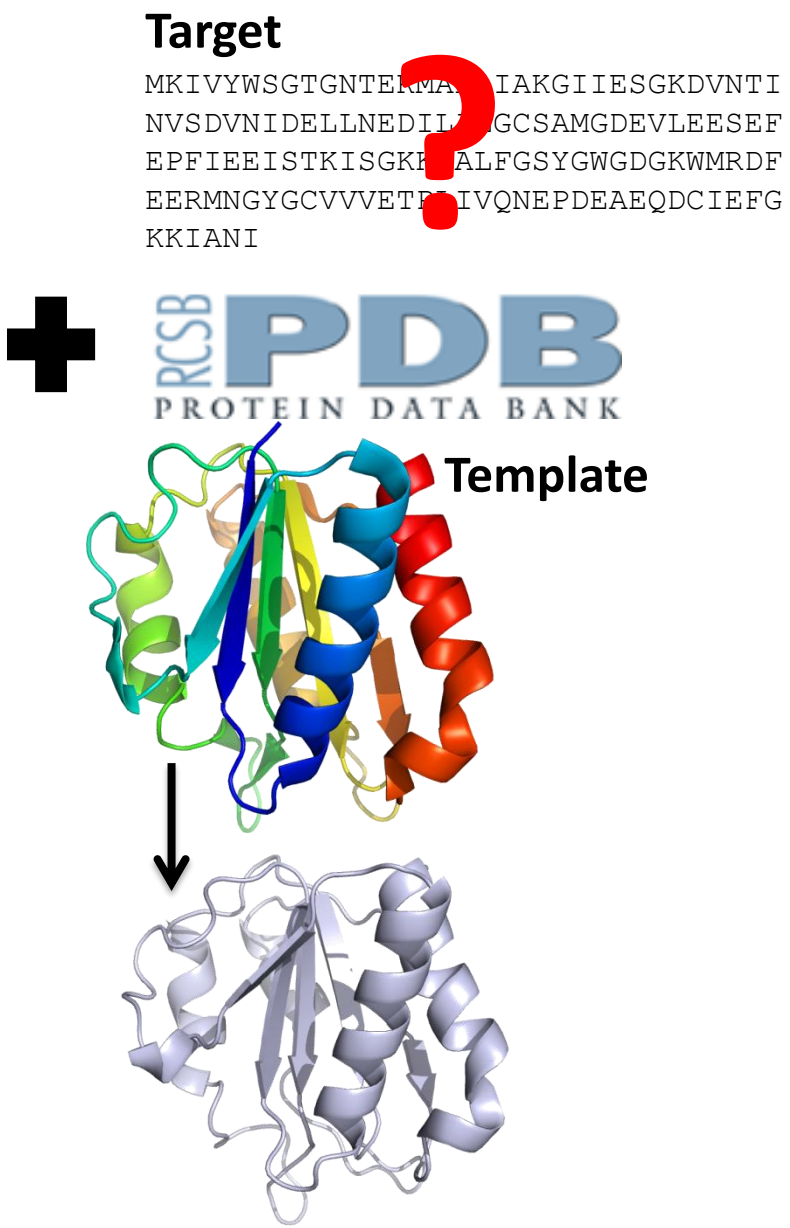
Comparative Modeling and Loop Building in Rosetta 3.2

Outline

- Introduction to comparative modeling
- General comparative modeling protocol
- Comparative modeling in Rosetta 3.2
- Loop building in Rosetta 3.2: CCD vs. KIC
- Clustering output models
- Analyzing results
- Modeling membrane proteins
- Useful references and websites

Introduction to Comparative Modeling



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

Threading: placing amino acids of the target sequence onto the coordinates of the template structure

Application of Models

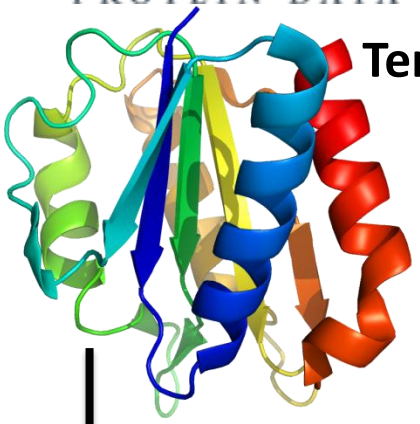
- Predict structure-function relationships
- Predict binding pockets for ligands for structure-based drug design
- Suggest site-directed mutagenesis experiments

Show Threading Video

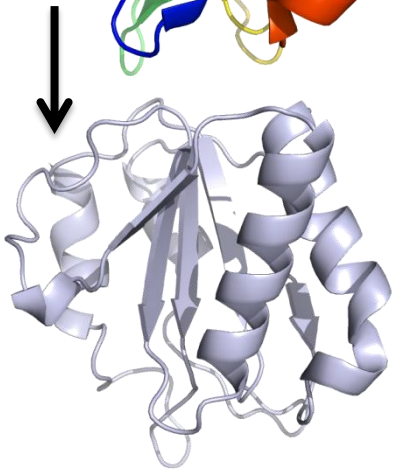
Identifying Template Structures

Target

MKIVYWSGTGNTERMA IAKGIIESGKDVNTI
NVSDVNIDELLNEDIL GCSAMGDEVLEESEF
EPFIEEISTKISGKK ALFGSYGWGDGKWMRDF
EERMNGYGCVVVETR IVQNEPDEAEQDCIEFG
KKIANI



Template



Identifying a Suitable Template:
Templates should ideally have >30% sequence identity to the target. There are two approaches to identifying templates:

1. Sequence Similarity: comparing proteins based on amino acid properties alone (BLAST, PSI-BLAST)

2. Fold Recognition: using predicted secondary structure information to detect proteins with similar 3D characteristics (DALI, PHYRE)

Comparative Modeling Protocol

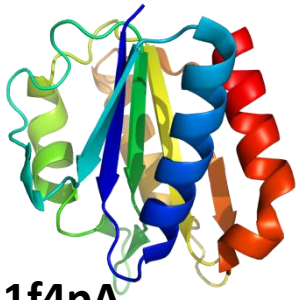
- **Step 1:** Align target sequence with sequence of template structure
- **Step 2:** Thread the target sequence onto the backbone of the template structure
- **Step 3:** Rebuild the loop regions of the model
- **Step 4:** Full-atom refinement of the model
- **Step 5:** Cluster models and analyze your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

```
2foxA 1 --MKIVYWSGTGNTEKMAELIAKGIIE
1f4pA 1 PKALIVYGSTTGNTEYTAETIARELAD
```

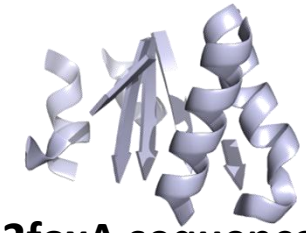
Comparative Modeling Protocol

- **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 of the model
- **Step 5:** Cluster models and analyze your results

```
2foxA 1 --MKIVYWSGTGNTEKMAELIAKGIIE
1f4pA 1 PKALIVYGSTTGNTEYTAETIARELAD
```



1f4pA

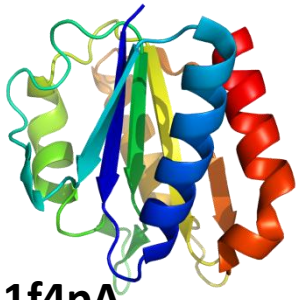


2foxA sequence threaded on backbone coordinates of 1f4pA

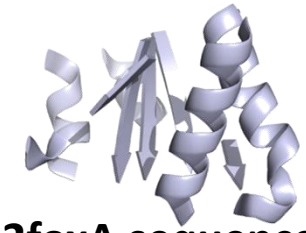
Comparative Modeling Protocol

- **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 of the model
- **Step 5:** Cluster models and analyze your results

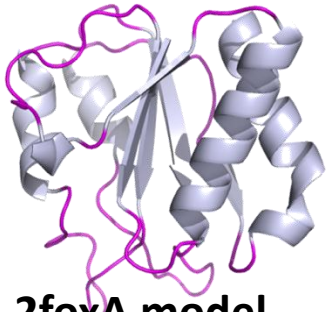
```
2foxA 1 --MKIVYWSGTGNTEKMAELIAKGIIE
1f4pA 1 PKALIVYGSTTGNTEYTAETIARELAD
```



1f4pA



2foxA sequence threaded on backbone coordinates of 1f4pA

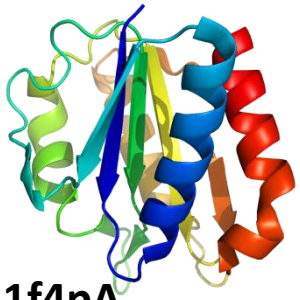


2foxA model with loops

Comparative Modeling Protocol

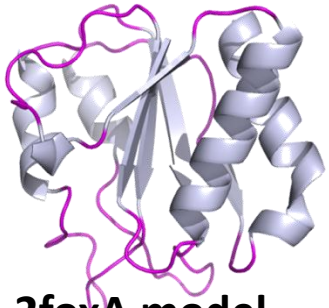
- **Step 1:** Align target sequence with sequence of template structure
- **Step 2:** Thread the target sequence onto the backbone of the template structure
- **Step 3:** Rebuild the loop regions of the model
- **Step 4:** Full-atom refinement of the model
- **Step 5:** Cluster models and analyze your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

```
2foxA 1 --MKIVYWSGTGNTEKMAELIAKGIIE
1f4pA 1 PKALIVYGSTTGNTEYTAETIARELAD
```

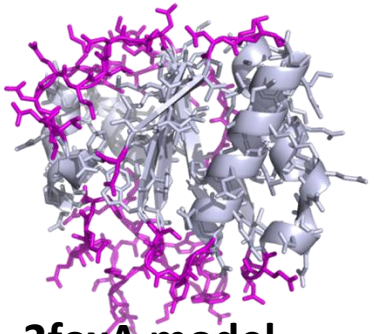


1f4pA

2foxA sequence threaded on backbone coordinates of 1f4pA



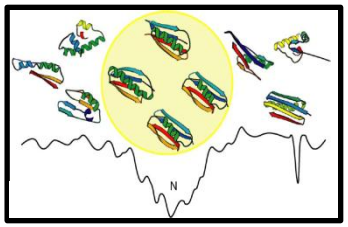
2foxA model with loops



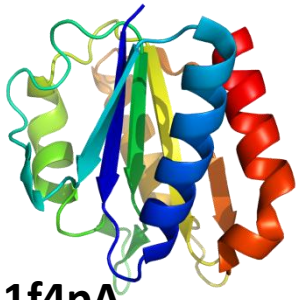
2foxA model with side chains

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 of the model
- **Step 5:** Cluster models and analyze your results

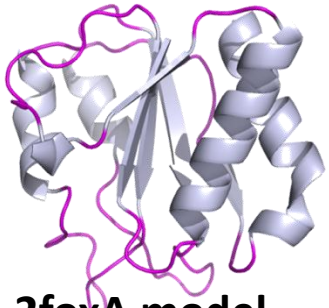


```
2foxA 1 --MKIVYWSGTGNTEKMAELIAKGIIE
1f4pA 1 PKALIVYGSTTGNTEYTAETIARELAD
```

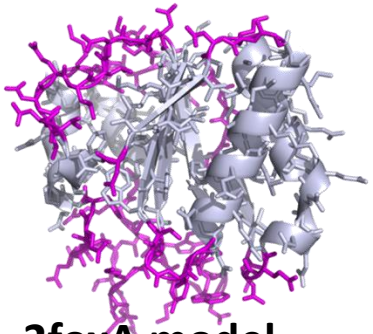


1f4pA

2foxA sequence threaded on backbone coordinates of 1f4pA



2foxA model with loops



2foxA model with side chains

Rosetta 3.2 Threading Protocol

Inputs

1. Target sequence **(2foxA)**
2. Template PDB **(1f4pA)**
3. Fragment files
4. Alignment of target and template sequences
5. Optional: Secondary structure file

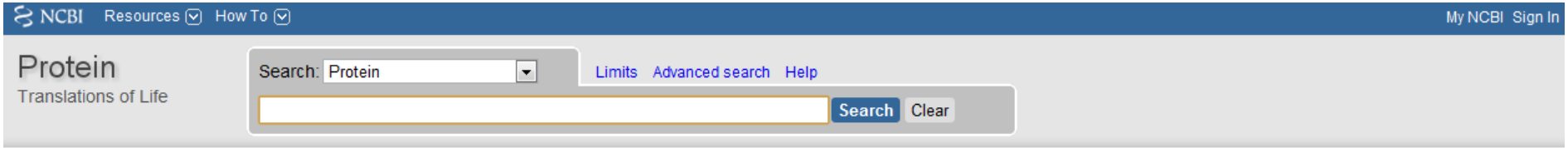
Output

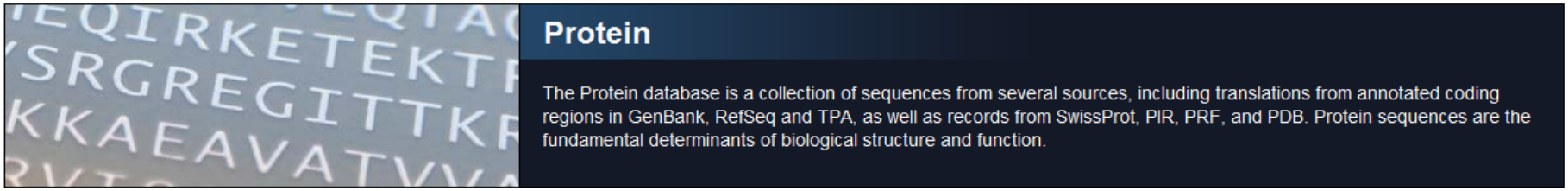
- Silent file containing models and corresponding scores

1: Target sequence

Find this file at $WORKSHOP_ROOT/tutorials/modeling/input_model/2foxA.fasta

```
>2foxA
MKIVYWSGTGNTEKMAELIAKGIIESGKDVNTINVSDVNIDELLNEDILILGCSAMG
DEVLEESEFEPFIEEISTKISGKKVALFGSYGWGDGKWMRDFEERMNGYGCVVVETP
LIVQNEPDEAEQDCIEFGKKIANI
```

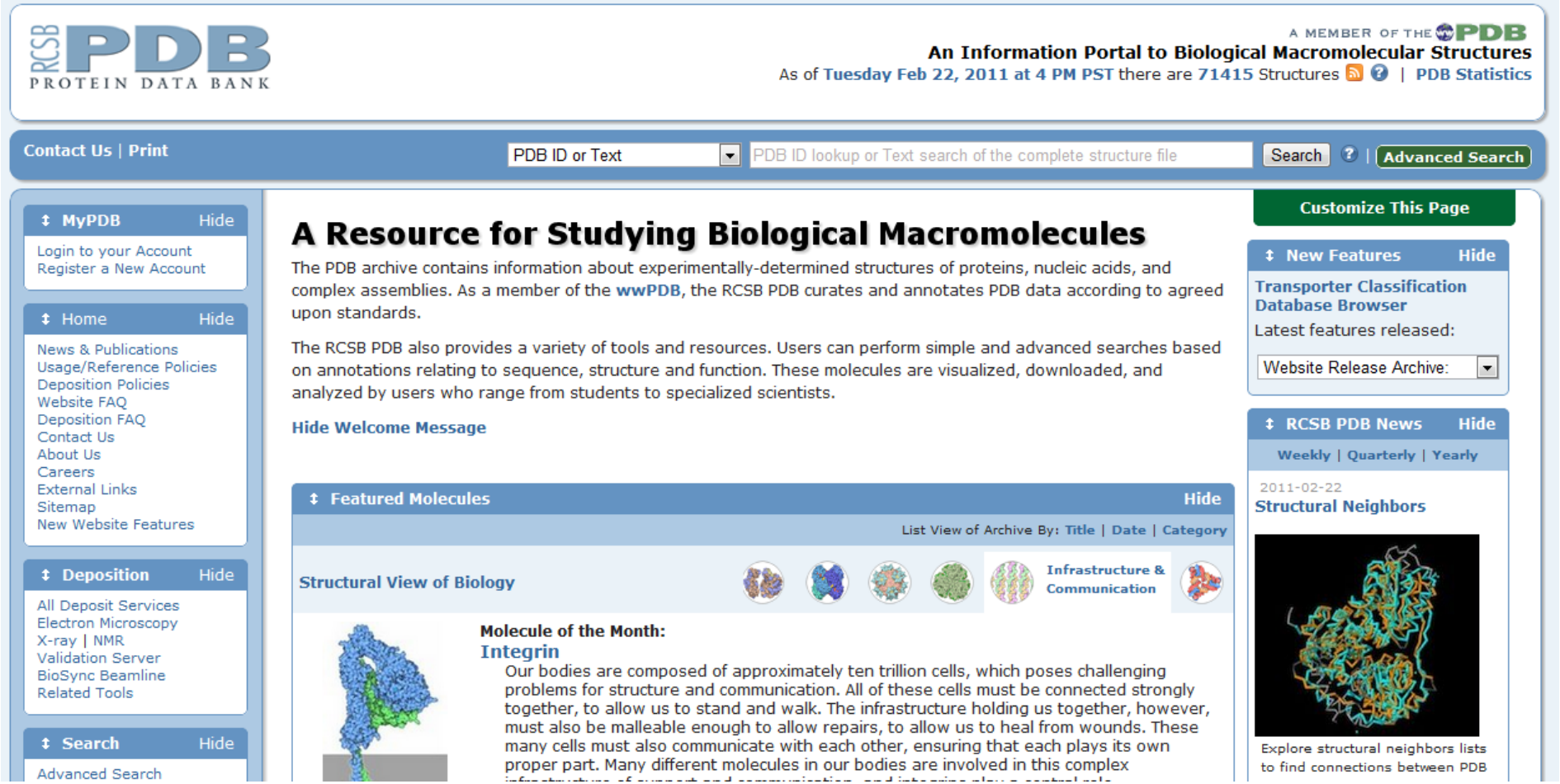




http://www.ncbi.nlm.nih.gov/protein

2: Template PDB

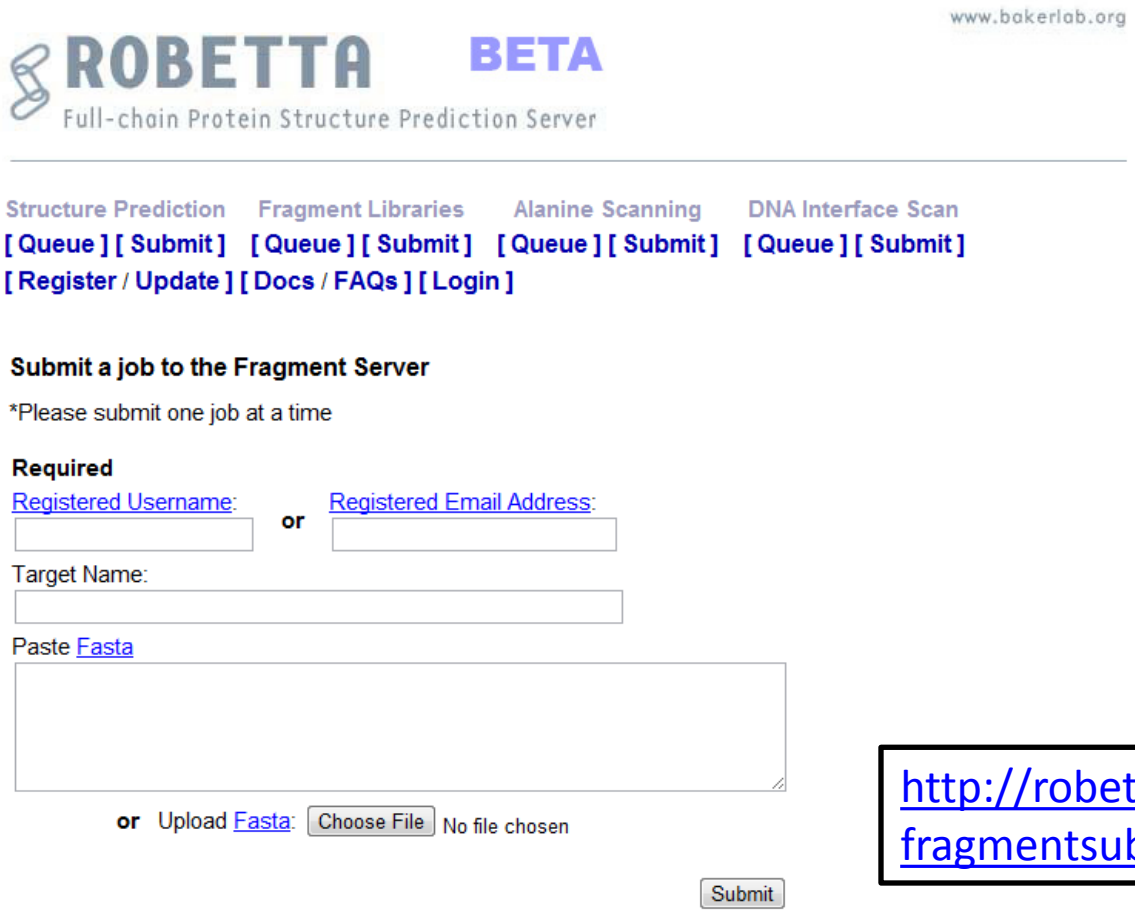
Find this file at $WORKSHOP_ROOT/tutorials/modeling/input_model/1f4pA.pdb



http://www.rcsb.org

3: Fragments

Find these files at
$WORKSHOP_ROOT/tutorials/modeling/input_model/aa2foxA03_05.200_v1_3 and
$WORKSHOP_ROOT/tutorials/modeling/input_model/aa2foxA09_05.200_v1_3

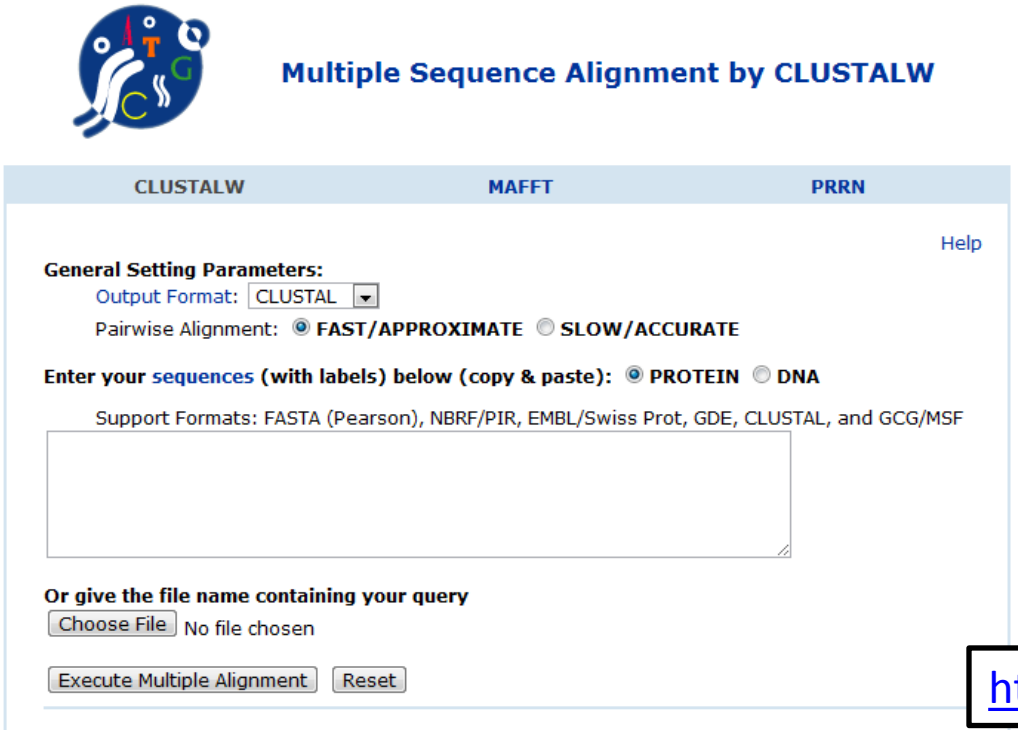


http://robetta.bakerlab.org/fragmentsubmit.jsp

4: Alignment

Find this file at $WORKSHOP_ROOT/tutorials/modeling/input_model/2foxA.1f4pA.aln

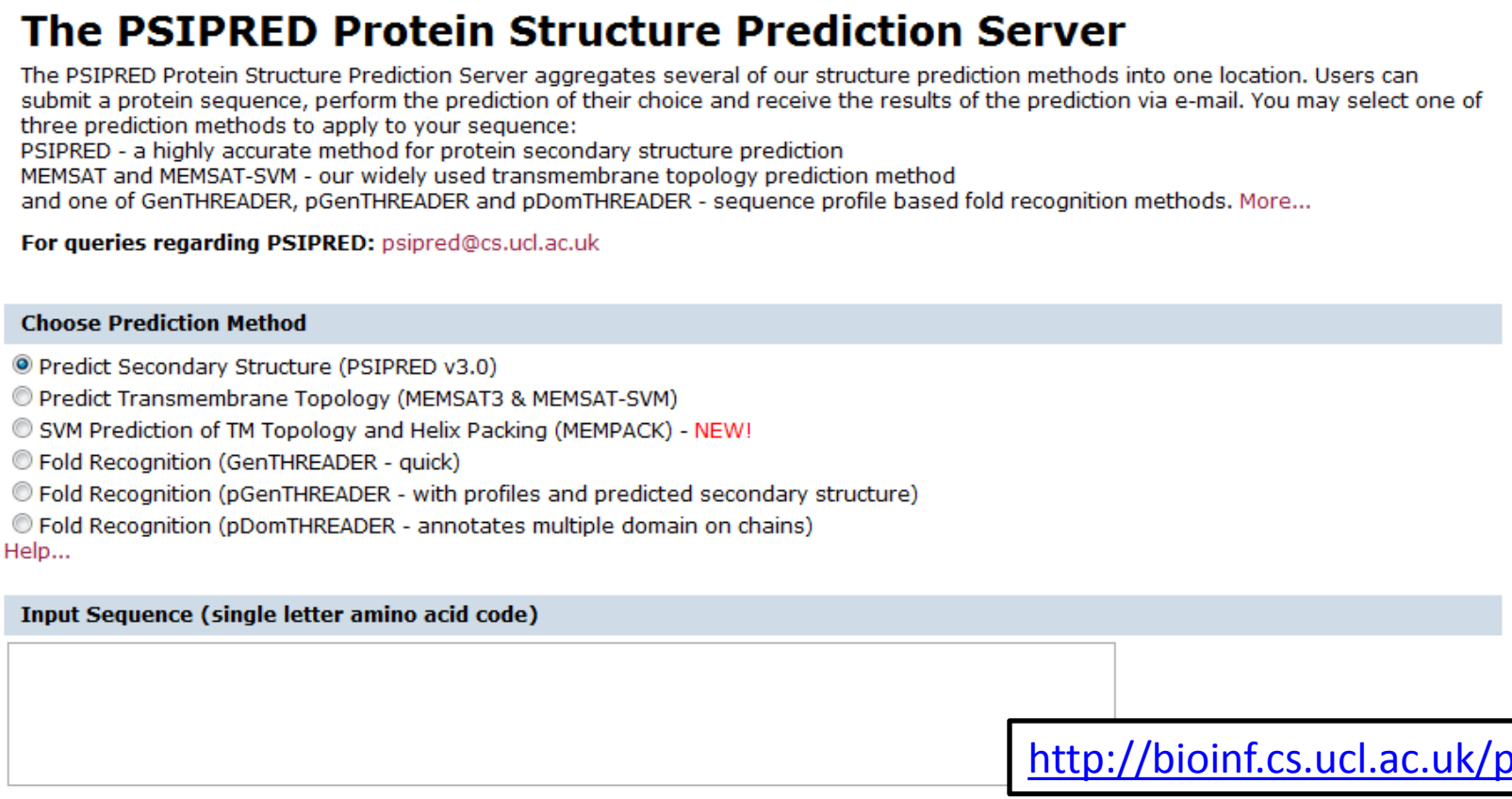
```
2foxA 1 --MKIVYWSGTGNTEKMAELIAKGIIE
1f4pA 1 PKALIVYGSTTGNTEYTAETIARELAD
```



http://align.genome.jp/

5: PSIPRED

Find this file at $WORKSHOP_ROOT/tutorials/modeling/input_model/2foxA.psipred_ss2



http://bioinf.cs.ucl.ac.uk/psipred/

Options: Input

Find this file at
$WORKSHOP_ROOT/tutorials/modeling/input_model/comparative_model.options

```
-run:protocol threading #call threading protocol


-in:file:fasta *.fasta #target sequence
-in:file:template_pdb *.pdb #template structure
-in:file:fullatom #input will be fullatom
-in:file:psipred_ss2 *.psipred_ss2 #optional: psipred secondary structure
-in:file:alignment *.aln #input alignment
```

Options: Loop Building (CCD)

-loops:frag_sizes 9 3 1 #sizes of fragments

-loops:frag_files <fragment files> #fragment files

-loops:remodel quick_ccd #use ccd to remodel loops

-idealize_after_loop_close #idealize structure after closing loops (structure will have ideal rosetta bond lengths and angles)

-loops:extended true #force extended on loops (phi-psi angles set to 180 degrees) independent of loop input file. for rebuilding loops entirely.

-loops:build_initial true #precede loop modeling with initial round of removing missing densities and building simple loops

-loops:relax fastrelax #fastrelax loops (5 cycles at default)

-cm:min_loop_size <int> #minimum size of loops to consider for rebuild

Options: Output

-out:nstruct <int> #number of models to build

-out:file:silent_struct_type binary #output file type

-out:file:silent *.out #output file name

-out:file:fullatom #output file will be fullatom

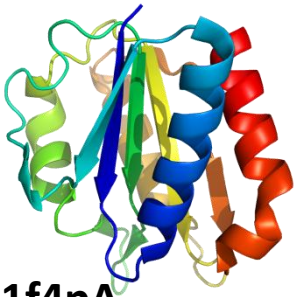
Running Rosetta

```
$ROSETTA_BIN/minirosetta.$ROSETTA_SUFFIX
@$WORKSHOP_ROOT/tutorials/modeling/input_model/comparative
_model.options -database $ROSETTA_DATABASE >&
$WORKSHOP_ROOT/tutorials/modeling/comparative_model.log &
```

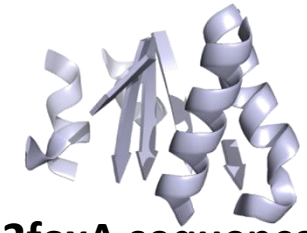
Behind the Rosetta 3.2 Threading Protocol

- **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 of the model
- **Step 5:** Cluster models and analyze your results

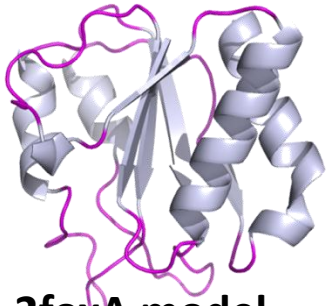
```
2foxA 1 --MKIVYWSGTGNTEKMAELIAKGIIE
1f4pA 1 PKALIVYGSTTGNTEYTAETIARELAD
```



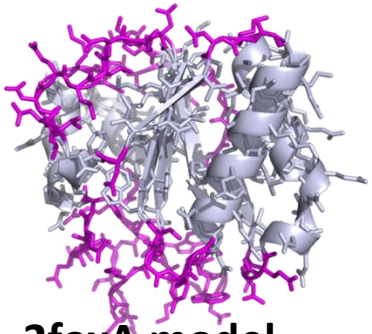
1f4pA



2foxA sequence threaded on backbone coordinates of 1f4pA

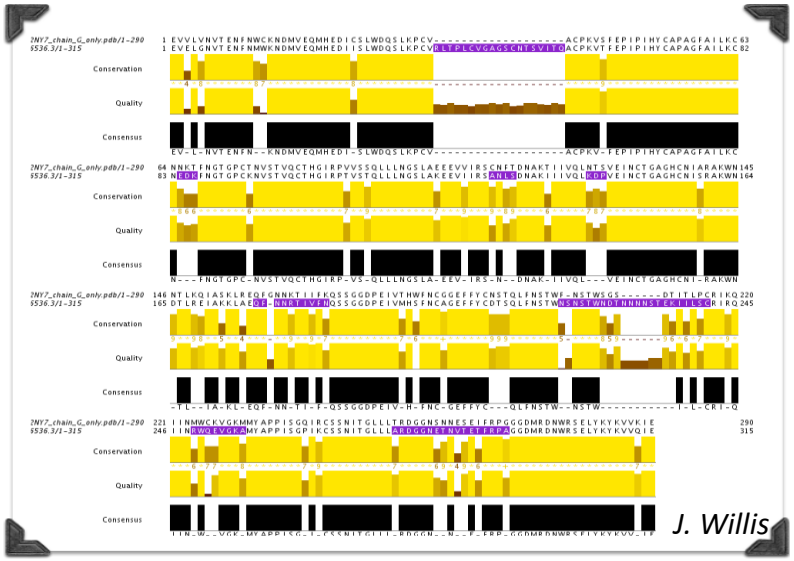


2foxA model with loops



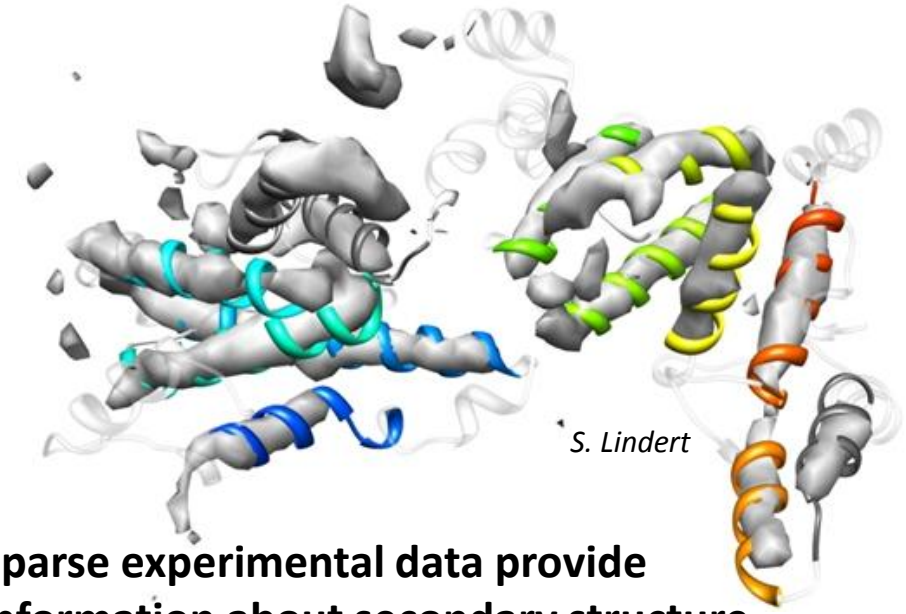
2foxA model with side chains

Why Loop Building is Important



J. Willis

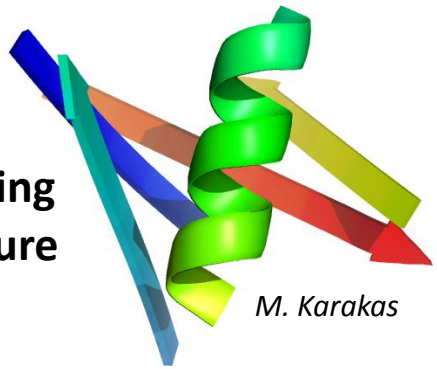
Sequence alignments in comparative modeling result in insertions & deletions



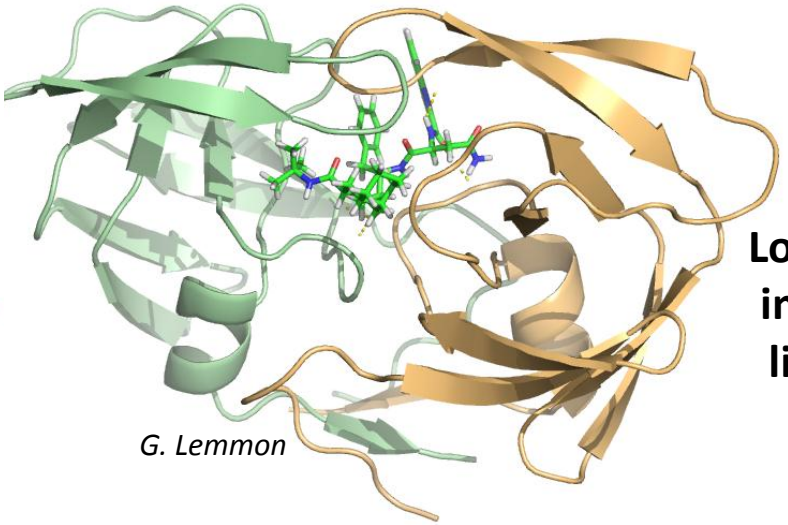
S. Lindert

Sparse experimental data provide information about secondary structure elements, but rarely loop conformations

De novo folding requires connecting secondary structure elements



M. Karakas



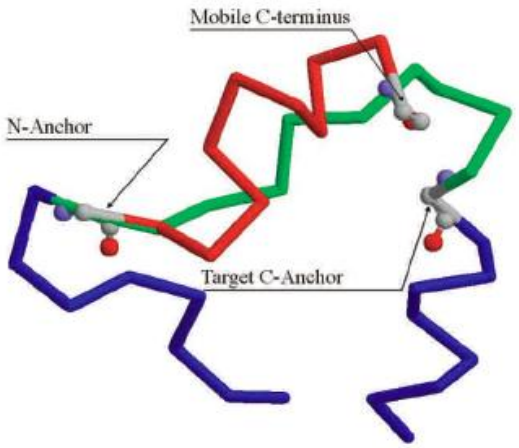
G. Lemmon

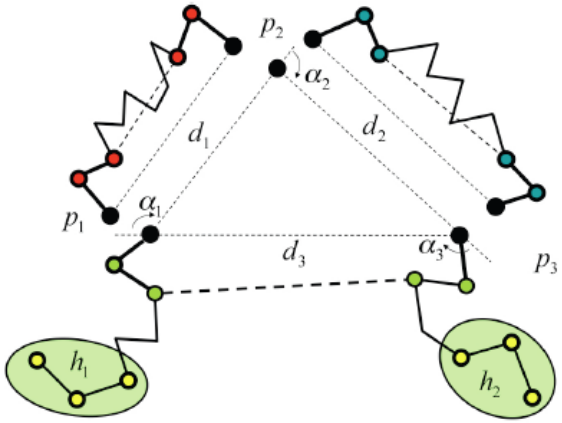
Loops can play an important role in ligand & peptide binding sites

Comparing Loop Building Methods

	Cyclic Coordinate Descent (CCD)	Kinematic Loop Closure (KIC)
Speed	★	
Accuracy in recovering crystal structures		★
Accuracy in homology modeling	★	

Stage 1: Remodel (CCD or KIC)
fast, broad sampling of backbone conformations, centroid
Stage 2: Refine (KIC)
side-chains are represented in all-atom detail, and together with backbone conformations, evaluated by Rosetta's high-resolution scoring function.





Rosetta 3.2 Loop Building Protocol

Inputs

1. Loop File
2. Model PDB with or without loop coordinates

CCD Only

3. Fragment files

Output

- Silent file containing models and corresponding scores

1: Loop File

Find this file at $WORKSHOP_ROOT/tutorials/modeling/input_loop/2foxA.loops

```
LOOP 6 11 0 0 0
```

Column 1	LOOP	The loop file identity tag
Column 2	<integer>	Residue number for starting loop anchor. NOTE: The starting structure must have real coordinates for all residues outside the loop definition, including the loop anchors (residues indicated in the loops file).
Column 3	<integer>	Residue number for ending loop anchor
Column 4	<integer>	Cut point residue number, >=startRes, <=endRes. default - let LoopRebuild choose cutpoint
Column 5	<float>	Skip rate (probability between 0 and 1). default - never skip
Column 6	<boolean>	Extend loop. Default false

Differences for KIC:

- For de novo reconstruction of protein loops, set 'extend loop' field in the loop definition file (the last column) to '1'.

2: PDB without loop coordinates

Find this file at
$WORKSHOP_ROOT/tutorials/modeling/input_loop/2foxA_no_loops.pdb

If you want to completely rebuild your loops, set the loop coordinates of your PDB file to zero with the following script:

```
$WORKSHOP_ROOT/py_protein_utils/scripts/remove_loop_coords.py
2foxA.loops 2foxA_start_model.pdb 2foxA_no_loops.pdb
```

The script sets loop x-y-z coordinates (as defined by your loop file) to 0.000 and the occupancy column to -1.00, as below:

```
ATOM   2253  CB  PHE A 231       0.000   0.000   0.000 -1.00  0.00           C
ATOM   2254  CG  PHE A 231       0.000   0.000   0.000 -1.00  0.00           C
ATOM   2255  CD1 PHE A 231       0.000   0.000   0.000 -1.00  0.00           C
ATOM   2256  CD2 PHE A 231       0.000   0.000   0.000 -1.00  0.00           C
ATOM   2257  CE1 PHE A 231       0.000   0.000   0.000 -1.00  0.00           C
```

Setting up CCD versus KIC

Find these files at $WORKSHOP_ROOT/tutorials/modeling/input_loop/kic.options and $WORKSHOP_ROOT/tutorials/modeling/input_loop/ccd.options

Common Options

```
-nstruct <int> #number of models to build. 1000 recommended for production runs.
-loops:input_pdb *.pdb #starting pdb with loops to rebuild
-loops:loop_file *.loops #loop file
-loops:relax fastrelax #does a minimization of the structure in the torsion space
-loops:extended #force phi-psi angles to be set to 180 degrees independent of loop input file (recommended for production runs)
-out:file:silent_struct_type binary #output file type
-out:file:silent *.out #output file name
-out:file:fullatom #output file will be fullatom
```

Options for CCD

```
-loops:frag_sizes 9 3 1
-loops:frag_files <fragment files>
-loops:remodel quick_ccd
-loops:refine refine_kic
```

Options for KIC

```
-loops:remodel perturb_kic
-loops:refine refine_kic
-ex1  #Include extra chi1 rotamers
-ex2
```

Running Rosetta

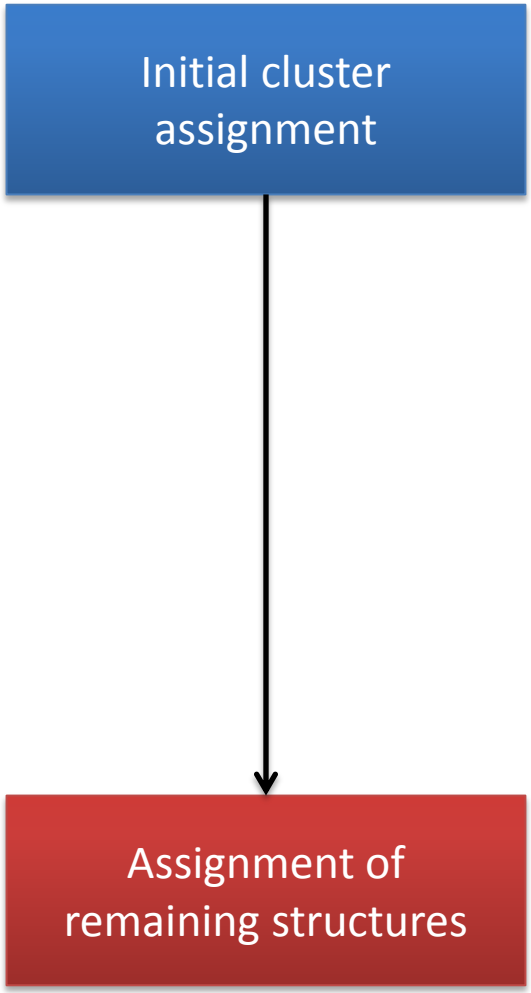
```
$ROSETTA_BIN/loopmodel.$ROSETTA_SUFFIX
@$WORKSHOP_ROOT/tutorials/modeling/input_loop/ccd.options
-database $ROSETTA_DATABASE >& $WORKSHOP_ROOT/
tutorials/modeling/ccd.log &

$ROSETTA_BIN/loopmodel.$ROSETTA_SUFFIX
@$WORKSHOP_ROOT/tutorials/modeling/input_loop/kic.options
-database $ROSETTA_DATABASE >& $WORKSHOP_ROOT/
tutorials/modeling/kic.log &
```

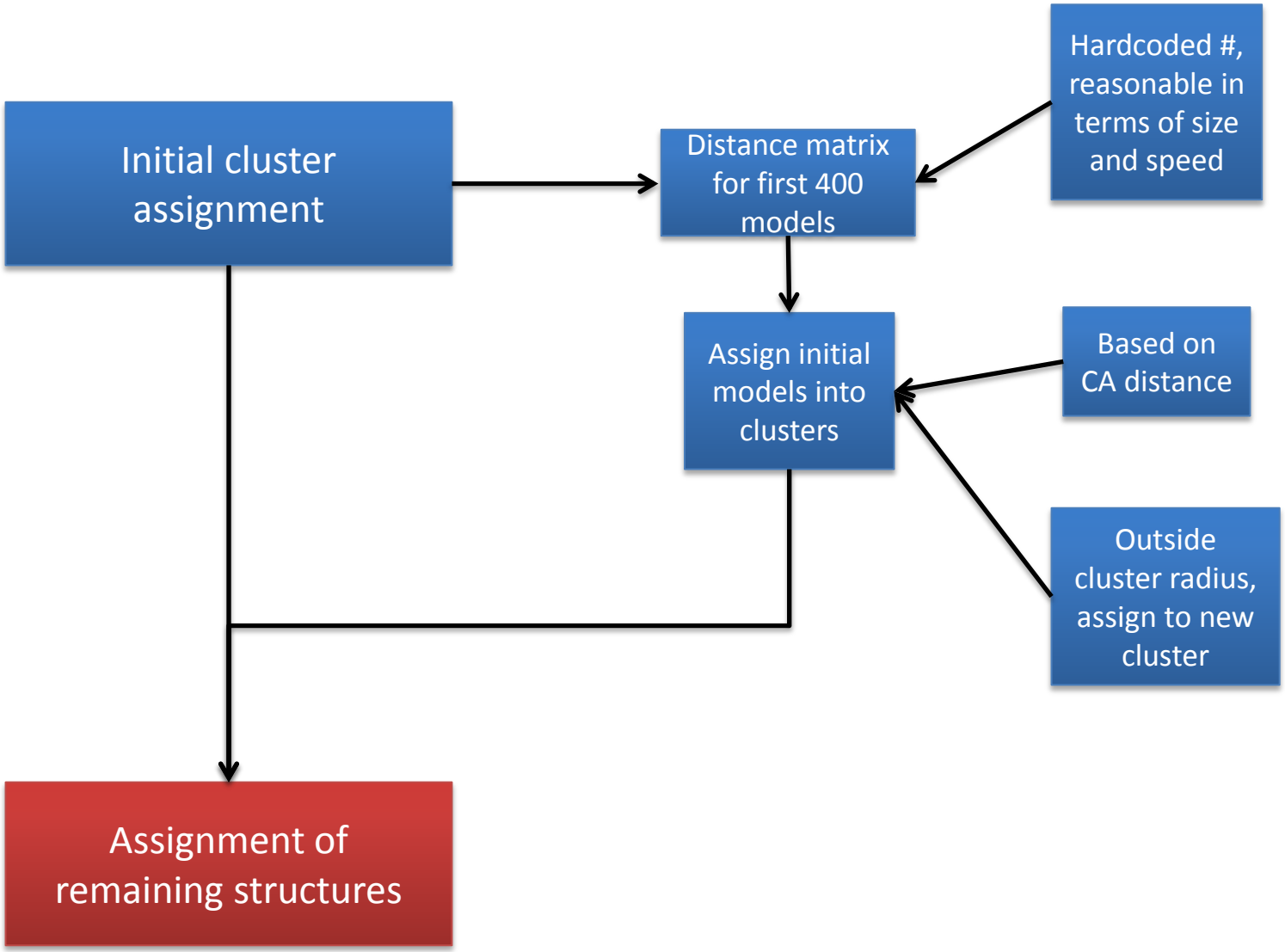
(Clustering in Rosetta)

- 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

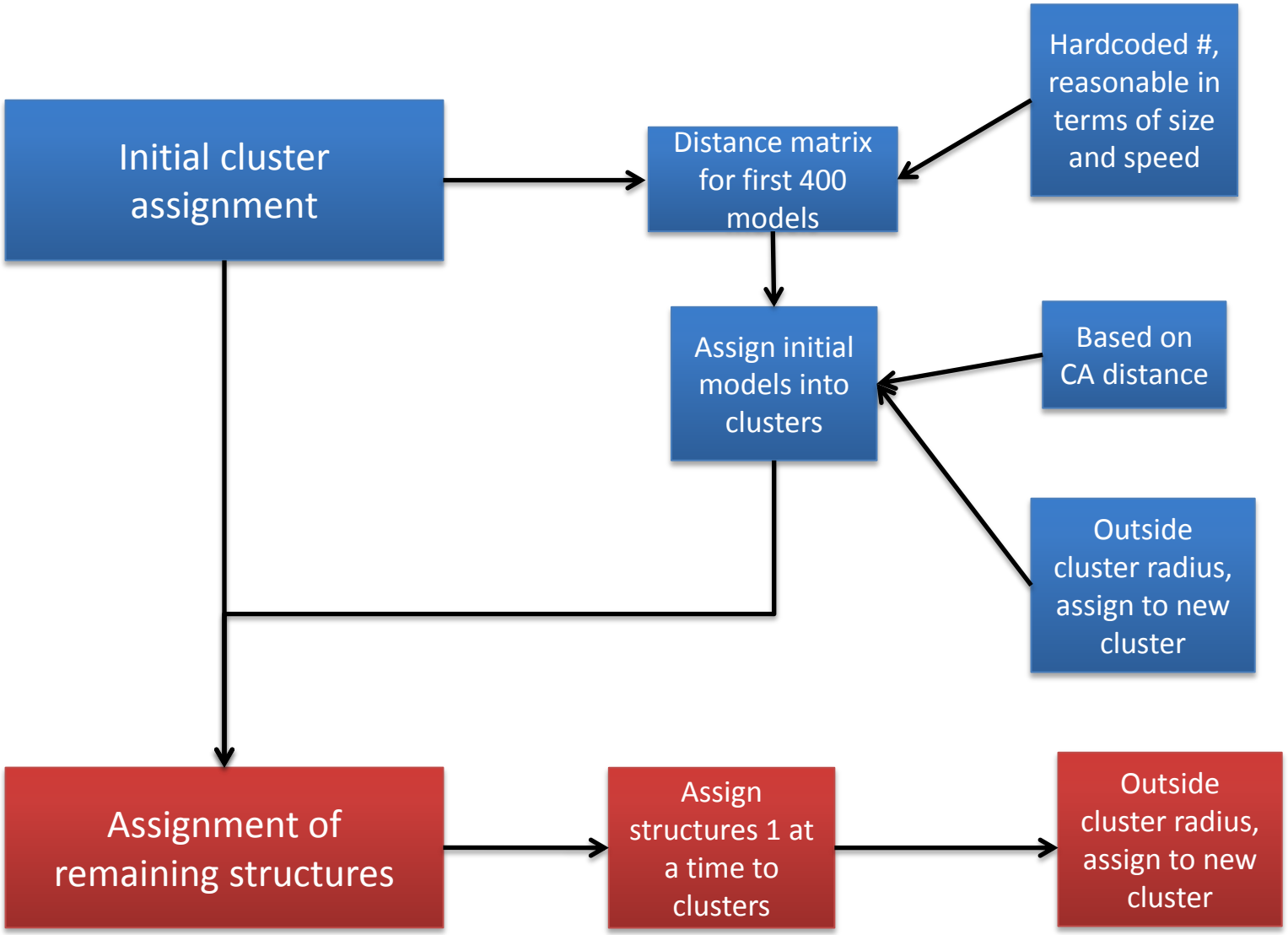
(Clustering In Rosetta)



(Clustering In Rosetta)



(Clustering In Rosetta)



Clustering : Options

Find this file at $WORKSHOP_ROOT/tutorials/modeling/input_cluster/cluster.options

-in:file:fullatom #Read as fullatom input structure

-out:file:silent #Output silent structures instead of PDBs

-run:shuffle #Use shuffle mode

-cluster:radius <float> #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:exclude_res <int> [<int> <int> ..] #Exclude residue numbers from structural comparisons

Clustering : Running

Before running the cluster application, combine all your silent files:

```
$ROSETTA_BIN/combine_silent.$ROSETTA_SUFFIX
  -database $ROSETTA_DATABASE
  -in:file:silent *.out
  -in:file:silent_struct_type binary
  -out:file:silent cluster_all.out
  -out:file:silent_struct_type binary
```

The clustering python script runs the Rosetta application and outputs summary files:

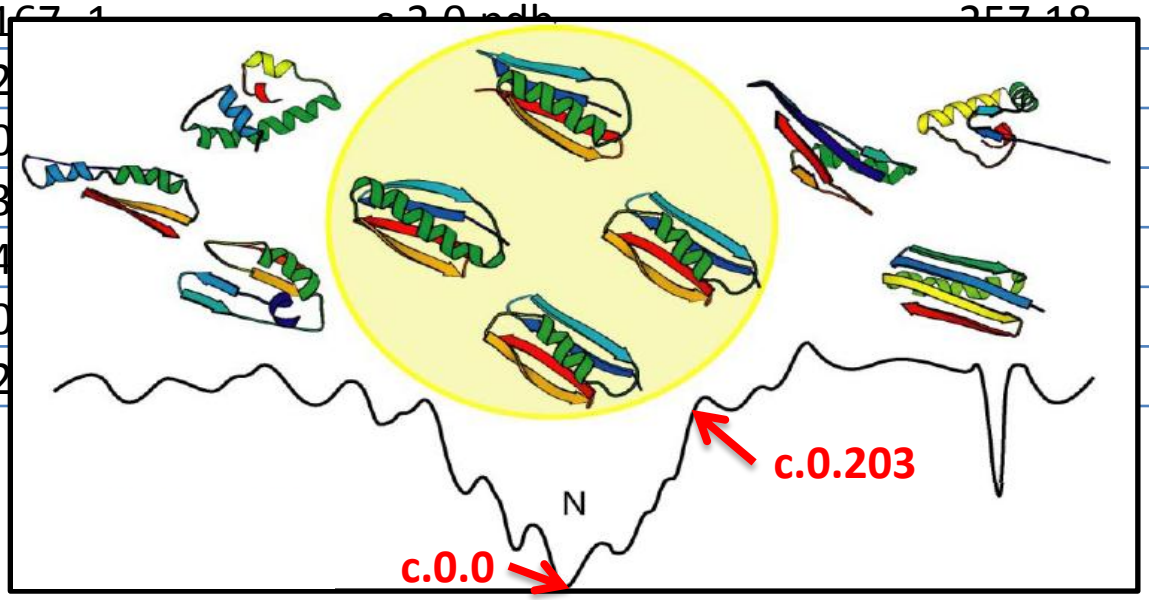
```
python $WORKSHOP_ROOT/py_protein_utils/scripts/clustering.py
--silent=cluster_all.out
--rosetta=$ROSETTA_BIN/cluster.$ROSETTA_SUFFIX
--database=$ROSETTA_DATABASE
--options=cluster.options
cluster_summary.txt cluster_histogram.txt
```

Clustering : Results

cluster number (random)

model number, sorted by energy

Tag	file_name	score	size
S_1F4PA_0410_1	**c.0.0.pdb**	**-267.131**	**203**
S_1F4PA_0356_1	**c.6.0.pdb**	**-252.855**	**40**
S_1F4PA_0036	**c.22.0.pdb**	**-248.465**	**29**
S_1F4PA_0127_1	**c.13.0.pdb**	**-251.634**	**24**
S_1F4PA_0116_1	**c.14.0.pdb**	**-251.295**	**24**
S_1F4PA_0281	**c.25.0.pdb**	**-248.026**	**24**
S_1F4PA_0162	c.29.0.pdb	-245.988	20
S_1F4PA_01			17
S_1F4PA_02			17
S_1F4PA_00			16
S_1F4PA_03			14
S_1F4PA_04			13
S_1F4PA_00			13
S_1F4PA_02			13

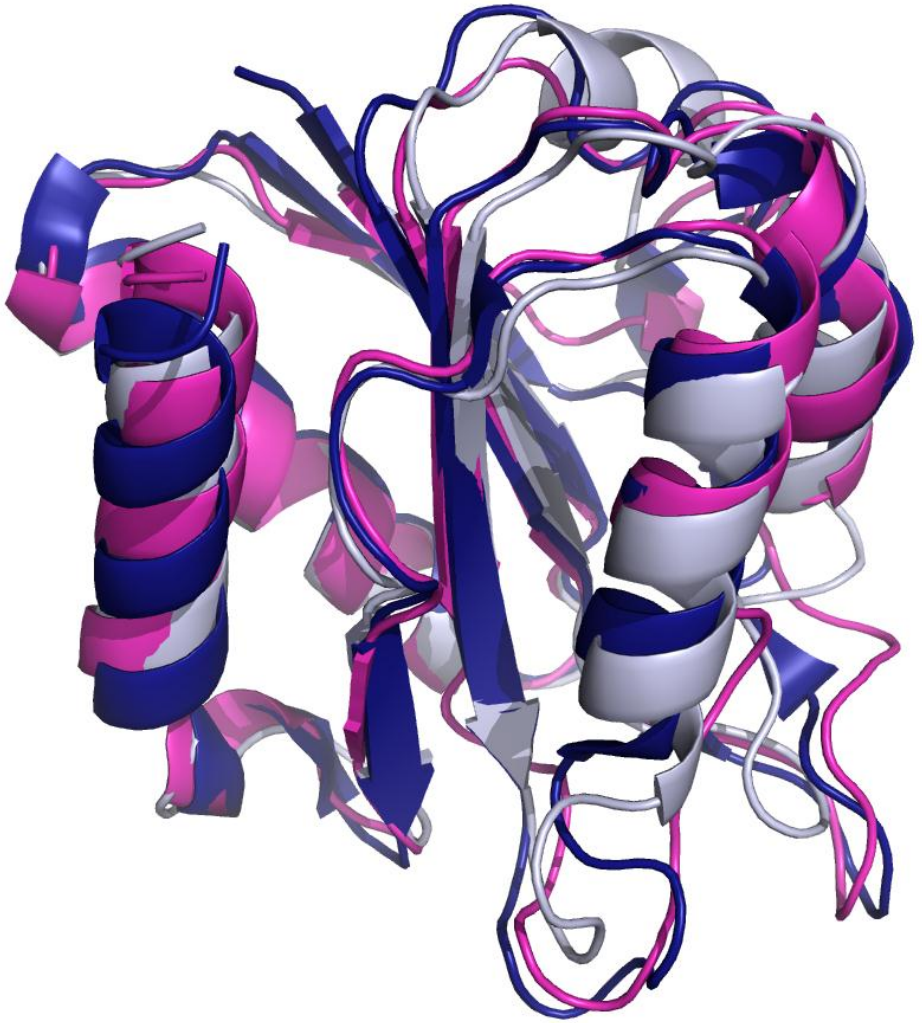


Analyze Your Results: See Workshop 1

- Generating Score vs. RMSD Plots
- Extracting and Looking at Models
- Looking at Models in PyMol
- Looking at RMSD Distributions

Results of Comparative Modeling Protocol

2foxA native structure
1f4pA template structure
2foxA model



RMSD of 2foxA model (c.0.0) to 2foxA native structure:
1.376 Angstroms

RMSD of 2foxA model (c.0.0) to 1f4pA template:
1.312 Angstroms

Comparative Modeling of Membrane Proteins

**this protocol is not yet benchmarked, use with caution*

- Generate LIPS and spanfile (see workshop 1)
- Add the following options to the options file:

```
-in:file:spanfile *.span # newly created spanfile
-in:file:lipofile *.lips4 # newly created lipo file
-membrane:no_interpolate_Mpair # membrane scoring specification
-membrane:Menv_penalties # turn on membrane penalty scores
-score:weights membrane_highres_Menv_smooth.wts
```

- Run protocol as before

It's Your Turn!

Three stand-alone tutorials are included in your worksheet:

- Comparative Modeling
- Loop Building (CCD and KIC)
- Clustering

Tips:

- BLUE text means that these files and/or this information is provided.
- RED text means that this material will NOT be conducted during the workshop
- Change all environment variables (anything starting with a $) to your local paths
- If you want to try making files that already exist (e.g., input files), write them to a new directory!
- See Tutorial 1 (De Novo Folding) for more details on analyzing your results.

References

- **Rosetta 3.2 User Guide**

http://www.rosettacommons.org/manuals/archive/rosetta3.2_user_guide/comparative_modeling.html

- **Comparative Modeling**

http://www.rosettacommons.org/manuals/archive/rosetta3.2_user_guide/comparative_modeling.html

Raman, S., Vernon, R., Thompson, J., Tyka, M., Sadreyev, R., Pei, J., Kim, D., et al. (2009). Structure prediction for CASP8 with all-atom refinement using Rosetta. *Proteins*, *77 Suppl 9*, 89-99.

- **Loop Building**

http://www.rosettacommons.org/manuals/archive/rosetta3.2_user_guide/ccd_loop_modeling.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**

http://www.rosettacommons.org/manuals/archive/rosetta3.2_user_guide/cluster_commands.html

- **Modeling Membrane Proteins**

http://www.rosettacommons.org/manuals/archive/rosetta3.2_user_guide/membrane_abinitio.html