SPROS: An SDP-Based Protein Structure Determination from NMR Data

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

[1] School of Computer Science, University of Waterloo; [2] BiPop Research Group, INRIA Grenoble Rhône-Alpes; [3] Department of Statistics and Actuarial Science, University of Waterloo; [4] Department of Combinatorics and Optimization, University of Waterloo; [5] Department of Biology, York University

Abstract

In protein NMR, the 3D structure is determined by making use of a set of distance restraints between proton pairs and exploiting the domain knowledge about proteins. Euclidean distance geometry methods based on semidefinite programming (SDP) provide a natural formulation for realizing a 3D structure from a set of given distance constraints. However, the complexity of SDP solvers is a major obstacle in their applicability to the protein NMR problem. We propose a novel SDP-based protein structure determination from NMR data, called SPROS, which is both fast and robust. By using a technique called 'semidefinite facial reduction,' the SDP matrix size and the number of equality constraints are approximately one quarter of the original problem. Using this technique results in a one hundred-fold decrease in the running time required by the SDP solver. SPROS is applied to proteins with a molecular mass less than 15 kDa, and the predicted structures are accurate.

1. Introduction

There are two major protein structure calculation methods:

- Simulated Annealing + Gradient-descent
  - Torsion angle molecular dynamics, such as CYANA
  - Cartesian coordinates molecular dynamics, such as XPLOR
- Euclidean Distance Matrix Completion (EDMC)
  - Directly filling in missing elements in the Euclidean Distance Matrix (EDM), such as EMBED and DISGEO.
  - Using the Gram matrix and completing the distance matrix by semidefinite programming (SDP), such as DAFGL and DISCO.

Semidefinite Programming

Euclidean distance matrix, D, and the Gram matrix or matrix of inner products, K, are linearly related:

$$\begin{aligned} D_{ij} &= (\boldsymbol{x}_i - \boldsymbol{x}_j)^\top (\boldsymbol{x}_i - \boldsymbol{x}_j) \\ &= K_{ii} - 2K_{ij} + K_{jj} \end{aligned}$$

Why SDP-based EDMC is superior?

- EDM D is valid (triangle inequalities, etc) *iff* K is Positive Semidefinite ($K \succeq 0$).
- Embedding dimension of $\{\boldsymbol{x}_i\}$ is rank of K. Therefore, we have control over dimensionality of $\{\boldsymbol{x}_i\}$.

The protein structure determination problem can be formulated as an instance of SDP and solved by off-the-shelf solvers:

$$\begin{array}{ll} \min\limits_{K} & \operatorname{Tr}\{CK\} \\ \text{subject to} & \operatorname{Tr}\{A_i K\} = d_i, \quad i \in \mathbb{C} \\ & K \succeq 0 \end{array}$$

Each iteration of SDP takes $\mathcal{O}\left(n^3 + m^3\right)$, where n is the size of the SDP matrix and m is the number of constraints. The iteration count should be very low and independent of the dimension, but it is dependent on the desired accuracy. It is known that for $n > 3,000$ and $m > 10,000$, SDP is not generally tractable.

2. Semidefinite Facial Reduction

If there are cliques in the data (a set of points that all pair-wise distances between them are known), complexity of SDP can be reduced.

Intuition: if we fix just $d+1$ points from a clique with embedding dimensionality d, the remaining points can be uniquely located. This concept is called *Semidefinite Facial Reduction* (Krislock and Wolkowicz, 2010):

$$K = UZU^\top,$$

where size of Z is smaller than K, and U is a predetermined matrix computed from the cliques information.

Proteins are made up of 2D (peptide planes, aromatic rings, etc) and 3D (tetrahedral carbons, etc) cliques.

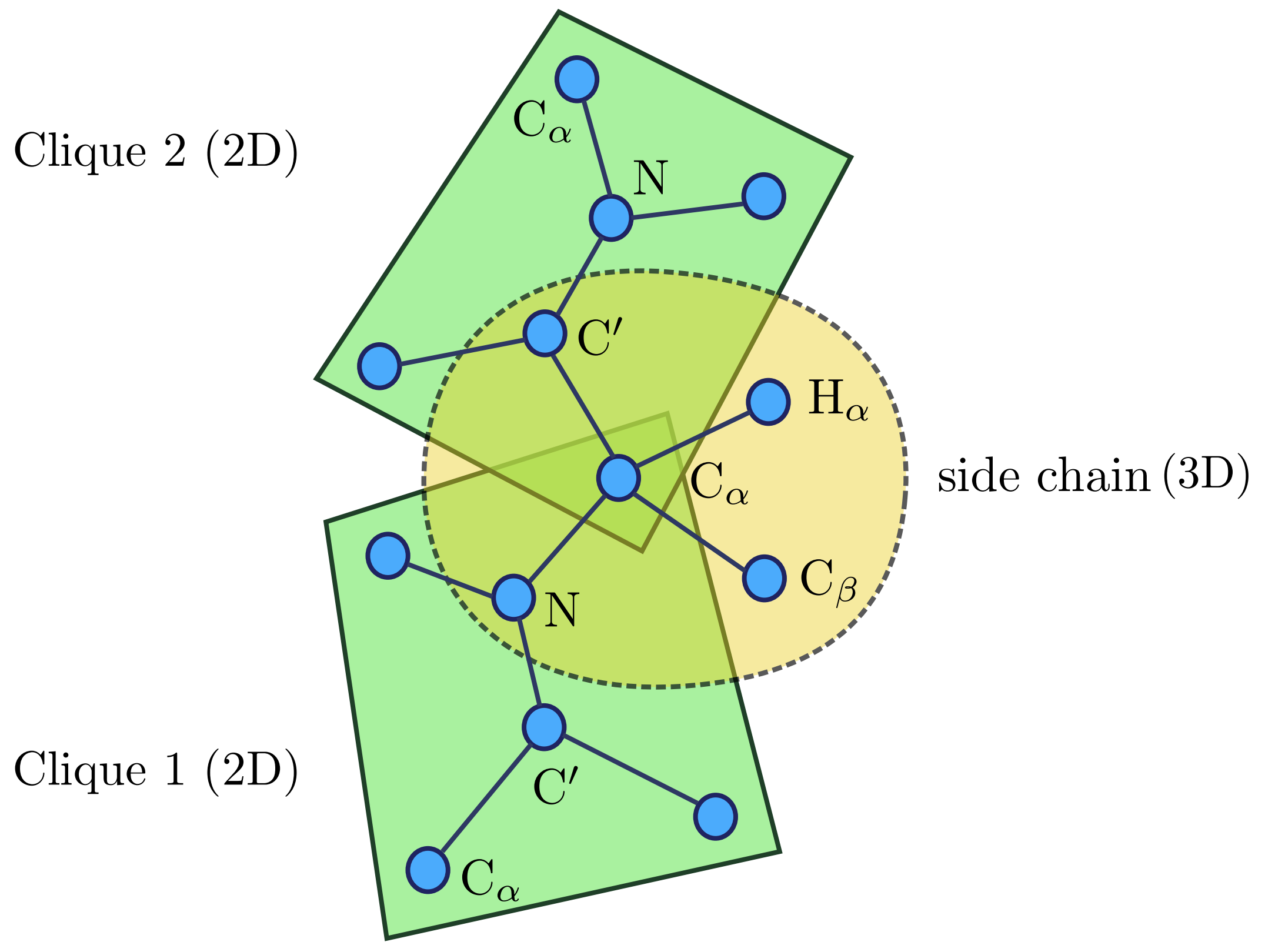

Figure 1: *Two peptide planes and the ALA side chain.*

SPROS Advantages

- After facial reduction, the Slater Condition (strict feasibility) is satisfied, and the problem size is reduced. As a result, the SDP problem is solved faster and is less prone to numerical problems.
- In contrast to CYANA and XPLOR, the objective function is Convex, and the best conformation is always found. Therefore, there is no need for Simulated Annealing. Moreover, the process is repeatable.
- Similar to the Torsion Angle space, adding each peptide plane increases the SDP problem size only by two. Similarly, each side chain dihedral angle increases the SDP problem size by one.
- In comparison to the unreduced SDP problem, m and n are reduced by a factor of three to four. Additionally, SDP iterations are nearly halved, which results in a 100-fold speed up.
- SPROS tolerates erroneous upper bounds by penalizing the ℓ_1-norm of deviations, which does not lock the structure like the ℓ_2-norm.

3. Results

SPROS is tested on three real protein datasets: STE50, the SAM domain of the yeast signaling regulator, AIDA1 neuronal signaling scaffolding protein PTB domain, and Fes SH2 domain.

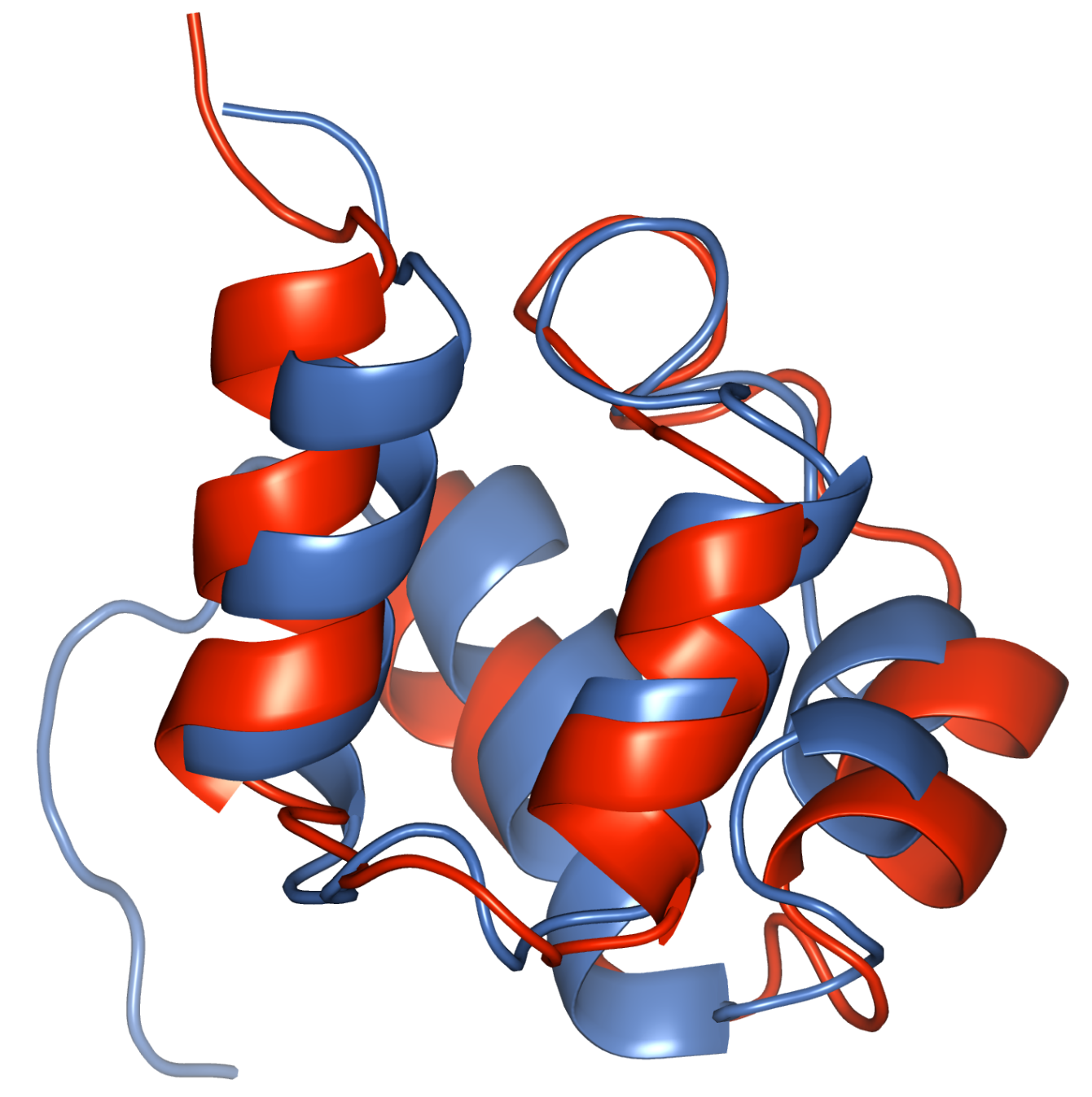

Figure 2: *Structure determined by SPROS (in red) and the reference structure in (blue) for STE50, the SAM domain of the yeast signaling regulator. RMSD: 3.3, ensemble RMSD: 2.5.*

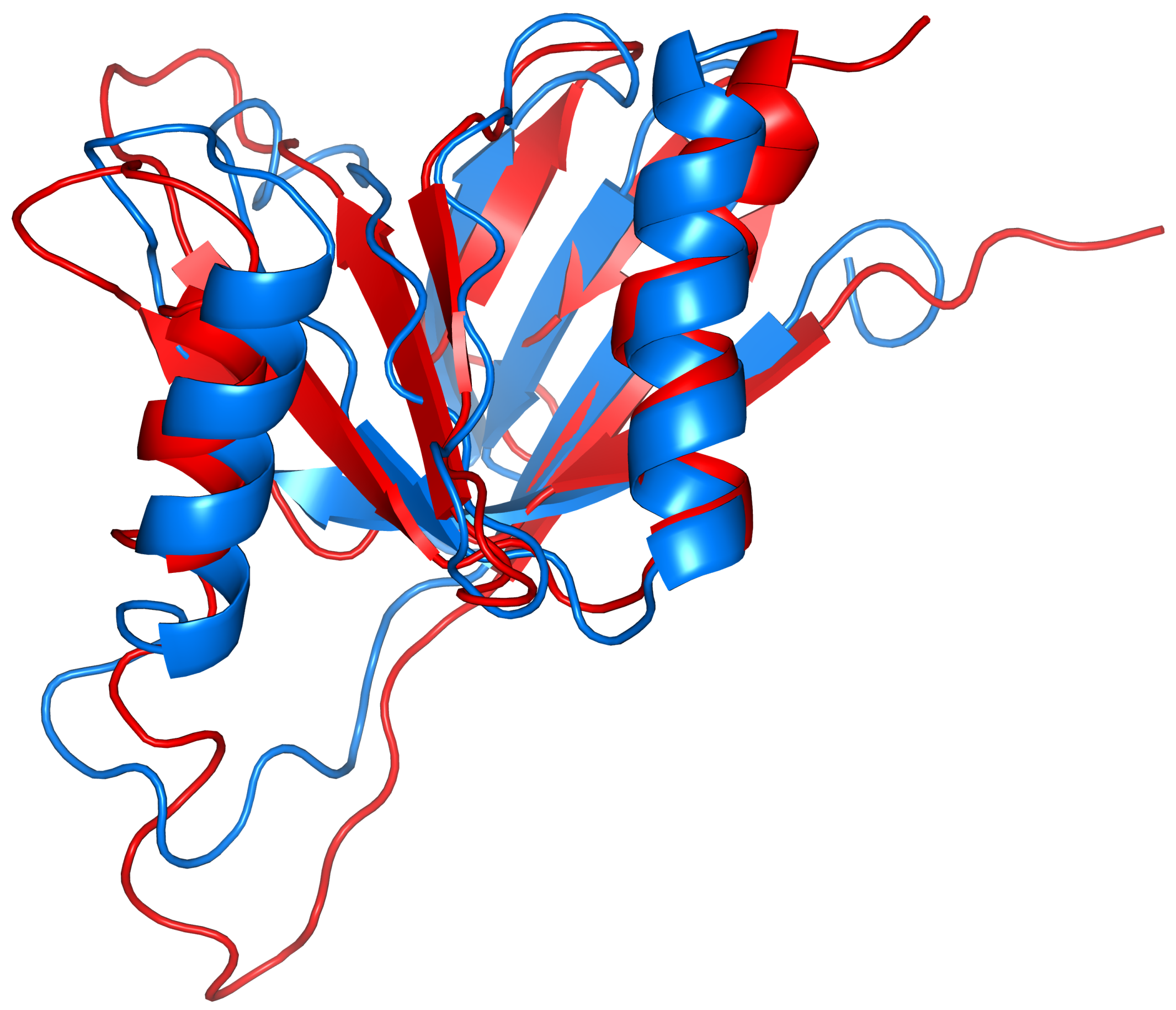

Figure 3: *Structure determined by SPROS (in red) and the reference structure in (blue) for AIDA1 neuronal signaling scaffolding protein PTB domain. RMSD: 1.7, ensemble RMSD: 1.8.*

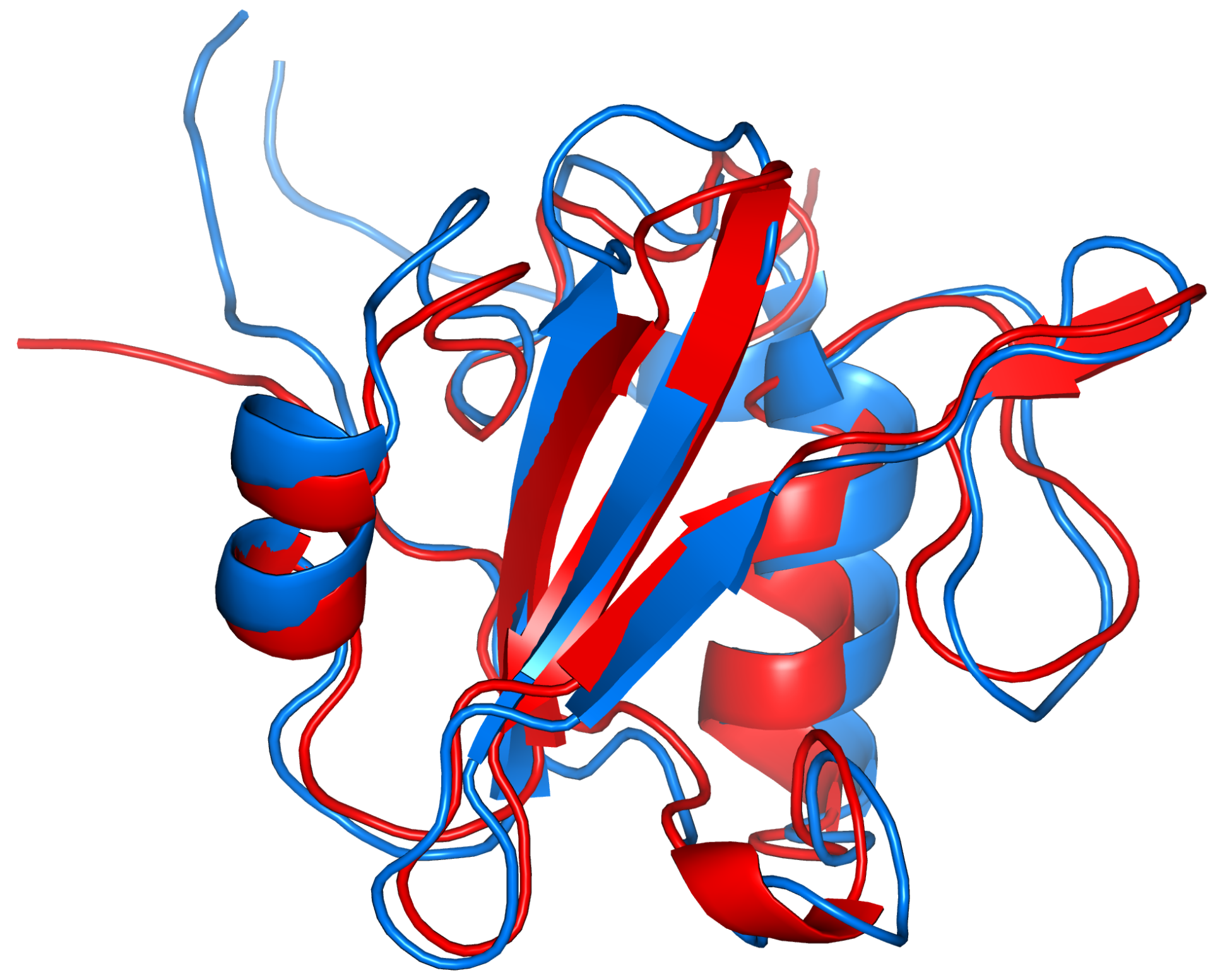

Figure 4: *Structure determined by SPROS (in red) and the reference structure in (blue) for Fes SH2 domain. RMSD: 1.5, ensemble RMSD: 2.1.*

Table 1: *Test proteins' information: n' and n are the original and the reduced SDP matrix size, respectively; m_u is the number of NMR constraints; and m'_e and m_e are the number of equality constraints for the original and the reduced SDP problem. For SPROS, the overall number of constraints is $m = m_e + m_u$ and for the original problem, it is $m' = m'_e + m_u$.*

Protein	length (a.a.)	time (s)	n'	m'_e	m_u	n	m_e
STE50	75	141	1403	5250	1286	405	1767
SH2	107	284	1976	8051	1944	541	2334
AIDA1	131	335	2343	8782	1538	665	2938

4. Conclusions

SPROS is an alternative to Simulated Annealing-based protein structure determination methods. It is very fast and finishes in a matter of minutes and always generates the same structure for a given input data. SPROS can be extended to docking applications. Moreover, by fixing the reliable parts of the structure, it can be used for further refinement of erroneous structures.

References

Krislock, N. and Wolkowicz, H. (2010). Explicit sensor network localization using semidefinite representations and facial reductions. *SIAM Journal on Optimization*, **20**(5), 2679–2708.