

# Fractal and Mathematical Morphology in Intricate Comparison between Tertiary Protein Structures

*Ranjeet Kumar Rout<sup>1</sup>, Pabitra Pal Choudhury<sup>1</sup>, Sk. Sarif Hassan<sup>2</sup>, Saurabh Singh<sup>1</sup>*

<sup>1</sup>*Applied Statistics Unit, Indian Statistical Institute, Kolkata, India,*

<sup>2</sup>*Institute of Mathematics and Application, Andharua, Bhubaneswar-751003, India,*

*Email: [ranjeetkumarrou@gmail.com](mailto:ranjeetkumarrou@gmail.com), [pabitrupalchoudhury@gmail.com](mailto:pabitrupalchoudhury@gmail.com), [sarimif@gmail.com](mailto:sarimif@gmail.com)  
[saurabhsingh2904@gmail.com](mailto:saurabhsingh2904@gmail.com).*

## ***Abstract***

A three-dimensional (3D) structure of a protein is one of the most important attribute to extract vital information about the protein. It can be used to predict functions of a protein or to classify a protein depending on its similarity with the other protein structures. Thus, computation of similarities and dissimilarities between 3D protein structures is highly important. Though, several algorithms have been devised to compute the similarity between protein structures. But, most of them compare proteins by structural alignment of the protein backbones. In this paper we attempt to compute the similarities and dissimilarities among 3D protein structures using the fundamental mathematical morphology operation and fractal geometry. To implement the same we propose two methods one to determine the superficial structural or global similarity and the other to compute the internal or local similarity in atom level of the protein molecules. Analyzing and aggregating the results obtained in the two methods mentioned above we ascertain the overall similarity between proteins.

Keywords: 3D-Protein Structure, Similarities, Mathematical Morphology, Geodesic Dilation, Skeleton, Fractal Dimension

## **1. Introduction:**

Proteins are made of amino acids chain with its length ranging from 50 to more than 3000. A carbon atom (called  $C_\alpha$ ) is connected to a carboxyl (-COOH) group, an amine (-NH<sub>2</sub>) group, a hydrogen atom and a residue (which depends on the specific amino acid) to formulate a single amino acid. The amine group of an amino acid is covalently bonded by polypeptide bond with the carboxyl group of another amino acid to form a protein. The sequence of  $C_\alpha$  carbon atoms forms the backbone of the protein. Whenever the protein is left in its natural environment, it folds to a specific 3D structure. This is due to the forces between the amino acids such that the total free energy is minimized [6]. This renders a stable 3D protein structure. Thus, a protein can either be considered as polypeptides sequence of 20 amino acids occurring naturally

or as a 3D structure into which a particular protein folds [9]. A protein's functional properties mainly depend on its 3D structure. This is because a proteins with similar 3D structure will react similarly; thereby, depicting highly similar functional properties. As a result, knowledge of the 3D structure of a protein can yield vital information about the functional properties of the protein. Since a protein's amino acid sequence determines the 3D structure of a protein, which significantly influences the functionality of a protein. This may lead to a conclusion that sequence similarity is also a very good predictor of functional similarity, but this turns out to be less the case. As, similar sequences sometimes yield dissimilar structures. Thus, sequence similarity is not a reliable predictor of functional similarity [8] [9]. This establishes the fact that 3D structure comparison is the most reliable alternative to compute similarity between proteins. Though, comparison of the three-dimensional structures of protein molecules is a challenging problem. The search for an effective solution for this problem is justified because such tools can be of aid to scientists for prediction of the functions of a newly found protein, in development of procedures for drug design, in the identification of new types of protein architecture, in the organization of the known database of protein structures by classifying them according to their structures and can help to discover unexpected evolutionary and functional inter-relations between proteins [12] [13]. Several algorithms have been devised to compute the similarity between protein structures but they present a difficult computational problem. These problems have been resolved resorting to different methods with diversified approach. But, each of these methods has one or the other limitation associated with them [11]. As, in many cases there is not even a single superposition that reveals all regions of similarity between compared proteins (RMSD, DALI, ProSup) [1]. Also, there are many conceptual difficulties associated with various methods (RMSD, ad hoc scores based on local secondary structure, hydrogen bonding pattern, burial status, or interaction environment) which have not been resolved [2]. Classical criteria such as the Root Mean Square Deviation (RMSD) fail to identify similar shapes in a consistent way [3]. To add on various systems have been proposed for structural classification, such as Structural Classification of Proteins (SCOP), Class Architecture Topology Homology (CATH), Families of Structurally Similar Proteins (FSSP), and others. The similarity in their cases is computed using structural alignment algorithms such as DALI, CE, VAST, SSAP and others. Most of these methods are computationally intensive and time-consuming, especially when searching large databases due to intrinsic complexity of structural alignment [7]. Also, the prevailing practice in the protein crystallographic community for computing structural differences is highly inappropriate, in particular when medium- and low-resolution structures are involved [4]. Geometrical feature like Fractal dimension of  $C_\alpha$  of the backbone structure of one peptide chain proteins are considered in [17]. Obviously, a more objective method is highly desirable. In this paper we attempt to compute the similarities among 3D protein structures using mathematical morphology and fractal dimension of all the atoms of protein molecule as an innovative method which may yield desired output.

The arrangement of the paper is as follows: in section 2, the literature of mathematical morphology relevant to our work is reviewed; in section 3, the method of fractal dimension and geodesic dilation and the experimental result are discussed; section 4, is the conclusion.

## 2. Mathematical Morphology

Mathematical Morphology is a widely used paradigm in the field of image processing. Morphological tools are already very popular for image segmentation, image decomposition etc. Morphological operations like erosion, dilation, opening, closing are used for processing images frequently and produce results with high accuracy. The definitions of these basic morphological operators are as follows [10]:

$$\text{Erosion: } A \ominus S = \{a - s : a \in A, s \in S\} = \bigcap_{s \in S} M_s$$

$$\text{Dilation: } A \oplus S = \{a + s : a \in A, s \in S\} = \bigcup_{s \in S} M_s$$

$$\text{Opening : } A \circ S = (A \ominus S) \oplus S$$

$$\text{Closing : } A \bullet S = (A \oplus S) \ominus S$$

Where A denotes the shape that is to be transformed and S denotes the structuring element that is used for the transformation.

## 3. Methodology

The Protein Data Bank (PDB) is the largest and most commonly used repository for any kind of information regarding proteins. Information like 3D structure, family, function of every protein found till date is available in PDB. Mainly the X-Ray crystallography and Nuclear Magnetic Resonance is used for determining the 3D structure of the protein. The 3D structure is represented in (x, y, z) coordinates (with respect to an arbitrary origin) of the atoms presented in the protein. The '.pdb' files available in the PDB database contain all the structural information of a protein. Any molecule structure viewer like PyMol, Jmol is able to simulate the 3D protein structure available in the .pdb file. In this section two different methods are proposed to computing the similarity between 3D protein structures.

### 3.1. Skeleton and Fractal Dimension

Morphological skeleton of every geometrical structure is a subset of the original structure which has the same connectivity as the original structure from which inference can be drawn. From each point of the skeleton the distance to the boundary of the original set is the radius of a maximal circle (whose center is at a point of the skeleton) which touches the boundary at least two different points. The skeleton of an

object gives a clear idea about the shape of the object. For the shape A, and the structuring element S, the skeleton can be constructed through the operation [5] [14]:

$$Sk_n = (A \ominus nS) \setminus (A \ominus nS) \circ S \text{ for } n = 1, 2, \dots, N$$

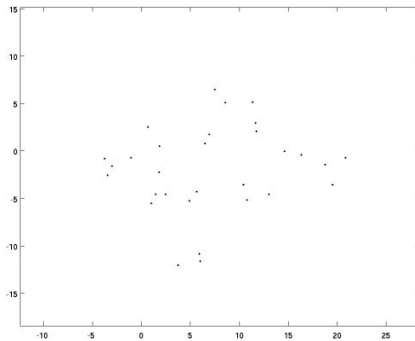
And the reverse process is as follows, Where, N is the number of performed iterations. Dilating the skeleton N times iteratively using the multi-scale structuring elements S a shape that is almost same to the original shape can be achieved.

$$A' = \bigcup_{n=0}^N Sk_n \oplus nS$$

Where,  $nS = S \oplus S \oplus \dots \oplus S$  (n times)

A fractal dimension is an index for characterizing fractal patterns or sets. The patterns illustrate self-similarity and the fractal dimension indicates the extent to which the fractal objects fills a particular Euclidean space in which it is embedded. These dimensions are usually non-integers. The proteins have an intrinsic self-similarity as they are hetero-polymers with a variable composition of twenty different amino acids. Thus, this protein backbone space curve consisting of  $C_\alpha$  atoms motivates us to compare 3D protein structure on the basis of their fractal features. [15]

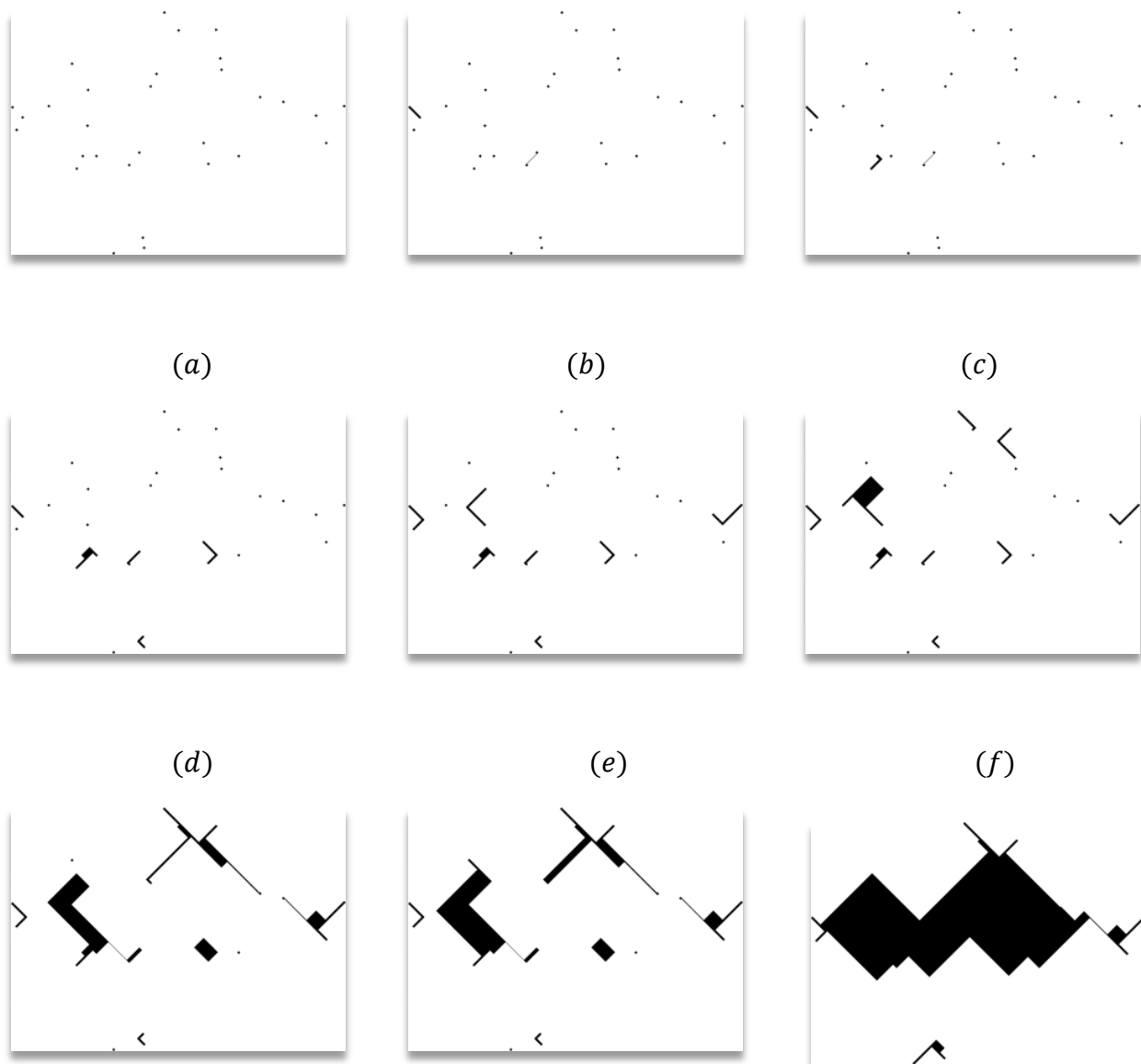
Now, for computing the local similarity the 3D structure is divided in some slices depending on the coordinates of the atoms. The atoms presented in the 3D structure are taken in terms of their z-coordinates. Any atom that has the same z-coordinate are in one slice irrespective of their x and y coordinates. A boundary value is introduced. Any atom that is located within the boundary value is taken under one slice along with other atoms. An example of such slice is shown in the following figure



*Figure 1 slice for protein 2LEP for  $z = -0.6$  to  $-0.699$*

In *Figure 1* a slice is shown for the protein 2LEP. Each “.” represents an atom. The slice contains all the atoms whose z coordinate is within  $-0.600$  to  $-0.699$ . Here the boundary value is taken as  $-0.1$  .

Depending on the slice interval the number of slice increases or decreases. And the number of atoms in a slice is inversely proportional to the number of slices. After acquiring the slices, the goal is to form a plane that contains all the atoms. A plane is needed because of the overall shape of the particular plane is important to us. Because if all the planes are stacked upon each other depending on the  $z$ -coordinates, then we get almost similar atom distribution as the original 3D protein structure. So each slice is important for describing the protein structure. To form the plane from a given slice, iterative multi-scale opening is used. For, each iteration the structuring element with which the opening is performed is increased by one. And for the opening, a primitive structuring element of size  $n \times n, n = 1, 2, \dots, N$  is used. The iterations for the slice shown in *Figure 1* are shown below.



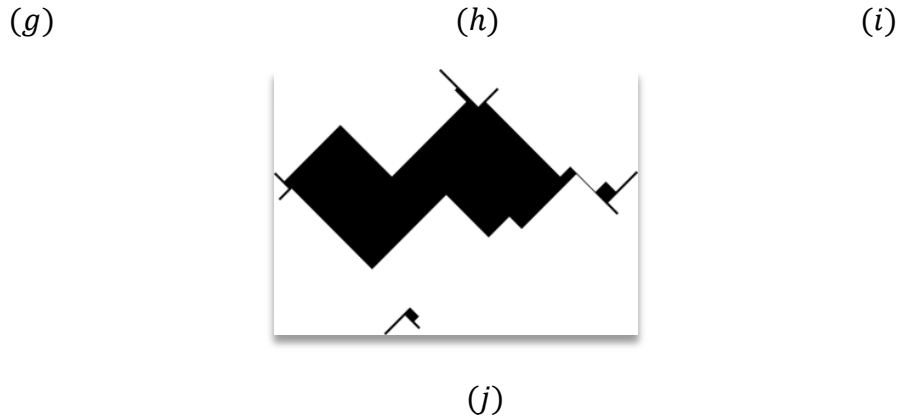


Figure 2. (a)- (j) shows the multi-scale opening using primitive structuring elements. Starting from size one, each figure shows the iteration with structuring element larger by ten units from the previous one.

The iterative opening may take a large number of iterations to contain all the atoms in a particular slice. And there may be more than one plane for a slice. So we dilate the plane with a primitive structuring element. This reduces the number of planes for each slice. The example of the plane after dilating with a disk shape structuring element of size 25, the resulting plane becomes as shown in Figure 3.

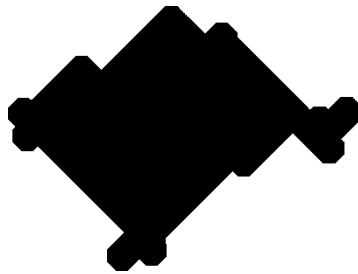


Figure 3. The plane shape after dilating the closed shape shown in Figure 2.(j)

After acquiring all the planes for a particular protein structure, our next aim is to find the skeletons for each of the plane shapes. Example of the skeleton for the plane shown in Figure 3 is given below.

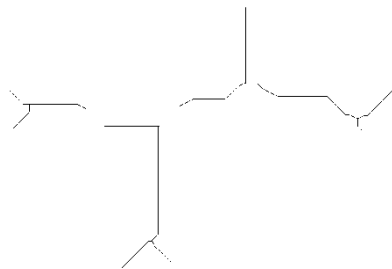


Figure 4. Skeleton for figure shown in Figure 3

If we stack the skeletons for all the planes over each other, then the resulting image gives us an idea of how the atoms form the overall protein structure in terms of the planes that are formed by the coordinates of the atoms. For the protein 2LEP the skeleton structure is like as follows,



Figure 5. Skeleton structure for protein 2LEP

From the skeleton we have an idea of fractal-like distribution of protein atoms of the 3D protein structure. The fractal dimensions for the skeleton can be computed through *Box Counting Method* which is briefly stated as follows.

**Box-Counting Method:** This method computes the number of cells required to entirely cover an object, with grids of cells of varying size. Practically, this is performed by superimposing regular grids over an object and by counting the number of occupied cells. The logarithm of  $N(r)$ , the number of occupied cells, versus the logarithm of  $1/r$ , where  $r$  is the size of one cell, gives a line whose gradient corresponds to the box dimension [15, 16]. To calculate the dimension for a fractal  $S$ , the Box-Counting dimension is defined as,

$$\text{Dim}_{\text{box}}(S) = \lim_{n \rightarrow 0} \frac{\log N(r)}{\log \frac{1}{r}}$$

Now we compute the fractal dimension for the skeleton obtained for a particular protein. In similar manner we compute the fractal dimension of all the proteins molecules. The similarity between two protein structures  $i$  and  $j$  can be computed by using the following equation:

$$\rho = |d_i - d_j|$$

Where  $\rho$  is the difference between the fractal dimensions of any two protein molecules and some experimental result shown in **Table 2**. The experimental result shows that if  $\rho \leq 0.008$ , two proteins are similar in structures and functions. Thus, lower difference between fractal dimensions will ensure high similarity between the proteins which are being compared. Thus, finally we compute the overall similarity taking into account the results obtained by both the methodologies.

### 3.2. Geodesic Dilation and Its Quantification:

Unlike the existing algorithms, this work does not work on primary or secondary structure of the proteins. As mathematical morphology considers size and shape of 2D objects, our first goal is to convert the 3D protein structure in terms of a collection of 2D objects. From the PDB database the protein structures are viewed by using Jmol and the protein structures are rotated depending on the 3-axis, from which we have collected the 6- faces or views (front, left, right, top, bottom, and back) of each 3D protein structure

respectively. To find out self similarity between two 2D images we use geodesic dilation which is a morphological transformation to operate only some part of the image (as marker) to grow until the boundary of the image and the advantages of this transformation is that the structuring element can vary at each pixel, according to the image.

The **Geodesic Dilation**  $\delta_X$  of an image Y inside X is defined as the intersection of the dilation of Y (with respect to a structuring element B) with the image X

$$\delta_X^n = (Y \oplus nB) \cap X \text{ where } n = 1, 2, \dots, N$$

So Geodesic dilation terminates when all the connected components of X are constructed i.e. idempotency is reached  $\forall n > n_0, \delta_X^{(n)}(Y) = \delta_X^{(n_0)}(Y)$ .

Let  $f$  and  $g$  are the front faces of two protein structure  $i$  and  $j$ , where  $f \cap g$  is the common connected components which is acting as the marker. Now we calculate the number of dilation  $\partial_1 = ((f \cap g) \oplus nB)$  until the entire connected component of  $f$  is constructed and also  $\partial_2 = ((f \cap g) \oplus nB)$  for  $g$  until the entire connected components are constructed. Similarly for all the faces (front, left, right, top, bottom, and back) of both the protein structures are computed. The similarity between two protein structures  $i$  and  $j$  can be computed by using the following equation:  $D = \sum_{i=1}^6 |\partial_i - \partial_j|$  where  $D$  is the difference between the numbers of dilation between two protein molecules. The experimental result shows that if  $D \leq 12$ , two proteins are similar in structures and functions. Thus, lower difference between geodesic dilation will ensure high similarity between the proteins which are being compared.

**Result and Discussion:** Now we take the front view of two different proteins to compute the similarity between them. For, this purpose we consider the front view of 3V2J and 3V2M. The images of front view of both the proteins are given below in figure 2.

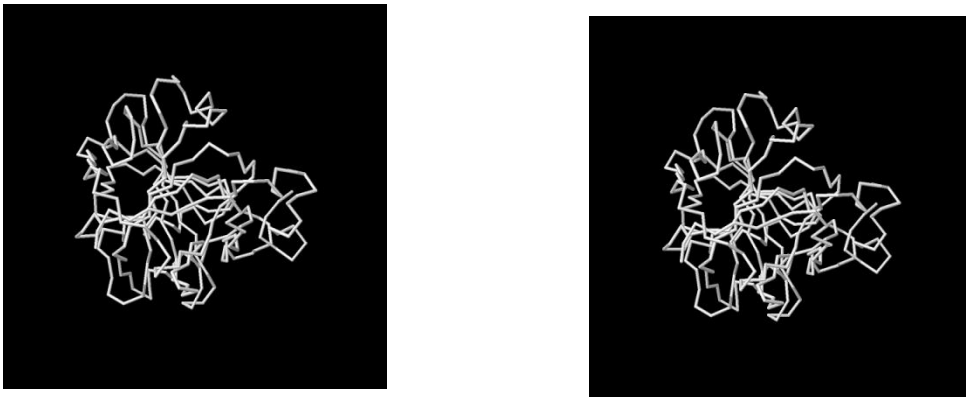


Figure 2: Front view of protein 3V2J and 3V2M

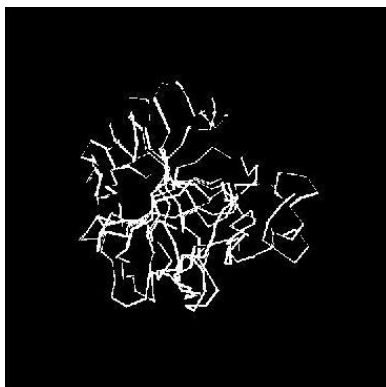


Figure 3: Intersection Front view of protein 3V2J and 3V2M

As discussed earlier in the methodology higher the difference between the dilations of two proteins for particular view the less becomes the similarity and vice-versa. This would become clearer with few examples for the same. The **Table 1** given below shows the different geodesic dilation of different protein molecule. From the **Table 1** we conclude that the proteins 3V2J and 3V2M are similar in structure and function as  $D=2$ ,

**Table 1: Geodesic Dilation  $\delta$  of different faces**

Protein ID	Geodesic Dilation $\delta$ of different faces						$\sum_{i=1}^6 \delta_i$	D
	Front	Left	Right	Top	Bottom	Back		
<b>3V2J</b>	4	4	5	6	5	3	27	<b>2</b>
<b>3V2M</b>	4	4	4	6	4	3	25	
2LE8	10	6	6	7	8	10	47	64
2LLS	19	19	21	18	15	19	111	

**Table 2: Difference between Fractal Dimensions of compared proteins pairs**

Protein ID 1 (pdb)	Fractal Dimension (FD) $d_i$	Protein-ID 2	Fractal Dimension (FD) $d_j$	Difference between FDs $\rho$	Geodesic Dilation  D	PDB Result
<b>3v2j</b>	1.661190e+000	3smk	1.620140e+000	0.04105	24	12%
		3t0o	1.646469e+000	0.014721	35	18%
		4ecs	1.649489e+000	0.011701	31	2%
		<b>3v2m</b>	1.656160e+000	<b>0.00503</b>	<b>2</b>	<b>100%</b>
		3sv1	1.605381e+000	0.055809	28	28%
		4ag2	1.695859e+000	0.034669	39	31%
<b>1cah</b>	<b>1.661085e+000</b>	<b>1cai</b>	1.660481e+000	<b>0.000604</b>	<b>5</b>	<b>100%</b>
		4bij	1.680213e+000	0.019128	21	41%
		2lep	1.549605e+000	0.11148	43	57%
		1cgi	1.635992e+000	0.025093	35	50%
		4eym	1.649456e+000	0.011629	23	39%
		<b>2cbc</b>	1.661399e+000	<b>0.000314</b>	<b>1</b>	<b>100%</b>

#### 4. Conclusion:

In this work, we presented a novel technique to compute the structural and shape similarity of 3D protein structure using fractal dimension and geodesic dilation in atom levels and proteins backbone structure level respectively. Compared with the existing methods, fractal dimension and geodesic dilation is easy to compute and efficient enough to eliminate the limitations encountered in the existing algorithms. In our experiments, atoms of all the protein structures are divided into slices by fixing the z co-ordinate value. So only the analysis of the x-y planes is done. This work can be further extended by fixing the x and y co-ordinate values, i.e. analysis of the x-z and y-z planes of the protein structure.

#### References:

- [1] A. Zemla, "LGA: a method for finding 3D similarities in protein structures"; Nucleic Acids Research, Vol.-31, Issue-13, pp. 3370-3374 (2003)
- [2] D. Goldman, C. H. Papadimitriou, S. Istrail; "Algorithmic Aspects of Protein Structure Similarity". *FOCS' 99 Proceedings of the 40<sup>th</sup> Annual Symposium on Foundations of Computer Science*, pp. 512.(1999)
- [3] F Guyon, P. Tufféry, "Assessing 3D scores for protein structure fragment mining ", Open Access Bioinformatics, Vol.-2, pp-67-77(2010).
- [4] G. J. Kleywegt, "Experimental assessment of differences between related protein crystal structures"; Acta Crystallogr. D Biol. Crystallogr, Vol.-55, pp. 1878-1884(1999).
- [5] G. Klette, "Skeletons in digital image processing", July, (2002).
- [6] G. Lancia, S. Istrail, "Protein Structure Comparison: Algorithms and Applications". In: Guerra, C., Istrail, S. (eds.) *Mathematical Methods for Protein Structure Analysis and Design*. LNCS (LNBI), Vol. 2666, pp. 1-33. Springer, Heidelberg (2003).
- [7] I.G. Choi, J. Kwon, S. H. Kim; *Local feature frequency profile: A method to measure structural similarity in proteins*; PNAS, Vol. 101; pp.-11(2004).
- [8] J. Galgonek, D. Hoksza, T. Skopal; "SProt: sphere-based protein structure similarity algorithm". Galgonek et al. *Proteome Science*, Vol.-9(Suppl 1):S20 (2011).
- [9] L. P. Chew; "Exact Computation of Protein Structure Similarity", Proceeding SCG '06 Proceedings of the twenty-second annual symposium on Computational geometry pp.468-474(2006).
- [10] L. L. Teo, B. S. Daya Sagar; "Modeling, description, and characterization of fractal pore via mathematical morphology"; *Discrete Dynamics in Nature and Society*, Vol. 2006, Article ID 89280; Pages 1-24 (2006).
- [11] L. Holm, C. Ouzounis, C. Sander, G. Tuparev, G. Vriend; "A database of protein structure families with common folding motifs"; Vol.-1, Issue-12, pp.1691-1698(1992).

- [12] N. Krasnogor, D. A. Pelta; “*Measuring the similarity of protein structures by means of the universal similarity metric*”; Bioinformatics, Vol.- 20,pp.1015-1021(2003).
- [13] P. Koehl, “*Protein structure similarities*”; Curr Opin Struct Biol, Vol.-11, pp.348-353(2001).
- [14] Petros A. Maragos, Ronald W. Schafer; “*Morphological skeleton representation and coding of binary images*”; IEEE Transactions on Acoustics, Speech and Signal processing, Vol. assp-34; pp. 5(1986).
- [15] D. Avnir, O. Biham, D. Lidar, O. Malcai, “*Is the geometry of Nature fractal*”, Science Vol.-**279**, pp. 39-40(1998).
- [16] K Develi, T Babadagli, “*Quantification of natural fracture surfaces using fractal geometry*”, Math. Geology, Vol.- **30** , pp. 971-998(1998).
- [17] C Cui, D. Wang, X. Yuan, “*3D protein structures similarity matching based on fractal features*”, SPIE Vol. 5637, pp-567-572(2005).