Distance matrix analysis of mutual secondary structure pairs for multiple structure alignment

Chiang-Man Sun , Tun-Wen Pai*, Jen-Chun Hung, Ying-Tsang Lo, and Po-Hung Chen

[1] Department of Computer Science and Engineering, National Taiwan Ocean University, No. 2,Pei-Ning Rd. Keelung, 20224, Taiwan, Republic of China

*twp@mail.ntou.edu.tw

Abstract

The main goal of the proposed system is to enhance the mutual correlation of secondary structure element (SSE) pairs for multiple structure alignment. The algorithm utilizes the local matching advantages through distance matrix approach to extract suitable candidates of SSE pairs. The similarity scores of compared distance matrices of mutual SSE pairs were calculated and ranked to decide representative segments as key anchors for multiple structure alignment. This strategy solves the misalignment problems caused by large number of SSEs in a protein structure in previous system. In this study, the experimental results showed the hitting scores are evidently increased in those proteins which possess low sequence identity.

Keywords: multiple structure alignment, secondary structure, geometrical correlation, distance matrix.

1. Introduction

The increasing growth of experimentally determined protein structures creates big challenges for comparing the similar structures. Up to date, protein data bank (PDB) contains 52402 determined structures[1], of which most non-redundant 34494 PDB entries are collected and classified in SCOP[2] and 30028 in CATH[3]. The SCOP aims to provide a detailed and comprehensive description of the structural and evolutionary relationships between all proteins and it is constructed manually by visual inspection and comparison of structures, and the CATH utilizes both automatic and manually classification techniques to construct a hierarchical classification of protein domain structures containing class(C), architecture (A), topology (T) and homologous superfamily (H) level. Most of proteins possess structural similarities with other proteins and, in some of these cases, share a common evolutionary origin. To discover the knowledge of these relationships is crucial to understand the protein evolution and development. Therefore, an efficient and effective structure alignment method is always urgently required for protein structure comparison.

The current methods of protein structure alignment can separate into pair-wise alignment and multiple alignments. Only a few pair-wise alignment methods were described here such as Dali which uses three-dimensional coordinates of each protein to calculate residue-residue distance matrices[4], FAST which employs a directionality-based scoring scheme to compare the intra-molecular residue-residue relationships in two structures[5],

and SuperPose which generates sequence alignments, structure alignments, PDB coordinates, RMSD statistics, difference distance Plots, and interactive images of the superimposed structures[6].

Multiple structure alignment analyzes the protein structures in spatial domain to find out the similar or individual characteristics. Several multiple methods have been proposed such as MASS which exploits the geometric method and clustering approaches on the basis of secondary structure information[7], SSM which removes the secondary structure elements with low heuristic score iteratively[8], and MultiProt which detects structurally similar segments for parts of input molecule using the geometric hashing method[9]. In addition to the previous introduced multiple structure alignment methods, a mechanism based on aligning mutual correlation of secondary structure element pairs was proposed[10], which focused on the geometrical positions and mutual relationship of secondary structure elements instead of evaluating residue contents. The method is powerful for comparing the proteins possessing with homologous structural but low sequence similarity. However, the mutual correlation of secondary structure elements based approach for multiple structure alignment holds a server misalignment problem when the structures under analysis possessing huge number of secondary structure elements. To overcome such problem, this study exploits the mutual correlation of amino acids in each second structure element as distance matrix and utilizes a local region search to identify similar SSE pairs. Through the comparison of the mutual relationship of SSE pairs, the prime three selected SSE pairs in each structure will be identified and performed as key anchors for constrained multiple structure alignment. According to the aligned results, our system can successfully solve the misalignment problem in its previous version, and confidently approve the application of mutual correlated SSE analysis being a better solution for low protein sequence similarity.

2. System architectures

Here we present a novel system configuration for multiple protein structure alignment and the system flowchart is depicted in Fig. 1. The standard definitions of secondary structure alignment of input protein data are obtained by Definition of Secondary Structure of Proteins, DSSP (kabsch and sander, 1983). In the proposed system, five major parts including vector transformation, intra-relationship analysis, target protein determination, inter-relationship analysis, and constrained structural alignment are developed. Firstly, it exploits a fast vector transformation technique to represent an SSE in its corresponding vector format. Secondly, the mutual geometrical relationship, angle, distance, length, and distance matrix among vectors are calculated as fundamental characteristics inside a protein structure. In the third component, a target protein determination process is performed by choosing a central protein which is the most structurally similar to all the others by local search method. After a target protein is decided, inter-relationship analysis compares the measuring scores of distance matrices among proteins and generates three SSE pairs with the highest scores. At the last component, based on this three key anchor segments, translation and rotation transformations are performed to obtain the experimental result.

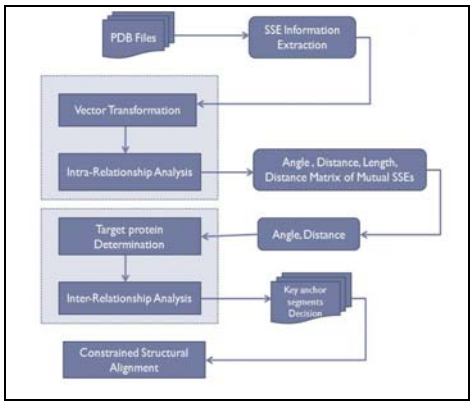

Fig. 1: the system flowchart for multiple structure alignment based on the mutual relationship of secondary structure element pairs.

3. Modules description and algorithm

3.1 Vector transformation

For reducing computational time complexity, a three dimensional regression technique is applied to transform SSEs into three dimensional regression vectors. The related C-α coordinates of a defined SSE segment are extracted from PDB files as fundamental data and transformed the data point set into a vector by creating three dimensional regression vectors from specified amino acids coordinates[10].

3.2 Intra-relationship analysis

Similar protein structures are composed of analogous SSEs with respect to their relative geometrical positions. Therefore, structure invariance becomes a reliable factor to analyze the structural similarity of SSEs among various protein structures. Every two SSEs in a protein structure form a SSE pair. Each SSE pair holds some attributes which include angle, distance, length, type combination and distance matrix. Angle is calculated from arccosine function and distance is measured as Euclidean distance of the centers of mass (CMs) of two SSE vectors (Fig. 2). Type combination includes helix to helix, helix to sheet, sheet to helix, and sheet to sheet. Distance matrix is constructed from the distance relationship among amino acids in an SSE pair. For instance, as in Fig. 3, SSE_i possesses seven amino acids and SSE_j for five. Hence, a 7x5 distance matrix is formulated for similarity evaluation. All the possible intra-relationships between two SSEs in a protein structure can be calculated in the similar processes.

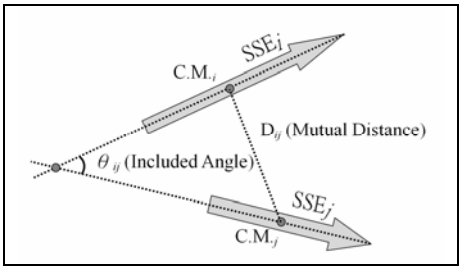

Fig. 2: The correlations of two pairs.

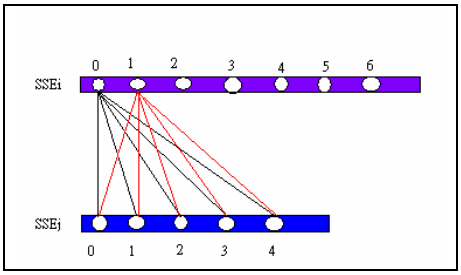

Fig. 3: The cross-distance relationship among amino acids in two different SSEs.

3.3 Target protein determination

This module determines a protein by choosing the most structurally similar protein among proteins of interested, and the others ready-to-align proteins will be aligned to the defined target protein respectively. The function of mutuality for each protein is defined as:

$$S_i = \frac{mutuality_i}{total_i} \quad (1)$$

, where the "*mutuality*" stands for the number of SSE pairs which contains SSE pairs of various proteins in a local region. The "*total*" parameter represents the total number of SSE pairs in a protein. If the score of a protein is high, the protein can be considered as high similarity to the others. Hence, the protein with highest mutuality score will be chosen as the target protein.

3.4 Inter-relationship analysis

The purpose of this module is to select three key anchor segments as constrained features. The flowchart of inter-relationship analysis is showed in Fig. 4. Firstly, each SSE pair represents a coordinate point in its angle-distance map. Regarding the SSE pairs in a target protein as a center point, it extends a local region to get the similar candidates from various proteins. Subsequently, the system compares the difference between the SSE pairs in target and query proteins through distance matrix matching and mutually matched scores are ranked in order. The mutual comparison of distance matrices has three different kinds which are showed in Fig. 5. The first type consists two distance matrices with different dimension that the size of SSE pair in the query protein is smaller than in the target protein both in width and length; the second type is the target protein and the query protein possess identical dimension of distant matrices; the last type represents the distance matrices within different dimension either in width or in length. When performing comparison, the system calculates the matching scores of all possible combinations and sorted in a ranked order. The matching score function is defined as:

$$RS = \frac{\sum_{i=x}^{i=x+M} \sum_{j=y}^{j=y+N} |T(i,j) - Q(i,j)|}{MN} \quad (2)$$

, where T is the SSE pair in the target protein, Q is the SSE pair in the ready-to-align protein, M is the width of distance matrix, and N is the length of distance matrix. The minimum matching score stands for the best fitting location between two SSE pairs. Hence, the best two matched candidates are chosen for ranking score accumulation. Finally, this module provides the final three best constrained features for further processes.

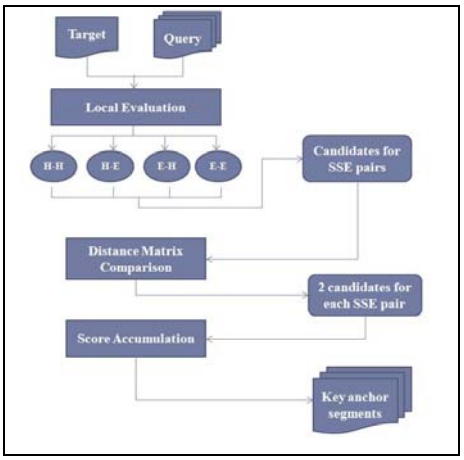

Fig. 4: The flowchart of inter-relationship analysis.

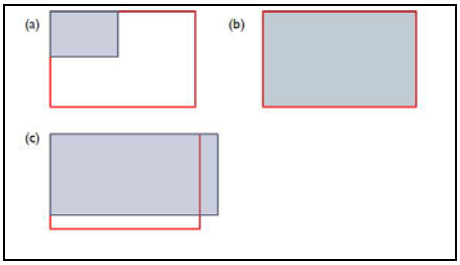

Fig. 5: Three different relations between two distance matrices of two SSE pairs.

3.5 Constrained structural alignment

Once the three key SSE pairs from the target and query proteins are identified, the alignment module is designed to extract three key pivot points (or residues) which are able to provide sufficient information for structure alignment from the three constrained features. All possible combinations of the three key anchors from two proteins are analyzed and the ones with lowest root-mean-square-deviance will be selected as the optimum pivot points. (Su, 2006)[11].

4. Results

In the previous structure alignment system based on the mutual SSE pair matching mechanism, we have discovered that some examples didn't provide satisfied alignment results. Taking the six PDBs of the Human RNase domain as an example, the 1RNF and 2HKY possessed a low hitting numbers and depicted in Table 1. Employing the improved algorithms proposed in this paper, the misalignment problem is solved, especially for the target sequences possessing low sequence identity. To verify the degree of improvement, we picked 10 different kinds of domains to compare the proposed method and previous one. The results were showed in Table 2. The proteins possess lower than 20% in sequence identity. However, the results of multiple structure alignment by the proposed method achieved 78.89% hitting score in average.

5. Conclusions

In this study, we have proposed a novel system which provides an efficient and effective algorithm for multiple structure alignment. If the protein sequences possess lower sequence identity, the sequence-based alignment is unable to provide a satisfied alignment results, and the

prevoius system

1GQV:A (target:135)	1E21:A (119)	1DYT:A (133)	1RNF:A (120)	1B1I:A (123)	2HKY:A (129)
hits/RMSD	96/1.10	114/1.22	54/1.28	86/1.24	57/1.34
RMSD Average:1.234576 Aligned Residues Average:81.400002					

present system

2HKY:A (target:129)	1E21:A (119)	1GQV:A (135)	1DYT:A (133)	1RNF:A (120)	1B1I:A (123)
hits/RMSD	94/1.19	111/1.29	98/1.25	96/1.26	91/1.25
RMSD Average:1.246055 Aligned Residues Average:98.000000					

Table 1: The results of human RNase domain alignment.

PDB identifiers	previous system (hits/RMSD)	present system (hits/RMSD)	sequence identity
1e21:a 1gqv:a 1dyt:a 1rnf:a 1b1i:a 2hky:a	81.4 / 1.23	98 / 1.25	11.97%
1dci:a 1pjh:a 1hnu:a 1xx4:a	190 / 1.14	199.6 / 1.04	5.71%
1cpz:a 1k0v:a 1p8g:a 1sb6:a	40.67 / 1.35	52 / 1.3	15.07%
1jub:a 1ep3:a 1f76:a 1d3g:a 1uum:a	237.75 / 1.17	263.75 / 1.05	7.75%
1f3y:a 1ktg:a 1xsa:a	90.5 / 1.21	116 / 1.18	10.53%
1uby 1rqj:a 1rtr:a	211.5 / 1.18	221.5 / 1.14	12.04%
1gtd:a 1t4a:a 1vq3:a	55 / 1.3	61 / 1.22	12.36%
1m6k:a 1k38:a 1k55:a 1h8z:a	217.33 / 0.92	211.67 / 1.07	13.73%
1doz:a 1hrk:a 1lbq:a	295 / 1.16	286 / 1.2	15.66%
1ed8:a 1ew2:a 1k7h:a	404.5 / 1.01	398 / 1.08	16.22%

Table 2: Results of previous system and the proposed system based on distant matrix analysis.

structure alignment only could be achieved information. Furthermore, if the proteins under analysis possess huge number of secondary structure elements, false matching of SSE pairs between two protein structures become inevitable. Hence, the spatial information containing distance, angle, length, type combination, and distance matrix of each SSE pair

should be considered simultaneously. This information facilitates important data fitting in the progress of multiple structure alignment.

The occurrence of misalignment is mainly due to global search in the angle – distance blocks of SSE pairs. It can't be expected to find the best candidates when the similar SSE pairs are located in different blocks or when there exist too many noisy pairs. The strategy in this paper is to apply local matching mechanism to match the target protein with similar candidates in its local region. The results have shown that the misalignment problem could be solved and the mutual correlated SSE analysis plays a better solution for dissimilar protein sequences.

6. References

[1] http://www.rcsb.org/pdb/home/home.do

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

[3] Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B., and Thornton, J.M, "CATH- A Hierarchic Classification of Protein Domain Structures," *Structure*. Vol 5. No 8. P.1093-1108, 1997.

[4] L Holm, C Sander, "Protein structure comparison by alignment of distance matrices," *J Mol Biol 233(1): 123-38,*1993.

[5] Zhu J, Weng Z, "FAST: a novel protein structure alignment algorithm," *proteins*, 58(3):pages 618 – 627,2004.

[6] Rajarshi Maiti, Gary H. Van Domselaar, Haiyan Zhang, and David S, "SuperPose: a simple server for sophisticated structural superposition," *Nucleic Acids Res. 2004 July 1*, 2004.

[7] Dror, O., et al., "MASS: multiple structural alignment by secondary structures," *Bioinformatics*, 19:p.i95-i104, 2003.

[8] Krissinel, E. and K. Henrick, "Mutiple Alignment of Protein Structures in Three Dimensions," *Computational Life Sciences: First International Symposium, CompLife,* 2005.

[9] Shatsky, M., R. Nussinov, and H. Wolfson, "A Method for Simultaneous Alignment of Multiple Protein Structures," *Proteins*, **56**(1):p.143-156, 2004.

[10] Chang R-H, Wang L-J, Chen J-M, Pai T-W, "Enhanced mutual Correlation of secondary structure elements for multiple structure alignment," *Pro Of the Joint Conference on Information Sciences,*2007.

[11] Su B-H, Pai T-W ,"Constrained multiple protein structure feature alignment for unique pattern detection," 2006.