Protein Fold Determination from Sparse Distance Restraints: The Restrained Generic Protein Direct Monte Carlo Method

Derek A. Debe,[†] Matt J. Carlson,[†] Jiro Sadanobu,[‡] S. I. Chan,[§] and W. A. Goddard III*,[†]

Materials and Process Simulation Center (MSC), Beckman Institute (139-74), California Institute of Technology, Pasadena, California 91125, Polymer and Materials Research Laboratories, Teijin Limited, 2-1 Hinode-cho, Iwakuni-shi, Yamaguchi 740, Japan, and Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, California 91125

Received: August 20, 1998; In Final Form: November 9, 1998

We present the *generate-and-select* hierarchy for tertiary protein structure prediction. The foundation of this hierarchy is the Restrained Generic Protein (RGP) Direct Monte Carlo method. The RGP method is a highly efficient off-lattice residue buildup procedure that can quickly generate the complete set of topologies that satisfy a very small number of interresidue distance restraints. For three restraints uniformly distributed in a 72-residue protein, we demonstrate that the size of this set is $\sim 10^4$. The RGP method can generate this set of structures in less than 1 h using a Silicon Graphics R10000 single processor workstation. Following structure generation, a simple criterion that measures the burial of hydrophobic and hydrophilic residues can reliably select a reduced set of $\sim 10^2$ structures that contains the native topology. A minimization of the structures in the reduced set typically ranks the native topology in the five lowest energy folds. Thus, using this hierarchical approach, we suggest that de novo prediction of moderate resolution globular protein structure can be achieved in just a few hours on a single processor workstation.

1. Introduction

Given the difficulty of protein structure prediction, it is important to simplify the problem using prediction approaches that incorporate predicted or experimentally determined structural information. For many prediction targets, distance restraints are available from labeling experiments, disulfide bond connectivity, or preliminary NMR data. Furthermore, methods exist for predicting local structural characteristics[1] such as residue contacts,[2,3] secondary structure,[4,5] accessible surface area,[6] and surface turns.[7]

Dewitte et al.[8] established the basic feasibility of obtaining fold predictions using a limited amount of distance information. They developed a method to exhaustively enumerate all walks on a diamond lattice consistent with a set of lattice pair restraint conditions. Their work demonstrated that as few as one restraint per residue could successfully limit the number of possible walks (conformations) to $\sim 10^3$. Unfortunately, the method was not computationally feasible when the number of restraints was small compared to the number of lattice steps (residues).

Since this original work, several different methods have been applied to the problem of structure prediction using a small number of distance restraints. Aszódi et al.[9] developed a distance-geometry-based approach that incorporated distance restraints and native secondary structure assignments as well as knowledge-based criteria such as backbone connectivity, hydrophobicity, and conservation data obtained from multiple alignments. This method efficiently generated structures with the correct topology using as few as $N/10$ restraints for very simple protein topologies, and $\sim N/4$ restraints for more complex folds.

Another approach is a dynamic Monte Carlo (MC) method, MONSSTER,[10] that folds random coil conformations using an energy function incorporating secondary structure and distance restraint information. Recently, this method achieved low-resolution structures (typically 5−6 Å CRMS) for a number of small proteins when used in conjunction with secondary structure and residue contact predictions.[11] However, since the algorithm is a dynamic procedure, generating a single structure that satisfies the restraints requires overnight simulation.

The results obtained by distance-geometry and dynamic MC suggest that knowing the correct secondary structure and $\sim N/4$ distance restraints leaves a very small number of possibilities for the topology of a polypeptide. In both methods, coupling this distance information with simple energy criteria usually resulted in an unambiguous determination of the native fold topology. Thus, each method capably finds the correct overall fold when the amount of distance knowledge specifies the correct topology with little ambiguity.

In this paper, we present a novel method that is useful in instances when very limited (sparse) structural information is available and the topology of the protein is far from uniquely specified. The method efficiently generates the complete set of topologies consistent with a set of interresidue restraints, even when the number of restraints is very small. We will show that as few as $N/24$ interresidue restraints reduce the number of topologies sufficiently so that a simple residue burial score can identify the native topology in a very small set of candidates (typically <5). We expect that improved contact prediction approaches will be capable of obtaining reliable sparse restraint information (at a level of $\sim N/12-N/24$) for a wide array of protein prediction targets. Furthermore, for many protein sequences, knowledge of simple biochemical information such as disulfide bond connectivity provides enough information to successfully apply our prediction hierarchy. With this hierarchi-

* Corresponding author.
[†] Materials and Process Simulation Center (MSC), Beckman Institute, California Institute of Technology.
[‡] Polymer and Materials Research Laboratories, Teijin Limited.
[§] Division of Chemistry and Chemical Engineering, California Institute of Technology.

10.1021/jp983429+ CCC: $18.00
Published on Web 04/01/1999

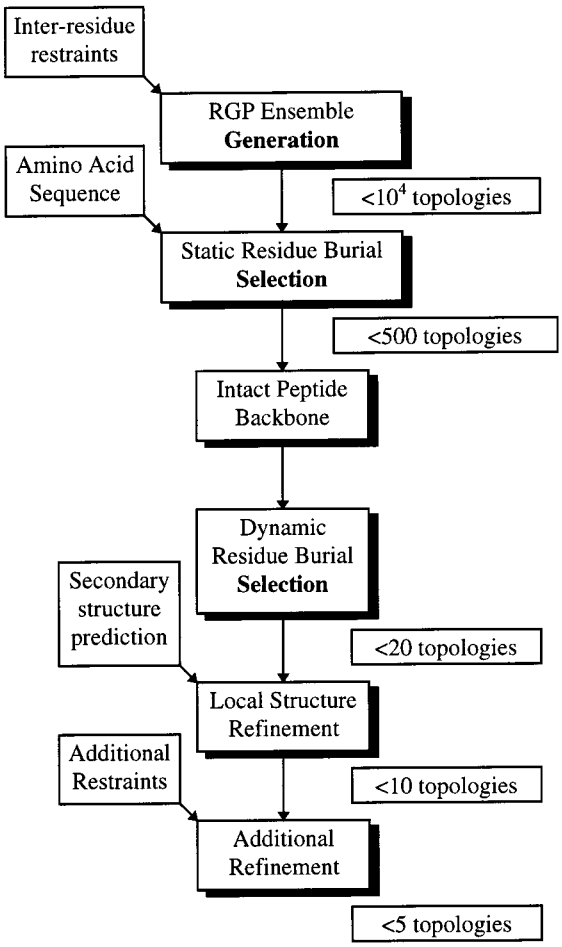

Figure 1. Flowchart diagram of the *generate-and-select* hierarchical method for predicting moderate resolution tertiary protein structure from sparse distance restraints.

cal approach, moderate resolution globular protein structure can be determined from sparse distance information in just a few hours on a single processor workstation.

2. Methods

The Restrained Generic Protein (RGP) Direct MC method is an off-lattice residue buildup procedure for generating all polypeptide topologies that are consistent with a set of inter-residue distance restraints. The RGP method is the first step in the *generate-and-select* hierarchical structure prediction procedure shown in Figure 1. In the second step of the hierarchy, a static residue burial (S-RB) scoring function is used to select a small set of candidates from the RGP ensemble. In the third hierarchical step, an intact peptide backbone representation is constructed for each fold in the selected set (the RGP method produces an α-carbon trace of each conformation). Following the construction of the intact peptide backbone, each of the selected conformations is minimized with respect to the residue burial function used in step 2. This *dynamic* residue burial (D-RB) selection process further reduces the set of remaining fold candidates. The final stage of the prediction hierarchy uses predicted secondary structure information or additional distance restraints to further reduce and refine the surviving set from the previous step.

A. Protein Representation. The RGP method employs a "ball-and-stick" protein model.[12] Each residue is connected to its neighboring residues by a fixed bond length, $l = 3.8$ Å, with fixed bond angle, $\theta = 120°$. Thus, the coordinates of residue i are precisely determined from the coordinates of residues $i - 1$, $i - 2$, and $i - 3$, given a single torsion, ϕ_i, about the central bond (Figure 2). The possible values of ϕ_i in an n-state per residue representation are $\phi_i = i \times (360°/n)$ for $i = 0,1,2,...n - 1$. Hence, for a 6-state per residue representation, $\phi_i = 0°$, 60°, 120°, 180°, 240°, or 300°.

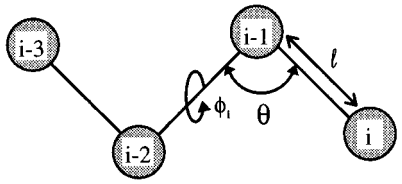

Figure 2. Ball-and-stick peptide representation used in the RGP-DMC method. Each residue is connected to its neighboring residues by a fixed bond length, $l = 3.8$ Å, with fixed bond angle, $\theta = 120°$. The possible values of ϕ_i in an n-state per residue representation are $\phi_i = i \times (360°/n)$ for $i = 0,1,2,...n - 1$.

B. Restraint Implementation. The RGP method is a residue buildup procedure. Residues are added one by one from the N- to C-terminus to construct a complete polypeptide. An efficient restraint technique ensures that the polypeptide conformations are consistent with a set of user-defined interresidue distance restraints. Consider using a buildup procedure to construct a polypeptide where residue j and residue k are less than 6 Å apart ($j < k$). The simplest approach is to randomly enumerate all possible conformations of residue j through residue k and discard the "dead end" conformations that do not satisfy this restraint.[8] Unfortunately, this approach becomes prohibitively expensive as the sequential distance between j and k increases, since the detection of a dead end occurs after the construction of residue k.

An algorithm that can determine if a conformation is a dead end prior to the addition of residue k yields a vast improvement in efficiency. The longest distance traversed by each residue addition step is a single bond length, $l = 3.8$ Å. Thus, it is impossible to place residue k within 6 Å of residue j if residue i ($j < i < k$) is greater than $6 + 3.8(k - i)$ angstroms from residue j. Thus, it is possible to predict at step i if a conformation must eventually result in a dead end at step k.

The restraint method incorporated into the RGP method is slightly more complex in that it also considers the angle between residues j, i, and $i - 1$. Figure 3 shows the possible positions for residue $i + 4$ in our peptide model when ϕ_{i+2}, ϕ_{i+3}, and ϕ_{i+4}, $= 0°$ or 180°. Consider a cylindrical coordinate system where the z-axis travels through the bond between residue $i - 1$ and residue i, and the z-axis origin is at residue $i - 1$. The radial axis, ρ, represents the perpendicular distance to the z-axis. In the figure, the solid line around the perimeter traces the maximum radial distance that residue $i + 4$ may be from the z-axis for a given value of z. Hence, this solid line represents the most extreme position in (z, ρ) space that residue ($i - 1 + 5$) may be placed from residues $i - 1$ and i in our polypeptide model. Similar diagrams lead to a general expression for the maximum value of ρ, ρ_{max}, for an arbitrary residue $(i - 1) + n$ at a specific z-coordinate. Defining

$$\rho_{peak} = (n - 1)(l \sin 60°) \tag{1}$$

we find that
(a) if n is even, then z must lie between

$$\{z_{min}, z_{max}\} = \{(-3l/4)(n - 4), (3l/4)(n)\} \tag{2}$$

and two cases define ρ_{max}:

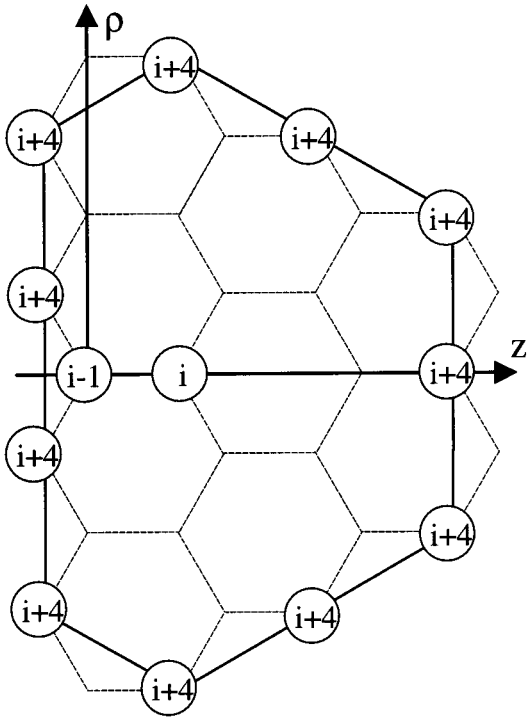

Figure 3. The allowed positions of residue $i + 4$ in relation to residue $i - 1$ and residue i when residues $i - 1$, i, $i + 1$, $i + 2$, $i + 3$, and $i + 4$ all lie in the same plane. For the cylindrical coordinate system (z, r), the maximum value of r for residue $i + 4$ may be expressed as a function of z. This is used to derive eqs 1−7.

(a.1) for $z \geq 3l/2$,

$$\rho_{max} = \rho_{peak} - (\tan 30°)(z - (3l/2)) \quad (3)$$

(a.2) for $z < 3l/2$,

$$\rho_{max} = \rho_{peak} + (\tan 30°)(z - (3l/2)) \quad (4)$$

(b) If n is odd, then z must lie in the range

$$\{z_{min}, z_{max}\} = \{(-l/4)(2 + 3(n - 5)), (l/4)(4 + 3(n - 1))\} \quad (5)$$

and two cases define ρ_{max}:
(b.1) for $z \geq l$,

$$\rho_{max} = \rho_{peak} - (\tan 30°)(z - l) \quad (6)$$

(b.2) for $z < l$,

$$\rho_{max} = \rho_{peak} + (\tan 30°)(z - l) \quad (7)$$

Thus, expressions 1−7 specify the greatest distance in (z, ρ) space that any residue ($i - 1 + n$) may be placed from residues $i - 1$ and i.

Now we return to the example of constructing a polypeptide conformation with restrained residues j and k. Assume that the restraint limits the distance between j and k to a maximum of 6.58 Å. This distance is equivalent to the maximum distance traveled in two residue addition steps, i.e., $2l(\sin \theta/2) = 6.58$ Å. Thus placing residue k within 6.58 Å of residue j is similar to requiring that residue j lies in allowed (z, ρ) space for residue $k + 2$. Thus if a candidate torsion ϕ_i places residue i ($j < i < k$) in a location such that residue j lies outside allowed (z, ρ) space for $n = k + 2 - (i - 1)$, the torsion will inevitably result in a dead end, and we can eliminate it.

In the above example, we assigned the distance restraint between residue j and residue k a *bond order* ($bo_{j,k}$), which represents the number of residue addition steps required to span the restraint distance. A single addition step of length l spans 3.8 Å; hence, $bo_{j,k} = 1$ represents this distance. Two residue addition steps span distances up to $2l(\sin \theta/2) = 6.58$ Å, hence for $d < 6.58$ Å, $bo_{j,k} = 2$.

The above discussion specifies how the RGP method satisfies a single interresidue restraint. A single restraint between two residues is called a *first-order* restraint. A first-order restraint occurs when residues j and k are restrained, and we seek to add residue i ($j < i < k$) such that

$$i - j \geq k - i + bo_{j,k} - 2 \quad (8)$$

If two restraints (j, k) and (p, q) are specified where $j < k < p < q$, then the restraint on (j, k) is satisfied before residue p is added. Thus, these restraints are separate. However, if $p < k$, then when adding new residues i, where $p \leq i \leq k$, we must simultaneously consider both restraints. We refer to this as a *second-order* restraint. Consider a polypeptide with a first-order restraint between residues ($bo_{5,39} = 2$) and a first-order restraint between residues 17 and 39 ($bo_{17,39} = 2$). Then there is effectively a second-order restraint between residues 5 and 17, with $bo_{5,17} = bo_{5,39} + bo_{17,39} = 4$. Thus, as we grow each residue i, such that $i - 5 \geq 17 - i + bo_{5,17} - 2$ (i.e., residues 12, 13, 14, 15, 16, and 17), residue 5 must lie in allowed (z, ρ) space for $n = 17 + bo_{5,17} - (i - 1)$. Thus, depending on the configuration of the interresidue restraints in the protein, there can be first- and second-order restraints that require attention at each growth step i.

C. Conformation Sampling Procedure. Now that we have described the restraint technique, it is possible to list the steps followed to construct a complete polypeptide by the RGP method.

The inputs required for the RGP method are the number of residues in the polypeptide (N), and a list of interresidue distance restraints with restraint bond orders, $bo_{j,k}$. The first- and second-order restraints for each residue addition step i are determined.

A three-residue starting fragment corresponding to the first three residues in the polypeptide sequence is constructed. Residues are added one at a time in one of $p = 6$ different torsional states to construct the complete N-residue polypeptide. For each residue addition step, q, the restraint conditions are evaluated. If the candidate torsion does not satisfy the restraints, the probability of selecting this torsion is zero. If a candidate torsion does satisfy the restraints, the probability of selecting this torsion is

$$P_q = \exp(-E_q/kT)/\sum_{i=1}^{p} \exp(-E_i/kT) \quad (9)$$

where p is the number of candidate torsions and E_q is the addition energy for a specific torsion candidate q, according to the CCB-DMC procedure.[13] The addition energy of a torsion candidate for residue i is given by the summation of the energy between residue i and each existing residue in the peptide fragment. For all residue types, the energy of a residue pair is taken as

$$E_{ij}(R) = E_0\left[\left(\frac{R}{R_0}\right)^{12} - 2\left(\frac{R}{R_0}\right)^{6}\right] \quad (10)$$

where $R_0 = 5.5$ Å, $E_0 = 0.15$ kcal/mol, R is the distance between the coordinates of each residue, and i and j are not nearest

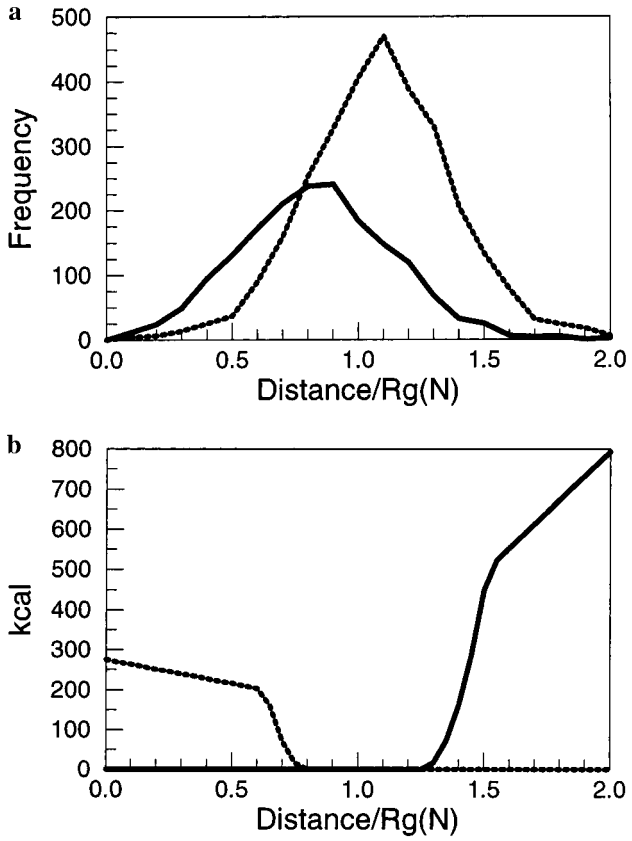

Figure 4. (a) Frequency histogram of the distance of hydrophobic (solid line) and hydrophilic (dotted line) residues to the center of mass for 61 nonhomologous, single-domain proteins (listed in Table S-1 of the Supporting Information). The distance is normalized by the factor $R_g(N)$, the expected minimum radius of gyration for a globular protein structure of N residues (see eq 11 in text). Figures S-1 through S-20 of the Supporting Information show the histograms for all 20 amino acids. (b) Burial bias potentials used for D-RB minimization. The solid line is the hydrophobic burial bias potential, E_{b-phob}, and the dotted line is the hydrophilic burial bias potential, E_{b-phil} ($R_g = 12$ Å). Using these potentials, hydrophobic residues are drawn toward the protein interior, while hydrophilic residues are excluded from the protein core.

neighbors in the sequence. This sequence independent energy function accounts for the excluded volume of each residue.

At a given residue addition step i, if no candidate torsion satisfies the restraint conditions, the polypeptide is re-grown from residue $i - 4$ in an attempt to satisfy this restraint. The current implementation allows one such *backtrack* before discarding the entire polypeptide and growing a new polypeptide from the starting fragment.

A *look-ahead* strategy may also be performed, where the placement of residue $i + 1$ determines the probability of selecting the torsion angle for residue i. That is, for a particular torsion candidate for residue i, if there is no torsion candidate ϕ_{i+1} that satisfies the restraints on residue $i + 1$, the probability of selecting that particular torsion candidate for residue i is zero.

D. Static Residue Burial (S-RB) Score. The RGP method generates the α-carbon coordinates for distance-restrained protein conformations without considering the identity of the amino acid sequence. To assign an energy to each of the RGP conformations, we developed a very simple, static residue burial (S-RB) score based on the observation that the α-carbon positions for the hydrophobic residues (Cys, Ile, Leu, Phe, and Val) lie closer to the protein center of mass than the hydrophilic residues (Arg, Asn, Asp, Gln, Glu, Lys, Pro, and Ser) (Figure 4a). Once the RGP method generates a complete polypeptide, the center of mass is calculated from the α-carbon coordinates. The distance from each hydrophilic and hydrophobic residue to the center of mass is calculated and expressed as a factor of

$$R_g(N) = -1.26 + 2.79N^{1/3} \tag{11}$$

where R_g represents the expected minimum radius of gyration for a globular protein of N residues.[14] Each hydrophobic and hydrophilic residue receives a residue burial score, W, that depends on its distance from the center of mass, $|R - R_{cm}|$.

For hydrophobic residues we take

$$W = -1, \text{ if } |R - R_{cm}| \le D_{phob} \tag{12a}$$
$$W = 2, \text{ if } |R - R_{cm}| > D_{phob}$$

where,

$$D_{phob} = 1.2R_g \text{ for Phe and Ile residues} \tag{12b}$$
$$D_{phob} = 1.25R_g \text{ for Leu and Val residues}$$
$$D_{phob} = 1.3R_g \text{ for Cys residues}$$

For hydrophilic residues we take

$$W = 2, \text{ if } |R - R_{cm}| \le D_{phil} \tag{13a}$$
$$W = -1, \text{ if } |R - R_{cm}| > D_{phil}$$

where,

$$D_{phil} = 0.85R_g \text{ for Asp residues} \tag{13b}$$
$$D_{phil} = 0.8R_g \text{ for Asn, Gln, Glu, Lys, Pro, and Ser residues}$$
$$D_{phil} = 0.75R_g \text{ for Arg residues}$$

The S-RB score for the polypeptide is the sum of the individual residue burial scores,

$$\text{S-RB} = \sum_{i=1}^{N} W_i \tag{14}$$

E. Intact Backbone Construction. The RGP method generates the α-carbon trace of a polypeptide. Thus, an intact peptide backbone must be constructed for each structure in the selected set. Since the RGP structures are very low-resolution folds, it is not critical that the backbone preserves the original trace exactly. To this end, we find that an algorithm developed by Park and Levitt[15] works well for quickly producing an intact backbone highly similar to the original RGP trace. A six-state per residue backbone representation[16] generates a full atom backbone from the α-carbon coordinates that is typically less than 3 Å CRMS from the original RGP conformation.

F. Dynamic Residue Burial (D-RB) Score. Once a peptide backbone has been constructed, each backbone is minimized with respect to the S-RB criteria using the simple burial bias potentials, E_{b-phob} and E_{b-phil} shown in Figure 4b. Letting $x = |R - R_{cm}|$, these have the form

$$E_{b-phob}(x) = 0 \text{ kcal for } x \le D_{phob} \quad (15)$$

$$E_{b-phob}(x) = 50(x - D_{phob})^2 \text{ kcal/Å}^2 \text{ for } D_{phob} < x \le D_{phob} + 3.16 \text{ Å}$$

$$E_{b-phob}(x) = (500 + 50(x - D_{phob})/\text{Å}) \text{ kcal for } x > D_{phob} + 3.16 \text{ Å}$$

$$E_{b-phil}(x) = (200 + 10(x - D_{phob})/\text{Å}) \text{ kcal for } x < D_{phil} - 2 \text{ Å} \quad (16)$$

$$E_{b-phil}(x) = 50(x - D_{phob})^2 \text{ kcal/Å}^2 \text{ for } D_{phil} - 2 \text{ Å} \le x \le D_{phil}$$

$$E_{b-phil}(x) = 0 \text{ kcal for } x > D_{phil}$$

Distance restraint bias potentials are added to preserve the original interresidue distance restraints between residue j and residue k. Letting $x = |R_j - R_k|$, these have the form

$$E_{j,k}(x) = (200 + 10(4 \text{ Å} - x)/\text{Å}) \text{ kcal for } x < 2 \text{ Å} \quad (17)$$

$$E_{j,k}(x) = 50(4 \text{ Å} - x)^2 \text{ kcal/Å}^2 \text{ for } 2 \text{ Å} \le x < 4 \text{ Å}$$

$$E_{j,k}(x) = 0 \text{ kcal for } 4 \text{ Å} \le x \le 7 \text{ Å}$$

$$E_{j,k}(x) = 50(x - 7 \text{ Å})^2 \text{ kcal/Å}^2 \text{ for } 7 \text{ Å} < x \le 7 \text{ Å} + 3.16 \text{ Å}$$

$$E_{j,k}(x) = (500 + 50(x - 7 \text{ Å})/\text{Å}) \text{ kcal for } x > 7 \text{ Å} + 3.16 \text{ Å}$$

Once the hydrophobic, hydrophilic, and distance restraint potentials are in place, 500 steps of conjugate gradient minimization are performed, where standard force field terms represent the peptide backbone.[17] After minimization, the S-RB score is recalculated, yielding the D-RB score.

G. Additional Restraints. Additional interresidue distance restraints can be incorporated into a structure by minimization using the restraint bias potential described in eq 17 between the newly restrained residues. Local structure can be refined to incorporate a secondary structure prediction using a restrained minimization procedure that refines the α-carbon coordinates in the original structure to comply with an optimally superimposed secondary structural unit.

3. Computational Protocol and Efficiency

In this paper, we consider the ability of the RGP algorithm to generate low-resolution tertiary folds given sparse interresidue distance information. A list of appropriate interresidue distance restraints was selected for each native protein by selecting pairs of residues known to be between 4 and 7 Å away in the native coordinate file. RGP treated each restraint pair (j, k) with $bo_{j,k} = 2$, until the addition of residue k. At this addition step, the RGP algorithm required that residue k be placed anywhere from 3.8 to 7.4 Å from residue j, rather than calculating z and ρ_{max} according to eqs 1−7. Thus the difference in the distance between the restrained residues in the generated structure and the native structure could be as large as 3.4 Å.

Figure 5 shows the probability of satisfying a first-order restraint pair ($bo_{j,k} = 2$) separated by N residues using the RGP method. For a sequence separation of 50 residues, the RGP method generates conformations that satisfy the restraint with >60% efficiency using just six states per residue without a look-ahead step. Without a restraint coupling, the probability of generating a 50-residue segment with the terminal residues between 3.8 and 7.4 Å apart is less than 0.005. Thus, by identifying dead ends at each step in the buildup procedure, RGP leads to an efficiency increase by a factor of 120 over a random or exhaustive approach. Using a Silicon Graphics R10000 processor, the RGP-DMC method can generate 1000 50-residue polypeptides with restrained termini in less than 3 min (six-state per residue representation).

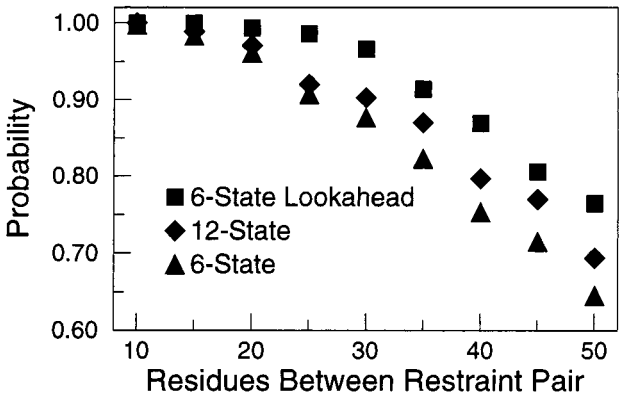

Figure 5. The probability of satisfying an interresidue restraint of bond order two plotted versus the sequence separation of the restrained residues using the RGP-DMC method. A six-state-per-residue representation coupled with a look-ahead step is denoted by (■) markers, a twelve-state-per-residue representation without a look-ahead step is denoted by (◆) markers, and a six-state-per-residue representation without a look-ahead step is denoted by (▲) markers.

4. Results

We will demonstrate how the RGP method can be used for making successful low-resolution tertiary structure predictions by applying the *generate-and-select* prediction hierarchy to two sequences with known tertiary structure. The first prediction target is the LexA repressor DNA binding domain from *Escherichia coli* (1lea),[18] a 72-residue protein with helical and β strand secondary structure. The second target is sea hare myoglobin (1mba),[19] a 146-residue protein with eight helices.

A. LexA Repressor DNA Binding Domain (72 AA). Table 1 shows the results obtained by applying the *generate-and-select* prediction hierarchy to the LexA repressor sequence using 2 (N/36), 3 (N/24), 6 (N/12), and 12 (N/6) interresidue distance restraints. For each restraint set (see Table 3), the RGP algorithm generated an ensemble of S structures (column 2 of Table 1) using a six-state per residue representation. The closest match to the experimental structure in this ensemble has a CRMS as given in column 3 of Table 1. From this original ensemble, a small subset of s structures (column 4) was selected according to the S-RB score. We then optimized the structure using D-RB minimization. Comparing these structures with the experimental structure, we found a near-native match with the rank given in column 5 and the CRMS given in column 6. We then took each structure in the selected set s and refined it to incorporate the results of a PHD[6] secondary structure prediction. We again applied D-RB and ranked the structures. We found a near-native match to the experimental structure with the rank given in column 7 and the CRMS given in column 8.

Table 1 shows that 3 (N/24) interresidue restraints combined with a secondary structure prediction is sufficient to identify a low-resolution native conformation (6.1 Å CRMS) in the top three of all conformations. Less than 30 min on a Silicon Graphics R10000 processor was required to generate and score the 5000 structure ensemble, and less than 2 h was required to refine and minimize the selected set, resulting in an overall prediction time of less than 3 h.

TABLE 1: Results of the *Generate-and-Select* Prediction Hierarchy for LexA Repressor (72 residues) Using Different Sets of Interresidue Restraints[a]

	RGP ensemble		selected set			sec. prediction	
	S[b]	CRMS[c]	s[d]	rank[e]	CRMS[f]	rank[g]	CRMS[h]
N/36	30000	6.85 Å	395	24t	7.46 Å	14t	6.67 Å
N/24	5000	6.57 Å	209	6t	6.76 Å	2t	6.11 Å
N/12	500	6.28 Å	271	1	6.43 Å	7t	4.45 Å
N/6			44	2	6.13 Å	1t	5.76 Å

[a] As an example, consider the case of three distance restraints (row *N*/24). The RGP method generated $S = 5000$ structures. One of the structures in this set has a CRMS of 6.57 Å from the native structure. Applying S-RB criteria to this set, we selected the $s = 209$ lowest energy conformations and performed a D-RB minimization. We found a near-native match in the top 7 (CRMS = 6.76 Å). We then incorporated a PHD[6] secondary structure prediction into the 209 conformations and ranked each structure according to its D-RB score. We found a near-native match in the top three structures (CRMS = 6.11 Å). The total time for this process was 180 min on an SGI workstation. [b] S denotes the total number of structures generated in the RGP ensemble (for *N*/6 constraints, the RGP method was not used to generate a conformation ensemble; the 271 structures in the selected set for *N*/12 were used as a starting set). [c] The lowest α-carbon CRMS structure in the RGP ensemble. [d] Set of s structures selected from the original RGP ensemble according to the SRB score. [e] Rank of the lowest energy structure possessing the native global fold using the DRB score (t denotes a tie). [f] CRMS of ranked structure from column d. [g] Rank of the lowest energy structure possessing the native global fold using the DRB score after incorporation of predicted sheet and helical regions from a PHD[6] secondary structure prediction. [h] CRMS of ranked structure from column f.

TABLE 2: Results of the *Generate-and-Select* Prediction Hierarchy for Myoglobin (146 residues) Using *N*/12 and *N*/6 Interresidue Restraints[a]

	RGP ensemble		selected set			sec. prediction	
	S	CRMS	s	rank	CRMS	rank	CRMS
N/12	50000	8.95 Å	117	11	8.77 Å	5	7.01 Å
N/6			23	1	9.28 Å	1	6.30 Å

[a] The definition of each column is similar to Table 1.

TABLE 3: List of Interresidue Restraints Used for the LexA Repressor and Myoglobin Structure Predictions[a]

	(a) LexA ($N = 72$ residues)		
N/36 (2)	1−72	25−64	
N/24 (3)	11−31		
N/12 (6)	8−50	28−44	55−68
N/6 (12)	2−53	11−47	18−26
	31−43	51−58	58−65
	(b) Myoglobin ($N = 146$ residues)		
N/12 (12)	1−84	6−129	10−75
	16−119	22−65	30−51
	46−54	88−137	93−144
	102−141	109−131	113−127
N/6 (24)	3−79	9−125	10−130
	13−123	17−116	26−60
	41−48	78−84	101−146
	105−138	117−122	141−146

[a] Each restraint set contained the restraints listed in the corresponding row along with the restraints listed in each prior row.

Given 6 (*N*/12) interresidue distance restraints, an RGP ensemble of only 500 structures resulted in a low-resolution structure that was 6.3 Å CRMS from the native. Adding the predicted secondary structure resulted in a 4.5 Å CRMS structure ranked in the top 10 of all candidates. Figure 6 shows this structure compared to the native.

Generating the 500 structure set with *N*/12 restraints required less than 10 min on our single processor workstation. Increasing the restraint density results in lower efficiency, and thus rather than generate structures with *N*/6 restraints using RGP, we began with the set of 271 lowest S-RB energy *N*/12 structures, and added the remaining six restraints during D-RB minimization. This led to just 44 structures that satisfied all 12 restraints. The second lowest D-RB energy structure had the same overall fold as the native (6.13 Å CRMS). Including the secondary structure prediction and carrying out D-RB minimization resulted in a near-native fold (CRMS = 5.76A) tied for the best D-RB score.

Figure 6. The *generate-and-select* backbone prediction (4.5 Å CRMS, dark strand) for 72-residue lexA repressor (light strand). Six interresidue distance restraints were used in conjunction with predicted secondary structure to obtain this prediction. The complete *generate-and-select* hierarchy required less than 3 h on a single processor R10000 Silicon Graphics workstation for this protein.

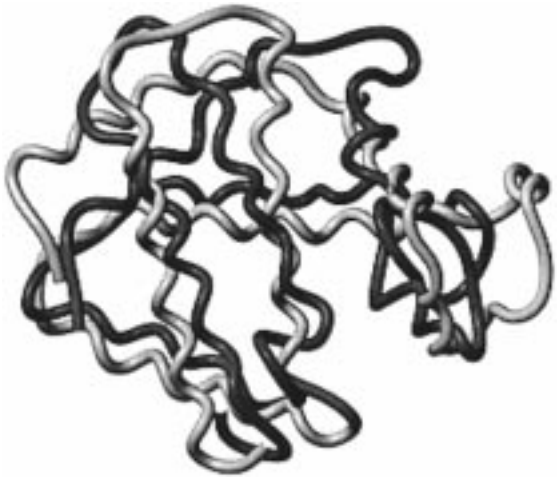

Figure 7. The *generate-and-select* backbone prediction (6.3 Å CRMS, dark strand) for 146-residue myoglobin (light strand). Twenty-four interresidue distance restraints were used in conjunction with predicted secondary structure to obtain this prediction. The complete *generate-and-select* hierarchy required less than 12 h on a single processor R10000 Silicon Graphics workstation for this protein.

B. Myoglobin (146 AA). The RGP method was also successful when applied to a much longer sequence using 12 (∼*N*/12) restraints (Table 2). After an ensemble of $S = 50\,000$ structures was generated using the RGP algorithm, a smaller set of $s = 117$ conformations was selected on the basis of the S-RB score (the nonminimized residue burial score). Applying D-RB to this set led to a near-native match as number 11. Incorporating the secondary structure predicted by PHD[6] into

each conformation and applying D-RB resulted in a near-native structure ranked fifth (7.01 Å CRMS).

Starting with the 117 structures with 12 restraints, we added 12 more (N/6 total) during D-RB minimization. This led to 23 structures that satisfied all 24 restraints. The lowest-energy fold in this set possessed the native overall fold topology. Incorporating the secondary structure prediction led to a highest-ranking structure with the correct overall fold (6.30 Å CRMS). Figure 7 compares this structure to native myoglobin.

5. Discussion

Both distance-geometry and dynamic MC methods produce correct low-resolution structure predictions given $\sim N/6$ interresidue restraints and accurate secondary structure assignments. We have demonstrated that a direct MC approach obtains predictions of similar precision with very little computational effort. Furthermore, since the RGP method employs a very simple packing force field, it can provide a coarse sampling of conformation space many orders of magnitude faster than more detailed dynamic simulation methods. Consequently, RGP is computationally feasible even when restraint information is very sparse.

Levitt and co-workers[20,21] have considered the ability of many different types of functions to "recognize" correct low-resolution (near-native) folds from large sets of incorrect decoys. To measure the success of a particular function, they devised a quality factor,

$$Q = \log_{10}\frac{M}{nr} \tag{18}$$

where M is the total number of structures in the set, r is the highest rank of a near-native structure, and n is the number of near-native structures in the set. While many selection functions recognize native crystal structures with Q-scores greater than 4, near-native structures ($<$4 Å CRMS by their definition) are far more difficult to recognize, with Q rarely exceeding 2. Thus, selecting a near-native structure from a set of greater than 10^5 decoys is not feasible with current recognition potentials.

To successfully recognize near-native folds, a selection function must possess two important properties. First, it should rank the native structure as one of the lowest-energy conformations. Second, the selection function should be insensitive to small structural changes, so that near-native structures appear similar in energy to the native. Unfortunately, selection functions that unambiguously identify the native typically do so at the cost of being highly sensitive to small changes in structure.

Though our S-RB selection function is very simple, it too is sensitive to small structural changes. By performing a short conformational minimization (D-RB) we greatly increase the structural invariance of the S-RB score, since each conformation is evaluated at a local minimum of the selection function.

Given just three restraints for the 72-residue 1lea, an ensemble of $\sim 10^4$ structures contains several native topology conformations. Thus, preserving the native topology in a smaller set of $\sim 10^2$ structures requires $Q \sim 1.5$, a level of recognition attained by our S-RB function, and possibly many previously developed recognition approaches. Most importantly, since this reduced set of structures is very manageable in size, it is computationally feasible to use a dynamic selection procedure to select an even smaller set that will still contain the native topology. In this manner, the most promising topologies are analyzed by a selection procedure that possesses both properties required for successful near-native structure recognition.

6. Conclusion

We have developed a direct Monte Carlo method for efficiently generating the complete set of protein topologies consistent with a set of interresidue distance restraints. We find that fewer than 10^4 distinct topologies are consistent with having three uniformly distributed restraints for a 72-residue protein. The RGP method can sample all of these topologies in less than 1 h using a single Silicon Graphics R10000 processor workstation. Using the simple S-RB and D-RB criteria, it is possible to preserve a small set of structures that contains native topology. Since this remaining set is typically $<$10 structures, we suggest it is computationally feasible to perform much more detailed structure analysis to uniquely determine the native topology.

Future work will apply the *generate-and-select* hierarchy to the de novo prediction problem using predicted interresidue contacts and biochemical structural restraints (such as disulfide bridges) as starting restraints for the RGP algorithm. A distinct benefit of the RGP method over previously developed methods is that only a very small number of restraints ($<N/12$) restraints are needed to generate the native topology. For many protein sequences, knowledge of disulfide bond connectivity may be sufficient to lead to a correct low-resolution structure prediction. Furthermore, because the set of restraints is small compared to other methods, only a few of the most reliably predicted restraints[2] must be used, greatly minimizing the chance of including incorrect constraints in the predictions. Even so, it will be critical to understand how well the RGP method performs when some of the supplied tertiary restraints are inaccurate. Our results suggest that there is a sizable margin for error, since the present work allowed a 7.4 Å distance between restrained residues.

Acknowledgment. This research was supported by the DOE (BCTR DE-FG36-93CH10581) and NSP (CHE 95-2219 and DURIP/ARO). The MSC is also supported by grants from BP Chemical, Exxon, Seiko-Epson, Beckman Institute, Owens-Corning, Avery Dennison, Chevron Petroleum Technology Co., Chevron Chemical Co., Asahi Chemical, and Chevron Research and Technology.

Supporting Information Available: List of 61 nonhomologous, single-domain proteins used in the residue burial study. This material is available free of charge via the Internet at http://pubs.acs.org.

References and Notes

(1) Bystroff, C.; Baker, D. *J. Mol. Biol.* **1998**, *281*, 565.
(2) Goebel, U.; Sander, C.; Schneider, R.; Valencia, A. *Proteins* **1994**, *18*, 309.
(3) Ortiz, A. R.; Kolinski, A.; Skolnick, J. *J. Mol. Biol.* **1998**, *277*, 419.
(4) Benner, S. A.; Cannarozzi, G.; Gerloff, D.; Turcotte, M.; Chelvanayagam, G. *Chem. Rev.* **1997**, *97*, 2725.
(5) Kneller, D. G.; Cohen, F. E.; Langridge, R. *J. Mol. Biol.* **1990**, *214*, 171.
(6) Rost B.; Sander, C. *J. Mol. Biol.* **1993**, *232*, 584.
(7) Kolinski, A.; Skolnick, J.; Godzick, A.; Hu, W. P. *Proteins* **1997**, *27*, 290.
(8) DeWitte, R. S.; Michnick, S. W.; Shakhnovich, E. I. *Protein Sci.* **1995**, *4*, 1780.
(9) Aszódi, A.; Gradwell, M. J.; Taylor, W. R. *J. Mol. Biol*. **1995**, *251*, 308.
(10) Skolnick, J.; Kolinski, A.; Ortiz, A. R. *J. Mol. Biol.* **1997**, *265*, 217.
(11) Ortiz, A. R.; Kolinski, A.; Skolnick, J. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 1020.
(12) Levitt, M. *J. Mol. Biol.* **1976**, *104*, 59.
(13) Sadanobu, J.; Goddard, W. A., III *J. Chem. Phys.* **1997**, *106*, 6722.

(14) Maiorov, V. N.; Crippen, G. M. *Proteins: Struct., Funct., Gen.* **1995**, *22*, 273.
(15) Park, B.; Levitt, M. *J. Mol. Biol.* **1995**, *249*, 493.
(16) Rooman, M. J.; Wodak, S. J. *Biochemistry* **1992**, *31*, 10239.
(17) Mayo, S. L.; Olafson, B. D.; Goddard, W. A., IIII *J. Phys. Chem.* **1990**, *94*, 8897.
(18) Fogh, R. H.; Ottleben, G.; Ruterjans, H.; Schnarr, M.; Boelens, R.; Kaptein, R. *EMBO J.* **1994**, *13*, 3936.
(19) Bolognesi, M.; Onesti, S.; Gatti, G.; Coda, A.; Ascenzi, P.; Brunori, M. *J. Mol. Biol.* **1989**, *205*, 529.
(20) Park, B.; Levitt, M. *J. Mol. Biol.* **1996**, *258*, 367.
(21) Park, B. H.; Huang, E. S.; Levitt, M. *J. Mol. Biol.* **1997**, *266*, 831.