A COMPUTER MODEL FOR THE 30S RIBOSOME SUBUNIT

IRWIN D. KUNTZ, *Department of Pharmaceutical Chemistry, University of California, San Francisco, California 94143*
GORDON M. CRIPPEN, *Department of Chemistry, Texas A & M University, College Station, Texas 77843 U.S.A.*

ABSTRACT We describe a computer-generated model for the locations of the 21 proteins of the 30S subunit of the *E. coli* ribosome. The model uses a new method of incorporating experimental measurements based on a mathematical technique called distance geometry. In this paper, we use data from two sources: immunoelectron microscopy and neutron-scattering studies. The data are generally self-consistent and lead to a set of relatively well-defined structures in which individual protein coordinates differ by ~ 20 Å from one structure to another. Two important features of this calculation are the use of extended proteins rather than just the centers of mass, and the ability to confine the protein locations within an arbitrary boundary surface so that only solutions with an approximate 30S "shape" are permitted.

INTRODUCTION

The 30S subunit from *E. coli* ribosomes has been the subject of quite a variety of experimental studies. These include immunoelectron microscopy, neutron scattering, cross-linking, fluorescence energy transfer, and direct x-ray measurements (1). A characteristic feature of these experiments is that they yield distances between pairs of "centers." Given data in this form, how does one proceed to define the "structure" of the subunit? First, there is direct model building. This method is quite straightforward when dealing with coordinates themselves (as in the immunoelectron microscopy experiment), but it is much more difficult when much distance information must be built in. The general problem of converting a set of distances into coordinates (that is, structures) does not have a simple closed-form solution if the distances are subject to experimental error. In the past few years several numerical procedures have been developed with specific application to the geometry of the 30S subunit. Among these are the triangulation and second moment methods used by Moore and Engelman (2, footnote 1) and the multidimensional scaling approach employed by Bolin (3) and most recently by Craven (4).

We have also been interested in the general geometric problem because of work on models geometry. While it has some similarity to multidimensional scaling, our approach differs from those above in that it: (*a*) accepts uncertain experimental results in the form of upper and lower bounds to the distances; (*b*) permits constraints that force all components of the subunit to lie within an arbitrary bounding surface or shape; (*c*) allows a primitive representation of

[1] Moore, P. B., and E. Weinstein. Manuscript submitted for publication.

the paired length distributions from neutron-scattering experiments; and (*d*) generates a population of structures meeting the constraints rather than a single "best" structure.

In this paper we report the results of applying our method to immunoelectron microscopy (IEM) and neutron-scattering (NS) data. We emphasize, at the onset, that model building activities of this type primarily provide self-consistency checks on the experimental data. It is always possible to add additional constraints arrived at through insight or theory, but our concern now is to examine the value of distance geometric methods in extracting the information currently available in these two sets of experimental data.

METHODS

The principal features of the distance geometry approach have been described in several publications (5–7). We represent the molecule or unit to be studied as a collection of spheres or beads of appropriate sizes. For the 30S subunit, each protein was represented by 3 beads if NS data were available and by the number of sites identified by electron microscopy if no NS results had been reported (Table I). Next, a set of upper and lower bounds for each interbead distance is proposed. Some of these distances are obtained directly from experiment. Others are only available as "self-consistent" limiting values. For instance, protein beads 7A and 18A were assigned directly to electron microscope sites of Lake and Kahan (8). Using coordinates taken from their model, the 7A,18A distance was 113 Å with an upper bound of 123 Å and a lower bound of 103 Å. The distance between protein beads 3B and 8B has not been established by direct experiment. Nonetheless, upper and lower bounds can be placed on this distance using two ideas: (*a*) no distance can be greater than some maximum set by the NS length distribution function (if available) or the largest subunit dimension, and (*b*) the upper bounds and lower bounds between all beads, taken three at a time, must obey the triangle inequality (5, 7). Using these conditions, an upper bound to the 3B,8B distance is 170 Å, and the lower bound is 0 Å.

Having established self-consistent bounds, the computer program next produces a set of trial distances randomly distributed between these bounds. The set of distances is converted to a set of three-dimensional coordinates (9) that are optimized to produce the best agreement with the upper and lower bounds. The optimization is performed on an error function, *E*, of the form

$$E = \sum_{i=1}^{n}\sum_{j=1}^{n}(D_{ij}^2 - B_{ij}^2)^2,$$

where *n* is the number of beads, *D*(*ij*) is the distance between beads *i* and *j*, and *B*(*ij*) is the appropriate bound (upper or lower) that is being violated. If *D*(*ij*) meets its boundary conditions, there is no contribution to the error function. A structure with an *E* of zero is one that meets all the constraints simultaneously. If more structures are found with zero error, the bounds are not sufficiently restrictive (that is, not sufficiently informative) to require a single conformation. In either case, there can be no claim that the program produces the "correct" structures: we can only say that the structures so generated meet the imposed constraints.

We can estimate in advance that many experimental measurements are required to provide strong constraints on a structure the size of the 30S subunit. To determine the positions of *n* spheres requires 4*n*–10 accurately known distances. As the accuracy of the data is reduced, more distances are required (6). Approximately 100 distances are likely to be needed to fix the positions of the centroids of the 21 proteins in the 30S subunit. If the structural information used as input to the program is not sufficient to determine the structure, the program produces families of conformations containing sites restricted to arcs, circles, or spherical shells instead of points. Local handedness ambiguities are also common in underdetermined or marginally determined data sets.

The mathematical approach we have described is quite general and useful for molecular modeling of many types. We now discuss several modifications of our earlier procedures that deal with special aspects of the ribosome structure problem.

TABLE I
REPRESENTATION OF 30S SUBUNIT PROTEINS

Site label	Experiment	Number of measurements (NS)
1	IEM	
2A	IEM,NS	1*
2B	IEM,NS	1*
2X	NS	1*
3A	IEM,NS	7
3B	NS	7
3X	NS	7
4A	NS	8
4B	NS	8
4X	NS	8
5A	IEM,NS	8
5B	NS	8
5X	NS	8
6A	IEM‡,NS	2*
6B	IEM‡,NS	2*
6X	NS	2*
7A	IEM,NS	5
7B	IEM,NS	5
7X	NS	5
8A	IEM,NS	7
8B	NS	7
8X	NS	7
9A	IEM‡,NS	6
9B	IEM‡,NS	6
9X	NS	6
10A	IEM,NS	6
10B	NS	6
10X	NS	6
11A	IEM,NS	6
11B	NS	6
11X	NS	6
12A	IEM,NS	6
12B	NS	6
12X	NS	6
13	IEM	
14	IEM	
15A	IEM‡,NS	1*
15B	IEM‡,NS	1*
16	IEM	
17	IEM	
18A	IEM‡	
18B	IEM‡	
19A	IEM	
19B	IEM	
20	IEM	
21	IEM	

IEM, immunoelectron microscopy; NS, neutron scattering.
*Insufficient data pairs for triangulation.
‡Large uncertainty in IEM location.

Shape Function

The 30S ribosome subunit has an irregular shape (8). We represented this by a set of cross sections at spacings of 25 Å. These cross sections are very roughly circular and were entered as 36 radial distances per plane. In some planes, the "platform" of the subunit required the use of two rough circles. Interpolative procedures were used to generate a smooth surface. The locations of all beads were checked against this surface, and if the bead was outside the surface, a term was added to the error function of the form

$$E_{\text{shape}} = \Sigma\,[R_i^2 - S(x,y,z)^2]^2,$$

where $R(i)$ is the radial distance of the bead from the centroid in the plane and $S(xyz)$ is the value of the shape function at that location. A set of 3 "outrigger" points along the x-, y-, and z-axes (10) served to map the coordinate system of the shape function onto that of the electron microscope model. A small expansion (10–20%) of the electron microscope model (8) was needed for low error solutions.

Length Distributions

The NS experiments (2, 11–13) yield spherically averaged radial distribution functions for pairs of proteins. These functions contain information about both the separation and the shapes of the proteins. We used the following method to extract some of this information: three beads were allocated to represent each protein. We distinguished between the two "outer" beads and the "center" bead, but we did not require that the three beads lie on a straight line; the center bead did not need to coincide with the center of mass of the three bead system. Nor did we use the molecular weight data that could fix the total volume of the beads (1). The use of three beads per protein gives rise to nine interprotein distances per pair of proteins. These distances were matched against three pairs of constraints taken from the length distributions.

(*i*) Close approach constraints: lower bound taken as closest approach, upper bound taken as the distance enclosing 25% of the normalized distribution. (*ii*) Peak approach constraints: lower and upper bounds taken from the peak of the distribution function with allowance for experimental errors (see below). (*iii*) Far approach constraints: lower bound taken as distance enclosing 75% of the normalized distribution; upper bound taken as distance of greatest separation.

We required that at least one of the nine distances meet the first and third constraints and that the distance between the two "center" beads of each pair of proteins meet the second constraint. The other (six) distances could have any value between the distance of closest approach and the distance of greatest separation unless additional information were available to confine them further. Note that no decision is made "ahead of time" as to which pair of beads was to meet conditions *i* or *iii*. The computer program, during the optimization phase, examined all nine interprotein distances, and, if some constraint was violated, the distance closest to meeting the constraint was selected for computation of the error function, the error gradient, and the new coordinates for the next step of the optimization.

Optimization

We encountered some difficulties in obtaining smooth optimization using constraints of the type just described with conventional optimizers such as the steepest descent and conjugate gradient methods (6, 7). We adopted, instead, an interactive optimizer that moved "one bead at a time." This is similar in spirit to the optimization procedures described by Hermans (14) and Goel (15), although there are significant differences in the detailed implementation. The algorithm we used proceeds along the steepest descent line, using analytical gradients. In practice, it is less efficient in computing time than the other optimizers we have tried, but it offered some superiority in avoiding high error local minima. Further, the optimization was performed on an in-house time-sharing computer (PDP-11/70; Digital Equipment Corp., Maynard, Mass.) so that we could inspect directly the optimization process. We found that small random displacements of the coordinates (±1 Å) were sufficient to proceed out of local minima.

DATA

We have used two data sets for the proteins of the 30S subunit: the IEM work of Lake et al. (8), and the NS results of Moore et al. (2, 11–13). For this first study we have not analyzed the important electron microscope experiments of Stoffler and Wittman (16) because of the lengths of proteins 6, 15, and 18 in their model.

Electron Microscope Data

The assignments of "beads" to the sites identified by Lake and Kahan used coordinates obtained from inspection of published diagrams (8) and a model provided by Lake. These were compared with a set of coordinates independently prepared by Lake.[2] Discrepancies were generally small (~ 20 Å) and were resolved in favor of the Lake coordinates. The coordinate bounds were initially set to ± 10 Å limits (~ 20 Å worst case uncertainty in distances) with the following exceptions: protein sites 6A,6B,15A,15B,18A, and 18B were allowed to be less extended than the original report indicated.[2] Protein sites 9A, and 9B were much less restricted (± 40 Å) because of uncertainty in location[2] and because a smaller amount of freedom was allowed for site 10A. No IEM data were used for protein 4.

Inspection of the NS data (see below) suggested several conflicts with the initial electron microscopic coordinates. These included the pairs 3-12, 6-11, 6-15, 8-9, and 9-10. Of these, the changes described above for proteins 6, 9, 10, and 15 were sufficient to meet the NS constraints for 6-11, 6-15, and 9-10. The difficulties with 3-12 and 8-9 involved the distances of closest approach. The IEM restrictions were relaxed slightly to permit the full range of NS distances. Other small (± 5 Å) relaxations were allowed during the initial phase of the investigation if some calculated coordinates were driven very close to one of the bounds.

The final set of coordinates and bounds used to represent the IEM data is given in Table II.

NS Results

As indicated earlier, the NS length distributions were represented by assigning three beads to each protein. If the protein was already described by two beads from the IEM data, the third bead was added by assuming it to lie "between" the two IEM sites. This restriction was generated by the assignment of this bead as one of the pair whose separation met the center-to-center NS distance (constraint *ii*, above), and by allowing the outer beads to be further apart than the center to outer bead upper bounds. This decision rests on the hypothesis that the IEM sites are near the ends of the proteins. It is certainly possible that some of the 30S proteins are considerably extended beyond the IEM sites. Two weak tests can be made: (*a*) the closest approach/greatest separation results clearly limit the maximal protein lengths, and (*b*) the computer program, itself, will identify mistaken assumptions of this type if they are geometrically inconsistent with the rest of the data. This issue is considered in more detail in the discussion of Table VIII.

[2] Lake, J. A. Private communication.

TABLE II
COORDINATES* AND BOUNDS‡ FROM EM DATA

Protein Site	X	Y	Z
1	−28 ± 12	207 ± 10	113 ± 12
2A	−20 ± 10	210 ± 12	40 ± 14
2B	30 ± 10	175 ± 12	66 ± 14
3A	45 ± 10	154 ± 12	30 ± 12
5A	45 ± 12	117 ± 10	35 ± 12
6A	−20 ± 20	165 ± 50	100 ± 40
6B	−14 ± 20	85 ± 25	80 ± 40
7A	−19 ± 14	230 ± 12	50 ± 12
7B	30 ± 10	195 ± 14	55 ± 10
8A	30 ± 12	163 ± 14	80 ± 12
9A	−20 ± 25	225 ± 50	60 ± 40
9B	35 ± 30	185 ± 40	60 ± 50
10A	31 ± 20	148 ± 25	−24 ± 20
11A	−30 ± 12	175 ± 12	110 ± 16
12A	30 ± 10	157 ± 16	75 ± 10
13	−20 ± 10	233 ± 12	40 ± 12
14	35 ± 10	165 ± 12	−25 ± 10
15A	−15 ± 20	170 ± 50	125 ± 40
15B	0 ± 20	30 ± 40	70 ± 40
16	−17 ± 20	220 ± 12	66 ± 12
17	25 ± 20	160 ± 12	100 ± 10
18A	−7 ± 30	150 ± 25	130 ± 40
18B	3 ± 30	23 ± 35	70 ± 30
19A	30 ± 10	150 ± 15	−25 ± 15
19B	−15 ± 10	225 ± 12	48 ± 12
20	−20 ± 10	205 ± 12	5 ± 12
21	−35 ± 12	180 ± 12	105 ± 12

*Coordinates in ångströms based on set supplied by J.A. Lake,[2] modified as described in text. The zero of the coordinate system is roughly at the lower tip of the 30S particle.

‡Errors represent bounds on coordinates. The smaller values do not correspond to experimental uncertainties, per se, but reflect relaxation of initial limits of ±10 Å, as described in text.

For those proteins where a single IEM site has been identified, and for which some NS results were available, we added two beads. One represented the "center"; the second bead represented the other "end" of the protein. These beads were allowed to position themselves quite freely subject to a generous estimate of the probable maximal length. The "center" beads, once again, were forced to meet the center-to-center limits from the NS experiment. No prior assumptions were made about which sites met the close approach or far approach limits (constraints *i* and *iii*, above).

At the time of this work, the only protein that had not been located by IEM was protein 4. For this protein all three beads were positioned only by NS data and a maximal length restriction.

The specific limits used from the NS data have been described for constraints *i* and *iii*. The peak approach limits were fixed from the distance at the peak of the distribution and the uncertainities generated by the statistical counting errors (11) with the following exceptions: (*a*) the minimum error was taken as ± 5 Å and (*b*) for those pairs where the coordinates calculated by Schlindler et al. (11) provide distances that did not fall within the above

TABLE III
BOUNDS DERIVED FROM NS DATA (REFERENCE 11)

Protein pair	Close approach		Peak approach		Far approach	
	Lower	Upper	Lower	Upper	Lower	Upper
2-5	50	90	95	140	40	180
3-4	20	50	64	74	85	125
3-5	0	45	59	70	80	130
3-7	N.A.	N.A.	90	115	N.A.	N.A.
3-8	0	65	86	99	105	170
3-9	0	40	52	63	70	115
3-10	0	35	35	70*	85	150
3-11	50	95	105	125	135	170
3-12	40	65	66	90	100	145
4-5	0	35	45	55	60	90
4-7	20	75	90	125	125	185
4-8	15	50	72	82	85	120
4-9	30	70	70*	125*	120	165
4-10	0	70	90	105	105	145
4-11	15	75	110*	150	155	190
4-12	10	40	38*	57	60	150
5-6	50	85	84	116	116	150
5-7	45	85	99	123	130	180
5-8	0	30	28	48	50	120
5-9	40	70	84	106	110	150
5-10	30	80	83	115*	125	180
5-11	40	70	85	99	105	140
5-12	0	35	43	60	65	110
6-11	10	40	52	62	70	110
6-15	0	30	43	53	60	110
7-8	25	95	109	121	125	170
7-9	0	25	31	51	50	75
7-12	0	85	95	120	120	145
8-9	30	70	90	130	120	170
8-10	N.A.	N.A.	90	140	N.A.	N.A.
8-11	0	80	100	120	130	180
8-12	0	45	45*	75	75	95
9-10	0	30	33*	48	55	80
11-12	0	95	99	121	130	180

Bounds in ångströms calculated from length distribution functions as described in text.
N.A., not available.
*Bound extended to include NS coordinate model distances (11); see text.

boundaries, the appropriate bound was extended so that the NS model distance could be accommodated.

The set of upper and lower bounds used for the NS data are given in Table III.

RESULTS

Test of Distance Geometry Methods

To test this general approach, and particularly to see what results could be expected with the extended protein representation, we used a model system prepared by Moore and Weinstein.[1]

TABLE IVA

STRUCTURES FROM ELLIPSOID MODEL SYSTEM. COMPARISON OF FIVE STRUCTURES (RMS ERROR IN ÅNGSTRÖMS PER COORDINATE AVERAGED OVER COORDINATES)*

	All	Center	All	Center	All	Center	All	Center
1	8	5	18	4	18	7	14	7
2			13	4	19	7	18	3
3					19	6	16	5
4							21	7

*Errors caluclated as: $E = \{1/n \Sigma_{i=1}^{n} [c(i) - C(i)]^2\}^{1/2}$, where c is the coordinate set from the calculated structure and C is the reference coordinate set. For this comparison, the equivalent positions of the outer beads for the ellipsoids were at distances of ± $a/2$ from the center of mass lying along the principal axis, with a being the semimajor radius. All, all beads; center, center beads.

This model consisted of 10 prolate ellipsoids of varying axial ratios. We were given length distribution data obtained from a Monte Carlo calculation for 36 of the 45 ellipsoid pairs.[3] We derived boundary conditions in the same way as described above for the 30S data.

Five structures were generated meeting all the constraints, that is, with the error function reduced to zero. These structures were all rotated to the same coordinate system using the algorithm of Ferro and Hermans (18). The average separation of equivalent points measures the agreement among a set of structures.

The following features are noteworthy: the structures differ among themselves by 20 Å (Table IVA). Since the largest distances in the test problem are > 200 Å, this implies that the major features of the five structures are similar. The largest discrepancies (~ 50 Å) arise from differing orientations of the longest ellipsoids (see below). The positions of the 10 "center" beads are much more restricted, with average differences of < 10 Å.

The coordinates of each structure can be compared with the known geometry of the system of 10 ellipsoids if we represent each ellipsoid with one bead at its center of mass and the two outer beads along the major axis at distances of plus or minus one-half the semimajor radius. The average errors are all ~ 18 Å for the 30 beads and < 10 Å for the 10 center beads (Table IVB). These values are very nearly those obtained amongst the five structures compared with each other. This suggests no significant bias in the structure-generating program that would restrict the structures to an inappropriately narrow segment of conformation space.

Finally, we examined the extension and orientation of each three-bead unit compared with the model ellipsoid (Table IVC). The separation of each pair of "outer" beads was consistently about one-half of the actual length of the ellipsoids. This effect arises largely from our use of the 25%, 75% cutoffs in the length-distribution data. Another contribution to the systematic underestimate is that the ellipsoid lengths are measured from "tip to tip," while the calculated distances run from bead center to bead center. Beads of appropriate size for ~ 25% of the mass would have radii approaching 10 Å, making a noticeable improvement in the comparison. The situation with respect to orientation is not good. The average angle between the major axis of the reference ellipsoids and the vectors linking the "outer" beads of each unit was found to be near 45° for all the units. Since the ends of the units and the ellipsoids are not

[3] Moore, P. A. Private communication.

TABLE IVB
STRUCTURES FROM ELLIPSOID MODEL SYSTEM. COMPARISON WITH REFERENCE COORDINATES (RMS ERROR IN ÅNGSTRÖMS PER COORDINATE)*

Structure	All	Center
1	17.2	5.6
2	18.3	5.6
3	16.8	5.3
4	18.0	6.0
5	17.8	6.7

*Same as footnote in Table IVA.

distinguishable, the maximum range of angle is 90°. Thus the calculated structures have near random orientation of each "protein" with respect to the reference ellipsoids. One source of the lack of proper orientation may be the relatively mild extension of the units.

We conclude from this test that the distance geometry methods can properly analyze data of the type available for the proteins of the 30S subunit. The limit on how well the structure can be specified rests on the amount and accuracy of the data, not on the methodology. The ellipsoid problem suggests that sufficient length-distribution functions will provide coordinates with rather good center-to-center separations as well as some useful information about extension of the nonspherical proteins. The test problem makes clear that we cannot hope to infer the orientation of the anisotropic proteins reliably at the present time, although the IEM data improves the situation for the 30S subunit calculation.

30S Subunit Results

Our representation of the 30S subunit proteins contains 47 beads for the 21 proteins (Table I). The boundary conditions derived from the IEM and NS data have been given (Tables II and III). We obtained five structures in five tries with values of the error function < 100, corresponding to average bound violations of ~ 0.001 Å. One set of coordinates is given in Table V and the structure is shown in Fig. 1. Comparing the five structures among themselves, we find that the average distance in the best fit rotation is 25 Å with the worst

TABLE IVC
STRUCTURES FROM ELLIPSOID MODEL SYSTEM. ELLIPSOIDAL EXTENSION

Ellipsoid	Calculated extension*	Semimajor axis of reference elipsoid
1	27	30
2	32	37
3	59	71
4	15	21
5	13	19
6	35	43
7	17	15
8	36	34
9	35	30
10	21	19

*Outer bead-outer bead separation averaged over the five structures.

TABLE V
COORDINATES OF STRUCTURE 1

		(ångströms)	
1	−27.5	202.3	120.7
2A	−22.3	205.9	48.7
2B	26.1	170.4	74.7
2X	−8.3	191.1	49.9
3A	41.7	146.7	40.0
3B	20.8	164.4	−15.5
3X	9.7	134.6	29.1
4A	37.6	118.0	64.2
4B	55.4	102.0	18.7
4X	67.0	110.6	44.3
5A	38.9	112.9	43.9
5B	51.8	88.2	81.6
5X	45.7	116.9	83.2
6A	−25.3	172.6	101.8
6B	−16.4	91.2	70.2
6X	−38.8	141.0	77.4
7A	−20.4	228.8	58.7
7B	25.4	190.7	62.4
7X	11.1	222.2	49.5
8A	28.8	159.4	86.3
8B	71.0	107.7	78.2
8X	62.3	142.9	112.6
9A	4.3	160.9	39.0
9B	8.3	209.5	6.3
9X	16.3	185.3	39.8
10A	20.7	171.7	−38.5
10B	22.8	142.6	27.0
10X	14.8	174.6	3.7
11A	−35.8	167.1	122.5
11B	−7.3	68.5	86.0
11X	−28.9	102.8	121.4
12A	47.6	185.5	69.2
12B	26.7	148.9	83.2
12X	69.3	149.7	68.7
13	−22.2	226.0	47.1
14	35.3	158.3	−16.1
15A	−2.7	140.0	101.1
15B	10.5	87.7	71.9
15X	2.9	129.1	107.9
16	−19.2	215.0	73.4
17	20.0	154.5	107.2
18A	−15.7	139.1	132.9
18B	13.4	23.5	102.0
19A	30.4	145.5	−17.3
19B	−17.0	220.4	56.1
20	−22.2	201.3	11.9
21	−33.8	177.1	113.9

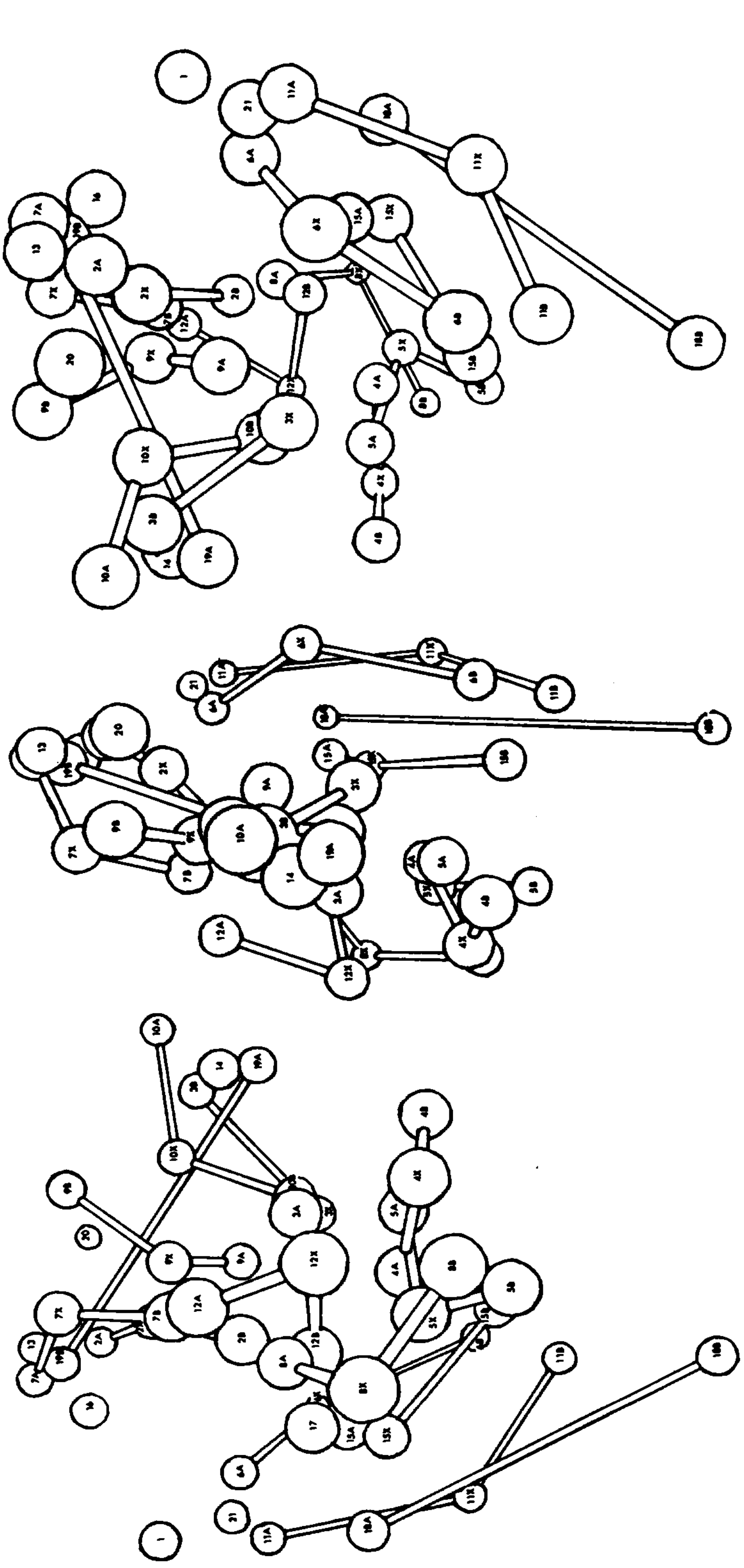

FIGURE 1 Drawing of 30S proteins in of Structure 1 (Table V). The protein subunits are labeled as indicated in Table I. Three views are shown approximating those in reference 8. The size of the circles represents the distance from the viewer, not the relative size of the proteins. As noted in the text, this structure is only one of several that meets the IEM and NS constraints used here, and it should not be taken to represent a unique conformation.

case deviations approaching 50 Å. The average displacements for each protein site are given in Table VI, column 2. Large values (30–50 Å, coded D in the table) are easily interpreted: they provide good evidence that the set of constraints used was not sufficient to restrict these sites more strongly. The eight sites that fall into this class are clearly those that would be judged *a*

TABLE VI
COMPARISON OF CALCULATED COORDINATES WITH IEM AND NS MODELS

Protein site	Average displacement‖			
	Calculated structures§	IEM model*	IEM input‡	NS model (11)
		(ångströms)		
1	A	A	A	—
2A	A	C	A	—
2B	A	C	A	—
2X	B	—	—	n.d.
3A	A	B	A	—
3B	D	—	—	—
3X	D	—	—	B
4A	C	—	—	—
4B	C	—	—	—
4X	B	—	—	B
5A	A	B	A	—
5B	B	—	—	—
5X	B	—	—	B
6A	C	N.A.	N.A.	—
6B	C	N.A.	N.A.	—
6X	D	—	—	n.d.
7A	A	B	A	—
7B	A	B	A	—
7X	C	—	—	B
8A	A	C	A	—
8X	B	—	—	B
9A	C	n.a.	n.a.	—
9B	C	n.a.	n.a.	—
9X	C	—	—	B
10A	C	D	B	—
10B	D	—	—	—
10X	C	—	—	B
11A	B	B	A	—
11B	D	—	—	—
11X	D	—	—	E¶
12A	D	—	—	—
12B	A	C	A	—
12X	B	—	—	C
13	A	A	A	—
14	A	B	A	—
15A	C	n.a.	n.a.	—
15B	B	n.a.	n.a.	—
15X	D	—	—	n.d.
16	A	A	A	—
17	A	C	A	—
18A	A	n.a.	n.a.	—
18B	B	n.a.	n.a.	—
19A	A	A	A	—
19B	A	A	A	—
20	A	B	A	—
21	A	B	A	—

*Initial coordinates supplied by Lake

‡Modified coordinates used as input (see text).

§RMS variation over five calculated structures, per coordinate.

‖RMS distance between average of calculated structures and coordinate set as indicated.

¶See text.

Abbreviations: A, <10 Å; B, 10–20 Å; C, 20–30 Å; D, 30–50 Å; and E, >50 Å displacements. n.a., not available; n.d., some data available but underdetermined.

priori to be the most poorly determined (Table I): 3B,3X,6X,10B,11B,11X,12A, and 15X. None has been assigned by IEM. 6X and 15X are underdetermined by NS and the rest of the sites are restricted only by the relatively weak conditions imposed by the wings of the length distribution functions. We note in passing that some of these sites are not randomly distributed over a 50-Å arc, but rather are clustered in two localized but well-separated positions. This is true of 3X,6X, and 11X. A larger number of low error structures would be needed to draw any firm conclusions from this observation.

Many sites are characterized by very similar positions (< 10 Å separations, coded A in Table VI) among the five structures. This result should not be overinterpreted. These sites were all assigned by IEM and more than half have not yet been studied by NS. Since we know, in advance, that the IEM data provide a self-consistent three-dimensional structure, these sites are expected to be near their "input" positions. Lack of variability cannot be construed as positive evidence fòr the "correctness" of the coordinates. Such a decision must be based on the accuracy of the input data, not merely on its self-consistency.

We compare the calculated structures with the experimental data in four ways: (*a*) average displacement of calculated coordinates from IEM coordinates; (*b*) average displacement of calculated coordinates from NS centers of mass coordinates; (*c*) calculated and observed second moments of the length distribution functions; and (*d*) calculated and observed lengths for individual proteins; calculated and observed closest approach and greatest separation distances.

IEM COMPARISON The IEM data set we used as input coordinates followed closely but was not identical with coordinates developed by Lake.[2] The exceptions were noted earlier and included modifications for proteins 6,9,10,15, and 18. If we compare the final coordinates for the five calculated structures with the initial Lake coordinates, omitting these proteins, the average mismatch is ~ 18 Å, with all five calculated structures being equally "distant" (Table VI, column 3). When we compare the final structures with the actual set of coordinates we used as input data, the average deviations are all 10 Å or less (Table VI, column 4) except for protein 10, which lay some 25 Å higher than the IEM position in all structures except one.

Generally, the agreement between the IEM coordinates and the calculated coordinates is far better than the suggested resolution of 50 Å. Hence, the discrepancies noted in Table VI are not worthy of extensive discussion, with the possible exception of protein 10.

NS COMPARISON Schindler et al. (11) have derived a set of coordinates for nine protein centers from their NS data (proteins 3,4,5,7,8,9,10,11, and 12). When we rotate our structures onto their coordinates, the average disagreement is 40 Å, with two of our structures having 50 Å differences taken over all nine protein centers (Table VI, column 5). The dominant part of the discrepancy lay in the location of protein 11, which was consistently 50–90 Å lower in the NS model than in ours. If protein 11 was omitted from the comparison, the mismatch dropped to an average of 18 Å for the eight protein centers. Most of the remaining residual was in the location of protein 12. Schindler et al. point out that the center of protein 11 is relatively poorly determined because the proteins used to locate it lie largely in a plane, making the triangulation procedure somewhat unstable (11). They note an alternative position for protein 11 that gave slightly better agreement with the structures here (average discrepancy 33 Å, worse case 40 Å), but the bulk of the disagreement was still in the

TABLE VII
SECOND MOMENTS

Protein pair	Experimental (11)	Calculated (see text)
2-5	14850 ± 4280	9870
3-4	5511 ± 0340	4494
3-5	4862 ± 0270	5084
3-7	12620 ± 0900	8506
3-8	8342 ± 0490	10189
3-9	3542 ± 0240	3500
3-10	5395 ± 0550	2936
3-11	15000 ± 0950	12853
3-12	7815 ± 1070	5227
4-5	2870 ± 0250	2406
4-7	2870 ± 0250	2406
4-8	6644 ± 0360	6510
4-9	10149 ± 0430	10388
4-10	9063 ± 0470	9192
4-11	14981 ± 1450	15511
4-12	2764 ± 0200	2836
5-6	11099 ± 2410	8252
5-7	13233 ± 1360	12244
5-8	1534 ± 0700	2287
5-9	9093 ± 1070	8180
5-10	12840 ± 1400	9895
5-11	8868 ± 0800	7866
5-12	3367 ± 0180	2172
6-11	3633 ± 0170	4102
6-15	2647 ± 0100	3160
7-8	13670 ± 0650	13317
7-9	1536 ± 0120	1828
7-12	12091 ± 1410	11396
8-9	12998 ± 1990	11008
8-11	12362 ± 1280	10824
8-12	3230 ± 0220	2478
9-10	1959 ± 0180	2300
11-12	13902 ± 1120	15203

protein 11 locations. We conclude that our coordinates for the centers of proteins 3,4,5,7,8,9,10, and 12 are the same as those of Schindler et al. within experimental error. The disagreement over the location of protein 11 lies outside these limits.

We next computed approximate second moments from our structures, assuming the beads represent prolate ellipsoids of revolution, with their long axes being given by the distance between the outer beads and the short axes to be 20 Å.[4] This description should consistently underestimate the radii of gyration, because the three-bead per protein representation does not yield the full extension of an equivalent ellipsoid. The calculated second moments (Table VII) average 1.8 SD (based on standard deviations calculated in reference 11) from the

[4] We use the formula of Moore and Weinstein (footnote 1): $M_{ij} -D^2_{ij}+r^2_i+r^2_j$, where $M(ij)$ is the second-moment of the i,jth length-distribution function, $D(ij)$ is the separation of the centers of mass of proteins i and j, and $r(i)$, $r(j)$ are the respective radii of gyration of the two proteins.

experimental values, with ~ 0.3 of this being due to underestimation. A few moments are fit particularly poorly: 3-7, 3-8, 3-10, 5-12, and 8-12. While we did not seek specifically to fit the second-moments, the second-moment data are dominated by the center of mass distances in most cases; the reasons for poor agreement in these five cases should be sought. We note that the 3-7 data have caused difficulties for some time. They are not consistent with the triangle inequality for proteins 3,7, and 9 (11) and the length distribution has not been published. For the other pairs, with the exception of 5-12, the length-distribution functions show strong shoulders or even two separate peaks. This condition will normally permit significant differences between the distance at the peak of the distribution function and the "true" distance between the centers of mass. We are attempting to model the actual mass distribution rather than locate the centers of mass, so that some of the disagreement is due to the difference in these two quantities. Additional calculations have shown that the 5-12 distance can, in fact, be increased, still maintaining the other constraints. With these five pairs removed and allowance made for the systematic underestimate, the calculated second-moments are within 1 SD of the experimental values.

Finally, we calculate the apparent extension of the 21 proteins in this model (Table VIII). These are compared with estimates of length deduced from experiment (Table VIIIA) and

TABLE VIIIA
MEASURES OF PROTEIN EXTENSION FOR 30S PROTEINS. COMPARISON OF OUTER BEAD SEPARATIONS WITH EXPERIMENTAL LENGTHS

Protein	Axial ratio*	Experimental lengths‡	Calculated extension§
1	10:1	240	n.a.
2	6:1	53‖	83
3	5:1	110	76
4		140	76
5	7:1		69
6	4:1	(166)‖	103
7		100, 51‖	75
8	2.3:1	100	90
9		54‖	78
10			94
11			107
12			70
13	2:1		n.a.
14			n.a.
15	5:1	120,(158)‖	69
16		90	n.a.
17	4:1		n.a.
18	10:1	(153)‖	123
19		110	115
20	6:1	100	n.a.
21	9:1		n.a.

*Taken from Table 5 of reference 21; averaged values based on neutron and x-ray scattering and hydrodynamic studies of isolated proteins.

‡Taken from Table 4 of reference 21; averaged values using techniques given above.

§Average of calculated outer bead separations + 20 Å to allow for end effects.

‖IEM results from Lake (refrence 8). Those in parentheses are now thought to be too extended.

n.a., not available.

TABLE VIIIB
MEASURES OF PROTEIN EXTENSION FOR 30S PROTEINS. COMPARISON OF OUTER BEAD SEPARATIONS WITH NEUTRON SCATTERING. PAIR DISTRIBUTION LIMITS AND ELECTRON MICROSCOPE RESULTS

Protein pair	NS limits	IEM*	Calculated	Calculated/NS‡
2-5	50 180	62 118	55 144	0.76
3-5	10 130	32 42	36 116	0.75
3-7	n.a.	56 104	52 115	
3-8	0 170	47 57	51 130	0.58
3-9	0 115	39 105	32 88	0.66
3-10	0 150	51 61	23 89	0.57
3-11	50 170	106 116	81 152	0.76
3-12	40 145	42 52	41 102	0.69
4-5	0 90		16 66	0.78
4-7	20 185		72 160	0.72
4-8	15 120		42 96	0.70
4-9	30 165		49 130	0.75
4-10	0 145		58 130	0.63
4-11	15 190		68 151	0.63
4-12	10 150		36 93	0.55
5-6	50 150	95 129	66 117	0.71
5-7	45 180	76 135	78 158	0.74
5-8	0 120	61 71	23 77	0.62
5-9	40 150	68 138	56 140	0.95
5-10	30 180	63 73	57 142	0.70
5-11	40 140	113 125	62 129	0.87
5-12	0 110	51 63	35 90	0.68
6-11	10 110	20 170	25 101	0.95
6-15	0 110	0 180	48 71	0.39
7-8	25 170	35 93	43 157	0.92
7-9	0 75	0 76	22 74	0.84
7-12	0 145	37 96	39 129	0.76
8-9	30 170	25 87	48 139	0.79
8-11	0 180	62 73	70 130	0.44
8-12	0 95	0 20	9 80	0.95
9-10	0 80	86 129	29 79	0.75
11-12	0 180	66 76	72 130	0.43

*Original coordinates of Lake.

‡Bead center to bead center separation + 20 Å.

with the extremes of the NS length distributions (Table VIIIB). Our model invariably yields extensions less than those measured directly or deduced from NS data. Approximately 50–75% of the range of the length distribution data are represented in these structures. This is not unexpected, because only part of the length distributions were forced upon the model.

DISCUSSION

Our major objective in this undertaking was to see what types of information could be extracted from the IEM and NS data to provide as complete and self-consistent a picture of the 30S subunit as possible. The calculated structures described here are our first attempt to model the protein constituents of the subunit. There are several interesting aspects.

First, it is clear that within some generous estimate of experimental error, the data from the two sources are essentially self-consistent. Second, most of the sites are rather well-located. Those that are not well-located are generally those that are simply underdetermined (Table I). Third, the major new feature in this model is the definition of an approximate relation between the protein centers (largely fixed by the NS peaks in the length distributions and/or the second-moments of the distributions) and the protein extensions (set by the IEM positions and/or the NS length-distribution extrema).

We comment briefly on the relation of the calculated structures to the data sets from which they were obtained. As we have seen, the IEM results are readily incorporated into distance geometry models. Three-dimensional coordinates are inherently extremely informative about the protein sites that reach the ribosome surface. The random errors have been estimated as 50 Å (8). Lake has pointed out several systematic errors; antigenic cross reactivity and projection problems are the most important (8). The calculations presented here generally confine the IEM sites to closer tolerances than 50 Å with the exceptions of proteins 4,6,9,10,15, and 18, as noted earlier. The general conclusion is that the IEM data are self-consistent and, with some small revisions, compatible with the NS results.

Turning to these latter experiments, we find that the most informative data are the length-distribution functions. Moore and Weinstein have shown that the separation of centers of mass is directly related to the second-moments of these functions.[1] Our current calculations accept interpoint distances. This permits a slightly different mode of analysis than that used by Schindler et al., who fit the second-moment results directly (11). The mathematical differences are somewhat subtle, involving implicit assumptions about the dominant error terms, the distinction between the center of mass separations and the most probable separation, and the mathematical stability of the solutions. In fact, the results of the two calculations are very similar with the sole exception of protein 11. Inspection of all the center-center distances shows the two models to be in agreement within ± 15 Å, except for the 3-10 and 3-12 distances, which differ by 21 Å. The rms error in distances is 9 Å. Hence we assign the discrepancy in the location of protein 11 to instability in the process of going from distances to coordinates. We note that this question is readily resolved experimentally. The 7-11, 9-11, and 10-11 distances are predicted by Schindler et al. to be 175–200 Å, whereas all our structures have these distances as 100–125 Å.

In summary, we feel that the calculated structures provide an acceptable fit to the IEM and NS data base. We have not included several other experiments such as cross-linking, energy transfer, etc. in the present calculation. We have nothing to add to the cogent discussion of Schindler et al. (11) regarding those pairs of proteins included in the NS experiments. As for the other protein pairs, the major areas of conflict between our results and the close proximity cases documented by Craven (4) are resolvable by postulating a rotation of protein 4 and extensions of proteins 13,20, and 21. Extension of protein 13 has been previously suggested (17). These points will be explored in a later paper.

There have been several mathematical models put forward for the 30S subunit. Bollen et al. used the multidimensional scaling technique on the relatively small amount of data available 5 years ago (3). They did not use extended proteins, so that their coordinates are not readily compared with those derived here or with the IEM models. At about the same time, Traut and co-workers summarized their cross-linking results in a model that also used spherical proteins (19). More recently, Cornick and Kretsinger (20) developed a novel lattice representation that permitted a more systematic search of possible protein arrangements. Again, it lacked extended proteins, although the method could be adapted to accept more than one site per protein. Finally, Gaffney and Craven (4) have quite recently used the multiscaling approach and the Stoffler and Wittman (16) IEM results to generate a self-consistent set of coordinates for the IEM sites. They did not directly include the NS results. The procedure we have used has several features that are not incorporated into these earlier models. We are able to combine quite diverse data sets, allowing both upper and lower bounds on the measurements. Further, we produce only structures that fit within the envelope of the shape of the 30S subunit. Finally, some effort is made to deduce the length and orientation of the proteins from the total data available. There are several ready extensions of the methodology we have outlined. Second moments can be fit directly. More sites per protein can be defined as the data warrant. Cross-linking and other experimental results can be incorporated. Most importantly, the new cross-linking experiments on 16S RNA and protein-RNA interactions can be used to produce a much more comprehensive picture of the 30S subunit (1, footnote 5). Mathematical modeling of a complex structure has several uses in manipulating data. When associated with computer graphics, it can produce vivid representations of complicated relationships. It cannot remove the need for accurate and precise data.

We are grateful to a number of people who provided us with manuscripts in advance of publication. These include Charles Cantor, Barry Cooperman, John Hearst, Jenifer Littlechild, and Paul Wollenzien. Peter Moore, Donald Engelman, and James Lake were most generous with data, interpretations, and encouragement. We thank Paul Weiner who supplied the hidden line program.

The facilities of the Computer Graphics Laboratory at University of California, San Francisco were used for much of this work. Partial financial support from the National Institutes of Health and the National Resource for Computation in Chemistry is acknowledged.

Received for publication 13 December 1979.

REFERENCES

1. Brimacombe, R., G. Stoffler, and H. G. Wittmann. 1978. Ribosome structure. *Annu. Rev. Biochem.* **47**:217–49.
2. Langer, J. A., D. M. Engelman, and P. B. Moore. 1978. Neutron-scattering studies of the ribosome of *Escherichia coli*: a provisional map of the locations of proteins S3,S4,S5,S7,S8, and S9 in the 30S Subunit. *J. Mol. Biol.* **119**:463–485.
3. Bollen, A., R. J. Cedergren, D. Sankoff,and G. Lapalme. 1974. Spatial configuration of ribosomal proteins: a computer-generated model of the 30S subunit. *Biochem. Biophys. Res. Commun.* **59**:1069–1078.
4. Gaffney, P. T., and G. Craven. 1978. Use of computerized multidimensional scaling to compare immunoelectron microscopy data with protein near-neighbor information:application to the 30S ribosome from *Escherichia coli. Proc. Natl. Acad. Sci. U.S.A.* **75**:3128–3132.
5. Crippen, G. M. 1977. A novel approach to calculation of conformation: distance geometry. *J. Comput. Phys.* **24**:96–107.
6. Havel, T. F., G. M. Crippen, and I. D. Kuntz. 1979. Effects of distance constraints on macromolecular conformation. II. simulation of experimental results and theoretical predictions. *Biopolymers.* **18**:73–81.
7. Kuntz, I. D., G. M. Crippen, and P. A. Kollman. 1979. Application of distance geometry to protein tertiary structure calculations. *Biopolymers.* **18**:939–957.
8. Lake, J. A. 1979. Antibody labeling studies on ribosomes. *In* Advanced Techniques in Biological Electron Microscopy. J. Koehler, editor. Springer-Verlag, GmbH., Berlin. 172–211.

[5] Wollenzein, P., J. Hearst, P. Thammana, and C. R. Cantor. Manuscript in preparation.

9. Crippen, G. M., and T. Havel. 1978. Stable calculation of coordinates from distance information. *Acta Crystallogr. Sect. A. Cryst. Phys. Diffract. Theor. Gen. Crystallogr.* **34**:282–284.

10. Crippen, G. M. 1979. Determination of protein conformation. I. The effectiveness of the experimental studies on tobacco mosaic virus protein. *Int. J. Pept. Protein. Res.* **13**:320–326.

11. Schindler, D. G., J. A. Langer, D. M. Engelman, and P. B. Moore. 1979. The positions of S10,S11,S12 in the 30S ribosomal subunit of *E. coli. J. Mol. Biol.* **134**:595–620.

12. Moore, P. B., J. A. Langer, B. P. Schoenborn, and D. M. Engelman. 1977. Triangulation of proteins in the 30S ribosomal subunit of *Escherichia coli. J. Mol. Biol.* **112**:199–234.

13. Engelman, D.M., P.B. Moore, and B.P. Schoenborn. 1975. Neutron scattering measurements of separation and shape of proteins in 30S ribosomal subunit of *Escherichia coli*:S2-S5,S5-S8,S3-S7. *Proc. Natl. Acad. Sci. U.S.A.* **72**:3888–3892.

14. Hermans, J., D.R. Ferro, J.E. McQueen, and S.C. Wei. 1976. *In* Environmental Effects on Molecular Structure and Properties. B. Pullman, editor. Reidel Publishing Co., Dordrecht. 459–483.

15. Ycas, M., N.S. Goel, and J.W. Jacobsen. 1978. On the computation of tertiary structure of globular proteins. *J. Theor. Biol.* **72**:443–57.

16. Stoffler, G., and H.G. Wittman. 1977. Primary structure and three-dimensional arrangement of proteins within the *Escherichia coli* ribosome. *In* Molecular Mechanisms of Protein Biosynthesis. Academic Press, Inc., New York. 117–202.

17. Changchien, L., and G.R. Craven. 1977. Proximity relationships among the 30S ribosomal proteins during assembly in vitro. *J. Mol. Biol.* **113**:103–122.

18. Ferro, D.R., and J. Hermans. 1977. Different best rigid-body molecular fit routine. *Acta Crystallogr. Sect. A. Cryst. Phys. Diffract. Theor. Gen. Crystallogr.* **33**:345–347.

19. Traut, R.R., R.L. Heimark, T.T. Sun, J.W.B. Hershey, and A. Bollen. 1974. Protein topography of ribosomal subunit from *Escherichia coli. In* Ribosomes. M. Nomura, A. Tissieres, and P. Lengyel, editors. Cold Spring Harbor Laboratory, New York, 270–308.

20. Cornick, G.C., and R.H. Kretsinger. 1977. The 30S subunit of the *Escherichia coli* ribosome topographical model of its component proteins. *Biochim. Biophys. Acta* **474**:398–410.

21. Wittmann, H.G., J.A. Littlechild, and B. Wittmann-Liebold. 1979. Steenbock Symposium on Ribosomes. In press.