Extracting Fractal Features for Analyzing Protein Structure

Yu Tao
Shenzhen Graduate School
Tsinghua University
taoyu@tsinghua-sz.edu.cn

Thomas R. Ioerger
Dept. of Computer Science
Texas A&M University
ioerger@cs.tamu.edu

James C. Sacchettini
Dept. of Biochemistry
Texas A&M University
sacchett@tamu.edu

Abstract

This paper is concerned with the development of a computational methodology based on fractal geometry for determining 3D structure of protein with imagery projection operations. In this investigation, the density-map image is a 2D projection of the 3D electron density map according to its depth of the density distribution along the projection direction. We extract fractal features of the density-map image in a region and use these features to look for candidate regions with similar patterns of density-map. We analyze its fractal signatures for determining 3D pattern of regions of density distribution. This contribution presents preliminary results of such a study, wherein the protein surface was assumed to be a fractal, and computed fractal feature (fractal dimension and fractal signature) were analyzed and found to possess fairly reasonable pattern for improving the discrimination abilities of the protein structure.

1 Introduction

A fundamental goal of research in molecular biology is to understand protein structure. To come to an understanding of complex biological macrolecules, we need to extract the features from many different sources. Once these features are recognized, this partial structure information can be used to improve the determination of crystal structure. A variety of computational methods have been proposed to assist the interpretation of electron density maps. One important method was proposed based on spatial reasoning (or computational imagery) to identify and categorize patterns in electron density maps [3]. They suggest using an array to represent the image, and then derive a description of its components in terms of their spatial relations, and it also provides an abstract representation without significant loss of information. However these representations offer only limited possibilities, therefore many alternative representations have developed to extend it for the interpretation of electron density maps. The TEXTAL system is a 3D pattern recognition approach to automating the interpretation of electron density maps [4]. It employs the pattern recognition technique of feature extraction to calculate numeric values that characterize various geometric aspects that make each local pattern of density unique. This novel approach was able to efficiently and quickly identify candidate regions that are likely to have truly similar patterns with a set of features. In order to explore the idea of pattern matching in X-ray crystallography for developing a set of features, we need to incorporate more powerful features into the TEXTAL system. In this paper, we propose extracting fractal features from the electron density maps to aid in the determination of new protein structure for improving the TEXTAL system.

The concept of fractals has recently been applied to a number of properties of proteins. As was the case for other new concepts, a fractal description appealed to many scientists. As a result, there has been a rapid accumulation of new information in this area, or at least the presentation of existing information in a new form. Surface representations of proteins have provided a powerful approach for characterizing the structure, folding, interactions, and properties [6, 7]. Fractal surface can be used to characterize the roughness of protein surfaces. Therefore, this paper is concerned with the development of a computational methodology based on fractal geometry for determining 3D structure of protein with imagery projection operations. We extract fractal features of the density-map image in a region and use these features to look for candidate regions with similar patterns of density-map. We analyze the fractal signatures of the density-map image for determining 3D pattern of regions of density distribution.

2 3D to 2D Projection

A protein consists of a set of points in three dimensions. Each point has an assigned identification number and a position defined by three coordinates in a Cartesian system based on the electron density map. The determination of molecular structures from x-ray diffraction data belongs to

the general class of image reconstruction exercises from incomplete and/or noisy data. Researchers in artificial intelligence and computer vision has long been concerned with such problems. In recent years, the digitized range data have become available from both active and passive sensors, and the quality of these data has been steadily improving. The range data can be produced in the form of an array of numbers, referred to as a range image (or density-map image), where the numbers quantify the distances from the projected plane to object surfaces within the field of view along an arbitrary viewpoint. The 2D image region approximates the 3D array data of the corresponding object surfaces in the field of view. Thus, considerable research has been carried out on extracting 3D information from one or more 2D images [1].

In this paper, we consider spatial relationships through the use of coordinate systems. In this method, each $x,\ y$ coordinate has an associated third dimension (z-coordinate) representing pixel intensity (e.g. density value). For reference purposes, we assume the existence of a world coordinate system that is placed at any convenient location. Objects are positioned in space relative to this coordinate system by means of translation and rotation parameters. We refer to the translation parameters of an object as the vector α and to the rotation parameters of an object as the vector θ. The number of parameters for each vector depends on the dimension of the depth map recognition problem. For example, the 2D problem requires a total of three parameters. For the 3D case, we write the necessary six parameters as follows: $\alpha = (\alpha,\ \beta,\ \gamma) \quad and \quad \theta = (\theta,\ \phi,\ \psi)$. We define our would model W as a set of ordered triples (object, translation, rotation): $W = (A_n,\ \alpha_n,\ \theta_n)\}^N{}_{n=1}$, where A_n is the vector of nth object with position α_n and orientation θ_n.

Let us assume an orthographic projection model. We write the projection as $f(x) = g_{A,\alpha,\theta}(x)$, where x is the vector of the spatial variables of the projection plane of the object. Since objects do not occupy all space, we need a convention for the value of the density-map function for values of the spatial vector $f(\rho) = (x, y, z)$. It maps the electron density distribution function ρ onto the 2D viewing plane from the desired point of view. If the point $f_{(x,y)}(\rho)$ cannot lie on an object surface, we assign the value of $-\infty$ to $f(\rho)$. Hence, we can write the three projections of a set of M, N, L objects surface that correspond to the X, Y, Z coordinates based on the orthographic projection model as follows, respectively:

$$f_{(x,y)}(\rho) = \sum_{k=1}^{L} g_{A_k,\alpha_k,\theta_k(\rho)},$$

$$f_{(x,z)}(\rho) = \sum_{j=1}^{N} g_{A_j,\alpha_j,\theta_j(\rho)},$$

$$f_{(y,z)}(\rho) = \sum_{i=1}^{M} g_{A_i,\alpha_i,\theta_i(\rho)}.$$

In this study, once the electron density function is known, it can represent the density-map as an image in three dimensions using the means of image processing. If we considered the pixel intensity as the height above a plane, then the intensity surface of a density-map image can be viewed as a fractal surface.

3 Computation of Fractal Dimension for Surfaces

For a fractal surface, there are relationships among fractal dimension, scaling, power spectrum and area size. While the definition of fractal dimension by self-similarity is straightforward, it is often difficult to estimate or compute, given the image data. However, a related measure of fractal dimension, the box dimension, can be computed more easily from the image data [2, 5]. In our algorithm, the means of estimating $N(L)$ is the key feature. Let $p(m, L)$ define the probability that there are m points within a box of size L (i.e. cube of side L), which is centered about a point on the image surface (see Figure 1). For all L, $p(m, L)$ is normalized, as $\sum_{m=1}^{N}(m, L) = 1$, where N is the number of possible points within the box. Let S be the number of image points (i.e. pixels in an image). If one overlays the image with boxes of side L, then the number of boxes with m points inside the box is estimated to be $(S/m)p(m, L)$. Therefore, the expected total number of boxes needed to cover the whole image is

$$\langle N(L)\rangle = \sum_{m=1}^{N} \frac{S}{m} p(m, L) = S \sum_{m=1}^{N} \frac{1}{m} p(m, L).$$

Hence, if we let $N(L) = \sum_{m=1}^{N} \frac{1}{m} p(m, L)$, this value

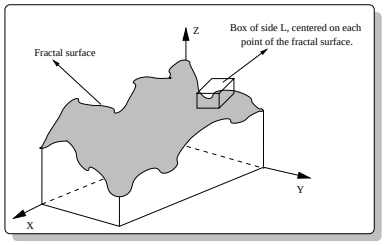

Figure 1. Principle of computing the density distribution that there are m points within a box of side L, $p(m, L)$.

is also proportional to L^{-D} and the box dimension can

be estimated by calculating $p(m, L)$ and $N(L)$ for various values of L, and by doing a least square fit on $[\log(L), -\log(N(L))]$. To estimate $p(m, L)$, one must center the cube of size L around an image point and count the number of neighboring points over the image gives the frequency of occurrence of m. This is normalized to obtain $p(m, L)$. Values of L are chosen to be even to simplify the centering process. Also, the centering and counting activity is restricted to pixels having all their neighbors inside the image. This obviously will leave out image portions of $width = L/2$ on the borders. This reduced image is then considered for the counting process. As is seen, large values of L results in increased image areas from being excluded during the counting process, thereby increasing uncertainty about counts near border areas of the image. This is one of the sources of errors for the estimation of $p(m, L)$ and thereby fractal dimension D. Additionally, the computation time grows with the L value. Hence, $L = 2, 4, 6, 8, \ldots, 32$ were chosen for this work. A number of factors could introduce uncertainty in the computation of $p(m, L)$ and thereby in the estimates of fractal dimension D. Using larger L values effectively reduce the image available for computation of $p(m, L)$ (border areas get excluded), and so statistical significance of $p(m, L)$ is reduced. This factor along with computation time (which increases with L) dictates the range of L values that could be used. "Effects of local slope" in the set $[\log(L), -\log(N(L))]$ has been shown to underestimate the value of fractal dimension D. For surfaces with large D values, it has been shown that there exists a lower bound on the smallest box size that could be used to obtain reasonable estimates of fractal dimension D.

4 Fractal Features for Recognition Purposes

Recognition and segmentation of objects and regions in natural scenes necessitates features which can provide unambiguous discrimination, and at the same time be insensitive to scene perturbations. Researchers have found that fractal feature are quite effective for this purpose [2, 5]. Among the various fractal features which could be computed from am image surface, the fractal dimension D is primary. Theoretically it is invariant to scaling, and known to characterize the roughness of the surface. However, it has been observed that two differently appearing surfaces could have the same value of D. To overcome this, Mandelbrot [8] introduced the term called lacunarity, which quantifies the denseness of an image surface. Many definitions of this term have been proposed and the basic idea in all these is to quantify the "gaps and lacunae" present in a given surface. Here, this term is called the fractal signature of the density-map image surface. One of the useful definitions of this term as suggested by Mandelbrot [8] is $(\frac{M}{E(M)} - 1)^2$, where M is the mass of the fractal and $E(M)$ the expected mass. In other words, this definition measures the deviation between the actual mass and the expected mass. Very similar to the case of measuring the length of a coastline, the mass of a fractal set is dependent on the length of the measuring yardstick and the power law $M(L) = KL^D$, and is a function of L. Calculation of the fractal signature $FS(L)$ is based on the second order statistics of $p(m, L)$. Defining $M(L)$ and $M^2(L)$, it is defined as the fractal signature (FS) as follow:

$$M(L) = \sum_{m=1}^{N} mp(m, L), \quad M^2(L) = \sum_{m=1}^{N} m^2 p(m, L)$$

$$FS(L) = (\frac{M^2(L) - [M(L)]^2}{[M(L)]^2}) \times 100$$

Both the fractal dimension and fractal signature are functions of the box size L.

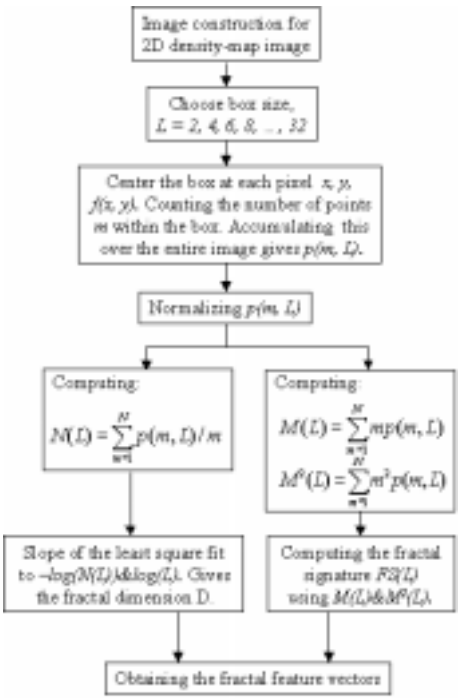

Figure 2. Procedure for computing fractal dimension and fractal signature

5 Discrimination of Crystal Structures

The protein structure data in the Protein Data Bank can be converted by Fourier transform into real 3D array data of an electron density map. The problem of determining the structure of a protein crystal from diffraction data belongs to the general class of image reconstruction problems.

Protein ID	Chosen Regions	D_{mean}	$Var(D)$	FS_{mean}	$Var(FS)$
1A0I.pdb	Atom ♯1-20	2.843	0.005	7.260	0.007
3NLL.pdb	Atom ♯1-20	2.600	0.008	5.436	0.003
3MDD.pdb	Atom ♯1-20	2.865	0.006	8.134	0.012
1AIE.pdb	Atom ♯1-20	2.132	0.006	1.271	0.003
3NUL.pdb	Atom ♯1-20	2.398	0.014	2.310	0.016

Table 1. Mean fractal dimensions D_{mean}, variances of the fractal dimension, mean fractal signature FS_{mean}, and variances of fractal signatures for different protein structures.

The goal of the reconstruction is to produce a complete image which contains the integration of the available features of the 3D representation in the protein structure. Figure 2 briefly illustrates the procedure of extracting fractal features from the electron density map. At first, a 3D array data of an electron density map is projected into three 2D density-map images with the imagery projection operation. Then, the density-map images are assumed to be the fractal surfaces. For a fractal surface, we calculate the value of fractal dimension for the fractal surfaces. The features of a perfect fractal surface can be represented by a single value fractal dimension. But, in practice, the real surface found in density-map images are not perfect fractal surfaces, and therefore do not have a constant fractal dimension over all scales, hence, the features of real surfaces cannot depend on only a single value for the fractal dimension. The normalized fractal signature provides a way to create the fractal feature vector for analysis of the density-map images in spite of the imperfect agreement with fractal theory. We have applied our method to extract the fractal features from the 3D array representation of five different electron density maps. Their feature vectors $\{D_{mean}, Var(D), FS_{mean}, Var(FS)\}^T$ are respectively presented in Table 1 .

Once we have these feature vectors, we compared them to the training set by calculating the weighted Euclidean distance (WED) between the different features vectors, the smallest distance was considered the match [4]. Results are presented demonstrating the utility of this approach for protein crystal structure determination.

6 Conclusions

In this short contribution, the authors have investigated and presented preliminary results on the use of fractal features for discrimination of protein structures. The success reported by researchers in using the imagery projection operation and the fractal features extraction for 3D object recognition in the field of image analysis and image processing has guided this work. Fractal analysis can be useful in the interpretation of the available (but complex) results. From the limited results on hand, it can be conjectured that this novel approach appears promising and further research is necessary to take full advantage of it for automated protein structure determination.

This work was supported in part by grant number R21-GM-59398 from the National Institutes of Health.

References

[1] P. J. Besl and R. C. Jain. "Invariant surface characteristics for 3D object recognition in range images," *Computer Vision, Graphics and Image Processing*, Vol. 33, pp. 33-80, 1986.

[2] S. Chen, J. M. Keller, and R. M. Crownover. "On the calculation of fractal features from images," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 15, No. 10, pp. 1087-1090, 1993.

[3] J. I. Glasgow and D. Papadias. "Computational imagery," *Cognitive Science*, Vol. 16, No. 3, pp. 355-394, 1992.

[4] T. R. Holton, J. A. Christopher, T. R. Ioerger, and J. C. Sacchettini. "Determining protein structure from electron density maps using pattern matching," *Acta Crystallographica*, D56, pp. 722-734, 2000.

[5] J. M. Keller, S. Chen, and R. M. Crownover. "Texture description through fractal geometry," *Comput. Vision Graphics Image Processing*, Vol. 45, pp. 150-166, 1989.

[6] N. N. Krasnogorskaya, E.F. Legushs, and N.J. Tsvileneva. "Fractal structure theory application in crystallography researches ", in *Proceedings of the 2nd International Workshop on Computer Science and Information Technologies CSIT'2000*, Ufa, Russia, Vol. 3, pp. 057-059, 2000.

[7] M. Lewis and D. C. Rees. "Fractal surfaces of proteins," *Science*, Vol. 230, pp. 1163-1165, 1985.

[8] B. B. Mandelbrot. *The Fractal Geometry of Nature*, New York: Freeman, 1983.