

Comparative Modeling using Rosetta Loop Building

Greg Sliwoski

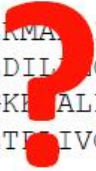
gregory.r.sliwoski@vanderbilt.edu

Rosetta Workshop 2014

Introduction to Comparative Modeling

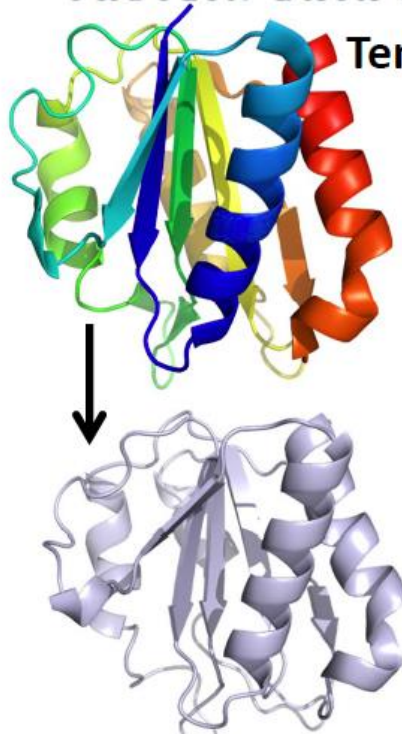
Target

MKIVYWSGTGNTERMAIAKGIIESGKDVNTI
NVSDVNIDELLNEDIIGCSAMGDEVLEESEF
EPFIEEISTKISGKIALFGSYGWGDGKWMRDF
EERMNGYGCVVETIVQNEPDEAEQDCIEFG
KKIANI



RCSB
PDB
PROTEIN DATA BANK

Template

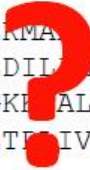


- **Comparative Modeling:** construction of an atomic-resolution model of a protein with no experimentally determined structure using the 3D structure of a related protein
- **Homology Modeling:** modeling a protein based on a template with common evolutionary origin

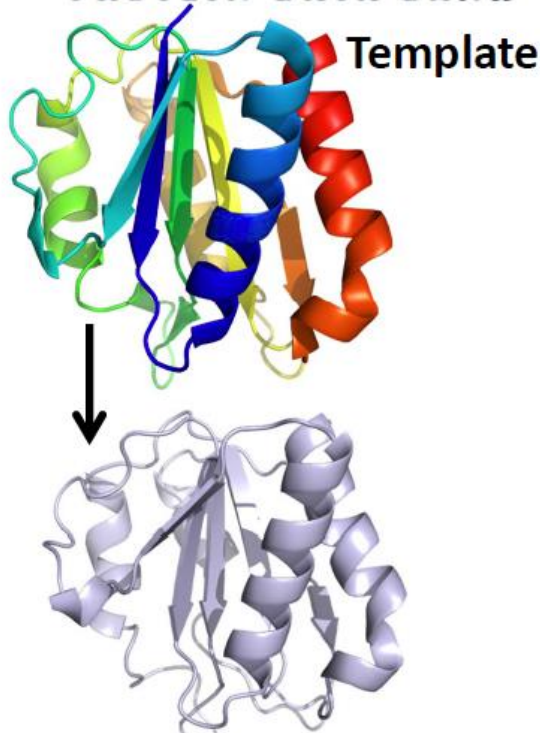
Identifying Template Structures

Target

MKIVYWSGTGNTERMAIAKGIIESGKDVNTI
NVSDVNIDELLNEDIIIGCSAMGDEVLEESEF
EPFIEEISTKISGKIALFGSYGWGDGKWMRDF
EERMNGYGCVVETIVQNEPDEAEQDCIEFG
KKIANI



RCSB **PDB**
PROTEIN DATA BANK



- **Suitable Templates:** ideally have >30% sequence identity to target
- **Identifying Templates:**
 - **Sequence Similarity:** compare proteins based on amino acid properties alone (BLAST, PSI-BLAST)
 - **Fold Recognition:** using predicted secondary structure information to detect proteins with similar 3D characteristics (DALI, PHYRE)

Some GPCR templates

Inactive

1u19: Bovine Rhodopsine
2rh1: Human β 2-adrenergic
2vt4: Turkey β 1-adrenergic
3eml: Human adenosine 2A
3odu: Human CXCR4
3pbl: Human dopamine D3
3rze: Human histamine H1
3uon: Human muscarinic M2
3v2w: Human sphingosine 1-phosphate
4daj: Rat muscarinic M3
4djh: Human kappa-opioid
4dkl: Mouse mu-opioid
4ea3: Human N/OFQ opioid
4ej4: Mouse delta-opioid

Potentially Active

2x72: rhodopsin (constitutively active mutant)
2y03: Turkey β 1-adrenergic
3qak: Human adenosine 2A
3sn6: Human β 2-adrenergic (bound to Gs)
4iar: Human 5HT-1B
4ib4: Human 5HT-2B

Note: Rosetta will sample active and inactive states
Both can be used for comparative modeling

Comparative Modeling Protocol

- **Step 1:** Align target sequence with sequence of template structure

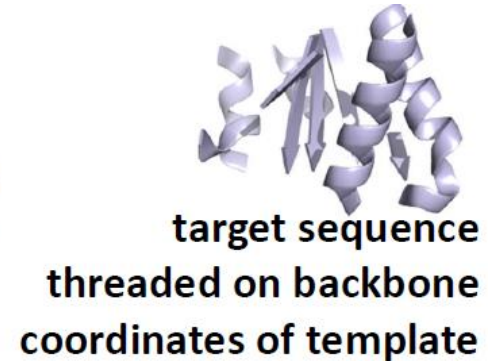
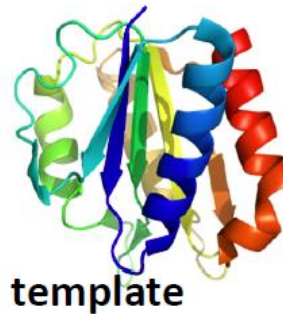
```
target    1  --MKIVYWSGTGNTEKMAELIAKGIIE
template  1  PKALIVYGSTTGNTHEYTAETIARELAD
```

- **Step 2:** Thread the target sequence onto the backbone of the template structure
- **Step 3:** Rebuild the loop regions of the model
- **Step 4:** Full-atom refinement
- **Step 5:** Cluster models and analyze results

Comparative Modeling Protocol

- **Step 1:** Align target sequence with sequence of template structure
- **Step 2:** Thread the target sequence onto the backbone of the template structure
- **Step 3:** Rebuild the loop regions of the model
- **Step 4:** Full-atom refinement
- **Step 5:** Cluster models and analyze results

```
target    1  --MKIVYWSGTGNTEKMAELIAKGIIE
template  1  PKALIVYGSTTGNTHEYTAETIARELAD
```



Threading

Template:

(0,0,0)	(1,1,1)	(2,2,2)	(3,3,3)	(4,4,4)	(5,5,5)	
L	K	R	N	N	H	-
(?,?,?)	(?,?,?)				(?,?,?)	(?,?,?)
L	K	-	-	-	H	V

Target:

*Thread
Coordinates*



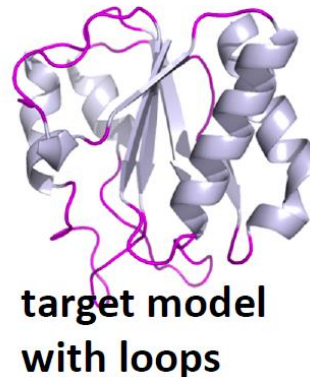
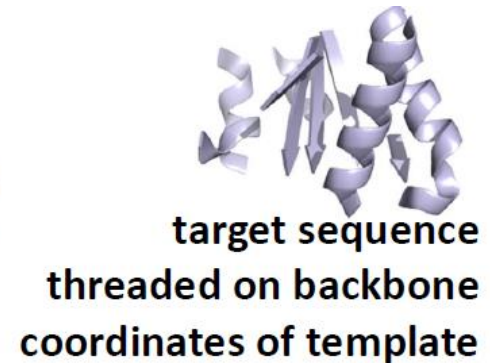
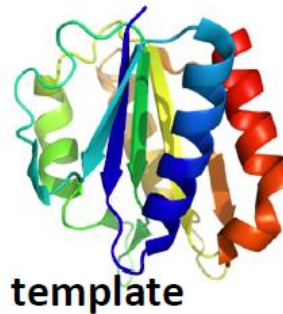
Target:

(0,0,0)	(1,1,1)				(5,5,5)	(?,?,?)
L	K	-	-	-	H	V

Comparative Modeling Protocol

- **Step 1:** Align target sequence with sequence of template structure
- **Step 2:** Thread the target sequence onto the backbone of the template structure
- **Step 3:** Rebuild the loop regions of the model
- **Step 4:** Full-atom refinement
- **Step 5:** Cluster models and analyze results

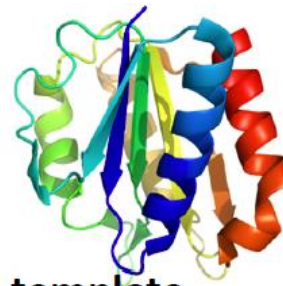
```
target    1  --MKIVYWSGTGNTEKMAELIAKGIIE
template  1  PKALIVYGSTTGNT EYTAETIARELAD
```



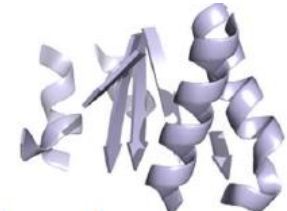
Comparative Modeling Protocol

- **Step 1:** Align target sequence with sequence of template structure
- **Step 2:** Thread the target sequence onto the backbone of the template structure
- **Step 3:** Rebuild the loop regions of the model
- **Step 4:** Full-atom refinement
- **Step 5:** Cluster models and analyze results

```
target    1  --MKIVYWSGTGNTTEKMAELIAKGIIE
template  1  PKALIVYGSTTGNTTEYTAETIARELAD
```



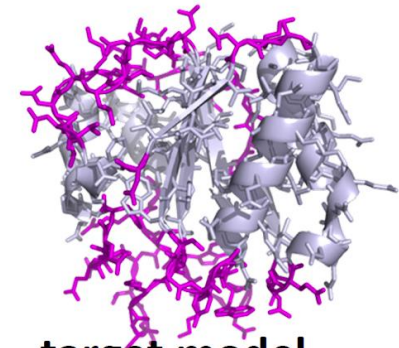
template



target sequence
threaded on backbone
coordinates of template



target model
with loops

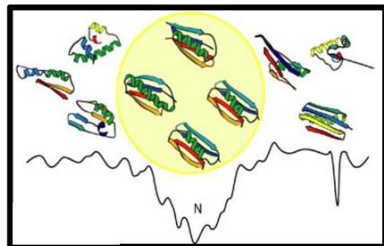
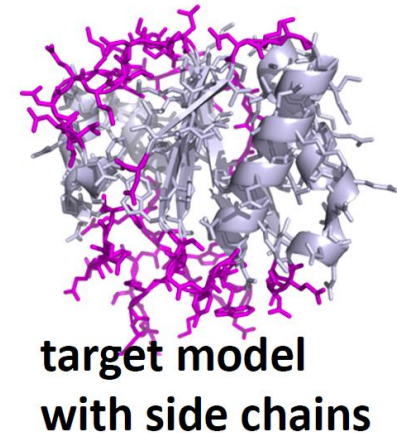
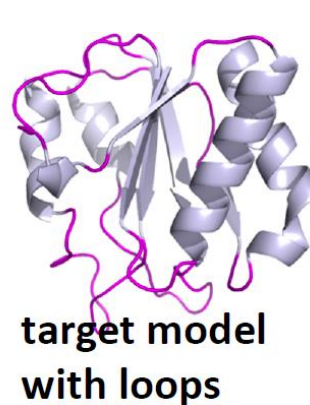
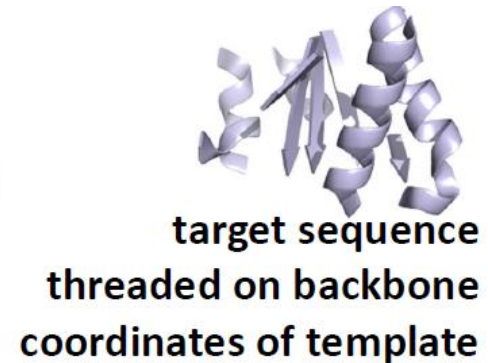
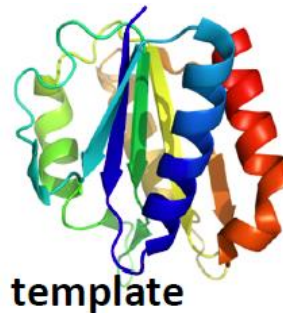


target model
with side chains

Comparative Modeling Protocol

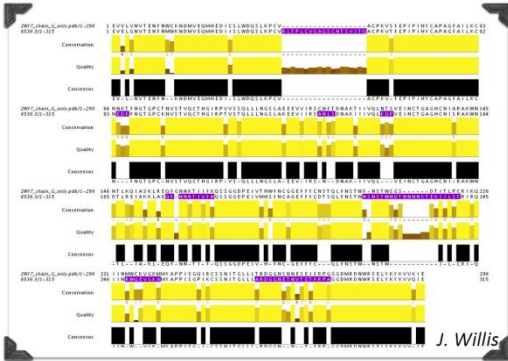
- **Step 1:** Align target sequence with sequence of template structure
- **Step 2:** Thread the target sequence onto the backbone of the template structure
- **Step 3:** Rebuild the loop regions of the model
- **Step 4:** Full-atom refinement
- **Step 5:** Cluster models and analyze results

```
target    1  --MKIVYWSGTGNTTEKMAELIAKGIIE
template  1  PKALIVYGSTTGNTTEYTAETIARELAD
```

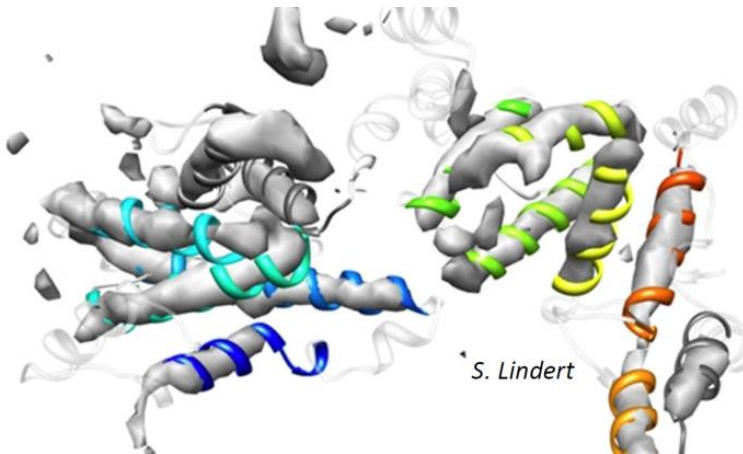
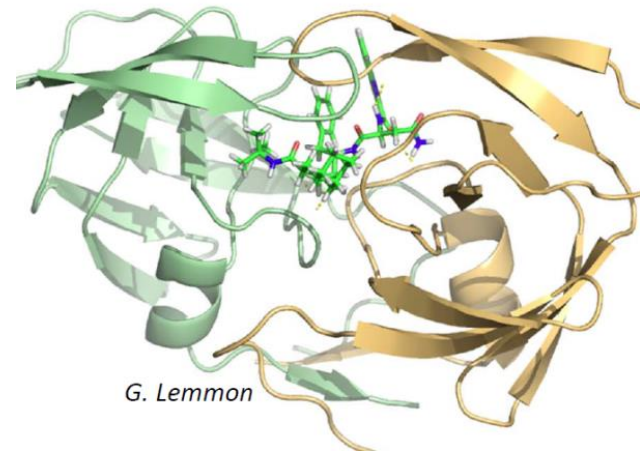


Loop Building is Important

- Low sequence identity in loop regions (gaps)
- Loops play important functional roles



- Experimental structures often have missing densities



Loop Building In Rosetta

Stage 1: Rebuild

1. Insert loop fragments (fragment files)
2. Cyclic Coordinate Descent (CCD) loop closure

Result: Low resolution centroid-based loop

Stage 2: Refine

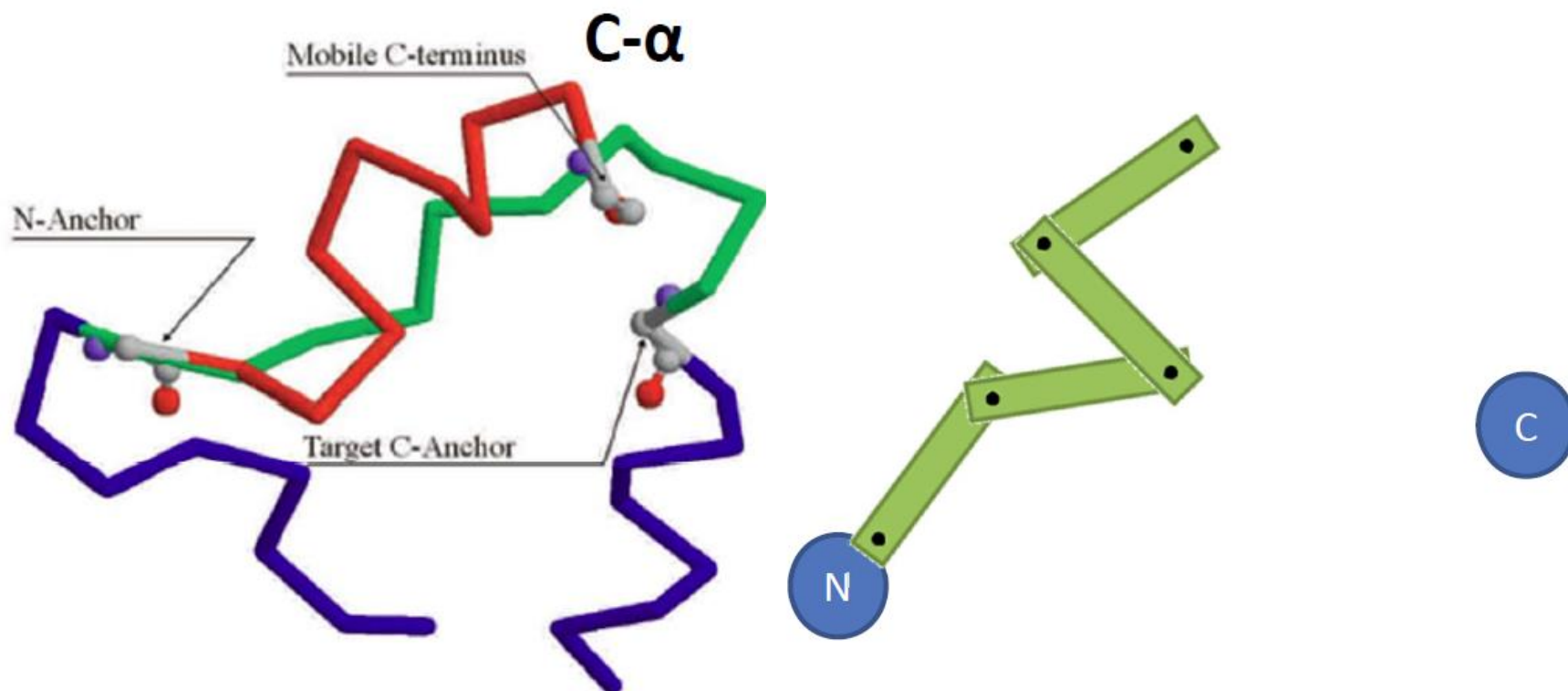
1. Kinematic Closure (KIC)

Result: High resolution full atom loops

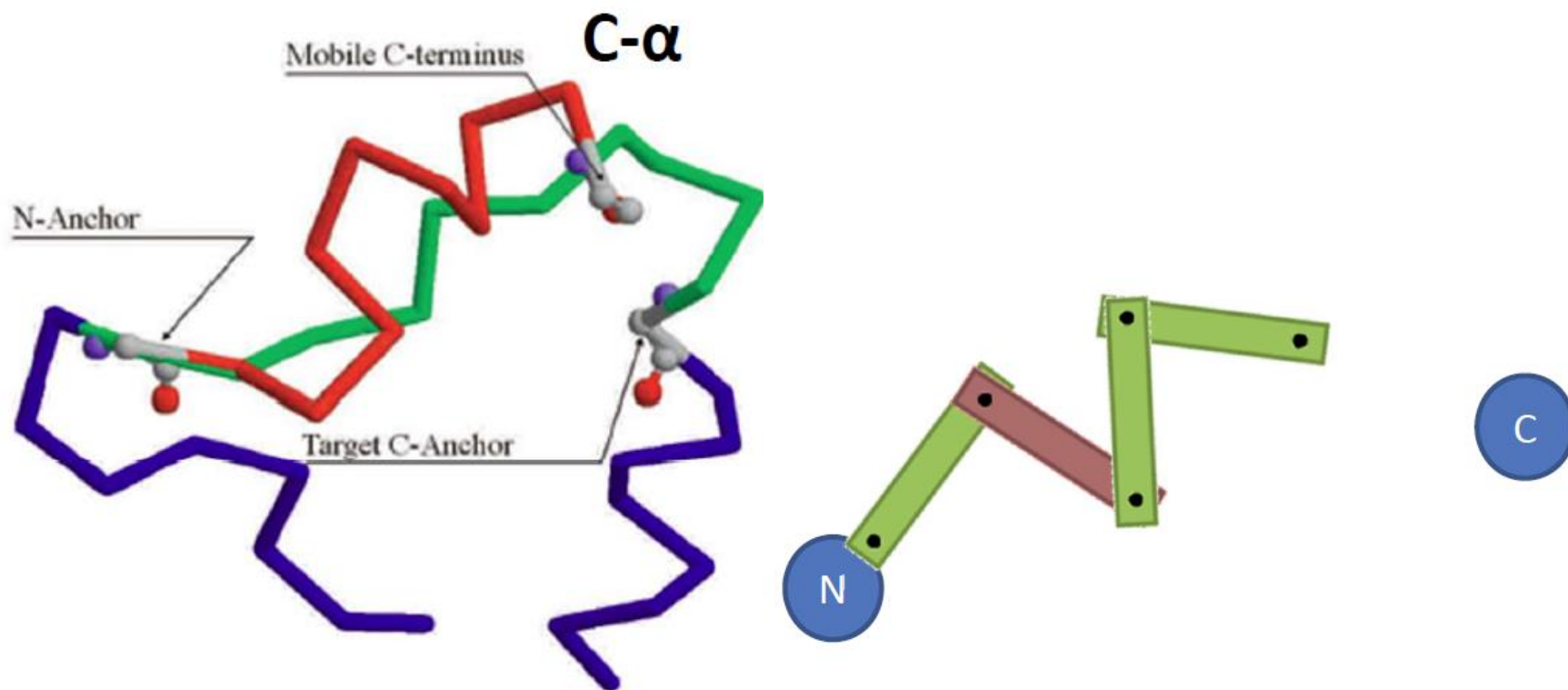
Cyclic Coordinate Descent

- Inverse kinematics method adopted from robotic engineering
- The loop is an arm and the amino acid centroids are joints
- Flexible and computationally fast
- Can close loops from nearly any starting configuration as long as the chain is long enough

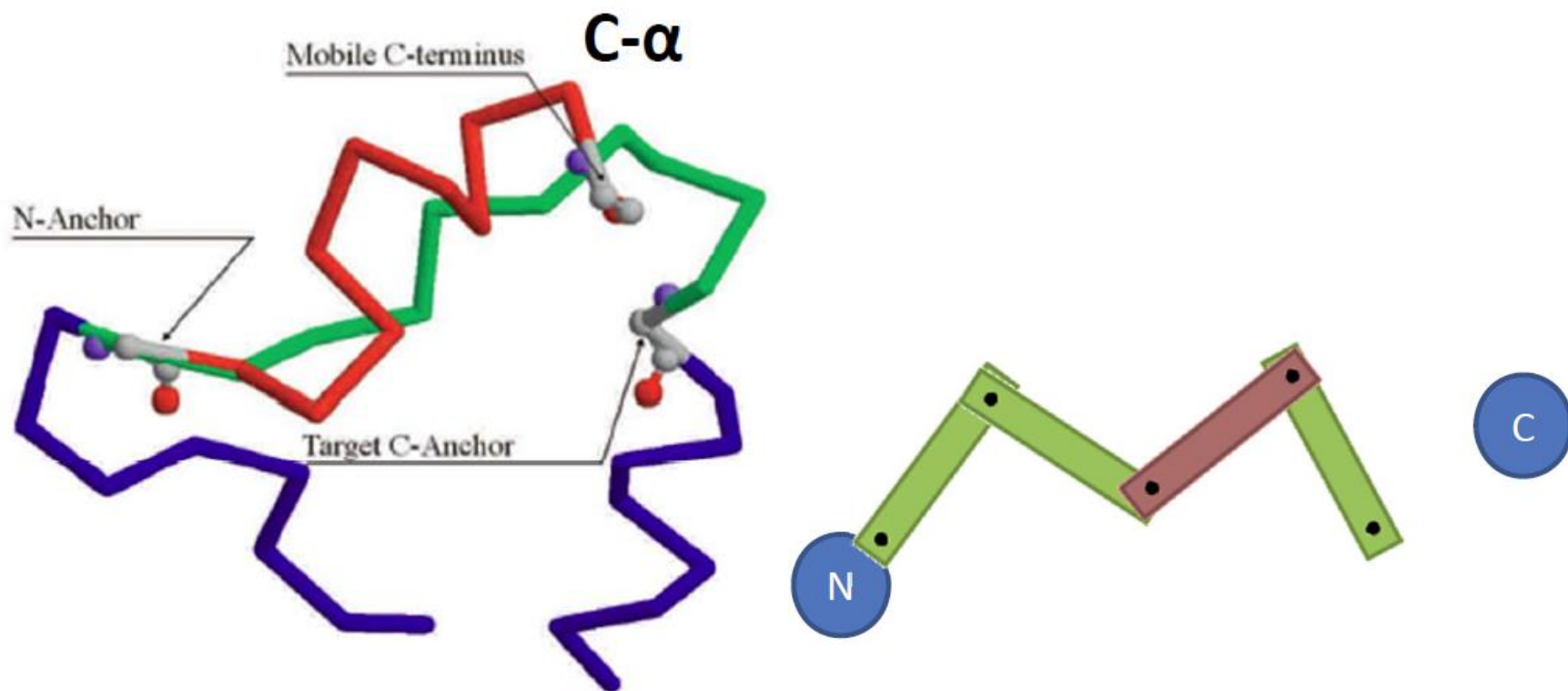
Loop Building with Cyclic Coordinate Descent



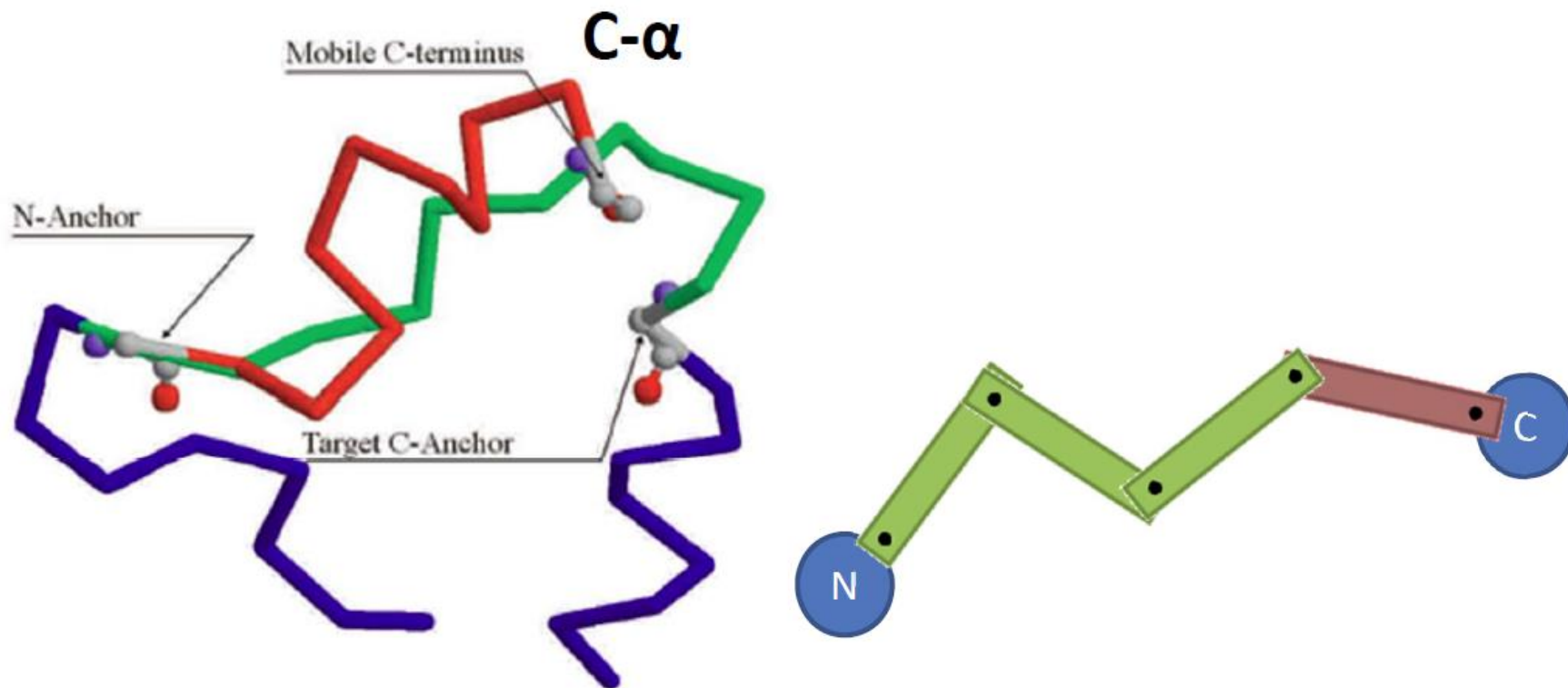
Loop Building with Cyclic Coordinate Descent



Loop Building with Cyclic Coordinate Descent

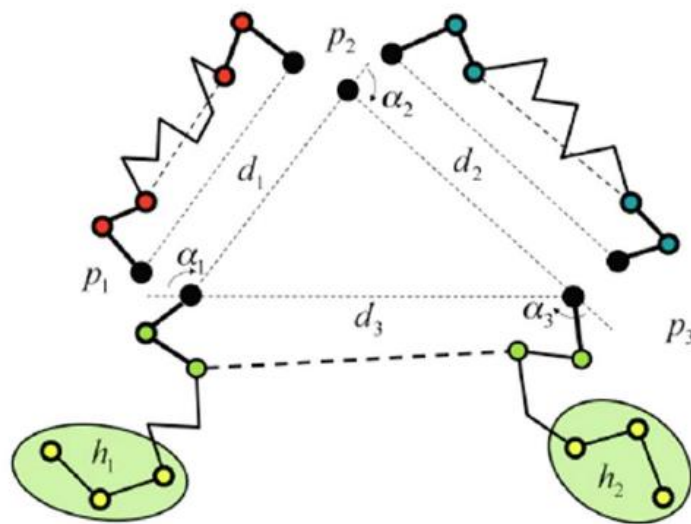
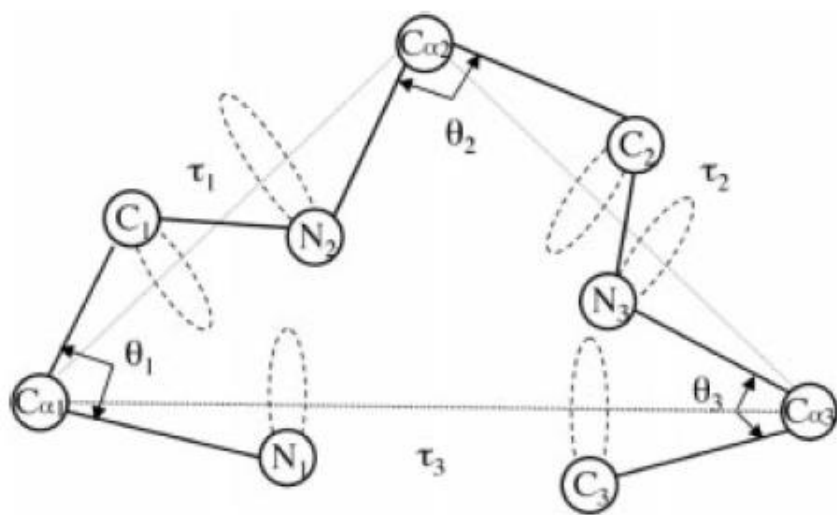


Loop Building with Cyclic Coordinate Descent



Kinematic Closure (KIC) Refinement

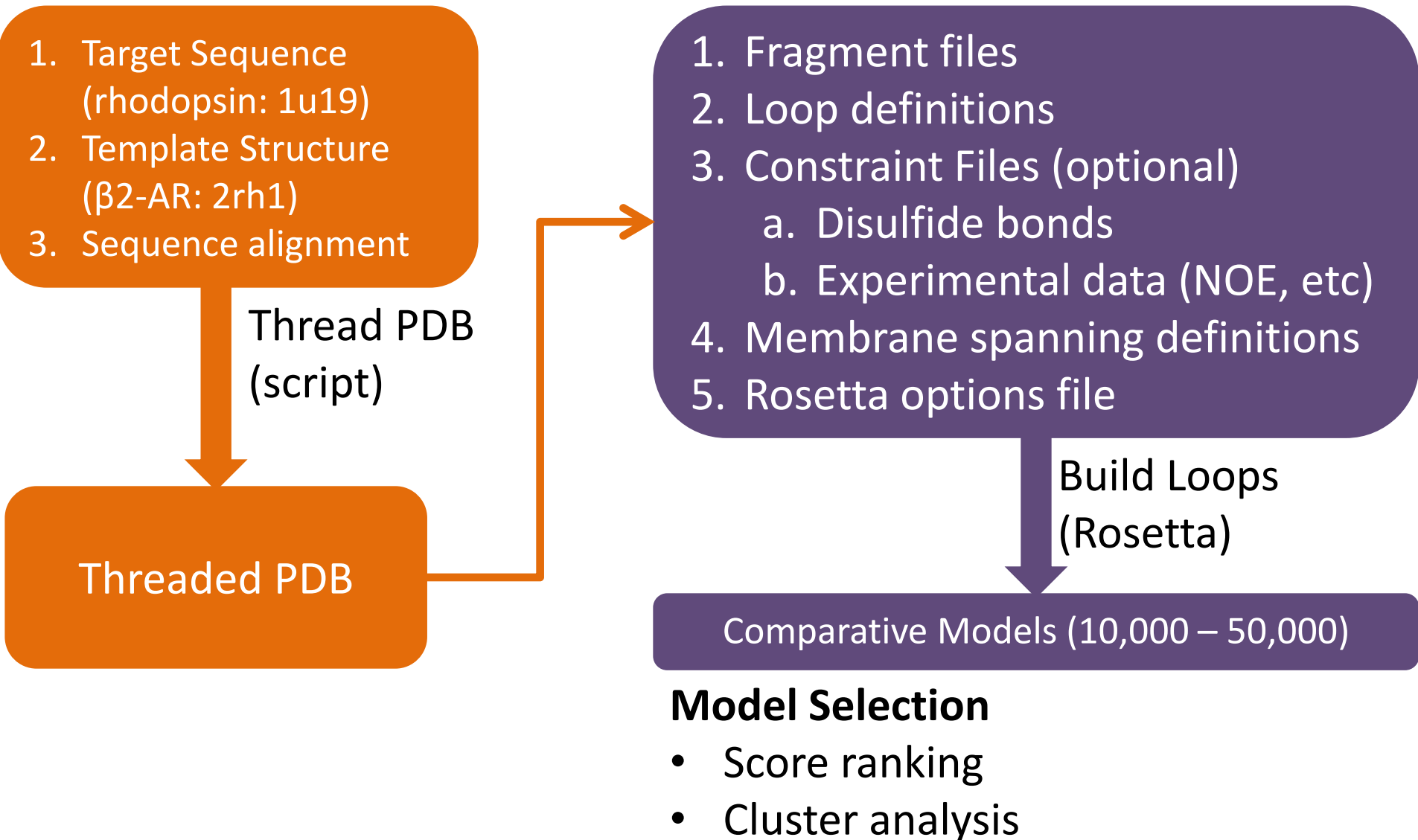
- Defines and fixes pivot points between which 3 phi-psi pairs are sampled at a time
- Samples side chain torsions using polynomial resultants providing full atom refinement.



CCD and KIC

- KIC may also be used for loop closure
 - Low resolution simulated annealing Monte Carlo
 - Can not sample terminal loops (N and C terminals are tethered)
 - More restricted than CCD
- CCD is designed for loop closure not sampling
 - Does not provide full-atom loops
 - Iterative nature favors large changes in first residues of loop
 - Less accurate without KIC refinement

Rosetta protein modeling in practice



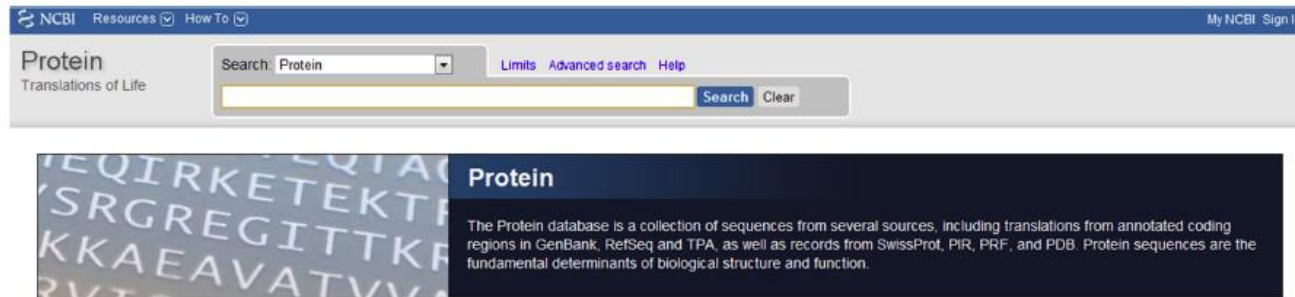
1. Target Sequence

Bovine rhodopsin (PDB: 1u19)

Find this file at /protein_modeling/input_model/1u19A.fasta

>1u19A

PWQFSMLAAYMFLLIMLGFPINFLTLYVTVQHKKLRTPLNYILLNLAVADLFMVFGG
FTTTLYTSLHGYFVFGPTGCNLEGFFATLGGEIALWSLVVLAIERVYVVCKPMSNFR
FGENHAIMGVAFTWVMALACAAPPLVGWSRYIPEGMQCSCGIDYYTPHEETNNESFV
IYMFVVHFIIPLIVIFFCYGQLVFTVKEAAAQQQESATTQKAEKEVTRMVIIMVIAF
LICWLPYAGVAFYIFTHQGSDFGPIFMTIPAFFAKTSAVYNPVIYIMMNKQFRNCMV
TTLCCG



<http://www.ncbi.nlm.nih.gov/protein>

2. Template PDB

Human β 2-adrenergic receptor (PDB: 2rh1)

Find this file at /protein_modeling/input_model/2rh1A.pdb

The screenshot shows the RCSB PDB website homepage. At the top, the logo for RCSB PDB (Protein Data Bank) is on the left, and a banner on the right states 'A MEMBER OF THE PDB' and 'An Information Portal to Biological Macromolecular Structures'. Below the banner, it says 'As of Tuesday Feb 22, 2011 at 4 PM PST there are 71415 Structures' with links to social media and PDB Statistics.

The main navigation bar includes 'Contact Us | Print', a search box for 'PDB ID or Text', and a link to 'PDB ID lookup or Text search of the complete structure file'. There are also links for 'Search' and 'Advanced Search'.

The left sidebar contains several sections: 'MyPDB' with links to login/register; 'Home' with links to news, policies, FAQ, contact, about, careers, external links, sitemap, and new website features; 'Deposition' with links to deposit services, electron microscopy, X-ray, NMR, validation server, BioSync Beamline, and related tools; and 'Search' with a link to advanced search.

The main content area is titled 'A Resource for Studying Biological Macromolecules'. It contains a paragraph about the PDB archive and its role in curating and annotating PDB data. Below this, there is a section for 'Featured Molecules' with a 'List View of Archive By: Title | Date | Category' and a 'Hide' button. The featured molecule is 'Integrin', described as the 'Molecule of the Month'. The description states: 'Our bodies are composed of approximately ten trillion cells, which poses challenging problems for structure and communication. All of these cells must be connected strongly together, to allow us to stand and walk. The infrastructure holding us together, however, must also be malleable enough to allow repairs, to allow us to heal from wounds. These many cells must also communicate with each other, ensuring that each plays its own proper part. Many different molecules in our bodies are involved in this complex infrastructure of structural and communication, and integration into a coherent whole.'

On the right side, there is a 'Customize This Page' button and a 'New Features' section with a 'Hide' button. Below this is a 'Transporter Classification Database Browser' with a 'Latest features released:' section and a 'Website Release Archive:' dropdown menu. Further down is the 'RCSB PDB News' section with a 'Hide' button and links for 'Weekly | Quarterly | Yearly'. The news section is dated '2011-02-22' and titled 'Structural Neighbors'. It features a 3D ribbon diagram of a protein structure and a link to 'Explore structural neighbors lists to find connections between PDB'.

<http://www.rcsb.org>

3. Sequence Alignment

Find this file at /protein_modeling/input_model/1u19A.2rh1A.aln

```
2rh1A  DEVWVVGMGIVMSLIVLAIVFGNVLVITAIKFERLQTVTNYFITSACADLVMGLAVVP
1u19A  -PWQFSMLAAYMFLLIMLGFPINFLTLYVTVQHKKLRTPLNYILLNLAVADLFMVFGGFT
      .  :.  *  *:::  .  *.*:  .  .:::*.~  **::  .**  ***.*  :.  ..
```



Multiple Sequence Alignment by CLUSTALW

CLUSTALW MAFFT PRRN

Help

General Setting Parameters:

Output Format:

Pairwise Alignment: ☒ FAST/APPROXIMATE ☐ SLOW/ACCURATE

Enter your **sequences** (with labels) below (copy & paste): ☒ PROTEIN ☐ DNA

Support Formats: FASTA (Pearson), NBRF/PIR, EMBL/Swiss Prot, GDE, CLUSTAL, and GCG/MSF

Or give the file name containing your query

No file chosen

www.genome.jp/tools/clustalw/

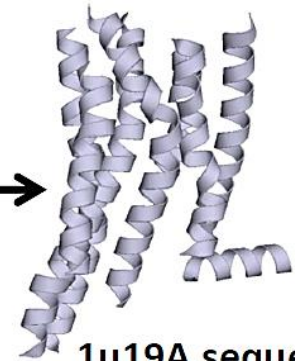
4. Threading

Find this file at /protein_modeling/input_model/1u19A_on_2rh1A.pdb

```
thread_pdb_from_alignment.py --template=2rh1A  
--target=1u19A --chain=A --align_format=clustal  
1u19A.2rh1A.aln 2rh1A.pdb 1u19A_on_2rh1A.pdb
```

2rh1A	DEVWVVGMGIVMSLIVLAIVFGNVLVITAIKFE
1u19A	-PWQFSMLAAYMFLLIMLGFPINFLTLYVTVQHK

+



**1u19A sequence
threaded on backbone
coordinates of 2rh1A**

5. Fragment Files

Find these files at

*/protein_modeling/input_model/aa1u19A03_05.200_v1_3 and
/protein_modeling/input_model/aa1u19A09_05.200_v1_3*

www.bakerlab.org

 **ROBETTA** **BETA**
Full-chain Protein Structure Prediction Server

Structure Prediction Fragment Libraries Alanine Scanning DNA Interface Scan
[\[Queue \]](#) [\[Submit \]](#) [\[Queue \]](#) [\[Submit \]](#) [\[Queue \]](#) [\[Submit \]](#) [\[Queue \]](#) [\[Submit \]](#)
[\[Register / Update \]](#) [\[Docs / FAQs \]](#) [\[Login \]](#)

Submit a job to the Fragment Server

*Please submit one job at a time

Required
[Registered Username:](#) or [Registered Email Address:](#)

Target Name:

Paste [Fasta](#)

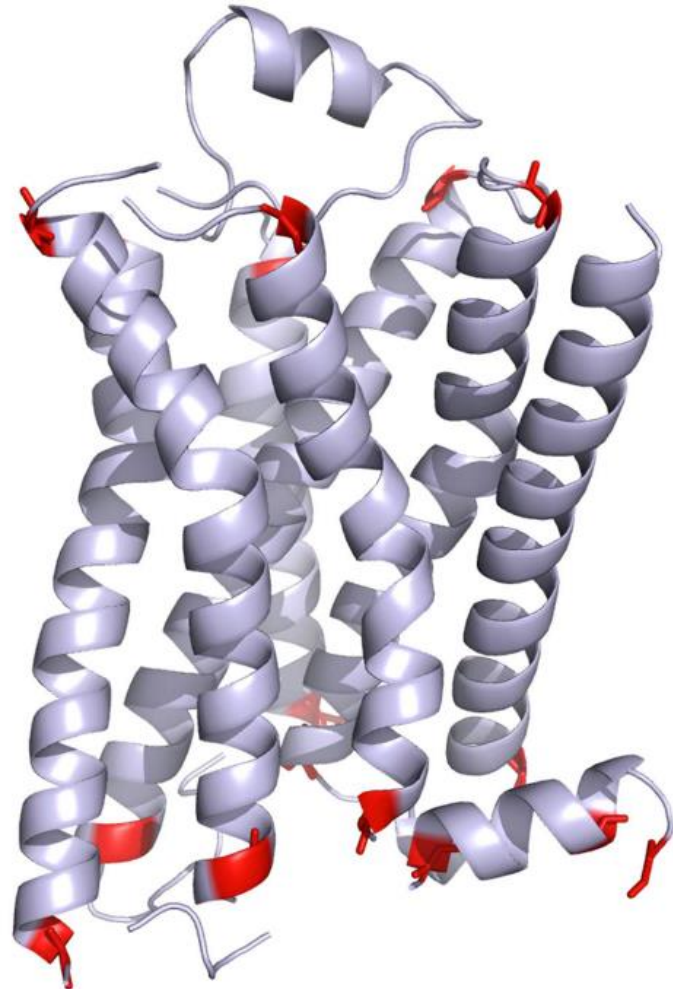
or Upload [Fasta](#): No file chosen

[http://robeta.bakerlab.org/
fragmentsubmit.jsp](http://robeta.bakerlab.org/fragmentsubmit.jsp)

6. Loop definitions

Find this file at /protein_modeling/input_model/1u19A.loops

LOOP 31 38
LOOP 67 74
LOOP 106 118
LOOP 139 169
LOOP 200 216
LOOP 244 252
LOOP 275 279
LOOP 286 290

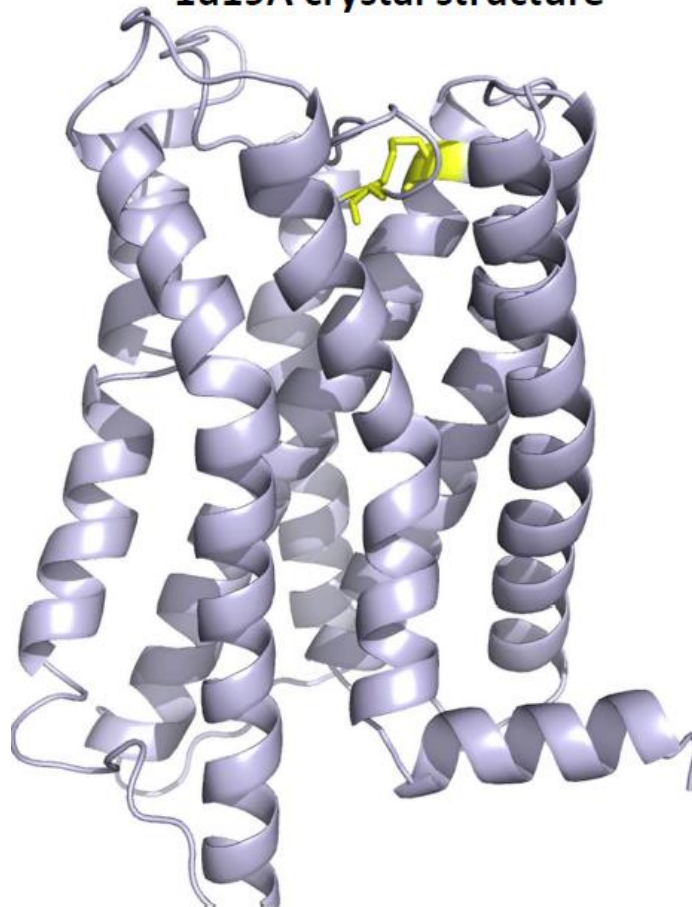


7. Disulfide constraints

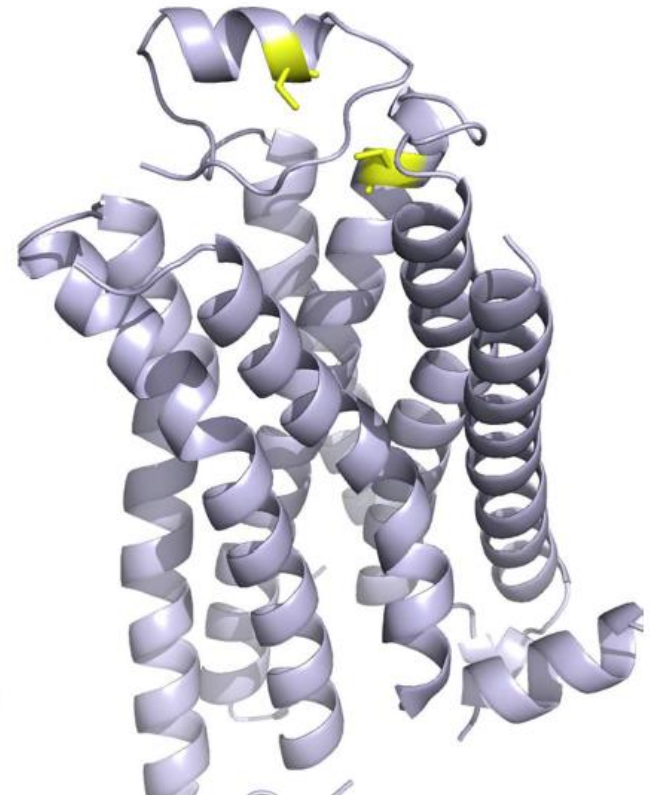
Find this file at /protein_modeling/input_model/1u19A.disulfide

77 154

1u19A crystal structure

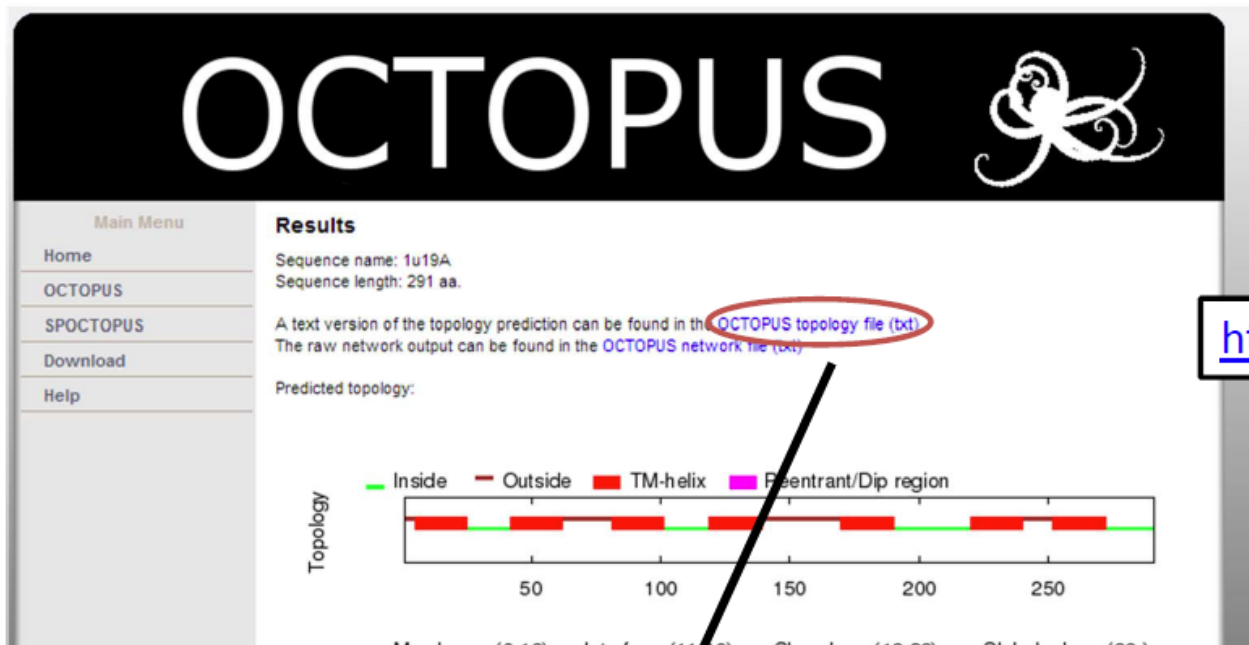


1u19A threaded onto 2rh1A



8. Membrane spanning definitions

Find this file at /protein_modeling/input_model/1u19A.span



<http://octopus.cbr.su.se/>

`octopus2span.pl 1u19A.octopus`

Rosetta Options File

Find this file at /protein_modeling/input_loop/ccd.options

Input/Output Options

-in:file:s 1u19A_on_2rh1A.pdb
-in:file:fullatom
-loops:loop_file 1u19A.loops
-in:fix_disulf 1u19A.disulfide
-in:file:spanfile 1u19A.span
-out:file:silent_struct_type binary
-out:file:silent 1u19A_on_2rh1A_ccd_test_01.out
-out:prefix 1u19A_on_2rh1A_ccd_test
-out:file:fullatom
-nstruct 2

Membrane Protein Options

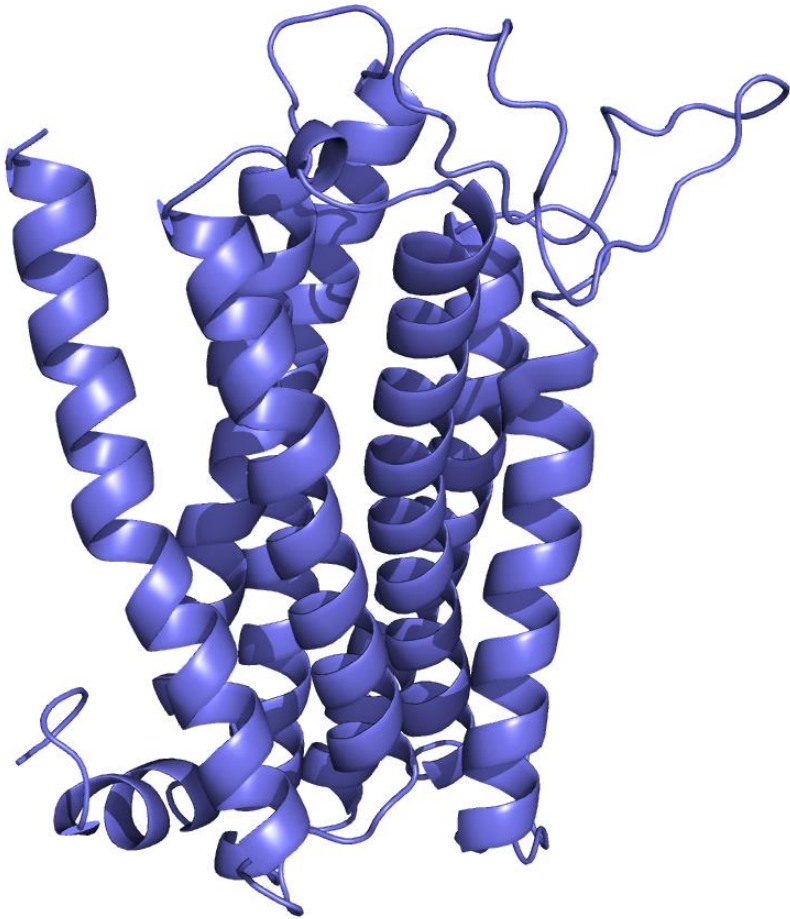
-membrane:no_interpolate_Mpair
-membrane:Menv_penalties
-score:weights membrane_highres_Menv_smooth.wts

Loop Building Options

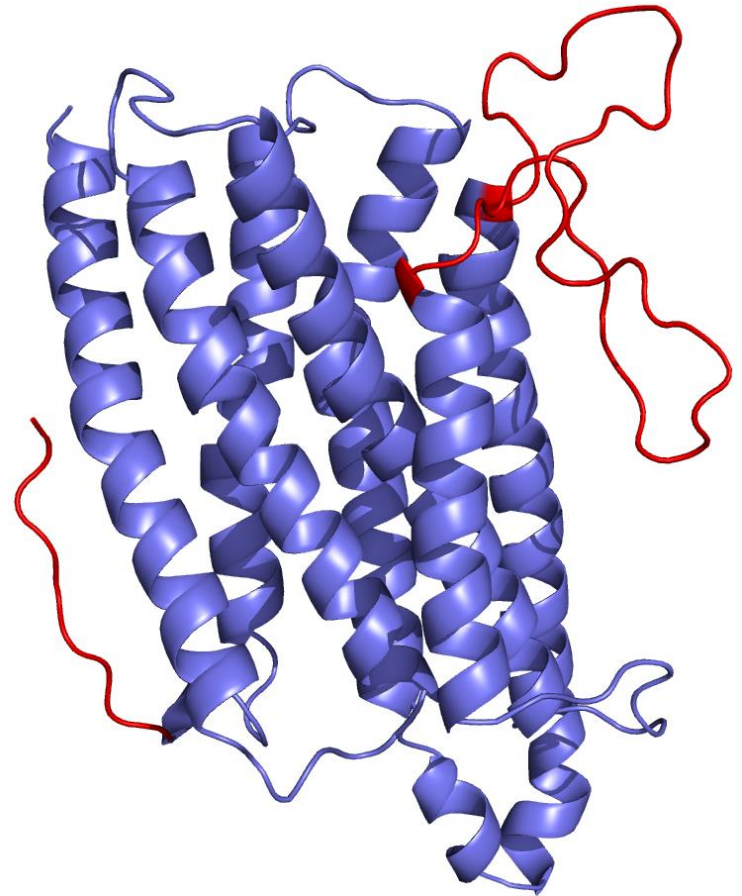
-loops:relax fastrelax
-loops:extended true
-loops:idealize_after_loop_close
-loops:fast
-ex1
-ex2
-loops:frag_sizes 9 3 1
-loops:frag_files <fragment files>
-loops:remodel quick_ccd
-loops:refine refine_kic
-loops:random_loop

Why use membrane scoring terms?

With membrane penalties/weights



Without membrane penalties/weights



Rosetta Membrane

- Rosetta will assign membrane position based on span file.
- Can visualize membrane position
- Rosetta output following span file loading:

```
core.scoring.MembranePotential: MembraneCenter -32.747 19.516 7.875
core.scoring.MembranePotential: MembraneNormal -0.044 -0.994 0.096
core.scoring.MembranePotential: ATOM 9999 X MEM A 999 -32.747 19.516 7.875
core.scoring.MembranePotential: ATOM 9999 Y MEM A 999 -33.410 4.600 9.314
core.scoring.MembranePotential: ATOM 9999 Z MEM A 999 -32.085 34.432 6.437
```

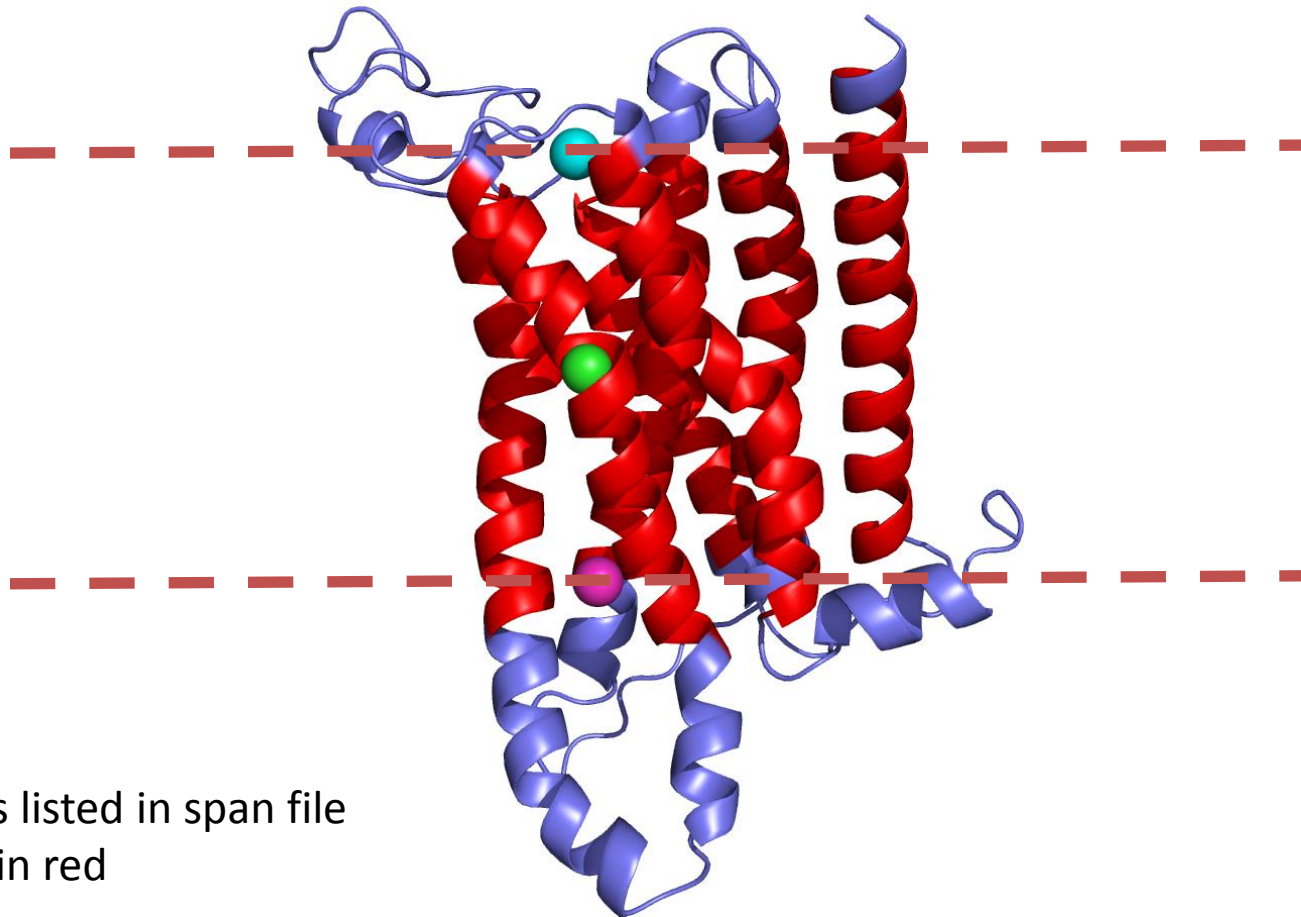
Create pdb file with 3 lines and visualize along with model:

ATOM	9999	X	MEM A 999	-32.747	19.516	7.875
ATOM	9999	Y	MEM B 999	-33.410	4.600	9.314
ATOM	9999	Z	MEM C 999	-32.085	34.432	6.437

Find this file at /protein_modeling/output_model/membrane.pdb

Rosetta Membrane

- In pymol, shown as spheres and colored:



Residues listed in span file
colored in red

RosettaCM

- Automated multi-template comparative modeling

High-Resolution Comparative Modeling with RosettaCM

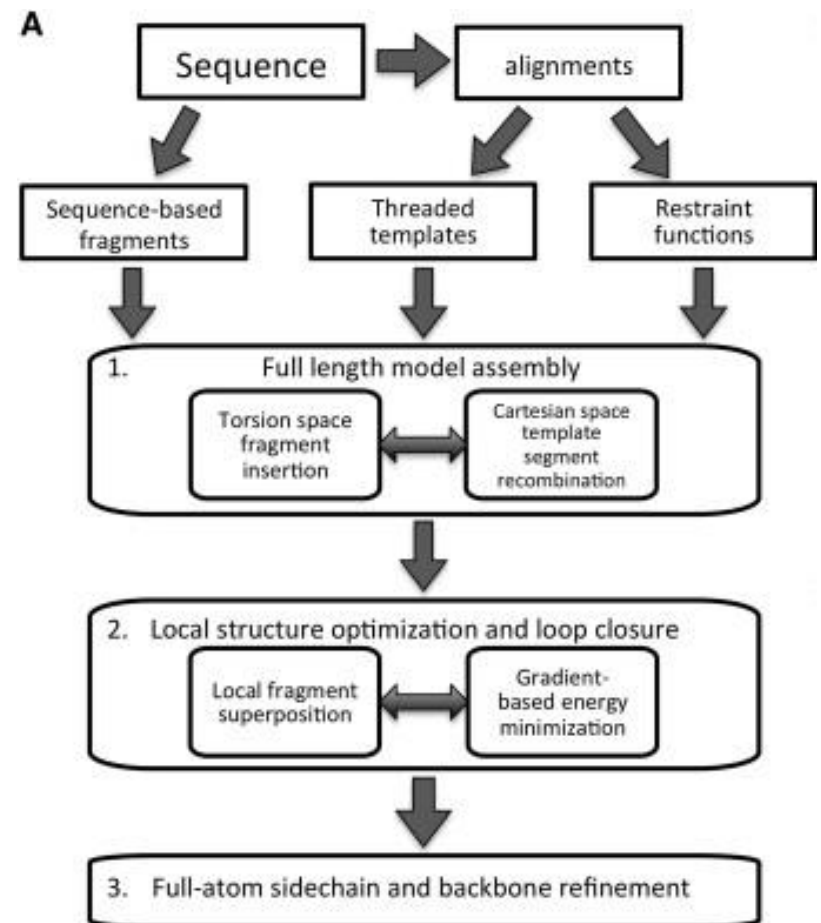
Yifan Song, Frank DiMaio, Ray Yu-Ruei Wang, David Kim, Chris Miles, TJ Brunette, James Thompson, David Baker

Structure

Volume 21, Issue 10, 8 October 2013, Pages 1735–1742

RosettaCM: Three Stages

1. Generate initial models from template alignments
- 2. Explore deviations from templates and close loops with 2-step MC:**
 - Randomly select de novo or template-based fragment and substitute into current conformation
 - Cartesian space full-backbone minimization
3. Full atom refinement and final relax



Song, et al. 2013

Using RosettaCM Locally

1. Generate config files (python script available from supplemental material)

```
setup_rosettaCM.py --fasta input.fasta --alignment input.alignment
```

Required Input:

- Target sequence
- Alignments (clustalW)

2. Run Rosetta with generated config files:

```
rosetta_scripts.linuxgccrelease -parser:protocol rosetta_cm.xml -in:file:fasta input.fasta @flags
```

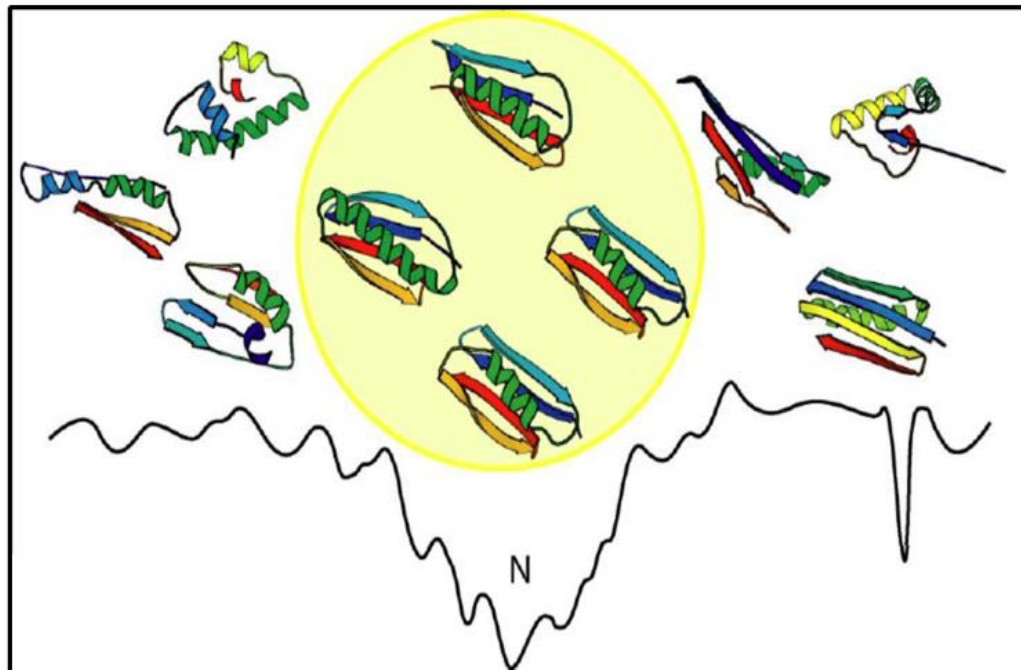
Using RosettaCM online

- ROBETTA web-based application uses RosettaCM to generate 3D models from sequence.
- Completely automated.
- Generates alignments with Hhsearch, SPARKS-X, and RaptorX

<http://robetta.bakerlab.org/submit.jsp>

Clustering

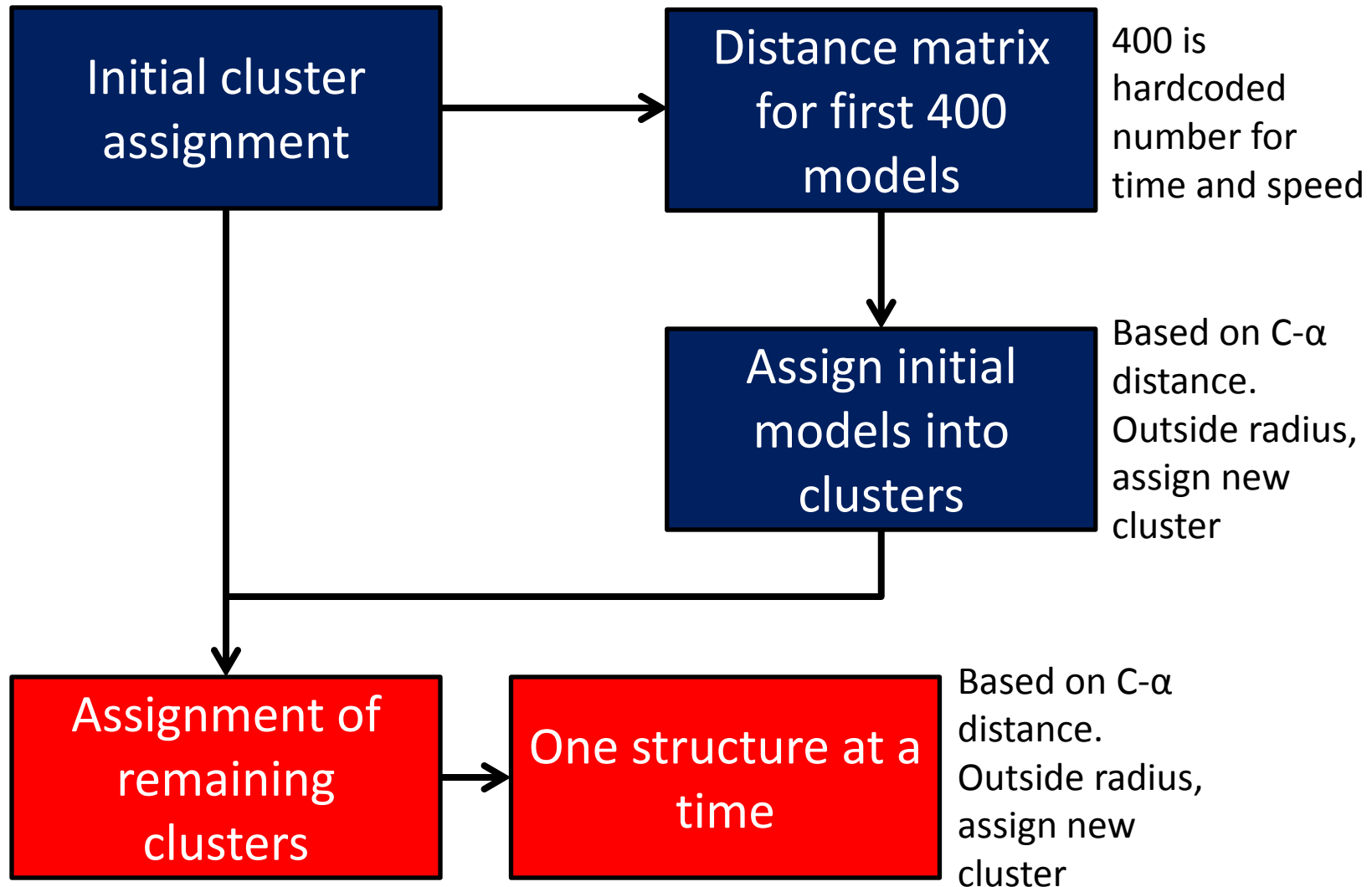
- Rosetta Scoring function is too coarse to sample deep minimum
- Many models have similar folds
- Take largest clusters for analysis



Clustering in Rosetta

- The Rosetta clustering algorithm is slightly unconventional
- Traditional clustering methods require the calculation of a pairwise distance matrix
 - The memory requirements of this method are n^2 where n is the number of models being clustered
 - For large numbers of models, these methods are therefore impractical

Clustering in Rosetta



Clustering Options File

Find this file at /protein_modeling/input_cluster/cluster.options

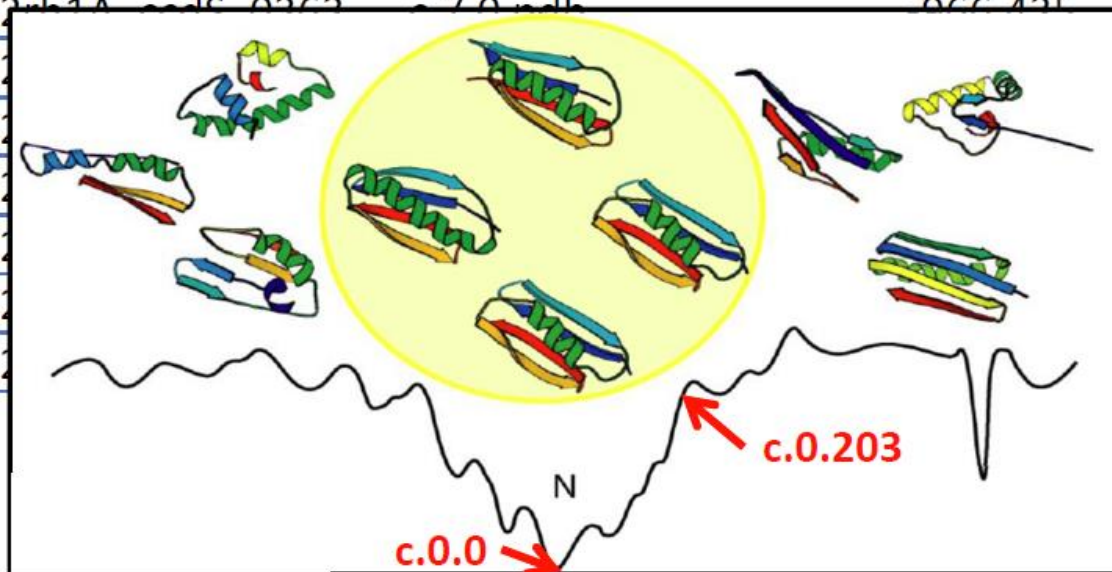
- **-in:file:fullatom** #Read as fullatom input structure
- **-out:file:silent** cluster.out #Output silent structures instead of PDBs
- **-run:shuffle** #Use shuffle mode
- **-cluster:radius -1** #Cluster radius in A for RMS clustering or in inverse GDT_TS for Global Distance Test score clustering.
Use "-1" to trigger automatic radius detection
- **-cluster:input_score_filter 0** #Ignore structures above certain energy

clustering.py: python script runs Rosetta clustering and outputs summary files

Output: cluster_summary.txt; cluster_histogram.txt

Clustering: Results

Tag	file_name	score	size
1u19A_on_2rh1A_ccdS_0241	c.0.0.pdb	-1053.97	403
1u19A_on_2rh1A_ccdS_0310	c.2.0.pdb	-1032.09	35
1u19A_on_2rh1A_ccdS_0394	c.1.0.pdb	-1015.93	24
1u19A_on_2rh1A_ccdS_0028	c.4.0.pdb	-1014.25	12
1u19A_on_2rh1A_ccdS_0286	c.10.0.pdb	-1007.26	10
1u19A_on_2rh1A_ccdS_0308	c.8.0.pdb	-987.543	5
1u19A_on_2rh1A_ccdS_0329	c.3.0.pdb	-976.877	3
1u19A_on_2rh1A_ccdS_0362	c.7.0.pdb	-966.425	2
1u19A_on_2rh1A_ccdS_0363	c.9.0.pdb	-966.425	1
1u19A_on_2rh1A_ccdS_0364	c.6.0.pdb	-966.425	1
1u19A_on_2rh1A_ccdS_0365	c.5.0.pdb	-966.425	1
1u19A_on_2rh1A_ccdS_0366	c.11.0.pdb	-966.425	1
1u19A_on_2rh1A_ccdS_0367	c.12.0.pdb	-966.425	1
1u19A_on_2rh1A_ccdS_0368	c.13.0.pdb	-966.425	1
1u19A_on_2rh1A_ccdS_0369	c.14.0.pdb	-966.425	1
1u19A_on_2rh1A_ccdS_0370	c.15.0.pdb	-966.425	1



Tutorials

Two stand-alone tutorials are included in your worksheet:

- Comparative Modeling with Loop Building
- Clustering

Tips:

- If you want to try making files that already exist (e.g. input files), write them to a new directory

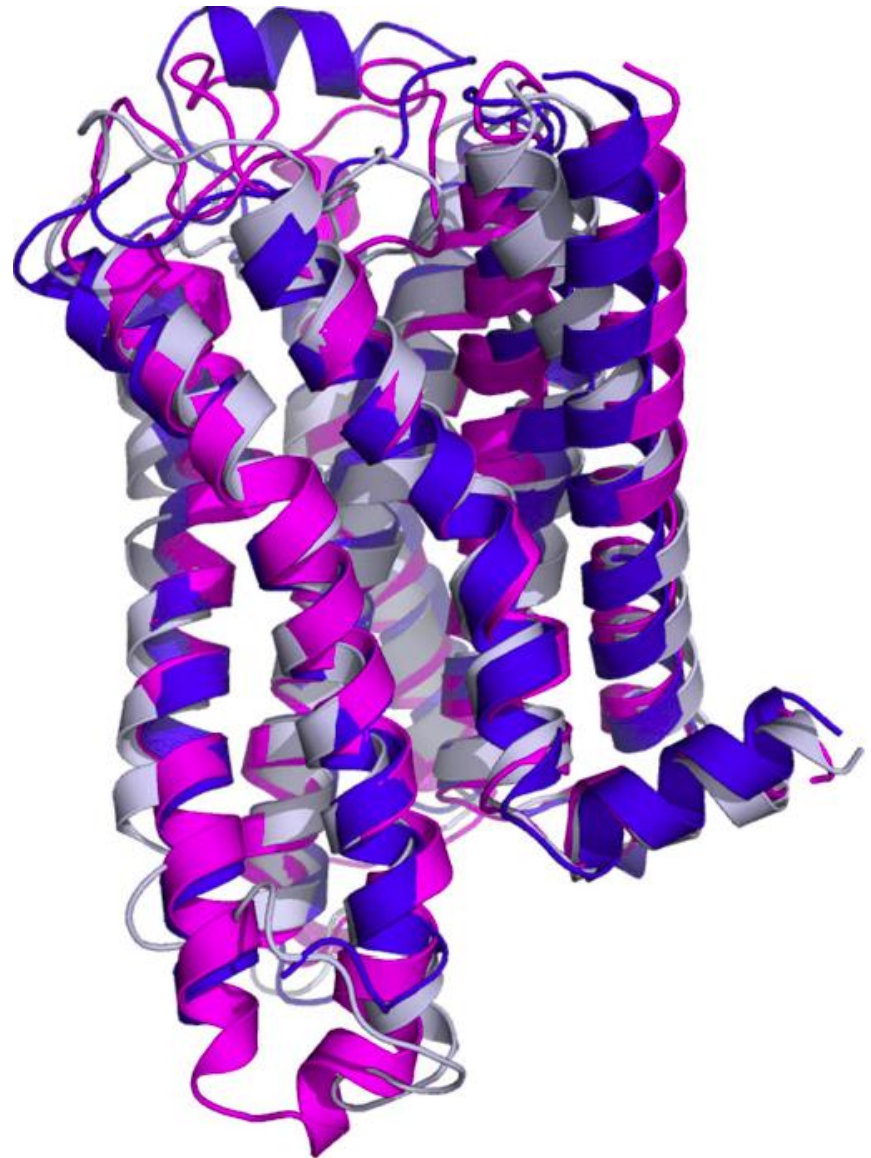
Tutorial Results

1u19A crystal structure

2rh1A template structure

1u19A model (lowest energy of largest
cluster, n=1300)

RMSD to native: 4.1Å



References

- **Rosetta 3.5 User Guide**

https://www.rosettacommons.org/manuals/archive/rosetta3.5_user_guide/index.html

- **Loop Building**

https://www.rosettacommons.org/manuals/archive/rosetta3.5_user_guide/dc/d67/loopmodel_ccd.html

https://www.rosettacommons.org/manuals/archive/rosetta3.5_user_guide/db/dbe/next_generation_k_i_c.html

Wang, C., Bradley, P., & Baker, D. (2007). Protein-Protein Docking with Backbone Flexibility. *Journal of Molecular Biology*, 373(2), 503-519.

Mandell, D. J., Coutsias, E. A., & Kortemme, T. (2009). Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. *Nat Meth*, 6(8), 551-552.

- **Clustering**

https://www.rosettacommons.org/manuals/archive/rosetta3.5_user_guide/db/d3c/cluster.html

- **Modeling Membrane Proteins**

https://www.rosettacommons.org/manuals/archive/rosetta3.5_user_guide/db/d38/membrane_abinitio.html

- **RosettaCM**

Yifan Song, et al. (2013). High-Resolution Comparative Modeling with RosettaCM. *Structure*, 21(10), 1735-1742.

Clustering: Other Options

- **-cluster:gdtmm** #Cluster by gdtmm instead of rms
- **-cluster:input_score_filter <float>** #Ignore structures above certain energy
- **-cluster:exclude_res <int> [<int> <int> ..]** #Exclude residue numbers from structural comparisons
- **-cluster:radius <float>** #Cluster radius
- **-cluster:limit_cluster_size <int>** #Maximal cluster size
- **-cluster:limit_clusters <int>** #Maximal number of clusters
- **-cluster:limit_total_structures <int>** #Maximal number of structures in total
- **-cluster:sort_groups_by_energy** #Sort clusters by energy

RosettaCM Energy Scores

- Incorporates similarity to template structures by generating template-derived restraints
- Restraint weights have been optimized over large training set
- Templates may be ranked with different weights based on alignment confidence