Protein tertiary structure retrieval algorithms: An efficient method with co-occurrence matrix of oriented gradient of distance matrices

Rezaul Karim [†]
Md. Momin Al Aziz[†]
Farhana Zaman[†]
Salman Kader[†] and
Md. Abul Kashem Mia[†]

[†] Department of Computer Science and Engineering,
Bangladesh University of Engineering and Technology
rezaulnkarim@gmail.com,momin.aziz.cse@gmail.com,farhana27glory@gmail.com
salmankaderrakin@yahoo.com,kashem@cse.buet.ac.bd

Abstract. The long history of life science research has encountered enormous discoveries of proteins, resulting tremendous growth of database storing protein structures and relevant information. Protein tertiary structures have been observed to share hierarchical evolutionary relationships. Again, there is empirical hypotheses that protein functionalities are correlated much more with tertiary structure than amino acid sequence. Therefore, efficient methods for faster retrieval of similar tertiary structures from protein database is of great significance for biologists,pharmacists and life science researchers to study on protein structure to function relationship, function prediction of novel structure, research of evolutionary information and in drug discovery and disease diagnosis. Exact measure for similarity of tertiary structure of two molecule for is CSP problem and so have time complexity of factorial order. In traditional methods, tertiary structure similarity were measured by alignment distance of alpha carbon distance matrix which is much time consuming. In our proposed method, we showed that, L2 norm of co-occurrence matrix of oriented gradient of distance matrix performs much better than alignment score of alpha carbon distance matrix in both performance and accuracy for tertiary structure retrieval.

Keywords: protein tertiary structure retrieval, novel feature of finding similar protein, function prediction of novel structure

1 Introduction

The functionalities of proteins are correlated to its tertiary structure not primary. Thats why when a protein is denatured at hight temperature it loses tertiary structure, most of functionality while its primary structure may still be present.

Moreover, analysis and study of similar tertiary structure is of great importance in disease diagnosis, drug discovery and many other fields. Therefore efficient approach for retrieval of protein tertiary structure is of great significance to speed up the research of biochemists, pharmacists, biologists and researchers of many other fields.

Protein tertiary structure is a protein's geometric shape which describes the complex 3 dimensional orientation of the peptide chain. It is the spatial organization of an entire protein molecule or other macromolecule consisting of a single chain [3]. Basically proteins are constructed by long chains of amino acid residues folding into complex $3D$ polypeptide chain structures. This $3D$ representation of a residue sequence and the way this sequence folds in $3D$ space are very important to understand the function of the protein. Also evolutionary evidence can be derived from the conserved protein [9].

- The linear sequence of amino acids are called the *primary structure*.
- *Secondary* structure is the specific geometric shape caused by intramolecular and intermolecular hydrogen bonding of amide groups.
- *Tertiary* structure refers to the three-dimensional structure of an entire polypeptide.
- *Quaternary* structure is the spatial arrangement of its subunits which is composed of two or more polypeptide chains.

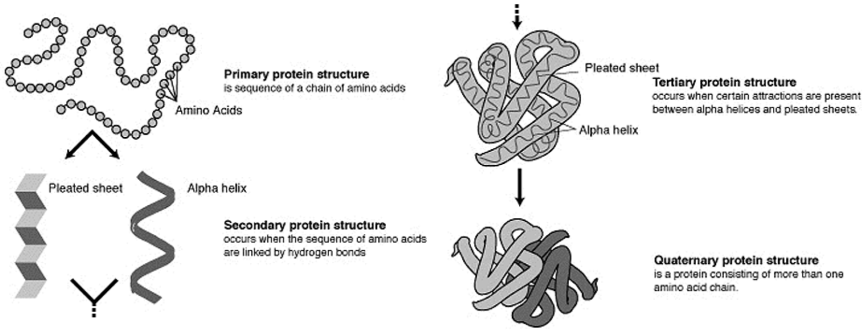

Primary protein structure
is sequence of a chain of amino acids
Amino Acids
Pleated sheet
Alpha helix
Secondary protein structure
occurs when the sequence of amino acids
are linked by hydrogen bonds
Pleated sheet
Tertiary protein structure
occurs when certain attractions are present
between alpha helices and pleated sheets.
Alpha helix
Quaternary protein structure
is a protein consisting of more than one
amino acid chain.

Fig. 1. All types of protein structure

Here we intend to work on the tertiary structure which reveals the folding of its secondary structural elements and specifies the position of each atom in the protein, including those of its side chains. Atomic coordinates of protein can be retrieved by X-Ray crystallography and nuclear magnetic resonance(NMR). This co-ordinate data for known proteins are stored on a world wide repository Protein Data Bank(http://www.wwpdb.org/).

The traditional way to compare structures, and the method that many structural alignment programs use, is to treat each one as a rigid three-dimensional object and superimpose one on the other. Differences are calculated using a least-squares method.

Most recently, alpha carbon distance matrix is widely used by many researcher as in [6] and [7] to map 3D coordinate data to 2D feature matrix to compare tertiary structure. This alpha carbon distance matrix resembles tertiary structure of a protein and secondary structure elements conserved in it. Alpha carbon distance matrix based exact solution is of order factorial N when matrix is in dimension NxN as the problem of aligning 2 matrix with minimum error is CSP problem. Matalign [1] proposed

$$O(N^4)$$

dp solution where N is the dimension of distance matrix and MASASW [7] proposed a sliding window based matrix alignment method with time complexity of

$$O(wWN^2)$$

where N is the dimension of distance matrix and w and W are window size. Still those comparison methods are computationally expensive as the database is huge. Our proposed solution is of

$$O(N^2)$$

time complexity where N is quantization bin size for gradient angle.

In our research, upon analyzing tertiary structure and alpha carbon distance matrix, we observed that not all data in alpha carbon distance matrix is equally important and co-occurrence matrix of oriented gradient of distance matrix is the most important feature to compare tertiary structure. Moreover, taking Euclidean distance of co-occurrence matrix of oriented gradient of distance matrix performs much better in accuracy than aligning the distance matrices and computationally more than 100 percent faster.

The rest of the paper is organized as follows: In section 2, some recent approaches for tertiary structure retrieval are briefly described. We have systematically formulated our problem in section 3. A precise description of our proposed solution is provided in section 4. Performance evaluation of our proposed algorithm is presented in section 6. Finally we have drawn a conclusion in section 7.

2 Existing approaches

2.1 DALI

DALI is global pair-wise structure alignment heuristic [4]. The traditional way to compare structures, and the method that many structural alignment programs use, is to treat each one as a rigid three-dimensional object and superimpose one on the other. Differences are calculated using a least-squares method. The Dali server uses a sum-of-pairs method, which produces a measure o f similarity by comparing intramolecular distances. Similarity is measured by Dali-Z scores. Structures that have significant similarities have a Z-score above 2, and usually have similar folds.

2.2 Wavelet based approach by Marsolo

Marsolo and Parthasarathy [6] presented two normalized, stand-alone representations of proteins that enabled fast and efficient object retrieval based on sequence or structure information [2]. For the range queries, they specified a range value r and retrieved all the proteins from the database which lied within a distance r to the query. In their work, they also used alpha carbon distance matrix to represent 3D structure as 2D matrix.

2.3 MatAlign

It is a two-step algorithm. Firstly, 3D protein structures are represented as 2D distance matrices, these matrices are aligned by means of dynamic programming in order to find the initially aligned residue pairs. Secondly, the initial alignment is refined iteratively into the optimal one, according to an objective scoring function. On the benchmark set of 68 protein structure pairs by Fischer et al., MatAlign provides better alignment results, according to four different criteria, than both DALI and CE in a majority of cases. MatAlign also performs well in structural database search as DALI does, and much better than CE does. MatAlign is about two to three times faster than DALI, and has about the same speed as CE.

2.4 MASASW

Matrix Alignment by Sequence Alignment [7] within Sliding Window also known as $MASASW$ is a faster approach than DALI, CE and MatAlign. They compared their approaches between themselves and with several existing algorithms, and they generally prove to be fast and accurate. They used alpha carbon distance matrix to represent 3D structure as 2D matrix. For comparison they aligned the distance matrices with using sliding window.

3 Problem formulation

For a given protein, the target is to retrieve proteins with nearest tertiary structure from database with ranked query. We divided the problem of protein tertiary structure retrieval into two sub problems and our proposed solution is organized accordingly. The sub problems are as follows:

3.1 Feature selection and comparison

The primary task is to formulate low level feature vector to represent high level features of tertiary structure. A good feature must have the property to have similarities in feature vector for similar structure and dissimilarity in feature vector for dissimilar structure. We also need to measure similarity or distance as a function of two feature vectors to compare two structures.

The exact solution to compare two 3D structure is CSP and thus time complexity is in factorial order.

3.2 Indexing database

Query to retrieve similar or nearest structure needs one vs. all comparison that is computationally expensive and the number of structure is increasing. So, we need to use hierarchical indexing of the database, based on similarity or distance metric so as to search in logarithmic order.

4 Our solution:

In this paper we focused on the first sub-problem, feature selection and comparison and propose our solution using these novel features.

4.1 Feature vector selection

With a view to define feature vector, the primary task is to select the features that uniquely represent the protein structure.

Cα-Cα distance matrix: Distance matrix of alpha carbons in residues is a good candidate to be chosen to transform 3D structure to 2D vector representation as shown in [7] and [6]. This distance matrix is the pairwise distance between all pair of alpha carbons in polypeptide chain. Proteins with similar tertiary structure will have similar distance matrix and proteins with dissimilar tertiary structure will have dissimilar distance matrix. The distance matrix also resembles secondary structure information altogether. If we consider the matrix as an monochromatic image, α-helices and parallel β-sheets will appear as dark lines parallel to main diagonal and antiparallel beta sheets appear as dark lines normal to main diagonal. This distance matrix feature also has a very appealing property that is invariance of translation, scaling and rotation of the protein.

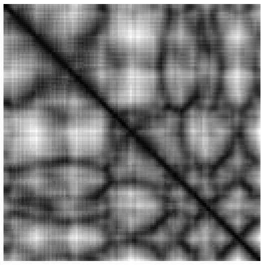


Fig. 2. Alpha carbon distance matrix as gray-scale image.

Again, as it is a two dimensional matrix and digital images are also two dimensional matrices; we have the opportunity to apply image processing and computer vision algorithms on it. Also, as graphs are also two dimensional matrices of adjacency structures, graph theory algorithms like graph isomorphism can also be applied to solve this problem with this feature. In this paper in our proposed system we applied ideas from the field of image processing and computer vision.

Wavelet coefficients of Cα-Cα distance matrix

Wavelet transform: Wavelet transform is an well known procedural approach in image processing and computer vision in order to image resizing and field compression. It was first proposed by Alfred Haar in 1909 [10]. Its based on small wavelets with limited duration. The wavelet transform is similar to the Fourier transform with a completely different merit function. The main difference is this: Fourier transform decomposes the signal into sine and cosine components, i.e. the functions are localized in Fourier space; in contrary the wavelet transform uses functions that are localized in both the real and Fourier space. Generally, the wavelet transform can be expressed by the following equation: $F(a,b) = \int_{-\infty}^{\infty} f(x)\psi^{*}_{(a,b)}(x)\,dx$

Wavelet coefficient as feature vector: We take wavelet transform approximate coefficients of normalized distance matrix. First we normalize the distance matrix to the nearest power of 2 with the help of bi-cubic interpolation.According to [5], nearest power of two for most of the structure is 128. For wavelet filter we used Daubechies-2 wavelet as [7] showed this filter outperforms other traditional wavelets for protein structure feature representation. With wavelet transform coefficients we re-sample all images to 128x128 dimension.

co-occurrence matrix of oriented gradient of wavelet coefficients: We compute co-occurrence matrix of oriented gradient and use that matrix as our novel feature. We found that L2 norm distance of this co-occurrence matrix of oriented gradient is much faster and more precise than alignment distance of wavelet coefficients of distance matrices.

Distance metric: In our research, for our proposed novel feature, we taken Euclidean distance or L2 norm of the feature as the distance measure of tertiary structures.

4.2 Motivation behind using co-occurrence matrix of oriented gradient

Upon Analyzing distance matrices and tertiary structure it is found that alpha helix and antiparallel beta sheets appear as dark lines parallel to diagonal and parallel beta sheets appear as dark lines normal to diagonal. Beta sheets of two strip appear as one dark line normal to diagonal, beta sheets of three strip appear as two dark line normal to diagonal and one dark line parallel to diagonal. In the same way, for a standard beta sheet of n strip ,the number of points of co-occurrence of parallel and antiparallel diagonal lines represent the number of strips in beta sheets. And number of only single parallel line represent number of standard alpha helix. In the similar way co-occurrence of gradient angles represent the secondary structure elements more precisely than just distance matrix image.This observation motivated us to take co-occurrence matrix of oriented gradient as representative feature. And after having our experiments, we found the results are amazing and awesome. And with this novel feature, now there is no need for align the distance matrix with computationally expensive algorithms, calculating Euclidean distance is enough now.

4.3 Our proposed algorithm

Step 1. Derive alpha carbon distance matrix from coordinate data generated by X-ray and nuclear magnetic crystallography. This distance matrices are symmetric matrix.
Step 2. Convert This matrices to 256 gray level gray scale image.
Step 3. Resize this images to the dimension nearest 2th power. Then We have images of dimensions 8x8, 16x16, 32x32, 64x64, 128x128, 256x256, 512x512, 1024x1024, 2048x2048 with bi-cubic interpolation.
Step 4. Take wavelet coefficient of the images. We use daubechies2 wavelet. Using Wavelet transform we re-sample all images to 128x128. Now we have all the images as 128x128. We needed to take all the images to same dimension. The reason for taking all the images to 128 is that most of the protein chains in the 152,487 protein chains are near the size of 128x128.
Step 5. Take gradient images. Calculate Gradient magnitude and angle images.
Step 6. This angle values are continuous and in within 0 to -360. Now create n bins of angles. We use 16 bins.
Step 7. Then Calculate Co-occurrence matrix of the bin matrix.
Step 8. Normalize the Co-occurrence matrix so that the sum is 1. This makes the matrix equivalent to probability distribution function. This matrix is our representative feature vector.
Step 9. We take Euclidean distance or L2 norm of the feature vector as distance metric.
Step 10. Sort the results according to ascending distance and get the closest array.

5 Experimental Results

We compared our results with scop(http://scop.berkeley.edu/) [8] domains classification which is well accepted as benchmark by many researchers as the classification was done by years of manual inspection. We have taken 152487 chains from scop domain. We have used ten randomly selected protein chains for matching tertiary structures from the total 152487 protein structures. We matched these 10 proteins with the total 152487 proteins that we have on our local database.

Our results in most cases are very well in accuracy and speed. Our method is more than 100 times faster than most of the automated retrieval methods. We First shown our results for varying the bin sizes for co-occurrence matrix of gradient orientation and compare those with MASASW. In figure 3 we have shown percentage of matches of Class for top 10,20,...,100 retrieval results for bin size 16,32 and 64 and for MASASW. In figure 4 we have shown percentage of matches of Fold for top 10,20,...,100 retrieval results for bin sizes 16,32 and 64. In the figure 5 we presented percentage of average matches for lower labels of scop classification in our top k retrieval results for k=10,20,30. This result shows better accuracy than [7] and [6].

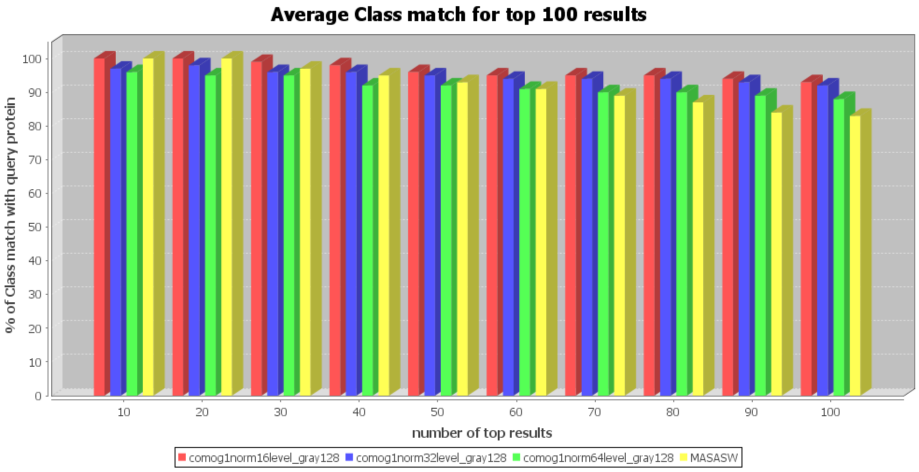


Fig. 3. percentage of matches of Class for top 10,20,...,100 retrieval results when image resized to 128x128 and bin size 16,32 and 64

From the figures, we see that the result of the three different bin size are almost the same. But bigger bin size the result is slightly off where smaller bin size has some better accuracy. In those charts in X axis we plot the number of top query results and Y axis denotes the accuracy percentage in class matching. From the experimental results we found bin size 16 is much better. It is clear from the figures that the results are much better than MASASW [7].

6 Performance Evaluation

The exact solution for the matching of tertiary structure with distance matrix is CSP and the time complexity is in factorial order of distance matrix dimension(N). The time complexity of MASASW is of

$$O(wWN^2)$$

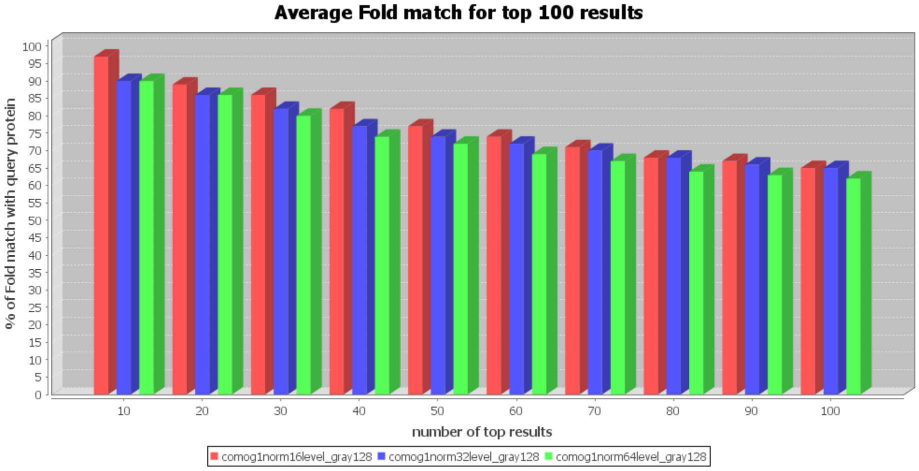
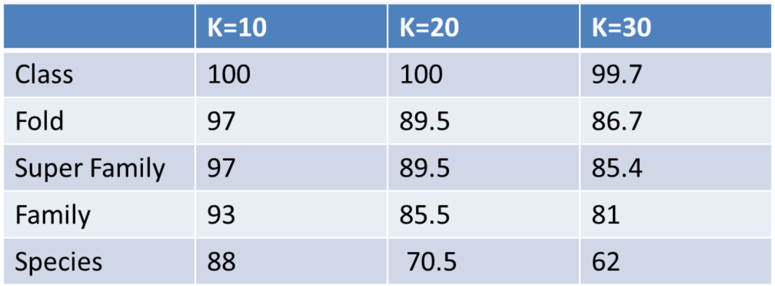


Fig. 4. percentage of matches of Fold for top 10,20,...,100 retrieval results when image resized to 128x128 and bin size 16,32 and 64



	K=10	K=20	K=30
Class	100	100	99.7
Fold	97	89.5	86.7
Super Family	97	89.5	85.4
Family	93	85.5	81
Species	88	70.5	62

Fig. 5. percentage of matches for class, fold, super family, family and species of scop classification in our top k retrieval results for k=10,20,30 to query protein.

and they taken N=32 as the size of distance matrix. W and w are the size of sliding window to align matrix and to align rows. They claimed that they empirically obtained most reasonable and suitable value for W=5 w=8.

In our method we have experimentally shown that taking co-occurrence matrix of oriented gradient with bin size 16 from distance matrix of size 128x128 is gives a feature matrix of 16x16 and using just L2 norm or Euclidean distance is much better in terms accuracy than aligning 32x32 distance matrices. So when this features are stored in database, run time is just

$$O(N^2)$$

and N=16. Therefore our proposed method is theoretically approximately (32x32x5x8)/(16x16)= 160 times faster than MASASW [7].

For accuracy, we experimentally showed that our results are more accurate than MASASW. The reason is that we designed a feature that is more representative to secondary structure elements in tertiary structure than previously used features. And to reduce loss of information we taken co-occurrence matrix of oriented gradient of 128x128 distance matrix where others used 32x32 distance matrix directly. We also experimentally showed that bin size 16 performs better than bin size 32 and 64. In the result set, we experimentally showed that, the accuracy of our approximate solution of this factorial order problem is much better than any state of the art distance matrix based approaches and the accuracy is acceptable as it conforms to scop classification label with good precision.

7 Conclusion

In this paper we presented a novel feature with computer vision oriented approach for faster retrieval of protein tertiary structure. We compared our results with labels of scop classification hierarchy. We showed average percentage of matching for class, fold, super family, family and species of our retrieval results with the query protein while most of the works showed similarity on class and fold and very few worked for automated similarity match for lower labels. Our method is surprisingly faster than any current methods available and its approximately 160 times more faster than the most recent and leading MASASW [7] algorithm. This creates the possibility of online web based service for protein tertiary structure retrieval while the present web services just gives email reply for the query results. We are working on automated clustering of protein tertiary structure currently and we plan to develop a online web service for protein tertiary structure retrieval. We expect our work will contribute to accelerate the speed of research of molecular biologists, biochemists, structural biologists and life science researchers for disease diagnosis, drug discovery and many other fields.

References

1. Aung, Z., Tan, K.L.: Matalign: precise protein structure comparison by matrix alignment. Journal of bioinformatics and computational biology 4(06), 1197–1216 (2006)
2. Brenner, S.E., Koehl, P., Levitt, M.: The astral compendium for protein structure and sequence analysis. Nucleic Acids Research 28(1), 254–256 (2000)
3. Gold, V., of Pure, I.U., Chemistry, A.: Basic terminology of stereochemistry (iupac recommendations) 68, 2220 (1996)
4. Holm, L., Sander, C.: Dali/fssp classification of three-dimensional protein folds. Nucleic acids research 25(1), 231–234 (1997)
5. Hwang, J.W., Lee, H.S.: Adaptive image interpolation based on local gradient features. Signal Processing Letters, IEEE 11(3), 359–362 (2004)
6. Marsolo, K., Parthasarathy, S., Ramamohanarao, K.: Structure-based querying of proteins using wavelets. In: Proceedings of the 15th ACM international conference on Information and knowledge management. pp. 24–33. ACM (2006)
7. Mirceva, G., Cingovska, I., Dimov, Z., Davcev, D.: Efficient approaches for retrieving protein tertiary structures. Computational Biology and Bioinformatics, IEEE/ACM Transactions on 9(4), 1166–1179 (July 2012)
8. Murzin, A.G., Brenner, S.E., Hubbard, T., Chothia, C.: Scop: A structural classification of proteins database for the investigation of sequences and structures. Journal of Molecular Biology 247(4), 536 – 540 (1995), http://www.sciencedirect.com/science/article/pii/S0022283605801342
9. Saier, M.: Computer-aided analyses of transport protein sequences: gleaning evidence concerning function, structure, biogenesis, and evolution. Microbiological reviews 58(1), 71–93 (1994)
10. Stanković, R.S., Falkowski, B.J.: The haar wavelet transform: its status and achievements. Computers & Electrical Engineering 29(1), 25–44 (2003)