COMPUTER FORUM 09

Modeling Structural Heterogeneity in Proteins From X-Ray Crystallography Data

Ankur Dhanik[1]

Henry van den Bedem[2]

Ashley Deacon[2]

Jean-Claude Latombe[1]

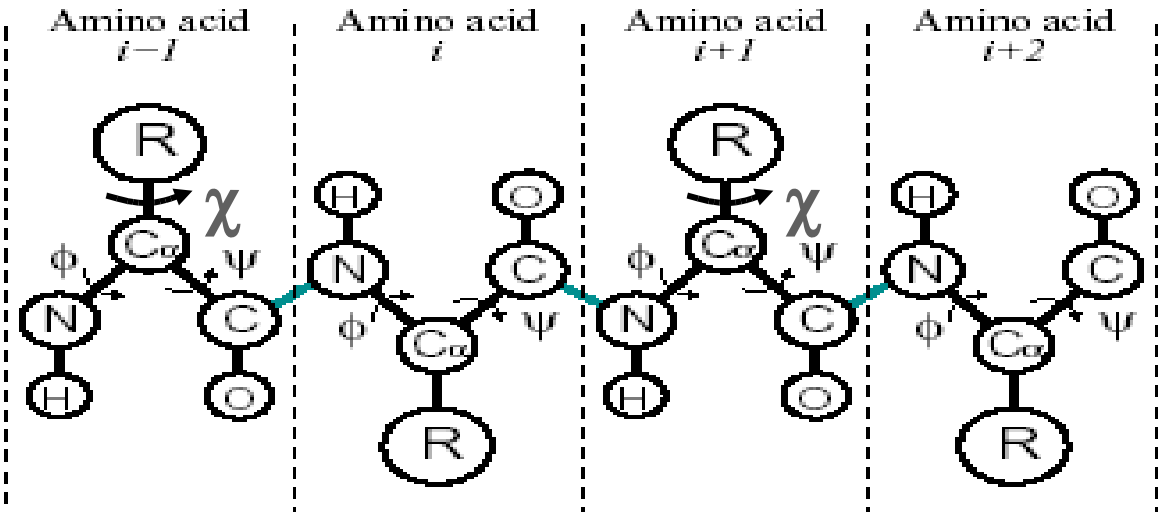
[1]Stanford University

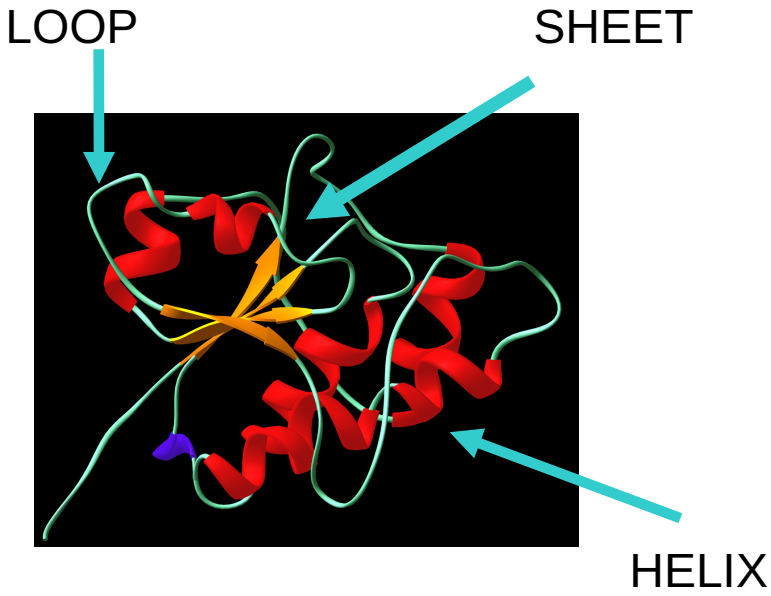
[2]Joint Centre for Structural Genomics

Protein

Sequence of amino acids (residues) connected with peptide bonds

- Forms kinematic chain with many DOFs
- Folds spontaneously into a compact structure

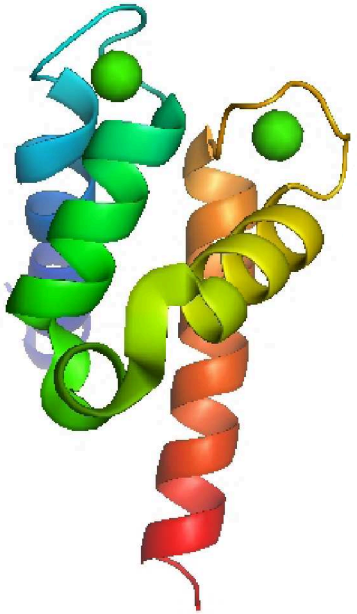




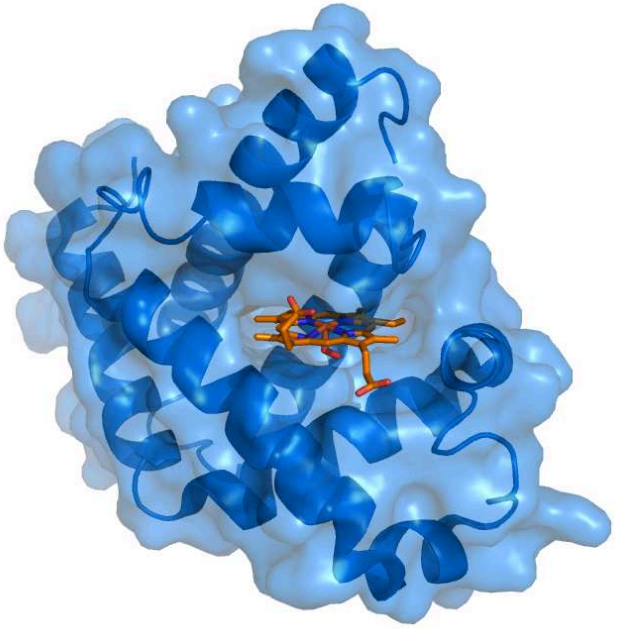
Protein

- Proteins are building blocks of human body
- The function of a protein depends on its folded structure
- Protein–ligand binding requires geometric/chemical complimentarity

→ Correctly determining protein's structure is critically important



Calcium binding protein



Hemoglobin

Image from wikipedia

X-Ray Crystallography

- Main experimental technique for determining protein structures
- About 90% structures deposited in Protein Data Bank have been determined using this technique

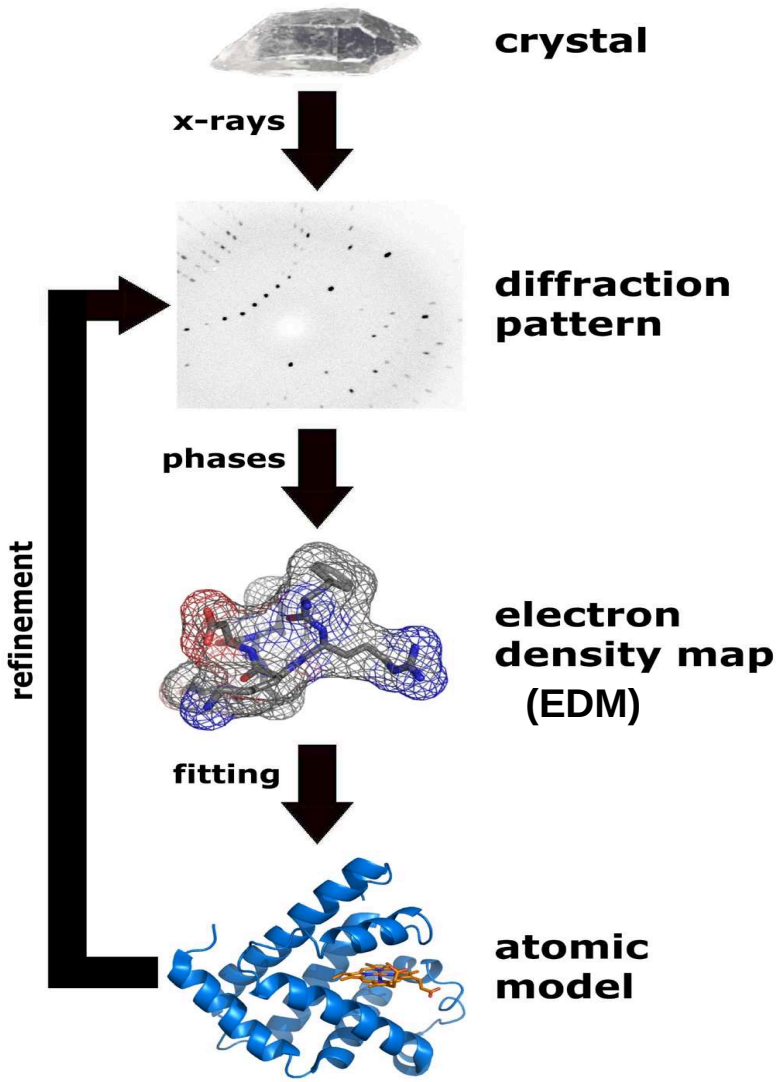
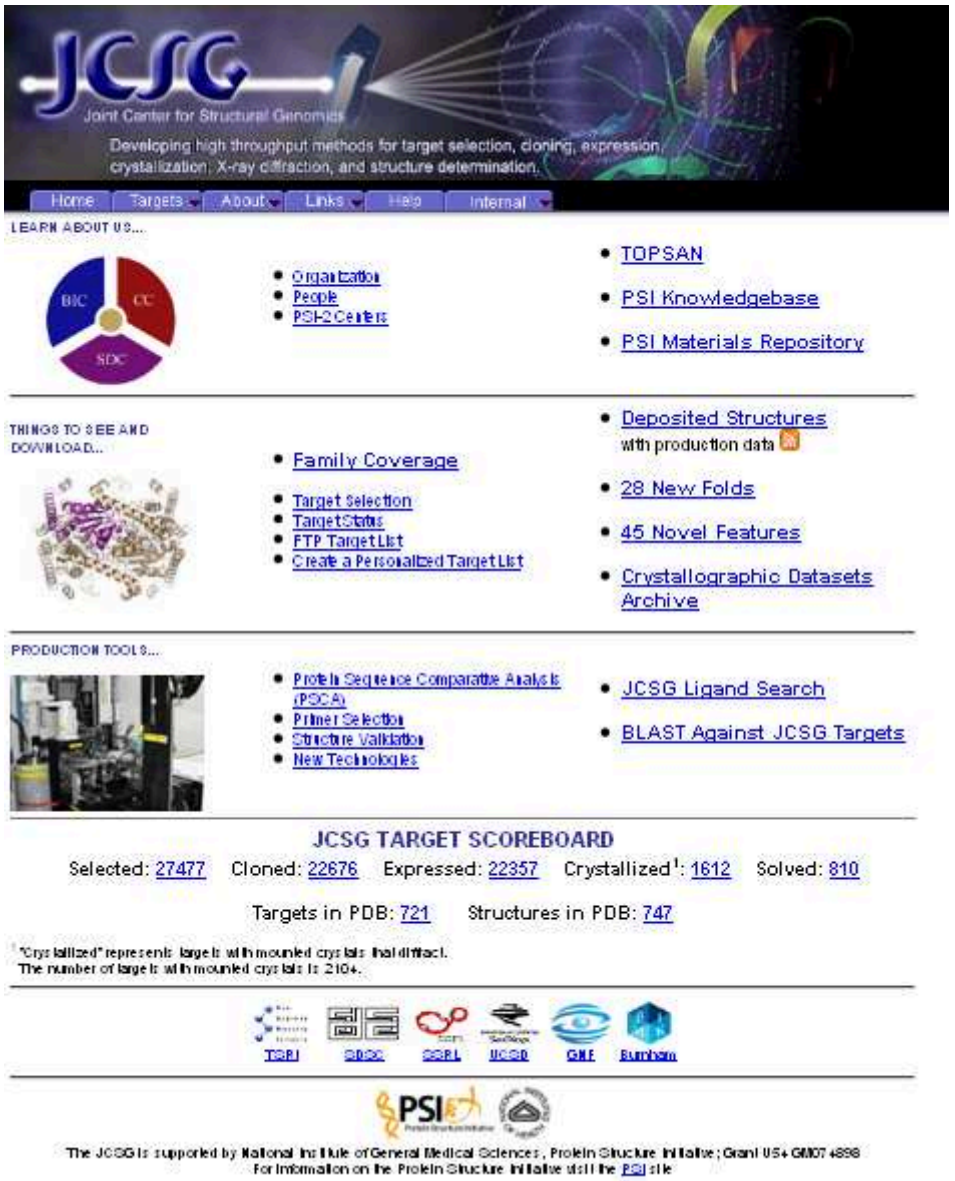


Image from wikipedia

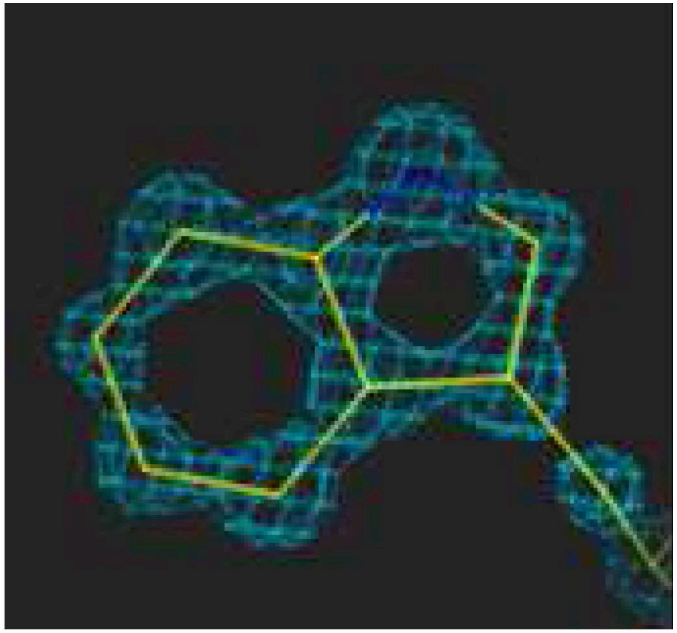
Joint Center for Structural Genomics

- Part of Protein Structure Initiative
- One of four large-scale protein structure determination centers funded by NIH
- As of now 747 structures deposited in PDB (Protein Data Bank)

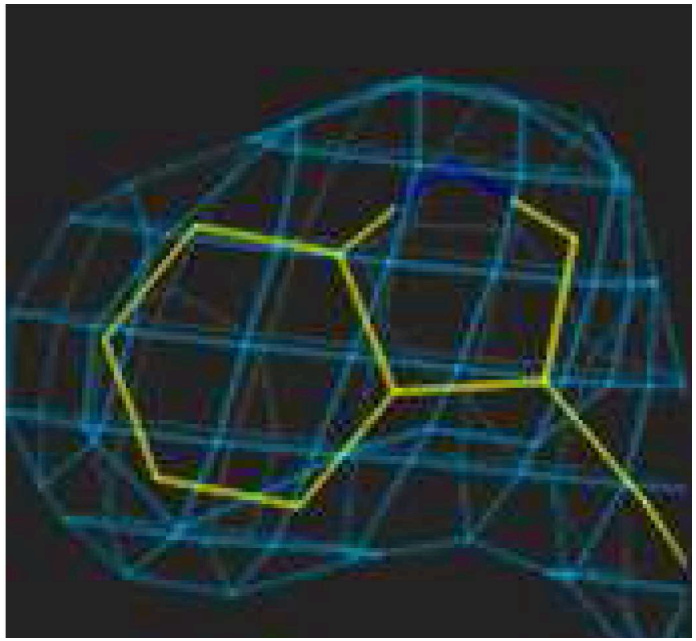


X-Ray Crystallography Challenges

- Low data resolution
- Data collection noise, impurities



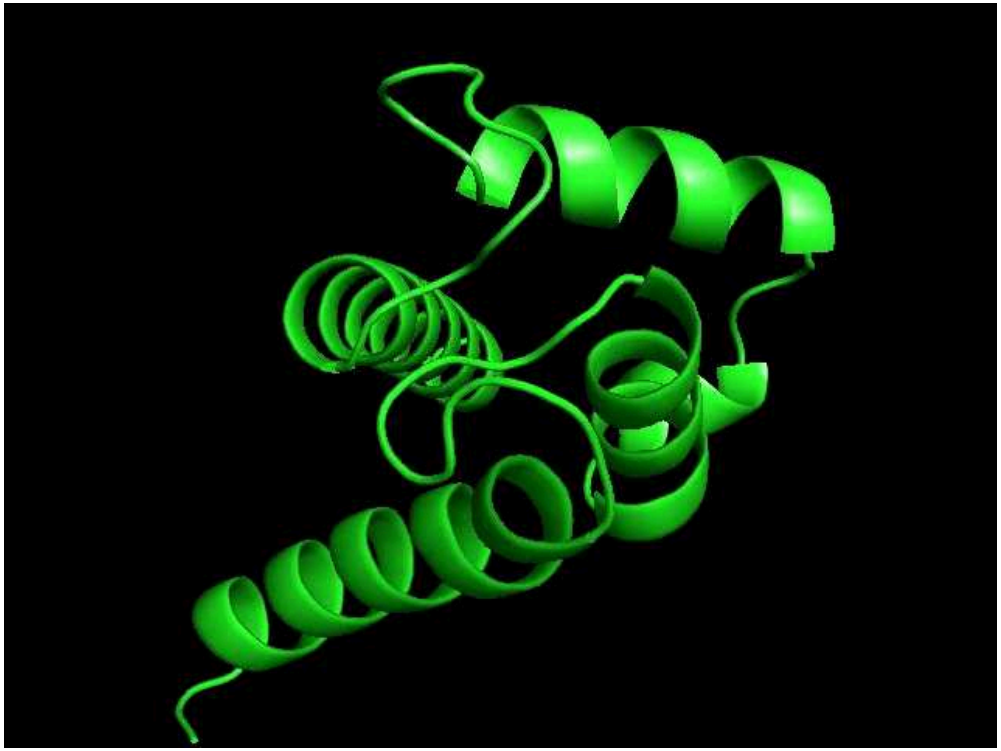
1.0Å



3.0Å

X-Ray Crystallography Challenges

- Structural heterogeneity
  - High-frequency thermal vibration of atoms
  - Low-frequency diffusive motions (coordinated motions of multiple atoms)
  - Diffusive motions are critical to the protein's function

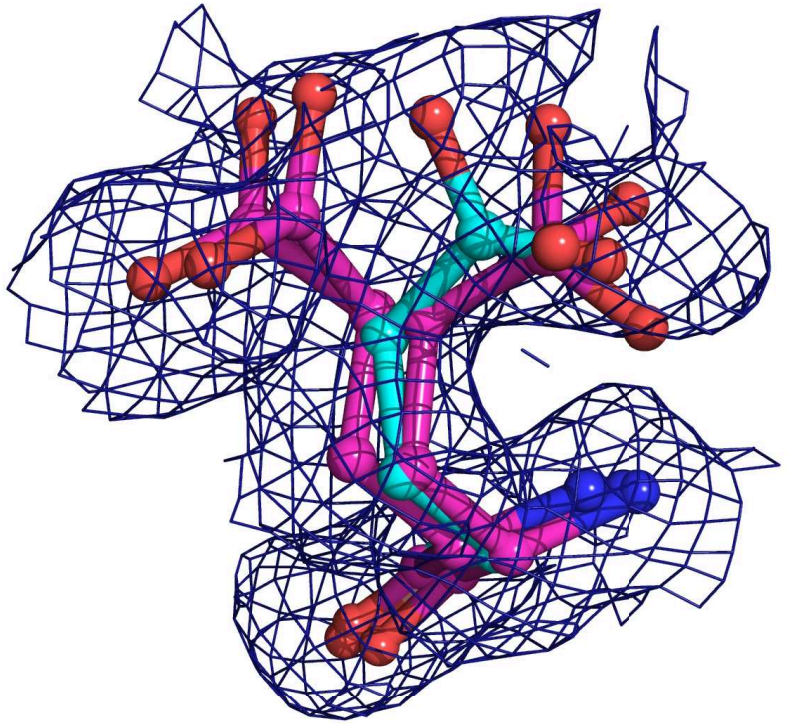


X-Ray Crystallography Challenges

- Structural heterogeneity
  - High-frequency thermal vibration of atoms
  - Low-frequency diffusive motions (coordinated motions of multiple atoms)
  - Diffusive motions are critical to the protein's function

→ Multiple conformations are present in a crystal

Diffraction pattern is occupancy weighted superposition of patterns (*Occupancy*: percentage of copies of a conformation)

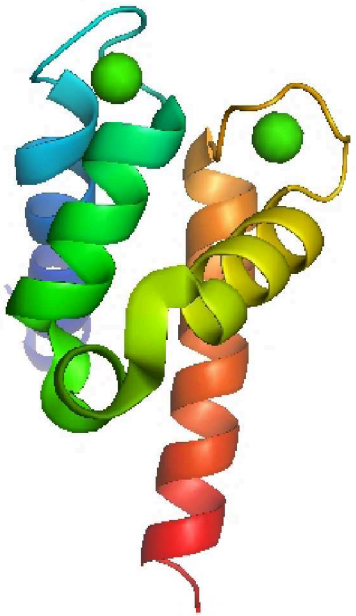


X-Ray Data Modeling

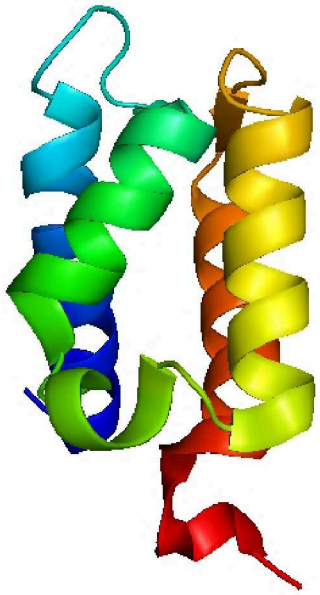
- Current practice
  - Compute **one** “average” protein conformation
  - Explain thermal vibrations with isotropic Gaussian distribution of atom position controlled by *temperature factor* or *B-factor*
  - Software suites: ARP/ RESOLVE/ TEXTAL
- But: **Is one solution enough?**
  - Several “average” conformations can explain the EDM equally well (Terwilliger, 2007)
  - Modelling structural heterogeneity is critically important (Furnham 2006, DePristo 2004)

Modeling Heterogeneity

- Our goal
  - Compute an ensemble of conformations (with occupancies and B-factors) that, collectively, provides a near-optimal explanation of the EDM
  - Current focus: fragments with deformable kinematics (mainly loops and side chains)



Calcium Bound State

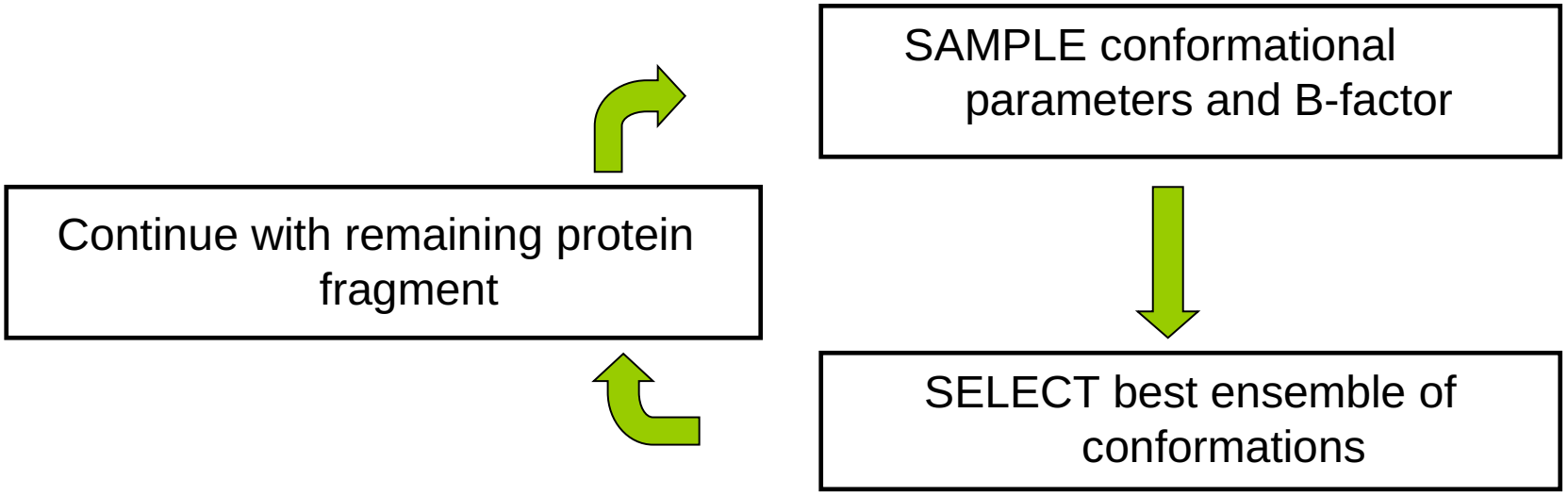


Calcium Free State

Modeling Heterogeneity

- Our Method

  Alternate SAMPLE (massive sampling) and SELECT (efficient selection) steps



Modeling Heterogeneity

SELECT

Uses linear programming to compute occupancies and select conformations with non-zero occupancies

Find $\alpha_1, \ldots, \alpha_N$ such that

$\sum_{p \varepsilon G} | E(p) - \sum_i \alpha_i E_i(p) |$ is minimum,

Under the constraints:

$\alpha_i >= 0$ for all i=1, …, N and $\sum_i \alpha_i = 1$

α_i = occupancy of conformation i

E(p) = electron density of given EDM at grid point p

$E_i(p)$ = electron density of EDM computed from conformation i at grid point p

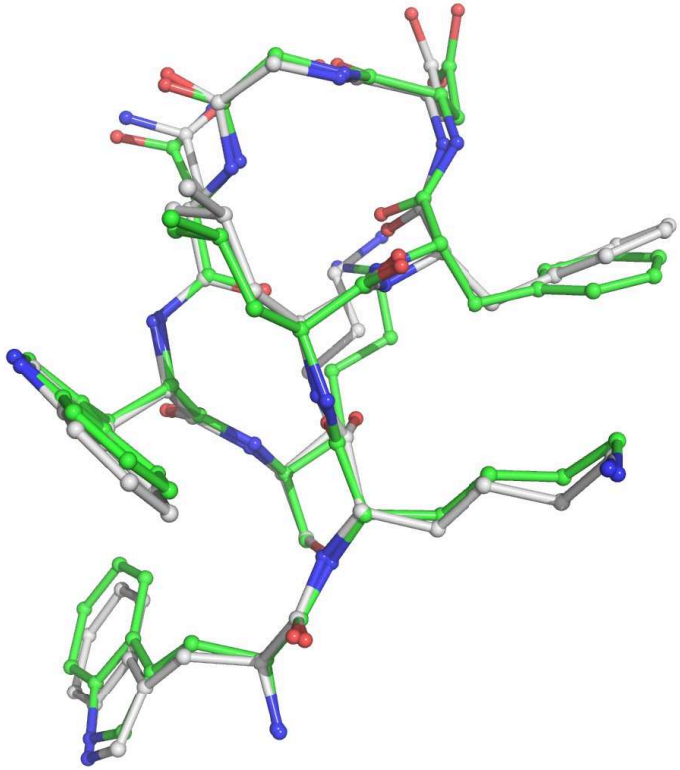
Results

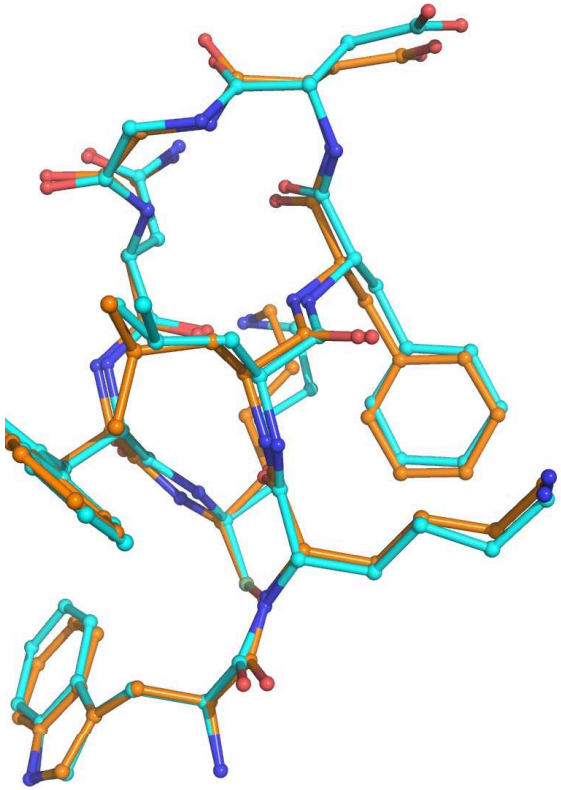
- Simulated and experimental EDMs
- Simulated EDMs are computed over a range of resolution levels, noise levels, occupancies, and B-factors
- Reasonably accurate conformations, occupancies, and B-factors are computed
- Libraries used: LoopTK, Clipper, Coin-OR

Results

- 2R4I (loop 104-112)
  - Two conformations, separated by 1.4 Å RMSD (root mean squared distance between corresponding atoms)
  - Average B-factor = 19.0 Å^2
  - Simulated EDM is generated at different resolutions and occupancies
  - Gaussian noise is added

Results





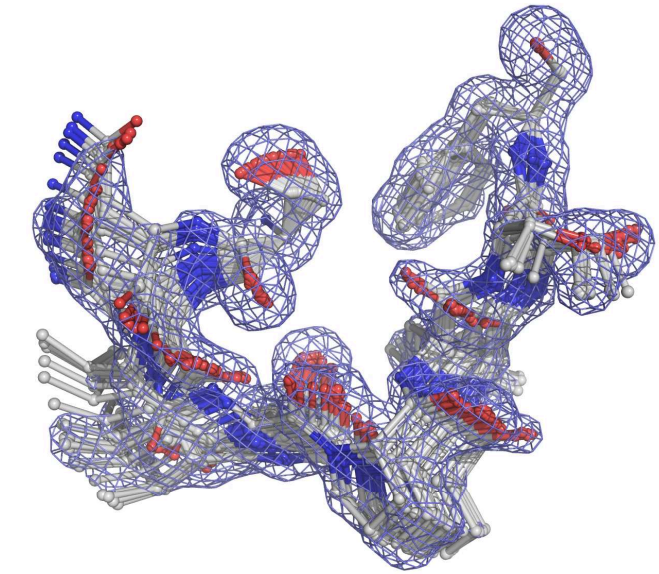
Occ	Res	RMSD	Calc Occ	Bcalc
0.7/0.3	1.3	0.27 0.34 0.34 0.34 0.64 0.48	0.70 0.30	24.1

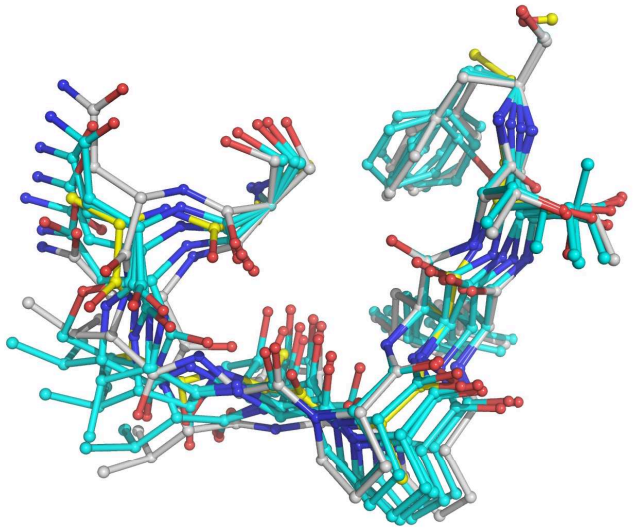
Results

- 1HFC (loop 142-149)
  - 20 conformations
  - Emulates coordinated motion of the loop
  - Start and finish conformations 2.7Å RMSD apart
  - Average B-factor = 11.7 $Å^2$
  - Simulated EDM is generated at equal occupancy (0.05) and 1.9 Å resolution

Results

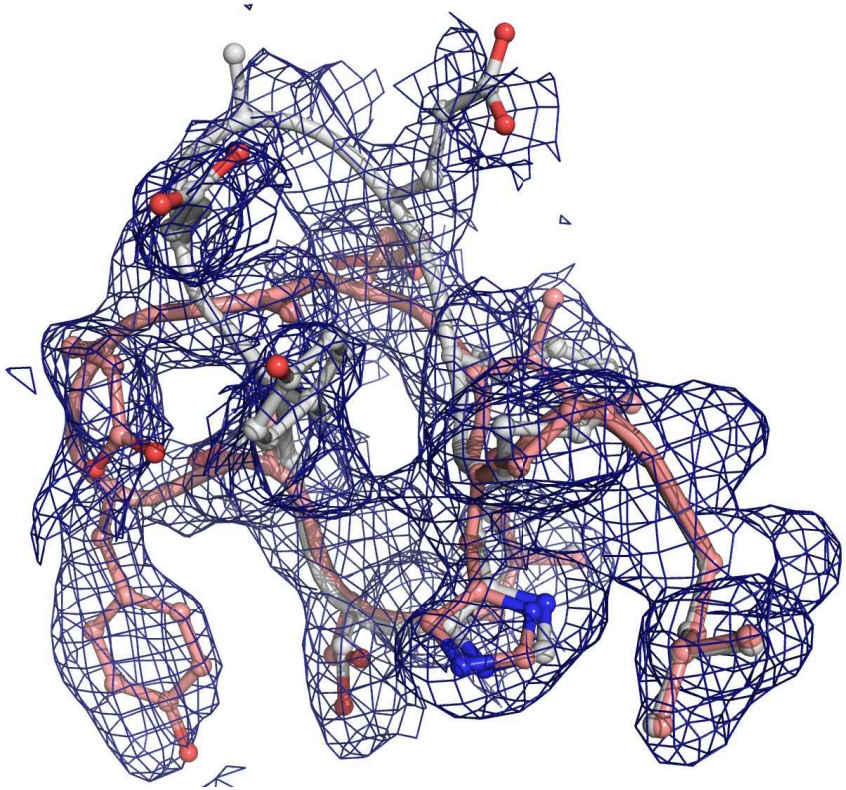
- Our method
  - 7 conformations, occupancies ranging from 0.10 to 0.23
  - Average B-factor = 15.0 Å^2
- RESOLVE
  - 1 conformation
  - Average B-factor = 35.7 Å^2





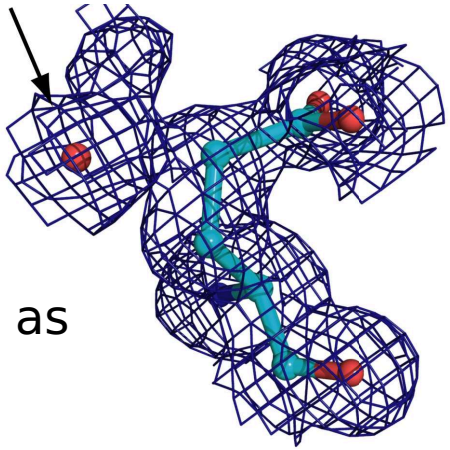
Results

- TM0755 (loop 316-325)
  - Experimental data from JCSG at 1.8 Å resolution
  - Difficult to model (even for trained crystallographers)
  - Two conformations
  - Average JCSG B-factor = 24.9 Å^2
  - 5 conformations computed
  - 0.47(0.15) , 0.64(0.27)
  - Average B-factor = 30.3 Å^2

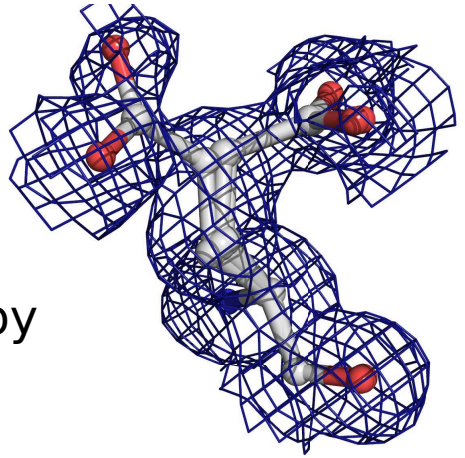


Results

- 2NLV (multiple side chain conformations)
  - Experimental data from JCSG at 1.3 Å resolution
  - PDB conformation contains water molecule in place of carbonyl oxygen
  - Our method modeled 12 multi-conformation alternatives for single conformation residues with improved fit to EDM



Residue 81GLU as modeled in the PDB



As modeled by our method

Is one solution enough?

- NO!

  JCSG has deposited two new structures in PDB using our method, including the highly heterogeneous protein 3EO6

Is one solution enough?

- NO!

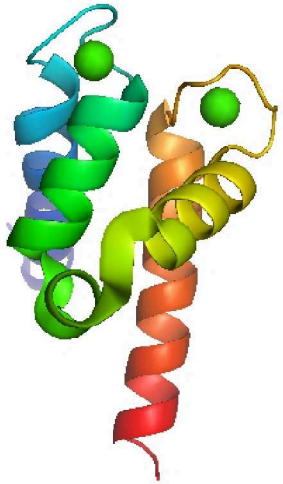
  JCSG has deposited two new structures in PDB using our method, including the highly heterogeneous protein 3EO6

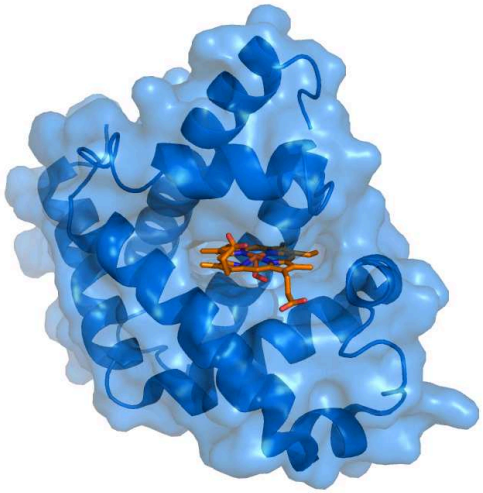
- Our method demonstrates that heterogeneity in EDM can be better explained by multiple conformations than by B-factors alone

Protein revisited

- Proteins are building blocks of human body
- The function of a protein depends on its folded structure
- Protein–ligand binding requires geometric/chemical complimentarity

→ Correctly determining protein's structure is critically important





Our method can lead to better determination and understanding of protein's structure, and provide insights into its functioning

Acknowledgements

- NSF award DMS-0443939, CNS0619926
- JCSG