Proc. Natl. Acad. Sci. USA
Vol. 75, No. 7, pp. 3128–3132, July 1978
Biochemistry

Use of computerized multidimensional scaling to compare immunoelectron microscopy data with protein near-neighbor information: Application to the 30S ribosome from *Escherichia coli*

(ribosome model/ribosomal proteins/multidimensional scaling/computer model)

PHILIP T. GAFFNEY AND GARY CRAVEN

Laboratory of Molecular Biology and Department of Genetics, University of Wisconsin, Madison, Wisconsin 53706

Communicated by C. B. Anfinsen, April 18, 1978

ABSTRACT **A three-dimensional model of the protein arrangement in the *Escherichia coli* 30S ribosome was constructed by using computerized multidimensional scaling of immunoelectron microscope data. This enabled data comparison between the new electron microscope technique and other methods such as crosslinking, chemical protection, affinity labeling, energy transfer, and assembly interactions. The immunoelectron microscopy data are reasonably consistent with those from other sources. Reasons for some inconsistent data are discussed and our calculation of the dimensions of the proteins, both globular and elongated, are summarized.**

One of the most interesting recent developments in the analysis of ribosome structure is the use of immunoelectron microscopy (immuno-EM) to determine the relative positions of the individual proteins on the surface of the subunit particle (1–5). In these experiments, dimers between ribosomal subunits are constructed by using antibodies bivalent for a single ribosomal protein. The location of the antigenic site on the surface of the ribosome can be visualized directly by electron microscopy using negative staining techniques. Because of the characteristic shape of the 30S ribosome, all protein antigenic sites can be oriented relative to special topographical features of the particle. Recently, the mapping of all 21 of the 30S ribosomal proteins has been completed (6).

Many other techniques have been used to gather information on the spatial relationships of the proteins in the ribosome. These include chemical crosslinking of neighboring proteins (e.g., refs. 7 and 8), chemical protection (9), fluorescent energy transfer (10, 11), affinity labeling (12–17), and neutron scattering (18). The question arises as to how the results of immuno-EM compare with those obtained by other methods. This study makes that comparison with the aid of a new computer modeling technique.

Outline of approach

Most of the techniques that have successfully produced valuable information about the spatial relationships of the ribosomal proteins have yielded data about protein pairs whose members are inferred to be close to each other within the ribosome structure. Thus, in order to make a comparison between these data and the immuno-EM data, it is necessary to generate a three-dimensional model with the proteins represented by volumes reflective of their actual polypeptide chains with definite dimensions rather than mere antigenic sites on the surface. With such a model it would be possible to calculate the nearest distances between all 210 pairs of the 21 ribosomal proteins.

We have built such a model by using the following two-step procedure. First, a three-dimensional plaster model was made based on the immuno-EM diagrams of Tischendorf *et al.* (1, 2) and the antigenic sites were marked on its surface. Second, the shortest distances between all pairs of these antigenic sites were measured and used as input for a multidimensional scaling program. Its output gives the three-dimensional coordinates for a "naked" protein model—one without the outline of the ribosome. The model provides a unique and insightful view of the 30S ribosome. The uneven distribution of proteins within the subunit and the elongation of 12 of them are strikingly apparent. Questions on the relationship of structure to function take on fresh meaning but, more importantly, the model allows us to make a systematic comparison of all the relevant published data regarding protein–protein relationships.

Computer program

Multidimensional scaling is a statistical technique that has been used for approximately 15 years in psychology and is now being used extensively in other fields, including an early attempt to generate a three-dimensional model of the 30S ribosome (19). It constructs a configuration of points in space from information about the distances between points. For example, if a matrix were made of the 703 distances in miles between any 38 cities in the United States, and used as input in a multidimensional scaling program, the output would be 38 points representing the cities in the correct two-dimensional configuration. The orientation (north, south, east, west) would be left for the subjective decision of the user. Missing data and tied data can be accommodated but present some problems.

The program places N points in a space of a given dimension so as to minimize STRESS, which measures the badness-of-fit between the configuration of points and the data. It represents the extent to which the data deviate in the least squares sense from the monotonic curve (see Fig. 3 *A* and *B*). The program finds the minimizing configuration by starting with some configuration (rational, arbitrary, or random), moving all the points to decrease the STRESS, and then iterating this procedure over and over again until no better fit can be made. Typically 15–50 iterations may be required.

We used the programs KYST and MINISSA 1. KYST offers many options, is portable, and has a simple input procedure. It is available on request from Bell Laboratories (20). MINISSA 1 can be obtained from James C. Lingoes, Department of Statistics, University of Michigan, Ann Arbor (21, 22). Its output gives a most convenient matrix of the distances between points in the model as well as reproducing the input matrix. MINISSA 1 substitutes the mean for any missing data whereas KYST merely ignores missing data. This latter approach is preferable when there is reason to believe that the data are not missing in a random way. Both programs include mechanisms to prevent the final configuration's being a merely local minimum.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviation: immuno-EM, immunoelectron microscopy.

The use of multidimensional scaling in this study was straightforward. All distances between a large number of points (38 antibody sites) were included in the input matrix. There was no concern over missing data, and the STRESS was 0.07%, indicating a near-perfect monotonic relationship between the input distances and the distances in the model (20). It should be emphasized that this very low STRESS value suggests not that our "naked" protein ribosome model is "right" in an absolute sense but that it accurately reflects the immuno-EM data as expected.

Description of model

Fig. 1 shows four views of our three-dimensional model of the 30S ribosome as derived from the multidimensional scaling analysis of the 38 antigenic sites found by Tischendorf and coworkers (1, 2, 6). Some small changes were made. Although Tischendorf *et al.* (2) found a single site for S19, two sites were demonstrated by Lake and Kahan (4). In immuno-EM studies, positive data must be given preference, so two sites were used and S19 is seen as an elongated protein.

The model in Fig. 1 has been given the dimensions 220 × 135 × 110 Å. These are a nice compromise between the estimations of Tischendorf *et al.* (maximum length, 180–200 Å) (2) and Lake (maximum length, 240–260 Å) (23) and are the dimensions given by Hill *et al.* (24) from x-ray scattering in dilute solutions.

Calculation of protein dimensions

The dimensions of ribosomal proteins can be calculated if their volumes are known. The hydrated volumes can be estimated by calculating the "dry volume," $M\bar{V}/N_o$, and adding a reasonable amount for the volume of the water of hydration. If the hydration is w g of water per g of protein and the density of the water of hydration is assumed to be the same as that of normal water, then the volume of the water of hydration is wM/N_o. Assuming a value of 0.3 g of water per g of protein for w (25), the total volume of the hydrated protein is given by $V = M\bar{V}/N_o + 0.3\,M/N_o$.

The shortest distances between antigenic sites on the model give the minimum length of elongated proteins. Their diameters, assumed constant, are determined by the formulae

$$V = \pi r^2 l = \frac{M\bar{V}}{N_o} + 0.3\frac{M}{N_o}.$$

The globular ribosomal proteins were considered to be spherical and their radii were calculated from their molecular weight by the formulae

$$V = \frac{4}{3}\pi r^3 = \frac{M\bar{V}}{N_o} + 0.3\frac{M}{N_o}$$

in which V = volume, M = molecular weight, $\bar{V}$ = partial specific volume, N_o = Avogadro's number, and l = length of protein.

The molecular weights of the proteins were obtained from two sources. Values from amino acid sequence data were used for S4, S6, S8, S9, S12, S13, S15, S16, S18, S20, and S21 (references in Table 1). The average of sedimentation equilibrium and sodium dodecyl sulfate gel determinations (26) was used for the remaining proteins.

The length of some proteins is surprising: S4, S7, S15, and S18 measure more than 200 Å. S18 consists of 74 amino acids (6) and is 240 Å long in the immuno-EM model. This implies an unusual secondary structure, because its maximum extension in α helical form is 111 Å. The secondary structure of S15 must also be unusually extended for similar reasons. However, if the antigenic sites have been correctly identified, the exceptionally extended nature of these proteins is inescapable.

The distribution of protein mass within the total ribosome

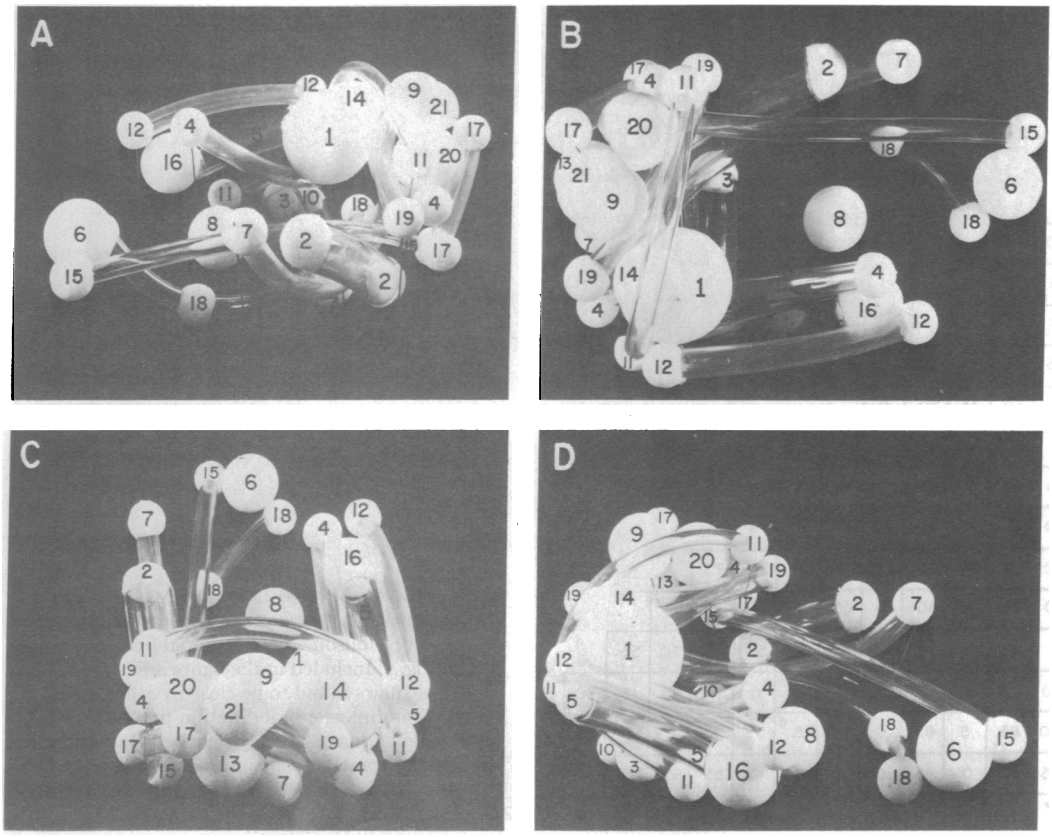

FIG. 1. Four different views of our model of the immuno-EM data. Proteins with more than one antigenic site are represented by styrofoam balls connected with plastic tubing. The dimensions of the balls and tubing closely reflect the dimensions of the actual proteins calculated from their established molecular weights. (*B*) Corresponds to the front view of the Tischendorf, Zeichardt, and Stöffler model (6).

Table 1. Shapes and sizes of the 30S ribosomal proteins derived from our immuno-EM model

Protein	Shape*	Length, Å	Diameter, Å	M_r × 10^{-4}	Ref.
S1	Globular		65	6.60	26
S2	Elongated	90	29	2.80	26
S3	Elongated	75	32	2.85	26
S4	Elongated	225	17	2.26	6
S5	Elongated	80	27	2.06	26
S6	Globular		40	1.50	6
S7	Elongated	205	17	2.25	26
S8	Globular		37	1.21	6
S9	Globular		39	1.46	6
S10	Elongated	85	21	1.42	26
S11	Elongated	180	16	1.61	26
S12	Elongated	125	17	1.36	27
S13	Globular		38	1.30	28
S14	Globular		39	1.45	26
S15	Elongated	210	11	1.00	6
S16	Globular		34	0.92	29
S17	Elongated	60	22	1.05	26
S18	Elongated	240	10	0.90	6
S19	Elongated	115	18	1.34	26
S20	Globular		34	0.96	6
S21	Globular		32	0.78	6

* A protein is termed "globular" solely on the basis of its having only one antigenic site on the ribosome (see text).

structure is interesting: 70% of the total protein mass resides in the upper portion of the ribosome model. This suggests that the bulk of the RNA must be organized in the lower half of the ribosome which contains only a small proportion of the protein.

Data comparison of distance between neighboring proteins

Various studies have been done to determine which proteins are near neighbors in the 30S ribosome. Fig. 2 and Table 2 contain 63 protein pairs which have been shown to be close by five different types of experiments. The short distance (0–15 Å) between the members of a pair is determined by crosslinking, chemical protection, fluorescent energy transfer, and affinity labeling. Comparison with the corresponding distance in our immuno-EM model is done in two ways. The computer printout gives the distances between all antigenic sites on the surface of our model in relative units. These are converted to angstroms and then adjustments are made by using the radii of proteins to determine the distances between their true centers (middles if elongated). The path taken by elongated proteins is also taken into consideration. It may pass very close to another protein even though its antigenic sites are at some distance from it (Fig. 1). This direct method is used to obtain distances between the edges of all proteins in our immuno-EM model (Fig. 2). Those pairs found to be "near neighbors" in previous studies are listed separately in Table 2.

A more indirect method of comparing the 63 protein "near neighbors" with their positions in our model makes use of the computer program in an unusual way. The 63 short distances are used in the input matrix in place of the corresponding distances in the immuno-EM model. A clear graphical representation of the data most in conflict is provided in the output of the KYST program. In Fig. 3 *upper*, the deviation of points from the monotonic line indicates the discrepancy between the immuno-EM data and the 63 bits from other sources. Fig. 3 *lower* shows the nearly perfect monotonic fit of the distances in the output model with the immuno-EM data alone.

This method is used only as a rough guide because it uses the

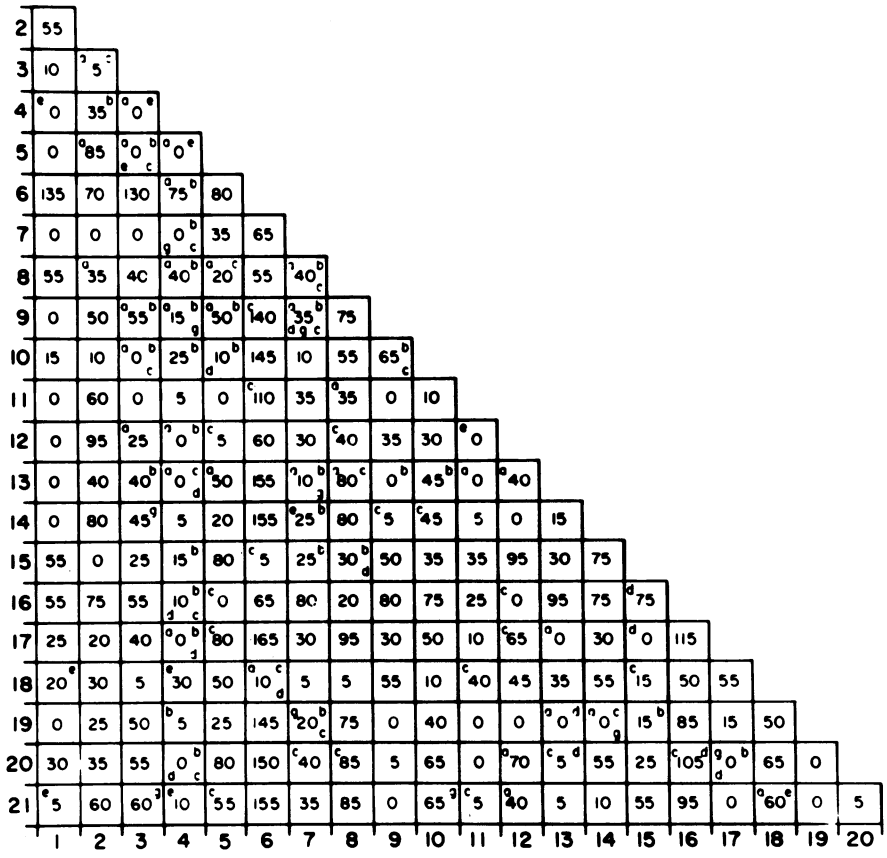

FIG. 2. Matrix summarizing all the nearest distances (Å) between the 210 protein pairs of the 30S ribosomes as calculated from the model in Fig. 1. Included in the figure are notations for those pairs found to be close by other methods. a, Crosslinkage; b, iodination protection; c, interdependent assembly relationships; d, singlet energy transfer; e, affinity label; g, trypsin protection. References for these proposed neighbor relationships are given in Table 2.

Table 2. Distances in our immuno-EM model between ribosomal proteins determined to be close by crosslinking, chemical protection, fluorescence energy transfer and affinity labeling

Protein pairs	Ref.	Separation, Å	Protein pairs	Ref.	Separation, Å
S1-S4	12	0	S7-S19	9	20
S3-S4	7	0	S3-S12	8	25
S3-S5	7, 8, 17	0	S4-S10	9	25
S3-S10	6, 7, 8	0	S7-S14	9, 15	25
S4-S5	6, 7, 20	0	S7-S15	9	25
S4-S7	8, 9	0	S4-S18	12, 14	30
S4-S12	7, 8, 18	0	S8-S15	9, 11	30
S4-S13	7, 11	0	S2-S4	9	35
S4-S17	7, 8, 11	0	S2-S8	7, 8	35
S4-S20	8, 11	0	S7-S9	7, 9, 11	35
S9-S13	8	0	S8-S11	8	35
S11-S12	12	0	S3-S13	9	40
S11-S13	7	0	S4-S8	8, 9	40
S13-S17	7	0	S7-S8	8, 9	40
S13-S19	6, 7, 11	0	S12-S13	8	40
S14-S19	7, 9	0	S12-S21	8	40
S15-S17	11	0	S3-S14	9	45
S17-S20	8, 9, 11	0	S10-S13	9	45
S1-S21	15	5	S5-S9	8, 9	50
S2-S3	6, 7	5	S5-S13	8	50
S4-S19	8	5	S3-S9	8, 9	55
S13-S20	11	5	S3-S21	9	60
S4-S16	9, 11	10	S18-S21	8, 15	60
S4-S21	13	10	S9-S10	9	65
S5-S10	9, 10	10	S10-S21	9	65
S6-S18	7, 8, 11	10	S12-S20	8	70
S7-S13	8, 9	10	S4-S6	8, 9	75
S4-S9	8, 9	15	S15-S16	11	75
S4-S15	9	15	S8-S13	8	80
S15-S19	9	15	S2-S5	7, 8	85
S1-S18	15	20	S16-S20	11	105
S5-S8	7, 8	20			

distances between antigenic sites rather than the more accurate distances between the centers or middles of the proteins. However, it does illustrate one of the powerful, convenient options of the multidimensional scaling program.

The data from the "near neighbor" experiments agree reasonably well with those from the electron microscope: 68% are in agreement with the EM results ($P = 0.02$) and only 10% are in definite disagreement. Twenty-two percent of the near-neighbor data are neither supported nor contradicted by the immuno-EM information. These classifications, which are somewhat arbitrary, are shown in Table 3.

There is a considerable degree of experimental uncertainty involved in identifying the position of a protein by the immuno-EM antibody-labeling technique. This is why the pair distances from 40 to 65 Å cannot be classified as "definitely not close" in the model. This error arises from many sources: the precise location of the combining site on the antibody is unknown; there is difficulty in determining the exact orientation of the ribosome with respect to its image in the electron micrograph; and there is doubt about which part of a protein bears the antigenic site. Unfortunately, there are no published data on the actual degree of experimental uncertainty. This is required to give a proper perspective to the protein maps and the model presented here.

Table 3. Summary of results given in Table 2 for all 66 protein pairs

Separation	Classification	No. of pairs	%
≤20 Å	Strong agreement	33	52
25–35 Å	Fair agreement	10	16
40–65 Å	Neither agreement nor disagreement	14	22
≥70 Å	Disagreement	6	10

There is also some degree of experimental uncertainty involved in the many types of experiments that show that 63 protein pairs are "close." This has been taken into consideration in finding the agreement with the immuno-EM studies.

It is interesting to take a detailed look at the immuno-EM data that disagree most with those obtained from other techniques and to propose reasons for the discrepancies.

Immuno-EM gives the distance between proteins S16 and S20 as 105 Å, and that between S16 and S15 as 75 Å (Table 2). Energy transfer studies (11) show S16 or S17 to be very close to S20 and close to S15. Therefore, it seems likely that S17, but not S16, is very close to S20 and close to S15. On examining the immuno-EM model for confirmation, S17 is seen to "touch" both S20 and S15.

Proteins S2 and S5 can be crosslinked (7, 8) but are 85 Å apart in the immuno-EM model. It could be that S5 has been placed on the wrong side of the 30S ribosome by an error in interpretation of the electron micrographs. A similar interpretational

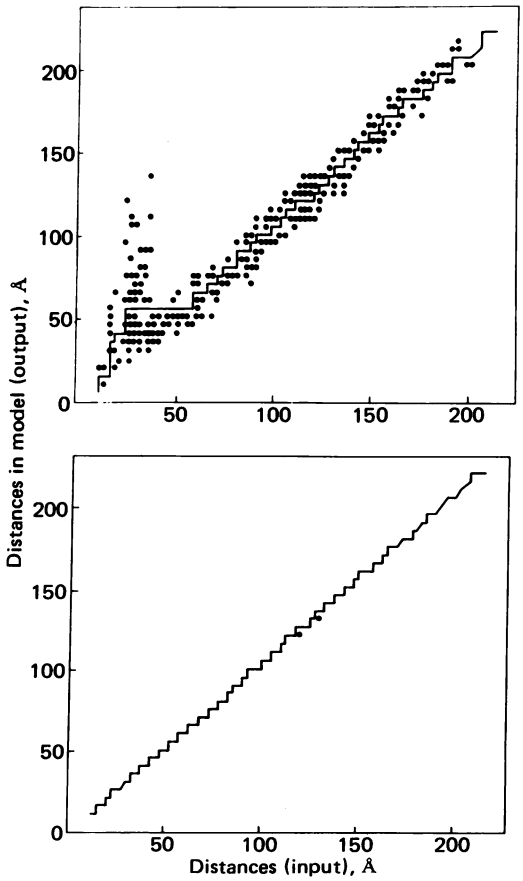

FIG. 3. Graphical representation of the computer program analysis of the input data versus the output information derived from the best-fit model. (*Upper*) Obtained from input data that included the 63 near-neighbor relationships as well as the immuno-EM relationships. (*Lower*) Obtained from input data by using the immuno-EM relationships alone. The solid circles represent the data that do not fall precisely on the monotonic line. In *Upper*, the vertical deviation of these points from the monotonic line indicates the discrepancy between immuno-EM data and those from other sources.

error would account for the 70-Å separation in the model between the crosslinked proteins S12 and S20 (8). If S5 and S12 were on the other side of the model, the S2-S5 and S12-S20 protein pairs would be within 15 Å. The distance of S5 and S12 from other "close" proteins would remain about the same.

The 80 Å between the crosslinked pair S8-S13 is difficult to explain without suggesting that S13 is elongated. Neutron diffraction experiments establish S8 as globular (18). Crosslinking, chemical protection, and energy transfer models show that S13 is close to at least 12 proteins in the 30S subunit. Several of these pairs are less than 10 Å apart in our immuno-EM model. The distance between the members of the other five pairs would generally be lessened if S13 were elongated (see Table 2 and model).

Crosslinking and chemical protection establish S4 as close to S6 (8). The reason they are 75 Å apart in our immuno-EM model is that S4 follows a straight line between antigenic sites B and C. In reality, S4 would have to follow the contour around the cleft in the 30S subunit. This would make it closer to S6 and in much better agreement with the other data.

The discovery that at least 12 of the 21 30S ribosomal proteins are elongated suggests a reassessment of the results of the neutron scattering method for determining distances between ribosomal proteins (18). This technique measures the distance between the centers of mass of a pair of proteins. If they are both spherical, the information is unambiguous. However, if one or both of the proteins are quite elongated, the distance between their centers of mass may not reflect the closest distance between them. Parts of the proteins can be touching while other parts are a considerable distance apart. The separation of the centers of mass of two elongated proteins is a valuable measure only if the proteins are approximately parallel.

It has been suggested that proteins closely related in the *Escherichia coli* 30S ribosomal assembly map are actually physically close within the ribosome (30, 31). There are 34 independent assembly relationships between protein pairs in the assembly map (32). These comprise 15 major and 19 minor assembly influences. The distances between these pairs were measured on our immuno-EM model. Most (59%) are close (within 35 Å), although the correspondence with the immuno-EM data is not statistically significant ($P = 0.35$). Some (23%) are neither close nor distant and others (18%) are definitely distant (>65 Å). The individual members of the six protein pairs in this distant group behave toward each other as might be predicted, exerting only minor influences in assembly. Thus, this model is consistent with the notion that proteins that exert an interdependence during assembly are organized close to one another in the final ribosome structure.

Two groups have used immuno-EM to map the proteins on the 30S ribosome. They derived substantially different models of the 30S subunit itself and they used different methods of assigning the orientation of the ribosome with respect to its EM image in locating antigenic sites. Tischendorf *et al.* (1, 2) used the staining of a central hollow, whereas Lake and Kahan (4) used a high shoulder. This leads to discrepancies between the results of the two groups. We have confined ourselves in this paper to analyzing the data of Tischendorf and his coworkers. They have mapped all the proteins in the 30S ribosome, whereas Lake and Kahan have published the positions of only six proteins thus far.

We thank Dr. Michael Subkoviak for advice on the use of multidimensional scaling programs and Drs. John Anderegg and Lawrence Kahan for their reading of the manuscript. This work was supported by the Graduate School and the College of Agriculture and Life Sciences, University of Wisconsin-Madison, and by Research Grant GM15422 from the National Institutes of Health.

1. Tischendorf, G. W., Zeichardt, H. & Stöffler, G. (1974) *Mol. Gen. Genet.* **134,** 187–208.
2. Tischendorf, G. W., Zeichardt, H. & Stöffler, G. (1975) *Proc. Natl. Acad. Sci. USA* **72,** 4820–4824.
3. Lake, J. A., Pendergast, M., Kahan, L. & Nomura, M. (1974) *Proc. Natl. Acad. Sci. USA* **71,** 4688–4692.
4. Lake, J. A. & Kahan, L. (1975) *J. Mol. Biol.* **99,** 631–644.
5. Wabl, M. R. (1974) *J. Mol. Biol.* **84,** 241–247.
6. Stöffler, G. & Wittmann, H. G. (1977) *Molecular Mechanisms of Protein Synthesis* (Academic, New York), pp. 117–202.
7. Lutter, L. C., Kurland, C. G. & Stöffler, G. (1975) *FEBS Lett.* **54,** 144–150.
8. Sommer, A. & Traut, R. R. (1976) *J. Mol. Biol.* **106,** 995–1015.
9. Changchien, L. M. & Craven, G. R. (1977) *J. Mol. Biol.* **113,** 103–122.
10. Huang, K. & Cantor, C. R. (1975) *J. Mol. Biol.* **97,** 423–441.
11. Huang, K., Fairclough, R. H. & Cantor, C. R. (1975) *J. Mol. Biol.* **97,** 443–470.
12. Pongs, O., Stöffler, G. & Lanka, E. (1975) *J. Mol. Biol.* **99,** 301–315.
13. Pongs, O. & Rossner, E. (1976) *Nucleic Acids Res.* **3,** 1625–1634.
14. Pongs, O., Stöffler, G. & Bald, R. W. (1976) *Nucleic Acids Res.* **3,** 1635–1646.
15. Fiser, I., Sheit, K. H., Stöffler, G. & Kuechler, E. (1975) *FEBS Lett.* **56,** 226–229.
16. Girshavich, A. S., Bochkarena, E. S. & Ovchinnikov, YuA. (1976) *Bioorg. Khim.* **2,** 1073–1084.
17. Schreiner, G. & Nierhaus, K. H. (1973) *J. Mol. Biol.* **81,** 71–82.
18. Engelman, D. M., Moore, P. B. & Schoenborn, B. P. (1975) *Proc. Natl. Acad. Sci. USA* **72,** 3888–3892.
19. Bollen, A., Cedergren, R. J., Sankoff, D. & Lapalme, G. (1974) *Biochem. Biophys. Res. Commun.* **59,** 1069–1078.
20. Kruskal, J. B., Young, F. W. & Seery, J. B. (1973) *How to Use KYST, a Very Flexible Program to Do Multidimensional Scaling and Unfolding* (Bell Laboratories, Murray Hill, NJ).
21. Lingoes, J. C. (1965) *Behav. Sci.* **10,** 183–184.
22. Roskam, E. & Lingoes, J. C. (1970) *Behav. Sci.* **15,** 204–205.
23. Lake, J. A. (1976) *J. Mol. Biol.* **105,** 131–159.
24. Hill, E. W., Thompson, J. D. & Anderegg, J. W. (1969) *J. Mol. Biol.* **44,** 89–102.
25. Kuntz, I. D. & Kauzmann, W. (1974) *Adv. Protein Chem.* **28,** 239–340.
26. Wittmann, H. G. (1974) in *Ribosomes,* eds. Nomura, M., Tissieres, A. & Lengyel, P. (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY), p. 104.
27. Funatsu, G., Yahuchi, M. & Wittmann-Liebold, B. (1976) *FEBS Lett.* **73,** 12–17.
28. Lindemann, H. & Wittmann-Liebold, B. (1976) *FEBS Lett.* **71,** 251–255.
29. Vanderkerckhove, J., Rombauts, W. & Wittmann-Liebold, B. (1977) *FEBS Lett.* **73,** 18–21.
30. Mizushima, S. & Nomura, M. (1970) *Nature* **226,** 1214–1218.
31. Morgan, J. & Brimacombe, T. (1973) *Eur. J. Biochem.* **37,** 472–480.
32. Held, W. A., Ballou, B., Mizushima, S. & Nomura, M. (1974) *J. Biol. Chem.* **249,** 3103–3111.