A Review on Linear encoding methods and Geometry based methods for protein structure alignment

Samira Fotoohifiroozabadi, Safaai Deris* and Mohd Saberi Mohamad

Artificial Intelligence and Bioinformatics Group, Faculty of Computing, Universiti Teknologi Malaysia, 81310 UTM Skudai, Johor, Malaysia.
samira.fotoohi@gmail.com, safaai@utm.my, saberi@utm.my

Abstract: In this work the linear encoding methods and geometry based methods and their applications in protein structure comparison analysis are presented and compared based on their time complexity and accuracy. The comparison results shows that linear encoding methods obtains higher running speed but with lower accuracy, while geometric based methods are much more accurate, but with slower speed.
[Samira Fotoohifiroozabadi, Safaai Deris and Mohd Saberi Mohamad. **A Review on Linear encoding methods and Geometry based methods for protein structure alignment.** *Life Sci J* 2014;11(3s):202-206]. (ISSN:1097-8135). http://www.lifesciencesite.com. 30

Keywords*:* protein structure alignment, linear encoding methods, geometry based methods

1.0 Introduction

An equivalence map between amino acid residues in different structures based on their relative position in a three dimensional space is a key step in protein structure analysis and is defined as structural alignment. Although, the comparison of a protein structure with other structures of the same or similar proteins is a difficult task but it reveals differences and similarities between related molecules and allows inferring how functional properties are implemented. The accuracy of this comparison may depend on the method or program used as well as what the user is trying to accomplish. Up to now, several methods are proposed and implemented to perform the comparison between protein structure and are categorized in different ways. As an example of two important categories, linear encoding methods and geometry based methods are compared in this paper.

Transforming 3D protein structural data into one-dimensional (1D) text strings or numerical series is defined as linear encoding method. This transformation converts complicated geometric problems of structural superimpositions to sequence comparison therefore the comparison can be done much easier.

On the other hand, geometric based methods refer to compare geometrical coordinates of the □□ backbone atoms in order to find the best optimal correspondence between residue pairs. In this work the existing methods and their applications in both categories are presented and compared based on their time complexity and accuracy. The comparison results shows that linear encoding methods obtains higher running speed but with lower accuracy, while geometric based methods are much more accurate, but with slower speed.

2.0 Geometry Based Method

Using heuristic strategies some algorithms are developed to compare geometrical coordinates of the Cα backbone atoms to find the best optimal correspondence between residue pairs. They are evaluated based on the extent of structural similarity that is recognized, where longer alignments and better rigid body superposition means better performance. A better score according to other geometric measures is also another examination to realize the performance of algorithm.

However, there are different algorithms designed for various applications (Holm and Sander 1993; Gibrat et al. 1996; Shindyalov and Bourne 1998; Ortiz et al. 2002; Krissinel and Henrick 2004; Zhang and Skolnick 2005; Kolbeck et al. 2006), none of the existing methods provide a complete solution for protein structure comparison. Typically, all the methods agree relatively well on the alignment of highly similar structures whereas they often disagree over details of low similar proteins.

In this study the most three famous methods are presented and the comparison is made based on these methods. The techniques include distance matrices comparison (DALI) (Holm and Sander 1993), combinatorial extension (CE) (Shindyalov and Bourne 1998), dynamic programming on TM-score rotation matrix (TM-align) (Zhang and Skolnick 2005).

2.1 DALI and DaliLite, Distance Matrix Alignment

DALI (Holm and Sander 1993) is based on computing the inter-residue distance matrix using 3D-coordinates of protein. In this method, each distance matrix is divided into fragments of hexa-peptide and, common local sub-matrices within the fragments of distance matrices are searched and after

retrieving the common fragments they are merged into larger overlapping segments.

By applying Monte Carlo algorithm, the aligned fragments are further optimized but the convergence to the global optimal correspondence is not guaranteed therefore, some alternative optimized alignments are applied. DALI is fully automatic and allows any length of gaps, reversal in direction of the chain, and free of topological connectivity among aligned segments.

DaliLite which is an independent package of the DALI algorithm has been implemented by Holm (Holm and Park 2000). Except the complicated database update protocol, DaliLite consists of all the functionality of the Dali server at European Bioinformatics Institute (EBI). Recently, DaliLite V.3 (Holm et al. 2008) is running which has some features for searching through database and updates also, throughput of the server and portability of the system is improved using the new protocols.

Even the graphical representation of the alignment is available in DaliLite V.3 but the alignment results are very difficult to be interpreted. Although the processing of the queries is fast, but sometimes the server is very busy. Structures with Z-score > 2 is returned by the server. DALI loses finding one similarity, which it finds in another query (Novotny et al. 2004).

2.2 CE, Combinatorial Extension of the Optimal Pathway

Using combinatorial extension (CE) of the alignment path among fragment pairs (Shindyalov and Bourne 1998), CE is supposed to optimally align two structures that satisfy certain constraints by considering structural similarity. After superposition of rigid body, root-mean-square deviation (RMSD) and inter-residue distances of the matched atoms are used to evaluate the similarity of the structures.

CE allows gaps, but it is restricted by a defined maximum size. A Z-score is computed for the optimal alignment as the significance measure. To this end, the probability of obtaining an alignment of the same length is calculated when comparing two structures. In CE, due to excessive traffic in network and server problem the waiting time for some queries is very long. It is also 10 times slower than secondary structure matching (SSM) method (Novotny et al. 2004). For homologous proteins, the CE mostly tends to give similar alignment with DALI (Mayr et al. 2007). In several studies CE and DALI are used for evaluating the performance of the other methods since they are considered as two standard tools for structural alignment of proteins.

2.3 TM-align, Protein Structure Alignment Based on the TM-score:

The convergence of dynamic programming algorithm is accelerated using TM-score which is a weighting scheme and a reasonable single measure to assess the quality of the alignment. The alignment is assessed by making a balance between length of alignment and accuracy according to (Zhang and Skolnick 2004):

$$TM-score = Max\left[\frac{1}{L_q}\sum_{i=1}^{L_n}\frac{1}{1+\left(\frac{d_i}{1.24\sqrt[3]{L_q-15}-1.8}\right)^2}\right]$$

where Lq is the length of query protein and Ln is the length of alignment and di is the distance between i-th pair of aligned residues. TM-score lies between (0,1] and noted that the higher value is better (Zhang and Skolnick 2004).

In order to improve the DALI and CE method's accuracy and speed, the TM-align was proposed by Zhang et.al (Zhang and Skolnick 2005). It defines TM-score as a novel measure to weight distance matrix and applies a dynamic programming algorithm to find best structural alignment. The server has a primary user interface to search in the database and provides a simple output for the alignment result.

3. Linear Encoding

The complexity of structure comparison and alignment problem has led researchers to use summarized representation of protein structure in their algorithms. Recently, several approaches are developed for linear encoding of protein 3-D structure in sequences of alphabets or other codes (Martin 2000; Guyon *et al.* 2004; Carpentier *et al.* 2005; Tung *et al.* 2007; Lo *et al.* 2007; Bauer *et al.* 2009; Budowski-Tal *et al.* 2010; Razmara *et al.* 2012). In this study three more famous methods are presented which includes: Structural similarity search by Ramachandran codes (SARST) (Lo et al. 2007) representing discrete internal angles of protein backbone as a sequence (YAKUSA) (Carpentier et al. 2005) kappa-alpha (κ, α) plot derived structural alphabet and BLOSUM-like substitution matrix (3D-BLAST) (Tung et al. 2007)

3.1. SARST, Structural Similarity Search Aided by Ramachandran Sequential Transformation

Based on Ramachandran map using nearest neighbor clustering, SARST maps structural

information of proteins into textual sequences. A regenerative scheme is also applied to make substitution matrices. A traditional sequence similarity search algorithm is used in the sequel so that the structural homologous proteins are retrieved (Lo et al. 2007). SARST obtained the best scores in terms of running speed and accuracy in comparison with YAKUSA, 3D-BLAST, and TOPSCAN (Martin 2000). But its accuracy is lower than CE and FAST methods. Its running speed is 18,000 and 240,000 times faster than FAST and CE methods, respectively (Lo et al. 2007).

In order to measure the protein structural similarity, iSARST which is the web server of SARST has provided an efficient search tool. Using two database searching tools, iSARST is implemented by integrating several structure comparison methods. For common structural homologs SARST and for homologs with circular permutations CPSARST is used. Once the target database is scanned by SARST/CPSARST, three traditional structure alignment methods including FAST, TM-align, and SAMO are used to refine and sorting the results.

Utilizing a multi-processor and batch-processing environment, the server obtains high running speed and high accuracy is achieved using refinement engines. A user friendly interface is provided by iSARST server and some options are available to define alignment details. The list of alignments and a functional summary of the best hits is included in the final output. An interactive and informative visualization tool is available at server to examine aligned structures.

3.2 YAKUSA, Fast Structural Database Scanning Method

YAKUSA (Carpentier et al. 2005) is a rapid program to search within a database of protein structures. The method describes protein backbone internal coordinates as encoded sequences and then applies a5-step algorithm to establish similarities between structures, where the first 3 steps are similar to BLAST (Altschul et al. 1990) steps: (1) making a deterministic finite automaton to represent all identical or similar patterns; (2) look for all patterns in the structures of database; (3) expand the patterns to find the longest common substructure (SHSPs); (4) choosing compatible SHSPs for all pairs of query and reference structure; and (5) scoring the matched SHSPs. YAKUSA obtains high accuracy as well as best related methods with a high running speed.

According to the evaluation study by Lo et al.(Lo et al. 2007) YAKUSA runs more than 35 and 4 times slower than SARST and3D-BLAST respectively but its running speed is more than 2300 and 170 times faster than CE and FAST.

Moreover, accuracy of the method in information retrieval assessment is similar to that of 3D-BLAST and SARST and lower than CE and FAST methods.

3.3 3D-BLAST, kappa-alpha plot extracted sequence of alphabets for Rapid Protein Structure Search

3D-BLAST (Tung et al. 2007) is a search tool within protein structure database that adopts features such as robust statistical bases, reliable and effective search abilities from BLAST (Altschul *et al.* 1990) which is a well-known sequence alignment tool. In 3D-BLAST, the protein structure is represented in kappa-alpha (κ, α) plot extracted sequence of alphabets. For searching within a database of structural alphabets, it uses a BLOSUM-like substitution matrix called structural alphabet substitution matrix (SASM).

In order to launch queries within the database, a simple user interface is provided by 3D-BLAST server. But occasionally the server is busy. The method is about 34,000 times faster than CE but with lower accuracy (Tung et al. 2007). 3D-BLAST performs almost the same as YAKUSA and SARST, but it is 9 times faster than YAKUSA and 26 times slower than SARST (Lo et al. 2007).

4. Comparison & Result

In this study we have compared three geometry based methods in terms of accuracy which are RMSD and length of alignment. In fact, the lower RMSD and the higher length of alignment lead to better alignment performance. Table 1 presents the comparison results of three geometry based methods which are CE, DALI and TM-ALIGN regarding to Structural alignments by different algorithms for 200 non-homologous PDB proteins.

Table 1. Comparison results of CE, DALI and TM-ALIGN based on accuracy

Geometry based methods	Length of Alignment	RMSD
CE	129.2	3.95
DALI	175.2	40
TM-ALIGN	165.8	4.44

As presented in Table 1, we can see that DALI has better length of alignment and CE has better RMSD performance between these three geometry based methods. Table 2 also shows the comparison between four linear encoding methods in terms of running speed considering 108 queries on a database of 34055 proteins.

Table 2. Comparison results of TOPSCAN, YAKUSA, 3D-BLAST and SARST based on running speed

Linear encoding methods	Average time per query (sec)
TOPSCAN	85.08
YAKUSA	35.6
3D-BLAST	9.07
SARST	0.34

As listed in Table 2, SARST has the best running speed among the other methods mentioned in the Table 2. In order to compare the geometry based and the linear encoding based methods, time and accuracy of each method as two features for testing the performance were considered. According to Tables 1 and 2, we selected TM-ALIGN and 3D-BLAST as candidates from geometry and linear based methods respectively, which they have intermediate performance related to others in their category.

Table 3-Comparision between geometry based and linear encoding methods in terms of accuracy and time regarding all-against-all comparison of 200 non-homologous proteins, considering all structure-pairs

	Category	Length of Alignment	RMSD	Time
TM-Align	Geometry	87.4	4.99	0.51
3D-BLAST	Linear encoding	65.7	6.69	0.002

As can we observed from table 3, TM-Align as a candidate from geometry based methods shows better length of alignments and RMSD against of 3D-BLAST which is a representative of linear encoding methods. But running speed of3D-BLAST is hundred times faster than TM-Align.

4.0 Conclusion

Linear encoding techniques adopted from these methods commonly reduce running time of the algorithms as they run hundreds of times faster than geometrical methods, however, 3D-structure conversion into 1D-sequence leads to lose some of the structural details of proteins. Consequently, these methods obtain lower alignment accuracy when compared to highly accurate geometrical search tools.

Acknowledgements

We would like to thank Universiti Teknologi Malaysia for sponsoring this research with a GUP Grant (Vot Number: Q.J130000.2507.04H34). This research is also funded by an e-science research grant (Grant number: 01-01-06-SF1234) from Malaysian Ministry of Science, Technology and Innovation.

Corresponding Author:

Safaai Deris,
Artificial Intelligence and Bioinformatics Group,
Faculty of Computing, Universiti Teknologi
Malaysia, 81310 UTM Skudai, Johor, Malaysia.
E-mail: safaai@utm.my

References

1. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., Basic local alignment search tool. Journal of molecular biology 1990; 215(3): 403-410.
2. Bauer, R.A., Rother, K., Moor, P., Reinert, K., Steinke, T., Bujnicki, J.M., Preissner, R., Fast structural alignment of biomolecules using a hash table, n-grams and string descriptors. Algorithms 2009; 2(2): 692-709.
3. Budowski-Tal, I., Nov, Y., Kolodny, R., FragBag, an accurate representation of protein structure, retrieves structural neighbors from the entire PDB quickly and accurately. Proceedings of the National Academy of Sciences 2010; 107(8): 3481-3486.
4. Carpentier, M., Brouillet, S., Pothier, J., YAKUSA: a fast structural database scanning method. Proteins: Structure, Function, and bioinformatics 2005; 61(1): 137-151.
5. Gibrat, J.-F., Madej, T., Bryant, S.H., Surprising similarities in structure comparison. Current opinion in structural biology 1996; 6(3): 377-385.
6. Guyon, F., Camproux, A.-C., Hochez, J., Tufféry, P., SA-Search: a web tool for protein structure mining based on a Structural Alphabet. Nucleic acids research 2004; 32(suppl 2): W545-W548.
7. Holm, L., Kääriäinen, S., Rosenström, P., Schenkel, A., Searching protein structure databases with DaliLite v. 3. Bioinformatics 2008; 24(23): 2780-2781.
8. Holm, L., Park, J., DaliLite workbench for protein structure comparison. Bioinformatics 2000; 16(6): 566-567.
9. Holm, L., Sander, C., Protein structure comparison by alignment of distance matrices. Journal of molecular biology 1993; 233(1): 123-138.

10. Kolbeck, B., May, P., Schmidt-Goenner, T., Steinke, T., Knapp, E.-W., Connectivity independent protein-structure alignment: a hierarchical approach. BMC bioinformatics 2006; 7(1): 510.
11. Krissinel, E., Henrick, K., Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. Acta Crystallographica Section D: Biological Crystallography 2004; 60(12): 2256-2268.
12. Lo, W.-C., Huang, P.-J., Chang, C.-H., Lyu, P.-C., Protein structural similarity search by Ramachandran codes. BMC bioinformatics 2007; 8(1): 307.
13. Martin, A.C., The ups and downs of protein topology; rapid comparison of protein structure. Protein engineering 2000; 13(12): 829-837.
14. Mayr, G., Domingues, F.S., Lackner, P., Comparative analysis of protein structure alignments. BMC Structural Biology 2007; 7(1): 50.
15. Novotny, M., Madsen, D., Kleywegt, G.J., Evaluation of protein fold comparison servers. Proteins: Structure, Function, and bioinformatics 2004; 54(2): 260-270.
16. Ortiz, A.R., Strauss, C.E., Olmea, O., MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. Protein Science 2002; 11(11): 2606-2621.
17. Razmara, J., Deris, S.B., Parvizpour, S., TS-AMIR: a topology string alignment method for intensive rapid protein structure comparison. Algorithms for Molecular Biology 2012; 7(4).
18. Shindyalov, I.N., Bourne, P.E., Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein engineering 1998; 11(9): 739-747.
19. Tung, C.-H., Huang, J.-W., Yang, J.-M., Kappa-alpha plot derived structural alphabet and BLOSUM-like substitution matrix for rapid search of protein structure database. Genome biology 2007; 8(3): R31.
20. Zhang, Y., Skolnick, J., Scoring function for automated assessment of protein structure template quality. Proteins: Structure, Function, and bioinformatics 2004; 57(4): 702-710.
21. Zhang, Y., Skolnick, J., TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic acids research 2005; 33(7): 2302-2309.

3/5/2014