A Method for Multiple Structural Alignment of Proteins Using Text Modeling Techniques

Jafar Razmara

***Abstract*—Here, an efficient method is introduced for multiple structural alignment of proteins. The method encodes geometry of protein secondary and tertiary structures in linear sequences and then uses a hierarchical procedure for superposition of proteins based on these sequences. To capture similarities between secondary structure sequences, the method utilizes n-gram modeling technique over entropy concept adopted from computational linguistics. Moreover, a step-by-step algorithm is used to align relative residue position sequences in tertiary structure level. A number of case studies are presented here to demonstrate the power of the method comparing with other structure alignment tools. The results provide evidence for efficiency and applicability of the proposed method.**

***Index Terms*—Multiple structure alignment, linear encoding methods, text modeling.**

I. Introduction

Structural comparison of biomolecules is a major step in structural biology. Biologists have believed that proteins with similar structures share common functions and properties. Therefore, structural comparison tools are widely utilized to classify all known proteins in the databases or search similarity of a newly discovered protein to known classified proteins. The tools are also used to determine evolutionary relationships between proteins that are difficult to detect by sequence similarity analysis.

Structural alignment tools are generally used to highlight similarities between various proteins functionalities. The algorithm looks for an optimal correspondence among atoms of two structures with a minimal distance between the matched pairs. There is not any initial knowledge about corresponding parts of two structures. Therefore, the algorithm needs an exhaustive heuristic search to superpose the best matched pair of atoms.

Several pairwise protein structure alignment methods have been proposed using heuristic strategies to compare geometrical coordinates of the C_α backbone atoms to find the best optimal correspondence between residues of two structures. The used techniques are distance matrices comparison (DALI) [1], vector alignment of secondary structure alignment (VAST) [2], combinatorial extension (CE) [3], secondary structure matching (SSM) [4], matching molecular models obtained from theory (MAMMOTH) [5], dynamic programming on TM-scorerotation matrix (TM-align) [6], genetic algorithm for non-sequential gapped protein structure alignment (GANGSTA) [7] and many others ([8]-[11]). Several comprehensive reviews and evaluation of the methods have been reported in literatures ([12]-[14]).

Multiple protein structure alignment as a typical alignment method is basically utilized to find correspondence between residues of a set of proteins through looking for their common substructures. It has wide applications in exploring evolutionary relationships between protein families [15], function prediction through analyzing conserved active sites in the structure of homologous proteins [16], and protein structure prediction through creating profiles and threading templates [17].

Structural comparison problem has been studied widely during past two decades and several methods have been proposed for pair-wise alignment. However, a few methods developed for multiple structure alignment. The programs for multiple alignment are mostly built on top of pairwise structure alignment methods and then extended to align multiple structures. MultiProt [18] is a fully automated efficient method which finds common geometrical cores between input molecules. It is an AFP-based program that uses rigid body superimposition. Mustang [19] is another program which uses a combination of short fragment alignment, contact maps, and consensus-based methods. Moreover, Matt [20] is an aligned fragment pair chaining algorithm which allows local flexibility between fragments. After a dynamic programming assembly of AFPs, Matt restores geometric consistency in the final step of the alignment. Despite proposition of these efficient techniques, the study for development of new alternative methods is still an active research area.

Linear encoding techniques have been recently developed for fast protein structure comparison [21]-[24]. The methods commonly encode protein structure in one-dimensional linear sequences and then, use sequence alignment techniques for structural alignment of proteins. These methods are more relevant to be extended for multiple structure alignment. The method firstly encodes geometry of secondary structure elements in a topology string and then, superposes these strings using n-gram modeling technique adopted from computational linguistics. The technique has been inspired from the scheme proposed by authors in [24]. After secondary structure superposition, the method encodes each protein structure in tertiary level into a linear sequence. Finally, the method employs n-gram modeling technique from computational linguistics to capture regularities between these sequences. The encoding technique is adopted from a method introduced by authors in [25]. To this end, a step-by-step procedure is utilized to locate identical n-gram words within protein sequences, and then extend the

Manuscript received October 1, 2014; revised December 22, 2014.

J. Razmara is with the Department of Computer Sciences, Faculty of Mathematical Sciences, University of Tabriz, Tabriz, Iran (e-mail: razmara@tabrizu.ac.ir).

DOI: 10.7763/IJFCC.2015.V4.373

alignment to the other residues along sequences.

II. Methods

The input of the algorithm is a set of n proteins P_1, P_2, …,P_n. For each protein, the geometry of secondary structure elements and 3D-coordinates of its atoms in PDB format are provided. The general strategy of the algorithm for multiple structure alignment is organized in two major steps: secondary structure superposition and 3D-structure alignment. The following is description of the algorithm.

A. Multiple Secondary Structure Superposition

The secondary structure is known as backbone of a protein and is made of highly regular substructures called α-helices and β-strands. In the first step, the method encodes geometry of secondary structure elements (SSEs) of each protein in a topology string called SSEs sequence. To this end, each element is assumed as a vector $r_{SSE}=r_b-r_e$ where for helices and for strands [26] (indices i and j denote the first and last residues). Based on the sign of the x, y, and z components, each vector is encoded to a letter as shown in Table I. Moreover, for each pair of consecutive SSE vectors, an inter-SSE vector is defined using end and start points of two SSEs. This vector determines relative position of an element with respect to its previous vector. Fig. 1 shows a typical example for SSEs representation as a set of vectors and their encoding in a topology string.

$$r_b = \left(0.74r_i + r_{i+1} + r_{i+2} + 0.74r_{i+3}\right)/3.48$$

$$r_e = \left(0.74r_{j-3} + r_{j-2} + r_{j-1} + 0.74r_j\right)/3.48 \quad (1)$$

$$r_b = \left(r_i + r_{i+1}\right)/2, r_e = \left(r_{j-1} + r_j\right)/2 \quad (2)$$

TABLE I: Secondary Structure Vectors Direction and Labels

Direction			Strand	Helix	Inter-SSEs
+x	+y	+z	A	I	Q
+x	+y	-z	B	J	R
+x	-y	+z	C	K	S
+x	-y	-z	D	L	T
-x	+y	+z	E	M	U
-x	+y	-z	F	N	V
-x	-y	+z	G	O	W
-x	-y	-z	H	P	X

TABLE II: Permutation of the Letters Based on 90 Degree Rotation around x, y, z Axes

	Strand	Helix	Inter-SSEs
Old	ABCDEFGH	IJKLMNOP	QRSTUVWX
Rotate around x	BDACFHEG	JLIKNPMO	RTQSVXUW
Rotate around y	EAGCFBHD	MIOKNJPL	UQWSVRXT
Rotate around z	EFABGHCD	MNIJOPKL	UVQRWXST

The 3D-coordinate of each protein is represented in an arbitrary relative direction. Therefore, finding a correspondence between two structures needs to rotate a structure around the other or use a coordinate independent representation of two structures. Having the topology string of secondary structure, now, the method applies a string permutation scheme to find an overlap between structures. The scheme generates 24 permuted strings from topology string of each protein by 90 degree rotation of a structure around the x, y, and z axes. For each rotation around axes, the letters in the topology string are permuted according to Table II.

The above 24 permuted strings are considered as estimation of different possible orientation of a query structure that can be matched with a reference protein. To find a match between two protein structures, the method applies cross entropy measure over n-gram modeling technique adopted from computational linguistics [27]. The technique firstly makes n-gram model by counting the words of one sequence in training phase, and then, measures predictability of the second sequence in recall phase via formula:

$$H(X, P_M) = -\sum_{(allw_i^n)} P(w_i^n) log\,(2 + P_M(w_{(i+n)} | w_i^{(n-1)}))$$
$$= -\frac{1}{N}\sum_{(allw_i^n)} Count(w_i^n) log\,(2 + P_M(w_{(i+n)} | w_i^{(n-1)})) \quad (3)$$

where the variable X is in the n-gram form $w_i^n=\{w_i, w_{i+1}, \ldots, w_{i+n-1}\}$. The summation runs over all the possible n-gram words w_i^n, and N is the number of n-grams. The term $Count(w_i^n)$ is computed by the word count within the first sequence. Moreover, the conditional probability in the summation makes relation between the n-th element of an n-gram and the preceding n-1 elements, which can be computed by counting the words of the second sequence and having the model estimated:

$$P(w_{(i+n)} | w_i^{(n-1)}) = Count(w_{(i+n)}) / Count(w_i^{(n-1)}) \quad (4)$$

The above cross entropy formula is used to measure similarity of each 24 different topology strings of the query structure to the topology string of a reference protein via the formula:

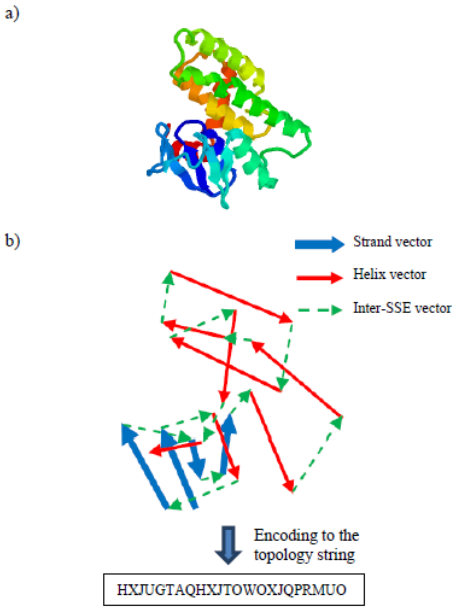

Fig. 1. A typical example for SSEs representation in a topology string: a) A protein 3D structure, b) Vector representation of secondary structure elements, and its generated topology string.

$$D(S_r, S_i) = \left|H\left(X_r, P_{M_i}\right) - PS\right| \quad (5)$$

where PS is the perfect score using the first sequence as

reference and model sequences. S_r and S_i also denote the reference topology string and *i*-th string of query structure respectively. The lower value of $D(S_r, S_i)$indicates higher similarity of the compared sequences.

Having the most similar topology string of two protein structures, now, a procedure looks for the most common and long identical n-gram words within topology strings of the protein set. The procedure applies an iterative task for decreasing size of n-gram from *m* (chosen empirically 6) down to basic size of n-gram (chosen at 3). After that, the procedure makes another effort to extend matches along the rest of elements among topology strings. The output of this procedure is the map of correspondence between SSEs of proteins.

The above procedure is used to find a correspondence map between SSEs of each pair of proteins. A matching score is calculated for each pair based on the number of matched elements. The protein with the maximum score is chosen as a reference to create the multiple superposition among proteins in the set. As a result, a multiple correspondence map is created based on the matched elements of each protein with the reference protein.

B. Multiple Structure Alignment at the Residue Level

After creation of an initial correspondence map between SSEs, the structure of each protein is rotated to achieve an initial overlap with the reference protein structure as chosen in the last procedure. To this end, a rotation matrix is created based on the average of the angles between matched SSE vectors of each protein in the set and the reference protein. Then, the rotation matrix of each structure is used to find the new coordinates of residues within protein.

In order to find a multiple structure alignment, first, an encoding procedure is used to convert protein structure in tertiary level into a sequence of letters. The encoding technique is inspired from a scheme introduced by authors in [24]. The scheme labels the relative position of each residue's $C\alpha$ atom with respect to the position of $C\alpha$ atom of its previous residue in 3D coordinates using the 26 letters of Alphabets. The resulting sequence is called relative residue position sequence [24] (The detailed information is available in [24]).

Having relative residue position sequence of each protein and the correspondence map of SSEs of all proteins, a step-by-step procedure is now applied to perform multiple structure alignment. The procedure runs as the following steps:

1) For each set of matched SSEs, locate identical words in relative residue position sequences and mark their residues as aligned. Expand the alignment to the ends of SSEs for corresponding residues leaving no unmatched residues between the matched ones.
2) For each set of remaining identical words along proteins, check connectivity of the aligned residues and distance between residues that are less than a predefined threshold, and then, mark their residues as aligned.
3) For each set of identical word from a subset of proteins, check connectivity of the aligned residues and general order of them along protein chains, and, select the best pair of identical words and mark their residues as aligned.
4) For the rest of unaligned residues, try to align residues, which satisfy the predefined distance threshold and connectivity with the aligned residues.

Having the set of aligned residues, the procedure applies a refinement task to find the optimal correspondence among structures. The method utilizes an iterative procedure to make a rotation matrix based on Kabsch's method [25] between each protein in the set and the reference protein.

TABLE III: PAIRWISE ALIGNMENT RESULTS OF TEN 'HARD TO DETECT' PAIRS OF STRUCTURES FROM FISCHER DATASET BY THREE METHODS

		CE		MultiProt		Our method	
Protein 1	Protein 2	Length	RMSD	Length	RMSD	Length	RMSD
1fxiA	1ubq	-	-	44	1.7	49	2.11
1ten	3hhrB	87	1.9	81	1.3	83	1.71
3hlaB	2rhe	85	3.5	60	1.8	72	2.71
2azaA	1paz	85	2.9	75	2.0	82	2.83
1cewI	1molA	69	1.9	76	1.8	74	2.01
1cid	2rhe	94	2.7	84	1.8	92	2.57
1crl	1ede	-	-	161	2.3	215	3.04
2sim	1nsbA	284	3.8	233	2.3	266	3.04
1bgeB	2gmfA	74	2.5	78	2.5	82	3.03
1tie	4fgf	82	1.7	95	2.1	103	2.52

TABLE IV: MULTIPLE STRUCTURE ALIGNMENT RESULTS OF 5 DIFFERENT PROTEIN FAMILIES BY OUR METHOD AND MULTIPROT THE RESULTS FOR MULTIPROT WERE TAKEN FROM [4])

Protein families	Average size	MultiProt Alignment Length	Our method Alignment Length
Serpins: 7apiA, 8apiA, 1hleA, 1ovaA, 2achA, 9apiA, 1psi, 1atu, 1kct, 1athA, 1attA, 1antl, 2antl	372	237	231
Serine Proteinase: 1cseE, 1sbnE, 1pekE, 3prkE, 3tecE	277	227	215
Calcium-binding: 4cpv, 2scpA, 2sas, 1top, 1scmB, 3icb	140	36	36
TIM-Barrels : 7timA, 1tml, 1btc, 4enl, 1pii, 6xia, 5rubA	391	44	40
Hclix-Bundle: 1flx, 1aep, 1bbhA, 1bgeB, 1le2, 1rcb, 256bA, 2ccyA, 2hmzA, 3inkC	140	27	27

III. RESULTS AND DISCUSSION

A. Pairwise Alignment Test

The first experiment is to test the method on a set of "hard to detect" pairwise alignments. The set consists of ten difficult to align pair of proteins as described in literatures [18]. The experiment compares the alignment results of our method with those of CE [3] and MultiProt [18]. The results for CE and MultiProt were taken from [18]. The results are shown in Table III based on RMSD and length of alignment. As can be seen from the table, the alignment results of our method are very competitive with two other methods.

B. Multiple Alignment Test

A comparative study has also done to test the ability of the method for multiple structure alignment. The study applies the introduced method over a set of protein families and compares its alignment results with those of MultiProt as a well-known multiple alignment method. Table IV represents the alignment results obtained by two methods. As can be seen from the table, in all cases, our method produces a comparable alignment outputs for the protein families. A sample multiple structure alignment prepared by our method

is represented in Fig. 2 for *Serine Proteinase* protein family.

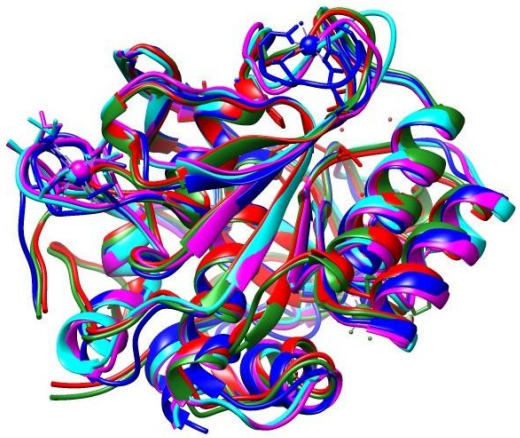

Fig. 2. Multiple structure alignment results prepared by our method for Serine proteinase protein family including PDB codes: 1CSEE (magenta), 1SBNE (CYAN), 1PEKE (red), 3PRKE (green), and 3TECE (blue).

IV. Conclusion

The introduced method applies a known powerful technique based on linear encoding of protein structure for multiple structure alignment. The advantage of the method is its simple hierarchical schemes for encoding secondary and tertiary structures and utilizing efficient techniques for their alignment. The experimental results demonstrate efficiency and applicability of the linear encoding scheme in structural alignment of biomolecules.

References

[1] L. Holm and C. Sander, "Protein structure comparison by alignment of distance matrices," *Journal of Molecular Biology*, vol. 233, no. 1, pp. 123-138, 1993.

[2] J. F. Gibrat, T. Madej, J. L. Spouge, and S. H. Bryant, "The VAST protein structure comparison method," *Biophysics Journal*, vol. 72, pp. MP298, 1997.

[3] I. Shindyalov and P. Bourne, "Protein structure alignment by incremental combinatorial extension (CE) of the optimal path," *Protein Engineering*, vol. 11, no. 9, pp. 739-747, 1998.

[4] E. Krissinel and K. Henrick, "Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions," *Acta Crystallographica Section D: Biological Crystallography*, vol. 60, pp. 2256-2268, 2004.

[5] A. R. Ortiz, C. E. Strauss, and O. Olmea, "MAMMOTH (matching molecular models obtained from theory): An automated method for model comparison," *Protein Science*, vol. 11, no. 11, pp. 2606-2621, 2002.

[6] Y. Zhang and J. Skolnick, "TM-align: A protein structure alignment algorithm based on the TM-score," *Nucleic Acid Research*, vol. 33, no. 7, pp. 2302-2309, 2005.

[7] B. Kolbeck and P. May, "Schmidt-Goenner T, Steinke T, Knapp EW: Connectivity independent protein-structure alignment: A hierarchical approach," *BMC Bioinformatics*, vol. 7, 2006.

[8] T. Kawabata, "MATRAS: A program for protein 3D structure comparison," *Nucleic Acids Research*, vol. 31, no. 13, pp. 3367-3369, 2003.

[9] M. L. Sierk and G. J. Kleywegt, "DEJAVU all over again: Finding and analyzing protein structure similarities," *Structure*, vol. 12, no. 12, pp. 2103-2111, 2004.

[10] R. Mosca and T. Schneider, "RAPIDO: a web server for the alignment of proteinstructures in the presence of conformational changes," *Nucleic Acids Research*, vol. 36, Web Server: W42-W46, 2008.

[11] S. Salem, M. J. Zaki, and C. Bystroff, "FlexSnap: Flexible non-sequential proteinstructure alignment," *Algorithms for Molecular Biology*, vol. 5, no. 12, 2010.

[12] G. Mayr, F. Domingues, and P. Lackner, "Comparative analysis of protein structure alignments," *BMC Structural Biology*, vol. 7, no. 50, 2007.

[13] R. Kolodny, P. Koehl, and M. Levitt, "Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures," *Journal of Molecular biology*, vol. 346, no. 4, pp. 1173-1188, 2005.

[14] M. Novotny, D. Madsen, and G. J. Kleywegt, "Evaluation of protein foldcomparison servers," *Proteins: Structure, Function, and bioinformatics*, vol. 54, no. 2, pp. 260-270, 2004.

[15] M. Menke, B. Berger, and L. Cowen, "Matt: local flexibility aids protein multiple structure alignment," *PLOS Computational Biology*, vol. 4, no. 1, pp. 88-99, 2008.

[16] J. A. Irving, J. C. Whisstock, and A. M. Lesk, "Protein structural alignments and functional genomics," *Proteins*, vol. 42, no. 3, pp. 378-382, 2001.

[17] R. L. Dunbrack, "Sequence comparison and protein structure prediction," *CurrOpinStructBiol*, vol. 16, no. 3, pp. 374-384, 2006.

[18] M. Shatsky, R. Nussinov, and H. Wolfson, "A method for simultaneous alignment of multiple protein structures," *Proteins* 56, no. 1, pp. 143-156, 2004.

[19] A. Konagurthu, J. Whisstock, P. Stuckey, and A. Lesk, "MUSTANG: A multiple structural alignment algorithm," *Proteins*, vol. 64, no. 3, pp. 559-574, 2006.

[20] M. Menke, B. Berger, and L. Cowen, "Matt: local flexibility aids protein multiple structure alignment," *PLoS Computational Biology*, vol. 4, no. 1, 2008.

[21] M. Carpentier, S. Brouillet, and J. Pothier, "YAKUSA: A fast structural database scanning method," *Proteins*, vol. 61, no. 1, pp. 137-151, 2005.

[22] C. H. Tung, J. W. Huang, and J. M. Yang, "Kappa-alpha plot derived structural alphabet and BLOSUM-like substitution matrix for rapid search of protein structure database," *Genome Biology*, vol. 8, no. 3, pp. R31, 2007.

[23] J. Razmara, S. Deris, and S. Parvizpour, "TS-AMIR: A topology string alignment method for intensive rapid protein structure comparison," *Algorithms for Molecular Biology*, vol. 7, no. 4, 2012.

[24] J. Razmara, S. Deris, and S. Parvizpour, "A rapid protein structure alignment algorithm based on a text modeling technique," *Bioinformation*, vol. 6, no. 9, pp. 344-347, 2011.

[25] W. Kabsch, "A discussion of the solution for the best rotation to relate two sets of vectors," *ActaCryst.*, vol. 34, pp. 827-828, 1978.

[26] A. P. Singh and D. L. Brutlag, "Hierarchical protein structure superposition using both secondary structure and atomic representations," in *Proc. International Conference on Intelligent Systems in Molecular Biology*, no. 5, pp. 284-293, 1997.

[27] A. B. Marta, A. Hategan, and I. Pitas, "Language engineering and information theoretic methods in protein sequence similarity studies," *Studies in Computational Intelligence*, vol. 85, pp. 151-183, 2008.

Jafar Razmara is an academic member (assistant professor) in the Department of Computer Sciences, Faculty of Mathematical Sciences, University of Tabriz. He received his BSc. and MSc. in software engineering from Isfahan University of Technology and Tarbiat Modares University, respectively. Also, he received his PhD. from Universiti Teknologi Malaysia in computer science.

His main research interests include computational intelligence and its applications. Specifically, his research interests are artificial intelligence applications in bioinformatics, computational linguistics, soft computing, data and text mining.