Randomized Algorithms for Three Dimensional Protein Structures Alignment

Yaw-Ling Lin* Ying-Hung Lin Po-Shun Yu Hsun-Chang Chang

Department of Comput. Sci and Info. Engineering, College of Informatics and Computing, Providence University, Shalu, Taichung County, Taiwan 433.
yllin@pu.edu.tw, {g9234020,peteryu,hcchang}@cs.pu.edu.tw

Abstract

The explosive growth of genetic sequence information has offered us comprehensive collections of the protein sequences found in many living organisms. The challenge of understanding these gene products has led to the development of functional proteomics methods, which collectively aim to imbue the raw sequence with biological understanding. Protein structure provides the opportunity to recognize homology that is undetectable by sequence comparison, and it represents a powerful means of discovering functions, yielding direct insight into the molecular mechanisms. Currently, there are several techniques available in attempting to find the optimal alignment of shared structural motifs between two proteins.

In this paper, we propose novel algorithms for pairwise alignment of protein structures. Methods of locating suitable isometric transformations of one structure and aligning it to the other structure are addressed. Our methods allow sequence gaps of any length, reversal of chain direction, and free topological connectivity of atom sequences. Sequential connectivity can be imposed as an option. The method is fully automatic to identify structural resemblances and common structural cores accurately and sensitively, even in the presence of geometrical distortions.

Keywords: proteomics, computational molecular biology, computational geometry, algorithms, structure alignments and comparisons.

1 Introduction

Protein structure alignment techniques have grown increasingly important as a means to quantitatively compare and classify all known protein structures. The number of structures in the Protein Data Bank [2] is currently (as of Jan 2005) more than 29,101. One of the primary goals of structural alignment programs is to quantitatively measure the level of structural similarity between all pairs of known protein structures. This data can provide several meaningful insights into the nature of protein structures and their functional mechanisms. For instance, the comparison of all structures against each other can show relationships, both functional and structural, between proteins that were previously not known to be related [8]. In addition, structure based distance measures are critical to constructing accurate phylogenies of proteins and classifying structures into families that share similar folds or motifs. Identifying these shared structural motifs using structural alignment techniques can provide significant insight into the functional mechanisms of the protein family.

There have been several methods proposed to compare protein structures and measure the degree of structural similarity between them. These methods have been based on alignment of secondary structure elements as well as alignment of intra and inter-molecular atomic distances [1, 5, 7]. The basic ideas are rapid identification of pair alignments of secondary structure elements, clustering them into groups, and scoring the best substructure alignment. For examples, the VAST system is based on continuous distribution of domains in the fold space. The FSSP/DALI system provides two levels of description – a coarse-grained one and one with a fine-grained resolution. The method, CATH, provides the complete PDB fold classification by domains and links to other sources of information. The two methods, CE and LGscore2, are based on a different idea. They focus on the local geometry rather than global features such as orientation of secondary structures

*Corresponding author. The work is supported in part by the National Science Council, Taiwan, R.O.C, grant NSC 93-2213-E-126-006.

and overall topology (as in the case of VAST or DALI) [3, 6, 9, 14, 17].

Our objective in this paper is to calculate the significance of score (rmsd) between spatial arrangements of Cα atom of protein backbone that are not necessarily adjacent in sequence. By matching the backbone Cα atoms between two sets of atoms, the algorithm can obtain lower (rmsd) scores comparing to these existed protein structure alignment systems like VAST or DALI.

2 Method

In this paper, methods of locating suitable isometric transformations of one structure, and align it to the other are addressed. We start with description of our general scheme for finding suitable isometric transformation from one configuration to the other. First, several initial setting of different *orientations* (transformations) of the point sets are uniformly distributed over the unit sphere. Each orientation, or *probe*, can be thought as a way of matching (super-positioning) one set of points upon the other. By finding the minimum bipartite matching between these two sets of points at the current orientation, we find a suitable matching between two configurations.

The match is served as an alignment setting; based upon the matching, the *rmsd* score can be calculated as a distance / difference measurement of the current orientation. The tricky part is to find a *good* orientation such that the associated rmsd score is small enough. Our method is to maintain a set of feasible (nice) orientations with associated rmsd scores. By randomly *perturbing* these orientations, some scores of the orientations can be improved step by step; finally, the orientation with smallest rmsd score is selected as the final score for the two given point sets.

Our method allows sequence gaps of any length, reversal of chain direction, and free topological connectivity of aligned segments. Sequential connectivity can be imposed as an option. The method is fully automatic and identifies structural resemblances and common structural cores accurately and sensitively, even in the presence of geometrical distortions.

2.1 Protein (molecular) structure distances, similarities, and scoring functions

We briefly explain the idea of the smallest root mean squared deviation (*rmsd*); it is a least-squares fitting method for two sequences of points, and was developed by several persons independently [15, 10]. The idea is to align atom vectors of the two given (molecular) structures, and use the common least averaged squared errors as a measurement of differences between these two (paired) sequences.

Let $P = \langle p_1, \ldots, p_n \rangle$ and $Q = \langle q_1, \ldots, q_n \rangle$ be two sequences of points. We assume that P is translated so that its centroid ($\frac{1}{n}\sum_{k=1}^{n} p_k$) is at the origin. We also assume that Q is translated in the same way. For each point or *vector* x, let $(x)_i (i = 1, 2, 3)$ denote the i-th (X,Y,Z) coordinate value of x, and $\|x\|$ denotes the length of x. Let $d(P, Q, R, \mathbf{a}) = \sqrt{\frac{1}{n}\sum_{k=1}^{n} \|Rp_k + \mathbf{a} - q_k\|^2}$ where R is a rotation matrix and $\mathbf{a}$ is a translation vector. Then, the *rmsd* value $d(P, Q)$ between P and Q is defined by $d(P, Q) = \min_{R,\mathbf{a}} d(P, Q, R, \mathbf{a})$. Although complicated as it might appear, the optimal rotation matrix and translation vector can be found simultaneously in $O(n)$ time. Schwartz [16] showed that $d(P, Q, R, \mathbf{a})$ is minimized when $\mathbf{a} = 0$ and $R = (A^t A)^{\frac{1}{2}} A^{-1}$ where the matrix $A = (A_{ij})$ $(i, j = 1, 2, 3)$ is given by $A_{ij} = \sum_{k=1}^{n} (p_k)_i (q_k)_j$, $A^{\frac{1}{2}} = B$ means $BB = A$, and $\mathbf{o}$ denotes the zero vector . Thus, $d(P, Q)$, R and $\mathbf{a}$ can be computed in $O(n)$ time.

Note that there must be an atom-pairing scheme before one can do the *rmsd* computation. The first atom of the first selection is compared to the first atom of the second selection, fifth to fifth, and so on. Usually, most existed protein alignment algorithms use *rmsd* to calculate the averaged squared different distances between Cα atoms of two protein backbones. Through *rmsd*, we can find the similarity between two protein structures. The *rmsd* algorithm is used by VAST,CE, and many other packages as the final refined measurement step. The trick, though, is how these algorithms to identify the suitable paired atoms selected from the two given structural elements.

2.2 Finding a suitable rigid transformation for matching structures

The main idea of our algorithm for finding a suitable matching between two sets of points before utilizing the RMSD procedure to fine-tune the final result is by space perturbation and the minimum bipartite matching between two sets. Let $P' = T \circ P$, and Q being translated to Q' such that the mass center of Q' is located at the origin. We construct a weighed graph $G = (V, E)$ with V being labelled with points of P' and Q', and each (p, q) in E being weighted the Euclidean 3D distance, for example, $w(p, q) = \|p, q\|$. We then solve the weighted minimum bipartite matching problem [4] to obtain the best matching of P' and Q'. After the matched pairings, we perturb and

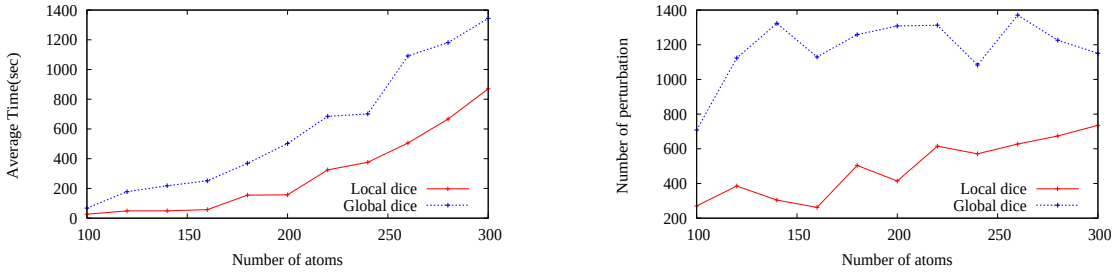

Figure 1: The average execution time and number of perturbations for alignment of two structures.

refine the final alignment by applying the algorithm MB-ALIGN and PERTURB to obtain lower *rmsd* ; for brevity of presentation, the detailed descriptions are omitted in the extended abstract.

For perturbing an orientation to its neighborhood, we need a formulation that provides the suitable transformation. Let p, q be two points on the (unit) sphere. The rotation matrix $M = R(p, q)$ rotates p to q such that the *rotation path* from p to q by M being the shortest. Here we describe how this matrix can be obtained. Let $r = p \times q$ (the *cross product* of p, q) be the normal vector, and θ be the angle of $\angle poq$ (o being the center of circle). Now let $c = \cos\theta, s = \sin\theta$, and let the coordinates of r be (x, y, z). It is interesting to know that the rotation matrix $M = R(p, q)$ can be calculated by the following formula: $M =$

$$\begin{bmatrix} c + x^2(1-c) & xy(1-c) - zs & xz(1-c) + ys \\ xy(1-c) + zs & c + y^2(1-c) & yz(1-c) - xs \\ xz(1-c) - ys & yz(1-c) + xs & c - z^2(1-c) \end{bmatrix}$$

Thus, for each orientation s, we can (uniformly) randomly pick one of its neighboring points, say s', such that the original rotation matrix is now composed with the rotation $R(s, s')$.

Note that MB-ALIGN uses PERTURB to rotate the Cα atoms of a protein, and performs minimum weighted bipartite matching to find good choices of atoms pairing between two structures before performing the refinement RMSD. First, the algorithm MB-ALIGN constructs a set S containing a set of orientations of the points set by the INIT-S procedure. INIT-S produces a set of orientations (rotation matrices) that are equally distributed upon a unit sphere. These orientations are to applied upon atoms of the structure before matching to the other structure.

Each orientation is called a *seed*, s; the set of seeds is denoted by the set S. For evenly distributing seeds upon a sphere, there are actually only five kinds of such initial seeding possible (such that $|S| = 4, 6, 8, 12, 20$ seeds.) Note that each $s \in S$ not only denotes the rotation matrix mat[s] but also is associated with a real value rms[s], representing the *badness* of the orientation. Through the PERTURB procedure, better orientations would be gradually obtained when better matchings are found. For each perturbed seed, MB-ALIGN finds the minimum bipartite matching, MBM, to decide the points pairing between point sets. Once it observes an improved seed s (smaller rms[s]), the seed is then put back to S. The algorithm stops either when a sufficiently small rms[s] is observed or when no further improvement is possible.

3 Experiments and Results

We have implemented these algorithms as several independent C programs and performed several experiments for validating methodologies discussed in the previous section. In implementing our system, we adapt the LEDA [13] package to perform the minimum weighted bipartite matching; the minimum bipartite matching algorithm is implemented by Dijkstra's algorithm with heuristics. In the worst case, the time complexity of this algorithm is $O(n(m + n \log n))$ [13].

For calculation of root mean square deviation (rmsd), we make use of the open source licensed software, PROFIT [11]; the system is designed to be the ultimate protein least squares fitting program. Adapting the McLachlan algorithm [12] in fitting points, PROFIT has many features including flexible specification of fitting zones and atoms, calculation of *rms* over different zones or atoms, and *rms*-by-residue calculation, and so on.

Experiments are performed as the following. First, a points set, P, of size varying from 50 to 1,000 are randomly generated as the tested case. The points set is then rotated and translated randomly to another set Q with the original ordering being randomly permuted. This structure alignment system shall be able to find the suitable re-

versed transformation so that the resulting *rmsd* ≈ 0. The experiment shows that the randomized MB-ALIGN algorithm does correctly discover the original pairing in most of the cases. Two versions of perturbation mechanism have been tested for the efficiency of the searching strategy. One way to perturb these 3D points is to uniformly rotate all orientations in S (*global dice*), while the other is to let each seed $s \in S$ retain its own rotated orientation (*local dice*). These experimental results are summarized in Figure 1; it clearly shows that the local, distributed, perturbation method is a more efficient way in finding the correct matchings.

Paired moleculars $(M_1 : M_2)$	VAST *rmsd* (A)	Improved *rmsd* (B)	Improved ratio (%) $(A-B)/A$
101M:2DHB-B	1.67	1.62	2.84
101M:1CH4-A	1.47	1.44	2.58
1MLL:1HLM	2.16	2.08	4.14
102M:1SPG-A	1.67	1.61	4.04
1SPG-A:1H1X-A	1.76	1.71	2.92
1SPG-A:1SCT-A	2.16	2.12	1.89
3HHB-A:1RSE	1.69	1.64	3.12
3HHB-A:1HRM	1.82	1.76	3.06
2DHB-A:1RSE	1.69	1.64	2.87
1OUT-A:1MOC	1.79	1.73	3.58
1OUT-A:1CH2-A	1.77	1.71	3.50
1H1X-A:1CH4-A	1.63	1.61	1.24

Table 1: Improvement ratios of our algorithm.

Other experiments also demonstrate that our structure alignment algorithms find better results comparing to some available structure comparison methods; e.g., NCBI's Vector Alignment Search Tool (VAST) [6]. Several sets of real protein structures are randomly picked from the PDB [2] for comparing the effectiveness of our algorithms. Among them, the first two protein structures we picked were the horse haemoglobin (PDB ID: 2DHB) and the sperm whale myoglobin (PDB ID: 101M). These are two of protein structures that were first solved. These two were recognizable as homologous even at low resolution, even though their sequences were more different than similar. While the VAST reports the 101M:2DHB pairing having the rmsd 1.67, our spatial minimum bipartite match algorithm immediately recognizes a better matching with rmsd 1.62, about 2.8% improvement. Many other similar improvements have been observed through the experiments; some of these improvements are summarized in Table 1.

Detailed experimental results, including thousands lines of C source code implementations in UNIX system and many parameter settings, can be obtained through e-mail request to the corresponding author.

References

[1] D.W. Barakat and P.M. Dean. Molecular structure matching by simulated annealing, iii. the incorporation of null correspondences into the matching problem. *J. Comp. Aided Mol. Design.*, 5:107–117, 1991.

[2] H.M. Berman, J. Westbrook, Z. Feng, et al. The protein data bank. *Nucleic Acids Research*, 28:235–242, 2000.

[3] S. Cristobal, A. Zemla, D. Fischer, L. Rychlewski, and A. Elofsson. A study of quality measures for protein threading models. *BMC Bioinformatics*, 2:5, 2001.

[4] Z. Galil. Efficient algorithms for finding maximum matching in graphs. *ACM Computing Surveys*, 18:1:23–38, 1986.

[5] M. Gerstein and M. Levitt. Using iterative dynamic programming to obtain accurate pair-wise and multiple alignments of protein structures. *In Proc. Fourth Int. Conf. on Intell. Sys. for Mol. Biol.* Menlo Park, CA: AAAI Press, pp 59-67, 1996.

[6] J.F. Gibrat, T. Madej, and S.H. Bryant. Surprising similarities in structure comparison. *Curr Opin Struct Biol.*, 6:377–385, 1996.

[7] L. Holm and C. Sander. Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.*, 233:123–138, 1993a.

[8] L. Holm and C. Sander. Structural alignment of globins, phycocyanins, and colicin. *FEBS Lett.*, 315:301–306, 1993b.

[9] L. Holm and C. Sander. Touring protein fold space with DALI/FSSP. *Nucleic Acids Res.*, 26:316–319, 1998.

[10] W. Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta. Cryst.*, A32:922–923, 1976.

[11] A.C.R. Martin. `http://www.bioinf.org.uk/software/profit/`.

[12] A.D. McLachlan. Rapid comparison of protein structres. *Acta Cryst*, A38:871–873, 1982.

[13] K. Mehlhorn and St. Naher. *The LEDA Platform of Combinatorial and Geometric Computing.* Cambridge University Press, 1999.

[14] C.A. Orengo, A.D. Michie, S. Jones, D.T. Jones, M.B. Swindells, and J.M. Thornton. CATH – a hierarchical classification of protein domain structures. *Structure*, 5:1093–1108, 1997.

[15] S.T. Rao and Rossmann M.G. Comparison of super-secondary structures in proteins. *J. Molecular Biology*, 76:241–256, 1973.

[16] J.T. Schwartz and M. Sharir. Identification of partially obscured objects in two and three dimensions by matching noisy characteristic curves. *Int. J. Robotics Research*, 6:29–44, 1987.

[17] I.N. Shindyalov and P.E. Bourne. Protein structure alignment by incremental combinatorial extension (ce) of the optimal path. *Protein Eng.*, 11:739–747, 1998.