Helix segment assignment in proteins using fuzzy logic

Shahriar Arab[1], Farzad Didehvar[2], Changiz Eslahchi[3], Mehdi Sadeghi[4*]

[1]Department of Bioinformatics, Institute of Biochemistry and Biophysics, University of Tehran, P.O. Box 13145-1384, Tehran, I.R. Iran [2]Institute for Studies in Theoretical Physics and Mathematics (IPM), Niavaran Square, P.O. Box 19395-5746, Tehran, I.R. Iran [3]Faculty of Mathematical Sciences, Shahid Beheshti University, Evin, Tehran, I.R. Iran [4]Department of Biochemistry, National Institute of Genetic Engineering and Biotechnology, P.O. Box 14155-6343, Tehran, I.R. Iran

Abstract
The automatic assignment of protein secondary structure from three dimensional coordinates is an essential step in the characterization of protein structure. Although, the recognition of secondary structures such as alpha-helices and beta-sheets seem straightforward, but there are many different definitions, each regarding different criteria. We have developed a new algorithm for protein helix assignment, by using fuzzy logic based on backbone torsion angles. In this method, each residue takes a number from 0 to 100 that indicates the helical membership degree of that residue. This method can be converted to a classical method whenever we assume that any residue with a membership degree greater than 83 is a helix. Comparison of the results with structures reported in protein data bank (PDB), dictionary of secondary structure of proteins (DSSP) and structure identification (STRIDE) for 324 proteins indicate that our algorithm works as well as DSSP showing 93% agreement. We believe that the fuzzy secondary structure assignment has more advantages than the other classical approaches used for protein structure comparisons and alignments.
Keywords: Protein structure; Secondary structure assignment; Fuzzy logic.

INTRODUCTION

The automatic assignment of protein secondary structure from three dimensional coordinates is an essential step in the characterization of protein structure. The secondary structure assignment plays an important role in structural genomics. The secondary structure segments are used in protein structure classification (Pearl *et al*., 2005; Andreeva *et al*., 2004; Hogue and Bryant, 1998), protein structure alignment (Sternberg *et al*., 1999; Marti-Renom *et al*., 2000; Sauder *et al*., 2000), comparative modeling and threading (Rost, 2000; Rice and Eisenberg, 1997; Kolinski *et al*., 1999; Xu *et al*., 1999), and also influence sequence alignment (Smith and Smith, 1992; Fischel-Ghodsian *et al*., 1993; Henneke, 1989). Although, the recognition of secondary structure such as alpha-helices and beta-sheets seem straightforward, there are still many different definitions, each regarding different criteria.

The main criteria used in secondary structure assignment are hydrogen bonding patterns known as dictionary of secondary structure of proteins (DSSP) (Kabsch and Sander, 1983), quantification of the back bone curvature (Richards and Kundrot, 1988), inter-c_α distances (Levitt and Greer, 1977) and combination of hydrogen bond energy and torsion angle information known as structure identification (STRIDE) (Frishman and Argos, 1995). Comparing these methods on a protein database showed only 63% agreement between the se three algorithms (Colloc'h *et al.*, 1993). Although, different methods may assign different secondary structure states to each residue, but they are similar in one aspect; each residue is defined in one state and we finally have a string of secondary structure states for the protein sequence. Despite the similarity between an assigned state such as the alpha-helix in different parts of a protein or different proteins, these structures are not exactly the same (Barlow and Thornton, 1988). For example, two alpha-helices with the same length in two different proteins may not have the exact geometrical similarity,

*Correspondence to: **Mehdi Sadeghi**, Ph.D.
Tel: +98 21 44580373; Fax: +98 21 44580399
E-mail: sadeghi@nrcgeb.ac.ir

but in the assignment methods this difference is not considered, since most of the protein structure comparison methods are based on secondary structure alignment, renouncement of their geometrical differences leads to an inexact three-dimensional comparison. Thus, it is necessary to define parameters for secondary structures so that different and similar structures can be compared more precisely. In this study, we use fuzzy logic and assign a membership degree to each residue by considering the geometry of consecutive residues with Phi and Psi angles that indicate regular or irregular turns for consecutive residues. These fuzzy numbers may vary from 0 to 100 and can be used to compare two helices for a better similarity or difference.

The exclusive use of backbone torsion angles is not sufficient for assignment of all the secondary structure elements, however, helices' geometry has enough information for detection of helices. Although the algorithm presented in this article is solely based on dihedral angles, results show that the assigned fuzzy numbers identify helical regions of protein structure as good as other classical methods.

MATERIALS AND METHODS

Representative set of X-ray and NMR protein structures with resolutions better than 2.5Å and without chain breaks were gathered from the protein data bank (PDB) based on the PDBSELECT list for proteins, with less than 25% sequence similarity. 324 proteins with 48644 amino acids were selected. These are listed in Table 1.

Alpha-helices assigned by PDB were chosen as standard assignment. Backbone dihedral angles (φ and Ψ) of each residue were taken as in DSSP. From a mathematical point of view, Δ and Δ^2 are approximations of the first and second derivatives. Since our fuzzy algorithm is based on the geometrical structure of helices, and first and second derivatives are tools for studying the plot of a structure, we therefore used Δφ and $\Delta^2\varphi$, Δψ and $\Delta^2\Psi$. To assign a helix fuzzy number to each residue, the following steps were carried out:

1. On all amino acids in the data set, Δφ, Δψ, $\Delta^2\varphi$ and

Table 1. Protein Data Bank (PDB) codes of the Data Set.

1a02N	1ezvA	1ig3A	1sfcA	4sgbE	1ul7A	1ykgA	1byqA	1dqeA	1ep0A	1gakA	1if1A	1qg7A
1aohB	1f4nA	1irdA	1sfcB	5cytR	1umqA	2asyA	1bywA	1dqgA	1eteA	1gd7A	1im0A	1qleA
1aoiA	1fl7A	1jcqA	1t3jA	1chvS	1uphA	2axlA	1c1kA	1dqiA	1ew0A	1gd8A	1irdB	1qmtA
1avyA	1fltV	1jnmA	1tafB	1ci5A	1ussA	2azvA	1c3mA	1ds1A	1excA	1gl2A	1irjA	1qsoA
1be3A	1g2cA	1jqcA	1tvxA	1cirA	1ut3A	1a2kA	1c5fA	1dxmA	1eypA	1gl2B	1j4xA	1qtoA
1bh8A	1g64A	1jqlA	1ty0A	1cixA	1uvfA	1a5oA	1c7kA	1e30A	1f2dA	1gnhA	1j75A	1quqA
1bmqA	1gcqA	1ktzA	1uixA	1cjgA	1uw0A	1afrA	1c94A	1e3kA	1f46A	1gr3A	1j90A	1sknP
1c8uA	1gd2E	1ljpA	1ur6A	1cl3A	1uw2A	1aihA	1c9iA	1e44A	1f5vA	1h6wA	1j9lA	1tafA
1cl7L	1gk4A	1llmC	1urqA	1gccA	1uzcA	1aohA	1cc8A	1e6iA	1fa2A	1hciA	1jejA	1tc3C
1cxzA	1gmjA	1ln1A	1urqB	1gd4A	1v06A	1aq4A	1cnoA	1ebuA	1fi2A	1hwwA	1jfiA	1tgxA
1dazA	1gqzA	1lqvA	1v54A	1gh1A	1v1cA	1atlA	1cqmA	1ec5A	1fl7B	1hxrB	1jgsA	1tiiD
1df4A	1gu4A	1mspA	1vkkA	1gh5A	1v1dA	1avoA	1cqxA	1ecsA	1flkA	1hziA	1ji6A	1tl2A
1dj7A	1guxA	1no4A	1wapA	1gh8A	1v2yA	1ayoA	1cqyA	1eczA	1flmA	1i07A	1jmvA	1tvxB
1dm9A	1h2sA	1oczA	1wmsA	1gh9A	1v31A	1b0nA	1cvmA	1ed1A	1fltW	1i4mA	1jyoA	1ukrA
1dp5A	1h3qA	1qg7B	1xbrA	1ghhA	1v32A	1b3aA	1d8eA	1ee6A	1fp2A	1i4sA	1k04A	1ycqA
1dpsA	1h80A	1qn2A	1xdtT	1gjtA	1v38A	1b4uA	1d8uA	1eggB	1fs7A	1i4wA	1k20A	2cpgB
1dtdA	1hcfA	1qnaB	1ycpL	1gjxA	1v3aA	1b66A	1d9uA	1ehkA	1fvzA	1i4zA	1k2fA	2eboA
1e1hA	1hxrA	1qnaA	1ytbA	1go1A	1v3fA	1b9xA	1dazC	1ej3A	1fx8A	1i5gA	1krqA	2hrvA
1e44B	1hynP	1qrvA	2cpgA	1uilA	1v5kA	1bfeA	1dcpA	1ej8A	1fzcA	1i8aA	1ktzB	2thiA
1e7kA	1hyrB	1r26A	2hddA	1ujdA	1v5lA	1bh9A	1debA	1ejeA	1fzhA	1i8nA	1mkaA	3ygsP
1eaiA	1i1rA	1r3jA	2occA	1ujoA	1v5mA	1bkrA	1dfnA	1ejfA	1fzrA	1i9bA	1mr8A	
1eayA	1i78A	1r4xA	2sivA	1ujrA	1v5rA	1bnlA	1dfuP	1elkA	1g5zA	1iazA	1p35A	
1eg4P	1i8lA	1r7jA	3caaA	1ujtA	1v61A	1bplA	1dm9B	1elwA	1g6uA	1ib5A	1pcfA	
1eggA	1ic2A	1ryhA	3ygsC	1ujvA	1v63A	1bqcA	1dmhA	1emvA	1g8kA	1ibyA	1qb3A	
1euvA	1idrA	1scjA	4fapA	1uk5A	1v65A	1bxaA	1dp7P	1eoiA	1g9zA	1id1A	1qftA	

$\Delta^2\psi$ for each residue were calculated as follow:

$$\Delta\varphi(n)=\begin{cases}\min\{|\varphi(n)-\varphi(n-1)|.|\varphi(n)-\varphi(n+1)|\} & \text{if } -100\le\varphi(n)\le 0\\ \frac{|\varphi(n)-\varphi(n\text{-}1)|+|\varphi(n)-\varphi(n+1)|}{2} & \text{Otherwise}\end{cases}$$

$$\Delta\psi(n)=\begin{cases}\min\{|\psi(n)-\psi(n-1)|.|\psi(n)-\psi(n+1)|\} & \textit{if } -100\le\psi(n)\le 0\\ \frac{|\psi(n)\text{-}\psi(n\text{-}1)|+|\psi(n)-\psi(n+1)|}{2} & \textit{Otherwise}\end{cases}$$

$$\Delta^2\varphi(n)=\frac{|\Delta(\varphi(n))-\Delta(\varphi(n-1)|+|\Delta(\varphi(n))-\Delta(\varphi(n+1))|}{2}$$

$$\Delta^2\psi(n)=\frac{|\Delta(\psi(n))-\Delta(\psi(n-1)|+|\Delta(\psi(n))-\Delta(\psi(n+1))|}{2}$$

Where n is denoted as the n^{th} amino acid in the protein.

2. Amino acids which are not located in the helix domain of the Ramachandran plot and with the following conditions were excluded from the data set.

$$\begin{cases}-180\le\varphi\le 0\\ 64.5\le\psi\le 180\end{cases}\ \text{or}\ \begin{cases}-180\le\varphi\le 0\\ -180\le\psi\le -153\end{cases}$$

These residues form the set A.

3. All of the segments assigned as alpha-helix by PDB, with lengths more than seven residues were selected. Three residues from the N-cap and three residues from the C-cap were excluded and averages of $\Delta^2\varphi$ and $\Delta^2\Psi$ for the remaining residues were calculated and denoted by α_φ and α_Ψ, respectively.
4. For all residues in the helix state, in the data set with $\Delta^2\varphi \ge \alpha_\varphi$, average of φ was calculated and named $\ell_{1,\varphi}$.$\ell_{1,\psi}$ was also calculated as above for Ψ angles. Hence, ℓ_φ and ℓ_ψ parameters are defined as:

$$\ell_\varphi = 2\ell_{1,\varphi} - \alpha_\varphi$$

$$\ell_\psi = 2\ell_{1,\psi} - \alpha_\psi$$

In fact α_φ and α_Ψ denote the maximum variations allowed for a helix to be considered as a standard helix. Similar to the rational behind a 95% confidence interval for a mean in a normal distribution, we consider a confidence region for an amino acid to be in a helix structure, based on ℓ_ψ and ℓ_φ simultaneously. It should be mentioned here that the information on amino acids discarded in step 3, is now being considered at this stage. This means no information has been missed. Since we are only interested in helix structure, therefore, all those amino acids considered in steps 3 (internal) and 4 (C- cap and N-cap) are not to be considered.

5. f_φ and f_Ψ functions were defined as follows:

$$f_\varphi(n)=\begin{cases}100 & \text{if } 0\le\Delta^2\varphi(n)\le\alpha_\varphi\\ \frac{100(\ell_\varphi-\Delta^2\varphi(n))}{\ell_\varphi-\alpha_\varphi} & \text{if } \alpha_\varphi\le\Delta^2\varphi(n)\le\ell_\varphi\\ 0 & \text{Otherwise}\end{cases}$$

$$f_\psi(n)=\begin{cases}100 & \text{if } 0\le\Delta^2\psi(n)\le\alpha_\psi\\ \frac{100(\ell_\psi-\Delta^2\psi(n))}{\ell_\psi-\alpha_\psi} & \text{if } \alpha_\psi\le\Delta^2\psi(n)\le\ell_\psi\\ 0 & \text{Otherwise}\end{cases}$$

finally function f gives the fuzzy value for helicity according to the following formulation:

$$f(n)=\begin{cases}\frac{f_\varphi(n)+f_\psi(n)}{2} & \text{if residue was not in set A}\\ 0 & \text{if residue was in set A}\end{cases}$$

RESULTS AND DISCUSSION

Analysis of helix regularity using variation in the consecutive residue dihedral angles φ and ψ gives the helix fuzzy number for each residue, between 0 to 100. Table 2 shows these numbers for two proteins. In this table helix assignment by PDB, DSSP, STRIDE, with fuzzy numbers greater than also 83 being compared. Usually the central residues of helices take numbers close to 100, and N- and C- terminal residues of each helix take lower values and show less regularity. Consecutive residues with the same or near fuzzy numbers show the regular helix turn, although it may be far from the standard helix structure. Segments with fuzzy numbers close to 100 are regular helices with standard helix geometries. Helix distortion has been studied in detail and can be attributed to factors such as solvent-side chain interactions, local sequence and side chain packing (Barlow and Thornton, 1998). However, these factors cause the residues in helices to have different major chain conformations and such distortions could be shown by differences in consecutive dihedral angles.

Figure 1 shows the superposition of fragments assigned as helices by PDB with the same length and different or same fuzzy numbers using the CE program (http://cl.sdsc.edu/) (Shindyalov and Bourne, 1998). Root mean square (RMS) calculation shows a relation between fuzzy numbers and geometry of compared helices. Two superposed helices with the same fuzzy numbers show less RMS which increases when the fuzzy numbers of two helices are different. These assigned fuzzy numbers for residue helicity, in addi-

Table 2. Fuzzy numbers for parts of two proteins and comparison of assigned helices by PDB, DSSP, STRIDE with fuzzy numbers greater than 83.

PDB Code	Residue No.	AA	Φ	ψ	PDB	DSSP	STRIDE	Fuzzy	fuzzy number
1tafA	1	P	360	-46.7			H		0
1tafA	2	K	-60.4	-46.67	H	H	H		24
1tafA	3	D	-71.63	-33.94	H	H	H	H	98
1tafA	4	A	-61.48	-40.57	H	H	H	H	95
1tafA	5	Q	-63.2	-42.66	H	H	H	H	100
1tafA	6	V	-60.52	-44.64	H	H	H	H	100
1tafA	7	I	-63.5	-38.7	H	H	H	H	100
1tafA	8	M	-66.31	-35.52	H	H	H	H	100
1tafA	9	S	-69.6	-37.58	H	H	H	H	100
1tafA	10	I	-64.63	-43.23	H	H	H	H	100
1tafA	11	L	-58.14	-44.84	H	H	H	H	100
1tafA	12	K	-68.43	-43.9	H	H	H	H	100
1tafA	13	E	-61.57	-31.72	H	H	H	H	95
1tafA	14	L	-92.67	18.81			H	H	94
1tafA	15	N	62.46	32.33					50
1tafA	16	V	-100.04	89.05					0
1tafA	17	Q	-71.16	-28.82					34
1tafA	18	E	-135.82	139.22					0
1tafA	19	Y	-157.47	154.29					0
1tafA	20	E	-66.29	138.8					0
1tafA	21	P	-51.5	-38.42	H		H		61
1tafA	22	R	-66.37	-10.64	H	H	H	H	94
1tafA	23	V	-62.41	-39.91	H	H	H	H	88
1tafA	24	V	-61.19	-43.74	H	H	H	H	100
1tafA	25	N	-60.59	-47.54	H	H	H	H	100
1tafA	26	Q	-56.4	-43.61	H	H	H	H	100
1tafA	27	L	-71.13	-30.13	H	H	H	H	99
1tafA	28	L	-71.39	-36.28	H	H	H	H	100
1tafA	29	E	-67.91	-38.07	H	H	H	H	100
1tafA	30	F	-64.85	-46.8	H	H	H	H	100
1tafA	31	T	-52.22	-48.2	H	H	H	H	96
1tafA	32	F	-63.38	-45.89	H	H	H	H	95
1tafA	33	R	-61.87	-43.52	H	H	H	H	100
1tafA	34	Y	-66.58	-51.07	H	H	H	H	100
1tafA	35	V	-62.33	-44.79	H	H	H	H	100
1tafA	36	T	-63.88	-38.44	H	H	H	H	100
1tafA	37	S	-62.64	-47.57	H	H	H	H	100
1tafA	38	I	-64.34	-44.48	H	H	H	H	100
1tafA	39	L	-66.32	-34.26	H	H	H	H	100
1tafA	40	D	-62.57	-39.36	H	H	H	H	100
1tafA	41	D	-77.4	-42.75	H	H	H	H	94
1tafA	42	A	-57.04	-35.34	H	H	H	H	95
1tafA	43	K	-62.26	-37.52	H	H	H	H	100
1tafA	44	V	-64.43	-44.47	H	H	H	H	100
1tafA	45	Y	-61.78	-43.66	H	H	H	H	100
1tafA	46	A	-61.67	-40.22	H	H	H	H	100
1tafA	47	N	-61.9	-50.3	H	H	H	H	98
1tafA	48	H	-65.91	-15.55	H	H	H	H	99
1tafA	49	A	-97.48	-5.11			H	H	85
1tafA	50	R	63.92	50.68					11
1tafA	51	K	-118.09	153.71					0
1tafA	52	K	-97.6	-24.84					44
1tafA	53	T	-115.29	129.16					0
1tafA	54	I	-65.54	131.73					0

Table 2. Continue

PDB Code	Residue No.	AA	Φ	ψ	PDB	DSSP	STRIDE	Fuzzy	fuzzy number
2hddA	1	R	360	98.14			C		0
2hddA	2	T	-56.81	156.48			C		0
2hddA	3	A	-122.02	109.92			C		0
2hddA	4	F	-64.56	141.64			C		0
2hddA	5	S	-78.2	150.34			C		0
2hddA	6	S	-60.71	-17.65	H	H	H		62
2hddA	7	E	-74.65	-39.48	H	H	H	H	87
2hddA	8	Q	-72.48	-47.52	H	H	H	H	100
2hddA	9	L	-57.22	-36.55	H	H	H	H	93
2hddA	10	A	-70.11	-34.6	H	H	H	H	97
2hddA	11	R	-75.58	-39.4	H	H	H	H	100
2hddA	12	L	-63.44	-44.31	H	H	H	H	99
2hddA	13	K	-64.21	-40.69	H	H	H	H	100
2hddA	14	R	-65.51	-37.31	H	H	H	H	100
2hddA	15	E	-71.47	-42.55	H	H	H	H	99
2hddA	16	F	-59.23	-39.56	H	H	H	H	100
2hddA	17	N	-65.6	-36.03	H	H	H	H	100
2hddA	18	E	-74.17	-42.78	H	H	H	H	100
2hddA	19	N	-150.93	110.74			T		0
2hddA	20	R	-75.14	-7.7		S	T	H	85
2hddA	21	Y	-122.12	129.73		S	T		0
2hddA	22	L	-84.17	146.57			T		0
2hddA	23	T	-100.18	163.94			C		0
2hddA	24	E	-60.51	-41.43	H	H	H		33
2hddA	25	R	-63.75	-46.86	H	H	H	H	100
2hddA	26	R	-66.22	-39.88	H	H	H	H	99
2hddA	27	R	-59.23	-38.93	H	H	H	H	99
2hddA	28	Q	-66.89	-44.56	H	H	H	H	96
2hddA	29	Q	-66.63	-38.52	H	H	H	H	100
2hddA	30	L	-67.87	-39.93	H	H	H	H	100
2hddA	31	S	-57.68	-50.17	H	H	H	H	97
2hddA	32	S	-59.15	-63	H	H	H	H	100
2hddA	33	E	-57.53	-35.04	H	H	H	H	95
2hddA	34	L	-93.27	-10.4	H	H	H	H	86
2hddA	35	G	68.66	49.19		T	C		15
2hddA	36	L	-134.64	163.67			C		0
2hddA	37	N	-77.92	148.18			C		0
2hddA	38	E	-58.92	-29.2	H	H	H		44
2hddA	39	A	-67.15	-33.47	H	H	H	H	99
2hddA	40	Q	-71.2	-37.64	H	H	H	H	100
2hddA	41	I	-63.97	-48.95	H	H	H	H	100
2hddA	42	K	-55.1	-47.67	H	H	H	H	100
2hddA	43	I	-70.52	-30.37	H	H	H	H	92
2hddA	44	W	-71.78	-42.86	H	H	H	H	97
2hddA	45	F	-64.27	-39.37	H	H	H	H	100
2hddA	46	K	-62.29	-46.72	H	H	H	H	100
2hddA	47	N	-70.3	-34.17	H	H	H	H	98
2hddA	48	K	-64.31	-49.57	H	H	H	H	99
2hddA	49	R	-53.59	-42.44	H	H	H	H	99
2hddA	50	A	-67.6	-36.55	H	H	H	H	96
2hddA	51	K	-65.26	-49.54	H	H	H	H	98
2hddA	52	I	-59.03	-34.39	H	H	H	H	94
2hddA	53	K	-60.14	-34.46	H	H	H	H	100
2hddA	54	K	-101.5	51.26		T	H		35
2hddA	55	S	-96.32	360			C		43

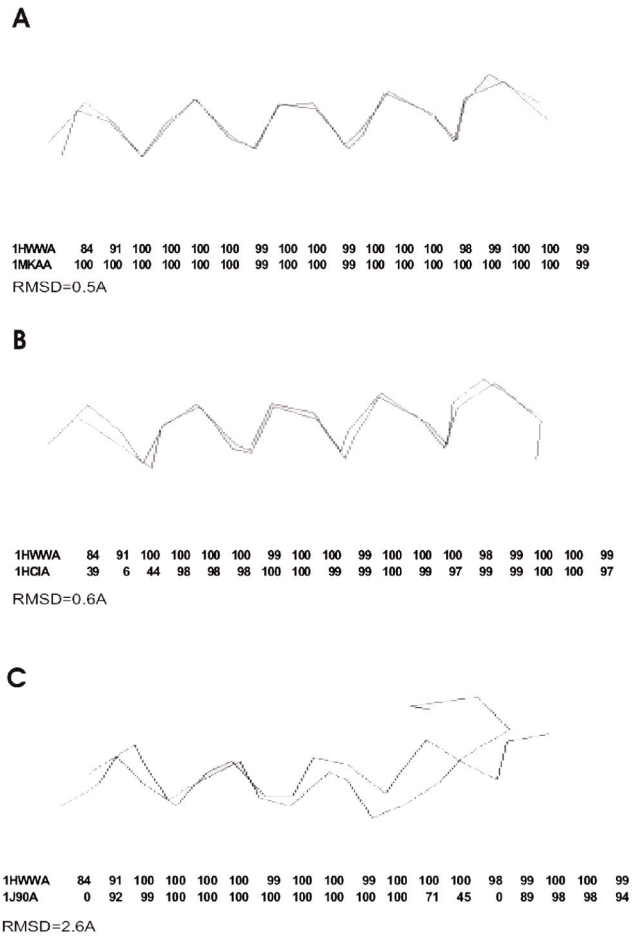

Figure 1. Superposition of helices with the same length and different or same fuzzy numbers, with their RMSD for residues 310-327 of 1HWWA, 79-96 of 1MKAA (A) 715-732 of 1HCIA (B) and 156-173 of 1J90A (C).

tion to showing helix regularity can be used for comparison and alignment of protein structures. Instead, those with are based on a string of secondary structure elements in which each residue is defined as belonging to one state or another, and where the regularity and geometry of secondary structure is ignored. Fuzzy numbers also show helicity for small segments with lengths of two or three residues that although are not classified as helices, but share similar geometry with the helix. However, the main goal of this method is assignment of a helical fuzzy number to each residue, but it can also be simply converted to the classical method involving the assignment of a residue with helical or non-helical structure. For this purpose, residues with fuzzy numbers greater than a threshold number k, were assigned as H and others as ***H***. In a five residue length window, if one H is surrounded by four ***Hs***, it can be converted to ***H*** and vice versa. Allowing k to vary, we can find all helix structures near to or far from the standard helix structure. For example, for k close to 100, the helix structures near to the standard are found and if k was far from 100, we detect the structure far from the standard. In order to compare with PDB, we look for a certain k for which the correlation coefficient of data generated by our algorithm after using the threshold number k and those generated by PDB are maximized. This leads to k= 83. Comparisons of the results with the crystallographer's assignments as percentage of correctly assigned residues in two states (helix or non-helix) are 90% for all amino acids in the dataset.

Comparison of DSSP with our method shows that they have 94% agreement for H and ***H***. Although many of the crystallographers define secondary structure based on the DSSP algorithm, comparison of DSSP and PDB assigned secondary structures in our dataset show 8% differences between them. Analysis of differences between results of this study and DSSP showed that 1342 residues were assigned by the method of this study to H, while DSSP assigned them to ***H***. There were 1783 residues that our method assigned to ***H***, while DSSP assigned them to H. Comparison of our method and STRIDE show approximately 94% agreement for H and ***H***. Table 3 shows the details of comparisons between the method described heae with PDB, DSSP and STRIDE and also comparisons of DSSP and STRIDE with PDB. Most of

Table 3. Comparison of results obtained by fuzzy logic and other methods.

Compared methods	TP[1]	TN[2]	FP[3]	FN[4]	Tot	%	Sensitivity	Specificity	CC[5]
Fuzzy and PDB	16822	26912	1200	3710	48644	89.9	81.9	93.3	0.79
Fuzzy and DSSP	15009	30510	1342	1783	48644	93.6	89.4	91.8	0.86
Fuzzy and STRIDE	15366	30175	1090	2013	48644	93.6	88.4	93.4	0.86
DSSP and PDB	16712	28035	80	3820	48644	91.98	81.4	99.5	0.84
STRIDE and PDB	17096	27832	283	3436	48644	92.36	83.3	98.4	0.85

[1]True positive (TP), [2]True negative (TN), [3]False Positive (FP), [4]False negative (FN), [5]Correlation Coefficient (CC).

the false positive and negative assignments between method of this study and PDB occurred at the edges of helices. Although the major assumption of this work is that helices can be defined by fuzzy logic and instead of assigning each residue to one state, it may be assigned by a fuzzy number which is far more valuable for comparing protein structures. However, this approach can also be used in the classical assignment of helix structure. The results obtained are as good as DSSP and STRIDE algorithms, which are the most widely used methods for secondary structure assignment.

In this article the main goal was only fuzzy number assignment to helices followed by demonstration of their regularities. Fuzzy number assignment to other secondary structures such as beta-strands and turns can be the subject of an independent work and in fact we are developing a method for fuzzy assignment of secondary structures. For this reason the title "Helix segment assignment in proteins using fuzzy logic" was selected for this article.

It is also believed that the combination of dihedral angles and other parameters such as H-bonds can lead to a different method with better results which can also be the subject of an other independent work.

References

Andreeva A, Howorth D, Brenner SE, Hubbard TJ, Chothia C, Murzin AG (2004). SCOP database in 2004: refinements integrate structure and sequence family. *Nucleic Acids Res*. 32: 226-229.

Barlow DJ, Thornton JM (1988). Helix geometry in proteins. *J Mol Biol*. 201: 601-619 Colloc'h N, Etchebest C, Thoreau E, Henrissat B, Mornon JP (1993). Comparison of three algorithms for the assignment of secondary structure in proteins: the advantages of a consensus assignment. *Protein Eng*. 6: 377-382.

Colloc'h N, Etchebest C, Thoreau E, Henrissat B, Mornon JP (1993). Comparison of three algorithms for the assignment of secondary structure in proteins: the advantages of a consensus assignment. *Protein Eng*. 6:377-82.

Fischel-Ghodsian F, Mathiowitz G Smith TF (1993). Alignment of protein sequences using secondary structure: a modified dynamic programming method. *Protein Eng*. 3: 577-81.

Frishman D, Argos P (1995). Knowledge-based protein secondary structure assignment. *Proteins* 23: 566-79.

Henneke CM (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: 141-50.

Hogue CW, Bryant SH (1998). Structure databases. *Methods Biochem Anal*. 39: 46-73.

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

Kolinski A, Rotkiewicz P, Ilkowski B, Skolnick J (1999). A method for the improvement of threading-based protein models. *Proteins* 37: 592-610.

Levitt M, Greer J (1977). Automatic Identification of Secondary Structure in Globular Proteins. *J Mol Biol.* 114 : 181-239.

Marti-Renom MA, Stuart A, Fiser A, Sanchez R, Melo F (2000). Comparative protein structure modeling of genes and genomes. *Annu Rev Biophys Biomol Struct*. 29: 291- 325.

Pearl F, Todd A, Sillitoe I, Dibley M, Redfern O, Lewis T, Bennett C, Marsden R, Grantm A, Lee D (2005). The CATH Doma4in Structure Database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. *Nucleic Acids Res*. 33: Database Issue D247-D251.

Rice DW, Eisenberg D (1997). A 3D-1D substitution matrix forprotein fold recognition that includes predicted secondary structure of thesequence. *J Mol Biol*. 267: 1026-1038.

Richards FM, Kundrot CE (1988). Identification of structural motifs from protein coordinate data: Secondary structure and first level super-secondary structure. *Proteins* 3:71-84.

Rost B (2000). TOPITS: Threading one-dimensional predictions into three-dimensional structures. *The third international conference on Intelligent Systems for Molecular Biology*, 314-321.

Sauder JM, Arthur JW, Dunbrack RL (2000). Large-scale comparison of protein sequence alignment algorithms with structure alignments. *Proteins* 40: 6-22.

Shindyalov IN, Bourne PE (1998). Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng*. 11: 739-747.

Smith RF, Smith TF (1992). Pattern-induced multi-sequence alignment (PIMA) algorithm employing secondary structure-dependent gap penalties for use in comparative protein modeling. *Protein Eng*. 5: 35-41.

Sternberg MJ, Bates PA, Kelley LA, MacCallum RM (1999). Progress in protein structure prediction: assessment of CASP3. *Curr Opin Str Biol*. 9: 368-73.

Xu Y, Xu D, Crawford OH, Einstein, Larimer F, Uberbacher E, Unseren MA, Zhang G (1999). Protein threading by PROSPECT: a prediction experiment in CASP3, *Protein Eng*. 12: 899-907.