International Journal of Modern Physics B
Vol. 17, Nos. 22, 23 & 24 (2003) 4367–4375
© World Scientific Publishing Company

World Scientific
www.worldscientific.com

ITERATED FUNCTION SYSTEM AND MULTIFRACTAL ANALYSIS OF BIOLOGICAL SEQUENCES*

ZU-GUO YU

Program in Statistics and Operations Research,
Queensland University of Technology, GPO Box 2434, Brisbane, Q 4001, Australia
Department of Mathematics, Xiangtan University, Hunan 411105, China
yuzg@hotmail.com,z.yu@qut.edu.au.

VO ANH

Program in Statistics and Operations Research,
Queensland University of Technology, GPO Box 2434, Brisbane, Q 4001, Australia

KA-SING LAU

Department of Mathematics, Chinese University of Hong Kong,
Shatin, Hong Kong, China

Received 8 August 2002

The fractal method has been successfully used to study many problems in physics, mathematics, engineering, finance, even in biology till now. In the past decade or so there has been a ground swell of interest in unravelling the mysteries of DNA. How to get more bioinformations from these DNA sequences is a challenging problem. The problem of classification and evolution relationship of organisms are the central problems in bioinformatics. And it is also very hard to predict the secondary and space structure of a protein from its amino acid sequence. In this paper, some recent results related these problems obtained through multifractal analysis and iterated function system (IFS) model are introduced.

Keywords: Measure representation; multifractal analysis; IFS (RIFS) model; complete genome; length sequence; protein.

1. Introduction

The concept of "fractal" was proposed by Benoit Mandelbrot[1] in the later of 1970s. Fractal geometry provides a mathematical formalism for describing complex spatial and dynamical structures[1,2] (e.g. the strange attractor of a chaotic dynamical system is usually a fractal). Multifractal analysis was initially proposed to treat

*The research was partially supported by QUT's Postdoctoral Research Support Grant No. 9900658 and the Youth Foundation of Chinese National Natural Science Foundation (No. 10101022), the RGC Earmarked Grant CUHK 4215/99P.

turbulence data. This kind of analysis is a useful way to characterise the spatial inhomogeneity of both theoretical and experimental fractal patterns[3] and play an important role in the fractal theory.

In the past decade or so there has been a ground swell of interest in unravelling the mysteries of DNA. The heredity information of most organisms is encoded in a universal way in long chains of nucleic acids formed by four different nucleotides, namely adenine (a), cytosine (c), guanine (g) and thymine (t). One of the challenges of DNA sequence analysis is to determine the patterns in these sequences. Problems related to the classification and evolution of organisms are also important. A significant contribution in these studies is to investigate the long-range correlation in DNA sequences.[4–20] The availability of complete genomes since 1995[21] induces the possibility to establish some global properties of these sequences. A time series model was proposed by Yu *et al.*[22–24] to study the correlation property of coding segments and length sequences of complete genome.

The global and visual methods can amplify the difference between a DNA sequence and a random sequence as well as to distinguish DNA sequences themselves in more details.[25] After the famous chaos game representation of DNA sequences proposed by Jeffrey *et al*,[26,27] Hao *et al.*[25] proposed a visualisation method based on counting and coarse-graining the frequency of appearance of substrings with a given length. They called it the *portrait* of an organism. They found that there exist some fractal patterns in the portraits which are induced by avoiding and under-represented strings. The fractal dimension of the limit set of portraits was also discussed.[28,29] The connection between the Hao's scheme and the chaos game representation is established through the multifractal property.[30] Yu *et al.*[31] proposed the measure representation of complete genomes followed by the multifractal analysis. The multifractal analysis of the length sequences based on the complete genome was performed.[32]

Twenty different kinds of amino acids are found in proteins. The three-dimensional structure of proteins is a complex physical and mathematical problem of prime importance in molecular biology, medicine, and pharmacology.[33,34] The central dogma motivating structure prediction is that: "the three dimensional structure of a protein is determined by its amino acid sequence and its environment without the obligatory role of extrinsic factors".[35,36] How to predict the high level structures (secondary and space structures) from the amino acid sequence is a challenge problem in science, in particular to the large proteins. A number of coarse-grained models have been proposed to provide insight to these very complicated issues.[36] A well known model in this class is the HP model proposed by Dill *et al.*[37] In this model 20 kinds of amino acids are divided into two types, hydrophobic (H) (or non-polar) and polar (P) (or hydrophilic). In last decade the HP model has been extensively studied by several groups.[34,38,39] After studying the model on lattices, Li *et al.*[38] found there are small number of structures with exceptionally high designability which a large number of protein sequences possess as their ground states. These highly designable structures are found to have protein-like secondary structures.[34,38,40] But the HP model may be too simple and lacks enough consideration on the heterogeneity and the complexity of the natural set of residues.[41] According to Brown,[42] in the HP model, one can divide the polar class into three classes: positive polar, uncharged polar and negative polar. So 20 different kinds of amino acids can be divided into four classes: non-polar, negative polar, uncharged polar and positive polar. In this model, one gets more details than in the HP model. We call this model a *detailed HP model.* In this paper we will adopt the detailed HP model.

The fractal method has been used to study the protein backbone,[43] the accessible surface of protein[43–46] and protein potential energy landscapes.[47] The multifractal analysis of solvent accessibilities in proteins was done by Balafas and Dewey.[48] The model used to fit the multifractal spectrum was also discussed.[48] But the parameters derived in their multifractal analysis cannot be used to predict the structural classification of a protein from its amino acid sequence.

The amino acid sequence of a protein is also called a *protein sequence* in this paper. Based the idea of DNA walk model and different mapping, a decoded walk model was proposed to study the correlation property of protein sequences by Pande *et al.*[49] using "Bridge analysis" and Straint and Dewey[50] using multifractal analysis. Deviations of the decoded walk from random behaviour provides evidence of memory.

Inspired by the idea of measure representation of DNA sequence,[31] we also proposed a visual representation — measure representation of protein sequences based on the detailed HP model.[51]

To our knowledge,[52] it is much harder to simulate a measure than to fit its multifractal spectrum (because different measures may have the same multifractal spectrum). The iterated function systems (IFS) model proposed by Barnsley and Demko[53] is a powerful tool in fractal theory (many fractals such as the Cantor set can be generated by the IFS model). We found that the recurrent IFS (RIFS) model can be used to simulate the measure representation of complete genomes while the IFS model can be used to simulate the measure representation of protein sequences. In this paper, the estimated parameters in RIFS or IFS model are used to discuss the classification of living organisms and the structural classification of large proteins.

2. Measure Representation of DNA and Protein Sequences

We call any string made of K letters from the set $\{g, c, a, t\}$ a K-string. For a given K, there are in total 4^K different K-strings. In order to count the number of each kind of K-strings in a given DNA sequence, 4^K counters are needed. We divide the interval $[0, 1]$ into 4^K disjoint subintervals, and use each subinterval to represent a counter. Using the observed frequencies of all 4^K kinds of K-strings in the complete genome, we can define a measure μ_K on the interval $[0, 1]$ in one dimensional space.[31] We call μ_K the *measure representation* of the organism corresponding to the given K.

Twenty different kinds of amino acids are found in proteins. In the detailed HP model they can be divided in to four classes: non-polar, negative polar, uncharged polar and positive polar. The eight residues designating the non-polar class are: ALA, ILE, LEU, MET, PHE, PRO, TRP, VAL; the two residues designating the negative polar class are: ASP, GLU; the seven residues designating the uncharged polar class are: ASN, CYS, GLN, GLY, SER, THR, TYR; and the remaining three residues: ARG, HIS, LYS designate the positive polar class.

For a given protein sequence with length L, $s = s_1 \cdots s_L$, where s_i is one of the twenty kinds of amino acids for $i = 1, \cdots, L$, we define

$$a_i = \begin{cases} 0, & \text{if } s_i \text{ is non-polar,} \\ 1, & \text{if } s_i \text{ is negative polar,} \\ 2, & \text{if } s_i \text{ is uncharged polar,} \\ 3, & \text{if } s_i \text{ is positive polar.} \end{cases} \tag{1}$$

So we can obtain a sequence $X(s) = a_1 \cdots a_L$, where a_i is a letter in the alphabet $\{0, 1, 2, 3\}$. Using the same idea as DNA sequences,[31] we can define the measure representation μ_K of K-strings of the given protein sequence.

3. Multifractal Analysis and IFS (RIFS) Model

The most common numerical implementations of multifractal analysis are the so-called *fixed-size box-counting algorithms.*[54] In the one-dimensional case, for a given measure μ with support $E \subset \mathbf{R}$, and q a real number, we can define the scaling exponent $\tau(q)$ and the generalized fractal dimensions D_q of the measure as which in Ref. 31. D_1 is called the *information dimension* and D_2 the *correlation dimension.* The D_q of the positive values of q give relevance to the regions where the measure is large, i.e. to the coding or noncoding segments which are relatively long. The D_q of the negative values of q deal with the structure and the properties of the most rarefied regions of the measure, i.e. to the segments which are relatively short.

By following the thermodynamic formulation of multifractal measures, Canessa[55] derived an expression for the "analogous" specific heat as $C_q \equiv -\frac{\partial^2 \tau(q)}{\partial q^2} \approx 2\tau(q) - \tau(q+1) - \tau(q-1)$. He showed that the form of C_q resembles a classical phase transition at a critical point for financial time series. The types of phase transitions are helpful to discuss the classification of bacteria.

In order to simulate the measure representation of the complete genome, Anh *et al.*[56] proposed the *iterated function systems* (IFS) model and the recurrent IFS model. IFS is the name given by Barnsley and Demko[53] originally to a system of contractive maps $w = \{w_1, w_2, \cdots, w_N\}$. Recurrent IFS (RIFS) is a kind of extension from IFS. Usually one can generate the attractor of an IFS or RIFS through the famous *Chaos game* process.[57] Let μ be the invariant measure on the attractor E of an IFS or RIFS, for the Borel subset $B \subset E$, $\mu(B)$ is the relative visitation frequency of B during the chaos game.[53]

The coefficients in the contractive maps and the probabilities in the IFS or RIFS model are the parameters to be estimated for a real measure which we want to simulate. Vrscay[57] introduced a moment method to perform this task. From the measure representation of a complete genome or protein sequence, we see that it is natural to choose $N = 4$ and

$$w_1(x) = x/4, \qquad w_2(x) = x/4+1/4, \qquad w_3(x) = x/4+1/2, \qquad w_4(x) = x/4+3/4$$

in the IFS or RIFS model. For a given measure representation of a complete genome or protein sequence, we obtain the estimated values of the probabilities p_1, p_2, p_3, p_4 in IFS model or the matrix of probabilities $\mathbf{P} = (p_{ij})$ by the moment method. Based on the estimated values of the probabilities, we can use the chaos game to generate a histogram approximation of the invariant measure of IFS or RIFS which we can compare with the real measure representation of the complete genome or protein sequence.

4. Applications to the Biological Sequence Analysis

Till now more than 50 complete genomes of Archaea and Eubacteria are available in public databases (for example in Genbank at web site ftp://ncbi.nlm.nih.gov/genbank/genomes/).

The multifractal analysis were performed on the measure representations of a large number of complete genomes.[31] For examples, the D_q and C_q curves of some organisms are shown in Fig. 1. From the measure representations and the values of the D_q spectra and related C_q curves, it was concluded that these complete genomes are not random sequences. For substrings with length $K = 8$, the D_q spectra of all organisms studied are multifractal-like and sufficiently smooth for the C_q curves to be meaningful. With the decreasing value of K, the multifractality lessens. The C_q curves of all bacteria resemble a classical phase transition at a critical point. But the 'analogous' phase transitions of chromosomes of non-bacteria organisms are different. Apart from Chromosome 1 of *C. elegans*, they exhibit the shape of double-peaked specific heat function.

We simulated the measure representations of the complete genomes of many organisms using the IFS and RIFS models.[56] We found that RIFS is a good model to simulate the measure representation of complete genome of organisms. For example, the histogram of substrings in the genome of *Buchnera sp. APS* for $K = 8$ is given in the left figure of Fig. 2. Self-similarity is apparent in the measure. The histogram approximation of the generated measure of *Buchnera sp. APS* using the RIFS model is shown in the right figure of Fig. 2. It is seen that the RIFS simulation traces very well the original measure representation of the complete genome.

It is well known that all statistical method and nonlinear scale method require enough data samples. The methods introduced in the previous sections can only

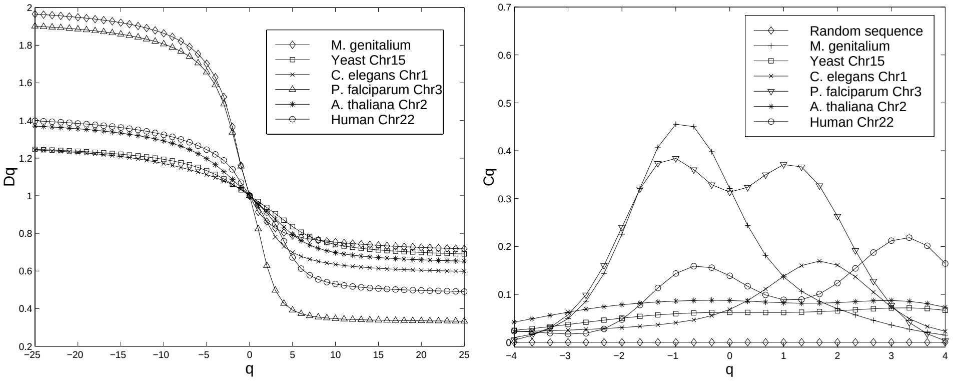

Fig. 1. Dimension spectra (Left) and "Analogous" specific heat (Right) of Chromosome 22 of Homo sapiens, Chromosome 2 of A. thaliana, Chromosome 3 of P. falciparum, Chromosome 1 of C. elegans, Chromosome 15 of S. cerevisiae and M. genitalium.

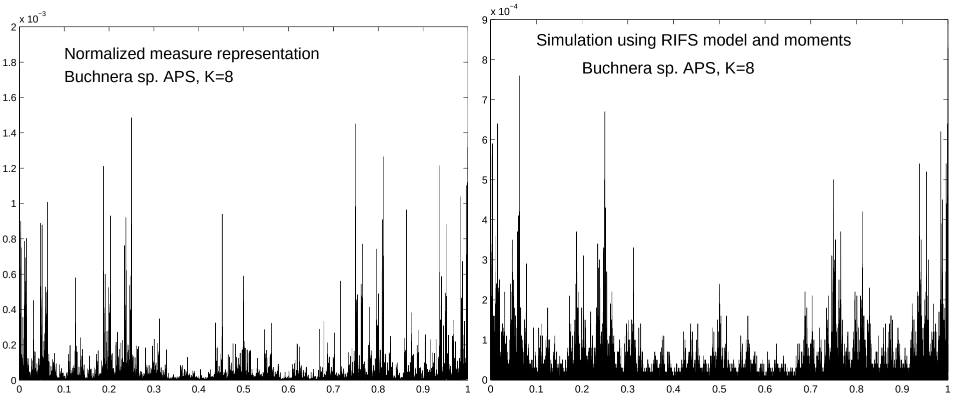

Fig. 2. The measure representation (left) and the RIFS simulation (right) of the complete genome of Buchnera sp. APS when $K = 8$.

be used to analyse long protein sequences (corresponding to large proteins). The amino acid sequence of 32 large proteins are selected from RCSB Protein Data Bank (PDB) (http://www.rcsb.org/pdb/index.html). These 32 proteins belong to five structure classes[58] according to their secondary structures: α, β, $\alpha + \beta$ (α,β alternate), α/β (α, β segregate) and others (no α and no β) proteins. First we convert the amino acid sequences of these proteins to their measure representations with $K = 5$ according to the method introduced in Section 2. If K is too small, there are not enough combinations of letters from the set $\{0, 1, 2, 3\}$, therefore there is no statistical sense. And if K is too big, the frequencies of most substrings are zero. So we cannot obtain any biological information from the measure representation. Considering the length of the selected proteins which ranges from 350 to 1000, we think it is suitable to choose $K = 5$.

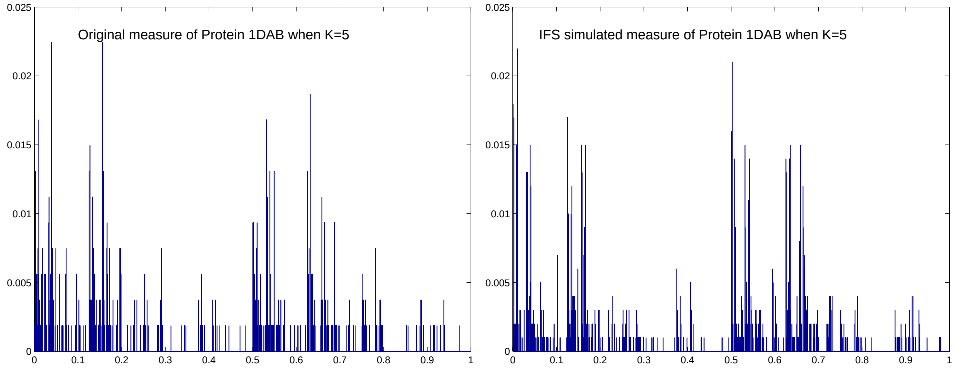

Fig. 3. The measure representation (left) and the IFS simulation (right) of protein *P.69 Pertactin* (PDB ID: 1DAB)

We found the IFS model is better than the RIFS model to simulate the measure representation of protein sequences.[51] For example, we show the histograms of measure representation and simulated measures of protein *P.69 Pertactin* (PDB ID: 1DAB) in Fig. 3.

From Fig. 3, one can see that the difference between measure representation and IFS simulated measure is very small. Once the probabilities are determined, the IFS model is obtained. Hence the probabilities obtained from the IFS model can be used to represent the measure representation of the protein sequence. From the estimated parameters in the IFS model for the 32 selected large proteins, we find the probability p_3 (which corresponding to the uncharged polar property) can be used to distinguish the structural class of proteins from α class and β class (values of p_3 of proteins in class α are less than those of proteins in class β), and the probability p_1 (which corresponds to the non-polar property) can be used to distinguish the structural class of proteins from class $\alpha + \beta$ and class α/β (values of p_1 of proteins in class α/β are less than those of proteins in class $\alpha + \beta$). Hence we believe that the non-polar residues and uncharged residues play a more important role than other kinds of residues in the protein folding process. This information is useful for the prediction of protein structure.

References

1. B. B. Mandelbrot, *The Fractal Geometry of Nature* (Academic, New York, 1983).
2. J. Feder, *Fractals* (Plenum, New York, 1988).
3. P. Grassberger and I. Procaccia, *Phys. Rev. Lett.* **50**, 346 (1983).
4. W. Li and K. Kaneko, *Europhys. Lett.* **17** , 655 (1992); W. Li, T. Marr and K. Kaneko, *Physica* **D75**, 392 (1994).
5. C. K. Peng, S. Buldyrev, A. L. Goldberg, S. Havlin, F. Sciortino, M. Simons and H. E. Stanley, *Nature* **356**, 168 (1992).
6. J. Maddox, *Nature* **358**, 103 (1992).
7. S. Nee, *Nature* **357**, 450 (1992).
8. C. A. Chatzidimitriou-Dreismann and D. Larhammar, *Nature* **361**, 212 (1993).

9. V. V. Prabhu and J. M. Claverie, *Nature* **359**, 782 (1992).
10. S. Karlin and V. Brendel, *Science* **259**, 677 (1993).
11. (a) R. Voss, *Phys. Rev. Lett.* **68**, 3805 (1992); (b) *Fractals* **2**, 1 (1994).
12. H. E. Stanley, S. V. Buldyrev, A. L. Goldberg, Z. D. Goldberg, S. Havlin, R. N. Mantegna, S. M. Ossadnik, C. K. Peng and M. Simons, *Physica* **A205**, 214 (1994).
13. H. Herzel, W. Ebeling and A. O. Schmitt, *Phys. Rev.* **E50**, 5061 (1994).
14. P. Allegrini, M. Barbi, P. Grigolini and B. J. West, *Phys. Rev.* **E52**, 5281 (1995).
15. S. V. Buldyrev, A. L. Goldberger, S. Havlin, R. N. Mantegna, M. E. Matsa, C. K. Peng, M, Simons and H. E. Stanley, *Phys. Rev.* **E51**(5), 5084 (1995).
16. A. Arneodo, E. Bacry, P. V. Graves and J. F. Muzy, *Phys. Rev. Lett.* **74**, 3293 (1995).
17. A. K. Mohanty and A. V. S. S. Narayana Rao, *Phys. Rev. Lett.* **84**(8), 1832 (2000).
18. L. Luo, W. Lee, L. Jia, F. Ji and L. Tsai, *Phys. Rev.* **E58**(1), 861 (1998).
19. Z. G. Yu and G. Y. Chen, *Comm. Theor. Phys.* **33**(4), 673 (2000).
20. C. L. Berthelsen, J. A. Glazier and M. H. Skolnick, *Phys. Rev.* **A45**(12), 8902 (1992).
21. C. M. Fraser *et al.*, *Science* **270**, 397 (1995).
22. Z. G. Yu and B. Wang, *Chaos, Solitons and Fractals* **12**(3), 519 (2001).
23. Z. G. Yu and V. V. Anh, *Chaos, Soliton and Fractals* **12**(10), 1827 (2001).
24. Z. G. Yu, V. V. Anh and Bin Wang, *Phys. Rev.* **E63**, 11903 (2001).
25. B. L. Hao, H. C. Lee and S. Y. Zhang, *Chaos,Solitons and Fractals*, **11**(6), 825 (2000).
26. H. J. Jeffrey, *Nucleic Acids Research* **18**(8), 2163 (1990).
27. N. Goldman, *Nucleic Acids Research* **21**(10), 2487 (1993).
28. Z. G. Yu, B. L. Hao, H. M. Xie and G. Y. Chen, *Chaos, Solitons and Fractals* **11**(14), 2215 (2000).
29. B. L. Hao, H. M. Xie, Z. G. Yu and G. Y. Chen, *Physica* **A288**, 10 (2001).
30. P. Tino, *Physica* **A304**, 480 (2002).
31. Z. G. Yu, V. V. Anh and K. S. Lau, *Phys. Rev.* **E64**, 031903 (2001).
32. Z. G. Yu, V. V. Anh and K. S. Lau, *Physica* **A301**(1-4), 351 (2001).
33. C. Chothia, *Nature* (London) **357**, 543 (1992).
34. C. T. Shih, Z. Y. Su, J. F. Gwan, B. L. Hao, C. H. Hsieh and H. C. Lee, *Phys. Rev. Lett.* **84**(2), 386 (2000).
35. C. Anfinsen, *Science* **181**, 223 (1973).
36. C. T. Shih, Z. Y. Su, J. F. Gwan, B. L. Hao, C. H. Hsieh and H. C. Lee, *Phys. Rev.* **E65**, 041923 (2002).
37. K. A. Dill, *Biochemistry* **24**, 1501 (1985); H. S. Chan and K. A. Dill, *Macromolecules* **22**, 4559 (1989).
38. H. Li, R. Helling, C. Tang and N. S. Wingreen, *Science* **273**, 666 (1996).
39. B. Wang and Z. G. Yu, *J. Chem. Phys.* **112**, 6084 (2000).
40. C. Micheletti, J. R. Banavar, A. Maritan and F. Seno, *Phys. Rev. Lett.* **80**, 5683 (1998).
41. J. Wang and W. Wang, *Phys. Rev.* **E61**, 6981–6986 (2000).
42. T. A. Brown, *Genetics* (3rd Edition) (CHAPMAN & Hill, London, 1998).
43. T. G. Dewey, *J. Chem. Phys.* **98**, 2250 (1993).
44. P. Pfiefer, U. Welz and H. Wipperman, *Chem. Phys. Lett.* **113**, 535 (1985).
45. B. A. Fedorov, B. B. Fedorov and P. W. Schmidt, *J. Chem. Phys.* **99**, 4076 (1993).
46. M. Lewis and D. C. Rees, *Science* **230**, 1163 (1985).
47. D. A. Lidar, D. Thirumalai, R. Elber and R. B. Gerber, *Phys. Rev.* **E59**, 2231 (1999).
48. J. S. Balafas and T. G. Dewey, *Phys. Rev.* **E52**, 880 (1995).
49. V. S. Pande, A. Y. Grosberg and T. Tanaka, *Proc. Natl. Acad. Sci. USA* **91**, 12972 (1994).
50. B. J. Strait and T. G. Dewey, *Phys. Rev.* **E52**, 6588 (1995).

51. Z. G. Yu, V. V. Anh and K. S. Lau, Fractal analysis of measure representation of large proteins based on the detailed HP model, (2002), submitted to *Physica A*.
52. V. V. Anh, K. S. Lau and Z. G. Yu, *J. Phys. A: Math. Gene.* **34**, 7127 (2001).
53. M. F. Barnsley and S. Demko, *Proc. Roy. Soc. London* **A399**, 243 (1985).
54. T. Halsy, M. Jensen, L. Kadanoff, I. Procaccia and B. Schraiman, *Phys. Rev.* **A33**, 1141 (1986).
55. E. Canessa, *J. Phys. A: Math. Gen.* **33**, 3637 (2000).
56. V. V. Anh, K. S. Lau and Z. G. Yu, *Phys. Rev.* **E66**, 031910 (2002).
57. E. R. Vrscay, in *Fractal Geometry and analysis*, Ed. J. Belair (NATO ASI series, Kluwer Academic Publishers, 1991).
58. R. B. Russell, in *Protein structure prediction: Methods and Protocls*, Ed. D. Webster (Humana Press Inc., Totowa, NJ, 2000).