Application of robotics algorithms for interpretation and modeling dynamics of protein structures

Liangjun Zhang

NSF Computing Innovation Fellow

Computer Science Department

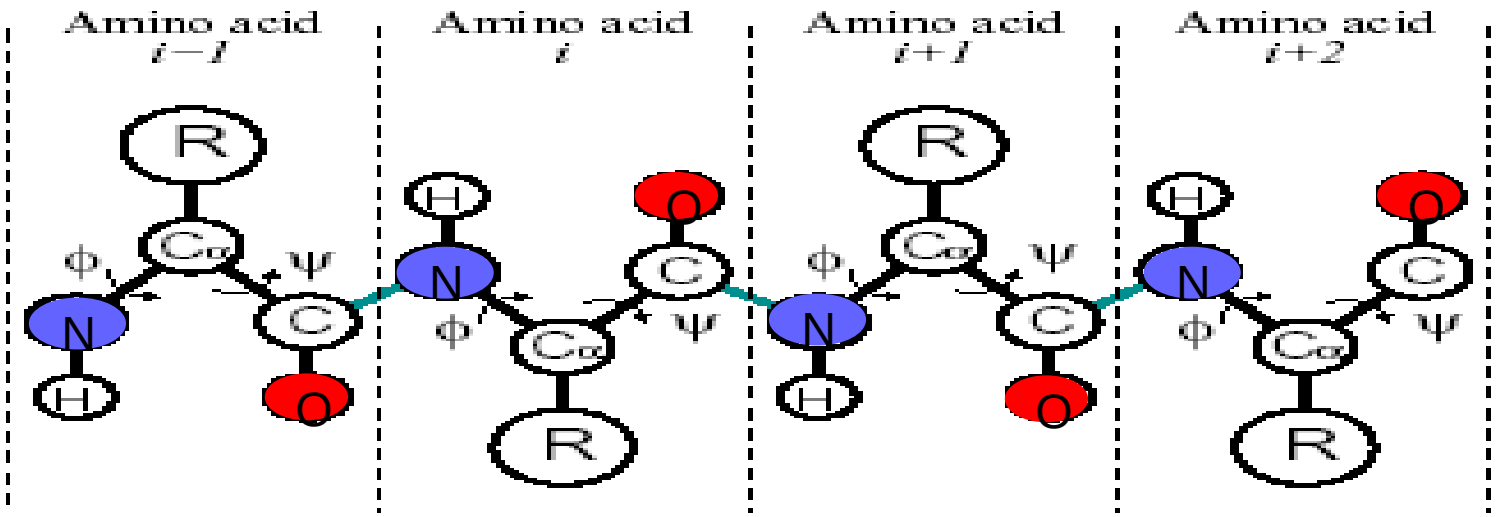
Stanford University

Work done with Dr. Peggy Yao and Prof. Jean-Claude Latombe

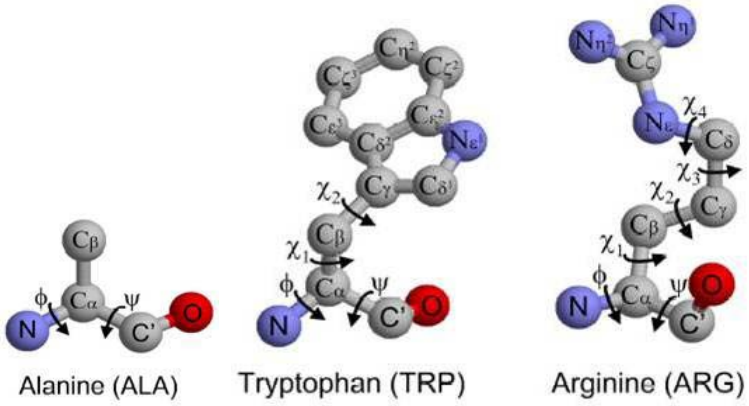
Thank Dr. Henry van den Bedem and Dr. Ashley Deacon in JCSG for the invitation

German Areospace Center's Justin robot emerging from a protein

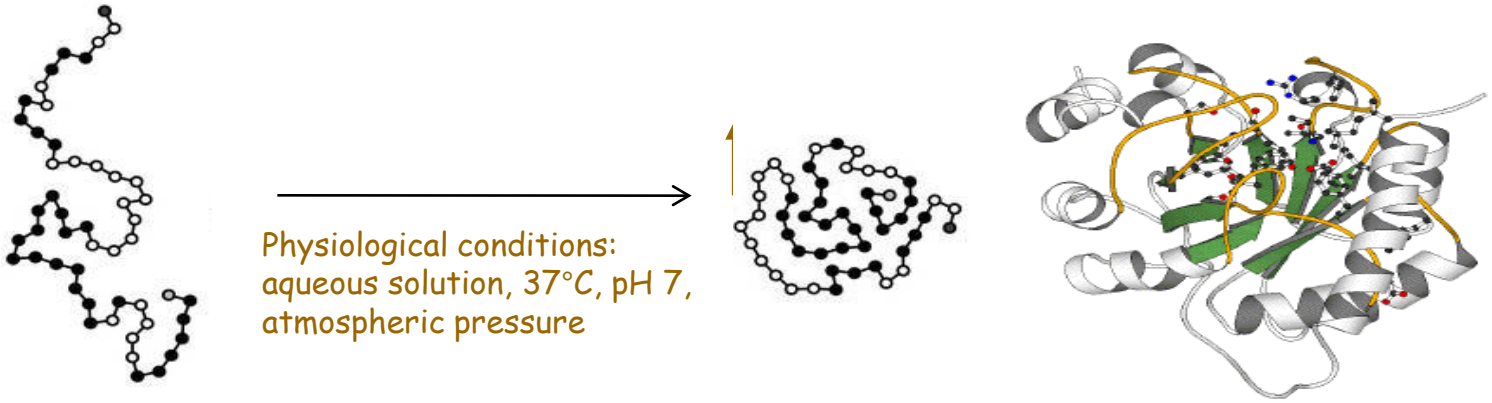
Protein Sequence



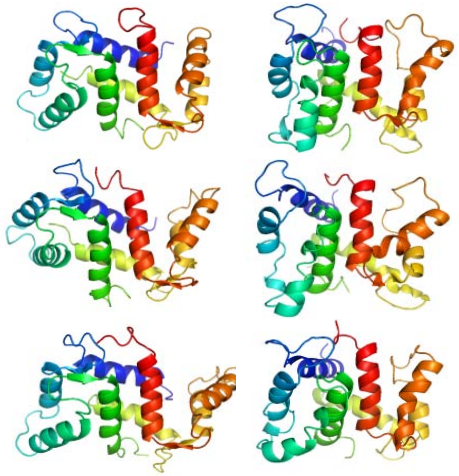
- Long sequence of amino-acids (dozens to thousands)
- Dictionary of 20 amino-acids



Protein Folding



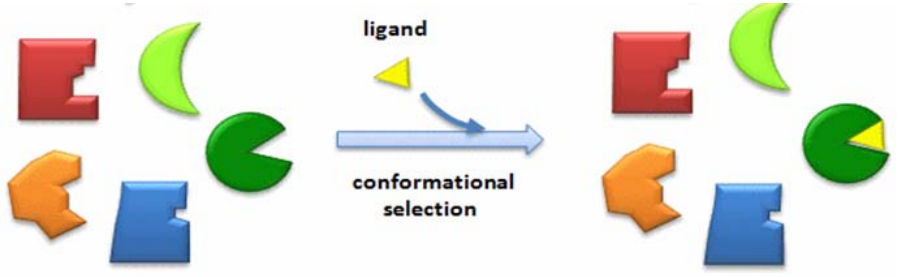
The folded structure is uniquely determined by the protein sequence but is not fully rigid

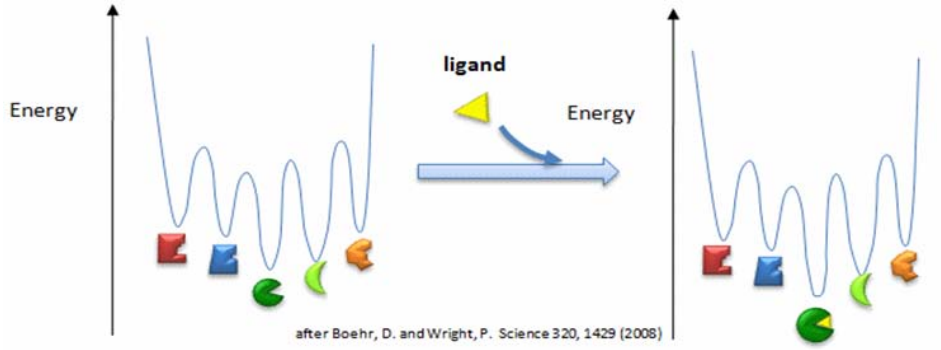


In fact, flexibility is necessary ...

... for a protein to achieve its functions by binding against other molecules (ligands)

- Conformational selection model





Why sampling folded protein conformations?

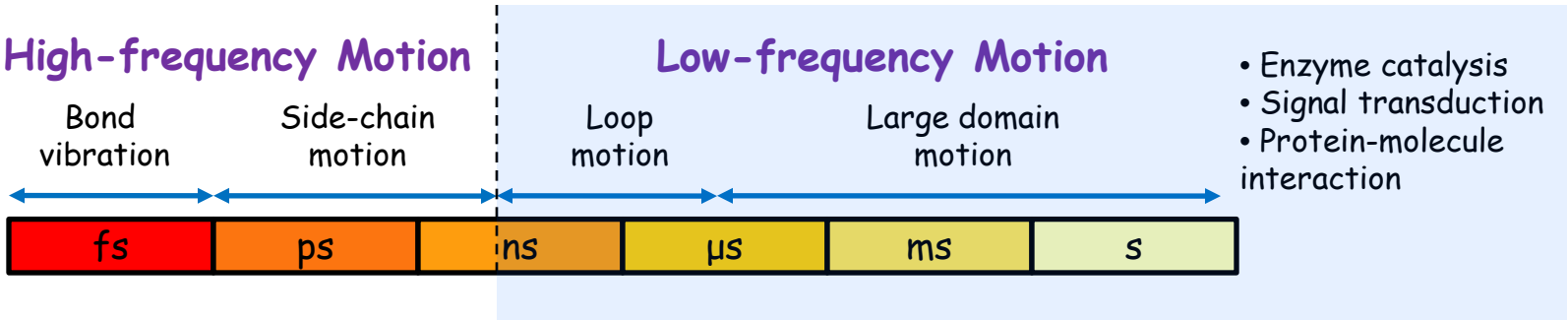
- Representation of molecular flexibility in the study of protein-ligand binding
- Interpretation of noisy experimental data (X-ray crystallography, NMR, Cryo-EM)
- Screening and design of pharmaceutical drugs
- Study of key determinants of protein stability

Approaches to Conformation Sampling

- Experimental:
  - X-ray crystallography
    → high resolution, but one or few conformations
  - NMR, Cryo-EM
    → small proteins and/or low resolution

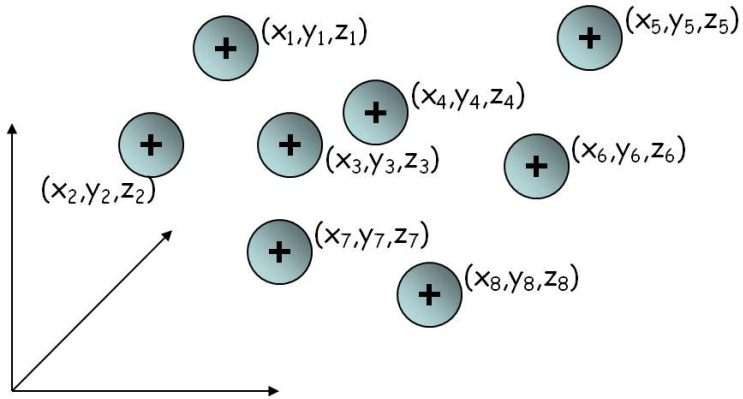
- Computational:
  - Energy-based (e.g., Molecular Dynamics and Monte Carlo simulation)
    → high computational cost
  - Kino-geometric

Kinematic Models



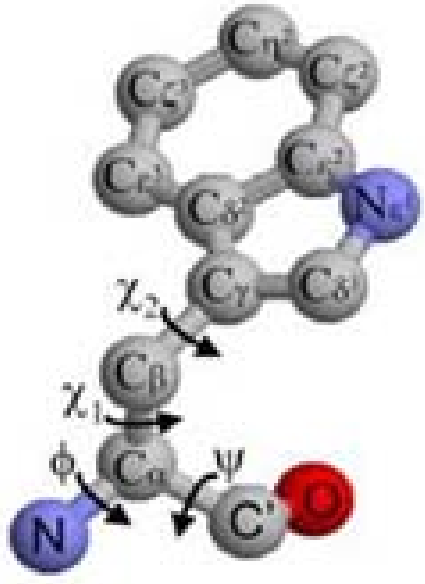
All frequencies:

Atoms can move independently

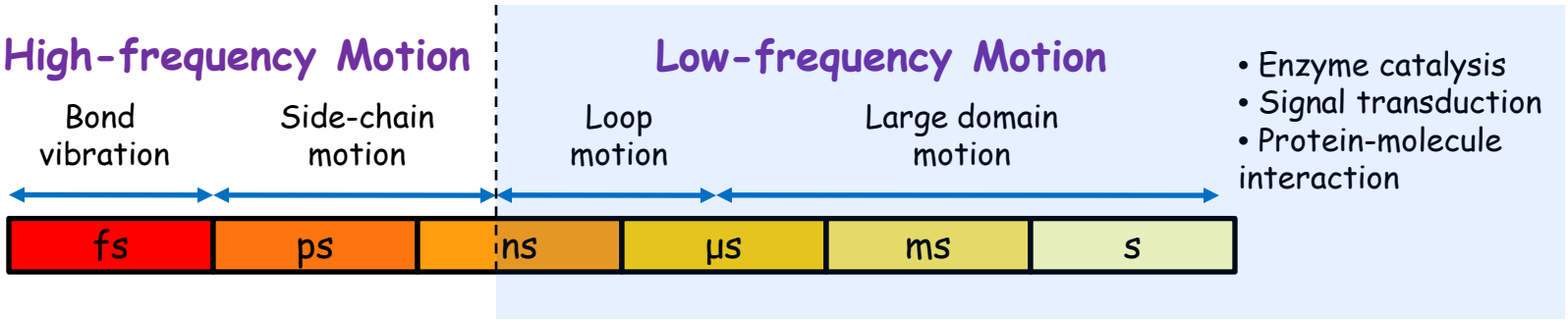


Low frequencies:

Atoms are connected by a rotatable kinematic linkage



Kinematic Models

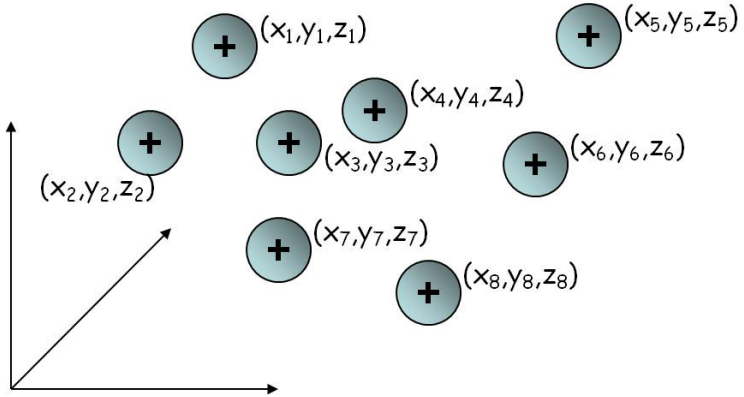


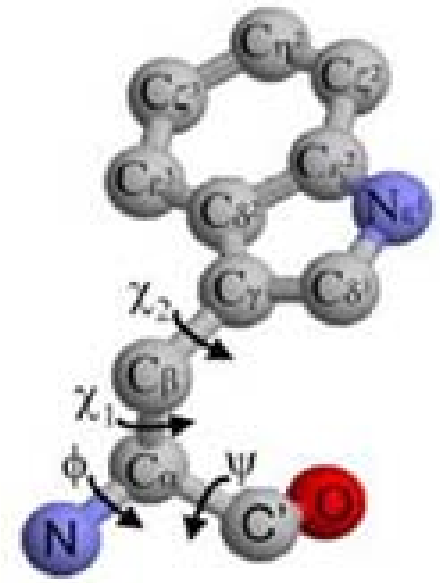
All frequencies:

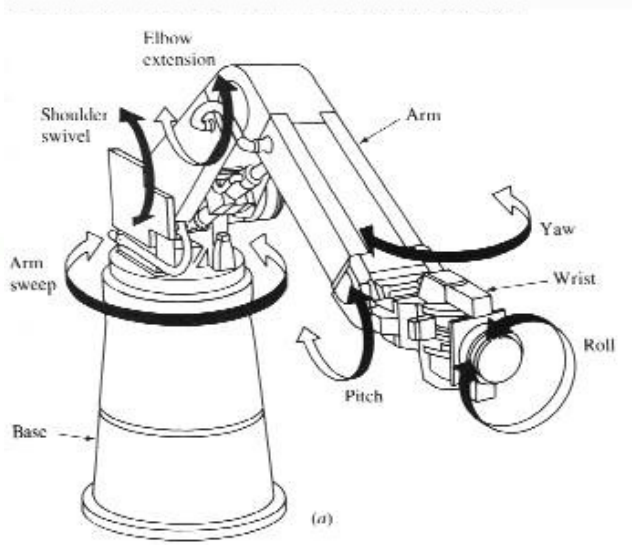
Atoms can move independently

Low frequencies:

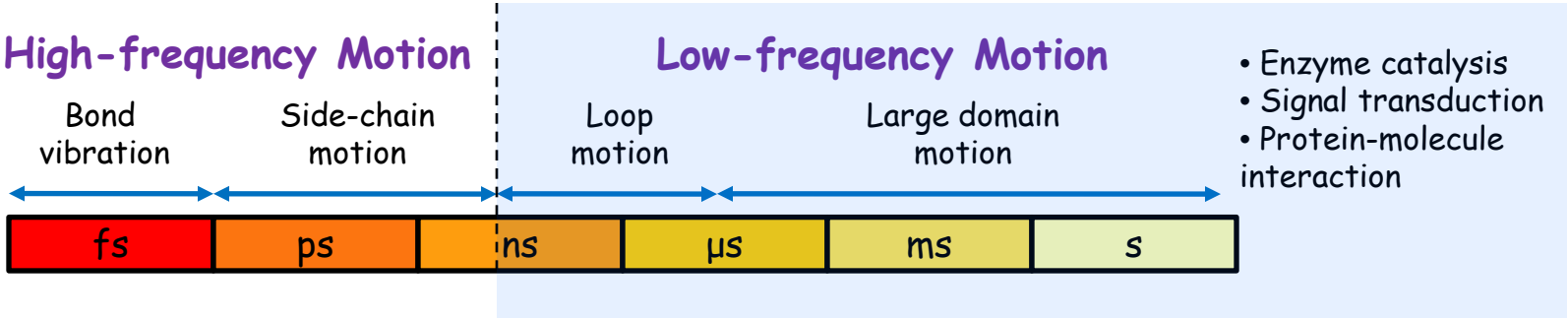
Atoms are connected by a rotatable kinematic linkage





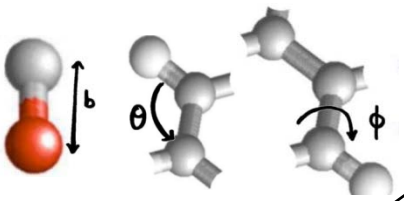


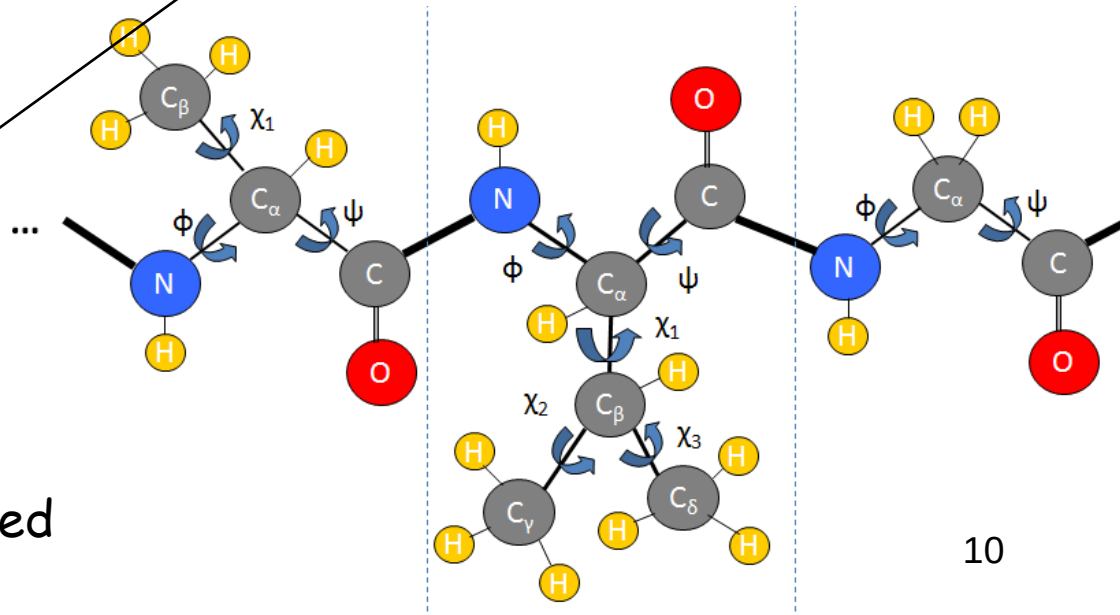
Kinematic Models



Low frequencies:

Atoms are connected by a rotatable kinematic linkage

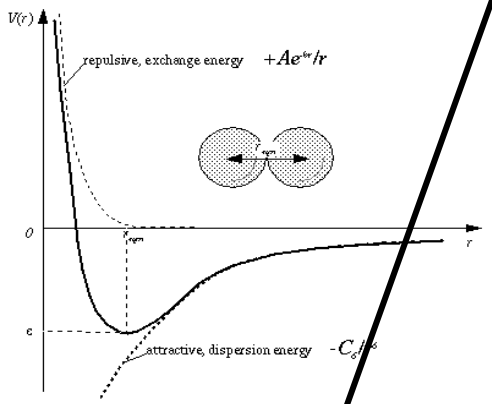




$$E = E_{bonded} + E_{non\text{-}bonded}$$

Geometric Model

- $E_{non\text{-}bonded} = E_{Coulomb} + E_{van\ der\ Waals}$



12-6 Lennard-Jones potential:

$$V(r) = 4\epsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right]$$

- In a folded conformation,
  vdW terms are dominant energy terms

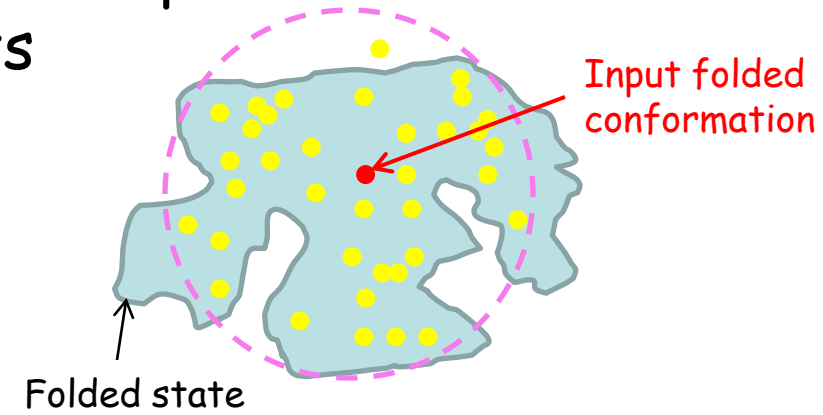
→ Atoms are modeled as hard spheres with <≈vdW radii

No two spheres may overlap

Kino-Geometric Sampling

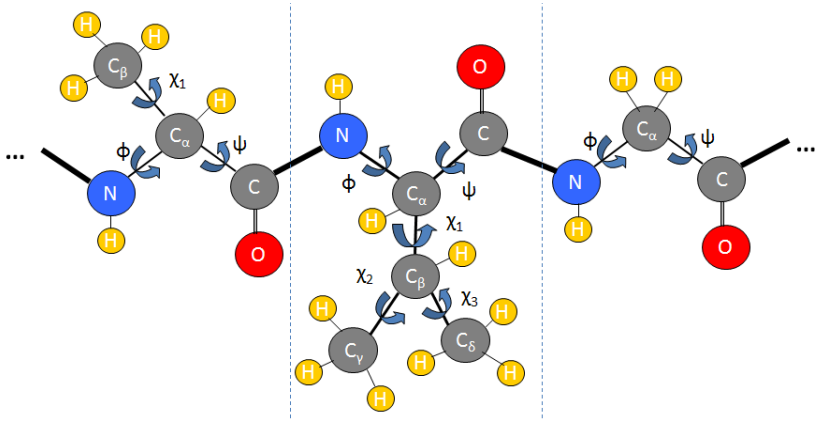
- Encode dominant energy terms of a folded conformation by means of relatively simple constraints:
  - Atoms and bonds form a kinematic linkage
    → kinematic constraints
  - Atoms are modeled as hard spheres
    → geometric constraints (volume exclusion)

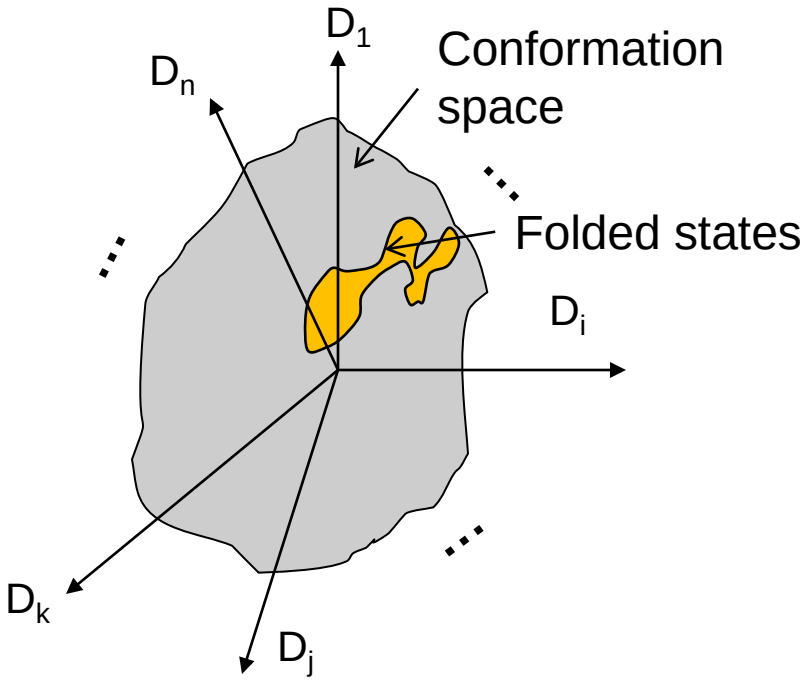
- Develop fast algorithms to sample conformations that satisfies these constraints



Computational Challenges

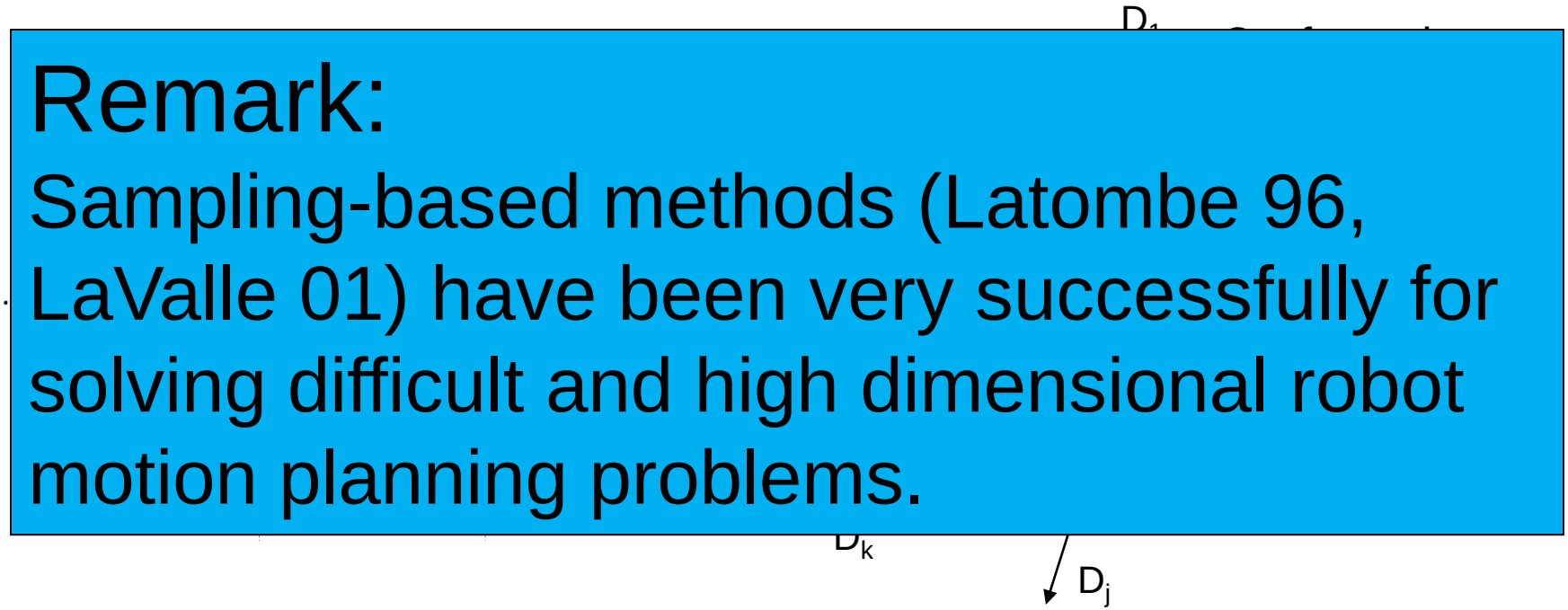
- High dimensionality, but small volume of the folded state





Computational Challenges

- High dimensionality, but small volume of the folded state



Remark:

Sampling-based methods (Latombe 96, LaValle 01) have been very successfully for solving difficult and high dimensional robot motion planning problems.

KGS Algorithm

for Sampling Folded Protein Conformations

Initialize distribution Δ to $\{q_{\text{init}}\}$

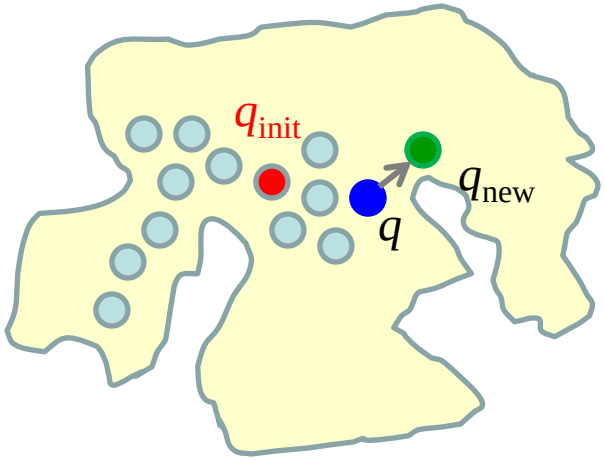
Given folded conformation, e.g., obtained from the Protein Data Bank (PDB)

Pick q from Δ with probability inverse to local sampling density

Make sure to diffuse well throughout the folded state

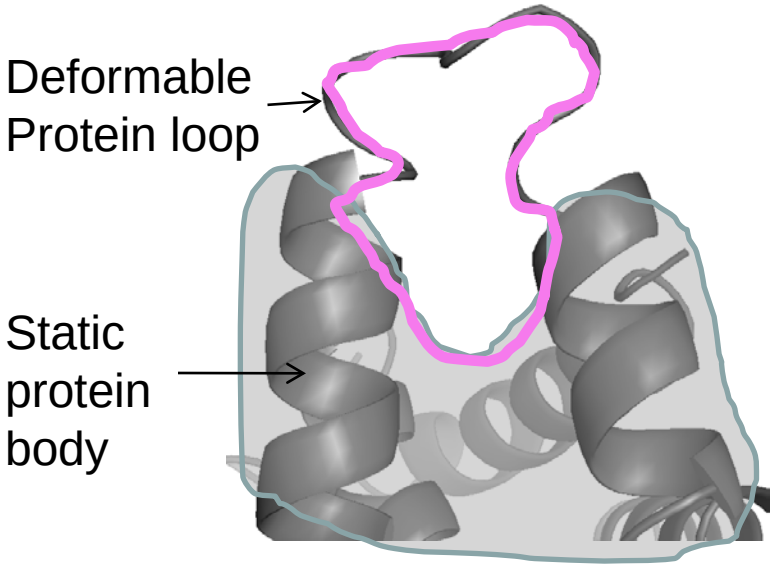
Deform q into q_{new} by perturbing angles ϕ, ψ, and χ of kinematic linkage at random

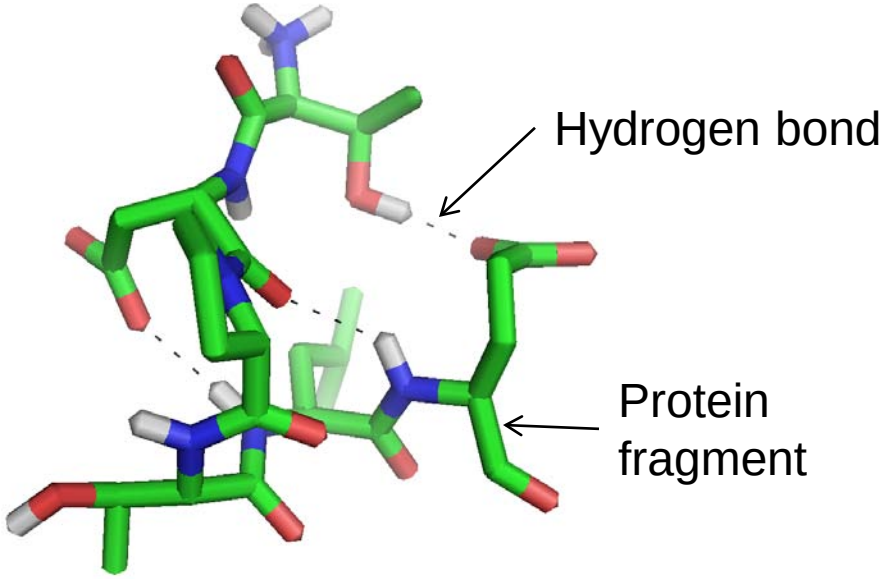
Test collisions in q_{new}.
If no collision insert q_{new} into Δ



Cycle Closure Constraints

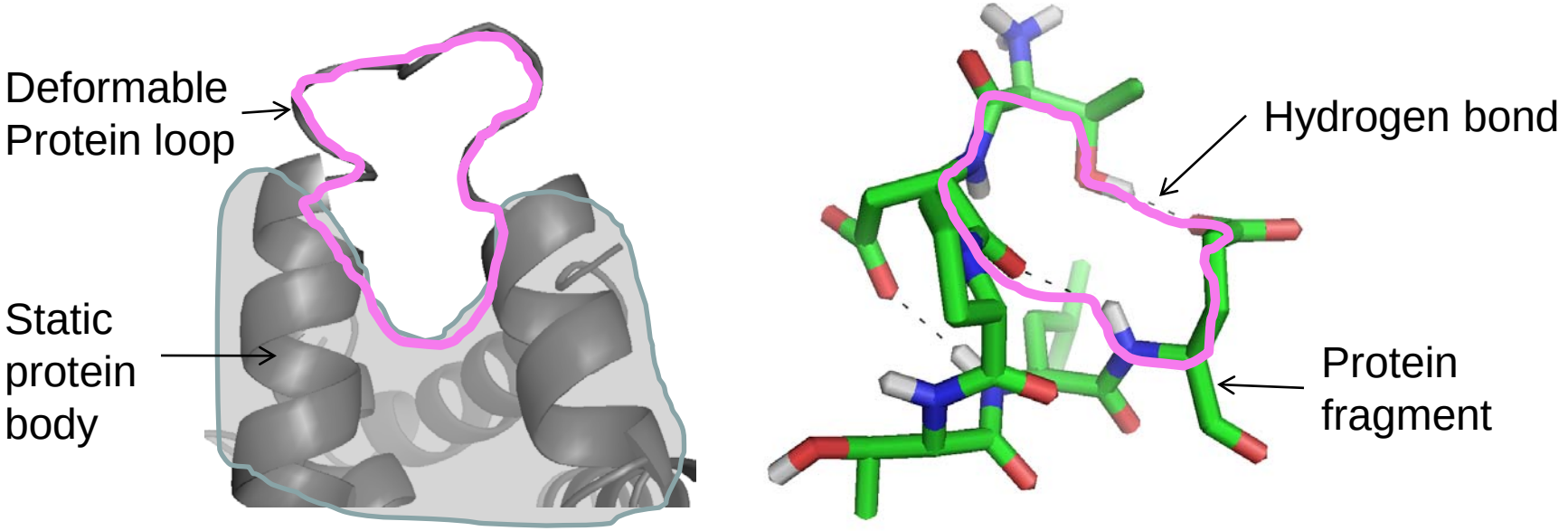
- **The linkage model may contain multiple closed deformable kinematic cycles**
  - Keep cycles closed during deformation
  - Move in tangent space





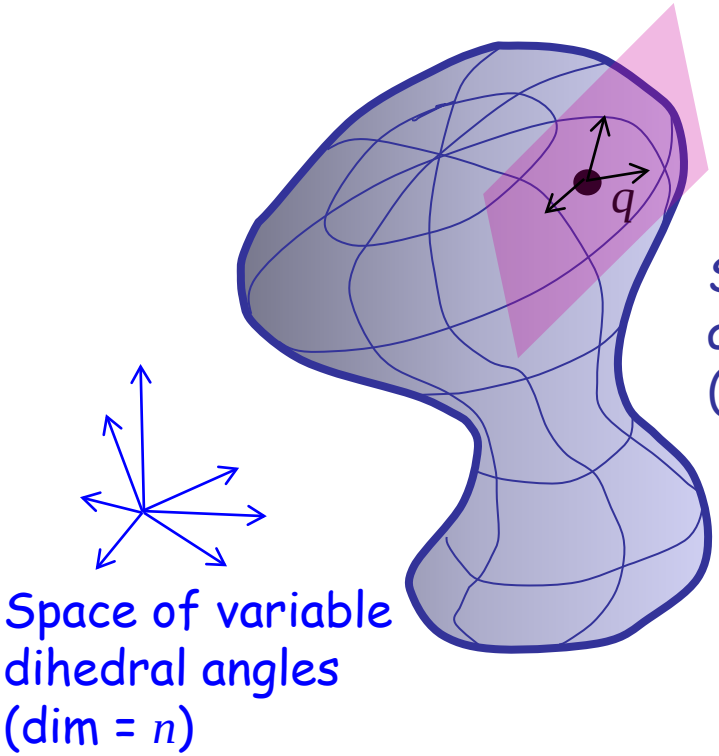
Cycle Closure Constraints

- **The linkage model may contain multiple closed deformable kinematic cycles**
  - Keep cycles closed during deformation
  - Move in tangent space



Tangent Space

- Cycle closure constraints → $F(q) = 0$
- Differentiation → $J_F \times dq = 0$
- SVD of J_F → Basis of tangent space at q

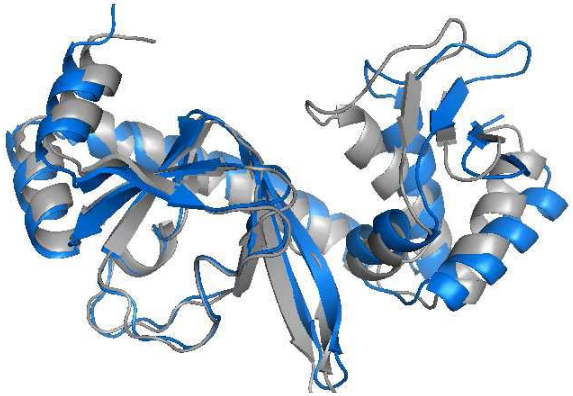


Subspace of rotatable angles that keep all cycles closed
(dim = $n - 5{\times}m$
where m = # non-redundant cycles)

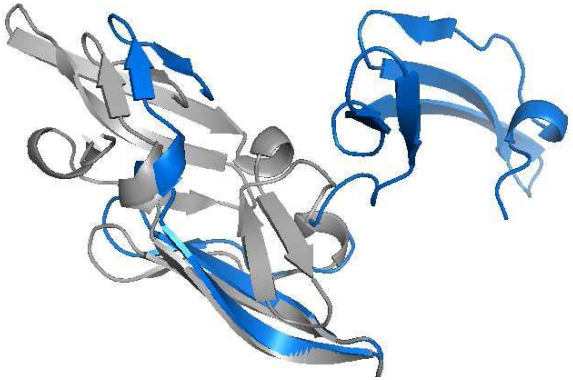
Space of variable dihedral angles
(dim = n)

→ fast + relatively large steps

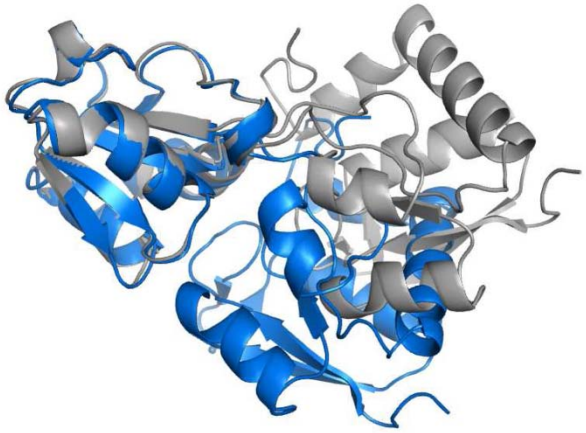
Test Proteins



1G6N, 3211 atoms, 1110 rotatable bonds
target conformation (blue) at RMSD 2.66Å from given conformation (grey)
Hinge motion



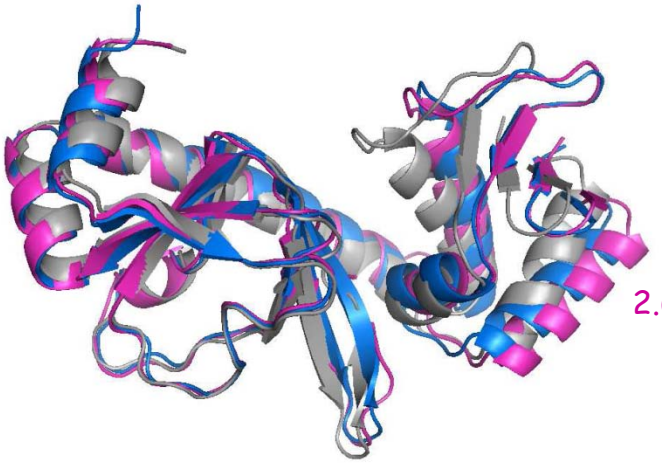
2EZM, 992 atoms, 502 rotatable bonds
target conformation (blue) at
RMSD 16Å from given conformation (grey)
Domain swapping

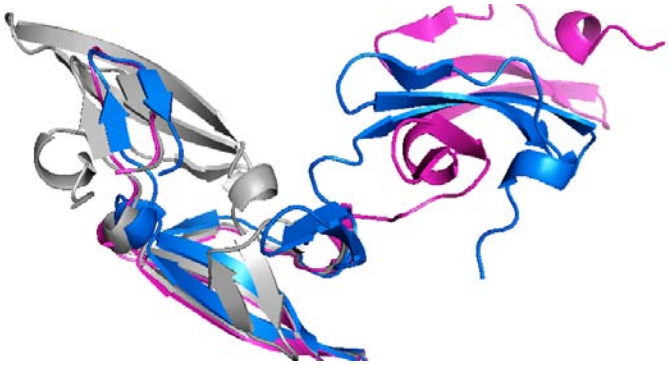


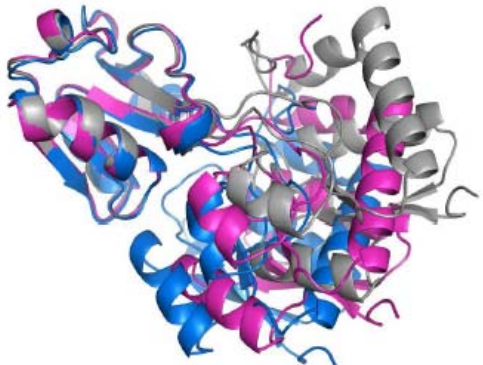
2LAO, 3649 atoms, 1183 rotatable bonds
target conformation (blue) at
RMSD 4.61 Å from given conformation (grey)
Hinge and twist motion

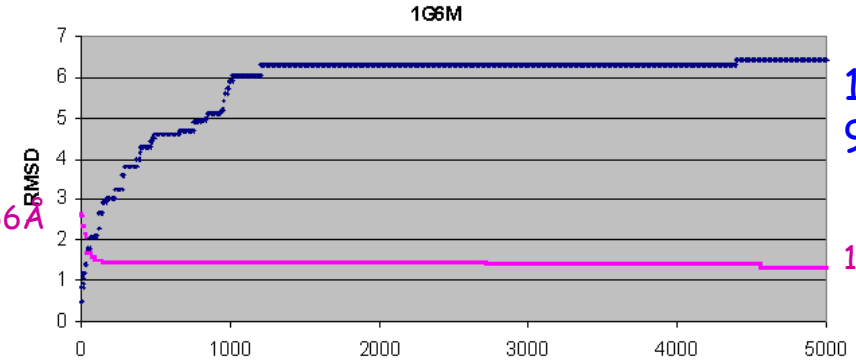
Grey: Initial
Blue: Target
Magenta: Closest sample to target

Results







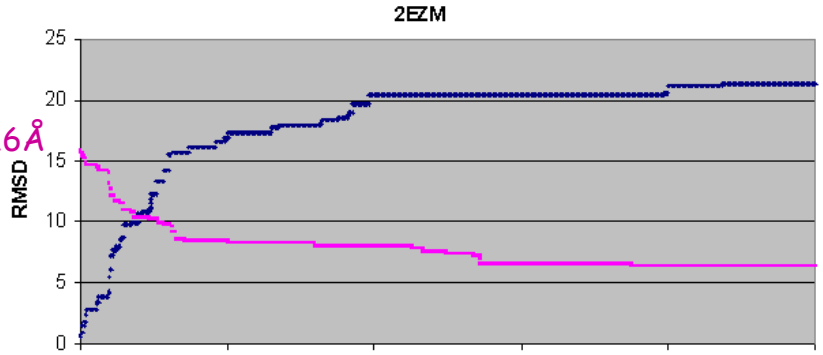


2.66Å

1.3Å

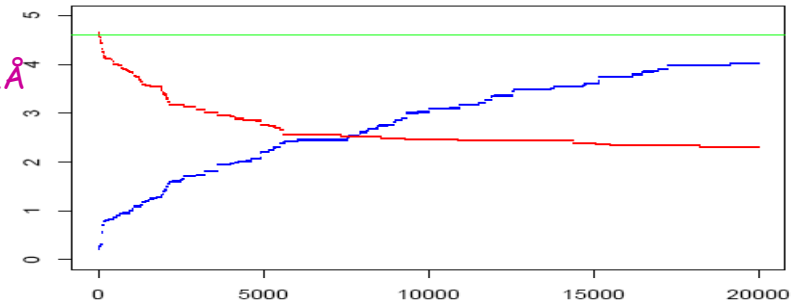
Running times (dual quad-core 3GHz)

13 hours
9.36 sec/sample



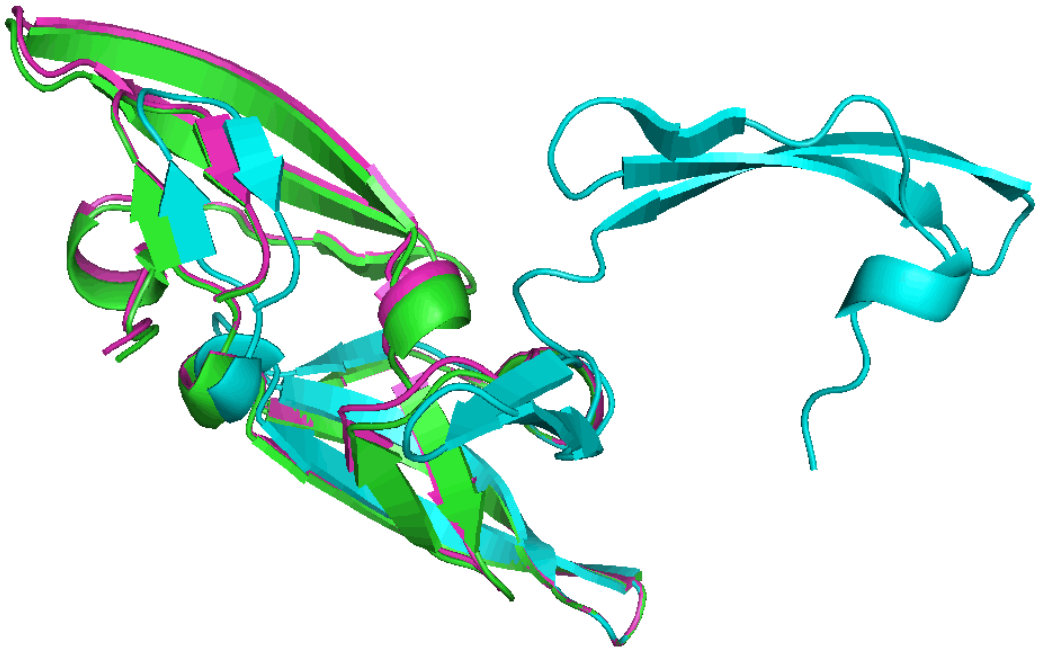
16Å

93 minutes
1.12 sec/sample
6.3Å



4.61Å

12 hours
2.29sec/sample
30Å



2EZM

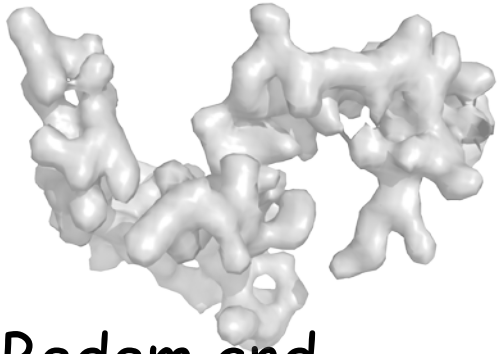
Conclusion

- Kino-geometric constraints characterize well the folded state of a protein
  - "A protein is modeled as a robot"

- Kino-geometric sampling is an efficient way to explore the folded state of a protein
  - Inspired by advancements in sampling-based robot motion planning algorithms

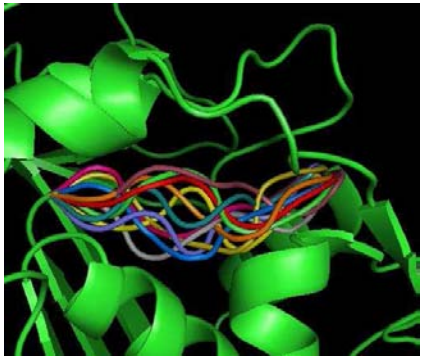
Future Work

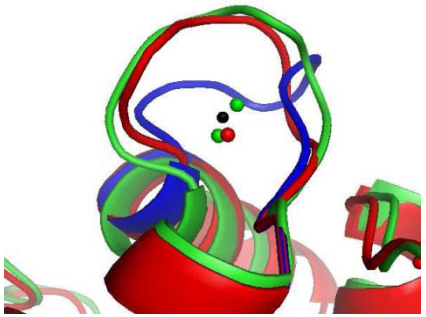
- Develop more efficient sampling algorithms
  - Robots usually have 3-40 degrees of freedom
  - Proteins can have a thousand DOFs
- Generation of graphical models of protein motion (e.g., Markov models)
  - Analyze the sampled conformations
- Interpret noisy experimental data
  - E.g. X-ray data
  - Low resolution
  - Collaborating with Dr. Henry van den Bedem and Dr. Ashley Deacon



Applications

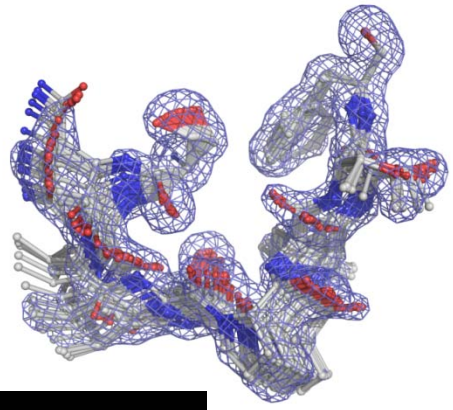
1. Loop sampling
   [Yao et al, 2007]



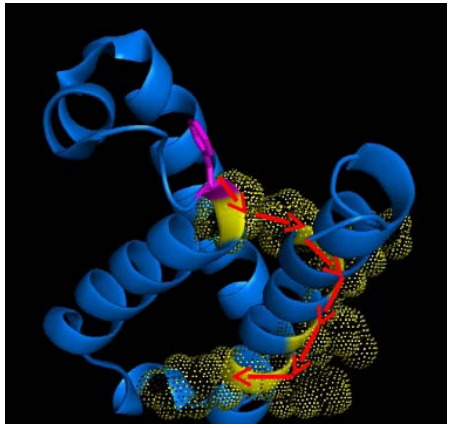


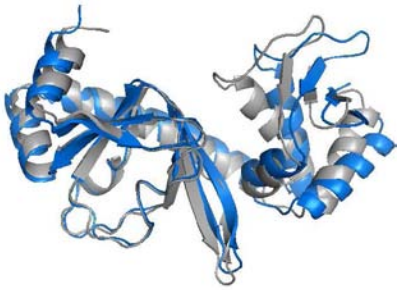
2. Whole conformation sampling
   [Yao, 2010]

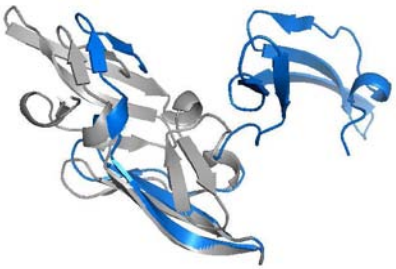
3. Modeling heterogeneity in X-ray crystallographic data
   [van den Bedem et al, 2009]

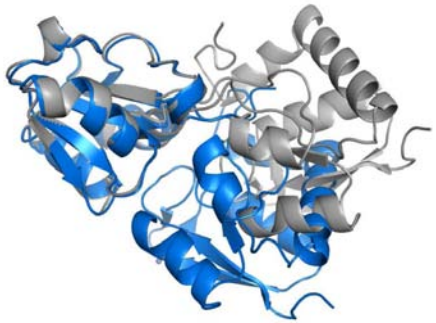


4. Identification of allosteric pathways
   [Dhanik, 2010]









Protein	# atoms	# rigid groups	# rotatable bonds	# cycles	# rotatable bonds in cycles
1G6N	3211	931	1110	80	577
2EZM	992	221	502	47	279
2LAO	3649	1023	1183	84	605