MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **65** (2011) 5-20

ISSN 0340 - 6253

An Information Theoretic Approach to Secondary Structure Assignment

Mahnaz Habibi[a,b], Changiz Eslahchi[a,1], Hamid Pezeshk[c], Mehdi Sadeghi[d]

a: Faculty of Mathematics, Shahid-Beheshti University, Tehran, Iran,
b: Bioinformatics Research Group, School of Computer Science ,Institute for Research in fundamental sciences (IPM), Tehran, Iran,
c: School of Mathematics, Statistics and Computer Sciences and Center of Excellence in Biomathematics, College of Science, University of Tehran, Tehran, Iran,
d: National Institute for Genetic Engineering and Biotechnology, Tehran, Iran.

(Received April 15, 2009)

Abstract

The automatic assignment of the protein secondary structure from three dimensional coordinates is an essential step in the characterization of protein structure. Although the recognition of secondary structure elements as alpha helices and beta sheets seem straightforward but there are many different definitions, each regarding different criteria. We introduce a new algorithm for the protein secondary structure assignment based on a number of geometric parameters and by using the entropy. A sequence is partitioned to segments. Then the secondary structure elements are assigned to each of these segments. It is shown that if the entropy of a segment increases then the regularity in the structure decreases. So it is concluded that the concept of entropy could be used as a measure of regularity of the secondary structure.

1 Introduction

Pauling and Corey predicted the existence of regular segments in the structure of protein [1]. The experimental determination of three-dimensional structure of proteins has confirmed the presence of these regular secondary structures. These regular secondary structures are classified into two classes: helix (which comprises three states: α-helix, 3_{10}-helix, π-helix) and β-strand. Five decades later, Berman realized that half of the residues in proteins participate in helices or sheets [2]. Pauling and Corey incorrectly predicted that 3_{10}- helices would not occur

[1] Corresponding author. E-mail address: ch-eslahchi@sbu.ac.ir Tel:+98-21-22431652

in proteins due to unfavorable bond angles [1]; although, approximately 4% of these residues are observed in this conformation [3].

Since the secondary structures present a simple and intuitive description of 3D structures, they are widely employed in a number of structural biology applications. They are used for protein structure classification [4, 5], better sequence alignment [6-8], comparative modeling and threading [9-12]. They also provide a natural framework for structure visualization [13,14]. Thus in order to achieve these tasks, precise assignment of secondary structures is required from three-dimensional atomic coordinates of proteins.

In fact, crystallographers assigned the secondary structure of proteins by eye from their 3D structures. At first, it was the only way for assigning secondary structures. However, there were occasionally disagreements among experts. To overcome this problem, the programs for assigning secondary structure automatically, were needed. There are a number of methods to perform automatic assignment of secondary structures.

One of the main criteria used for secondary structure assignment is hydrogen-bonding pattern. Pauling established the hydrogen bond as an important principle in chemistry [15]. One method in this class is Dictionary of Secondary Structure of Proteins called DSSP [16]. DSSP performs sheet and helix assignments solely based on backbone-backbone hydrogen bonds. DSSP still remains as one of the most important programs for secondary structure assignment. One of the commonly used software related to DSSP is the secondary STRuctural IDEntification method (STRIDE) by Frishman and Argos [17]. This program uses an empirically derived hydrogen bond energy and phi-psi torsion angle criterion to assign secondary structure.

Other methods use geometric criteria for identification of secondary structures. The geometric features employed are various. The algorithm DEFINE is developed based on C_α distances [18].

It has been shown that all the three methods assign similar secondary structure only to 63% of residues.

Other methods have been developed to use different criteria to assign secondary structures. P-CURVE is based on a mathematical analysis of protein curvature [19]. PSEA only uses C_α atoms and is based on distance and angle criteria [20]. VOTAP employs the concept of Voronoi tessellation [21] and KAKSI uses C_α distances and (φ /ψ) angles [22].

In this work, we introduce a new method based on geometric features for secondary structure assignment. This method is implemented in a program called PSE (Protein Segmentation with Entropy). We use geometry of consecutive residues and assign a sequence of two codes to the residue of the sequence. Then using new sequence, the entropy values are assigned to all segments in a protein. To check the performance of our method, we compare the results of secondary structure assignments obtained by DSSP, STRIDE, PSEA and PSE with some standard assignments of PDB. Furthermore, we show that our method, by using the entropy value assigned to each segment, gives more information about helix and β-strand geometry. This leads to a more accurate secondary structure assignment.

2 Material and Method

The assignment of secondary structure by PSE is based on distances, torsion angles of backbone and entropy.

2.1 *Dataset*

Representative set of X-ray protein structures with resolution <1.7 Å was gathered from PDB by using advanced search in RCSB (http://www.rcsb.org). The structures with more than 40% similarities in sequences were excluded. Taking these criteria into account, 1988 proteins were selected for this dataset.

2.2 *The parameters*

The regularities of secondary structure of a protein and the geometry of its backbone are used as criteria to produce some parameters. In this section, we introduce the torsion angles and distances between consecutive α-carbons in a protein.

Let A be a protein consisting of n residues with numeral labels $1,2,...,n$. Let s_i denote the coordinate of the α-carbons of the i-th residue. For each residue i, $1 \le i \le n-6$ of protein A, we associate a distance vector $\overrightarrow{D_i} = (\overrightarrow{d_1^i}, \overrightarrow{d_2^i}, \overrightarrow{d_3^i}, \overrightarrow{d_4^i}, \overrightarrow{d_5^i})$ and for each residue i, $7 \le i \le n$ of protein A, we set $\overrightarrow{D_i} = (\overrightarrow{d_1^i}, \overrightarrow{d_2^i}, \overrightarrow{d_3^i}, \overrightarrow{d_4^i}, \overrightarrow{d_5^i})$, where $\overrightarrow{d_j^i}$, $j=1,2,3,4,5$ is the distance between s_i and s_{i+j} ,and $\overrightarrow{d_j^i}$ is the distance between s_i and s_{i-j}. Let $\vec{ij}$ be the vector from α-carbon of residue i to the α -carbon of residue j. To each residue i, $1 \le i \le n-3$ we assign a triple of angles

Table 1: The mean and the standard deviations of C_α -distance and torsion angles obtained from the database.

	Structure	$\overline{d_1}$	$\overline{d_2}$	$\overline{d_3}$	$\overline{d_4}$	$\overline{d_5}$	$\overline{\phi_1}$	$\overline{\phi_2}$	$\overline{\phi_3}$
Mean	α-helix	3.80	5.475.30	5.65	6.85	8.92	92.6°	92.6°	40.6°
	π-helix	3.81	5.65	6.11	8.02	10.1	91.2°	91.2°	54.8°
	3_{10}-helix	3.81	6.60	5.53	6.72	8.50	82.5 °	82.5 °	52.8°
	β-strand	3.79		9.24	11.4	12.2	131°	131°	132°
Standard deviation	α-helix	10^{-4}	0.35	1.22	1.62	1.68	8.2°	8.2°	51.6°
	π-helix	10^{-4}	0.25	1.19	1.58	1.98	6.6°	6.6°	52.8°
	3_{10}-helix	10^{-4}	0.27	1.27	1.63	1.87	7.3 °	7.3 °	56 °
	β-strand	10^{-4}	0.53	1.31	2.26	3.15	14.5°	14.5°	49.1°

$\overrightarrow{TA_i} = (\overrightarrow{\phi_1^i}, \overrightarrow{\phi_2^i}, \overrightarrow{\phi_3^i})$, where $\overrightarrow{\phi_1^i}$ is the angle between the vectors $\overrightarrow{(i+1)i}$ and $\overrightarrow{(i+1)(i+2)}$; $\overrightarrow{\phi_2^i}$ is the angle between the vectors $\overrightarrow{(i+2)(i+1)}$ and $\overrightarrow{(i+2)(i+3)}$, and $\overrightarrow{\phi_3^i}$ is defined by

$$\overrightarrow{\phi_3^i} = \frac{\overrightarrow{(i+2)(i+3)}.(\overrightarrow{i(i+1)} \times \overrightarrow{(i+1)(i+2)})}{|\overrightarrow{(i+2)(i+3)}.(\overrightarrow{i(i+1)} \times \overrightarrow{(i+1)(i+2)})|}\gamma_{i3} \quad (1)$$

where '.' and '×' are inner and outer products, respectively. γ_{i3} denotes the angle between the plane passing through from three points s_i, s_{i+1}, s_{i+2} and the plane passing through from three points s_{i+1}, s_{i+2} and s_{i+3}. We also assign a triple of angles $\overrightarrow{TA_i} = (\overrightarrow{\phi_1^i}, \overrightarrow{\phi_2^i}, \overrightarrow{\phi_3^i})$, to each residue i, $3 \leq i \leq n$, where $\overrightarrow{\phi_1^i}$ is the angle between $\overrightarrow{(i-1)i}$ and $\overrightarrow{(i-1)(i-2)}$; $\overrightarrow{\phi_2^i}$ and $\overrightarrow{\phi_3^i}$ are defined in a similar fashion as discussed above.

2.3 *Distributions of distance and torsion angle*

The main goal of this section is to determine some intervals that with some good probabilities, the unknown parameters, $\overline{d_j}$, *j=1, 2, 3, 4,5* and $\overline{\phi_j}$, *j=1, 2, 3*, would be located in these intervals. By considering the random samples of proteins in dataset, we obtain the estimates of parameters ($\overline{d_j}$ and $\overline{\phi_j}$) for residues corresponding to one of the four secondary structure categories: α-helix, 3_{10}-helix, π-helix and β-strand. The means and the standard deviations of these distributions are shown in table 1. For each secondary structure categories, the parameters $\overline{d_j}$ (*j=1, 2, 3, 4, 5*) and $\overline{\phi_j}$ (*j=1, 2, 3*) for consecutive residues are obtained. It is noticeable that the

Table 2: Periodic features of regular secondary structures.

Structure	α-Helix	310-Helix	π-Helix	β-Strand
Number of Cα atoms per turn	3.6	3	4.4	2
T	3.6	3	4.4	2
Ω	2π/3.6	2π/3	2π/4.4	Π

statistical distributions are different. For each secondary structure categories only one parameter ($\overrightarrow{d_1}$) has almost constant distribution. Therefore, it is concluded that the mean of categories are equal to d_1's.

Using Anderson-Darling test (with p-values less than 0.1), it is concluded that for all categories of structures the random variable $\overrightarrow{d_2}$ has a normal distribution. In the same fashion, it is concluded that $\overrightarrow{\phi_1}$ (or $\overrightarrow{\phi_2}$) has also a normal distribution. Using

$$[\overline{X} - Z_{0.005}\frac{S}{\sqrt{n}}, \overline{X} + Z_{0.005}\frac{S}{\sqrt{n}}] \quad (2)$$

we define 99.5% confidence intervals for $\overrightarrow{d_2}$, $\overrightarrow{\phi_1}$ and $\overrightarrow{\phi_2}$, in each category. Note that $Z_{0.005}$=2.58 is a point from standard normal distribution, Z, for which $p(Z>Z_{0.005})=0.0025$. $\overline{X}$ and *S* are, respectively, the mean and the standard deviation of the random sample of size *n*.

We also use a geometric based approach to obtain some other intervals for $\overrightarrow{d_3}$, $\overrightarrow{d_4}$, $\overrightarrow{d_5}$ and $\overrightarrow{\phi_3}$.

General parametric equation for an ideal helix is defined by:

$$r(t) = (R\sin wt, R\cos wt, Bt) \quad (3)$$

where *R* is the radius of hypothetical base area of the ideal helix, *w* is the angular frequency of oscillation (see table 2) and *B* is the pitch of the helix.

Therefore:

$$\overrightarrow{d_j^t} = \sqrt{2R^2(1-\cos((j-t)w) + (j-t)^2B^2}\,. \quad (4)$$

So $\overrightarrow{d_{t+1}^t} = \overrightarrow{d_1} = \sqrt{2R^2(1-\cos w) + B^2} = 3.81$. Therefore

$$B^2 = 14.5161 - 2R^2(1-\cos w)\,. \quad (5)$$

Table 3: Intervals of distance and torsion angle in helices and β-strands.

	α-helix	3_{10}-helix	π-helix	β-strand
$\overrightarrow{d_2}$	[5.10, 5.85]	[5.30, 6.0]	[5.0, 5.60]	[6.10, 7.10]
$\overrightarrow{d_3}$	[3.65, 6.52]	[1.37, 5.98]	[5.59, 7.10]	[9.43, 10.74]
$\overrightarrow{d_4}$	[4.0, 8.44]	[0.0, 5.29]	[8.17, 9.92]	[12.20, 14.20]
$\overrightarrow{d_5}$	[6.44, 11.12]	[0.0, 6.72]	[9.90, 12.20]	[15.42, 17.80]
$\overrightarrow{\phi_1}$	[87.8°, 97.4°]	[80°, 85°]	[85.8°, 95.4°]	[120°, 142°]
$\overrightarrow{\phi_2}$	[87.8°, 97.4°]	[80°, 85°]	[85.8°, 95.4°]	[120°, 142°]
$\overrightarrow{\phi_3}$	[36.5°, 63.4°]	[31°, 59°]	[50.4°, 78.3°]	[160°,228°]

According to (4), (5) and intervals obtained for $\overrightarrow{d_2}$, we can find other intervals, for $\overrightarrow{d_j}$, *j=3, 4, 5*. Using (5) and intervals obtained for $\overrightarrow{\phi_1}$ and $\overrightarrow{\phi_2}$, we obtain the intervals for $\overrightarrow{\phi_3}$, for each secondary structure categories of helices.

It should be noted here that, β-strand can be considered as a helix with two C_α atoms in each turn (i.e., the period of C_α atom for β-strand is 2); therefore, using the above procedure, we can find the intervals corresponding to β-strand. In table 3, we present the intervals of distances and torsion angles in helices and β-strand.

2.4 *Protein segmentation algorithm*

Let's suppose that we receive a message W of length n, from an alphabet set $A=\{A_1, A_2, \ldots, A_N\}$. Let the frequency of A_i in W be $|A_i|$, $1 \le i \le N$. The entropy of sequence W is given by

$$H(W)=\sum_{i=1}^{N} -\frac{|A_i|}{n} \log(\frac{|A_i|}{n}). \quad (6)$$

By using entropy, we introduce an algorithm for partitioning the sequence of amino acids of a protein. Using intervals shown in table 2 the functions $F_A(i)$ and $\tilde{F}_A(i)$ are defined by

$$F_A(i)=\begin{cases} o & \overrightarrow{d_i} \text{ or } \overleftarrow{d_i} \in I^H \text{ or } \overrightarrow{\phi_i} \text{ or } \overleftarrow{\phi_i} \in J^H \\ 2 & \text{otherwise} \end{cases} \qquad \tilde{F}_A(i)=\begin{cases} 1 & \overrightarrow{d_i} \text{ or } \overleftarrow{d_i} \in I^S \text{ or } \overrightarrow{\phi_i} \text{ or } \overleftarrow{\phi_i} \in J^S \\ 2 & \text{otherwise} \end{cases}$$

where I^H and J^H are intervals for helices and I^S and J^S are intervals for β-strand. Sequences A_0 and A_1 are defined by $A_0 = \{F_A(i)\}_{i=1}^n$, $A_1 = \{\tilde{F}_A(i)\}_{i=1}^n$. We would like to find subsequences of A_0 and A_1 with entropy of at most 0.26 and with the maximum length.

Let S be a sequence of $\{0,2\}$. Assume that 0 is the most frequent letters of S. The sequence is called 0-regular if:

1) the two ends of S are 0's,
2) for each subsegment S' of S, which its ends are 0's, we have $H(S') \le 0.26$.

Similarly, we can define 1-regular sequence if S is a sequence of $\{1,2\}$. In our algorithm, the 0-regular and 1-regular segments are considered as helix and β-strand, respectively. In order to assign secondary structure to a protein, we should obtain the maximum 0-regular segments of A_0 and the maximum 1-regular segments of A_1.

For this purpose, let $A_0 = a_1,\ldots,a_n$ and $A_1 = b_1,\ldots,b_n$. Define two graphs $G(A_i)$, i=0, 1 as

$V(G(A_i)) = \{(a_j, a_{j+1}, a_{j+2}, a_{j+3}) \mid 1 \le j \le n-3\}$,

$E(G(A_i)) = \{uv \mid u = (a_j, a_{j+1}, a_{j+2}, a_{j+3}),\ v = (a_{j+1}, a_{j+2}, a_{j+3}, a_{j+4}),\ H(u) \le 0.26, H(v) \le 0.26\}$.

It is obvious that $G(A_i)$, i=0,1 is a union of disjoint paths. In the following theorem, we show that each path of graph has the entropy value of at most 0.26.

Theorem 1

Let $u = u_j,\ldots,u_{k+j-3}$ be the path of $G(A_i)$ for which $u_j = (a_j, a_{j+1}, a_{j+2}, a_{j+3})$ and $u_{k+j-3} = (a_{k+j-3}, a_{k+j-2}, a_{k+j-1}, a_{k+j})$. If $a_j = a_{k+j} = i$ then $a_j,\ldots,a_{k+j}$ is i-regular segment of A_i.

Proof: Suppose $i = 0$, $u = u_j,\ldots,u_{k+j-3}$ and $a_j = a_{k+j} = 0$ such that for each $j \le l \le k+j-3$, $H(u_l) \le 0.26$. Let S be the segment constructed by $a_j, a_{j+1}, \ldots, a_{k+j}$. It is obvious that between any consecutive 2's, in segment S, there exists at least three 0's (otherwise there is a t, such that $H(u_t) > 0.26$). Therefore the maximum number of 2's appear in S is $\left\lceil \frac{k-1}{4} \right\rceil$. Now,

$$H(S) \le f(k) = -\frac{\left\lceil \frac{k-1}{4} \right\rceil}{k+1} \log \frac{\left\lceil \frac{k-1}{4} \right\rceil}{k+1} - \frac{\left\lfloor \frac{3k+5}{4} \right\rfloor}{k+1} \log \frac{\left\lfloor \frac{3k+5}{4} \right\rfloor}{k+1}.$$

But the maximum of $f(k)$ occurs at $k=6$ and $f(6)=0.2597$. For the case of $i=1$, the proof is the same as for $i=0$. □

The following algorithm finds the paths with the maximum length in the graph. As we discussed before, these paths are the secondary structure segments of the protein.

2.5 *Identification of turns*

Due to the shortness of the length of turns, and the similarity of them with helices, identification of the segments of a protein as turns is more complicated. In DSSP, a minimal size of helix is set to have two consecutive hydrogen bonds, leaving out single helix hydrogen bonds, which are assigned as turns. But by using the distance vector $\overrightarrow{D_i}$ we can clearly distinguish between turns and helices. According to distance vector, we consider three consecutive residues as a turn, if the sum of entries of the following matrix is less than *25*.

$$\begin{bmatrix} |d_1^i - 3.8| & |d_1^i - 5.6| & |d_1^i - 5.3| & |d_1^i - 5.4| & |d_1^i - 6.8| \\ |d_1^{i+1} - 3.8| & |d_2^{i+1} - 5.4| & |d_2^{i+1} - 7.5| & |d_2^{i+1} - 10| & |d_2^{i+1} - 10.3| \\ |d_3^{i+2} - 3.8| & |d_3^{i+2} - 6| & |d_3^{i+2} - 9.5| & |d_3^{i+2} - 10.6| & |d_3^{i+2} - 14| \end{bmatrix}.$$

2.6 *Validation criteria*

In order to study the relative performance of the PSE algorithm against existing algorithm, we need to define criteria that determine the agreement between our method and the existing algorithms. We use a known measures as $S_n=(TP/(TP+FN))$ and $S_P=(TN/(TN+FP))$.

These measures are based on the relation between the number of residues correctly assigned positive (TP), the number of residues correctly assigned negative (TN), the number of residues incorrectly assigned positive (FP), the number of residues incorrectly assigned negative (FN).

There is a combination of two parameters called correlation coefficient (CC). This is shown by:

$$CC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TN + FP)(TP + FN)(TP + FP)(TN + FN)}}. \quad (7)$$

Pseudo code of the PSE algorithm
Input: C_α coordination in 3 dimensional apace.
Step1://produce two distance vectors for each residue, except the boundary points
For point $i \leftarrow 1$ to point $n-6$
$\overrightarrow{D_i}$ = Get distance vector from the Cartesian coordinates of C_α.
For point $i \leftarrow n$ to point 7
$\overrightarrow{D_i}$ = Get distance vector from the Cartesian coordinates of C_α.

Step2:// produce two torsion angle vectors for each residue, except the boundary points
For point $i \leftarrow 1$ to point $n-3$
$\overrightarrow{TA_i}$ = Get triple angle vector from the Cartesian coordinates of C_α.
For point $i \leftarrow n$ to point 4
$\overrightarrow{TA_i}$ = Get triple angle vector from the Cartesian coordinates of C_α.
Step3: using $F_A(i)$ and $\tilde{F}_A(i)$, produce sequences A_0 and A_1.
Step4: // to each sequence A_0 and A_1, we perform this step
$start \leftarrow -1$
$end \leftarrow -1$
For window $i \leftarrow 1$ to window $n-4$
if $H(a_i, a_{i+1}, a_{i+2}, a_{i+3}) > 0.26$ AND $H(a_{i+1}, a_{i+2}, a_{i+3}, a_{i+4}) \le 0.26$
if $a_i = 2$
$start \leftarrow i+1$
elseif $a_{i+1} = 2$
$start \leftarrow i+2$
elseif $a_{i+2} = 2$
$start \leftarrow i$
elseif $H(a_i, a_{i+1}, a_{i+2}, a_{i+3}) \le 0.26$ AND $H(a_{i+1}, a_{i+2}, a_{i+3}, a_{i+4}) > 0.26$
if $a_i = 2$
$end \leftarrow i+2$
elseif $a_{i+1} = 2$
$end \leftarrow i$
elseif $a_{i+2} = 2$
$end \leftarrow i+1$
if $end > start$
add a regular segment from start to end to the list HELIXLIST.

Agreement (A), is defined as the number of residues for which both methods agree (TP+TN), divided by the total number of residues.

$$A = \frac{TP + TN}{TP + TN + FN + FP} \tag{8}$$

Table 4: Comparison of helix assignment results obtained by PSE and other methods.

Method	TP	TN	FP	FN	Total Residues	A%	Sensitivity%	Specificity%	CC
PSE _DSSP	*153107*	*313813*	*26156*	*21534*	*514610*	*90.73*	*87.67*	*92.31*	*0.79*
PSE_ STRIDE	*158223*	*313205*	*21040*	*22142*	*514610*	*91.61*	*87.72*	*93.70*	*0.82*
PSE_PSEA	*159386*	*323636*	*19877*	*11711*	*514610*	*93.86*	*93.15*	*94.21*	*0.86*

Table 5: Comparison of strand assignment results obtained by PSE and other methods.

Method	TP	TN	FP	FN	Total Residues	A%	Sensitivity%	Specificity%	CC
PSE _DSSP	*72421*	*334473*	*55970*	*51747*	*514610*	*79.06*	*58.32*	*85.66*	*0.43*
PSE_ STRIDE	*72984*	*331949*	*55407*	*54270*	*514610*	*78.68*	*57.35*	*85.70*	*0.42*
PSE_PSEA	*85433*	*343706*	*42958*	*42513*	*514610*	*83.39*	*66.77*	*88.90*	*0.56*

3 Results and discussion

3.1 *Comparison of secondary structure assignment methods*

Comparing the methods of protein secondary structure assignment is not a trivial task. Since there is not any standard way for the comparison of several methods, so it is difficult to show that one method is better than the other. Therefore, in order to check the validity of our method, we compare it with assignments performed by DSSP, STRIDE, and PSEA. We also compare the results of DSSP, STRIDE, PSEA and PSE with the assignments reported by PDB file as a standard test set.

In table 4 and 5, we show the results of the comparison between our results and other methods (DSSP, STRIDE, and PSEA). We find that there are a strong agreement between PSE and other methods for helix assignment, but in table 5, the agreements between PSE and DSSP and STRIDE for β-strand and coil are not higher than agreements between the PSE and these methods for helix assignments.

It is the consequence of the fact that we use geometric and structural criteria. Hence, we find few β-strands that do not participate in the hydrogen bonds with other β-strands. While hydrogen bonds are used by many methods (DSSP, STRIDE) for defining the elements of secondary structure. The β-sheet residues are defined as either having two hydrogen bonds in the β-sheet, or being surrounded by two hydrogen bonds in the β-sheet. For this reason, the β-strand segments are longer than β-sheet segments.

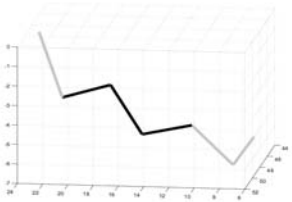

Figure 1. A β-strand obtained by PSE. Note that the bold segment is reported as a strand in a β-sheet in STRIDE and DSSP.

Table 6: Comparison of turn assignment results obtained by PSE and other methods.

Method	TP	TN	FP	FN	Total Residues	A%	Sensitivity%	Specificity%	CC
PSE_DSSP	*16328*	*489326*	*3347*	*5608*	*514610*	*98.26*	*74.43*	*99.32*	*0.77*
PSE_STRIDE	*18102*	*492558*	*1573*	*2377*	*514610*	*99.23*	*88.39*	*99.68*	*0.89*

For example, the figure 1 shows the comparison between the β -strand assignment which reported by PSE and one reported by DSSP and STRIDE. All residues in segment of *Thermosulfurigenes* (PDB code '1A0C', residues 226 to 232) are assigned as a β-strand by PSE, while the residues 227 to 230 in this segment are reported in STRIDE and DSSP as β-sheet.

Finally, we have compared our results on turn structures with the assignments performed by other methods. Among the mentioned methods, only DSSP and STRIDE assign turn structures. So we compare our results with the turn assignments performed by DSSP and STRIDE in table 6. Table 6 shows that there is a strong agreement between the results of our assignments with those of DSSP and STRIDE.

The assignments performed by the crystallographers in PDB files are the most popular and frequently used assignments. Therefore, in order to check the validity of PSE, in tables 7 and 8, we compare the assignments reported by PDB files as the standard test set. Although many of crystallographers identify secondary structure based on DSSP algorithm, these tables show that the highest agreement between PDB and PSE is in the helix assignments. However, the PDB and the STRIDE have good agreement in the β-sheet assignments.

In figure 2, we present three illustrative examples to point out differences in five assignment schemes. Most similarity between these methods occurs in the core of segments. The disagreements among methods occur in terminals of segments.

3.2 *Protein geometry analysis by using entropy*

A systematic analysis of geometries of proteins was first reported by Barlow and Thornton [23]. They found that not all α-helices in proteins have the same geometry and they differ in their shapes. Some of the α-helices are linear (normal α-helix) whereas some of them have distortion in their shapes. Most of the current methods of secondary structure assignments are only able to assign secondary structures to residues, and they do not give any information about the differences between the geometry of α-helices. The PSE gives more information about the geometry of α-helices. Distortions cause the residues in helices not to be in α-helix intervals and these distortions could be shown by difference codes. In fact, if deviation of each residue happens in helix structure, it is shown in two parameters, the distance vector and the triple of angles. Therefore, the entropy value of each regular segment calculated by PSE can be used as a parameter to describe the geometry of α-helix at the protein. On the other hand, if distortions happen in helix structures, then the entropy value related to helix will increase. For example in figure 3A, all residues have the same codes, and the entropy value of this segment is equal to zero, thus we expect that the geometry of this helix should be nearly the same as normal α-helix and the figure 3B shows the segment of helix assigned in PSE with different entropy value (0.08).

Table 7 Comparison of helix assignment results reported by PDB and other methods.

Method	TP	TN	FP	FN	Total Residues	A%	Sensitivity%	Specificity%	CC
PSE _PDB	*165271*	*318449*	*9370*	*21520*	*514610*	***93.99***	*88.47*	***97.14***	***0.87***
DSSP_PDB	*161342*	*314520*	*13299*	*25449*	*514610*	*92.47*	*86.37*	*95.94*	*0.84*
STRIDE_PDB	*164931*	*312385*	*15434*	*21860*	*514610*	*92.75*	*88.29*	*95.29*	*0.84*
PSEA_PDB	*159741*	*322889*	*11356*	*20624*	*514610*	*93.78*	***88.56***	*96.60*	*0.86*

Table 8 Comparison of strand assignment results reported by PDB and other methods.

Method	TP	TN	FP	FN	Total Residues	A	sensitivity	specificity	CC
PSE _PDB	*78432*	*346144*	*49959*	*40075*	*514610*	*82.50*	*66.18*	*87.38*	*0.52*
DSSP_PDB	*81848*	*353784*	*42319*	*36659*	*514610*	*84.65*	*69.06*	***89.31***	*0.57*
STRIDE_PDB	*83922*	*352771*	*43332*	*34585*	*514610*	***84.86***	***70.83***	*89.06*	***0.58***
PSEA_PDB	*79925*	*348082*	*48021*	*38582*	*514610*	*83.17*	*67.44*	*87.88*	*0.54*

```
PDB code: 1BZY
Residue number 67.....................................................................143
PDB            SCCCHHHHHHHHHHHHHHHCCCCCCCCSSSSSSSSSCCCCCCCCCCCSSSSCHHHHHHCCCSSSSSSCCCCCCHHHH
STRIDE         STTHHHHHHHHHHHHHHHHCCCCCCCSSSSSSSSSSSTTSSSSSSSSSTTTHHHTTTTSSSSSSSSSTTTHHHHH
DSSP           STTTHHHHHHHHHHHHHHTTSSCCCCSSSSSSSSSSSSTTSSSSSSSSSCHHHHHHTTCSSSSSSSSSSHHHH
PSE            CCCHHHHHHHHHHHHHHHHHCSSSSSSSSSSSSSSSSTTTCCCSSSSSSHHHHHHHHHHSSSSSSCCTTTHHHH
PSEA           CCCCHHHHHHHHHHHHHHHHHHCCCCCSSSSSSSSSSSSCCCCCCCCCCCCCCCCCCCCCSSSSSSSCCCCCHHHH

Residue number 67.............................179
PDB            HHHHHHHHCCCSSSSSSCSSSCCCCCCCCCCCCC
STRIDE         HHHHHHHHCCTTSSSSSSSSSCTTTTCCCCTTSSS
DSSP           HHHHHHHTTCCSSSSSSSSSSSCCTTCCCCCCSSSS
PSE            HHHHHHHHHTTSSSSSSSSSSSTTTCCSSSSSSSS
PSEA           HHHHHHHCCCCSSSSSSCCCCCCCCCCCCCCCCC
```

```
PDB code: 1E8Y

Residue number 143....................................................................219
PDB            CHHHHHHHHHHHHHHHHCCCCCCCCCCCHHHHHHHHHHHHHHHHHHHHHHHHHHHCCCCCCCCHHHHHHCCCCCC
STRIDE         CCHHHHHHHHHHHHHHCCTTTTTTTTCCHHHHHHHHHHHHHHHHHHHHHCCHHHHHCCCHCCCCCCHHHHHTTTTTC
DSSP           CCHHHHHHHHHHHHHHTHCTTCCTTCHHHHHHHHHHHTHHHHHHHHHHCHHHHHCCCHCCHCCCHHHHTTCHHHC
PSE            CHHHHHHHHHHHHHHHHTTTTTTTCCCHHHHHHHHHHHHHHHHHHHHHHHHHHHCEETTTCCCHHHHHHHTTTTE
PSEA           CCHHHHHHHHHHHHHHCCCCCCCCCCCCHHHHHHHHHHHHHHHHHHHHCCHHHHHCEEEECCCCCCCCCCCCCCCE

Residue number 220..........................253
PDB            CCCCCCCCCCCCCCCCCCCCHHHHHHHHHHHHH
STRIDE         EEEEEEETTEEECCEETTTTCHHHCHHHHHHTTT
DSSP           EEEEEEETTEECCCEECTTCCHHHTHHHHHHHHC
PSE            EEEEEEEEEEEEEEEEEECHHHHHHHHHHHHHC
PSEA           EEEEEEECCEEEEEEEECCCCHHHHHHHHCCCCC
```

```
PDB code: 1AOC

Residue number 20.....................................................................86
PDB            CCCCHHHHHHHHHHHHHHHHHCCCCCCCCCHHHCCCCHHHHHHCCCCCCCCCHHHHCHHHCCCCCCCCCCCSSSSCCC
STRIDE         CCCCHHHHHHHHHHHHHHHHHHHCTTTTCCHHHHTTTHHHHHHHCCCTTTTCHHHHCHHHTTTTCCTTTCCCSSSSSSC
DSSP           SCCCHHHHHHHHHHHHHHHHTTCSSTTSHHHTTTSHHHHHHCCCCTTTCCHHHHSHHHSTTCCCCSSCCCSSSSCCC
PSE            SCCCHHHHHHHHHHHHHHHHHHHCCCCTTHHHHHTTHHHHHHCCCCTTTTCCCHHHHHHHHTTTSSSSSSSSSSSSSS
PSEA           CCCCHHHHHHHHHHHHHHHHHHHHCCCCCCCCCCCCCCHHHHHCCCCCCCCCCCCCCCCCCCCCCCCCCSSSSSSSSC
```

Figure 2. Comparison of assignments obtained by PDB, STRIDE, DSSP, PSEA and PSE preformed on three proteins. Here H represents the helix class, S represents the β-strand class T represents the turn class and C represents the coil class.

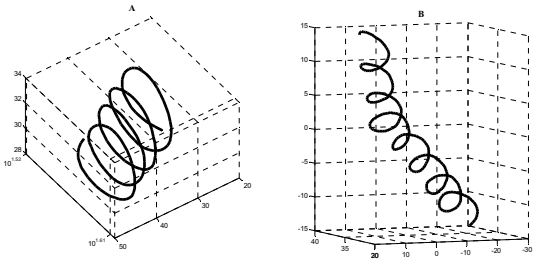

Figure 3. Segments of helices with various entropies: (A) and (B) indicate that two helices with different shapes have different entropies. A normal helix with zero entropy is shown in (A), and (B) shows a distortion helix with entropy of 0.08.

Similar to the calculation of entropy value for normal helix, we can calculate the entropy value for each β-strand segment reported in PSE. Figure 4 shows two segments of β-strands assigned by PSE with different entropies. As shown by the figure, when the entropy increases, then the regularity in the structure decreases. So the concept of entropy could be used as a measure of regularity of the secondary structure.

We also find that there are strong agreement between PSE and other methods for normal helix or β-strand assignments. Figure 5a shows the α-helix assigned by PSE with entropy value of 0.09 on *2new insights into mechanisms of transcriptionalcontrol* (PDB code '1HLO' residues, 2130 to 2343). A comparison with PDB assignments reveals that this segment matches with the assignment of PDB file. It is also interesting that the α-helix reported by other methods is shorter than α-helix obtained by PSE. In fact, there is a deviation in this segment. The entropy value demonstrates this deviation. In figure 5b we also show the segment obtained by PSE on the *human hgprtase with transition state inhibitor* (PDB code '1BZY', residues 70 to 87) as a normal helix with entropy value zero. The figure illustrates that this segment has regular structure, while PDB file do not consider the whole the segment as a helix. The main discrepancies between the assignments are observed at the terminal ends of the segment.

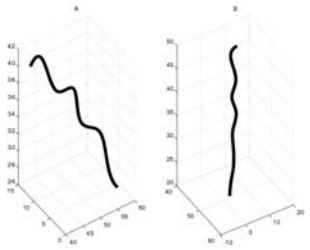

Figure4**.** Segments with various entropies: (A) and (B) indicate that two β-strands with different shapes have different entropies (H (A) =0.21 and H (B) = 0).

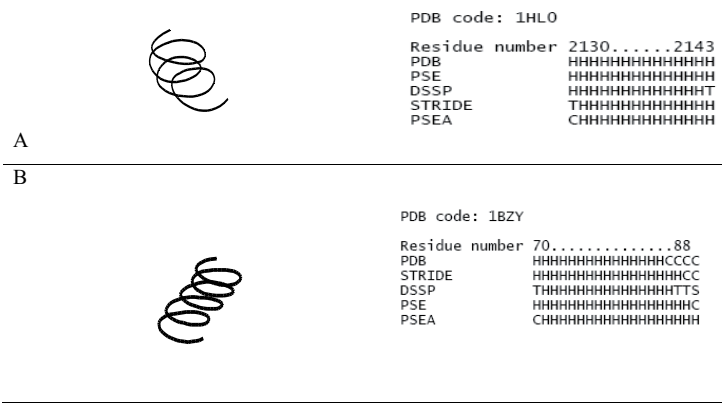

Figure 5. Two segments performed by PSE as a normal helix, and comparison of it with other method. The main discrepancies between these methods are observed at terminals end of the secondary structure.

Acknowledgments

Mahnaz Habibi and Changiz Eslahchi are grateful to the Faculty of Mathematics of Shahid Beheshti University. Hamid Pezeshk would like to thank the Department of Research Affairs of University of Tehran. We are very grateful to anonymous referee for careful reading and valuable comments and suggestions. This research was in part supported by a grant from IPM (No. CS1385-1-02).

References

[1] Pauling, L., Corey, R.B., 1951. The pleated sheet, a new layer configuration of polypeptide chains. Proc. Natl. Acad. Sci. USA. 37, 251-256.

[2] Berman, H.M., Westbrook, J., Feng, Z., Gilllliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E., 2000. The Protein Data Bank. Nucleic Acids Res 28, 235–42.

[3] Andersen, C.A.F., 2001. Protein structure and the diversity of hydrogen bonds. The Technical University of Denmark Ph.D. Thesis.

[4] Murzin, A.G., Berner, S.E., Hubbard, T., Chothia, C., 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. 247, 536-540.

[5] Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B., Thornton, J.M., 1995. CATH- a hierarchic classification of protein domain structures. Structure 5, 1093-1108.

[6] Fischel-Ghodsian, F., Mathiowitz, G., Smith, T.F., 1993. Alignment of protein sequences using secondary structure: a modified dynamic programming method. Protein Eng. 3, 577-581.
[7] Henneke, C.M., 1989. A multiple sequence alignment algorithm for homologous proteins using secondary structure information and optionally keying alignments to functionally important sites. Comput. Appl. Biosci. 5, 141-150.
[8] Smith, R.F., Smith, T.F., 1992. Pattern-induced multi-sequence alignment (PIMA) algorithm employing secondary structure-depentent gap penalties for use in comparative protein modeling. Protein Eng. 5, 35- 41.
[9] Fischer, D., Eisenberg, D., 1996. Fold recognition using sequence-derived properties. Protein Sci. 5, 947-955.
[10] Kolinski, A., Rotkiewoicz, P., Ilkowski, B., Skolnik, J., 1999. A method for the improvement of threading- based protein models. Proteins 37, 592-610.
[11] Rice, D.W., Eisenberg, D., 1997. A 3D-1D substituation matrix for protein fold recognition that include predicted secondary structure of the sequence. J. Mol. Biol. 267, 1026-1038.
[12] Russell, R.B., Copley, R.R., Barton, G.J., 1996. Protein fold recognition by mapping predicted secondary structures. J. Mol. Biol. 259, 349-365.
[13] Sayle, R.A., Milner-White, E.J., 1995. RASMOL: biomolecular graphics for all. Trends. Biochem. Sci. 20, 374.
[14] Humphrey, W., Dalke., A., Schulten, K., 1996. VMD: visual molecular dynamics. J. Mol. Graph. 14, 33-38.
[15] Pauling, L., 1939. The Nature of the Chemical Bond. New York. Cornell University Press.
[16] Kabsch, W., Sander, C., 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen- bonded and geometrical features. Biopolymers 22, 2577-2637.
[17] Frishman, D., Argos, P., 1995. Knowledge-based protein secondary structure assignment. Proteins 23, 566-579.
[18] Richards, F.M., Kundrot, C.E., 1988. Identification of structural motifs from protein coordinate data: secondary structures and first-level supersecondary structure. Proteins 3, 71-84.
[19] Sklenar, H., Etchebest, C., Lavery, R., 1989. Describing protein structure: a general algorithm yielding complete helicoidal parameters and a unique overall axis. Proteins 6, 46-60.
[20] Labesse, G., Colloch, N., Pothier, J., Mornon, J.P., 1997. P-SEA: a new efficient assignment of secondary structure from C alpha trace of proteins. Comput. Appl. Biosci. 13, 291-295.
[21] Dupuis, F., Sadoc, J.F., Mornon, J.P., 2004. Protein secondary structure assignment through Voronoi tessellation. Proteins 55, 519-528.
[22] Martin, J., Letellier, G., Marin, A., Taly, J.F., Brevern, A.G., Gibrat, J.F., 2005. Protein secondary structure assignment revisited: a detailed analysis of different assignment methods. BMC Struct.Biol. 5, 17.
[23] Barlow, D., Thornton, J.M., 1998. Helix geometry in proteins. J.Mol.Biol. 201, 601-619.