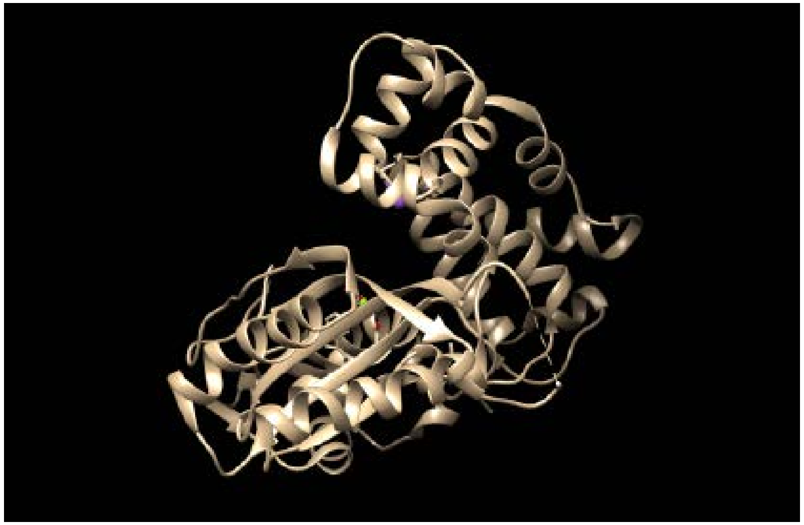


Integrating NOE and RDC using semidefinite programming for protein structure determination

Yuehaw Khoo

Stanford University

Joint work with:

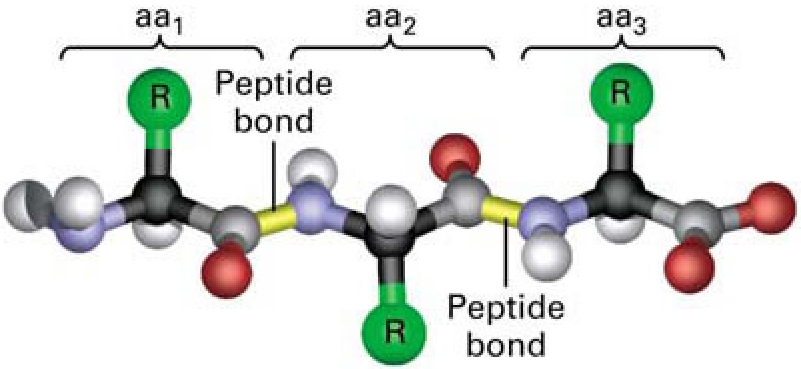
- Amit Singer (Princeton University)

- David Cowburn
  (Albert Einstein College of Medicine)

NMR Spectroscopy

Nuclear Magnetic Resonance

- Goal: Determine the position of every atom in the protein
- From chemistry:
  - Amino-acid sequence
  - Bond lengths, bond angles
  - …



- From NMR data:
  - NOE: Pairwise hydrogen-hydrogen distances <6A
  - Torsion angles
  - RDC: Residual dipolar couplings
  - …

Protein structural calculation

- Geometric constraints between atoms from NMR spectra

Structure calculation problem

Find 3D coordinates of atoms satisfying geometric constraints.

Classical approach: Distance geometry I

- NOE spectra provides pairwise distances d_{kl} of atoms (Wuthrich 82)

- Find coordinates $X = [x_1, \dots, x_K] \in \mathbb{R}^{3\times K}$ for K atoms such that

$$(d_{kl}^{low})^2 \leq || \, x_k - x_l \, ||_2^2 \leq (d_{kl}^{up})^2$$

- Non-convex. Easy when having complete measurement
  - Classical multidimensional scaling (Shoenberg 35)

Classical approach: Distance geometry II

- Global optimization
  - Xplor-NIH: Simulated annealing (Schwieters et al. 02), Majorize-minimize (De-Leeuw 77), DGSOL: Gaussian smoothing (More & Wu 99), Branch and Prune (Liberti et al. 07)….
- Convex relaxation
  - SDP relaxation on Gram matrix (Alfakih et al. 99, So & Ye 06, Biswas et al. 07)

Classical approach: Distance geometry III

- Speed up for protein structuring:
  - Divide-and-conquer: ABBIE (Hendrickson 95), DISCO (Leung et al. 07), 3D-ASAP (Cucuringu et al. 12), GRET (Chaudhury, K., Singer 15)
  - Semidefinite facial reduction (Krislock & Wolkowicz 10, Alipanahi et al. 12)
- From sensor network localization:
  - Divide-and conquer: PATCHWORK (Koren et al. 05), LRE (Singer 08), ARAP (Zhang et al.).
  - Further relaxation: ESDP (Wang et al. 07)

Classical approach: Distance geometry IV

- And many more…

Bad news

- Large molecules:
  - Missing NOE peaks
  - Wrong resonance assignments
  - Few or wrong NOE distances (Xu et al. 06)

- Good news – other data:

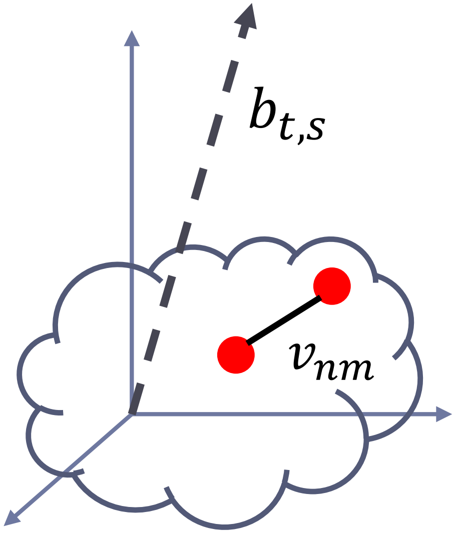
Residual Dipolar Coupling (RDC)

Residual Dipolar Coupling I

- RDC:

$$r_{nm} = < \left(v_{nm}^T b\right)^2 - \frac{1}{3} >_{t,s}$$

  - b: unit vector for magnetic field direction
  - v_{nm}: direction between atoms (n, m)



- Fix a coordinate frame such that molecule is static:

$$r_{nm} = v_{nm}^T S v_{nm}$$

$$S = < bb^T >_{t,s} - \frac{I}{3} \in \mathbb{R}^{3\times 3}$$

Residual Dipolar Coupling II

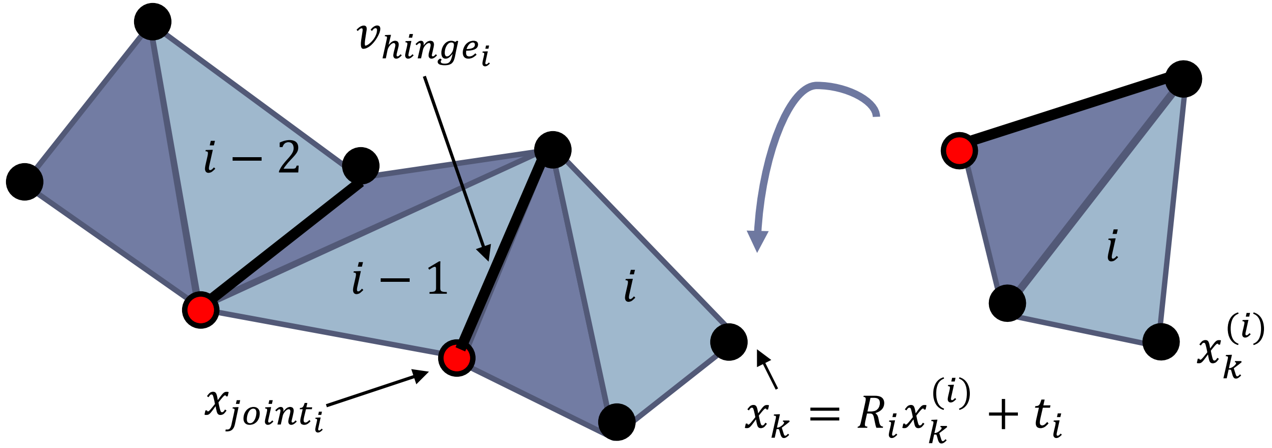
- In principle, both Saupe tensor S and all v_{nm}'s are unknown.

- Assuming S can be pre-estimated.

- Bond directions v_{nm} depends on the underlying protein structure.

Use RDC to get protein the structure

Molecule Structure from RDC



- Model protein as M rigid units chained together by hinges.

- Structure of each unit is known: Know local coordinate of point k in i-th rigid-unit, $x_k^{(i)}$

- Global coordinate of atom k in i-th unit is determined by rotations:

$$x_k = R_i \left(x_k^{(i)} - x_{joint_i}^{(i)} \right) + x_{joint_i}\,, \qquad R_i \in SO(3)$$

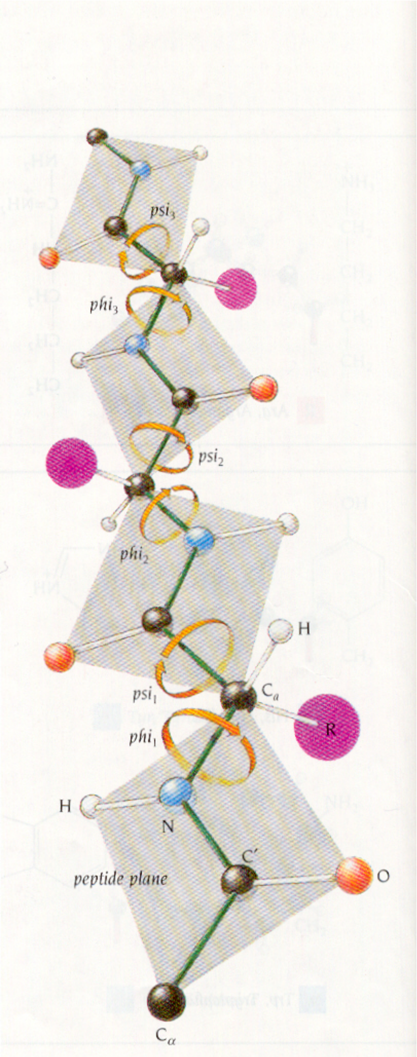
Molecule Structure from RDC

- Determine R_i from RDC gives structure

- $v_{nm} = R_i v_{nm}^{(i)}$
  - $v_{nm}^{(i)}$: bond (n, m) in the i-th rigid unit frame

- $r_{nm}^{(j)} = v_{nm}^{(i)\,T}\; R_i^T\; S^{(j)}\; R_i\; v_{nm}^{(i)}.$
  - Superscript j: RDC for different alignment

- R_i's are not independent
  - $R_i v_{hinge_i}^{(i)} = R_{i-1} v_{hinge_i}^{(i-1)}, \quad i = 2, \dots, M$



Previous works

- Degeneracy with one Saupe tensor (Hus et al. 07)
- Mainly used for refinement.
- Branch and Prune (Zeng et al. 09), Torsion angle sampling (Bryson et al. 08)…

Optimization over rotations

- Minimize w.r.t. R_i

$$\sum_{j=1}^{N}\sum_{i=1}^{M}\sum_{(n,m)\in E_{RDC_i}} \left| v_{nm}^{(i)\,T} R_i^T\, S^{(j)}\, R_i\, v_{nm}^{(i)} \; - r_{nm}^{(j)} \right|$$

- Hinge constraints: $R_i v_{hinge_i}^{(i)} = R_{i-1} v_{hinge_i}^{(i-1)}$

- Cost and domain (product of $SO(3)$) non convex, search space exponentially large

- Propose convex relaxation to quadratic problem on $SO(3)$.

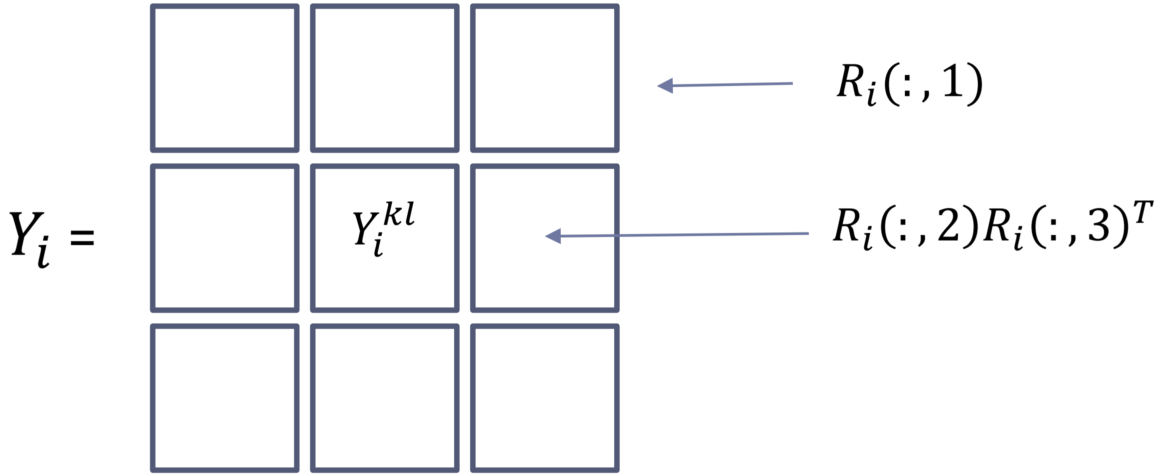
SDP relaxation: RDC-SDP

- Main idea: Cost is convex in the rank 1 PSD variable

$$Y_i = vec(R_i)vec(R_i)^T \in \mathbb{R}^{9\times 9}$$

- Write optimization problem in Y_i and relax to

$$Y_i \succcurlyeq vec(R_i)vec(R_i)^T$$



SO(3) constraint

- $R_i^T R_i = I \quad \Rightarrow \quad Tr\left(Y_i^{kl}\right) = \delta_{kl}$
- $R_i R_i^T = I \quad \Rightarrow \quad \left(Y_i^{11} + Y_i^{22} + Y_i^{33}\right) = I$

- Treat unit quaternions as $q \in \mathbb{R}^{4\times1}$: $q^T q = 1$
  - Linear relation between rotation R and qq^T
  - $qq^T qq^T = qq^T$
  - Gives linear constraints on Y_i and R_i

Full program

$$\min_{Y_i, R_i} f(Y_i, R_i)$$

$$s.t. \qquad Y_i \succcurlyeq vec(R_i)\ vec(R_i)^T$$

$$Tr\left(Y_i^{kl}\right) = \delta_{kl}$$

$$\left(Y_i^{11} + Y_i^{22} + Y_i^{33}\right) = I_3$$

Quaternion constraints

- Hinge constraints coupled Y_i's: $R_i v_{hinge_i}^{(i)} = R_{i-1} v_{hinge_i}^{(i-1)}$,
- Redundant linear constraints on Y_i, Y_{i-1}:

$$v_{hinge_i}^{(i)\,T} R_i^T e_k e_l^T R_i \mathrm{v}_{hinge_i}^{i} = v_{hinge_i}^{(i-1)\,T} R_{i-1}^T e_k e_l^T R_{i-1} v_{hinge_i}^{(i-1)}$$

  - e_k: Canonical basis in $\mathbb{R}^3$.

Other details

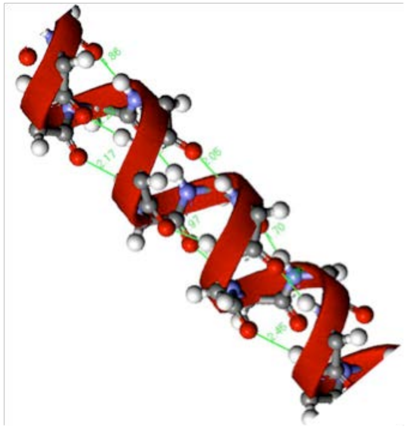
- Rounding
  - Rank 1 projection of Y_i
  - Polar decomposition (change sign if determinant<0)

- Manopt refinement (Boumal13)

- Can also incorporate distance constraints - RDC-NOE-SDP :
  - Atom coordinates linearly related to rotations
  - Can be expressed in dimension 9M gram matrix

Simulation setting

- Noise model: $r_{nm}^{(j)} = v_{nm}^T S^{(j)} v_{nm} + \sigma \epsilon_{nm}^{(j)},$
  $$\epsilon_{nm}^{(j)} \sim \mathcal{N}(0,1)$$



- Alpha helix of ubiquitin, $M = 18, K = 80.$

- Two different Saupe tensors $S^{(1)}, S^{(2)}$, 3 bonds per rigid units.

- RMSD $= \sqrt{\frac{|| X-\bar{X}||_F^2}{K}}$, $\quad X = [x_1, \dots, x_K]$

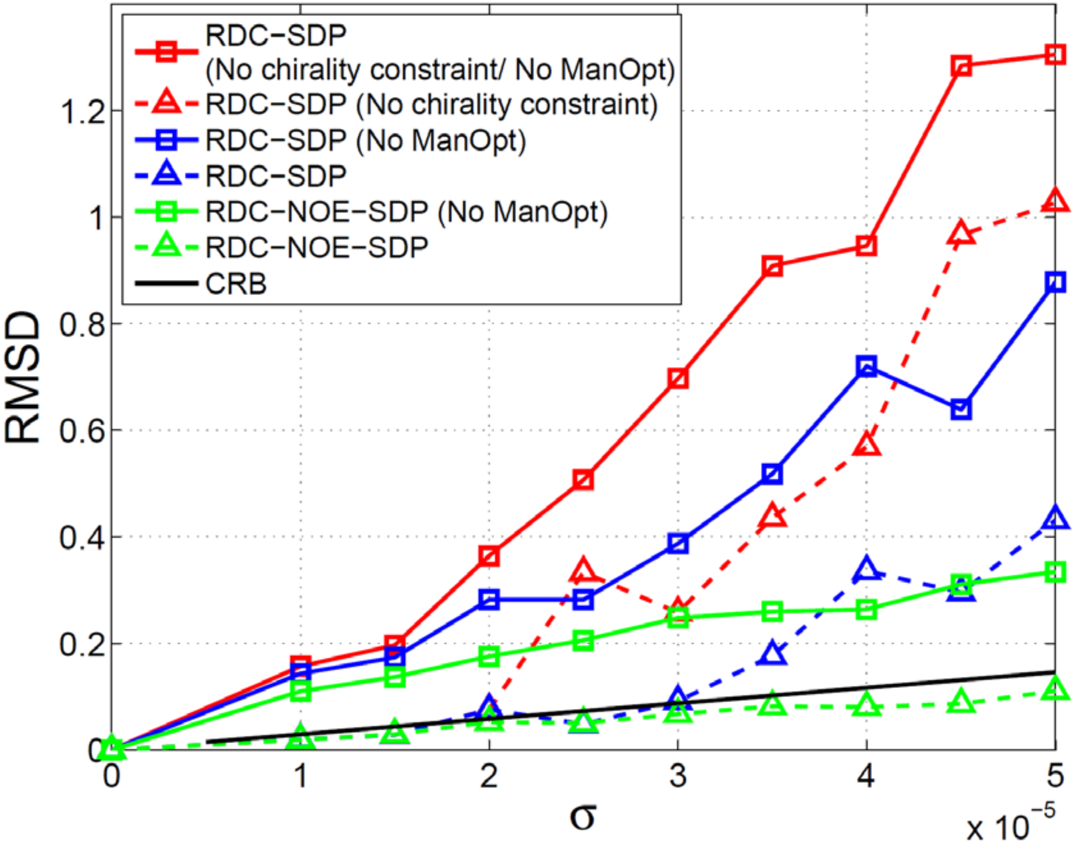
- Atomic resolution if RMSD is within 1 Angstrom

Simulation results: Atomic resolution

- $\sigma \sim 4 \times 10^{-5}$

  Realistic noise

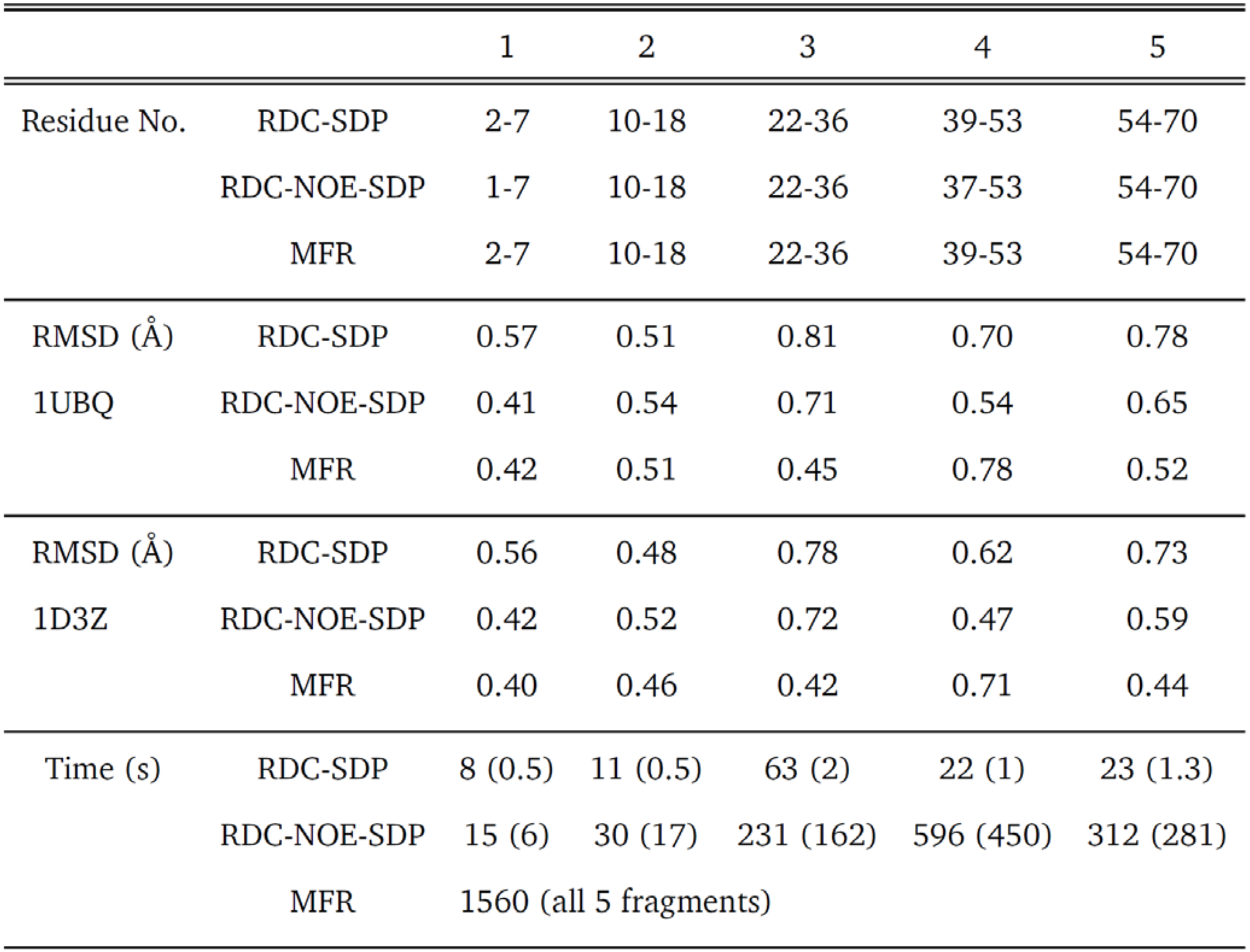
- Average over 30 realizations



Experimental data

- Ubiquitin: 500 backbone atoms
- Divide ubiquitin into 5 fragments
- Run RDC-SDP and RDC-NOE-SDP on each fragment and combine them using inter-fragment distances.

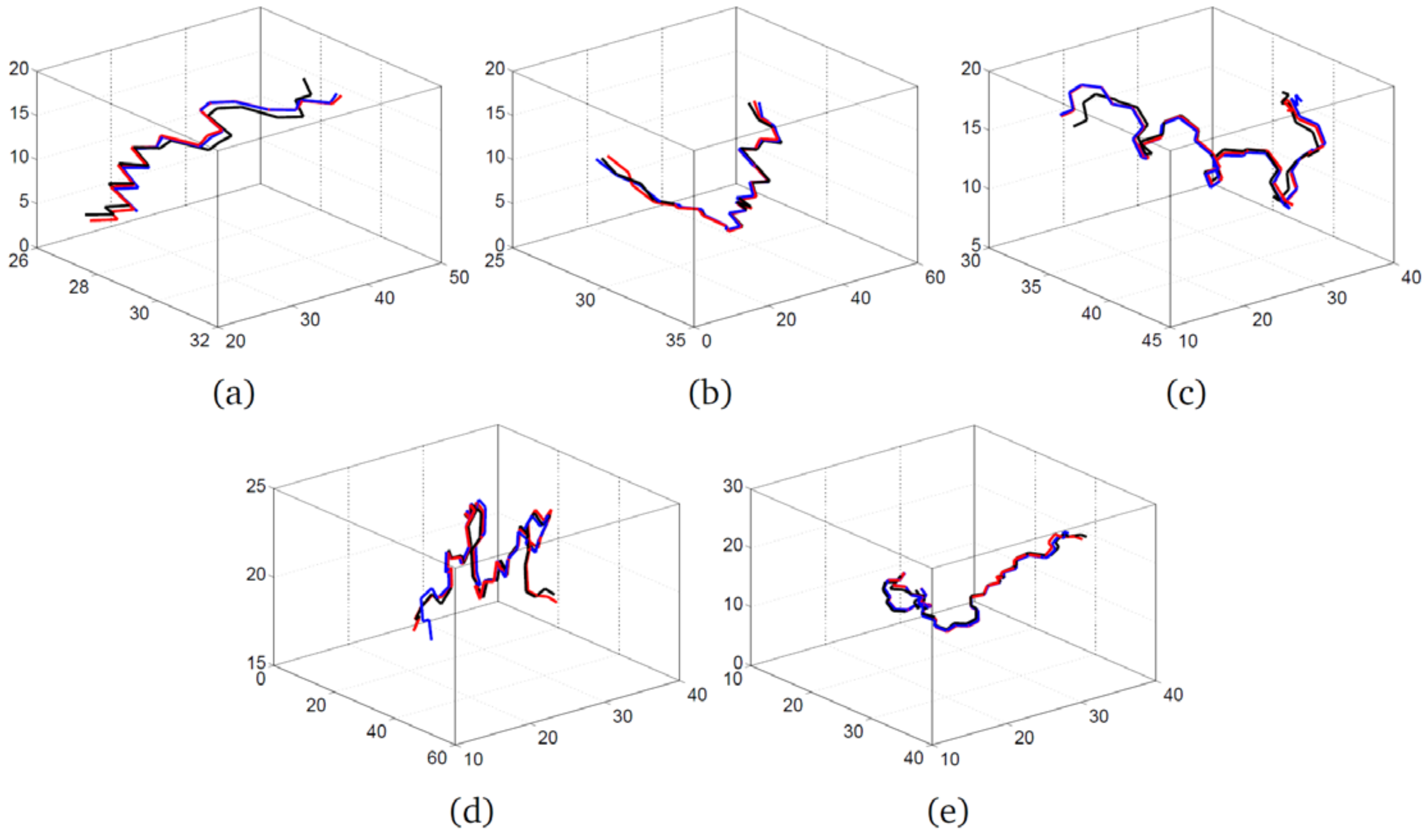
Comparison:



		1	2	3	4	5
Residue No.	RDC-SDP	2-7	10-18	22-36	39-53	54-70
	RDC-NOE-SDP	1-7	10-18	22-36	37-53	54-70
	MFR	2-7	10-18	22-36	39-53	54-70
RMSD (Å)	RDC-SDP	0.57	0.51	0.81	0.70	0.78
1UBQ	RDC-NOE-SDP	0.41	0.54	0.71	0.54	0.65
	MFR	0.42	0.51	0.45	0.78	0.52
RMSD (Å)	RDC-SDP	0.56	0.48	0.78	0.62	0.73
1D3Z	RDC-NOE-SDP	0.42	0.52	0.72	0.47	0.59
	MFR	0.40	0.46	0.42	0.71	0.44
Time (s)	RDC-SDP	8 (0.5)	11 (0.5)	63 (2)	22 (1)	23 (1.3)
	RDC-NOE-SDP	15 (6)	30 (17)	231 (162)	596 (450)	312 (281)
	MFR	1560 (all 5 fragments)				

Comparison:

- Black: X-ray structure. Blue: RDC-SDP. Red: RDC-NOE-SDP



Summary

- Y. Khoo, A. Singer, D. Cowburn, "Integrating NOE and RDC using semidefinite programming for protein structural calculation", Submitted

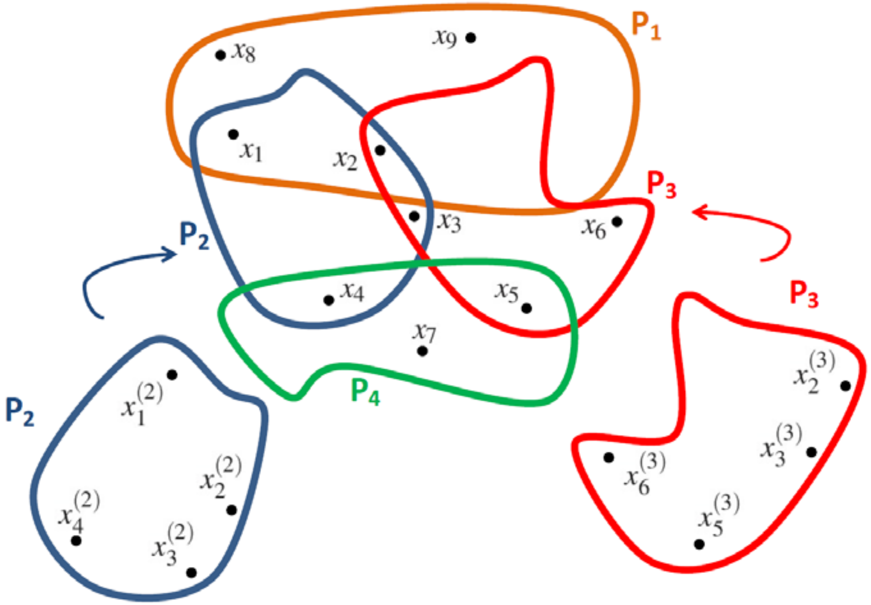
Thank you!

- Questions?

The stitching problem

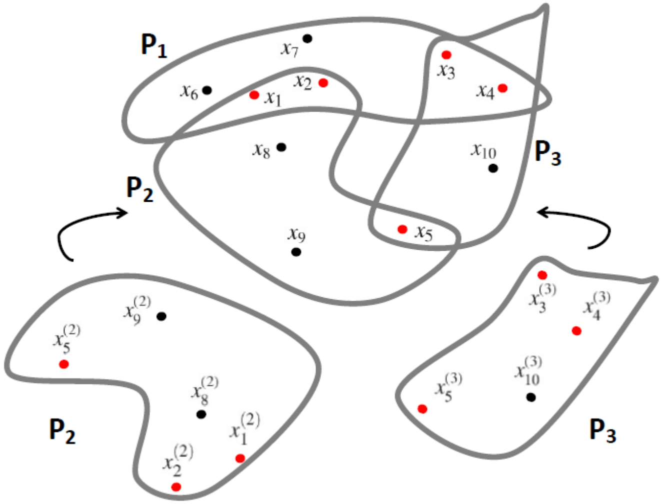
- The solution to distance geometry problem has rigid transform ambiguity
- For local coordinate $x_k^{(i)}$ of point k in fragment P_i

Find x_k, O_i, t_i such that $x_k \approx O_i x_k^{(i)} + t_i$



Global registration

- Sequential approach: greedy and may not work



- GRET: Minimize

$$\sum_i \sum_{k\in P_i} || x_k - O_i x_k^{(i)} - t_i ||_2^2$$

  - O_i : orthogonal matrix variable in d dimension.
  - x_k, t_i : free variables

GRET: Max-cut relaxation

- First order optimality condition:

$$\partial_{x_k}\phi = 0, \qquad \partial_{t_i}\phi = 0$$

  - $[x_1, \dots, x_N, t_1, \dots, t_M] = [O_1, \dots, O_M]\, A$

- Optimization solely in $O = [O_1, \dots, O_M]$:

$$\min_{O} Tr\left(CO^T O\right)$$

  - Search space non-convex and exponentially large

- Let $G = O^T O$, relaxing rank constraint:

$$\min_{G \succcurlyeq 0} Tr(CG) \quad s.t. \quad G_{ii} = I_d$$

Initial results

- RMSD $= \sqrt{\frac{\| X - \bar{X} \|_F^2}{K}}$,

  $X = [x_1, \dots, x_K]$
- Simulations ⇒

η	GRET	ASAP	DISCO
0	1.11(0.47)	1.22(0.51)	1.26(0.34)
0.2000	1.36(0.73)	1.33(0.73)	1.39(0.56)
0.4000	1.34(0.72)	1.34(0.73)	1.51(0.87)
0.6000	1.67(1.09)	1.83(1.24)	2.00(1.54)
0.8000	1.80(1.24)	2.03(1.49)	2.32(1.97)
1.0000	1.84(1.32)	1.94(1.36)	2.57(2.19)

- Real data: 1gb1 with ~1000 NOE constraints
- RMSD = 2.23Å (1.61Å for backbone), running time: 2 min