CORRELATION OF AMINO ACID SEQUENCE AND CONFORMATION IN TOBACCO MOSAIC VIRUS

MARIANNE SCHIFFER *and* ALLEN B. EDMUNDSON

From the Division of Biological and Medical Research, Argonne National Laboratory, Argonne, Illinois 60439

ABSTRACT Correlation of the amino acid sequence with the conformation in tobacco mosaic virus protein is considered in this article. After division of the sequence into groups with helical or nonhelical potential, the segments likely to be helical were related to the X-ray diffraction patterns obtained by Franklin, Caspar, Holmes, and Klug. The approximate locations of these segments within the known boundaries of the subunit were predicted from the radial distribution and helical projection of electron density. As a result of these assignments, the number of possible conformations was also reduced for the nonhelical segments. The structure of the subunit was simulated by flexible models of rubber and electrical tubing, as well as by space-filling Corey-Pauling-Koltun models. These models were used to locate the protein segments impinging upon the ribonucleic acid of the virus. The two pairs of carboxyl groups believed to be responsible for the binding of lead were also tentatively identified on these models as aspartic acid residues 64 and 66 (first pair) and glutamic acid residues 131 and 145 (second pair).

INTRODUCTION

In an earlier article (Schiffer and Edmundson, 1967) we described the use of projections called "helical wheels" to divide the amino acid sequences of proteins into categories with helical and nonhelical potential. This division substantially reduces the number of conformations a protein can assume. This number can be decreased still further by the imposition of chemical and X-ray information. However, the quantity of ancillary information required to deduce an accurate and unique conformation is not clear. This problem, examined in the present article, is of special importance in those cases in which the proteins cannot be crystallized or are otherwise not suitable for X-ray analysis.

Because our knowledge at this stage is limited to results for only a few proteins, we chose to investigate tobacco mosaic virus, for which both chemical and X-ray data are available. The complete sequences of the 158 amino acid residues in two strains (*vulgare* and *dahlemense*) are known, and the amino acid replacements have

been identified in many mutants (Funatsu et al., 1964; Anderer et al., 1960, 1965; Anderer and Handschuh, 1962; Wittmann-Liebold and Wittmann, 1963; Fraenkel-Conrat, 1965). With X-ray techniques, oriented gels have been used to obtain fiber diagrams from which the symmetry of the virus and the size and shape of its individual subunits have been deduced (Franklin et al., 1959; Klug and Caspar, 1960; Caspar, 1963; Holmes and Klug, 1963). Unfortunately, the lack of either single crystals or suitable isomorphous derivatives has thus far made it impossible to determine the conformation of either the protein or the RNA in each subunit. Caspar (1956) and Franklin (1956; see also Franklin et al., 1959) did succeed in calculating the radial distribution of electron density. They related this distribution to the contributions of the protein and the RNA.

The density profile contained peaks with maxima at radial distances of 25, 40, 66, and 78 A from the particle axis (i.e., from the middle of the central cavity in the virus). When the protein alone was examined, a broad, asymmetrical peak with a maximum at about 48 A was detected, and a minimum density was found at 40 A. The region of high density at a radius of 40 A was accordingly assigned to the nucleic acid, and the remainder of the peaks to segments of protein. In 1963, Holmes and Klug provided a more detailed projection of electron density at a resolution of 10 A. The major difference between the two maps is the presence of two peaks shifted toward a radius of 60 A in the helical projection used by Holmes and Klug. There are also two peaks at different levels in the region between radii of about 70 and 77 A. In the areas of disagreement we tentatively assumed that the more recent projection obtained by Holmes and Klug superseded the earlier calculations.

To construct our model, we combined the X-ray results with the use of helical wheels (Schiffer and Edmundson, 1967) and auxiliary chemical data. The estimated helical percentage of about 40% (Schiffer and Edmundson, 1967) suggests that the application of methods designed to distinguish helical from nonhelical segments should be effective. However, it should be emphasized that an unequivocal interpretation of electron density maps is not possible at a resolution of only 10 A. The present model is consequently offered only as one plausible and approximate fit to the experimental results and is intended to be used as a starting point for further investigation.

CONSTRUCTION OF MODELS

The TMV protein was first represented by a flexible model, consisting of electrical tubing (spaghetti-type) and rubber tubing. Both the length of the electrical tubing and the diameter of the rubber tubing were selected to fit a scale of $\frac{1}{8}$ inch = 1 A. It was assumed that the polypeptide chain in nonhelical segments was in an extended form corresponding to approximately 3.5 A of length per amino acid residue. Each peptide unit in these segments was represented by black rubber tubing equivalent to 4 A in diameter. Side chains of polar residues were generally omitted, but those of

hydrophobic residues were constructed of plastic beads and tied to the rubber tubing as needed. Cylinders of white rubber tubing, equivalent to 10 A in diameter to include side chains, were used to represent segments previously predicted to be helical (Schiffer and Edmundson, 1967). Each was assumed to be an α-helix, with 3.6 residues per turn and a pitch of 1.5 A per residue.

Spatial relations among the various segments were determined with the aid of ancillary information, such as the molecular dimensions (Franklin et al., 1959). The known shape of the subunit was outlined on cardboard to the same scale as that used for the flexible model. The radial distances were marked off at intervals of 10 A, and the flexible model was fitted to the shape. Wherever possible the hydrophobic side chains were turned away from the surfaces exposed to solvent and directed toward the "inside" of the molecule (see Kendrew et al., 1961; Perutz et al., 1965).

Major advantages of these models, with the lengths and diameters of the components reduced to average values, are the ease and simplicity of construction and the flexibility. They have proved useful in the testing of different over-all conformations, but it should be emphasized that they are not designed for detailed investigation of specific regions of the proteins. For this purpose we have employed the Corey-Pauling-Koltun (CPK) models, which were supported on a lattice-work of ½ inch rods and oriented within the boundaries assigned to the subunit. Short segments of RNA were also simulated with the CPK models.

INTERPRETATION OF ELECTRON DENSITY

Implicit in the interpretation of the electron density is the assumption that the peaks are mainly attributable to the presence of compact structures, probably helices. To represent the peaks, there are six segments with helical potential (Schiffer and Edmundson, 1967). The correlation of density and structure will be described for each peak. The results of this selection process are shown in Fig. 1.

On the basis of the cylindrical Patterson function, Franklin (1955) concluded that the predominant directions of the polypeptide chain were perpendicular to the long axis of the particle (the so-called "tangential direction"). Franklin and Holmes (1958) further suggested that the principal directions lay in a series of coaxial cylindrical surfaces. After considering infrared and birefringence measurements, Caspar (1963) favored an alignment of the helices along the long axis of the subunit (the radial direction). The subunit is relatively thin at radial distances of 55–65 A, and Caspar (1963) proposed that the electron density in this region may be accounted for by two helices oriented side by side in the radial direction. Using the additional data obtained by Holmes and Klug (1963), we have concluded that both types of alignment occur in TMV.

For the model, we assumed that the helix numbered *2* in Fig. 1 runs in a tangential direction. This helix (residues 19–33; Schiffer and Edmundson, 1967) was assigned to the peak with a maximum of 48 A. The placement of helix *2* was dictated by the

location of the single sulfhydryl group at a radial distance of 56 A (Franklin and Holmes, 1958). This distance corresponds to a crevice between two peaks on the map of radial density. The cysteine residue in question (No. 27; Tsugita et al., 1960; Funatsu et al., 1964; Anderer and Handschuh, 1962) is a constituent of helix *2*. It is the only polar residue in what is otherwise a completely hydrophobic arc on the

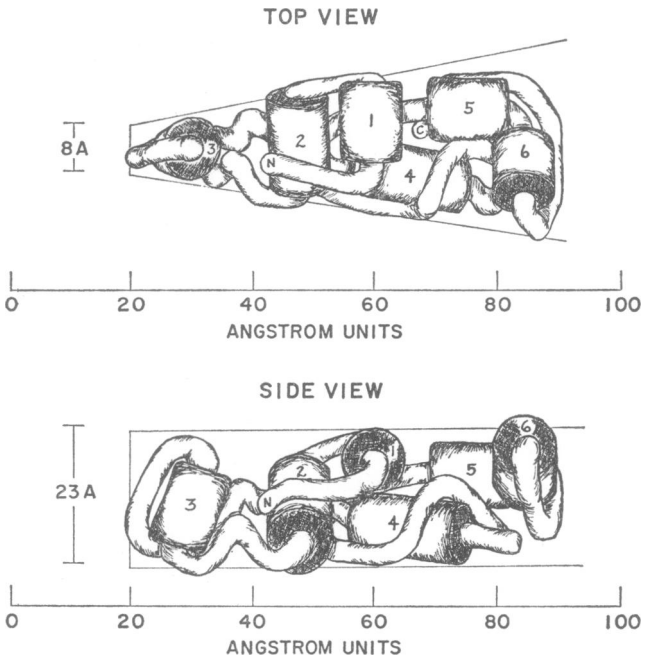

FIGURE 1 Model of TMV protein, top and side views. The drawings by Mrs. Kathryn Ely depict two views of a model of the polypeptide chain in one subunit of TMV. The proposed helical segments, represented by cylinders, are numbered as in the text. Individual side chains are not shown in either the helical or nonhelical segments. The radial distances from the axis of the virus particle are listed below each drawing. The outlines indicate the maximum amount of space available for each subunit. In the vicinity of 40 A the protein segments are in contact with the RNA of the virus, but the RNA is not shown in the drawings. The single sulfhydryl group is located in helix 2 at a radius of 56 A. The two pairs of carboxyl groups responsible for the binding of lead are found at radii of 25 and 84 A. Unequivocal assignments of the positions of the N- and C-terminal residues are still not possible, and they have arbitrarily been placed at radii of about 42 A (N-terminus) and 67 A (C-terminus). For further details, consult text.

helical wheel. The shielding afforded by this hydrophobic region, oriented toward the "inside" of the molecule, probably explains the inaccessibility of the cysteine residue to bulky sulfhydryl reagents (Fraenkel-Conrat, 1965). The chain must turn if helix *1* (residues 10–18) is to remain within the boundaries of the subunit. Such a turn usually requires one or more residues arranged in a nonhelical conformation, and these residues would probably be derived from helix *2*. The model presented in

Fig. 1 reflects the assignment of four of these residues (Nos. 19, 20, 21, and 22) to this "corner" segment. We cannot rule out the possibility that the structural requirements peculiar to this region result in residues 10–18 forming a compact, but not necessarily α-helical segment. The N-terminal nine residues precede helix *1*, but the scarcity of information about the environment prevents unequivocal orientation of both helix *1* and the end of the molecule.

Assignment of one of the helical segments to the peak at 25 A also requires consideration of the space available, since the subunit tapers to a width of only 8 A at the edge of the central cavity (i.e. at a radius of 20 A from the particle axis). As Caspar (1963) points out, an α-helix is not likely to be present in the space bordering this cavity. At a radial distance of 25 A, however, the width of the subunit has expanded to about 10 A, large enough to accommodate an α-helix in an upright or slightly tilted position. In the selection of this particular helix, we were restricted because the height of the subunit (23 A) had to accommodate the short segments connected to the ends of the helix, as well as the length of the helix itself. Helix *3* (residues 45–53) best meets these requirements. The tilted orientation (see Fig. 1) was chosen to obtain the closest fit to the helical projection of electron density (Holmes and Klug, 1963).

The polypeptide chain continuing from helix *3* was assumed to proceed to the central cavity and to loop back under one side of helix *3* toward the middle of the subunit. After this segment of chain crosses helix *2* at radii of about 43–53 A, it leads into helix *4* (residues 77–89). The chain has sufficient flexibility and length to permit the alignment of the center of helix *4* at peaks of density at either 60 or 66 A, but not at 78 A. By difference, the latter peak on the map of radial density (Franklin et al., 1959) has been attributed to helices *5* and *6* (residues 117–125 and 126–135). The helical projection (Holmes and Klug, 1963) indicates that a long L-shaped segment of density, with one maximum at a radius of about 74 A, extends from about 70 to 90 A. The segment of density oriented in the radial direction was ascribed to helix *5* and the segment in the tangential direction to helix *6* (see Fig. 1). As in the corner between helices *1* and *2*, residues (asparagines-126 and -127) had to be shifted from the helical to nonhelical category to make the turn between helices *5* and *6*.

In the model helix *6* extends to about 90 A, and the remaining weak density from 90 to about 93 A is probably attributable to a single loop of chain in a nonhelical conformation. A section emanating from helix *6* and composed of residues 136–148 has been designated as this loop in the model. This assignment is based largely on the necessity for the presence of a carboxyl group at a radius of 84 A (see next section). We believe that this carboxyl group is contributed by glutamic acid-145. This residue can be brought into position by swinging the loop outward to approximately 93 A and back toward 84 A. The next problem is the orientation of the carboxyl end, which is a continuation of the loop. The C-terminal threonine residue is known to be accessible to reagents used for end-group determinations (Fraenkel-Conrat, 1965). It is probably located in the outer regions where the subunits are in limited

contact in the virus. For these reasons, as well as to provide two layers of chain to give the subunit the proper thickness, the C-terminal segment was arbitrarily placed under helix *5*.

A similar lack of specific information hampers the positioning of the nonhelical segment between helices *4* and *5*. Again the known depth of the subunit (Franklin et al., 1959) requires at least two layers of polypeptide chain, but the relations between the nonhelical and helical segments cannot be deduced with the present model. In contrast the two nonhelical segments connecting helices *2* and *3* and *3* and *4* are in relatively fixed positions between the radii of 25 and 48 A. These segments are of particular importance because a section of the viral RNA crosses the subunit in the region centered at a radial distance of 40 A (Franklin, 1956; Franklin et al., 1959). The structural features of the possible binding sites and the interactions will be discussed in the following section.

ASSOCIATION OF PROTEIN AND NUCLEIC ACID

The RNA in the virus is believed to be single-stranded and to have its conformation determined by its interactions with the protein (Franklin et al., 1959). Each subunit of the protein is associated with a strand of three nucleotides. Since the length available at the radius of 40 A is about 15 A, the average distance between successive phosphorus atoms of the RNA is 5 A. The nucleotides are expected to follow a flat helix with the bases roughly perpendicular to the direction of the backbone of phosphate and ribose units (Franklin et al., 1959). Using the CPK models, we found that all of these requirements were met if the chain was folded into a helix in which the nucleotides were related by a threefold screw axis; i.e., the bases were separated by 120° of rotation. The planes of the bases were kept parallel both to each other and to the particle axis.

To examine the binding sites, a section of an RNA model was oriented along the 40 A radius of the corresponding CPK model of the protein. Although the interactions cannot be deduced with certainty, the general features of the binding sites are quite clear if the model is correct in this region. Like the active site of ribonuclease, the segments in TMV protein contain several basic residues, including three arginine residues (41, 46, and 71) and the only two lysine residues (53 and 68) in the protein.

There are also glutamine (Nos. 36, 38, and 39), asparagine (Nos. 25 and 29), and threonine residues (Nos. 37 and 42) available for hydrogen bonding with the appropriate groups on the bases or ribose units. While the binding sites probably do not change appreciably from one subunit to the next, the sequence of bases in each set of trinucleotides is expected to be extremely variable. The interaction energies will consequently be different in some cases, but the factors responsible for the specificity of the association between the RNA and the protein are not as yet clearly defined (see Caspar, 1963). The binding sites in the present model do not include

many residues suitable for apolar bonding with the bases, although tyrosine residues 70 and 72 are present in the area.

LEAD BINDING SITES IN THE SUBUNIT

While not of comparable interest, two other binding sites at radii of 25 and 84 A are also important, both in the aggregation of the protein and in the binding of lead (Caspar, 1956, 1963; Franklin and Holmes, 1958; Fraenkel-Conrat and Narita, 1958). Two atoms of lead are presumed to be bound to two pairs of interacting carboxyl groups. While these may not be located in the same subunit, it was tentatively assumed that they were. The first pair were identified as aspartic acid residues (Nos. 64 and 66), the only two acidic groups in the vicinity of 25 A in the model. Beyond a radius of about 75 A all of the potential participants except two glutamic acid residues (Nos. 131 and 145) are in the amide form. As noted in a previous section, these two residues are found in different segments of the polypeptide chain but are brought into apposition at a radius of 84 A in the model.

POSSIBLE INTERACTIONS BETWEEN SUBUNITS

In addition to the study of interactions in the binding sites within each subunit, the model can be used to investigate the interactions along the opposing surfaces of adjacent subunits. Such interactions are of particular significance in the assembly of TMV. Moreover, the protein can aggregate even in the absence of the RNA (see Fraenkel-Conrat, 1965). The interactions resulting in this aggregation are likely to occur mainly in the anhydrous regions in which the subunits are in close contact. Many of these regions are located between radii of 20 and 60 A. For example, Caspar (1963) visualized the sides of the tapered inside end of the subunit as being composed principally of hydrophobic residues. Conversely, the side chains are expected to be predominantly polar in all segments exposed to the solvent. An example of this type is the segment facing the central cavity at 20 A (Franklin et al., 1959).

These criteria are generally met in the model, particularly at the tapered end. Between radii of 20 and 30 A and bordering on the central cavity, for example, are the polar residues, glutamine-57, threonine-59, arginine-61, aspartic acid-64, and glutamine-45. These are listed in order of alignment from the top to the bottom of the model. At the level of 25–35 A proline-54, serine-55, and proline-56, are on top. Opposite these on the bottom are phenylalanine-67, aspartic acid-66, and serine-65. Aspartic acid residues 64 and 66, the proposed carboxylate pair, are believed to be un-ionized and to bind a proton firmly. When they are ionized, the protein will not form the helical aggregate in the absence of RNA (for discussion, see Caspar, 1963).

On the left side at radii of 20–30 A are residues with hydrophobic side chains, including tryptophan-52, phenylalanine-48, and valine-44. On the right side are valine-58, valine-60, phenylalanine-62, and proline-63. Since the subunits are in

contact in the tapered zone, the hydrophobic residues on the left side may interact with those of the right side of the next subunit. A similar relation exists between the top of one subunit and the bottom of each of two subunits above it. The properties of the helical array are such that each subunit is beneath one-third of one subunit and two-thirds of another. Since all subunits are identical, it is possible in theory to identify all of these interacting groups on one subunit. As the subunit widens at radii greater than 30 A, however, the space restrictions decrease. Consequently, it is no longer possible to use the model for such detailed analysis of the structure.

CALCULATION OF THE RADIAL DISTRIBUTION OF ELECTRON DENSITY FROM THE MODEL OF TMV PROTEIN

As a preliminary test of the credibility of the model, the radial distribution of electron density of the protein was calculated and compared with the values determined experimentally. For these calculations we divided the model into sectors at

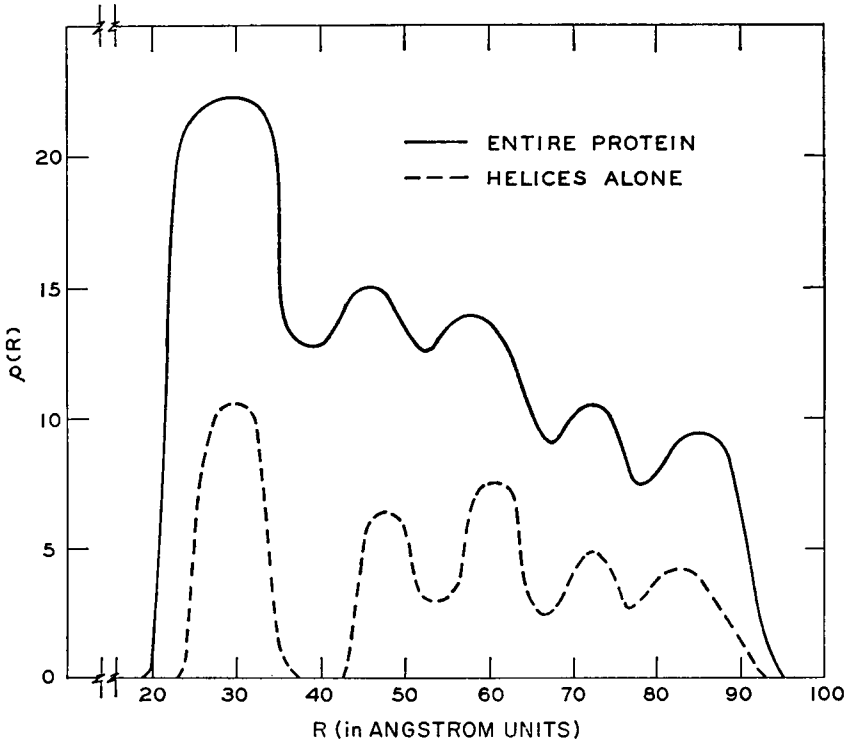

FIGURE 2 Radial distribution of electron density in the present model of TMV protein. The calculated values of $\Sigma e/R$ [designated as $\rho(R)$ on the ordinate] for the protein and for the helical segments are plotted against the radial distances (R) from the long axis of the virus particle. These values are proportional to the radial electron density.

intervals of 5 A along the radial axis, and summed the number of electrons for all amino acid residues within the sectors. Each sum, Σe, was divided by the radius, R (in A), at the midpoint of the sector to give an average value of $\Sigma e/R$ [defined as $\rho(R)$]. These values, which are proportional to the electron density ($\Sigma e/A^3$), were plotted against the radii. The density curves for the protein and for the helical segments are presented in Fig. 2.

No attempt was made to distinguish the contributions of electrons in side chains from those in the polypeptide backbone, or between electrons in helical and nonhelical segments. Such factors as thermal motion of side chains on the "surface" of the molecule lead to unequal contributions, and the curves are therefore considered only approximate guides to the true electron density.

The curve for the protein tends to follow the contours corresponding to the helices alone, although the moving of one or more nonhelical segments to equally plausible positions would alter the agreement. The relative locations of the peaks and troughs are similar to those found by Franklin et al. (1959) and by Holmes and Klug (1963) for radii of 20 to 60 A, but reflect the differences in the two sets of experimental results beyond 60 A (see Introduction). In contrast to the earlier map of density (Franklin et al., 1959), for example, both the helical projection of density (Holmes and Klug, 1963) and the curves in Fig. 2 have peaks at 60 A and general shifts of density toward larger radii. The relative heights of the peaks in Fig. 2 are similar to the experimental values in most sectors. However, the density in the region of 25–35 A is high when compared with the remaining density in the protein. Nevertheless, the general closeness of fit is encouraging in the development of a working model of the protein.

PERSPECTIVES

An opportunity to assess the procedures used for our study of TMV is afforded by the determination of the structure of bovine ribonuclease A in three laboratories (Kartha et al., 1967; Avey et al., 1967; Wyckoff et al., 1967). A model of the enzyme was constructed by the same methods. Fewer assumptions were necessary because the chemical information was more extensive than that for TMV. However, this advantage was offset by the low helical content, since the degree of helicity is the prime concern of our approach. Both the thermodynamic (Schiffer and Edmundson, 1967) and the statistical (Prothero, 1966) procedures led to high estimates of the percentage of α-helical structure. Our predictions for segments with helical potential (residues 3–11, 24–31, 43–49, 50–58, and 104–112) proved surprisingly accurate for the three regions (residues 3–11, 26–33, and 50–58) found to be helical by X-ray diffraction (Wyckoff et al., personal communication). For reasons that are not completely clear our method was not sufficiently restrictive in the two remaining cases in the enzyme.

One of the disadvantages in our type of selection is the need to consider each

segment as though it were isolated and uninfluenced by interactions with other regions of the protein. For example, the same short sequence in one protein may give rise to a different conformation in another environment.

Although the imposition of the two extra helices made parts of the model of ribonuclease too compact, the chain directions of the segments containing residues 1–58 and 84–102 were closely similar to those found in the structures proposed by Kartha et al. (1967) and Wyckoff et al. (1967). The discrepancies between the predicted and observed conformations in the carboxyl end were greater, mainly because the segment (residues 104–112) leading into the region of the active site was predicted to be in the helical rather than extended form. In regard to the small differences in the results obtained at the present stage of refinement by the three groups of crystallographers, our model is generally in closest agreement with the conclusions of Wyckoff and his colleagues.

It is clear from this discussion that the present approach should be applied mainly to proteins with helical content at least as high as that of ribonuclease. For completion of plausible models, large quantities of auxiliary chemical and X-ray information are also required, and the desirable level has not yet been reached in the case of TMV. Despite the absence of such information, the present model and its successors can still be used to interpret existing and future X-ray results and to design chemical experiments. Included in the latter category are attempts to test the tentative identifications of the carboxylate pairs and the binding sites for RNA.

We thank Miss Nancy Hutson, Miss Florence Sheber, and Mrs. Kathryn Ely for assistance and advice in the preparation of the manuscript; Mrs. Kathryn Ely for the drawings of the model; and Drs. John P. Schiffer, Robert W. Wolfgang, and John F. Thomson for encouragement and helpful discussions.

This work was supported by the U. S. Atomic Energy Commission.

Received for publication 17 April 1967.

REFERENCES

ANDERER, F. A., and D. HANDSCHUH. 1962. *Z. Naturforsch.* **17**:536.

ANDERER, F. A., H. UHLIG, E. WEBER, and G. SCHRAMM. 1960. *Nature* **186**:922.

ANDERER, F. A., B. WITTMANN-LIEBOLD, and H. G. WITTMANN. 1965. *Z. Naturforsch.* **20B**:1203.

AVEY, H. P., M. O. BOLES, C. H. CARLISLE, S. A. EVANS, S. J. MORRIS, R. A. PALMER, B. A. WOOLHOUSE, and S. SHALL. 1967. *Nature* **213**:557.

CASPAR, D. L. D. 1956. *Nature* **177**:928.

CASPAR, D. L. D. 1963. *Adv. Protein Chem.* **18**:37.

FRAENKEL-CONRAT, H. 1965. The Proteins. Academic Press, Inc., New York. **3**:99.

FRAENKEL-CONRAT, H., and K. NARITA. 1958. Symposium on Protein Structure. John Wiley & Sons, Inc., New York. 249.

FRANKLIN, R. E. 1955. *Nature* **175**:379.

FRANKLIN, R. E. 1956. *Nature* **177**:929.

FRANKLIN, R. E., D. L. D. CASPAR, and A. KLUG. 1959. Plant Pathology Problems and Progress 1908–1958. University of Wisconsin Press, Madison. 447.

FRANKLIN, R. E., and K. C. HOLMES. 1958. *Acta Cryst.* **11**:213.

FUNATSU, G., A. TSUGITA, and H. FRAENKEL-CONRAT. 1964. *Arch. Biochem. Biophys.* **105**: 25.

HOLMES, K. C., and A. KLUG. 1963. *Acta Cryst.* **16A**:79.
KARTHA, G., J. BELLO, and D. HARKER. 1967. *Nature* **213**:862.
KENDREW, J. C., H. C. WATSON, B. E. STRANDBERG, R. E. DICKERSON, D. C. PHILLIPS, and V. C. SHORE. 1961. *Nature* **190**:666.
KLUG, A., and D. L. D. CASPAR. 1960. *Advan. Virus Res.* **7**:225.
PERUTZ, M. F., J. C. KENDREW, and H. C. WATSON. 1965. *J. Mol. Biol.* **13**:669.
PROTHERO, J. W. 1966. *Biophys. J.* **6**:367.
SCHIFFER, M., and A. B. EDMUNDSON. 1967. *Biophys. J.* **7**:121.
TSUGITA, A., D. T. GISH, J. YOUNG, H. FRAENKEL-CONRAT, C. A. KNIGHT, and W. M. STANLEY. 1960. *Proc. Natl. Acad. Sci. U. S.* **46**:1463.
WITTMANN-LIEBOLD, B., and H. G. WITTMANN. 1963. *Z. Vererbungslehre* **94**:427.
WYCKOFF, H. W., T. INAGAMI, L. N. JOHNSON, K. D. HARDMAN, N. M. ALLEWELL, and F. M. RICHARDS. 1967. *Federation Proc.* **26**(2):385.