors are grateful to the anonymous reviewers for their valuable comments and help in improving the manuscript.

FUNDING

National Natural Science Foundation of China (NSFC) [20675057, 20705024]. Funding for open access charge: NSFC.

Conflict of interest statement. None declared.

REFERENCES

1. Holm,L. and Sander,C. (1993) Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.*, **233**, 123–138.
2. Shindyalov,I.N. and Bourne,P.E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.*, **11**, 739–747.
3. Theobald,D.L. and Wuttke,D.S. (2006) Empirical Bayes hierarchical models for regularizing maximum likelihood estimation in the matrix Gaussian Procrustes problem. *Proc. Natl Acad. Sci. USA*, **103**, 18521–18527.
4. Tung,C.H., Huang,J.W. and Yang,J.M. (2007) Kappa-alpha plot derived structural alphabet and BLOSUM-like substitution matrix for rapid search of protein structure database. *Genome Biol.*, **8**, R31.
5. Friedberg,I., Harder,T., Kolodny,R., Sitbon,E., Li,Z. and Godzik,A. (2007) Using an alignment of fragment strings for comparing protein structures. *Bioinformatics*, **23**, e219–224.
6. Budowski-Tal,I., Nov,Y. and Kolodny,R. (2010) FragBag, an accurate representation of protein structure, retrieves structural neighbors from the entire PDB quickly and accurately. *Proc. Natl Acad. Sci. USA*, **107**, 3481–3486.
7. Przytycka,T., Aurora,R. and Rose,G.D. (1999) A protein taxonomy based on secondary structure. *Nat. Struct. Biol.*, **6**, 672–682.
8. Krissinel,E. and Henrick,K. (2004) Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr. D Biol. Crystallogr.*, **60**, 2256–2268.
9. Fontana,P., Bindewald,E., Toppo,S., Velasco,R., Valle,G. and Tosatto,S.C. (2005) The SSEA server for protein secondary structure alignment. *Bioinformatics*, **21**, 393–395.
10. Ramachandran,G.N., Ramakrishnan,C. and Sasisekharan,V. (1963) Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.*, **7**, 95–99.
11. Hovmöller,S., Zhou,T. and Ohlson,T. (2002) Conformations of amino acids in proteins. *Acta Crystallogr. D Biol. Crystallogr.*, **58**, 768–776.
12. Ison,R.E., Hovmöller,S. and Kretsinger,R.H. (2005) Proteins and their shape strings. An exemplary computer representation of protein structure. *IEEE Eng. Med. Biol. Mag.*, **24**, 41–49.
13. Zhou,T., Shu,N. and Hovmöller,S. (2010) A novel method for accurate one-dimensional protein structure prediction based on fragment matching. *Bioinformatics*, **26**, 470–477.
14. Tang,Z., Li,T., Liu,R., Xiong,W., Sun,J., Zhu,Y. and Chen,G. (2011) Improving the performance of beta-turn prediction using predicted shape strings and a two-layer support vector machine model. *BMC Bioinformatics*, **12**, 283.
15. Zhu,Y., Li,T., Li,D., Zhang,Y., Xiong,W., Sun,J., Tang,Z. and Chen,G. (2012) Using predicted shape string to enhance the accuracy of gamma-turn prediction. *Amino Acids*, **42**, 1749–1755.
16. Li,D., Li,T., Cong,P., Xiong,W. and Sun,J. (2012) A novel structural position-specific scoring matrix for the prediction of protein secondary structures. *Bioinformatics*, **28**, 32–39.
17. Rost,B., Sander,C. and Schneider,R. (1994) Redefining the goals of protein secondary structure prediction. *J. Mol. Biol.*, **235**, 13–26.
18. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
19. Zhang,Z., Schaffer,A.A., Miller,W., Madden,T.L., Lipman,D.J., Koonin,E.V. and Altschul,S.F. (1998) Protein sequence similarity searches using patterns as seeds. *Nucleic Acids Res.*, **26**, 3986–3990.
20. Xiong,W., Li,T., Chen,K. and Tang,K. (2009) Local combinational variables: an approach used in DNA-binding helix-turn-helix motif prediction with sequence information. *Nucleic Acids Res.*, **37**, 5632–5640.
21. Bernstein,F.C., Koetzle,T.F., Williams,G.J., Meyer,E.F. Jr, Brice,M.D., Rodgers,J.R., Kennard,O., Shimanouchi,T. and Tasumi,M. (1977) The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.*, **112**, 535–542.
22. Wang,G. and Dunbrack,R.L. Jr (2003) PISCES: a protein sequence culling server. *Bioinformatics*, **19**, 1589–1591.
23. Rost,B. (2001) Review: protein secondary structure prediction continues to rise. *J. Struct. Biol.*, **134**, 204–218.
24. Koh,I.Y., Eyrich,V.A., Marti-Renom,M.A., Przybylski,D., Madhusudhan,M.S., Eswar,N., Grana,O., Pazos,F., Valencia,A., Sali,A. *et al.* (2003) EVA: Evaluation of protein structure prediction servers. *Nucleic Acids Res.*, **31**, 3311–3315.
25. Schneider,T.D. and Stephens,R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100.