Integrated Modelling of PROTEIN Complexes VIA Single Shot Registration using DREAM (IMPROVISeD)

iGem IISc-Software
Ayush Raina, Rahul Chavan, Anirudh Gupta, Niladri
September 24, 2024

Presentation at ThermoFisher Scientific


Kotak IISc AI–ML Centre

Introduction

iGem: International Genetically Engineered Machine

Figure 1: Logo

iGEM is a synthetic biology competition that gathers the young minds from around the world to collaborate, innovate and tackle complex challenges in the field of biotechnology.

iGem empowers teams of undergradudate and graduate students to design, build and test novel biological systems and applications.

iGem: Software and AI

An iGem competition track dedicated to projects based on computational methods.

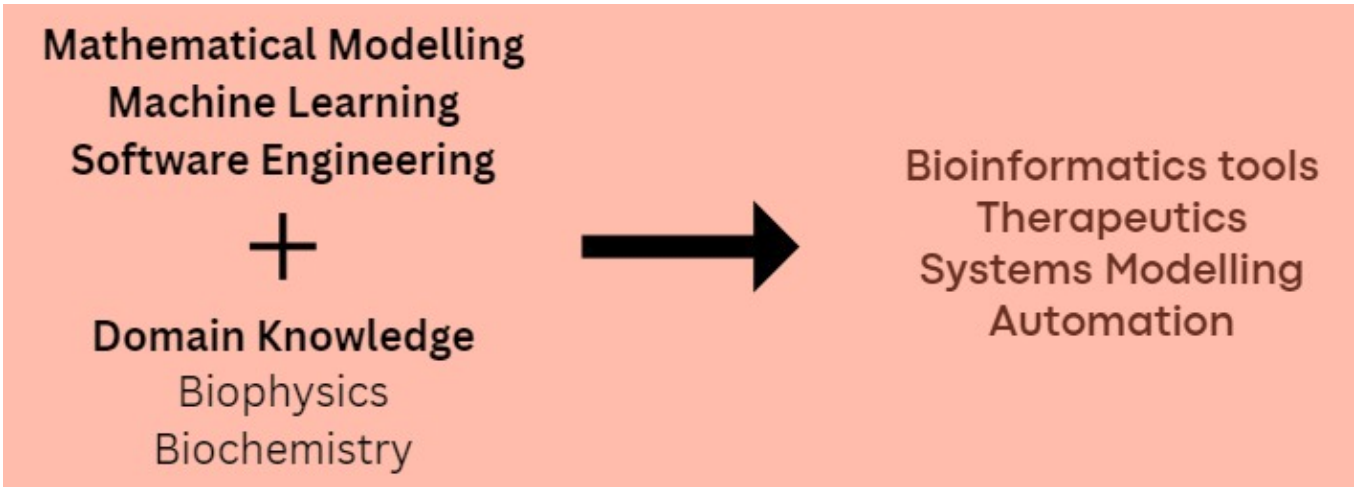


Figure 2: Logo

Computational Biology

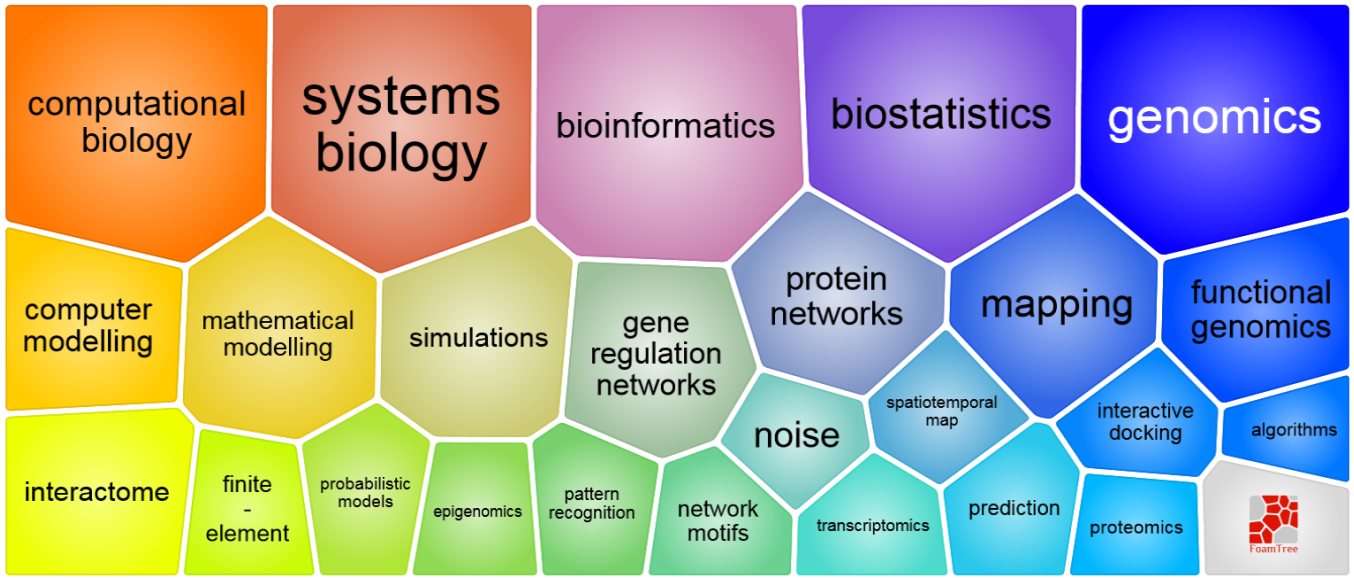


Figure 3: Comp Bio

Computational biology refers to the use of data analysis, mathematical modeling and computational simulations to understand biological systems and relationships.

Integrated Modelling of Protein Complexes

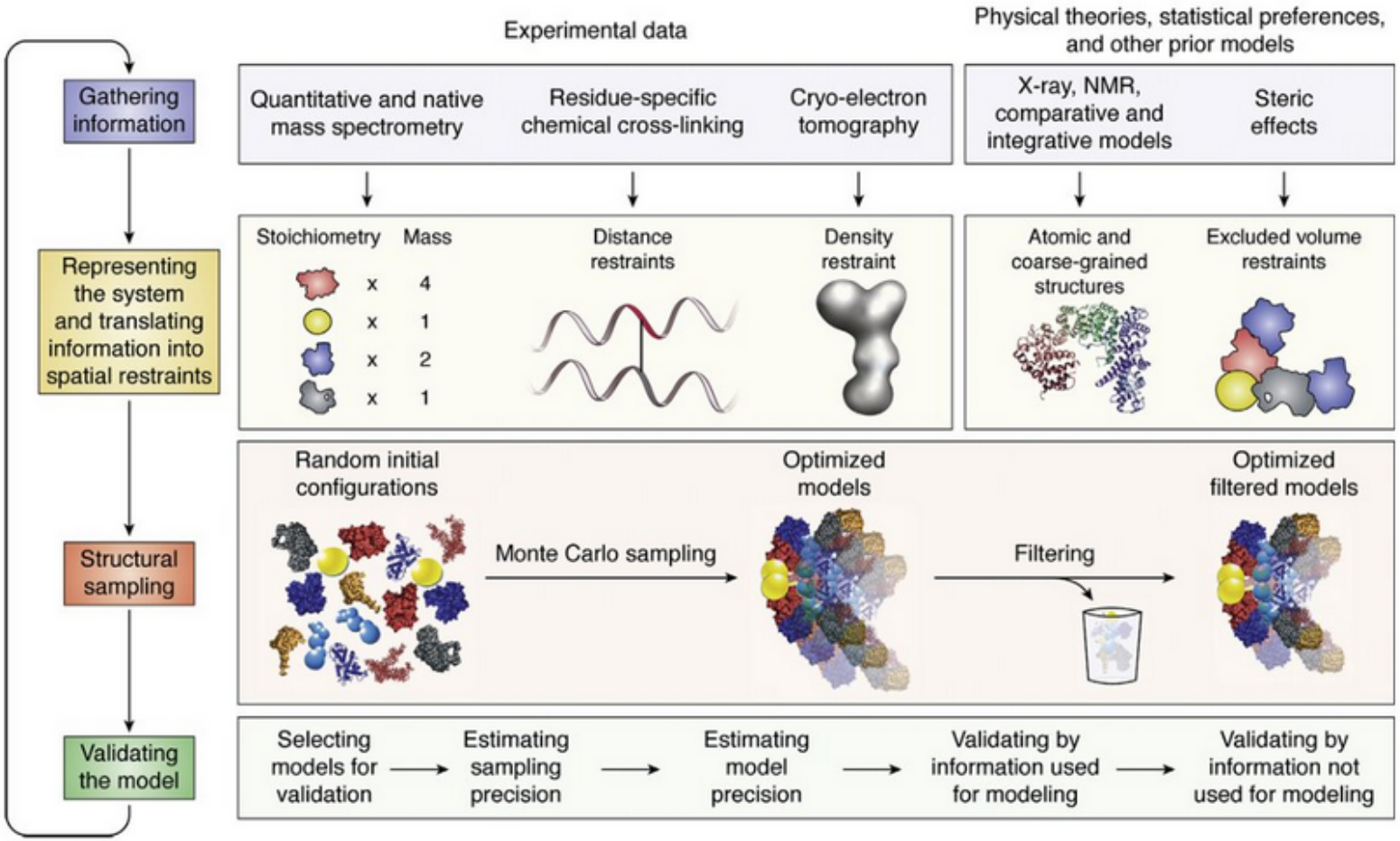
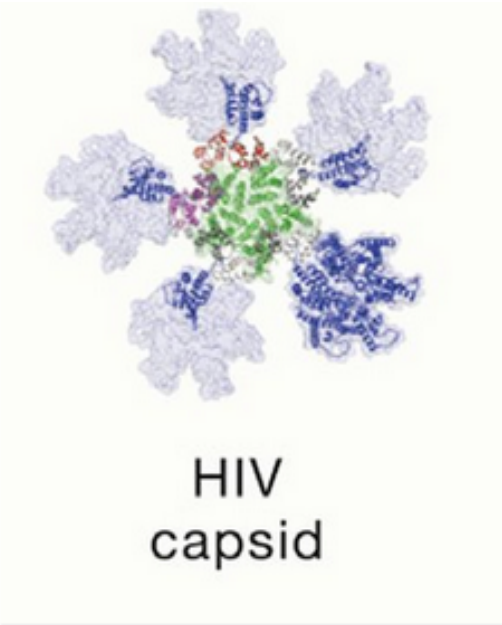
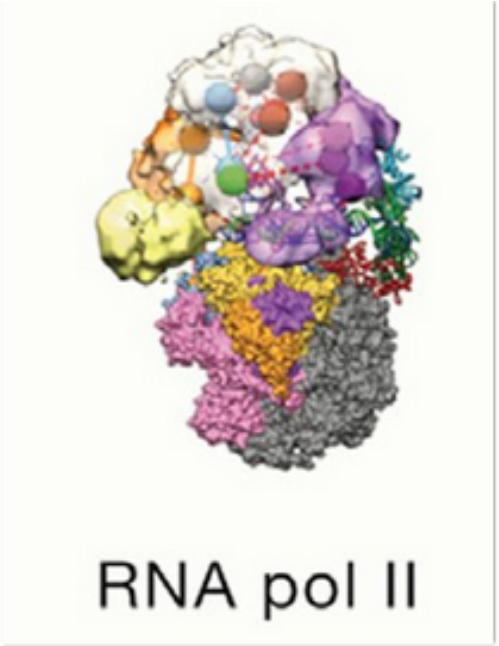


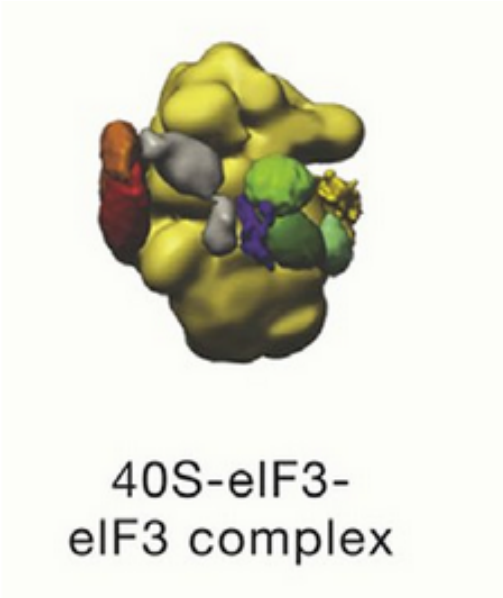
Figure 4: Flowchart representing the IMP



(a) Deshmukh et al., 2013



(b) Murakami et al., 2013

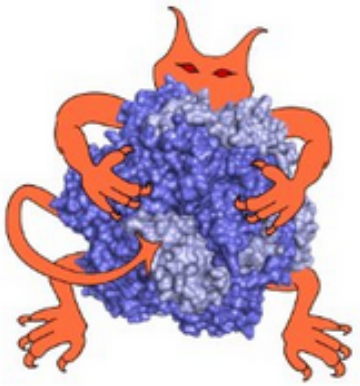


(c) Erzberger et al., 2014

Why Integrated Modelling

1. Using new information
2. Maximizing accuracy, precision and completeness
3. Planning experiments

Present Landscape



(a) IMP,the integrative modelling platform

(b) rosetta

(c) haddock

Program	Functionality	Web Site	Reference
ISD	Bayesian modeling on the basis of NMR data	N/A	Rieping et al., 2005
IMP	Integrative modeling	integrativemodeling.org	Russel et al., 2012
Rosetta	Integrative modeling	rosettacommons.org	Das and Baker, 2008
ISDB	Integrative modeling	plumed.org	Bonomi and Camilloni, 2017
pow^{er}	Integrative modeling	lbm.epfl.ch/resources/	Degiacomi and Dal Peraro, 2013
cMNXL and Jwalk/MNXL	Integrative modeling	topf-group.ismb.lon.ac.uk/Software	Bullock et al., 2018a; Bullock et al., 2018b
PyRy3D	Integrative modeling	genesilico.pl/pyry3d/	J. M. Kasprzak, M. Dobrychłop, and J. Bujnicki
PGS	Modeling genome structure	github.com/alberlab/PGS	Hua et al., 2018
TADBit	Modeling genome structure	sgt.cnag.cat/3dg/tadbit/	Serra et al., 2017
MDFF/NAMD	Fitting of molecular models into EM maps using MD simulations	ks.uiuc.edu/Research/mdff	Trabuco et al., 2008
ATSAS	Integrative modeling using SAXS	embl-hamburg.de/biosaxs	Franke et al., 2017
iFoldRNA	Integrative modeling of RNA	iFoldRNA.dokhlab.org	Sharma et al., 2008
HADDOCK	Integrative modeling using docking and data derived restraints	haddock.science.uu.nl	Dominguez et al., 2003
ATTRACT-EM	Integrative modeling using docking and EM	attract.ph.tum.de	de Vries and Zacharias, 2012
DireX	Flexible fitting of EM maps with data derived distance restraints.	schroderlab.org/software/direx/	Wang and Schröder, 2012
MDFit	MD based integrative modeling using EM maps	smog-server.org/SBMextension.html#mdfit	Ratje et al., 2010
FPS	Integrative modeling using FRET data	www.mpc.hhu.de/en/software/fps.html	Kalinin et al., 2012
XPLOR-NIH	Structure determination using NMR data	nmr.cit.nih.gov/xplor-nih/	Schwieters et al., 2018
PatchDock	Molecular docking by shape complementarity	bioinfo3d.cs.tau.ac.il/PatchDock/	Schneidman-Duhovny et al., 2005
iSPOT	Structure determination using SAS, footprinting and docking	www.theyanglab.org/ispot/	Hsieh et al., 2017
BCL	Various servers for integrative modeling	meilerlab.org/index.php/servers	Woetzel et al., 2011
ChimeraX	Model visualization	rbvi.ucsf.edu/chimerax	Goddard et al., 2018
VMD	Model visualization	ks.uiuc.edu/research/vmd	Humphrey et al., 1996
Protein Model Portal	Portal to atomic models of proteins	proteinmodelportal.org	Haas et al., 2013
PDB-Development	Archiving of integrative structures	pdb-dev.wwpdb.org	Burley et al., 2017

Figure 7

Our Improvement

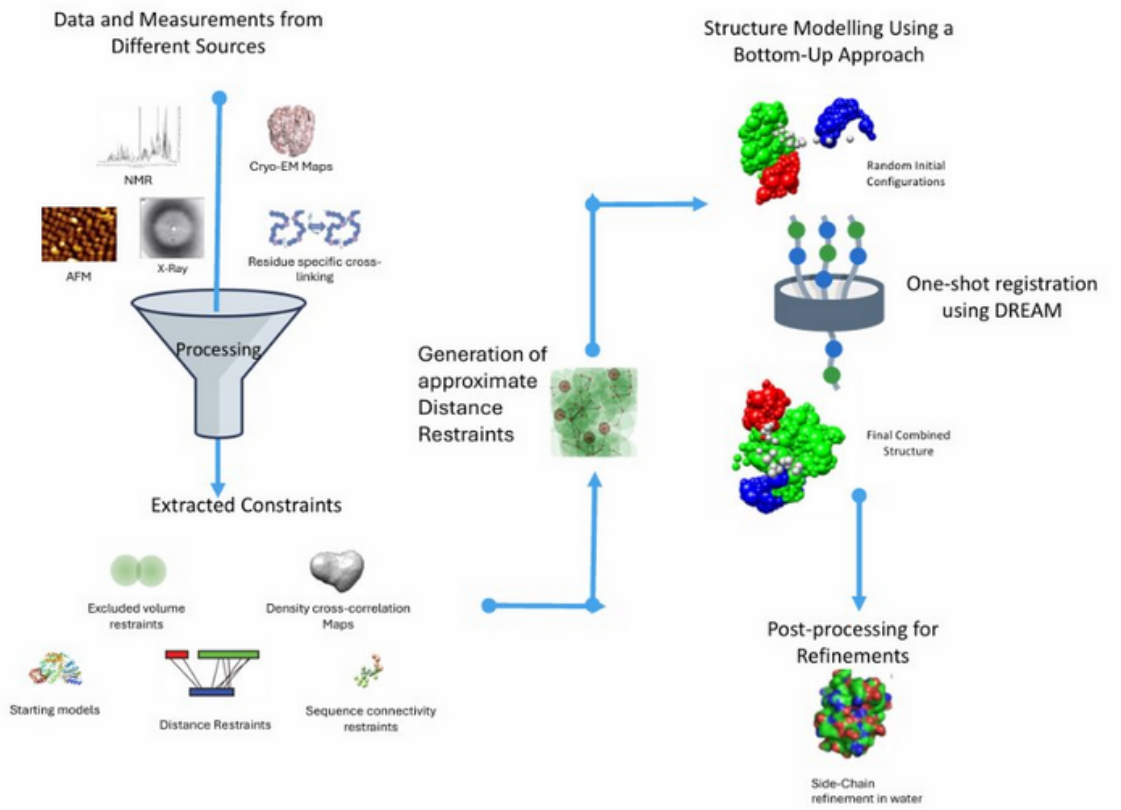


Figure 8: Flow Chart

1. Single Shot Registration
2. Scalability

2 Methodology

Methodology

Distance Restraints and Energy Assisted Modelling

DREAM algorithm uses distance restraints obtained from NMR data to model the structure of proteins in 3 steps:

1. **Construction of Substructures:** We divide the available distance restraints data into dense fragments and model their structure first.

Methodology

Distance Restraints and Energy Assisted Modelling

DREAM algorithm uses distance restraints obtained from NMR data to model the structure of proteins in 3 steps:

1. **Construction of Substructures:** We divide the available distance restraints data into dense fragments and model their structure first.

Methodology

Distance Restraints and Energy Assisted Modelling

DREAM algorithm uses distance restraints obtained from NMR data to model the structure of proteins in 3 steps:

1. **Construction of Substructures:** We divide the available distance restraints data into dense fragments and model their structure first.

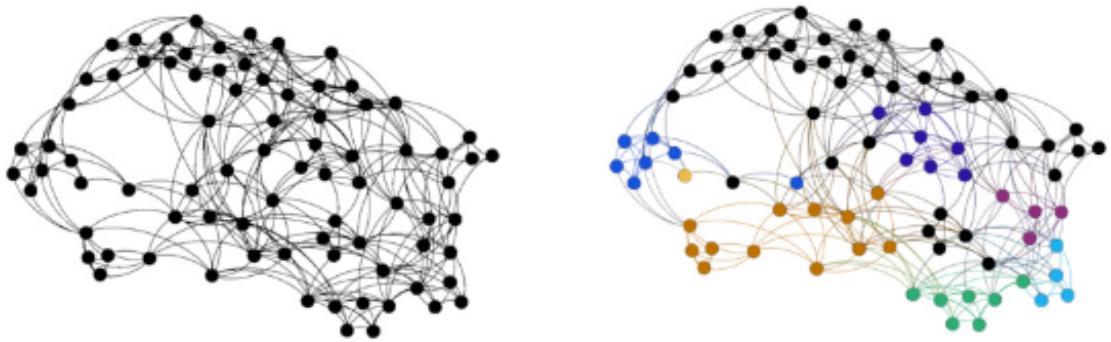


Figure 9: Dividing into dense fragments

DREAM

2. **One Shot Registration:** We then join all the substructures into a single structure at once instead of sequential registration.

DREAM

2. **One Shot Registration:** We then join all the substructures into a single structure at once instead of sequential registration.

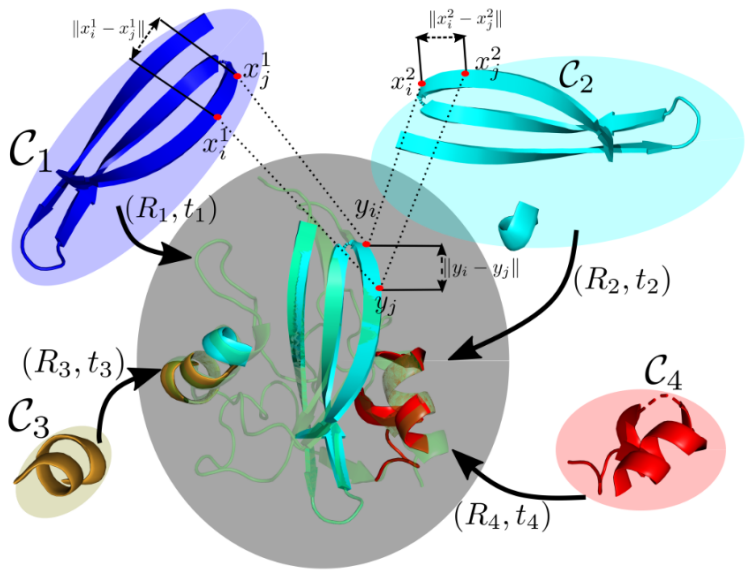


Figure 10: One Shot Registration

Here (R, t) denotes the rotation and translation of the substructure.

DREAM

3. **Gap Filling:** Here many hybrid approaches are used to model the missing regions in the structure. This includes energy minimization, water refinement etc.

DREAM

3. **Gap Filling:** Here many hybrid approaches are used to model the missing regions in the structure. This includes energy minimization, water refinement etc.

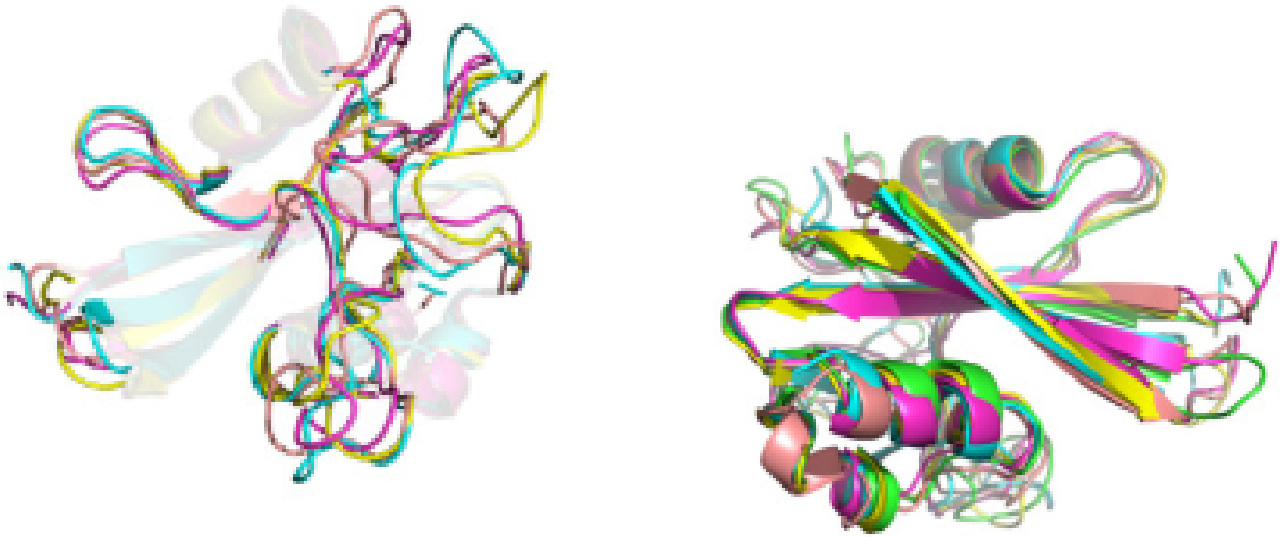


Figure 11: Gap Filling

But why DREAM ?

- Uses divide and conquer approach which increases robustness and scalability.

But why DREAM ?

- Uses divide and conquer approach which increases robustness and scalability.
- Reduces numerical instabilities because of the use of dense fragments.

But why DREAM ?

- Uses divide and conquer approach which increases robustness and scalability.
- Reduces numerical instabilities because of the use of dense fragments.
- In sequential registration, the error keeps on accumulating which is not the case in one shot registration.

Sequential vs One Shot Registration

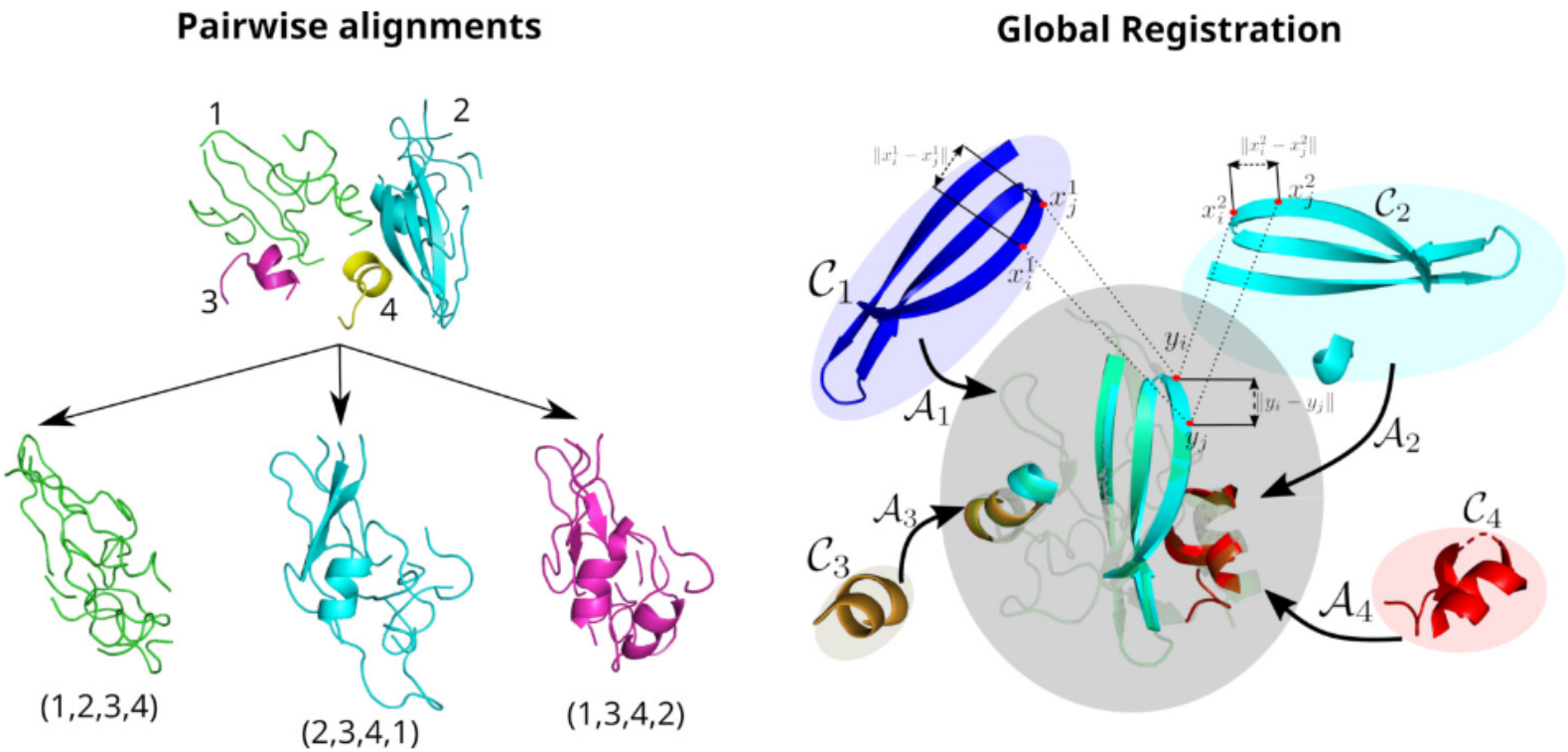


Figure 12: Sequential vs One shot registration

Integrative Modelling Platform

Now we already know what is IMP. It is computationally expensive and time consuming due Markov Chain Monte Carlo (MCMC) sampling.

Integrative Modelling Platform

Now we already know what is IMP. It is computationally expensive and time consuming due Markov Chain Monte Carlo (MCMC) sampling.

We wish to replace the computationally expensive sampling techniques to paradigms used in DREAM:

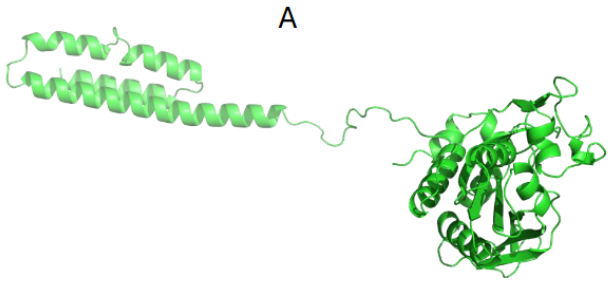
- Orientate the structures of subunits based on experimental evidence which is similar to substruture computation in DREAM
- Register the subunits in one shot while respecting the experimental evidence. (an enhancement of DREAM's registration)

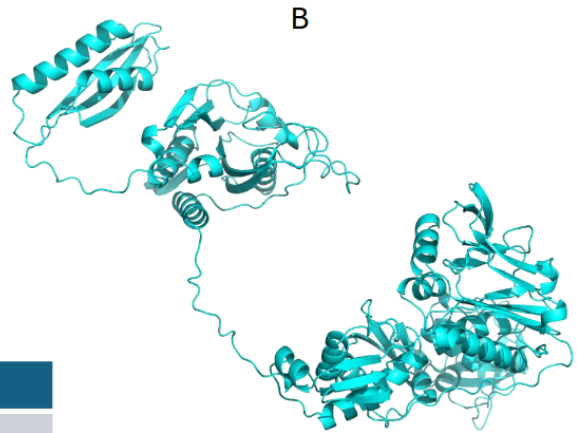
How will this happen ?

- Our substructures in this case are different kinds of proteins.
- We have cross-links data available these proteins.

Given this information, we need to do one shot registration of these proteins to model the structure of complex.

Inputs





Coordinates	
75	CA
85	CA
181	CA
237	CA
321	CA
342	CA

Crosslinks			
78	A	650	B
85	A	650	B
87	A	679	B
321	A	502	B
342	A	610	B
181	A	593	B
75	A	650	B
237	A	502	B

Coordinates	
502	CA
593	CA
610	CA
650	CA
679	CA

Figure 13: Example of 2 proteins with cross-links data

Problem

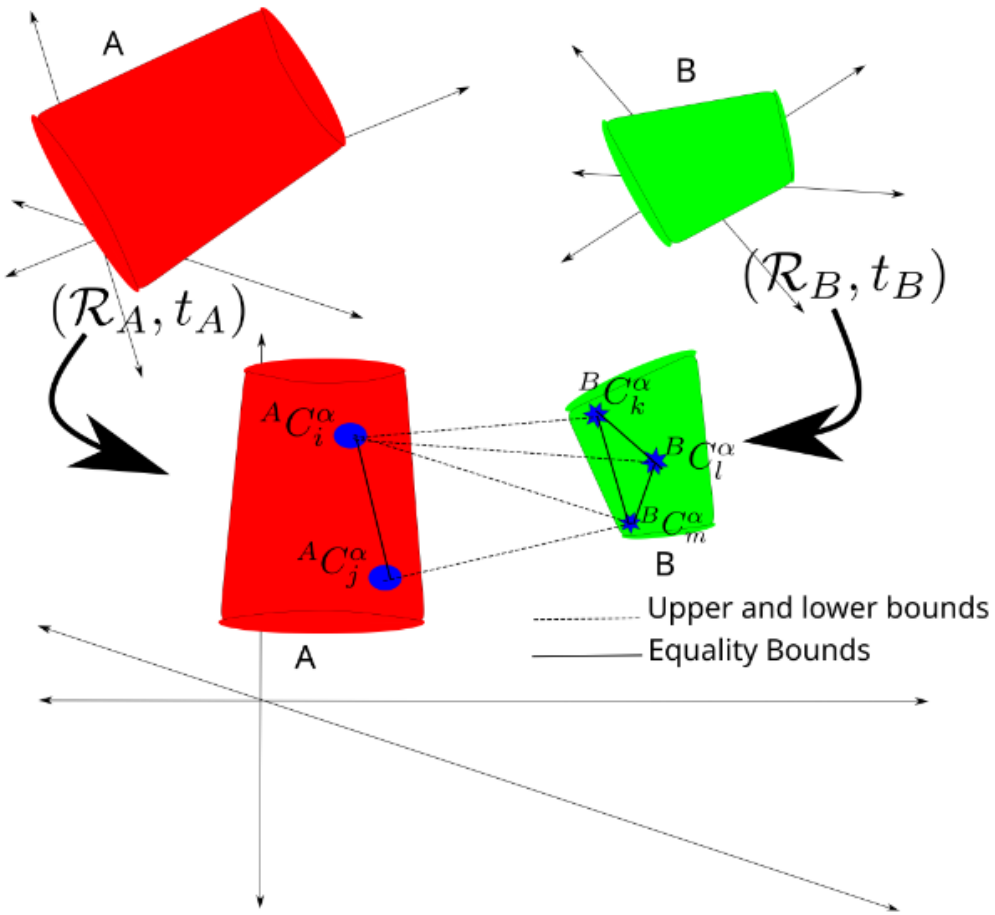
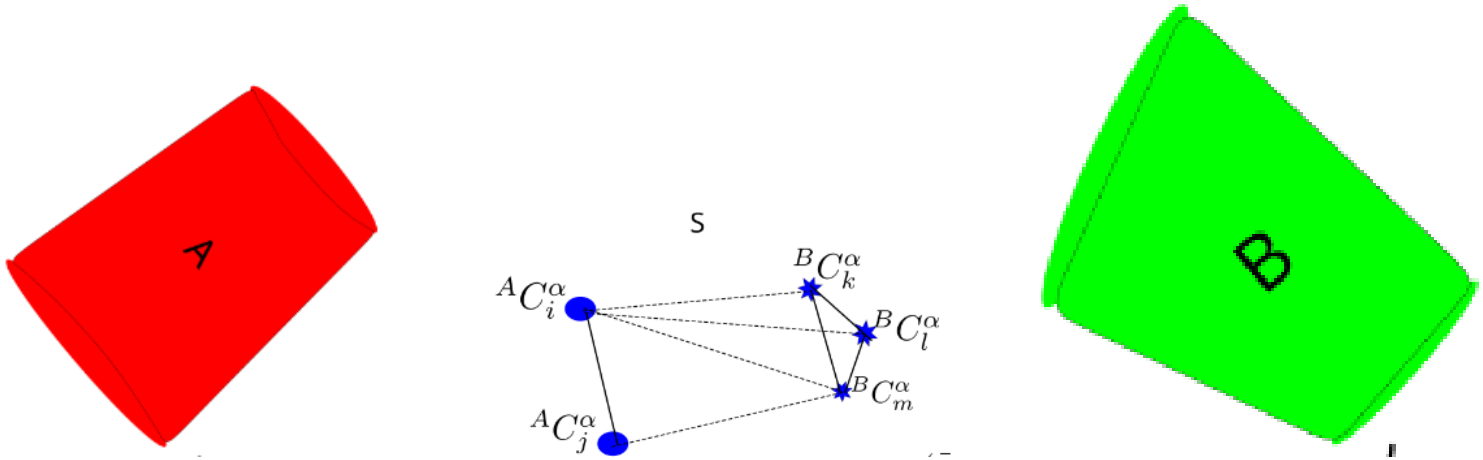


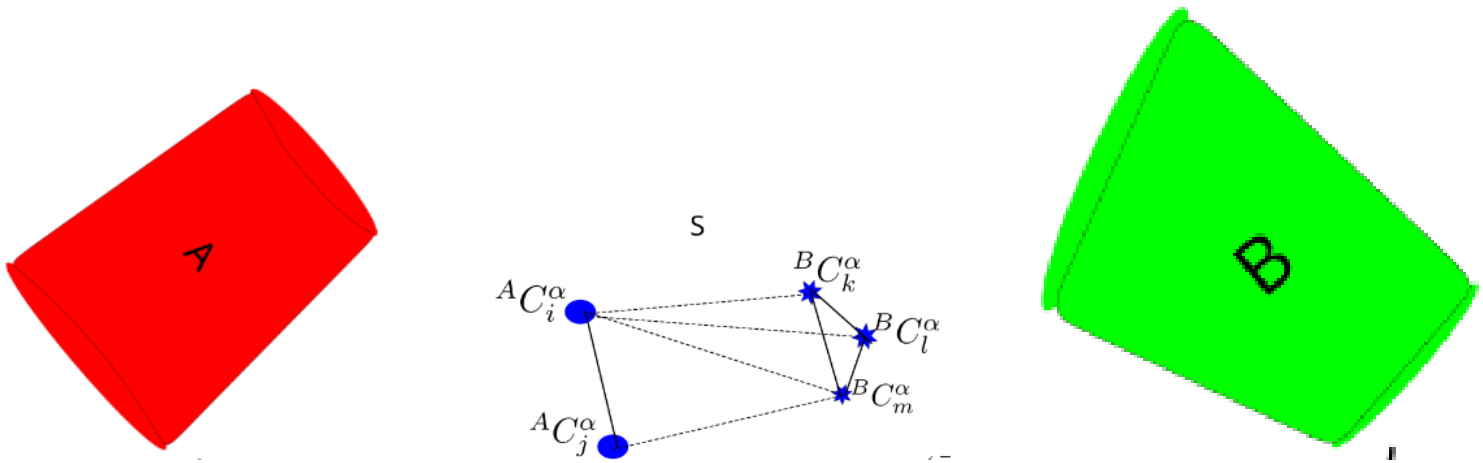
Figure 14: Consider A,B as proteins and the lines as cross-links

Solution



Consider S as hypothetical framework.

Solution



Consider S as hypothetical framework. Then we can do one shot registration of A,S and B

Solution

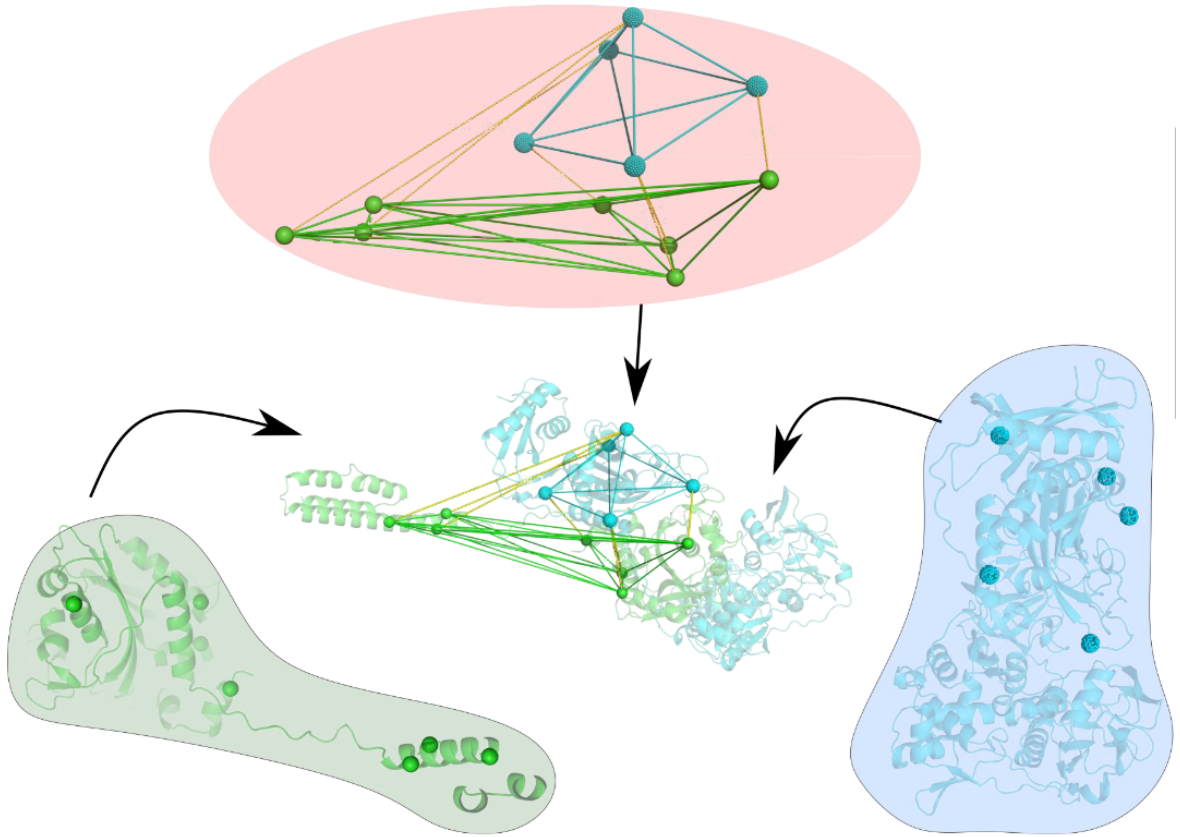


Figure 15: One shot registration

Some Observations

1. In the hypothetical framework, we have all pairs of distances between C-alpha atoms in each protein.

Some Observations

1. In the hypothetical framework, we have all pairs of distances between C-alpha atoms in each protein.
2. For registering n proteins, only 1 hypothetical fragment is needed. So registration of $n + 1$ proteins is done.

Implementation

Consider the case of "1dfj" which has E and I chain. We have crosslink data available. Here is the true structure:

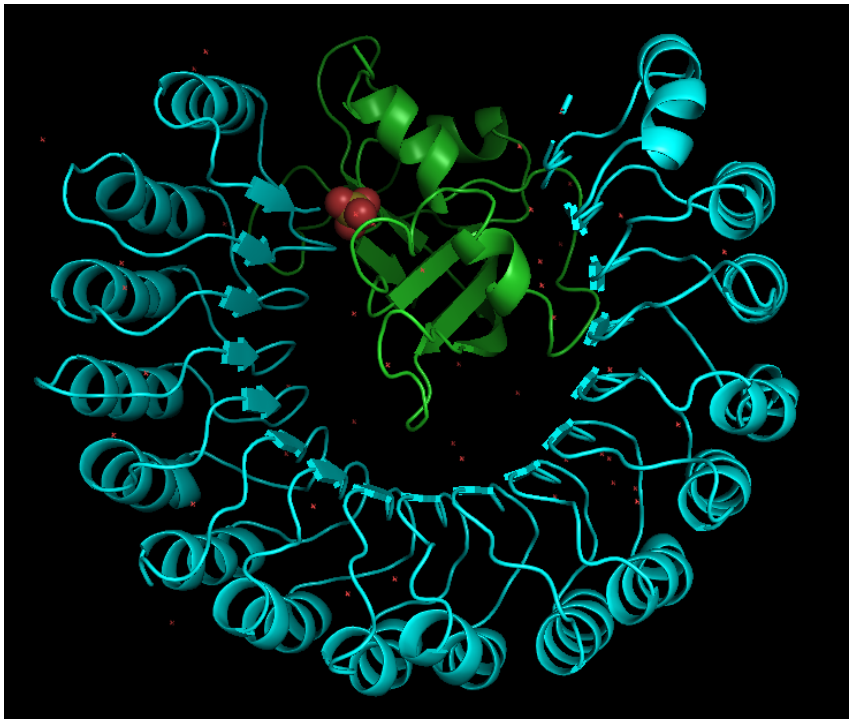


Figure 16: True structure of 1dfj

Structure obtained using our method

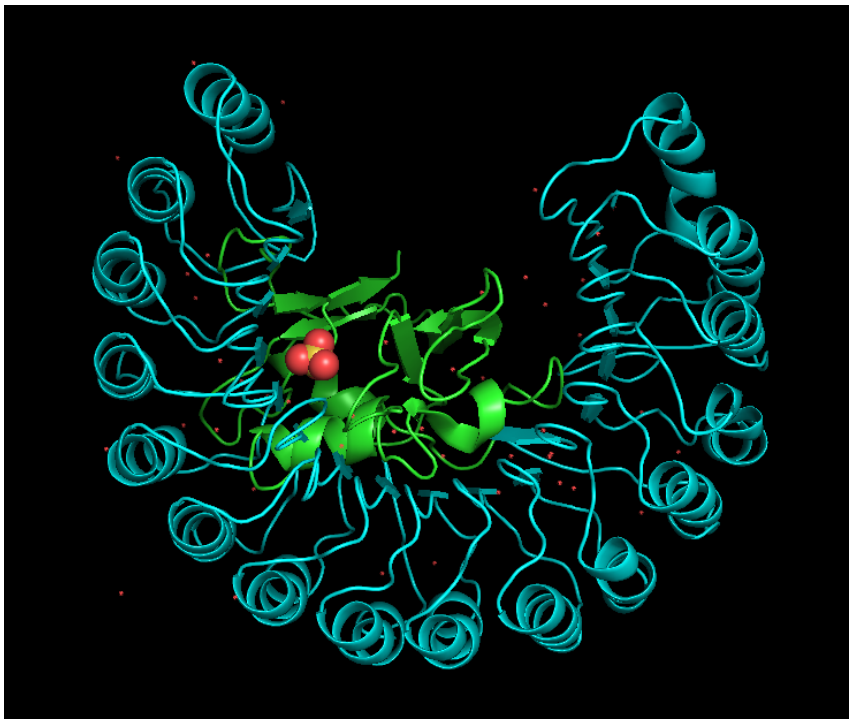


Figure 17: Structure obtained using our method

Structure obtained using our method

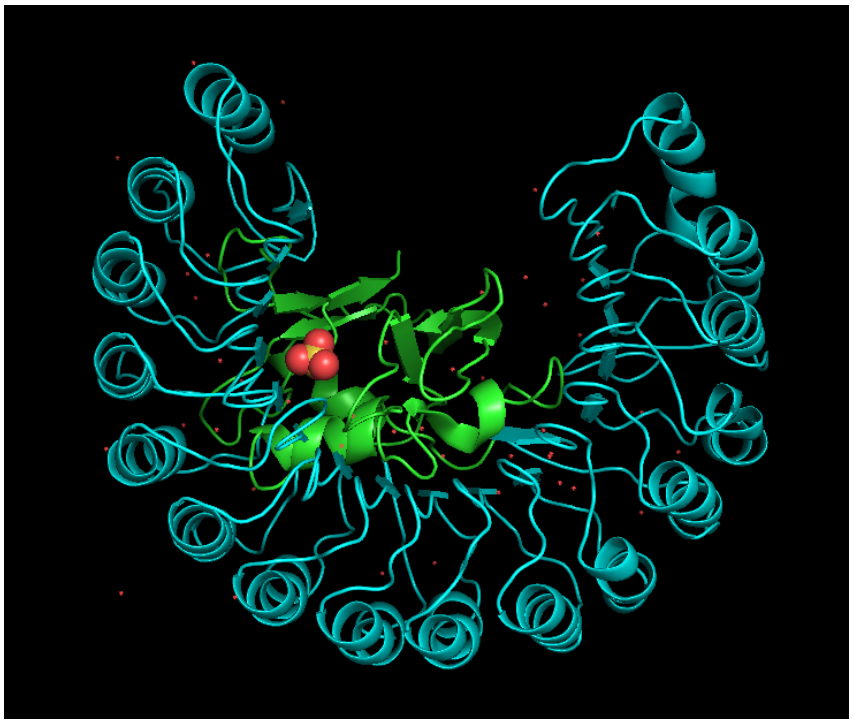


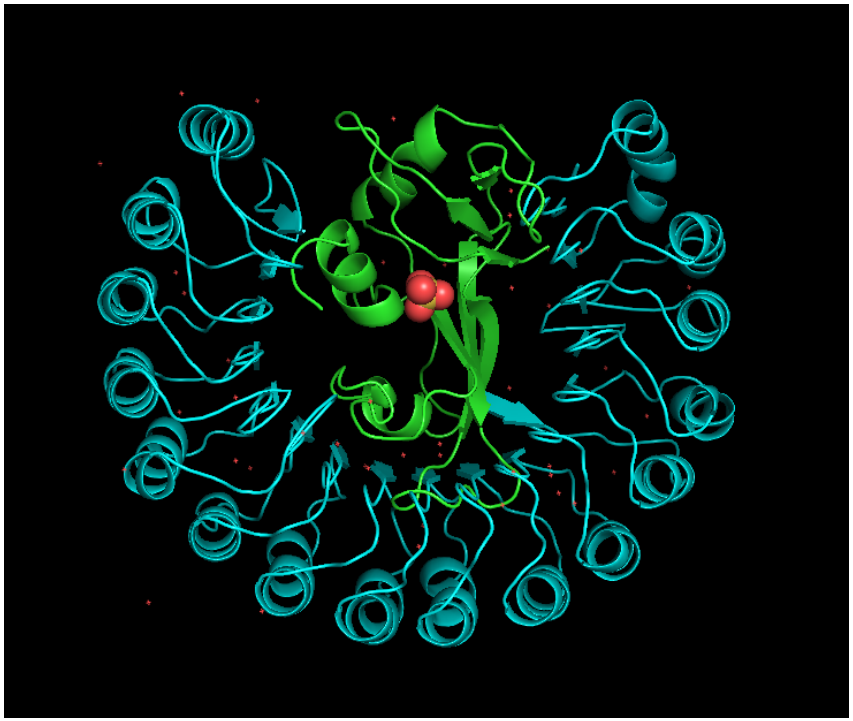
Figure 17: Structure obtained using our method

There are 2 problems:

1. Clashes between proteins
2. Flipped Ramachandran angles

Possible Fix for Clashes

Take random subset of crosslink data instead of all crosslink data. Here is the result for random subset of 6 crosslinks out of 12:



Again there is flipped Ramachandran angles, but clashes have reduced.

Next Steps

1. Fix the flipped Ramachandran angles
2. If this does not work, then translate to remove the clash keeping the orientation same
3. Energy minimization

In this way we are taking experimental data into consideration instead of random sampling of initial configurations.

Energy Minimization

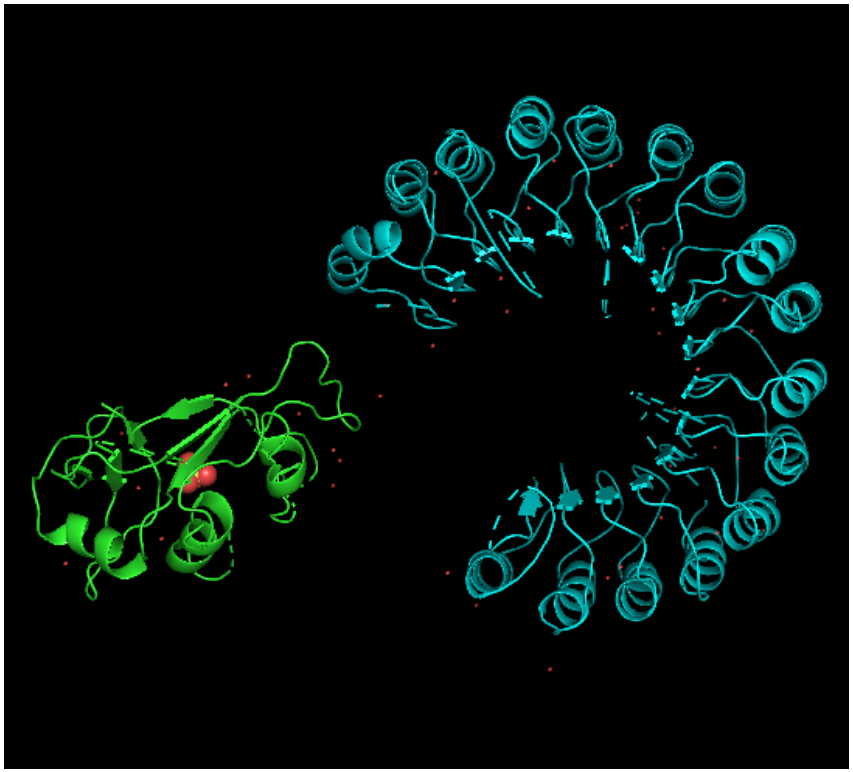


Figure 18: Translating with same orientation

Now we run Energy minimization to get the final structure, which will satisy the given experimental (crosslink) data.

3 Human Practices

Human Practices

The 3R's

- Reflection
- Responsibility
- Responsiveness

Interaction with stakeholders

- Professors
- Industries and Business Professionals
- Research Students

Education

1. Online and Live talks with school students about Computational Biology and AI
2. Inclusivity of college women students, village school students across India
3. Working with Foundations like Steam Vision Foundation and Ladakh Science Foundation
4. Computational Biology Ideathon among colleges
5. Structural Biology workshops for iGEMers
6. Talks by professors from Academia

Our Mentors

(a) Prof Debnath Pal (Dept for Computational and Data Sciences)

(b) Dr. Shruthi(NCBS)

(c) Dr. Manjula Das (MSMF)

To Summarize

1. Extract the distance restraints from various experimental data.
2. Use the principle of DREAM algorithm to model the complex.
3. Generate the PDB file.

Thank You!

Thank You!

Here are some references:

1. DREAMweb, https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/10.1002/pmic.202300379?af=R
2. Improved NMR-data-compliant protein structure modeling captures context-dependent variations and expands the scope of functional inference, https://onlinelibrary.wiley.com/doi/full/10.1002/prot.26439?_gl=1*1f1toyi*_gcl_au*MjcyNjE4NzcOLjE3MTg5MDgxOTg.
3. Figure 1: https://erc.europa.eu/projects-statistics/science-stories/computational-biology-spotlight-erc-projects

Thank You

4. Figure 2,5: https://www.sciencedirect.com/science/article/pii/S002192582100538X and https://www.sciencedirect.com/science/article/pii/S002192582100538X