J. Biosci., Vol. 19, Number 2, June 1994, pp 255–266. © Printed in India.

Periodicities in protein sequences

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

MS received 10 August 1993; revised 25 February 1994

Abstract. The correlation between various amino acid residues (either same or different), along the polypeptide chain have been studied using a large data base. A table of preference values for pairs having strong correlations has been constructed, which can be used to study any sequence and by calculating the weight of these sequences based on these preference values, a rough distinction between a "natural" and a "random" sequence can be made, One can further comment on the evolutionary status of proteins based on these weights.

Keywords. Polypeptide sequences; correlations; pair preferences; periodicities.

1. Introduction

The linear sequence of amino acid residues in a polypeptide is called its primary structure, The primary structure is responsible for the final three dimensional folded structure of the protein (Anfinsen 1973) which is the biologically active form. There have been several attempts to relate primary structure to protein folding. Simon (1985) has shown that globular proteins possess sub-sequences called X-sequences, which are the only structures in which short overlapping segments of the polypeptide chain are in one of their significantly stable conformation and these subsequences play an important role in the initiation of the folding process. Vonderviszt *et al* (1986) have shown that generally valid short range regularities exist in protein sequences which result in characteristic residue environment for every amino acid. They also observed that though amino acid residues show a nearly random distribution, the di- and tripeptide segments reveal a non-random distribution. They have also suggested that medium and long range correlations among nucleotide at DNA level for the sake of DNA stability may result in amino acid preferences, We have also shown that protein sequences are a special subset of the set of polypeptides and have attempted to distinguish between random sequences and natural sequences (Meeta Rani, Cserzo M, Mitra C K and Simon I, unpublished results), Some studies show that the distribution of the residues in protein sequences has some fractal nature too (Mitra and Meeta Rani 1993).

We have tried to study the primary structures of natural proteins by techniques used in analyzing time-series. A time-series is the recording of events according to a horizontal axis along which equal intervals correspond to equal intervals of time. The techniques used to study time-series can be extended to spatial situations also, as in the case of protein primary structures where the horizontal axis is the

*Corresponding author.

positional axis instead of the time-axis. The serial correlation test measures the dependencies between successive terms, for series which are not random. The array of correlations of different orders obtained from the distributions are converted by fourier transformation into a spectrum. From such a spectrum for any given pair we can understand the nature of relationship between them. The most intense peak in the spectrum is generally selected as corresponding to the highest of the preference values between the residues of the pair. The position of this peak on the positional axis corresponds to the most common frequency and division of 360 by this frequency gives the periodicity at which the two pairs occur most frequently in the data base. Hence from the spectrum the highest preference value, as well as the corresponding periodicity, can be calculated easily. Therefore, by analysing the spectra of the various pairs we can understand and make general conclusions about the relationship between various amino acid residues, *i.e.*, mutual preferences and repulsions, and also the preferred distances from each other within the sequences.

2. Methodology

The swiss-prot protein sequence data bank, release 10·0, March 1989, containing 10,008 sequences and about 3 million residues, was used to study periodicities of amino acids in protein sequences.

Using the data base the positional distribution of each residue up to 200 positions was obtained. Only those sequences which had at least 200 residues were considered (5034 sequences). This was done because the modal value of the sequence length is close to 200, as clearly seen from figure 1. Thus, about one million residues were used to obtain the positional distributions. Residues beyond 200th position

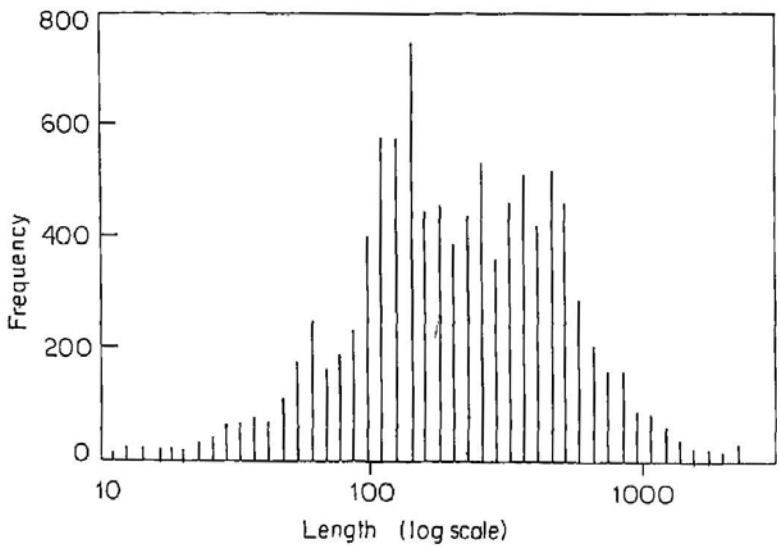

Figure 1. The distribution of the sequence lengths of proteins in the data base. The lengths are plotted on the x-axis, in the logarithmic scale, against their respective frequencies, on the y-axis.

were ignored. A program was written in turbo pascal to do the positional countings for each residue (Meeta Rani 1990).

The serial correlation test (Kendall *et al* 1983) was performed on the positional distribution of the residues with respect to themselves as well as other residues. The correlations obtained can give us the intra-dependencies (*i.e.,* autocorrelations) as well as the inter-dependencies (*i.e.,* cross-correlations) among the positional distributions of residues in protein chains. In addition, one expects to detect the periodicities of amino acids, if any, in the protein sequences. The total number of residues of type i, where i takes values from 1 to 20 and denotes the ordinality of one letter code of the amino acid residue and l denotes the positions from 1 to 200 along the protein sequences is denoted by $u_{i,l}$. u is a 20 × 200 matrix and each row of this matrix denotes a positional distribution for one of twenty amino acids. Considering any two residues l and j, the correlation between them, if any, can be calculated from their covariances (Kendall 1976). The correlations as mentioned above are two kinds. When $l = j$ (both are same amino acid residues) the covariances are called autocovariance and correlations are called autocorrelations, When $l <> j$ (the two amino acid residues are distinct) the covariances are called cross-covariances and correlations are called cross-correlations. The autocorrelations give a measure of self-preferences among residues, while the cross-correlations give a measure of the preferences between various residues for each other. The correlation between residue l and j are calculated as follows:

The mean frequency $E(u_{i,l}) = \mu_i. = \Sigma(u_{i,l}/200)$. (1)

Similarly, the mean frequency $E(u_{j,l}) = \mu_j = \Sigma(u_{j,l}/200)$. (2)

These frequencies are essentially the observed proportions in the data base for the respective amino acid residues.

Let s denote a positional lag between the two series $u_{i,l}$ and $u_{j,l}$ *i,e.,* we compare $u_{i,l}$ and $u_{j,l+s}$ always in our calculations. The positional lag s is allowed to vary from 0 to 99 (hence a maximum periodicity of 100 can be detected).

The sth covariance $c(i, j, s)$ between i and j is defined as

$$c(i, j, s) = E\,[(u_{i;j} - \mu_i)*(u_{j,l+s} - \mu_j)], \quad (3)$$

A correlation of order s between residues i and j is the correlation when they are separated by s number of residues along the chain. The sth order correlation between residues i and j, $r(i, j, s)$, is derived from the sth covariance, $c(i,j,s)$ as follows:

$$r(i, j, s) = c(i, j,s)/(\text{var } U_i * \text{var} U_j)^{1/2}, \quad (4)$$

where var U_i is the variance of the ith positional distribution. This can be rewritten as,

$r(i, j, s) =$

$$\frac{\frac{1}{n-s}\sum_{l=1}^{n-s}\left(u_{i,l}-\frac{1}{n-s}\sum_{l=1}^{n-s}u_{i,l}\right)*\left(u_{j,l+s}-\frac{1}{n-s}\sum_{l=1}^{n-s}u_{j,l+s}\right)}{\left[\frac{1}{n-s}\sum_{l=1}^{n-s}\left(u_{i,l}-\frac{1}{n-s}\sum_{l=1}^{n-s}u_{j,l}\right)^2*\frac{1}{n-s}\sum_{l=1}^{n-s}\left(u_{j,l+s}-\frac{1}{n-s}\sum_{l=1}^{n-s}u_{j,l+s}\right)^2\right]^{1/2}}. \quad (5)$$

In this equation n denotes the total number of data points, *i.e.,* 200 in the present case.

Let i and j denote a pair of residues. When $i = j$. let such a pair be called a homo-pair else be called a hetero-pair. Out of the 400 possible pairs 380 are hetero pairs and 20 are homo-pairs. We obtained the array of 100 correlations *i.e.,* from 0 to 99, for each pair (i, j, The correlations are transformed by fourier transformation into spectral densities, The array of spectral densities plotted against the frequency is called a spectrum. The spectrum and the correlogram uniquely determine each other.

The symbols $c(i, j, a)$ and $r(i, j, a)$ are used to denote cross-covariance and cross-correlation respectively of a sample, whereas $\gamma(i, j, a)$ and $\rho(i, j, a)$ denote the cross-covariance and cross-correlation of the population, *i.e.,* the complete series.

Since we are also dealing with cross-correlations in addition to autocorrelations (a case of multiple time-series), it was considered appropriate to construct the coherence spectra for analysing the relationship of hetero-pairs; for the homo-pairs the squares of the spectral densities are plotted instead of coherences as they were equal to one, at all angles, as expected.

For any pair of series say u_i and u_j and positional lag of s number of residues varying from $-\infty$ to $+\infty$ we obtain a set of correlations $r(i, j, s)$. Then the spectral density is defined over the range of angles from 0 to π. In actual computation s was taken from 0 to 99 only. The angle a is in radians. The spectral densities are obtained by fourier transformations of the correlations ranging from $-\infty$ to $+\infty$, and is written as follows:

$$w(i, j, a) = \Sigma\, \rho(i, j, s) * e^{i*s*a}. \quad (6)$$

In the above expression $e = 2{\cdot}731$ and $i = (1)^{1/2}$. $w(i, j, a)$ is also written alternatively and equivalently as $w_{i,j}(a)$. Whereas the summation over s theoretically runs from $-\infty$ to $+\infty$, because of the symmetric nature of the autocorrelations, we can consider only half of them, *i.e.,* from 0 to $+\infty$. In addition, we can use a finite sum, in this case up to $s = 99$, as an approximation. In the case of homo-pairs the coherence is always equal to unity and we plotted the squares of the spectral densities (intensities) to obtain the spectra. The amplitudes of these intensities were compared to obtain the five highest peaks, as there were often several significantly high peaks with only small differences in heights. From the positions corresponding to the peaks we obtained the respective frequencies and the periodicities were obtained by dividing 360 by the corresponding frequencies.

Since the cross-correlations are not symmetric, *i.e.*, $\rho(i, j, s) <> \rho(j, i, s)$, but $\rho(i, j, s) = \rho(j, i, -s)$ therefore the expression obtained for $w(i, j, a)$ becomes,

$$w(i, j, a) = 1 + \Sigma\{\rho(i, s) * \cos(s, a) + \rho(i, j, -s) * \cos(s, a)\}$$

$$+ \Sigma \{\rho(i, j, s) * \sin(s, a) - \rho(j, i, s) * \sin(s, a)\} \quad (7)$$

$$= c(a) + i * q(a). \quad (8)$$

The quantity $c(a)$ is called co-spectrum and $q(a)$ is called the quadrature spectrum. The sum of their squares *i.e.,* $c^2 + q^2$, is called amplitude of the spectrum. The standardized quantity "coherence", *i.e.,* $C[i, j, a]$ is defined as,

$$C(i, j, a) = [c^2(a) + q^2(a)]/[w_i(a) * w_j(a)] \quad (9)$$

$$= [w_{i,j}(a)]^2/[w_i(a) * w_j(a)]. \quad (10)$$

Where w_i and w_j are the spectral densities of u_i and u_j respectively. (The difference between the capital C and the lower case c are to be noted). The coherence spectra are symmetrical, *i.e.,* $C[i, j, a] = C[j, i, a]$. Coherence spectra of each pair has 180 values, one corresponding to each angle in 1 steps. There are 380 hetero pairs (*i.e.,* when $i <> j$) and due to symmetry we have considered only 190 spectra. For each of these pairs, the maximum amplitude of the coherences and the angle corresponding to the maximum amplitude was found using a program written in turbo pascal. We obtained 190 values, all the maximum coherence values for each pair.

These coherences were arranged in a decreasing order of magnitude, using a sort program. The 50 highest values *i.e.,* the 50 most intense peaks out of 90 were selected for further study. (The remaining peaks with smaller values indicate negligible correlations). The value of the most intense peak was noted down and also the corresponding frequency on the horizontal axis. These positions correspond to the highest occurrences or frequencies of the respective pair *i*, *j*. Corresponding periods were obtained by division of 360 by these frequencies. The periods give the periodicity *i.e.,* the number of amino acids separating the residues *i* and *j*. The length of the series being 200, periodicities greater than 100 cannot be detected, hence periods greater than 100 were ignored.

Next the data base were searched directly to see the accuracy of the period calculated from the coherence spectra. For every pair of residues *i* and *j* separated by a distance of *s,* the frequency of occurrence in the whole of data base was actually found out by direct counting. The observed proportion of finding the pair was found out by multiplying the observed proportions for the individual residues *i* and *j* (*i.e.,* the product of μ_i and μ_j). That is, the percentage occurrence obtained for each pair was divided by frequency of the constituent residue to get the preference for the pair. This quantity is called the pair preference. This was done for all the pairs and the values are presented in table 1 for self-preferences and in table 2 for cross preferences. It is to be noted that these values are essentially an average preferences and has been obtained from the whole of the data base and can be considered as attributes or characteristics of the data base. As one can see in the tables, in case of self-preferences, for each residue up to 5 different

Table 1. Preferences for various amino acid residues (self preferences for the twenty amino acid residues).

Residue	1st		2nd		3rd		4th		5th	
	Count	Perd.	Count	Perd.	Count	Perd.	Count	Perd.	Count	Perd.
Ala	18646[b]	(18)[c]	18134[b]	(23)[c]	20107[b]	(10)[c]	19142[b]	(16)[c]	18471[b]	(15)[c]
(7·791)[a]	9·071[d]	(1·16)[e]	9·005[d]	(1·16)[e]	9·476[d]	(1·22)[e]	9·227[d]	(1·18)[e]	8·869[d]	(1·14)[e]
Cys	2124	(12)	1671	(36)	938	(90)	1208	(60)	1785	(2)
(1·875)	4·248	(2·27)	3·778	(2·01)	2·784	(1·48)	3·109	(1·66)	3·415	(1·82)
Asp	8269	(5)	8286	(3)	8635	(2)	8544	(8)		
(5·213)	5·697	(1·09)	5·677	(1·09)	5·898	(1·13)	5·944	(1·14)		
Glu	13491	(8)	12889	(6)	12838	(10)	14531	(3)		
(6·150)	7·982	(1·30)	7·571	(1·23)	7·650	(1·24)	8·443	(1·37)		
Phe	4562	(5)	5332	(4)	4423	(2)	5663	(3)		
(3·927)	4·175	(1·06)	4·866	(1·24)	4·007	(1·02)	5·149	(1·31)		
Gly	21950	(9)	17021	(20)	21443	(18)	18509	(4)		
(7·301)	10·905	(1·49)	8·773	(1·20)	10·966	(1·50)	9·040	(1·24)		
His	2395	(4)	2148	(28)	2190	(2)	2219	(3)		
(2·287)	3·757	(1·64)	3·659	(1·60)	3·405	(1·49)	3·465	(1·52)		
Ile	9519	(4)	7428	(2)	7876	(5)				
(5·298)	6·423	(1·21)	4·982	(0·94)	5·333	(1·01)				
Lys	9873	(45)	12346	(3)	12022	(4)				
(5·778)	7·296	(1·26)	7·654	(1·32)	7·488	(1·30)				
Leu	24997	(10)	23531	(2)	22561	(24)	27965	(3)		
(9·060)	10·073	(1·11)	9·221	(1·02)	9·527	(1·05)	10·989	(1·21)		
Met	821	(180)	1647	(7)	1592	(2)	1662	(5)	1606	(4)
(2·481)	2·568	(1·04)	2·632	(1·06)	2·505	(1·01)	2·641	(1·06)	2·544	(1·03)
Asn	6462	(6)	6666	(2)	5902	(9)	6346	(4)		
(4·344)	5·376	(1·24)	5·475	(1·26)	4·958	(1·14)	5·246	(1·21)		
Pro	10541	(9)	11676	(4)	12019	(3)	10734	(12)		
(5·207)	7·393	(1·42)	8·061	(1·55)	8·274	(1·59)	7·605	(1·46)		
Gln	7190	(4)	6420	(5)	7180	(2)	7595	(3)		
(4·099)	6·273	(1·53)	5·621	(1·37)	6·228	(1·52)	6·608	(1·61)		
Arg	10240	(3)	9709	(6)	9892	(7)	10367	(2)		
(5·208)	7·025	(1·35)	6·743	(1·29)	6·896	(1·32)	7·081	(1·36)		
Ser	17185	(2)	16839	(3)	16956	(4)				
(6·996)	8·797	(1·26)	8·655	(1·24)	8·745	(1·25)				
Thr	10798	(4)	11853	(2)	10376	(5)	5185	(180)		
(5·855)	6·670	(1·14)	7·266	(1·24)	6·432	(1·10)	6·328	(1·08)		
Val	11142	(40)	12412	(12)	12928	(4)	12019	(17)	12894	(3)
(6·509)	7·029	(1·08)	7·046	(1·08)	7·114	(1·09)	6·945	(1·07)	7·073	(1·08)
Trp	820	(7)	753	(3)	666	(12)	579	(5)	591	(90)
(1·347)	2·198	(1·63)	1·992	(1·48)	1·820	(1·35)	1·542	(1·14)	1·728	(1·28)
Tyr	3646	(7)	3363	(4)	3691	(5)	3182	(2)		
(3·203)	4·132	(1·29)	3·769	(1·17)	4·149	(1·30)	3·542	(1·11)		

[a]Per cent probability of occurrence of the residue.
[b]Count refers to the actual number of pairs as determined from the data base.
[c]Perd. is the periodicity for the residue.
[d]Conditional probability in per cent.
[e]Relative (compared to randomized) degree of occurrence; preference.

periods had been selected and 80 preference values whereas in case of hetero-pairs, though there were 190 coherence spectra only 44 significant cross-preference values existed. The numbers in tables 1 and 2 reflect the total number of pairs of residues

Table 2. Preferences for various amino acid pairs (attractions).

AA pair	Pref.	Period	AA pair	Pref.	Period	AA pair	Pref.	Period
Gly-Pro	1·270	3	Pro-Gln	1·093	7	Asn- Ile	1·062	12
Trp-Cys	1·262	51	Ile-Lys	1·090	73	Trp-Pro	1·062	38
Phe-Trp	1·227	31	Ile-Trp	1·078	46	Trp-Asn	1·060	38
Cys-Tyr	1·168	16	Trp-Lys	1·074	46	Gln-Glu	1·060	2
Phe-Asn	1·167	156	Leu-Trp	1·073	10	His-Tyr	1·060	1
Tyr-Trp	1·157	40	Leu-Ile	1·073	17	Pro-Cys	1·059	172
Leu-His	1·156	17	Asn-Asp	1·072	1	Lys- Asp	1·058	7
Lys-Glu	1·133	11	Ser-Asn	1·071	1	Trp-Tyr	1·057	40
Asn-Lys	1·120	180	Cys-Pro	1·071	172	Trp-Phe	1·055	31
Cys-Phe	1·118	19	Gly-His	1·071	10	Glu-Ile	1·055	2
Phe-Ile	1·114	30	His-Ile	1·068	1	Tyr-Asn	1·054	18
Tyr-Cys	1·109	16	Phe-His	1·067	46	Ile-Tyr	1·053	44
Tyr-His	1·107	1	Lys-Ile	1·066	73	Asn-Glu	1·053	2
Glu-Lys	1·103	11	Thr-Asn	1·066	19	Ala-Gly	1·052	7
Trp-His	1·102	7	Ala-Pro	1·065	4	Glu-Asp	1·050	2
			His-Trp	1·065	7	Tyr-Phe	1·050	18
			Asp-Glu	1·062	2			
	> 10%				> 5%			

Preference is determined by dividing the observed proportions of occurrence by the theoretical probability of occurrence (*e.g.,* the probability of occurrence for the pair in a completely random sequence).
Period refers to the number of amino acid residues separating the given pair under consideration.
Preferences involving Met have deliberately been left out because of unusual distribution of methionine.

in the data base. For this computation all the sequences in the data base were used. Hence the total number of residues that were used in the determination of periodicities were much higher. Since all the residues were considered and the residues after 200 were not ignored, the total number of residues that were considered is much larger than the number of residues that are used to arrive at the periodicities.

Based on these values of self-preferences and cross-preferences, a "weight" was defined as the natural logarithm of the observed preference values. The weight of a sequence was the sum of the weights of the constituent weights due to the pairs that occur in the sequence (those pairs which didn't exist in the table of self or cross-preferences did not contribute to the weight of the sequence; it is negligible compared to the weight of the sequence). The weight of each sequence in the data base was calculated as follows: The 80 self-preferences and the 44 cross-preferences were fed in a program to calculate the weight of each sequence in the data base, based on the presence of these pairs in the sequences at the expected periodicity (periodicity here refers to the number of amino acids separating pair within a sequence). The weight of each sequence was initialized to zero. If any of these pairs were found in the sequence at the expected periodicities then the weight was incremented by the natural logarithm of the preference value associated with the pair at that periodicity. Finally the total weight of the sequence was obtained. Simultaneously, a random sequence of the same length as the sequence being analysed was generated and its weight was also calculated in the same manner. This process was repeated for all sequences in the data base as well as for the random sequences which had been generated. The weights of the natural and random

sequences were compared to check the extent of randomness and significant non-randomness, if any, in the natural sequences (see figure 3).

Using our table of preferences and corresponding periods, we have extended this technique to calculate the weights of sequences of a few protein families and have been able to show the presence of marked difference between the weight of the sequences of a family in question and random simulated sequence of the same length. We have also attempted to characterize these families by comparing them with weights of random sequences of the same length.

Since the observed patterns are expected to be valid in general for all sequences, we have tried these patterns on several families. These families have lengths less than 200 and were not considered for determination of the pattern. We have chosen several sequences from the data base for the following families for weight calculation – calmodulin, cytochrome c, heat shock proteins, histones and myoglobins. All sequences of a given family present in the data base were considered but fragments were excluded. The weight per residue in each case and also the mean weight per residue of the sequences of the family was calculated. A random sequence corresponding to each actual sequence (of same length) had been generated and the mean weight per residue and standard error in case of random sequences were calculated. These results are presented in table 3.

Table 3. The weight/residue ratio of real and simulated random sequences.

Protein family	Number of sequences	Length of sequences	Weight/residue real sequences	Weight/residue random sequences
Histone (H3)	24	134–136	0·058	0·047 ± 0·0022
Histone (H4)	10	102	0·055	0·046 ± 0·0026
Heat shock proteins	18	143–174	0·049	0·047 ± 0·0024
Calmodulin	15	138–162	0·048	0·047 ± 0·0016
Myoglobins	68	146–154	0·061	0·048 ± 0·0011
Cytochrome c (plants)	7	104–113	0·054	0·047 ± 0·0030
Cytochrome c (animals)	11	102–107	0·053	0·047 ± 0·0036

3. Results and discussion

The distribution of lengths of protein sequences in the data base was studied. Figure 1 gives the graphical representation of the distribution of lengths in the logarithmic scale. As seen in the figure, the modal length is close to 200. The mutual preference of residues in the various pairs were calculated as described earlier. Pairs having strongest preference values and the distance from each other at which at the strongest preference is seen (at some most likely unique distance from each other within the sequences) as calculated from spectral analysis of the serial correlations and verified by checking in the sequences of the data base, are presented in the tables 1 and 2. Table 1 contains preferences and the distances (periodicities) corresponding to the preferences. Since several strong preferences are seen in case of homo-pairs up to 5 strong preferences have been mentioned along with respective preferred distances between them.

Table 2 contains the preferences in case of hetero-pairs. Only the strongest peak along with the preferred distance within the sequence has been mentioned. There are 44 hetero-pairs with relatively strong preferences. The remaining hetero-pairs have not been considered for calculation of the weights of the sequences as they carry relatively small weight. Only pairs having more than 5% preferences have been listed (and considered) as compared to a random sequence. The random and natural sequence were compared by plotting the graphs of sequence weights *vs* log(length) and (sequence weight)/(sequence length) *vs* log(length).

3.1 *Sequence weight vs log*(*length*)

In the case of random sequences, as clear from figure 2a, the weight is found to be a function of length, as it increased with length. This is expected because as the length increases, more number of residues are present and the probability of occurrence of a pair which has not yet occurred in the sequence increases. Since this fact holds true for the natural sequences also, one does see in figure 2b a general increase in the weight with increase in length. But one also finds fluctuations in this pattern in figure 2b. Any additional weight than expected (as in a random case) at any length is due to strong intrinsic preferences among the residues in the sequence. The weight of a natural sequence of length 1000 is 60 units, whereas for a random sequence it is about 50. This deviation from the graph of random sequences is a measure of non randomness of the natural sequences, which though small does exists and is probably sufficient to impart biological functionality. It may be necessary to mention that non-randomness need not always impart biological activity but biological activity mostly implies non-randomness.

3.2 *Sequence weight/sequence length vs log*(*length*)

When we compare the weights/residue in the case of a random sequence it is found to be 0·046 whereas in case of a natural sequence it is found to be 0·063. The additional weights are due to the non-randomness in the primary sequences of the natural sequences.

In case of random sequences as well as actual sequences we find that the ratio of sequence weight to sequence length converges to a constant value of 100. But below the length of 100 one finds fluctuations and not a constant value. This is because we have not considered the preferences which correspond to periods greater than 100. Beyond 100 all the correlations have been ignored and a constant value of weight/residue is obtained. Using this technique, we have a rough way of distinguishing a random sequence from a natural one based on these weights alone. This method can possibly be made more accurate by introducing more number of pairs, even though they may carry less weight.

We calculated the weights of a number of sequences of the following proteins – histones, calmodulin, heat shock proteins, myoglobins and cytochrome c. In all these cases, it was seen that the natural sequences had more weight than random sequences. This confirms the correctness of our inferences made earlier regarding the weights of natural and random sequences as deduced from figure 2. In table 3, we have summarized the comparison of weights of sequences of five selected

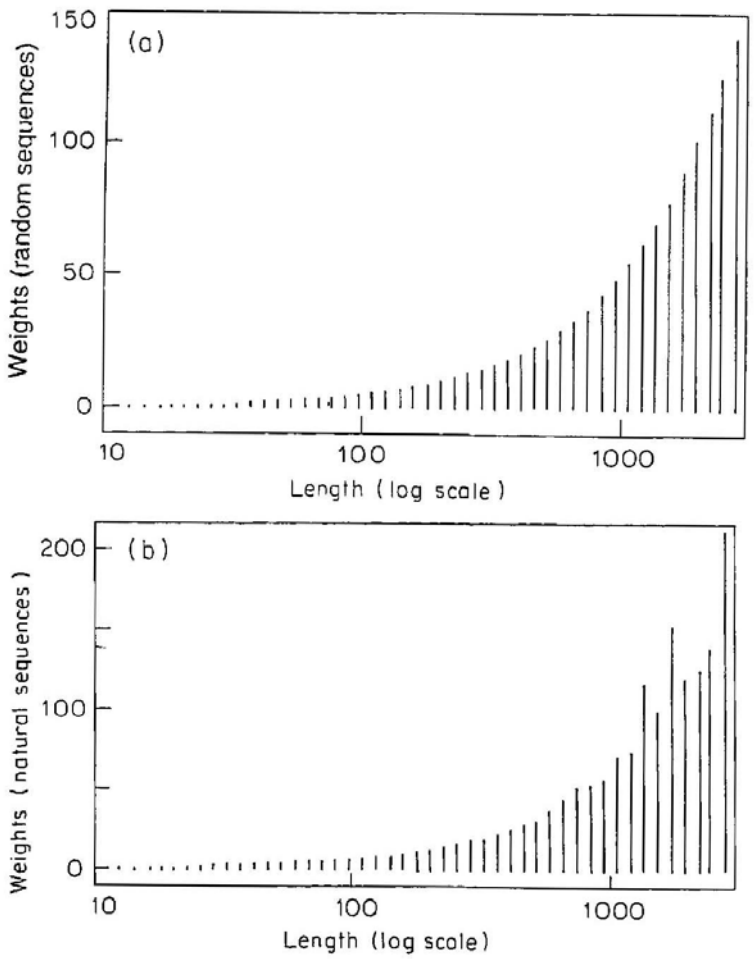

Figure 2. (a) The lengths of simulated random polypeptides are plotted on the x-axis against their respective "weights" on the y-axis. Weight is defined as the sum of weights of all those amino acid pairs which are present at a given periodicity in these sequences as required as in tables 1 and 2. (**b**) The lengths of natural polypeptide sequences, as found in the whole of the data base, arc plotted on the x-axis as against their "weights" on the y-axis. In this plot all sequences have been considered (not only the ones with length $\geq$ 200 residues and all residues, even beyond 200 have been included).

families. It can be seen that all of them have weights greater than random sequences. The difference of the weights from that of a random sequence gives a measure of the non-randomness of the protein. In that sense myoglobin sequences have maximum non-randomness and calmodulin sequences have least non-randomness among the five sets of families of proteins. However, less amount of non-randomness does not imply that the sequence does not have a conserved structure. It only suggests that in the process of evolution from a random sequence to a meaningful sequence it acquired its meaning or biological functionality much earlier compared to other sequences.

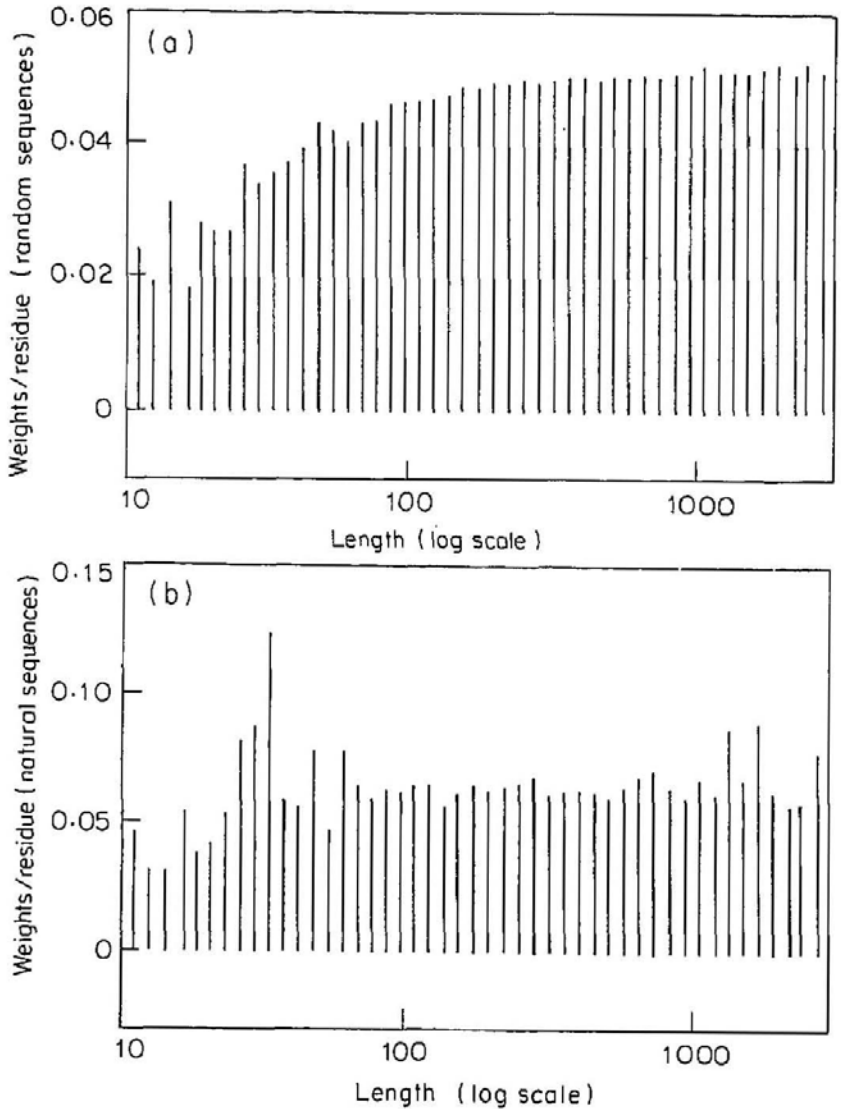

Figure 3. The weight/residue factor of all lengths [(**a**) simulated random polypeptides and (**b**) natural polypeptides of the data base] are plotted on the x-axis against the lengths in the logarithmic scale on the y-axis.

Ptitsyn (1984) has remarked that proteins are largely random polypeptides which have been edited in the course of evolution to impart biological meaning. We have found that weights of simulated random polypeptides are not markedly lower from those of the natural ones but off course they are always lower than them. The difference in their weights then is a measure of the non-randomness. For smaller lengths it is not expected to be very significant but for larger lengths one finds considerable differences. If one assumes that protein have emerged by edition of random polypeptides then those sequences which are older in the evolutionary scale must have smaller weight differences than the most recent ones. We have selected 5 groups of proteins and calculated their weights and based on the difference of

the weights of the simulated random polypeptides from them, have calculated the evolutionary status of these proteins in the increasing order of evolution as follows:

Calmodulin < heat shock proteins < cytochrome c < histones < myoglobin.

It is important to add in conclusion that real data used in the computations of these preference values may have "trends" and "seasonalities" that we have not made any explicit effort to eliminate. These may have some influence on our results, but since the finally used values are experimental ones obtained from the data base itself, their efforts are expected to be minimal (Beran 1992).

Acknowledgement

One of the authors (MR) wishes to thank the University Grants Commission, New Delhi for a senior research fellowship.

References

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

Kendall M 1976 *Time series,* 2nd edition (London: Charles Griffin)

Kendall M, Stuart A and Ord J K 1983 *The advanced theory of statistics,* 4th edition (London, High Wycombe: Charles Griffin)

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

Ptitsyn O B 1984 Proteins as edited polymers; *Mol. Biol. USSR* **18** 574–590

Simon I 1985 Investigation of protein refolding: a special feature of native structure responsible for refolding ability; *J. Theor. Biol.* **113** 703–710

Vonderviszt F, Matrai Gy and Simon I 1986 Characteristic residue environment of amino acid residues; *Int. J. Pep. Prot. Res.* **27** 483–492

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

Beran J 1992 Statistical methods for data with long range dependence; *Stat. Sci.* 7 404–427