

# PROTEIN MODELING FROM SPARSE DISTANCE-CONSTRAINTS DERIVED FROM NMR

NILADRI RANJAN DAS, KUNAL N. CHAUDHURY, AND DEBNATH PAL  
 MATHEMATICAL SCIENCE(NMI), INDIAN INSTITUTE OF SCIENCE  
 8TH EECS RESEARCH STUDENTS SYMPOSIUM



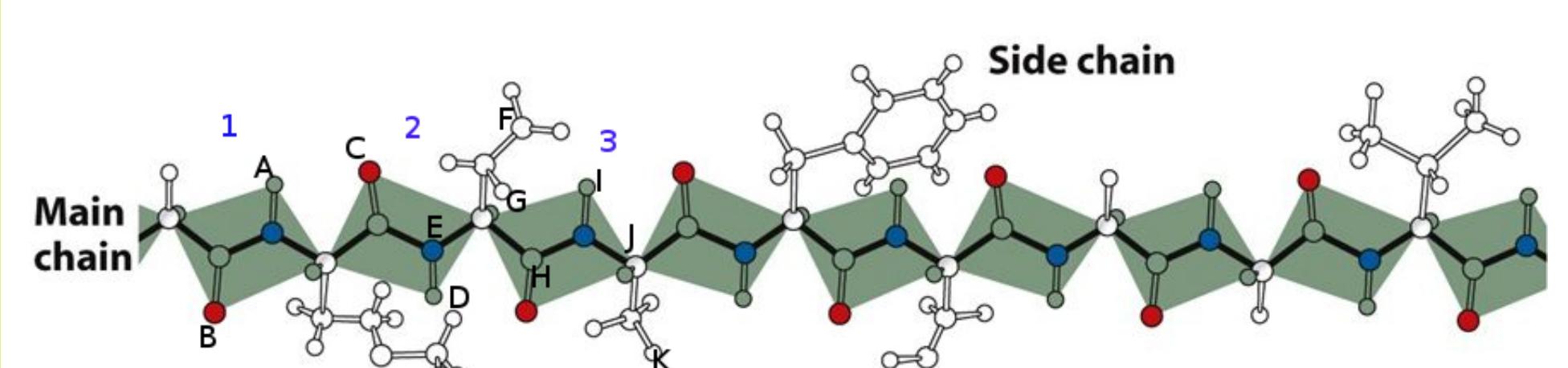
## PROTEIN CONFORMATION

- NMR experiments give:
  - a sparse set of bounds on interatomic distances (NOESY),
  - bounds in torsional angle between amino acid residues ( $J$ -coupling).
- Calculate protein confirmation(s) which respects distance bounds.

## CHALLENGES

- 1. Distance geometric methods
- 2. Energy minimization coupled with simulated annealing
- 1. Multiple local minima
- 2. Non-convex
- 3. Conformations satisfying distance bounds are exponential
- 4. Molecular conformation problem is NP-hard.

## GRAPH MODELING



1. Residue graph
2. Atom graph
- Dense regions in the residue correspond to core regions of protein.

## PROPOSED METHOD

Get the input files

Preprocess

Model the data as graph

Extract dense subgraphs  
(stable regions in protein)

Calculate structure  
for dense  
subgraphs

Calculate structure  
for dense  
subgraphs

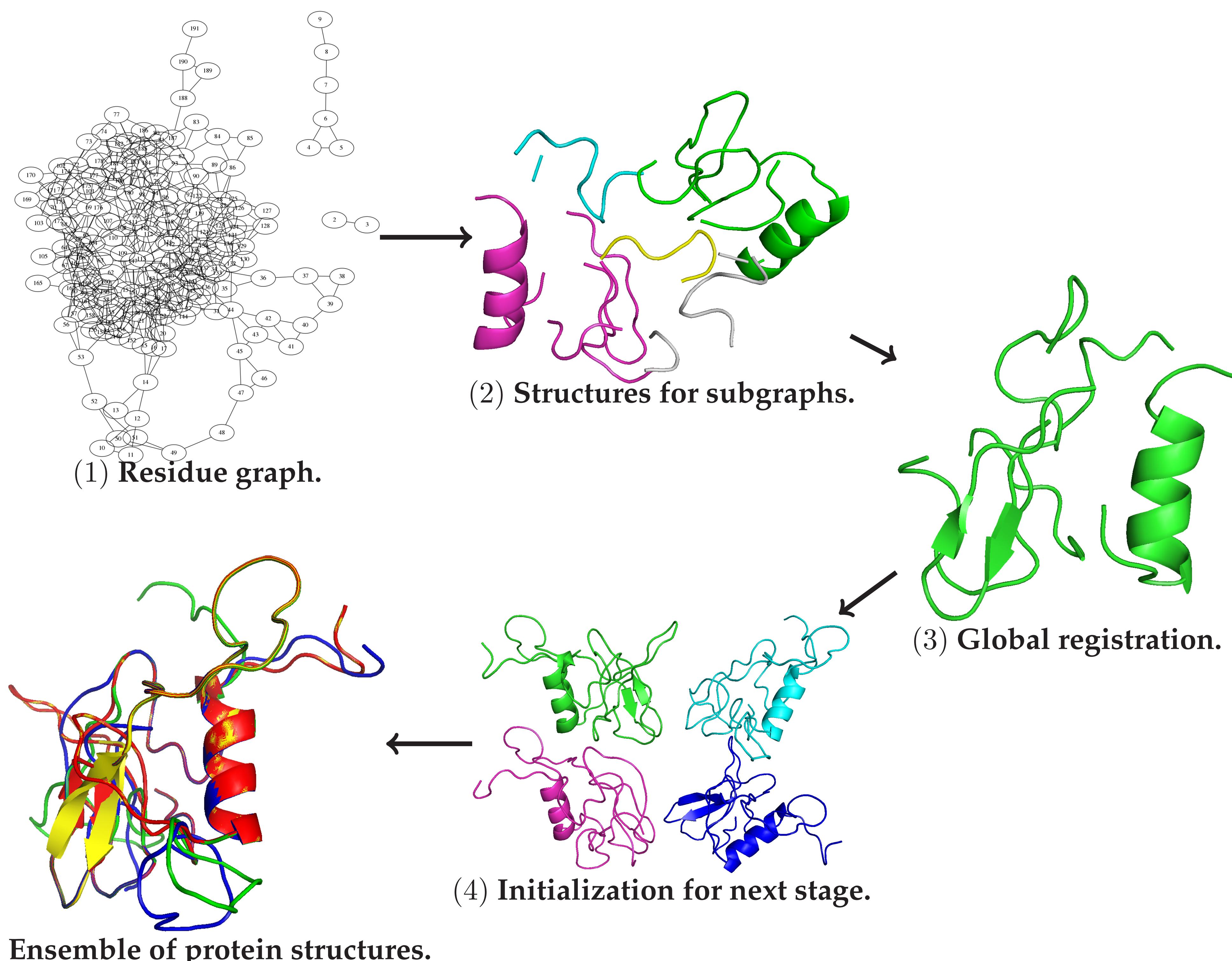
Globally register struc-  
tures in one step

Use stable regions as anchors

Find additional con-  
straints from input files

Derive ensemble of  
protein conformations

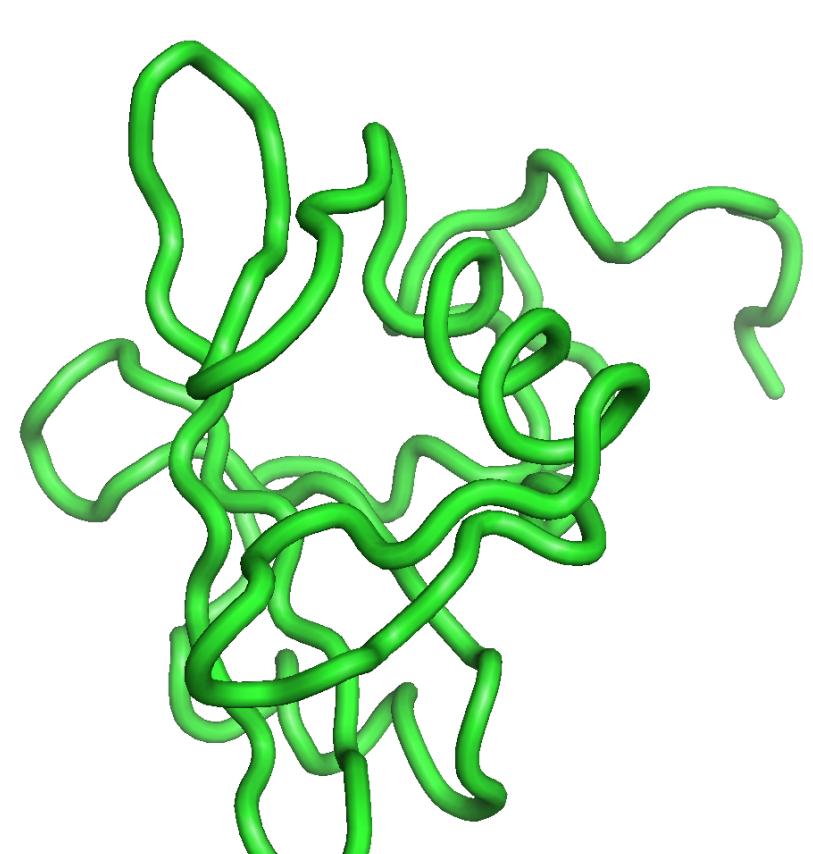
## PROPOSED WORKFLOW



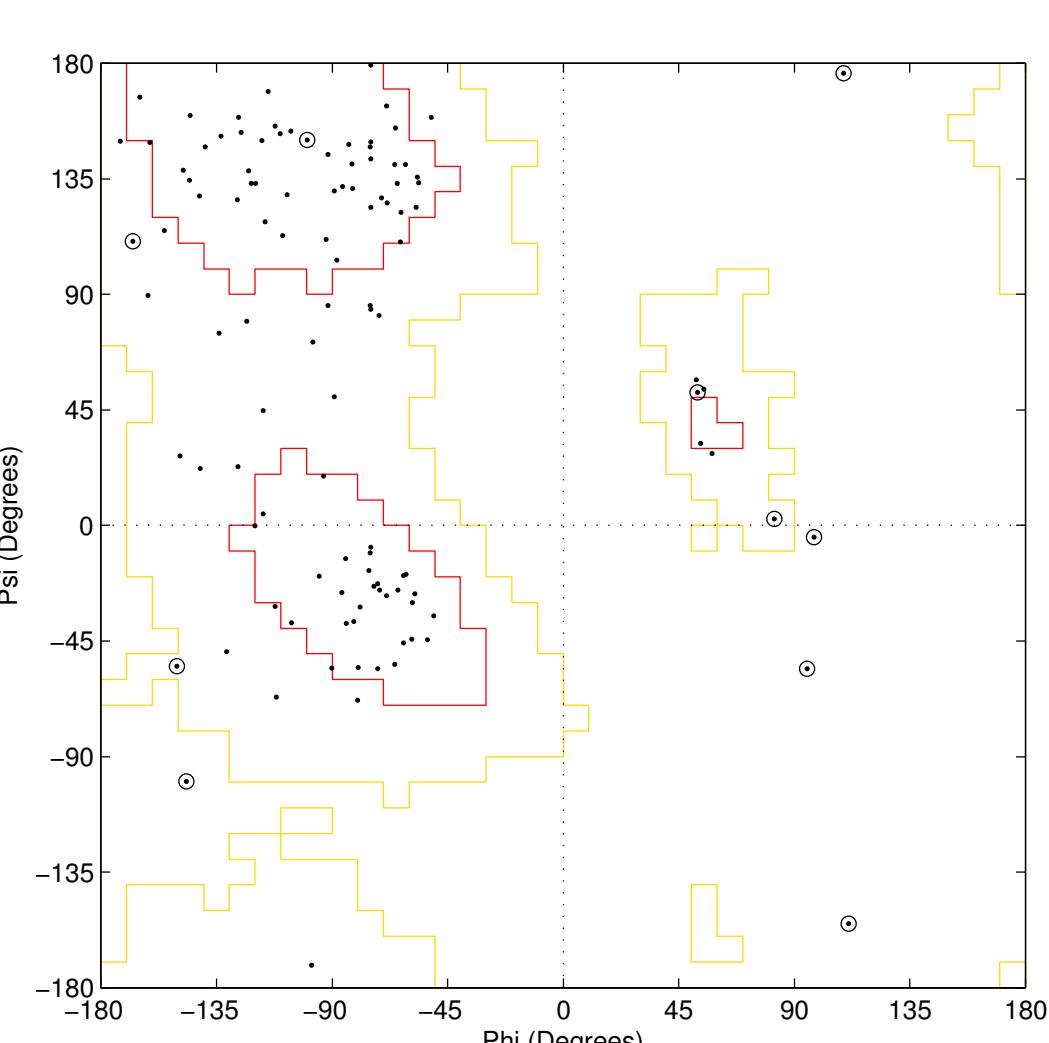
## CONCLUSION

- Follows a divide and conquer approach
- Uses the natural packing of protein in core (helices and  $\beta$ -sheets) and free regions (loops)
- Conquer step can proceed in parallel
- Minimizes error by registering in single step
- Scalable
- Work with large data sets with inadequate constraints

## RESULT



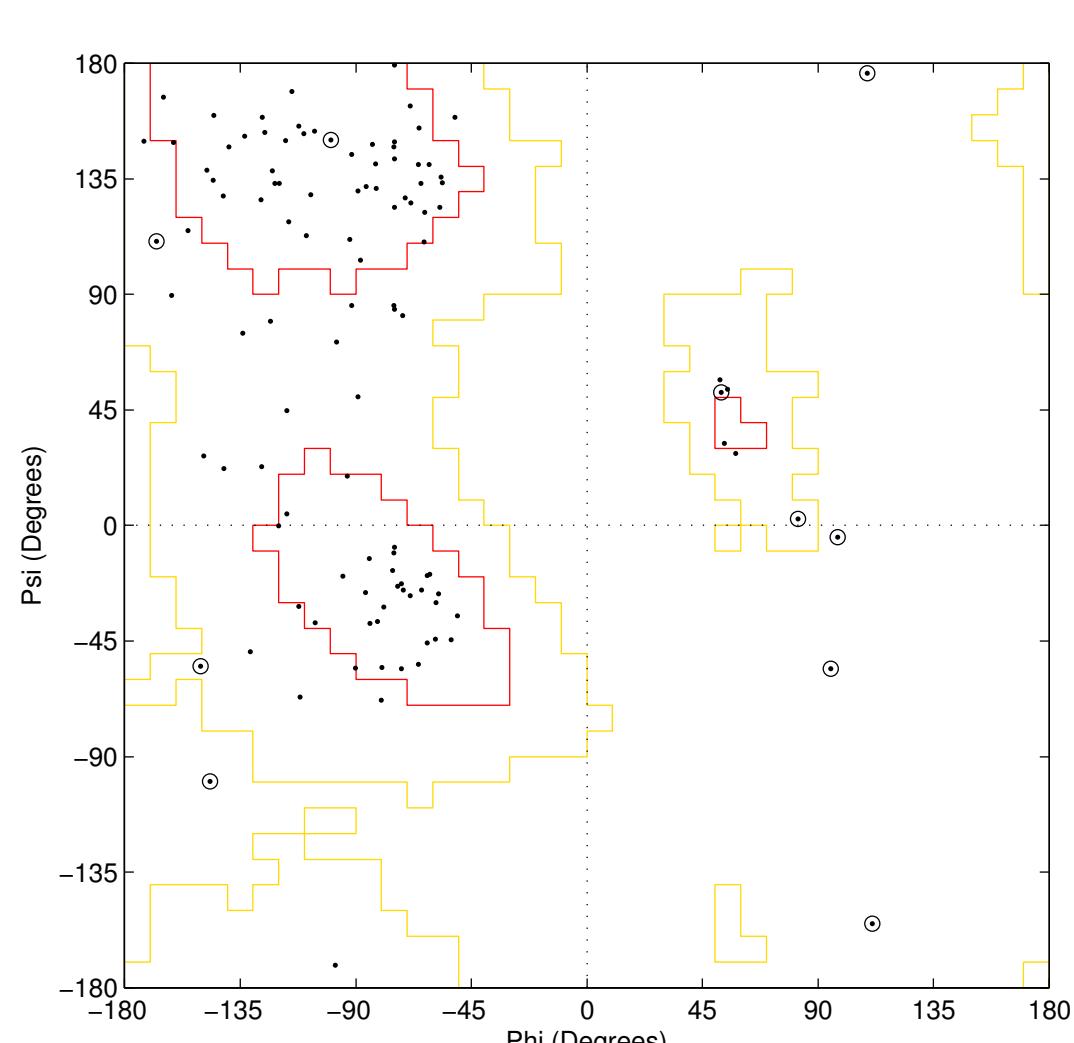
(a) CYANA.



(b) Ramachandran plot.



(c) Proposed method.



(d) Ramachandran plot.

Figure 1: Comparison of structures determined by CYANA and our method (only one conformation shown).

# 3D protein modeling from sparse distance-constraints derived from NMR experimental data

Niladri Ranjan Das

Indian Institute of Science

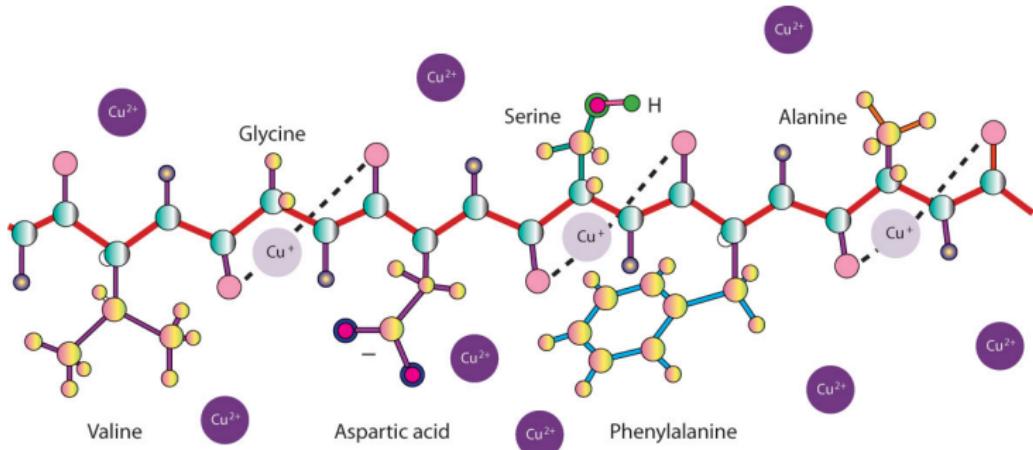
*Mathematical Science (NMI)  
EECS Symposium*

April 4, 2017



# Outline

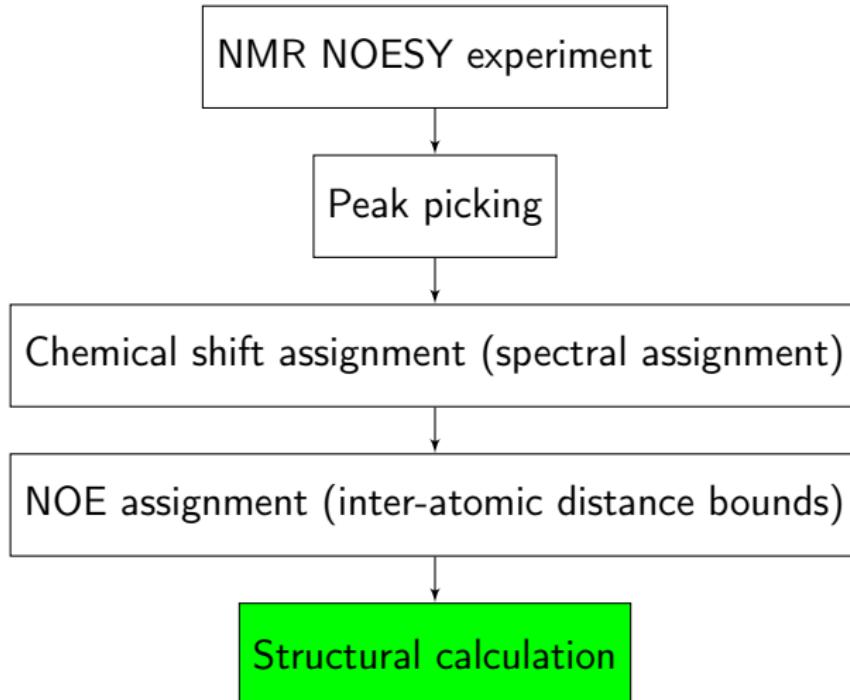
- 1 NOESY experiment
- 2 Problem Definition
- 3 Motivation
- 4 Our method
- 5 Conquer
- 6 Consolidate
- 7 Complete 3D structure
- 8 Results



**Figure:** Small fragment of a protein structure.

- NMR NOESY experiments: sparse set of bounds on inter-atomic distances.
- J coupling: bound on torsional angles.
- Derive 3-D protein structure(s) respecting the constraints.

# Brief pipeline (NMR)



**Figure:** Pipeline for obtaining protein conformation from NMR.

# Problem Definition

## Input

- Protein sequence.
- $\mathcal{E}$ : Equality bounds (covalently bonded atoms).
- $\mathcal{U}$ : NOESY and J coupling experimental data.
- $\mathcal{L}$ : NOESY, J coupling experimental data; van der Waals radii.

## Problem

Find  $X = [x_1, \dots, x_n] \in \mathbb{R}^{3 \times n}$  such that:

$$\|x_i - x_j\|_2 = d_{ij} \quad \forall (i, j) \in \mathcal{E},$$

$$\|x_i - x_j\|_2 \leq \bar{d}_{ij} \quad \forall (i, j) \in \mathcal{U},$$

$$\|x_i - x_j\|_2 \geq \underline{d}_{ij}, \quad \forall (i, j) \in \mathcal{L}.$$

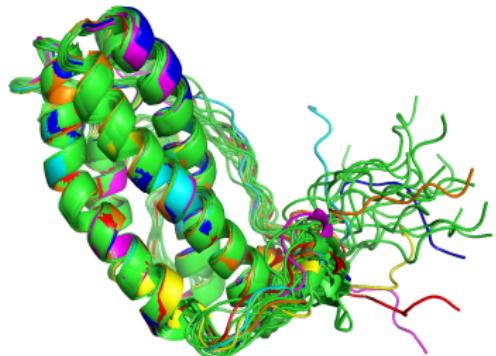
## Contemporary approaches

- Distance geometry.
- Molecular dynamics.
- Energy minimization coupled with simulated annealing.

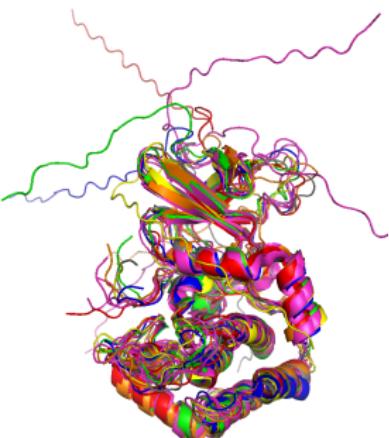
## Challenges

- ① Sparse distance constraints.
- ② Objective function: Non convex.
- ③ Conformations satisfying distance bounds are exponential.
- ④ Distance geometric approaches do not scale.
- ⑤ Molecular conformation problem is NP-hard.

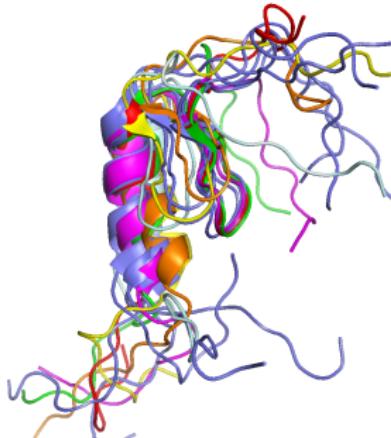
# Stable and free regions



(a) PDB ID: 1A7M.



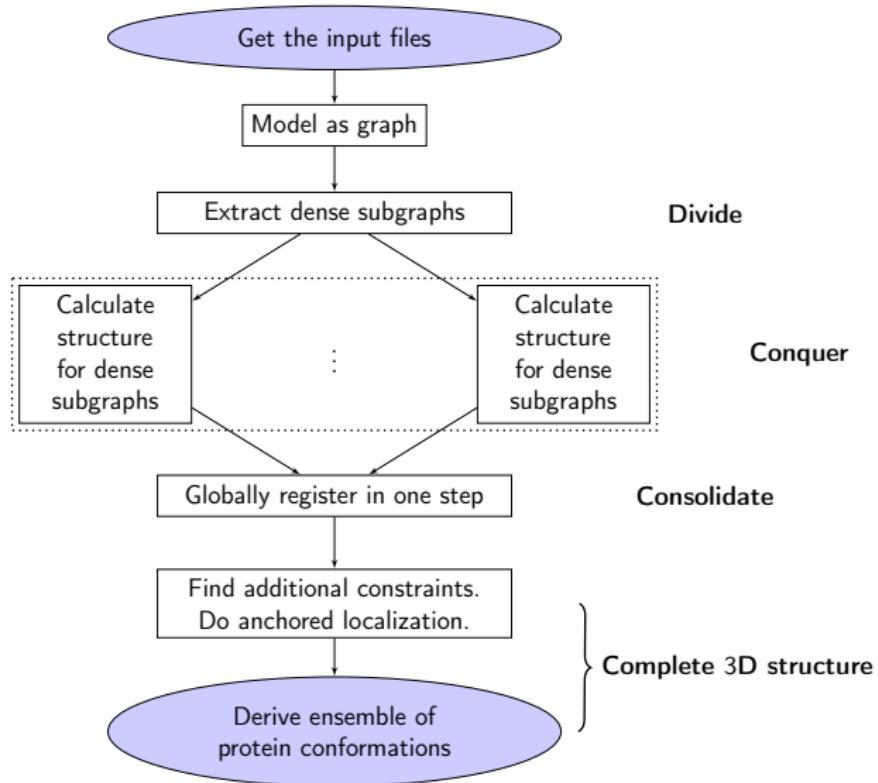
(b) PDB ID: 2KUL.



(c) PDB ID: 2YTO.

- Protein molecules contain
  - stable core regions (e.g. helices and  $\beta$ -sheets, buried regions).
  - regions which are free to move (e.g. loops).
- Formulate a divide and conquer approach.

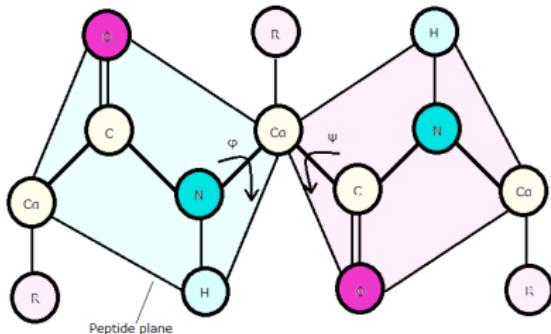
# Our method



# Residue graph.

Graph:  $G(V, E)$  such that

- $V$  : the residues in the amino acid sequence of the protein.
- $E$  :  $(v_i, v_j)$  if any atom in residue  $v_i$  shares an upper or lower bound with residue  $v_j$ .

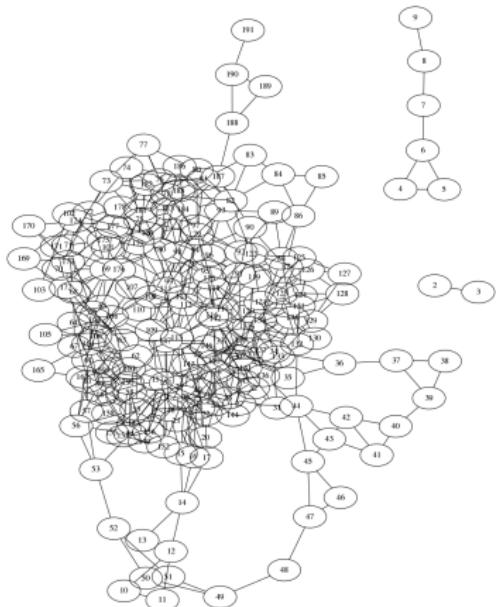


## Graph density $\eta$

Let  $S \subseteq V$ .

$$\eta(S) = \frac{|E(S)|}{|S| \times (|S| - 1)}$$

# Localization



(1) Residue graph.



(2) Structures for subgraphs.

- Extract dense subgraph from the residue graph [2]
- Solve structure for each of the subgraph

# Localization...

## Problem

- Find  $X = \{x_1, \dots, x_n\} \in \mathbb{R}^{3 \times n}$  : 3D satisfying
  - $d_{ij}$  distance bounds between  $x_i$  and  $x_j$ .
- Least square minimization of violations: Non-convex.
- Graph embedding in  $d$  dimension: known NP complete problem.

## Gram matrix

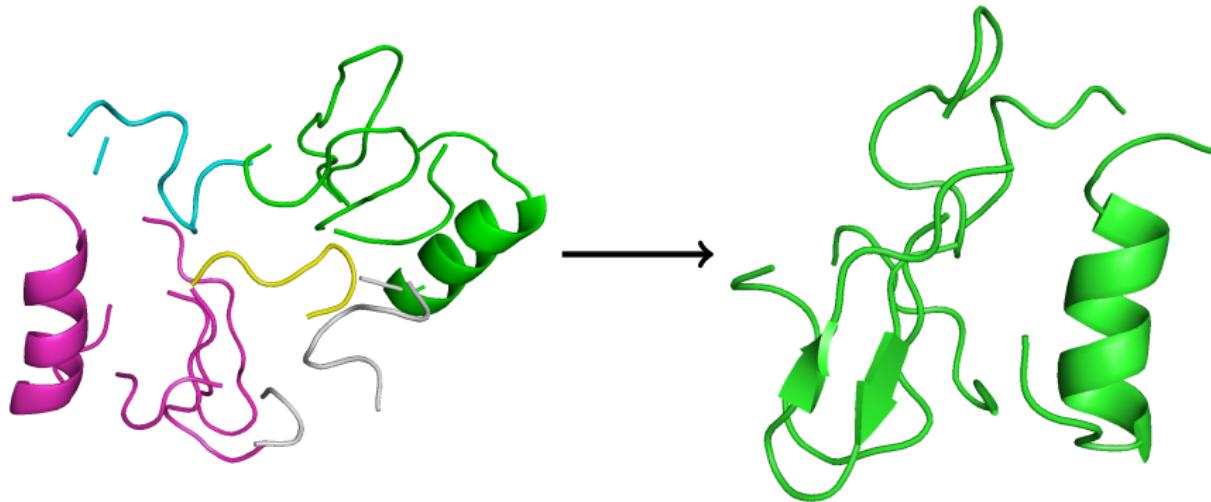
$$G := X^T X$$

$\text{Rank}(G) = \text{Rank}(x)$ .  $G \in \mathbb{S}^n$ , it is of low rank ( $d \ll n$ ).

## Low rank approximation

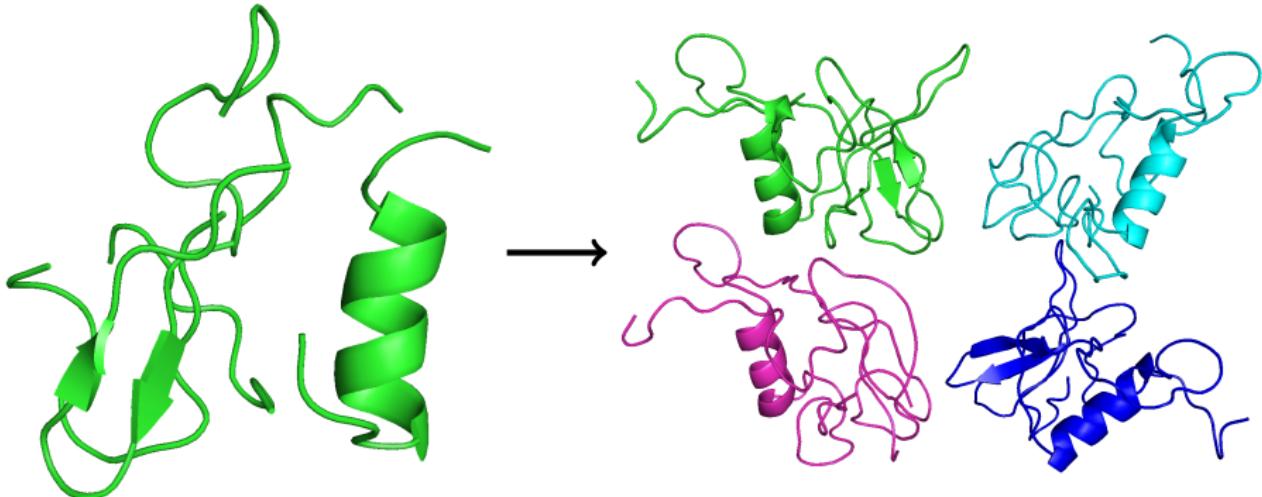
Localization problem: low-rank  $G$  satisfying the distance constraints[3][4].

# Global registration



(2) Structures for subgraphs. (5) Structure after global registration.

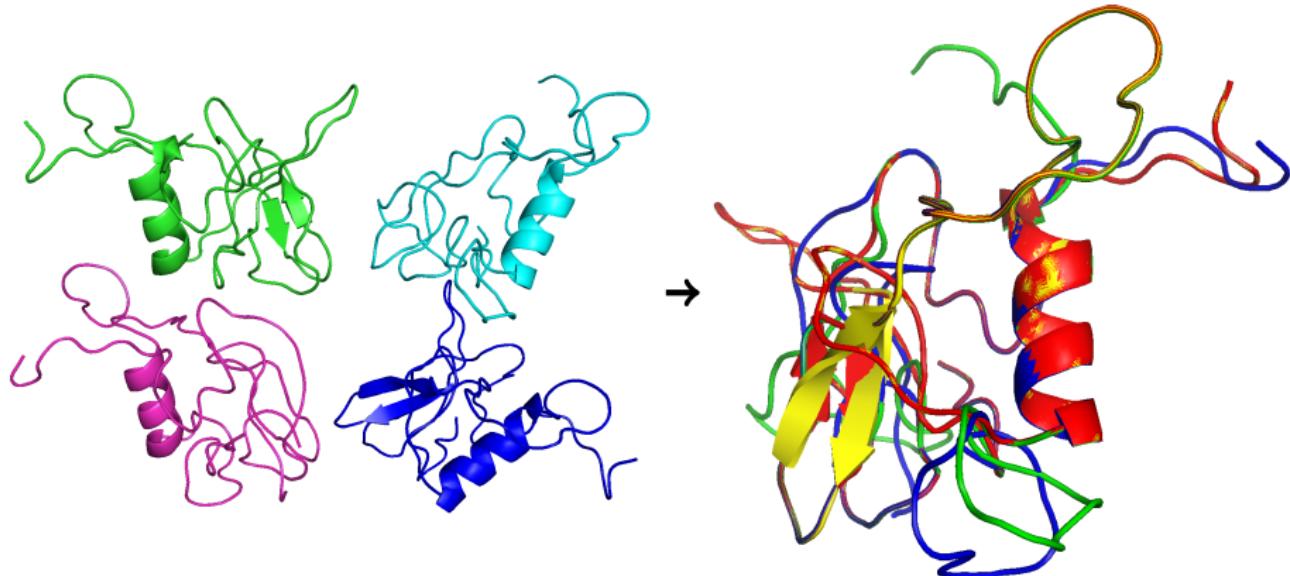
- ① Associate rigid transformation with each structure.
- ② Register the structures globally in a single step.
- ③ Use semidefinite relaxation [5].



### Fill “gaps”

- Structures for “dense” regions already calculated (anchors).
- Fill “gaps” (homology and loop modeling).
- Get additional distance bounds.
- Used as initial point for optimization.

# Ensemble of conformations

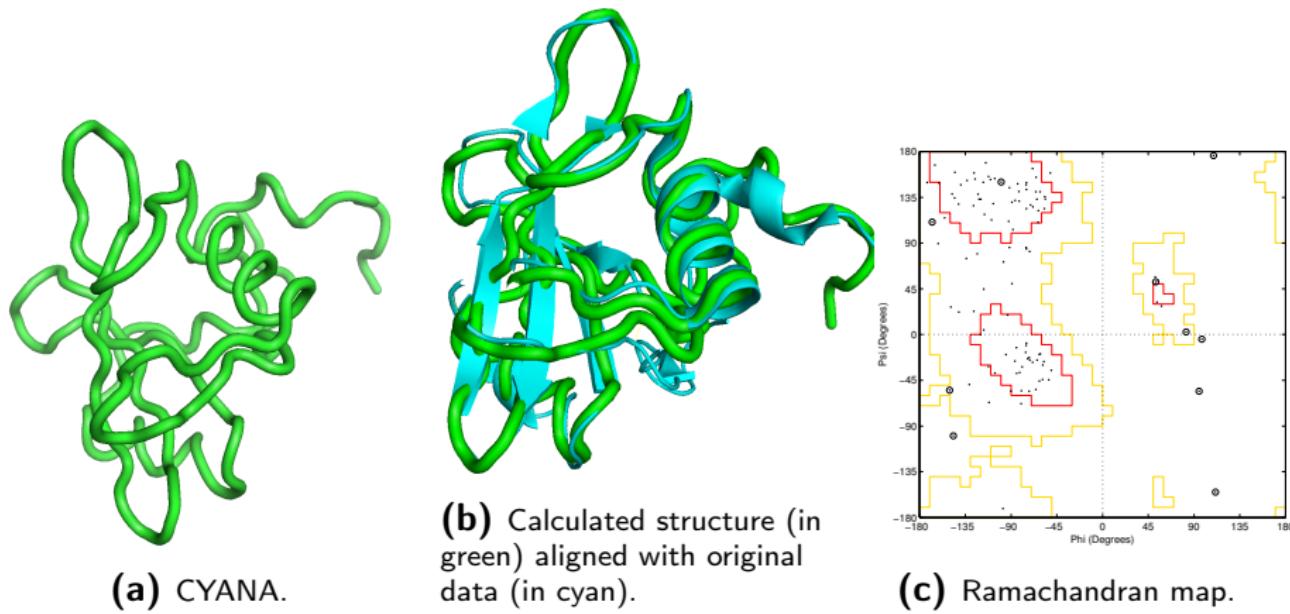


(4) Initialization for next stage.

(5) Ensemble of protein structures

- ① Each structure is an initial point for optimization.
- ② Do anchored localization to derive ensemble of structure.

# Results I

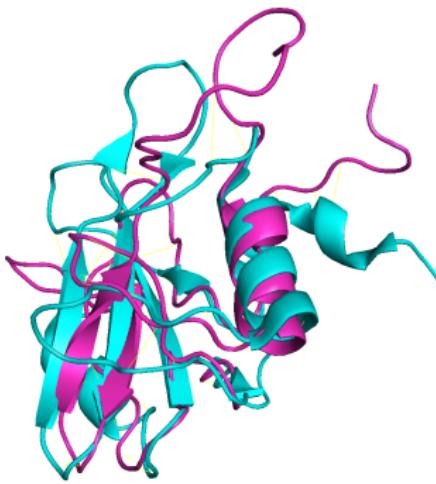


**Figure:** Structure determined by CYANA (ver 2.1).

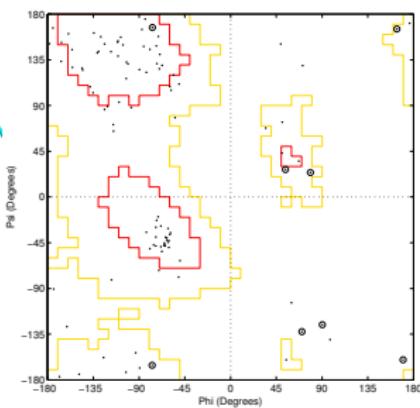
## Results II



(a) CYANA.



(b) Calculated structure (in magenta) aligned with original data (in cyan).



(c) Ramachandran map.

Figure: Structure determined by proposed method.

# Thank You

-  Alipanahi, B., Krislock, N., Ghodsi, A., Wolkowicz, H., Donaldson, L., and Li, M. *Determining protein structures from noesy distance constraints by semidefinite programming*. J. Comput. Biol. vol. 20(4), pp. 296-310 (2013).
-  Jie Chen, Yousef Saad. *Dense subgraph extraction with application to community detection*. IEEE Transactions on Knowledge and Data Engineering, 24(7), pp. 1216-1230 (2012).
-  Weinberger, K. Q., Sha, F., and Saul, L. K. *Learning a kernel matrix for nonlinear dimensionality reduction*. ICML'04: Proceedings of the twenty-first international conference on Machine learning., pp.106 (2004).
-  Biswas, P. and Ye, Y. *A distributed method for solving semidefinite programs arising from ad hoc wireless sensor network localization*. Multiscale Optimization Methods and Applications, volume 82 of Nonconvex Optim. Appl., pp. 69-84 (2006).



Chaudhury, k., Khoo, Y., Singer, A., and Cowburn, D. *Global registration of multiple point clouds using semidefinite programming.* arXiv:1306.5226, (2013).