A Branch and Bound Algorithm for Protein Structure Refinement from Sparse NMR Data Sets

Daron M. Standley*, Volker A. Eyrich, Anthony K. Felts
Richard A. Friesner and Ann E. McDermott

*Department of Chemistry and
Center for Biomolecular
Simulation, Columbia
University, New York
NY 10027, USA*

We describe new methods for predicting protein tertiary structures to low resolution given the specification of secondary structure and a limited set of long-range NMR distance constraints. The NMR data sets are derived from a realistic protocol involving completely deuterated ^{15}N and ^{13}C -labeled samples. A global optimization method, based upon a modification of the αBB (branch and bound) algorithm of Floudas and co-workers, is employed to minimize an objective function combining the NMR distance restraints with a residue-based protein folding potential containing hydrophobicity, excluded volume, and van der Waals interactions. To assess the efficacy of the new methodology, results are compared with benchmark calculations performed *via* the X-PLOR program of Brünger and co-workers using standard distance geometry/molecular dynamics (DGMD) calculations. Seven mixed α/β proteins are examined, up to a size of 183 residues, which our methods are able to treat with a relatively modest computational effort, considering the size of the conformational space. In all cases, our new approach provides substantial improvement in root-mean-square deviation from the native structure over the DGMD results; in many cases, the DGMD results are qualitatively in error, whereas the new method uniformly produces high quality low-resolution structures. The DGMD structures, for example, are systematically non-compact, which probably results from the lack of a hydrophobic term in the X-PLOR energy function. These results are highly encouraging as to the possibility of developing computational/NMR protocols for accelerating structure determination in larger proteins, where data sets are often underconstrained.

© 1999 Academic Press

Keywords: protein folding; global optimization; annealing; distance restraints

*Corresponding author

Introduction

Energy minimization in the absence of long-range constraints has not yet contributed to the determination of new protein structures. Two requirements for a solution to this problem are (1) a potential energy function which is accurate enough to distinguish the native from all other conformations; and (2) a means of globally mini-

mizing such a function. In the absence of an accurate energy function, some effort has been spent on augmenting existing functions with experimentally derived constraints. Provided the quality and quantity of the constraints are sufficient to guarantee a global minimum near the native conformation, the augmented energy functions (or target functions) may be useful for testing and developing minimization algorithms. The resulting constrained energy minimization (CEM) techniques avoid one serious problem associated with unconstrained minimization, which is that the global minimum is not, in general, known.

In addition to providing an objective target for optimization, CEM is immediately applicable to the problem of NMR structure refinement. Typi-

Abbreviations used: DGMD, distance geometry/molecular dynamics; CEM, constrained energy minimization; 3D, three-dimensional; NOE, nuclear Overhauser enhancement; CTF, C-terminal fragment; BPTI, bovine pancreatic trypsin inhibitor.

E-mail address of the corresponding author: rich@chem.columbia.edu

cally, constrained minimization has been applied to NMR structure refinement for cases in which the experimental constraints, in conjunction with steric and covalent interactions, are sufficient to define the structure. Long-range energy terms, such as the hydrophobic interaction, are not usually considered. The primary goal is to refine structures within a narrow region of conformational space defined by the constraints, rather than to select one of many competing regions. It is not clear that algorithms which are efficient at refining NMR structures with a high degree of accuracy and precision, such as distance geometry (Havel & Wüthrich, 1984), molecular dynamics (Brünger, 1992), or variable target function with conjugate gradient minimization (Braun & Gö, 1985), will perform well as global optimizers in cases where the experimental constraints alone do not uniquely define a three-dimensional (3D) structure.

Global optimization algorithms have been employed to estimate the minimum number of constraints needed to produce native-like structures using current energy functions. Smith-Brown *et al.* (1993) used a Monte Carlo approach to predict the tertiary structures of four proteins (hemerythrin, the variable light domain of an immunoglobulin, flavodoxin, and BPTI). Using a target function consisting of a Lennard-Jones potential supplemented with restraints on selected distances and on dihedral angles within helices and strands, tertiary structures were built up, one secondary structure at a time, by minimizing the target function *via* a Metropolis Monte Carlo algorithm (Levy *et al.*, 1989) implemented in the program IMPACT (Kitchen *et al.*, 1990). Here, three distances per secondary structure pair were sufficient to generate medium-low resolution structures (3–5 Å RMS from the native).

Skolnick *et al.* (1997) used a lattice model to produce medium-low resolution structures for seven proteins: the C-terminal fragment of the L7/L12 ribosomal protein (CTF), protein G, thioredoxin, flavodoxin, myoglobin, plastocyanin, and BPTI. The target function, in addition to short-range (dihedral angle) and long-range (distance) restraints, consisted of a hydrophobic term, a hydrogen-bonding term, and an overlap term. Short-range restraints included turn regions as well as helices and strands. Minimization was carried out using two stages of Monte Carlo-simulated annealing on a high-resolution lattice. Results similar to those reported by Smith-Brown *et al.* (1993) were obtained (4.0–5.6 Å RMS) using fewer distance constraints (10–35 total).

Using a hierarchical Monte Carlo-based algorithm we have obtained low-resolution (5–6 Å RMS) structures for proteins such as myoglobin (Gunn *et al.*, 1994), BPTI (Standley *et al.*, 1998), and CTF (Monge *et al.*, 1995) given a knowledge of secondary structure alone. Our results confirm that applying the statistical energy function we used to typical protein folds leads to a small number of low energy basins of attraction, one of which is

generally in the native region; however, the native region is not always lowest in energy.

The encouraging results discussed above suggest that CEM (constrained energy minimization) would be a useful way of predicting new protein structures, provided the constraints are sufficient to transform a competitive native region into a global minimum. Insufficient constraint sets occur in solution NMR experiments on larger proteins where line broadening and spectral crowding make complete assignment of NOE cross-peaks difficult or impossible. Our objective in this paper is to test this possibility by examining a constraint set based upon a realistic NMR protocol, described below. This is to be contrasted with the work discussed above, in which constraints were selected either randomly (Skolnick *et al.*, 1997) or by hand (Smith-Brown *et al.*, 1993) rather than from real or simulated data.

A recently proposed procedure for simplifying very crowded NOE spectra of larger proteins involves preparation of completely deuterated, ^{15}N and ^{13}C -labeled proteins. The hydrogen atoms on the backbone amide groups (and on some sites on the side-chains) are then exchanged for protons. The resulting spectra (H, N and C) can be assigned using heteronuclear 3D methods (Grzesiek *et al.*, 1995; Venters *et al.*, 1995), with two important advantages over the more commonly performed heteronuclear assignment protocol (Clore & Gronenborn, 1991; Bax & Grzesiek, 1993) for protonated proteins: not only is the spectral crowding reduced, but also the carbon line-widths are significantly narrower and the relaxation times significantly longer, leading to dramatically enhanced transfer efficiencies. The N-H...N-H NOE and C^α chemical shift data obtained in this protocol often allow identification of secondary structure elements. However, the issue has been raised repeatedly that the long-range NOE data from such predeuterated proteins are far too few to define the fold (Gardener *et al.*, 1997); as a result, hybrid and rather complex methods involving partial deuteration of the aliphatic groups are now under investigation. Thus, the problem of determining protein structures using underconstrained distance sets is now of central importance for solution NMR of large proteins, and probably will be in other magnetic resonance arenas as well. Here, we focus specifically on the clearly underconstrained case of a moderate sized protein for which only N-H...N-H NOE data would be available.

Our goal is to assess the prediction of seven mixed α/β proteins (Table 1) using only secondary structure and amide proton distance constraints which would be observable in an idealized N-H...N-H experiment where only amide groups are protonated. This set of proteins excludes simple β -sheets and barrels, which, in general, are completely defined by the amide constraint sets. Helical proteins were excluded as well, because there are generally too few constraints to guarantee a global minimum in the native region.

In order to deal with the multiple minimum problem, we introduce a new optimization algorithm, which, along with an improved potential energy function, is also applicable to protein structure prediction by unconstrained minimization. The optimization algorithm is based on the α BB branch and bound algorithm developed by Floudas and co-workers (Androulakis *et al.*, 1995), but has several new features which are critical in order to generate low-energy structures in complex proteins such as 1acf, 1npg, and 1xnb. We describe the optimization algorithm in detail in Protein Structure Refinement Protocols.

The distance constraints used in this work are idealized, and probably could not be measured to the accuracy and precision assumed. Nevertheless, the results represent a preliminary assessment of a global minimization algorithm which may have applications for protein structure determination in cases where there are few or no experimental constraints.

Protein Structure Refinement Protocols

Overview

Here, we describe the geometric model of the protein, distance constraints, target function, and optimization algorithm. The model and a previous version of the potential have been described elsewhere (Gunn *et al.*, 1994). Significant modifications to the energy function, as well as the new optimization algorithm, are presented for the first time. In order to generate structures consistent with the constraints, but independent of the energy function, distance geometry/molecular dynamics (DGMD) was run for each of the cases. The structures resulting from the DGMD runs were also used as alternative starting configurations for CEM runs. We also describe how disulphide bonding

constraints can be incorporated in the minimization without knowledge of the actual bonding pattern.

The model

Each residue was represented by six atoms: N, H, C α , C β , C, and O. Covalent bond lengths and bond angles were fixed to ideal values. The peptide torsional angle was fixed to 180°. The dihedral angles ϕ and ψ were, thus, the only independent variables. The Cartesian and dihedral representations are equivalent, and each can be constructed from the other.

Distance constraints

The simulated NOEs (nuclear Overhauser enhancements) consisted of all H_{amide} – H_{amide} distances under 5 Å and separated by more than three residues. Upper and lower bounds were derived by expanding ± 1.75 Å from the native distances. For the CEM runs, distances within secondary structures were removed. The distance constraint sets are plotted in Figure 1 and listed explicitly in the Appendix.

The target function

The target function consisted of five parts: a hydrophobic term, a van der Waals term, an overlap penalty, a torsional term, and a distance constraint term:

$$T = E_{\text{hyd}} + E_{\text{vdw}} + E_{\text{ovr}} + E_{\text{tor}} + E_{\text{dis}} \tag{1}$$

We give below a brief description of each term in the potential; a more detailed exposition will be presented in another publication.

Table 1. Structural data for the seven β -containing structures considered in the present work

Code	Reference	Protein	N_{res}	N_{dis}	Secondary structure	Fold
1acf	Fedorov <i>et al.</i> (1994)	Actin-binding protein	123	53	4 α /7 β	$\alpha + \beta$, 2-Layer sandwich (beta-lactamase)
1ctf	Leijonmarck & Liljas (1987)	L7/L12 50 S ribosomal protein (C-terminal domain)	67	23	3 α /3 β	$\alpha + \beta$, 2-Layer sandwich (alpha-beta plaits)
1npg	Morera <i>et al.</i> (1994)	Nucleoside diphosphate kinase	143	43	11 α /4 β	$\alpha + \beta$, 2-Layer sandwich, (alpha-beta plaits)
4pti	Wlodawer <i>et al.</i> (1987)	Proteinase inhibitor (trypsin)	56	12	2 α /2 β	$\alpha + \beta$, Few secondary structures (factor Xa inhibitor)
1ris	Lindahl <i>et al.</i> (1994)	Ribosomal protein S6	92	48	2 α /4 β	$\alpha + \beta$, 2-Layer sandwich (alpha-beta plaits)
1ubi	Ramage <i>et al.</i> (1994)	Ubiquitin	72	36	2 α /5 β	$\alpha + \beta$, Roll (ubiquitin-like)
1xnb	Campbell <i>et al.</i> (1994)	Endo-1,4-beta-xylanase	183	148	1 α /16 β	Mainly β , sandwich (jelly roll)

Coordinates are taken from the Protein Data Bank (Berstein *et al.*, 1977; Abola *et al.*, 1987); secondary structure assignments were taken from the PDB files (with the exception of the 4pti secondary structure, which was taken from the PDB file 5pti); descriptions of the fold were taken from the CATH website (www.biochem.ucl.ac.uk/bsm/cath).

Figure 1. Flowchart for the Branch & Bound algorithm used for constrained minimization. The Figure illustrates the basic steps of the three phases of the program (initialization, search, and convergence) both textually (left) and symbolically (right).

The hydrophobic term is based on the potential of mean force derived by Sippl and co-workers (Casari & Sippl, 1992) and has the following form:

$$E_{hyd} = C_{hyd} \sum_{i=1}^{N_{res}-10} \sum_{j=i+10}^{N_{res}} \langle h_{ij} \rangle d_{ij}^{\beta\beta} \quad (2)$$

where $d_{ij}^{\beta\beta}$ is the $C^\beta - C^\beta$ distance between residues i and j , and $\langle h_{ij} \rangle$ is the average hydrophobicity for each residue pair. The $\langle h_{ij} \rangle$ value is given by $1/2(h_{ij} + h_{ji}) + 2 h_0$, where h_{ij} and h_{ji} can be found in Table 2 of Casari & Sippl (1992) and h_0 is the

background hydrophobicity which must be empirically optimized. Here $h_0 = 0.3$ in all cases.

The van der Waals term has the following form:

$$E_{vdw} = C_{vdw} \sum_{i=1}^{N_{res}-1} \sum_{j=i+1}^{N_{res}} \sum_{k=1}^2 \sum_{l=1}^2 M(d_{ij}^{kl}) \quad (3)$$

where k and l denote the atoms within the residue (C^α or C^β), i and j are the residue numbers, and $M(d_{ij})$ is a truncated Morse function of the distance d_{ij}^{kl} . For a given distance d , M is given by:

$$M(d) = \begin{cases} \varepsilon(1 - e^{-\beta(d-\sigma)})^2 - \varepsilon & \text{if } d > d_0 \\ ad^4 + b & \text{if } d \leq d_0 \end{cases} \quad (4)$$

where ε , β , and σ are functions of i , j , k and l ; a and b are chosen so that M and its derivative are continuous at the x intercept ($d_0 = \sigma - \ln 2/\beta$) and the slope of M is zero at the y intercept. The parameters ε , β , and σ were determined from functional fits of the Morse function to data obtained from a database of known protein structures (not including the seven proteins considered here) through a four-step procedure: first, the all-atom van der Waals energies between each residue pair in the database were calculated with a Lennard-Jones potential using AMBER (Weiner *et al.*, 1984, 1986) parameters; second, each energy was binned according to the residue pair and the four non-bonded distances: $C^\alpha - C^\alpha$, $C^\alpha - C^\beta$, $C^\beta - C^\alpha$, $C^\beta - C^\beta$; third, the energies in each bin were averaged; fourth, the parameters ε , β , and σ of the Morse function were fit to the averaged energies along each of the four non-bonded distances for each residue pair.

The overlap term has the following form:

$$E_{ovr} = C_{ovr} \sum_{i=1}^{N_{res}-3} \sum_{j=i+3}^{N_{res}} \sum_{k=1}^2 \sum_{l=1}^2 e^{-(d_{ij}^{kl}/d_0)^{10}} \quad (5)$$

where the minimum distance d_0 is identical to that used in the Morse potential (4).

The torsional term is essentially an overlap penalty for C^β atoms on neighboring residues.

$$E_{tor} = C_{tor} \sum_{i=1}^{N_{res}-1} \begin{cases} e^{-(d_{ii+1}^{22}/d_0')^{10}} & \text{if } d < 5 \text{ \AA} \\ 0 & \text{if } d \geq 5 \text{ \AA} \end{cases} \quad (6)$$

The parameter d_0' is chosen so that the minimum observed $C^\beta - C^\beta$ distance for a given residue pair corresponds to a negligible contribution from E_{tor} .

The distance constraint term is given by

$$E_{dis} = \frac{1}{N_{dis}} \sum_{i=1}^{N_{dis}} C_{dis} \begin{cases} d^i - d_{max}^i & \text{if } d^i > d_{max}^i \\ d_{min}^i - d^i & \text{if } d^i < d_{min}^i \\ 0 & \text{if } d_{min}^i < d^i < d_{max}^i \end{cases} \quad (7)$$

where N_{dis} is the number of distances in the constraint set. The weights of each distance constraint, C_{dis} , were set to 1000.

The coefficients for the overlap and torsional terms were varied during the optimization process

(see below), but the final values were: $C_{\text{hyd}} = C_{\text{vdw}} = 1.0$, and $C_{\text{ovr}} = C_{\text{tor}} = 100$. In order to determine the final weights, C_{hyd} was set to 1.0, and a grid search was performed over C_{vdw} and C_{ovr} . (C_{tor} was set equal to C_{ovr} because the functional form of these two terms was identical.) At each value of C_{vdw} and C_{ovr} , the full potential (without distance constraints) was minimized, starting from a randomly perturbed native structure†. The perturbed structures were generated by randomly changing one dihedral angle (chosen at random) in the native structure. By inspecting the distribution of structures resulting from more than 100 such minimizations, it was possible to find a range of values for C_{vdw} and C_{ovr} over which the native region was relatively stable. Based on these observations, the overlap weight C_{ovr} was initially set to 500; however, after repeated attempts to globally optimize the potential with $C_{\text{ovr}} = 500$, we found that native-like energies could only be achieved if C_{ovr} was set to 100. Because the native structure has a very low overlap and torsional energy, changing C_{tor} and C_{ovr} by a factor of 5 will not significantly affect the potential surface close to the native.

A calculation of the target function along with its gradient consists of roughly N^2 evaluations of the exponentials in the van der Waals and overlap penalty expressions. Since exponential functions are computationally expensive, local minimization times can be dominated by the energy and gradient terms. To avoid this problem, a more efficient, four-point "Newton-Gregory" interpolation scheme (Allen & Tildesley, 1987) was used to calculate the energies between each residue pair. The interpolation was performed over a grid of precalculated values of these two terms over equally spaced distance intervals. The exponentials in the van der Waals and overlap expressions are encountered only while making this grid, which is done in the initialization stage of the global minimizer.

The local minimizer uses an updated Hessian (starting from the unit matrix). Due to the linear terms in the constraint penalty, the true Hessian would have infinite values at points where the first derivative is not continuous. However, the updated Hessian interpolates over these points, and thus does not ever go to infinity. For this reason, the minimization is stable.

Constrained energy minimization

Overview

The minimization is carried out *via* a branch and bound algorithm coupled with dynamic smoothing of the target function surface. Bounds are defined in terms of a set of distances including, but not limited to, the experimental distance constraints. The basic steps are outlined in Figure 2. Branching involves two steps: the first step is to partition (bisection) the bounds on a particular distance, thus creating two "subspaces", each with non-overlapping bounds for the bisected distance; the second step is to minimize the smoothed target function inside each subspace. "Smoothing" in the context of this paper means augmenting the target function in order to reduce the number of local minima. As with the αBB algorithm developed by Floudas and co-workers (Androulakis *et al.*, 1995), it is required that for a given conformation, the smoothed target function is always less than or equal to the original target function. Furthermore, as the size of the subspace goes to zero, the smoothing must vanish. Ideally, the minimum of the smoothed target function constitutes a "lower bound" on the minimum energy of the original function within a given subspace.

Following partitioning and minimization in each half of the partition, the bounds, coordinates, and energies associated with each subspace are saved. The next step is to locate the lowest saved smoothed energy. The bounds corresponding with this energy are removed from the saved list. The partitioning, minimization, and saving steps are repeated until a convergence criterion is reached. Early in the minimization, when the smoothing effect is greatest, subspaces are selected primarily on the basis of size. As the subspaces grow in number and shrink in size, however, the selection becomes more influenced by the shape of the original target function.

The third phase is convergence. When a subspace reaches a minimum size, the coordinates of the structure are written out and a penalty function is constructed which repels conformations in subsequent minimizations from being attracted to previously located minima.

The key features of the αBB algorithm (Androulakis *et al.*, 1995; Maranas & Floudas, 1994a,b; Adjiman & Floudas, 1996; Adjiman *et al.*, 1998a,b) which were implemented in the present work are: (1) construction of a smoothed function which returns to the original target function as the size of the bounds decreases; and (2) selection of subspaces for further partitioning based on the minima of the smoothed function.

There are, in addition to the similarities, several important differences between our method and the αBB algorithm. (1) Smoothing of the potential surface is accomplished primarily by annealing in the overlap, torsional, and constraint terms, with a smaller contribution from the smoothing function

† This procedure was carried out on 50 different proteins. Although 1acf and 1ctf were included among the structures used in this Monte Carlo procedure, we expect that the influence of these structures on our final choice of C_{vdw} and C_{ovr} was not significant, since such a large number of structures was considered and since the goal was only to determine the right order of magnitude of the weights.

Figure 2. Experimentally derivable distance constraints are plotted in red on the lower half of the distance matrix. Blue regions show poorly defined regions in the ensemble of DGMD structures. For all DGMD structures with two or less distance or dihedral angle constraints (combined) the RMS deviation of each $C^\alpha - C^\alpha$ distance was calculated. Blue marks indicate an RMS > 10.0 Å. Secondary structure cartoons are also shown on the x-axis, which correspond with the residue numbers.

used in the $\alpha\mathbf{BB}$ algorithm (equation (12)), (2) The smoothed function is not, in general, convex; it is merely smoother than the original target function; the minima of the smoothed target function are not, therefore, a rigorous lower bound. (3) Convergence is defined in terms of subspace size rather than by bracketing the target function between upper and lower bounds; thus, there is no need for calculating an "upper bound" to the global minimum. (4) Minima which have converged are "filled in" by constructing a Gaussian-shaped penalty function at the location of the minimum, thus many distinct minima may be located because the converged structures repel similar structures in subsequent minimizations.

Conformational space

The conformational space is defined in terms of a set of distances d_j^i and their associated upper and lower bounds $d_{\max,j}^i$ and $d_{\min,j}^i$. Here, i refers to a particular residue pair and j is the index of a given subspace; $j = 0$ refers to the initial (global) bounds. Both distance constraints, and distances for which the bounds are essentially unknown, which we will refer to as "free distances," are used in the definition of the conformational space. Free distances are chosen from regions in the structure where constraints are lacking (Figure 3). There are a minimum of two free distances per helix which tether the helices to the sheets. Subdividing bounds on these distances enables sampling of different helix-sheet orientations. Additional free distances tether loops to sheets and helices to one another. The free distances are listed in Table 2.

The distance constraint function (equation (7)) is used to enforce bounds on the free distances, but the weights C_{dis} are 100, rather than 1000. Only the constraints are included in the final evaluation of the target function (e.g. in the convergence, phase). In this way, the free distances are used only to drive the conformational search with the penalty function acting to prevent minimizations in one subspace from falling into a neighboring subspace.

Partitioning

The first step in the search phase is to bisect the current subspace B along the distance with the widest bounds. This distance d^k is given by $d_{\max}^k - d_{\min}^k \geq d_{\max}^i - d_{\min}^i$ for all i . Bisection occurs at the midpoint: $d_{\text{mid}}^k = d_{\min}^k + 1/2(d_{\max}^k - d_{\min}^k)$. The two new subspaces created by bisection B_1 and B_2 are identical, except for the k th distance bound which is given by $(d_{\min}^k, d_{\text{mid}}^k)$ for B_1 and $(d_{\text{mid}}^k, d_{\max}^k)$ for B_2 .

Smoothing

The second step of the search phase is to perform a local minimization of the smoothed target function in each of the new subspaces, B_1 and B_2 . While the target function is rather smooth when

plotted as a function of the distances, the angular dependence is much more rugged. From inspection of the potential surface in angle space, it was found that most of the minima could be removed by reducing the weights of the overlap, torsional, and constraint terms. In keeping with the philosophy of the $\alpha\mathbf{BB}$ algorithm, the smoothing should diminish with box size, restoring the potential surface to its original state in the limit of zero subspace size. In order to anneal the original target function in gradually with each subdivision, the weights of the reduced terms in the potential must be functions of the size of the subspace. We will define the fractional width of the j th subspace as:

$$f_j = \sum_{i=1}^{N_{\text{free}}} \frac{(d_{\max,j}^i - d_{\min,j}^i)}{(d_{\max,0}^i - d_{\min,0}^i)} \quad (8)$$

where the sum is only over the free distances (since bounds on the constraints are never subdivided). If $f_{\min}$ is the fractional width at which we want the weights to be completely restored, the following relationships will ensure that the weights grow linearly as f_j decreases:

$$C_{\text{our},j} = C_{\text{our},0} \left(\frac{1 - f_j}{1 - f_{\min}} \right) \quad (9)$$

$$C_{\text{tor},j} = C_{\text{tor},0} \left(\frac{1 - f_j}{1 - f_{\min}} \right) \quad (10)$$

$$C_{\text{dis},j} = C_{\text{dis},0} \left(\frac{1 - f_j}{1 - f_{\min}} \right) + 1.0 \quad (11)$$

Note that the initial overlap and torsional weights are zero, while the initial constraint weight is 1.0. Thus, the constraint term is never completely turned off. The value at which the coefficients are restored $f_{\min}$, along with the size of the protein and the number of constraints, affects the length of the minimization: The smaller $f_{\min}$ is, the longer the program will run. The values used in the present work ($1\text{acf} = 10.5$; $1\text{ctf} = 0.6$; $1\text{npk} = 0.8$; $4\text{pti} = 0.4$; $1\text{ris} = 0.7$; $1\text{ubi} = 0.5$, $1\text{xbn} = 0.6$) were chosen based on an acceptable convergence time. If the first structure converged too quickly (<3 hours), or too slowly (>72 hours), $f_{\min}$ was adjusted accordingly. In several cases, the above $f_{\min}$ values were reduced without a decrease in the energy of the lowest-energy converged structure, suggesting that the minimization could not be improved by further reducing $f_{\min}$. (This does not mean that the minimization could not be improved, for example, by increasing the number of free distances or by using alternate starting configurations.) By examining a large number of minimizations, we hope to arrive at an automated method of picking $f_{\min}$ based on the size of the structure and the associated constraint set.

The target function can be further augmented by adding a convex function of the type used in the $\alpha\mathbf{BB}$ algorithm (Androulakis *et al.*, 1995; Maranas

Figure 3. RMS deviation from native *versus* energy for structures resulting from three CEM minimizations: Hybrid DGMD-CEM, green ($\diamond$); CEM, red ($\square$). After each of the three runs, the structures were clustered (Protein Structure Refinement Protocols section) at a cutoff RMS of 3.0 Å. Only the lowest-energy structure of each cluster is shown.

Table 2. Free distances

Residue pair						
1acf	2 W-20 A 49 F-97 V	8 T-87 I 27 N-123 Q	53 D-60 F	58 G-69 T	94 A-114 E	40 A-62 L
1ctf	54 F-77 G 92 A-113 A	60 A-64 N 98 V-100 K	64 N-88 E 100 K-120 K	77 G-79 G 113 A-116 E	79 G-98 V	88 E-92 A
1npk	9 E-88 V 57 A-77 V	16 K-95 I 59 H-83 E	25 V-78 V 68 L-110 G	35 K-83 E 78 V-128 E	49 K-121 I 88 V-143 L	55 H-124 S 98 T-150 N
4pti	1 R-24 N 31 Q-50 D	7 E-18 I 3 D-20 R	29 L-56 G 8 P-12 G	35 Y-47 S 14 C-17 R	3 D-50 D 37 G-41 K	4 F-56 G 42 R-46 K
1ris	25 I-42 E 25 I-76 A	61 L-76 A 25 I-89 M	76 A-89 M 42 E-76 A	6 V-42 E 42 E-89 M	6 V-76 A	25 I-61 L
1ubi	13 I-28 A 4 F-49 Q	28 A-42 R 4 F-57 S	49 Q-57 S 13 I-42 R	57 S-68 H 13 I-49 Q	4 F-28 A 13 I-57S	4 F-42 R 13 I-68 H
1xnb	28 A-49 Q 21 G-61 N	28 A-57 S 21 G-101 D	28 A-68 H 41 G-81 V	42 R-57 S 121 D-161 G	49 Q-68 H 141 N-181 N	

Bounds on the free distances, along with the distance constraints, define the conformational space of each molecule. The bounds on free distances are $[1 \text{ \AA}, d_j^i + 20 \text{ \AA}]$, where d_j^i is the distance of the starting conformation following a local minimization of the target function.

& Floudas, 1994a,b; Adjiman & Floudas, 1996; Adjiman *et al.*, 1998a,b):

$$E_{\text{con},j} = \sum_{i=1}^{N_d} -\alpha^i (d_{\text{max},j}^i - d_j^i)(d_j^i - d_{\text{min},j}^i) \tag{12}$$

where α^i , is a number greater than zero, and the sum includes all of the distances which define the conformational space of the molecule. For all i corresponding to distance constraints, $\alpha^i = 1.0$; for the free distances, $\alpha^i = 0.05$. The complete smoothed target function, with the exception of the Gaussian repelling function (see below) is thus given by:

$$T_{\text{smooth},j} = T_j + E_{\text{con},j} \tag{13}$$

where the T_j now implicitly depends on the fractional width f_j . Note that $T_{\text{smooth},j}$ is always less than or equal to the unsmoothed target function within a given subspace, as required.

Enforcing distance bounds

The independent variables in our model are the dihedral angles, but the bounds themselves are on distances. Before a structure can be minimized within a subspace, it must first be placed within the bounds of that subspace and then constrained within the subspace in subsequent minimizations. When a subspace is divided, the structure associated with the parent bounds will only satisfy one of the bounds on the two newly created subspaces (assuming the parent bounds were satisfied). Therefore, at each subdivision there is a need for placing a structure within one of the new bounds. It turns out that E_{dis} (equation (7)) and E_{con} (equation (12)) are generally (but not always) convex provided the number of distances is not too large (as is usually the case with $H_{\text{amide}} - H_{\text{amide}}$ distances). For this reason, a local minimization of the potential with all terms except E_{dis} and E_{con} turned off and all $\alpha^i = 10.0$, is usually sufficient to generate an initial conformation for which each

distance is no more than 1 Å outside of the bounds. In general, the structures do not move further from the bounds upon minimization of the full smoothed target function; in fact, minimization of the smoothed target function often results in a lower distance penalty (probably due to the perturbation of the $E_{\text{dis}} + E_{\text{con}}$ potential surface and the additional minimization steps which follow). Thus in each cycle of the search phase, there are generally three local minimizations: one local minimization of E_{dis} and E_{con} and two local minimizations of E_{smooth} .

Selecting the current subspace

After two local minimizations of the smoothed target function in the search phase have been performed, the resulting bounds, coordinates, and energies are saved. The next step is to find the lowest saved energy $T_{\text{save}}(i^*)$. The current bounds are then set to those associated with the i^* th energy $B_{\text{save}}(i^*)$. The coordinates $X_{\text{save}}(i^*)$ are used as initial conditions for the next local minimization of T_{smooth} . Finally, the i^* th energy, coordinates, and bounds are removed from the saved list.

Convergence

The branch and bound cycle is repeated until the fractional width falls below a certain value for a given subspace. This is the convergence criterion. In each of the runs, the convergence criterion was given by $f_{\text{min}} - 0.1$, where the f_{min} are listed above. That is, for a given subspace, after the potential is fully annealed, the branch and bound cycle continues until the fractional width decreases by 0.1. (This value was chosen somewhat arbitrarily to ensure that bounds on the fully annealed potential were subdivided at least once before convergence.) Any subspace for which the fractional width equals $f_{\text{min}} - 0.1$ is said to have converged. In the convergence phase, a local minimization of the unsmoothed target function is performed. As mentioned above, the free distances are not included in

this minimization. Following minimization, the target function and coordinates of the structure are recorded. A Gaussian penalty, which prevents similar structures from falling into minima located previously, is then constructed. The Gaussian penalty for a structure j is given by:

$$E_{\text{Gauss}} = H_G \sum_i^{N_{\text{conv}}} e^{-W_G R_{ij}^2} \quad (14)$$

where H_G and W_G are the height and width of the Gaussian, respectively (in the present work $H_G = 100.0$ and $W_G = 1.5 \text{ \AA}$); N_{conv} is the number of converged structures, R_{ij} is the RMS deviation between $C^\alpha - C^\alpha$ distances in structures i and j :

$$R_{ij} = \sqrt{\frac{1}{N_D} \sum_k^{N_D} (d_i^{\alpha\alpha,k} - d_j^{\alpha\alpha,k})^2} \quad (15)$$

where $N_D = 1/2N_{\text{res}}(N_{\text{res}} - 1)$ and the sum goes over all $C^\alpha - C^\alpha$ distances within structures i and j . E_{Gauss} (equation (14)) is thus a function of the coordinates of the converged structures. E_{Gauss} is included in the smoothed target function in all subsequent minimizations, and must be updated after each new structure is converged. E_{Gauss} effects the smoothed target function associated with each of the saved structures, and thus the energy rank of each structure. Because the subspace with the lowest smoothed target function is chosen for further bisection, and because the Gaussian penalty penalizes structures similar to previously converged structures, unconverged basins of attraction get an improved energy rank relative to regions already located. Thus, E_{Gauss} works in conjunction with the convex function E_{con} (equation (12)), which favors large subspaces over small ones. (Small subspaces would be more likely to contain structures which are similar to other saved structures). In the final evaluation of the target function, neither E_{Gauss} nor E_{con} is included; each is used only to guide the exploration of subspaces. While E_{con} may be thought of as an "entropic reward" because it is a function only of the size of the subspace, equation E_{Gauss} is a "redundancy penalty" and is a function of the converged structures.

Disulphide bonds

Because the conformational space is defined in terms of distances, it is relatively straightforward to explore different disulphide bonding patterns. First, each possible disulphide pair is included in the list of free distances. Then, instead of starting with the global bounds at the beginning of the search phase, the global bounds are prepartitioned into one set of bounds for every possible disulphide bonding pattern. If there are $2N_{di}$ cystine residues involved in disulphide bonds, then there are $\{2N_{di} - 1\}\{2N_{di} - 3\} \dots 1$ possible ways of forming N_{di} disulphide bonds. For example, there are three disulphide bonds in BPTI, and thus 15 differ-

ent possible ways of forming them. Since all 15 are included at the beginning of the program, there is no bias for the correct pattern to be selected. The selection is based solely on the basis of the smoothed energy associated with each subspace. Note that we are assuming knowledge of the number of disulphide bonds, but that this is not in general, required. One could, in principle, prepartition the bounds to include variation in the number of bonds as well.

Because disulphide bonds were not included in the DGMD runs, they were introduced only in the second CEM stage (which was run on both CEM and DGMD structures). A special procedure was used to include the disulphide constraints into the starting configurations in stage 2. First, the initial configurations (CEM or DGMD) were locally minimized with the full potential without disulphide constraints. Second, the structure was locally minimized using the full potential and one of the disulphide bonding patterns. This was followed by a Monte Carlo minimization which consisted of initially perturbing all of the angles by a random value (not more than ± 0.5 radians), and then locally minimizing of the resulting structure. Following ten such Monte Carlo minimizations, the lowest energy configuration was retained. This whole process (local minimization followed by ten Monte Carlo minimizations) was repeated once for each disulphide bonding pattern and this, in turn, was repeated once for each of the initial starting configurations. Thus, if there were originally ten starting configurations and 15 disulphide bonding patterns, there would be 150 starting configurations at the end of this process. The remainder of the CEM run was identical to the other stage 2 CEM runs.

Clustering

Since large numbers of structures are generated by both DGMD and CEM, similar structures are combined into clusters following each DGMD or CEM run. We follow the method described by Romesburg (1984), in which structures are iteratively combined into increasingly large and increasingly dissimilar clusters. Initially, each structure constitutes a cluster; the clustering proceeds until a cutoff in the average RMS deviation of the non-loop residues between members of each cluster is reached. The loops are ignored in our implementation so that clusters are defined in terms of the arrangement of secondary structures alone. In all runs, a cutoff of $3.0 \text{ \AA}$ was used.

Hybrid distance geometry-molecular dynamics

DGMD calculations were performed using the program X-PLOR (Brünger, 1992), with the input files controlling sub-embedding, distance geometry simulated annealing, and refinement taken directly from tutorials (Nilges *et al.*, 1988, 1991; Kuszewski *et al.*, 1992). Distance constraints were the same as

those used in the CEM runs, with the exception that the distances within secondary structures were not removed. Secondary structures were fixed by constraining the dihedral angles to native values $\pm 5^\circ$ (i.e. calculated from the PDB file). The peptide torsional angle ω was constrained to $180(\pm 5)^\circ$. (The one exception is residue 91 in 1ctf, which is the only instance of a *cis* amide in the present work. Since we would not have known the location of a *cis* amide *a priori*, this correction was not made in the CEM runs.)

Results

Figure 2 is a two-dimensional representation of the distance constraints. In the lower right triangle, the constraints are shown in red. In the upper left triangle, blue regions correspond to $C^\alpha - C^\alpha$ distances for which the RMS deviation (within the ensemble of DGMD structures) is 10 Å or greater. Note that the secondary structures were well defined ($RMS < 1.5$ Å) in both the CEM and DGMD runs.

Table 3 lists the results of the DGMD, CEM, and hybrid DGMD-CEM runs. The DGMD results are listed in columns 1-4. For the CEM runs, the minimization process was done in two stages. In stage 1, the overlap, torsional, and distance constraint terms were annealed in and the van der Waals term was turned off. In stage 2, the overlap, torsional, and distance constraint weights were fixed to their final values, and the van der Waals term was turned on. The first stage started from extended conformations. Structures that converged in the first stage were used as starting configurations in the second stage. The CEM results are listed in columns 10-15. For the hybrid DGMD-CEM runs, DGMD was used to generate the starting configurations for CEM. The CEM run was otherwise identical with the second stage described above. In each DGMD run, 50 structures were generated. Only those clusters that had less than two distance or dihedral angle violations (combined), were used as starting configurations for CEM. Note that the target functions used in the DGMD and CEM methods are different, so the energies of the DGMD and CEM structures cannot be compared. The data from the CEM runs, as well as the hybrid DGMD-CEM runs, is plotted in Figure 3. Only the five lowest energy clusters for each run are listed. The DGMD structures were dissimilar enough that many clusters consisted of one structure. In the CEM runs, there were generally multiple structures in clusters and the number of structures is listed (Table 3, columns 9 and 15). For all runs the RMS deviations from the native of the C^α atoms (R_T) and of C^α atoms located in secondary structures (R_S) refer only to the lowest-energy member of each cluster. For the CEM and hybrid DGMD-CEM results, the average RMS deviations ($\bar{R}_T$ and $\bar{R}_S$) refer to the entire cluster. Similarly, the X-PLOR energies (E_X) and CEM energies (E_C) refer

only to the lowest energy members of each cluster, while the average CEM energies, (E_C) refer to the cluster.

By comparing the CEM and DGMD runs (Table 3 and Figures 4 to 10), we can assess the importance of the potential energy terms. Figure 3, on the other hand, emphasizes the importance of the global minimization algorithm. Both the energy terms and the minimization algorithm are essential to the success of this method.

Discussion

From the DGMD results we see that for the proteins 1acf, 1ctf, 1npk, and 4pti, there is not enough information in the distance constraints alone to define an accurate structure. If we consider only the arrangement of secondary structure, DGMD produces a low-resolution structure for 1xnb. The primary error in the 1xnb structure is the placement of a single loop. In the native structure, this loop is buried, while in the DGMD structure it projects into what would be the solvent (see Figure 10). Generally, all of the DGMD structures are expanded, with helical regions projecting out into space, resulting in a larger radius of gyration than the native. Even 1ris and 1ubi, which have essentially the correct topology, do not form compact structures.

The lack of compactness in the DGMD structures is consistent with the potential used in the X-PLOR program; namely the interaction between an unconstrained segment, such as a helix, and the rest of the molecule is governed by only a hard-core repulsive term (although other terms could, in principle, be added). Without a long-range attractive term such as a hydrophobic interaction, there is no driving force for compactness. From the hybrid DGMD-CEM results, the improvement in RMS deviation from the native and in compactness upon minimization with a more complete energy function is evident. With the exception of 1npk, the secondary structure RMS deviations are all less than 6 Å; moreover, the average RMS deviations are very close to the minimum-energy values, meaning that the vast conformational space consistent with the distances has been reduced to a narrow low-energy region.

In the cases of 1acf and 1npk, however, the hybrid DGMD-CEM method fails completely. This is apparently the result of being trapped in local minima, as the energies are much higher than the more accurate structures produced by CEM (Figures 4 and 6). These minima are probably due to the fact that when distance geometry is used to refine sparse NMR data, mirror images, which satisfy the distance constraints, are often generated; since molecular dynamics (and presumably CEM) cannot recover from such severe mistakes, the structures remain high in RMS (G. Warren, personal communication). All of the hybrid DGMD-CEM results prove to be local minima when com-

Table 3. Energies and RMS deviatons (relative to the native) of low-energy structures resulting from DGMD, CEM and hybrid DGMD-CEM runs

DGMD			Hybrid DGMD-CEM						CEM					
R_T	R_S	E_X	R_T	R_S	$\overline{R_S}$	E_C	$\overline{E_C}$	N	R_T	R_S	$\overline{R_S}$	E_C	$\overline{E_C}$	N
A. <i>1acf</i>														
15.89	16.86	120.21	11.35	12.01	11.96	1243.22	1252.31	2	5.08	4.97	4.83	-143.93	82.35	8
15.38	16.35	121.51	9.70	9.36	9.21	1300.85	1358.13	2	4.00	3.81	4.43	-74.63	63.92	6
12.40	13.28	125.86	10.20	9.13	9.33	1385.04	1436.93	2	6.95	7.16	7.14	-68.20	31.36	5
12.61	13.29	127.51	11.61	11.06	11.01	1412.57	1426.62	3	7.02	7.26	7.36	-40.28	-3.29	3
14.93	16.05	128.35	12.19	11.81	11.71	1415.57	1742.11	7	4.63	4.39	4.34	-14.60	118.94	3
B. <i>1ctf</i>														
9.60	9.34	72.93	3.40	3.41	3.90	1845.36	-1610.15	3	1.96	1.82	2.91	-1873.47	-1750.32	7
9.42	9.60	73.84	3.24	3.25	2.69	-1775.57	-1705.89	4	2.99	2.91	3.01	-1852.51	-1730.04	3
8.95	9.24	75.26	5.30	5.25	4.87	-1734.66	71555.25	5	4.68	4.61	4.25	-1807.06	-1726.75	7
10.86	10.88	75.71	3.61	3.61	4.31	-1701.02	-1632.64	3	9.16	8.55	8.50	-1745.54	-1669.53	3
8.71	8.94	77.09	5.85	5.95	5.64	-1639.93	-1621.52	3	9.09	9.05	9.18	-1737.15	1664.10	2
C. <i>1npk</i>														
17.76	15.65	196.27	9.14	8.36	8.25	71420.99	-1107.69	4	5.40	4.40	4.43	-2169.11	-1931.28	17
16.61	16.08	197.00	7.96	7.66	7.30	-1405.72	-1112.42	4	6.54	5.74	5.76	-2063.45	-1802.82	14
15.74	14.89	198.03	6.89	6.34	6.56	-11269.63	1073.15	3	6.38	5.63	5.52	-2027.38	-1807.47	7
15.21	14.44	199.27	8.94	8.15	8.21	-1014.53	3998.76	2	4.81	4.28	4.43	-1987.41	-1714.54	6
15.06	14.05	202.99	7.76	7.52	7.52	-968.84	2.968.84	1	6.52	6.05	5.90	-1963.37	-1841.00	6
D. <i>4pti</i>														
11.68	12.36	71.58	6.41	4.54	4.87	-968.00	-864.43	9	5.68	4.37	4.92	-1109.69	-932.89	4
8.28	9.05	71.96	8.53	6.20	6.73	-887.25	-786.29	8	8.50	7.10	7.11	-1076.31	-1013.79	3
9.45	9.73	72.68	8.69	8.15	7.85	-841.26	-731.94	7	5.13	4.29	4.55	-1061.73	-993.36	4
9.95	10.68	74.40	7.60	5.51	5.51	-812.58	7,812.58	1	7.71	6.45	7.09	-1058.23	-967.32	5
12.45	12.23	75.78	6.47	4.80	5.19	-791.13	-620.09	3	4.44	3.22	3.59	-1047.52	-1025.86	3
E. <i>1iris</i>														
6.52	6.42	125.60	3.86	3.52	3.61	-3262.49	-3155.09	5	3.71	3.35	3.66	-3425.40	-3121.43	10
5.91	5.68	126.90	2.69	2.47	2.86	-3248.76	-2984.99	7	3.23	3.06	3.29	-3314.51	-3007.25	6
6.07	5.79	127.78	4.11	3.89	3.77	-3179.63	-3138.41	3	2.91	2.69	2.81	-3293.80	-3011.34	11
6.39	6.01	127.82	4.16	4.17	4.17	-2808.87	-2808.87	1	4.41	3.86	3.77	-3289.96	-3234.51	4
6.37	6.01	136.32	4.30	3.81	3.81	-2747.36	-2747.36	1	3.46	3.19	3.36	-3272.72	-3109.63	9
F. <i>1ubi</i>														
5.62	5.93	90.70	4.44	4.39	4.68	-1972.46	-1738.92	12	2.53	2.29	2.66	-2784.25	-2313.44	24
5.85	6.05	91.08	3.79	2.81	2.74	1953.77	-1501.82	13	3.95	3.58	3.63	-2632.25	-2314.41	9
6.09	6.16	91.83	4.33	3.89	4.03	7,1812.28	-1431.76	7	4.32	4.00	3.65	-2561.25	-2344.45	19
5.92	6.31	92.07	6.02	5.94	6.01	-1604.01	-1147.97	16	3.23	3.04	3.28	72558.12	-2293.39	11
5.27	4.97	96.08	5.84	5.57	5.57	-1480.63	-1480.63	1	3.67	3.32	3.38	-2556.40	-2208.72	14
G. <i>1xnb</i>														
8.50	5.72	198.17	6.09	5.07	4.87	3273.13	3753.14	24	5.16	4.53	4.57	1604.29	2410.28	32
9.74	7.07	214.40	5.38	4.22	4.36	3907.49	4968.80	42	5.79	4.97	4.92	1886.67	2502.03	27
9.84	7.81	221.77	5.19	4.13	4.30	3984.57	4448.82	21	6.33	5.60	5.75	2101.21	2410.86	14
9.27	7.50	224.96	6.18	4.93	4.84	4637.15	4808.90	4	6.22	5.28	5.28	2409.23	2780.05	15
9.42	6.85	238.97	5.49	4.16	4.91	4685.15	5171.93	9	8.44	7.19	7.05	2838.31	3044.48	6

The five lowest-energy clusters from each of the runs are listed. Each cluster is represented by the lowest-energy structure within that cluster. For the DGMD runs, the RMS deviations from the native of the C^α atoms (R_T) and of C^α atoms located in secondary structures (R_S) are listed, as well as the X-PLOR energy (E_X). The CEM and hybrid DGMD-CEM runs differ in the first stage only; the final target function is the same in both cases. For columns 4-9, DGMD was used to generate the starting configurations for CEM. In columns 10 ~ 15, starting configurations for stage 2 were generated by running CEM without the van der Waals term. Along with R_T and R_S , the average RMS deviation of the secondary structures within each cluster ($\overline{R_S}$), CEM energy (E_C), average CEM energy ($\overline{E_C}$), and number of structures in each cluster (N) are listed. The hybrid DGMD-CEM and CEM data correspond with that plotted in Figure 3.

pared with the energies of the best CEM structures. The key to locating these lower energy structures appears to be the annealing. The local minima are primarily due to the van der Waals, overlap, torsional, and constraint terms. With the exception of the van der Waals term, these functions are positive semidefinite, and can be annealed into the smoothed energy surface. By removing the van der Waals term from the first stage, smoothing is enhanced. Thus, while the converged structures from stage 1 may not have a low van der Waals energy, they are, in general, compact, self-avoiding, with few or no constraint violations, and have a relatively low hydrophobic energy. In all cases, there were a small number (<10) of structures which fit this criteria. This result is consistent with the observation that the native can be distinguished from decoys based on a relatively simple screening function (Huang *et al.*, 1996). Apparently since the van der Waals energy is primarily a local effect, compactness and self-avoidance are retained while the van der Waals term is optimized in the second stage.

When there is a significant contribution to the target function from long-range energy terms, there is a risk that the influence of the constraints will be weakened. If we consider only violations greater than 0.5 Å (the criterion used in the DGMD calculations), we find that the low-energy CEM predictions of 1ris and 1xnb have violations. These were all less than 1.0 Å, and were all upper bound violations: 1ris had two (0.6, 0.6) and 1xnb had seven (0.67, 0.66, 0.55, 0.52, 0.51, 0.52, 0.63). It is not clear whether these violations could be reduced or eliminated through further minimization, or whether they represent limitations in the model (i.e. rigid secondary structure and idealized peptide torsional angle; $\omega = 180^\circ$). This is a question which will be addressed in future work. For the present, the violations appear to be small compared with the gain in accuracy relative to DGMD alone.

The least accurate CEM predictions, 1npk and 4pti, are also the least constrained, and the third least accurate, 1xnb, is significantly larger than the rest of the proteins (183 residues). In all three of these cases, the secondary structure RMS (R_s) is more than 0.5 Å lower than the total RMS (R_T). This can partially be explained by the presence of long or unusual loop configurations: 4pti has two loops which are ten residues or longer, 1xnb has four, and 1npk is the only case in which a terminal coil was included (as explained below).

The greatest topological error occurs in the BPTI prediction where the orientation of the short N-terminal helix is incorrect (see Figure 7). If this helix is ignored, the secondary structure RMS drops from 5.7 to 2.3 Å. From Table 3 and Figure 3, we see that this conformation is closely followed in energy by a much less accurate group of clusters (C^α RMS = 8.50). In the misfolded structures, both helices are oriented incorrectly, although their centers of mass are in approximately the correct

location. In our previous work on BPTI using a similar potential without distance constraints or the van der Waals term, the lowest energy structure was misfolded. The constraints and/or changes in the potential thus lead to more accurate results but the misfolded region is apparently still very competitive.

In 1npk, the main error in the CEM prediction is in a C-terminal segment comprising a loop and a polyproline helix. The 1npk protein is a hexamer in solution, and one of the major contacts between the monomers is at the edge of the sheet where two pairs of leucine residues meet (Morera *et al.*, 1994). Consequently, these residues are exposed in the monomer. In the CEM structure, the C-terminal segment, which also contains two leucine residues, folds out of the native conformation and into one which brings the C-terminal leucine residues and several other hydrophobic residues into contact with those at the edge of the sheet (see Figure 6). Thus, in the prediction of 1npk, the methodology makes a reasonable decision in protecting the exposed hydrophobic residues which are buried in the hexamer.

Although it is less certain in the case of BPTI, the accuracy of the CEM results strongly suggest that in these seven cases, the global minimum of the target function is within 2-6 Å of the native region. When low-RMS structures consistent with our model are generated by minimizing the RMS directly, they are generally not low in energy, due to local barriers in the energy function. We have carried out local minimizations on low-RMS structures and have found that the energies are not generally competitive with those from the CEM runs. For this reason, the best estimate of the global minimum, and thus the accuracy of the target function, comes from CEM.

Conclusion

The results presented above represent significant progress in the utilization of folding algorithms and potentials to assist in practical NMR structure determination. Firstly, the employment of NMR distance constraints derived from a realizable NMR protocol is considerably more meaningful than that of constraint sets which have no experimental correlate. Secondly, the comparison with the X-PLOR results is critical: if the constraint set is such that the standard DGMD protocol yields structures comparable with the methodology being tested, the latter cannot be said to provide any improvement. Moreover, the DGMD calculations are clearly sensitive to both the number and type of constraints: one cannot simply argue that a "small" number of constraints precludes obtaining reasonable results from this approach. The qualitative improvement we obtain in every case as compared with DGMD, even when DGMD does produce a reasonable structure, is convincing evidence that we are not looking at trivial cases

where the constraints, secondary structure, and packing forces are sufficient to specify the topology with reasonable accuracy. There can be no question that, in the examples that we have provided, our residue-based potential is playing a major role in discriminating among a substantial range of structures consistent with the distance constraints. Thirdly, both the minimization algorithm and the potential function display a high degree of reliability, providing accurate structures in every

case that we have tested, despite the complex topologies present in many of the examples. The success for the larger systems is also significant, and there does not appear to be any barrier to extending the methodology to the 200-300 residue regime (although this needs to be tested). The power of the overall global optimization scheme is suggested by the hybrid DGMD-CEM results: it is easy to become trapped in high-lying minima, but the combined protocol that we have developed appears to avoid this for the cases examined here.

Figure 4. (legend on page 1707)

Figure 5. (legend on page 1707)

However, more tests need to be carried out, and head-to-head benchmarking against alternative approaches, such as the Monte Carlo methodology used by Skolnick *et al.* (1997), is essential to objectively assess the cost/performance characteristics of each methodology.

Based on the results presented here, the only requirement for this methodology to work is that there be enough distance constraints to guarantee a global minimum near the native region. At this

point we do not have a good sense of what "enough" means. Presumably, after looking at more cases, we will be able to set limits on the RMS of the lowest-energy structure given the number of constraints and size of the protein.

We have chosen to focus on mixed α/β proteins because long-range constraints would be observable in NMR experiments with predeuterated samples. Other experiments which would generate constraints for all alpha, or membrane-bound proteins could also be envisioned. The accuracy of the

Figure 6. (legend on page 1707)

Figure 7. (legend on page 1707)

predicted structures is influenced by the number of constraints which, in turn, reflects the limitations in the underlying model and potential function. The next step in refining these structures would be to add side-chains and optimize the full protein using a more detailed potential energy function. Methodology for carrying out such refinement has been discussed (Johnson *et al.*, 1992).

The αBB algorithm, on which our CEM algorithm was based, uses local minimization in an intelligent way to search a very large confor-

mational space. Based on our experience, αBB is not, by itself, efficient at overcoming rugged local minima and the inclusion of annealing is a significant improvement. Moreover, it appears that for difficult cases, such as 1npk, generating a set of compact, self-avoiding structures as a first step is crucial in locating what appears to be the global minimum.

It is not clear how the accuracy of the CEM method discussed in this work will be affected when real data sets are used. The precision and

Figure 8. (legend on page 1707)

Figure 9. (legend on page 1707)

Figure 10.

Figures 4 to 10. Comparison between native (top), lowest-energy CEM (center), and lowest-energy DGMD conformations for the seven β -containing structures considered in the present work. The CEM and DGMD structures are shown in their optimal rotation relative to the native. Red lines, indicating $H_{amide} - H_{amide}$ distance constraints, are drawn between C^α atoms for simplicity. The radius of gyration (R_G) is included for all conformations. The RMS deviation from the native of all C^α atoms (R_T) and of C^α atoms located in secondary structures (R_S) is shown for the CEM and DGMD structures. In the case of 4pti, the correct disulphide bonds are shown in green in the CEM structure. In 1npg, the hydrophobic residues Leu39, Val40, Leu42, Lys43, and

accuracy of the predictions could be compromised if some of the expected NOEs are not observable or if some secondary structures are not well defined. Since selective deuteration schemes are designed primarily for NMR studies of large proteins (e.g. >25 kDa), it will be especially important to determine how the performance of the minimizer scales with the size of the problem. Significant progress has been made in speeding up the energy calculations since these results were generated. A typical two-stage CEM run which would have taken two to eight days on a Pentium Pro processor in the past, can be done in about a third of this time now. A parallel version of the program is also in place, which decreases the time to convergence approximately linearly with the number of processors. Using these improvements, we are currently focusing on applying this CEM method to real data sets and to minimizations in the absence of distance constraints.

Acknowledgments

R.A.F. acknowledges support from NIH grant GM52018. A.E.M. acknowledges support from the Kanagawa Academy of Science and Technology and NIH grant GM50291.

References

- Abola, E. E., Bernstein, F. C., Bryant, S. H., Koetzle, T. F. & Weng, J. (1987). Protein Data Bank. In *Crystallographic Databases-Information Content, Software Systems, Scientific Applications*, pp. 107-132, Data Commission of the International Union of Crystallography, Bonn, Cambridge, Chester.
- Adjiman, C. S. & Floudas, C. A. (1996). Rigorous convex underestimators for general twice differentiable problems. *J. Global Optimization* **9**, 23-40.
- Adjiman, C. S., Dallwig, S., Floudas, C. A. & Neumaier, A. (1998a). A global optimization method, αBB , for general twice differentiable NLPs: I. Theoretical advances. *Comput. Chem. Eng.* **22**, 1137-1158.
- Adjiman, C. S., Androulakis, I. P. & Floudas, C. A. (1998b). A global optimization method, αBB , for general twice differentiable NLPs: II. Implementation and Computational Results. *Comput. Chem. Eng.* **22**, 1159-1179.
- Allen, M. P. & Tildesley, D. J. (1987). *Computer Simulation of Liquids*. Clarendon Press, Oxford.
- Androulakis, I. P., Maranas, C. D. & Floudas, C. A. (1995). αBB : a global optimization method for general constrained nonconvex problems. *J. Global Opt.* **7**, 337-363.
- Bax, A. & Grzesiek, S. (1993). Methodological advances in protein NMR. *Acc. Chem. Res.* **26**(4), 131-138.

Leu45 in β -strand 2, and Leu143, Leu144, Val147, and Lys148 in the C-terminal segment are shown in blue. The Figures were produced using Molscrip (Kraulis, 1991).

- Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535-542.
- Braun, W. & Gö, N. (1985). Calculation of protein conformations by protein-proton distance constraints, a new efficient algorithm. *J. Mol. Biol.* **186**, 611-626.
- Brünger, A. T. (1992). *X-PLOR, Version 3.1. A system for X-ray Crystallography and NMR*, Yale University Press, New Haven, CT.
- Campbell, R. L., Rose, D. R., Wakarchuk, W. W., To, R. J., Sung, W. & Yaguchi, M. (1994). High-resolution structures of xylanases from *B. circulans* and *T. harzianum* identify a new folding pattern and implications for the atomic basis of the catalysis.
- Casari, G. & Sippl, M. (1992). Structure-derived hydrophobic potential hydrophobic potential derived from X-ray structures of globular proteins is able to identify native folds. *J. Mol. Biol.* **224**, 725-732.
- Clore, G. M. & Gronenborn, A. M. (1991). Structures of larger proteins in solution: Three- and four-dimensional heteronuclear NMR spectroscopy. *Science*, **252**, 1390-1399.
- Fedorov, A. A., Magnus, K. A., Graupe, M. H., Lattman, E. E., Pollard, T. D. & Almo, S. C. (1994). X-ray structures of isoforms of the actin-binding protein profilin that differ in their affinity for phosphatidylinositol phosphates. *Proc. Natl Acad. Sci. USA*, **91**(18), 8636-8640.
- Gardener, K. H., Rosen, M. K. & Kay, L. E. (1997). Global folds of highly deuterated, methyl-protonated proteins by multidimensional NMR. *Biochemistry*, **36**, 1389-1401.
- Grzesiek, S., Wingfield, P., Stahl, S., Kaufman, J. D. & Bax, A. (1995). Four-dimensional N^{15} -separated NOESY of slowly tumbling predeuterated N^{15} -enriched proteins, application to HIV-1 Nef. *J. Am. Chem. Soc.* **117**(37), 9594-9595.
- Gunn, J., Monge, A., Friesner, R. & Marshall, C. (1994). Hierarchical algorithm for computer modeling of protein tertiary structure: folding of myoglobin to 6.2 Å resolution. *J. Phys. Chem.* **98**, 702-711.
- 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 internuclear $H^1 - H^1$ proximities in solution. *Bull. Math. Biol.* **46**, 673-698.
- Huang, E. S., Subbiah, S., Tsai, J. & Levitt, M. (1996). Using a hydrophobic contact potential to evaluate native and near-native folds generated by molecular dynamics simulations. *J. Mol. Biol.* **257**, 716-725.
- Johnson, L., Monge, A. & Friesner, R. A. (1992). A hierarchical algorithm for polymer simulations. *J. Chem. Phys.* **97**, 9355-9365.
- Kitchen, D. B., Hirata, F., Westbrook, J. D., Levy, R. M., Kofke, D. & Yarmush, M. (1990). Conserving energy during molecular dynamics simulations of water, proteins, and proteins in water. *J. Comp. Chem.* **11**, 1169-1180.
- Kraulis, P. J. (1991). Molscript: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallog.* **24**, 946-950.
- Kuszewski, J., Nilges, M. & Brünger, A. T. (1992). Sampling and efficiency of metric matrix distance geometry: a novel "partial" metrization algorithm. *J. Biomol. NMR*, **2**, 33-56.
- Leijonmarck, M. & Liljas, A. (1987). Structure of the C-terminal domain of the ribosomal protein L/L12 from *Escherichia coli* at 1.7 Å. *J. Mol. Biol.* **195**, 555-579.
- Levy, R. M., Bassolino, D. A., Kitchen, D. B. & Pardi, A. (1989). Solution structures of proteins from NMR data and modeling: alternative folds for neutrophil peptide 5. *Biochemistry*, **28**, 9361-9372.
- Lindahl, M., Svensson, L. A., Liljas, A., Sedelnikova, S. E., Eliseikina, I. A., Fomenkova, N. P., Nevskaya, N., Nikonov, S. V., Garber, M. B., Muranova, T. A. & Rykonova, A. I. (1994). Crystal structure of the ribosomal protein s6 from thermus thermophilus. *EMBO J.* **13**, 1249-1254.
- Maranas, C. D. & Floudas, C. A. (1994a). A deterministic global optimization method for molecular structure determination. *J. Chem. Phys.* **100**, 1247-1261.
- Maranas, C. D. & Floudas, C. A. (1994b). Global minimum potential energy conformations of small molecules. *J. Global Optimization* **4**, 135-170.
- Monge, A., Lathrop, E., Gunn, J., Shenkin, P. & Friesner, R. (1995). Computer modeling of protein folding: conformational and energetic analysis of reduced and detailed protein models. *J. Mol. Biol.* **247**, 995-1012.
- Morera, S., Lebras, G., Lascu, I., Lacombe, M. L., Veron, M. & Janin, J. (1994). Refined X-ray structure of dictyostelium nucleoside diphosphate kinase at 1.8 Å resolution. *J. Mol. Biol.* **243**, 873-890.
- Nilges, M., Clore, G. M. & Gronenborn, A. M. (1988). Determination of three-dimensional structures of proteins from interproton distance data by, hybrid distance geometry-dynamical simulated annealing calculations. *FEBS Letters*, **229**, 317-324.
- Nilges, M., Kuszewski, J. & Brünger, A. T. (1991). *Computational Aspects of the Study of Biological Macromolecules by NMR*, Plenum Press, New York, NY.
- Ramage, R., Green, J., Muir, T. W., Ogunjobi, O. M., Love, S. & Shaw, K. (1994). Synthetic structural and biological studies of the ubiquitin system. Part 1. *Biochem. J.* **299**(24), 151-158.
- Romesburg, H. (1984). *Cluster Analysis for Researchers*, Lifetime Learning Publications, Belmont, CA.
- Skolnick, J., Kolinski, A. & Ortiz, A. (1997). MONSTER: a method for folding globular proteins with a small number of distance restraints. *J. Mol. Biol.* **265**, 217-241.
- Smith-Brown, M. J., Komminos, D. & Levy, R. (1993). Global folding of proteins using a limited number of distance constraints. *Protein Eng.* **6**, 605-614.
- Standley, D. M., Gunn, J. R., Friesner, R. A. & McDermott, A. E. (1998). Tertiary structure prediction of mixed α/β proteins via energy minimization. *Proteins: Struct. Funct. Genet.* **33**, 240-252.
- Venters, R. A., Metzler, W. J., Spicer, L. D., Mueller, L. & Farmer, B. T. (1995). Use of $H_N^1 - H_N^1$ NOEs to determine protein global folds in predeuterated proteins. *J. Am. Chem. Soc.* **117**(37), 9592-9593.
- Weiner, S. J., Kollman, P. A., Case, D. A., Singh, U. C., Ghio, C., Alagona, G., Profeta, S., Jr & Weiner, P. (1984). A new force field for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc.* **106**, 765-784.
- Weiner, S. J., Kollman, P. A., Nguyen, D. T. & Case, D. A. (1986). An all atom force field for simulations of proteins and nucleic acids. *J. Comp. Chem.* **7**, 230-252.
- Wlodawer, A., Deisenhofer, J. & Huber, R. (1987). Comparison of two highly refined structures of bovine pancreatic trypsin inhibitor. *J. Mol. Biol.* **193**, 145-156.

Distance Constraints

to the residue numbers of the original PDB files. In the CEM runs, constraints within helices or strands or in loops preceding (following) the first (last) helix or strand were ignored.

Res. i	Res. j	d_{ij}	Res. i	Res. j	d_{ij}	Res. i	Res. j	d_{ij}	Res. i	Res. j	d_{ij}
4 Y	8 N	4.840	5 V	9 L	4.979	7 T	11 G	4.433	16 T	97 V	4.442
16 T	98 Y	4.826	16 T	99 N	4.363	17 Q	97 V	3.424	18 A	31 S	3.246
18 A	97 V	4.914	19 A	95 V	3 040	19 A	97 V	4.924	20 I	28 W	3.565
20 I	29 A	2.919	21 L	28 W	4.630	21 L	93 I	2.871	21 L	94 L	4.749
21 L	95 V	4.590	22 G	26 N	3.516	22 G	27 T	4.984	22 G	28 W	4.370
22 G	93 I	4.558	37 T	40 G	4.199	37 T	61 A	3.669	57 G	67 T	3.971
58 F	65 Y	3.133	58 F	66 V	4.983	58 F	67 T	4.647	59 D	65 Y	4.928
60 L	65 Y	4.311	66 V	78 K	2.585	66 V	79 K	4.952	68 L	76 Y	4.142
68 L	77 G	4.572	68 L	78 K	4.189	69 R	76 Y	2.856	70 A	76 Y	4.839
71 D	75 I	4.913	74 S	88 K	4.834	75 I	86 T	3.394	75 I	88 K	4.667
76 Y	86 T	4.896	77 G	84 V	2.877	77 G	85 I	4.969	77 G	86 T	4.628
78 K	84 V	4.725	83 G	98 Y	3.025	84 V	98 Y	4.721	85 I	96 G	3.045
85 I	97 V	4.798	85 I	98 Y	4.825	86 T	96 G	4.961	87 V	94 L	3.037
88 K	94 L	4.888	89 T	93 I	4.787	89 T	94 L	4.200	103 Q	106 A	4.973

[illegible]

Res. i	Res. j	d_{ij}	Res. i	Res. j	d_{ij}	Res. i	Res. j	d_{ij}	Res. i	Res. j	d_{ij}
9 E	84 G	3.765	9 E	86 G	4.754	10 R	84G	4.936	11 T	82 F	2.942
12 F	124 S	3.210	13 L	80 M	3.098	13 L	81 V	4.948	13 L	82 F	4.464
14 A	80 M	4.838	14 A	122 H	2.624	14 A	123 G	4.970	15 V	78 V	3.130
15 V	79 A	4.884	15 V	80 M	4.792	15 V	122 H	4.824	16 K	20 V	4.927
16 K	121 I	4.320	16 K	122 H	3.817	19 G	120 I	4.555	38 V	83 E	2.861
38 V	84 G	4.975	39 L	83 E	4.868	40 V	81 V	3.768	40 V	82 F	4.831
40 V	83 E	4.177	41 G	81 V	2.744	41 G	144 L	4.983	42 L	81 V	4.768
43 K	79 A	3.425	45 L	77 V	2.666	45 L	78 V	4.972	45 L	79 A	4.074
46 V	77 V	4.671	48 T	52 A	4.272	95 I	108 I	4.666	96 G	108 I	4.880
96 G	109 R	4.993	97 V	103 S	4.381	97 V	104 A	4.458	106 G	111 D	4.276
107 S	111 D	4.146	125 D	130 A	4.952	126 S	130 A	3.539			

[illegible]

Table A5. 1ris distance constraints

Res. <i>i</i>	Res. <i>j</i>	<i>d_{ij}</i>	Res. <i>i</i>	Res. <i>j</i>	<i>d_{ij}</i>	Res. <i>i</i>	Res. <i>j</i>	<i>d_{ij}</i>	Res. <i>i</i>	Res. <i>j</i>	<i>d_{ij}</i>
2 R	67 M	2.787	2 R	69 E	3.902	3 R	93S	4.677	4 Y	65 V	3.122
4 Y	66 E	4.900	4 Y	67 M	4.408	4 Y	93 S	4.457	5 E	91 V	2.800
5 E	92 K	4.935	5 E	93 S	4.105	6 V	63 Y	2.756	6 V	64 Q	4.950
6 V	65 V	4.384	7 N	89 M	2.809	7 N	90 V	4.866	7 N	91 V	4.784
8 I	61 L	3.138	8 I	62 W	4.997	9 V	86 R	4.096	9 V	87 R	2.680
9 V	88 V	4.991	10 L	59 Y	2.837	10 L	60 F	4.885	10 L	61 L	4.673
10 L	86 R	4.553	11 N	59-Y	4.905	11 N	86 R	3.856	15 D	19 L	4.430
31 E	35 A	4.861	36 R	66 E	2.682	36 R	67 M	4.906	38 E	64 Q	4.095
38 E	65 V	4.649	38 E	66 E	4.399	39 K	64 Q	2.819	40 V	64 Q	4.806
41 E	62 W	2.766	41 E	63 Y	4.918	43 L	60 F	4.600	43 L	61 L	4.647
43 L	62 W	4.036	44 G	60 F	3.018	45 L	60 F	4.700	46 R	58 G	2.592
46 R	60 F	4.619	48 L	57 Q	4.414	48 L	58 G	4.209	52 1	57 Q	4.938

Table A6. 1ubi distance constraints

Res. <i>i</i>	Res. <i>j</i>	<i>d_{ij}</i>	Res. <i>i</i>	Res. <i>j</i>	<i>d_{ij}</i>	Res. <i>i</i>	Res. <i>j</i>	<i>d_{ij}</i>	Res. <i>i</i>	Res. <i>j</i>	<i>d_{ij}</i>
2 Q	17 V	U.8 5-2	2 Q	64 E	3.921-	3 I	15 L	2.819	3 1	17 V	4.211
3 I	64 E	4.964	4 F	65 S	4.275	4 F	66 T	4.781	4 F	67 L	3.800
5 V	13 I	3.004	5 V	14 T	4.836	5 V	15 L	4.655	5 V	67 L	4.695
6 K	13 I	4.791	6 K	67 L	3.602	6 K	68 H	4.908	6 K	69 L	3.862
7 T	11 K	3.653	7 T	12 T	4.953	7 T	13 I	4.384	7 T	69 L	4.805
21 D	56 L	3.907	21 D	57S	4.878	22 T	26 V	4.748	23 I	54 R	3.978
23 I	55 T	4.988	41 Q	72 R	4.830	42 R	70 V	3.121	42 R	72 R	4.668
43 L	50 L	4.0	4 L	70 V	4.981	44 I	50 L	4.883	44 I	68 H	2.759
44 I	69 L	4.909	44 I	70 V	4.580	45 F	50 L	4.269	45 F	68 H	4.905

Table A7. 1xnb distance constraints

Res. <i>i</i>	Res. <i>j</i>	<i>d_{ij}</i>	Res. <i>i</i>	Res. <i>j</i>	<i>d_{ij}</i>	Res. <i>i</i>	Res. <i>j</i>	<i>d_{ij}</i>	Res. <i>i</i>	Res. <i>j</i>	<i>d_{ij}</i>
4 D	41 G	4.195	4 D	42 W	4.828	4 D	43 T	4.050	5 Y	41 G	3.013
5 Y	42 W	4.918	5 Y	43 T	4.511	7 Q	39 G	3.083	7 Q	40 K	4.928
7 Q	41 G	4.881	8 N	39 G	4.985	9 W	37 V	3.166	9 W	38 V	4.980
10 T	37 V	4.952	11 D	34 G	4.929	11 D	35 N	3.555	13 G	34 G	3.661
13 G	35 N	4.365	14 G	34 G	4.926	15 I	31 S	3.099	15 I	33 T	4.843
16 V	31 S	4.993	17 N	29 N	3.009	17 N	30 W	4.882	18 A	29 N	4.892
19 V	27 S	2.854	19 V.	28 V	4.877	21 G	26 Y	4.621	21 G	27 S	4.047
24 G	184 V	4.560	25 N	184 V	4.523	26 Y	182 V	3.074	28 V	180 S	2.809
30 W	178 G	2.871	30 W	179 S	4.952	31 S	178 G	4.968	32 N	176S	4.774
32 N	177S	4.591	33 T	175 Q	4.614	33 T	176 S	2.723	34 G	175 Q	4.371
35 N	175 Q	4.814	36 F	173 G	2.794	38 V	171 T	2.916	39 G	171 T	4.963
40 K	169 M	3.257	41 G	169 M	4.976	42 W	167 Q	3.066	42 W	168 V	4.903
45 G	165 A	4.119	45 G	166 Y	4.802	45 G	167 Q	4.469	51 I	144 I	3.286
51 I	146 F	3.460	52 N	144 I	4.966	53 Y	142 A	2.998	54 N	142 A	4.856
54 N	183 T	2.682	55 A	139 G	2.942	55 A	140 S	4.222	55 A	183 T	4.856
56 G	181 N	4.507	56 G	182 V	4.909	56 G	183 T	4.033	57. V	181 N	2.786
59 A	179S	3.428	59 A	180 S	4.915	61 N	177S	3.617	61 N	178 G	4.549
61 N	179 S	4.537	62 G	87 T	3.688	62 G	177S	4.901	63 N	87 T	4.340
63 N	174 Y	4.746	63 N	177 S	4.823	64 G	86 G	2.595	64 G	87 T	3.605
64 G	88 Y	3.935	65 Y	172 E	2.727	65 Y	173 G	4.911	66 L	83 D	4.275
66 L	84S	3.130	66 L	85 W	4.786	67 T	83 D	4.909	67 T	170 A	3.404
67 T	172 E	4.560	68 L	81 V	2.753	68 L	82 V	4.826	68 L	83 D	4.362
68 L	170 A	4.765	69 Y	168 V	2.979	69 Y	169 M	4.965	69 Y	170 A	3.842
70 G	79 Y	2.968	70 G	168 V	4.964	71 W	79 Y	4.886	71 W	165 A	4.003
71 W	166 Y	4.328	72 T	77 I	2.623	72 T	165 A	4.748	73 R	77 I	4.956
73 R	161 G	4.992	73 R	162 S	4.813	73 R	163 N	3.075	73 R	164 W	4.871
73 R	165 A	4.457	78 E	126 T	4.711	78 E	127 Q	4.971	78 E	128 Y	4.061
79 Y	128 Y	4.725	80 Y	128 Y	4.022	80 Y	129 W	4.858	80 Y	130 S	4.124
81 V	130 S	4.738	82 V	130 S	3.390	82 V	131 V	4.802	82 V	132 R	4.109
84 S	136 R	4.799	93 T	109 T	2.653	95 K	107 I	3.780	95 K	108 Y	4.913
96 G	107 I	2.933	97 T	107 I	4.825	98 V	105 Y	2.762	99 K	105 Y	4.835
100 S	105 Y	4.450	101 D	145 T	4.887	104 T	133 Q	3.856	104 T	134S	4.833
105 Y	133 Q	4.917	106 D	131 V	2.960	106 D	132 R	4.839	106 D	133 Q	4.175
107 I	131 V	4.990	108 Y	129 W	3.122	108 Y	130S	4.885	108 Y	131 V	4.653
110 T	127 Q	2.898	111 T	127 Q	4.825	112 R	125 F	2.953	112 R	126 T	4.954
112 R	127 Q	4.620	114 N	123 T	4.414	114 N	124 T	4.141	114 N	125 F	4.908
115 A	123 T	2.603	115 A	124 T	4.989	117S	123 T	4.568	154 K	158 M	4.692