Published in final edited form as:

Optim Methods Softw. 2009 August ; 24(4-5): 837–855. doi:10.1080/10556780902753486.

Enhanced Bounding Techniques to Reduce the Protein Conformational Search Space

Scott R. McAllister and Christodoulos A. Floudas*

Department of Chemical Engineering, Princeton University, Princeton, NJ 08544-5263, U.S.A.

Abstract

The complexity and enormous size of the conformational space that must be explored for the protein tertiary structure prediction problem has led to the development of a wide assortment of algorithmic approaches. In this study, we apply state-of-the-art tertiary structure prediction algorithms and instead focus on the development of bounding techniques to reduce the conformational search space. Dihedral angle bounds on the ϕ and ψ angles are established based on the predicted secondary structure and studies of the allowed regions of ϕ/ψ space. Distance bounds are developed based on predicted secondary structure information (including β -sheet topology predictions) to further reduce the search space. This bounding strategy is entirely independent of the degree of homology between the target protein and the database of proteins with experimentally-determined structures. The proposed approach is applied to the structure prediction of protein G as an illustrative example, yielding a significantly higher number of near-native protein tertiary structure predictions.

1 Introduction

A multitude of approaches yielding significant advances have defined the field of protein structure prediction in recent years. The first major class uses comparative modeling techniques to predict structures by making comparisons with experimentally-determined homologs. A second category applies fold recognition methods that rely upon evolutionary distant homologs and the knowledge that protein folds are more conserved than sequence. The final class of protein structure prediction is made up of first principles approaches. Knowledge-based first principles methods incorporate information or constraints from known structures into statistical models. These approaches can be contrasted with physics-based first principles approaches, which try to predict protein structure based solely upon the primary sequence and the application of detailed force fields and energy models. The recent development of a novel physics-based first principles approach, ASTRO-FOLD, has shown tremendous promise in this area [1]. Three recent reviews provide a thorough overview of the methods for each class of protein structure prediction [2,3,4].

An interesting question that arises prior to the tertiary structure prediction of a protein is “How can known information about a protein be included to narrow the conformational search space?” Typically, the source of this information is due to homologous proteins that have already been experimentally determined. This homology can be exploited to determine likely ϕ and ψ angles, likely interatomic distance bounds, and other information using techniques as varied as fold recognition and fragment assembly. The goal of this paper is to introduce as much relevant information as possible for the tertiary structure prediction of a protein that is independent of the similarity to known protein structures.

* Author to whom all correspondence should be addressed; Tel: +1-609-258-4595; Fax: +1-609-258-0211. floudas@titan.princeton.edu .

The backbone conformation of a protein is typically represented by the ϕ and ψ dihedral angles. The allowed conformations of these angles have been consistently studied for decades. The classical treatment of allowed backbone dihedral angle space defines boundaries using hard sphere atomic overlaps [5]. The approximate location of favorable, low-energy conformations, such as right and left-turning α -helices, β -strands, and 3_{10} helices were identified in a similar fashion [6]. A study of the free energies of an alanine dipeptide in water using molecular dynamics showed agreement with these defined regions, but the boundaries between allowed and disallowed conformations were much less distinct [7]. As the number of available protein structures has increased and the protein structures have been determined to a higher resolution, the definition of these boundaries have become more refined. The structure validation program ProCheck used 463 protein structures to define $10^\circ \times 10^\circ$ degree regions of the Ramachandran plot as either favored, allowed, generously allowed, or outside for non-Glycine and non-Proline residues [8,9]. Kleywegt and Jones [10] showed that 98% of the non-Glycine residues exist within less than 20% of the area of a Ramachandran plot by using a set of 403 high resolution protein models. Lovell et al. [11] used a database of 500 high quality protein structures and established a set of allowed and favorable contours for four types of residues (a) Glycine, (b) Proline, (c) Pre-Proline and (d) other.

The definition of allowed and favorable regions of the ϕ and ψ space in these applications has been intended for structure validation applications. When using X-ray crystallography or NMR techniques to experimentally determine the structure of a protein, a model of the protein structure must be built that fits the data. The use of defined ϕ and ψ dihedral angle regions ensures a protein model not only fits the experimental data but also satisfies generally observed rules for these dihedral angles within an acceptable tolerance. The goal of the dihedral angle information in this paper, however, is not to validate a structure determined with the aid of experimental data. Instead, this defined region of the ϕ/ψ space will be used to establish strict bounds on the allowable backbone conformations for protein structure prediction applications. This idea of restricting or using the ϕ and ψ dihedral angle information is not new. A number of methods, including the Rosetta structure prediction approach [12,13], rely on the conformational statistics of known protein structures as an integral part of their energy function. The conformational space annealing (CSA) approach has limited the sampling of dihedral angles with disfavored ϕ and ψ angles by introducing bounds during the creation of random and mutated trial conformations [14]. Floudas and co-workers [15] suggested that the physically feasible values for the backbone ϕ and ψ dihedral angles are $[-180, -50]$ and $[-75, 175]$, respectively, for their study of the 20 residue peptide melittin. Although these specific ranges are restrictive because they ignore the favorable left-handed α -helix and also the greater conformational flexibility of glycine, the introduction of dihedral angle bounds is still a reasonable method for narrowing the conformational search space. The introduction of generic bounds based upon the studies of Lovell et al. [11] will be presented in Section 2.1.

The introduction of distance bounds for interatomic distances within protein structures has been considerably less studied. In specific protein folding applications, comparative modeling approaches, such as the MODELLER program, can build spatial restraints from homologous proteins of known structure and subsequently predict a protein structure that satisfies them [16,17]. With the notable exceptions of bond lengths and other intra-residue distances, very little information can be applied to the problem for an entirely unrestricted search of the protein conformational space.

A first principles framework that predicts protein secondary structure information prior to a final tertiary structure prediction stage, such as the ASTRO-FOLD approach [1], may benefit from the introduction of bounds on interatomic distances. A critical area of interest is the need for quicker and more accurate methods for tertiary structure prediction, given secondary structure information and β -sheet topologies. A carefully restrained problem can significantly

reduce the feasible search space of the protein. The search space reduction can result in both an increase in the speed of the algorithm, as well as improved results through a more focused problem statement. The introduction and refinement of distance bounds given secondary structure location and topology information will be the focus of Section 2.2.

2 Methods

The secondary structure of a protein can be predicted using a variety of algorithmic techniques [18,19,20]. In addition, one secondary structure prediction approach has demonstrated the ability to predict the topology of protein secondary structure, specifically the β -sheet topologies and the disulfide bridges that may form between Cysteine residues [21]. The goal of the proposed work, given this secondary structure information, is to develop a set of meaningful dihedral angle bounds and distance bounds between C^α atoms within the protein backbone.

The introduction of dihedral angles bounds can be done using only the predicted secondary structure location and the residue identity, as presented in Section 2.1. The bounds on interatomic distances will then be discussed in Section 2.2. In Section 2.3, the use of torsion angle dynamics is discussed to generate an ensemble of protein structures that satisfies these bounds. A hybrid global optimization method for tertiary structure prediction is described in Section 2.4 to identify low-energy protein structures that satisfy these bounds.

2.1 Dihedral Angle Bounds

There is a variety of ways to constrain the dihedral angles based on the prediction of secondary structure elements. The ideal, right-handed α -helix adopts a ϕ angle of approximately -57° and a ψ angle of approximately -47° [22]. In a folded natural protein, the geometry of an α -helix may deviate from these values depending upon its environment. As shown in Table 1, Klepeis and Floudas have used two similar sets of dihedral angle bounds for residues predicted to be part of an α -helix. The ideal values of the α -helix and the α -helical bounds are plotted against the allowed and strictly allowed ϕ and ψ angles (for all residues except Glycine and Proline) as determined by examining a database of 500 high quality protein structures [11] in Figure 1.

The ideal antiparallel β -strand adopts a ϕ angle of approximately -139° and a ψ angle of approximately -135° , whereas the parallel β -strand has a ϕ angle of -119° and a ψ angle of 113° [22]. The geometry of the β -strands also deviates from these values in folded proteins depending on the environment and other factors. As shown in Table 1, Klepeis and Floudas have used two similar sets of dihedral angle bounds for residues predicted to be part of a β -strand. These values are used regardless of the existence of a parallel or an antiparallel orientation.

If the structure of the loop residues is not predicted using available methodologies [23], the dihedral angles of these residues can still be restrained based on the strictly allowed regions of the ϕ and ψ angles [24]. These restraints, based upon the identity of the residue at a given position, are presented in Table 1.

Rotation around the $C'-N$ bond is represented by the ω dihedral angle. Unlike the relatively flexible backbone ϕ and ψ dihedral angles, the feasible conformations of the ω angle are greatly reduced due to the partial double bonded nature of the $C'-N$ peptide bond. Although the cis form ($\omega=0^\circ$) of the peptide bond is achievable, this form disfavorably places neighboring C^α atoms in close proximity. As a result, the trans form ($\omega=180^\circ$) of the peptide bond is approximately 1000 times more common [22]. For the protein ensemble generation and tertiary structure prediction applications discussed here, only the trans form of the peptide bond will be considered. In these applications, the ω dihedral angle will be restricted to the range of [160,200] to allow for limited flexibility around the trans conformation of the peptide bond.

For most residues, further information about the dihedral angles associated with the side chain atoms (the χ angles) is needed to fully specify the position of all the atoms within the residue. Although protein side chains often adopt a discrete set of favorable conformations, there are very few generic bounds that can be introduced based only upon the residue identity. The main exception to this rule are dihedral angles that correspond to the placement of symmetric atoms. An example of type of symmetry is the χ_1 angle of alanine, which controls the placement of the $H^{\beta 1}$, $H^{\beta 2}$, and $H^{\beta 3}$ atoms. These three atoms are indistinguishable, so popular convention assigns the $\beta 1$ label to the atom placed by a dihedral angle in the range of $[0,120]$. In the implementation of this algorithm, symmetric atoms will be restricted to ranges of $[0,120]$ or $[0,180]$ depending on the number of symmetric atom positions that are defined by a single χ angle.

2.2 Distance Bounds

A number of distance bounds can be developed from basic secondary structure information that may be known about a protein. These types of bounds can be divided into two categories. The first is local interactions that occur within a single element of secondary structure. A number of bounds can also be established on non-local interactions between elements of secondary structure. To maintain a manageable representation of all of these distance bounds, only the distances between the C^α carbon residues will be considered throughout the rest of this paper.

The previous strategies for introducing bounds on the distances within a protein will be discussed first. The presentation of a dataset of proteins used to establish and validate some of the lower and upper bounding values will follow. Improved strategies for bounding local interactions and non-local interactions will then be discussed.

2.2.1 Previous Distance Bounding Strategies—Distance constraints can be introduced based on predicted secondary structure elements or predicted topology. The formation of an α -helix is dependent upon the formation of a hydrogen bond between the carbonyl oxygen at residue i and the backbone -NH at position $i + 4$. Klepeis and Floudas [1] introduce bounds on the distances between the C^α atoms of these helical residues to enforce the formation of the hydrogen bonding network. Knowledge of β -sheet topology can also be used to introduce bounds on the distances before the application of the tertiary structure prediction algorithm. Klepeis and Floudas [1] again use bounds on the distances between the C^α atoms for β -sheets. These bounds are non-local and help to enforce the formation of a β -sheet between two opposing β -strands. The lower and upper distances used for these bounds are presented in Table 2. Other predicted non-local tertiary contacts, such as disulfide bridges, can be introduced as distance bounds as well. Disulfide bridges are enforced by bounding the distance between two Cysteine sulfur atoms with a range of 2.01 to 2.03 Å.

2.2.2 Dataset—In order to obtain a representative set of protein structure tendencies, a large set of non-homologous protein structures that span the Protein Data Bank [25] should be selected. For this implementation, a set of structures that contain no more than 25% sequence similarity, denoted as the PDBSelect25 [26], has been used to develop the distance bounds based on geometric tendencies. The recent release of the PDBSelect25 contains 2216 protein chains and a total of 352855 residues.

For each protein, there are a number of distance bounds that will be based on the secondary structure at a particular residue position. This secondary structure will be assigned using the Dictionary of Protein Secondary Structure (DSSP) [27]. The DSSP method assigns secondary structure through the identification of hydrogen bonding patterns indicative of α -helices, β -sheets, and turns.

The derivation of distance bounds results from a threshold that can be established on the relative number of violations that may occur for a given value of the distance bounds. For example, a lower bound of 4.5 Å may be such that 99.5% of the occurrences in the PDBSelect25 dataset satisfy such a bound. By statistically quantifying the establishment of lower and upper bounds, we may extend these bounds to be representative of the overall tendencies of protein structures.

2.2.3 Local Interactions—There are three main types of local interactions that should be considered, based on the type of secondary structure. The first type of local interactions involves helical regions of the protein. α -helical residues are characterized by an intra-helical hydrogen bond between residues i and $i + 4$, four positions away in the amino acid sequence. Instead of merely characterizing the helix by this single intra-helical distance, let us consider all $i, i + n$ distances within the helix. Equations 1-2 will be used as a method to derive bounds from the dataset in Section 2.2.2. The lower and upper bounds of Equation 1 are developed specifically for each value of $n \leq 8$. In contrast, for $n > 8$ in Equation 2, the helical interactions are generalized into distance bounds normalized by residue separation to more easily account for longer helices.

$$\alpha_n^L \leq d_{i,i+n} \leq \alpha_n^U \quad \forall 2 \leq n \leq 8 \quad (1)$$

$$n \cdot \alpha_\infty^L \leq d_{i,i+n} \leq n \cdot \alpha_\infty^U \quad \forall n > 8 \quad (2)$$

Consider the example of Figure 2. This figure considers the case of $n = 3$ from Equation 1, analyzing the distances that commonly occur for residues 3 positions away within a helix. A single-peaked distribution of distance occurrences is clearly visible. In this case, a lower bound of 4.0 Å and an upper bound of 6.0 or 6.2 Å appear to be appropriate to capture the typical intrahelical $i, i + 3$ distance for protein structures.

A similar set of distance bounds can be established for local β -strand structure. Equations 3-4 are used to derive bounds on the upper and lower distance bounds of β -strands within a protein. Equation 3 specifically addresses distances between residues $i, i + n$ for cases of $2 \leq n \leq 6$. For intra-strand distances separated by a large number of residues, $n > 6$, Equation 4 defines per residue characteristic bounds.

$$\beta_n^L \leq d_{i,i+n} \leq \beta_n^U \quad \forall 2 \leq n \leq 6 \quad (3)$$

$$n \cdot \beta_\infty^L \leq d_{i,i+n} \leq n \cdot \beta_\infty^U \quad \forall n > 6 \quad (4)$$

The typical distances between the C^α atoms of any two neighboring residues is approximately independent of the secondary structure or residue identity, due to the planar nature of the peptide bond. Thus, for the residue pair $i, i + 1$, a single set of distance bounds can be established for all residue classifications in Equation 5.

$$N^L \leq d_{i,i+1} \leq N^U \quad (5)$$

It is important to distinguish between distance bounds that further constrain the conformational space for tertiary structure prediction applications and redundant bounds that are only

introduced for modeling applications when protein coordinates are not available. The bounds on the local interactions are not typically directly useful for protein tertiary structure prediction, but can be useful in optimization-based models where no coordinate system is defined, such as one proposed for residue contact prediction in mixed α/β proteins [28]. The values of the bounds recommended for use in this section are presented in Table 3. For neighboring residues, lower and upper bounds of 3.7 Å and 3.9 Å, respectively, are recommended.

2.2.4 Nonlocal Interactions—When a secondary structure prediction method that provides topology information is used [21], tight distance bounds can be developed on contacts between β -sheets and disulfide bridges. Hydrogen bonds form between residues within the β -sheets, bounding the distance between them. The distance bound form for the main β -contact residues is illustrated in Equation 6 where i, j is a contact between two β -sheets.

$$BC_0^L \leq d_{ij} \leq BC_0^U \quad \forall i, j \in \beta\text{-contact} \quad (6)$$

Some additional distance bounds can be applied to residues within β -sheets that form layered topologies. Thus, if two distinct β -strands form β -sheets with a third strand, the distance between them can be tightly bound in a fashion distinct from Equation 6. These so-called “double” contacts, and even “triple” contacts, can be used to bound such residue-to-residue distances as defined by Equations 7-8. The type of residue-residue pairs that are considered double and triple contacts are illustrated in Figure 3.

$$DC^L \leq d_{ij} \leq DC^U \quad \forall i, j \in \text{double contact} \quad (7)$$

$$TC^L \leq d_{ij} \leq TC^U \quad \forall i, j \in \text{triple contact} \quad (8)$$

Further bounds on contacting β -sheets can be introduced based on the assumption of approximately parallel (in an undirected sense) β -strands that are spanned by approximately perpendicular contacting residues. The development of these “cross” contact bounds can then be aided by the use of the Pythagorean Theorem and introduced as shown in Equations 9-10. These cross contact distance bounds are only imposed across β -strands connected by a single β -contact or a double contact. The type of residue-residue pairs that are considered single cross and double cross contacts are illustrated in Figure 4.

$$SX^L \leq d_{ij} \leq SX^U \quad \forall i, j \in \text{single cross contact} \quad (9)$$

$$DX^L \leq d_{ij} \leq DX^U \quad \forall i, j \in \text{double cross contact} \quad (10)$$

The values of the bounds recommended for use in this section are presented in Tables 4-5.

2.3 Protein Ensemble Generation

The introduction of bounds on both the dihedral angles and interatomic distances requires a method to transform these constraints into valid protein structures. Metric matrix distance geometry can be used to uniquely determine the Cartesian coordinates of all atoms if the exact distance values are available for all pairs of atoms [29,30,31]. For the applications discussed in this paper, however, the distance values are neither exact nor completely specified between

all pairs of atoms. Distance geometry algorithms such as EMBED [31] and dgsol [32] can be used to produce molecular conformations that satisfy a set of sparse distance constraints. A number of alternatives to distance geometry algorithms exist, including molecular dynamics [33,34], variable target function methods [35,36], and molecular dynamics in torsion angle space (torsion angle dynamics) [37,38,39]. The various algorithms for protein structure determination are reviewed in further detail elsewhere [40].

Torsion angle dynamics fixes the bond lengths and bond angles of the protein structure and defines a protein solely by its torsion angles. While this representation has the advantage of reducing the number of variables in the system and allowing for longer integration time steps than traditional molecular dynamics approaches, it has the disadvantage of increasing the complexity of equations of motion that must be solved at each time step. The development of a torsion angle dynamics algorithm that scales linearly with the number of dihedral angles helped overcome this complexity issue [37]. This algorithm has been combined with simulated annealing approaches and successfully applied to NMR structure determination problems [38,39]. A study of the conformational sampling properties of torsion angle dynamics algorithms showed that good sampling can be achieved in both well-constrained proteins and long unconstrained polypeptides [41].

The new, tighter set of dihedral angle and distance bounds can be used as input to a torsion angle dynamics approach to generate an ensemble of protein conformations that satisfy the bounds. A program developed for NMR structure refinement, CYANA, is used for the initial ensemble generation [38]. This algorithm performs molecular dynamics in torsion angle space to minimize a scoring function that sums the violation of the dihedral angle bounds, the violation of the lower and upper distance bounds, and the violation of steric constraints based upon the van der Waals radii of interacting atoms.

2.4 Tertiary Structure Prediction

The new set of dihedral angle and distance bounds can also be used as part of a tertiary structure prediction framework. The tertiary structure prediction problem has previously been formulated as a constrained nonlinear minimization problem [15]. The basic formulation is the minimization of the ECEPP/3 force field [42] energy over torsion angle space, subject to upper and lowering bounding constraints on these torsion angles as well as distance constraints in Cartesian space.

The ASTRO-FOLD tertiary structure prediction approach applies a hybrid algorithm that consists of the α BB global optimization algorithm, a stochastic global optimization method, and a molecular dynamics approach in torsion angle space [1,15]. The torsion angle representation of the model significantly reduces the size of the independent variable set, while only modestly increasing the model complexity.

The use of the α BB global optimization algorithm [43,44,45,46,47,48,49,50,51] provides a theoretical guarantee of convergence to the global minimum solution by establishing a valid sequence of upper and lower bounds on the potential energy minimum. The upper bound on the global minimum is obtained by constrained nonlinear minimization on any protein conformation. The lower bound is determined by creating a valid convex underestimating function and identifying its minimum function value. The algorithm converges by partitioning regions of conformational space at every level of a branch and bound tree to create subregions where tighter lower bounding functions can be derived.

Stochastic optimization methods can further improve the performance of the upper bounding approach of the formulation. One such hybrid global optimization method, described as an alternating hybrid, has been proposed [15,52]. It combines the deterministic α BB approach

with the stochastic approach of conformational space annealing [53,54,55,56,57]. Conformational space annealing balances genetic algorithm techniques of mutations and crossovers with simulated annealing to identify low energy conformers. A hybrid algorithm of this form retains the deterministic guarantees of convergence while yielding a much more efficient search for the native state [15,52].

The local minimization strategy for this hybrid algorithm is composed of three stages (1) torsion angle dynamics, (2) rotamer optimization, and (3) sequential quadratic programming. The identification of initial feasible points is critical to the success and efficiency of the constrained nonlinear minimization algorithm. The basic torsion angle dynamics routine has been integrated into the hybrid optimization algorithm by interfacing with the CYANA software package [38]. The number of structures that satisfy both the distance and dihedral angle constraints after this routine is significantly increased, especially in the cases of distance constraints with tight bounds or large numbers of distance constraints. The goal of the rotamer optimization stage is to remove any steric clashes that may exist between protein side chains in conformations produced by torsion angle dynamics or trial conformations from the conformational space annealing algorithm. The rotamer optimization is implemented as a three stage approach, combining the FASTER algorithm [58] with a cyclical search and a random search. With the backbone fixed in its initial conformation, the rotamers are varied one residue at a time and a replacement is made if any energetic improvement is realized. In this way, rotamer optimization acts as an efficient local minimizer, with no guarantee that the optimal set of rotamers is chosen. After the rotamer optimization stage is completed, this feasible, low-energy conformation is subject to local minimization using sequential quadratic programming techniques. The NPSOL package [59] is especially attractive for the protein structure prediction minimization problem as it requires relatively few evaluations of the computationally expensive ECEPP/3 potential energy function. The minimization using NPSOL is repeated with multiple values of the line search tolerance to improve convergence to feasible protein conformations.

An improved parallel implementation of this hybrid global optimization approach has been developed [52]. In this approach, a single processor is allocated as a control processor, to control both the α BB iterations and the CSA iterations. The control processor maintains the lower and upper bounds and the lower-bounding subregions for the α BB portion of the approach as well as the bank of conformers necessary for the CSA iterations. This allocation of a single processor is effective because the work distributed to the work processors requires several orders of magnitude more time for processing than for the communication associated with delivering the work and receiving the results. In fact, the idle time of this control processor can be used to further explore the conformational space with perturbations of existing structures. Initially, all of the remaining processors are assigned α BB iterations. Once, the CSA bank is full, a pre-specified number of these processors are assigned CSA iterations to perform instead. A complete description of the implementation of this improved hybrid global optimization algorithm is available in [52].

The ASTRO-FOLD methodology has been successfully applied to a varied set of proteins throughout the range of small to medium-sized proteins [1]. The recent success of the ASTRO-FOLD method in a double blind prediction of a four-helix bundle reinforces the value of the approach [60].

3 Illustrative Example: Protein G

A 56 amino acid protein, the immunoglobulin binding domain of protein G from the *Streptococcus* species, was selected to illustrate the bounding strategy proposed in Section 2. Both NMR spectroscopy and X-ray crystallography have been used to determine the structure

of this protein, which folds into a $\beta\beta\alpha\beta\beta$ motif. The four β -strands form a single β -sheet and a single α -helix is located above the plane of the β -sheet. The single α -helix spans residues 23-35 and the four β -strand span the residues 2-8, 13-19, 42-46, and 51-55, with an experimental topology as shown in Figures 3-4. Protein G is a well-studied system due to its tightly-packed hydrophobic core and extensive hydrogen bonding network. The following sections will describe the development of distance and dihedral angle bounds, the application of the bounding strategy using torsion angle dynamics, and the application of the bounding strategy using a hybrid global optimization method for tertiary structure prediction.

3.1 Distance and Dihedral Angle Bounds

The ϕ and ψ dihedral angles of the α -helix region are restricted to $[-90, -40]$ and $[-60, 10]$, respectively. The residues in the β -strands have their ϕ and ψ dihedral angles constrained to the ranges $[-155, -75]$ and $[110, 180]$, respectively. The ϕ and ψ dihedral angles of all the remaining residues are both allowed to vary from $[-180, 180]$. The imposed distance constraints for the first study correspond to the bounding strategy of Klepeis and Floudas [1], as shown in Table 2. In Table 6, the 9 distance bounds associated with the formation of the hydrogen bonding network in the α -helical structure are presented. An additional 17 lower and upper distance bounds are introduced to enforce the formation of the β -sheet topology as shown in Table 7. These bounds use an expanded distance range, allowing the distance between a pair of contacting β -strand residues to be as close as 4.25 Å instead of 4.5 Å. This set of dihedral angle and distance constraints will be denoted as the “original” bounding strategy and will be the basis for comparing the results of the new bounding strategy.

The methods of Sections 2.1-2.2 are then used to develop additional dihedral angle and distance bounds on the Protein G structure given this information. The loop residues are constrained according to the methodology of Section 2.1, as shown in Table 8. The three Glycine residues have a restricted range for the value of ϕ , and the total area of the ϕ/ψ dihedral angle space that must be explored for each Glycine residue is approximately 86% of the original conformational space. The remaining 14 non-Glycine loop residues have restricted ϕ and ψ angle bounds, which narrows the total area of the ϕ/ψ dihedral angle space to approximately 58% of the original area.

The β -sheet contacts from the original bounding strategy, established above in Table 7, can be examined to extract double and triple β -sheet contacts. The distances between the C^α carbons in these double and triple contacts can be bounded as shown in Table 9. The ten double contacts, with the interactions $\beta1-\beta3$ and $\beta2-\beta4$ are restricted to the range 8.5-13.0 Å, while the five triple contacts between $\beta2$ and $\beta3$ are bounded between 13.0 and 19.5 Å.

The β -sheet contacts from the original bounding strategy can also be used to create bounds on single cross and double cross contacts. The set of eight single cross contacts introduced for Protein G is presented in Table 10. The four single cross contacts between $\beta1$ and $\beta2$ are established as two sets with a separation of 5 (the maximum separation allowed) and distance bounds between 12.8 and 18.7 Å. The single cross contacts for the pairs $\beta1, \beta4$ and $\beta3, \beta4$ have a separation of 4 and distance bounds of 10.1-15.4 Å. A total of 36 double cross contacts with offsets of 1-3 are enumerated in Table 11. Of these 36 double cross contacts, 16 have an offset of 1 with distance bounds of 9.0-13.5 Å, 12 have an offset of 2 with distance bounds of 9.5-14.8 Å, and 8 have an offset of 3 with distance bounds of 10.8-16.7 Å.

Tables 6-11, along with the ϕ and ψ dihedral angle restrictions on the secondary structure elements, represent the new bounding strategy for Protein G. In total, 85 lower and upper distance bounds are introduced using this new strategy, compared to 26 lower and upper distance bounds in the original strategy. The number of distance bounds, however, is not an adequate method to judge the utility of a set of distance constraints. Clearly, the distance

between every atom can be bounded by $(0, \infty)$, but these bounds would not add any meaningful information to the model. The ensemble generation method of Section 2.3 and the tertiary structure prediction method of Section 2.4 will instead be used with the original and new bounding strategies to evaluate the utility of the new sets of bounds.

3.2 Ensemble Generation with Bounds

For each bounding strategy, an ensemble of 1000 protein conformations was generated. These protein structures will be evaluated by comparing the root mean squared deviation (RMSD) of the ensembles to the native protein G structure. An RMSD value of 0.0 Å indicates perfect agreement between the native structure and the predicted protein conformer, and the RMSD value increases as the prediction deviates from the native structure. The evaluation of the RMSD values throughout this paper are calculated using the vectors representing the C^α coordinates of the predicted protein structure and the experimentally-determined structures. The RMSD of the protein ensembles will be measured against the violation value of the imposed distance bounds, which will be the only measure of structural quality discussed in this section.

Figure 5 presents the distance constraint violation values of the proteins in the ensemble versus their RMSD to the native protein G structure for both bounding strategies. The RMSD values achieved by the new bounding strategy range from 3.41-10.55 Å, while the RMSD values from the original bounding strategy vary from 3.73-10.68 Å. If we define 5.0 Å as the RMSD cutoff for a near-native protein structure, the new bounding strategy identifies 65 near-native structures, 16 more than from the original bounding strategy. In Figure 5 there is a clear separation of conformers generated using the new bounding strategy into two distinct clusters centered around RMSD values of 9.5 Å and 5.5 Å. This separation results from the placement of the α -helix relative to the β -sheet. If the β -sheet region is fixed in a given orientation, sometimes the helix appears in front of the β -sheet and sometimes the helix appears behind the β -sheet.

One of the major focuses of the distance bounds in this paper is to introduce constraints related to β -sheet formation given a set of contacts between β -strands. A better illustration of the utility of the new bounding strategy can be obtained if the RMSD values are computed only for the residues in the β -strand regions of protein G. Figure 6 shows the RMSD values for the selected regions of these proteins plotted against the distance constraint violation values. The new bounding strategy achieves a low RMSD of 1.40 Å and a maximum RMSD of 5.72 Å, while the RMSD values using the original bounding strategy vary from 1.62-6.93 Å. If we define 2.0 Å as the RMSD cutoff for a near-native protein structure, the new bounding strategy identifies 79 near-native structures, 54 more than from the original bounding strategy.

3.3 Tertiary Structure Prediction with Bounds

The utility of the proposed bounding strategy was validated by using torsion angle dynamics to create an ensemble of conformations of protein G in Section 3.2. This proposed bounding strategy led to increased conformational sampling of near-native structures, as defined by their RMSD from the native protein G structure. The ability to generate near-native structures based on torsion angle dynamics is important, but it is more advantageous to identify these same near-native structures as energetically favorable using the proposed tertiary structure prediction algorithm. For the analysis in this section, a distinct conformation is defined to be a protein structure that has a RMSD of at least 1.0 Å from all other conformations in the ensemble

The proposed tertiary structure prediction algorithm of Section 2.4 was applied to the structure prediction of protein G using both the original bounding strategy of Klepeis and Floudas [51, 1] and the new bounding strategy of Section 2. As described in Section 3.1, 85 lower and upper

distance bounds are introduced using the new strategy, compared to 26 lower and upper distance bounds for the original strategy. Tighter ϕ and ψ dihedral angles bounds are also introduced for the 17 loop residues in the new bounding strategy. The quality of each ensemble of predicted protein conformers can be evaluated by presenting the RMSDs from the native protein G structure as a function of the potential energy, as shown in Figure 7. Table 12 compares the ensembles generated by both bounding strategies. Although the RMSD of the lowest energy conformation is slightly lower using the original bounding strategy, there is a clear advantage to implementing the new bounding strategy. The best single conformations, measured by both RMSD and GDT TS score [61,62], are found using the new strategy. Of even more importance is the much larger number of near-native conformations that are identified using the new bounding strategy. The new bounding strategy is able to identify ten times as many distinct protein conformations below 3.0 Å RMSD to the native protein G structure when it is combined with the ASTRO-FOLD tertiary protein structure prediction method. This difference further illustrates the need to impose any available additional bounding information for the tertiary structure prediction problem, allowing the proposed algorithm to focus its efforts in the most fruitful regions of the conformational space.

4 Summary and Discussion

There exists a clear trade-off between the selection of worst-case bound values and tight bounds that exclude large portions of the feasible space while introducing small deviations from some native protein structures. The numerical values of these bounds can be adjusted depending on the application and the confidence in the secondary structure location and topology predictions. For example, in Figure 1, the allowed and strictly allowed dihedral angle values are presented as determined by Lowell et al. The allowed dihedral angle region encompasses 99.95% of the dihedral angles in a database of high resolution structures. The strictly allowed dihedral angle regions significantly shrink the conformational space, but only include 98% of the angles in the database. While the values of the bounds are not as important as the strategy, the bounds proposed here are fairly conservative and tended to fall in the 90-95% range. In fact, the largest contributions to the RMSD in the examples studied was the spatial relationship of the alpha-helix to the beta-strand and the local loop and coil conformations, which were not the result of incorrect bounds on the distances and dihedral angles.

The utility of the bounding strategy described here is dependent upon the quality of the secondary structure location and topology predictions. It is inherent in a framework-based approach that errors in the earliest stages of the framework will propagate to the later stages. While the proposed bounding strategy cannot “fix” an incorrect secondary structure prediction, it can focus the tertiary structure prediction on the most native-like regions of the conformational space for a protein with the predicted topology. If the limitations of available computational resources can be overcome, the bounding strategy and tertiary structure prediction can be applied to a number of the highest ranking secondary structure topology predictions. For mixed-integer linear optimization approaches [21,64,28], such a rank-ordered list can easily be produced using integer cut techniques [65].

5 Conclusions

A new strategy for establishing bounds on the backbone dihedral angles and C^α - C^α distances within the protein structure has been presented. The dihedral angle bounds on the ϕ and ψ angles within the loop regions of a protein structure have been refined to incorporate known information on the allowed backbone conformations of a residue. The use of distance bounds to enforce knowledge of secondary structure elements and hydrogen bonding patterns was previously explored, but has been further investigated and augmented here. A number of additional distance bounds were introduced, especially related to the formation of β -sheet

topologies, that enabled a torsion angle dynamics approach to generate a significantly higher number of near native protein conformations for protein G than the original bounding strategies. Similar improvements were observed when the bounding strategy used for the tertiary structure prediction of protein G. The restrictions on the dihedral angles and distances within the protein structures allows both of these conformational search algorithms to spend more time sampling the correct (i.e., near-native) regions of space, without relying upon homology between the target protein and other known protein structures.

Acknowledgments

CAF gratefully acknowledges financial support from the National Science Foundation, the National Institutes of Health (R01 GM52032, R24 GM069736), and the US EPA (GAD R 832721-010). This work has not been reviewed by and does not represent the opinions of the funding agencies.

References

- [1]. Klepeis JL, Floudas CA. ASTRO-FOLD: A combinatorial and global optimization framework for ab initio prediction of three-dimensional structures of proteins from the amino acid sequence. *Biophys. J* 2003c;85:2119–2146. [PubMed: 14507680]
- [2]. Floudas CA, Fung HK, McAllister SR, Mönnigmann M, Rajgaria R. Advances in protein structure prediction and de novo protein design: A review. *Chem. Eng. Sci* 2006;61:966–988.
- [3]. Floudas CA. Computational methods in protein structure prediction. *Biotechnol. Bioeng* 2007;97:207–213. [PubMed: 17455371]
- [4]. Zhang Y. Progress and challenges in protein structure prediction. *Curr. Opin. Struct. Biol* 2008;18:342–348. [PubMed: 18436442]
- [5]. Ramachandran GN, Sasisekharan V. Conformations of polypeptides and proteins. *Adv. Protein Chem* 1968;23:283–437. [PubMed: 4882249]
- [6]. Ramakrishnan C, Ramachandran GN. Stereochemical criteria for polypeptide and protein chain conformations. ii. allowed conformations for a pair of peptide units. *Biophys. J* 1965;5:909–933. [PubMed: 5884016]
- [7]. Anderson AG, Hermans J. Microfolding: Conformational probability map for the alanine dipeptide in water from molecular dynamics simulations. *Prot. Struct. Funct. Gen* 1988;3:262–265.
- [8]. Morris AL, Macarthur MW, Hutchinson EG, Thornton JM. Stereochemical quality of protein structure coordinates. *Prot. Struct. Funct. Gen* 1992;12:345–364.
- [9]. Laskowski RA, Macarthur MW, Moss DS, Thornton JM. ProCheck—a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr* 1993;26:283–291.
- [10]. Kleywegt GJ, Jones TA. Phi/psi-chology: Ramachandran revisited. *Struct* 1996;4:1395–1400.
- [11]. Lovell SC, Word JM, Richardson JS, Richardson DC. The penultimate rotamer library. *Prot. Struct. Funct. Gen* 2000;40:389–408.
- [12]. Simons KT, Kooperberg C, Huang C, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol* 1997;268:209–225. [PubMed: 9149153]
- [13]. Rohl CA, Strauss CEM, Chivian D, Baker D. Modeling structurally variable regions in homologous proteins with Rosetta. *Prot. Struct. Funct. Bioinf* 2004;55:656–677.
- [14]. Lee J, Scheraga HA. Conformational space annealing by parallel computations: Extensive conformational search of met-enkephalin and the 20-residue membrane-bound portion of melittin. *Int. J. of Quant. Chem* 1999;75:255–265.
- [15]. Klepeis JL, Pieja MT, Floudas CA. Hybrid global optimization algorithms for protein structure prediction : Alternating hybrids. *Biophys. J* 2003b;84:869–882. [PubMed: 12547770]
- [16]. Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol* 1993;234:779–815. [PubMed: 8254673]
- [17]. Marti-Renom MA, Stuart A, Fiser A, Sanchez R, Melo F, Sali A. Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct* 2000;29:291–325. [PubMed: 10940251]

- [18]. Rost B, Sander C. Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol* 1993;232:584–599. [PubMed: 8345525]
- [19]. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol* 1999;292:195–202. [PubMed: 10493868]
- [20]. Klepeis JL, Floudas CA. Ab initio prediction of helical segments in polypeptides. *J. Comput. Chem* 2002;23(2):245–266. [PubMed: 11924737]
- [21]. Klepeis JL, Floudas CA. Prediction of beta-sheet topology and disulfide bridges in polypeptides. *J. Comput. Chem* 2003a;24:191–208. [PubMed: 12497599]
- [22]. Creighton, TE. *Proteins: Structures and Molecular Properties*. second edition. W. H. Freeman; New York, NY: 1993.
- [23]. Mönnigmann M, Floudas CA. Protein loop structure prediction with flexible stem geometries. *Prot. Struct. Funct. Bioinf* 2005;61:748–762.
- [24]. Lovell SC, Davis IW, Arendall WB III, de Bakker PIW, Word JM, Prisant MG, Richardson JS, Richardson DC. Structure validation by $C\alpha$ geometry: ϕ , ψ , and $C\beta$ deviation. *Prot. Struct. Funct. Gen* 2003;50:437–450.
- [25]. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. *Nuc. Acids Res* 2000;28(1):235–242.
- [26]. Hobohm U, Sander C. Enlarged representative set of protein structures. *Prot. Sci* 1994;3:522–524.
- [27]. Kabsch W, Sander C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 1983;22:2577–2637. [PubMed: 6667333]
- [28]. McAllister, SR.; Floudas, CA. α -helical residue contact prediction in mixed α/β proteins using mixed integer linear programming. 2008. In preparation
- [29]. Blumenthal, LM. *Theory and Applications of Distance Geometry*. Cambridge University Press; Cambridge, UK: 1953.
- [30]. Crippen GM. A novel approach to the calculation of conformation: Distance geometry. *J. Comp. Phys* 1977;26:449–452.
- [31]. Crippen, GM.; Havel, TF. *Distance Geometry and Molecular Conformation*. John Wiley and Sons; New York, NY: 1988.
- [32]. Moré JJ, Wu Z. Distance geometry optimization for protein structures. *J. Global Optim* 1999;15:219–234.
- [33]. Allen, MP.; Tildesley, DJ. *Computer Simulation of Liquids*. Clarendon Press; Oxford, UK: 1987.
- [34]. Brünger, AT. X-PLOR, Version 3.1. A System for X-Ray Crystallography and NMR. Yale University Press; New Haven, CT: 1992.
- [35]. Braun W, Go N. Calculation of protein conformations by proton-proton distance constraints. a new efficient algorithm. *J. Mol. Biol* 1985;186:611–626. [PubMed: 2419572]
- [36]. Güntert P, Wüthrich K. Improved efficiency of protein structure calculations from NMR data using the program DIANA with redundant dihedral angle constraints. *J. Biomol. NMR* 1991;1:446–456.
- [37]. Jain A, Vaidehi N, Rodriguez G. A fast recursive algorithm for molecular dynamics simulation. *J. Comp. Phys* 1993;106:258–268.
- [38]. Güntert P, Mumenthaler C, Wüthrich K. Torsion angle dynamics for NMR structure calculation with the new program DYANA. *J. Mol. Biol* 1997;273:283–298. [PubMed: 9367762]
- [39]. Stein EG, Rice LM, Brünger AT. Torsion angle dynamics as a new efficient tool for NMR structure calculation. *J. Magn. Reson* 1997;124:154–164. [PubMed: 9424305]
- [40]. Güntert P. Structure calculation of biological macromolecules from NMR data. *Q. Rev. Biophys* 1998;31:145–237. [PubMed: 9794034]
- [41]. Güntert P, Wüthrich K. Sample of conformational space in torsion angle dynamics calculations. *Comp. Phys. Comm* 2001;138:155–169.
- [42]. Némethy G, Gibson KD, Palmer KA, Yoon CN, Paterlini G, Zagari A, Rumsey S, Scheraga HA. Energy parameters in polypeptides. 10. Improved geometrical parameters and nonbonded interactions for use in the ECEPP/3 algorithm, with application to proline-containing peptides. *J. Phys. Chem* 1992;96:6472–6484.
- [43]. Adjiman CS, Androulakis IP, Maranas CD, Floudas CA. A global optimization method, α BB, for process design. *Comp. Chem. Eng* 1996;20:S419–S424.

- [44]. Adjiman CS, Androulakis IP, Floudas CA. Global optimization of MINLP problems in process synthesis and design. *Comp. Chem. Eng* 1997;21:S445–S450.
- [45]. Adjiman CS, Dallwig S, Floudas CA, Neumaier A. A global optimization method for general twice-differentiable NLPs. i. theoretical advances. *Comp. Chem. Eng* 1998b;22:1137–1158.
- [46]. Adjiman CS, Androulakis IP, Floudas CA. A global optimization method for general twice-differentiable NLPs. ii. implementation and computational results. *Comp. Chem. Eng* 1998a;22:1159–1179.
- [47]. Androulakis IP, Maranas CD, Floudas CA. α BB: A global optimization method for general constrained nonconvex problems. *J. Global Optim* 1995;7:337–363.
- [48]. Androulakis IP, Maranas CD, Floudas CA. Prediction of oligopeptide conformations via deterministic global optimization. *J. Global Optim* 1997;11:1–34.
- [49]. Klepeis JL, Floudas CA. Free energy calculations for peptides via deterministic global optimization. *J. Chem. Phys* 1999;110:7491–7512.
- [50]. Klepeis JL, Floudas CA. Predicting peptide structures using NMR data and deterministic global optimization. *J. Comput. Chem* 1999;20:1354–1370.
- [51]. Klepeis JL, Floudas CA. Ab initio tertiary structure prediction of proteins. *J. Global Optim* 2003b;25:113–140.
- [52]. McAllister, SR.; Floudas, CA. An improved hybrid global optimization algorithm for protein tertiary structure prediction. 2008. Submitted for publication
- [53]. Lee J, Scheraga HA, Rackovsky S. Conformational analysis of the 20-residue membrane-bound portion of melittin by conformational space annealing. *Biopolymers* 1998;46:103–115. [PubMed: 9664844]
- [54]. Lee J, Scheraga HA, Rackovsky S. New optimization method for conformational energy calculations on polypeptides : Conformational space annealing. *J. Comput. Chem* 1997;18:1222–1232.
- [55]. Lee J, Scheraga HA. Conformational space annealing by parallel computations: Extensive conformational search of met-enkephalin and the 20-residue membrane-bound portion of melittin. *Int. J. of Quant. Chem* 1999;75:255–265.
- [56]. Ripoll D, Liwo A, Scheraga HA. New developments of the electrostatically driven Monte Carlo method: Tests on the membrane-bound portion of melittin. *Biopolymers* 1998;46:117–126. [PubMed: 9664845]
- [57]. Lee J, Pillardy J, Czaplowski C, Arnautova Y, Ripoll DR, Liwo A, Gibson KD, Wawak RJ, Scheraga HA. Efficient parallel algorithms in global optimization of potential energy functions for peptides, proteins and crystals. *Comp. Phys. Comm* 2000;128:399–411.
- [58]. Desmet J, Spriet J, Lasters I. Fast and accurate side-chain topology and energy refinement as a new method for protein structure optimization. *Prot. Struct. Funct. Gen* 2002;48:31–43.
- [59]. Gill, PE.; Murray, W.; Saunders, M.; Wright, MH. NPSOL 4.0 User's Guide. Systems Optimization Laboratory, Department of Operations Research; Stanford University, CA: 1986.
- [60]. Klepeis JL, Wei YN, Hecht MH, Floudas CA. Ab initio prediction of the three-dimensional structure of a de novo designed protein: A double-blind case study. *Prot. Struct. Funct. Bioinf* 2005;58:560–570.
- [61]. Zemla A, Venclovas C, Moult J, Fidelis K. Processing and analysis of CASP3 protein structure predictions. *Prot. Struct. Funct. Gen* 1999;S3:22–29.
- [62]. Zemla A. LGA: A method for finding 3D similarities in protein structures. *Nuc. Acids Res* 2003;31:3370–3374.
- [63]. Ramachandran GN, Ramakrishnan C, Sasisekharan V. Stereochemistry of polypeptide chain configurations. *J. Mol. Biol* 1963;7:95–99. [PubMed: 13990617]
- [64]. McAllister SR, Mickus BE, Klepeis JL, Floudas CA. A novel Approach for alpha-helical topology prediction in globular proteins: generation of interhelical restraints. *Prot. Struct. Funct. Bioinf* 2006;65:930–952.
- [65]. Floudas, CA. Nonlinear and Mixed-integer Optimization: Fundamentals and Applications. Oxford University Press; 1995. 1995

Figure 1.

A Ramachandran plot [63] comparing idealized secondary structure to dihedral angle bounds on secondary structure and allowed dihedral angles for all residues except Glycine and Proline. The allowed (black, solid) and strictly allowed (black, dashed) as defined by Lovell et al. [24] are included for reference. The ideal right-handed α -helix (blue, square) and the ideal 3_{10} -helix (blue, circle) are shown compared to the bounds of Klepeis and Floudas, 2003b (blue, dotted) and Klepeis and Floudas, 2003c (blue, dashed). The ideal antiparallel β -strand (red, square) and the ideal parallel β -strand (red, circle) are shown compared to the bounds of Klepeis and Floudas, 2003b (red, dotted) and Klepeis and Floudas, 2003c (red, dashed).

Figure 2. Distribution of distance occurrences for intrahelical $i, i + 3$ distances within the PDBSelect25 dataset

Figure 3.

An illustration of a double (blue) and triple (red) contact, given the contacts that represent the β -sheet topology for Protein G

Figure 4.

An illustration of a set of single cross contacts with a separation of 5 (red) and a set of double cross contacts with an offset of 2 (blue)

Figure 5.
The constraint violation versus the RMSD from the native protein G structure (PDB:2GB1) using torsion angle dynamics and both the original (black) and new (red) bounding strategies.

Figure 6. The constraint violation versus the RMSD from the β -strands of the native protein G structure (PDB:2GB1) using torsion angle dynamics and both the original (black) and new (red) bounding strategies.

Figure 7. The ECEPP/3 potential energy of each protein conformer and its RMSD from the native protein G structure (PDB:2GB1) using both the original (black) and new (red) bounding strategies.

Table 1

Dihedral angle bounds for residues based on type and secondary structure

Residue	ϕ^L	ϕ^U	ψ^L	ψ^U
α -helix [51]	-85	-55	-50	-10
α -helix [1]	-90	-40	-60	-10
β -sheet [51]	-155	-75	110	180
β -sheet [1]	-180	-80	80	180
Loop - Glycine	25	335	-180	180
Loop - Proline	n/a	n/a	-75	195
Loop - others	-180	100	-75	195

Table 2

Bounds on distances based secondary structure

Secondary structure type	d^L	d^U
α -helix residue ($i, i + 4$)	5.5	6.5
β -sheet contact	4.5	6.5

Table 3

Lower and upper intrahelix and intrasheet distance bounds

α -helix				β -sheet			
Offset	d^L (Å)	d^U (Å)	Offset	d^L (Å)	d^U (Å)	Offset	d^L (Å)
2	5.0	6.5	2	6.0	7.2		
3	4.5	6.0	3	8.0	10.5		
4	5.5	6.5	4	10.0	14.0		
5	7.5	9.0	5	11.5	17.9		
6	9.0	10.5	6	13.0	21.0		
7	10.0	11.5	$n > 6$	$1.5 \cdot n$	$3.5 \cdot n$		
8	11.5	13.0	-	-	-		
$n > 8$	$1.3 \cdot n$	$1.8 \cdot n$	-	-	-		

Table 4Lower and upper β -contact distance bounds

Type	d^L (Å)	d^U (Å)
Single	4.2	6.5
Double	8.5	13.0
Triple	12.7	19.5

Table 5

Lower and upper single and double cross contact distance bounds

Separation	Single		Double	
	d^L (Å)	d^U (Å)	d^L (Å)	d^U (Å)
1	4.5	7.4	8.5	13.5
2	5.4	9.6	9.5	14.8
3	7.5	12.4	10.8	16.7
4	10.1	15.4	-	-
5	12.8	18.6	-	-

Table 6

The set of intrahelical distance bounds for Protein G

Residue pair	d ^L	d ^U	Residue pair	d ^L	d ^U
(23,27)	5.5	6.5	(28,32)	5.5	6.5
(24,28)	5.5	6.5	(29,33)	5.5	6.5
(25,29)	5.5	6.5	(30,34)	5.5	6.5
(26,30)	5.5	6.5	(31,35)	5.5	6.5
(27,31)	5.5	6.5			

Table 7

The set of intersheet distance bounds for Protein G

Residue pair ($\beta 1 - \beta 2$)	d^L	d^U	Residue pair ($\beta 1 - \beta 4$)	d^L	d^U	Residue pair ($\beta 3 - \beta 4$)	d^L	d^U
(2,19)	4.25	6.5	(4,51)	4.25	6.5	(42,55)	4.25	6.5
(3,18)	4.25	6.5	(5,52)	4.25	6.5	(43,54)	4.25	6.5
(4,17)	4.25	6.5	(6,53)	4.25	6.5	(44,53)	4.25	6.5
(5,16)	4.25	6.5	(7,54)	4.25	6.5	(45,52)	4.25	6.5
(6,15)	4.25	6.5	(8,55)	4.25	6.5	(46,51)	4.25	6.5
(7,14)	4.25	6.5						
(8,13)	4.25	6.5						

Table 8

Dihedral angle bounds for Protein G loop residues

Residue type	Residue positions	ϕ^L	ϕ^U	ψ^L	ψ^U
Glycine	9,38,41	25	335	-180	180
Non-Glycine	10-12,20-22,36-37,39-40,47-50	-180	100	-75	195

Table 9

The set of double and triple contact distance bounds for Protein G

Residue pair ($\beta 1 - \beta 3$)	d^L	d^U	Residue pair ($\beta 2 - \beta 4$)	d^L	d^U	Residue pair ($\beta 2 - \beta 3$)	d^L	d^U
(4,46)	8.5	13.0	(13,55)	8.5	13.0	(13,42)	13.0	19.5
(5,45)	8.5	13.0	(14,54)	8.5	13.0	(14,43)	13.0	19.5
(6,44)	8.5	13.0	(15,53)	8.5	13.0	(15,44)	13.0	19.5
(7,43)	8.5	13.0	(16,52)	8.5	13.0	(16,45)	13.0	19.5
(8,42)	8.5	13.0	(17,51)	8.5	13.0	(17,46)	13.0	19.5

Table 10

The set of single cross contact distance bounds for Protein G

Residue pair ($\beta 1 - \beta 2$)	d^L	d^U	Residue pair ($\beta 1 - \beta 4$)	d^L	d^U	Residue pair ($\beta 3 - \beta 4$)	d^L	d^U
(2,14)	12.8	18.7	(4,55)	10.1	15.4	(42,51)	10.1	15.4
(3,13)	12.8	18.7	(8,51)	10.1	15.4	(46,55)	10.1	15.4
(7,19)	12.8	18.7						
(8,18)	12.8	18.7						

Table 11

The set of double cross contact distance bounds for Protein G

Offset	Residue pair ($\beta 1 - \beta 3$)	d^L	d^U	Residue pair ($\beta 2 - \beta 4$)	d^L	d^U
1	(4,45)	9.0	13.5	(13,54)	9.0	13.5
1	(5,44)	9.0	13.5	(14,53)	9.0	13.5
1	(5,46)	9.0	13.5	(14,55)	9.0	13.5
1	(6,43)	9.0	13.5	(15,52)	9.0	13.5
1	(6,45)	9.0	13.5	(15,54)	9.0	13.5
1	(7,42)	9.0	13.5	(16,51)	9.0	13.5
1	(7,44)	9.0	13.5	(16,53)	9.0	13.5
1	(8,43)	9.0	13.5	(17,52)	9.0	13.5
2	(4,44)	9.5	14.8	(13,53)	9.5	14.8
2	(5,43)	9.5	14.8	(14,52)	9.5	14.8
2	(6,42)	9.5	14.8	(15,51)	9.5	14.8
2	(6,46)	9.5	14.8	(15,55)	9.5	14.8
2	(7,45)	9.5	14.8	(16,54)	9.5	14.8
2	(8,44)	9.5	14.8	(17,53)	9.5	14.8
3	(4,43)	10.8	16.7	(13,52)	10.8	16.7
3	(5,42)	10.8	16.7	(14,51)	10.8	16.7
3	(7,46)	10.8	16.7	(16,55)	10.8	16.7
3	(8,45)	10.8	16.7	(17,54)	10.8	16.7

Table 12

A comparison of the ensembles generated for protein G using the original and new bounding strategies with the hybrid global optimization algorithm for protein tertiary structure prediction

Evaluation metric	Original	New
Lowest Energy (kcal/mol)	−419.386	−426.539
RMSD of Lowest Energy (Å)	3.26	3.48
Best GDT-TS Score	70.54	72.77
Lowest RMSD (Å)	2.62	2.24
Distinct Conformers, E<0	20064	19066
Distinct Conformers, RMSD<3.0Å	10	159
Distinct Conformers, GDT-TS>65	13	68