COMPARISON OF AMINO ACID SEQUENCES AMONG RIBOSOMAL PROTEINS OF *ESCHERICHIA COLI*

B. WITTMANN-LIEBOLD and M. DZIONARA
Max-Planck-Institut für Molekulare Genetik, Berlin-Dahlem, Germany

Received 17 November 1975

1. Introduction

E. coli ribosomes consist of three RNA species and 55 proteins, all of which have been isolated and characterized by chemical, physical and immunological methods (for recent reviews see [1–3]). From immunochemical studies with antisera against individual proteins from both ribosomal subunits, it was concluded [3–5] that there are no *extensive* sequence similarities among the ribosomal proteins with the exception of two pairs of proteins, namely L7/L12 and S20/L26.

Much more detailed information concerning possible homologous regions among the ribosomal proteins can be obtained from investigations on the primary structures of these proteins. Studies in our laboratory, and in collaboration with other groups, have so far resulted in the sequence determination of more than 3300 of the approx. 8000 amino acid residues present in the *E. coli* ribosome. This paper summarizes the results obtained by systematic searching for identical regions among these *E. coli* ribosomal proteins for which sequences are available. Identical regions consisting of three to six amino acids were found to be present within these proteins with a much higher frequency than could be expected on a random basis. However, no identical regions longer than six amino acids occur among the proteins investigated, except in the pairs S20/L26 and L7/L12.

2. Materials and methods

2.1. *Nomenclature*

The amino acids in the figures are abbreviated according to the one-letter code proposed by the IUPAC-IUB Commission on Biochemical Nomenclature [6].

2.2. *Compared sequences*

The amino acid sequences from a total of 50 ribosomal proteins (table 1) were used for the comparison in this paper. So far the primary structures of 14 proteins have completely been determined and those of a further 8 proteins are almost complete. For the remaining proteins, partial sequences, mainly obtained from sequenator degradations, are available (table 1).

2.3. *Searching for identical regions*

For the sequence comparison described in this paper the DEC system 10 (Digital Equipment Corporation), located at the Fritz-Haber-Institute of the Max-Planck-Gesellschaft in Berlin-Dahlem, was used.

3. Results and discussion

The number of amino acids whose sequences were determined in our laboratory and in collaboration with other groups amounts to more than 3300, i.e. approx. 40% of the total number of amino acid residues present in both subunits of *E. coli* ribosomes. The comparison based on these data can, therefore, be regarded as representative for the whole situation a final analysis of which clearly has to await the complete determination of the primary structures of all the *E. coli* ribosomal proteins.

North-Holland Publishing Company – Amsterdam

Table 1
Data used for the sequence comparison of E. coli ribosomal proteins

Protein	Sequenced amino acids	Sequence completed	Ref.	Protein	Sequenced amino acids	Sequence completed	Ref.
S1	20	–	29	L5	40	–	28
S2	35	–	27	L6	27	–	29
S3	48	–	27	L7	120	+	13
S4	203	+	7	L9	46	–	29
S5	160	(+)	21	L10	46	–	29
S6	135	+	8	L12	120	+	13
S7	32	–	2	L13	46	–	28
S8	109	+	9	L14	55	–	29
S9	128	+	10	L15	46	–	29
S10	42	–	2	L16	25	–	29
S11	32	–	27	L17	35	–	28
S12	110	(+)	27,22	L18	117	+	14
S13	71	–	27,30	L19	16	–	29
S14	36	–	27	L20	50	–	28
S15	85	(+)	27,22	L21	45	–	29
S16	56	(+)	27,23	L22	60	–	28
S17	97	(+)	27,24	L23	50	–	28
S18	74	+	11	L24	55	–	27,28
S19	72	(+)	27,24	L25	94	+	15,16,28
S20	88	(+)	27,25	L26	43	–	29
S21	70	+	12	L27	84	+	17,28
L1	47	–	28	L29	63	+	18,28
L2	40	–	28	L30	58	+	19
L3	32	–	28	L32	56	+	20
L4	38	–	29	L33	50	(+)	26,29

+ : sequence completed
(+) : sequence almost completed
– : sequence partially known

3.1. *Identical tripeptides*

Systematic searching for identical regions comprising three amino acids ('tripeptides') among the proteins listed in table 1 resulted in a high number of tripeptides (approx. 600) most of which occur in two proteins. However, a few tripeptides, e.g. Lys–Ala–Gly, are present in up to nine proteins and further in some of the proteins, the identical tripeptides are located at the same amino acid positions of the protein chain. An example of the latter situation is provided by proteins S6 and L17, whose sequences both begin with Met–Arg–His, whereas this sequence was not found at all in any other protein. The same is true for the N-terminal tripeptide Ala–Thr–Val in proteins S2 and S12, as well as for Ala–Lys–Leu in proteins L1 and L5. However, the occurrence of tripeptides at almost identical positions of two protein chains is not limited to the N-terminal regions; for instance the tripeptide Glu–Ala–Gly is present in positions 112–114 of proteins L7/L12 and of protein L18 (as well as occurring in positions 125–127 of protein S6).

The probability for the tripeptides to occur in the various proteins by chance varies between 10^{-1} to 10^{-3}, i.e. a given tripeptide would be expected to occur on a random basis in one out of 10–1000 proteins of the same amino acid composition and length as the protein under consideration. The probability that two identical tripeptides are located by chance at identical (or almost identical) positions of two protein chains is even lower (see below).

Table 2
Tetrapeptides occurring in three proteins

peptide	protein positions		protein positions		protein positions	
Lys–Ala–Ala–Val	S8	21 - 24	L22	42 - 45	L23	44 - 47
Gly–Lys–Val–Lys	L14	50 - 53	L24	22 - 25	L27	64 - 67
Ala–Lys–Phe–Val	S7	23 - 26	S13	60 - 63	L1	35 - 38
Ile–Arg–Glu–Lys	S4	142 - 145	L1	10 - 13	L33	4 - 7
Val–Glu–Lys–Ala	S17	92 - 95	S19	18 - 21	L18	54 - 57
Leu–Arg–Lys–Ala	S9	97 - 100	S9	117 - 120	L10	40 - 43

3.2. *Identical tetrapeptides*

The search for identical tetrapeptides among the proteins listed in table 1 gave the following results: each of 97 tetrapeptides was found in two or more proteins. 91 of them are present in only two proteins (data not shown) whereas six tetrapeptides occur in three proteins (table 2).

By considering the amino acid positions at which the tetrapeptides are located in the various proteins, it was found that each of 16 tetrapeptides is present at an almost identical position in at least two proteins. The peptide Lys–Ala–Ala–Val occurs not only at almost identical positions in two proteins, namely positions 42–45 of protein L22 and positions 44–47 of L23, but also in a third protein, S8, although in this case at a different position (table 2).

One can calculate the probability for each of the 97 identical tetrapeptides to occur *on a random basis* in the proteins in which they were found. These probabilities are different for each tetrapeptide, since they depend on the number of residues and the amino acid composition of the protein in which a given tetrapeptide is found. The probabilities which were calculated for each of the 97 tetrapeptides (data not shown here) range from 10^{-2} to 10^{-4}, with a few exceptions. The value for a given peptide represents the probability that a tetrapeptide with this sequence occurs by chance and in any position within the protein concerned. The probability that the tetrapeptide is located at a specified position of the protein chain is lower than the value given above, by a factor which depends on the length of the protein. It is between 20 and 200 for the proteins listed in table 1.

The probability that a given tetrapeptide occurs independently in two (or more) proteins by chance is the product of the probabilities for its occurrence in each of the two (or more) proteins. This calculation will be illustrated for the tetrapeptide Lys–Ala–Ala–Val which was found in three proteins. The calculated probability for its occurrence in protein S8 is 8.6×10^{-3}, in protein L22 is 2.1×10^{-2} and in protein L23 is 2.1×10^{-2}. Therefore the probability that this tetrapeptide occurs by chance in all three proteins is only 3.8×10^{-6}. This very low value becomes even smaller with the additional restriction that this tetrapeptide occurs in two proteins, namely L22 and L23, at almost identical positions (table 2). Calculations for the random occurrence of other tetrapeptides in three proteins gave similar results.

3.3. *Identical pentapeptides*

Each of ten pentapeptides was found in two proteins as listed in table 3, which also gives the probabilities for the occurrence of the pentapeptides on a random basis. A calculation for the pentapeptide Asp–Asp–Ala–Glu–Ala, similar to that made above for the tetrapeptide Lys–Ala–Ala–Val, results in a probability of 2.6×10^{-6} for the random occurrence of this pentapeptide in two proteins. Here again the chances are very low that the experimental finding can be explained as a random phenomenon.

3.4. *Identical hexapeptide*

The largest identical region found in two proteins is the hexapeptide Val–Val–Ala–Asp–Ser–Arg present in proteins L10 and S16. The

Table 3
Occurrence of pentapeptides

peptide	protein	$p/10^{-3}$	protein	$p/10^{-3}$
Asp–Asp–Ala–Glu–Ala	S6	1.2	L7/L12	2.2
Thr–Val–Lys–Gly–Gly	S5	1.1	S9	0.3
Leu–Gly–Leu–Arg–Arg	S15	1.3	L30	0.3
Ala–Ala–Ala–Ala–Val	L7/L12	40.0	L18	9.4
Ala–Val–Ile–Lys–Ala	L7/L12	4.8	L20	1.2
Val–Ile–Arg–Glu–Lys	S4	0.5	L1	0.1
Lys–Ser–Val–Glu–Glu	S5	0.4	L29	0.2
Glu–Leu–Arg–Lys–Ala	S9	0.6	L10	0.4
Val–Glu–Lys–Ala–Val	S17	0.7	S19	0.3
(Asp)–Gly–Lys–Val–Lys	L14	0.3	L27	0.6

p: Probability of finding the peptide in any position of the protein.

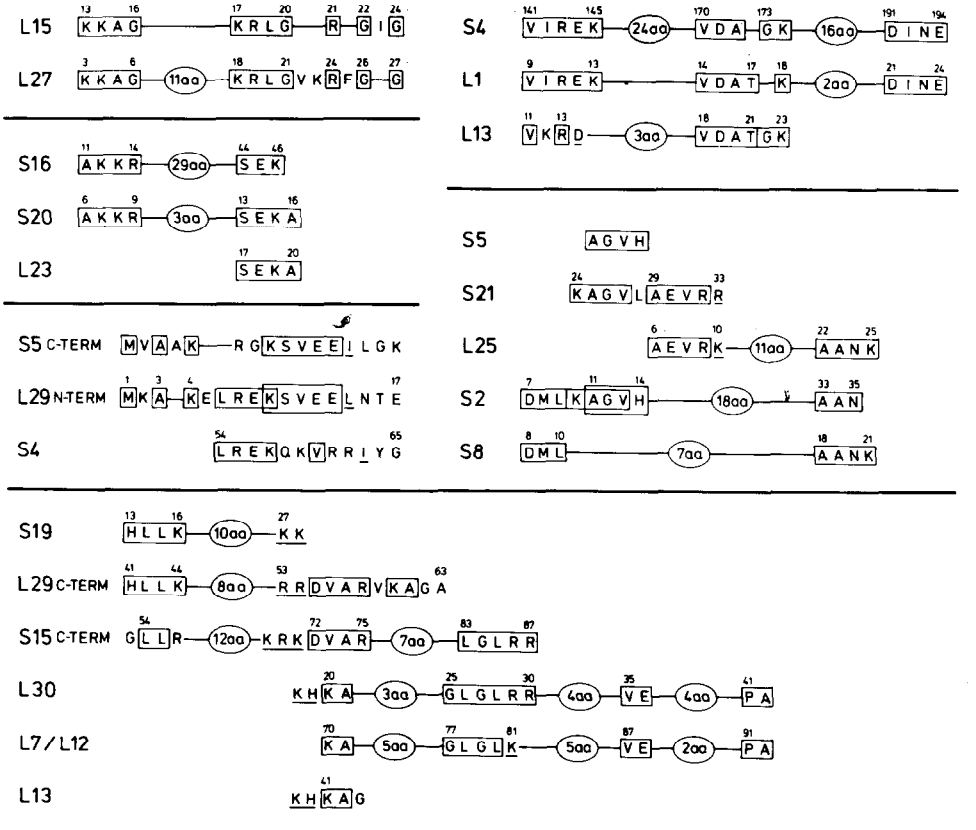

Fig.1. Similarities in the structures of *E. coli* ribosomal proteins. The amino acids are abbreviated according to [6]. Structurally related amino acids in corresponding positions are underlined, e.g. I (Ile) and L (Leu) or K (Lys), R (Arg) and H (His) or D (Asp) and E (Glu).

probability that this peptide occurs in the two proteins by chance is extremely low, namely 9.1×10^{-10}. This value does not even take into account the fact that the hexapeptide occupies similar positions in the amino acid sequence of the two protein chains.

3.5. *Patterns of similarity*

As mentioned above, identical regions up to a size of six amino acids were found in pairs of proteins. In some of these protein pairs, identical or similar regions occur at several points. The most pronounced example is the pattern of similar regions among proteins S6 and L7/L12 as illustrated in [8]. Additional, although weaker, patterns of similarity are shown in figs.1 and 2. These patterns can in many cases only be arranged if one 'shortens' or 'stretches' the intervening non-homologous regions. It is interesting that structure similarities exist between the N-terminal and C-terminal regions of some proteins as illustrated in fig.1.

It is difficult at the present time to make correlations between the structural similarities among ribosomal proteins on the one hand and other known properties of these proteins, e.g. involvement in certain functions of the ribosomes, on the other. Nevertheless, there is no doubt that similarities in the primary structure of some proteins, as shown in this paper, occur at much higher frequency than can be expected on a random basis.

Considering the various factors which determine whether a protein region is recognized as an antigenic determinant (minimal size, amino acid composition, spatial structure, individual specifity, etc.; for a review see reference [31] one could expect that

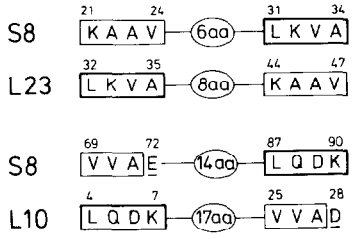

Fig.2. Identical peptides in two pairs of proteins, in which the positions of the peptides along the protein chain are inversed.

the degree of sequence homology described here would lead to only rare, if any, cross-reaction among most ribosomal proteins. This is in agreement with the results from immunochemical studies [3–5].

Acknowledgement

We are grateful to Dipl. Chem. W. Oppermann for his very helpful advice in the computer work.

References

[1] Wittmann, H. G. (1974) in: Ribosomes, p. 93–114 Cold Spring Harbor Monograph Series, New York.
[2] Wittmann, H. G. and Wittmann-Liebold, B. (1974) in: Ribosomes, p. 115–140 Cold Spring Harbor Monograph Series, New York.
[3] Stöffler, G. (1974) in: Ribosomes, p. 615–647 Cold Spring Harbor Monograph Series, New York.
[4] Stöffler, G. and Wittmann, H. G. (1971) Proc. Nat. Acad. Sci. USA 68, 2283–2287.
[5] Stöffler, G. and Wittmann, H. G. (1971) J. Mol. Biol. 62, 407–409.
[6] IUPAC-IUB-Commission on Biochemical Nomenclature (1968) J. Biol. Chem. 243, 3557–3559.
[7] Schiltz, E. and Reinbolt, J. (1975) Eur. J. Biochem. 56, 467–481.
[8] Hitz, H., Schäfer, D. and Wittmann-Liebold, B. (1975) FEBS Lett. 56, 259–262.
[9] Stadler, H. (1974) FEBS Lett. 48, 114–116.
[10] Chen, R. and Wittmann-Liebold, B. (1975) FEBS Lett. 52, 139–140.
[11] Yaguchi, M. (1975) FEBS Lett. 59, 217–220.
[12] Vandekerckhove, J., Rombauts, W., Peeters, B. and Wittmann-Liebold, B., Hoppe-Seyler's Z., in press.
[13] Terhorst, C. P., Möller, W., Laursen, R. and Wittmann-Liebold, B. (1973) Eur. J. Biochem. 34, 138–152.
[14] Brosius, J., Schiltz, E. and Chen, R. (1975) FEBS Lett. 56, 359–361.
[15] Dovgas, N. V., Markova, L. F., Mednikova, T. A., Vinokurow, L. M., Alakhov, Y. B., Ovchinnikov, Y. A. (1975) FEBS Lett. 53, 351–354.
[16] Bitar, K. G. and Wittmann-Liebold, B. (1975) Hoppe-Seyler's Z. 356, 1343–1352.
[17] Chen, R., Mende, L. and Arfsten, U. (1975) FEBS Lett. 59, 96–99.
[18] Bitar, K. G. (1975) Biochim. Biophys. Acta 386, 99–106.
[19] Ritter, E. and Wittmann-Liebold, B. (1975) FEBS Lett. 60, 153–155.
[20] Wittmann-Liebold, B., Greuer, B. and Pannenbecker, R., (1975) Hoppe-Seyler's Z. physiol. Chemie in press.

[21] Wittmann-Liebold, B., Greuer, B. and Lehmann, A., unpublished.
[22] Funatsu, G., unpublished.
[23] Rombauts, W. and Vandekerckhove, J., unpublished.
[24] Yaguchi, M., unpublished.
[25] Wittmann-Liebold, B., Marzinzig, E., Greuer, B. and Lehmann, A., unpublished.
[26] Wittmann-Liebold, B. and Pannenbecker, R., unpublished.
[27] Wittmann-Liebold, B. (1973) FEBS Lett. 36, 247–249.
[28] Wittmann-Liebold, B., Geissler, A. W. and Marzinzig, E. (1975) J. Supramolecular Structure, Vol. 3, issue 5/6, in the press.
[29] Wittmann-Liebold, B., Geissler, A. W., Graffunder, H. and Marzinzig, E., unpublished.
[30] Lindemann, H., unpublished.
[31] Arnon, R. (1971) in: Current Topics in Microbiology and Immunology (Springer-Verlag, Heidelberg) Vol. 54, p. 47–93.