J Mol Evol (1983) 19:355–361

Journal of Molecular Evolution

Linear Repetitions of Amino Acids and Convergent Evolution Inside Protein Subregions of Ordered Secondary Structures

Christian Wuilmart[1] and Philippe Delhaise[2]

[1] Laboratory of Animal Physiology, [2] Laboratory of Chemical Biology, Université Libre de Bruxelles, 67, rue des Chevaux, 1640 Rhode-St-Genèse, Belgium

Summary. 51 polypeptides of known 3-dimensional structures have been submitted to a search for internal similarities. It is shown that the frequency of proteins displaying significant amounts of internal similarities is higher than predicted by chance. A non-negligeable part of those similarities probably occurs in connection with the existence of ordered secondary structures. Indeed, similarity occurs at a much more important rate when analyses are restricted to protein subsequences corresponding to α helices or β pleated sheets. Furthermore, the correlation existing between the rates at which linear and inverted repeats occur inside protein subregions of ordered secondary structures suggests that a significant part of short similarities are analogies rather than homologies. An hypothesis is put forward suggesting that the regular alternations of hydrophobicity which characterize most of α helices and β strands could provoke the occurrence of significant amounts of similarities inside protein sequences.

Key words: Analogies – Ordered secondary structures

Introduction

The biological properties of a protein are determined by its amino acid sequence. Since there are 10^{130} possible sequences for a protein of 100 amino acid lengths, it seems plausible that primitive proteins were short peptides. Because of their small length, those primitive peptides could exhibit many different conformational states. It is then possible that the lengthening of the ancestral sequences has been selected for in order to stabilize these conformations endowed with some biological activity. Our present understanding of the mechanisms which allowed the lengthening of ancestral sequences is based on several non-exclusive hypotheses.:

- proteins are lengthened by duplicating pre-existing segments of their amino acid sequences (see Woese 1971).
- proteins are lengthened at random: short amino acid subsequences are added randomly and the additions improving the stability of the primitive active site are selected for (McLachlan 1972).

Since Eck and Dayhoff's initial discovery (1966) showing that Clostridium ferredoxin was displaying two homologous halves, the former kind of hypothesis is accepted. Moreover, a close inspection of numerous proteins shows that homologous subsequences can be pointed out in many sequences of different origins (see Dayhoff 1978). Most subsequence similarities are then considered as relics of partial intragenic duplications.

However the existence of short similar subsequences in genetically unrelated proteins (Greller and Erhan 1974; Wuilmart et al. 1982), indicates that convergent evolutionary pressures could also provoke the occurrence of similarities inside protein sequences.

The goal of the present study has been to evaluate this hypothesis by looking for privileged locations for similarities. 51 proteins of known 3-dimensional structure have therefore been analyzed. Our results show that similarity occurs much more significantly inside or between protein subregions of identical ordered secondary structures, suggesting that an important part of the short internal similarities could result from intrinsic properties of α helices or β pleated sheets.

Offprint requests to: C. Wuilmart

Methods

Comparison Criteria. Amino acids are compared on the basis of two criteria. First, the similarity between two amino acids i and j is taken as the minimum number of mutations, mr (i, j) necessary for their interconversion according to the genetic code table. Second, similarity is measured by scores, s (i, j) derived from the relative frequencies of amino acid substitution in homologous proteins (McLachlan 1972).

Matching Experiments. The rate at which similarity occurs is derived by using a statistical test based on Fitch's polypeptide comparison method (1966). A protein of length L is analysed by comparing spans of n contiguous amino acids; a span starting at position i is compared to every span starting at position j (j > i); the total number of comparisons is then given by equation (1)

$$NSC = 0.5\,(\,(L-n)^2 + L - n) \qquad (1)$$

A score is derived for each of those NSC comparisons, for example, the minimum number of mutations (= MR) which is required to interconvert the two compared spans of n amino acids. The expected distribution of the different values of MR is approximatively gaussian whereas a marked deviation from randomness is observed when similarity occurs at a significant rate.

Probability Distribution. Suppose that a protein of length L is analyzed by comparing spans of n amino acids and let N_i and N_j be the numbers of positions respectively occupied by amino acids i and j. Then the probability of comparing i and j for one random matching position is given by Equation (2) if i and j are different and by Equation (3) if i and j are identical amino acids:

$$P_{i,j} = \frac{N_i}{L} \times \frac{N_j}{(L-1)} \qquad (2)$$

$$P_{i,i} = \frac{N_i}{L} \times \frac{(N_i - 1)}{(L-1)} \qquad (3)$$

The probabilities p_{mr} or p_s for obtaining by chance each mr or s value (with $0 \leqslant mr \leqslant 3$ and $0 \leqslant s \leqslant 9$) when two randomly selected amino acids are matched, are derived by summing all the possible $p_{i,j}$ corresponding to each mr or s value. The probability distribution of the different values of mr or s can then be obtained by combining the different p_{mr} or p_s values in a polynomial form:

$$\sum_{i=0}^{t} p_i x^i \qquad (4)$$

In polynomial (4), t has the value 3 if mr (i, j) are used to measure similarity and the value 9 if s (i, j) are used. For each term of (4), the coefficient of x^i is the probability p^i that mr (i, j) or s (i, j) has the match value i when two randomly selected amino acids are matched. The global result of a span to span comparison, MR or S is derived by summing the values of mr (i, j) or s (i, j) relative to each of the n matches and its probability P_{MR} or P_S can then be derived by raising polynomial (4) to the power n.

$$\left(\sum_{i=0}^{t} p_i x^i\right)^n \qquad (5)$$

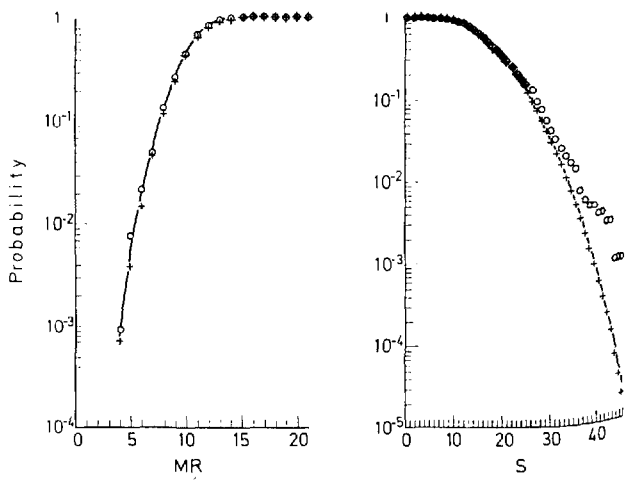

Fig. 1. Cumulative probability distributions of MR and S obtained by analyzing Rubredoxin with a comparison span of 7 amino acids (o) Real sequence analysis; (+) averaged results obtained by analyzing 400 MTCA sequences with the same length and amino acid composition as Rubredoxin; *smooth curves*: results predicted by the theoretical probability distribution.

Polynomial (5) describes the probability distribution of the possible global results ($0 \leqslant MR \leqslant 3n$ or $0 \leqslant S \leqslant 9n$): the coefficient of $x^{i.n}$ gives the probability that the global result, MR or S, has the match value i.n when two randomly selected spans of n amino acids are compared.

The relevance of polynomial (5) in deriving the probability distribution of MR or S has been established as follows: 400 Monte Carlo (= randomly constructed) sequences with the same length and amino acid composition as Rubredoxin have been generated. Each Monte Carlo sequence is analysed and the averaged frequency distributions of MR or S are then compared to their probability distributions derived from polynomial (5). Figure 1 shows the comparison between the theoretical probability distribution (derived from polynomial (5)) and the experimental probability distribution obtained by analyzing the shuffled sequences. It is seen that the theoretical and experimental probability distributions match perfectly. On the other hand, it is shown that a marked deviation from randomness occurs when the real sequence is analyzed, meaning that similar spans of amino acids are more numerous than predicted by chance in Rubredoxin sequence.

Statistical Significance of Results. The statistical significance of the deviations between the frequency distributions of MR or S obtained from matching experiments and their probability distributions is derived by using the χ^2 apparent test (Fitch 1970).

$$\chi^2_{app} = \frac{2 \times \sqrt{NIO \times NSC}}{1+(0.0038 \times n \times i)} - (f \times \ln(f/e) + (1-f) \times \ln(1-f)/(1-e)) \qquad (6)$$

The different parameters used in equation (6) can be described as follows:

NSC : total number of comparisons (given by eqn (1));
n : length of the comparison span;
f : cumulative frequency of MR or S;
e : cumulative probability of MR or S;
i : arbitrary value allowing a perfect matching between the distributions of χ^2_{app} and χ^2 for one degree of freedom. (=12 when the similarity between amino acids is measured by using mr (i, j) or 20 when s (i, j) are used);

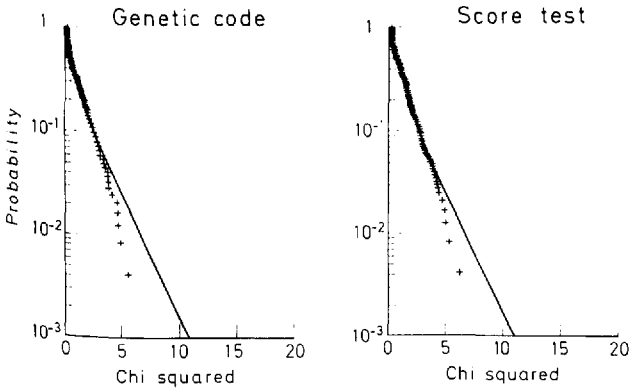

Fig. 2. Distribution of the highest values of χ^2_{app} obtained by analyzing Monte Carlo sequences of random amino acid composition and lengths varying between 50 and 400 amino acids. The genetic code or the score tests have been used as amino acid match values. *Smooth curves:* χ^2 distribution for one degree of freedom; (+) distribution of the 240 highest values of χ^2_{app} which are plotted in rank order from the largest to the smallest, their probability being equal to 1/240 th of their rank order

NIO: maximum number of independent observations. Two observations are independent if they have no amino acid match in common : the maximum number of independent observations can then be derived by using eqn (7)

$$NIO = NC \times (L - n) - n \times F(NC - 1) \quad . \qquad (7)$$

The different parameters used in eqn (7) are as follows:

L : length of the protein ;
n : length of the comparison span ;
NC : integer (L/n) ;
F(x) : $0.5\ (x^2 + x)$.

The relevance of eqn (6) in deriving the statistical significance of our results has been established by analyzing 240 Monte Carlo sequences constructed with random amino acid compositions and lengths varying between 50 and 400 amino acids. Each Monte Carlo sequence is analyzed by using comparison spans of 7 amino acids and the frequency distributions of MR and S are compared to their probability distributions by using eqn (6). χ^2_{app} values are therefore derived from the deviations between the cumulative frequency of occurrence of each value of MR or S and its respective cumulative probability. The two highest values of χ^2_{app} (one for each of the two criteria used to measure similarity between amino acids) are then selected to characterize each of the 240 Monte Carlo analyses. Those two randomly obtained distributions of χ^2 are then compared to the theoretical distribution of χ^2 for one degree of freedom (see Fig. 2): the 240 experimental values of χ^2_{app} are therefore plotted in rank order from the highest to the smallest, their probability corresponding to 1/240 th of their rank order. Figure 2 shows that the experimental χ^2 distributions obtained by analyzing Monte Carlo sequences using mr (i,j) (Fig. 2A) of s (i,j) (Fig. 2B) to measure similarity between amino acids are practically superimposed to the theoretical distribution of χ^2 for one degree of freedom.

Results

Linear repeats of amino acids are usually defined as spans possessing a high degree of similarity and occurring more often than predicted by chance. Fitch's polypeptide comparison method (1966) has then been used to analyse the sequences of 51 globular proteins of known 3-dimensional structure (Levitt and Greer 1977). Analyses have been performed by comparing spans of 7 amino acids. The resulting distributions of MR and S (the minimum mutation distance, mr (i,j) and scores derived from the relative frequencies of substitution, s (i,j) have been used to measure similarity between amino acids) are compared to their respective probability distributions by using the χ^2_{app} test (Fitch 1970): χ^2_{app} are derived to characterize the deviations between the frequency of occurrence of the different values of MR or S and their respective probability. χ^2_{app} are not derived for all values of MR or S. A cut-off arbitrary value of probability has been chosen in order to restrict calculations to the MR or S values obtained by matching spans possessing a sufficiently high degree of similarity. A probability of 0.1 has been chosen as cut-off value because the frequency of comparisons involving similar spans never exceeds this proportion in the 51 proteins which have been analysed (of course, an identical cut-off value has also been used in the Monte Carlo analyses described in Method section).

The 2 highest values of χ^2_{app} (one characterizing the distribution of MR, the other characterizing the distribution of S) are then selected as the result of the analysis. Those individual results are presented on Table. 1:

It is seen that χ^2_{app} values as high as 7.8 and 6.6 are obtained when ferredoxin and the Fab region of immunoglobulin New heavy chain are analysed. An internal duplication can be easily demonstrated in the ferredoxin amino acid sequence; in the same way, the homology existing between variable and constant regions of immunoglobulins is illustrated by the likeness of their cristallographic models (see Amzel et al. 1974). Lower values of χ^2_{app} are obtained when the analysis is restricted to the V region (4.63 and 4.3 for Rei V region) which indicates that this domain could also result from an intragenic duplication. Indeed, β strands 1, 2, 3 and 4 can be aligned with β strands 6, 7, 8 and 9 when the heavy chain is taken into account. However, that alignment is not very significant ($\cong 10^{-3}$, which is probably due to the occurrence of hypervariable subregions that span one third of the immunoglobulin V region) but it appears that numerous shorter similarities occur. They are found between parts of different pairs of β strands and even between partially overlapping spans inside a β strand.

Identical observations can be made when the globin sequences are analysed: similar sequences are relatively short and seem to be randomly distributed along the sequence. Moreover, the significance of the overall results indicate that no common trends appear within one given family of related sequences: χ^2_{app} values as different as 0.55 and 4.14 are obtained when methymoglobin and cyanmethemoglobin are analysed. These re-

Table 1. Upper χ^2_{app} obtained for each analysis using the genetic code (columns 1 to 3) and the score tests (columns 4 to 6) as amino acid match values. Analyses have been performed for the total sequences (columns 1 or 4) or restricted to subregions exhibiting identical ordered secondary structures: α subregions (columns 2 or 5) and β subregions (columns 3 or 6)

	Genetic code test			Score test		
	Total	Alpha	Beta	Total	Alpha	Beta
1 α-Chymotrypsin	0.0	–	0.0	0.0	–	0.05
2 Adenylate kinase	1.36	4.37	–	3.49	7.95	–
3 Alcohol dehydrogenase	2.27	0.41	0.63	2.72		
4 Alkaline serine protease B	0.0	–	3.20	1.33	–	9.09
5 Aquomethemoglobin 1	0.88	0.84	–	0.77	2.19	–
6 Aquomethemoglobin 2	1.65	13.38	–	1.88	6.35	–
7 Azomyohemerythrin	0.99	1.24	–	0.47	0.32	–
8 Bence-Jones MCG 1	0.0	–	3.15	0.42	–	0.19
9 Bence-Jones MCG 2	0.0	–	14.35	0.42	–	4.40
10 Ca^{2+} binding protein	0.45	11.57	–	0.0	8.38	–
11 Carbonic anhydrase C	1.16	0.15	2.02	2.12	0.08	2.35
12 Carboxyhemoglobin	0.09	0.48	–	2.06	2.15	–
13 Carboxypeptidase A	0.0	0.53	0.09	0.0	0.64	3.27
14 Carboxypeptidase B	0.0	0.36	0.32	0.02	0.0	0.86
15 Concanavalin A	0.51	–	3.26	1.07	–	7.39
16 CuZn superoxyde dismutase	0.0	–	0.13	0.01	–	0.06
17 Cyanmethemoglobin	4.14	0.0	–	1.57	0.0	–
18 Cytochrome C_{550}	0.58	0.0	–	1.37	0.0	–
19 D-Glyceraldehyde-3-phosphate dehydrogenase	0.00	2.56	0.72	0.02	7.05	4.25
20 Deoxyhemoglobin 1	0.86	0.66	–	0.14	2.02	–
21 Deoxyhemoglobin 2	3.29	5.92	–	4.08	7.56	–
22 Elastase	0.0	–	0.0	0.0	–	0.0
23 Ferredoxin	7.78	–	–	9.51	–	–
24 Ferricytochrome b_5	0.10	0.68	–	0.35	0.04	–
25 Ferricytochrome C	0.01	0.06	–	0.13	1.20	–
26 Ferricytochrome C_2	0.40	1.12	–	0.02	0.01	–
27 Hemerythrin	0.02	1.21	–	1.34	2.60	–
28 IgG Fab' (new) 1	0.16	–	3.76	1.02	–	0.81
29 IgG Fab' (new) 2	6.61	–	2.89	5.57	–	1.36
30 Insulin dimer 1	0.38	0.0	–	1.07	0.0	–
31 Insulin dimer 2	0.38	0.0	–	1.07	0.05	–
32 Lactate dehydrogenase	0.0	0.0	0.0	0.0	0.32	0.73
33 Lysozyme	0.0	0.06	–	0.72	1.11	–
34 Metmyoglobin	0.55	0.04	–	1.14	0.64	–
35 Nuclease	0.0	0.02	0.31	1.09	1.68	0.11
36 Oxidized flavodoxin	0.14	0.02	0.55	1.54	5.28	0.44
37 Oxidized high-potential iron protein	0.21	–	–	0.74	–	–
38 Oxidized thioredoxin	0.19	1.59	–	0.01	4.86	–
39 Papain	0.91	0.54	–	0.57	2.37	–
40 Prealbumin dimer 1	0.03	–	1.51	0.0	–	2.56
41 Prealbumin dimer 2	0.03	–	1.48	0.0	–	2.19
42 Ribonuclease S	2.52	0.00	0.0	0.49	0.06	0.61
43 Rubredoxin	0.72	–	–	3.09	–	–
44 Subtilisin BPN'	3.65	0.78	0.0	1.11	1.44	0.0
45 Thermolysin	0.03	0.70	0.00	0.59	2.45	0.26
46 Triose phosphate isomerase 1	0.03	0.0	0.0	0.09	3.36	0.0
47 Triose phosphate isomerase 2	0.03	0.0	0.01	0.09	2.24	0.0
48 Trypsin (diisopropylphosphate-treated)	0.0	–	1.35	0.0	–	2.18
49 Trypsin inhibitor	0.63	–	0.72	0.44	–	3.27
50 Variable part of human myeloma Rei 1	4.63	–	6.91	4.30	–	3.07
51 Variable part of human myeloma Rei 2	4.63	–	5.73	4.30	–	2.25

sults suggest that an important part of the short internal repeats are not shared by homologous sequences.

The distribution of the experimental χ^2_{app} values obtained in this analysis can be compared to the distribution of χ^2_{app} for one degree of freedom. Figure 3 shows that the theoretical distribution of χ^2 for one degree of freedom underestimates the χ^2_{app} distributions which are obtained from our analysis; suggesting that, the frequency of proteins exhibiting significant amounts of internal repeats is higher than predicted by chance.

Relics of ancient intragenic duplications involving a large part of the primitive sequence cannot be found in the majority of actual protein families. On the contrary, the interspersing and the relatively short lengths of

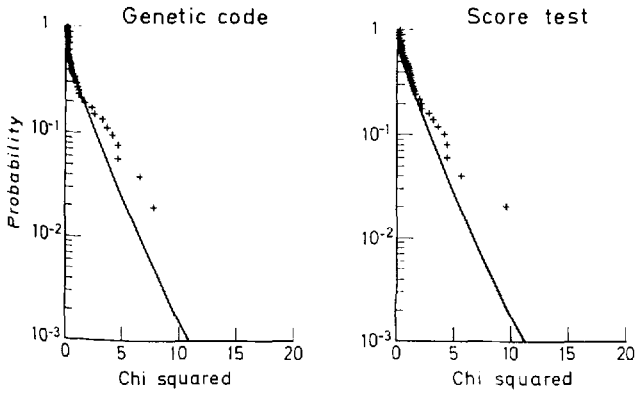

Fig. 3. Distribution of the 51 highest χ^2_{app} values obtained by analyzing 51 real sequences with a comparison span of 7 amino acids and using mr (i,j) or s (i,j) as amino acid match values. *Smooth curves:* distribution of χ^2 for one degree of freedom. (+) distribution of the 51 highest χ^2_{app} values

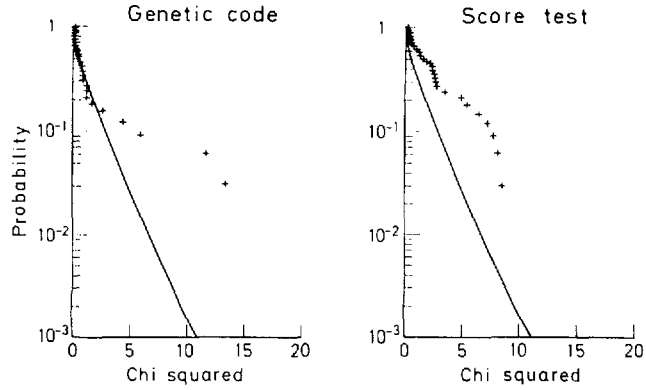

Fig. 4. Distribution of the 33 highest χ^2_{app} values obtained by analyzing the α subregions of 33 real sequences with a comparison span of 7 amino acids and using mr (i,j) or s(i,j) as amino acid match values. *Smooth curves:* distribution of χ^2 for one degree of freedom; (+) distribution of the 33 highest χ^2_{app} values

most of the internal repeats observed in numerous actual sequences could mean that most of the original peptides have been lenghtened by addition of short pre-existing subsequences. Nevertheless, a close inspection of the coordinates of the similar subsequences shows that they very often involve subregions of identical secondary structures. The statistical significance of this observation has been investigated by restricting our analyses to protein subregions displaying the same type of ordered secondary struture. α helices and β pleated sheets were taken into account. Those protein fragments are analysed in the same way as complete sequences by using comparison spans of 7 amino acids but equations (1) and (7) are no more relevant in deriving the total number of observations and the highest number of independent observations. In this case, the total number of comparisons, NSC, was derived by counting the comparisons one by one and the highest number of independent observations, NIO, has been derived by summing the NIO's relative to inter and intra-subsequences comparisons. Of course, the NIO values corrsponding to intra-subsequence comparisons can be derived from equation (7). In the case of inter-subsequence comparisons, the NIO were derived from equation (8) which can also be used when different proteins are compared:

$$NIO = NC \times (L + N - 2n + 1) - 2n \times F(NC - 1) \quad . \quad (8)$$

The different parameters used in eq. (8) are as follows:

L,N : lengths of the 2 proteins (or the 2 subsequences) which are compared :
n : length of the comparison span ;
NC : integer (minimum (L,N)/n) ;
F(i) : $0.5 \times (i^2 + i)$.

Analyses are restricted to proteins allowing a minimum of 30 comparisons inside their α or β subregions. χ^2_{app} derived from those analyses are presented on table I: the

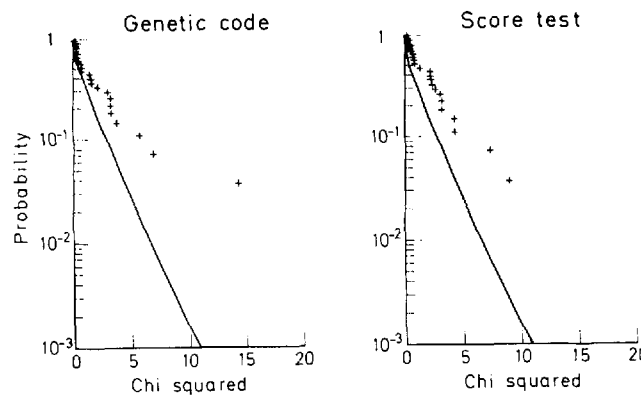

Fig. 5. Distribution of the 28 highest χ^2_{app} values obtained by analyzing the β subregions of 28 real sequences with a comparison span of 7 amino acids and using mr (i,j) or s (i,j) as amino acid match values. *Smooth curves:* distribution of χ^2 for one degree of freedom; (+) distribution of the 28 highest χ^2_{app} values

χ^2_{app} obtained by analyzing the complete sequence is given together with χ^2_{app} values obtained by restricting comparisons to subregions displaying identical secondary structures. It can be seen that similarity occurs generally at a much more significant rate inside the α or β subregions. χ^2_{app} as high as 13.4 or 11.6 are obtained when α helical subregions of Aquomethemoglobin or Ca^{++} binding protein are analysed. High values of χ^2_{app} characterize also the analysis of the β pleated sheet subregions of the variable part of immunoglobulin MCG (= 14.4). Those results are significant and relevant: – probabilities are derived by using the amino acid compositions of the analyzed subregions. – NSC strongly decreases when comparisons are restricted to subregions displaying identical types of ordered secondary structure. χ^2_{app} distributions have then been constructed by analyzing Monte Carlo sequences allowing few comparisons ($30 \leqslant NSC \leqslant 500$). Those distributions were shown to match with the theoretical distribution of χ^2 for one degree of freedom (result not shown).

Figures 4 and 5 describe the distributions of χ^2_{app} which are obtained when the analyses are restricted to the α and β subregions. It is clear that the frequency of proteins exhibiting high amounts of internal repeats inside their α and β subregions is significantly higher than predicted by chance.

This significant occurrence of similarity seems to be at least partly connected with properties of the α and β secondary structures. Indeed, we have shown that significant amounts of similarities are also observed when comparisons between unrelated proteins are restricted to the subsequences corresponding to a given type of secondary structure (Wuilmart et al. 1982). An ubiquitous physical property of subsequences corresponding to ordered secondary structures is the regular alternation of hydrophobicity which is displayed by the amino acid side chains. In α helices, 2 polar residues are often followed by 2 hydrophobic residues (Schiffer and Edmundson 1967) whereas polar and hydrophobic residues altern in β strands (Richardson 1977; Sternberg and Thornton 1977). Obviously, such a property of the ordered secondary structures could generate internal repeats: amino acid of related hydrophobicity are coded for by related codons and substitutions are practically always conservative. Yet such an origin for an important part of internal repeats implicits the existence of another kind of amino acid regularity: the symmetrical arrangements of amino acids. This second kind of amino acid regularity has been repeatly pointed out in different polypeptides (Urbain 1969; Bauer 1971; Wuilmart et al. 1975; Delhaise et al. 1980). Its occurrence cannot be put in connection with genetic "accidents" because of the high degree of polarisation of the genetic apparatus. We have recently shown that symmetrical arrangements of amino acids also occur preferentially inside or between the protein subregions sharing the same type of ordered secondary structure (Delhaise et al. 1980). The χ^2_{app} obtained in the present analysis can be compared with those which have been obtained in a search for symmetrical arrangements of amino acids. A significant occurrence of linear repetitions of amino acids is often in connection with a significant occurrence of symmetrical arrangements of amino acids and reversely. For examle, the analysis of the α subregions of aquomethemoglobin is characterized by a χ^2_{app} of 13.4 when the search is made for internal repetitions of amino acides and by a χ^2_{app} of 7.0 when the search is made for symmetrical arrangements of amino acids. On the contrary, both kinds of internal regularities do not significantly occur ($\chi^2 = 0$) when the lactate dehydrogenase sequence is analysed. χ^2_{app} values characterizing the occurrence of those 2 kinds of internal regularities can be correlated: liniear correlation coefficients as high as 0.6 and 0.5 are obtained by respectively considering the α subregions of 31 proteins and the β subregions of 27 proteins. We can note that the probability of exceeding such values of r is only 5%. Such a correlation between the occurrences of those two kinds of internal regularities strongly suggests that an important part of internal repeats and symmetrical arrangements of amino acids have the same origin, the regular alternation of hydrophobicity which is displayed by the protein subsequences of ordered secondary structure.

Discussion

51 amino acid sequences have been analyzed for the occurrence of internal repeats. Long amino acid repetitions compatible with intragenic duplications have only been observed in Ferredoxin, Fab region of immunoglobulin New Heavy chain, Subtilisin BPN', Carboxypeptidase A and Rubredoxin. In those polypeptides, the repeated amino acid subsequence represents approximately 40% of the total length whereas the repeated subsequences which are the most frequently found in our analyses rarely exceed 10 amino acids lengths. Complete intragenic duplications are then found in approximately 20% of the protein families represented in our sample. Which is understandable if one considers the influence of a complete intragenic duplication on the three-dimensional structure and the necessity for the resulting polypeptide to keep its original biological function. On the contrary, complete intragenic duplications involving closed structural domains naturally allow the conservation of the previous function simultaneously with the acquisition of new biological functions. It is for example the case of immunoglobulins, the constant regions of which are independent structural domains endowed with different biological functions.

However, the results obtained in this analysis show that the frequency of proteins exhibiting significant amounts of short amino acid repetitions is slightly higher than predicted by chance. Short repeats could thus constitute one of the properties of polypeptide primary sequences.

It seems obvious that partial intragenic duplication is not the only phenomenon able to explain the occurrence of short amino acid repetitions. Convergent evolutionary pressures could also provoke the occurrence of short internal repetions of amino acids.

In this respect, it is shown that short internal repeats are more often located in subregions of ordered secondary structure. Results obtained by analyzing α or β subregions are more significant than those obtained by analyzing complete sequences. Figures 4 and 5 clearly show that the frequency of proteins exhibiting amino acid repetitions inside their α or β subregions is significantly higher than predicted by chance. χ^2_{app} values as high as 11.5 or 14.3 are obtained when the α subregions of Ca^{++} binding protein the β subregions of immunoglobulin McG are analyzed. The relevance of those results is clearly established if one considers the comparison between hemoglobins α and β (Fitch 1970);

the upper value of χ^2_{app} is 14. Moreover, the length of most of the amino acid repetitions practically never exceeds the length of the protein subregions involved in a given type of ordered secondary structure: short amino acid repetitions have lengths varying between 5 and 20 residues whereas α helices and β sheets have respective lengths varying between 7 and 30 and between 5 and 20 residues.

This preferential location of internal repetitions of amino acids could difficultly be attributed to intragenic duplication events unless one admits that protein subregions of ordered secondary structure do evolve much more slowly than the other subregions. On the contrary, we know that an important property of the ordered secondary structures is the regular alternation in hydrophobicity which occurs at the level of their primary sequences. Such regular alternations of polar and hydrophobic amino acids could provoke the occurrence of repeated short spans of amino acids. The correlation which is found between the rates at which linear and inverted arrangements of amino acids occur strenghtens this hypothesis. Moreover, it has already been shown (Greller and Erhan 1974) that short similarities can be detected by comparing unrelated proteins. We were able to demonstrate that their frequency of occurrence significantly increases when the α or β subregions of 190 pairs of unrelated proteins are compared (Wuilmart et al. 1982).

Acknowledgements. This work has greatly benefited from helpful discussions with Professor Jacques Urbain.

This work was carried out under the association contract between Belgian State and the University of Brussels.

References

Amzel LL, Chen BL, Phizackerley RP, Poljak RJ, Saul G (1974) High resolution X-Ray diffraction studies of the Fab' Fragment of IgG NEW. Progr Immunol 1:85–92

Bauer K (1971) Homology of a partial sequence of calf thymus histone IV with several non-histone proteins. Int J Pept Protein Res 3:165–172

Dayhoff MO (1978) Atlas of protein sequences. National Biomedical Research Foundation Suppl 3, pp 359–362

Delhaise P, Wuilmart C, Urbain J (1980) Relationships between alpha and beta secondary structures and amino acid pseudosymmetrical arrangments. Eur J Biochem 105:553–564

Eck RV, Dayhoff MO (1966) Evolution of the structure of Ferredoxin based on living relics of primitive amino acid sequences. Science 152:363–366

Fitch WM (1966) An improved method of testing for evolutionary homology. J. Mol Biol 16:9–16

Fitch WM (1970) Further improvements in the method of testing for evolutionary homology among proteins. J Mol Biol 49:1–14

Fitch WM (1973) Aspects on molecular evolution. Ann Rev Genet 7:343–380

Greller LD, Erhan S (1974) Short length amino acid sequence homology among ancestrally unrelated proteins. Int J Pept Protein Res 6:165–174

Levitt M, Greer J (1977) Automatic identification of secondary structure in globular proteins. J Mol Biol 114:181–239

McLachlan AD (1972) Repeating sequences and gene duplication in proteins. J Mol Biol 64:417–437

Richardson J (1977) Beta-sheet topology and the relatedness of proteins. Nature 268:495–500

Schiffer M, Edmundson AB (1967) Use of helical wheels to represent the structures of proteins and to identify segments with helical potential. Biophys J 7:121–135

Sternberg M, Thornton JM (1977) On the conformation of proteins: hydrophobic ordering of strands in beta-pleated sheets. J Mol Biol 115:1–17

Urbain J (1969) Evolution of immunoglobulins and ferredoxins and the occurrence of pseudosymmetrical sequences. Biochem Genet 3:249–269

Woese CR (1971) Evolution of macromolecular complexity. J Theor Biol 33:29–34

Wuilmart C, Wijns L, Urbain J (1975) Linear and inverted repetitions in protein sequences. J. Mol Evol 5:259–278

Wuilmart C, Delhaise P, Urbain J (1982) The sharing of amino acid short spans by ancestrally unrelated proteins may be the result of ubiquitous alpha and beta secondary structures. Bio system 15:221–232

Received September 1981/Revised February 15, 1983