POSSIBLE STRUCTURAL HOMOLOGIES AMONG 30S RIBOSOMAL PROTEINS

Lawrence I. Slobin

Section of Microbiology, Division of Biological Sciences

Cornell University, Ithaca, New York

Received March 23, 1970

SUMMARY

A comparison of the amino acid compositions of 30S ribosomal proteins was made using data already published. Similar comparisons were made for a variety of different cytochrome C's as well as for a group of thirty-five functionally unrelated proteins. On the basis of these comparisons it is concluded that some of the 30S ribosomal proteins are likely to possess considerable similarities in amino acid sequence.

Recently several laboratories have reported analyses on the purified protein components of the 30S ribosomal subunit from E. coli (1-4). Moore et al. in Geneva (1) have obtained thirteen different 30S ribosomal proteins in purities between 80 and 95%. Hardy et al. in Madison (3) report that the same 30S subunit contains 21 proteins which are chromatographically and electrophoretically unique. Both groups have characterized their purified components by amino acid composition, molecular weight, and tryptic peptide maps. The Geneva group concluded (1) that "by all criteria, these proteins differ greatly in primary structure." The Madison group found (4) that "detailed comparisons of the... proteins reveal that with the exception of proteins 5 and 9, all have unique peptide maps. This taken with the amino acid compositions...suggests that we have succeeded in purifying 19 unique proteins."

Given the complexity of ribosomal protein structures (perhaps a total of 50 different proteins in an intact 70S ribosome) it would be desirable to assess whether there are certain groups of these proteins which are structurally related. Unfortunately, sequence data is not yet available. Other possible methods for comparing ribosomal proteins,

such as molecular size, tryptic fingerprints and functional criteria, all have serious deficiencies. Recently Metzger et al. (5) have compared proteins on the basis of amino acid composition. The rationale for this approach is that although a small number of sequence changes might exert a profound influence on protein structure and function, it might be expected that the overall composition of evolutionarily related proteins would remain quite similar. I have applied the technique employed by Metzger et al. to the available data on the amino acid composition of ribosomal proteins. The results indicate that several of the ribosomal proteins show a strong degree of compositional relatedness. Comparison with a similar analysis of families of proteins known to be evolutionarily related indicates that it is highly likely that some of the 30S ribosomal proteins possess considerable similarities in amino acid sequence.

The method of analysis (5) is based on a tabulation of the mole fraction of each amino acid per 10^5 g of protein. Two proteins are compared by determining the absolute value of the difference in the fractional content of each amino acid, summing the differences, and multiplying the sum by 50 to obtain a difference index (DI). Two proteins with identical composition have a DI of zero, two totally unrelated polypeptides, e.g., polylysine and polyalanine, have a DI of 100. Most nonidentical protein pairs have a DI between 1 and 50; the smaller the DI the greater the degree of compositional similarity. An analysis of n proteins gives $(n^2-n)/2$ comparisons or DI's.

The matrix comparing the 21 different 30S ribosomal proteins found by the Madison group was obtained with the aid of a computor program (Table 1). Similar matrices were obtained from the data of Moore et al. A comparison was also made for the two sets of data on the amino acid composition of 30S ribosomal proteins. The data is plotted in the form of histograms, giving number of pairs having a DI between two integer

EVALUATION OF DIFFERENCE INDEX (DI) FOR TWENTY-ONE DIFFERENT RIBOSOMAL PROTEINS

	16	15a	15	14	13	12b	12a	12	11	10	9	8	7	6	5	4a	4	3	2a	2	1
1	26.2	18.1	18.5	18.0	14.5	16.9	20.1	12.5	19.3	11.3	11.5	14.3	15.0	13.1	10.4	8.9	17.4	12.9	9.5	20.5	0
2	22.5	20.7	27.4	23.3	22.3	19.6	21.6	21.2	27.5	20.2	18.3	17.1	24.7	17.8	19.3	18.3	16.3	21.0	19.8		
2a	20.7	13.8	17.1	18.0	12.0	11.5	14.5	11.1	18.1	9.7	5.0	9.7	13.6	10.4	4.5	8.6	17.0	9.9			
3	23.6	18.7	16.5	22.7	15.6	16.9	21.3	14.0	12.9	16.8	12.3	14.6	10.5	10.6	11.0	12.6	22.2				
4	26.4	19.7	21.5	17.1	17.7	18.8	18.1	17.7	24.2	15.4	15.4	15.9	20.9	18.1	14.4	18.1					
4a	21.2	14.4	19.1	15.7	10.1	11.5	15.0	12.0	16.9	10.8	9.9	13.0	13.1	9.8	9.3						
5	20.5	10.8	13.9	16.6	9.8	11.0	12.4	9.7	15.4	9.3	3.1	10.3	12.0	8.8							
6	17.8	12.9	16.8	17.7	12.8	11.4	15.4	11.5	15.6	11.7	8.6	11.3	10.8								
7	24.3	14.8	16.2	15.6	13.7	16.9	20.4	11.5	6.4	15.3	13.2	18.1									
8	20.4	11.1	15.7	15.2	15.6	7.4	14.3	12.0	20.5	10.1	9.5										
9	19.5	8.9	13.8	16.0	10.8	9.0	11.8	9.2	17.3	8.1											
10	22.6	11.7	15.7	11.5	11.8	9.6	13.6	8.3	19.2												
11	30.2	16.3	15.2	19.3	13.6	19.0	24.6	16.3													
12	24.7	11.6	12.9	14.6	12.5	11.8	12.3														
12a	21.2	13.6	19.1	17.2	17.1	12.1															
12b	17.8	8.7	16.4	14.6	11.2																
13	22.2	12.6	11.3	14.7																	
14	26.3	14.1	17.9																		
15	29.6	11.3																			
15a	23.2																				
16	0																				

TABLE I. Amino acid composition data were taken from Craven et al.(4). DI values less than ten are underscored.

values as a function of DI (Figure 1a and 1b). In order to compare the DI values of ribosomal protein pairs with a family of proteins known to be evolutionarily related, a similar matrix was generated for various cytochrome C's. These results are given in histogram form in Figure 2, along with a plot of the matrix generated by Metzger et al. in a comparison of thirty-six unrelated proteins.

In order to focus on pairs of proteins having similar amino acid compositions, I have compiled the number of pairs (and percent of the total) of proteins in each figure with DI values less than 10. The results are given in Table 2. Although the number 10 is an arbitrary cut-

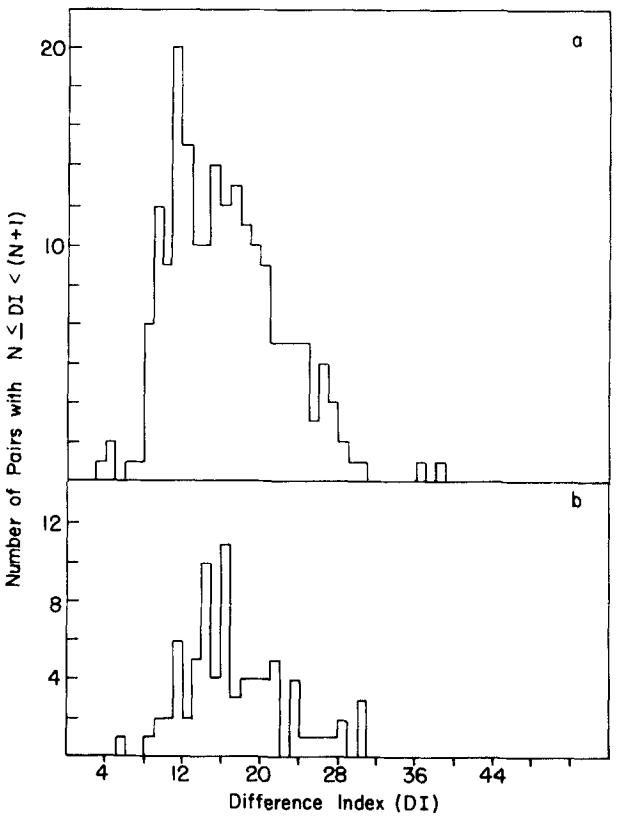

Fig. 1. Plot of number of protein pairs with DI's between two integer values $(N \leq DI < (N + 1)$ as a function of DI

(a) 30S ribosomal proteins. Data taken from Craven et al. (4).
(b) 30S ribosomal proteins. Data taken from Moore et al. (1).

TABLE II. Number and percent of total protein pairs having a DI of less than 10 for different families of proteins.

Proteins compared	Number of comparison with DI 10	Percent of total comparisons
30S ribosomal proteins[a]	27	12.9
30S ribosomal proteins[b]	4	5.1
Unrelated proteins[c]	3	0.48
Cytochrome C[d]	67	43.8

[a] Data taken from Figure 1a
[b] Data taken from Figure 1b
[c] Data taken from Figure 2a
[d] Data taken from Figure 2b

off point, it may be seen that only 3 out of 628 comparisons tabulated for ostensibly unrelated proteins (Figure 2a) gave DI values less than 10. None of the DIs for the unrelated protein comparisons was less than 9.

An examination of Figures (1) and (2) reveals that, whereas the histogram for unrelated proteins (2a) is approximately gaussian in appearance (with a medium DI of about 24), all of the other figures appear rather unsymmetrical with a decided skewing to the left, and median DI's of less than 20. This deviation from a random pattern might be expected for families of cytochrome C's. It further justifies the conclusion that the overall composition of evolutionarily related proteins can be expected to remain quite similar. The unsymmetrical distribution for ribosomal proteins, however, comes as somewhat of a surprise.

Using the data of the Madison group, it may be seen that 27 pairs (out of 210 comparisons) have DI values less than 10 (Figure 1a). In addition, 15 out of the 21 proteins are part of pairs with DI values less than 10. Some proteins such as 3 and 12a figure in only one pair with a DI less than 10; others such as 2a, 5 and 9 figure in 7 or more such comparisons. Omitting all proteins that figure in only one comparison (with DI less than 10) leaves a total of 11 proteins (1, 2a, 4a, 5, 6,

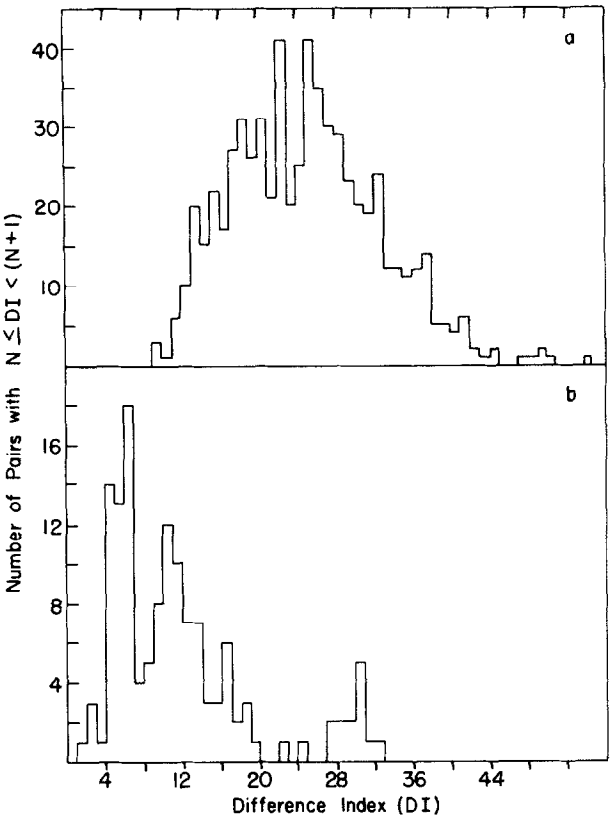

Fig. 2. Plot of number of protein pairs with DI's between two integer values $(N \leq DI < (N+1))$ as a function of DI.

(a) Comparison of unrelated proteins. Data taken from (5). Proteins compared include tryptophan synthetase, bovine; ribonuclease, bovine; chymotrypsinogen A, bovine; carboxypeptidase A, bovine; papain, papaya; pepsin, bovine; insulin, bovine; insulin, α chain; insulin, β chain; glucagon, bovine; haemoglobin, α chain, human; haemoglobin, β chain, human; apoferritin, horse; cytochrome c, horse; β-lactoglobulin AB, bovine; lactalbumin A, bovine; lysozyme, hen egg; thyroglobulin, bovine; hyaluronidase, bovine; DNase, bovine; ATP-creatinetransphosphorylase, rabbit; glycogen phosphorylase, human; leucine aminopeptidase, porcine; α-glycoprotein-Zn, human; avidin, hen egg yolk; tobacco mosaic virus; trypsinogen, bovine; myosin, rabbit; actin, rabbit; serum albumin, human; trypsin inhibitor, bovine; carbonic anhydrase B, human; biothrombin, bovine; fibrinogen, bovine; α-ACTH, bovine; γ G-light chains, human. Comparisons between human haemoglobin α and β chains and bovine fibrinogen and biothrombin were omitted from the tabulation.

(b) Cytochrome c. Data taken from (8). Cytochrome c from the following species were compared: human, Rhesus monkey, horse, pig, dog, rabbit, kangaroo, chicken, Pekin duck, snapping turtle, rattlesnake, tuna fish, moth, neurospora, bakers yeast, Candida krusei, and Pseudomonas fluorescens.

8, 9, 10, 12, 12b, 15a) that have a high degree of compositional relatedness to at least two other proteins in a 30S ribosome. It might be

pointed out here that only proteins 5 and 9 show a striking similarity in their peptide maps.[1/]

A comparison of both sets of composition data indicates that every protein with the exception of number 3, that was found by Moore et al. has a compositional counterpart in the proteins found by the Madison team, i.e., for every protein in the one group there is at least one matchup in the other, with a DI less than 7.5 for the related pair. However, there are many proteins (7, 11, 2A, 12B, 15, 15A and 16) that have been analyzed by the Madison team that do not have compositional relatives in the data obtained by Moore et al. The Geneva group indicates that approximately 20% of the protein material from a 30S ribosome was not subjected to analysis. They suggest that these proteins comprise large numbers of species present in the 30S mixture in small amounts. In addition, as has been already pointed out (4), there are serious discrepancies in the molecular weight data obtained by the two groups, for individual 30S proteins.

Whatever the reasons for these discrepancies, it has been shown (6) that two minor components (4 and 12b) are absolutely required for the in vitro activity of reconstituted ribosomes. One of these proteins, 12b, is related to four other components with a DI less than 10; one of these related components (number 8) is a major and unambiguously unique protein. Thus one of the closely comparable pairs involves two proteins which are almost certain to be functionally important.

It has been suggested, because of the large number of different ribosomal proteins, the sum of whose weight adds up to about twice the amount that can be accommodated to a 30S particle, that ribosomes are

---

[1/] Craven et al. (6) states that "the striking similarities between proteins 5 and 9 taken with the variable quantities of protein 9, have suggested that one of these proteins may be an enzymatically derived fragment of the other or...both of a third portion that we have not yet fully recovered."

heterogeneous. If this is true it may be speculated that among the protein pairs that share low DI values are some that are structurally homologous, each participating uniquely in a different 30S particle. These homologous pairs presumably arose as a result of gene duplication and subsequent evolution. In this regard it is known that several 30S ribosomal proteins map in a single region of the E. coli genome (7).

Obviously only direct sequence determinations can firmly establish some of these suggestions. However, it is hoped that the compositional comparisons presented will call attention to the good possibility that some 30S ribosomal proteins are structurally related. If subsequent sequence analysis should validate this proposal, it would provide further evidence for the utility of the method of compositional comparison for predicting the existence of structurally homologous proteins.

ACKNOWLEDGEMENT. I would like to thank Dr. R. E. MacDonald for calling my attention to some of the problems of ribosome structure, and for many stimulating discussions on this topic. I am also indebted to Mr. Paul Goldberg for the writing and execution of the computor programs as well as for aid in the tabulation of the data.

This work was sponsored by a research grant from the National Institutes of Health, U. S. Public Health Service (AI-08242-02).

REFERENCES

1. Moore, P. B., Traut, R. R., Moller, H., Pearson, P. and Delius, H., J. Mol. Biol. 31, 441 (1968).

2. Fogel, S. and Sypherd, P.S., Proc. Natl. Acad. Sci. U.S. 59, 1329 (1968).

3. Hardy, S. J. S., Kurland, C. G., Voynow, P. and Mora, G., Biochem. 8, 2897 (1969).

4. Craven, G. R., Voynow, P., Hardy, S.J. S., and Kurland, C. G., Biochem. 8, 2906 (1969).

5. Metzger, H., Shapiro, M. B., Mosimann, J. E. and Vinton, J. E., Nature 219, 1166 (1968).

6. Traub, P., Hosokawa, K., Craven, G. R., and Nomura, M., Proc. Natl. Acad. Sci. U. S. 58,2430 (1967).

7. Guthrie, C., Nashimoto, H. and Nomura, M. Proc. Nat. Acad. Sci. U.S. 63, 384 (1969).

8. Eck, R. U. and Dayhoff, M. O., Atlas of Protein Sequence and Structure, Nat. Biomed. Res. Foundation, Silver Spring, Maryland, 1966.