J. Biosci., Vol. 18, Number 2, June 1993, pp 213-220. © Printed in India.

Protein sequences as random fractals

CHANCHAL K MITRA* and MEETA RANI
School of Life Sciences, University of Hyderabad, Hyderabad 500 134, India

MS received 20 February 1992; revised 28 September 1993

Abstract. The analysis of primary sequences from a protein sequence data base suggests that the sequences can be considered as examples of constrained random fractals. Fractal dimensions of the positional distributions of the 20 residues along the chain have been calculated. These fractal dimensions can be used as indices of intrinsic preferences of various residues.

Keywords. Fractal dimension; protein sequences; proteins and fractals.

1. Introduction

Whether naturally occurring protein sequences are random polypeptides, or there are some general rules, yet undiscovered, governing their sequences, has been an intriguing question since long. Several workers have tried to find an answer to this question; some workers have discovered elements of randomness while others could see deterministic elements as well.

The primary structures of proteins appear far from random if we consider the sequential occurrence of residues *e.g.* the occurrences of the 400 dipeptide residues are often anamolous. The distribution of X (any given residue) in a subsequence -A-X-(A = alamine for example) is not in accordance with its average representation in the organism. In other words, the positional frequency of a residue in a given sequence is influenced to a significant extent by the previous residue. This may be interpreted as a pair preference (which includes both attractions and repulsions). Kolaskar and Ramabrahmam (1983) have shown that the distribution of pairs of amino acids cannot be regarded as linear combinations of frequencies of occurrences of the constituent individual amino acids. Statistically, various amino acid pairs (20 × 20 pairs or possibilities) are not present in a data base as predicted by simple probability (Meeta Rani 1990). The preferential occurrence of amino acid pairs is not perfect, which is what gives rise to the diversity of protein sequences. 'Neighbourhood' preferences, albeit imperfect, may be considered as introducing a deterministic element in the sequence (Vonderviszt *et al* 1986). Dose (1976) pointed out that the composition and primary structure of thermal protenoids differ from the composition of the reactant amino acid mixture, again showing a non-random distribution of residues.

A far greater deterministic element is introduced in native sequences by the forces of evolution and by the requirement of biological functionality. A sequence which is not meaningful in a biological system will either be eliminated or will never be found in nature. We also note" that such deterministic forces necessarily must be "imperfect" because they allow for diversities and mutations. The next question that naturally arises is of the extent to which these two forces — the random and deterministic—influence the protein sequences. Ptitsyn (1984) suggested the

*Corresponding author.

presence of random elements by remarking that natural proteins are mainly random sequences which have been edited during the course of evolution to impart biological functionality. If this is true then the more evolved proteins should have accumulated more of such 'editions' than the lesser evolved ones and this may be why they are indeed far from random. We may therefore conclude that all naturally occurring proteins have random as well as deterministic elements. The more evolved proteins would have resulted from deterministic forces to a greater extent than from random and their evolutionary status can be correlated to this.

If we are to have a model for protein sequences in order to study them qualitatively, the "random fractal" model appears most appropriate. The most common fractals, the geometrical fractals, have no random elements present and are generated by repeated application of some simple geometrical rules. Natural fractals, on the other hand, always have some random elements present and are therefore different from classical geometrical fractals (Saupe 1988). We have therefore considered protein sequences to be examples of random fractals, or more precisely constrained random fractals, to stress the presence of deterministic forces *(i. e.* pair preferences, requirement of biological functionality, etc.). The random elements are the cause of great variety within themselves (*i.e.* protein families) and the deterministic elements impart to them the characteristic features by which they are recognizably the same fractals and can be characterized by their fractal dimensions. By modelling protein sequences as random fractals it is possible to study them by using conventional fractal techniques. We have used fractal dimension as a parameter for the study of the 20 different positional distributions, each corresponding to the twenty different amino acid residues naturally occurring in proteins. Although several kinds of dimensions of fractals have been reported in the literature, they do not offer significant advantage over the 'conventional' fractal dimension for demonstrating fractal properties. A computationally simple algorithm, the box counting algorithm (Barnsley 1988) has been adopted to find the fractal dimension of the positional distribution of residues in protein sequences. As expected, the dimensions calculated in this way are fractional and depend on the residue for which the distribution was considered. Fractal dimensions calculated thus are often referred to as the box dimension.

2. Methodology

The Swiss-Prot Protein Sequence Data Bank (Release 10·0, March 1989) was taken for our analysis. The data base had 10,008 polypeptide sequences containing about 3 million residues.

At this point it may be useful to consider the nature of the data base used. The data base contains all protein sequences that were available in the literature at that time. Naturally, it contains a number of highly homologous proteins. The question here arises whether this is going to introduce any perceptible bias in our analysis. Extracting a suitably curtailed dataset from the entire bank is to a significant extent a subjective exercise. A basic question at this point is whether the data bank used can be considered as a random sample (of all the known and unknown sequences) without replacement. This task has been attempted by Doolittle (1981). When a particular protein is sequenced, it is selected based on several considerations: (i) general interest in the protein, (ii) biological significance, (iii) ease of purification

and a host of other factors. General interest in a protein gives rise to certain homologies that are present in the data bases. However their total number is relatively small and we do not expect that this will bias the data base significantly. It is rather obvious that homologies that are present as a result of evolution cannot be considered as a source of bias in the data base. The major causes of bias in the data base we have used comes from "fragments" that should have been excluded. We feel, however, that screening of the data base may introduce more bias than it is supposed to remove. Hence we have used the data base without any modifications.

The median value of the sequence length is close to 200 and hence we have considered all sequences that had at least 200 residues (5034 sequences). All residues beyond the 200th position were ignored. The positional distribution of the 20 different residues (indicated by their standard one letter codes) were computed as follows:

If A^i_{jk} indicates the presence (1) or absence (0) of a residue of type i at the jth position for the Kth sequence, the positional distribution is given as

$$p^i_j := \Sigma_k A^i_{jk}, \tag{1}$$

where $A^i_{jk} = 1$ if residue i is present in the fcth sequence at the j th position and $A^i_{jk} = 0$ otherwise. The resulting p^i_j array is stored in a 20 × 200 matrix and is called the positional distribution matrix.

This p^i_j matrix was "normalized" by dividing each element by the total number of sequences. The resulting array was stored in the same p^i_j array and was denoted as the normalized positional distribution matrix.

The total number of a residue of type i was obtained as

$$p^i := \Sigma_{jk} A^i_{jk}, \tag{2}$$

where, j=1 ·· 200 positions, k= 1·· 5034 sequences and the probability of finding a residue of the type i was obtained as:

$$p(i) := p^i/(200 * 5034). \tag{3}$$

These probabilities are shown in figure 1. This also gives us the abundances of various amino acid residues. These values agree with the values reported in the literature (Doolittle 1981; Vonderviszt 1986).

Each element in row i of the normalized positional distribution matrix was divided by $p(i)$ to obtain the relative normalized positional distribution and was stored in the same matrix p^i_j where

$$p^i_j := p^i_j / p(i). \tag{4}$$

The relative normalized positional distribution matrix was used for computation of fractal dimensions for various residues. The p^i_j values all lie in the neighbourhood of unity, a value > 1 suggesting that a residue is preferred (over the average) at the particular position j and a value <1 suggesting that the particular residue is not favoured at that particular position. This, however, is to be understood in a relative sense; if a particular residue is strongly favoured at any position, the preference for other residues automatically is reduced at the same position.

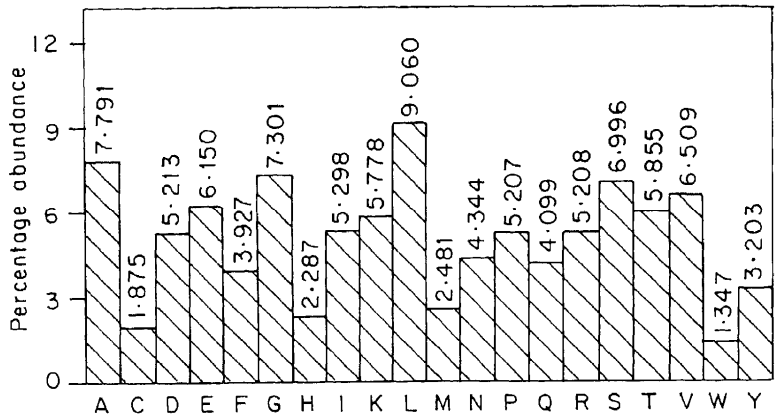

Figure 1. The probability of occurrence in percentages of the twenty amino acids in the sequences of the data base (10,008 sequences and 2,932,613 residues). The amino acids are indicated by their one letter codes. The values obtained agree with the literature values commonly available.

For comparison, 5034 chains were simulated by a Monte Carlo technique using the probabilities (abundances) of various amino acids as given in figure 1. These sequences were also analysed using the same procedure.

2.1 *Box counting algorithm*

We have followed the box counting algorithm to determine the fractal dimensions of the relative normalized positional distribution curves for the 20 amino acid residues. Although details of the principles of this technique is available in the literature, a brief description will not be out of place here.

Consider a set of points *(i. e.* the relative normalized frequencies at different positions, a total of 200 points) distributed in a plane. Divide the plane into a number of square grids (taken for computational convenience as 4^1 4^2,4^3,4^4 and 4^5) and count the number of boxes that include at least one point. The limiting slope of the straight line relating

In (*number of boxes containing a point*) *to n In* (*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, subdivisions beyond 32 × 32 = 1024 boxes are not physically meaningful and hence were not carried out. In addition, instead of actually calculating the limiting slope, we measured the least squares slope for computational reasons. Mathematically,

$$D = \lim_{n \Rightarrow \infty} \frac{\ln \text{ (box count)}}{n \cdot \ln 2}$$

where n is the order of subdivision. Although we have used a least squares technique for calculating the limiting slopes, the correlation coefficients were very high (> 99%) indicating that within a physically meaningful scale, our values are meaningful. As expected, the dimensions calculated this way are fractional in nature and are often called box dimensions. The box dimensions calculated for the random sequences are all equal to one. This is expected since in absence of preferences, the distribution graphs are straight lines parallel to the position axis and a straight line has a dimension of unity. The fractal dimensions obtained as above are shown in figure 2, for the twenty different amino acid residues.

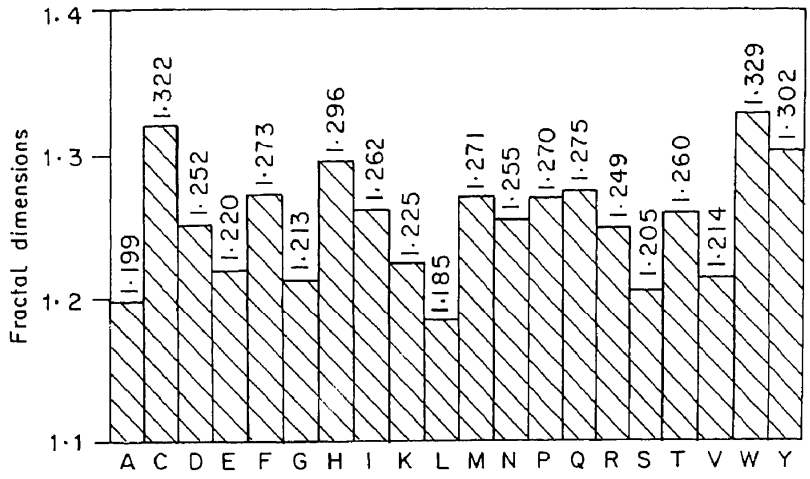

Figure 2. The fractal dimensions, determined by box counting algorithm, of the positional distributions of the twenty residues along 200 positions in 5034 proteins of the data base. The amino acids are indicated by their one letter codes.

The relative normalized positional distributions were studied by simple statistical tests for randomness. A simple test, the number of turning points (Kendall 1976), reveals that the distributions are random. Although this test cannot detect several kinds of non-randomness, it is useful to demonstrate the fractal nature of the distributions. Further proof of the fractal nature of the distribution is given by studies on the scaling behaviour.

All the computations were performed on an IBM compatible PC-AT computer and the required programs were written in Turbo-Pascal.

2.2 *Fractal nature of distributions*

If we plot log $(1\text{-}P_{xx}/P_x)$ against Dx, the fractal dimension for residue x, where P_{xx} is the probability of finding a homodipeptide (xx) and P_x is the probability of finding a single residue x, we obtain a straight line (figure 3). The straight line has a slope of 0·250 and an intercept of –0·341. The correlation coefficient is almost

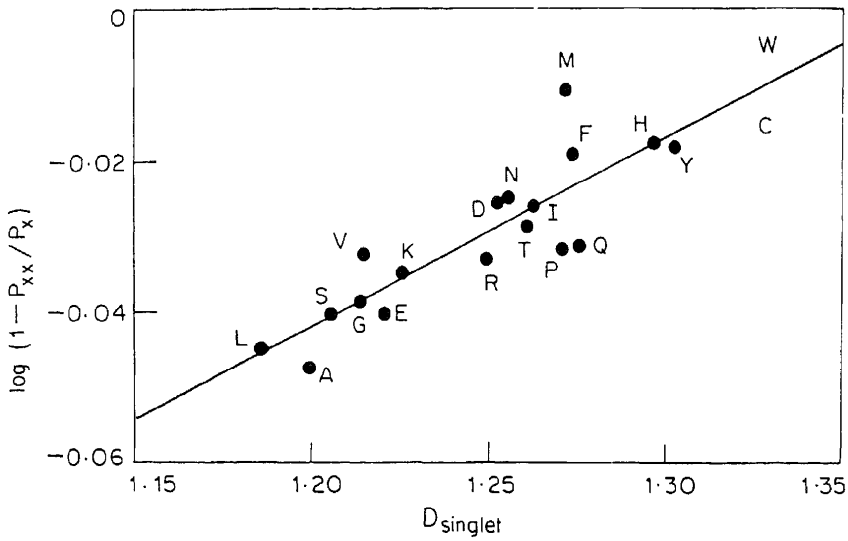

Figure 3. The plot of log(l — P_{xx}/P_x), where P_x is the probability of finding a singlet (x) and P_{xx} is the probability of finding a doublet (xx) where x is any residue and xx is its corresponding homodipeptide as function of D_x, where D_x is the fractal dimension of x as determined by box counting algorithm. The points show a correlation of almost 89·9%.

90%. This suggests that we can write an empirical relation,

$$\log (1\text{-}P_{xx}/P_x) = 0{\cdot}250 * D - 0{\cdot}341$$

$$\text{or} \quad 1\text{-}p_{xx}/p_x = 10^{(\cdot 25 * D - 0{\cdot}341)} = 0{\cdot}4560\ 1{\cdot}778^D$$

$$\text{or} \quad P_{xx}\ /p_x = 1 - 0{\cdot}4560\ (1{\cdot}778)^D .$$

This gives us a simple relationship between the dipeptide and single residue frequencies, based on their fractal dimensions. This also shows the qualitative dependence of the fractal dimension on the abundance of the various residues.

3. Results and discussion

Our results show that the distribution of a given residue along a polypeptide chain is fractal in nature. The fractal dimension D of the distribution is different for different residues and lies in the range of 1·18 to 1·35, a range common in various naturally occurring fractal objects (Voss 1988). We also observe a qualitative negative correlation between the probabilities of occurrence of various residues and their fractal dimensions. For example, leucine (L) has the highest abundance (9·061%) and the lowest fractal dimension (1·185) while tryptophan (W) has the lowest abundance (1·347%) and the highest fractal dimension (1·342).

The origin of this correlation is not clear at this moment. The abundances of various residues are determined by the genetic code and it is necessary to study the fractal nature of the nucleic acid. While this paper is under revision there has been a

report of the fractal nature of DNA sequences (Voss 1992). But the negative correlation is certainly not due to statistical effects, as studies using pseudo random sequences show a fractal dimension of one when calculated using the same program, for all the residues. Thus the fractal dimension is most likely determined by intrinsic pair preferences.

The distribution graphs do not show any obvious end effects, except, perhaps for methionine, which is abnormal. This is because many of the sequences in the data base have been determined from cDNA sequences and the start codon and the methionine codon happen to be the same. However, this materially does not affect the fractal dimension determined for methionine.

In the determination of the fractal dimensions of various residues, the graphs were "normalized" so that the relative abundances do not effect the general pattern of distributions. We have also shown, using pseudo-random sequences, that the abundances do not play any direct role in the fractal dimensions determined. Therefore we propose that the fractal dimensions so determined are indicative of the "intrinsic preferences" of various residues. This is apparent from the following justification:

A residue having a low box dimension (*e.g.* leucine) tends to be distributed more evenly (randomly) and has a "relatively" low preference for itself (1·093 times that expected for a completely random distribution). Tryptophan, in contrast, has a high box dimension and therefore does not distribute itself evenly but has high preferences and repulsions. For example it has a relatively high preference for itself (1·377 times than that expected in case of a random distribution).

We therefore conclude that the fractal dimensions as determined above are indicative of the intrinsic preferences in a general sense and can be used to quantify such preferences for a residue in a native sequence.

The process of renaturation, *i.e.*, the formation of the correct 3-D structure from the random one, suggests that 3-D structural information of the protein is already present in the primary sequence (Anfinsen 1973). Primary sequences of proteins have been studied in great detail with the major objective of predicting the 3-D folding pattern of the final protein by Blout *et al* (1960), Davies (1964), Dirkx (1972), Finkelstein and Ptitsyn (1971), Chou and Fasman (1974), etc. The Chou-Fasman algorithm, or its variants, have been the most successful in this regard. But still no clear insight is available on the protein folding mechanism. Based on the ideas presented here it appears that fractal structure of the primary sequence causes a fractal structure of the 3-D structure of proteins also. Analysing proteins at all levels of structures by fractal techniques might throw more light on the mysterious relation between the primary sequence and the 3-D structure.

A final point we would like to stress is that by using the techniques of fractal interpolation, it is statistically possible to predict sequences that are biologically meaningful. The details of this have not been worked out, but since the doublet and singlet probabilities are seen to be correlated, it is intuitively obvious that such a possibility exists.

Acknowledgement

M R wishes to thank the University Grants Commission, New Delhi, for the award of a Junior Research Fellowship.

References

Anfinsen C B 1973 Principles that govern the folding of protein chains; *Science* **181** 223–230

Barnsley M F 1988 *Fractals everywhere* (New York: Academic Press) pp 176–177

Blout E R de Loze C, Bloom S M and Fasman G D 1960 The dependence of the conformations of synthetic polypeptides on amino acid composition; *J. Am. Chem. Soc.* **82** 3787–3789

Chou P Y and Fasman G D 1974 Prediction of protein conformation; *Biochemistry* **13** 222–244

Davies D R 1964 A correlation between amino acid composition and protein structure; *J. MoL Biol.* **9** 605–609

Dirkx J 1972 Une methode semi-empirique de prediction des regions L-helicoidales des chaines polypeptidiques d'apres leur structure primaire; *Arch. Int. Physiol. Biochim.* **80** 185–187

Doolittle R F 1981 Similar amino acid sequences: chance or common ancestry?; *Science* **214** 149–159

Dose K 1976 Ordering process; in *Protein structure and evolution* (eds) J L Fox, Zdenek Deyl and Anton Blazej (New York: Marcel Dekker) pp 165–166

Finkelstein A V and Ptitsyn O B 1971 Statistical analysis of the correlation among amino acid residues in helical, *ß*-structural and non-regular regions of globular proteins; *J. Mol Biol.* **62** 613–624

Kendall M 1976 *Time-series* 2nd edition (London: Charles Griffin) pp 21-28

Kolaskar A S and Ramabrahmam V 1983 Conformational properties of pairs of amino acids; *Int. J. Peptide Protein Res.* **22** 83–91

Meeta Rani 1990 *Fractal dimensions of protein sequences,* M..Phil, dissertation, University of Hyderabad, Hyderabad

Ptitsyn O B 1984 Protein as an edited copolymer; *Mol Biol. USSR* **18** 574–590

Saupe D 1988 Random fractal algorithms; in *The science of fractal images* (eds) H O Peitgen and D Saupe (New York: Springer-Verlag) pp 71–113

Vonderviszt F, Matrai Gy and Simon I 1986 Characteristic residue environment of amino acids in proteins; *Int. J. Peptide Protein Res.* **27** 483–492

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