STUDIES ON THE SIGNIFICANCE OF SEQUENCE HOMOLOGIES AMONG PROTEINS FROM *ESCHERICHIA COLI* RIBOSOMES

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

Received 5 April 1976

1. Introduction

Sequence investigations of the ribosomal proteins from *E. coli* have led to the finding that these proteins possess unique primary structures without extensive regions of homology [1,2]. Only two protein pairs (L7/L12 and S20/L26) are a notable exception to this. Among some of the proteins, however, certain short identical or similar regions were found (see [2] for a summary).

In a previous paper [3] the amino acid sequences of the *E. coli* ribosomal proteins and protein fragments known at the time were compared and searched systematically for short identical regions. A large number of identical tripeptides and tetrapeptides as well as several pentapeptides and hexapeptides were found in at least two different proteins. For example, proteins S16 and L10 contain the peptide Val–Val–Ala–Asp–Ser–Arg.

This poses the question whether a similar number of identical regions as observed in ribosomal proteins can also be found among unrelated proteins (e.g. hemoglobin, lysozyme, ribonuclease) or among artificial proteins which have the same amino acid compositions and lengths as the ribosomal proteins but whose amino acid sequences are randomly generated by a computer program. In this paper we report the results of such a comparison.

2. Materials and methods

2.1. *Ribosomal proteins*

The amino acid sequences of 50 *E. coli* ribosomal proteins listed in table 1 of our previous paper [1], the complete sequence of protein L34 [4] and the N-terminal region of L28 (unpublished results) were included in this comparison.

2.2. *Unrelated proteins*

The primary structures of the following proteins were taken from the literature [5,6]: coat protein of RNA phage fr; histone III from calf thymus; myoglobin from sperm whale; β-chain of human hemoglobin; tobacco mosaic virus coat protein, strain vulgare; turnip yellow mosaic virus coat protein; bovine ribonuclease A; nuclease from *Staphylococcus aureus;* hen egg lysozyme; tryptophan synthetase from *E. coli* (up to position 250); horse heart cytochrome; kappa-chain of human immunoglobulin; bovine chymotrypsinogen; papain; subtilisin from *Bacillus subtilis* BPN′ (up to position 250); phospholipase A from honey bee; trypsin inhibitor from soy bean; ferredoxin from *Scenedemus;* hemerythrin from Sipunculid worm; B-chain of bovine insulin; bovine adrenocorticotropin; mouse epidermal growth factor.

2.3. *Artificially generated proteins*

52 artificial protein sequences, having the same chain lengths and amino acid compositions as those of the native ribosomal proteins, were generated by means of a computer program which distributed the amino acids in each protein randomly. In this manner, twelve different sets of the 52 artificial proteins were obtained.

3. Results and discussion

The occurrence of the various identical peptides which were found in at least two different native ribosomal proteins is presented in table 1. About 25%

Table 1
Frequency of identical peptides within three groups of proteins

Peptides	Rib.	Lit.	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	$\bar{M}$
Hepta-	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0.08
Hexa-	3	0	2	0	1	0	1	1	0	0	0	0	2	1	0.7
Penta-	11	4	6	6	10	6	10	5	8	8	7	6	5	7	7
Tetra-	86	58	71	82	89	88	81	84	85	85	68	75	86	100	83
Tri-	657	603	642	631	674	645	640	677	664	672	680	663	657	633	656

The table gives the numbers of identical peptides within the following three groups of proteins: 'Rib.' = *E. coli* ribosomal proteins; 'Lit.' = unrelated proteins whose sequences were taken from the literature; I–XII: twelve sets of artificially generated proteins (see text for details). The average values ($\bar{M}$) of I–XII are given in the last column.

of these identical regions were found to be situated in identical or almost identical positions in the protein chains (data not shown).

Three approaches were made to assess the 'randomness' of occurrence of such identical peptides. In the first approach, the actual compositions of the identical peptides were compared with the most probable compositions of peptides which could be derived from the parent proteins. The probabilities for the occurrence of the peptides were calculated on the basis of the amino acid compositions of the ribosomal proteins. In most cases, the identical sequences found were not of the same compositon as those most frequently expected (data not shown).

In the second approach, the ribosomal proteins were compared for identical peptides with the group of unrelated proteins whose amino acid sequences were taken from the literature. In both groups of proteins the total number of amino acids was the same. A lower number of identical peptides was observed among the sequences of the unrelated proteins than among the ribosomal proteins (table 1). This was especially true for the number of pentapeptides and hexapeptides. These differences may be expected when the characteristic differences in the amino acid compositions of proteins in both groups are examined. Ribosomal proteins are much richer in basic amino acids [7] than are the unrelated proteins taken from literature. Furthermore, both groups of proteins also differ considerably in other amino acids (data not shown). The more equal the distribution of different amino acids is in a given set of proteins, the less identical peptides can be expected.

In the third approach, we compared native ribosomal proteins with artificially generated proteins which have the same amino acid compositions and lengths as ribosomal proteins but whose amino acid sequences were randomly distributed by a computer program. The randomization was done twelve times (table 1; I–XII) and each time for a number of proteins and amino acids identical with the native ribosomal proteins. The results show that about the same numbers of identical tripeptides and tetrapeptides occur in the artificially generated proteins as in the native ribosomal proteins while pentapeptides and hexapeptides were found less frequently. In one set (number IX) of the artificial proteins an identical heptapeptide was obtained, to the exclusion of any hexapeptides from that set. Similarly to the native ribosomal proteins, a high number of identical peptides were found in almost identical positions in the artificially generated protein chains (data not shown).

In conclusion, the frequency pattern of the occurrence of identical peptides is similar for the native ribosomal proteins and their artificially generated isomers. From the results presented in this paper, there is no indication of strong structural homologies among the so far known sequences of *E. coli* proteins. As already mentioned, the only exceptions are the protein pairs L7/L12 and S20/L26 [2,8].

Acknowledgements

We are grateful to Drs B. Jacrot, C. G. Kurland and S. Robinson for helpful discussions.

References

[1] Wittmann-Liebold, B. (1973) FEBS Lett. 36, 247–249.
[2] Wittmann, H. G. and Wittmann-Liebold, B. (1974) in: Ribosomes, Monograph Series, Cold Spring Harbor Laboratory, p. 115–140.
[3] Wittmann-Liebold, B. and Dzionara, M. (1976) FEBS Lett. 61, 14–19.
[4] Chen, R. and Ehrke, G. (1976) FEBS Lett. 63, 215–217.
[5] Dayhoff, M. O. (1972) in: Atlas the Protein Sequence and Structure. Nat. Biomed. Res. Found., Washington, USA.
[6] Croft, L. R. (1973) in: Handbook of Protein Sequences, Joynson-Bruvvers Ltd., Oxford, England.
[7] Kaltschmidt, E., Dzionara, M. and Wittmann, H. G. (1970) Mol. Gen. Genet. 109, 292–297.
[8] Stöffler, G. (1974) in: Ribosomes, Monograph Series, Cold Spring Harbor Laboratory, p. 615–667.