J. Biosci., Vol.20, Number 5, December 1995, pp 579-590. © Printed in India.

Proteins as special subsets of polypeptides

MEETA RANI, CHANCHAL K MITRA†, M CSERZO* and I SIMON*
Department of Biochemistry, School of Life Sciences, University of Hyderabad, Hyderabad 500 046, India
*Institute of Enzymology, Biological Research Centre, Hungarian Academy of Sciences, Budapest, H-1518, PO Box 7, Hungary

MS received 18 February 1995; revised 20 July 1995

Abstract. A large protein sequence database with over 31,000 sequences and 10 million residues has been analysed. The pair probabilities have been converted to entropies using Boltzmann's law of statistical thermodynamics. A scoring weight corresponding to "mixing entropy" of the amino acid pairs has been developed from which the entropies of the protein sequences have been calculated. The entropy values of natural sequences are lower than their random counterparts of same length and similar amino acid composition. Based on the results it has been proposed that natural sequences are a special set of polypeptides with additional qualification of biological functionality that can be quantified using the entropy concept as worked out in this paper.

Keywords. Proteins; sequence analysis; pair-preferences; statistical thermodynamics; mixing entropy.

1. Introduction

A protein sequence differs from another sequence only in the number and kind of amino acid residues and the sequence of their arrangement. In principle, sufficient number of changes, including additions, deletions or substitutions, in the sequence of any protein can change it into another protein. The central question of protein evolution is how mutational changes in amino acid sequence leads to change in the structure and stability and thereby change in protein function. The three-dimensional structure of a protein is uniquely encoded in the primary sequence and in principle it is possible to predict the structure and function of a protein based solely on the primary sequence (Anfinsen 1973). In reality this has remained an unsolved problem even today although several empirical treatments of the problem are available in the literature (Chou and Fasman 1974a,b).

In his work on the *H* theorem, Boltzmann deduced a relation between entropy and probability: $S = k \ln W$, where S is the entropy (of a given system) and W is the fraction of all possible arrangements where the given configuration is realized (k is the Boltzmann constant). This statistical mechanical derivation of entropy as a measure of order (or randomness) is formally equivalent to the classical definition as given by the second law of thermodynamics. A rigourous derivation including quantum mechanical must include an additional factor or statistical weight while calculating the probability W. The probability W need not correspond to the equilibrium state of the system but the equilibrium state can be identified with the

†Corresponding author (Fax: 091-40-258120; Email: ckmsl@uohyd.ernet.in).

most probable probability distribution. This gives us a very powerful but simple technique to measure the degree of order in a given system. The equilibrium state has the most probable distribution and the highest entropy. The most ordered system has the lowest probability and the lowest entropy. As a system moves towards equilibrium, the randomness increases and the entropy also increases.

We have used the above idea to compute the entropies of protein sequences. In a protein sequence, it is well known from simple statistical frequency analysis that the various possible pairs (with 20 amino acids, it is possible to have 400 pairs) do not occur as expected if they were completely random. This indicates that the residues are not distributed randomly along the sequence. We can define a pair-preference as follows: Let $p(A)$ be the probability of finding a given residue A in the database. Thus $p(A)$ is equal to the total number of residues of type A divided by the total number of residues (of all types). This is nothing but the abundance of A in the given database. We can similarly define $p(B)$ to be the probability of finding a given residue B in the same database. This is equivalent to the abundance of B in the database. The abundances of all the common twenty residues have been obtained by direct enumeration. We now define a pair probability $p(AB)$ as the total number of AB pairs in the database divided by the total number of all possible pairs present in the database. A sequence of length N has a total of N - 1 pairs. In general, the total number of AB pairs is not equal to the total number of BA pairs and hence $p(AB) \neq p(BA)$.

If the distributions of the residues are completely random, then we have from the law of probability that $p(AB) = p(A).\ p(B)$. If there is any correlation between the residues A and B then the above relation is not valid. When the two residues A and B are not correlated, $p(AB) = p(BA)$, as evident from the above formula.We define a quantity called pair preference between A and B as $p(AB)/[p(A)\ .\ P(B)]$. If the distributions of A and B are independent (no correlation) then the pair preference value is equal to unity. If the pair preference is significantly different from unity, this indicates the presence of correlation between the respective pairs. The above principles can be used even when the residues A and B are separated by one or more residues (*vide infra* §2.1).

The pair preferences (as seen in the protein database) decrease the entropy or increase the order. A completely randomized sequence is not expected to show any such preferences and has the highest entropy. The completely randomized sequences were obtained by a Monte Carlo technique with a constrain that the overall distribution of the amino acids may not be changed. The randomized sequences form therefore a database that has the highest probability and has the same length distribution as the real database. Since entropy is an extensive property, we also expect that the mixing entropy (i.e., the entropy due to the presence of pair preferences) calculated can be divided by the sequence length to obtain an entropy per residue and that may be taken as an index for comparison with other sequences.

1.1 *Statistical analysis of primary sequences*

With the accumulation of a large number of known protein sequences in various databases (over 31 thousand sequences in a recent release of Swiss Prot Protein Sequence Databank at the time of working on the problem) and availability of

good computing facilities, it has become possible to study the primary sequences in detail.

A key question is whether there are some unknown rules governing the sequences and structures of proteins or they are random copolymers of the 20 amino acid residues. Scientists have studied this problem applying theoretical as well as experimental approaches. Sorm and Knichal (1962) have applied a theoretical approach in order to find whether all structural similarities of peptidic fragments contained in the primary sequences are real or not while Williams *et al* (1961) used an experimental approach for the same (however, their database consisted only of a few proteins). They concluded that even model structures drawn at random may exhibit regularities (symmetrical arrangement, repetition of sequences with interchange of individual elements, etc.,). They argue that only a comprehensive mathematical treatment of all observed regularities within the structure of an individual protein and also between different proteins will make it possible to judge the ignificance of such regularities. Such an attempt has been made by Meeta Rani and Mitra (1994), where they have applied Fourier analysis to detect periodicities.

2. Methodology

The frequencies of occurrence of the various amino acid pairs (20 × 20, i.e., 400 pairs) were obtained from the Swiss Prot Protein Sequence Databank (Release 26, 1994) containing 31,808 sequences and 10,875,091 residues. For reasons of convenience, all sequences that are shorter than 20 residues were excluded (802 sequences). No selective filtering of the database was attempted. We feel that screening of the database may introduce a fresh and additional bias (rather than to remove any bias already present). It is difficulty to remove any existing bias in the database when the sources of the bias are not well known or completely haracterized. The sequences present in the database is obviously not a representative random sample and there exists a strong bias introduced due to several factors: (i) biological importance, (ii) simplicity and ease of structure and sequence determination, (iii) personal interests/preferences in the class of molecules and (iv) the current trends and thrusts in the area. At present no known technique exists to remove all the above sources of bias. Therefore, apart from eliminating all sequences smaller than twenty residues, no other conditioning of the database was done.

2.1 *Calculation of frequencies*

The frequencies, i.e., actual counts, were obtained for the pairs separated by 0..9 residues (1st to 10th neighbours) by direct counting. If we consider a pair of residues A and B that are separated by any number (0..9) of any residue (denoted as X), i.e., a sequence of $A{\cdot}X_n{\cdot}B$, where n assumes a value between 0 and 9, then the frequencies of occurrences of such sequences are determined. The observed frequencies were normalized by the average frequencies, i.e., frequencies of occurrence assuming no correlation, of the respective pairs. This was calculated using the formula for the conditional probability for residues A and B, assuming independence. In other words, the average probability for the pair $A{\cdot}X_n{\cdot}B$ was taken as $p(A)\cdot p(B)$, where $p(A)$ is the probability of finding the residue A in the whole database (see § 1).

This was done for all the 20 × 20, i.e., 400 elements of the 10 matrices (corresponding to the 1st to 10th neighbours). A value greater than 1 suggests that the particular pair $A{\cdot}X_n{\cdot}B$ is preferred more than average. Similarly, a value less than unity suggests that the corresponding pair is less preferred. These values have been reported by Cserzo and Simon (1989) and Meeta Rani (1990). This matrix is referred to as the pair preference matrix. It is important to note that pair preferences beyond the tenth neighbour have not been explicitly taken into consideration. Hence contribution to the entropy (weights) from the long range correlations are absent in this calculation. It is also important to note here that for a completely random sequence, the normalized probability $p(A{\cdot}Xn{\cdot}B)$ can be greater than (in case of attractions) of less than (in case of repulsions) or equal to unity (for the completely random distribution). For the completely random case, all these normalized probabilities are equal to 1 for all values of n (number of intervening residues).

2.2 *Calculation of mixing entropies of the pairs*

As mentioned, the entropy is given by the statistical formula: $S = k$ In W, where W is the probability of finding the given configuration. In this work, we have explicitly calculated the pair preferences (from the database) and approximated W as $1/p(AB)$, apart from a multiplicative constant. The details of calculations of W are available in standard text books on statistical mechanics (ter Haar 1954). As the probability is multiplicative, the total mixing entropy will be additive (because of the presence of the logarithmic term). We note that the constant may include any statistical weights that may be applicable. The total mixing entropy is thus the negative sum total of all the pair preferences present in any given sequence. For the present application, we are simply interested in the relative entropy values (with reference to the completely randomized sequence) and the additive constant can be safely ignored.

As mentioned in the previous section the pair preference values are considered as conditional probabilities. Strictly speaking, these are not probabilities in the true sense,as they can be greater than unity. However, these are quantities that are normalized (i.e., a value of unity signifies a uniform or random distribution) and are easy for comparison. We expect, accordingly zero entropy for a completely random sequence and a negative entropy for a sequence that shows overall pair preferences and positive entropy for sequences that show overall pair repulsions.

The combined entropy or "mixing entropy" for each of these pairs was obtained by taking the natural logarithm of their conditional probabilities (*vide supra*), i.e., In *Pi.* Multiplication by the gas constant R is required for conversion of the result (and a change of sign) to mixing entropy. This was done as described below.

By direct counting, the total number of pairs $A{\cdot}X_n{\cdot}B$, where X is any residue and n varied from 0 to 9, were obtained from the database. All the possible residues (for A and B) were considered, i.e., ten 20 × 20 matrices were obtained. Each element of the ten matrices were divided by $p(A)$ and $p(B)$, where $p(A)$ is the probability of occurrence of the given residue A. The set of ten matrices are now referred to the pair preference matrices. Each element of the ten matrices were next converted into the entropy matrices by a transformation (– In x). It is to be noted that the value of n varied from 0 to 9 corresponding to a neighbourhood of 1 to 10. The total mixing entropy for each protein sequence of the database

was computed using the mixing entropy matrices. To calculate the weight of a given sequence, we add (since probabilities are multiplicative, their logarithms are additive) the entropy values for all the pairs present in the sequence using the table of themixing entropy. For comparison, for every natural protein sequence, a random chain of the same length was simulated using a Monte Carlo technique, taking into consideration the natural abundances of various amino acids. The mixing entropies of these random sequences were also calculated using the same procedure.

In the earlier work, using Fourier transformation of the correlations, we have shown that long range correlation does exist in significant amount in protein sequences. Therefore we cannot say that the mixing entropy calculated by this technique is unique, because neighbours beyond the tenth have been ignored. Nevertheless, we have a consistent and uniform procedure to compare various sequences. All the computations have been performed on an IBM compatible PC-AT/486 and all the programs were written in Turbo Pascal 6·0.

3. Results

To find out the distribution of sequence lengths in the database, we have plotted a histogram of distribution as shown in figure 1. For a clear picture, a logarithmic scale was chosen for the x-axis and the sequences were divided into 100 classes (equal interval in the logarithmic scale). It is to be noted that equal intervals in

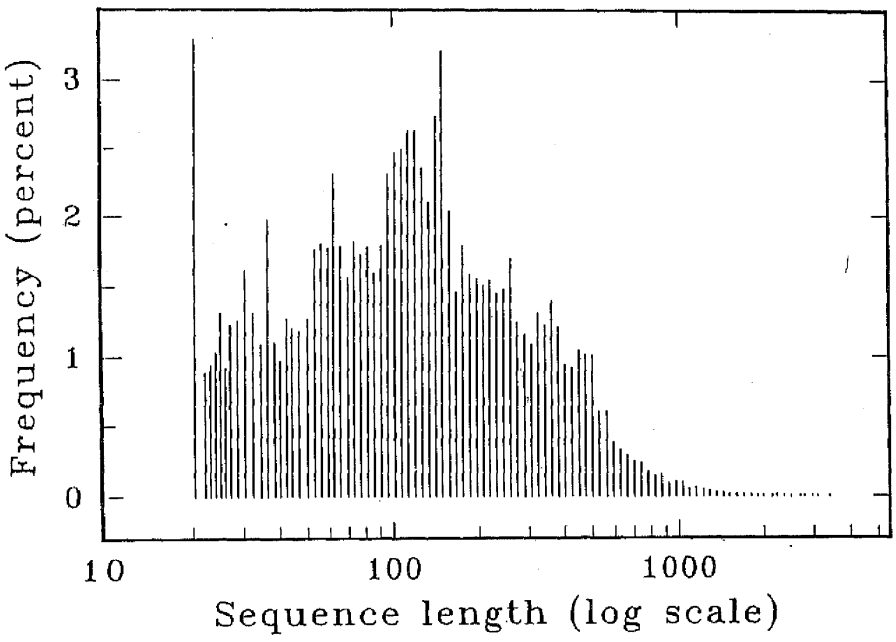

Figure 1. A histogram showing the distribution of the lengths of the sequences present in the database used. For reasons of convenience, lengths are plotted on a logarithmic scale on the x-axis. The modal sequence length is around 200 residues and the maximum length is around 4000 and sequences shorter than 20 have been ignored.

the linear scale do not correspond to equal units in the logarithmic scale. Therefore the number of sequences in each interval was counted and the frequency density was obtained by dividing the actual frequency by the class interval multiplied by 100. The histogram was finally plotted using these per cent frequencies. As mentioned, sequences shorter than twenty residues were not considered. The most probable (the most common) sequence length appears to be close to 200. This graph is important because the frequencies of various classes can be directly obtained from this graph. We also computed the amino acid composition for the whole database and the values are in good agreement with our earlier results in which a much smaller database was used. These values (frequencies of occurrences of various residues) were used in the simulation of random sequences by Monte Carlo technique.

To calculate the entropy of a given sequence, the mixing entropy matrix is required. As explained earlier, the matrix was obtained by direct enumeration of the database followed by normalization and taking logarithms. Since neighbours up to 10th position were considered, this is a 20 × 20 × 10 matrix, i.e., a matrix with 4000 elements. This matrix has been sorted and the pair preferences (attractions and repulsions) have been reported in the tables 1 and 2 for the first 100 strongest attractions and repulsions.

In these tables, positive values (of log (*P*) represent attractions and negative values represent repulsions. *F* refers to the relative separation (i.e., *F* equals to *n*+l, i.e., the *F*th neighbour) of the two residues *A* and *B*. It is to be noted that the interaction is not symmetric, i.e., preference of *A* for *B* is not the same as the preference of *B* for *A,* as has been explained earlier. These tables lists preferences only up to 10th neighbours and hence are not exhaustive.

In figure 2A, we present the total weight of a sequence in the database plotted against its length. As in the earlier case, the 31,000 weights were divided into 100 classes and the mean ± SD for each class were plotted as a function of the logarithm of mean sequence length of the class. As before, all classes have equal width in the logarithmic scale used. Figure 2B presents the same information for the randomized sequences. The difference between the two graphs is significant. This is somewhat expected, as the natural sequences follow positive preferences (attractions) and the total weight increases with the chain length. Both the graphs follow an exponential pattern (since the x-axis is logarithmic), i.e., the weight is linearly related to the sequence length. However, the slopes are quite different, as expected. Also, the random (simulated) sequences show far less scatter of weights, as indicated by the lower standard deviations.

The linear dependence of the weights can be seen more clearly if the weight per residue is plotted against the sequence length. This can be obtained simply by dividing the calculated weight by its length. For natural sequences, this is shown in figure 3A and the corresponding graph for random sequences is shown in figure 3B. As expected, the weight/residue for random sequences is a constant for all practical purposes. For natural sequences, the "noises" are considerable. As the class interval increases (because of the logarithmic nature of the abscissa), the frequency increases, and the standard deviation decreases for the randomized sequences. However, for normal sequences, such a pattern is not seen suggesting that variabilities are not governed by random processes. The reasons for this are not very clear at this moment.

Table 1. Preference values for the various aa residue pairs in the decreasing order (attractions).

No.	log(*P*)	*A*	*B*	*F*	No.	log(*P*)	*A*	*B*	*F*	No.	log(*P*)	*A*	*B*	*F*
1	1·0562	Cys	Cys	3	2	0·81894	Cys	Cys	7	3	0·78686	Cys	Cys	5
4	0·75369	Cys	Cys	6	5	0·72849	Cys	Cys	4	6	0·72542	Cys	Cys	10
7	0·6826	Cys	Cys	9	8	0·67796	Cys	Cys	8	9	0·56974	Cys	Cys	2
10	0·49977	His	His	1	11	0·49912	His	His	4	12	0·49121	Cys	Cys	1
13	0·48353	Gln	Gln	1	14	0·43523	Pro	Pro	3	15	0·41914	Trp	Trp	7
16	0·41029	His	His	3	17	0·40854	His	His	2	18	0·40831	Gln	Gln	3
19	0·4044	Pro	Pro	4	20	0·39975	His	His	5	21	0·39198	Pro	Pro	2
22	0·3893	Gln	Gln	2	23	0·385	Pro	Pro	6	24	0·38432	Trp	Trp	3
25	0·38171	Gln	Gln	4	26	0·37395	Trp	Trp	8	27	0·36899	Trp	Trp	4
28	0·36883	Gly	Gly	3	29	0·36544	Gln	Gln	7	30	0·34392	Glu	Glu	1
31	0·3421	Gly	Gly	6	32	0·33378	Pro	Pro	5	33	0·33198	Trp	Trp	6
34	0·33066	Gln	Gln	6	35	0·3221	Ala	Ala	4	36	0·31353	Trp	Trp	9
37	0·31341	Gly	Gly	9	38	0·31274	Glu	Glu	7	39	0·31109	His	His	6
40	0·30968	Trp	Trp	10	41	0·30843	Arg	Arg	1	42	0·3053	Ala	Ala	1
43	0·30341	Trp	Trp	2	44	0·30226	Pro	Pro	7	45	0·3011	Gln	Gln	8
46	0·29136	Trp	Cys	7	47	0·29126	Trp	Trp	1	48	0·2865	Pro	Pro	1
49	0·28538	Gln	Gln	10	50	0·28369	Glu	Glu	3	51	0·28243	Arg	Arg	2
52	0·28152	Glu	Lys	3	53	0·2806	Gln	Gln	9	54	0·27785	Pro	Pro	9
55	0·27695	His	His	9	56	0·27565	Lys	Lys	1	57	0·27328	Gln	Gln	5
58	0·26749	Pro	Pro	8	59	0·26719	His	His	10	60	0·26235	Arg	Arg	3
61	0·26186	Lys	Lys	3	62	0·26116	Pro	Pro	10	63	0·2606	Glu	Glu	4
64	0·25864	Ala	Ala	3	65	0·25765	Ala	Ala	2	66	0·25367	Ser	Ser	1
67	0·25192	Arg	Arg	4	68	0·2496	Lys	Lys	4	69	0·24875	Glu	Lys	4
70	0·24584	Lys	Glu	4	71	0·24391	Lys	Lys	2	72	0·2424	Lys	Lys	5
73	0·24146	Glu	Glu	8	74	0·23953	Lys	Lys	8	75	0·23614	Asn	Asn	2
76	0·23422	Arg	Arg	7	77	0·23287	Asn	Asn	1	78	0·23133	Lys	Lys	7
79	0·23021	Cys	His	4	80	0·22976	Glu	Glu	2	81	0·22864	His	Cys	8
82	0·22789	His	His	7	83	0·22578	Tyr	Cys	2	84	0·22542	Ser	Ser	4
85	0·22423	His	His	8	86	0·22391	Tyr	Tyr	5	87	0·22382	Cys	His	1
88	0·22356	Ser	Ser	2	89	0·22203	Asn	Asn	4	90	0·21979	Gly	Gly	2
91	0·21951	Lys	Lys	6	92	0·21886	His	Cys	1	93	0·21884	Glu	Arg	3
94	0·21865	Thr	Thr	2	95	0·21841	Tyr	Tyr	1	96	0·21471	Cys	Gly	4
97	0·21229	Ser	Ser	3	98	0·21217	Asn	Asn	8	99	0·2111	Gly	Gly	4
100	0·21063	Lys	Lys	9	101	0·20919	Gly	Gly	8	102	0·20826	Phe	Phe	4

For a more clear picture, we have plotted the distribution of weight/residue for both natural and random sequences in figure 4. The two distributions overlap considerably, but a clear pattern is apparent and the peaks are distinct. The distribution corresponding to the natural sequences shows a greater variance (width) compared to the random sequences.

These weights obtained can be correlated with the mixing entropy after multiplication with R (the gas constant). The entropy values so obtained may be considered as mixing or preference entropies based on a standard state which is given by the database used. It may therefore be observed that the most probable value of the average entropy (figure 4) is close to 0·1 (actually 0·08) for the natural sequences. On the other hand, the random chains have approximately 0·05 unit higher entropy per residue because of the lack of positional preferences (about 0·13). This can be an appreciable amount for chains of larger lengths (see figure 2A). The presence of residual entropy of 0·1 units for the natural sequences is

Table 2. Preference values for the various aa residue pairs in the decreasing order (repulsions)·

No.	log(*P*)	*A*	*B*	*F*	No.	log(*P*)	*A*	*B*	*F*	No.	log(*P*)	*A*	*B*	*F*
1	–0·28743	Trp	Pro	1	2	–0·27652	Pro	Met	1	3	–0·27069	Glu	Pro	1
4	–0·24067	Cys	Met	6	5	–0·23983	Cys	Met	1	6	–0·22593	Glu	Ser	1
7	–0·22484	Cys	Met	9	8	–0·22439	Cys	Glu	3	9	–0·21927	Cys	Met	2
10	–0·21806	Glu	Met	3	11	–0·21535	Glu	Met	7	12	–0·21498	His	Asp	1
13	–0·21317	Cys	Glu	1	14	–0·21133	Gly	Glu	3	15	–0·21129	Cys	Glu	7
16	–0·20999	Tyr	Ala	1	17	–0·20649	Pro	Met	5	18	–0·2037	Gln	Asp	2
19	–0·19875	Cys	Met	5	20	–0·19623	His	Glu	1	21	–0·19581	Pro	Met	4
22	–0·19446	Pro	Met	8	23	–0·19266	Gly	Lys	5	24	–0·19242	Lys	Met	4
25	–0·19133	Tyr	Met	5	26	–0·19055	Glu	Gly	4	27	–0·18874	Cys	Met	8
28	–0·18617	Cys	Glu	4	29	–0·18581	Pro	Met	3	30	–0·18493	Cys	Glu	8
31	–0·18444	Cys	Gln	3	32	–0·18404	His	Lys	1	33	–0·18403	Gly	Met	4
34	–0·18392	Gly	Lys	3	35	–0·18312	Glu	Ser	2	36	–0·18176	Asn	Ala	2
37	–0·18121	Ser	Met	3	38	–0·1807	Glu	Cys	4	39	–0·18054	Lys	Met	3
40	–0·18015	Gly	Glu	2	41	–0·17845	Asp	Gln	1	42	–0·17781	Ser	Met	1
43	–0·17535	Lys	Pro	2	44	–0·17369	Pro	Ile	1	45	–0·17351	Ala	Asn	4
46	–0·17315	Cys	Ala	10	47	–0·17243	Pro	Met	7	48	–0·17232	Cys	Ala	1
49	–0·1722	Leu	Met	2	50	–0·17141	Gly	Lys	4	51	–0·1712	Tyr	Ala	4
52	–0·17051	Ala	Cys	7	53	–0·17019	Pro	Lys	2	54	–0·17015	Gly	Asn	3
55	–0·1682	Asn	Gly	2	56	–0·16764	Asp	Gly	2	57	–0·16731	Phe	Met	1
58	–0·16701	Thr	Met	3	59	–0·16654	Glu	Trp	3	60	–0·16644	Asn	Gly	5
61	–0·16545	Lys	Met	7	62	–0·16545	His	Met	6	63	–0·16506	Trp	Met	8
64	–0·16467	Trp	Met	10	65	–0·16464	Ala	Cys	9	66	–0·16418	Tyr	Ala	3
67	–0·16293	Glu	Pro	2	68	–0·16237	Cys	Ala	8	69	–0·1613	Asp	Met	4
70	–0·16114	Cys	Met	4	71	–0·1605	Ser	Met	7	72	–0·16041	Glu	Gly	3
73	–0·1604	Ala	Asn	1	74	–0·15936	Tyr	Met	2	75	–0·15904	Arg	Met	9
76	–0·15882	Phe	Ala	1	77	–0·15876	Tyr	Ala	7	78	–0·15892	Gly	Gln	3
79	–0·15823	Cys	Ala	4	80	–0·1577	Trp	Lys	3	81	–0·15619	Pro	Asn	5
82	–0·15559	Asn	Gly	4	83	–0·15529	Glu	Cys	7	84	–0·15528	Gln	Met	6
85	–0·15526	Glu	Met	4	86	–0·15484	Gln	Met	7	87	–0·15483	Thr	Arg	1
88	–0·15403	Val	Met	5	89	–0·15343	Gly	Met	10	90	–0·15256	Pro	Asn	3
91	–0·15185	Tyr	Met	9	92	–0·15167	Ser	Met	4	93	–0·15131	Glu	Pro	5
94	–0.15089	Pro	Lys	6	95	–0.15082	Thr	Met	9	96	–0.15043	Asp	Met	10
97	–0.15024	Tyr	Ala	10	98	–0.15002	Val	Met	6	99	–0.14968	Cys	Ile	6
100	–0.14956	Arg	Met	4	101	–0.14893	Ala	Trp	3	102	–0.14864	Met	Trp	1

possibly due to the neglect of long range attractions and repulsions. As we have shown in an earlier work, such long range attractions and repulsions are actually present in natural sequences.

The overlapping part of the two curves is considerably large. We may offer several reasons for such overlap. From graph it is clearly seen that more than 95% of the values fall within 0·0 to 0·2 e.u. An examination of the values show that the overlapping region is dominated by relatively short sequences. Most short sequences, although naturally occurring, do not have a well defined 3-D structure and are close to random sequences. Secondly, the difference between the two peaks of the distributions is about 0·05 e.u. and as the standard deviations are relatively large, considerable overlap exists as a natural consequence. A large standard deviation (also seen in figures 1 and 2) for natural sequences perhaps confers robustness to the sequence against spontaneous mutations. Finally, the nature of the database itself may be sufficiently skewed so that the standard deviations of the natural

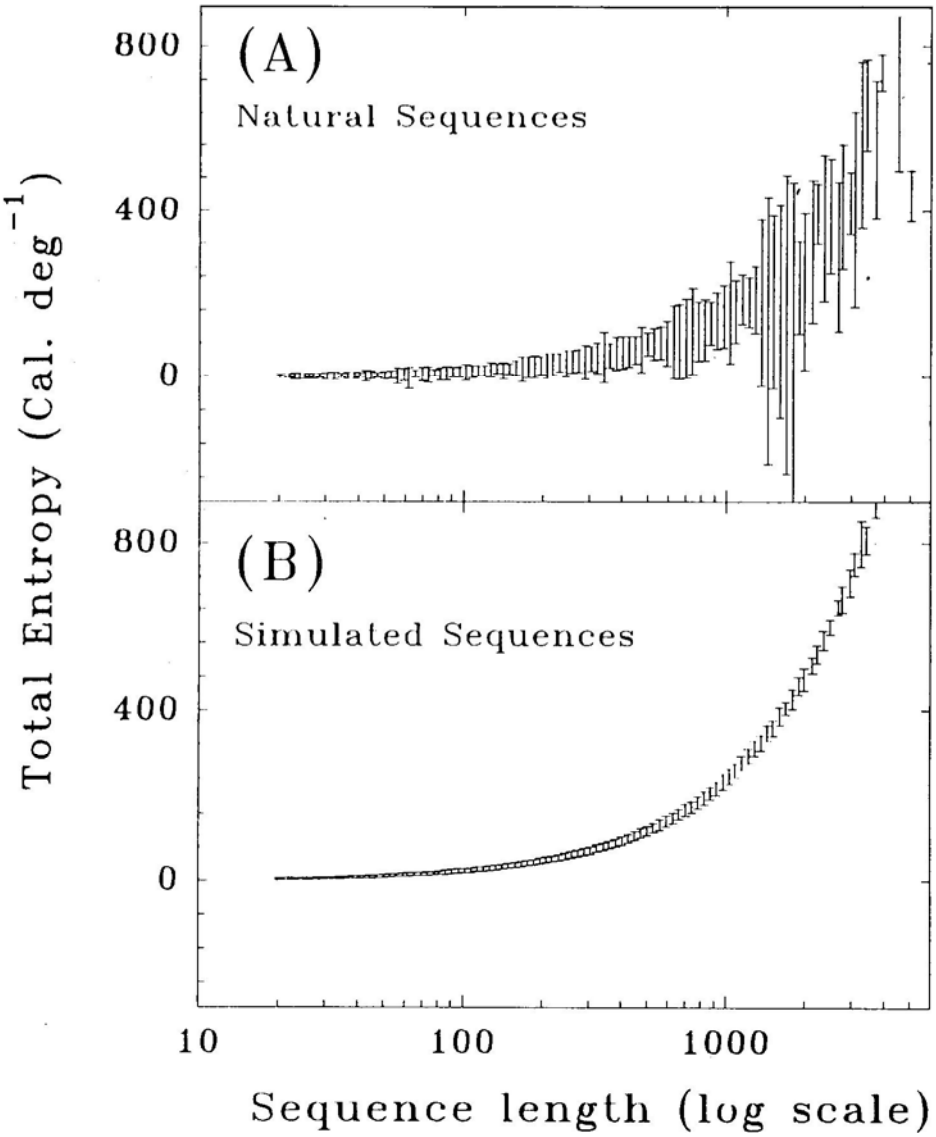

Figure 2. The total entropy due to pair attractions and repulsions present in the sequences in the database are plotted against their lengths. The sequences have been grouped into 100 classes on a logarithmic scale based on their lengths. The vertical bars represent standard deviations of the samples in their respective classes. **(A)** The total entropy of the natural sequences and **(B)** and of the random sequences generated using Monte Carlo procedure having an identical length distribution as the sample database. For case of comparison the two graphs have been plotted using the same scale.

sequences may be unrealistic. However, the confidence of distinction between real and random sequences increases with the increasing sequence length. Hence, the reliability of the computation improve as and when larger database and better analytical techniques becomes available. The natural and random-like amino acid

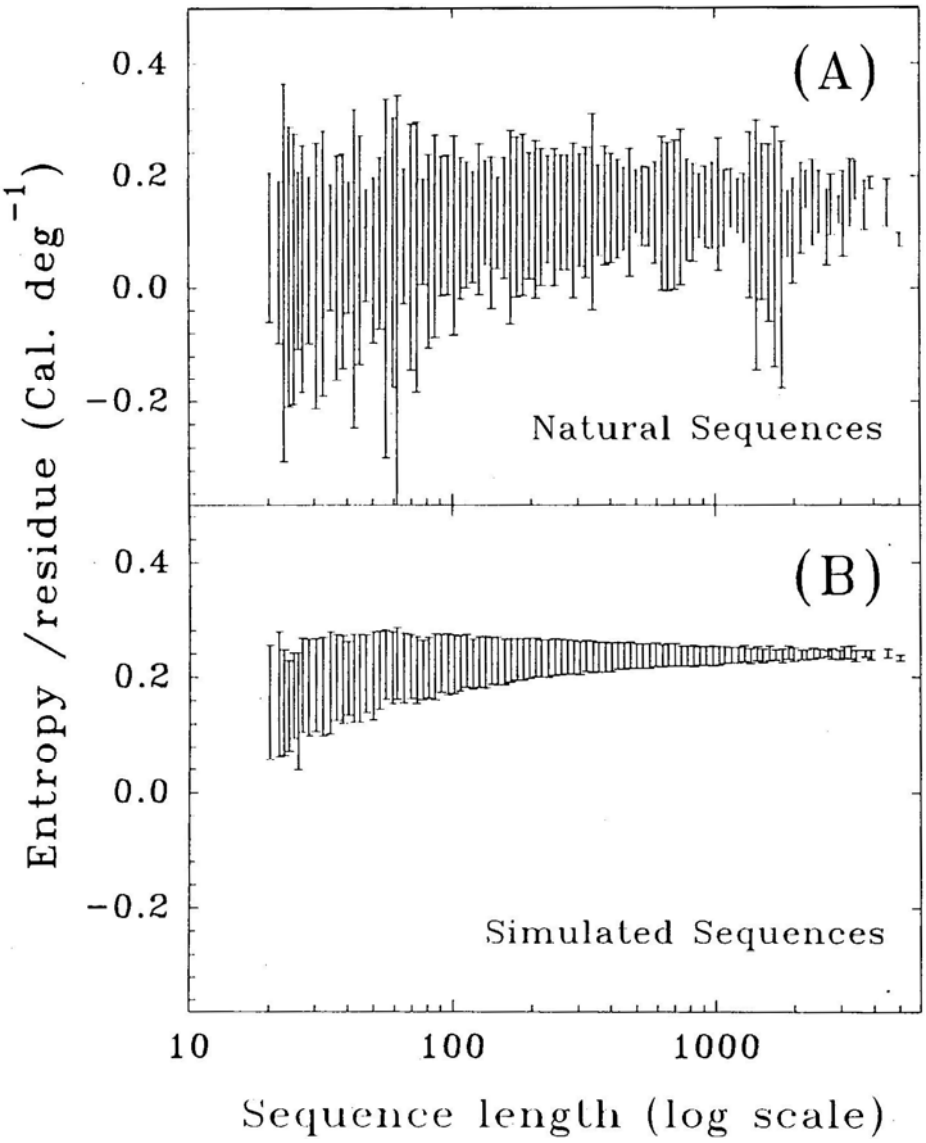

Figure 3. The entropy per residue plotted as a function of the sequence length plotted on a logarithmic scale, using 100 classes. The vertical bars represent mean ± SD calculated for the class. In **(A)** the entropy corresponds for natural sequences and in **(B)** corresponds to the randomized sequences generated by a Monte Carlo technique.

sequences are distinguishable if the chain is longer than 70 residues. Note that this length reminds the lower limit of length of protein structural domains. The procedure described above can be a test for identification of protein-coding open reading frames.

4. Discussions

The values of entropies shown in all the g raphs are in standard entropy units (i.e.,

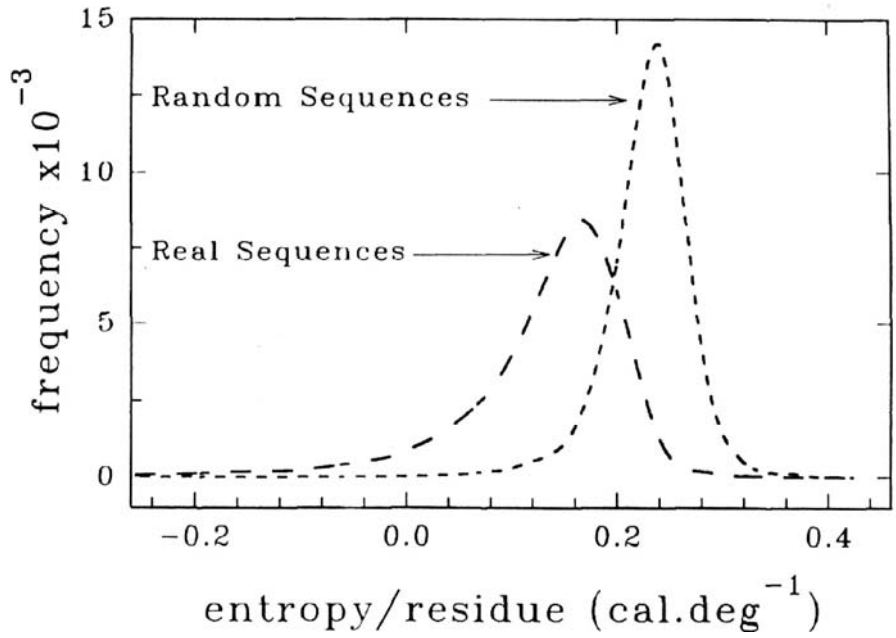

Figure 4. A plot of the frequency distribution of the mixing entropies for the natural and simulated sequences. The 31,006 weights are classified into 100 classes and their frequencies were obtained. The two distributions arc clearly different. Only the important region of the graph is presented and the tails extend beyond the graph shown. The natural sequences have, on an average, 0·05 e.u. less entropy/residue compared to the simulated sequences.

cal deg^{-1}) obtained by multiplication with gas constant *R*. However, it is to be noted further that the values are relative and not absolute. The point of reference (i.e., the zero of the entropy scale) depends on the database used. From the results mentioned above we propose that proteins are a special subset of polypeptides, characterized by the above mixing entropies. The characteristic that distinguishes a natural polypeptide sequence from a random one is the presence of pair preferences. Such preferences, we presume, is vital to the 3-D structure and the biological functionality of the polypeptide sequence. It may be argued that a random (e.g., a completely synthetic polypeptide) sequence may also have a well defined 3-D structure and a significant (unspecified) biological activity. We suspect that such sequences may not be truly random in the sense we have outlined above (zero mixing entropy or complete absence of pair preferences). However, the technique developed is rather poor for small sequences. Simon has earlier pointed out that proteins can be considered as generalized crystals, where the regular 3-D structure is analogous to the crystal state (Simon 1986). It is clear that the primary sequences of proteins is significantly governed by the pair preferences; this is expected since the folding of the primary sequence into a 3-D compact structure requires a considerable degree of co-operativity amongst the residues. It is to be noted in this connection that pair preference in a given region manifests as a pair repulsion elsewhere. Thus, both attractions and repulsions contribute to order in a given protein sequence. This interaction is mostly of van der Waal's kind; a very strong

interaction may preclude the diversity of protein sequences. In this light, the smallness of the "mixing entropy" values is not surprising. It is to be noted that the results obtained are dependent on the database used, because the pair preference matrix is obtained directly from the database. There is no reason to believe that the database used in our computation is a "random representative sample". However, it may not be desirable to doctor the database by selective elimination of sequences-this may introduce more bias than we intend to remove. On the other hand, since the database is sufficiently large in size, bias is likely to be relatively of lesser importance. Another point to be noted that the computations used do not take into account pair preferences beyond the 10th neighbour. In an earlier work, we have demonstrated that such preferences do exist to a significant extent (Meeta Rani and Mitra 1994, 1995) and may not be ignored. However, long range preferences "were not considered in this work because the actual counts may be rather small and based on the present calculations may not be very significant. We do not imply that their actual contribution to the total mixing entropy is negligible; however, the present technique needs computational improvements before long range correlations can be accurately incorporated in the calculations. This merits further detailed study. Direct counting from the database is a very straightforward exercise; however, if we wish to explore long range preferences, e.g., up to 50th neighbour, then the computations become excessively slow. Other techniques, e.g., correlation analysis, may be employed for such purposes. Nevertheless, the natural proteins retain the characteristic pair preferences and are therefore qualified to be described as a special subset of the set of all possible sequences of polypeptides (simulated or randomized).

Acknowledgements

This work was supported in part by grant OTKA 318 and 1361 from the Hungarian Academy of Sciences. CKM wishes to thank the Indian National Science Academy, New Delhi, for a travel grant and MR acknowledges receipt of Senior Research Fellowship from the University Grants Commission, New Delhi.

References

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

Chou P Y and Fasman G D 1974a Conformational parameters for amino acids in helical, β-sheets and random coil regions calculated from proteins; *Biochemistry* **13** 211–221

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

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

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

Meeta Rani and Mitra C K 1994 Periodicities in protein sequences; *J. Biosci.* **19** 255–266

Meeta Rani and Mitra C K 1995 Correlation analysis of the distribution amino acid residues in protein sequences; *J. Biosci.* **20** 7–16

Simon I 1986 Proteins as general crystals; *J. Theor. Biol.* **123** 121–124

Sorm F and Knichal V 1962 On proteins: Mathematical approach to the evaluation of similarities in protein structures; *Collection Czechoslav. Chem. Commun.* **27** 1988

ter Haar D 1954 *Elements of statistical mechanics* (New York: Rinehart) pp 70-84

Williams J, Clegg J B and Mutch M U 1961 Coincidence and protein structure; *J. Mol. Biol.* **3** 533

Corresponding editor: SAMIR K BRAHMACHARI