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
[2] 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,

April 22, 2012

Protein Structure?

Protein three-dimensional structure is key to deciphering its function and biological role.

Nuclear Magnetic Resonance (NMR)

- Determining structure of bio-macromolecules in aqueous solution
- Studying molecular dynamics
- Analyzing protein folding pathways
- Drug screening and design

Problem Definition

In short: Compute the 3D structure of a protein given a set of upper bounds on the distances between spatially proximate (closer than 5 Å) hydrogen atoms.

More formally, for a protein with n atoms, find $X = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n]$, $\boldsymbol{x}_i \in \mathbb{R}^3$ such that:

$$\begin{aligned} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2 &= e_{ij}, \quad \forall (i,j) \in \mathbb{E}\,, \\ \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2 &\leq u_{ij}, \quad \forall (i,j) \in \mathbb{U}\,, \\ \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2 &\geq l_{ij}, \quad \forall (i,j) \in \mathbb{L}\,. \end{aligned}$$

- $\mathbb{E}$: bond lengths, bond angles, and so on.
- $\mathbb{U}$: information *inferred* from NMR experiments.
- $\mathbb{L}$: mostly steric constraints.

Major Challenges

- Structure determination problem is NP-hard.
- Number of distance constraints is small, $|\mathbb{E}|$ and $|\mathbb{B}|$ are $O(n)$.
- Any proposed method should handle a large number of severely-violated bounds (~25 Å) and an even larger number of slightly-violated bounds.

Structure Determination

Major protein structure determination methods:

- Euclidean Distance Matrix Completion (EDMC)
  - Directly filling in missing elements in EDM
  - Using the Gram matrix and completing the EDM by Semidefinite Programming (SDP)
- Simulated Annealing
  - Torsion angle molecular dynamics (CYANA)
  - Cartesian coordinates molecular dynamics (XPLOR)
- Fragment Assembly
  - CS-RDC-NOE-Rosetta: uses distance constraints in the sampling
  - FALCON-NMR: uses distance constraints in picking top decoys

The Gram Matrix

Working with the Gram matrix $K = X^\top X$ has many advantages:

1. The EDM D and the Gram matrix are linearly related:

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

2. Instead of enforcing all of the triangle inequality constraints, it is sufficient to enforce that the Gram matrix is positive semidefinite.
3. The embedding dimension and the rank of the Gram matrix are directly related.

SDP Formulation

We can solve the EDMC problem by SDP:

$$\begin{aligned} \text{minimize} \quad & \langle C, K \rangle \\ \text{subject to} \quad & \langle A_i, K \rangle = d_i, \quad i = 1, \ldots, m \\ & K \in \mathbb{S}_+^n \end{aligned}$$

Challenges

1. SDP solvers run in $O(n^3 + m^3)$ and problems with $n > 2,000$ and $m > 10,000$ are not tractable.
2. The SDP problem does not satisfy Slater's condition, or strict feasibility, causing numerical problems.

Semidefinite Facial Reduction

If for the feasible set we have:

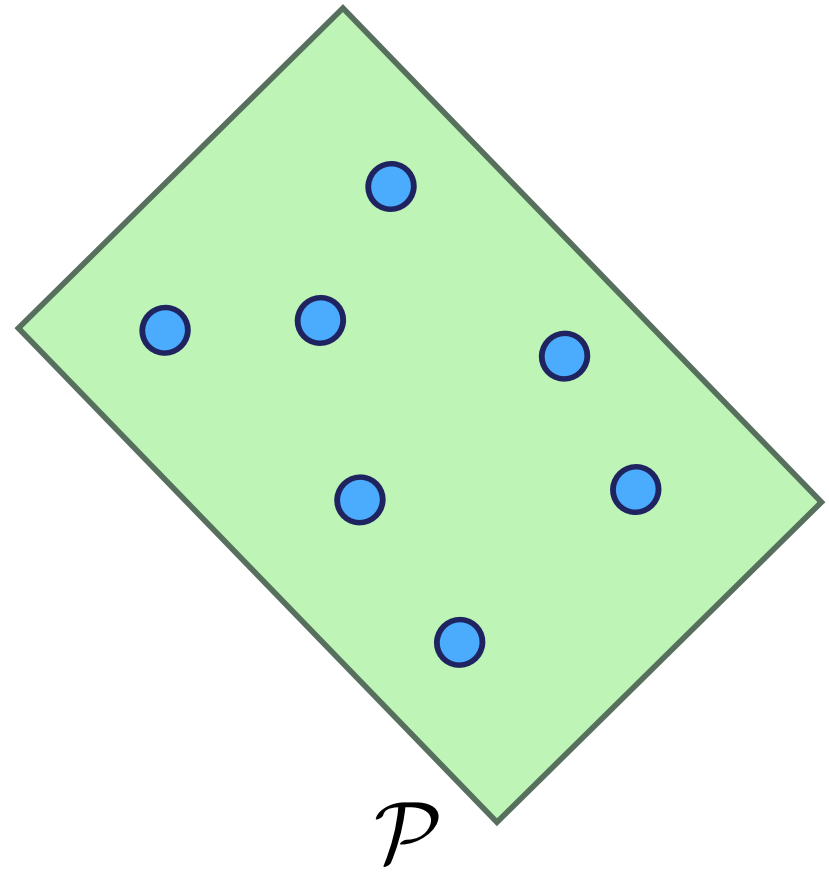
$$\left\{K \in \mathbb{S}_+^n : \langle A_i, K\rangle = d_i, \forall\, i\right\} \subseteq \underbrace{U\mathbb{S}_+^{k}U^\top}_{\text{face of } \mathbb{S}_+^n}$$

then $K = UZU^\top$, $k < n$, for some $Z \in \mathbb{S}_+^k$.

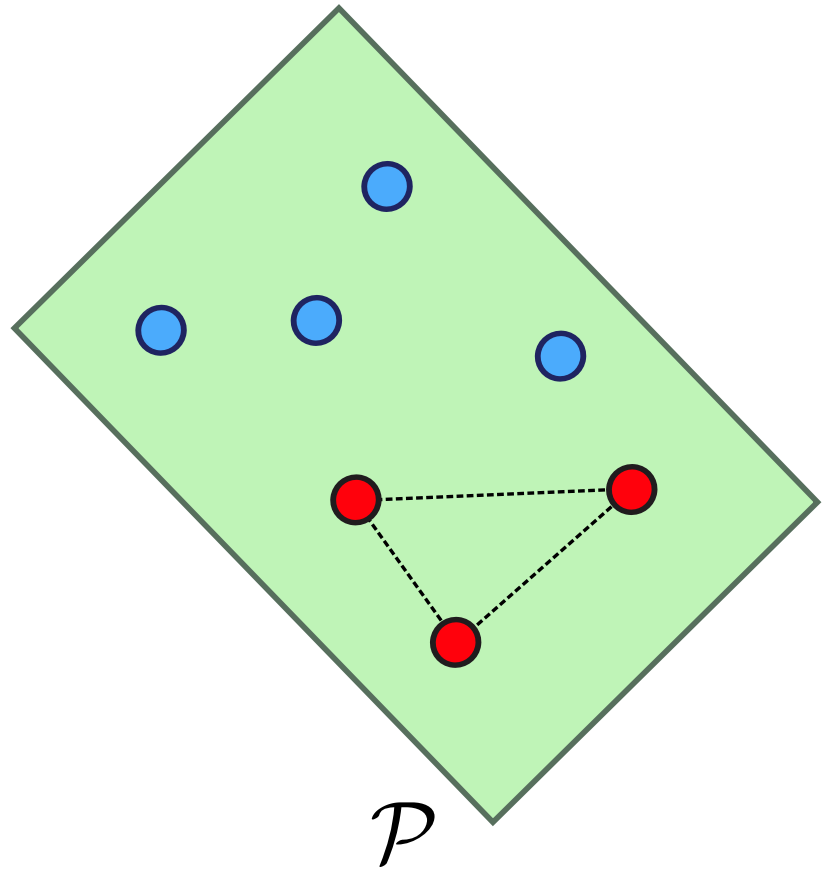
In EDMC, if there are cliques in the data (a set of points with all pair-wise distances between them known), K can be decomposed.

- Intuition: if we fix just $d+1$ points from a clique with embedding dimensionality d, the remaining points can be located.

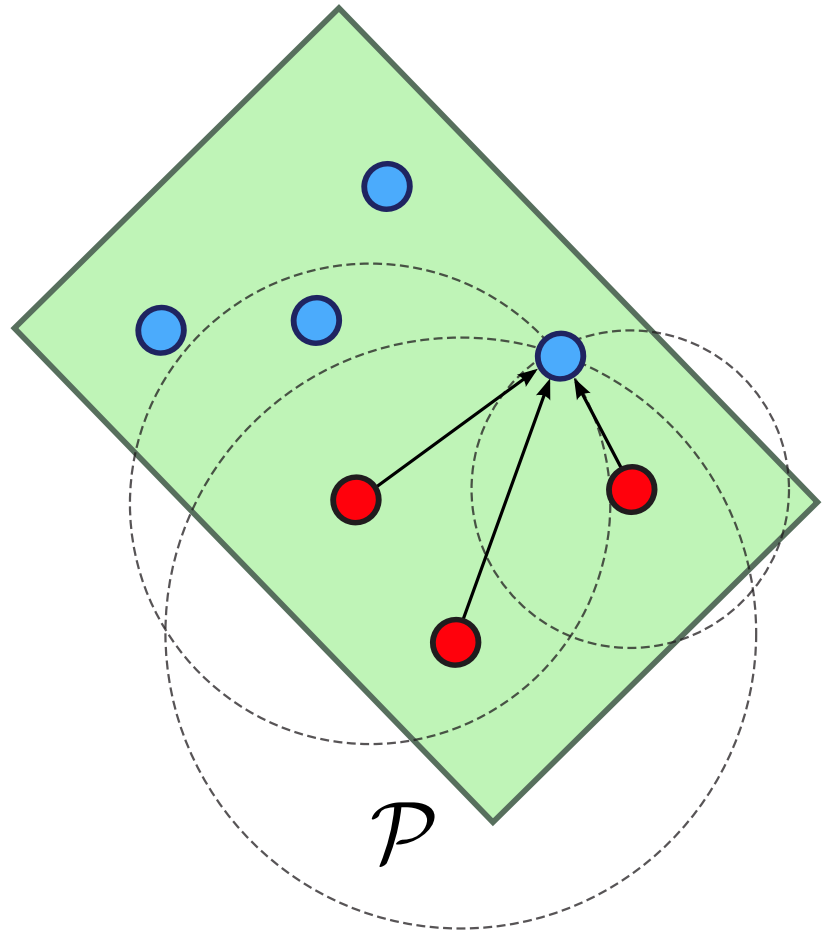
A 2D Clique



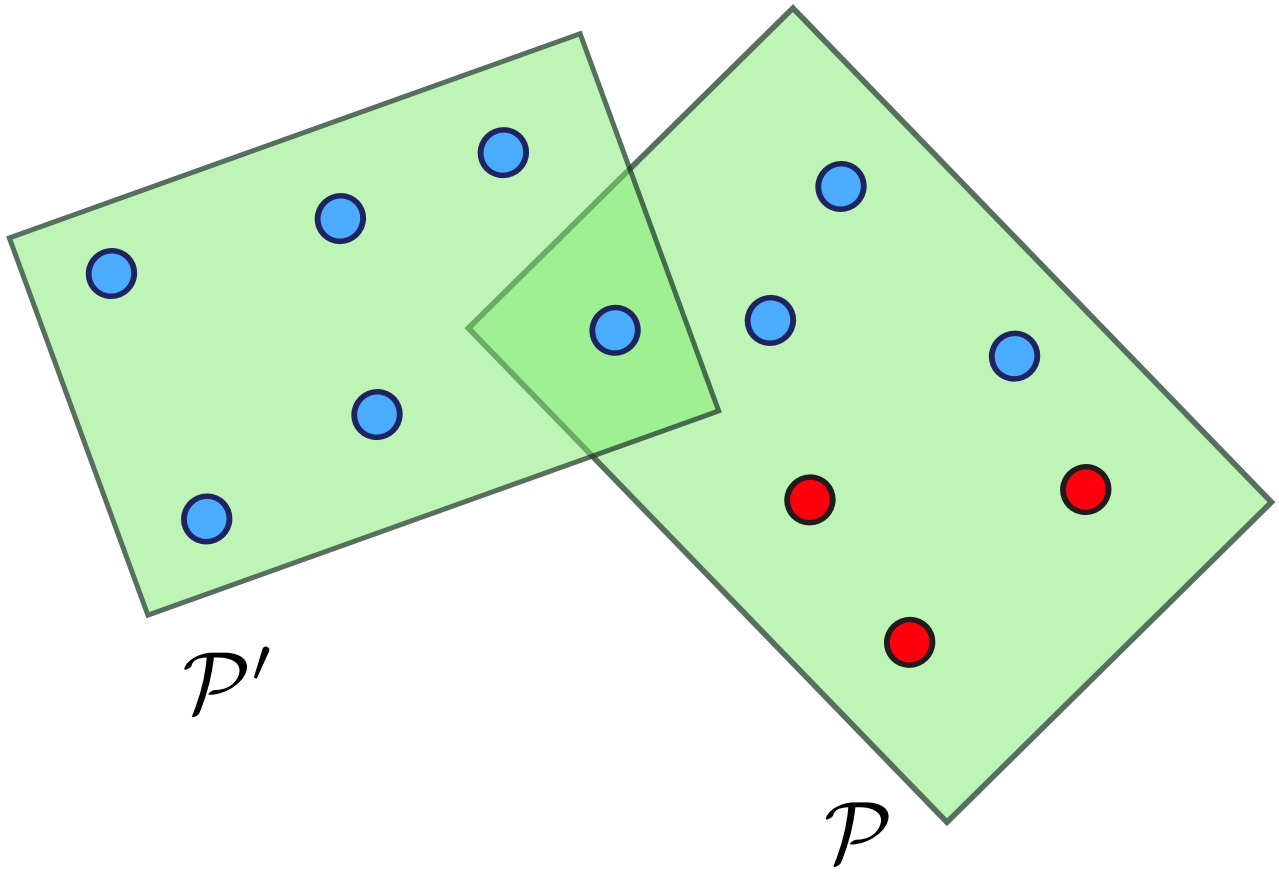
Base Points



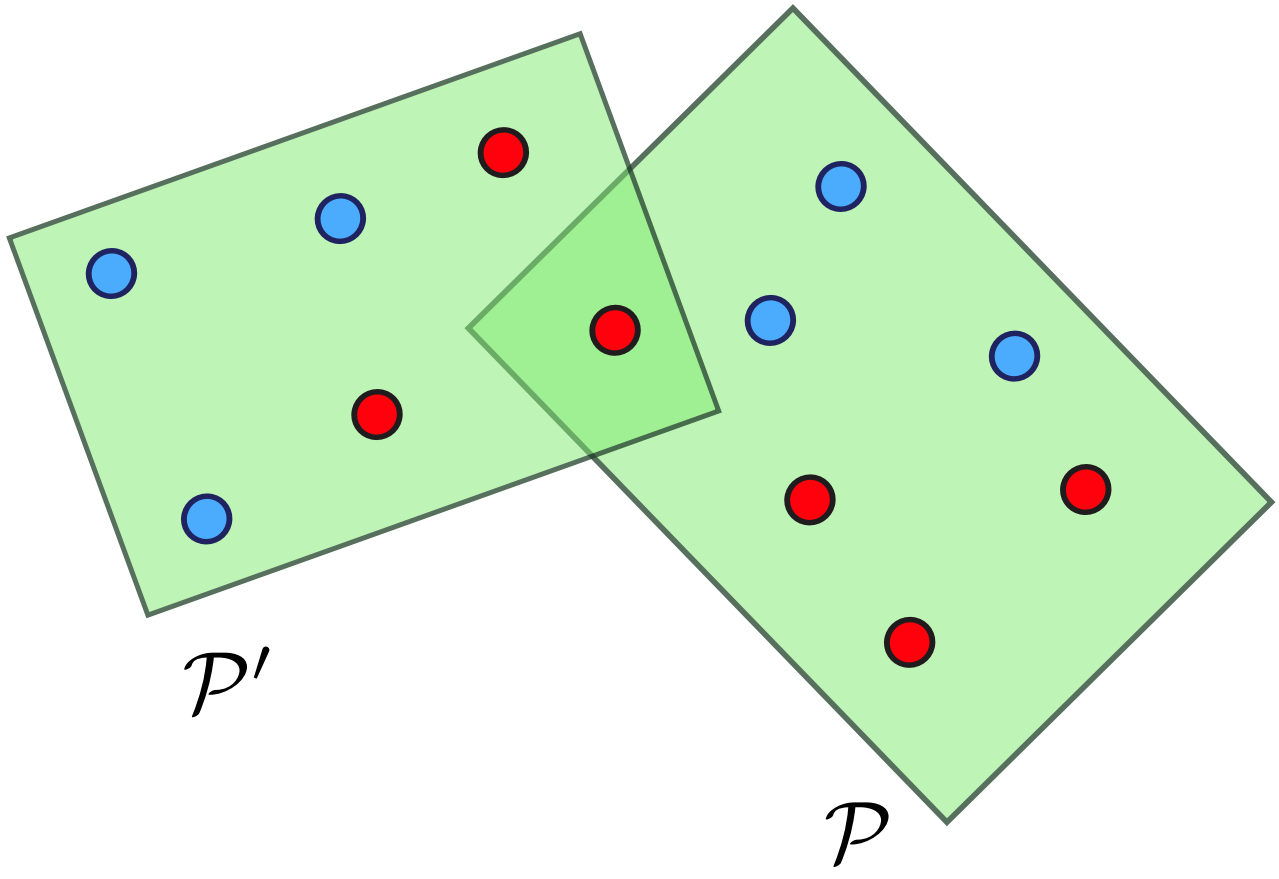
Reconstruction



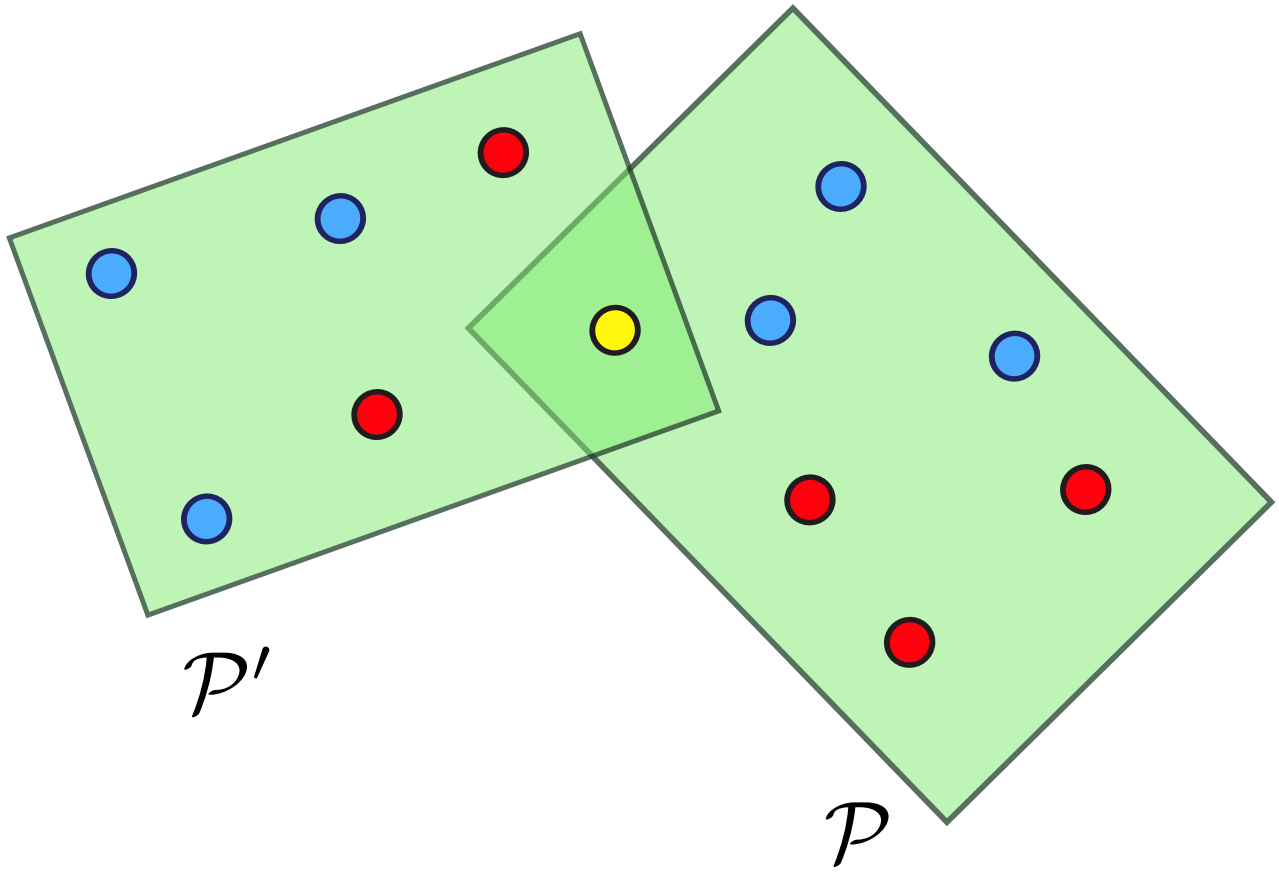
Intersecting Cliques



Base Points



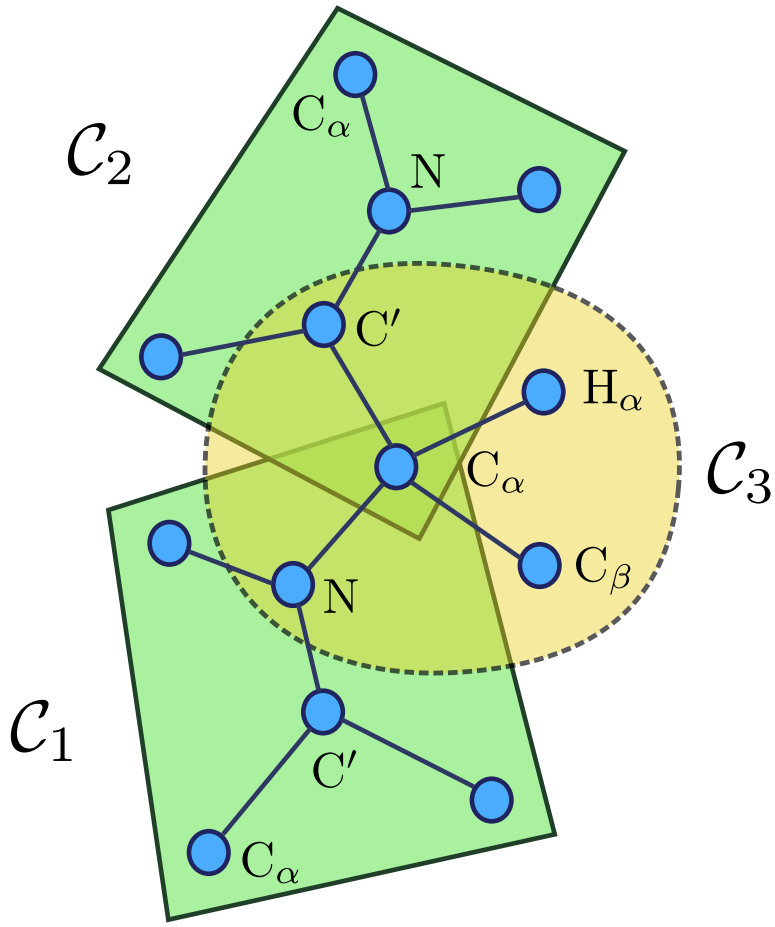
Base Points



SPROS

"SPROS" (SDP-based Protein Structure determination), models the protein molecule as a set of intersecting 2D and 3D cliques.

- For example, peptide planes or aromatic rings, are 2D cliques, and tetrahedral carbons form 3D cliques.



SPROS

- After facial reduction, the Slater Condition is satisfied.
- The objective function is Convex.
- ℓ_1-norm of violations are penalized
  - Enforces sparsity in the number of violated constraints.
  - Does not prevent correct folding like ℓ_2-norm.
- Similar to the Torsion Angle space, adding each peptide plane increases the SDP problem size only by two.
- 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 Steps

1. Sample a random structure.
2. Simplify side chains.
3. Form the cliques and the U matrix.
4. Solve the SDP problem.
5. Perform structure refinement.

Test Proteins

- SPROS is tested on 18 proteins: 15 protein data sets from the DOCR database (NMR Restraints Grid) and three protein data sets from Donaldson's laboratory at York University.
  - 5 A, 4 B, 5 a+b, and 4 a/b topologies.
  - Sequence lengths: 76-307
  - Molecular weights: 8.58 to 35.30 kDa.
- SDP matrix size was reduced by a factor of 3.6 on average.
- Number of equality constraints was reduced by a factor of 4.7 on average.
- Input files are the same as CYANA.

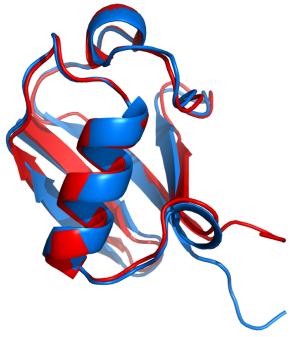
SPROS Results

SPROS is implemented in MATLAB (water refinement is done by XPLOR-NIH).

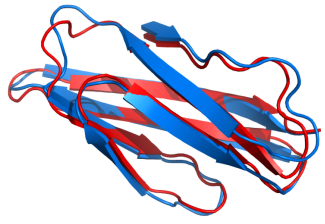
- Average backbone RMSD: 1.15 ± 0.37 Å (heavy atoms RMSD: 1.4 ± 0.44 Å).
- Average run time: 500 s (SDP time: 185 s).
- Average percentile of allowed torsion angles: 96%.

Note: A speedup of ~50–100X can be achieved if the code is transfered to C++, parallelized, optimized and more efficient BLAS libraries such as GotoBLAS2 are used.

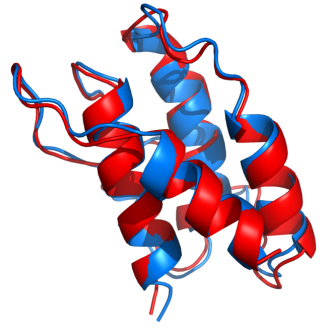
SPROS Structures



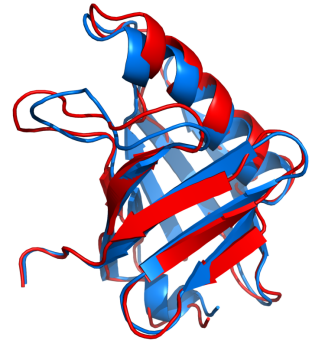
(a) 1G6J



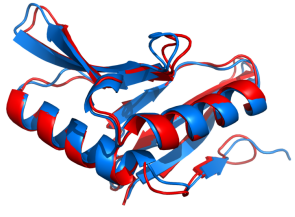
(b) 1B4R



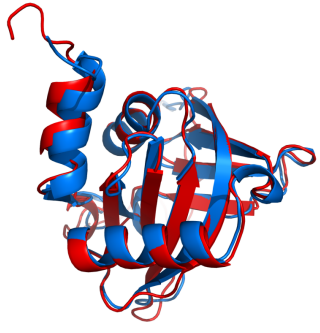
(c) 2L30



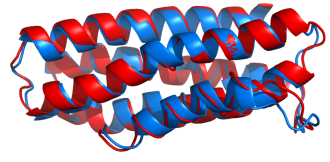
(d) 2KTS



(e) 2K49



(f) 2YTO



(g) 2KVP

(h) 2LJG

Comparison with X-ray

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.

- `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.
- `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.

Huge Opportunity!

The Semidefinite Facial Reduction method can be very effective!

Ex. Fragment Assembly

Assume a protein with 270 residues is composed of 30 *rigid* 9-mers. It will have around 5,000 atoms, while after reduction the matrix size will be just 100 (every 9-mer is a large 3D clique). A contact map can be verified in just a couple of seconds.

- Conclusions
  - SPROS is the first practical SDP-based protein structure determination method.
  - SPROS is a fast and robust alternative to the Simulated Annealing-based protein structure determination methods.
- Future Work
  - Design an iterative protocol for SPROS
  - Extend SPROS to virtual screening (docking) applications.

Acknowledgements

- My supervisors: Prof. Ming Li and Prof. Ali Ghodsi
- My collaborators: Prof. Logan Donaldson, Prof. Henry Wolkowicz, and Dr. Nathan Krislock
- David R. Cheriton Graduate Scholarship $\times 2$

- Thank you!
- Comments or questions?