A New Approach for Examining the Similarity of Protein 3D Shape

Min Hu, Wei Chen, Tao Zhang, Qunsheng Peng and Liguang Xie

State Key Lab of CAD & CG, Zhejiang University
Hangzhou, 310058, China

{humin,chenwei,zhangtao,peng,xieliguang}@cad.zju.edu.cn

Abstract

A novel technique for estimating the potential similarity of protein 3D shape is proposed. It is found that the structures of proteins are of self-similarity, such a phenomenon can be well described by their fractal dimensions. We then adopt the size and the volume fractal dimensionality of proteins to characterize the spatial complexity of the protein shape. Such quantitative measurement is helpful to reduce the search space when we retrieve protein structures from large data sets. Experimental results demonstrate the potential of our proposed approach.

Keyword: Volume Fractal dimension, Protein structure, similarity.

I. Introduction

Since the first three dimensional structure of a protein molecule in atomic detail was explored in 1960[10,13], the number of explored protein structures has dramatically increased to more than two thousand by Dec. 2004 in the PDB[3] owing to the recent advances in protein engineering, crystallography, and spectroscopy. It is still accumulating at a fast rate, demanding effective approaches for processing the tremendous amounts of protein structure information. One important research issue is how to estimate the 3D shape similarity between different protein molecules and to develop rapid retrieval mechanism for similarity queries.

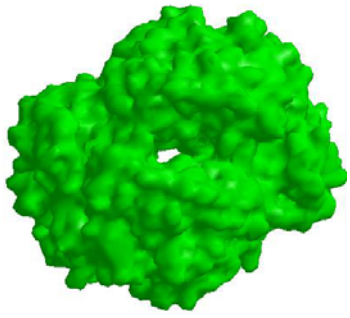


Fig1 Surface image of PDB1a00 (drawn by Chem3D)

3D shape similarity has been studied extensively in the field of computer vision and object recognition. Nevertheless, most of these efforts focus on 3D solid models that are represented by polygons and surfaces. They can hardly be employed to analyze complex protein structures. Fig.1 shows an example of accessible surface image of the hemoglobin protein(PDB 1a00). It is found that there are a lot of surface corrugation and roughness in the protein, even with tiny holes through the body. Although traditional 3D shape parameters can hardly characterize the spatial features of proteins directly, there exist other approaches which try to keep track of the shape similarity of known protein structures based on their spatial atomic position. These approaches usually examine the distances between their inner elements, such as C-alpha atoms, residues or secondary structure elements. Their measure of similarity is usually determined by the following methods: 1) RMSD [1], which adopts means of rigid body motions and least square fitting to find the minimum value of the root-mean-square deviation between corresponding atoms; 2) Distance Matrices [9] , also called distance plots or distance maps, in which, all pair-wise distances

between residue centers are represented, similar 3D structures have similar inter-residue distance; 3) Contact Map [11,15], an N×N matrix S for a protein of N residues whose element $S_{i,j} = 1$ if residues i and j are in contact and $S_{i,j} = 0$ otherwise. Similar sub-structures can be detected by patterns hidden in the contact maps; 4) The graph based method [7,19], in which, a protein molecule is represented as a graph, while the graph isomorphism is examined to investigate the similarity of proteins. There are also some other geometrical, topological and statistical methods [2,4,5,17,18]. Although they provide quite good results in some aspects, most of the exist shape comparison methods face great challenges in terms of the tremendous computational cost since the molecular data set expands dramatically. It is therefore necessary to develop new approaches that can extract the features of the protein shape and solve for the similarity retrieval between different proteins in a more effective way.

In this paper, a set of global quantitative parameters, the radius and volume fractal dimensionality (R, VFD) of a protein, are introduced as features to help screen out a small set of candidates from the large source data set. Further identification needs only to be performed within the small candidate data set. We also describe a framework with feedback pathway to conduct rough to fine similarity comparisons. Such mechanism greatly speeds up the retrieval process.

II. Method

Although all proteins are constructed from a common set of just 20 kinds of amino acids, the configuration of these amino acids for arrangement of a particular protein is many. In general, proteins in nature are composed of tens to thousands of amino acids. Even for a small protein with 100 amino acids in length, there will be 20^{100} combinatorial cases. On the other hand, the biological function of a protein depends completely on its native conformation, which corresponds to a unique three dimensional structure. Proteins with similar functions have similar spatial distribution of the proteins' components. Further more, it is found that the protein molecules possess the symmetry of self-similarity [6,8,12,16,20,22]. The symmetry of 3D shape ensures its invariance to rotation, reflection, inversion and translation. Based on the above observation, we adopt the parameter of volume fractal dimensionality to characterize the geometrical complexity of spatial distribution of the protein components and put forward a framework for fast similarity retrieval.

A. Volume Fractal Dimensionality of proteins

A fractal is a shape composed of parts similar to the whole at any scale. We assume that the volume of a protein can be represented as spatial occupation of its atoms. The volume fractal dimension of a protein can be determined by measuring its volume V with a ruler of fixed length l. If the following form is satisfied:

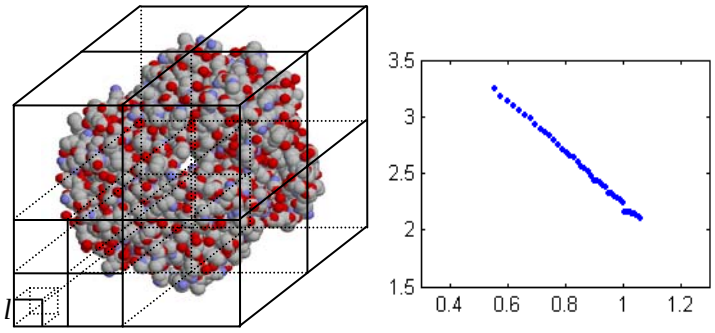
$$\mathrm{V}(l) \propto l^{D} \qquad (1)$$

then the fractal dimensionality of the volume is D.

We employ the simple "Box Counting" method to analyze and obtain the VFD [14]. In this approach, the fractal shape lying in a 3-dimensional space is covered by 3-dimensional grids with elements of size l (length scale), as depicted in Fig.2(a). The result of box counting or capacity dimensionality is given by

$$D_B = - \lim_{l \to 0} (\log(N(l)) / \log(l)) \qquad (2)$$

where $N(l)$ is the number of non-empty grid elements. The volume fractal dimensionality D_B can be easily obtained by the bivariate linear regression model $y = a + bx$. Taking into account the n samples of the form $\{(x_i, y_i), i ? [0, n \quad 1]\}$, in which, x_i stands for $\log(l_i)$ and y_i stands for $\log(N(l_i))$, D_B corresponds to the slope of the line. Fig.2(b) illustrates the fractal diagram of PDB1a00.



(a) The box counting method (b) The plots of $\log(N(l_i))$ - $\log(l_i)$

Fig.2 The shape of a hemoglobin and its calculation of VFD

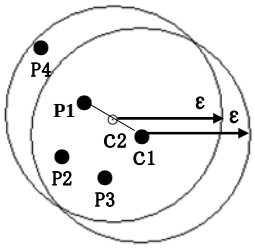


Fig. 3. Reachable_objects

B. *Similarity evaluation of 3D protein shape*

Note that several factors may affect the result of our similarity evaluation. Firstly, it is found that some local mutations do not affect the protein' function significantly. Secondly, the components of similar proteins, even proteins with same structure may have different spatial positions [21]. Thirdly, as the protein structures data preserved at PDB are obtained in different experimental environments, the 3D data may vary in accuracy. The former two factors tell us that similar proteins may have some difference in size and spatial distribution of atoms. The third one indicates the test data we use to evaluate the similarity may include experimental error.

Due to the complicated function-structure mapping mechanism of proteins as well as the tolerance of the test data, the volume fractal dimensionality can be significant only in an average or statistical sense. Although it is impossible to exclusively identify the unique protein structure pattern by VFD, we can rapidly pick up a small set of candidates from a large database.

To perform shape similarity comparison for proteins, we propose a rough to fine procedure with a feedback pathway. First, we adopt two feature parameters to characterize the protein spatial structure, i.e., the maximum radius of the protein and the corresponding VFD, notated by (R, VFD). This is because the VFD strongly depends on the size of the cube which surrounds the protein. Proteins are roughly considered to have similar shape if they have close R and VFD under a threshold. The above process effectively screens out a similar candidate proteins set. Further comparison method can be used to get finer result.

In order to quickly identify proteins falling into a specified range of (R, VFD) , an ordered table is built up, with each entry composed of the radius of the protein, the VFD of the protein and the PDB id, notated as (R, VFD, ID). Every protein has a unique entry in the table. We sort all the entries recursively first by R then by VFD to set up an ordered table. Then we cluster the adjacent entries in the (R, VFD) space.

Before we present a complete similarity retrieval algorithm, some of the notations used in the algorithm are explained below:

Seed_object: Let P be a protein entry, let ε be a threshold value, we select a small set of candidates for further similarity comparison from the ε-neighborhood of P and define object P as the Seed_object.

Reachable_object: All entries within the ε-neighborhood of entry C are considered as the Reachable_objects of C. As depicted in Fig.3, points P1, P2, P3 are Reachable_objects of the Seed_object C1 and P4 is not, but P4 is a Reachable_object of the object C2. Here, C2 is a virtual point, which is the centroid of P1 and C1.

A framework based on an effective searching strategy with a feedback pathway is shown in Fig.4. When a query shape is given, we firstly calculate the radius and VFD of the protein. Then the query object is assigned as the Seed_object so as to find the neighbors in the (R,VFD) space by looking up the ordered table. The candidates of potential similarity are those whose size and VFD are close to the Seed_object within the range of a threshold.

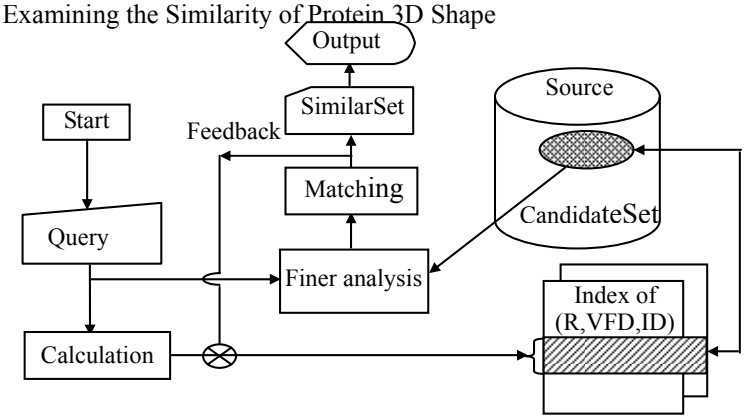


Fig. 4 Overview of the rapid matching process

To choose an appropriate value of the threshold, a feedback pathway is set up to provide opportunities for further adjustment. Such a framework can be implemented with the algorithm outlined below:

```
Procedure FR(QueryObject,ε)
{
    CandidateSet = Null;
        // Store all candidates of similar objects
    SimilarSet = Null ;
                // Store similar Objects
    Let Seed_object := QueryObject;
    LOOP:
          Search ordered index of Pi(R, VFD, ID)  DO
          {
                If  (||Pi-Seed_object|| <ε ) THEN
                      Put Pi into the CandidateSet;
                      // Put all the Reachable_objects of the
                      // current Seed_object into CandidateSet;
                      Pi.Processed := False;
                      // Marking for later processing;
          }
     For each Qi in the candidateSet DO
     {
                if (RMSD(QueryObject, Qi) < GivenThreshold )
                       THEN
                Put Qi into the SimilarSet;
     }
     If （there are still any objects NOT Qi.Processed in the SimilarSet) THEN
     {
                Produce a new Seed_object;
                     with its index location at the Centroid of the current Seed_object and Qi;
                Qi.Processed := True;
                Feedback to LOOP;
      }
    Else {
                 Output the SimilarSet;
                 EXIT(1); // End the retrieving procedure
          }
}
```

Our similarity retrieval strategy greatly reduces the search space. The feedback mechanism makes it possible to find all the proteins with similar shape within a conservative threshold.

III. Experiments and Discussion

To illustrate the above algorithm clearly, we selected a small data source which contains all the samples with the first digit of PDB ID is 7 from the PDB(Jan.2004, Release #107). There are a total of 74 model-ID individuals, representing 74 source objects. Table 1 lists all these PDB IDs and their radius and VFD values.

Table 1. 74 PDB IDs and their volume radius and VFDs

No.	PDBID	RADIUS	VFD	No.	PDBID	RADIUS	VFD
1.	7kmeI	11.89	0.72	38.	7gsp	30.86	2.21
2.	7r1rDEFP	86.33	1.25	39.	7atj	32.62	2.21
3.	7kmeL	16.3	1.72	40.	7tln	35.49	2.21
4.	7pti	19.24	1.96	41.	7taa	43.46	2.21
5.	7wga	40.4	1.99	42.	7prc	75.88	2.21
6.	7rxn	15.37	2.02	43.	7pcy	20.68	2.22
7.	7ame	20.36	2.02	44.	7tli	35.37	2.22
8.	7rnt	21.23	2.03	45.	7lyz	23.33	2.23
9.	7ceiB	23.41	2.05	46.	7hbi	30.43	2.24
10.	7hsc	39.93	2.05	47.	7enl	36.44	2.24
11.	7ceiA	22.82	2.06	48.	7ca2	29.54	2.25
12.	7znf	23.67	2.06	49.	7r1rABC	90.29	2.25
13.	7msi	20.31	2.07	50.	7paz	20.81	2.26
14.	7rat	26.61	2.08	51.	7ccp	30.56	2.26
15.	7cat	56.91	2.09	52.	7ptd	32.18	2.26
16.	7reqBD	92.88	2.09	53.	7cpp	36.76	2.26
17.	7dfr	26.02	2.1	54.	7cpa	31.47	2.27
18.	7prn	37.17	2.1	55.	7tim	40.13	2.27
19.	7pck	88.73	2.1	56.	7nse	50.21	2.27
20.	7ins	27.89	2.11	57.	7mdh	65.92	2.27
21.	7upj	34.07	2.11	58.	7a3h	28.41	2.28
22.	7rsa	27.86	2.12	59.	7cel	35.75	2.28
23.	7icd	42.79	2.13	60.	7cgt	43.56	2.29
24.	7kmeH	31.12	2.14	61.	7i1b	30.27	2.3
25.	7fab	43.52	2.14	62.	7aat	50.74	2.3
26.	7reqAC	93.22	2.15	63.	7nn9	34.86	2.31
27.	7hvp	33.12	2.17	64.	7acn	48.35	2.31
28.	7lzm	29.22	2.18	65.	7fd1	20.55	2.32
29.	721p	23.99	2.19	66.	7yas	27.2	2.33
30.	7lpr	25.74	2.19	67.	7ahl	63.77	2.33
31.	7adh	42.27	2.19	68.	7jdw	30.51	2.34
32.	7odc	43.89	2.19	69.	7gss	32.08	2.36
33.	7at1	60.4	2.19	70.	7gep	34.33	2.37
34.	7gch	26.72	2.2	71.	7fdr	20.38	2.38
35.	7est	28.52	2.2	72.	7std	34.18	2.4
36.	7abp	36.3	2.2	73.	7gpb	76.63	2.42
37.	7api	40.26	2.2	74.	7xim	53.01	2.44

An ordered table of 74 entries(R, VFD, PDBID) is set up. The 2D map of (R, VFD) for the test data set is depicted in Fig.5, where the abscissa represents the radius R and the ordinate represents the volume fractal dimensionality. When we give a query molecule structure PDB1b7i, for example, and search for the similar protein molecules, we first calculate the radius and VFD of this protein, which are (20.82, 2.01). We then select a conservative threshold, equal to 30% fluctuated scale of the query protein's radius and allow a VFD deviation of ±0.10, the searching space falls into the range of {(r, v); r∈ (17.70,23.94), v∈ (1.91, 2.11) }. Fig.6 depicts a zoomed plot of such a region, marked with ABCD. The CandidateSet is then obtained with our similarity retrieval framework. It can be seen there are only seven entries in the rectangle ABCD, which is less than 10% of the initial data set. This dramatically reduces the searching space. Further matching may be carried out with traditional method. We adopt the RMSD method to perform finer comparison. There are two proteins with similar structure, PDB 7ame and PDB7msi. The two similar proteins are then used as new Seed_objects to launch a new loop. In our experiments, all the proteins of similar shape are selected in the first loop correctly.

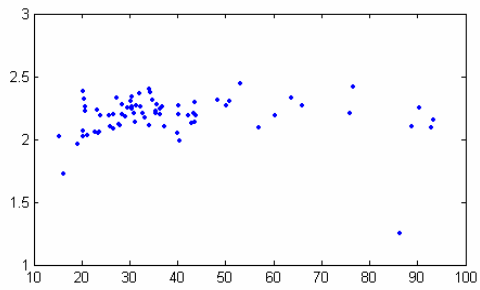


Fig5. the map of (R, VFD)

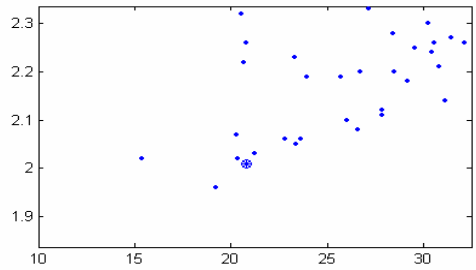


Fig.6 CandidateSet, zoomed in on the area including the query object from Fig.5

It can be seen from the above test that a much smaller data set is obtained from the source database with a Seed_object. This is true even in a large source database. For another example, we used the same query protein PDB1b7i and the same threshold as above, but we selected another data source which contains all the samples with the first digit of PDB ID is 4 from the PDB, total 231 model-ID individuals. Seventeen candidates are screened out after the first loop in the procedure FR. So we only need to find similar structures from the reduced searching space, which is less than 10% size of the search space. By further comparison, we found the actual similar proteins, PDB 4msi and PDB 4ame.

Ⅳ. Conclusion

We have integrated the natural volume fractal feature with conventional comparison approach to finding globally similar 3D shape proteins in large data sets. The main advantage of VFD is its simplicity of calculation and its invariance to the transformation of the scaling, translation and rotation of proteins. With VFD and the ordered indices table, we have provided an effective means to find all the potential candidates of similar shape by reducing the searching space. Fine similarity result can be obtained by further employing the traditional method.

Biological molecules search is an emerging application of content-based retrieval. The explosive increasing amount of the 3D structure database of bio-molecules makes it imperative to develop faster retrieval techniques. The approach presented in this paper is a significant exploration.

Acknowledgments

This research is supported by the NSFC Funds for the key project No. 60533050.

References

[1] Tatsuya AKUTSU, Protein Structure Alignment Using Dynamic Programming and Iterative Improvement, IEICE TRANS. INF. & SYST., VOL, E78-D, 1996

[2] Birgit Albrecht, Guy H.Grant and W.Graham Richards, Evaluation of structural similarity based on reduced dimensionality representations of protein structure, Protein Engineering, Design and Selection, 2004, 17(5), 425-432

[3] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne: The Protein Data Bank. Nucleic Acids Research, (2000)28,235-242

[4] David Bosticka and Iosif I. Vaismanb, A new topological method to measure protein structure similarity, Biochemical and Biophysical Research Communications, (2003) 304,320–325

[5] Oliviero Carugo, Sandor Pongor, Protein Fold Similarity Estimated by a Probabilistic Approach Based on Cα- Cα Distance Comparison, J. Mol. Biol(2002)315, 887-898

[6] L. Elber , Fractal analysis of proteins, Fractal approach to heterogeneous chemistry, 1989, 345-361, John Wiley and Sons LTD

[7] Helen M. Grindley, Peter J. Artymiuk, David W. Rice and Peter Willwtt, Identification of Tertiary Structure Resemblance in Proteins Using a Maximal a Common Subgraph Isomorphism Algorithm, J. Mol. Biol. 1993, 229, 707-721

[8] Thomas Goetze and Jürgen Brickmann, Self similarity of protein surfaces, Biophysical Journal, 1992, 61, 109-118

[9] Liisa Holm, Chris Sander, Protein Structure Comparison by Alignment of Distance Matrices, J. Mol. Biol, ,1993, 233, 123-138

[10] Liisa Holm, Chris Sander, etc., SEARCHING PROTEIN STRUCTURE DATABASES HAS COME OF AGE, Proteins ,1994, 19:165-173

[11] Jingjing Hu, Xiaolan Shen, Yu Shao, Chris Bystroff , MoHammed J. Zaki, Mining Protein Contact Maps, Workshop on Data Mining in Bioinformatics (SIGKDD02) , 2002, 3-10

[12] Min HU and Qunsheng PENG, Volume Fractal Dimensionality: A Useful Parameter for Measuring the Complexity of 3D Protein Spatial Structures, 20th ACM Symposium on Applied Computing,Bioinformatics Track， New Mexico. USA，March, 2005, 172-176

[13] Kendrew, J. C., Strandberg, B. E., Hart, R. G., Davies, D. R., Phillips, D. C. , Shore, V. C. Structure of myoglobin. A three-dimensional Fourier synthesis at 2 Å resolution. Nature , 1960, 185: 422-427

[14] Paul Meakin, Fractals, scaling and growth far from equilibrium, CAMBRIDGE UNIVERSITY PRESS, 1998

[15] Mohammed J. Zaki, Shan Jin, Chris Bystroff, Mining residue contacts in proteins using local structure predictions, IEEE Inte. Symp. on Bioinformatical Enginering, Nov. 2000,168-175

[16] Mitchell Lewis, D.C.Rees, Fractal surfaces of Proteins, SCIENCE, 1985, 230, 1163-1165

[17] Patra SM, Vishveshwara S, Backbone cluster identification in proteins by a graph theoretical method, BIOPHYSICAL CHEMISTRY, , 2000, 84(1), pp 13-25

[18] Ilya N. Shindyalov and Philip E.Bourne, Protein structure alignment by incremental combinatorial extension(CE) of the optimal path, Protein Engineering , 1998,11(9), 739-747

[19] SARASWATHI VISHVESHWARA, K.V. BRINDA and N.KANNAN, PROTEIN STRUCTURE: INSIGHTS FROM GRAPH THEORY, J. of Theoretical and Computational Chemisty, 2002,1(1), 187-211

[20] Cun Xin Wang, Yun Yu Shi, Fu Hua Huang, Fractal study of tertiary structure of proteins, PHYSICAL REVIEW A, 1990, 41(12)

[21] An-Suei Yang and Barry Honig, An Integrated Approach to the Analysis and Modeling of Protein Sequences and Structures. Ⅱ. On the Relationship between sequence and structural Similarity for proteins that are Not Obviously Related in Sequence, J. Mol. Biol. (2000) 301, 679-689

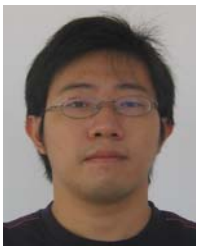
[22] CARL-DIETER ZACHMANN and JÜRGEN BRICKMANN, Hausdorff Dimension as a Quantification of Local Roughness of Protein Surfaces, J. Chem. Inf. Comput. Sci., 1992, 32(1), 120-122

A. Min Hu is a graduate student at Zhejiang University. Her research interests include scientific data visualization and bioinformatics.

B. Dr. Wei Chen is an associate professor at the State Key Lab of CAD&CG at Zhejiang University, P.R. China. He received his PhD degree in 2002 from the Department of Applied Mathematics of Zhejiang University. He has performed research in volume rendering and related technical areas for the past three years. His current interests include hardware accelerated visualization, photo-realistic rendering, bio-medical imaging, and digital geometry processing.

C. Tao Zhang is a graduate student at Zhejiang University. His research interests include computer graphics and bioinformatics.

D. Qunsheng Peng is a professor of computer graphics at Zhejiang University. His research interests include realistic image synthesis, computer animation, scientific data visualization, virtual reality, bio-molecule modeling. Prof. Peng graduated from Beijing Mechanical College in 1970 and received a Ph.D from the Department of Computing Studies, University of East Anglia in 1983. He serves currently as a member of the editorial boards of several international and Chinese journals.



E. Liguang Xie is an undergraduate student at Zhejiang University. He majors in Computer Science.