Proc. Natl. Acad. Sci. USA
Vol. 92, pp. 557-560, January 1995
Biochemistry

Periodic recurrence of methionines: Fossil of gene fusion?

(amino acid sequence/segmented protein/methionine distribution/recombination)

EUGENE KOLKER* AND EDWARD N. TRIFONOV*†

*Department of Structural Biology, The Weizmann Institute of Science, Rehovot 76100, Israel; and †Institute of Molecular Medical Sciences, 460 Page Mill Road, Palo Alto, CA 94306

Communicated by Peter B. Dervan, California Institute of Technology, Pasadena, CA, October 25, 1994 (received for review July 20, 1994)

ABSTRACT As we have recently shown, ≈20% of proteins are made of uniform size units of ≈123 aa for eukaryotes and ≈152 aa for prokaryotes. Such regularity may reflect certain past events in protein evolution by fusion (molecular recombination) of a spectrum of standard-size protein-coding DNA segments—the early genes. Consequently, methionines, as start residues, would mark those locations in proteins that correspond to the DNA recombination sites—the borders between the fused genes. This positional preference of the methionines may still survive as a fossil of the early protein sequence organization. In this study we address the question how methionines are distributed in modern protein sequences. This analysis of eukaryotic sequences shows that methionine residues do preferentially appear at the positions corresponding to the multiples of the unit size, as predicted.

The early notion that proteins might have a standard modular structure (1) has been recently confirmed (2, 3). A well-detectable (≈20%) excess of certain protein sequence sizes has been found. The preferred polypeptide chain lengths are multiples of 123 ± 3 aa for eukaryotes and 152 ± 4 aa for prokaryotes (3). Most (if not all) species and most protein types and families contribute to this general trend. The protein sequence size regularity is particularly strong among more conserved proteins (e.g., enzymes) (3). This could be taken as an indication that the observed underlying order in the protein sequence lengths is a reflection of an ancient regular organization of proteins.

At some early stages, proteins might evolve by combinatorial fusion (shuffling) of genes initially of the same elementary size. It is proposed that the early genes could have been fused through molecular recombination events—insertions of the protein-coding DNA circles of certain optimal size (3, 4). The size of 300–500 bp both ensures maximal efficiency of the DNA circularization (5) and has appropriate coding capacity (100–170 aa). The recombination events—that is, insertions and deletions of the standard unit-size DNA circles—might have occurred anywhere along the sequences. However, the locations of the sites of recombination preferentially at the ends of already existing protein-coding segments would be perhaps more likely. In this case the new composite genes would carry intact component genes (segments).

One consequence of the described hypothetical process of gene fusion would be the more frequent (initially perhaps strictly regular) appearance of the initiation triplets at the borders between the unit-length sequence segments. If the newly acquired methionines (initiation residues) were not in a conflict with the performance of the new recombination protein product, the presence of the methionines at the segment borders would be retained during later stages of protein evolution, and they still might be detectable in modern sequences (4). Similar behavior would be expected from termination triplets as well, or rather from mutational derivatives thereof. However, unlike the unique initiation codon (which did not necessarily have to change to be accommodated into the recombination product), the termination codons had to mutate to sense codons—i.e., to one of at least 18 possible mutational derivatives of the terminators. These derivatives would thus show substantially weaker positional preferences than the unique initiation triplet.

Most proteins (≈80%) do not follow the standard size regularity, making a rather smooth background to the overall size distribution. This perhaps can be explained by the combined action of various changes of the sequence lengths at later stages—e.g., an accumulation of numerous insertions and deletions in the originally segmented proteins. Clearly, in these proteins the borders between the internal segments would have drifted away from their regular positions. Exon reshuffling (6) would also destroy the initial pattern. One may expect that if only standard-size proteins (multiples of the segment size) were taken, the segment borders would be more frequently found at their appropriate regular locations. In this work we compiled three eukaryotic nonredundant protein sequence sets of the lengths of approximately two, three, and four standard segment sizes. The analysis of the distributions of methionines along these sequences shows indeed that these residues do retain the regular pattern with the expected period. This regular recurrence of the methionines can be considered as fossil evidence in support of both the general notion of gene fusion (excision–reinsertion) (7–9) and the particular mechanism of combinatorial fusion of the early genes (3, 4).

SEQUENCES AND METHODS

Sequence data (>13,000 eukaryotic protein sequences) were taken from the Swiss-Prot protein sequence data bank (10) release 24.0 and subjected to an extensive "cleaning." First, all irrelevant sequences, such as fragments, open and unidentified reading frame sequences, duplicated and almost duplicated entries, and functionally ill-defined sequences were removed (3). Then the sequences of proper lengths were selected. This resulted in reduction to a total of 1300 eukaryotic sequences, with lengths of 245 ± 15 aa, 370 ± 15 aa, and 490 ± 15 aa, multiples of the standard segment size. Two additional cleaning procedures were then applied to these sequence sets: an original technique based on the dipeptide compositions and performed by the CONTRAST program (11) and n-peptide (we used $n = 7$) sequence comparisons by the PROSET program (12). All closely homologous sequences were discarded, except for one representative. Finally, only those sequences that contained at least two methionines, one in the first position, were retained. As a result, the data set taken for further analysis contained 509 nonredundant sequences: 171, 188, and 150 sequences in the classes of two-, three-, and four-segment sizes, respectively. This 25-fold reduction of the original data base was necessary to ensure that the remaining sequences were indeed unrelated and thus suitable for the sequence analysis, as described below. The positional distributions of all amino acids were derived from this sequence collection.

For the distribution of methionine clusters, any two methionines found within the 8-aa window size were considered as a cluster. Similar results were obtained for other window sizes, as long as the window was small compared with the sequence segment size.

The smoothed positional distributions were obtained by calculating running averages with a 40-aa window.

RESULTS

The sequence sizes of three nonredundant sets of eukaryotic proteins (≈245, ≈370, and ≈490 aa) correspond to polypeptide chains that would be formed by the fusion of two, three, or four elementary sequence segments. The chains of all three sets would contain the border region between the first two segments. Thus, the entire data set of 509 sequences was used to analyze the methionine distribution around this region. Apart from a high concentration of methionines at the origin due to the initiation residues, a single maximum is observed at the position 120 ± 10 aa (Fig. 1, curve A), indistinguishable from the expected 123-aa position. Similarly, for the proteins of three- and four-segment sizes (Fig. 1, curve B), the 120 ± 10 aa peak is present. In addition, one more maximum of methionine occurrences is observed at ≈240 aa, which would correspond to the border between second and third segments (Fig. 1, curve B). The four-segment sequences (Fig. 1, curve C) show less pronounced maxima due to the smaller sample size, indicating, however, the presence of one more methionine peak at the boundary between third and fourth units. Thus, in the data set of sequences enriched by two-, three-, and four-segment-size proteins, the methionines are preferentially located at the multiples of 120 ± 6 aa, apparently marking the borders between the internal segments. This value is derived by cos–Fourier transforms of the curves in Fig. 1.

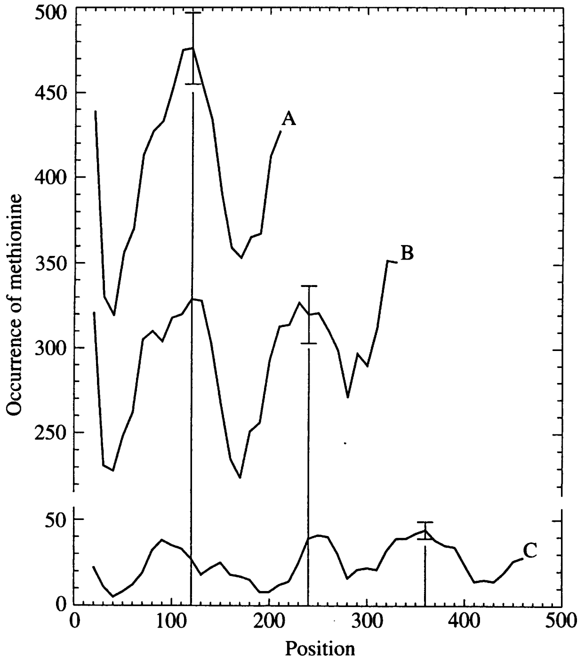

FIG. 1. Distribution of methionines along eukaryotic polypeptide chains. The curves correspond to three sequence sets: the complete set, with chain lengths close to 245, 370, and 490 aa (two, three, and four segments) (curve A), three- and four-segment-long sequences (curve B), and four-segment-long sequences (curve C). The two-, three-, and four-segment sets contain 171, 188, and 150 sequences, respectively. The smooth curves correspond to running averages with windows of 40 residues. Error bars were calculated as square roots of corresponding mean values of the methionine occurrences.

In addition to the above analysis of the methionine distributions along the sequences, we examined distributions of all other amino acids as well. Fig. 2 presents the relative occurrences of the 20 aa residues for all three sequence sets combined. Most residues manifest rather small variations around respective mean levels, with the exception of asparagine (N), cysteine (C), methionine (M), proline (P), and tryptophan (W). The methionines display a single dominating peak in the 100- to 150-aa region of interest. There is some preference for asparagine in this area (90–110 aa). However, only the methionines maintain statistically solid periodic recurrence at 120-aa multiples (Fig. 1). Analysis of positional distributions of the amino acids in a broader range (three- and four-segment proteins) also showed that only methionines have significant preference for the position at ≈240 aa (data not shown). The large scatter of occurrences of some amino acids close to the amino termini (Fig. 2) is due to the unusual composition of the leader ends, which are normally rich in nonpolar residues.

The effect of positional preference of the methionines is enhanced when the distribution of methionine clusters rather than of single residues is analyzed. The relative occurrences of both are shown in Fig. 3. This figure demonstrates not only that there is a tendency of the methionines to cluster, which was also found independently (13), but that the clusters have the same positional preferences at the borders between consecutive 120-aa sequence segments. Similar analysis of prokaryotic sequences (data not shown) indicates that in this case methionines also appear to be clustered at 145 ± 10 aa from the start, in agreement with the size of the sequence segments in prokaryotic proteins. The sequence sample size of available nonredundant prokaryotic sequences, however, is not large enough to make a statistically firm conclusion.

DISCUSSION

The results confirm a crucial prediction about the periodical recurrence of the methionines that follows from the hypothesis on recombinational mechanism of protein evolution (4). According to this hypothesis, the larger genes had been formed by fusion of small genes of standard sizes in a form of DNA rings. As a result, the proteins developed a segmented structure, still detectable in their lengths and in the periodical recurrence of the methionines.

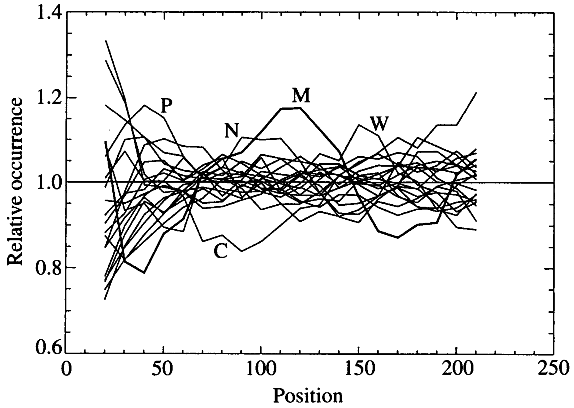

FIG. 2. Distribution of the relative occurrences of methionines (thick line) and all other amino acids. The curves correspond to running averages with windows of 40 residues. Mean values of amino acid occurrences are calculated without the first 20 aa residues, where the amino acid composition is strongly biased. Curves corresponding to asparagine (N), cysteine (C), methionine (M), proline (P), and tryptophan (W) are labeled.

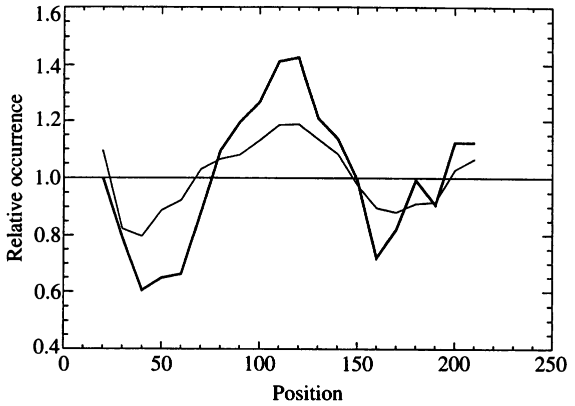

FIG. 3. Distribution of the relative occurrences of methionines and methionine clusters (thick line) in the eukaryotic sequences. Methionine clusters are present in 329 of 509 sequences. The curves correspond to running averages with windows of 40 residues.

One can think of several possible reasons for the apparent segmented structure of protein sequences. First, as discussed above, it may be a fossil of the segmented organization of primordial protein sequences. Second, exon reshuffling (6, 14) could potentially lead to the segmented structure. We believe, however, that the observed protein sequence regularity reflects the shuffling of our standard size segments rather than exons. The length of the segment, ≈120 aa, is not a typical exon length, which is ≈40 aa (15, 16). Neither exons nor introns show any periodicity size-wise; hence the exon reshuffling could only smear any periodic size pattern.

There also could be some other selection pressure(s) in favor of the regular segmentation with preferential positioning of some amino acids within and at the borders between the segments—e.g., to ensure optimal folding of the segments. In such a case, a compromise between structural diversity of the polypeptide chains (the longer, the better) and efficiency of their folding (the shorter, the better) may lead to the formation of distinct segments. This mechanism, however, is unlikely to explain the observed recurrence of methionines. Indeed, the border regions between the formed segments would be expected to have some distinct structural properties as the linkers between independently folding optimal size units. It does not seem to be the case, however. First, our analysis indicates that none of the amino acids other than methionine shows periodic appearance at the 120-aa repeat distances. Second, the methionine residue alone would not be sufficient to provide any special folding properties to the border region, and, third, not all protein sequences even have the single methionine at the segment borders. Thus, the possible pressure of optimal folding does not seem to cause the regularity in the distribution of methionines. It is, however, worth noting that the methionines possess higher flexibility and "stickiness" than other nonpolar amino acids (17). In the border regions the methionine residues may thus survive as points of contact with nonpolar residues, including the methionine itself (18).

The contour lengths of the early circular genes are assumed to correspond to integer numbers of triplets, in order to keep the reading frames of the fusion products intact. Molecular recombination, generally, is a very complex process that involves diverse systems of recognition and various mechanisms of interaction. In the context of our hypothesis only one aspect of the process matters, the conservation of total DNA length during recombination events. Generally speaking, the recombination could result in some loss of the sequences involved, as in illegitimate recombination in mammals (19) or in chromosomal translocations in protozoa (20). Similarly, the fusion product could also acquire some additional sequence, apart from the integrated sequence unit *per se*, as in the case of the insertion sequences in prokaryotes (21). The recombination events also could involve both loss and some gain of the sequence material simultaneously, as in known cases of retroviral DNA transfer (22). We believe that early recombination mechanisms perhaps did not have that degree of sophistication and occurred without any loss or gain of the sequences, very much like site-specific recombination (integration) of bacteriophages (23–25) and plasmids (25, 26), for example. Also, in homologous recombination during meiosis the conservation of the length of the exchanging sections is a dominant rule (27).

Interestingly, in the case of the standard DNA rings with contour lengths modulo 3, a striking quantitative agreement between calculated and experimental estimations of the average free DNA helical repeat is observed (4). This comes about when the rings with the contour lengths divisible by 3 are conditioned to be torsionally relaxed (flat), which would perhaps be optimal for efficient recombination.

Strong independent evidence in favor of the segmented structure of proteins is provided by the modular organization of many proteins, with the modules appearing in various linear combinations (for reviews, see refs. 28 and 29). Although the known modules do not all have the same size, those of them which deviate from the above sequence sizes may well be later derivatives thereof due to insertions and deletions.

It is quite likely that the observed regularities in protein sizes and in methionine distribution could be the result of a combined action of all the above-mentioned factors or even some other mechanisms. The fusion of the standard-size early genes was perhaps a dominant factor, in which case the preferred protein sizes of ≈120-aa multiples (3), recurrence of methionines with that period, correspondence of the segment to the optimal DNA ring closure size (5), and correctly predicted DNA helical repeat (4) all provide a strong factual basis to the general idea of gene fusion by excision–reinsertion (7–9) and to the particular segmentation mechanism (3, 4). The observed periodicities in protein sizes and in methionine distribution thus appear as fossils of those early days of protein evolution.

We are grateful to Drs. S. Schwarzbaum, S. Pietrokovski, I. Orr, J. B. O. Mitchell, A. S. Girshovich, J. Durup, and J. S. Beckmann for valuable comments. We thank the Biological Computing Division, the Bioinformatics Unit, and the Computer Center's UNIX group of the Weizmann Institute of Science for help. This work was partly supported by the Kimmelman Center for Biomolecular Structure and Assembly and by the Genome Committee of the Weizmann Institute of Science.

1. Svedberg, T. (1929) *Nature (London)* **123,** 871.
2. Savageau, M. A. (1986) *Proc. Natl. Acad. Sci. USA* **83,** 1198–1202.
3. Berman, A. L., Kolker, E. & Trifonov, E. N. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 4044–4047.
4. Trifonov, E. N. (1994) *J. Mol. Evol.* **38,** 543–546.
5. Shore, D., Langowski, J. & Baldwin, R. L. (1981) *Proc. Natl. Acad. Sci. USA* **78,** 4833–4837.
6. Gilbert, W. (1978) *Nature (London)* **271,** 501.
7. Bonner, D. M., DeMoss, J. A. & Mills, S. E. (1965) in *Evolving Genes and Proteins*, eds. Bryson, V. & Vogel, H. J. (Academic, New York), pp. 305–318.
8. Zuckerkandl, E. (1975) *J. Mol. Evol.* **7,** 1–57.
9. Cunningham, B. A., Hemperly, J. J., Hopp, T. P. & Edelman, G. M. (1979) *Proc. Natl. Acad. Sci. USA* **76,** 3218–3222.
10. Bairoch, A. & Boeckmann, B. (1993) *Nucleic Acids Res.* **21,** 3093–3096.
11. Pietrokovski, S., Hirshon, J. & Trifonov, E. N. (1990) *J. Biomol. Struct. Dyn.* **7,** 1251–1268.
12. Brendel, V. (1992) *Math. Comput. Modelling* **16,** 37–43.
13. White, S. H. & Jacobs, R. E. (1993) *J. Mol. Evol.* **36,** 79–95.
14. Patthy, L. (1991) *Curr. Opin. Struct. Biol.* **1,** 351–361.
15. Hawkins, J. D. (1988) *Nucleic Acids Res.* **16,** 9893–9908.
16. Dorit, R. L., Schoenbach, L. & Gilbert, W. (1990) *Science* **250,** 1377–1382.
17. Gellman, S. H. (1991) *Biochemistry* **30,** 6633–6636.

18. Durup, J. (1991) *J. Phys. Chem.* **95,** 1817–1829.
19. Meuth, M. (1989) in *Mobile DNA*, eds. Berg, D. E. & Howe, M. M. (Am. Soc. for Microbiol., Washington, DC), pp. 833–860.
20. Cowman, A. F. & Kemp, D. J. (1992) in *Mechanisms of Eukaryotic DNA Recombination*, eds. Gottesman, M. E. & Vogel, H. J. (Academic, New York), pp. 197–208.
21. Iida, S., Meyer, J. & Arber, W. (1983) in *Mobile Genetic Elements*, ed. Shapiro, J. A. (Academic, New York), pp. 159–221.
22. Varmus, H. & Brown, P. (1989) in *Mobile DNA*, eds. Berg, D. E. & Howe, M. M. (Am. Soc. for Microbiol., Washington, DC), pp. 53–108.
23. Pato, M. L. (1989) in *Mobile DNA*, eds. Berg, D. E. & Howe, M. M. (Am. Soc. for Microbiol., Washington, DC), pp. 23–52.
24. Nagaraja, R. & Weisberg, R. A. (1990) *J. Bacteriol.* **172,** 6540–6550.
25. Campbell, A. M. (1992) *J. Bacteriol.* **174,** 7495–7499.
26. Brown, D. P., Idler, K. B. & Katz, L. (1990) *J. Bacteriol.* **172,** 1877–1888.
27. Haber, J. E. (1992) *Curr. Opin. Cell Biol.* **4,** 401–412.
28. Doolittle, R. F. (1992) *Protein Sci.* **1,** 191–200.
29. Bork, P. (1992) *Curr. Opin. Struct. Biol.* **2,** 413–421.