J. Biosci., Vol. 20, Number 1, January 1995, pp 7–16. © Printed in India.

Correlation analysis of frequency distributions of residues in proteins

MEETA RANI and CHANCHAL K MITRA*

School of Life Sciences. University of Hyderabad, Hyderabad 500 134, India

MS received 16 May 1994; revised 24 October 1994

Abstract. Autocorrelation and spectrum analyses of amino acid residues along protein chains in a large data base has been performed. Results reveal the presence of general long range correlations. Similar analyses of simulated (random) peptides do not exhibit any such long range correlations. Based on the results of nur analysis, an attempt has been made to model the distribution of residues in protein sequences on a fractional Brownian motion and individual sequences *as* multi-fractals. For this purpose, the characteristics of an fractional Brownian motion namely, the scaling parameter *H.* the spectral exponent β and the fractal dimension *D,* have been described.

Keywords. Protein sequences; long range correlations; fractals; fractional Brownian motions.

1. Introduction

Proteins are linear biopolymers made up of 20 different amino acids. The linear sequence of amino acids in a protein is called its primary structure. The primary structure folds to form a 3-dimensional, roughly spherical structure, but with a definite pattern that is called its tertiary structure. Just as the 26 letters of the (English) alphabet can make hundreds of thousands of meaningful words, these 20 amino acids can make between 10^{10} to 10^{12} kinds of naturally occurring proteins each with a unique sequence (assuming a typical length of 200 residues). The sequences that are biologically significant (naturally occurring) are not random sequences but are quasi-random; they represent a minute fraction of the total number of theoretically possible sequences. The typical length of a protein sequence ranges from 10 5000 and the average (modal) length is about 250. Since any given random sequence of polypeptide does not necessarily form a biologically meaningful active protein (though it may have a unique 3-D structure), the secret of the biological activity of proteins lies in their specific primary structures, i.e., the specific distribution of amino acid residues in their sequences.

A series of random numbers has no correlations between successive terms. Since we believe that protein sequences are not random, we have tried to study the positional correlation between successive amino acid residues in the primary sequences of proteins. This would help us distinguishing in general between sequences that do not exist in nature (i.e., random polypeptides that are not biologically meaningful) and sequences that are common in nature (i.e., natural polypeptides). In other words, we would be able to characterize and distinguish, at least in principle, the 'allowed' primary structures in natural proteins. A number of short range correlations

*Corresponding author.

have been observed by several workers (Cserzo and Simon 1989); these correlations mostly correspond to the regular secondary structures of the proteins. We have attempted to find existence of long range correlations in protein sequences.

As we are interested only in the general differences between proteins (natural polypeptides) and random polypeptide sequences, our results are expected to reveal only their average statistical properties.

2. Methodology

For our correlation analysis we used the Swis-Prot Protein Sequence Data Bank. Release 10.0, March 1989. It contains over 10,000 sequences and about 3 million residues. Lengths of protein sequences vary from a few residues to several thousands. Since the modal length is close to 200 (Mitra and Meeta Rani 1993) we have considered only those sequences that have at least 200 residues (5034 sequences) and residues beyond 200 have been ignored. As we are looking only for correlations, this is not expected to effect the results significantly, since correlations beyond 200 are disregarded in our calculations.

2.1 *Estimation of spectral density from autocorrelation function*

Consider alanine: if we count the total number of alanines in first position of all these sequences, it gives the positional count of alanine at First position. Similarly we can find the positional count of alanine at all the positions. What we then get is the positional distribution count of alanine in protein sequences (up to length 200).

We can make such positional distribution counts-for all the 20 amino acids. Thus we obtain 20 series of positional distribution counts, one for each amino acid. For each of the amino acids, these series have been used for calculation of position autocorrelation. The techniques used to analyse time-series were applied (Kendall *et al* 1983). In a time-series a variabie changing with time is studied but here a variable (i.e., the frequency of a residue) changing with position in protein sequences (represented on the x-axis) is studied. The methodology for finding correlations, however, is exactly the same.

To compare our results we simulated 5034 random sequences (same as the number being analysed) having the same amino acid composition as in the data base. These random sequences were analysed in the same way as real sequences.

Each of the 20 distributions may be denoted as U_t^r where t denotes the position and varies from 1 to 200 and r is an index denoting the amino acid residue that varies from 1 to 20. Each amino acid is therefore represented by a series of 200 numbers.

The mean of the series may be denoted as $\mathrm{E}(U_t^r) = \mu^r$.

The variance then will $E(U_t^r - \mu^r) = \sigma_r^2 = \mathrm{var}\ U^r$.

The *k*th autocovariance is defined as $E[(u_t^r - \mu^r) * (U_{t+k}^r - \mu^r)] = \gamma_k^r$

and the *k*th autocorrelation is defined as $\rho_k^r = \rho^r{}_{-k} = \gamma_k^r / \sigma_r^2$.

The Fourier transformation of the autocorrelations yields the spectral densities.

Spectral density is defined as $w^r(\alpha) = \sum_{k=-\infty}^{\infty} \rho_k^r \cdot e^{i \cdot k \cdot \alpha} = 1 + 2 \, . \sum_{k=-\infty}^{\infty} \rho_k^r \cdot \cos(\alpha \cdot k)$.

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 [$w(\alpha)$]. The graph of log [$w(\alpha)$] vs log (α) is called the spectrum. The spectrum is again a series of 180 numbers that have been plotted directly after smoothening for visual presentation. The slopes are least square approximations using all these points. The least square regression slopes were obtained directly from the plotting package (Sigmaplot 4·01). There are reasonably prominent peaks in these spectra suggesting long range correlations. However, these peaks were not explicitly eliminated for the computation of the spectral exponent (see below).

2.2 *Studying the spectrum*

The spectrum reveals a good deal about the relationship between various positions occupied by a residue. The characteristics of this spectrum are related to several important parameters of the positional correlations of the residue being studied. The high peaks always refer to strong correlations and less intense peaks refer to weak or no correlation depending upon the relative heights of the peaks in the spectrum. The position on x-axis corresponding to the peak in the graph refers to the periodicity of the residue, i.e., high frequencies corresponds to short periodicities (and *vice versa*). Hence presence of high peaks in the low frequency region indicates long range correlations and their presence in high frequency region indicates short range correlations. In general, short range correlations are reasonably well established as originating from the regular secondary structures of the proteins (e.g, alpha helices and beta pleated sheets). Long range correlations on the other hand give rise to the specific folding patterns that are not very well understood at present.

Since the graph was very noisy, the peaks were not very clear. In order to find the true peaks we smoothened the curve using a spectral window of order nine. The smoothening function is a rectangular window covering nine points. The smoothening process is described below is as prescribed by Daniell (1946). The smoothening process does eliminate some high frequency components of the spectrum but this does not affect our results.

2.3 *Smoothening the spectra*

Smoothening a sample spectra removes the inconsistencies of the estimators. The spectra obtained were smoothened by using a spectral window of order nine (rectangular function) as per the procedure described by Daniell (1946). Essentially, it involves taking an average of the nine values falling within the window and assigning the average to the middle point in the smoothened spectrum.

A program was written in Turbo Pascal for smoothening the spectra. The smoothened graph was analysed and the frequencies corresponding to the highest peaks correlate to the most common frequency of distribution of the residue. Broad peaks refer to a range of occurring frequencies. Low frequencies indicate long range correlations in the distribution of the residue, whereas high frequencies indicate short range correlations and medium magnitudes of the frequencies indicate intermediate range correlations. The negative slope of this graph gives the spectral exponent, β.

2.4 *Spectral exponent* β *and scaling parameter H*

In a typical case, it is not expected that only one periodicity (or frequency) is present. In such a case, multiple peaks are generally expected. In this particular situation, a very large number of peaks are observed with different intensities. As a comparison, white noise contains all frequencies in equal amount. Hence the spectrum of a white noise is expected to be a horizontal straight line. Other kinds of noise patterns are possible in which all frequencies are present in unequal amounts. The distribution of the intensities (for various frequencies) is measured by the spectral exponent. The spectral exponent β is the negative slope of the graph of the spectrum plotted as log (spectral density) *vs* log (frequency). In other words, β measures the decrease of the intensity with frequency on a double logarithmic plot. The spectral density varies inversely as the spectral exponent β. If β is negative, the intensity increases with frequency. In other words, long range correlations are predominant, β is related to the scaling parameter *H.* The scaling parameter reflects how the distribution function changes when the independent variable is scaled (i.e., by multiplying with a given constant *r*). The scaling parameter *H* characterizes the scaling behaviour of fractal traces, random processes, noises, etc. Defined explicitly, if we consider a distribution function $y = f(x)$, then for a scale *r* for *x* (*i.e.,* $x{:}= r\dot{}x$)*y* scales as r^H *y* (*i.e.,* $y{:} = r^{H}\dot{}y$). For a white noise, *H* is 0·5. Qualitatively, it gives an idea about the correlations in space (or time, as the case may be). In a random walk, the mean distance travelled scales as the square root of the number of steps taken. *H* lies in general between 0 and 1. *H* and β are related by the simple relation *H* = (1—ß)/2 for a distribution with a Euclidean dimension of unity. So, *H* = 0 and *H* = 1 correspond to the cases β = 1 and β = 1. Thus both β and *H* provide important information about the positional correlations of amino acid residues in protein sequences. One can find nature of distributions, whether they are linear, random or fractal (see discussions) (Vass 1988).

3. Results

The graphs of the 20 spectra (corresponding to the 20 amino acids) for the real sequences are presented in figure 1. The graphs of the spectra for the amino acids in the case of random sequences are presented in figure 2.

3.1 *Real sequences*

The spectral exponent β was found to lie between 0˙33 and 0˙05 for all residues (except methionine, which had a β value of 0·07). β has a significant non-zero

Figure 1. Logarithm of the spectral densities of the positional frequency distributions of the twenty amino acids. The frequency data have been collected from the data base of 5034 sequences. The angular frequency is plotted on the x-axis on a logarithmic scale. The slope of the graph is designated 'm' and is indicated on the left hand side (bottom).

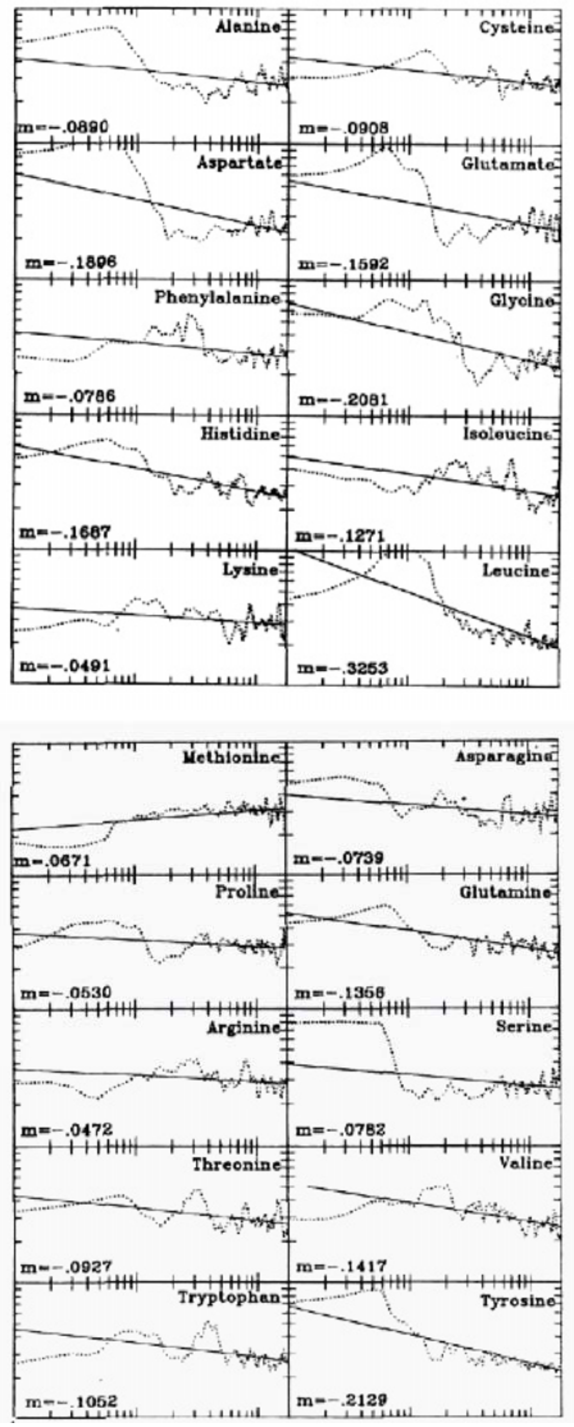
Alanine
m=-.0890
Cysteine
m=-.0908
Aspartate
m=-.1896
Glutamate
m=-.1592
Phenylalanine
m=-.0786
Glycine
m=-.2081
Histidine
m=-.1687
Isoleucine
m=-.1271
Lysine
m=-.0491
Leucine
m=-.3253
Methionine
m=.0671
Asparagine
m=-.0739
Proline
m=-.0530
Glutamine
m=-.1356
Arginine
m=-.0472
Serine
m=-.0782
Threonine
m=-.0927
Valine
m=-.1417
Tryptophan
m=-.1052
Tyrosine
m=-.2129

Figure 1.

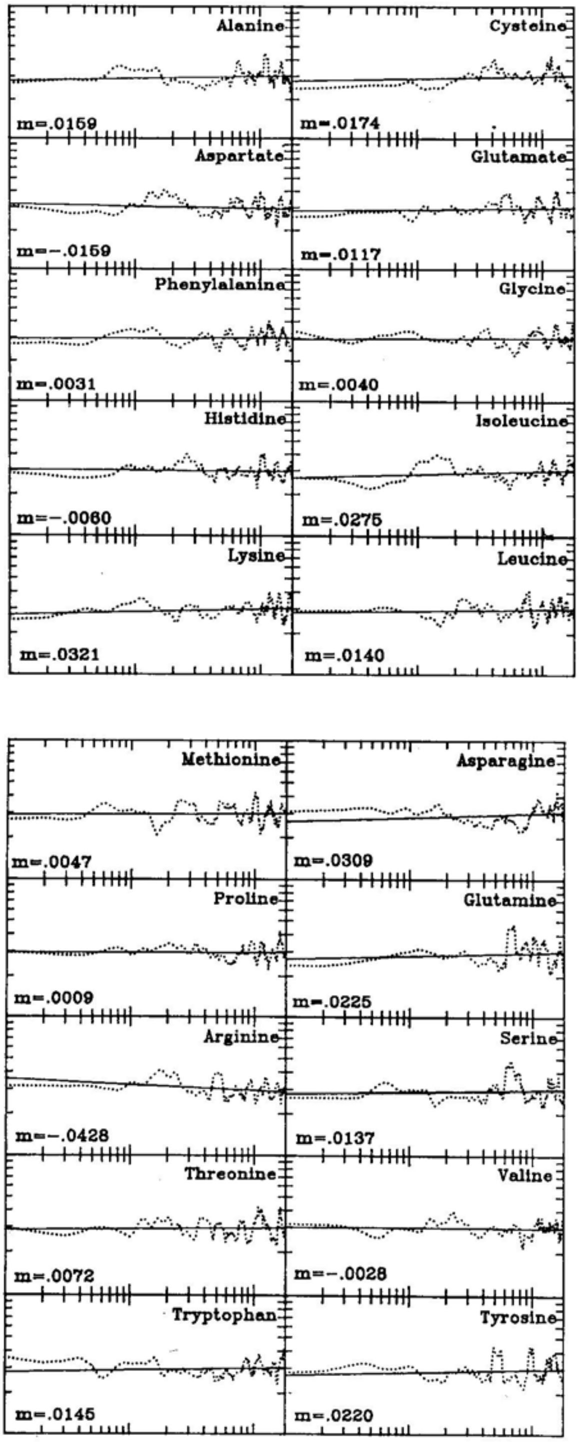
Alanine
m=.0159
Cysteine
m=.0174
Aspartate
m=-.0159
Glutamate
m=.0117
Phenylalanine
m=.0031
Glycine
m=.0040
Histidine
m=-.0060
Isoleucine
m=.0275
Lysine
m=.0321
Leucine
m=.0140
Methionine
m=.0047
Asparagine
m=.0309
Proline
m=.0009
Glutamine
m=.0225
Arginine
m=-.0428
Serine
m=.0137
Threonine
m=.0072
Valine
m=-.0028
Tryptophan
m=.0145
Tyrosine
m=.0220

Figure 2.

value compared to the random sequences (see random sequences). The average value of β turns out to be 0·13 (excluding methionine) and 0·03 (including methionine)· Presence of high peaks in the low frequency regions indicates long range autocorrelations· This is also apparent from figure 1.

The scaling parameter *H* in case of all the residues for real sequences was found to be in the range 0·48 and 0·34, with an average of 0·45 (including methionine). For a completely random sequence *H* is expected to be 0·5.

3·2 *Random sequences*

Small values of β for the random sequences suggest that both low and high frequencies are equally probable (as s een in case of all residues except methionine). High values of β would indicate that high frequencies are far less common compared to low frequencies.

In case of simulated random sequences β values are in the range -0·043 to 0·03. Since the values are for random sequences, the expected values are zero and the average slope turns out to be 0·007 (very close to 0 as expected). Also these spectra showed a lack of any long range correlation in any residue distribution as indicated by lack of high peaks in the low frequency regions. This is clear from figure 2·

The scaling parameter in case of random sequences was in the range 0·48 to 0·52 (with an average of 0·5), as expected in case of random noises. For a completely random sequence *H* is expected to be 0·5.

The fractional Brownian motion (fBm) characteristics, i.e., β, *H* and *D* for the distributions of all residues are presented in table 1· The relation between the spectral exponent β and the fractal dimension *D* has been plotted in figure 3. The least square regression line and the corresponding 99% confidence limits are also shown· We note that the spectral exponent is linearly related to the fractal dimension, as expected for a fBm· A similar relation is seen with the scaling parameter, *H* and the fractal dimension *D,* but since the spectral exponent and the scaling parameter are linearly related, this graph has not been presented.

4. Discussions and conclusions

4·1 *Fractal geometry: the geometry of nature?*

Fractals are· geometrical objects showing self-similarity (Mandelbrot 1982) at all magnifications (or scales), i.e., they appear similar at all magnifications. The self-similarity may be geometrically exact or statistically apparent. But self-affine fractals differ from both of these· Self-affine fractals show statistical self-similarity only when the *x* and *y* axes are magnified by different scales. The importance of fractal geometry is due to the fact that it has become a very convenient and

Figure 2. Logarithm of the spectral densities of the positional frequency distributions of the twenty amino acids obtained from the 5034 simulated random sequences· The log spectral density is plotted on the *y*-axis and on the x-axis are plotted the angular frequencies also in a logarithmic scale· The slope of the graph, designated 'm' and is indicated on the bottom left hand side.

Table 1. fBm characteristics of positional distributions of residues in protein sequences·

No.	Residue	β	*H*	*D*	*No.*	Residue	β	*H*	*D*
1.	Ala (A)	0·0890	0·4555	1·199	2.	Cys (C)	0·0908	0·4546	1·322
3.	Asp (D)	0·1896	0·4052	1·252	4.	Glu (E)	0·1592	0·4204	1·220
5.	Phe (F)	0·0786	0·4607	1·273	6.	Gly (G)	0·2081	0·3960	1·213
7.	His (H)	0·1687	0·4157	1·296	8.	Ile (I)	0·1271	0·4365	1·262
9.	Lys (K)	0·0491	0·4755	1·225	10.	Leu (L)	0·3253	0·3374	1·185
11.	Met (M)	−0·0671	0·5336	1·271	12.	Asn (N)	0·0739	0·4631	1·255
13.	Pro (P)	0·0530	0·4735	1·270	14.	Gln (Q)	0·1356	0·4322	1·275
15.	Arg (R)	0·0472	0·4764	1·249	16.	Ser (S)	0·0782	0·4609	1·205
17.	Thr (T)	0·0927	0·4537	1·260	18.	Val (V)	0·1417	0·4292	1·214
19.	Trp (W)	0·1052	0·4474	1·329	20.	Tyr (Y)	0·2129	0·3936	1·302

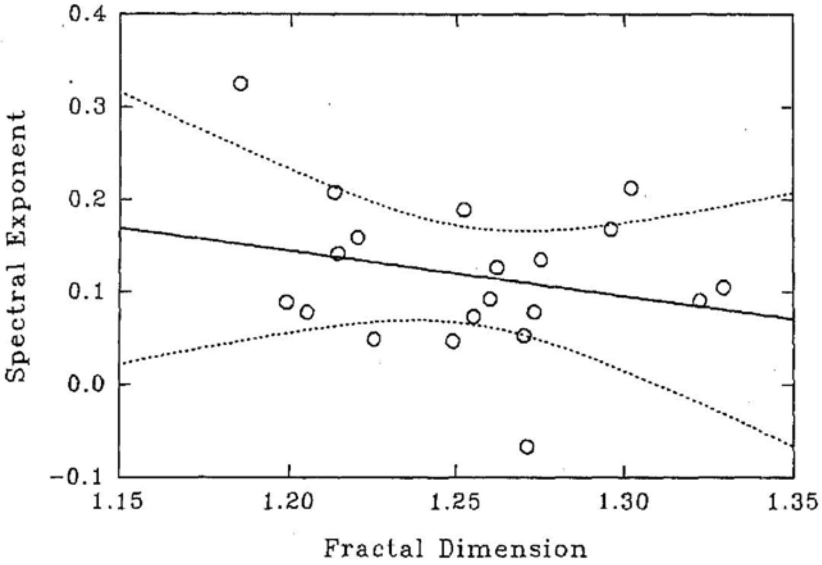

Figure 3. A plot of the spectral exponent β against the fractal dimension *D* of the positional distributions of the 20 amino acids The straight line shows the least square regression line and the two dotted lines represent the 99% confidence interval based on *D*. We do not expect a perfect correlation between the two considering the limited sample size used in our analysis. The distribution of methionine is abnormal (as several sequences in the data base used have been determined from the cDNA sequences) and this results in the negative spectral exponent for this residue.

popular method for analysing objects, processes and phenomena occurring in nature. This is because fractal geometry is capable of providing the language and formalism for studying physical processes like diffusion limited aggregation, condensation of matter on microscopic scale, etc. (Orbach 1986); for describing biological phenomena like patterns existing in long DNA sequences (Peng *et al* 1992; Voss 1992; Nandy 1994), branching patterns in the network of neurons, arteries, bronchioles, trees, symmetries and patterns of plants and flowers etc. (Oppenheimer 1986)· The origin of fractal patterns in nature is due to chaotic dynamics of non-linear deterministic systems (Devaney 1988).

4.2 *Fractal model for amino acid distribution and protein sequences*

In the usual Brownjan motion or random walk, the sum of the independent increments or steps leads to a variation that scales as the square root of the total number of steps. Thus $H = 1/2$ corresponds to a normal Brownian motion. A Brownian motion where $H \neq 1/2$ is called fBm (Voss 1988). An fBm trace is characterized by scaling parameter *H,* fractal dimension *D* and spectral exponent . An fBm trace repeats statistically only when the *x* and *y* co-ordinates are magnified by different amounts. If *x* is magnified by a factor *r* (*x* becomes *rx*), then *y* must be magnified by a factor r^{H} , (*i.e., y* becomes r^{H} *y*). This non-uniform scaling where shapes are statistically invariant under transformations that scale different co-ordinates by different amounts is called self- *affinity* as already mentioned.

Such self-affine processes are not random but have some correlations. A statistically self-affine fractional Brownian function V_H provides a good model for many natural scaling processes and shapes. As a function of one variable ($E = 1$), it is a good model for noises, random processes, music (Voss and Clarke 1975) etc. We have modelled the residue distribution in protein sequences as an fBm and its spectral exponent β, the scaling parameter *H* and fractal dimension *D,* have been calculated. We also note the expected linear relation between the spectral exponent β and the fractal dimension *D.* The observed slope of this line (–0·5) is in agreement with the theoretical considerations (Voss 1988).

Based on our studies and results we find that the distribution of amino acids in protein sequences can be modelled as a fBm. Hence an fBm model has been proposed. The characteristics of an fBm relate to the fBm integrated over an interval of time. Differentiating such an fBm gives what is called fractional Gaussian noise. Similarly the characteristics of the distribution of a residue relate to the average Statistical properties of the residue in the protein sequences. Differentiating such an fBm would give us the distribution of a residue in an individual sequence. Hence distribution of a residue in an individual sequence is a fractional Gaussian noise and its distribution in the complete set of natural proteins is a fBm. Since an individual sequence consists of several residues (the distribution of each being a fractal) an individual protein sequence can be considered a multi-fractal. Shapes and measures requiring more than one fractal dimension are known as multi-fractals. If a sequence contains all the 20 residues then it requires 20 fractal dimensions (one for each residue) to describe it. Proteins like collagen having fewer kinds of residues would be multi-fractals requiring lesser number of fractal dimensions. Different sequences of a given family, say myoglobin for example, have similar primary sequences, hence based on our model they are multi-fractals of one kind. After characterising this family of multi-fractals one can in principle predict unknown sequences belonging to the family yet undiscovered, but existing in nature, by computer simulations (utilising the parameters already characterized for the protein family).

Also one can study the evolution of protein families by observing how the multi-fractal (i.e., primary sequence of a protein) slowly changes with change in the parametres of the multi-fractal. It may also be extended to the tertiary structure of proteins. Work is in progress to sort out the details of such an endeavour.

Acknowledgement

MR wishes to thank the University Grants Commission, New Delhi for a senior research fellowship.

References

Cserzo M and Simon L 1989 Regularities in primary structures of proteins; *Int J. Peptide Protein Res.* **34** 184-195

Daniell P J 1946 Discussion on “Symposium on auto-correlation in time series"; *Suppl. J. R. Statist. Soc.* **8** 88

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 137-168

Kendall M, Stuart A and Ord J K 1983 *The advanced theory of statistics* (High Wycombe: Charles Griffin) pp 422-694

Mandelbrot B 1982 *The fractal geometry of nature* (New York: W H Freeman)

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

Nandy A 1994 Recent investigations into global characteristics of long DNA sequences; *Indian J. Biochem· Biophys·* **34** 149-155

Oppenheimer, P E 1986 Real time design and animation of fractal plants and trees; *Comput· Graphics* **20 4**

Orbach R 1986 Dynamics of fractal networks; *Science* **231** 814-819

Peng C K, Buldyrev S V, Goldberger A L, Havlin S, Sciortino F, Simons M and Stanley H E 1992 Long range correlations in nucleotide sequences; *Nature* (*London*) **356** 168-170

Voss R F and Clarke J 1975 1/F noise in music and speech; *Nature* (*London*) **258** 317-318

Voss R F 1988 Fractals in nature; in *The science of fractal images* (eds) H O Peitgen and D Saupe (New York: Springer-Verlag) pp 21-70

Voss R F 1992 Evolution of long-range fractal correlations and 1/f noise in DNA base sequences; *Phys Rev Lett* **68** 3805-3808