A combinatorial distance-constraint approach to predicting protein tertiary models from known secondary structure

Gareth Chelvanayagam*, Lukas Knecht, Thomas Jenny, Steven A Benner† and Gaston H Gonnet

Address: Computational Chemistry Group, Universitätstrasse 16, ETH Zentrum, Zürich, CH 8092, Switzerland.

Present addresses: *John Curtin School of Medical Research, Australian National University, Canberra ACT, Australia. †Department of Chemistry, University of Florida, Gainesville, FL 32611, USA.

Correspondence: Gareth Chelvanayagam
E-mail: gareth.chelva@anu.edu.au

Background: Distance geometry methods allow protein structures to be constructed using a large number of distance constraints, which can be elucidated by experimental techniques such as NMR. New methods for gleaning tertiary structural information from multiple sequence alignments make it possible for distance constraints to be predicted from sequence information alone. The basic distance geometry method can thus be applied using these empirically derived distance constraints. Such an approach, which incorporates a novel combinatoric procedure, is reported here.

Results: Given the correct sheet topology and disulfide formations, the fully automated procedure is generally able to construct native-like Cα models for eight small β-protein structures. When the sheet topology was unknown but disulfide connectivities were included, all sheet topologies were explored by the combinatorial procedure. Using a simple geometric evaluation scheme, models with the correct sheet topology were ranked first in four of the eight example cases, second in three examples and third in one example. If neither the sheet topology nor the disulfide connectivities were given *a priori*, all combinations of sheet topologies and disulfides were explored by the combinatorial procedure. The evaluation scheme ranked the correct topology within the top five folds for half the example cases.

Conclusions: The combinatorial procedure is a useful technique for identifying a limited number of low-resolution candidate folds for small, disulfide-rich, β-protein structures. Better results are obtained, however, if correct disulfide connectivities are known in advance. Combinatorial distance constraints can be applied whenever there are a sufficiently small number of finite connectivities.

Key words: distance geometry, protein folding, small β-sheet proteins, structure prediction

Received: **14 August 1997**
Revisions requested: **23 September 1997**
Revisions received: **30 January 1998**
Accepted: **06 February 1998**

Published: **19 March 1998**
http://biomednet.com/elecref/1359027800300149

Folding & Design 19 March 1998, **3**:149–160

Introduction

The hierarchical approach towards predicting the tertiary structure of a protein begins by predicting the local conformation (secondary structure) of segments of the polypeptide chain, followed by exploring suitable packings for these segments held as rigid units. Until recently, tools for predicting secondary structure have been insufficiently accurate to sustain efforts to model tertiary structure. Methods that predict protein secondary structure from alignments of families of homologous protein sequences [1,2], however, have recently been shown in *bona fide* prediction settings to provide an accuracy that should, at least in principle, be able to sustain tertiary structure modeling [3]. Thus, it is timely to explore methods to assemble a set of predefined structural elements into a tertiary fold.

Computational methods for assembling tertiary structures can be broadly classified as minimization-based, empirical (distance geometry), or combinatoric [4]. Energy minimization [5–9] and molecular dynamics [10,11] methods seek to minimize a set of potential functions, corresponding to observed physical and chemical effects, through direct adjustments of conformational parameters or, in the latter case, through trajectory calculations. Despite ever increasing computational power, these methods are limited by their complexity and difficulties in accurately modeling the potential functions representing solvent.

Distance geometry algorithms [12–15] are best known in association with proton NMR experiments. They generally fix standard bond lengths and angles, thereby reducing the number of independent variables to only the torsion angles. Using the data obtained from an NMR experiment, a protein conformation can be calculated from a set of distance constraints by minimizing a target function that is zero if all distances are satisfied and increases monotonically as constraints are violated. The reliability of the generated structures depends upon the quality and quantity of the experimental data used as input [14]. More recently, the distance geometry approach has been used in a variety of methods for the folding of polypeptide chains into compact globular structures. For example, Nishikawa and coworkers

[16] introduced constraints from known secondary structure and from potential residue contacts derived from the number of residue neighbours within a 14 Å radius of each residue. Aszódi and Taylor [17,18] have used empirical distance constraints obtained by examining known crystal structures. Likewise, Mumenthaler and Braun [19], have developed an approach for packing helical structures by self-correcting distance geometry. Although these methods share much in common, such as constraints based on the hydrophobic effect, they each add novel aspects to the general approach, testifying to its versatility. Here, a further new development is considered: the use of combinatorics.

Combinatoric methods were pioneered by Cohen and coworkers [4,20,21]. These procedures generate the set of all possible standard packings of secondary structure elements and remove from the set all structures that violate stereochemical rules, leaving a small set of residual 'topologies' to be evaluated. Combinatorial heuristics have been generated for α/α, β/β and α/β protein folds.

The approach presented here follows on from the previous work in that it uses a set of empirical distance constraints to collapse models of protein structure into a compact form. It is different in that it couples a distance-constraint minimization technique with a combinatoric procedure to predict simple Cα models for small proteins containing a few known strands in a single β sheet. The method provides a unified description of loop and secondary structure regions, allows structural flexibility in the pre-assigned secondary structure, and can directly incorporate distance-constraint information that can be derived from multiple sequence alignments. It uses the regular nature of hydrogen bonding in helices and β sheets, the hydrophobicity and the general globular disposition of protein structures to generate empirical distance constraints, which can be enhanced by assignments of active-site residues and disulfide-bond connectivities. In this work, such constraints were determined for the class of small, disulfide rich, β proteins and were applied with a combinatorial procedure to generate Cα model structures.

Methodology

Minimizing distance constraints

In the simple model used here, each amino acid residue is represented by a single point corresponding to its Cα atom. Thus, for a sequence of length n, there are n points that must be positioned in three dimensions to define a Cα model structure. If bond lengths and secondary structures are idealized, some of the $n(n-1)/2$ pairwise distances between points can be estimated to a high degree, particularly if the secondary structure is known. Other distances, for example between internal residues, can also be estimated, but with a much larger uncertainty. The uncertainty associated with each distance estimate can be described by a variance. Using a complete set of pairwise distance estimates, expressed in an $n \times n$ symmetric distance matrix D, and the related set of variances, expressed in an $n \times n$ symmetric matrix V, it is possible to compute a set of coordinates in k dimensions ($k < n - 1$) for each point by minimizing the following function:

$$\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{(\mathrm{E}(\mathrm{x}_i,\mathrm{x}_j) - \mathrm{D}_{ij})^2}{\mathrm{V}_{ij}} \qquad (1)$$

where $\mathrm{E}(\mathrm{x}_i,\mathrm{x}_j)$ is the Euclidean distance between points i and j, where x_i and x_j are vectors in k dimensions. D_{ij} is the empirically determined distance between points i and j and V_{ij} is the variance of the distance between points i and j. For protein structures, a solution in three dimensions is required. In practice, a better result is often obtained if a solution is first found in a higher dimension and only then embedded in three dimensions. In this work an initial solution is generated in five dimensions before squeezing the result into three dimensions.

Distance constraints

Given the mathematical model above, it remains for empirical distances and variances to be estimated for the input data, consisting of the sequence string, residue three-state accessibility assignments (internal, surface or ambivalent), disulfide-bond connectivities, active-site residue assignments, secondary structure assignments and sheet topology. If only the first two input types are available any further assignments can be omitted, resulting in a system that will try to pack solely on the basis of hydrophobicity. Taylor [22] recently showed, however, that using only hydropathic information, a correct fold is unlikely.

In the present work, estimates for the distance constraints (the contents of the matrices D and V) were applied hierarchically. The looser constraints (e.g. repulsive forces) were applied first, whereas constraints that enforced tighter restrictions (the relatively fixed distance between the Cα atoms of amino acids adjacent in the chain) were applied later. Distance constraints are described below in the order in which they are applied. All distances and variances used are summarized in Table 1.

Repulsion

The distance and variance matrices were first initialized with a general repulsion constraint to discourage points from overlapping. A large distance of 20 Å, corresponding to the approximate average Cα–Cα distance of globular proteins of lengths of up to 60 amino acids, was used to fill the matrix D. A very high variance of 120 Å^2, so that points violating this restriction were not heavily penalized, was used to fill the matrix V.

Surface and interior

An additional artificial point, corresponding to the center of mass of the model, was created and used to define

Table 1

A summary of the distances and variances used.

Constraint type	Distance (Å)	Variance	Points applied to*
Repulsion	20.00	120.00	*i*–*j*
Exterior	12.50	15.00	*i*E–(centre of mass)
Interior	5.00	10.00	*i*I–(centre of mass)
Active site	6.50	10.25	*i*site–*j*site
Next adjacent	5.20	0.30	*i*–(*i* + 2)
	7.00	0.50	*i*–(*i* + 3)
Disulfide	5.50	0.20	*i*cys–*j*cys
Sheet	4.54	0.10	*i*cSx–*j*cSy
	9.08	0.10	*i*cSx–kcSz
Strand	6.74	0.08	*i*Sx–(*i*Sx + 2)
	10.10	0.10	*i*Sx–(*i*Sx + 3)
	13.30	2.00	*i*Sx–(*i*Sx + 4)
Helix 3/10	5.26	0.09	*i*3x–(*i*3x + 2)
	6.68	0.01	*i*3x–(*i*3x + 3)
Helix	5.48	0.02	*i*H–(*i*H + 2)
	5.20	0.02	*i*H–(*i*H + 3)
	6.28	0.07	*i*H–(*i*H + 4)
	8.75	0.07	*i*H–(*i*H + 5)
Adjacent	3.81	0.01	*i*–(*i* + 1)

*Centre of mass, pseudo point representing the centre of mass; *i*E, all points assigned as surface; *i*I, all points assigned as interior; *i*site, all points assigned as active site; *i*Cys–*j*Cys, all pairs of disulfide bonds; *i*cSx, centre point of strand x; *j*cSy, centre point of strand directly bonded to x; kcSz, centre point of strand directly bonded to y but is not x; *j*Sx, point in the strand x; *i*3x, point in the 3/10 helix x; *i*Hx, point in the helix x; and *i*, all points.

Table 2

Protein statistics.

Structure	Protein	R	D	H	S	Sheet	Topology	Resolution
BDS	Anti-viral protein	43	3	0	3	1 3	−1 1	NMR
						2 3	−1 1	
CBH	Cellobiohydrolase I,	36	2	0	3	1 3	−1 1	NMR
	C-terminal domain					2 3	−1 1	
CPA	Carboxypeptidase	39	3	0	3	1 3	−1 1	2.5
	inhibitor					2 3	−1 1	
CRN	Crambin	46	3	3	2	1 2	−1 1	1.5
IL8	Interleukin 8	72	2	2	3	1 2	−1 1	NMR
						2 3	−1 1	
OVO	Ovomucoid	56	3	1	3	1 2	−1 1	1.5
	Third domain					1 3	−1 1	
TGS	Trypsinogen	56	3	1	3	1 2	−1 1	1.8
	inhibitor					1 3	−1 1	
PTI	Trypsin	58	3	2	3	1 2	−1 1	1.0
	inhibitor					1 3	−1 1	

R, number of residues; D, number of disulfide bonds; H, number of helices; and S, number of strands.

constraints for points corresponding to residues assigned as either surface or interior. Thus, the matrices D and V are actually of dimension $n + 1$ by $n + 1$. This additional point is used only in the calculations and is thereafter ignored. For the set of proteins listed in Table 2, inspection of the experimental structure shows that the average distance from the center of mass to the Cα atoms of residues with a relative surface accessibility of < 33% is ~7.5 ± 5 Å. Likewise, the average distance from the center of mass of those residues with a relative surface accessibility of > 66% is ~12 ± 5 Å. Consequently, similar values were used here for the distances and related variances between the center of mass of the model and any residue assigned as surface or interior, respectively. No constraints were applied to residues with intermediate surface accessibility.

Active site

The residues forming the catalytic active site of a protein may be dispersed along the one-dimensional sequence string, but usually occur near in space in the tertiary fold. To simulate this effect, all pairs of residues assigned as participating in the active site were constrained to lie 6.5 Å apart. A relaxed variance of 10.25 Å^2 was selected to enable some positional flexibility.

Next adjacent residues

Using standard bond lengths and bond angles, the distance between adjacent Cα atoms is fixed. The distance between Cα atoms separated by one residue depends upon the torsion angles along the backbone chain, however. A distance of 5.2 Å with a variance of 0.3 Å^2 was used here to fill the second off diagonals in the matrices D and V, respectively. The third off diagonals were set to a distance of 7.0 Å with a variance of 0.5 Å^2. These values lie between those found for equivalent residues in helical (tightly coiled backbone chain) and strand (fully extended backbone chain) conformations (Table 1).

Disulfide bonds

Disulfide constraints were estimated by applying a distance of 5.5 Å with a variance of 0.2 Å^2 between the pairs of points representing the cysteine residues. In real protein structures, although the covalent-bond length between the sulfur atoms of the participating cysteines has a small variance, the separation distance between the respective Cα atoms can range from 4 Å to 7 Å.

Sheet

If a sheet topology is specified, then the adjacency and relative orientation (parallel or antiparallel) of the constituent strands are known. This provides a wealth of distance constraints. Here, this information is captured by positioning the middle residues of adjacent strands at a separation distance of 4.54 Å apart with a small variance of 0.1 Å^2. This guarantees a large overlap of potentially hydrogen-bonding residues, although the variance still allows some slippage between the strands depending upon the distance constraints of the system as a whole. The remaining residues in the shorter of the strands were then tethered to their expected hydrogen bonding partner with the same distance and variance as the middle residue. Furthermore, an additional constraint of 9.08 Å with a variance of 0.1 Å^2, between the middle residue of each strand (q) and the middle residue of a strand two away ($q+2$ or $q-2$ for edge strands and strands one in from the edge; $q \pm 2$ for other strands), was introduced to induce a more planar-type structure.

Strands and helices

If secondary structure information is available as input, local structures can be generated by applying a small number of constraints to consecutive positions within the secondary structure unit. Given a strand, depending upon its length, distance constraints were applied to each residue pair within the strand, relating positions i to $i+2$, i to $i+3$, and i to $i+4$, as reported in Table 1. Similarly, analogous constraints were applied to pairs of residues occurring in regular α helices, with an additional constraint applied between positions i to $i+5$. With 3/10 helices, only the pairwise distances i to $i+2$ and i to $i+3$ were constrained (Table 1). For these constraints the variance increases with the separation distance along the chain to allow gentle curvature. The distances chosen generally reflect idealized geometries for the helices and strands, whereas the variances allow a small degree of flexibility.

Adjacent residue constraints

The most rigid constraint specifies the distance between adjacent residues in the sequence, corresponding to the first off diagonal in the matrices. Therefore, the distance constraint between adjacent Cα points was applied last. These distances were set to a value of 3.81 Å with a very small variance of 0.01 Å^2 to restrict the stretching or compression of the virtual Cα–Cα bond.

Smoothing

Any two residues that are distant along the polypeptide chain, but are near in space, force their immediate neighbours along the chain also to be close in space. To reflect this fact, a local smoothing was applied to the matrices, where adjacent cells in the matrix were assigned the same distance with an increment (+ 0.5 Å) and a larger variance. As implemented here, distances $<$ 6.5 Å with a variance of $<$ 0.5 Å^2 would induce smoothing to the neighbour matrix cells, but only if the variance in that cell was $>$ 10 Å^2. The adjusted cells were set to have a new variance of 2.0 Å^2.

Sheet combinatorics

Sheet topologies are expressed here as a set of four tuples, where each tuple consists of two strand indices, a binary relative orientation (1 for parallel and –1 for antiparallel) and a sheet number, where each strand and sheet in the protein has a unique index. Thus, for example, a protein consisting of a single mixed sheet of three strands, where strands 1 and 2 are parallel and strands 2 and 3 antiparallel, would have a sheet topology expressed by {(1,2,1,1) (2,3,–1,1)}.

This description is not able to specify a sheet conformation uniquely because it contains no information on the precise hydrogen bonding partners between separate strands. It also does not specify the handedness of the pairs of strands (whether the second strand is bonded to the right or to the left of the first strand after the first strand is unambiguously oriented). Such a description is, however, consistent with the inherent inability to determine the handedness of a system from pure distance information. Furthermore, this allows adjacent strands to position themselves as best as possible within the global context of all the constraints imposed.

For any given sheet topology, an ensemble of model structure is possible. Depending on the restrictiveness of the constraints, the actual number of models fulfiling the constraints can vary from two, the left-handed and right-handed solutions for a completely defined system, to infinity. Because a β sheet is a tertiary structure, a large number of restrictive constraints are available. Consequently, a sample size of five independent models was deemed sufficient to represent the ensemble. Larger sample sizes (see below) did improve the results but only marginally, not warranting the additional computation. If a sheet topology is specified as part of the input, distance constraints were applied directly to yield the five independent models, otherwise, all possible combinations of sheet topologies are determined, and five models were generated independently for each. Models were calculated in parallel on several machines in a networked cluster.

A protein containing only a single sheet of N strands has $N!(2^{N-1})/2$ possible sheet topologies, because the handedness cannot be distinguished. Thus, combinations of up to four strands in a single sheet can be run routinely.

Disulfide combinatorics

Disulfide bonds form valuable distance constraints, which are particularly suited to the current methodology. It is often easy to predict residues involved in disulfide bonds by examining multiple sequence alignments, looking for conserved cysteine residues. If there are more than two conserved cysteine residues, however, it is generally not obvious which of these are paired. Fortunately, it is relatively straightforward to determine these connectivities experimentally. If experimental information is not available, it is still possible to make use of these potential distance constraints by trying all possible combinations of disulfide connectivities with the distance constraint procedure. There are $[N-1]!/[2^{[(N-2)/2]}[(N-2)/2]!]$ possible disulfide connectivities between N cysteine residues, if N is even. Combinations of these disulfide connectivities are tried here in conjunction with all combinations of sheet topologies, leading to a large pool of potential structures.

Model evaluation

To evaluate a sample solution, a simple evaluation function (E) based on geometric properties of the models was used. The function consists of a van der Waals overlap term (E_v) and a hydrogen bonding term (E_h). The model for which E is the minimum is presumed to be the best. The function is given by:

$$E = \begin{cases} \infty \text{ if } E_v > 0.4 \\ E_h \text{ otherwise} \end{cases} \tag{2}$$

Models involving a large value for E_v become difficult to evaluate because they will often have very favourable E_h values, but nonetheless yield incorrect solutions that are physically unrealistic. Consequently, any model where $E_v > 0.4$ is immediately discarded. Most native folds have $E_v = 0$, although for BDS it is as high as 0.2.

The van der Waals term is given by the sum of the inverse of the squared distances of all pairs of atoms *i,j* that overlap; that is, for which the distance between their centres is less than the sum of their van der Waals radii. For the N Cα atoms in each sequence, C_{vdw} is taken to be 2 Å. Thus:

$$E_v = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} vdw_{ij} \tag{3}$$

where

$$vdw_{ij} = \begin{cases} 1/d_{ij}^2 \\ 0 \text{ if } d_{ij} \geq 2C_{vdw} \end{cases} \tag{4}$$

and

$$d_{ij} = [(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2]^{1/2} \tag{5}$$

The hydrogen bonding term is given by the normalised sum of the distances between a residue in a strand and the nearest residue in an adjacent strand, for all resides in each pair of strands. Thus:

$$E_h = \sum_{k=1}^{Q} [1/(1+(b-a)) \sum_{r=a}^{b} \sum_{s=c}^{d} \min(d_{rs})] \tag{6}$$

where the sum is over all Q unique pairs of adjacent strands *r,s* with residue extents a to b and c to d.

An additional solvent accessibility term (E_a) was also considered, where E_a is a simple step function that assigns a score for each residue depending upon its accessibility state (internal or surface) and the number of neighbouring residues it has at a distance of < 12 Å. Thus, a residue assigned to the interior with > 20 neighbours increments the score by 3; with > 14 neighbours but < 20 neighbours by 1; and with < 15 neighbours by –1. A residue assigned to the surface with < 20 neighbours increments the score by 3; with < 26 neighbours but > 20 neighbours by 1; and with > 25 neighbours by –1. These values were selected based on a plot of accessibility versus number of neighbours for the examples shown in Table 2. This term was dropped, however, because the values obtained from the models were not sufficiently discriminatory.

Equipment

Programs were implemented in the programming languages C [23] and DARWIN (G.H. Gonnet, unpublished software) and incorporated as part of the DARWIN package. Calculation of the models was performed in parallel over a network of workstations using the DARWIN Parexec feature (L.K. and G.H.G., unpublished software).

Dataset

The Brookhaven databank [24] contains a number of small, disulfide-rich proteins containing a single β sheet. A total of eight structures (Table 2) involving four distinct sheet topologies were used to test the method presented in this work. The sequence, secondary structure assignments, disulfide connectivities and surface accessibilities were obtained from the output of the DSSP program [25] using default parameterization. Isolated hydrogen bonded residues (assigned as B by DSSP) were also taken as strand assignments. Furthermore, only strands involving hydrogen bonds to the same subunit were considered. No active-site assignments were used in this work. Surface and interior assignments were derived from the DSSP accessibility values for the known tertiary structures and taken by choosing an accessible surface cutoff of < 33% accessible for interior residues and > 66% accessible for surface residues. Accessible surface normalization was performed by dividing the DSSP value by the values reported by Zielenkiewicz and Saenger [26] for Gly–X–Gly tripeptides in an extended form. In a more general application of the algorithm, these values could be derived from multiple alignments [27].

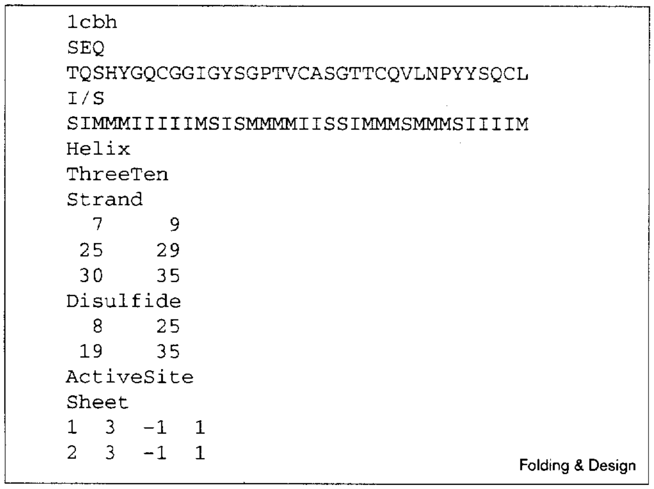
Results

Constraint minimization

To test the distance-constraint minimization, the method was applied to each of the folds in Table 2, using knowledge of the correct sheet topology, secondary structure assignments, disulfide bonds and approximate surface accessibility. A typical input file for the procedure is indicated in Figure 1. For each structure, 10 models were built by repeated application of the procedure using random initial positions for each run. The models were then scored with the evaluation function and ranked. The top scoring model for each structure was then superimposed on the native fold by minimizing the root mean squared deviation (rmsd) between the equivalent strand residues. The superpositions are illustrated in Figures 2 and 3, and are grouped according to folds that are essentially correct and folds that are mostly correct but involve some incorrect placements of some structural elements, often the termini. Only three of the eight structures fall into the category with folds that are mostly correct.

The folds that show an overall chain trace very similar to the native fold include all the structures with the shortest sequence lengths: BDS (anti-viral protein), CBH

Figure 1



```
1cbh
SEQ
TQSHYGQCGGIGYSGPTVCASGTTCQVLNPYYSQCL
I/S
SIMMMIIIIIMSISMMMMIISSIMMMSMMMSIIIIM
Helix
ThreeTen
Strand
  7    9
 25   29
 30   35
Disulfide
  8   25
 19   35
ActiveSite
Sheet
1  3  -1  1
2  3  -1  1
```

Folding & Design

An example input file for CBH (cellobiohydrolase I, C-terminal domain) for the distance constraint procedure. The file indicates the amino acid sequence, the approximate surface accessibility (I, internal; S, surface; M, ambivalent), secondary structure residue extents, disulfide connections and sheet topology (strand identifier 1; strand identifier 2; 1, parallel or −1, antiparallel; sheet identifier). There are no helical or active-site residues described.

(cellobiohydrolase I, C-terminal domain), CPA (carboxypeptidase inhibitor) and CRN (crambin). OVO (ovomucoid) is longer, however. BDS, CBH and CPA have the same sheet topology involving a three-stranded antiparallel sheet with strand 3 in the middle. The topology of CRN is distinctly different, containing only a single pair of antiparallel strands in addition to some helical structures. Likewise, the topology of OVO is also distinct and contains a helix and a three-stranded antiparallel sheet, with strand 1 being the central strand. The model of CRN highlights a limitation of the current methodology. Although the backbone traces a similar path through space to that of the native fold, one of the helices has a left-handed (and therefore incorrect) chirality. A similar situation is found in the OVO helix. The models for BDS, CBH and CPA, on the other hand, represent well the general native fold, as is seen in Figure 2. Nevertheless, the models are still too crude for an overall rmsd value to be truly meaningful (~8 Å average rmsd).

Knowledge of the correct sheet topology and location of the disulfide bonds is sufficient to restrict the best model for IL8 to be close to that of the native fold, as illustrated in Figure 3. The exact orientation of the structural segments and terminal regions, however, show considerable flexibility in the Cα description used here. The model for IL8 shows a different positioning of the C-terminal helix and the N-terminal extension. In the native fold the C-terminal helix lies across one face of the β sheet, perpendicular to the strand direction, whereas in the model structure it occurs on the other face of the sheet. Using only Cα atoms, hydrophobic packing of the helix to the correct face of the sheet was not possible. The general shape of the IL8 model is also more globular than that of the native fold, as is emphasized by the folding back of the hairpin formed by strands 1 and 2. This phenomenon arises from the parameterization of the distance and variance constraints, which are biased towards the smaller structures that form the majority of the data set. Figure 3 also shows the best scoring models for TGS (trypsinogen inhibitor) and PTI (trypsin inhibitor) superimposed on their respective native folds. Although not obvious from the figure, the core of TGS is essentially correct. The N terminus (residues 1–11), however, packs to the wrong side of the model whereas the helix packs with a different orientation to the sheet. Surprisingly, TGS and OVO have very similar structure, but they differ in sequence (~33% identity), solvent-accessibility assignments and secondary structure definitions. These changes clearly influence the correctness of the predicted models. For PTI, strands 1 and 2 form a reasonable hairpin in the core of the model. As with TGS, however, the N and C termini pack to the wrong side of the model. Strand 3, which consists of only one residue, is also not well placed. Models with longer strand definitions tend to yield better results.

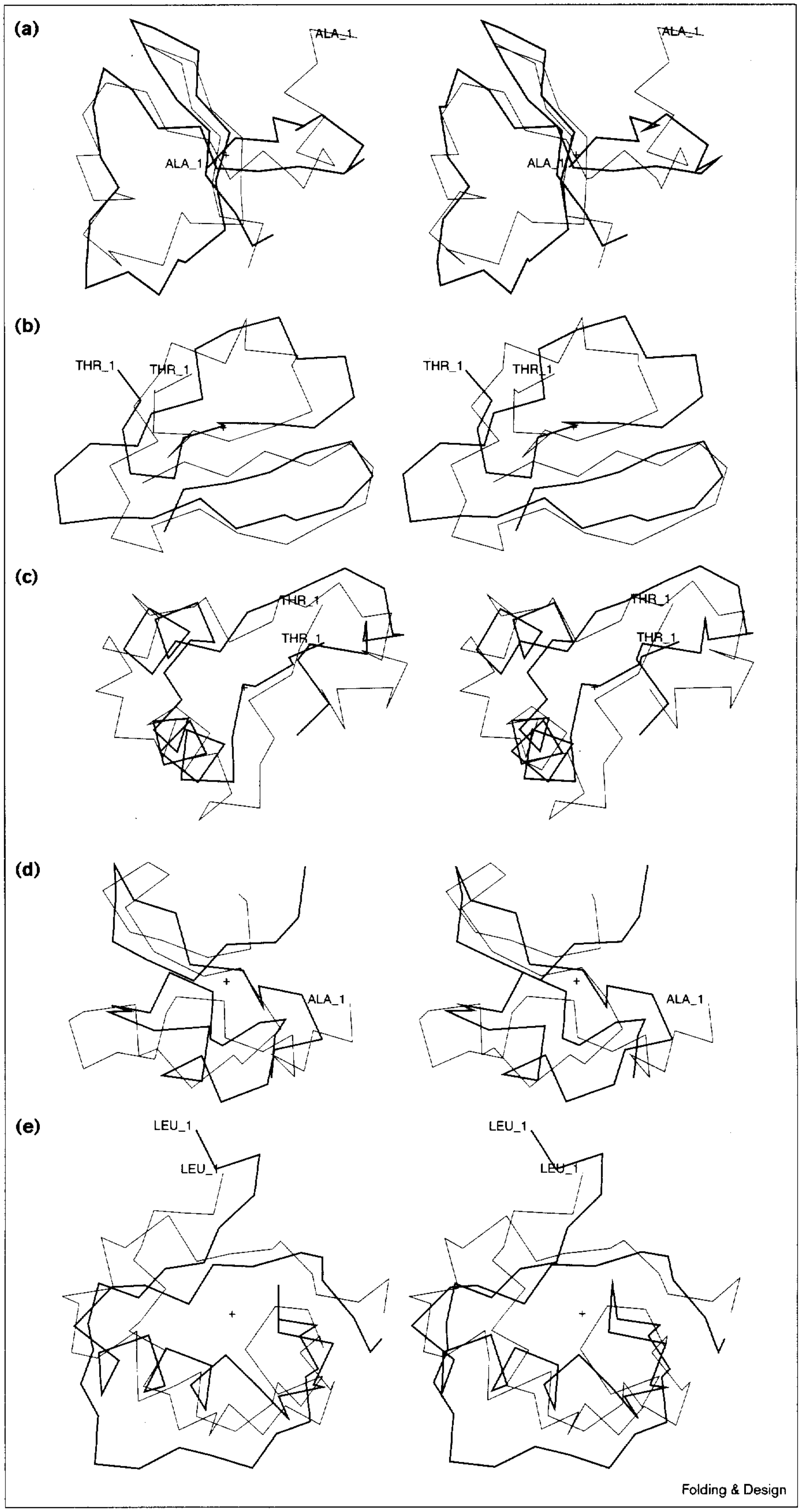
Sheet combinatorics

The results from the previous section indicate that when given the correct sheet topology and disulfides, native-like chain traces are generally induced by the method. Specifying the correct sheet topology, however, enforces considerable constraints on the protein models. Furthermore, this information is not usually obtainable from sequence information. The same applies to disulfide connectivities. To generalize the method such that models can be obtained from quantities that may be predicted from multiple sequence alignments, a combinatorial approach was adopted, whereby all possible sheet topologies were used as input to the procedure, assuming that the correct secondary structure extents and, for the meantime, the correct disulfide bonding have been predicted.

Seven of the eight structures investigated have sheets consisting of three strands, thereby allowing 12 possible sheet topologies. CRN, with only two strands, gives rise to only two possible topologies. For each sheet topology, five models were generated resulting in a total of 60 model structures for each protein in Table 2, except CRN for which there are only 10 models. For each protein, the model structures were then ranked by the evaluation scheme. Table 3 summarizes the results. In four out of eight cases, models with the correct sheet topology scored the highest, three were ranked second and one third. The quality of the correct models, generated here from only five random starting configurations, is generally as good as the models generated from 10 starting configurations as performed in the constraint minimization section above,

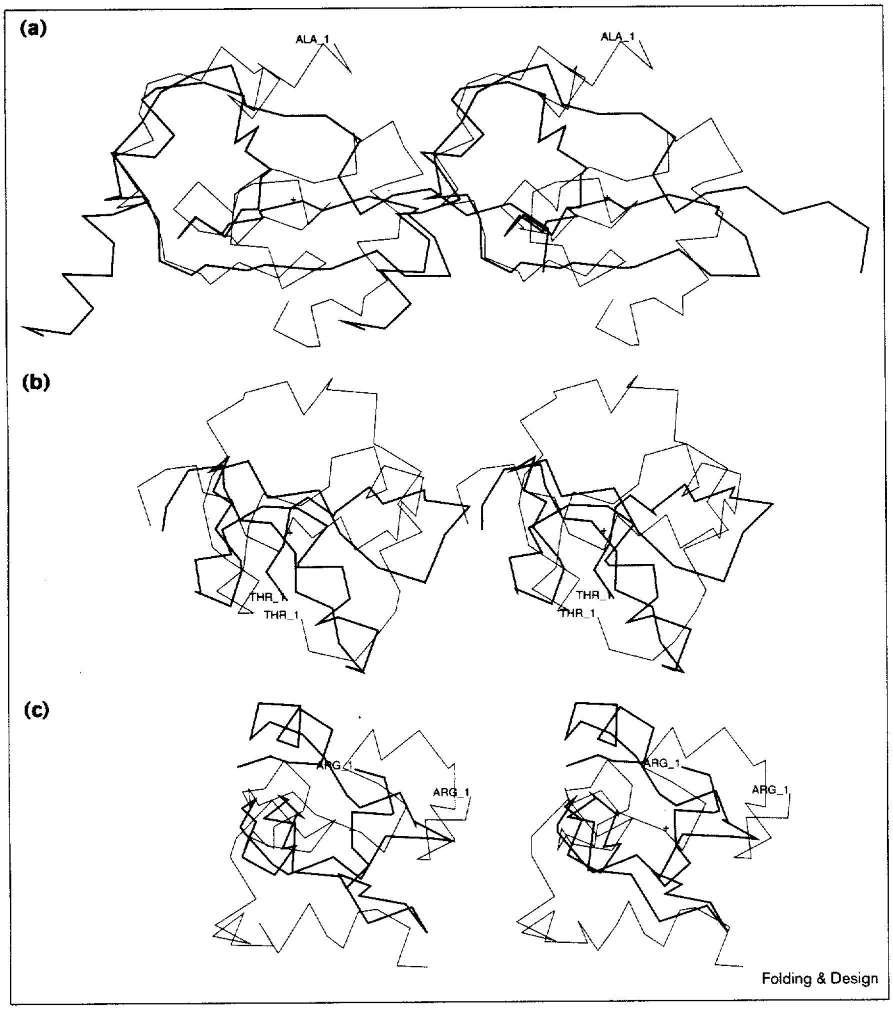
Figure 2

Stereo diagrams of model folds for **(a)** BDS (anti-viral protein), **(b)** CBH (cellobiohydrolase I, C-terminal domain), **(c)** CRN (crambin), **(d)** CPA (carboxypeptidase inhibitor) and **(e)** OVO (ovomucoid, third domain) superimposed on the native structure (bold) using the residues in the β sheet.



Folding & Design

Figure 3



Folding & Design

Stereo diagrams of models for **(a)** IL8 (interleukin 8), **(b)** TGS (trypsinogen inhibitor) and **(c)** PTI (trypsin inhibitor) superimposed on the native structure (bold) using the residues in the β sheet.

although with more trials, closer approximations to the native fold do usually result.

Table 3 indicates the number of trial sheet topologies used in generating the models. Also listed are the number of intuitive sheet topologies, generated by applying a loop length and disulfide connectivity filter. Strands separated by very short loops (≤ 4 residues) indicate that the strands will be antiparallel if they occur in the same sheet (providing that the loop is indeed a loop and not a bulge and the strand lengths flanking the loop are longer than the loop itself). There is, however, no guarantee that such strands will necessarily be directly hydrogen bonded together if the loop is longer than one residue, because it is possible for another strand to intervene [28]. Disulfide connectivities, where the cysteine residues are in or are at least no more than two residues away from two strands, indicate that the strands are adjacent in the sheet, although the orientation might not be able to be deduced. For example, in CBH (Figure 1), the two-residue loop between strands 2 and 3 make these strands antiparallel, but not necessarily adjacent [28], reducing the maximum number of folds to six: [1 2 (1), 2 3 (−1)]; [1 2 (−1), 2 3 (−1)]; [2 1 (1), 1 3 (−1)]; [2 1 (−1),

Table 3

Prediction rankings.

Structure	Rank	Maximum number of folds*	Number of intuitive folds†
BDS	2	12	6
CBH	1	12	4
CPA	1	12	4
CRN	2	2	1
IL8	2	12	6
OVO	1	12	12
TGS	1	12	12
PTI	3	12	6

*Computed by $N!\,2^{(N-1)}/2$. N, number of strands. †As determined by loop length and disulfide bonding as described in the text.

Table 4

Combinatoric disulfide prediction rankings.

Structure	Rank	Maximum number of folds*
BDS	60	180
CBH	2	36
CPA	22	180
CRN	16	30
IL8	4	36
OVO	5	180
TGS	21	180
PTI	5	180

*Computed by $[n!\,2^{(n-1)}/2]\ [(N-1)!/[2^{[(N-2)/2]}\ [(N-2)/2]!]$, where n is the numbers of strands and N is the number of cysteine residues.

1 3 (1)]; [2 3 (–1), 3 1 (1)]; and [2 3 (–1), 3 1 (–1)]. The fact that strands 1 and 3 are linked by a disulfide rules out the topologies [1 2 (1), 2 3 (–1)] and [1 2 (–1), 2 3 (–1)], but is not able to define a relative orientation of the strands. Thus, four topologies are intuitively likely. Such filters are used by Cohen *et al.* [21] to reduce the number of combinatorics. Taken together the results indicate that the prediction potential of the method is not as strong as the first impression because, in some cases, up to 66% of the sheet topologies could be discarded by applying the filters. Surprisingly, the evaluation scheme failed to select the correct sheet topology for CRN when all sheet combinations were explored, even though the intuitive filters could. Examination of the selected model shows that the sheet is not well formed, with an inter-strand twist angle of ~90°, so it is difficult to decide if the strands are parallel or antiparallel. This stems from the very short nature of the strand definitions. Overall, however, the method performs well and is able to select a native-like fold in almost all cases.

Sheet and disulfide combinatorics

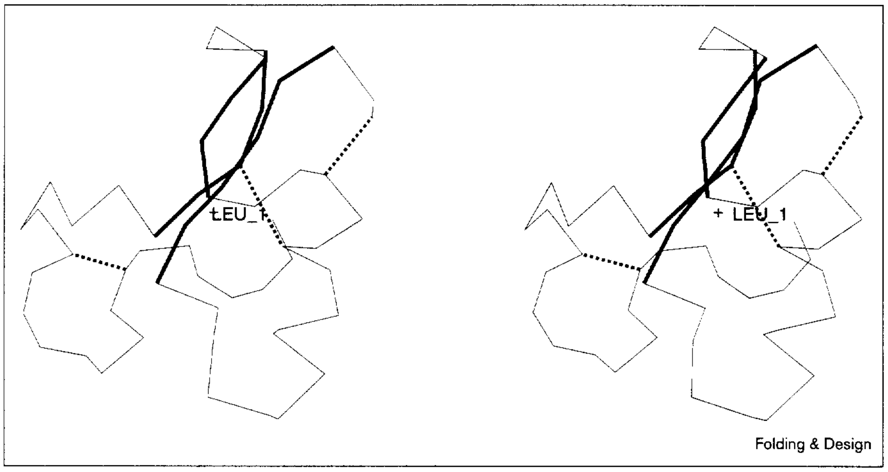
A more difficult test for the procedure is when all sheet topologies are explored without the benefit of using correct disulfide pairings. Although disulfide connectivities can usually be obtained by experiment, they are not easy to predict from sequence alignments alone. Thus, the combinatorial procedure was run again, this time using all combinations of sheet topology and all possible combinations of disulfide pairings as constraints. The results are indicated in Table 4. The table shows that for half the example cases, the correct fold ranked amongst the top five fold types. The correct fold was never ranked first, however. This suggests that folds with the same secondary structure extents but different disulfide connectivities are viable, if they have a different topology. This is akin to the virus capsid jellyroll folds and the immunoglobulin family having similar secondary structural extents but being 'wired' differently. For example, Figure 4 shows a novel but plausible fold for OVO, which has disulfide bonds 8–16, 24–38 and 35–56 and has strands 1 and 3 parallel in contrast to the native fold, which has disulfide bonds 8–38, 16–35 and 24–56 and a fully antiparallel sheet. Although this opens interesting possibilities for protein engineering, exploring combinations of disulfide pairings does not appear generally useful at the low resolution of the models considered here. Higher resolution models, in which the geometry of the disulfide bond is considered, may benefit from such an approach, however.

Discussion

The method presented here addresses the problem of packing known secondary structure elements into globular and compact native folds. Taylor and Aszódi [17,18,22] have already shown that in the use of distance geometry techniques, if the distances are derived empirically to reflect the hydrophobic effect, the resulting distance matrix is generally underdetermined and the

Figure 4

Stereo diagrams of a putative novel fold that involves the same strand extents and cysteine residues as OVO. Strands are highlighted in bold and disulfide bonds by dashes.



Folding & Design

hydrophobicity packing measure, in this context, is alone insufficient to determine a correct tertiary fold from its native sequence. Nevertheless, such a technique has many useful properties, including the immediate incorporation of any experimentally determined distance constraints, the ability to compute solutions in a multidimensional space (thereby freely passing through high-energy barriers or protein knots brought about by the initial random placement of points) and the ability to assign a variance or softening parameter to the distance constraints. This general approach has been used previously for the placement of nodes in the drawing of tree structures (G.H. Gonnet, unpublished software). The current procedure extends from previous approaches in several novel ways.

The most significant extension is the introduction of a combinatorial procedure akin to that proposed by Cohen and coworkers [4,20,21]. Any particular combinatoric of a β structure enforces a large number of long-range distance constraints, thereby complimenting a distance geometry technique. In the original combinatoric approach, candidate structures were selected from the pool of all possible packing combinations based on filters involving loop length and the packing of hydrophobic faces. Loops connecting packed secondary structure units were added at a later stage. The current procedure offers a unified approach to generate a hydrophobic core, construct loop regions (which can now also be included in the designation of the core) and introduce experimental distance constraints. A drawback of the current method is the computational time required to generate the set of candidate model structures. This is offset, however, by the further advantage that the precise orientation of packing units is no longer rigidly fixed and some flexibility (given by the variances) is permitted to the structure to position its constituents in a globally optimal sense.

A second extension makes use of the fact that if any pair of residues are separated by a fixed distance, then their immediate neighbours must also be separated by a similar distance, through chain connectivity. This is most advantageous for long-range distance constraints. Such information may be deduced from multiple alignments (the conservation of functional residues is often indicative of them being localized in an active-site region) or from experiment (disulfide bonds or other residue interactions). More recently, covariation has been explored as a tool for deducing distance constraints [29–36]. In the current procedure, a smoothing process is introduced, so that any pair of residues separated by a well-defined distance constraint can propagate distance information to adjacent residues. Smoothed distance matrices tend to converge faster because the system as a whole is more constrained, thereby enhancing the usefulness of the approach.

A consequence of the combinatorial procedure is that several families of model solutions are generated: one family for each sheet topology attempted. This necessitates a method for differentiating between the models and selecting the solution that displays the most protein-like characteristics. Several methods have been developed for this purpose [37–39]. These methods, however, work best with a more detailed description of the structures than the Cα models presented here. Taylor [22] circumvents this problem by constructing backbone and sidechain coordinates from the Cα coordinates. To avoid this added complexity, a different approach was developed here that relies only upon the Cα coordinates. The method evaluates candidate folds based upon how well they pack in a hydrophobic core while avoiding atomic overlap, as well as considering idealized hydrogen bonding. Despite the simplicity of the evaluation scheme, it performs reasonably well, correctly identifying six out of eight native folds as the best scoring and scoring second in the two remaining cases. When the same structures were evaluated by summing over pairwise potentials of mean force [40], a similar level of success was achieved, with the procedure selecting native-like but not native folds for IL8, PTI and TGS. When applied to the combinatoric set of model structures, however, the prediction accuracy using pairwise potentials was dramatically reduced, correctly finding a native-like model for only CRN and CPA. This is intuitive for two reasons: the size of the structures and the atomic overlap generated by incorrect sheet topologies. Sippl and coworkers [38,40] show that their method works best on larger protein structures and less well for structures smaller than ~60 residues, the typical size used in this work. Furthermore, steric considerations are not considered and as a consequence the Sippl potentials usually select folds that give rise to a low energy simply because they present a dense hydrophobic core, unfortunately involving far too many atomic overlaps to be protein like. Thus, given the rudimentary nature of the Cα models produced, the evaluation scheme presented here is a fast and efficient method for selecting good model candidates.

Clearly, the level of success of the method is governed by the quality of the input constraints. For example, an incorrect set of secondary structures, predicted from an alignment, will generate an incorrect fold. If, however, the secondary structures are essentially correct and err only in the extents of the elements, realistic folds can be achieved. In fact, increasing the length of the DSSP secondary structure definitions by one residue each side resulted in better quality predictions. This was particularly noteworthy in example cases that included a strand consisting of only a single residue. The additional residues help by defining the direction of the strand but can also shorten loop lengths, possibly precluding some sheet topologies. Thus, care must be exercised in deciding when to extend the lengths of structural elements. For long strands the method is quite robust with respect to variations in strand length and even shifts of one or two residues, providing that the loops are not unduly shortened.

The potential of this method has been demonstrated, but the current implementation still suffers from a number of limitations. At present, there is no facility to accommodate models for proteins with more than ~70 residues because the method has been parameterized for small proteins. Further work is required to establish a useful function relating the length of the amino acid sequence to the matrix parameterization. Taylor [22] has made some progress in this area. A related issue is the number of expected surface and interior residues. No checks are made at present to consider if a sufficient number of internal and surface assignments have been made. The importance of this is exemplified by a comparison of the models produced for TGS and OVO (see above). It must also be pointed out that in the current implementation, combinatoric sheet structures are arrived at by aligning the centers of adjacent strands, guaranteeing maximal hydrogen bonding between the strands. Thus, although there is some freedom for strand slippage, not all explicit hydrogen bonding topologies are explored. Usually, this simplification is acceptable, but β sheets are complicated tertiary structures and the bonding of a very short strand to a very long strand could disrupt the procedure. Finally, the issue of chirality must be considered. It is trivial to interconvert between a structure and its mirror image, but difficulty arises in resolving folds consisting of some left-handed segments and some right-handed segments. In the current method, no effort was made to distinguish between left-handed and right-handed substructures. Instead, emphasis was placed on the overall chain trace and evaluating the usefulness of the combinatorial procedure. Usually the hydrophobic constraints are able to guide the fold into a native-like structure (Figures 2 and 3). Methods for enforcing the correct handedness have been discussed and generally involve introducing the calculation of a triple scalar product for the vectors connecting three consecutive Cα atoms [41].

Other approaches are emerging for packing secondary structures elements into tertiary folds. A Monte Carlo method has been developed [42], which, when given the secondary structure and a small number of long-range distance constraints, produces folded structures. The method was applied to hemerythrin, flavodoxin, bovine pancreatic trypsin inhibitor and an immunoglobulin domain and the resulting models were shown to have an rmsd of 3–5 Å for the backbone coordinates. Although this is particularly encouraging, it is important to note that the quality of the models depends heavily on the choice of distance constraints, as does the method presented here. A set of nine non-redundant Cα–Cα distance constraints were used to generate a structure for PTI with the Monte Carlo method. In the current approach, only three explicit distance constraints were used (corresponding to experimentally determined disulfide bonds) to generate the model depicted in Figure 3. The remaining constraints were all derived using the empirical formulations described in the Methods section. Using another recent distance geometry method [43] the authors found that an approximate structure (rmsd ~4 Å) required at least one additional distance constraint for each amino acid in the protein.

Dandekar and Argos [44] have used the Genetic Algorithm to fold the mainchain of several small proteins using only predicted secondary structures. In this approach, mainchain dihedrals were fixed to a small set of possibilities. Populations of strings representing the backbone dihedrals are allowed to vary by means of mutations in and between the strings, guided by a specific fitness criteria that selects the fittest individual as the solution on termination of the algorithm. Although not directly comparable to the current methodology, the Genetic Algorithm approach is complementary in the sense that the fitness function incorporates some of the empirical constraints used here, particularly the hydrophobic interactions.

Recently, an algorithm to generate low-resolution protein tertiary structures from known secondary structure was described [45]. The algorithm uses a simplified representation of the polypeptide chain and a potential based on hydrophobicity. Low-resolution structures were generated for two four-helix bundle proteins, but no other topologies were explored. Again, this emphasizes the importance of hydrophobicity.

Acknowledgements

The authors wish to thank Martina Chelvanayagam for assistance with preparation of the manuscript. This work was supported by the Swiss National Science Foundation. G.C. is the recipient of an Australian Research Council Postdoctoral Fellowship.

References

1. Benner, S.A., Badcoe, I., Cohen, M.A. & Gerloff, L.D. (1993). Bona fide prediction of aspects of protein conformation. Assigning interior and surface residues from patterns of variation and conservation in homologous protein sequences. *J. Mol. Biol.* **235**, 926-958.
2. Rost, B. & Sander, C. (1993). Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.* **232**, 584-599.
3. Thornton, J.M., Flores, T.P., Jones, D.T. & Swindells, M.B. (1992). Protein structure. Prediction of progress at last. *Nature* **354**, 105-106.
4. Cohen, F.E., Sternberg, M.J.E. & Taylor, W.R. (1982). Analysis and prediction of the packing of α helices against a β sheet in the tertiary structure of globular proteins. *J. Mol. Biol.* **156**, 821-826.
5. Levitt, M. & Warshel, A. (1975). A computer simulation of protein folding. *Nature* **253**, 694-698.
6. Burgess, A.W. & Scheraga, H.A. (1975). Assessment of some problems associated with prediction of the three-dimensional structure of a protein from its amino-acid sequence. *Proc. Natl Acad. Sci. USA* **72**, 1221-1225.
7. Hagler, A.T. & Honig, B. (1978). On the formation of protein tertiary structure on a computer. *Proc. Natl Acad. Sci. USA* **75**, 554-558.
8. Abagyan, R.A. (1993). Towards protein folding by global energy optimization. *FEBS Lett.* **325**, 17-22.
9. Skolnick, J., Kolinski, A., Brooks, C.L.III, Godzik, A. & Rey, A. (1993). A method for predicting protein structure from sequence. *Curr. Biol.* **3**, 414-422.
10. van Gunsteren, W.F. & Berendsen, H.J.C. (1987). *Groningen Molecular Simulation (GROMOS) Library Manual.* pp. 1-229, Biomos B.V., Groningen, Germany.

11. Karplus, M. & McCammon, J.A. (1983). Dynamics of proteins: elements and function. *Annu. Rev. Biochem.* **52**, 263-300.
12. Crippen, G.M. (1978). Rapid calculation of coordinates from distance matrices. *J. Comput. Phys.* **26**, 449-452.
13. Braun, W., Bösch, C., Brown, L.R., Go, N. & Wüthrich, K. (1981). Combined use of proton-proton Overhauser enhancements and a distance geometry algorithm for determination of polypeptide conformations. Application to micelle-bound glucagon. *Biochim. Biophys. Acta* **667**, 377-396.
14. Havel, T., Kuntz, I.D. & Crippen, G.M. (1983). Theory and practice of distance geometry. *Bull. Math. Biol.* **45**, 665-720.
15. Havel, T. & Wüthrich, K. (1984). A distance geometry program for determining the structures of small proteins and other macromolecules from nuclear magnetic resonance measurements of intramolecular ^{1}H-^{1}H proximities in solution. *Bull. Math. Biol.* **46**, 673-698.
16. Saito, S., Nakai, T. & Nishikawa, K. (1993). A geometrical constraint approach for reproducing the native backbone conformation of a protein. *Proteins* **15**, 191-204.
17. Aszódi, A. & Taylor, W.R. (1994). Secondary structure formation in model polypeptide chains. *Protein Eng.* **7**, 633-644.
18. Aszódi, A. & Taylor, W.R. (1994). Folding polypeptide α carbon backbones by distance geometry methods. *Biopolymers* **34**, 489-505.
19. Mumenthaler, C. & Braun, W. (1995). Predicting the helix packing of globular proteins by self-correcting distance geometry. *Protein Sci.* **4**, 863-871.
20. Cohen, F.E., Richmond, T.J. & Richards, F.M. (1979). Protein folding: evaluation of some simple rules for the folding of helices into tertiary structures with myoglobin as an example. *J. Mol. Biol.* **132**, 275-288.
21. Cohen, F.E., Sternberg, M.J.E. & Taylor, W.R. (1980). Analysis and prediction of protein beta-sheet structures by a combinatorial approach. *Nature* **285**, 378-382.
22. Taylor, W.R. (1993). Protein folding refinement: building models from idealized folds using motif constraints and multiple sequence data. *Protein Eng.* **6**, 593-604.
23. Kernighan, B.W. & Ritchie, D.M. (1978). *The C programming language.* Prentice-Hall Inc., Englewood Cliffs, NJ, USA.
24. Bernstein, F.C., *et al.*, & Tansumi, M. (1977). The protein data bank: a computer based archive file for macromolecular structures. *J. Mol. Biol.* **112**, 535-542.
25. Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure. Pattern recognition of hydrogen bonded and geometrical features. *Biopolymers* **22**, 2577-2637.
26. Zielenkiewicz, P. & Saenger, W. (1992). Residue solvent accessibilities in the unfolded polypeptide chain. *Biophys. J.* **63**, 1483-1486.
27. Benner, S.A. & Gerloff, D. (1990). Patterns of divergence in homologous proteins as indicators of secondary and tertiary structure: a prediction of the structure of the catalytic domain of protein kinase. *Adv. Enzyme Regul.* **31**, 121-181.
28. Hutchinson, E.G. & Thornton, J.M. (1993). The Greek key motif: extraction, classification and analysis. *Protein Eng.* **6**, 233-245.
29. Altschuh, D., Lesk, A.M., Bloomer, A.C. & Klug, A. (1987). Correlation of coordinated amino acid substitutions with function in Tobamoviruses. *Protein Eng.* **1**, 228-236.
30. Altschuh, D., Lesk, A.M., Bloomer A.C. & Klug, A. (1987). Correlation of coordinated amino acid substitutions with function in viruses related to Tobacco mosaic-virus. *J. Mol. Biol.* **193**, 693-707.
31. Altschuh, D., Vernet, T., Moras, D. & Najai, K. (1988). Coordinated amino-acid changes in homologous protein families. *Protein Eng.* **2**, 193-199.
32. Taylor, W.R. & Hatrick, K. (1994). Compensating changes in protein multiple sequence alignments. *Protein Eng.* **7**, 341-348.
33. Shindyalov, I.N. Kolchanov, N.A. & Sander, C. (1994). Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? *Protein Eng.* **7**, 349-358.
34. Göbel, U., Sander, C., Schneider, R. & Valencia, A. (1994). Correlated mutations and residue contacts in proteins. *Proteins* **18**, 309-317.
35. Neher, E. (1994). How frequent are correlated changes in families of protein sequences? *Proc. Natl Acad. Sci. USA* **91**, 98-102.
36. Chelvanayagam, G., Eggenschwiler, A., Knecht, L., Gonnet, G.H. & Benner, S.A. (1997). An analysis of simultaneous variation in protein structures. *Protein Eng.* **10**, 307-316.
37. Chiche, L., Gregoret, L.M., Cohen, F.E. & Kollman, P.A. (1990). Protein model structure evaluation using the solvation free energy of folding. *Proc. Natl Acad. Sci. USA* **87**, 3240-3243.
38. Hendlich, M., *et al.*, & Sippl, M.J. (1990). Identification of native protein folds amongst a large number of incorrect models. The calculation of low energy conformations from potentials of mean force. *J. Mol. Biol.* **216**, 167-180.
39. Lüthy, R., Bowie, J.U. & Eisenberg, D. (1992). Assessment of protein models with three-dimensional profiles. *Nature* **356**, 83-85.
40. Sippl, M.J. (1990). Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol.* **213**, 859-883.
41. Crippen, G.M. & Havel, T. (1988). Distance geometry and molecular conformation. Wiley, New York.
42. Smith-Brown, M.J., Kominos, D. & Levy, R.M. (1993). Global folding of proteins using a limited number of distance constraints. *Protein Eng.* **6**, 605-614.
43. Lund, O., Hansen, J., Brunak, S. & Bohr, J. (1996). Relationship between protein structure and geometrical constraints. *Protein Sci.* **5**, 2217-2225.
44. Dandekar, T. & Argos, P. (1994). Folding the main chain of small proteins with the genetic algorithm. *J. Mol. Biol.* **236**, 844-861.
45. Monge, A., Friesner, R.A. & Honig, B. (1994). An algorithm to generate low-resolution protein tertiary structures from knowledge of secondary structure. *Proc. Natl Acad. Sci. USA* **91**, 5027-5029.

Because *Folding & Design* operates a 'Continuous Publication System' for Research Papers, this paper has been published on the internet before being printed. The paper can be accessed from http://biomednet.com/cbiology/fad – for further information, see the explanation on the contents pages.