PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Assessing secondary structure assignment of protein structures by using pairwise sequence-alignment benchmarks

Wei Zhang,[1,2,3] A. Keith Dunker,[1,2] and Yaoqi Zhou[1,2]*

[1] Indiana University School of Informatics, Indiana University-Purdue University, Indianapolis, Indiana 46202
[2] Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, Indiana 46202
[3] Institute of Applied Physics and Computational Mathematics, Beijing 100088, People's Republic of China

ABSTRACT

How to make an objective assignment of secondary structures based on a protein structure is an unsolved problem. Defining the boundaries between helix, sheet, and coil structures is arbitrary, and commonly accepted standard assignments do not exist. Here, we propose a criterion that assesses secondary structure assignment based on the similarity of the secondary structures assigned to pairwise sequence-alignment benchmarks, where these benchmarks are determined by prior structural alignments of the protein pairs. This criterion is used to rank six secondary structure assignment methods: STRIDE, DSSP, SECSTR, KAKSI, P-SEA, and SEGNO with three established sequence-alignment benchmarks (PREFAB, SABmark, and SALIGN). STRIDE and KAKSI achieve comparable success rates in assigning the same secondary structure elements to structurally aligned residues in the three benchmarks. Their success rates are between 1–4% higher than those of the other four methods. The consensus of STRIDE, KAKSI, SECSTR, and P-SEA, called SKSP, improves assignments over the best single method in each benchmark by an additional 1%. These results support the usefulness of the sequence-alignment benchmarks as a means to evaluate secondary structure assignment. The SKSP server and the benchmarks can be accessed at http://sparks.informatics.iupui.edu

Proteins 2008; 71:61–67.

Key words: alignment benchmarks; protein structure prediction; secondary structures.

INTRODUCTION

The secondary structure of a protein refers to the local conformation of its polypeptide backbone. Knowing secondary structures of proteins is essential for their structure classification,[1,2] understanding folding dynamics and mechanisms[3–5] and discovering conserved structural/functional motifs.[6,7] Secondary structure information is also useful for sequence and multiple sequence alignment,[8,9] structure alignment,[10,11] and sequence to structure alignment (or threading).[12–15] As a result, predicting secondary structures from protein sequences continues to be an active field of research[16–18] 56 years after Pauling and Corey[19,20] first predicted that α-helix and β-sheet are the most common regular patterns of protein backbones. Prediction and application of protein secondary structure rely on prior assignment of secondary structure elements from a given protein structure by human inspection or automatic computational methods.

Many computational methods have been developed to automate the assignment of secondary structures. Examples are DSSP,[21] STRIDE,[22] DEFINE,[23] P-SEA,[24] KAKSI,[25] P-CURVE,[26] XTLSSTR,[27] SECSTR,[28] SEGNO,[29] and VoTAP.[30] These methods are based on either the hydrogen-bond pattern, geometric features, expert knowledge or their combinations. However, they often disagree on their assignments. For example, disagreement among DSSP, P-CURVE, and DEFINE can be as large as 25%.[31] More β sheet is assigned by XTLSSTR[27] and more π-helix by SECSTR[28] than by DSSP. The discrepancy among different methods is caused by nonideal configurations of helices and sheets.[32–34] As a result, defining the boundaries between helix, sheet, and coil is problematical and a significant source of discrepancies between different methods.

Inconsistent assignment of secondary structures by different methods highlights the need for a criterion or a benchmark of "standard" assignments that could be used to assess and compare assignment methods. One possibility is to use the secondary structures assigned by the authors who solved the protein structures. STRIDE,[22] in fact, has been optimized to achieve the highest agreement with the authors' annotations. However, it is not clear what is the criterion used for manual or automatic assignment of secondary structures by different authors. Indeed, Levitt and Greer,[35]

Grant sponsor: NIH; Grant numbers: R01 GM 966049, R01 GM 068530.
*Correspondence to: Prof. Yaoqi Zhou, Indiana University School of Medicine, Walkar Plaza, 719 Indian Ave. Suite 319, Indianapolis, IN 46202. E-mail: yqzhou@iupui.edu
Received 9 April 2007; Revised 25 May 2007; Accepted 6 June 2007
Published online 11 October 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21654

who developed the first algorithm for assignment of secondary structure from 3D structure, pointed out the subjective and variable nature of author assignments and the increased consistency of objective, algorithm-based methods. Another possibility is to treat the consensus prediction by several methods as the gold standard.[31,30] However, there is no obvious reason why each method should weigh equally in assigning secondary structures. Furthermore, it not even clear which methods should be chosen for the development of a consensus. Other criteria include helix-capping propensity,[36] the deviation from ideal helical and sheet configurations,[29] and structural accuracy produced by sequence-to-structure alignment guided by secondary structure assignment.[29]

In this article, we propose to use sequence-alignment benchmarks for assessing secondary structure assignments. These benchmarks are produced by 3D-structure alignment of structurally homologous proteins. Instead of assessing the accuracy of secondary structure assignment directly, which is not yet feasible, we compare the two assignments of secondary structures in structurally aligned positions. We assume that the best method should assign the same secondary structure element to the highest fraction of structurally aligned positions. Certainly, structurally aligned positions do not always have the same secondary structures. Moreover, different structure-alignment methods do not always produce the same result.[15,37,38] Nevertheless, this criterion provides a means to locate a secondary structure assignment method that is most consistent with tertiary structure alignment. We suggest that this approach provides an objective evaluation of secondary structure assignment methods.

Here, we assessed six methods (DSSP, STRIDE, SECSTR, KAKSI, P-SEA, and SEGNO) using three established sequence alignment benchmarks based on structural alignment. By this measure, the top two methods are STRIDE and KAKSI. A consensus method based on four methods further improves the assignments compared to STRIDE and KAKSI.

METHODS

Secondary structure assignment methods

We obtained six secondary structure methods that are available from internet.

DSSP[21] is, perhaps, the most popular program for secondary structure assignment. This program assigns secondary structures according the pattern of hydrogen bonds. DSSP was obtained from http://www.cmbi.kun.nl/swift/dssp/.

STRIDE[22] is another widely used program for secondary structure assignment. This program uses hydrogen bonds patterns and the backbone torsion angle (Φ/Ψ) and was optimized to be consistent with expert assignments in protein structures. It was obtained from ftp://ftp.ebi.ac.uk/pub/software/unix/stride/src.

SECSTR[28] applies the criteria similar to those used in DSSP for defining hydrogen bonds. This program uses additional structural information to assign π-helix. SECSTR was obtained from http://www.mbfys.lu.se/Services/SecStr/.

P-SEA[24] makes use of several geometric criteria for secondary structure assignment. This program includes the C_α—C_α distances as well as their associated angles. P-SEA was obtained from ftp://ftp.lmcp.jussieu.fr/pub/sincris/software/protein.

SEGNO[29] is also a geometric-based approach. Helices are defined based on whether residues fit inside an imaginary cylinder: they must be within the correct radius of a central axis. β-strands are assigned based on backbone dihedrals and with alternating peptide bonds. SEGNO was obtained directly from Dr. Lovell.

KAKSI[25] is based on the geometric information of a set of characteristic values of C_α—C_α distances and Φ/Ψ backbone-dihedral angles. The parameters of KAKSI have been chosen to best fit the secondary structure assignment obtained from PDB files. KAKSI was obtained from http://migale.jouy.inra.fr/mig/mig_fr/servlog/kaksi/.

Sequence-alignment benchmarks

PREFAB 4.0 contains 1682 pairs of proteins.[39] Each pair of structures are aligned using the CE aligner,[40] and only those pairs for which FSSP[41] and CE agreed on 50 or more positions are retained. We have removed those pairs of proteins for which the sequences from PDB files are inconsistent with those used in the alignment. Also, there were some proteins for which some of the six methods failed to produce secondary structures for unknown reasons (only executables are available, which prevented further investigation). Thus, a common set of 416 pairs of proteins (with an average sequence identity of 29.6%) for which all six methods produce secondary structure assignments is used.

SABmark[42] (Sequence Alignment Benchmark) is a multiple sequence alignment benchmark. It contains two sets. The superfamily set contains sequences that have low-to-intermediate sequence identities while the twilight set is made of sequences with low sequence identities. Reference alignments are from consensus structure alignments by SOFI[43] and CE.[40] From SABmark, we select one pair of proteins from each group in the twilight set. A common set of 167 pairs of proteins for which all six methods made assignment are kept. The average sequence identity for this set is 14.9%.

SALIGN benchmark[44] contains 200 selected pairs with an average pair sharing 20% sequence identity and 65% of structurally equivalent C_α atoms superposed with an rmsd of 3.5Å.[44] Reference alignment is obtained from the structural alignment obtained from the TMalign

Table I
The Average Fractions of Structurally Aligned Residues with the Same Secondary Structure Assignment (and Standard Deviations) Given by Six Secondary Structure-Assignment Methods for Three Sequence-Alignment Benchmarks

	STRIDE (%)	DSSP (%)	SECSTR (%)	KAKSI (%)	P-SEA (%)	SEGNO (%)
PREFAB	86.4 ± 0.4	85.4 ± 0.4	85.8 ± 0.4	86.1 ± 0.4	83.2 ± 0.4	83.2 ± 0.4
SALIGN	79.9 ± 0.6	78.8 ± 0.6	79.2 ± 0.6	80.2 ± 0.6	76.0 ± 0.7	76.3 ± 0.6
SABmark	81.2 ± 0.8	80.0 ± 0.8	80.0 ± 0.8	82.0 ± 1.0	77.8 ± 1.0	81.1 ± 1.0

program[38] that optimizes Template Modeling score between two structures. For the reason described above, we used 143 pairs (of 200 pairs) of proteins as the benchmark for secondary structure assignment. The average sequence identity for this set is 16.4%.

All structure-alignment methods (CE,[40] FSSP,[41] TM-align,[38] and SOFI[43]) used for above benchmarks do not use secondary structures to assist pairwise structural alignment. Thus, there is no intrinsic bias toward a specific secondary structure assignment method.

To assess the possible overlaps between the three above-described benchmarks, we calculate sequence identity between sequences in different benchmarks. Two benchmarks share a homologous sequence pair if the two sequences of a pair in one benchmark have more than 40% sequence identity separately with the two sequences in a pair in another benchmark. We found that there is overlap of one pair between SABmark and SALIGN, five pairs between PREFAB and SALIGN, and four pairs between PREFAB and SABmark benchmark. These overlap sequence pairs in different benchmarks were not removed because (1) the number of these pairs is too small to make any difference in the results presented here and (2) all benchmarks are used here for testing rather than for training.

Assessment

We assess each assignment method by the success rate or the fraction of structurally aligned residues (in the reference alignment provided by each benchmark) having the same secondary structure assignment. A three-state [Helix (H), Strand (E), and Coil (C)] definition is used. In the DSSP and STRIDE methods, we use the following assignment: (H,G,I)→ H, (E,B) → E, and others → C. For SECSTR, we apply the following conversion: (H,G,I)→ H, (E,e) → E, and others → C. For SEGNO, (H,G,I,h)→ H, (E,e,B,b,P,p) → E, and others → C. KAKSI and P-SEA only assign three states.

Consensus methods

For an equally weighted consensus method, its prediction is the state predicted by more methods than all others. If two states are assigned by an equal number of assignment methods, the secondary structure is chosen according to the priority of H, E, then C. We use this priority because we prefer a consensus method to give more structural information and the helical residues have more recognizable structural characteristics than strand residues which have more than coil residues. The effect of this priority is small. For example, the difference between the EHC and HEC priorities is only 0.1% in success rate by SKSP for the PREFAB benchmark.

We also test an unequally weighted consensus method where the prediction given by different methods is weighted differently. The weights are optimized by a simple grid search.

RESULTS

Table I shows the fractions of structurally aligned residues with the same secondary structure assignment (success rates) given by six secondary structure-assignment methods for three benchmarks. The highest success rates are 86.4% by STRIDE in PREFAB, 80.2% by KAKSI in SALIGN, and 82.0% by KAKSI in SABmark. However, the difference in success rates given by STRIDE and KAKSI is statistically insignificant for all three benchmarks. The results of DSSP and SECSTR are very similar and only about 1% less in success rates than either STRIDE or KAKSI. P-SEA and SEGNO are 3% less in success rates than either STRIDE or KAKSI for PREFAB and SALIGN benchmarks. In SABmark, the success rate of P-SEA is 4% less from the best and that of SEGNO is 1% from the best.

Above results indicate that different methods perform differently. Thus, it is of interest to know similarity between the assignments made by different methods. Table II displays the percent of agreement in secondary structure assignment given by a pair of methods for the PREFAB benchmark. DSSP, STRIDE, and SECSTR are most similar to each other and their pairwise agreement ranges from 92 to 95%. The likely origin of this agreement is that all three methods use a similar definition of hydrogen bonds. These three methods differ considerably from the other three methods. The agreement between these two groups of methods is between 81.9% and 85.0%. In addition, the other three methods (KAKSI, P-SEA, and SEGNO) also differ considerably from each other (their agreement is between 82.8% and 86.2%).

Table II
The Pairwise Agreement in Secondary Structure Assignment Given by Six Methods for the PREFAB Benchmark

	DSSP (%)	STRIDE (%)	KAKSI (%)	SECSTR (%)	P-SEA (%)	SEGNO (%)
DSSP	100.0	95.4	83.6	94.0	82.1	82.3
STRIDE		100.0	85.0	92.1	83.3	81.9
KAKSI			100.0	83.2	82.8	85.0
SECSTR				100.0	81.9	83.9
P-SEA					100.0	86.2
SEGNO						100.0

They differ because they use different geometric parameters for assigning secondary structures.

KAKSI and STRIDE differ in assignment by 15% (Table II), yet they have the highest success rates in agreeing with 3D-structure alignment (Table I). This suggests that a consensus method might be useful for improving the consistency between secondary structure assignment and structure alignment. Table III shows the performance of various possible equally weighted combinations of three to five methods. We do not include DSSP for consensus because DSSP assignment is very similar to STRIDE (95%) and SECSTR (94%). The top two best combinations are made by STRIDE, KAKSI, and P-SEA (SKP) or by STRIDE, KAKSI, SECSR, and P-SEA (SKSP). Both yield a success rate of 87.5%, which is 1% better than STRIDE alone. This improvement is statistically significant, judged by the fact that the standard deviations of the success rates are only 0.3–0.4%. Many other consensus methods have very similar performance. However, some perform worse than STRIDE. We further found that optimizing weight factors for different methods (STRIDE, KAKSI, SECSR, and P-SEA) can only further increase the success rate from 87.5% to 87.8%–a statistically insignificant improvement. Thus, weight factors were not considered further.

A consensus method such as SKSP improves the consistency with 3D-structure alignment. We found that it changes the distribution of helical or strand fractions somewhat. For the PREFAB benchmark, the fractions of helical, strand, and coil residues are 45.0%, 26.3%, and 28.6% for the SKSP method, respectively. The corresponding fractions are 42.9%, 24.6%, and 32.5% for the DSSP method, respectively. That is, there is an increase in helix (2%) and strand (2%) residues and a decrease in coil (4%) residues. These changes reflect a slight increase in average lengths of helices and strands (from 8.6 by DSSP to 9.4 residues by SKSP for helices and from 4.5 to 4.8 residues for strands).

To test the consensus methods, we apply the best two to SALIGN and SABmark. For the SKP consensus method (the best in the three-method consensus), the success rate is 81.8% for SALIGN and 82.1% for SABmark (the corresponding highest success rates for a single method are 80.2 and 82.0, respectively). The SKSP consensus method yields 81.4% for SALIGN benchmark and 83.0% for SABmark benchmark. Thus, SKSP consistently produces a higher success rate (∼1%) than the best single method for a given benchmark and 2–3% higher than DSSP for all three benchmarks.

What is the source for only 81–88% consistency between secondary structure assignment and 3D-structure alignment? To answer this question, we evaluate the success rates given by DSSP and the SKSP consensus method according to the sequence identity between two aligned sequences. Results are shown in Table IV. It is clear that both DSSP and SKSP make highly consistent assignments (93–94% in success rates) with structure alignment for homologous proteins (>50% in sequence identity). As the sequence identity between two sequences decreases (remote and structural homologs), the discrepancy between secondary structure assignment and structural alignment increases. Structure alignment for remote homologs is not as reliable as structure alignment for close homologs because the structures for close homologs are more conserved and thus easier to align with consistency. Moreover, the consistency of secondary structure assignment should be independent of sequence identity, if the structures are aligned correctly. Thus, somewhat arbitrary structure alignment between structurally variable regions of remote homologs is the main source for the low success rates of secondary structure assignment methods at low sequence identities. The excellent agreement between secondary structure assignment and structure alignment for homologs indicates that the error due to inconsistency of secondary structure assignment is less than 6%. It is of interest to note that the overlap between DSSP and SKSP is 92%, regardless of sequence identity.

Table III
The Success Rates (and Standard Deviations) in the PREFAB Benchmark Given by Various Combinations of Equally-Weighted Methods

Consensus component					
STRIDE	KAKSI	SECSTR	P-SEA	SEGNO	Succ. rate (%)
x			x	x	85.7 ± 0.4
	x		x	x	86.1 ± 0.4
		x	x	x	85.2 ± 0.3
	x	x		x	86.3 ± 0.3
x		x		x	86.5 ± 0.4
x	x			x	86.7 ± 0.3
	x	x	x		87.2 ± 0.4
x		x	x		86.6 ± 0.4
x	x		x		87.5 ± 0.4
x	x	x			86.7 ± 0.4
	x	x	x	x	86.7 ± 0.3
x		x	x	x	86.1 ± 0.4
x	x		x	x	86.9 ± 0.3
x	x	x		x	87.3 ± 0.3
x	x	x	x		87.5 ± 0.3
x	x	x	x	x	87.3 ± 0.3

Table IV
Success Rates and Assignment Errors in the PREFAB Benchmark by DSSP and The SKSP Consensus Method

		Succ. rate		DSSP/SKSP	Misassignment (SKSP)			
Seq. identity (%)	# pairs	DSSP (%)	SKSP (%)	similarity	H–C (%)	H–E (%)	C–E (%)	All (%)
<15	98	79.7	82.9	92.3	7.9	0.4	8.8	17.1
15–30	177	83.8	86.0	92.8	6.9	0.4	6.7	14.0
30–50	78	89.6	91.4	92.4	4.3	0.0	4.3	8.6
>50	63	93.3	94.2	92.6	3.2	0.0	2.6	5.8
All	416	85.4	87.5	92.6	6.1	0.3	6.1	12.5

Table IV also displays the distribution of errors (percent of inconsistent secondary structure assignment of structurally aligned residues). The majority of errors involves misidentification between helix and coil residues and between strand and coil residues. Interestingly, misidentification between helix and strand residues is small, regardless the sequence identity between the sequences that are aligned structurally. This suggests that the majority of structure-alignment errors also involves coil residues.

To further analyze inconsistent secondary structure assignment of structurally aligned residues in more detail, we define two different boundaries. For a given structurally aligned residue pair, it is not in the boundary of aligned regions if both nearest neighboring pairs (both before and after this pair) are also structurally aligned. Similarly, for a given structurally aligned residue pair in the same secondary structure, it is not in the boundary of a secondary structure element if both nearest neighboring pairs (both before and after this pair) are also structurally aligned in the same secondary structure element. For the SKSP method, 58.3% of inconsistent assignments occur at the boundary (the beginning or ending) of helix or strand elements and 30.5% at the boundary of aligned regions. There is only 11.2% of inconsistency is in the middle of a helix or strand, or coil.

The criterion proposed in this article for evaluating secondary structure assignments may favor a method whose assignment is dominated by one secondary structure type. For example, if there were only one secondary structure type, the structurally aligned pair will always have the same secondary structure type. Thus, any secondary structure assignment method that reduces the diversity of secondary structure types will improve the agreement with structural alignment. The diversity can be measured by d ($d = 1 - (|f_H - f_E| + |f_H - f_C| + |f_E - f_C|)/2$) where f_H, f_E, and f_C are fractions of helix, strand, and coil residues, respectively. $d = 0$ if there is only one state, $d = 0.5$ if there is only two equally distributed states, and $d = 1$, the largest diversity, if three states are equally distributed ($f_H = f_E = f_C$).

Table V shows the success rates and the diversity values given by DSSP and SKSP. The d value decreases for all three benchmarks when the DSSP assignment changes to the SKSP assignment. It decreases by 0.004 in PREFAB, 0.007 in SALIGN, and 0.065 in SABmark. The corresponding improvements in consistency between secondary structure assignment and structural alignment are 2.1%, 2.6%, and 3.0% for the three benchmarks, respectively. A ten-times greater reduction in diversity in SABmark than in SALIGN did not lead to a dramatic increase in consistency. Thus, the relation between diversity and consistency is not simple.

DISCUSSION

We have proposed a criterion to assess the methods for secondary structure assignment. This criterion is based on a structure assignment comparison, rather than the use of individually assigned secondary structures. It assumes that the best method for secondary structure assignment is the method that achieves the highest success rate of assigning the same secondary structure element to structurally aligned residues. This criterion works only if the structure-alignment program aligns structures without the use of secondary structure information. Many existing programs (such as CE, SOFI, TM-align) indeed perform structure alignment between two proteins using the three-dimensional coordinates only. This makes the application of this criterion possible. Three established sequence alignment benchmarks are used in this article to assess the methods for secondary structure assignment. Obviously, not all structurally aligned residues should have the same secondary structure assignment because structural alignment focuses on

Table V
The Success Rates and the Diversity (d) of the Assigned Secondary Structures Given by DSSP and the SKSP Consensus Method

	DSSP		SKSP		SKSP-DSSP	
Method	Succ. (%)	*d*	Succ. (%)	*d*	Succ. (%)	*d*
PREFAB	85.4	0.817	87.5	0.813	2.1	−0.004
SALIGN	78.8	0.797	81.4	0.790	2.6	−0.007
SABmark	80.0	0.909	83.0	0.844	3.0	−0.065

global rather than local structure. It is not possible to objectively identify which structurally aligned residue pair should or should not have the same secondary structure assignment. We avoid this problem by concentrating on the result based on all residue pairs rather than specific residue pairs.

We found that this criterion can indeed distinguish different secondary structure assignment methods by their different success rates. The difference is small (<4%) but statistically significant. STRIDE and KAKSI have the highest success rates among six methods compared. A consensus of four methods (STRIDE, KAKSI, SECSTR, and P-SEA) can further improve the assignments over any single method by an additional 1%. However, not all consensus methods improve assignments over the best single method.

About 30% of the inconsistent assignments occur at the boundaries of structurally aligned segments. This 30% error is likely caused by somewhat arbitrary structure alignment between structurally variable regions in close or remote homologs. Remote homologs are likely to have more unconserved regions than close homologs. This leads to the highest success rate of assigning the same secondary structure for close homologs (94% by SKSP) and the lowest for remote homologs (83% for sequence identity of 15% or less). Excluding the errors at the boundaries of structurally aligned segments increases the success rate for close homologs from 94 to 95%. If secondary structures were aligned perfectly in structure alignment of close homologs, there would be 5% remaining for further improvement of secondary structure assignment.

The largest inconsistency of secondary structure assignment is misclassification between helix and coil residues or between strand and coil residues and is located mostly at the boundary of a helix, strand, or coil segment (58%). This agrees with previous findings.[31] The smallest misassignment error (<0.4%), which is between helix and strand residues, is likely the result of physical constraints that prohibit helix and strand residues to locate next to each other.[45] Moreover, helices and strands are located in geometrically distinct locations in the Ramachandran diagram.[46]

The proposed criterion for secondary structure assignment can be used to discover new methods for secondary structure assignment, methods that are more consistent with the 3D structure alignments. Such an assignment method would likely be useful for improving the accuracy of sequence-to-structure threading. For example, one can develop a method for secondary structure prediction based on the SKSP consensus assignment method, rather than based on the DSSP assignment.[18] Also, the matching between predicted secondary structures of a query sequence and actual secondary structure of a template (assigned by SKSP) will likely serve as a better constraint for identification of structurally aligned regions than the method based on the DSSP assignment.[15,47] We found that SKSP is 2–3% higher in agreement with structurally aligned residues than DSSP for all three benchmarks, which is a significant result in the light of the fact that small improvement in alignment often leads to large improvement in predicted structures.[47]

ACKNOWLEDGMENT

We would like to thank Dr. Chi Zhang for many helpful discussions.

REFERENCES

1. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
2. Orengo C, Michie A, Jones S, Jones D, Swindells M, Thornton J. CATHY: a hierarchic classification of protein domain structures. Structure 1997;5:1093–1108.
3. Karplus M, Weaver DL. Protein-folding dynamics. Nature 1976;260: 404–406.
4. Kim PS, Baldwin RL. Intermediates in the folding reactions of small proteins. Annu Rev Biochem 1990;59:631–660.
5. Dobson CM, Sali A, Karplus M. Protein folding: a perspective from theory and experiment. Angew Chem Int Ed 1998;37:868–893.
6. Mizuguchi K, Go N. Comparison of spatial arrangements of the secondary structural elements in proteins. Protein Eng 1995;8:353–362.
7. Gibrat JF, Madej T, Bryant SH. Surprising similarities in structure comparison. Curr Opin Struct Biol 1996;6:377–385.
8. Simossis VA, Heringa J. Integrating secondary structure prediction and multiple sequence alignment. Curr Protein Pept Sci 2004;5: 1–15.
9. Zhou H, Zhou Y. SPEM: improving multiple-sequence alignment with sequence profiles and predicted secondary structures. Bioinformatics 2005;21:3615–3621.
10. Dror O, Benyamini H, Nussinov R, Wolfson H. Multiple structural alignment by secondary structures: algorithm and applications. Protein Sci 2003;12:2492–2507.
11. Krissinel E, Henrick K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. Acta Cryst D 2004;60:2256—2268.
12. Gribskov M, McLachlan AD, Eisenberg D. Profile analysis: detection of distantly related proteins. Proc Natl Acad Sci USA 1987;84:4355–4358.
13. Marti-Renom M, Stuart A, Fiser A, Sanchez R, Melo F, Sali A. Comparative protein structure modeling of genes and genomes. Annu Rev Biophys Biomol Struct 2000;29:291–325.
14. Zhou H, Zhou Y. Single-body residue-level knowledge-based energy score combined with sequence-profile and secondary structure information for fold recognition. Proteins 2004;55:1005–1013.
15. Zhou H, Zhou Y. Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. Proteins 2005;58:321–328.
16. Rost B. Review: protein secondary structure prediction continues to rise. J Struct Biol 2001;134:204–218.
17. Birzele F, Kramer S. A new representation for protein secondary structure prediction based on frequent patterns. Bioinformatics 2006;22:2628–2634.
18. Dor O, Zhou Y. Achieving 80% ten-fold cross-validated accuracy for secondary structure prediction by large-scale training. Proteins 2007;66:838–845.

19. Pauling L, Corey RB, Branson HR. The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. Proc Natl Acad Sci USA 1951;37:205–211.
20. Pauling L, Corey RB. The pleated sheet, a new layer configuration of polypeptide chains. Proc Natl Acad Sci USA 1951;37:251–256.
21. Kabsch W, Sander C. Dictionary of protein structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
22. Frishman D, Argos P. Knowledge-based protein secondary structure assignment. Proteins 1995;23:556–579.
23. Richards FM, Kundrot CE. Identification of structural motifs from protein coordinate data: secondary structure and first level supersecondary structure. Proteins 1988;3:71–84.
24. Labesse G, Colloc'h N, Pothier J, Mornon JP. P-SEA: a new efficient assignment of secondary structure from C alpha trace of proteins. Comput Appl Biosci 1997;13:291–295.
25. Martin J, Letellier G, Marin A, Taly JF, Brevern AGD, Gibrat GF. Protein secondary structure assignment revisited: a detailed analysis of different assignment methods. BMC Strut Biol 2005;5:17.
26. Sklenar H, Etchebest C, Lavery R. Describing protein structure: a general algorithm yielding complete helicoidal parameters and a unique overall axis. Proteins 1989;6:46–60.
27. King SM, Johnson WC. Assigning secondary structure from protein coordinate data. Proteins 1999;3:313–320.
28. Fodje MN, Al-Karadaghi S. Occurrence, conformational features and amino acid propensities for the π-helix. Protein Eng 2002;15: 353–358.
29. Cubellis MV, Cailliez F, Lovell SC. Secondary structure assignment that accurately reflects physical and evolutionary characteristics. BMC Bioinformatics 2005;6:S8.
30. Dupuis F, Sadoc JF, Mornon JP. Protein secondary structure assignment through Voronoi tessellation. Proteins 2004;55:519–528.
31. Colloc'h N, Etchebest C, Thoreau E, Henrissat B, Mornon J-P. Comparison of three algorithms for the assignment of secondary structure in proteins: the advantages of a consensus assignment. Protein Eng 1993;6:377–382.
32. Schulz GE, Barry CD, Friedman J, Chou PY, Fasman GD, Finkelstein AV, Lim VI, Pititsyn OB, Kabat EA, Wu TT, Levitt M, Robson B, Nagano K. Comparison of predicted and experimentally determined secondary structure of adenyl kinase. Nature 1974;250:140–142.
33. Robson B, Garnier J. Introduction to proteins and protein engineering. Amsterdam: Elsevier Press; 1986.
34. Barlow DJ, Thornton JM. Helix geometry in proteins. J Mol Biol 1988;201:601–619.
35. Levitt M, Greer J. Automatic identification of secondary structure in globular proteins. J Mol Biol 1977;114:181–239.
36. Richardson JS, Richarson DC. Amino acid preferences for specific locations at the ends of alpha helices. Science 1988;240:1648–1652.
37. Kolodny R, Koehl P, Levitt M. Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures. J Mol Biol 2005;346:1173–1188.
38. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucl Acids Res 2005;33:2302–2309.
39. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucl Acids Res 1994;32:1792–1797.
40. Shindyalov IN, Bourne P. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng 1998;11:739–747.
41. Holm L, Sander C. The FSSP database of structurally aligned protein fold families. Nucl Acids Res 1994;22:3600–3609.
42. Walle IV, Lasters I, Wyns L. SABmark: a benchmark for sequence alignment that covers the entire known fold space. Bioinformatics 2005;21:1267–1267.
43. Boutonnet NS, Rooman MJ, Ochagavia ME, Richele J, Wodak SJ. Optimal protein structure alignments by multiple linkage clustering: application to distantly related proteins. Protein Eng 1995;8:647–662.
44. Marti-Renom MA, Madhusudhan M, Sali A. Alignment of protein sequences by their profiles. Protein Sci 2004;1071–1087.
45. Fitzkee NC, Rose GD. Steric restrictions in protein folding: an α-helix cannot be followed by a contiguous β-strand. Protein Sci 2004;13:633–639.
46. Ramachandran GN, Sassiekharan V. Conformation of polypeptides and proteins. Adv Prot Chem 1968;28:283–437.
47. Liu S, Zhang C, Liang S, Zhou Y. Fold recognition by concurrent use of solvent accessibility and residue depth. Proteins 2007;68:636–645.