# Fidelity of the Protein Structure Reconstruction from Inter-Residue Proximity Constraints

**Yiwen Chen,[†,‡] Feng Ding,[†] and Nikolay V. Dokholyan*[,†]**

*Department of Biochemistry and Biophysics, School of Medicine, and Department of Physics and Astronomy, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599*

*Received: December 27, 2006; In Final Form: April 10, 2007*

Inter-residue proximity constraints obtained in such experiments as cross-linking/mass spectrometry are important sources of information for protein structure determination. A central question in structure determination using these constraints is, What is the minimal number of inter-residue constraints needed to determine the fold of a protein? It is also unknown how the different structural aspects of constraints differentiate their ability in determining the native fold and whether there is a rational strategy for selecting constraints that feature higher fidelity in structure determination. To shed light on these questions, we study the fidelity of protein fold determination using theoretical inter-residue proximity constraints derived from protein native structures and the effect of various subsets of such constraints on fold determination. We show that approximately 70% randomly selected constraints are sufficient for determining the fold of a domain (with an average root-mean-square deviation of $\leq$3.4 Å from their native structures). We find that random constraint selection often outperforms the rational strategy that predominantly favors the constraints representing global structural features. To uncover a strategy for constraint selection for the optimal structure determination, we study the role of the topological properties of these constraints. Interestingly, we do not observe any correlation between various simple topological properties of the selected constraints, emphasizing different *global* and *local* structural features, and the performance of these constraints, suggesting that accurate protein structure determination relies on a composite of *global* and *local* structural information.

## Introduction

Structure determination of a protein is traditionally accomplished by X-ray crystallography or nuclear magnetic resonance (NMR) based on inter-proton nuclear Overhauser enhancement (NOE). At the cost of months or years of laborious work, these methods can produce high-resolution structures at the atomic level. The structures of many proteins or protein complexes cannot be determined using these methods because of the specific physical and chemical properties of these proteins or complexes. On the other hand, the knowledge of these protein structures is very important to elucidate their biological functions due to the close relationship between structure and function. Therefore, there is a pressing need for developing new methods to both increase the speed and broaden the target spectrum of protein structure determination with the rapid growth of the number of the identified proteins from genomic[1,2] and proteomic studies.[3]

Emerging experimental methods, such as intramolecular cross-linking[4,5] coupled with mass spectrometry (MS),[6−9] fluorescence resonance energy transfer (FRET),[10,11] electron paramagnetic resonance (EPR),[12,13] and paramagnetic relaxation enhancement (PRE)[14,15] allow for the determination of inter-residue proximity constraints, i.e., the typical distance separation range between two residues. These constraints have been used for fold identification,[16,17] as well as structure determination when other information is combined.[18,19] With as few as 18 intramolecular constraints derived from cross-linking and MS experiments, Young et al.[17] identified the fold of the bovine basic fibroblast growth factor (FGF)-2 by selecting the structures consistent with the constraints from a pool of models generated by threading program. With the constraints obtained by PRE[19] and the known secondary structure constraints obtained from analysis of the nuclear Overhauser enhancement spectroscopy (NOSEY) data, the barnase structure was calculated with backbone root-mean-square-deviation (rmsd) less than 3 Å from the crystal structure.

In practice, the number of inter-residue proximity constraints is often limited. Therefore, an important question is, What is the minimal number of inter-residue constraints needed to determine the fold of a protein? Both analytical[20] and computational[21] efforts have been devoted to addressing related problems in early studies in the context of general polymer theory and lattice protein models. The simplifications of polymer/protein models used in these studies are not directly extendable to real proteins. Several other studies[22−24] used more realistic protein models and offered significant insights into the relationship between protein structure and geometric constraints. However, *a priori* known secondary structures of protein were assumed in these studies, and this assumption significantly limited their applications. More recent studies pioneered the combination of *de novo* structure prediction methods utilizing knowledge-based force field with limited NMR constraints to facilitate protein structure determination.[25−28] These studies exemplified that the combination of a well-developed knowledge-based force field and experimental constraints is able to greatly improve the efficiency of structure determination. Here, we aim to offer a more general insight into the problem regarding a minimal number of inter-residue constraints required for deter-

* Corresponding author. Nikolay V. Dokholyan, Department of Biochemistry and Biophysics, University of North Carolina at Chapel Hill, North Carolina 27599. Tel.: 919-843-2513. Fax: 919-966-2852. E-mail: dokh@med.unc.edu.

[†] Department of Biochemistry and Biophysics, School of Medicine.

[‡] Department of Physics and Astronomy.

10.1021/jp068963t CCC: $37.00 
Published on Web 06/02/2007

mining protein fold. First, we treat all distance constraints equally, irrespective of the secondary or tertiary constraints and, therefore, have no *a priori* assumption about the constraints sets. Second, other than the distance constraints, bonded terms for chain connectivity, and steric exclusions between atoms, we do not apply any external force field when determining protein structure, thereby ensuring the independence of the results on any specific force field. Therefore, we are able to obtain general insights into this problem with these minimal assumptions.

Other important questions are how the distinct structural features of constraints differentiate their ability in determining the native fold and whether there is a rational strategy to select the inter-residue constraints that feature higher fidelity in structure determination. In such approaches as cross-linking/MS, FRET, EPR, and PRE, one can in principle choose various sets of constraints by engineering the chemically active residues into different positions in the protein. Thus, if such a strategy for constraint selection exists, it can help decide what constraints best determine the protein structure.

We perform discrete molecular dynamics (DMD)[29−32] of eleven structurally diverse protein domains subject to various sets of inter-residue proximity constraints and generate the corresponding structural ensembles. The typical inter-residue constraints obtained from experiments only contain the information of the upper bound within which two atoms are distant from each other. Within the upper distance bound, the exact distance between two atoms is undetermined. In different types of experiments, this upper distance bound can vary between ∼3 Å and >20 Å. Here, we do not consider the constraints with a large upper distance bound (>10 Å), since such constraints are subject to larger uncertainties than more proximal constraints, and, consequently, without additional more precise structural information, these large separation constraints are not useful in structure determination. For simplicity, we focus on the inter-residue constraints with a uniform upper distance bound 7.5 Å between $C_\beta$ atoms ($C_\alpha$ for Gly). This upper distance bound is usually used to define the contact map of a protein structure, from which the native structure can be faithfully determined.[29] We study the dependencies of rmsd (from the native structure) of constructed structural ensembles on varying numbers of randomly selected constraints. We also attempt to identify rational strategies for selecting the constraints that result in higher fidelity in structure determination. A feasible rational strategy requires a quantitative measure that can distinguish the performance of constraints in structure determination. Thus, we calculate several topological properties of the selected constraints and test whether these properties dictate the performance of these constraints.

## Methods

**DMD Simulation and Four-Bead Protein Model.** We perform DMD[30−33] simulations using a four-bead protein model,[34] in which each residue is represented by three backbone beads N, C, $C_\alpha$, and one side-chain bead $C_\beta$ (only $C_\alpha$ for Gly). The detailed implementation of the covalent bonds and constraints that maintain the correct geometry of each residue and the peptide connectivity is described by Ding et al.[34]

**Clustering Methods.** Clustering is the procedure that categorizes or groups similar entities together on the basis of quantitative distance (similarity) measures. To perform clustering of structures in the present study, we define the distance between two structures as their mutual rmsd. The smaller the rmsd from each other, the higher the similarity there is between two structures. We perform hierarchical agglomerative clustering[35] by first finding the two entities that have the minimal distance between them. After joining those two entities into a cluster, the method then searches for the minimal distance between two entities, but taking those entities that have already been clustered as a single unit. This process is repeated until there are no more entities to cluster. The process of the hierarchical agglomerative clustering is summarized in a treelike diagram, called a dendrogram.[35] The root of the dendrogram, which is at the top (zeroth) level, is one cluster containing all the entities (structures). The leaf nodes at the bottom level of the dendrogram correspond to isolated entities before clustering. From the top to the bottom level of the dendrogram, the number of emerging clusters increases. There are three different hierarchical agglomerative clustering methods applied: single linkage, complete linkage, and average linkage. In single linkage, the minimal distance between members of the two clusters is taken as the cluster distance. In complete linkage, the maximal distance between members of clusters is taken. In average linkage, the average distance between members in the clusters is taken. All the clustering in this work is performed using the program OC.[36]

**Selection of Constraints from Contact Map.** A protein contact map is a set of contacts defined as follows: if the distance between $C_\beta$ atoms ($C_\alpha$ for Gly) from two residues $i$ and $j$ in the native structure is within 7.5 Å, residues $i$ and $j$ are considered to form a native contact. Here, an inter-residue proximity constraint corresponds to a native contact in the contact map. We will use native contacts and constraints interchangeably hereafter. For each native contact between residues $i$ and $j$, the contact distance is defined as $|i - j|$. We exclude the constraints with the contact distance $|i - j| \leq 2$, since these constraints are mainly defined by polypeptide connectivity.

The contact order[37] of a set of constraints is defined as the average $|i - j|$ taken over all constraints in the set. In simulations, we use different strategies for selecting constraints from the contact map. In the random selection procedure, the constraints are randomly selected from the contact map. In the contact order ranked (COR) method, the constraints are selected from the contact map sequentially according to the descending order of the corresponding $|i - j|$ values ($i, j = 1, 2, ..., N$).

**Generation of Coarse-Grained Structural Ensembles Satisfying a Given Set of Constraints.** We incorporate a given set of inter-residue constraints into the simulations as effective interactions between $C_\beta$ atoms ($C_\alpha$ for Gly) of corresponding residues

$$U_{ij} = \begin{cases} +\infty, |\mathbf{r}_i - \mathbf{r}_j| \leq a_0 \\ -\Delta_{ij}, a_0 < |\mathbf{r}_i - \mathbf{r}_j| \leq a_1 \\ 0, |r_i - r_j| > a_1 \end{cases} \quad (1)$$

where $a_0$ is the hard core diameter, and $a_1$ is the upper bound of distance constraints, which is 7.5 Å. All $\mathbf{r}_i$ and $\mathbf{r}_j$ are the Cartesian coordinates of $C_\beta$ atom ($C_\alpha$ for Gly) of $i$th and $j$th residue, respectively. $||\Delta_{ij}||$ is a matrix with elements $\Delta_{ij} = 1$ if there is a constraint between residue $i$ and $j$, and $\Delta_{ij} = 0$ otherwise. For each set of constraints, prior to the production simulations, we perform simulations from the fully extended protein state at temperature $T = 2.0$ to $T = 0.1$ (in units of the inverse Boltzmann constant $k_B^{-1}$). Then, we perform production simulations at $T = 0.1$ for $10^5$ time units. We choose five trajectories starting from different initial conditions in which all given constraints are satisfied and there is no "mirror structure"—a structure that satisfies all given constraints, but the transformation that superimposes it with the native structure is an improper rotation.[38] The trajectories of mirror structures

**TABLE 1: Nine Protein Domains**

| length | PDB code | CATH name | CATH code | class(C) |
|---|---|---|---|---|
| 60 | 1GO3(48:107) | DNA-directed RNA polymerase subunit f | 1.10.150.80 | mainly-α |
| 49 | 1DD3(1:49) | 50s ribosomal protein 17/112 | 1.20.58.20 | mainly-α |
| 62 | 1NXB(1:62) | neurotoxin b | 2.10.60.10 | mainly-$\beta$ |
| 60 | 1VIE(19:78) | dihydrofolate reductase | 2.30.30.60 | mainly-$\beta$ |
| 60 | 1JDC(358:417) | 1,4-α maltotetrahydrolase | 2.60.40.1180 | mainly-$\beta$ |
| 61 | 1IGD(1:61) | protein g | 3.10.20.10 | α-$\beta$ |
| 60 | 1BXY(1:60) | ribosomal protein 130 | 3.30.70.700 | α-$\beta$ |
| 65 | 1E4F(237:301) | cell division protein ftsa | 3.90.640.10 | α-$\beta$ |
| 60 | 1D0D (1:60) | anticoagulant protein | 4.10.410.10 | few secondary structures |

are automatically eliminated by clustering the structures from trajectories that satisfy all the given constraints (clustering methods). Regardless of the different clustering methods applied, the "mirror structures" always appear as a distinct cluster at the first level of dendrograms[35] (clustering methods). In coarse-grained representation, the "mirror structures" have the same energy as the native ones, since they satisfy the same set of constraints. By contrast, we found that the energies of the reconstructed all-atom models of the "mirror structures" tend to be higher than the native ones that satisfy the same set of constraints. The details of the all-atom reconstruction and the energy function used for calculation is described in a recent work by Ding and Dokholyan.[39] Therefore, we can recognize "mirror structures" without *a priori* knowledge of the native structure.

**Topological Properties of Constraints in the Contact Map Graph.** We construct the contact map graph by representing each residue as a node and each constraint in the contact map as an edge. Each set of constraints represents a subgraph of the contact map graph. There are four topological properties of constraint sets in the graph inspected in the current study.

The shortest path[40] between two nodes $i$ and $j$ in the network are defined as the paths traversing a minimal number of edges among all paths connecting $i$ and $j$. The shortest path length between two nodes is the number of edges traversed by the shortest paths. The shortest path length of a set of constraints is the average of shortest path lengths between all nodes in the set.

The betweenness centrality $C_B(l)$ of a node or an edge $l$ in the network is defined as follows:[41]

$$C_B(l) = \sum_{s \neq t \in V} \frac{\sigma_{st}(l)}{\sigma_{st}} \tag{2}$$

where $\sigma_{st}$ is the total number of shortest paths connecting node $s$ and $t$; $\sigma_{st}(l)$ is the number of shortest paths connecting node $s$ and $t$ that pass through the node/edge $l$. The betweenness centrality of a set of constraints is the average of betweenness centrality taken over the set.

The clustering coefficient of a node[42,43] $i$ in the network is defined as the ratio between the number $E_i$ of edges that actually exist between all $k_i$ nodes directly connected to node $i$, and the maximal possible number $k_i(k_i - 1)/2$ of edges between these $k_i$ nodes

$$C_i = \frac{2E_i}{k_i(k_i - 1)} \tag{3}$$

The clustering coefficient of a set of constraints is the average of clustering coefficients taken over all nodes in the set.

The degree of a node $i$[40] in the network is defined as total number of nodes having edges directly connected to it. The degree of a set of constraints is the average of degrees taken over all nodes in the set.

## Results and Discussion

We study nine protein domains chosen from the CATH protein structure classification database.[44] The CATH database is a hierarchical domain classification of protein structures in the PDB. There are four major "top to bottom" levels of classification of protein structures in the CATH database, which are class, architecture, topology, and homologous superfamilies. For the broad coverage of the structural space of all proteins, we choose nine protein domains representing all categories: mainly-α, mainly-$\beta$, α−$\beta$, and few secondary structures (Table 1) at the top level class in CATH. To minimize the length dependence of the result, all protein domains are chosen to be approximately 60 residues long. For each protein domain, we first determine the contact map based on its native structure (Methods section). The contact map contains all inter-residue proximity constraints with a cutoff distance of 7.5 Å between $C_\beta$ atoms ($C_\alpha$ for Gly). For each fraction of constraints, 10%, 30%, 50%, 70%, and 90%, we then generate six constraint sets with the corresponding number of constraints selected from the contact map. One set is generated by the COR method, and the other five are by a random selection (Methods section). Then, we perform DMD simulations of the simplified protein model (Methods section) to generate the coarse-grained conformational ensembles satisfying these constraint sets.

**Approximately 70% of all proximity constraints** derived from the native structures are sufficient to determine the protein folds within an average rmsd of 3.4 Å. We determine the average and standard deviation of rmsd of the structural ensembles, subject to five randomly selected constraint sets, as the function of the fraction of constraints applied (Figure 1). For each fraction of constraints, 10%, 30%, 50%, 70%, and 90%, in Figure 1 we only show the largest and the smallest average rmsd with their corresponding standard deviations. We observe a sharp decrease of the average rmsd as the fraction of constraints increases. Although the exact fraction at which this cooperative transition occurs varies for different domains, this transition typically occurs at less than ∼30% fraction of constraints. For the protein domains under study, the 30% fraction of constraints corresponds to $0.65 \pm 0.05$ constraints per residue (the number of constraints applied per length of a domain). In addition, we find that 70% fraction of all proximity constraints derived from the native structure are sufficient to determine the fold of a domain with an average rmsd of ≤3.4 Å. The 70% fraction of constraints corresponds to ∼90 constraints for a 60 residue protein and $1.51 \pm 0.11$ constraints per residue. Noticeably, for different folds, the minimal fraction of constraints required to determine the native fold varies significantly (Figure 1), reflecting unique topological characteristics of distinct folds. For example, for protein domains 1GO3 (mainly-α), 1NXB (mainly-$\beta$), 1VIE (mainly-$\beta$), 1BXY

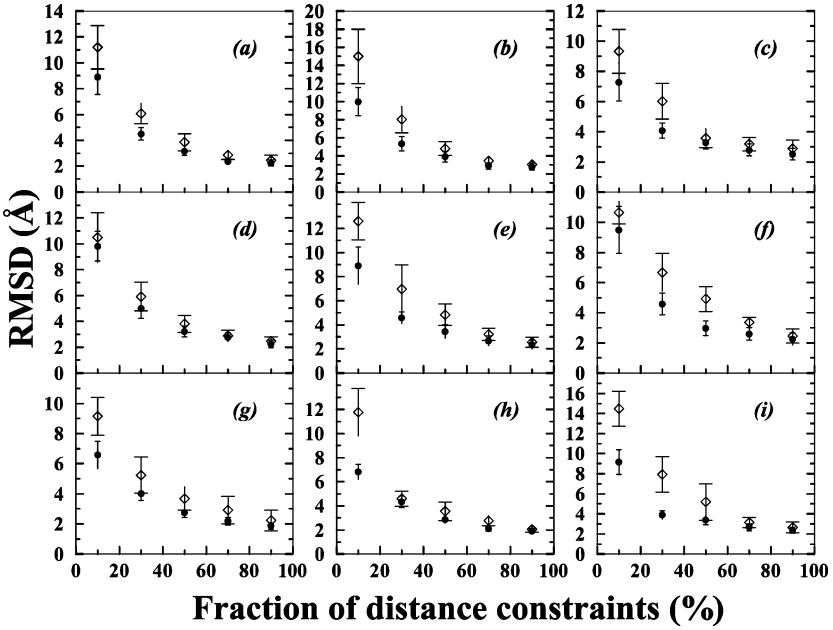


**Figure 1.** The average and the standard deviation of rmsd of the structural ensembles subject to 10%, 30%, 50%, 70%, and 90% randomly selected constraints for nine domains: (a) 1GO3 (48−107), (b) 1DD3 (1−49), (c) 1NXB (1−62), (d) 1VIE (19−78), (e) 1JDC (358−417), (f) 1IGD (1−61), (g) 1BXY (1−60), (h) 1E4F (237−301), (i) 1D0D (1−60). For each fraction of constraints, we only show the largest (◇) and smallest (■) average rmsd and their corresponding standard deviations (out of five ensembles).

($\alpha-\beta$), and 1E4F ($\alpha-\beta$), 50% fraction constraints are sufficient to determine the structural ensembles within an average rmsd of 4 Å from the native structures, while it is not the case for other domains. Therefore, 70% is only a conservative estimate.

According to the theoretical work by Reva et al.[45] and others,[46] the rmsd distribution for a ∼60 residue protein with randomly selected/constructed globular proteinlike structure is Gaussian with the average value of 11 Å and the standard deviation of 2 Å. Therefore, the probability of observing a structure within 3.4 Å rmsd from the native structure by chance is less than $7 \times 10^{-5}$. It is statistically significant to conclude that a structural ensemble satisfying 70% native contacts is close to the native structure.

The current study focuses on the relatively small domains as test cases. For larger domains, the total number of the constraints required for determining the native fold will increase accordingly. To see whether the current results can be extrapolated to larger domains, we study two other domains with the length of approximately 105 residues (Supporting Information). We find that, for these two domains, 70% randomly selected native contacts are also sufficient to determine the native folds (Supporting Information). Therefore, we expect that our results will also hold for larger single-domain proteins. However, it remains to uncover in future studies to what extent our result can be extended to the case of multidomain proteins.

Here, we do not study how the cutoff distance, which is used to define constraints, affects the quality of the determined structure. This question is of practical importance, because in cross-linking/mass spectrometry experiments, cross-linkers of various lengths[47,48] may be used in protein structure determination. An early computational study by Vendruscolo et al.[49] suggested that the quality of the determined structure is significantly degraded if the cutoff distance used to define constraints is below a certain threshold. Therefore, in cross-linking/mass spectrometry experiments, both the number of constraints and the length of cross-linkers used in protein structure determination have a significant impact on the quality of the determined structure.

**Random constraint selection** often outperforms rational contact order-based selection strategy. For each fraction of constraints, 10%, 30%, 50%, 70%, and 90%, we find that the structural ensembles subject to six constraint sets (five randomly selected sets and one selected by the COR method) have distinct rmsd from the native structure. For example, in the case of the protein domain 1E4F (237−301) (Figure 1h), the structural ensembles, subject to various constraint sets utilizing 10% of the constraints, have an average rmsd ranging from 7 to 12 Å. In 1E4F (237−301), 10% of the constraints corresponds to thirteen constraints. These results indicate that, using a sparse number of constraints (∼13), it is possible to reconstruct a structural ensemble with an average rmsd of 7 Å. When other information, such as homology or secondary structure information, is incorporated, it is not surprising that the structure ensembles can be determined with higher accuracy.[17,19] The large performance variation of different constraint sets also suggests that proper selections of constraints can significantly improve the efficiency of structure determination.

The constraint sets that have large values of contact order (i.e., have more long-range constraints) offer more information about global rather than local structural properties. Correspondingly, the constraint sets that have small values of the contact order offer more information about local rather than global structural protein properties. Intuitively, global structural information plays a major role in determining protein folds. Hence, we speculate that the constraint sets containing more long-range constraints lead to structural ensembles with smaller rmsd from the native structure. Since the separation of residues in the constraint sets can be quantified by contact order (Methods section), we develop a rational strategy, COR method (Methods section), for constraint selection to predominantly favor long-range constraints. Surprisingly, we find that the conformational ensembles, corresponding to the constraint sets selected by the COR method, often have larger rmsd than those ensembles corresponding to the constraint sets selected randomly (Figure 2). Clearly, randomly selected sets have a broader distribution of contact distance than the sets selected by the COR method

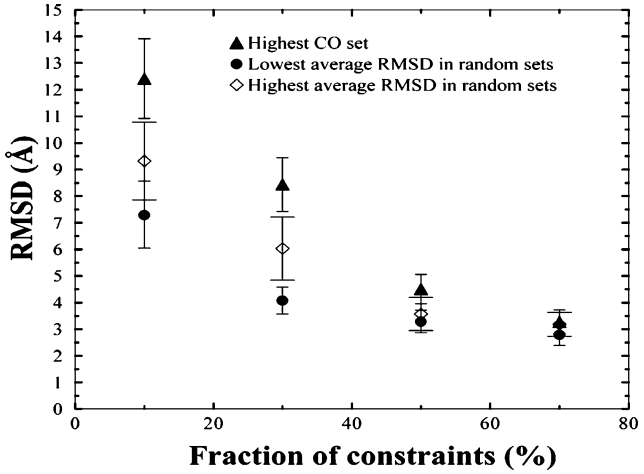


**Figure 2.** Neurotoxin b (PDB code: 1NXB). The average and standard deviation of rmsd of the structural ensembles subject to 10%, 30%, 50%, 70%, and 90% constraints selected by the contact order ranked method and random strategy: triangle (▲), the structural ensembles subject to constraint sets selected by the contact order-ranked method (CO); diamond (◇), the highest average rmsd and corresponding standard deviation of the structural ensembles subject to randomly selected constraint sets; filled circle (●), the lowest average rmsd and corresponding standard deviation of the structural ensembles subject to randomly selected constraint sets.

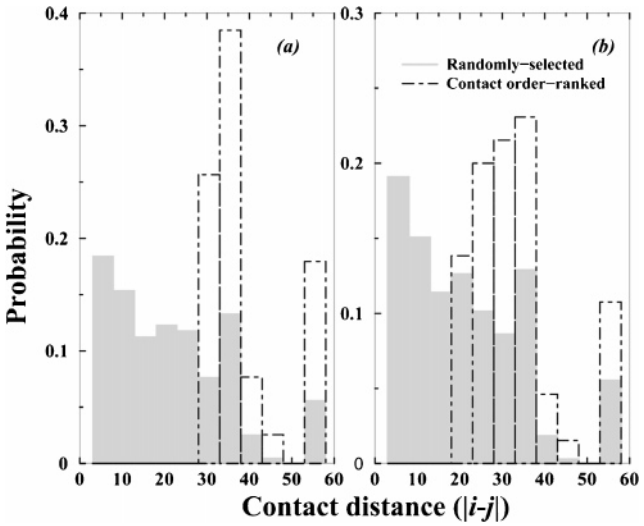


**Figure 3.** Neurotoxin b (PDB code: 1NXB). The histograms of contact distance ($|i - j|$) for (a) 30% constraints selected by random strategy (gray) and the contact order ranked method (dot-dashed); (b) 50% constraints selected by random strategy (gray) and the contact order ranked method (dot-dashed).

(Figure 3), which is because random selection tends to include a mixture of short-range and long-range constraints, while the COR method favors long-range constraints. The observed difference of the performance between random selection and the COR method suggests that accurate structure determination depends on a blend of global and local structural information rather than global structural information alone.

**Simple topological properties of the selected constraints** do not correlate with the performance of these constraints in structure determination. To explore other potential factors of structure determination performance, we examine several topological properties of constraint sets in the contact map graph (Methods section), where each residue is represented as a node and each constraint is represented as an edge. The spatial relationships between the residues in the protein structure are mapped to the connections between the nodes in the contact map graph. Each set of selected constraints represents a subgraph of the contact map graph. First, we study the correlation between the edge betweenness centrality (Methods section) of a constraint set and the performance of this set in structure determination. The betweenness centrality of an edge measures how often this edge lies on the shortest path between all node pairs in the graph and indicates its significance in the connectivity in the whole graph.[41] The higher betweenness centrality value an edge has, the more important it is in connecting the whole graph. We expect that the constraints with larger betweenness centrality contain more structural information on the spatial connection between different elements of the protein structure and, thus, have more prominent effects on protein structure determination. Although the betweenness centrality is a sensitive measure of network topology,[50] we do not observe a correlation between the edge betweenness centrality and the performance of a constraint set (Table 2). For example, in the case of domain 1NXB (1−62) a constraint set with a larger average edge betweenness centrality leads to an ensemble with smaller average rmsd, while in the case of domain 1IGD (1−61) (Table 2), a constraint set with a larger average edge betweenness centrality leads to an ensemble with larger average rmsd.

**TABLE 2: Differences in the Edge Betweenness Centrality between the Constraint Sets Corresponding to the Ensembles with the Largest Average Rmsd and the Constraint Sets Corresponding to the Ensembles with the Smallest Average Rmsd among the Randomly Selected Set[a]**

| constraint sets (random) | 1NX8 (10%) | 1IGD (10%) | 1NX8 (30%) | 1IGD (30%) |
|---|---|---|---|---|
| largest average rmsd | 18.86 | 26.58 | 24.50 | 24.72 |
| smallest average rmsd | 23.37 | 19.46 | 26.18 | 22.04 |

[a] For illustration, we only show the result of the constraint sets containing 10% and 30% constraints for two domains (1NXB (1−62), 1IGD (1−61)), respectively.

It was found in protein folding studies that certain residues, called key residues,[29,51−53] are critical for forming protein folding nucleus[52] and play essential roles in folding kinetics. Given the important roles of these residues in determining overall native topology during protein folding, we test whether the constraints between these key residues play more important roles in structural determinations than other constraints for the case of domain 1VIE (19−78). In comparison with other residues in a protein, the key residues have larger node betweenness centrality in the contact map graphs of both the native state and folding transition state.[51] Node betweenness centrality was shown to predict key residues in transition states,[51] because there are much fewer nonkey residues that have a relatively high value of node betweenness centrality in the transition state compared with the native state. Thus, since there is no experimental structural information on transition states of the proteins studied here, we use node betweenness centrality of a residue in the native state as a proxy of its importance in folding kinetics. We first rank all residues in domain 1VIE (19−78) by their node betweenness centrality values in descending order. We then choose constraints between the residues with higher ranks and use the selected constraints in simulations to determine the domain's structural ensemble. We find that this procedure has improved performance than the COR method in structure determination, but it has comparable performance to random selection strategy. We use rmsd to measure the structural similarity between the determined structures and the native state (Table 3). We also employ an alternative measure, the Q-value (the fraction of native contacts formed in a given structure),[54] to evaluate how structurally similar the determined structures are to the native state: a larger Q-value[54] suggest higher structural similarity to

**TABLE 3: Average and Standard Deviation of the rmsd and the Q-Value[54] of the Structural Ensembles of the Domain 1VIE (19−78) Determined Using 30% of All Constraints That Are Chosen by Random Selection, between Key Residues,[29,51] and by COR Method[a]**

| 30% constraints | random selection (largest rmsd) | random selection (smallest rmsd) | between key residues | COR method |
|---|---|---|---|---|
| rmsd (Å) | 5.92 ± 1.11 | 5.00 ± 0.76 | 5.86 ± 0.79 | 7.36 ± 1.54 |
| Q-value | 0.46 ± 0.07 | 0.54 ± 0.06 | 0.44 ± 0.06 | 0.40 ± 0.05 |

[a] For the ensembles determined using randomly selected constraints, we only show the largest and the smallest average rmsd, their corresponding standard deviations, and the Q-value of the corresponding ensembles.

the native state. We find that the structures determined using constraints between kinetically residues have larger Q-values than the ones determined using the COR method (Table 3). These results suggest that, although selecting constraints between kinetically important residues is a better strategy than the COR method, it does not show superior performance to random constraint selection in structure determination.

In addition, we study the correlation between the performance of a given constraint set and other topological properties, such as the node degree, the clustering coefficient, and the shortest path length[55] (Methods section). However, similar to the case of edge betweenness centrality, we do not observe any correlation between any of these topological properties and the performance of constraint sets (data not shown).

These findings indicate that the rmsd of a conformation to the native structure is a multivariable function of different topological properties of constraints. Therefore, the accurate structure determination requires a combination of constraints with composite structural features, which are not characterized by any single topological property.

## Conclusions

For eleven structurally diverse protein domains, we have shown that approximately 70% of all proximity constraints derived from their native structures are sufficient for determining the fold of domains with an average rmsd of ≤3.4 Å. This finding by no means offers a comprehensive answer to the question, what is the minimal number of inter-residue proximity constraints needed to determine the fold of a protein? Rather, it offers a theoretical estimation of the lower bound of this minimal number. We believe that this estimation is an important starting point for further studies. It is important to note that in the current study we do not consider any experimental errors of constraints in the simulations. It is expected that different types of experimental errors can have distinct effects on the determined protein structure, and these effects deserve further study.

In addition, we find that randomly selected constraints often outperform the constraints representing global structural features, suggesting that both local and global structural features are important in determining the fold of a protein. Further, we do not observe any correlation between various topological properties of the selected constraints, emphasizing different structural features and the performance of these constraints. Both these findings indicate that a rational strategy based on a quantitative measure that can distinguish the performance of constraints in structure determination is yet to be developed. Due to the significant complexity inherent in the mapping from constraints to the protein structures, more work is needed to understand how to build a minimum set of structure determining constraints.

**Acknowledgment.** We thank Sagar D. Khare, Deanne Sammond, and Kyle Wilcox for suggestions on the manuscript. This work is supported in part by the Muscular Dystrophy Association grant MDA3720, research grant no. 5-FY03-155 from the March of Dimes Birth Defect Foundation, the American Heart Association grant no. 0665361U, and the North Carolina Biotechnology Center grant no. 2006-MRG-1107 to N.V.D. Simulations are performed by utilizing computational resources provided by National Partnership for Advanced Computational Infrastructure (NPACI) grant MCB040055.

**Supporting Information Available:** The descriptions (Table S1) and simulation results (Figure S1) of two larger protein domains included in the study. This material is available free of charge via the Internet at http://pubs.acs.org.

## References and Notes

(1) Stoesser, G.; Tuli, M. A.; Lopez, R.; Sterk, P. *Nucleic Acids Res.* **1999**, *27*, 18−24.
(2) Benson, D. A.; Boguski, M. S.; Lipman, D. J.; Ostell, J.; Ouellette, B. F. F.; Rapp, B. A.; Wheeler, D. L. *Nucleic Acids Res.* **1999**, *27*, 12−17.
(3) Zhu, H.; Bilgin, M.; Snyder, M. *Annu. Rev. Biochem.* **2003**, *72*, 783−812.
(4) Cohen, F. E.; Sternberg, M. J. *J. Mol. Biol.* **1980**, *137*, 9−22.
(5) Swaney, J. B. *Methods Enzymol.* **1986**, *128*, 613−626.
(6) Fenn, J. B.; Mann, M.; Meng, C. K.; Wong, S. F.; Whitehouse, C. M. *Science* **1989**, *246*, 64−71.
(7) McLafferty, F. W.; Fridriksson, E. K.; Horn, D. M.; Lewis, M. A.; Zubarev, R. A. *Science* **1999**, *284*, 1289−1290.
(8) Qin, J.; Chait, B. T. *Anal. Chem.* **1996**, *68*, 2108−2112.
(9) Wang, R.; Chait, B. T. *Curr. Opin. Biotechnol.* **1994**, *5*, 77−84.
(10) Dong, W. J.; Xing, J.; Chandra, M.; Solaro, J.; Cheung, H. C. *Proteins* **2000**, *41*, 438−447.
(11) Cardullo, R. A.; Parpura, V. *Methods Cell Biol.* **2003**, *72*, 415−430.
(12) Hubbell, W. L.; Gross, A.; Langen, R.; Lietzow, M. A. *Curr. Opin. Struct. Biol.* **1998**, *8*, 649−656.
(13) Lakshmi, K. V.; Brudvig, G. W. *Curr. Opin. Struct. Biol.* **2001**, *11*, 523−531.
(14) Gillespie, J. R.; Shortle, D. *J. Mol. Biol.* **1997**, *268*, 170−184.
(15) Gillespie, J. R.; Shortle, D. *J. Mol. Biol.* **1997**, *268*, 158−169.
(16) Albrecht, M.; Hanisch, D.; Zimmer, R.; Lengauer, T. *In Silico Biol.* **2002**, *2*, 325−337.
(17) Young, M. M.; Tang, N.; Hempel, J. C.; Oshiro, C. M.; Taylor, E. W.; Kuntz, I. D.; Gibson, B. W.; Dollinger, G. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 5802−5806.
(18) Perozo, E.; Cortes, D. M.; Cuello, L. G. *Nat. Struct. Biol.* **1998**, *5*, 459−469.
(19) Gaponenko, V.; Howarth, J. W.; Columbus, L.; Gasmi-Seabrook, G.; Yuan, J.; Hubbell, W. L.; Rosevear, P. R. *Protein Sci.* **2000**, *9*, 302−309.
(20) Gutin, A. M.; Shakhnovich, E. I. *J. Chem. Phys.* **1994**, *100*, 5290−5293.
(21) Dewitte, R. S.; Michnick, S. W.; Shakhnovich, E. I. *Protein Sci.* **1995**, *4*, 1780−1791.
(22) Aszodi, A.; Gradwell, M. J.; Taylor, W. R. *J. Mol. Biol.* **1995**, *251*, 308−326.
(23) Smithbrown, M. J.; Kominos, D.; Levy, R. M. *Protein Eng.* **1993**, *6*, 605−614.
(24) Lund, O.; Hansen, J.; Brunak, S.; Bohr, J. *Protein Sci.* **1996**, *5*, 2217−2225.
(25) Li, W.; Zhang, Y.; Kihara, D.; Huang, Y. P. J.; Zheng, D. Y.; Montelione, G. T.; Kolinski, A.; Skolnick, J. *Proteins: Struct., Funct., Genet.* **2003**, *53*, 290−306.
(26) Meiler, J.; Baker, D. *J. Magn. Reson.* **2005**, *173*, 310−316.
(27) Meiler, J.; Baker, D. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 15404−15409.
(28) Skolnick, J.; Kolinski, A.; Ortiz, A. R. *J. Mol. Biol.* **1997**, *265*, 217−241.
(29) Dokholyan, N. V.; Borreguero, J. M.; Buldyrev, S. V.; Ding, F.; Stanley, H. E.; Shakhnovich, E. I. *Methods Enzymol.* **2003**, *374*, 616−638.
(30) Dokholyan, N. V.; Buldyrev, S. V.; Stanley, H. E.; Shakhnovich, E. I. *Folding Des.* **1998**, *3*, 577−587.
(31) Smith, S. W.; Hall, C. K.; Freeman, B. D. *J. Comput. Phys.* **1997**, *134*, 16−30.
(32) Zhou, Y. Q.; Karplus, M. *Mol. Phys.* **1996**, *89*, 1707−1717.
(33) Rapaport, D. C. *The art of molecular dynamics simulation*; Cambridge University Press: Cambridge, 1997.
(34) Ding, F.; Borreguero, J. M.; Buldyrey, S. V.; Stanley, H. E.; Dokholyan, N. V. *Proteins: Struct., Funct., Genet.* **2003**, *53*, 220−228.

(35) Everitt, B. S.; Landau, S.; Leese, M. *Cluster analysis*; Oxford University Press: Oxford, 2001.
(36) Barton, G. J. *OC - A cluster analysis program*; University of Dundee, Scotland, 2002.
(37) Plaxco, K. W.; Simons, K. T.; Baker, D. *J. Mol. Biol.* **1998**, *277*, 985−994.
(38) Kabsch, W. *Acta Crystallogr., Sect. A: Cryst. Phys., Diffr., Theor. Gen. Crystallogr.* **1976**, *5*, 922−923.
(39) Ding, F.; Dokholyan, N. V. *PLoS Comput. Biol.* **2006**, *2*, e85.
(40) Bollobás, B. *Modern Graph Theory*; Springer: New York, 1998.
(41) Freeman, L. *Sociometry* **1977**, *40*, 35−41.
(42) Watts, D. J.; Strogatz, S. H. *Nature (London)* **1998**, *393*, 440−442.
(43) Wasserman, S.; Faust, K. *Social network analysis: methods and applications*; Cambridge University Press: Cambridge, 1994.
(44) Orengo, C. A.; Michie, A. D.; Jones, S.; Jones, D. T.; Swindells, M. B.; Thornton, J. M. *Structure* **1997**, *5*, 1093−1108.
(45) Reva, B. A.; Finkelstein, A. V.; Skolnick, J. *Folding Des.* **1998**, *3*, 141−147.
(46) Ding, F.; Buldyrev, S. V.; Dokholyan, N. V. *Biophys. J.* **2005**, *88*, 147−155.
(47) Back, J. W.; de Jong, L.; Muijsers, A. O.; de Koster, C. G. *J. Mol. Biol.* **2003**, *331*, 303−313.
(48) Sinz, A. *J. Mass Spectrom.* **2003**, *38*, 1225−1237.
(49) Vendruscolo, M.; Kussell, E.; Domany, E. *Folding Des.* **1997**, *2*, 295−306.
(50) Dokholyan, N. V. *Gene* **2005**, *347*, 199−206.
(51) Vendruscolo, M.; Dokholyan, N. V.; Paci, E.; Karplus, M. *Phys. Rev. E* **2002**, *65*, 061910.
(52) Dokholyan, N. V.; Buldyrev, S. V.; Stanley, H. E.; Shakhnovich, E. I. *J. Mol. Biol.* **2000**, *296*, 1183−1188.
(53) Vendruscolo, M.; Paci, E.; Dobson, C. M.; Karplus, M. *Nature (London)* **2001**, *409*, 641−645.
(54) Clementi, C.; Jennings, P. A.; Onuchic, J. N. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 5871−5876.
(55) Albert, R.; Barabasi, A. L. *Rev. Mod. Phys.* **2002**, *74*, 47−97.