Hierarchical minimization with distance and angle constraints

John R. Gunn
Département de Chimie, Université de Montréal
C.P. 6128, Succ. Centre-ville, Montréal, Québec H3C 3J7
and
Centre de Recherche en Calcul Appliqué
5160 boul. Décarie, bureau 400, Montréal, Québec H3X 2H9
gunnj@cerca.umontreal.ca

Abstract

The incorporation of experimentally-determined constraints into structure-prediction methods based on energy minimization leads to both improved selectivity with empirical potential functions and structure determination with far fewer constraints than are required for distance-geometry calculations. Some methods will be described for using both distance and angle constraints with the hierarchical minimization algorithm. The simulation is based on a combination of Monte Carlo Simulated Annealing and Genetic Algorithm techniques which are integrated into a single framework. The selection cycle of the genetic algorithm is carried out at the same temperature as the mutations, or alternatively the crossover cycle can be considered as a type of Monte Carlo trial move, such that each temperature annealing step corresponds to a new generation. The sequence is divided up into segments, and the mutation step consists of replacing an entire segment with a choice from a pre-selected list. This list is in turn constructed from a list of smaller segments, and the number of overall conformations can thus be pruned at each level of selection. Results will be shown for test cases using a small number of flexible distance constraints used as an additional term in the potential, and for restrictions placed on backbone dihedral angles used as an additional screening criterion for constructing trial moves.

Keywords: Protein folding, Genetic algorithm, Distance geometry, Dihedral angle constraints.

Introduction

For many years now, people have been working on the so-called *ab initio* protein-folding problem, that is the prediction of the structure using only a universal energy function and minimization strategy which can be applied to any protein. It is becoming more and more apparent, however, that this idealistic approach is unlikely to be useful in practice in the near future. There is currently a rapid development of techniques, both in experiment and in bioinformatic analysis, which are more and more effective at finding structures. Energy minimization is therefore evolving into one of many complementary tools which are applied to structure determination.

The major challenge faced by any structure-prediction algorithm is the need to search through an astronomically large number of possible conformations, which requires the use of simplified models to describe both the geometric structure and free energy, and therefore implies a trade-off between the accuracy of the model and the associated computational expense. Currently, most protein-folding simulations use empirical statistically-derived energy functions, which have been shown to perform poorly in identifying native structures (Wang *et al.* 1995).

We have applied the hierarchical algorithm to the problem of finding the minimum-energy conformations of a simplified representation of a protein backbone (Gunn 1996; 1997). In tests with several different empirical potential functions, the minimized structures were found to be lower in energy than the correct native structure, thus showing that such potentials do not include sufficient information to properly identify the native structure in an unrestricted minimization. Since the hierarchical algorithm is designed to be scaled up to larger systems, for which this problem is generally worse due to the rapidly increasing number of low-scoring incorrect structures, it is therefore better suited to problems where at least some additional data can be included.

In contrast to the strictly *ab initio* approach to structure prediction, most current applications of computer modelling are found in situations where a great deal is already known about a structure, either from experimental X-ray diffraction and NMR data or from sequence homology. The problem of structure determination from experimental data is much more straightforward, and standard methods exist which perform quite well (Brünger *et al.* 1986). Recent results, however, both in our group and elsewhere (Aszódi, Gradwell, & Taylor 1995; Skolnick, Kolinski, & Ortiz 1997) have shown that a much smaller number of distance

constraints is often sufficient to obtain reasonable results when combined with more general prediction algorithms. These results suggest that the key to overcoming weaknesses in generic potential functions is the use of site-specific constraints, in addition to amino-acid type dependent interactions.

The goal of the present work is to study the performance of the hierarchical algorithm using a very simple empirical potential in combination with different types of constraint. This type of hybrid approach will allow us to extend the scope of current modelling techniques further towards practical structure prediction, and potentially reduce considerably the total effort required to obtain useful structures. The hierarchical algorithm has a fundamental advantage which makes it particularly suited to this approach. Structural features are evaluated at several different levels of detail, which provides a very flexible mechanism for introducing different types of information at different scales, as either constraints or as part of an energy function.

In addition, the constraints can be applied in such a way as to bias the structure minimization without requiring that they be strictly satisfied. This means that experimentally-determined constraints containing error bars can be supplemented with guesses of likely structures taken from sequence and homology analyses. This flexibility of the hierarchical algorithm will allow the generation of databases of plausible structures with a confidence level and degree of resolution which is appropriate to the data provided as input. This will ultimately allow the assembly of model structures to be added to the list of information-management techniques used in the application of bioinformatics to drug design.

The Reduced Model Representation

Structural Model

We use a reduced model representation of the protein molecule introduced previously (Gunn 1996; 1997). The description of each residue consists of the positions of six atoms, N, H, C_α, C_β, C′, and O. All internal coordinates are held fixed at standard geometries with the exception of the ϕ and ψ dihedral angles around the N–C_α and C_α–C′ bonds, respectively, which are thus the only degrees of freedom in the model. The possible values of ϕ and ψ are further restricted to a discrete set of points on the two-dimensional Ramachandran map, corresponding to representative conformations in the energetically allowed regions. For the present work we have chosen a uniform distribution within the allowed regions with a spacing of roughly seven degrees. Since there is no explicit potential energy term which depends on ϕ and ψ, this corresponds to a hard-wall-type effective torsional potential which is used to bias the choice of trial moves. The conformation of the molecule can thus be simply represented by a series of integers labeling the conformational state of each residue. The secondary structure is imposed simply by assigning specified residues to conformations corresponding to idealized helix or strand geometries ($\{-65, -40\}$ and $\{-147, 145\}$, respectively) and holding them fixed throughout the simulation. In this model all residues (including glycine) are equivalent from a geometric perspective, corresponding to a physical model of poly-alanine (thereby neglecting the ring closure and possible *cis* conformation of proline.) This is not however a necessary restriction, and is used here simply for convenience.

Potential Functions

The amino-acid sequence appears only in the potential function, in which the effective interactions between the atoms differ according to residue type. Potential interactions exist only between residues separated by more than four in the primary sequence, representing in principle potentials of mean force averaged over solvent and sidechain conformations with local correlations due to covalent bonding neglected.

The primary potential function is a residue-based hydrophobic pair potential based on statistics obtained from the Protein Data Bank (PDB) (Casari & Sippl 1992). This is an effective free energy function where the effects of sidechain conformation and solvent are implicitly included by assuming that the structures in the Data Bank represent an equilibrium distribution of residues under the conditions found in native proteins. The potential is of the form

$$E = \sum_{i-j \geq 4}^{N} (h_i + h_j + h_0)|\mathbf{r}_i - \mathbf{r}_j|$$

where the relative hydrophobicities h_i depend on the residue types at each sites and are otherwise independent of the separation in the sequence. The residue coordinates $\mathbf{r}_i$ are taken to be those of the C_β atoms. The parameter h_0 is the net hydrophobicity which provides the overall driving force towards compactness, and can be adjusted according to the sequence. In this model, hydrophobic residues will have positive values of h_i and an attractive interaction while hydrophilic residues will have negative h_i and be repelled. A positive value of h_0 ensures that the average force in the molecule will be attractive.

The hydrophobic potential is supplemented with an excluded-volume term which is based on a table of distances of closest approach for each pair of C_α and C_β atoms. These distances have been fit to observed values in the PDB.[1] A constant penalty is added to the

[1] A.K. Felts, private communication.

energy for each pair of atoms which violate this condition. An additional term is added which is proportional to the radius of gyration in order to ensure the correct overall density. The coefficients of both of these additional terms are varied during the simulation in order to provide a simple smoothing of the potential by reducing the impact of short-range interactions in the initial stages of the minimization.

This potential function was chosen because it represents the simplest form which gives plausible results, even though it lacks the detail needed to accurately identify the native structure. It is therefore a useful reference point with which to evaluate the impact of any additional constraints.

The Hierarchical Algorithm

Segment Lists

In the hierarchical algorithm, trial moves are generated and evaluated in three different steps. At the simplest level, segments of three residues (triplets) are generated by choosing three conformational states at random. Each triplet is immediately accepted or rejected according to whether or not the orientation of its endpoints falls into an allowed region of triplet conformational space. Three residues contain a sufficient number of degrees of freedom to make this non-trivial in the sense that different sequences of backbone dihedral angles can produce the same endpoints, however the segment is small enough that the screening of random conformations can be done fairly cheaply. The second level consists of complete loop segments as determined by the secondary structure. These loops are evolved from previously existing structures by using the set of triplets from the first level as trial moves and by evaluating new loops based on the difference in overall geometry from the starting loop. The final level then corresponds to the entire molecule, for which the trial moves consist of substituting entire loops with the new loops generated in the second level. It is only at this final level that the structure is evaluated by calculating the full potential using the table of pairwise contact energies.

For the purposes of evaluating triplets and loops in the first two levels, an endpoint geometry is introduced which is specified by five coordinates that determine the orientation of the final residue relative to the first one. The objective is to estimate the 'lever-arm' effect, that is, the amount by which the ends of the molecule will move following a change in the conformation of a series of interior residues, since this will dominate the total energy of the molecule. The trial moves in the simulation are carried out by using the atomic coordinates of the rigid H–N–C_α unit to define an absolute coordinate system for each residue. The five geometrical parameters can then be thought of as pseudo bond angles and dihedrals between the coordinate axes of two residues as shown in Figure 1. This allows the net rotation between any two residues to be determined and therefore makes it trivial to splice together different structures and incorporate new segments into existing structures, including the necessary rigid-body rotations.

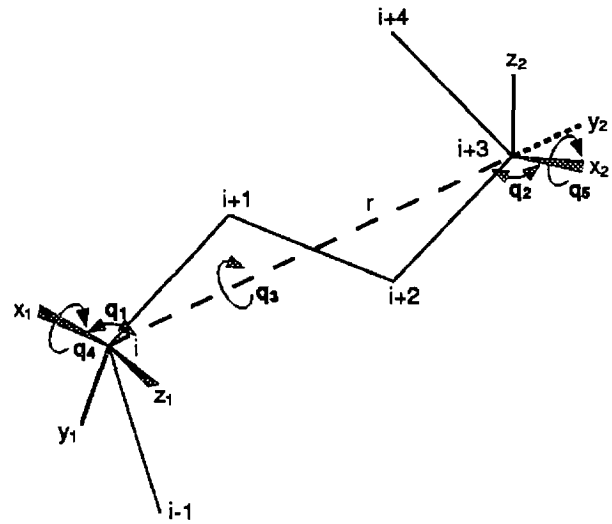

Figure 1: Schematic illustration of the angular parameters, q_i, used to define the relative orientation of two C_α centers. The structure shown here is not intended to be realistic, since in general the C_α atoms would not be co-planar.

The advantage of using the hierarchical algorithm is that the Monte Carlo step size, in this case the maximum difference between loops, can be controlled by the choice of cutoff in the loop generation thereby eliminating trial moves which would be too large to have a reasonable chance of being accepted. Although the set of loops in the list is a limited subset of all possible conformations satisfying the same cutoff, it is still many orders of magnitude larger than the number of attempted Monte Carlo moves and therefore imposes a negligible restriction on the conformational sampling. The hierarchical screening of loops and triplets by geometry, however, is a much more efficient method of restricting the trial moves than simply applying a geometry cutoff to randomly generated loop conformations. The acceptance rate at each level is adjusted by tightening the selection criteria as the number of degrees of freedom increases. In the present work, the geometrical parameters are each divided into four bins, which corresponds to 90 degrees for the dihedral angles. This gives a total of 1024 bins on the 5-dimensional grid, which means that triplets can be found for each bin in

a reasonable time with a random search. The cutoff for accepting new loops is about 20 degrees, which can be easily achieved using the screened triplets as trial moves. In a typical run, the acceptance rate is about one in 200 for the triplets and about one in 15 for the loops. The odds of a randomly generated loop meeting the 20-degree cutoff directly would be on the order of 2 million to one.

The Genetic Monte Carlo Minimization

The genetic algorithm consists of three main components: mutation, hybridization, and selection. The 'genome' corresponding to each member of the ensemble of structures is the series of loop conformations defined by the loop lists as described above. The mutation step involves making a random change to the genome of an individual structure, and is therefore equivalent to a standard Monte Carlo move in which the trial moves consists of changing the conformation of a single loop. The new loop is accepted with a probability of one if the new structure is lower in energy, and with a probability of $\exp(-\beta\Delta E)$ if the new structure is higher in energy, with β being the inverse annealing temperature.

The hybridization and selection steps, on the other hand, couple different members of the ensemble and give the genetic algorithm its additional advantage. Hybridization consists of creating a new structure by combining the genomes of two existing structures. One of the secondary structural elements is randomly selected as the splice point, and a new structure is created by taking all loop conformations on the N-terminal side of the splice point from one parent structure and all loop conformations on the C-terminal side from the other, with at least two loops coming from each parent. The two parent structures are chosen randomly among the existing ensemble according to a Boltzmann distribution. A population of new structures is created by keeping those distinct hybrids whose energies are less than a maximum cut-off determined by the average energy of the two parent structures.

A selection algorithm is then used in which members of the newly created hybrid population are allowed to compete with and displace existing structures from the original ensemble of parents, thereby modifying the new ensemble which will be passed on to the next iteration. This is carried out during an equilibration cycle in which two structures, one from each of the old and new populations, are randomly chosen and their exchange is accepted or rejected according to the probability $\exp(-\beta\Delta E)$ exactly as in the mutation step. Trial exchanges are attempted until the two populations have been completely scrambled and the total energy of the new ensemble is stable. The new generation is thus a combination of old and new structures selected from a Boltzmann distribution, and will not generally correspond to the set of lowest-energy structures. However, the best structures will have increasingly higher probabilities of being selected, with the single lowest-energy structure almost certain to be one of those in the new ensemble.

One iteration is defined as a complete cycle of loop-list generation followed by one generation of the genetic algorithm, as summarized in Figure 2. Each iteration is carried out at a specified temperature and within a limited region of the total conformational space. This region is defined by the cut-off used to select new loops for the loop lists, and the space within this region is further discretized by selecting a fixed number of loops to be used during the subsequent iteration. At the beginning of each iteration, the ensemble is used to construct new loop lists. The mutation and selection steps of the genetic algorithm then constitute a trajectory within this space which satisfies the Metropolis algorithm at a specified temperature. The end result is a new ensemble located somewhere within this space which is then used as a starting point for the following iteration. As the annealing temperature is reduced the population of structures converges to a set of low-energy local minima.

Results

Evaluation of the Method

The ability of the algorithm to handle a distance-constraint potential was tested with a simple model in which the only energy was -1 unit for a pair of residues within 1 Å of the crystal structure value and zero otherwise. This is a highly frustrated system in which there are no long-range forces to guide the minimization. This somewhat artificial system was chosen to provide a more stringent test of the ability of the algorithm to successfully mova around a large number of local minima with sharp barriers. This model potential was minimized both with the hierarchical algorithm and with a simple Monte Carlo algorithm in order to provide a basis for comparison. Each simulation was run for approximately the same total time.

Results are shown in Table 1 for adenylate kinase (3ADK), a mixed α/β protein with 193 residues. The nature of the potential function in this case causes the simulation to get trapped in local minima and leads to significant differences between different runs. The lowest-r.m.s.-deviation structures in each case, however, are close to the best resolution that can be achieved with the model using a discrete set of residue conformations and an idealized secondary structure, thereby suggesting that at least some of the time the minimization is finding the best structures. The native score is always much lower in this case because the geome-

Update annealing temperature, overlap penalty, and other parameters

Structure build-up cycle

Generate triplet lists by screening random conformations
16 per triplet per structure

Generate loop lists by modifying existing loops using triplet lists
32 per loop per structure

Generate lists of loop-helix-loop fragments using triplet lists
32 per pair of loops per structure

Monte Carlo minimization cycle

Mutation cycle using loop list trial moves
64 structures, 1000 trials per structure

Hybridization cycle using pairwise crossover
448 new structures, 50 trials per hybrid

Unrestricted mutation cycle using all lists
512 structures, 200 trials per structure

Monte Carlo selection cycle
64 structures selected

Repeat
100 annealing steps

Figure 2: Summary of a cycle of the hierarchical algorithm.

try of the model structures makes it impossible to simultaneously satisfy all constraints, which is the case for the crystal structure by definition. In the case of 3ADK, however, slightly better scores are actually obtained with higher-r.m.s.-deviation structures showing that even in the case of a designed potential there are alternate minima. The hierarchical algorithm produces substantially better distance-constraint scores than the simple Monte Carlo, with the Monte Carlo structures being exclusively higher in r.m.s. deviation as well as in potential score. This demonstrates the improved efficiency of the conformational search with the hierarchical algorithm.

Distance Constraints

A more flexible constraint function was introduced to be used in combination with the hydrophobic potential function. Constraints are defined as being any pair of residues with a C_β–C_β distance of 8 Å or less. The constraint function is equal to a constant value for distances up to 8 Å and decays to zero exponentially for larger distances with a characteristic length of 3 Å. This adds

Table 1: Comparison of results for the minimization of 3ADK using the Hierarchical and Monte Carlo algorithms. Each run represents an ensemble of 64 structures.

Method	Low r.m.s	Ave. r.m.s.	Ave. score
MC	8.3	15.8	-530
MC	9.7	15.7	-518
MC	7.8	16.1	-548
HA	3.3	3.4	-888
HA	3.3	3.8	-948
HA	4.3	4.6	-969
HA	3.6	4.3	-870
HA	6.9	7.3	-978

a long-range contribution which favors nearly-satisfied constraints, but at the same time there is no penalty for gross violations since the energy remains bounded.

This function was tested for two small globular proteins which have been previously studied in the literature: calcium-binding protein (3ICB), an α protein with 72 residues, and tendamistat (3AIT), a β protein with 62 residues. In each case, a total of ten constraints were chosen at random from among the eligible pairs of residues in the crystal structure. This was repeated for 20 simulations, each using a different set of constraints, and compared with the results of a similar experiment carried out by Aszódi, Gradwell, & Taylor (1995). In that work, a more sophisticated potential function was used, which combines residue hydrophobicities with a sequence-dependent contact potential based on residue conservation among families of related structures, in addition to the explicit constraints discussed below. Additional terms for solvent exposure and secondary structure packing were also included.

The results are summarized in Tables 2 and 3, although the definition of the r.m.s. deviation differs slightly between the two sets of results. The numbers given for the present work refer to the α-carbon r.m.s. deviation, whereas those given by Aszódi, Gradwell, & Taylor (1995) are weighted averages which take into account the R-factors in the crystal structures. Results using the bare potential (no constraints) as well as using all of the eligible constraints included for reference. Although the results obtained using only the potential are generally quite poor, they are nonetheless significantly better than random. The simple form used here is not expected to have a large influence, with the major contribution coming from the excluded volume and radius of gyration terms which act primarily as a sanity check. Preliminary result indicate, however, that although the hydrophobic potential is largely irrelevant with large

Table 2: Results of simulations with constraints for 3ICB.

	Constraints	Low r.m.s.	Ave. r.m.s.	Std. deviation
Present	0	4.6	9.8	1.9
work	10	3.0	4.9	1.3
	89	3.0	3.3	0.2
Aszódi	0		10.0	1.5
et al.	10		6.3	2.0
	86		2.9	0.2

Table 3: Results of simulations with constraints for 3AIT.

	Constraints	Low r.m.s.	Ave. r.m.s.	Std. deviation
Present	0	8.4	9.7	0.4
work	10	4.8	8.4	1.3
	116	3.6	6.8	1.6
Aszódi	0		9.4	0.7
et al.	10		5.8	0.6
	120		3.7	0.2

numbers of constraints, it does play a positive role with fewer constraints.

For 3ICB, 10 constraints are sufficient to find as good a structure as was found using all of the constraints. Because of the use of ideal β-strands without any sort of strand-pairing potential, 3AIT proved to be much more difficult, although the addition of 10 constraints does also lead to a significant improvement. Aszódi, Gradwell, & Taylor (1995) do however obtain better results by using an explicit strand-pairing term in addition to the distance constraints.

Other published simulations (Lund *et al.* 1996) show better results when all of the constraints are used, but fail completely for small numbers of constraints. A test was also carried out with a larger molecule, myoglobin (1MBA), an α protein with 140 residues. Using 20 constraints in this case, a structure with an r.m.s. deviation of 4.5 Å was obtained, comparable to 4.9 reported elsewhere (Skolnick, Kolinski, & Ortiz 1997) for the same set of constraints. This result improved to 3.2 Å with a random selection of 30 constraints.

Random sets of constraints were used in order to test the robustness of the method, since some choices will perform better than others depending on the distribution of the constraints. This is an important consideration, since experimentally-obtained distances may be redundant or unevenly distributed. A test was also carried out, though, for a specific set of constraints as a counter-example. One of the weaknesses of the model is the neglect of the calcium ions in the calcium-binding protein, 3ICB. A simulation was thus carried out using only constraints associated with residues directly bonded to the calcium ions. This forces the correct formation of the binding sites and corrects a specific error in the hydrophobic potential, namely that negative residues normally repel one another. This set of 26 constraints generated a low-r.m.s. structure of 2.3 Å, which is actually superior to that obtained using the complete set of constraints. This suggests that constraints determined based on functional or binding information may be even more effective.

Angle Constraints

In constrast to the distance constraints, angle constraints were applied directly to the choices of residue conformation used in the construction of the triplet lists for loop residues. The usual distributions were augmented with an additional grid of points centered around the crystal-structure value. The size of the distribution was chosen to approximate the range of values that could be predicted from NMR spectroscopy (Hong, Gross, & Griffin 1997), although the practical resolution has not been studied in detail.

Test calculations were carried out for myoglobin (1MBO) using a modified potential function in which the hydrophobic pair potential was replaced by a more accurate contact potential (Bryant & Lawrence 1993). In order to have consistent structures using crystallographic values, the native structure was also used for the α-helices in this case, with all values optimized to take into account small differences in other geometric parameters which were still treated as constant.

Results are shown in Table 4 for an additional grid with a width of 10 degrees. A significant difference is seen in the results even when the additional points represent less than half of the available conformations, indicating that only a partial bias is needed to influence the outcome of the simulation.

Reasonable results were also obtained when the size of the grid is increased without changing the number of points. Table 5 shows the results for a weight of 50 % as a function of resolution. There is a loss of precision associated with the larger grids, however the minimization of the potential function is unaffected, suggesting that there is a large potential minimum around the native structure. The application of the constraints can therefore be adapted to correspond to the resolution and reliability of the experimental data.

Table 4: Results of simulations of 1MBO using angle constraints with different relative weights.

grid weight	Low r.m.s	Ave. r.m.s.	Ave. score
0 %	8.1	11.1	-172
6 %	7.4	11.7	-172
20 %	4.9	9.8	-173
30 %	5.1	6.5	-218
50 %	2.5	4.1	-226
100 %	1.7	2.7	-226

Table 5: Results of simulations of 1MBO using angle constraints with different grid sizes for the constraints.

grid size	Low r.m.s	Ave. r.m.s.	Ave. score
10°	2.5	4.1	-226
20°	3.1	6.4	-225
30°	3.6	7.2	-228

Constraint Flexibility

Some preliminary calculations have been done to test the performance of the algorithm in the presence of faulty or misleading data, with the goal of developing a method capable of managing larger numbers of possible constraints even when some are unreliable or cannot be simultaneously true. A simulation of 1MBO was carried out with 75 distance constraints representing pairs of residues closer than 8 Å in the crystal structure, to which an additional 25 false constraints were added which attempted to force contacts between residues known to be at least 20 Å apart. This produced a structure with an r.m.s. deviation of 3.9 Å, which compares well with the value of 3.7 Å obtained with 50 true constraints. This implies that good and bad data more or less cancel one another out, and that the results depend on the 'net' quantity of good information.

A simulation was also carried out with angular constraints using a 10° grid of points weighted 50%, as in the previous section, with the constraint values chosen entirely at random. The results are only slightly worse than those obtained with no constraints at all, which suggests that the additional points are more or less ignored by the simulation when they do not lead to a reasonable structure. The error tolerance of the method is a significant advantage if the peak assignments are ambiguous and there is a set of possible values associated with each residue rather than a single constraint. The simulation can than be carried out using all possible values and will be capable of converging to a structure that is self-consistent.

Conclusions

The hierarchical simulation algorithm has been shown to work effectively with geometrical constraints, both in the limiting case with no potential function and in more realistic situations where a limited amount of data is used to augment a simple empirical energy. Both the use of non-local contacts included as an additional scoring function and local conformational constraints used to bias the choice of trial moves have been studied. An immediate objective is now to use both types of constraint simultaneously to take advantage of their complementary nature.

The ability of the method to tolerate error has also been studied by deliberately adding false information to the set of constraints, and the results have been found to be relatively robust. This opens up the possibilty of incorporating less reliable or even contradictory information as constraints and producing structures consistent with the net quality of the constraints.

Acknowledgments

Special thanks are given to Pierre-Jean L'Heureux and Geneviève Dufresne, who made many important contributions to the work presented here and carried out many of the calculations which have been discussed. This work has been supported financially by FCAR (Québec) and NSERC (Canada), whose contributions are gratefully acknowledged.

References

A. Aszódi, M.J. Gradwell, and W.R. Taylor, *J. Mol. Biol.* **251**, 308 (1995).

A.T. Brünger, G.M. Clore, A.M. Gronenborn, and M. Karplus, *Proc. Natl. Acad. Sci. USA* **83**, 3801 (1986).

S.H. Bryant and C.E. Lawrence, *Proteins Struct. Funct. Genet.* **16**, 92 (1993).

G. Casari and M.J. Sippl, *J. Mol. Biol.* **224**, 725 (1992).

J.R. Gunn, *J. Phys. Chem.* **100**, 3264 (1996).

J.R. Gunn, *J. Chem. Phys.* **106**, 4270 (1997).

M. Hong, J.D. Gross, and R.G. Griffin, *J. Phys. Chem. B* **101**, 5869 (1997).

O. Lund, J. Hansen, S. Brunak, and J. Bohr, *Prot. Sci.* **5**, 2217 (1996).

J. Skolnick, A. Kolinski, and A.R. Ortiz, *J. Mol. Biol.* **265**, 217 (1997).

Y. Wang, H. Zhang, W. Li, and R.A. Scott, *Proc. Natl. Acad. Sci. USA* **92**, 709 (1995).