THE USE OF HELICAL NET-DIAGRAMS TO REPRESENT PROTEIN STRUCTURES

P. DUNNILL

From the Department of Botany, University College London, London, England

ABSTRACT The arrangement of side chains in α-helical regions of protein can be represented by helical net-diagrams. The addition of quantitative indices of the character of the residue at each net-point facilitates a straightforward examination of the groupings of residues of different character. As such the method complements that of Schiffer and Edmundson (1967). In addition, the reasons for nonhelicity in some regions may be seen to be correlated with particular side-chain distributions. The method is illustrated with respect to myoglobin and lysozyme.

INTRODUCTION

Recently several newly determined protein structures have been added to those of lysozyme (Blake et al., 1967) and myoglobin (Kendrew et al., 1961) (see for example, Matthews et al., 1967 and Ludwig et al., 1967). All these structures are very complex and indeed much more so than myoglobin, the first described protein structure. In seeking to understand why such structures are needed for enzyme action, how they become folded into the known conformations after synthesis and what forces stabilize them, it is useful and perhaps even essential that methods of representing them in two dimensions be developed.

Previous Two-Dimensional Representations of Protein Structure

Several approaches to this problem have already been made differing in character according to the object of the investigation. Dickerson's drawing of sperm whale myoglobin (Dickerson, 1964) elegantly conveys the general shape and broad relationships of the α-helices and non-α-helical regions of chain with the haem position shown. It is, however, still a fairly complicated drawing and the addition of all the side chains as named entities will overload it; even more so if any index of their character is to be appended.

The representation of helical and nonhelical regions of the globins by Perutz, Kendrew, and Watson (1965) is a step towards a two-dimensional convention permitting discussion of the factors involved in stabilizing particular secondary structures such as the α-helix. It does not attempt to convey relative dimensions of sec-

ondary structure or to include quantitative indices of the character of individual residues. One of the principal features which it discloses in α-helices is the tendency for nonpolar groups to occur on one side and polar groups on the other side of the helix. The effect is particularly critical for the strongly hydrophobic groups which evidently must nearly all be able to exist within the interior of the molecule.

The above principle is at the heart of the third and last representation which will be mentioned. Schiffer and Edmundson (1967) have used what they term helical wheels to summarize the arrangement of residues about an α-helix. The wheels represent the view looking down an α-helix with residues at the correct angular displacement from one another and are numbered to show the order of residues. Their representation is accurate as regards the angular dispositions of residues about α-helices and could be extended to represent the dimensions of residues in relation to the diameter of the helix core. The method is not at present intended to convey the character of residues in any detail or their longitudinal arrangement.

Net-Diagrams as a Means of Representing Protein Structure

From the examples already quoted it is clear that no single method can represent in two dimensions all the various facets of a three dimensional structure. The basic component in the present method is a helical net-diagram, Fig. 1. If a hollow cylinder represents the α-helix with residues winding around it in a spiral, the net dia-

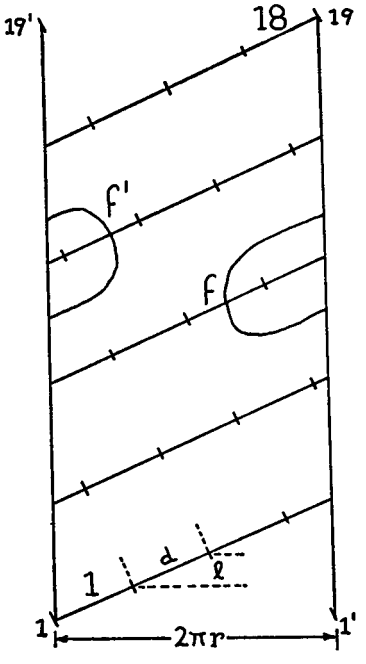

FIGURE 1 Net-diagram describing an α-helix.

gram is simply the cylinder slit open along a line parallel to its axis and laid flat. Unless stated, 'α-helix' will be taken to mean right-handed α-helix in this paper. The existence of two types of right-handed α-helix (see Nemethy et al., 1967) will be referred to later but it does not influence the present discussion. The use of similar diagrams to represent a helical array is widespread, for example, the radial projections of Fraser et al. (1965) used in discussion of synthetic polypeptide structure. The radius of the cylinder, r, was chosen as the distance from the center of the α-helix to the alpha-carbon atom. The distance between adjacent alpha-carbons, d, and the longitudinal shift per residue, l, were then given values such that the net-diagram represented the known dimensions of the α-helix. Any superimposed feature or contour which runs across the net-diagram and off one side of it must reappear on the other side, see for example feature f-f'. Aside from this need to visualize the carrying over of features across the line where the cylinder is cut, the net-diagram faithfully represents the arrangement of features about the α-helix. It may be adequate and it is sometimes useful just to write the abbreviation for each amino acid residue at each net point. When one is concerned with analyzing the effect of the character of the residues, it is more useful to have some way of representing this.

An Index of the Character of Amino Acid Residues

The protein amino acids have side chains of a very diverse nature. Apart from the aliphatics and aliphatic hydroxyl and carboxylic residues, they are not easily grouped. If one is to represent them by a few simple indices it is inevitable that some of the subtle features of each will be lost. Furthermore, one must decide whether to attempt to represent such common features as can be found, either in a qualitative, or a quantitative way. In a recent comparison of egg white and T4 lysozymes (Dunnill 1967), the symbols H = hydrophobic, p = polar, + or − = may be positively or negatively charged, and s = small were used. It was pointed out that hydrophobic residues are of critical importance. Charged residues are usually external in the molecule and small ones are often critical in terms of the packing together of separate parts of the chain. Polar residues are significant because of their ability to hydrogen bond and for their hydrophilic nature. Some residues are members of more than one class. Serine is both polar and small; lysine has charged and hydrophobic components. In addition the letters representing the various properties cannot be usefully related to one another: one cannot ask what is the effect of three H's on two p's. One would like a quantitative index representing hydrophobic character, polar character, and charge which preferably gives some indication of the size of the residue.

The basis chosen for such an index was the list made by Tanford (1962) of free energy changes. These values, which are shown in the third column of Table I, represent the change in free energy when transferring an amino acid from ethanol

TABLE I

SIDE-CHAIN CONTRIBUTIONS TO THE FREE ENERGY CHANGE FOR TRANSFER FROM ETHANOL TO WATER

	Abbreviation	Δf_u per side chain*	Hydrophobic term	Polar term	Aromatic term
		cal/molecule			
Tryptophan	Try	3000	5850	−600	−3200
Isoleucine	Ileu	2970	2600	—	—
Tyrosine	Tyr	2870	4500	−600	−1600
Phenylalanine	Phe	2650	4500	—	−1600
Proline	Pro	2600	1950	—	—
Leucine	Leu	2420	2600	—	—
Valine	Val	1690	1950	—	—
Lysine	Lys	1500	2600	−600	—
Methionine	Met	1300	1950	—	—
Alanine	Ala	730	650	—	—
Arginine	Arg	730	1950	−1800	—
Threonine	Thr	440	1300	−600	—
Serine	Ser	40	650	−600	—
Glutamic acid	Glu	550	1300	−1200	—
Glutamine	GlN	−100	1300	−1200	—
Aspartic acid	Asp	540	650	−1200	—
Asparagine	AsN	−10	650	−1200	—
Glycine	Gly	0	—	—	—
Histidine	His	—	2600	−1200	—
Cysteine	CysH	—	650	—	—

* Tanford, 1962.

to water. The value for glycine has been subtracted from each to leave the free energy term associated with the side chain. The justification for this step and the determination of the over-all values are discussed in detail in Tanford's paper. Basically, the values of Δf represent the hydrophobic character or lack of it in each residue. Without further modification they represent quite interesting simple indices of character. They were employed by Phillips (1967) in discussing the character of different parts of the lysozyme chain. In this instance, the values were plotted as ordinates against residue number as abscissae and a number of secondary symbols were added to amplify the description of each residue. In putting the Δf values on a net-diagram, one must aim at the representation which is most quickly assimilated. Numerals are thereby ruled out. Representation in terms of areas should be symmetrical about the net-points since, although many side chains are asymmetric, the path along which the longer extension lies cannot be defined a priori. Circles having an area directly related to the values were the first choice. The actual scale of the diameters was chosen to avoid overlap between the largest of them and was not directly related to the scale of the net diagram. As in the case of the diagram of Phillips (1967) described above, it was immediately found necessary to add further symbols. For example the small equal values of Δf for alanine and arginine arise

from quite different causes. That for alanine is due to the presence of one small hydrocarbon group. The value for arginine is the summation of a larger hydrophobic term due to the $-(CH_2)_3-$ chain and of the effect of polar end groups having a Δf term of opposite sign. With this fault in mind, an attempt has been made to break down the values into hydrophobic terms (hydrocarbon and sulphur) and hydrophilic terms (polar nitrogen and oxygen) by ascribing a fixed contribution per atom (Table I). Consistency with Tanford's over-all Δf values was improved by including a third factor, aromaticity, set to reduce the hydrophobic character of hydrocarbon groups. This has a physical basis in the greater electron mobility in aromatic rings. Kauzmann (1959) in a rigorous discussion of thermodynamic fac-

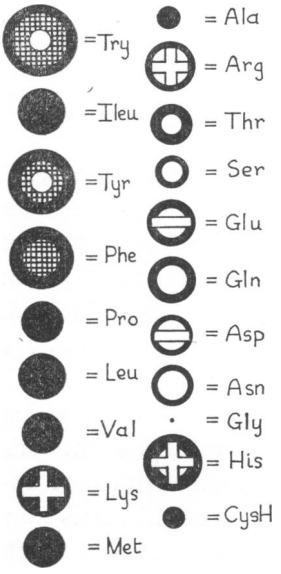

FIGURE 2 Indices of the character of individual amino acid residues.

tors involved in protein structure concluded that aliphatic hydrocarbons would be more hydrophobic than aromatic ones and that the two would therefore aggregate separately for maximum stabilization. The division of the Δf term in the above manner is of limited accuracy but it must be borne in mind that one is already quite far removed from the physical reality of the side chains so that a very precise treatment is not worthwhile. Fig. 2 shows the composite circles used to represent each side chain. The area of the inner white circles represents the polar term. The outer black circular bands denote the aliphatic hydrocarbon term, and cross hatched bands represent the aromatic hydrocarbon contribution. The only extra symbols indicate charged groups. It appears from the structures of myoglobin and lysozyme that charged groups are very rarely internal. Even lysine with a considerable aliphatic term is normally found on the surface. The fact that this consideration does not enter the Δf term may perhaps be explained by its derivation from small mole-

cule data. The instability caused by burial of charged groups within a hydrophobic core cannot arise in small molecules: micelle formation of a number of molecules with charged groups inside is not analogous. The division of the Δf term into components permits the theoretical construction of values for histidine and cysteine which were not available from Tanford's paper. It was stated earlier that the indices should if possible indicate the relative size of each residue. It is inherent in the present procedure that the circles will represent relative sizes in a qualitative way.

USE OF NET-DIAGRAMS IN THE ANALYSIS OF PARTICULAR PROTEIN STRUCTURES

The use of net-diagrams will be considered in relation to myoglobin and egg white lysozyme since these are structures for which the most detailed descriptions have been provided.

Myoglobin

The first diagram is for sperm whale myoglobin, Fig. 3. On it are superimposed contour lines to indicate the internal, external polar and external nonpolar residues specified in the paper of Perutz, Kendrew, and Watson (1965). (The present account should not be taken as the definitive presentation of the environment of globin residues.) Internal regions are contoured with a full-line and are cross-hatched. External regions are divided into polar and nonpolar by a broken line and the former are speckled.

The present method of net-diagram representation has evolved from one which the author and Prof. D. C. Phillips employed in seeking to formulate a framework within which to study the folding of proteins (see Dunnill, 1965). Any such approach will in future have to take into account the several enzyme structures now known but the net-diagram representation of myoglobin illustrated features which appeared consistent with the framework envisaged. The newer net-diagram amplifies these features.

Briefly, it was suggested that (*a*) proteins would be likely to commence folding from the *N*-terminal end before completion of synthesis, (*b*) the *N*-terminal sections would fold into well-defined conformations based particularly on hydrophobic aggregates and (with less emphasis at that time) on hydrogen-bonded aggregates, (*c*) first-formed aggregates would be expected to act as internal templates for the folding of later sections, and (*d*) simple rearrangements of sections of chain might occur in response to interaction with new groups of residues as they are synthesized but one would anticipate being able to recognize aggregates in the final molecule which are formed during synthesis.

It is in the recognition of possible aggregates in myoglobin that net diagrams are useful. Groups of hydrophobic residues occur throughout myoglobin (Fig. 3) but one may concentrate on the first third of the sequence as being most important

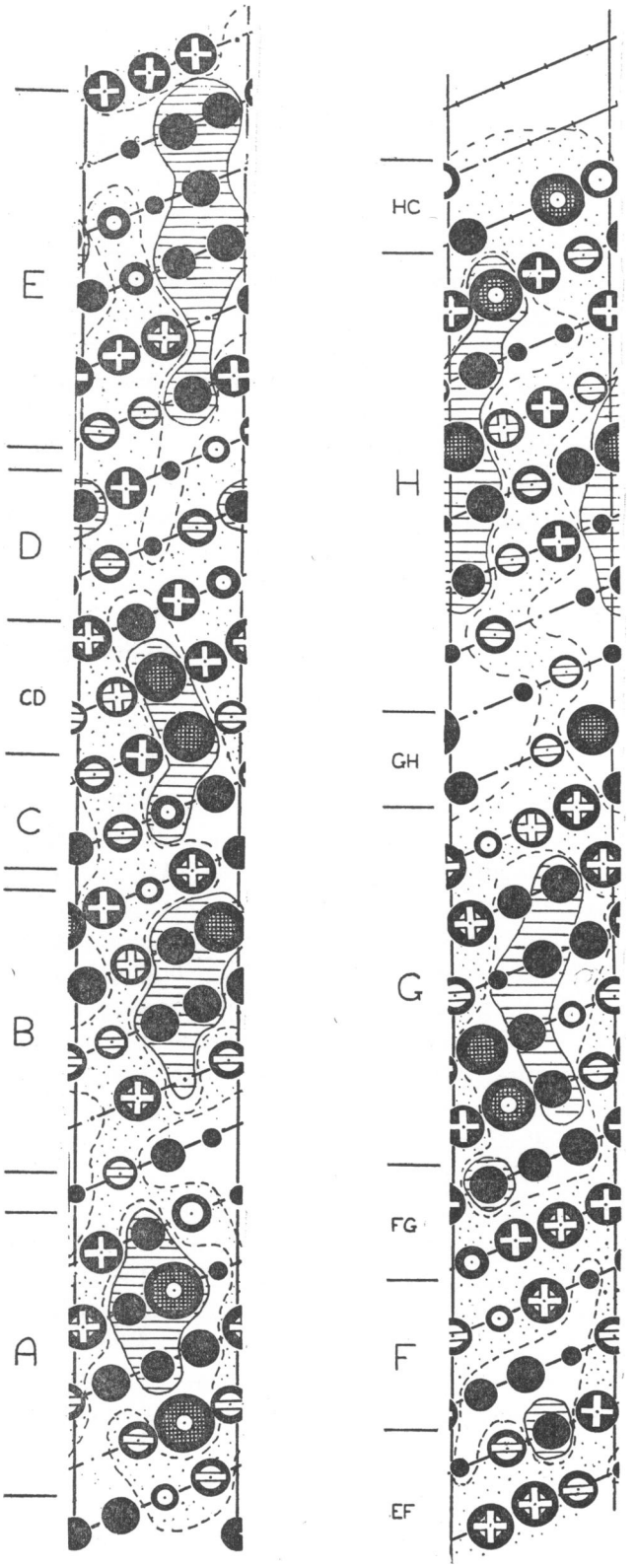

FIGURE 3 Net-diagram for sperm whale myoglobin. Single letters indicate α-helix; two letters indicate irregular folding.

as a potential internal template and least influenced by the addition of the haem group. The groups of hydrophobic residues between the *N*-terminal residue and the end of helix *D* are strikingly simple in shape when represented on a helical net. The groups in sections *A* and *B* extend in a broad band around one side of each α-helix and these are the actual conformations as they occur in native myoglobin. The group of hydrophobic residues commencing in *C* and continuing in section *CD* is markedly different. It is long, narrow and takes a helical path opposite in sense to that of the right-handed α-helix. Moreover, it is bounded by groups of charged residues similarly extended. Since sections *A* and *B* can gain both from the stabilizing effect of α-helix formation and that of compact grouping of hydrophobic residues, one may suppose that these conformations are taken up immediately following synthesis. It would be logical to surmise that the two hydrophobic groups, one in *A* and one in *B*, rest against one another at this stage. The groups of hydrophobic residues of *CD* cannot be both part of an α-helix and compact. In the final conformation, *CD* is folded to permit compact grouping and the net-diagram indicates the nature of the driving forces causing this type of folding. The discussion of subsequent steps in folding is complicated in myoglobin by the presence of the prosthetic group.

Egg White Lysozyme

The structure of egg white lysozyme as determined by Blake et al. (1967) contains much less α-helix than myoglobin (see Fig. 4). Even the sections indicated are not all of the classic type described by Pauling, Corey, and Branson (1951). Some sections have the dimensions associated with α-helix but have differently oriented bonds (Nemethy, Phillips, Leach, and Scheraga, 1967) while others are close to a 3-10 helix though again there are two possible types of bond orientation. As in the case of myoglobin this account should not be taken as a definitive presentation of the environment of side chains. The figure was prepared following the inspection of a wire model of lysozyme by courtesy of Prof. D. C. Phillips. Residues, particularly large ones, which were partially exposed are indicated by breaks in the contours. As one would expect, the collections of hydrophobic groups are mostly not of the simple compact shape found widely in myoglobin.

Helix net-diagrams do not indicate anything directly about nonhelical regions. Attempts to incorporate such data have been made. For example, the dihedral angles ϕ, ψ, which define the orientation of each N-C_α and C_α-C bond were superimposed over the α-helical net-diagram for lysozyme. Their presence tended to overload the diagram and they proved to be of no immediate value. Where one is not concerned with relating the actual structure to the hypothetical structure as α-helix, the representation used by Phillips (1967) which has already been described will be better suited.

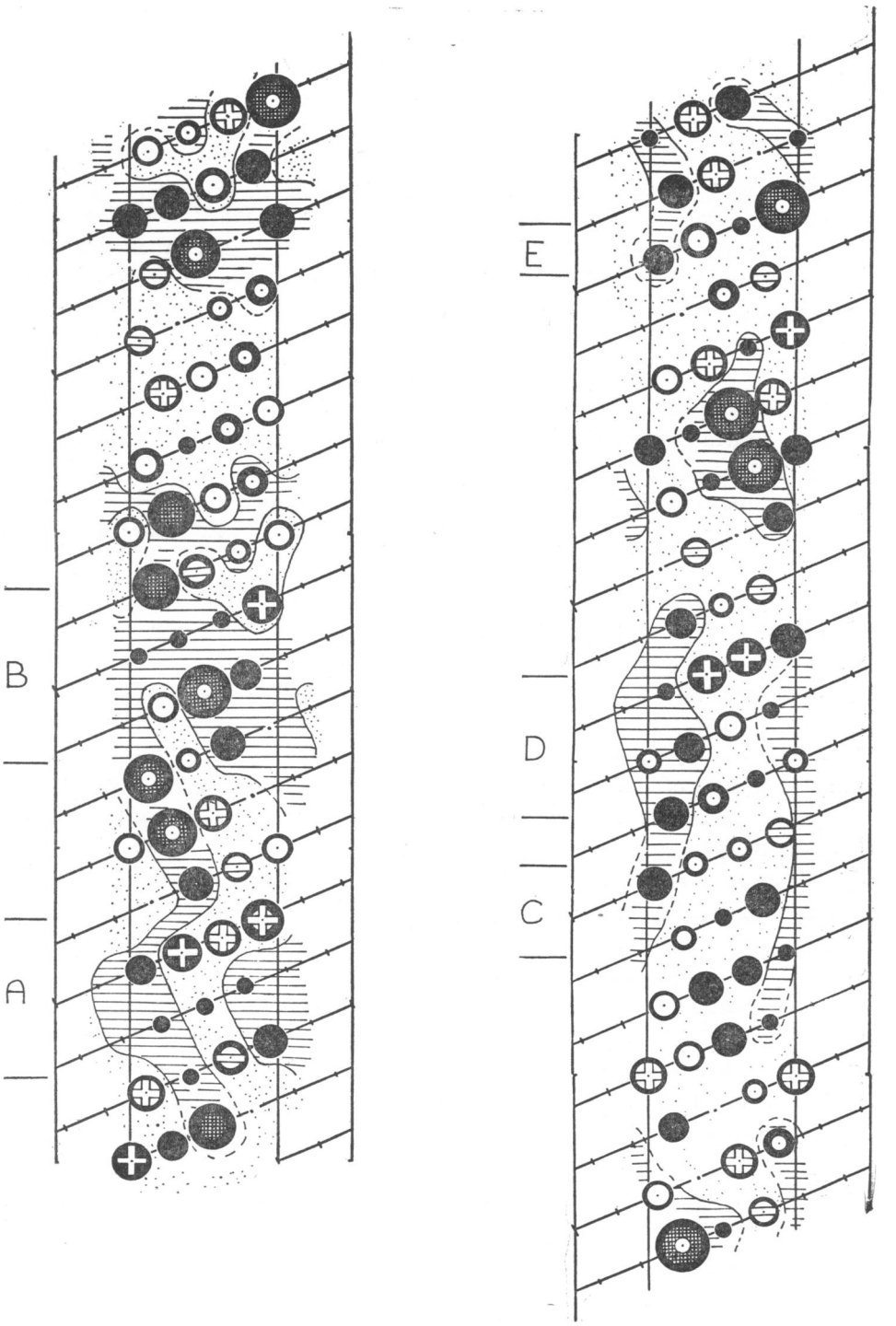

FIGURE 4 Net-diagram for egg white lysozyme. Single letters indicate helical sections.

DISCUSSION

The difficulty of representing the actual conformation of nonhelical regions places a limit on the use of net-diagrams. They can however complement the methods of Perutz et al. (1965) and Schiffer and Edmundson (1967) in aiding the detection of α-helices and do perhaps combine some of their features. In addition it is possible from them to gain some idea of the relative magnitudes of the hydrophobic and polar terms for various residues and groups of residues. It is possible in some instances to see why regions are not α-helical, for example, region *CD* in myoglobin mentioned earlier. The α-helix can be considered the "ground state" of secondary polypeptide structure in proteins. Pauling and Corey's classic studies (Pauling, Corey, and Branson, 1951) indicated that, on the basis of packing, hydrogen bonding, and bond staggering, the right-handed α-helix is one of the most favoured structures when the polypeptide chains possess β-carbon substituents on the alpha-carbons. The point is emphasized by the quantitative studies of Van der Waal's forces and hydrogen bonding by De Santis et al. (1965). All the recent discussions of factors influencing protein structure including those already mentioned, papers by Guzzo (1965) and Prothero (1966) and the present one, are principally concerned with the conditions permitting or barring this ground state. The creation of a framework for discussion of non-α-helical regions is a much more intractable problem.

The ideas of Guzzo (1965), Prothero (1966), and those concerning protein folding all imply that at least some sections of polypeptide chain have significant structural autonomy. According to ideas embodied in the former two papers, α-helices may represent such sections: if α-helicity can be confirmed to be principally dependent on primary sequence within the helix it cannot be greatly influenced by other regions of the structure. A central contention of the above folding ideas is that some regions of chain and particularly those based upon a compact group of hydrophobic residues can act as stable nuclei around which other sections fold. It is implicit that, though there may be simple rearrangements of some such nuclei, they will often maintain to a significant degree their independent conformations. If the autonomy of most α-helical conformations and hydrophobic nuclei is established in the cases of proteins whose structures are known, the identification of such regions from primary sequence data for other proteins would be useful both for setting up trial structures for X-ray analysis and in testing such ideas. An interesting feature of the net-diagrams is that the groups of like-residues seem to be arranged in fairly simple patterns. One may hope that the representation of all the known protein tertiary structures in a two-dimensional form, such as the one proposed, may show up common types of pattern of which the hydrophobic groupings associated with α-helix and hydrophobic nuclei may just be the most obvious.

The author is indebted to Prof. D. C. Phillips for encouragement and valuable comment on this subject and to Miss J. Lee for assistance in preparing the net-diagrams.

Received for publication 5 February 1968 and in revised form 18 March 1968.

REFERENCES

BLAKE, C. C. F., G. A. MAIR, A. C. T. NORTH, D. C. PHILLIPS, and V. R. SARMA. 1967. *Proc. Roy. Soc. (London), Ser. B.* **167**:365.

DESANTIS, P., E. GIGLIO, A. M. LIQUORI, and A. RIPAMONTI. 1965. *Nature.* **206**:456.

DICKERSON, R. E. 1964. *In* The Proteins. H. Neurath, editor. Academic Press, Inc., New York 2nd edition. II.

DUNNILL, P. 1965. *Sci. Prog. (London).* **53**:607.

DUNNILL, P. 1967. *Nature.* **215**:621.

FRAZER, R. D. B., B. S. HARRAP, T. P. MACRAE, F. H. C. STEWART, and E. SUZUKI. 1965. *J. Mol. Biol.* **14**:423.

GUZZO, A. V. 1965. *Biophys. J.* **5**:809.

KAUZMANN, W. 1959. *Advan. Protein Chem.* **14**:1.

KENDREW, J. C., H. C. WATSON, B. E. STRANDBERG, R. E. DICKERSON, D. C. PHILLIPS, and V. C. SHORE. 1961. *Nature.* **190**:666.

LUDWIG, M. L., J. A. HARTSUCK, T. A. STEITZ, H. MUIRHEAD, J. C. COPPOLA, G. N. REEKE, and W. N. LIPSCOMB. 1967. *Proc. Natl. Acad. Sci. U.S.* **57**:511.

MATTHEWS, B. W., P. B. SIGLER, R. HENDERSON, and D. M. BLOW. 1967. *Nature.* **214**:652.

NEMETHY, G., D. C. PHILLIPS, S. J. LEACH, and H. A. SCHERAGA. 1967. *Nature.* **214**:363.

PAULING, L., R. B. COREY, and H. R. BRANSON. 1951. *Proc. Natl. Acad. Sci. U.S.* **37**:205.

PERUTZ, M. F., J. C. KENDREW, and H. C. WATSON. 1965. *J. Mol. Biol.* **13**:669.

PHILLIPS, D. C. 1967. *Proc. Natl. Acad. Sci. U.S.* **57**:484.

PROTHERO, J. W. 1966. *Biophys. J.* **6**:367.

SCHIFFER, M., and A. B. EDMUNDSON. 1967. *Biophys. J.* **7**:121.

TANFORD, C. 1962. *J. Am. Chem. Soc.* **84**:4240.