Protein Structure by Semidefinite Facial Reduction

Babak Alipanahi[1], Nathan Krislock[2], Ali Ghodsi[3], Henry Wolkowicz[4], Logan Donaldson[5], and Ming Li[1⋆]

[1] David R. Cheriton School of Computer Science, University of Waterloo, Waterloo, Ontario, Canada
[2] INRIA Grenoble Rhône-Alpes, France
[3] Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada
[4] Department of Combinatorics and Optimization, University of Waterloo, Waterloo, Ontario, Canada
[5] Department of Biology, York University, Toronto, Ontario, Canada

Abstract. All practical contemporary protein NMR structure determination methods use molecular dynamics coupled with a simulated annealing schedule. The objective of these methods is to minimize the error of deviating from the NOE distance constraints. However, this objective function is highly nonconvex and, consequently, difficult to optimize. Euclidean distance geometry methods based on semidefinite programming (SDP) provide a natural formulation for this problem. However, complexity of SDP solvers and ambiguous distance constraints are major challenges to this approach. The contribution of this paper is to provide a new SDP formulation of this problem that overcomes these two issues for the first time. We model the protein as a set of intersecting two- and three-dimensional cliques, then we adapt and extend a technique called semidefinite facial reduction to reduce the SDP problem size to approximately one quarter of the size of the original problem. The reduced SDP problem can not only be solved approximately 100 times faster, but is also resistant to numerical problems from having erroneous and inexact distance bounds.

Key words: Molecular structural biology, nuclear magnetic resonance

1 Introduction

Computing three-dimensional protein structures from their amino acid sequences has been one of the most widely studied problems in bioinformatics because knowing the structure of protein structure is key to understanding its physical, chemical, and biological properties. The protein nuclear magnetic resonance (NMR) method is fundamentally different from

⋆ All correspondence should be addressed to mli@uwaterloo.ca.

the X-ray method: It is not a "microscope with atomic resolution"; rather it provides a network of distance measurements between spatially proximate hydrogen atoms [12]. As a result, the NMR method relies heavily on complex computational algorithms. The existing methods for protein NMR can be categorized into four major groups: (*i*) methods based on Euclidean distance matrix completion (EDMC) [6, 15, 5, 20], (*ii*) methods based on molecular dynamics and simulated annealing [24, 8, 28, 14, 13], (*iii*) methods based on local/global optimization [7, 23, 34], and (*iv*) methods originating from sequence-based protein structure prediction algorithms [29, 25, 2].

1.1 Gram Matrix Methods

Using the Gram matrix, or the matrix of inner products, has many advantages: (*i*) The Gram matrix and Euclidean distance matrix (EDM) are linearly related to each other. (*ii*) Instead of enforcing all of the triangle inequality constraints, it is sufficient to enforce that the Gram matrix is positive semidefinite. (*iii*) The embedding dimension and the rank of the Gram matrix are directly related. Semidefinite programming (SDP) is a natural choice for formulating the problem using the Gram matrix. However, the major obstacle is the computational complexity of SDP solvers.

1.2 Contributions of the Proposed SPROS Method

Most of the existing methods make some of the following assumptions: (*i*) assuming to know the (nearly) exact distances between atoms, (*ii*) assuming to have the distances between any type of nuclei (not just hydrogens), (*iii*) ignoring the fact that not all hydrogens can be uniquely assigned, and (*iv*) overlooking the ambiguity in the NOE cross-peak assignments. In order to automate the NMR protein structure determination process, we need a robust structure calculation method that tolerates more errors. We give a new SDP formulation that does not assume (*i*–*iv*) above. Moreover, the new method, called "SPROS" (Semidefinite Programming-based Protein structure determination), models the protein molecule as a set of intersecting two- and three-dimension cliques. We adapt and extend a technique called semidefinite facial reduction which makes the SDP problem strictly feasible and reduces its size to approximately one quarter the size of the original problem. The reduced problem is more numerically stable to solve and can be solved nearly 100 times faster.

2 The SPROS Method

2.1 Euclidean Distance Geometry

Euclidean Distance Matrix A symmetric matrix D is called a Euclidean Distance Matrix (EDM) if there exists a set of points $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, $\boldsymbol{x}_i \in \mathbb{R}^r$ such that:

$$D_{ij} = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2, \quad \forall i, j. \tag{1}$$

The smallest value of r is called the *embedding dimension* of D, and is denoted $\mathbf{embdim}(D) = r$. The space of all $n \times n$ EDMs is denoted $\mathcal{E}^n$.

The Gram Matrix If we define $X := [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n] \in \mathbb{R}^{r \times n}$, then the matrix of inner-products, or *Gram Matrix*, is given by $G := X^\top X$. It immediately follows that $G \in \mathcal{S}^n_+$, where $\mathcal{S}^n_+$ is the set of symmetric positive semidefinite $n \times n$ matrices. The Gram matrix and the Euclidean distance matrix are linearly related:

$$D = \mathsf{K}(G) := \mathbf{diag}(G) \cdot \mathbf{1}^\top + \mathbf{1} \cdot \mathbf{diag}(G)^\top - 2G, \tag{2}$$

where $\mathbf{1}$ is the all-ones vector of the appropriate size. To go from the EDM to the Gram matrix, we use the $\mathsf{K}^\dagger \colon \mathcal{S}^n \to \mathcal{S}^n$ linear map:

$$G = \mathsf{K}^\dagger(D) := -\tfrac{1}{2} HDH, \quad D \in \mathcal{S}^n_H, \tag{3}$$

where $H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ is the *centering* matrix, $\mathcal{S}^n$ is the space of symmetric $n \times n$ matrices, and $\mathcal{S}^n_H := \{A \in \mathcal{S}^n : \mathbf{diag}(A) = \mathbf{0}\}$, is the set of symmetric matrices with zero diagonal.

Schoenberg's Theorem Given a matrix D, we can determine if it is an EDM with the following well-known theorem [27]:

Theorem 1. *A matrix $D \in \mathcal{S}^n_H$ is a Euclidean distance matrix if and only if $\mathsf{K}^\dagger(D)$ is positive semidefinite. Moreover,* $\mathbf{embdim}(D) = \mathbf{rank}(\mathsf{K}^\dagger(D))$ *for all $D \in \mathcal{E}^n$.*

2.2 The SDP Formulation

Semidefinite optimization or, more commonly, semidefinite programming is a class of convex optimization problems that has attracted much attention in the optimization community and has found numerous applications in different science and engineering fields. Notably, several diverse convex optimization problems can be formulated as SDP problems [31]. Current state-of-the-art SDP solvers are based on *primal-dual interior-point* methods.

Preliminary Problem Formulation There are three types of constraints in our formulation: (i) equality constraints, which are the union of equality constraints preserving bond lengths ($\mathcal{B}$), bond angles ($\mathcal{A}$), and planarity of the coplanar atoms ($\mathcal{P}$), giving $\mathcal{E} = \mathcal{E}_{\mathcal{B}} \cup \mathcal{E}_{\mathcal{A}} \cup \mathcal{E}_{\mathcal{P}}$; (ii) upper bounds, which are the union of NOE-derived ($\mathcal{N}$), hydrogen bonds ($\mathcal{H}$), disulfide and salt bridges ($\mathcal{D}$), and torsion angle ($\mathcal{T}$) upper bounds, giving $\mathcal{U} = \mathcal{U}_{\mathcal{N}} \cup \mathcal{U}_{\mathcal{H}} \cup \mathcal{U}_{\mathcal{D}} \cup \mathcal{U}_{\mathcal{T}}$; (iii) lower bounds, which are the union of steric or van der Waals ($\mathcal{W}$) and torsion angle ($\mathcal{T}$) lower bounds, giving $\mathcal{L} = \mathcal{L}_{\mathcal{W}} \cup \mathcal{L}_{\mathcal{T}}$. We assume the target protein has n atoms, $a_1, \ldots, a_n$. The preliminary problem formulation is given by:

$$\begin{array}{ll} \text{minimize} & \gamma\langle I, K\rangle + \sum_{ij} w_{ij}\xi_{ij} + \sum_{ij} w'_{ij}\zeta_{ij} \\ \text{subject to} & \langle A_{ij}, K\rangle = e_{ij},\ (i,j) \in \mathcal{E} \\ & \langle A_{ij}, K\rangle \leq u_{ij} + \xi_{ij},\ (i,j) \in \mathcal{U} \\ & \langle A_{ij}, K\rangle \geq l_{ij} - \zeta_{ij},\ (i,j) \in \mathcal{L} \\ & \xi_{ij} \in \mathbb{R}_+,\ (i,j) \in \mathcal{U}, \quad \zeta_{ij} \in \mathbb{R}_+,\ (i,j) \in \mathcal{L} \\ & K\mathbf{1} = \mathbf{0}, \quad K \in \mathcal{S}^n_+, \end{array} \tag{4}$$

where $A_{ij} = (\boldsymbol{e}_i - \boldsymbol{e}_j)(\boldsymbol{e}_i - \boldsymbol{e}_j)^\top$ and $\boldsymbol{e}_i$ is the ith column of the identity matrix. The *centering* constraint $K\mathbf{1} = \mathbf{0}$, ensures that the embedding of K is centered at the origin. Since both upper bounds and lower bounds may be inaccurate and noisy, non-negative penalized slacks, ζ_{ij}'s and ξ_{ij}'s, are included to prevent infeasibility and manage ambiguous upper bounds. The heuristic rank reduction term, $\gamma\langle I, K\rangle$, with $\gamma < 0$, in the objective function, produces lower-rank solutions [33].

Challenges in Solving the SDP Problem Solving the optimization problem in (4) can be challenging: For small to medium sized proteins, the number of atoms, n, is 1,000-3,500, and current primal-dual interior-point SDP solvers cannot solve problems with $n > 2,000$ efficiently. Moreover, the optimization problem in (4) does not satisfy *strict feasibility*, causing numerical problems; see [32].

It can be observed that the protein contains many small intersecting cliques. For example, peptide planes or aromatic rings, are 2D cliques, and tetrahedral carbons form 3D cliques. As we show later, whenever there is a clique in the protein, the corresponding Gram matrix, K, can never be full-rank, which violates strict feasibility. By adapting and extending a technique called *semidefinite facial reduction*, not only do we obtain an equivalent problem that satisfies strict feasibility, but we also significantly reduce the SDP problem size.

2.3 Cliques in a Protein Molecule

A protein molecule with ℓ amino acid residues has $\ell + 1$ planes in its backbone. Moreover, each amino acid has a different side chain with a different structure; therefore, the number of cliques in each side chain varies. We assume that the i-th residue, r_i, has s_i cliques in its side chain, denoted by $\mathcal{S}_i^{(1)}, \ldots, \mathcal{S}_i^{(s_i)}$. For all amino acids (except glycine and proline), the first side chain clique is formed around the tetrahedral carbon CA, $\mathcal{S}_i^{(1)} = \{\mathrm{N}_i, \mathrm{CA}_i, \mathrm{HA}_i, \mathrm{CB}_i, \mathrm{C}_i\}$, which intersects with two peptide planes $\mathcal{P}_{i-1}$ and $\mathcal{P}_i$ in two atoms: $\mathcal{S}_i^{(1)} \cap \mathcal{P}_{i-1} = \{\mathrm{N}_i, \mathrm{CA}_i\}$ and $\mathcal{S}_i^{(1)} \cap \mathcal{P}_i = \{\mathrm{CA}_i, \mathrm{C}_i\}$. There is a total of $q = \ell + 1 + \sum_{i=1}^{\ell} s_i$ cliques in the distance matrix of any protein. To simplify, let $\mathcal{C}_i = \mathcal{P}_{i-1}$, $1 \leq i \leq \ell + 1$, and $\mathcal{C}_{\ell+2} = \mathcal{S}_1^{(1)}$, $\mathcal{C}_{\ell+2} = \mathcal{S}_1^{(2)}$, $\ldots$, $\mathcal{C}_q = \mathcal{S}_\ell^{(s_\ell)}$.

2.4 Algorithm for Finding the Face of the Structure

For $t < n$ and $U \in \mathbb{R}^{n \times t}$, the set of matrices $U\mathcal{S}_+^t U^\top$ is a face of $\mathcal{S}_+^n$ (in fact every face of $\mathcal{S}_+^n$ can be described in this way); see, e.g., [26]. We let $\mathbf{face}(\mathcal{F})$ represent the smallest face containing a subset $\mathcal{F}$ of $\mathcal{S}_+^n$; then we have the important property that $\mathbf{face}(\mathcal{F}) = U\mathcal{S}_+^t U^\top$ if and only if there exists $Z \in \mathcal{S}_{++}^t$ such that $UZU^\top \in \mathcal{F}$. Furthermore, in this case, we have that every $Y \in \mathcal{F}$ can be decomposed as $Y = UZU^\top$, for some $Z \in \mathcal{S}_+^t$, and the reduced feasible set $\{Z \in \mathcal{S}_+^t : UZU^\top \in \mathcal{F}\}$ has a strictly feasible point, giving us a problem that is more numerically stable to solve (problems that are not strictly feasible have a dual optimal set that is unbounded and therefore can be difficult to solve numerically; for more information, see [32]). Moreover, if $t \ll n$, this results in a significant reduction in the matrix size.

The Face of a Single Clique Here, we solve the Single Clique problem, which is defined as follows: Let D be a partial EDM of a protein. Suppose the first n_1 points form a clique in the protein, such that for $\mathcal{C}_1 = \{1, \ldots, n_1\}$, all distances are known. That is, the matrix $D_1 = D[\mathcal{C}_1]$ is completely specified, where for an *index set* $\mathcal{I} \subseteq \{1, \ldots, n\}$, $B = A[\mathcal{I}]$ is the $|\mathcal{I}| \times |\mathcal{I}|$ matrix formed by rows and columns of A indexed by $\mathcal{I}$. Moreover, let $r_1 = \mathbf{embdim}(D_1)$. We now show how to compute the smallest face containing the feasible set $\{K \in \mathcal{S}_+^n : \mathsf{K}(K[\mathcal{C}_1]) = D_1\}$.

Theorem 2 (Single Clique, [17]). *Let $U_1 \in \mathbb{R}^{n \times (n - n_1 + r_1 + 1)}$ be defined as follows:*

- *let $V_1 \in \mathbb{R}^{n_1 \times r_1}$ be a full column rank matrix such that* $\mathbf{range}(V_1) = \mathbf{range}(\mathsf{K}^\dagger(D_1))$*;*

– *let* $\bar{U}_1 := \begin{bmatrix} V_1 & \mathbf{1} \end{bmatrix}$ *and* $U_1 := \begin{array}{c} \\ n_1 \\ n-n_1 \end{array} \overset{\begin{array}{cc} r_1+1 & n-n_1 \end{array}}{\begin{bmatrix} \bar{U}_1 & 0 \\ 0 & I \end{bmatrix}} \in \mathbb{R}^{n \times (n-n_1+r_1+1)}$.

Then U_1 *has full column rank,* $\mathbf{1} \in \mathbf{range}(U)$*, and*

$$\mathbf{face}\{K \in \mathcal{S}_+^n : \mathsf{K}(K[\mathcal{C}_1]) = D[\mathcal{C}_1]\} = U_1 \mathcal{S}_+^{n-n_1+r_1+1} U_1^\top.$$

Computing the V_1 Matrix In Theorem 2, we can find V_1 by computing the eigendecomposition of $\mathsf{K}^\dagger(D[\mathcal{C}_1])$ as follows:

$$\mathsf{K}^\dagger(D[\mathcal{C}_1]) = V_1 \Lambda_1 V_1^\top, \quad V_1 \in \mathbb{R}^{n_1 \times r_1},\ \Lambda_1 \in \mathcal{S}_{++}^{r_1}. \tag{5}$$

It can be seen that V_1 has full column rank (columns are orthonormal) and also $\mathbf{range}(\mathsf{K}^\dagger(D[\mathcal{C}_1])) = \mathbf{range}(V_1)$.

2.5 The Face of a Protein Molecule

The protein molecule is made of q cliques, $\{\mathcal{C}_1, \ldots, \mathcal{C}_q\}$, such that $D[\mathcal{C}_l]$ is known, and we have $r_l = \mathbf{embdim}(D[\mathcal{C}_l])$, and $n_l = |\mathcal{C}_l|$. Let $\mathcal{F}$ be the feasible set of the SDP problem. If for each clique $\mathcal{C}_l$, we define $\mathcal{F}_l := \{K \in \mathcal{S}_+^n : \mathsf{K}(K[\mathcal{C}_l]) = D[\mathcal{C}_l]\}$, then

$$\mathcal{F} \subseteq \left(\bigcap_{l=1}^{q} \mathcal{F}_l \right) \cap \mathcal{S}_C^n, \tag{6}$$

where $\mathcal{S}_C^n = \{K \in \mathcal{S}^n : K\mathbf{1} = \mathbf{0}\}$. For $l = 1, \ldots, q$, let $F_l := \mathbf{face}(\mathcal{F}_l) = U_l \mathcal{S}_+^{n-n_l+r_l+1} U_l^\top$, where U_l is computed as in Theorem 2. We have [17]:

$$\left(\bigcap_{l=1}^{q} \mathcal{F}_l \right) \cap \mathcal{S}_C^n \subseteq \left(\bigcap_{l=1}^{q} U_l \mathcal{S}_+^{n-n_l+r_l+1} U_l^\top \right) \cap \mathcal{S}_C^n = (U \mathcal{S}_+^k U^\top) \cap \mathcal{S}_C^n, \tag{7}$$

where $U \in \mathbb{R}^{n \times k}$ is a full column rank matrix that satisfies $\mathbf{range}(U) = \bigcap_{l=1}^{q} \mathbf{range}(U_l)$.

We now have an efficient method for computing the face of the feasible set $\mathcal{F}$. After computing U, we can decompose the Gram matrix as $K = UZU^\top$, for $Z \in \mathcal{S}_+^k$. However, by exploiting the centering constraint, $K\mathbf{1} = \mathbf{0}$, we can reduce the matrix size one more. If $V \in \mathbb{R}^{k \times (k-1)}$ has full column rank and satisfies $\mathbf{range}(V) = \mathbf{null}(\mathbf{1}^\top U)$, then we have [17]:

$$\mathcal{F} \subseteq (UV) \mathcal{S}_+^{k-1} (UV)^\top. \tag{8}$$

For more details on facial reduction for Euclidean distance matrix completion problems, see [16].

Constraints for Preserving the Structure of Cliques If we find a *base* set of points $\mathcal{B}_l$ in each clique $\mathcal{C}_l$ such that $\mathbf{embdim}(D[\mathcal{B}_l]) = r_l$, then by fixing the distances between points in the base set and fixing the distances between points in $\mathcal{C}_l \setminus \mathcal{B}_l$ and points in $\mathcal{B}_l$, the entire clique is kept rigid. Therefore, we need to fix *only* the distances between base points [3], resulting in a three- to four-fold reduction in the number of equality constraints. We call the reduced set of equality constraints $\mathcal{E}_{\mathrm{FR}}$.

2.6 Solving and Refining the Reduced SDP Problem

From equation (8), we can formulate the reduced SDP problem as follows:

$$\begin{aligned} \text{minimize} \quad & \gamma\langle I, Z\rangle + \textstyle\sum_{ij} w_{ij}\xi_{ij} + \sum_{ij} w'_{ij}\zeta_{ij} \\ \text{subject to} \quad & \langle A'_{ij}, Z\rangle = e_{ij},\ (i,j) \in \mathcal{E}_{\mathrm{FR}} \\ & \langle A'_{ij}, Z\rangle \le u_{ij} + \xi_{ij},\ (i,j) \in \mathcal{U} \\ & \langle A'_{ij}, Z\rangle \ge l_{ij} - \zeta_{ij},\ (i,j) \in \mathcal{L} \\ & \xi_{ij} \in \mathbb{R}_+,\ (i,j) \in \mathcal{U}, \quad \zeta_{ij} \in \mathbb{R}_+,\ (i,j) \in \mathcal{L} \\ & Z \in \mathcal{S}_+^{k-1}, \end{aligned} \tag{9}$$

where $A'_{ij} = (UV)^\top A_{ij}(UV)$.

Post-Processing We perform a refinement on the raw structure determined by the SDP solver. For this refinement we use a BFGS-based quasi-Newton method [21] that only requires the value of the objective function and its gradient at each point. Letting $X^{(0)} = X_{\mathrm{SDP}}$, we iteratively minimize the following objective function:

$$\begin{aligned} \phi(X) = w_{\mathcal{E}} \sum_{(i,j)\in\mathcal{E}} \left(\|\boldsymbol{x}_i - \boldsymbol{x}_j\| - e_{ij}\right)^2 + w_{\mathcal{U}} \sum_{(i,j)\in\mathcal{U}} f\left(\|\boldsymbol{x}_i - \boldsymbol{x}_j\| - u_{ij}\right)^2 \\ + w_{\mathcal{L}} \sum_{(i,j)\in\mathcal{L}} g\left(\|\boldsymbol{x}_i - \boldsymbol{x}_j\| - l_{ij}\right)^2 + w_r \sum_{i=1}^{n} \|\boldsymbol{x}_i\|^2, \end{aligned} \tag{10}$$

where $f(\alpha) = \max(0, \alpha)$ and $g(\alpha) = \min(0, -\alpha)$, and w_r is the weight of the regularization term. We also employ a hybrid protocol from XPLOR-NIH that incorporates thin-layer water refinement [22] and a multidimensional torsion angle database [18, 19].

3 Results

We tested the performance of SPROS on 18 proteins: 15 protein data sets from the DOCR database in the NMR Restraints Grid [10, 11] and

three protein data sets from Donaldson's laboratory at York University. We chose proteins with different sizes and topologies, as listed in Table 1. Finally, the input to the SPROS method is exactly the same as the input to the widely-used CYANA method.

3.1 Implementation

The SPROS method has been implemented and tested in MATLAB 7.13 (apart from the water refinement, which is done by XPLOR-NIH). For solving the SDP problem, we used the SDPT3 method [30]. For minimizing the post-processing objective function (10), we used the BFGS-based quasi-Newton method implementation by Lewis and Overton [21]. All the experiments were carried out on an Ubuntu 11.04 Linux PC with a 2.8 GHz Intel Core i7 Quad-Core processor and 8 GB of memory.

3.2 Determined Structures

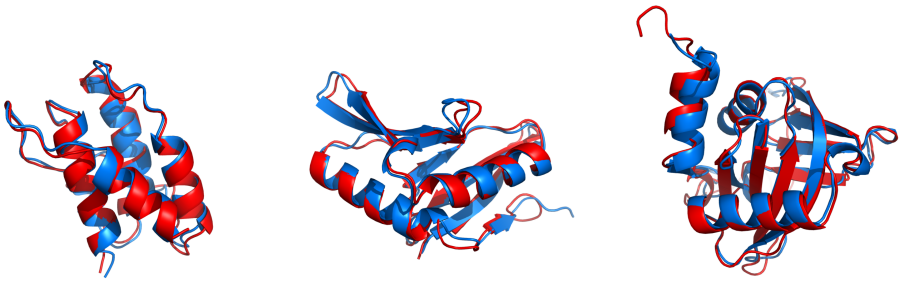
From the 18 test proteins, 9 of them were calculated with backbone RMSDs less than or equal to 1 Å, and 16 have backbone RMSDs less than 1.5 Å. Detailed analysis of calculated structures is listed in Table 2. The superimposition of the SPROS and reference structures for three of the proteins are depicted in Fig. 1. More detailed information about the determined structures can be found in [4].

To further assess the performance of SPROS, we compared the SPROS and reference structures for `1G6J`, Ubiquitin, and `2GJY`, PTB domain of Tensin, with their corresponding X-ray structures, `1UBQ` and `1WVH`, respectively. For `1G6J`, the backbone (heavy atoms) RMSDs for SPROS and the reference structures are 0.42 Å (0.57 Å) and 0.73±0.04 Å (0.98±0.04 Å), respectively. For `2GJY`, the backbone (heavy atoms) RMSDs for SPROS and the reference structures are 0.88 Å (1.15 Å) and 0.89 ± 0.08 Å (1.21 ± 0.06 Å), respectively.

3.3 Discussion

The SPROS method was tested on 18 experimentally derived protein NMR data sets of sequence lengths ranging from 76 to 307 (weights ranging from 8 to 35 KDa). Calculation times were in the order of a few minutes per structure. Accurate results were obtained for all of the data sets, although with some variability in precision. The best attribute of the SPROS method is its tolerance for, and efficiency at, managing many incorrect distance constraints (that are typically defined as upper bounds).

Our final goal is a fully automated system for NMR protein structure determination, from peak picking [1] to resonance assignment [2], to



(a) 2L3O (b) 2K49 (c) 2YTO

Fig. 1. Superimposition of structures determined by SPROS in blue and the reference structures in red.

Table 1. Information about the proteins used in testing SPROS. The second, third, and fourth columns, list the topologies, sequence lengths, and molecular weight of the proteins, the fifth and sixth columns, n and n', list the original and reduced SDP matrix sizes, respectively. The seventh column lists the number of cliques in the protein. The eights and ninth columns, $m_{\mathcal{E}}$ and $m'_{\mathcal{E}}$, list the number of equality constraints in the original and reduced problems, respectively. The 10^{th} column, $m_{\mathcal{U}}$, lists the total number of upper bounds for each protein. The 11^{th} column, bound types, lists intra-residue, $|i-j|=0$, sequential, $|i-j|=1$, medium range, $1<|i-j|\leq 4$, and long range, $|i-j|>4$, respectively, in percentile. The 12^{th} column, $\bar{m}_{\mathcal{U}} \pm s_{\mathcal{U}}$, lists the average number of upper bounds per residue, together with the standard deviation. The 13^{th} column, $m_{\mathcal{N}}$, lists the number of NOE-inferred upper bounds. The 14^{th} column, $p_{\mathcal{U}}$, lists the fraction of pseudo-atoms in the upper bounds in percentile. The last two columns, $m_{\mathcal{T}}$ and $m_{\mathcal{H}}$, list the number of upper bounds inferred from torsion angle restraints, and hydrogen bonds, disulfide and salt bridges, respectively.

ID	topo.	len.	weight	n	n'	cliques (2D/3D)	$m_{\mathcal{E}}$	$m'_{\mathcal{E}}$	$m_{\mathcal{U}}$	bound types	$\bar{m}_{\mathcal{U}} \pm s_{\mathcal{U}}$	$m_{\mathcal{N}}$	$p_{\mathcal{U}}$	$m_{\mathcal{T}}$	$m_{\mathcal{H}}$
1G6J	a+b	76	8.58	1434	405	304 (201/103)	5543	1167	1354	21/29/17/33	31.9±15.3	1291	32	63	0
1B4R	B	80	7.96	1281	346	248 (145/103)	4887	1027	787	26/25/6/43	17.1±10.8	687	30	22	78
2E8O	A	103	11.40	1523	419	317 (212/105)	5846	1214	3157	19/29/26/26	71.4±35.4	3070	24	87	0
1CN7	a/b	104	11.30	1927	532	393 (253/140)	7399	1540	1560	46/24/12/18	23.1±13.4	1418	31	80	62
2KTS	a/b	117	12.85	2075	593	448 (299/149)	7968	1719	2279	22/28/14/36	34.6±17.4	2276	25	0	3
2K49	a+b	118	13.10	2017	574	433 (291/142)	7710	1657	2612	22/27/18/38	40.9±21.1	2374	27	146	92
2K62	B	125	15.10	2328	655	492 (327/165)	8943	1886	2367	21/32/15/32	33.9±18.6	2187	32	180	0
2L3O	A	127	14.30	1867	512	393 (269/124)	7143	1492	1270	24/38/20/18	22.5±12.7	1055	25	156	59
2GJY	a+b	144	15.67	2337	639	474 (302/172)	8919	1875	1710	7/30/19/44	25.0±16.6	1536	29	98	76
2KTE	a/b	152	17.21	2576	717	542 (360/182)	9861	2089	1899	17/31/22/30	24.3±20.8	1669	30	124	106
1XPW	B	153	17.44	2578	723	541 (355/186)	9837	2081	1206	0/31/11/58	17.0±10.8	934	37	210	62
2K7H	a/b	157	16.66	2710	756	563 (363/200)	10452	2196	2768	29/33/13/25	30.3±11.3	2481	19	239	48
2KVP	A	165	17.28	2533	722	535 (344/191)	9703	2094	5204	31/26/23/20	59.2±25.0	4972	22	232	0
2YTO	a+b	176	19.17	2940	828	627 (419/208)	11210	2404	3357	23/28/14/35	34.9±22.3	3237	30	120	0
2L7B	A	307	35.30	5603	1567	1205 (836/369)	21421	4521	4355	10/30/44/16	27.6±14.4	3459	23	408	488
1Z1V	A	80	9.31	1259	362	272 (181/91)	4836	1046	1261	46/24/18/13	28.6±16.3	1189	15	0	72
HACS1	B	87	9.63	1150	315	237 (156/81)	4401	923	828	46/21/5/27	20.2±14.2	828	20	0	36
2LJG	a+b	153	17.03	2343	662	495 (327/168)	9009	1909	1347	40/29/8/22	16.4±11.9	1065	28	204	78

Table 2. Information about determined structures of the test proteins. The second, third, and fourth columns list SDP time, water refinement time, and total time, respectively. For the backbone and heavy atom RMSD columns, the mean and standard deviation between the determined structure and the reference structures is reported (backbone RMSDs less than 1.5 Å are shown in bold). The seventh column, CBd, lists the number of residues with "CB deviations" larger than 0.25 Å computed by MolProbity, as defined by [9]. The eighth and ninth columns list the percentage of upper bound violations larger than 0.1 Å and 1.0 Å, respectively (the numbers for the reference structures are in parentheses). The last three columns, list the percentage of residues with favorable and allowed backbone torsion angles and outliers, respectively.

				RMSD			violations		Ramachandran		
ID	t_s	t_w	t_t	backbone	heavy atoms	CBd.	0.1 Å	1.0 Å	fav.	alw.	out.
1G6J	44.5	175.5	241.0	**0.68±0.05**	0.90±0.05	0	4.96 (0.08±0.07)	0.85 (0)	100	100	0
1B4R	21.4	138.0	179.0	**0.85±0.06**	1.06±0.06	0	20.92 (13.87±0.62)	6.14 (2.28±0.21)	80.8	93.6	6.4
2E8O	129.8	181.3	340.9	**0.58±0.02**	0.68±0.01	0	31.33 (31.93±0.14)	9.98 (10.75±0.13)	96.2	100	0
1CN7	75.0	230.1	339.7	1.53±0.11	1.80±0.10	0	10.27 (7.63±0.80)	3.18 (2.11±0.52)	96.1	99.0	1.0
2KTS	116.7	231.0	398.5	**0.92±0.06**	1.13±0.06	0	25.36 (27.44±0.58)	6.49 (10.36±0.68)	86.1	95.7	4.3
2K49	140.7	240.7	422.7	**0.99±0.14**	1.24±0.16	0	13.75 (15.79±0.67)	2.80 (4.94±0.46)	93.8	97.3	2.7
2K62	156.1	259.0	464.2	**1.40±0.08**	1.72±0.08	1	33.74 (42.92±0.95)	10.79 (21.20±1.20)	87.8	95.9	4.1
2L3O	61.7	212.0	310.0	**1.28±0.15**	1.59±0.15	0	21.53 (19.81±0.58)	7.33 (7.61±0.31)	80.4	92.8	7.2
2GJY	113.7	285.9	455.7	**0.99±0.07**	1.29±0.09	0	11.67 (8.36±0.59)	0.36 (0.49±0.12)	85.4	92.3	7.7
2KTE	139.9	297.7	503.2	**1.39±0.17**	1.85±0.16	1	35.55 (31.97±0.46)	11.94 (11.96±0.40)	79.4	90.8	9.2
1XPW	124.8	297.1	489.7	**1.30±0.10**	1.68±0.10	0	9.74 (0.17±0.09)	1.20 (0.01±0.02)	87.9	97.9	2.1
2K7H	211.7	312.0	591.0	**1.24±0.07**	1.49±0.07	0	17.60 (16.45±0.30)	4.39 (4.92±0.35)	92.3	96.1	3.9
2KVP	462.0	282.4	814.8	**0.94±0.08**	1.05±0.09	0	15.15 (17.43±0.29)	4.01 (5.62±0.21)	96.6	100	0
2YT0	292.1	421.5	800.1	**0.79±0.05**	1.04±0.06	1	29.04 (28.9±0.36)	6.64 (6.60±0.30)	90.5	97.6	2.4
2L7B	1101.1	593.0	1992.1	2.15±0.11	2.55±0.11	3	19.15 (21.72±0.36)	4.23 (4.73±0.23)	79.2	91.6	8.4
1Z1V	30.6	158.8	209.2	**1.44±0.17**	1.74±0.15	0	3.89 (2.00±0.25)	0.62 (0)	90.9	98.5	1.5
HACS1	17.4	145.0	176.1	**1.00±0.07**	1.39±0.10	0	20.29 (15.68±0.43)	4.95 (3.73±0.33)	83.6	96.7	3.3
2LJG	94.7	280.4	426.3	**1.24±0.09**	1.70±0.10	1	28.35 (25.3±0.51)	10.76 (8.91±0.49)	80.6	90.7	9.3

protein structure determination. An automated system, without the laborious human intervention will have to tolerate more errors than usual. This was the initial motivation of designing SPROS. The key is to tolerate more errors. Thus, we are working towards incorporating an adaptive violation weight mechanism to identify the most significant outliers in the set of distance restraints automatically.

Acknowledgments

This work was supported in part by NSERC Grant OGP0046506, Canada Research Chair program, an NSERC Collaborative Grant, David R. Cheriton Graduate Scholarship, OCRiT, Premier's Discovery Award, and Killam Prize.

References

1. Alipanahi, B., Gao, X., Karakoc, E., Donaldson, L., Li, M.: PICKY: a novel SVD-based NMR spectra peak picking method. Bioinformatics 25(12), i268–275 (2009)
2. Alipanahi, B., Gao, X., Karakoc, E., Li, S., Balbach, F., Feng, G., Donaldson, L., Li, M.: Error tolerant NMR backbone resonance assignment and automated structure generation. Journal of Bioinformatics and Computational Biology 0(1), 1–26 (2011)
3. Alipanahi, B., Krislock, N., Ghodsi, A.: Manifold learning by semidefinite facial reduction (2011), unpublished manuscript (in preparation)
4. Alipanahi, B.: New Approaches to Protein NMR Automation. Ph.D. thesis, University of Waterloo (2011)
5. Biswas, P., Toh, K.C., Ye, Y.: A distributed SDP approach for large-scale noisy anchor-free graph realization with applications to molecular conformation. SIAM J. Sci. Comput. 30, 1251–1277 (March 2008)
6. Braun, W., Bösch, C., Brown, L.R., Go, N., Wüthrich, K.: Combined use of proton-proton overhauser enhancements and a distance geometry algorithm for determination of polypeptide conformations. application to micelle-bound glucagon. Biochimica et biophysica acta 667(2), 377–396 (Feb 1981)
7. Braun, W., Go, N.: Calculation of protein conformations by proton-proton distance constraints. a new efficient algorithm. Journal of molecular biology 186(3), 611–626 (1985)
8. Brünger, A.T.: X-PLOR Version 3.1: A System for X-ray Crystallography and NMR. Yale University Press (1993)
9. Chen, V.B., Arendall, W.B., Headd, J.J., Keedy, D.a., Immormino, R.M., Kapral, G.J., Murray, L.W., Richardson, J.S., Richardson, D.C.: MolProbity: all-atom structure validation for macromolecular crystallography. Acta crystallographica. Section D, Biological crystallography 66(Pt 1), 12–21 (Jan 2010)
10. Doreleijers, J.F., Mading, S., Maziuk, D., Sojourner, K., Yin, L., Zhu, J., Markley, J.L., Ulrich, E.L.: BioMagResBank database with sets of experimental NMR constraints corresponding to the structures of over 1400 biomolecules deposited in the protein data bank. Journal of biomolecular NMR 26(2), 139–146 (Jun 2003)
11. Doreleijers, J.F., Nederveen, A.J., Vranken, W., Lin, J., Bonvin, A.M., Kaptein, R., Markley, J.L., Ulrich, E.L.: BioMagResBank databases DOCR and FRED containing converted and filtered sets of experimental NMR restraints and coordinates from over 500 protein PDB structures. Journal of biomolecular NMR 32(1), 1–12 (May 2005)
12. Güntert, P.: Structure calculation of biological macromolecules from NMR data. Quarterly reviews of biophysics 31(2), 145–237 (1998)
13. Güntert, P.: Automated NMR structure calculation with CYANA. Methods in Molecular Biology 278, 353–378 (2004)
14. Güntert, P., Mumenthaler, C., Wüthrich, K.: Torsion angle dynamics for NMR structure calculation with the new program DYANA. Journal of molecular biology 273, 283–298 (1997)
15. Havel, T.F., Wüthrich, K.: A Distance Geometry Program for Determining the Structures of Small Proteins and Other Macromolecules From Nuclear Magnetic Resonance Measurements of Intramolecular H-H Proxmities in Solution. Bulletin of Mathematical Biology 46(4), 673–698 (1984)
16. Krislock, N.: Semidefinite Facial Reduction for Low-Rank Euclidean Distance Matrix Completion. Ph.D. thesis, University of Waterloo (2010)

17. Krislock, N., Wolkowicz, H.: Explicit sensor network localization using semidefinite representations and facial reductions. SIAM J. Optimiz. 20, 2679–2708 (2010)
18. Kuszewski, J., Gronenborn, A.M., Clore, G.M.: Improving the quality of NMR and crystallographic protein structures by means of a conformational database potential derived from structure databases. Protein Sci 5(6), 1067–1080 (1996)
19. Kuszewski, J., Gronenborn, A.M., Clore, G.M.: Improvements and extensions in the conformational database potential for the refinement of NMR and x-ray structures of proteins and nucleic acids. Journal of magnetic resonance 125(1), 171–177 (Mar 1997)
20. Leung, N.H.Z., Toh, K.C.: An SDP-based divide-and-conquer algorithm for large-scale noisy anchor-free graph realization. SIAM J. Sci. Comput. 31, 4351–4372 (2009)
21. Lewis, A., Overton, M.: Nonsmooth optimization via BFGS. Submitted to SIAM J. Optimiz. (2009)
22. Linge, J.P., Habeck, M., Rieping, W., Nilges, M.: ARIA: automated NOE assignment and NMR structure calculation. Bioinformatics 19(2), 315–316 (Jan 2003)
23. Moré, J.J., Wu, Z.: Global continuation for distance geometry problems. SIAM J. Optimiz. 7, 814–836 (Mar 1997)
24. Nilges, M., Clore, G.M., Gronenborn, A.M.: Determination of three-dimensional structures of proteins from interproton distance data by hybrid distance geometry-dynamical simulated annealing calculations. FEBS letters 229(2), 317–324 (Mar 1988)
25. Raman, S., Lange, O.F., Rossi, P., Tyka, M., Wang, X., Aramini, J., Liu, G., Ramelot, T.A., Eletsky, A., Szyperski, T., Kennedy, M.A., Prestegard, J., Montelione, G.T., Baker, D.: NMR structure determination for larger proteins using Backbone-Only data. Science 327(5968), 1014–1018 (Feb 2010)
26. Ramana, M.V., Tunçel, L., Wolkowicz, H.: Strong duality for semidefinite programming. SIAM J. Optimiz. 7(3), 641–662 (1997)
27. Schoenberg, I.J.: Remarks to Maurice Fréchet's article "Sur la définition axiomatique d'une classe d'espace distanciés vectoriellement applicable sur l'espace de Hilbert". Ann. of Math. (2) 36(3), 724–732 (1935)
28. Schwieters, C., Kuszewski, J., Tjandra, N., Clore, G.: The Xplor-NIH NMR molecular structure determination package. Journal of Magnetic Resonance 160, 65–73 (2003)
29. Shen, Y., Lange, O., Delaglio, F., Rossi, P., Aramini, J.M., Liu, G., Eletsky, A., Wu, Y., Singarapu, K.K., Lemak, A., Ignatchenko, A., Arrowsmith, C.H., Szyperski, T., Montelione, G.T., Baker, D., Bax, A.: Consistent blind protein structure generation from NMR chemical shift data. Proceedings of the National Academy of Sciences of the United States of America 105(12), 4685–4690 (2008)
30. Tütüncü, R., Toh, K., Todd, M.: Solving semidefinite-quadratic-linear programs using SDPT3. Math. Program. 95(2, Ser. B), 189–217 (2003)
31. Vandenberghe, L., Boyd, S.: Semidefinite programming. SIAM Review 38(1), 49–95 (1996)
32. Wei, H., Wolkowicz, H.: Generating and measuring instances of hard semidefinite programs. Mathematical Programming 125, 31–45 (2010)
33. Weinberger, K.Q., Saul, L.K.: Unsupervised learning of image manifolds by semidefinite programming. Computer Vision and Pattern Recognition, IEEE Computer Society Conference on 2, 988–995 (2004)
34. Williams, G.A., Dugan, J.M., Altman, R.B.: Constrained global optimization for estimating molecular structure from atomic distances. Journal of computational biology 8(5), 523–547 (2001)