Predicting how proteins fold up

In current and pending publications there are signs of a new, more ambitious phase in the struggle to predict the biologically active conformation of a protein from its covalent structure. Certainly there is much incentive for ambitious investigations in this field. In the right conditions, many proteins spontaneously and reversibly fold into their biologically active conformation maintained by intramolecular non-bonding forces and reducible disulphide bridges. Such observations have appropriately been interpreted as evidence that the linear sequence of amino acid residues which comprises the covalent structure of the protein carries all the necessary information for directing the folding process, and it is just a short step to speculation that artificial synthesis or genetic engineering of novel sequences will lead to conformations with novel catalytic and control functions. But which sequences will lead to the desired conformation, and achieve it in reasonable time?

The latest contenders in pursuit of the answer are Ralston and De Coen (*J. molec. Biol.*, **83,** 393; 1974) and Nagano (*ibid.*, **84,** 337; 1974). Both follow current opinion in seeking the answer in nucleation, that is, in centres of conformational structuralisation which initiate and guide the subsequent folding up of the protein molecule.

Ralston and De Coen carry out conformational energy calculations on a variety of possible nucleating structures of varying complexity, though generally on sequence fragments of six or less residue units. Principally, the object of the exercise is to find the favoured (lowest energy) conformation for a variety of short sequences. In order to reduce the computation to reasonable proportions, some severe approximations are made. In particular, these workers exploit the Liquori concept (*Q. Rev. Biophys.*, **2,** 1; 1969) of a stereochemical alphabet in which only a small number of specified conformations is allowed for any amino acid unit and interactions between units merely determine which of these is the favoured one. Furthermore, not all sidechain conformations are adjusted in the search for the lowest energy confirmation, and neither are geometric variables other than rotations around single bonds. A true minimisation of the energy of the overall system could yield rather different results, and it is surprising that these workers do not apply established minimisation procedures even where these could be easily and fruitfully applied. Nevertheless, interesting conclusions are drawn concerning the favoured conformations of many short sequences which could, within much longer sequences, present the first appearance of structural organisation during the folding process and so act as nucleation sites.

It could be argued, however, that work of this kind is premature. Perhaps another, less ambitious phase should properly precede this one—the phase of consolidation of the methods used to calculate the conformational energies. Conclusions concerning nucleation are likely to be no more credible than the least reliable conformational energy function used. Complete substantiation of the results of Ralston and De Coen awaits full minimisation and the advent of more realistic conformational energy functions, or at least the substantiation of those already in use.

For this reason the calculations of Nagano would seem to be on safer ground. Nagano belongs to the school of investigators who predict nucleating structures not on the basis of conformational energy calculations but on the basis of statistical analysis of proteins of known sequence and conformation. A crude example of this type of approach might be the observation that glutamic acid tends to begin a run of residue units in an α-helical conformation, and therefore it might be predicted to do so in a protein of unknown conformation. The approach of Nagano, however, follows current trends in being much more sophisticated and quantitative, and takes account of the effect of pairs of residue units at different separations along the sequence. In his current paper, Nagano reports on the predictability of α helix, extended chain regions which are potential candidates for β-pleated sheet structures, and looped conformations of the backbone, in proteins of recently determined conformation which do not appear in the data of the initial statistical analysis. How good are the predictions for new proteins of completely unknown conformation or when the conformation, if known, is not taken into account.

Nagano reports that the accuracy of the predictions "seems to depend very strongly on the actual data used", and concludes that his predictions are less reliable for the new proteins. The predictions of α helix are good, however, providing the protein is not rich in β-pleated sheet, and the predictions of β-pleated sheet are good if the protein is not rich in α helix. The 'goodness' of a prediction is assessed in terms of a number of different accuracy statistics.

The results reported are at first glance disappointing to those who have learned from past publications in this field to expect steadily increasing degrees of accuracy. The important point is, however, that Nagano goes on to show that predictability can be enhanced by taking into account interactions between α helix, potential β-pleated sheet and loops. In particular, the consequences of β-pleated sheet structures occurring in the vicinity of α helices are considered. Nagano is therefore taking the first really serious look at the role in the folding of proteins of interactions between nucleating structures. Connoisseurs of the art of predicting protein conformation may feel that there is nothing entirely new in his reasoning, but at the very least Nagano presents a comprehensive preview of the problems to be encountered.

Nagano concludes his analysis of nucleation with a brief discussion of the final stage of protein folding after the topological pathway of the protein backbone in three-dimensional space is roughly determined. Recognising the importance of the close packing of the hydrophobic sidechains during this stage, he would do well to consider the recent papers of Richards (*J. molec. Biol.*, **82,** 1; 1974) and Wishnia and Lappi (*ibid.*, **82,** 77; 1974). Richards presents the most detailed analysis so far of the tight packing of atoms in proteins of known conformation, and he concludes that simple geometrical packing may provide useful criteria in guiding and evaluating trial structures in simulations of protein folding. Wishnia and Lappi, on the other hand, have investigated experimentally the binding of apolar ligands to hydrophobic sites, using the interaction of cyclohexane and n-heptane with α and β-cyclodextrins as a model. If one considers the final stages of the folding of a protein as a folding of hydrophobic sidechains onto a previously established structure with a hydrophobic surface, then it is clear that work of this type will play an important part in providing thermodynamic parameters for the simulation of such a process.

What does seem clear is that the final stages of folding will be the most difficult to consider theoretically and therefore to predict. For simulating this stage of folding, the concepts and parameters obtained through investigations like those of Richards, Wishnia and Lappi will be necessary but not sufficient. The number of conformations to be treated and compared is clearly astronomical, and success will owe as much to the size and speed of future generations of computers as it does to the programs and data fed to them. But it is remarkable that one can now discuss the difficulty of predicting the final stages of folding, not the folding as a whole. This position has been achieved because of workers like Ralston, De Coen and Nagano who have done much towards reducing the number of initial and intermediate conformations in the folding process. B.R.