THE STRUCTURAL PERIODICITY OF E. COLI RIBOSOMAL PROTEINS

O.Ch. Ivanov, P.S. Kenderov* and J.P. Revalski*

Institute of Organic Chemistry with Centre of Phytochemistry, Bulgarian Academy of Sciences, Sofia 1040
*Institute of Mathematics, Bulgarian Academy of Sciences, Sofia 1090, P.O.Box 373, Bulgaria

Abstract. It is established that the sequences of all different proteins from E. coli ribosome as well as two protein biosynthesis initiation factors, two ribosome-associated DNA-binding proteins, and the elongation factor EF-Tu from the same source possess a periodicity expressed more weakly and different from that found earlier for a number of proteins representatives of 18 superfamilies. The statistical significance of the periodicity observed was checked by comparing the area below the periodicity curve of every protein examined with that of computer generated sequences having the same amino acid composition and length. The results concerning the proteins from small and large ribosomal subunit are compared. The conclusions support and supplement the concept about the presence of a trend in protein molecular evolution from universal (Gly, Ala) to specialized (Phe, Tyr, Trp, Cys) amino acids.

INTRODUCTION

On the basis of the study of a considerable number of protein sequences representing different levels of evolutionary development evidence for the universality of structural periodicity in proteins were presented (1). Now, most of the 55 E. coli ribosomal protein sequences are known. These proteins perform important mutually connected functions in protein biosynthesis. The sequences of two of the three protein biosynthesis initiation factors, that of elongation factor EF-Tu, and those of two ribosome-associated DNA-binding proteins from E. coli were also determined. For the sake of brevity we shall call all these proteins ribosomal. Having in mind all stated above we set ourselves the task to elucidate whether ribosomal proteins possess periodicity,

Origins of Life **14** (1984) 557–564. 0302-1688/84/0141-0557$01.20

and if yes - to compare it with that of other proteins examined for periodicity (1) and, if possible, to reach a general conclusion about the molecular evolution of these important proteins.

METHODS

59 ribosomal proteins: 20 from the small and 34 from the large subunit, two protein biosynthesis initiation factors, the elongation factor EF-Tu and two ribosome-associated DNA-binding proteins from E. coli were tested for periodicity. As it has been stated in a previous paper (1) with a shortening of the length of the repeated segment the probability of mutational alteration decreases abruptly. That leads to a correspondingly significant increase in the importance of the identical residues occupying corresponding positions within the two segments compared. In view of the above mentioned reasons we directed our attention to the earlier suggested (2, 3) group analyses of the distances between identical residues along the protein chain. This analysis lies in the following. The cardinal numbers of each type of amino acid residue are written in an ascending order. By substracting each number from the following one (or ones - so that the difference should not exceed 10) the distances between identical residues are obtained. For each protein of interest the percentage of all identity cases (the identity cases for all amino acids multiplied by 100/N, where N is the length of the chain) corresponding to each distance from 1 to 10 are plotted along the ordinate. Thus, the periodicity curves were obtained (Figure). The percentage of periodicity was also calculated. For a given protein this is the sum of all identity cases for all distances up to 10 residues multiplied by 100/N. Actually this is the area below the corresponding periodicity curve.

To evaluate the statistical significance of the periodicity observed for each of 53 ribosomal proteins (the exceptions are: EF-Tu, S1, S2, L2, L9, L17) 100 artificial protein sequences having the same composition and length were generated by computer. A generator for random numbers uniformly distributed in the unit interval was used. For both periodicity curves and percentages of periodicity the arithmetical mean was calculated with a view to be compared with corresponding quantities of the artificially generated and of the previously examined proteins (1). The difference between the percentage of periodicity of the native protein and the corresponding arithmetical mean for the set of 100 generated proteins of the same composition and length was calculated. To determine the sensitivity of the method used, the difference between the mean percentage of periodicity for two sets of proteins generated on the basis of the same initial composition and length was computed. Two couples of sets based on the identical proteins L7=L12 and S20=L26 were used. The differences obtained

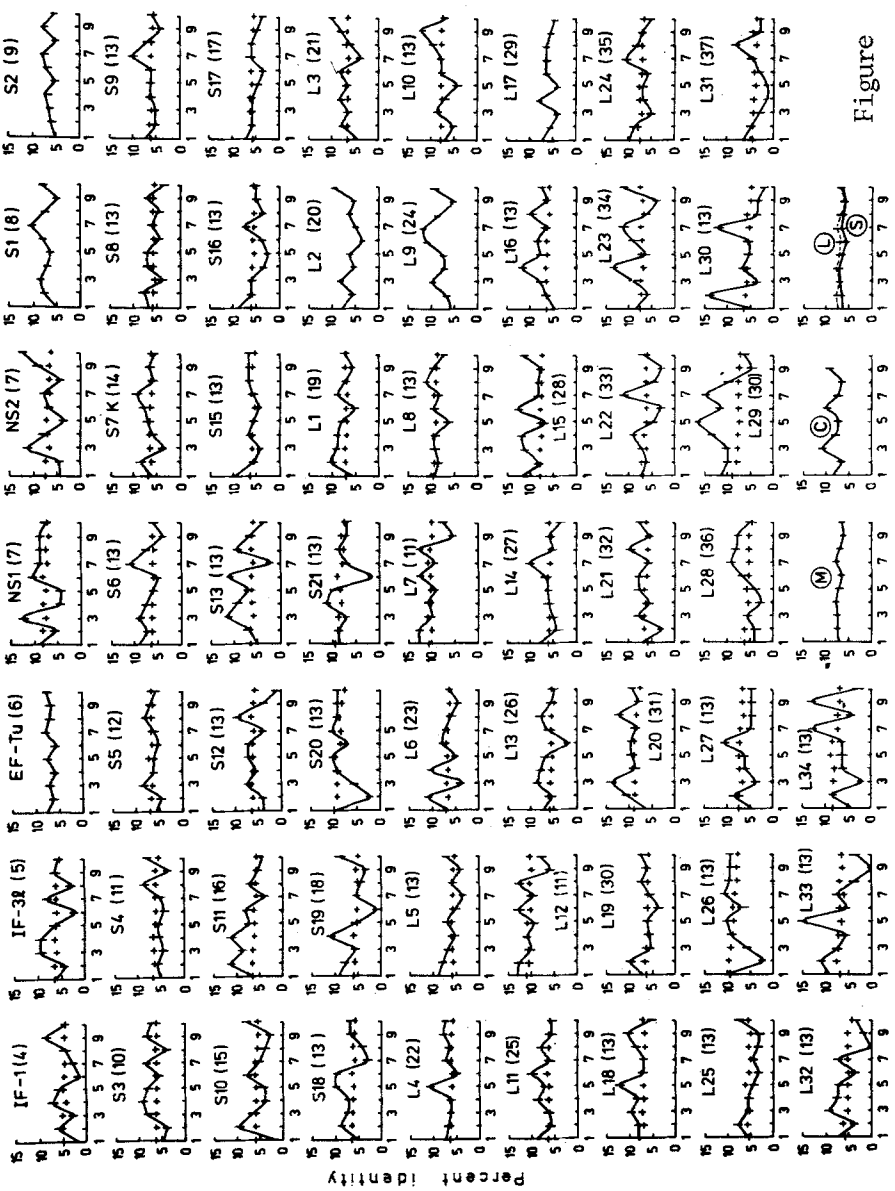
were 0.91% and 3.7% correspondingly. The corresponding differences by generation of 50, 100, 150 and of 50, 100 proteins were calculated for 2 and 6 cases correspondingly. It proved that the differences decrease when the number of generations increases. All possible differences were not greater than 0.76%. The value of 4% was accepted as a measure of sensitivity of the method.

RESULTS AND DISCUSSION

In 32 out of 53 cases of (22 proteins from the small ribosomal subunit and 31 - from the large one) the absolute value of the difference between the percentage of periodicity for the native protein and the mean percentage of periodicity for its generated proteins exceeds 4%. These cases were considered as relatively statistically significant.

It turned out that the mean percentage of periodicity for the 59 ribosomal proteins was 68.9, which is close to that of 72.5 for the proteins previously examined (1) representing 18 superfamilies [according to the classification of Dayhoff (38) from sources with a different level of evolutionary development]. However, the juxtaposition of the mean periodicity curves shows considerable differences (Figure): the proteins previously examined possess clearly expressed maxima corresponding to distances of 1, 3, 6, and 9 residues, while the maxima of ribosomal proteins are weakly expressed and correspond to distances of 4, 7, and 10. Obviously, the weakly expressed maxima of the mean periodicity curve for ribosomal proteins are caused by superposition of maxima and minima. Therefore, there is no genaral rule in the periodicity observed. That is why we decided to juxtapose the results obtained for the proteins from the small ribosomal subunit (S-group) with those for the proteins from the large ribosomal subunit (L-group). The proteins IF-1, IF-3, EF-Tu, NS1 and NS2 were included in the S-group on the basis of the data on their role in protein biosynthesis (39).

The mean periodicity curves for S-group (25 proteins) and L-group (34 proteins) show a certain similarity in shape (Figure). At the same time considerable differences are observed. For the small distances (1 to 3 residues) the L-group curve possesses a decreasing character and the S-group curve - an increasing one. For distances from 1 to 5 residues only the L-group curve shows an alternative character (regular alternation of neighbouring maxima and minima). A detailed examination of the individual periodicity curves again manifested a stronger alternativity for the L-group. All the values for the L-group curve lie higher than the corresponding ones for the S-group: the mean percentage of periodicity for the L-group is higher (72.1) compared to that for the S-group (64.7). The cases of relatively statistically significant periodicity (difference above 4% - see the Methods) are almost equally


IF-1(4)
IF-3I (5)
EF-Tu (6)
NS1 (7)
NS2 (7)
S1 (8)
S2 (9)
S3 (10)
S4 (11)
S5 (12)
S6 (13)
S7 K (14)
S8 (13)
S9 (13)
S10 (15)
S11 (16)
S12 (13)
S13 (13)
S15 (13)
S16 (13)
S17 (17)
S18 (13)
S19 (18)
S20 (13)
S21 (13)
L1 (19)
L2 (20)
L3 (21)
L4 (22)
L5 (13)
L6 (23)
L7 (11)
L8 (13)
L9 (24)
L10 (13)
L11 (25)
L12 (11)
L13 (26)
L14 (27)
L15 (28)
L16 (13)
L17 (29)
L18 (13)
L19 (30)
L20 (31)
L21 (32)
L22 (33)
L23 (34)
L24 (35)
L25 (13)
L26 (13)
L27 (13)
L28 (36)
L29 (30)
L30 (13)
L31 (37)
L32 (13)
L33 (13)
L34 (13)
Percent identity
Distance

Figure

Figure. Periodicity curves of 59 ribosomal proteins from E. coli. S-group: initiation factors - IF-1, IF-3ℓ; elongation factor EF-Tu; ribosome-associated DNA-binding proteins - NS1, NS2; ribosomal proteins from the small subunit - S1-S13, S15-S21; L-group: ribosomal proteins from the large subunit - L1-L34. Citation numbers are in brackets. Symbol + denotes the values forming the mean periodicity curve for 100 proteins of same composition and length generated from the corresponding protein. M - mean periodicity curve of the proteins examined. C - mean periodicity curve of 38 proteins, representatives of 18 superfamilies. S - mean periodicity curve of S-group ribosomal proteins. L - mean periodicity curve of L-group ribosomal proteins. See text for details.

Table. Comparison of S- and L-group ribosomal proteins

a	b	c	d	e	f	g	h	i	j	k	l	m	n
S	25	64.7	63.6	0	30.5	17.4	7.1	10.3	6.0	71.4	35.7	62.5	3.0
A	314	-	-	-	27.1	17.0	8.4	8.6	11.2	-	-	-	-
L	34	72.1	58.1	100	31.2	20.0	7.9	12.1	5.4	88.9	88.9	76.9	3.2

a: S - group of the small subunit E. coli ribosomal proteins together with proteins IF-1, IF-3, EF-Tu, NS1, NS2; A - average protein amino acid composition (40); L - group of the large subunit E. coli proteins; b: number of proteins; c: mean percentage of periodicity (%); d: cases of relatively statistically significant periodicity (%); e: cases of especially statistically significant periodicity (%); f: total percent of Gly, Ala, Pro, Arg; g: total percent of Gly and Ala; h: percent of Gly; i: percent of Ala j: total percent of Phe, Tyr, Trp, Cys; k: correspondence of the relative statistical significance with a total percent of Gly, Ala, Pro, Arg above that of the average protein (%); l: the same as k for the total percent of Gly and Ala (%); m: correspondence of the "unsufficient" statistical significance with a total percent of Gly and Ala below 18.5 (%); n: mean extent of coincidence (42).

distributed in both L- and S-group (Table, d). But in all 6 cases of especially significant periodicity (difference above 10%) corresponding proteins belong to the L-group.

In this situation it was logical to extend the comparison on amino acid composition level. We directed our attention to the amino acids of extreme share (max or min) in proteins. On one hand Gly, Ala, Pro, and Arg were chosen. We had in mind not only their average percent in contemporary proteins (40): 8.4, 8.6, 5.2, and 4.9 correspondingly (relatively small for Pro and Arg), but also the correspondence between the share of amino acids in proteins

and that of corresponding codons in the genetic code (41). The codons of the amino acids with highest C + G content: Gly, Ala, Pro, Arg possess 4, 4, 4, 6 positions in the genetic code and a considerable total average percent in contemporary proteins: 27.1. On the other hand attention was paid to the rearly destributed Phe, Tyr, Trp, and Cys, whose average percent in contemporary proteins (40) is 3.6, 3.4, 1.3, and 2.9 correspondingly, and their total percent is 11.2.

On this basis the cases of relatively statistically significant periodicity for the two groups were examined. The combination of statistical significance and a total percent of Gly, Ala, Pro, and Arg above 27.1 was considered as a correspondence. The same was performed with the total percent of Gly and Ala for the critical value of 17.0. The results show that the correspondence for group L is considerably better (Table, k, 1). Moreover, for this group it does not change during the transition from Gly + Ala + + Pro + Arg to Gly + Ala. For the S-group not only the correspondence is weaker, but also it becomes worse after above transition. Therefore, more clearly expressed periodicity of the L-group is due mostly to Gly and Ala. This was confirmed also by an examination of the remaining cases when the difference is less than 4% (see the Methods). In the majority of cases both for L- and S-group the total percent of Gly and Ala is around and below that for the average protein (less or equal to 18.5 - Table, m). So, it follows again that the ribosomal protein periodicity is closely connected with the share of Gly and Ala. The average percent of Gly + Ala + Pro + Arg for both L- and S-group exceeds that for the average protein and is about 30 (Table, f), while this of Gly + + Ala for the L-group (20.0) is sufficiently higher than that for the S-group (17.4). The same is true for the individual components (Table, h, i). As to the average pervent of Phe + Tyr + + Trp + Cys for the L-group it is smaller (5.4) than that for the S-group (6.0), but in both two cases it is considerably smaller than that for the average protein.

In a previous paper (42) we grounded the evolutionary relationship of the E. coli ribosomal proteins. It is known that the position of the root (the point of earliest divergence) in a phylogenetic tree is difficult to be established without additional information (43). The results show a higher extent of coincidence for the L-group ribosomal proteins (Table, n). Besides, in 5 out of 6 cases of especially statistically significant periodicity the extent of coincidence is very high. It is reasonable to consider that in a large group of proteins these showing highest extent of coincidence with the remaining ones are closer to the evolutionary ancestor. Therefore, the L-group ribosomal proteins possessing more clearly expressed periodicity are more conservative. This is in a conformity with the previously established stronger periodicity for nodal ancestor proteins, and also

with the data that the most conservative contemporary proteins possess most clearly expressed periodicity. Besides, it is difficult to accept the conversion from aperiodic to periodic structure as a general principle having in mind our knowledge about the mechanism of evolutionary changes (1). Then, the arrow from L-group proteins to S-group proteins should coincide with the evolutionary direction. So, far after some extrapolation we arrive at the following conclusion:

ancient proteins	Evolution →	contemporary proteins
extreme composition		more uniform composition
great share of universal amino acids		smaller share of universal amino acids
small share of specialized		greater share of specialized
clearly expressed periodicity		weakly expressed periodicity
non-specific functions		apecific functions

example:
→ L-group ribosomal proteins
→ S-group ribosomal proteins

Indeed, as far as it is known, the S-group ribosomal proteins are more deeply involved in protein biosynthesis than L-group ribosomal proteins. In our opinion, for a relatively short period the S-group ribosomal proteins have undergone more changes connected with the new function than L-group ones. Then, most probably, a functional conservatism corresponding to the importance of the function performed should have predominated. A reason for such a judgement is the relatively preserved extreme amino acid composition in both two groups.

As earlier by searching for an evolutionary relationship (42) and now, by investigation of the periodicity of objects with exclusive role in protein biosynthesis, we have reached the same conclusion about the trend in molecular evolution of proteins from universal to specialized amino acids. The extreme extrapolation of this trend towards to the past reveals that ancient proteins were strongly periodic and contained small number of universal amino acids.

REFERENCES

(1) Ivanov, O. Ch. and Ivanov, Ch. P. 1980, J. Mol. Evol. 16, pp. 47-68. (2) Ivanov, Ch.P. and Ivanov, O.Ch. 1973, Compt. rend. Acad. bulg. Sci. 26, pp. 1641-1644. (3) Ivanov, O. Ch. and Ivanov, Ch. P. 1976, In "Protein Structure and Evolution" (eds J. L. Fox, Z. Deyl and A. Blažej), pp. 413-428. M. Dekker. (4) Pon, C. L., Wittmann-Liebold, B. and Gualerzi, C. 1979, FEBS Lett. 101, pp. 157-160. (5) Brauer, D. and Wittmann-Liebold, B. 1977, FEBS Lett. 79, pp. 269-275. (6) Jones, M. D., Petersen, T. E., Nielsen, K.M., Magnusson, S., Sottrup-Jensen, L., Gausing, K. and Clark, B. F. C. 1980, Eur, J. Biochem. 108, pp. 507-526. (7) Mende, L., Timm, B.

and Subramanian, A. R. 1978, FEBS Lett. 96, pp. 395-398. (8) Schnier, J., Kimura, M., Foulaki, K., Subramanian, A.R., Isono, U. and Wittmann-Liebold, B. 1982, Proc. Natl. Acad. Sci. USA 79, pp. 1008-1011. (9) Wittmann-Liebold, B. and Bosserhoff, A. 1981, FEBS Lett. 129, pp. 10-16. (10) Brauer, D. and Röming, R. 1979, FEBS Lett. 106, pp. 352-357. (11) Hunt, L.T. and Dayhoff, M.O. 1976, In "Atlas of Protein Sequence and Structure" (ed. M.O. Dayhoff), Vol. 5, Suppl. 2, pp. 225-232. Nat. Biomed. Res. Found. (12) Wittmann-Liebold, B. and Greuer, B. 1978, FEBS Lett. 95, pp. 91-98. (13) Hunt, L. T., Schwartz, R. M. and Dayhoff, M.O. 1978, In "Atlas of Protein Sequence and Structure" (ed. M.O. Dayhoff), Vol. 5, Suppl. 3, pp. 251-264. Nat. Biomed. Res. Found. (14) Reinbolt, J., Tritsch, D. and Wittmann-Liebold, B. 1978, FEBS Lett. 91, pp. 297-301. (15) Yaguchi, M., Roy, C. and Wittmann, H.G. 1980, FEBS Lett. 121, pp. 113-116. (16) Kamp, R. and Wittmann-Liebold, B. 1980, FEBS Lett. 121, pp. 117-122. (17) Yaguchi, M. and Wittmann, H.G. 1978, FEBS Lett. 87, pp. 37-40. (18) Yaguchi, M. and Wittmann, H.G. 1978, FEBS Lett. 88, 227-230. (19) Brauer, D. and Öchsner, I. 1978, FEBS Lett. 96, pp. 317-321. (20) Kimura, M., Mende, L. and Wittmann-Liebold, B. 1982, FEBS Lett. 149, pp. 304-312. (21) Muranova, T. A., Muranov, A. V., Markova, L.E. and Ovchinnikov, Yu.A. 1978, FEBS Lett. 96, pp. 301-305. (22) Kimura, M. and Wittmann-Liebold, B. 1980, FEBS Lett. 121, pp. 317-322. (23) Chen, R., Arfsten, U. and Chen-Schmeisser, U. 1977, Hoppe-Seyler's Z. Physiol. Chem. 358, pp. 531-535. (24) Kamp, R.M. and Wittmann-Liebold, B. 1982, FEBS Lett. 149, pp. 313-319. (25) Dognin, M.J. and Wittmann-Liebold, B. 1980, Eur. J. Biochem. 112, pp. 131-151. (26) Mende, L. 1978, FEBS Lett. 96, pp. 313-316. (27) Morinaga, T., Funatsu, G., Funatsu, M., Wittmann-Liebold, B. and Wittmann, H. G. 1978, FEBS Lett. 91, pp. 74-77. (28) Giorginis, S. and Chen, R. 1977, FEBS lett. 84, pp. 347-350. (29) Rombauts, W., Feytons, V. and Wittmann-Liebold, B. 1982, FEBS Lett. 149, pp. 320-327. (30) Brosius, J. and Arfsten, U. 1978, Biochemistry 17, pp. 508-516. (31) Wittmann-Liebold, B. and Seib, C. 1979, FEBS Lett. 103, pp. 61-65. (32) Heiland, I. and Wittmann-Liebold, B. 1979, Biochemistry 18, pp. 4605-4612. (33) Wittmann-Liebold, B. and Greuer, B. 1980, FEBS Lett. 121, pp. 105-112. (34) Wittmann-Liebold, B. and Greuer, B. 1979, FEBS Lett. 108, pp. 69-74. (35) Wittmann-Liebold, B. 1979, FEBS Lett. 108, pp. 75-80. (36) Wittmann-Liebold, B. and Marzinzig, E. 1977, FEBS Lett. 81, pp. 214-217. (37) Brosius, J. 1978, Biochemistry 17, pp. 501-508. (38) Dayhoff, M.O., Barker, W.C. and Hunt, L.T. 1976, In "Atlas of Protein Sequence and Structure" (ed. M. O. Dayhoff), Vol. 5, Suppl. 2, pp. 9-19. Nat. Biomed. Res. Found. (39) Liljas, A. 1982, Prog. Biophys. molec. Biol. 40, pp. 161-228. (40) Dayhoff, M. O., Hunt, L. T. and Hurst-Calderone, S. 1978, In "Atlas of Protein Sequence and Structure" (ed. M. O. Dayhoff), Vol. 5, Suppl. 3, pp. 363-373. Nat. Biomed. Res. Found. (41) King, J. L. and Jukes, T. H. 1969, Science 164, pp. 788-798. (42) Ivanov, O. Ch. 1983, J. Mol. Evol. (in press). (43) Penny, D. 1976, J. Mol. Evol. 8, pp. 95-116.