J. Mol. Biol. (1991) 217, 1–7

COMMUNICATIONS

A New Method for Building Protein Conformations from Sequence Alignments with Homologues of Known Structure

Timothy F. Havel

Biophysics Research Division
University of Michigan
Ann Arbor, MI 48109-2099, U.S.A.

and Mark E. Snow

Scientific Computation Group
University of Michigan
535 West William St, Ann Arbor, MI 48103, U.S.A.

(Received 4 June 1990; accepted 23 August 1990)

We describe a largely automatic procedure for building protein structures from sequence alignments with homologues of known structure. This procedure uses simple rules by which multiple sequence alignments can be translated into distance and chirality constraints, which are then used as input for distance geometry calculations. By this means one obtains an ensemble of conformations for the unknown structure that are compatible with the rules employed, and the differences among these conformations provide an indication of the reliability of the structure prediction. The overall approach is demonstrated here by applying it to several Kazal-type trypsin inhibitors, for which experimentally determined structures are available. On the basis of our experience with these test problems, we have further predicted the conformation of the human pancreatic secretory trypsin inhibitor, for which no experimentally determined structure is presently available.

Because of the difficulties and uncertainties that are often involved in the determination of the three-dimensional structures of proteins by either X-ray crystallography or nuclear magnetic resonance (n.m.r.†) spectroscopy, considerable effort has been devoted to methods of predicting their conformations by means of analogy with homologous proteins whose structures are already available (for reviews, see Blundell *et al.*, 1988; Feldmann *et al.*, 1985; Greer, 1985*b*; Swenson *et al.*, 1978). The basic idea on which these predictions are based is simply that, since the three-dimensional structures of proteins tend to be conserved while their sequences evolve, the conformations and relative positions of the amino acid residues of the unknown structure should be similar to those of the corresponding homologous residues in the known structures.

Although this idea is clearly sound and has been successfully applied to a number of proteins (e.g. see Blundell *et al.*, 1983; Chothia *et al.*, 1986; Greer, 1985*a*; Palmer *et al.*, 1986), there is no general consensus on what is the best way to put it into practice. In the first place, the decision as to which pairs of residues, one in the unknown structure and the other in the known structure, "correspond" depends on how the protein sequences have been aligned. In the second, since there are usually insertions and deletions between these sequences, it is rarely possible to make the backbone conformation of the unknown protein coincide exactly with those of the known proteins, and considerable care is required in order to model the conformations of inserted loops correctly (Bruccoleri & Karplus, 1987; Moult & James, 1986). Although it is generally easier to predict the conformations of the

† Abbreviations used: n.m.r., nuclear magnetic resonance; r.m.s., root-mean-square; BUSI, bull seminal inhibitor, PPSTI, porcine pancreatic secretory trypsin inhibitor; HPSTI, human pancreatic secretory trypsin inhibitor; JQOM3, Japanese quail ovomucoid third domain; SPOM3, silver pheasant ovomucoid third domain; r.m.s.d., root-mean-square co-ordinate deviation; PPAD, ϕ,ψ angle difference.

0022-2836/91/010001-07 $03.00/0

side-chains of mutated residues, even this problem is not entirely trivial (Snow & Anzel, 1986; Summers & Karplus, 1989).

Sequence alignment algorithms, though based on heuristic weights, at least provide us with an objective criterion by which to obtain the correspondence between the sequences. Other considerations, for example the fact that the spatial structure of the interior core of a protein is usually better conserved than the surface (Hubbard & Blundell, 1987; Sutcliffe *et al.*, 1987), can also be incorporated into the results. The process of actually building a "similar" three-dimensional model, however, remains relatively laborious and somewhat subjective despite substantial efforts at automating the procedure (see above references). In addition, in most methods it is necessary to refine the resultant conformation *versus* a potential energy function in order to obtain good covalent geometry and atom packing.

Here, we report our preliminary experience with a new method of deriving the three-dimensional conformation of a protein from a given (and possibly multiple) sequence alignment with homologue(s) of known structure. This method is based on the EMBED distance geometry algorithm (Crippen & Havel, 1988; Havel *et al.*, 1983), which has been used extensively for the calculation of protein structures from n.m.r. data (Wüthrich, 1986). As a general-purpose tool for protein model-building, the EMBED algorithm offers the following distinct advantages over the usual procedure of r.m.s. superposition of substructures followed by manual adjustment of the result.

(1) It is efficient, not just in terms of computer time but also, and more importantly, in terms of people's time.

(2) It requires its users to state their geometric hypotheses concisely, and it helps them to discover contradictions in these hypotheses.

(3) Once these hypotheses have been stated, the computation proceeds automatically and completely free of possible user bias.

(4) Multiple conformations consistent with the hypotheses are generally found, and the magnitude of the differences between them provides an estimate of the reliability and precision of the predictions.

(5) Although these conformations do not necessarily have a low total energy, they are at least free of significant covalent distortions and unacceptable steric overlaps before energy minimization.

The input to the EMBED algorithm consists of a list of lower and upper bounds on the possible values of the interatomic distances in the molecule, together with the chiralities of its asymmetric centres. This type of information will henceforth be referred to as distance and chirality constraints. As described by Crippen & Havel (1988), these constraints are available in abundance from the primary structure of a protein, and further constraints, e.g. hydrogen bond lengths, can be derived from the secondary structure. The output of the EMBED algorithm consists of a set of conformations that are consistent with the given information, but that are otherwise "random". Hence, any geometric features that are uniformly present in all members of such a *conformational ensemble* can be inferred to be necessary consequences of the given constraints. In addition, if the algorithm fails to find any conformations consistent with the constraints, we can infer that the constraints are mutually contradictory, i.e. at least one of the assumptions on which they are based must be in error. When dealing with the many unknowns present in evolution, such checks can be extremely valuable.

The calculations reported here were done using the implementation of the EMBED algorithm known as the DISGEO program (Havel & Wüthrich, 1984) version 3.0.3 running on an IBM 3090/600E. Only the heavy (i.e. non-hydrogen) atoms were included in these calculations, and neither substructures nor metrization were used, since it has been shown that the implementation of these procedures in the DISGEO program can lead to sampling problems (Havel, 1990). To improve the convergence, a preliminary refinement of the error function was performed in four dimensions before projection into three dimensions and a final refinement. With this procedure together with the large quantities of uniformly distributed distance information available from evolutionary studies, we have obtained reasonably good convergence, with a failure rate of less than one in four attempts to calculate a conformation.

The first step of the prediction procedure is to align the sequence of the unknown protein structure with the sequences of the known structure(s). In the calculations reported here, this was done using the GCG sequence analysis package (Devereux *et al.*, 1984), and iteratively applying the Needleman–Wunsch algorithm (Needleman & Wunsch, 1970) with identity matrix weights to all pairs of sequences. In order to translate the resulting multiple sequence alignment into distance constraints, we first identified a subset of the atoms consisting of all C^{α} atoms together with the C^{γ} atoms in those residues that were the same in all of the aligned sequences. Lower and upper bounds on the distances among these atoms of the unknown structure were then derived from the formula:

$$\overline{d_{ij}} \pm (\sigma_d \Lambda_{ij}(1+\Delta_{ij}/4) + \tau_d(1+\Delta_{ij}/2)),$$

where d_{ij} is the average distance between atoms i and j in the known structures, Λ_{ij} was half of the observed range in value of this distance, Δ_{ij} is the sum of the lengths of any intervening gaps in the alignment, and σ_d and τ_d are adjustable parameters. In addition, the spatially aligned known structures were examined in order to identify hydrogen bonds and disulphides that were conserved during evolution, and when identical residues were present in the unknown protein, distance constraints were imposed that ensured that these same interactions were present also in the computed conformations.

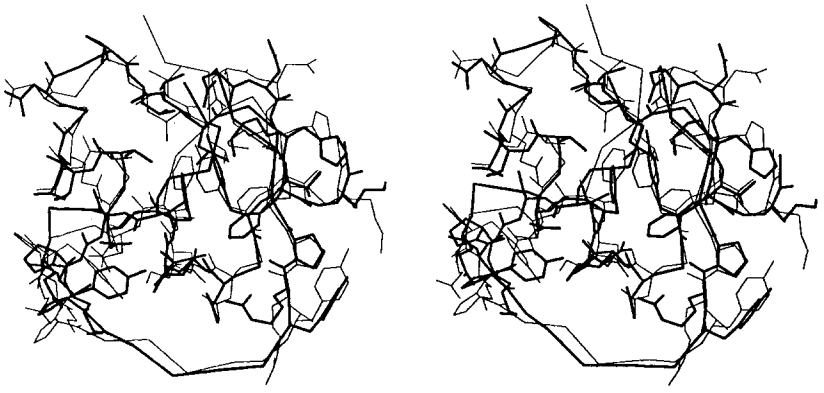

Figure 1. Stereoview of the best-convergent SPOM3 structure from DISGEO (heavy lines) superimposed on the crystal structure (light lines) so as to minimize the C^{α} r.m.s.d. between the C-terminal 52 residues. Only C^{α} atoms and side-chains of the residues that are the same in JQOM3 have been included.

Finally, in order to ensure that the secondary structure of the computed conformations coincides with the secondary structures of the known conformations to the extent that the secondary structures have been conserved in them, lower and upper bounds on the oriented volumes among each four consecutive C^{α} atoms along the sequence were derived by the formula:

$$\overline{v_{hijk}} \pm (\sigma_v \Lambda_{hijk} + \tau_v),$$

where $\overline{v_{hijk}}$ is the average oriented volume spanned by the four C^{α} atoms (i.e. half the triple product of the virtual bond vectors connecting them: Crippen & Havel, 1988; Kabsch & Sander, 1983), and the parameters σ_v and τ_v will generally be different from the corresponding parameters σ_d and τ_d for the distance bounds.

Clearly, these rules for deriving geometric information from sequence alignments are somewhat arbitrary, and we anticipate that they can and will be improved as our experience with them accumulates. Our immediate concern has been simply to demonstrate that, when used with any reasonable set of such rules, the EMBED algorithm is capable of efficiently generating reasonable protein conformations. For this purpose, we have used the Kazal-type trypsin inhibitors BUSI, PPSTI, SPOM3 and JQOM3. A low-resolution n.m.r. structure is available for BUSI (with a mean backbone r.m.s.d. among the ensemble of 3 Å: Williamson *et al.*, 1985), and high-resolution crystal structures are available for PPSTI (1·8 Å resolution: Bolognesi *et al.*, 1982), SPOM3 (1·5 Å resolution: Bode *et al.*, 1985) and JQOM3† (1·9 Å resolution: Papamŏkos *et al.*, 1982).

As our first problem, we treated the conformation of SPOM3 as unknown and predicted it from that of JQOM3. These two proteins have a sequence identity of 89% and can be aligned without any gaps. In the above rules for deriving geometric constraints from this alignment, we set $\tau_d = 2$ Å and $\tau_v = 20$ Å^3 (the values of σ_d and σ_v are irrelevant, since we had only 1 known structure). In addition to the distance and chirality constraints obtained by these rules, "supplementary" distance constraints were imposed in order to ensure that the three disulphide and 27 hydrogen bonds one would expect to be conserved were also present in the computed conformations. With this geometric information as input, the EMBED algorithm easily produced an ensemble of ten SPOM3 conformations with a mean C^{α} r.m.s.d. among them of 0·76 Å, an all-atom r.m.s.d. of 1·43 Å, and a PPAD of 60·9°.

The corresponding mean differences between the computed conformations and the crystal structure of JQOM3 were 0·78 Å for the C^{α} r.m.s.d. and 94·8° for the PPAD, thus showing that the DISGEO program faithfully produced structures whose backbone conformation is quite similar to JQOM3. The mean differences between the computed conformations and SPOM3 were 2·36 Å for the C^{α} r.m.s.d., 2·60 Å for the all-atom r.m.s.d. and 63·6° for the PPAD. These differences are comparable to the corresponding differences between the JQOM3 and SPOM3 crystal structures, which are 2·28 Å for the C^{α} r.m.s.d. and 84·6° for the PPAD. Most of these differences, however, are due to the drastically different conformations of the N-terminal four residues: when these are deleted, the C^{α} r.m.s.d. between the crystal structures falls to only 0·76 Å, while the C^{α} and all-atom r.m.s.d. values between the computed conformations and the SPOM3 crystal structure falls to only 1·10Å and 1·82 Å, respectively. A detailed drawing of one of the computed structures superimposed upon the SPOM3 crystal structure may be found in Figure 1, from which it may be seen that despite the overall similarity, significant differences still exist, especially in the orientations of the side-chains. These differences, however, could be observed also between the computed structures themselves, thus showing that they are allowed by our geometric hypotheses and should be considered as possible alternatives in the course of any careful structure prediction.

† Since the unit cell of the crystal structure of JQOM3 contains 4 non-equivalent but highly similar copies of the same molecule, we arbitrarily chose one of these for this study. (1 Å = 0·1 nm.)

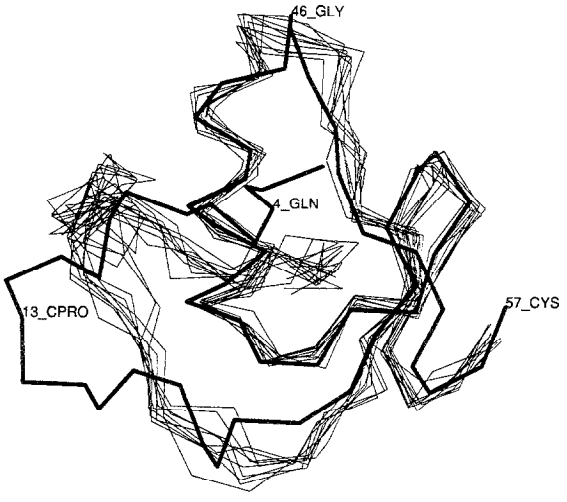

Figure 2. The C^{α} trace of the 10-structure ensemble computed for BUSI superimposed upon the C^{α} trace of the n.m.r. conformation (heavy line) so as to minimize the C^{α} r.m.s.d. A few residues have been labelled for reference.

In our next computational experiment, we predicted the conformation of BUSI from a multiple sequence alignment with the ovomucoids JPOM3 and SPOM3. In this case, the sequence alignments of BUSI with the ovomucoids began at residue number 2 and contained a single gap where the ovomucoids were missing residues 11 and 12 of BUSI. The overall identity of BUSI with JQOM3 and SPOM3 was 47% and 48%, respectively. As previously noted, the C^{α} r.m.s.d. between the ovomucoid structures was 2·28 Å, primarily because of the differences in the N-terminal four residues. The distance and chirality constraints were obtained by the same formula, with $\tau_d = 2$ Å and $\tau_v = 20$ Å^3 as before, and $\sigma_d = \sigma_v = 2$. In addition, the three disulphides and 16 invariant hydrogen bonds were included as supplementary constraints.

The mean C^{α} and all-atom r.m.s.d. values among the ten resulting BUSI conformations were 1·38 Å and 2·25 Å, respectively, while the mean PPAD was 74°. The mean C^{α} and all-atom r.m.s.d. values with the n.m.r. conformation, on the other hand, were 3·88 Å and 5·00 Å, while the PPAD was 79°. The large r.m.s.d. values were due in part to the N-terminal four residues, but also to the significant differences that exist between the n.m.r. conformation of BUSI and the ovomucoid conformations in the extended loop from residues 9 to 16 of BUSI (where the deletion has occurred). As expected, the backbone of our computed BUSI conformations tended to follow those of the ovomucoids in this loop as well as elsewhere, giving a mean C^{α} r.m.s.d. between the BUSI ensemble and the JQOM3 and SPOM3 crystal structures of 1·44 Å and 2·04 Å, respectively. The C^{α} r.m.s.d. values between the BUSI n.m.r. conformation and the corresponding residues of the JQOM3 and SPOM3 crystal structures were 3·56 Å and 3·83 Å. A drawing of the C^{α} trace of the ten computed BUSI conformations superimposed on the n.m.r. conformation is shown in Figure 2.

A final run was made with BUSI using the above distance and chirality constraints, together with a collection of 202 distance constraints derived from n.m.r. spectroscopy (Williamson *et al.*, 1985). Since n.m.r. experiments usually yield large numbers of short-range† distance and chirality constraints, but are sometimes lacking in sufficient numbers for the long-range constraints that are available in abundance from evolutionary studies, such studies may prove to be an extremely useful supplement to n.m.r. structure determinations (and *vice versa*). Although the mean C^{α} r.m.s.d. among the computed structures decreased a little (to 1·31 Å), the mean C^{α} r.m.s.d. with the n.m.r. structure remained high (3·78 Å). Furthermore, the residual violations of the distance constraints, though not much larger than in the previous run, were far more numerous. This indicates that the evolutionary constraints and the n.m.r. constraints were not entirely compatible, and hence that our choice of σ and τ as above were probably smaller than they should be for the degree of sequence identity that exists here.

In order to obtain some feeling for the relation between the parameters σ and τ, and the accuracy and precision with which a structure is determined when the sequence identity is quite low, we next performed a series of three runs in which we attempted to derive the PPSTI structure from those of the same two ovomucoids, and compared the resulting ensembles with its crystal structure (Bolognesi *et al.*, 1982). The sequence alignment used contained no gaps, and the sequence identities of PPSTI with JQOM3 and SPOM3 were 30% and 21%, respectively. The first of these runs (I) used $\sigma_d = \sigma_v = 2$, $\tau_d = 1$ and $\tau_v = 10$, while the second (II) used $\sigma_d = \sigma_v = 2$, $\tau_d = 2$ and $\tau_v = 20$ (which are the values used elsewhere in this paper), and the third (III) used $\sigma_d = \sigma_v = 3$, $\tau_d = 3$ and $\tau_v = 20$. In addition the constraints in all three runs included 15 hydrogen bonds and three disulphides.

The mean C^{α} r.m.s.d. and PPAD values among the resultant ensembles and with the PPSTI, JQOM3 and SPOM3 X-ray structures are shown in Table 1. As expected, the greatest differences between the computed structures and the PPSTI crystal structure occurred in the N-terminal 20 residues, where the sequence identity is lowest and the two ovomucoid crystal structures also exhibit significant differences. Another significant difference occurred in the position of the triple-stranded β-sheet, which was displaced by about 4 Å towards the N terminus in the computed structures, a difference that could be seen also in the ovomucoid structures. The C^{α} traces of the computed conformations superimposed on those of the ovomucoid crystal structures are shown in Figure 3.

† The terms short-range and long-range refer to the number of covalent bonds separating the pair of atoms in question, and not to their spatial proximities.

Table 1
Mean C^{α} r.m.s.d. and PPAD values

Average taken over	Difference	Run I	Run II	Run III
All pairs of computed structures	r.m.s.d.	0·98	1·22	1·32
	PPAD	68·1	73·7	76·6
PPSTI X-ray plus computed structures	r.m.s.d.	3·74	3·63	4·00
	PPAD	76·4	75·0	78·2
JQOM3 X-ray plus computed structures	r.m.s.d.	1·33	1·38	1·61
	PPAD	91·1	89·9	93·4
SPOM3 X-ray plus computed structures	r.m.s.d.	2·27	2·16	2·68
	PPAD	82·8	82·1	87·7

See the text for explanations.

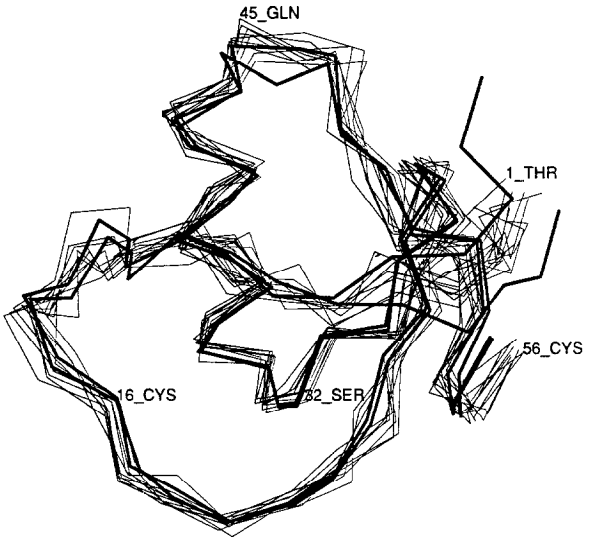

Figure 3. The C^{α} trace of the 10-structure ensemble computed for PPSTI in run II, together with the C^{α} trace of the SPOM3 crystal structure, all superimposed on the crystal structure of JQOM3 so as to minimize the C^{α} r.m.s.d. The 2 ovomucoids JQOM3 and SPOM3 have been drawn with a heavy line, and a few residues have been labelled for reference.

The fact that the differences to the PPSTI and ovomucoid structures increased more rapidly than the differences of the computed conformations with one another in going from (I) to (III) is a consequence of the fact that with loose constraints the distance geometry program used here tends to produce uniformly expanded structures. This tendency could be eliminated by using a new distance geometry program recently developed by the first author, and a more thorough study of the relation between the degree of sequence identity, the rules by which sequence alignments are translated into distance constraints, and the accuracy and precision of the results with this new program is in progress. Nevertheless, in order to keep the results of this preliminary study comparable with one another, we decided to finish our work with the Kazal family of trypsin inhibitors using the DISGEO program. As we have seen, with the σ and τ parameters of run II the r.m.s.d./PPAD among the members of the computed ensembles and with the structures from which the constraints were derived are roughly comparable over a wide range of sequence identities, indicating that reasonably good sampling is obtained within the allowed range. In addition, our experience with the ovomucoids indicates that this choice of parameters is reasonable when the sequence identity is sufficiently high, in that the actual structure (whatever it may be) largely satisfies the resultant constraints and hence could be found by the DISGEO program.

As our final project, we proceeded to predict the structure of the human inhibitor HPSTI, for which no experimental structure currently is available, from the PPSTI crystal structure. The sequence alignment again contained no gaps, with an overall sequence identity of 73%. Using the values of σ and τ from run II above, the distance constraints implied by this alignment were combined with the three disulphide bonds and the 24 invariant hydrogen bonds, and used as input for the DISGEO program. The resulting ensemble of ten conformations had a mean C^{α} r.m.s.d. of 0·83 Å, a mean all-atom r.m.s.d. of 1·74 Å and a mean PPAD of 68·3°. The mean C^{α} r.m.s.d. to the PPSTI crystal structure was 1·00 Å, and the corresponding PPAD was 93·4°. On the basis of our experience with predicting SPOM3 from JQOM3, we believe that if the N-terminal eight residues preceding the first disulphide are deleted, these structures are all within 1·5 Å of the "actual" HPSTI structure in their C^{α} r.m.s.d., and perhaps 2·5 Å in their all-atom r.m.s.d.

In order to optimize the atom packing, torsional angle distributions and other fine details of the structure that are difficult to model by means of geometric constraints alone, the individual members of this ensemble were subjected to restrained energy minimization. These minimizations were performed using a version of the AMBER program (Weiner *et al.*, 1986) that was modified so that consistency with the distance constraints could be enforced by adding an appropriate pseudopotential onto the energy. Although the energies of the structures obtained from the DISGEO program all exceeded 1000 kcal (1 kcal = 4·184 J), after only 500 cycles of conjugate gradient minimization the energies ranged from −593 to −244 kcal. As can be seen from the superposition of the lowest energy conformation before and after minimization (Fig. 4), these minimizations did not make any significant changes to the conformations obtained directly from the DISGEO program, i.e. the geometric constraints used as input for that program were already sufficient to ensure that the resulting conformations were all very close to low-energy conformations. The four lowest energy structures again differed by exactly 0·83 Å in their mean C^{α} r.m.s.d., 1·74 Å in their mean all-atom r.m.s.d., and 67·3° in their mean PPAD. We conclude that any successful attempt to reduce the range of structural possibilities by means of energetic considerations will have to be based on

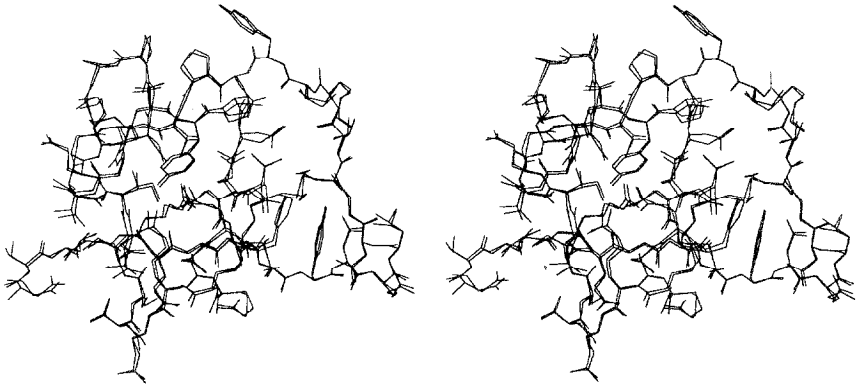

Figure 4. Stereoview of the best-convergent HPSTI conformation from DISGEO superimposed on the same structure after restrained energy minimization so as to minimize the all-atom r.m.s.d.

a much more careful analysis, including solvation, electrostatic and entropic considerations.

In conclusion, we have shown that the EMBED distance geometry algorithm (Crippen & Havel, 1988) provides protein modellers with a powerful and as yet under-utilized tool by which complete protein structures can be built with relatively little effort. Although the validity of the results obtained depends upon many other factors, e.g. the percentage identity, the sequence alignment weights and the rules by which the alignments are translated into distance constraints, this saving in effort enables one to spend more time experimenting with these other parameters. In addition, the result of these calculations consists of an entire ensemble of conformations consistent with the geometric hypotheses, which enables one to explore systematically the necessity and sufficiency of different hypotheses.

The methodology described here should generalize easily to much larger proteins, and many improvements in our procedure for deriving distance constraints from sequence alignments are clearly possible. In addition, the rotomeric preferences of the individual amino acids (Ponder & Richards, 1987) and energetic considerations (Summers *et al.*, 1987) could be used to predict the conformations of side-chains that are underdetermined by geometric constraints. We are in the process of refining the procedure used here with the larger and more complicated plastocyanin protein family, whose structures and evolutionary history have been studied extensively (Chothia & Lesk, 1982). In addition, we are working on a *de novo* prediction of the structure of the flavodoxin from *Escherichia coli*, which has been cloned and whose crystal structure is now in the process of being determined (M. L. Ludwig, R. G. Matthews & C. Osborne, unpublished results).

This work has been supported by NIH grant R01 GM37708. We thank the Computer Allocation for Computational Sciences Program of the University of Michigan for computer time.

References

Blundell, T. L., Sibanda, B. L. & Pearl, L. (1983). Three-dimensional Structure, Specificity and Catalytic Mechanism of Renin. *Nature (London)*, **304**, 273–275.

Blundell, T. L., Carney, D., Gardner, S., Hayes, F., Howlin, B., Hubbard, T., Overington, J., Singh, D. A., Sibanda, B. L. & Sutcliffe, M. (1988). Knowledge-based Protein Modelling and Design. *Eur. J. Biochem.* **172**, 513–520.

Bode, W., Epp, O., Huber, R. & Laskowski, M., Jr & Ardelt, W. (1985). The Crystal and Molecular Structure of the Third Domain of Silver Pheasant Ovomucoid. *Eur. J. Biochem.* **147**, 387–399.

Bolognesi, M., Gatti, G., Menegatti, E., Guarneri, M., Marquart, M., Papamokos, E. & Huber, R. (1982). Three-dimensional Structure of the Complex between Pancreatic Secretory Trypsin Inhibitor (Kazal Type) and Trypsinogen at 1·8 Å Resolution. *J. Mol. Biol.* **162**, 839–868.

Bruccoleri, R. E. & Karplus, M. (1987). Conformational Sampling. *Biopolymers*, **26**, 137–168.

Chothia, C. & Lesk, A. M. (1982). Evolution of Proteins Formed by β-sheets. I. Pastocyanin and Azurin. *J. Mol. Biol.* **160**, 309–323.

Chothia, C., Lesk, A. M., Levitt, M., Amit, A. G., Mariuzza, R. A., Phillips, S. E. V. & Poljak, R. J. (1986). The Predicted Structure of Immunoglobulin D1.3 and Its Comparison with the Crystal Structure. *Science*, **233**, 755–758.

Crippen, G. M. & Havel, T. F. (1988). *Distance Geometry and Molecular Conformation*. Research Studies Press, Letchworth, U.K., ISBN 0-86380-073-4. Publisher in the U.S. is J. Wiley & Sons, New York, NY, ISBN 0-471-92061-4.

Devereux, J., Heberli, P. & Smithies, O. (1984). A Nucleic Acid Analysis Program for the VAX. *Nucl. Acids Res.* **12**, 387–395.

Feldmann, R. J., Bing, D. H., Potter, M., Mainhart, C., Furie, B., Furie, B. C. & Caporale, L. H. (1985). On the Construction of Computer Models of Proteins by the Extension of Crystallographic Structures. In *Macromolecular Structure and Specificity: Computer-assisted Modeling and Applications* (Venkataraghavan, B. & Feldmann, R. J., eds), vol. 439, pp. 12–30, Annals New York Academy of Sciences, New York.

Greer, J. (1985*a*). Molecular Structure of the Inflammatory Protein C5a. *Science*, **228**, 1055–1060.

Greer, J. (1985*b*). Protein Structure and Function by Comparative Model Building. In *Macromolecular Structure and Specificity: Computer-assisted Modeling and Applications* (Venkataraghavan, B. & Feldmann, R. J., eds), vol. 439, pp. 44–63, Annals New York Academy of Sciences, New York.

Havel, T. F. (1990). The Sampling Properties of Some Distance Geometry Algorithms Applied to Unconstrained Polypeptide Chains: A Study of 1830 Independently Computed Conformations. *Biopolymers*, **29**, 1565–1585.

Havel, T. F. & Wüthrich, K. (1984). A Distance Geometry Program for Determining the Structures of Small Proteins and Other Macromolecules from Nuclear Magnetic Resonance Measurements of ^1H–^1H Proximities in Solution. *Bull. Math. Biol.* **46**, 673–698.

Havel, T. F., Kuntz, I. D. & Crippen, G. M. (1983). Theory and Practice of Distance Geometry. *Bull. Math. Biol.* **45**, 665–720.

Hubbard, T. J. P. & Blundell, T. L. (1987). Comparison of Solvent-inaccessible Cores of Homologous Proteins: Definitions Useful for Protein Modelling. *Protein Eng.* **1**, 159–171.

Kabsch, W. & Sander, C. (1983). Dictionary of Protein Structure: Pattern Recognition by Hydrogen-bonded and Geometrical Features. *Biopolymers*, **22**, 2577–2637.

Moult, J. & James, M. N. G. (1986). An Algorithm for Determining the Conformation of Polypeptide Segments in Proteins by Systematic Search. *Proteins*, **1**, 146–163.

Needleman, S. B. & Wunsch, C. D. (1970). A General Method Applicable to the Search for Similarities in the Amino-acid Sequence for Two Proteins. *J. Mol. Biol.* **48**, 443–453.

Palmer, K. A., Scheraga, H. A., Riordan, J. F. & Vallee, B. L. (1986). A Preliminary Three-dimensional Structure of Angiogenin. *Proc. Nat. Acad. Sci., U.S.A.* **83**, 1965–1969.

Papamokos, E., Weber, E., Bode, W. & Huber, R. (1982). Crystallographic Refinement of Japanese Quail Ovomucoid, a Kazal Type Inhibitor, and Model Building Studies of Complexes with Serine Proteases. *J. Mol. Biol.* **158**, 515–537.

Ponder, J. W. & Richards, F. M. (1987). Tertiary Structure Templates for Proteins: Use of Packing Criteria in the Enumeration of Allowed Sequences for Different Structural Classes. *J. Mol. Biol.* **193**, 775–791.

Snow, M. E. & Amzel, L. M. (1986). Calculating Three-dimensional Changes in Protein Structure Due to Amino-acid Substitutions: The Variable Region of Immunoglobulins. *Proteins*, **1**, 267–279.

Summers, N. L., Carlson, W. D. & Karplus, M. (1987). Analysis of Side-chain Orientations in Homologous Proteins. *J. Mol. Biol.* **196**, 175–198.

Summers, N. L. & Karplus, M. (1989). Construction of Side-chains in Homology Modelling. *J. Mol. Biol.* **209**, 785–811.

Sutcliffe, M. J., Haneef, I. D. C. & Blundell, T. L. (1987). Knowledge-based Modelling of Homologous Proteins. Part I: Three-dimensional Frameworks Derived by the Simultaneous Superposition of Multiple Structures. *Protein Eng.* **1**, 377–384.

Swenson, M. K., Burgess, A. W. & Scheraga, H. A., (1978). Conformational Analysis of Polypeptides : Application to Homologous Proteins. In *Frontiers in Physiochemical Biology* (Pulman, B., ed.), Academic Press, New York.

Weiner, S., Kollman, P., Nguyen, D. & Case, D. (1986). An All Atom Force Field for Simulations of Proteins and Nucleic Acids. *J. Comput. Chem.* **7**, 230–252.

Williamson, M. P., Havel, T. F. & Wüthrich, K. (1985). Tertiary Structure of the Proteinase Inhibitor IIA from Bull Seminal Plasma. *J. Mol. Biol.* **182**, 295–315.

Wüthrich, K. (1986). *NMR of Proteins and Nucleic Acids*, J. Wiley & Sons, New York, NY. ISBN 0-471-82893-9.

Edited by P. E. Wright