BIOINFORMATICS ORIGINAL PAPER Vol. 26 no. 4 2010, pages 470–477
doi:10.1093/bioinformatics/btp679

Sequence analysis

A novel method for accurate one-dimensional protein structure prediction based on fragment matching

Tuping Zhou†, Nanjiang Shu† and Sven Hovmöller*

Division of Structural Chemistry, Stockholm University, Stockholm SE-106 91, Sweden

Received on August 17, 2009; revised on November 9, 2009; accepted on December 4, 2009

Advance Access publication December 9, 2009

Associate Editor: Limsoon Wong

ABSTRACT

Motivation: The precise prediction of one-dimensional (1D) protein structure as represented by the protein secondary structure and 1D string of discrete state of dihedral angles (i.e. Shape Strings) is a prerequisite for the successful prediction of three-dimensional (3D) structure as well as protein–protein interaction. We have developed a novel 1D structure prediction method, called Frag1D, based on a straightforward fragment matching algorithm and demonstrated its success in the prediction of three sets of 1D structural alphabets, i.e. the classical three-state secondary structure, three- and eight-state Shape Strings.

Results: By exploiting the vast protein sequence and protein structure data available, we have brought secondary-structure prediction closer to the expected theoretical limit. When tested by a leave-one-out cross validation on a non-redundant set of PDB cutting at 30% sequence identity containing 5860 protein chains, the overall per-residue accuracy for secondary-structure prediction, i.e. Q3 is 82.9%. The overall per-residue accuracy for three- and eight-state Shape Strings are 85.1 and 71.5%, respectively. We have also benchmarked our program with the latest version of PSIPRED for secondary structure prediction and our program predicted 0.3% better in Q3 when tested on 2241 chains with the same training set. For Shape Strings, we compared our method with a recently published method with the same dataset and definition as used by that method. Our program predicted at 2.2% better in accuracy for three-state Shape Strings. By quantitatively investigating the effect of data base size on 1D structure prediction we show that the accuracy increases by ∼1% with every doubling of the database size.

Availability: The program is available for download at http://www.fos.su.se/∼nanjiang/Frag1D/download. Supplementary data are available at http://www.fos.su.se/∼nanjiang/Frag1D/supplement/suppl.html

Contact: svenh@struc.su.se

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Predicting protein structures from their amino acid sequences remains far from solved in spite of decades of efforts by researchers from various disciplines. This problem becomes increasingly important due to the widening gap between the known protein sequences and determined protein structures as deposited in the Protein Data Bank (PDB) (Berman *et al.*, 2002). The prediction of the secondary structure of proteins has long been considered as an important stage for three-dimensional (3D) structure prediction. Accurate prediction of secondary structure can improve the sensitivity of threading methods (Jones, 1999a) and is critical to many *ab initio* structure prediction methods (Bradley *et al.*, 2003). However, for on average ∼40% of all residues in random coils, the classical secondary-structure representation carries no structural information. On the other hand, the backbone protein structure is precisely described by a series of Φ/Ψ torsion angle pairs, one pair for each residue, due to the planarity of the peptide bond. The Φ/Ψ torsion angle pairs of protein structures are actually clustered into distinct regions. Therefore the backbone protein structure can be rather accurately described by a one-dimensional (1D) string of symbols representing the clustered regions of Φ/Ψ torsion angle pairs, called Shape Strings (Ison *et al.*, 2005). Shape Strings describe not only the conformations of residues in regular secondary-structure elements, e.g. shape A corresponds to regular α-helix (centered at $\Phi = -61°$, $\Psi = -41°$ on the Ramachandran plot) and shape S corresponds to regular β-sheet (centered at $\Phi = -116°$, $\Psi = 128°$ on the Ramachandran plot) (Hovmöller *et al.*, 2002). Shape Strings also classifies residues in random coils into several states thus containing much richer conformation. It has been shown that Shape Strings can be used for efficient searching for similar structures in a database (Shu *et al.*, 2008a) and the precise backbone structure can be reconstructed from Shape Strings (Gong *et al.*, 2005; Ison *et al.*, 2005).

Since the first protein structures were solved by X-ray crystallography, attempts have been made to predict the secondary structure of proteins as α-helix, β-sheet and random coil from their amino acid sequences. Chou and Fassman (1974) carried out the prediction based on simple statistics of the probabilities of each individual amino acid appearing at each of the three states, namely H (helix), S (sheet) and R (random coil). Later, sophisticated algorithms such as neural networks were employed and the prediction accuracy improved significantly (Kneller *et al.*, 1990; Rost and Sander, 1993). The accuracy of protein secondary-structure prediction jumped 5–10% by taking into account the evolutionary data (Rost and Sander, 1994) which were derived from large families of homologous sequences. Such sequence families can now be obtained in an automated manner by sequence searching programs

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First authors.

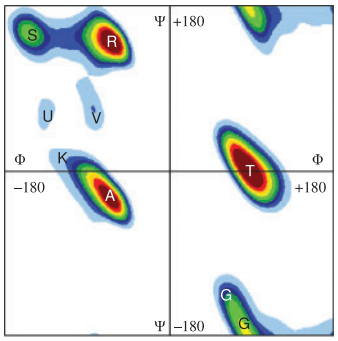

Fig. 1. The definitions of eight-state Shape Strings (S, R, U, V, K, A, T, G) on Ramachandran plot (Ison *et al.*, 2005). The typical Shapes for α-helices and β-sheets are A and S, respectively. Shape R represents the so-called polyproline type II structure. Shape K is often found at ends of helices or in 3_{10} helices. T denotes the turn region and G is special for glycine. Three-state Shape Strings are obtained by mapping S, R, U and V to S, K and A to H, T and G to T. The Ramachandran plot shown here is a montage from two plots; the left part shows the Ramachandran plot for all amino acids found in random coil, while the left half of the figure is that found for all Gly residues. Both are taken from Hovmöller *et al.* (2002) with permission.

such as PSI-BLAST (Altschul *et al.*, 1997). Recently developed methods (Dor and Zhou, 2007; Homaeian *et al.*, 2007; Jones, 1999b; Wood and Hirst, 2005) are almost without exception based on sequence profiles generated by PSI-BLAST. The Q3 (overall three-state per-residue accuracy) for those methods is approaching 80% and slightly better result may be obtained by combining several of these methods (Cheng *et al.*, 2007). Only recently, attempts to predict the conformation of the protein backbone in segments of random coil have also been made.

Bystroff *et al.* (2000) predicted 11-state Shape Strings with an overall MDA score of 58.8%, using a Hidden Markov Model (HMM). The MDA score is defined as the fraction of residues that are found in predicted eight-residue segments in which no predicted Φ/Ψ angle differs by more than 120° from the true structure. Kuang *et al.* (2004) predicted three-state Shape Strings with overall per-residue accuracy of 79.5% and for four-state Shape String, 78.4%, using Support Vector Machines (SVM). Note that slightly different definitions on how to discretize clustered regions of Φ/Ψ angle pairs on the Ramachandran plot have been used for these works [see the comparison of different definition in the review by (Shu *et al.*, 2008a)]. In this study, Shape Strings are defined according to Figure 1, if not especially mentioned.

We here present a novel method, called Frag1D, for 1D protein-structure prediction based on a straightforward segment matching. The basic idea of the fragment matching method is the same as the nearest-neighbor approach (Yi and Lander, 1993). Both approaches predict the secondary-structure state of the central residue of a test segment based on the secondary structure of high-scoring candidate segments from proteins with known structures. The difference is how these candidate segments are matched. For the nearest-neighbor approach, e.g. Yi and Lander's method, the segment similarity score is calculated based on a scoring table derived from local structural environment (Bowie *et al.*, 1991). Those secondary-structure prediction methods based on the nearest-neighbor approach generally predicted three-state secondary structure with Q3 ~70%. For our fragment matching method, candidate segments are selected by a profile–profile score [Equation (2)] derived from PICASSO score (Mittelman *et al.*, 2003) and the profile is created by taking the advantage of PSI-BLAST (Altschul *et al.*, 1997). Fragment matching approach has also been used in tertiary-structure prediction, such as Rosetta (Simons *et al.*, 1999). We for the first time applied segment matching method based on profile–profile scores to 1D structure prediction and obtained satisfactory results. A standard leave-one-out cross-validation on a non-redundant dataset of PDB chains cutting at 30% sequence identity shows that our method predicts 82.9% of all residues correctly as α-helix, β-sheet or random coils. For Shape String predictions, S3 (overall three-state Shape String per-residue accuracy) is 85.1% and S8 (overall eight-state Shape String per-residue accuracy) is 71.5%. Three-state Shape Strings are better predicted than three-state secondary structure. This is because the baseline for three-state Shape String prediction is higher than that for three-state secondary structure. The average abundance of the three secondary-structure states H, S and R are 38.1, 21.7 and 40.3%, respectively (Table 3). Therefore, a random guess of the secondary structure will yield $Q3 = (0.381^2 + 0.217^2 + 0.403^2) = 35.5\%$. For three-state Shape Strings, the average compositions for H (A + K), S (S + R + U + V) and T (T + G) are 51.7, 42.6 and 5.7%, respectively, and thus the S3 of a random guess is $(0.517^2 + 0.426^2 + 0.057^2) = 45.2\%$. It has long been a topic of discussion that the accuracy of secondary-structure prediction increases as the size of the database increases, even if the method has not been improved. We show here for the first time, that the accuracy of the secondary-structure prediction increases by ~1% with every doubling of the database, thus giving a quantitative answer to the question of the effect of the size of the database on the prediction accuracy.

2 MATERIALS AND METHODS

2.1 Database preparation

The dataset we used in this work was a non-redundant set of protein chains in PDB (as of June 2007) culled at 30% sequence identity by the PISCES server (Wang and Dunbrack, 2003), containing 5860 chains (1 480 756 amino acids). Out of these 5860 chains, two subsets, one with 4103 chains culled at 25% sequence identity, and another with 3255 chains at 20%, were also created (see Supplementary Material) to test the uniqueness and the size of the database on the prediction accuracy of the method. The three-state secondary structure (H: helix, S: sheet and R: random coil) of proteins was defined by converting the eight-state DSSP (Kabsch and Sander, 1983) definition with the scheme: H, G and I to H, B and E to S and the rest to R. The eight-state Shape String is defined according to Figure 1. The three-state Shape String was transformed from eight-state Shape String with the following scheme: S, R, U and V to S, K and A to H, T and G to T.

2.2 Profiles

Sequence profiles were obtained by running PSI-BLAST (Altschul *et al.*, 1997) against the NCBI nr database (July 2007) with three iterations and the *e*-value of 0.001. Moreover, for the protein sequences of the training set, we have enriched sequence profiles by structural profiles generated from blocks of Shape Strings (Ison *et al.*, 2005) (called FragAcc). Sequence and structural profiles were combined linearly according to Equation (1) by following the work of Teodorescu *et al.* (2004)

$$F_{ij} = (1 - weight) * Q_{ij} + weight * S_{ij} \quad (1)$$

where Q_{ij} is the sequence profile generated by PSI-BLAST, S_{ij} is the FragAcc and F_{ij} is the combined profile. The weight was set to 0.4 in this work. The effect of combining FragAcc in profiles is discussed in the discussion section.

The block of Shape Strings for a nine-residue fragment was built by searching the Shape String of this fragment in all other nine-residue long Shape Strings in the training set for similar Shape Strings. The corresponding amino acids of the Shape Strings in the block are used to build the substitution matrix [see the Supplementary Material in Shu *et al.* (2008b) for more details]. FragAcc carries the amino acid substitution information among similar local structures.

2.3 Structure prediction

A leave-one-out cross-validation procedure was carried out to evaluate the prediction. For each target chain, a sliding window of nine amino acids with their respective profiles, in this chain to be predicted, was searched among all the 1.48-million nine-residue segments in the other 5859 protein chains. At each position of a target sequence, the 100 segments with the highest profile–profile scores were kept, together with the accompanying PDB chain ID and positions in the sequence. The profile–profile score between two compared nine-residue segments was defined by

$$Score(\alpha,\beta)=\sum_{n=1}^{9}\left(\sum_{i=1}^{20}(\alpha_{ni}\log(\beta_{ni}/P_i)+\beta_{ni}\log(\alpha_{ni}/P_i))\right) \quad (2)$$

where α and β are profiles for the two compared nine-residue segments respectively and P is the background frequency for 20 standard amino acids. This profile–profile score was derived from PICASSO score (Mittelman *et al.*, 2003). For the target sequence, the profile was just the sequence profile generated by PSI-BLAST, while for candidate sequences, profiles were combined from sequence and structural profiles, i.e. FragAcc, according to Equation (1). These top 100 segments with highest profile–profile scores were further sorted by the weighted profile–profile score and only the top 10 were kept after re-sorting. The weighted profile–profile score is defined as

$$Score2(\alpha,\beta)=\sum_{n=1}^{9}\left\{Pinfo_n*\left(\sum_{i=1}^{20}(\alpha_{ni}\log(\beta_{ni}/P_i)+\beta_{ni}\log(\alpha_{ni}/P_i))\right)\right\} \quad (3)$$

where $Pinfo_n$ is the information score which is defined as

$$Pinfo_n=(1-\sum_{i=1}^{20}(X_{ni}*X_{ni})*(1-\sum_{i=1}^{20}X_{ni}*X_{ni})) \quad (4)$$

where $X_{ni}=(q_{ni}/p_i)/\sum_{i=1}^{20}(q_{ni}/p_i), i=1,2,3,\ldots,20, q_{ni}$ denotes the probability for amino acid i at position j in the profile, p_i is the background frequency for amino acid i. Equation (4) is empirical; the closer the profile is to the background composition, the larger the P_{info} score is. This score ranges from 0 to 0.90. Score2 [Equation (3)] was assigned to each of these selected segments.

Not all of these 10 selected nine-residue segments were used to predict the local structure of the query segment, nor were they used with equal weights. Albeit the dataset was culled at 30% (or 25 or 20%) sequence identity, homologues to the target chain may still exist in the training set and our Segment Matching Method can detect them very accurately [see a brief description of the Segment Matching Method for finding homologues below and see also Shu *et al.* (2008b) for the definition of homology score. Details about this method are described in the Supplementary Material]. The number of segments which were actually used for secondary structure and Shape String prediction depended on whether presumed homologue(s) were detected or not for the target chain. If a homologue to the target chain was predicted, only the top five segments were used for predicting the secondary structure, since the conformation of the selected segments were believed to be closer to the native conformation of the target protein to be predicted at that position. Otherwise the top 10 were used. Among these 5 or 10 segments actually used for local structure prediction, some may belong to the predicted homologues. Their scores [Score2 defined by Equation (3)] were multiplied by a factor between 1 and 3 based on the homology score (i.e. how sure we are that this is really a homologue to our target protein).

The probability for a residue of the target appearing at each state (H, S or R for the three-state secondary structure and S, R, U, V, K, A, G or T for the eight-state Shape Strings) was predicted as the sum of weighted scores of all matched segments with the state of the residue aligned at that position equaling that state. As mentioned above, if there are homologues detected for the target chain, the top five candidate fragments for each position are used for prediction, and otherwise the top 10 are used. Since a residue in a nine-residue target segment may be aligned to at most nine positions of a candidate segment, there are in total at most either 45 or 90 candidate segments aligned to a target segment depending on whether there are homologues predicted for this target chain or not (see Supplementary Fig. 1 for an example). The state with the highest probability was predicted as the secondary structure or Shape String state for that residue. In case of equal probability, the secondary structure was predicted in descending order as R, S and H, and the Shape String in the order of G, T, V, U, K, S, R and A. We have noted that S was often under-predicted. In order to remedy this, an empirical 3% probability score was added to the S state. The thus calculated probability for the residue to be predicted on each state was taken as the raw confidence of the prediction. However, the Q3, S3 and S8 were on average 5–10% better than this raw confidence. We thus normalized this raw confidence, such that for a prediction with a given confidence, one might on average expect the Q3, S3 and S8 accuracy to be the same as the confidence. The raw confidence was normalized by a linear function: $y = ax+b$, where x is the raw confidence and y is the normalized confidence. The parameters a and b were obtained by first plotting raw confidence against the real Q3, or S3 or S8, and then made a linear regression (see also Supplementary Fig. 3 for the relationship of raw confidence and real Q3 as well as the linear function).

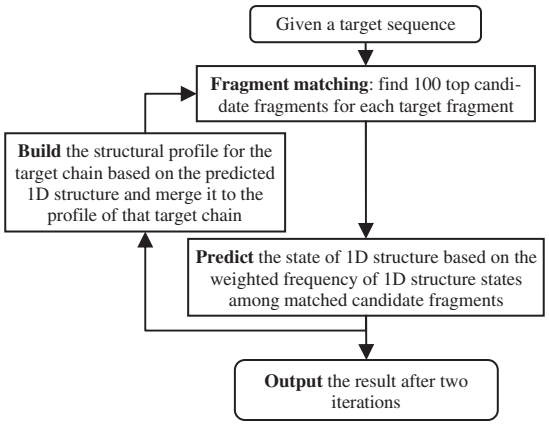

Fig. 2. Outline of the 1D structure prediction procedure.

After the prediction was made, we obtained the 1D structure of target chains with an expected high accuracy. Therefore, we can again build structural profiles for target chains from the predicted 1D structure and then enrich profiles of target chains by these structural profiles. A second round of prediction was thus carried out with the same setting as the first round, but now the profiles of target chains were enriched by structural profiles built from predicted 1D structure. The whole prediction procedure is outlined in Figure 2. In principle this procedure can be iterated many rounds until it converges. However, we noted that Q3 dropped already at the third round. This is most probably because the inaccuracy of structural profiles embedded in the predicted 1D structure accumulates quickly as the iteration procedure progresses and thus the gain by using such structural profiles is soon counteracted by the loss caused by the accumulated inaccuracy. Therefore, the final results were obtained from the second round.

The principle of the Segment Matching Method for homology detection is that among the tens of thousands of nine-residue segments (100 per each segment) with high profile–profile scores to the corresponding query fragment in the target protein chain, many are from a few candidate chains. Consider, for example, a 200 amino acid long protein chain; 192×100 = 19 200 segments will be selected. With nearly 6000 protein chains in the data

set, the average chain will be represented by about three segments. However, we very often found a few candidate chains having very many segments (from 10 to sometimes over 100) with high profile–profile scores to segments of the target chain. These candidate chains are potential homologues to the target chain. The positions of the fragments in the target sequence were plotted against the positions of the matched fragments in the candidate sequence on a two-dimensional (2D) dot-plot diagram. Only when these dots formed long consecutive lines were they a strong indication of homologous chains. A homology score is derived from the pattern of these 2D diagrams. More than 90% of the predicted homologues with a homology score larger than 30 were indeed (remote) homologues (Shu *et al.*, 2008b).

3 RESULTS AND DISCUSSIONS

3.1 Secondary-structure prediction

A solid leave-one-out cross-validation on 5860 non-redundant chains shows our program predicted three-state secondary structure with Q3 of 82.9% (Table 1). For 90% of all amino acids with the highest confidence, 86% are correctly predicted as being part of α-helix, β-sheet or random coil (Fig. 3). Note that these residues were identified only from their predicted confidence. This means that not only do we get a good overall score of Q3, but we have also identified quite well at which parts of the sequence the prediction is unsure.

Segment overlapping measure (SOV) score is considered as a more precise measure of the prediction of the secondary structure since it treats the secondary-structure elements as whole units (Rost *et al.*, 1994; Zemla *et al.*, 1999). For an α-helix predicted wrongly at every other residues, the Q3 will still be 50% but the prediction is actually meaningless since the whole α-helix at this location is missed. The SOV score will be zero for such a prediction since the prediction at this location is not identified as an α-helix. On the other hand, SOV neglects errors at the beginning or end of the α-helix and β-sheet if the core region is correctly predicted. Our program predicted these 5860 chains with an SOV score of 82.6% and Q3 of 82.9% (Table 1), which means our program not only predicted overall residues correctly but also the core regions of secondary structure elements correctly.

Many of the amino acids that are hard to predict are at the beginnings or ends of α-helices or β-strands (Fig. 4), having nearly equal indications for random coil or α-helix or β-sheet, respectively. In strong contrast, the nearest amino acids before or after are usually correctly predicted. The ends of α-helices and β-strands are often distorted. It is a matter of definition which secondary structure to assign for amino acids at these distorted regions. DSSP (Kabsch and Sander, 1983) puts a high weight on the hydrogen-bonding scheme for defining α-helix or β-sheet. In contrast, when Shape Strings (Ison *et al.*, 2005) are used for defining the conformations of amino acids, only the torsion angles of the polypeptide backbone are used. For Shape Strings, the Φ/Ψ torsion angles are clustered in eight regions (Fig. 1). For each conformation state, the standard deviations for torsion angles of amino acids are ~15–20° around the center of each region. As many as 47.2% of all α-helices are ended with an amino acid in K-shape, while most (~50.9%) of all β-strands are ended with an amino acid in R-shape. Indeed, for those

Table 1. Q3 and SOV for protein secondary-structure prediction on 5860 chains (numbers given in percentages)

	Helix	Sheet	Random coil	All
Q3	89.0	78.8	79.4	82.9
SOV[a]	89.7	81.9	76.4	82.6

[a]The normalized SOV was calculated based on a method introduced by Zemla *et al.* (1999).

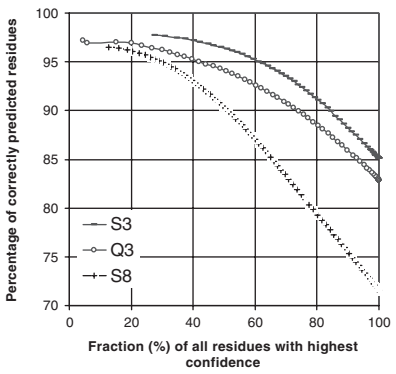

Fig. 3. Correctly predicted secondary-structure (Q3) and Shape Strings (S3 and S8) as a function of all residues above a certain confidence. For example, for ~80% amino acids predicted with highest confidence, Q3, S3 and S8 are roughly 88, 92 and 79%, respectively.

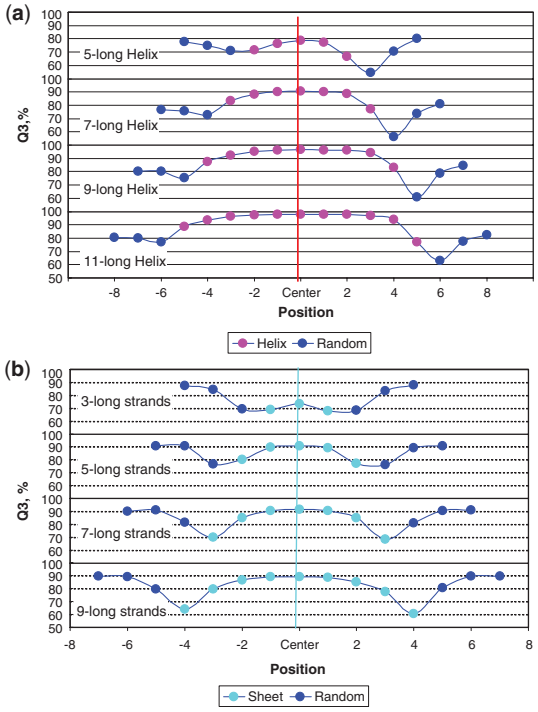

Fig. 4. The distribution of Q3 at the beginning, middle and end of (**a**) α-helices and (**b**) β-strands. For α-helices, the ends were predicted at lower Q3 accuracy than the beginnings while for β-strands, the beginnings and ends were equally well-predicted, resulting in a symmetric pattern. In general, the longer the α-helices or β-strands were, the more accurately their central residues were predicted. While short β-strands, i.e. three residues long, are hard to predict, the central amino acids in α-helices of average length (~10.5 amino acids) are correctly predicted in over 95% of the cases.

Table 2. Comparison between PSIPRED and our method Frag1D (numbers given in percentages)

		Helix	Sheet	Random coil	All
Q3	Frag1D	88.4	75.1	76.1	80.8
	PSIPRED (version 2.61)	86.1	72.8	79.1	80.5
SOV	Frag1D	88.3	78.6	73.5	80.4
	PSIPRED (version 2.61)	86.6	77.9	75.6	80.4

amino acids with wrongly predicted secondary-structure, many are predicted with correct Shape String.

We also benchmarked our method with one of the most successful secondary-structure predictors, PSIPRED (Bryson *et al.*, 2005) version 2.61. For the training set, we used the same 6598 chains that were used to build PSIPRED2.61 weighting files. For getting the testing set, all chains from PDB (as of June 2009) were filtered with the criteria that those chains with more than 30% sequence identity to any chain in the training set were removed. The remaining chains were further cut down to <30% sequence identity. This resulted in 2421 testing chains, with half of them (53%) submitted to PDB after 2007. See also the Supplementary Material for lists of training and test sets and a detailed description of how the testing set was generated. As shown in Table 2, our method predicted the secondary-structure 0.3% better than PSIPRED in Q3. A closer analysis of the 6598 chains in the training set shows that they have high redundancy. When cutting at 30% sequence identity, these 6598 chains were reduced to 3643. We carried out the secondary-structure prediction also for the same test set on this new training set of 3643 chains and obtained the same Q3, 80.8%. This means that the addition of many redundant, closely related sequences in a training set does not improve the secondary-structure prediction, at least not for our method. Although Frag1D predicted only 0.3% better Q3 compared to PSIPRED, it predicted 2.3% better for helices and sheets (Table 2). The fact that Frag1D and PSIPRED predicted differently on helices, sheets and random coils, may benefit consensus methods such as JPred (Cole *et al.*, 2008) which combine the results of other original, independent methods to take the merits of these two methods to obtain a higher accuracy. Frag1D is also computationally comparable to PSIPRED. It takes ∼10 min for Frag1D to predict a protein sequence with 300 amino acids running on a PC with 2 GHz CPU and 1 GB memory, while it takes ∼9.5 min for PSIPRED, given the current NCBI nr database (with 6.5-million sequences) and PSI-BLAST v2.2.17.

3.2 Shape string prediction

The 1D string of secondary-structure describes the protein backbone structure concisely, but for on average ∼40% of all residues locating in so-called random coils, the secondary-structure description carries no information about the conformation. The 'random coil' is an unfortunate wording, since there are many kinds of distinct conformations in this category. The Shape String concept describes accurately the conformations of all amino acids, including those in random coil. Thus, the prediction of Shape Strings makes it possible to build a tentative native 3D structure. Although each shape symbol represents a rather large area (Fig. 1), with torsion angles ϕ and ψ spreading in the order of $\pm 20^\circ$, many native local structure fragments of proteins with the same Shape Strings are quite similar in 3D (see Supplementary Fig. 2 for an example). This is probably because the conformation space for native structures of proteins is limited. Gong *et al.* (2005) also proved that it is possible to rebuild native protein conformation from highly approximated torsion angles grouped into 36 labelled, $60^\circ \times 60^\circ$ grid squares, each called a mesostate. Therefore, if the predicted Shape Strings are correct enough (since we have a reliable confidence estimation of the prediction, more accurately predicted chains can be identified), it is possible to build the backbone structure of the protein, or part of the protein. The proportions of the eight different shapes are, on average 45.2% A, 24.4% S, 16.1% R, 6.3% K and 4.4% T and just above 1% each of the three less common shapes; 1.3% U, 1.2% V and 1.2% G (Supplementary Table 1). The shapes of regular α-helices and β-strands are A and S, respectively. Although only 38.1% of all amino acids are in α-helices, 45.2% have A shape, because many individual amino acids in random coils also have A-shape, but they are not considered being α-helical unless at least four consecutive amino acids have A shape. Sometimes, three consecutive amino acids with A-shape are considered forming a 3_{10} helix, denoted G in DSSP. Many β-strands are distorted from the ideal S-shape. Thus, many β-strands contain one or more amino acids with R, U or V shape, although they are annotated as E-state in DSSP. The secondary structure can be defined by Shape Strings, using the scheme: (i) three or more consecutive A-shape is helix; (ii) two or more consecutive S-shape is sheet. The relationships between eight-state DSSP and eight-state Shape String are shown in Table 3. It is obvious that the DSSP and Shape String definitions are quite different.

Table 3. The relationship between the three-state DSSP and three-state Shape String

	DSSP			
Shape	Helix	Sheet	Random coil	Sum
Shape H (A + K)	37.8	0.7	13.2	51.7
Shape S (S + R + U + V)	0.1	20.8	21.7	42.6
Shape T (T + G)	0.2	0.2	5.3	5.7
Sum	38.1	21.7	40.3	100

All numbers are given as percentage of all combinations. For example, almost all amino acids being Helix or Sheet according to DSSP have H or S shape, respectively, but the reverse is not true. As many as half of the amino acids with S shape are actually found in stretches of random coil.

We predicted Shape Strings of all the 5860 protein chains, concomitantly with and in a similar way as the secondary-structure prediction. Obviously, the prediction of Shape Strings is on the one hand harder but on the other hand more informative, with eight rather than three categories. A random guess of secondary structure, given the condition that the proportions must be correct, i.e. 38.1% α-helix, 21.7% β-sheet and 40.3% random coil, results in 35.5% Q3 ($0.381^2 + 0.217^2 + 0.403^2$), but for eight-state Shape Strings, only 29.6% S8. We predicted secondary-structure at 82.9% Q3 and eight-state Shape Strings at 71.5% S8. An intermediate description is to define only three different shapes (A+K 51.7%, S + R + U + V 42.6% and T+G 5.7%). Here a random guess gives 45.2% S3, while we obtained 85.1% S3. Kuang *et al.* (2004) predicted three-state Shape Strings on a smaller dataset containing 1296 protein chains, cutting at ≤20% sequence identity and obtained 79.5% S3. The definition used by Kuang *et al.* is rather similar to that defined in Figure 1

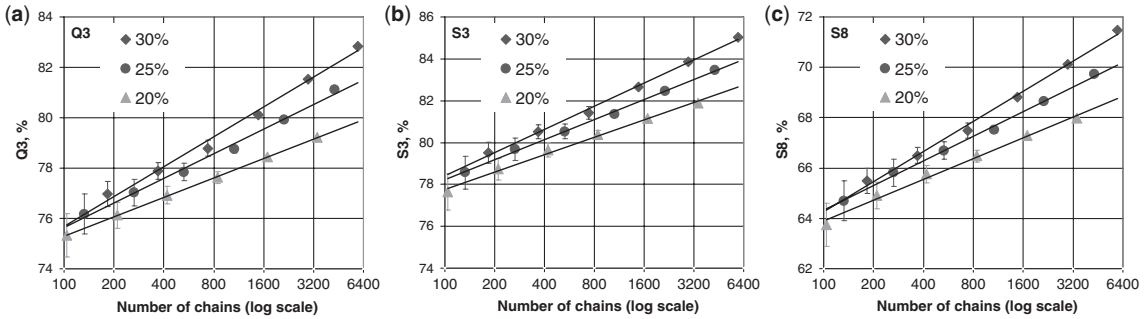

Fig. 5. The local structure prediction results (Q3, S3 and S8) increase by ∼1% for every doubling of the number of used non-redundant protein structures. The relation between (**a**) Q3, (**b**) S3, (**c**) S8 and the size of dataset. For a given number of protein chains used, the prediction decreases if stricter criteria are used for non-similarity between the used proteins. When all the 5860 protein chains with <30% sequence identity were used, Q3, S3 and S8 are 82.9, 85.1 and 71.5%, respectively.

Table 4. Results of Shape String prediction on 5860 chains (numbers given in percentages)

	At helix	At sheet	At random coil	Total
S3	94.7	89.9	72.6	85.1
S8	91.4	72.8	50.8	71.5

(agreed on 98.5% of all residues for three-state Shape Strings). Our method predicted 81.1% S3 on that dataset with our definition and 81.7% with Kuang's definition. This means our method Frag1D predicted three-state Shape Strings with 2.2% higher in S3. The relation between three-state Shape Strings and three-state secondary-structure from DSSP is shown in Table 3. For 80% of amino acids with highest confidence, the three- and eight-state Shape Strings were predicted at 92% S3 and 79% S8, respectively (S3 and S8 in Fig. 3).

3.3 Improved predictions with larger data sets

To investigate the effects of the sequence identity cutting level and the size of the dataset on Q3, S3 and S8, the 5860 chains having ≤30% sequence identity were further culled at 25% (4227 chains) and 20% (3338 chains) sequence identity levels (Wang and Dunbrack, 2003). For each of these data sets, the chains were divided randomly into two groups with the same number of chains. These two subgroups were further divided into 4, 8, 16 and 32 equal groups. Predictions were made for each subgroup. As shown in Figure 5, every doubling of the number of sequences used for the training set leads to on average ∼1% increase in Q3, S3 and S8.

3.4 Comparison of the predicted secondary structure and Shape String

The secondary-structure prediction predicts the protein backbone structure as a set of α-helices, β-sheets and random coils. For the Shape String prediction, the protein backbone is predicted as a set of discretized torsion angles, one symbol for each residue. The main advantage of the Shape String prediction is that, unlike secondary structure prediction, it predicts the conformation for every amino acid. The S3 and S8 for Shape Strings at random coils is 72.6 and 50.8%, respectively (Table 4).

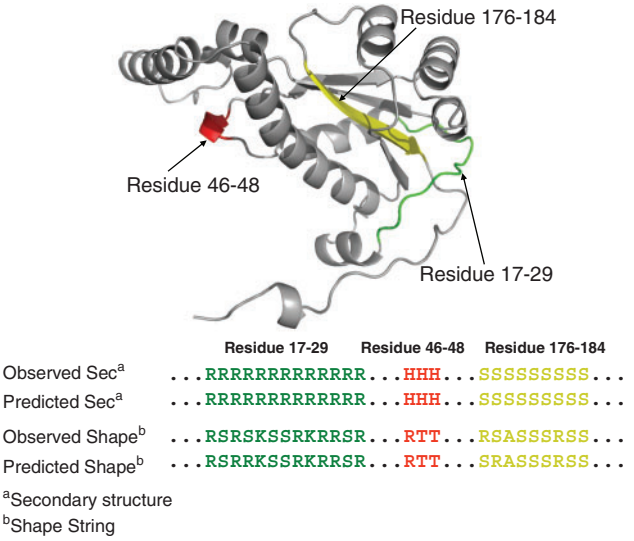

Fig. 6. An example (chain 1NOXA) for predicted secondary structure and predicted eight-state Shape Strings. The chain 1NOXA was predicted at 90% Q3 based on the leave-one-out cross-validation on 5860 chains, which is better than the average Q3 of 82.9%. However, ∼20% of all chains were predicted at better Q3 than this example. Our program predicted both the secondary structure and eight-state Shape String correctly at three segments highlighted in the figure. Note that for the random coil segment (residues 17–29), the Shape String is actually quite rich in conformation but our program predicted such detailed conformation correctly. The illustration of the 3D structure of 1NOXA was drawn by PyMOL (DeLano, 2002).

Moreover, in the secondary-structure description, a unique symbol denotes α-helices or β-sheets. This means α-helices and β-sheets are treated as straight rods and strands, but in reality the torsion angles for α-helices or β-sheets are not always the same throughout the backbone. In contrast, Shape Strings depict the possible distortions within α-helices or β-sheets (see an example in Fig. 6) and thus may facilitate the 3D structure modelling from predicted 1D structure.

The existence of remote homologues (if detected), even at a dataset cutting at 30% (or 25 or 20%) sequence identity, do improve the prediction of secondary structures and Shape Strings. As shown in Table 5, although the Q3 for all 5860 chains cutting at 30% sequence identity level was 82.9%, the Q3 for those 4194 chains

Table 5. The prediction on those chains with and without homologues predicted, for datasets cutting at 30, 25 and 20% sequence-identity level, respectively

Dataset	≤30%			≤25%			≤20%		
Homologue[a]	Yes	No	All	Yes	No	All	Yes	No	All
No. of chains	4194	1666	5860	2394	1833	4227	1327	2011	3338
Q3 (%)	84.7	77.1	82.9	83.2	77.6	81.1	81.1	77.8	79.2
S3 (%)	86.6	80.2	85.1	85.2	80.6	83.5	83.5	80.7	81.9
S8 (%)	73.0	66.6	71.5	71.4	66.9	69.7	69.3	66.9	68.0

We used 'predicted homologues' instead of 'real homologues' as defined by SCOP (Andreeva *et al.*, 2004) to show the effect of the existence of potential homologues on prediction results because for one thing, about half of all 5860 chains are not annotated in the latest SCOP database, making it impossible to do the statistics; and for another, only the knowledge of 'predicted homologues' can be obtained for unknown structures.
[a]In the row of homologue: 'Yes' means for chains with at least one homologue predicted; 'No' means for chains with no homologue predicted; 'All' means for all chains.

of which at least one homologue was predicted, was as high as 84.7%. For the other 1666 chains, for which no homologue was predicted in the training set, the Q3 was only 77.1%. Q3 decreases as the dataset is cut at lower sequence identity level. For the chains with homologues predicted, Q3 drops from 84.7 to 83.2 to 81.1% as the sequence identity cutting level is lowered from 30 to 25 to 20%. This is because fewer homologues exist and they are also becoming more distantly related as the sequence identity cutting level becomes more restricted. However, for those chains without any homologues detected, Q3 varies insignificantly for datasets cutting at different sequence levels. For details, see Table 5. The results also show that for really hard cases where no homologues exist, our method can predict the secondary structure at an average Q3 of ∼77.5%. Note that for most published works tested on so called non-redundant datasets cutting at either 30% sequence identity or zero HSSP distance, many homologues still exist.

3.5 How to use the potential homologues in 1D structure prediction?

As mentioned in 'Materials and methods' section, different numbers of candidate segments were used to predict the conformation state of the target depending on whether there were predicted homologues for the target or not. One may argue that the 1D structure can be directly predicted from homology modelled structures if a homologue exists and can be detected. However, we observed that the secondary-structures predicted by our method were significantly more accurate than those directly generated from homology modelled structures. For 50 randomly selected chains predicted at Q3 varying from 75 to 90%, the average Q3 for our method Frag1D was 85.8% but the Q3 for the secondary-structure generated from homology modelling was only 72.8%. The homology modelling was carried out by MODELLER9v6 (Marti-Renom *et al.*, 2000) with default settings (the list of these 50 randomly selected chains and templates used to build models can be found in the Supplementary Material). The structure variations between remote homologues are usually quite big so that their secondary structures do not agree very well. However, the secondary-structure prediction based on segment matching predicts the conformation of local structures from a broad range of similar local structures and may thus recover the variation between individual homology pairs.

Table 6. Comparison of predictions using sequence profiles (Q_{ij}) and enriched profiles (F_{ij}) (numbers given in percentages)

	Q3	S3	S8
Q_{ij} profile	82.0	83.4	69.5
F_{ij} profile	82.9	85.1	71.5

3.6 Incorporation of structural information in profiles

Sequence profiles built from large families have been used extensively for the effective detection of homologues and structurally similar local structures (Eddy, 1996; Madera and Gough, 2002; Rangwala and Karypis, 2005; Sadreyev and Grishin, 2003; Soding, 2005). Therefore, profiles were used to search for candidate segments instead of the amino acid sequence alone in our methods, just as in most other recently developed methods. However, for a protein whose existing family contains too few sequences or are strongly biased, the profile built from that sequence family tend to be so poor that true homologues cannot be detected. Moreover, for the segment matching method, all candidate segments with similar local structures to the target can be used. It is not necessary that the whole protein is homologous to the target chain. Therefore, we enriched the sequence profile by FragAcc which were derived from blocks of similar Shape Strings. As shown in Table 6, the average Q3, S3 and S8 for all 5860 chains increased by ∼1, 1.6 and 2%, respectively, by using enriched profiles.

4 CONCLUSIONS

We have presented a new method for predicting 1D protein structures, i.e. secondary structures and Shape Strings, based on a straightforward segment matching. Our method predicted protein secondary-structure at 82.9% Q3 when tested on 5860 chains, a non-redundant set of PDB, cutting at 30% sequence identity. It also predicted Shape Strings with 85.1% S3 (overall three-state Shape String per-residue accuracy) and 71.5% S8 (overall eight-state Shape String per-residue accuracy) on the same dataset. At this level of accuracy, the predicted secondary-structures, together with predicted Shape Strings will be very helpful for tools to build 3D structure models. By performing our program on a series of evenly divided datasets, we showed that the accuracy of the secondary-structure prediction increases by ∼1% with every doubling of the database size. Since Shape Strings describe not only the conformation of regular secondary-structure elements, but also random coils in detail, the predicted Shape Strings might be used as a powerful starting point for 3D structure modeling. This is our further goal.

ACKNOWLEDGEMENTS

The authors thank Prof. David Jones for providing the training set for PSIPRED2.61.

Funding: Calidris, Sweden.

Conflict of interest: none declared

REFERENCES

Altschul,S.F. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Andreeva,A. *et al*. (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res.*, **32**, D226–D229.

Berman,H.M. *et al*. (2002) The Protein Data Bank. *Acta Crystallogr. D Biol. Crystallogr.*, **58**, 899–907.

Bowie,J.U. *et al*. (1991) A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, **253**, 164–170.

Bradley,P. *et al*. D. (2003) Rosetta predictions in CASP5: successes, failures, and prospects for complete automation. *Proteins*, **53**(Suppl. 6), 457–468.

Bryson,K. *et al*. (2005) Protein structure prediction servers at University College London. *Nucleic Acids Res.*, **33**, W36–W38.

Bystroff,C. *et al*. (2000) HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins. *J. Mol. Biol.*, **301**, 173–190.

Cheng,H. *et al*. (2007) Consensus Data Mining (CDM) Protein secondary structure prediction server: combining GOR V and fragment database mining (FDM). *Bioinformatics*, **23**, 2628–2630.

Chou,P.Y. and Fasman,G.D. (1974) Prediction of protein conformation. *Biochemistry*, **13**, 222–245.

Cole,C. *et al*. (2008) The Jpred 3 secondary structure prediction server. *Nucleic Acids Res.*, **36**, W197–W201.

DeLano,W.L. (2002) The PyMOL Molecular Graphics System on World Wide Web.

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

Eddy,S.R. (1996) Hidden Markov models. *Curr. Opin. Struct. Biol.*, **6**, 361–365.

Gong,H. *et al*. (2005) Building native protein conformation from highly approximate backbone torsion angles. *Proc. Natl Acad. Sci. USA*, **102**, 16227–16232.

Homaeian,L. *et al*. (2007) Prediction of protein secondary structure content for the twilight zone sequences. *Proteins*, **69**, 486–498.

Hovmöller,S. *et al*. (2002) Conformations of amino acids in proteins. *Acta Crystallogr. D Biol. Crystallogr.*, **58**, 768–776.

Ison,R.E. *et al*. (2005) Proteins and their shape strings. An exemplary computer representation of protein structure. *IEEE Eng. Med. Biol. Mag.*, **24**, 41–49.

Jones,D.T. (1999a) GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *J. Mol. Biol.*, **287**, 797–815.

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

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

Kneller,D.G. *et al*. (1990) Improvements in protein secondary structure prediction by an enhanced neural network. *J. Mol. Biol.*, **214**, 171–182.

Kuang,R. *et al*. (2004) Protein backbone angle prediction with machine learning approaches. *Bioinformatics*, **20**, 1612–1621.

Madera,M. and Gough,J. (2002) A comparison of profile hidden Markov model procedures for remote homology detection. *Nucleic Acids Res.*, **30**, 4321–4328.

Marti-Renom,M.A. *et al*. (2000) Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.*, **29**, 291–325.

Mittelman,D. *et al*. (2003) Probabilistic scoring measures for profile-profile comparison yield more accurate short seed alignments. *Bioinformatics*, **19**, 1531–1539.

Rangwala,H. and Karypis,G. (2005) Profile-based direct kernels for remote homology detection and fold recognition. *Bioinformatics*, **21**, 4239–4247.

Rost,B. and Sander,C. (1993) Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.*, **232**, 584–599.

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

Rost,B. *et al*. (1994) Redefining the goals of protein secondary structure prediction. *J. Mol. Biol.*, **235**, 13–26.

Sadreyev,R. and Grishin,N. (2003) COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance. *J. Mol. Biol.*, **326**, 317–336.

Shu,N. *et al*. (2008a) Describing and comparing protein structures using shape strings. *Curr. Protein Pept. Sci.*, **9**, 310–324.

Shu,N. *et al*. (2008b) Prediction of zinc-binding sites in proteins from sequence. *Bioinformatics*, **24**, 775–782.

Simons,K.T. *et al*. (1999) Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins*, **34**, 82–95.

Soding,J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics*, **21**, 951–960.

Teodorescu,O. *et al*. (2004) Enriching the sequence substitution matrix by structural information. *Proteins*, **54**, 41–48.

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

Wood,M.J. and Hirst,J.D. (2005) Protein secondary structure prediction with dihedral angles. *Proteins-Struct. Funct. & Bioinformatics*, **59**, 476–481.

Yi,T.M. and Lander,E.S. (1993) Protein secondary structure prediction using nearest-neighbor methods. *J. Mol. Biol.*, **232**, 1117–1129.

Zemla,A. *et al*. (1999) A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment. *Proteins-Struct. Funct. Genet.*, **34**, 220–223.