Dynamics of protein evolution*

MEETA RANI†

Department of Biochemistry, School of Life Sciences, University of Hyderabad, Hyderabad 500 046, India
†Present address: Bioinformatics Centre, Lilly Lobby, National University Hospital, 5 Lower Kent Ridge Road, 119074 Singapore
(Fax, 65-7790724; Email, meeta@bic.nus.edu.sg)

Protein sequences of the SWISS-PROT data bank were analysed by fractal techniques and harmonic analysis. In both cases, the results show the presence of self-affinity, a kind of self-similarity, in the sequences. Self-similarity is a sign of fractality and fractality is a consequence of a chaotic dynamical process. The evolution of the protein sequences is modelled as a dynamical system. The abundance of the fractal form in biology and creation of fractal forms as a result of "chaos" is already established. It may be noted that the word "chaos" here implies that most predictable processes can also become unpredictable under certain conditions, and that the most unpredictable processes are not as unpredictable as they are expected to be. In evolutionary dynamics, this allows scope for mutations and variations in otherwise predictable situations, potentially leading to increased diversity.

1. Introduction

Life on Earth is believed to have originated over 3 billion years ago and various life forms have evolved since then. There are several theories which aim to explain the process of evolution and more are being developed. However, what is intuitively clear is that the ultimate theory must be a general and universal one capable of explaining the evolution of matter (physical and chemical evolution) as well as the evolution of life (biological evolution). I have been interested in developing such a theory using concepts of "dynamical systems". For this living systems are treated as non-equilibrium systems and their evolution as a dynamical system. The concept of dynamical systems to try to explain the origin of common patterns and forms in biological systems (i.e., fractals) as a result of the same dynamics is being proposed in this study.

Evolution of a biological process (or function) is also dependent on the evolving form and structure of the participating molecules; one such example is the structure-function relationship of globular proteins. We have studied primary sequences of proteins as a case-study to find patterns in them. Observing the general structure of proteins at its various levels it is proposed that a protein qualifies to be a fractal at all structural levels. Moreover, it is suggested that protein sequences have acquired this fractal nature as a result of the dynamics of the evolution of its sequence.

2. Evolution of a system

When we talk about evolution of a system we refer to the manner in which the observable parameters of the system change with time, space or both. The system whose evolution we are talking about could refer to a system such as the universe, Earth, living organisms or the physiological processes in living systems.

2.1 *Dynamical systems*

According to Devaney, 1988, a dynamical system is that which *evolves* (please note that "evolution" is different from "change" in the sense that evolution is an

*Part of this work was presented at the National Symposium on Evolution of Life.

Keywords. Evolution; dynamical systems; chaos; fractals; proteins

irreversible change and it is analogous to a spontaneous process in thermodynamics) with time according to a specific rule being iterated. For this, an *initial point* is chosen to start the iterative process and successive *iterates* are generated. The successive iterates are said to form an *orbit or trajectory*. If these orbits or trajectories always tend to land up on a particular set, then that set is called an "attractor", as it attracts orbits. *Stable orbits* are those in which a slight change in the initial input does not significantly affect the behaviour of the resulting orbit. However, it has been found that sometimes a point very close to the initial point behaves very unpredictably and such a condition is called "chaos". Very often the set of all such points (with *unstable orbits*) which lead to chaos form a "fractal". Such dynamical systems are also called *chaotic dynamical systems*. Thus the unpredictability of a dynamical system is due to chaos whereas in a stable dynamical system a small change in the initial input does not alter the eventual outcome.

2.2 *Evolution of biological systems*

Any system that evolves in time is a dynamical system. For example, if we consider a chemical reaction, we find that reactants are consumed and products are formed with time. But once the reaction reaches an equilibrium state it does not evolve further as no change occurs with time. On the other hand, in biological systems evolution is irreversible. For example, in case of a living organism equilibrium implies death of the organism. It is well accepted that biological systems are non-equilibrium systems but maintain a quasi-steady state as they are open systems. Hence, in principle, the trends in evolution of life can be understood by studying it as a dynamical system. The field of dynamical systems deals with understanding the rules governing an evolving process and predicting the overall future behaviour of a system on the basis of its past history and dynamics. It may be noted that, for such prediction, only slowly varying parameters can be meaningful and can be studied conveniently.

2.3 *Self-similarity in evolution*

Often, we find similarities between various organisms. Similarities are known to exist among the sequences of their nucleotides or proteins, among their organs (e.g., homologous and analogous organs), their physiological processes or biochemical pathways and even in their morphologies. Mostly these similarities are due to common ancestry as the beneficial and adaptive variations in the organisms which are stable are preserved by the genetic mechanism and are passed on to the next generation. However, it can also be due to adaptive processes (parallel evolution) as seen in the case of analogous organs. This indicates the importance of structural and morphological characteristics in their functional efficiency. As a result, a structure which is successful in this regard is found to be more common and leads to self-similarity in nature. Thus similarity of beneficial structures or processes in evolution may be either due to common ancestry or due to adaptation. This self-similarity gives rise to fractal patterns in living organisms. Hence I propose that both the evolutionary process (trajectory of evolution) and its product i.e., a specific structure or function, are fractals.

2.3a *"Chaotic conditions" in evolution*: If we consider two organisms which had a common ancestor but now have diverged enormously this reminds us of the condition of "chaos" in dynamical systems where two points situated closeby diverge due to very different trajectories. One such example in the evolution of life, could be when a plant cell and animal cell diverged from their immediate common ancestor. Similar examples can be found at various levels of evolution as well as in population dynamics.

2.4 *Fractals in biology—the products of evolution of life*

As observed by Benoit B Mandelbrot, 1983, in his book, "—the patterns of nature are very irregular and fragmented compared with Euclid—a term used in this work to denote all of standard geometry—Nature exhibits not simply a higher degree but an altogether different level of complexity and the number of distinct scales of length of natural patterns is for all practical purposes infinite." Mandelbrot has described these irregular and fragmented patterns in nature by identifying them by a family of shapes called *fractals*. He further mentions that fractals involve chance and both their regularities and irregularities are statistical. But the degree of their irregularity is identical at all scales.

The above mentioned property of statistical self-similarity is the essential quality of fractals in nature (i.e., natural fractals).

Geometrically, a fractal is an entity which possesses invariance or self-similarity at all magnifications. This self-similarity is as a result of a recursive rule being applied to an initial geometrical entity. For example, the well-known geometrical fractal—"von Koch's curve"—is obtained easily by a recursive rule applied to an equilateral triangle, as shown in figure 1. Interestingly, fractals are very common entities in nature also and they arise naturally from the dynamics of chaotic dynamical system. For a lucid reading on this subject please refer to Devaney (1992). Biology particularly abounds in fractal structures. Some examples are the branching patterns of trees; networks of blood vessels; neuronal networks;

shapes of corals, shells, conches; arrangement of vascular bundles in plants; folds of intestinal villi; convolutions in the brain of mammals, etc. According to Goldberger *et al* (1990), though the fractal anatomies in the human body serve apparently disparate functions in different organ systems, several common anatomical and physiological themes emerge. Fractal branches or folds greatly amplify the surface area for absorption (as in the intestine), distribution and collection (in blood vessels), and in information processing (by the nerves). Fractal structures, partly by their redundancy and irregularity are robust and resistant to injury. The robustness allows the fractal structures in the human body to operate under a wide range of conditions and are therefore adaptable and flexible. This plasticity allows systems to cope with the exigencies of an unpredictable and changing environment.

2.5 *Protein evolution: A case study*

We have mentioned that dynamical systems which undergo chaotic dynamics yield fractals. Also fractality is usually reflected as self-similarity or invariance in the system. Hence presence of self-similarity (along with other evidences) is an indication of the fractal nature of the system. We have carried out certain tests on protein sequences and have discovered fractality, suggesting that it evolved as a result of the dynamics of protein structure and function.

It is well accepted that all biological species have developed continuously starting from a single or a very limited number of ancestral species. The great diversity in biological species have resulted as an outcome of numerous natural experiments. According to Dayhoff (1978), a renowned scientist and expert on biomolecular evolution, the "relics" of ancient organisms can be found in the biochemical systems of their living descendants and this dynamic preservation of the biochemical components of the living cells is quite often as rigorous as the preservation of a sedimentary fossil. The exceedingly conservative nature of the evolutionary process has preserved such relics in all living species. Chemical characteristics of organisms particularly the sequence of DNA and proteins can be readily quantified and correlated using logical and statistical methods. The sequences already known contain as much as information as thousands of morphological traits. Each amino acid position in each protein may be considered to be a variable trait with twenty potential levels of distinction. Thus it is possible to construct evolutionary trees based on analysis of the proteins and DNA in various organisms and to establish the evolutionary closeness or distance between various living organisms.

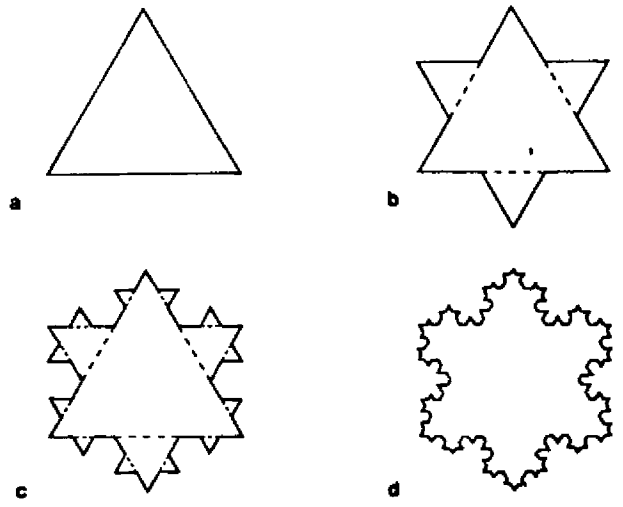

Figure 1. The von Koch curve can be obtained from an equilateral triangle by a simple recursive rule. Each of the three side of the triangle is divided into three equal segments. Then the middle segment in all of them is replaced by a bent line. After a few iterations we obtain the simple von Koch curve. More complex versions can be obtained by taking non-equilateral triangles or by unequal division of the sides of the triangles. Please note that the coastline of islands, a natural fractal, is usefully represented by this geometrical fractal.

2.5a *Proteins as fractals*: Protein molecules are indispensable to the living cell as they run its complete machinery by performing various functions such as enzymes, hormones, carrier molecules, antibodies, regulators of gene expression, signal transducers and structural elements. There are between 10^{10} to 10^{12} proteins in nature, each with a unique sequence of amino acids. The sequence of the amino acids in a protein determines its 3-D structure and the 3-D structure determines its function. Hence, protein activity can be considered as a fractal, formed through the dynamics of its structure, as follows:

Primary structure (amino acids, in a unique sequence)

↓ folds

Secondary structure (α-helices, β-sheets)

↓ folds

Super secondary structures (hairpin loops, α–β–α-structures)

↓ folds

Tertiary structures (well defined 3-D conformation, biologically active).

This process of folding and superfolding of the protein structure reminds us of a recursive process in a dynamical system. Here the *primary structure* may be treated as analogous to the "initial point" and the *process of folding* as "the recursive rule", the *folding path* as the "trajectory or orbit" and the final folded *3-D native*

structure as a "fractal" in a dynamical system. In protein folding, it may be assumed that at each folding stage there are several folding rules which allow large variations and diversity in protein structures. However, for each protein there is a unique path of the successive folding steps, which determines its 3-D structure.

2.5b *Protein evolution—A dynamical system*: Chaotic dynamical systems often result in the formation of fractal structures. If protein evolution has elements of a chaotic dynamical process then one can expect to find fractality in them. Evidence of fractality requires the demonstration of self-similarity in sequences and calculation of the associated fractal dimensions. We also found long range autocorrelations in the distribution of amino acid residues.

We have performed an analysis of the sequences of proteins from a large data base and have found that they possess fractal patterns. Just as in chaotic systems where there exist points (or conditions) whose orbits (i.e., evolution) can give rise to different final states depending upon the choice of the initial conditions, in proteins too, the folding of the protein depends upon the preceding structure and ultimately on the primary sequence of amino acid residues.

3. Studies on protein sequences

Earlier studies on protein sequences (Meeta Rani 1994), also described here relate to demonstration of fractality in protein sequences and to long range correlations as elicited by harmonic analysis. Long range correlations are normally associated with fractals. These two studies are described very briefly below.

3.1 *Methods and results*

I have used the concepts of dynamical systems to model the evolution of protein sequences. The SWISS-PROT Protein Sequence Databank has been used as the source of the protein sequences. The software programs were written in Turbo-Pascal ver. 6.0 and the calculations were performed on a IBM compatible PC-AT 486. All the graphs have been plotted using the plotting package Sigmaplot for Windows ver. 2.01.

3.1a *Discovery of fractal patterns by box-counting method*: Simple statistical tests confirmed that the distribution of amino acids in proteins is non-random. In the process of finding general patterns in them we discovered fractality. Fractality was demonstrated by determination of the box-dimensions using the box-counting algorithm (Barnsley 1988). Further we also discovered general long range correlations in them; such long range correlations are normally associated with fractals.

The box-counting algorithm was carried out as follows. The positional frequencies of each of the 20 amino acids in the data base were calculated and were normalized. Consider the set of points i.e., the normalized positional frequencies of a residue distributed in a rectangular plane. The plane is divided into 4" (figure 2) equal grids

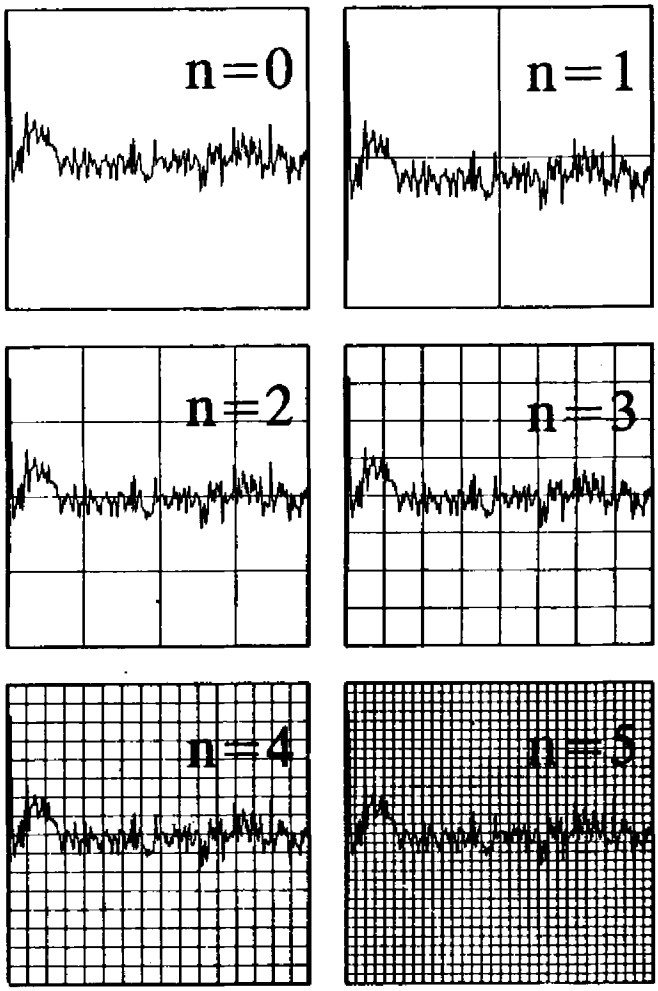

Figure 2. The box-counting algorithm is a method to calculate the box-dimensions of objects and is illustrated here using the example of the positional distribution of alanine as the object. The rectangular plane containing the points corresponding to alanine's frequency (normalized frequency in this case) at the various positions in protein sequences was treated as follows. The plane was divided into 4 equal grids and the number of non-empty grids was counted. This process was repeated till the meaningful physical dimensions of the grids (the x-co-ordinate of the grid corresponded to sequential positions in a protein and hence could not be less than one). The number of non-empty grids is related to the box-dimension calculated as $D = \log(\text{non-empty grids})$ divided by $n(\log 2)$ as the number of divisions approach infinity. The box-dimensions thus calculated, if at all are fractional in nature, then t indicates the fractal nature of the object which was subjected to the test. The tests showed the distribution of all 20 amino acids to be fractals with their box-dimensions lying between 1·2 and 1·3.

(taken for computational convenience as $4^1, 4^2, 4^3, 4^4$ and 4^5) and the total number of boxes containing at least one point (non-empty boxes) are counted. The limiting slope of the line, relating log (number of non-empty boxes) to $n \cdot \log 2$, where n is the order of subdivision (i.e, 1, 2, 3, 4 or 5) gives the fractal dimension D of the set. Since our set is finite, subdivision beyond 4^5 is not physically meaningful and hence were not carried out. Mathematically,

$$D = \lim_{n \to \infty} \frac{\log (box\text{–}count)}{n \cdot \log 2} \qquad (1)$$

where n is the order of iteration and '*box-count*' is the number of boxes that contain at least one point at a given iteration n. The correlation coefficients were very high (>99%) indicating that within a physically meaningful scale, our values are meaningful. The values of the fractal dimensions of the distributions of the amino acid residues are presented in figure 3. The box-dimensions calculated for simulated random sequences are all equal to one. This is expected, as in the absence of preferences, the distribution graphs are straight lines parallel to the position axis and a straight line has dimension of unity. For details of the calculations please refer to Mitra and Meeta Rani (1993).

3.1b *Harmonic analysis indicates long range correlation—sign of fractality*: The presence of fractality in the distribution of amino acids in protein sequences shows that they are not random. This indicates that correlations between residues may exist. We carried out correlation analysis and discovered long range correlations in the positional distributions of the residues. For the correlation analysis, we have treated protein sequences as a time-series (Meeta Rani and Mitra 1995).

A time-series is a set of values recorded for a given parameter at different times. In that sense, it is similar to a dynamical system. The techniques used to study time-series can be extended to spatial situations and we have done that in the case of protein sequences. A protein sequence can be considered as a time-series as the type of amino acid residue changes with sequential position along the polypeptide chain. The autocorrelation and cross-correlation function can be used to find the correlations between residues at any two positions. This is useful in finding patterns, symmetry, periodicity, etc., in protein sequences. To detect the presence of long range correlations, we tried out only autocorrelation analysis for the sake of simplicity. The steps are as follows. The 20 positional distributions were treated as a time-series and their autocorrelations (up to order 99) were calculated. The Fourier transformation of the autocorrelations yields the spectral densities. The x-axis denotes frequencies (in terms of the angular frequency (α) and is plotted as log (α). On the y-axis, the spectral densities are plotted as log [w (α)]. The graph of log [w (α)] vs. log (α) is called the spectrum. The characteristics of the spectrum are related to the parameters of the positional correlation between the members of the pair of residues being studied. The high peaks always refer to strong correlations and lower peaks refer to weak correlations. The presence of these high peaks in the low frequency regions indicates long range correlations and their presence in high frequency regions indicates short range correlations. We found that predominant long-range correlations (figure 4) were seen in the case of each of the 20 homo-pairs of residues. However, the spectra of the pairs from simulated random sequences totally lacked long-range correlations (figure 5). The characteristics of these spectra are also related to fractality. The negative slope of the graph of the spectrum plotted as log (spectral density) vs. log (frequency) is called the spectral exponent and is denoted as β. The spectral exponent β and scaling parameter H of the distribution of the amino acids were calculated. Both β and H reflect how the distribution function changes when the independent variable is scaled [i.e., by multiplying with a given constant (r)]. This kind of scaling is called *self-affinity* (self-affinity is a kind of statistical self-similarity in which one object can be superimposed on the other by applying suitable affine transformations). Thus both β and H provide important information about positional correlations between amino acids within the sequences.

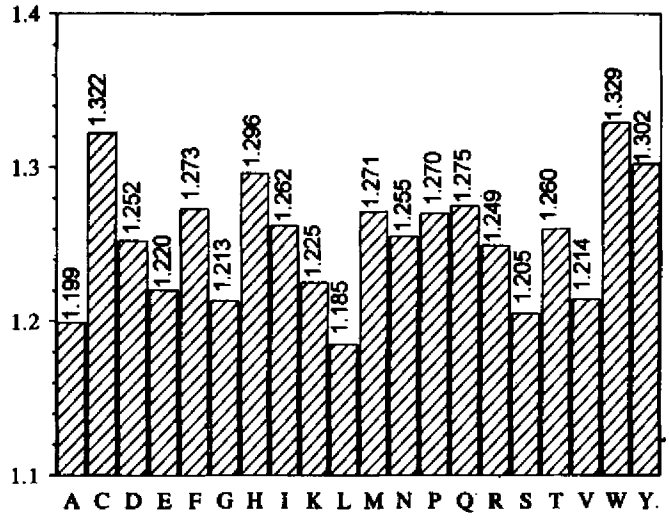

Figure 3. The fractal dimensions, determined by the box-counting algorithm, of the positional distributions of the 20 residues along 200 positions in 5034 proteins of the database. The amino acids are indicated by their one letter codes on the bottom of the x-axis and the fractal dimensions are shown at the top on the x-axis.

Based on the above studies and the results obtained, the distribution of amino acids in protein sequences can be modelled as fractional Brownian motion (fBm) showing self-affinity. The values of β and H indicate self-affinity in the distributions. Since an individual protein sequence consists of several residues (the distribution of each being a fractal), an individual protein sequence can be considered as a multi-fractal. Hence, we have elsewhere proposed a multi-fractal model for protein sequences (Meeta Rani and Mitra 1995).

4. Discussion

One may say that proteins are fractals as they possess self-similarity. Though their self-similarity is statistical, it can be seen at all structural levels. They are all made up of the same 20 amino acids. Self-similarity is least obvious at the primary sequence level. However, at this level we may consider all the conserved sequences, repeat sequences, palindrome sequences, etc., as self-similar. Repeat sequences often exist due to genetic mechanisms such as gene duplications. The secondary structures possess regular geometrical structures such as α-helices and β-sheets. Supersecondary structures with definite structures like the Greek key motif, β-hairpin, four-helix bundle, etc., are also seen in several proteins with different primary sequences. A small number of folding patterns describe in outline most of the known

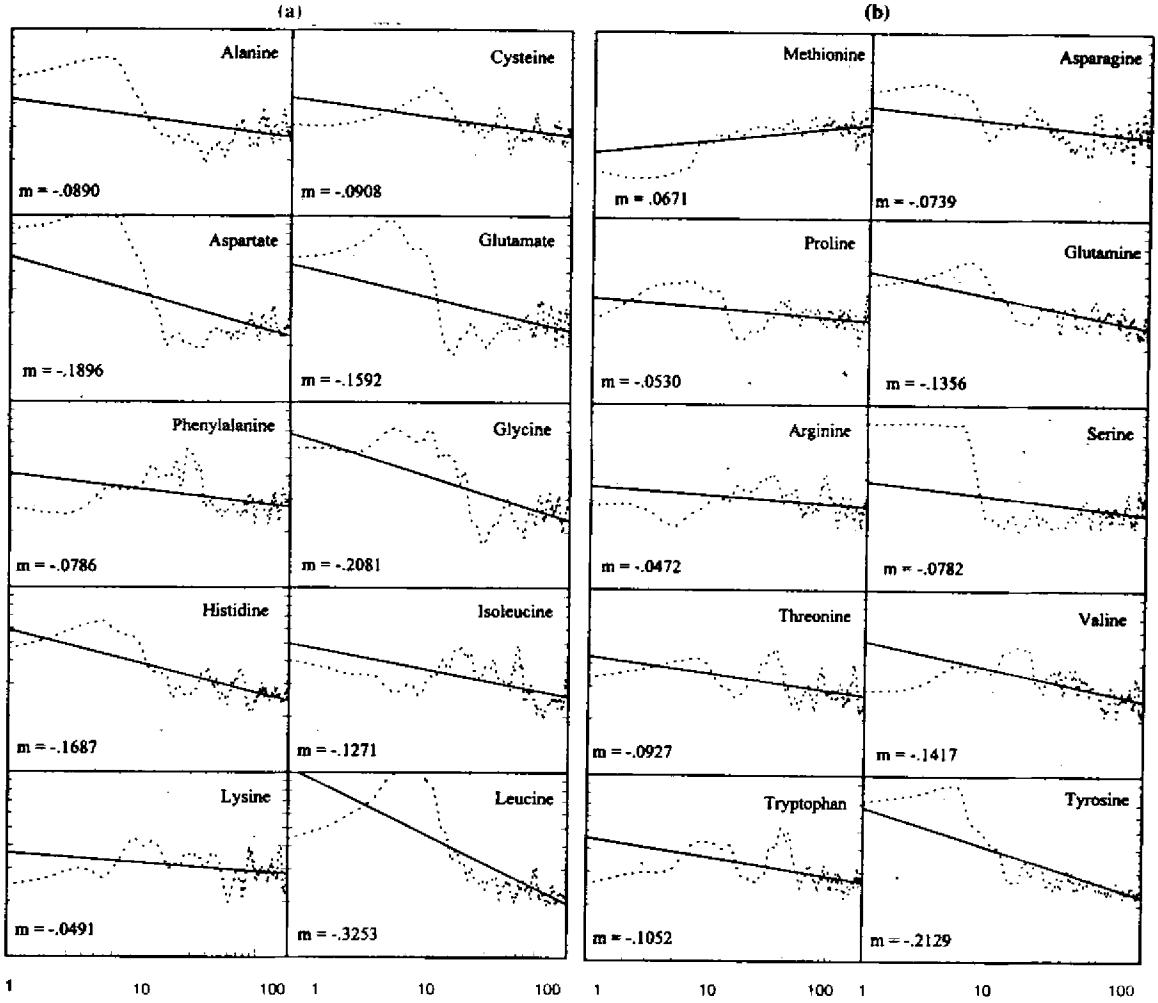

Figure 4. (a) The spectra of the homo-dipeptide pairs of the amino acids obtained as a result of the Fourier transformation of the autocorrelations (order 0 . . . 99). The spectra for the amino acids from alanine to leucine have been shown here. (b) The spectra of the homo-dipeptide pairs of the amino acids obtained as a result of the Fourier transformation of the autocorrelations (order 0 . . . 99). The spectra for the amino acids from methionine to tyrosine have been shown here.

On the x-axis we have the logarithm of angular frequency and on the y-axis we have the logarithm of the spectral densities of the homo-dipeptide pair at various angular distances from 0–180°. We find that at lower values of angular frequencies there is a broad but conspicuous peak. This is the general pattern in the case of all the amino acid homo-dipeptide pairs. This suggests long range correlations as explained in the text.

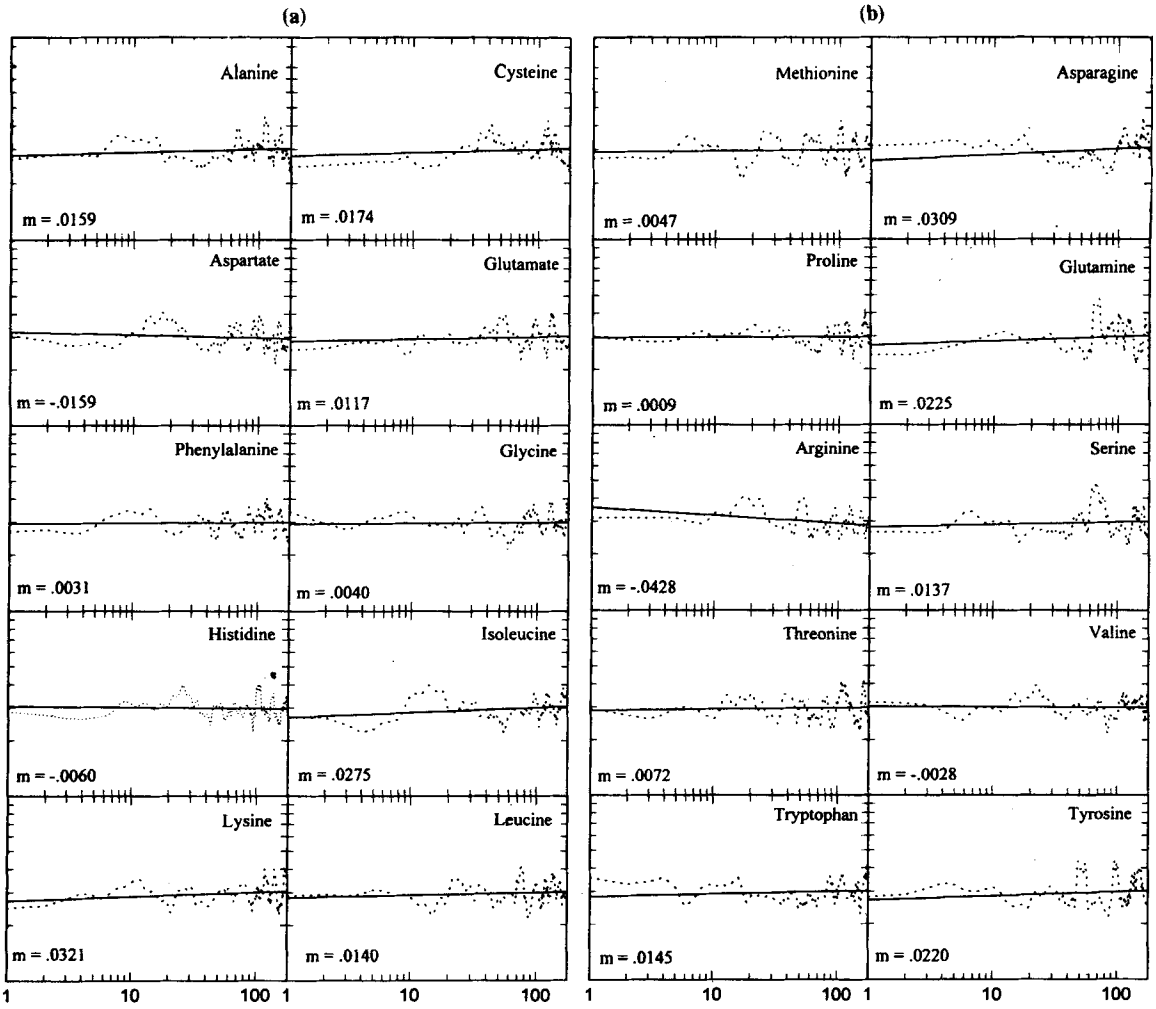

Figure 5. (**a**) The harmonic analysis as described in the text was also performed on Monte Carlo simulated random sequences having similar amino acid composition and lengths. The spectra of these random sequences were calculated exactly as in the case of the real sequence and the spectra for the amino acids from alanine to leucine are presented here. Noteworthy is the remarkable absence of the conspicuous broad peaks which were seen in the real sequences. This clearly indicates the absence of long range correlations in the simulated sequences. Since long range correlations are associated normally with fractal objects it also indicates a fractal nature in protein sequences. (**b**) This graph presents the spectra and concludes similarly for the amino acid residues methionine to tyrosine as taken from the Monte Carlo simulated random sequences mentioned in case of the earlier amino acids (alanine to leucine).

protein structures, as simple combinations of a few geometric arrangements (motifs) have been found to occur frequently in protein tertiary structures. This imparts a statistical self-similarity (self-affinity) to the protein structures and the protein structures resemble multi-fractals. It has been found that though each sequence corresponds to a unique 3-D structure the reverse may not be true. Only 1000 different folds are known to describe all known tertiary structures of proteins (Branden and Tooze 1991).

Our analysis has shown to be multi-fractals possessing self-affinity which agrees well with the above mentioned known information on them.

Acknowledgements

I thank Prof. Chanchal K Mitra for useful discussions and help in preparation of this manuscript and the late Prof. B K Bachhawat, Dr. S Ghosh and other organisers of the National Symposium on Evolution of Life for giving me a chance to present part of this work. I gratefully acknowledge the Council of Scientific and Industrial Research, New Delhi, for a Research Associateship. I also wish to thank COSTED-IBN, Chennai, and CSIR, New Delhi for a partial travel grant to attend an International Workshop on Dynamics of Non-Equilibrium Systems at ICTP, Trieste, Italy.

References

Barnsley M F 1988 *Fractals everywhere* (New York: Academic Press)

Branden C and Tooze J 1991 *Introduction to protein structure* (New York: Garland)

Dayhoff M O 1978 *Atlas of protein sequences* vol. 3 (Washington DC: National Biomedical Research Foundation)

Devaney R L 1988 Fractal patterns arising in chaotic dynamical systems; in *The science of fractal images* (eds) H O Peitgen and D Saupe (New York: Springer-Verlag) pp 137–167

Devaney R L 1992 *A first course in chaotic dynamical systems: Theory and experiment* (Massachussets: Addison-Wesley)

Goldberger A, Rigney and West B J 1990 Chaos and fractals in human physiology; *Sci. Am.* **262** 35–41

Mandelbrot B B 1983 *The fractal geometry of nature* (New York: W H Freeman)

Meeta Rani 1994 *Theoretical studies on protein sequences*, Ph.D. Thesis, University of Hyderabad, Hyderabad

Meeta Rani and Mitra C K 1995 Correlation analysis of the frequency distributions of amino acids in protein sequences; *J. Biosci.* **20** 7–15

Mitra C K and Meeta Rani 1993 Protein sequences as random fractals; *J. Biosci.* **18** 213–220

MS received 28 April 1997; accepted 27 February 1998

Corresponding editors: AMITABH JOSHI and RAGHAVENDRA GADAGKAR