De novo Folding in Rosetta 3.2

Rosetta Workshop

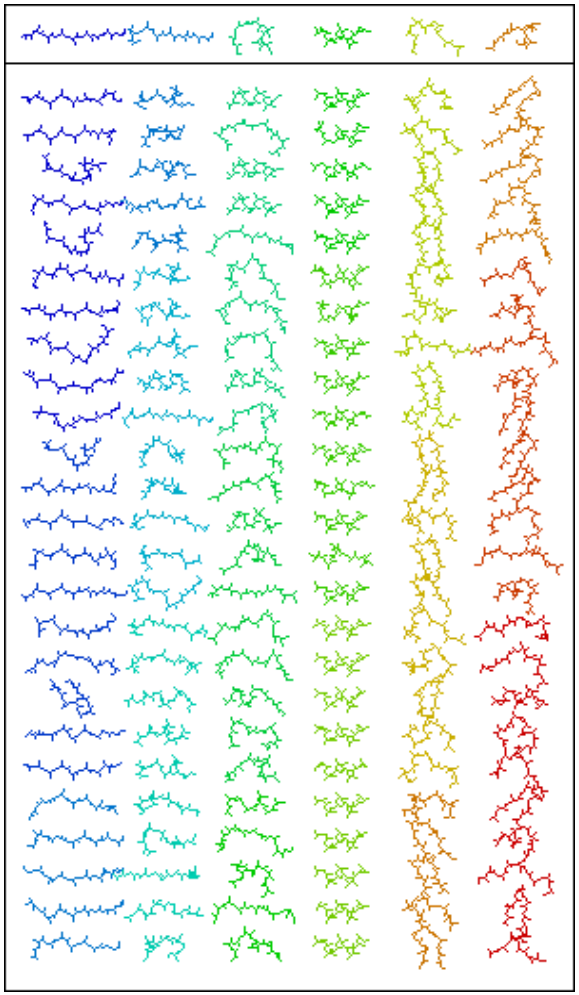
March 12, 2011

Vanderbilt University

Tutorial Outline

- What is *de novo* folding?
- AbinitioRelax Protocol
  - Necessary input files
  - Running AbinitioRelax
  - Extracting data and analyzing output
- Folding membrane proteins with MembraneAbinitio
  - How is it different from AbinitioRelax?
- Folding with restraints
- Useful references and websites

What is *De novo* Folding?

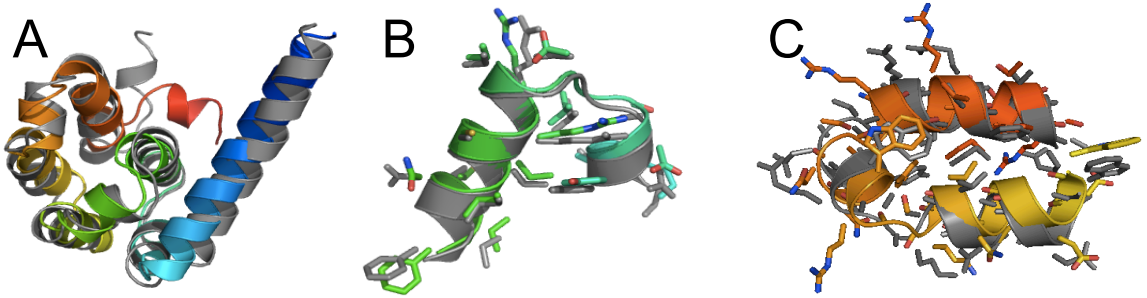


- We are folding *from the primary sequence* using secondary structure prediction and peptide fragments from the PDB
- Use the fragments to change the geometry of the protein and score to keep good fragment insertions

Figure courtesy of Jens Meiler

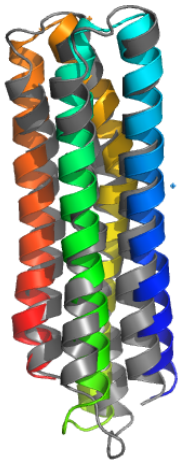
What can Rosetta3.2 Actually Fold?

T4-lysozyme C-terminal domain

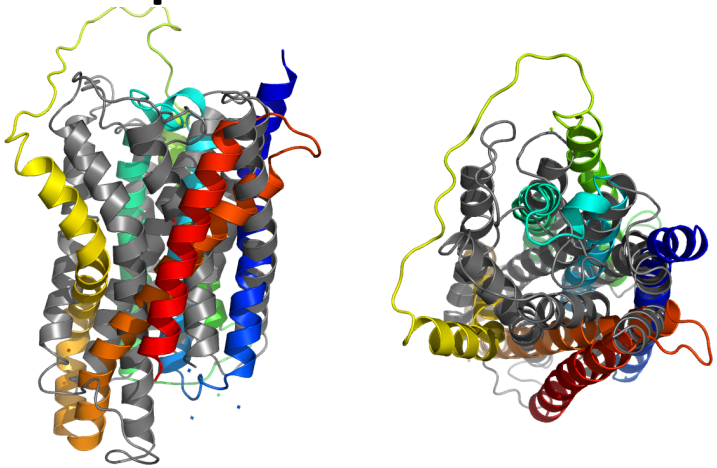


small, globular, soluble proteins

V-type Na^+ ATP synthase subunit



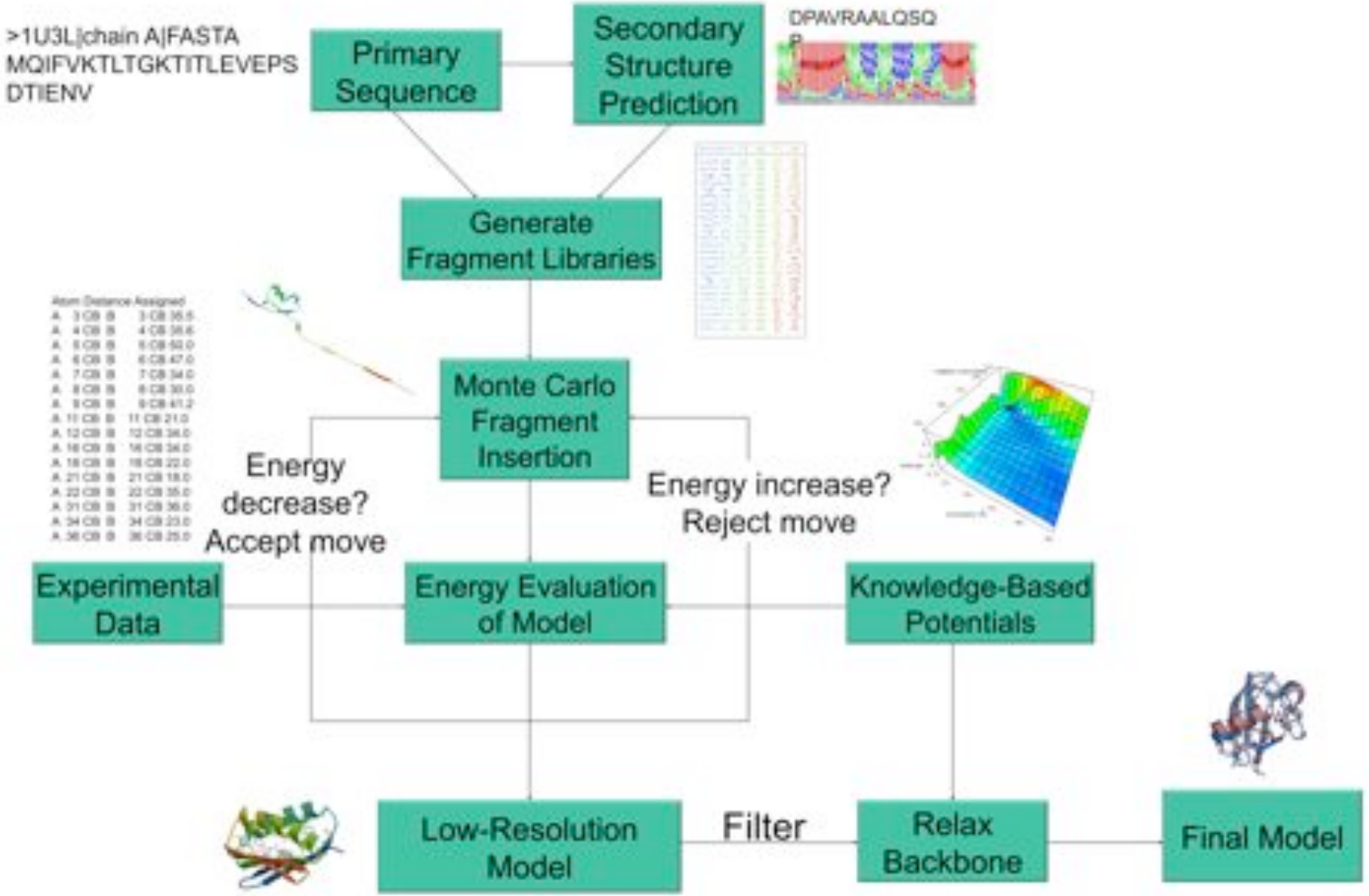
rhodopsin



small, simple membrane proteins

…but not large, complex proteins

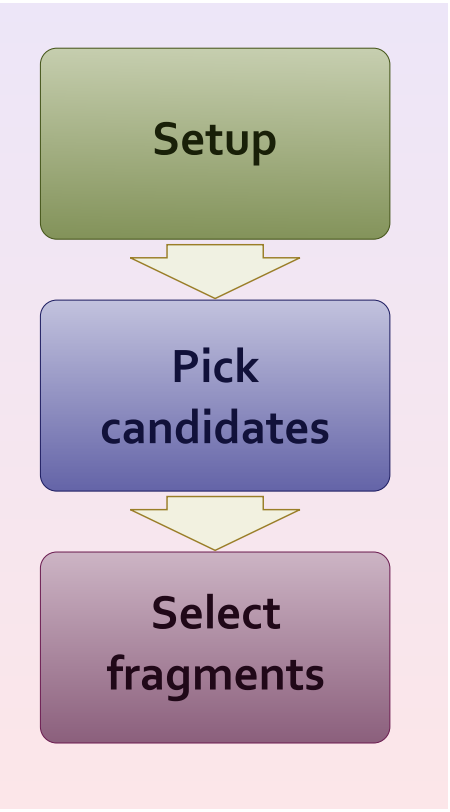
Rosetta *De novo* Folding Protocol



Necessary Input Files for AbinitioRelax

- FASTA file of your protein sequence
- “Clean” PDB file of native structure (optional)
- Fragment library files
- Options file

What's Happening When we Make Fragments?



- Vall database
- Primary sequence
- Secondary structure prediction
- NMR data (if applicable)

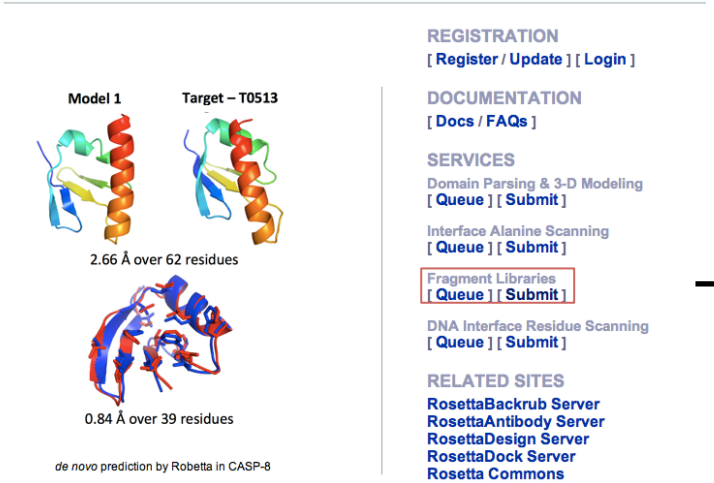
- Gather all possible fragments
- Score candidates based on input

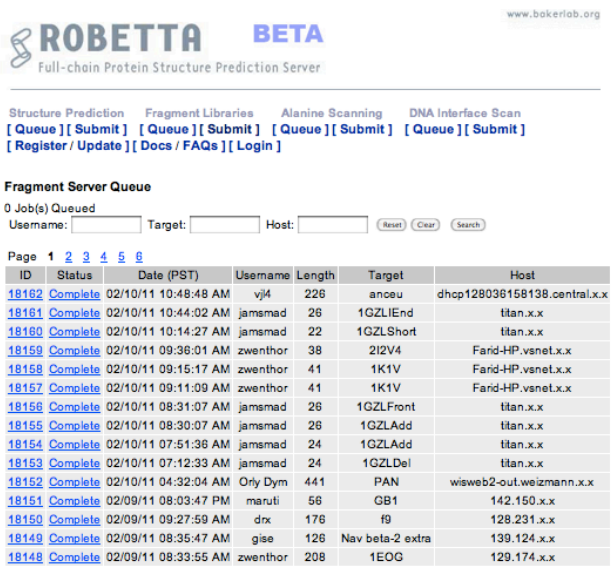
- Keep the best *N* fragment
- default=200 per sequence
- Write to fragment files

Figure courtesy of Dominik Gront

Making Fragments with Robetta

http://robetta.bakerlab.org/





ID	Status	Date (PST)	Username	Length	Target	Host
18162	Complete	02/10/11 10:48:48 AM	vjl4	226	anceu	dhcp128036158138.central.x.x
18161	Complete	02/10/11 10:44:02 AM	jamsmad	26	1GZLIEnd	titan.x.x
18160	Complete	02/10/11 10:14:27 AM	jamsmad	22	1GZLShort	titan.x.x
18159	Complete	02/10/11 09:36:01 AM	zwenthor	38	2I2V4	Farid-HP.vsnet.x.x
18158	Complete	02/10/11 09:15:17 AM	zwenthor	41	1K1V	Farid-HP.vsnet.x.x
18157	Complete	02/10/11 09:11:09 AM	zwenthor	41	1K1V	Farid-HP.vsnet.x.x
18156	Complete	02/10/11 08:31:07 AM	jamsmad	26	1GZLFront	titan.x.x
18155	Complete	02/10/11 08:30:07 AM	jamsmad	26	1GZLAdd	titan.x.x
18154	Complete	02/10/11 07:51:36 AM	jamsmad	24	1GZLAdd	titan.x.x
18153	Complete	02/10/11 07:12:33 AM	jamsmad	24	1GZLDel	titan.x.x
18152	Complete	02/10/11 04:32:04 AM	Orly Dym	441	PAN	wisweb2-out.weizmann.x.x
18151	Complete	02/09/11 08:03:47 PM	maruti	56	GB1	142.150.x.x
18150	Complete	02/09/11 09:27:59 AM	drx	176	f9	128.231.x.x
18149	Complete	02/09/11 08:35:47 AM	gise	126	Nav beta-2 extra	139.124.x.x
18148	Complete	02/09/11 08:33:55 AM	zwenthor	208	1EOG	129.174.x.x

Setting Up Options for AbinitioRelax

- First, create a new file called 2LZM_abrlx.options

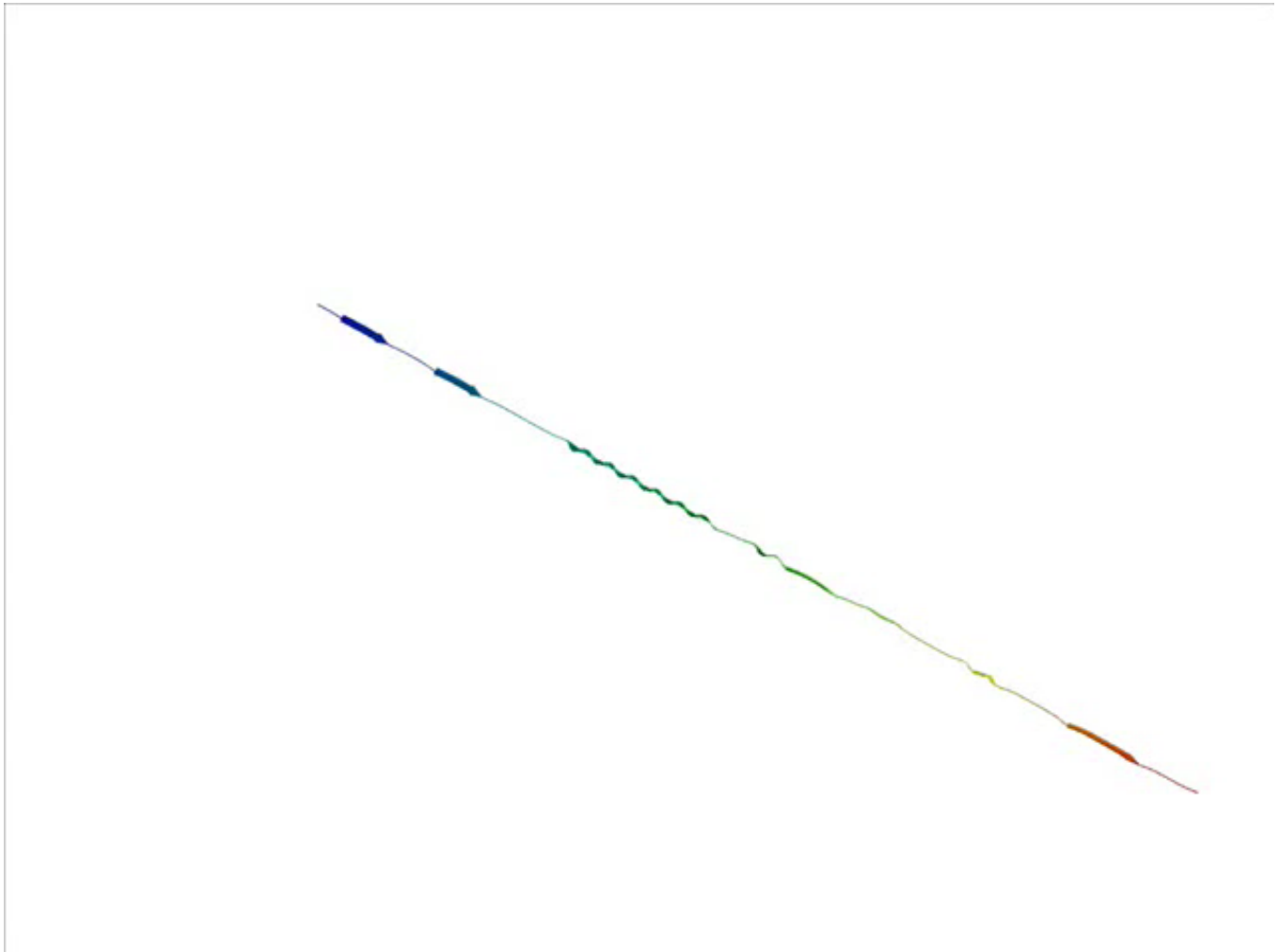
```
-in
    -file
        -native <native PDB file>        # native PDB file (optional)
        -fasta <primary sequence in FASTA format>     # protein sequence in fasta format
        -frag3 <3mer fragment file>      # protein 3-residue fragments file
        -frag9 <9mer fragment file>      # protein 9-residue fragments file
-psipred_ss2 <PSIPRED secondary structure prediction file>     # psipred_ss2 secondary structure
definition file (required for -use_filters)
-abinitio
    -increase_cycles 10     # Increase the number of cycles at each stage in AbinitioRelax by this factor
    -rg_reweight 0.5        # Reweight contribution of radius of gyration to total score by this scale factor
    -rsd_wt_helix 0.5       # Reweight env, pair, and cb scores for helix residues by this factor
    -rsd_wt_loop 0.5        # Reweight env, pair, and cb scores for loop residues by this factor
    -relax              # At the end of de novo folding, do a relax step
-relax
    -fast   # Type of relax protocol.  This has been shown to be the best deal for speed and robustness.
# Use radius of gyration (RG), contact-order, and sheet filters. This option conserves computing by not
continuing with refinement if a filter fails. A caveat is that for some sequences, a large percentage of
models may fail a filter. The filters are meant to identify models with non-protein like features
-use_filters true
```

Setting Up Options for AbinitioRelax Cont.

```
--run
      -reinitialize_mover_for_each_job # Job distributor generates fresh copy of its mover before each apply (once per job)
      -constant_seed  # Use a constant seed (1111111 unless specified with -jran)
      -jran 1111111  # Specify seed. Should be unique among jobs (requires -constant_seed)
-score
      -find_neighbors_3dgrid  # Use a 3D lookup table for doing neighbor calculations. For spherical, well-distributed conformations
-evaluation
      -rmsd <file to compute RMSD against> <column name> <file defining residues over which to compute RMSD>       # compute CA-RMSD for model comparing to native structure
-output        # use this to tell Rosetta you actually want output
      -nstruct 1    # how many structures do you want to generate?  Minimum of 1000 recommended
      -sf <scorefile>      # full path to scorefile
      -file
            -silent <silent output file>    # full path to silent file output
            -silent_struct_type binary      # we want binary silent files
-overwrite      # overwrite any existing output with the same name you may have generated
```

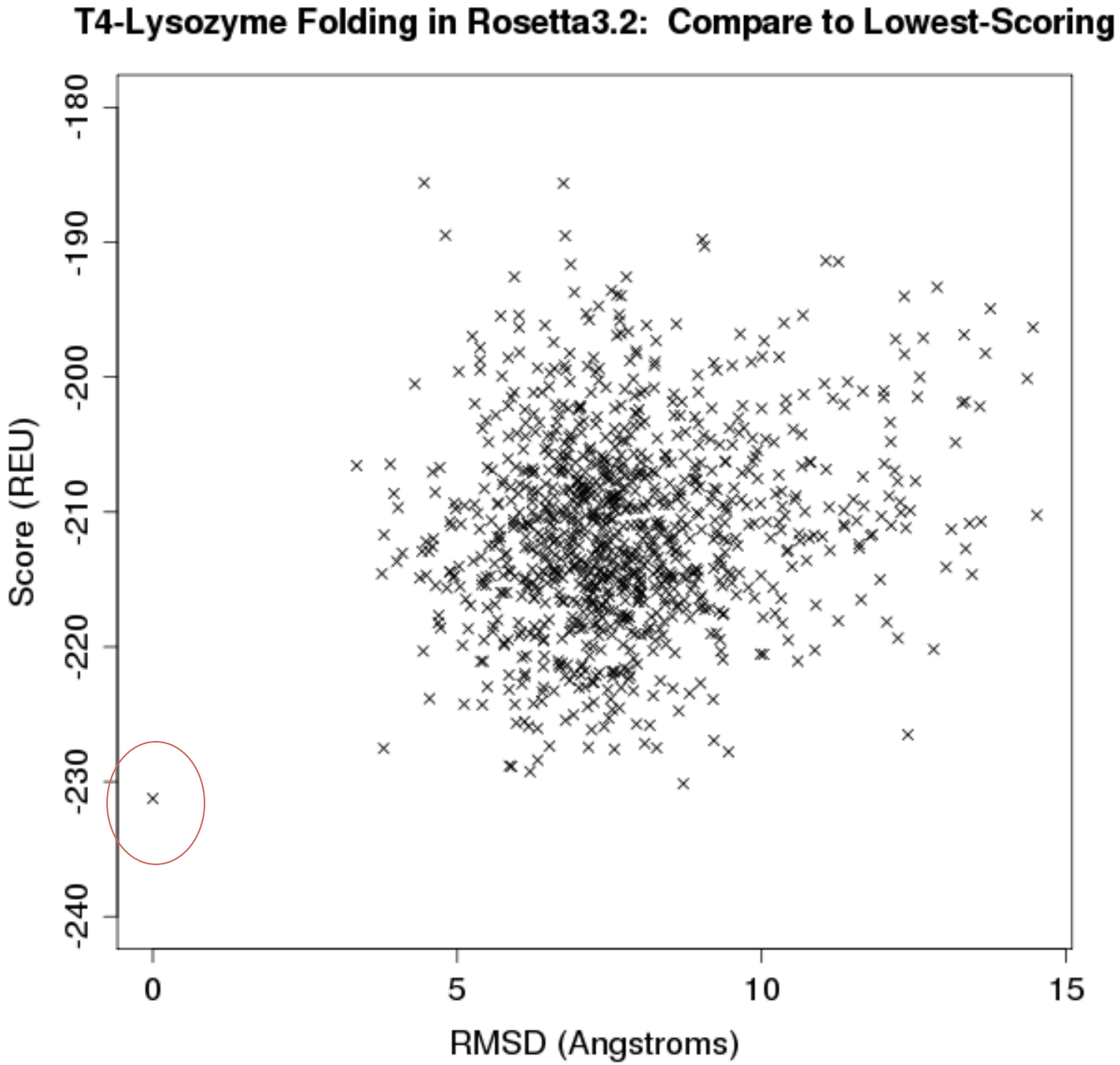
To run: `$ROSETTA_BIN/AbinitioRelax.$ROSETTA_SUFFIX @2LZM_abrlx.options –database $ROSETTA_DATABASE >& 2LZM_abrlx.log &`

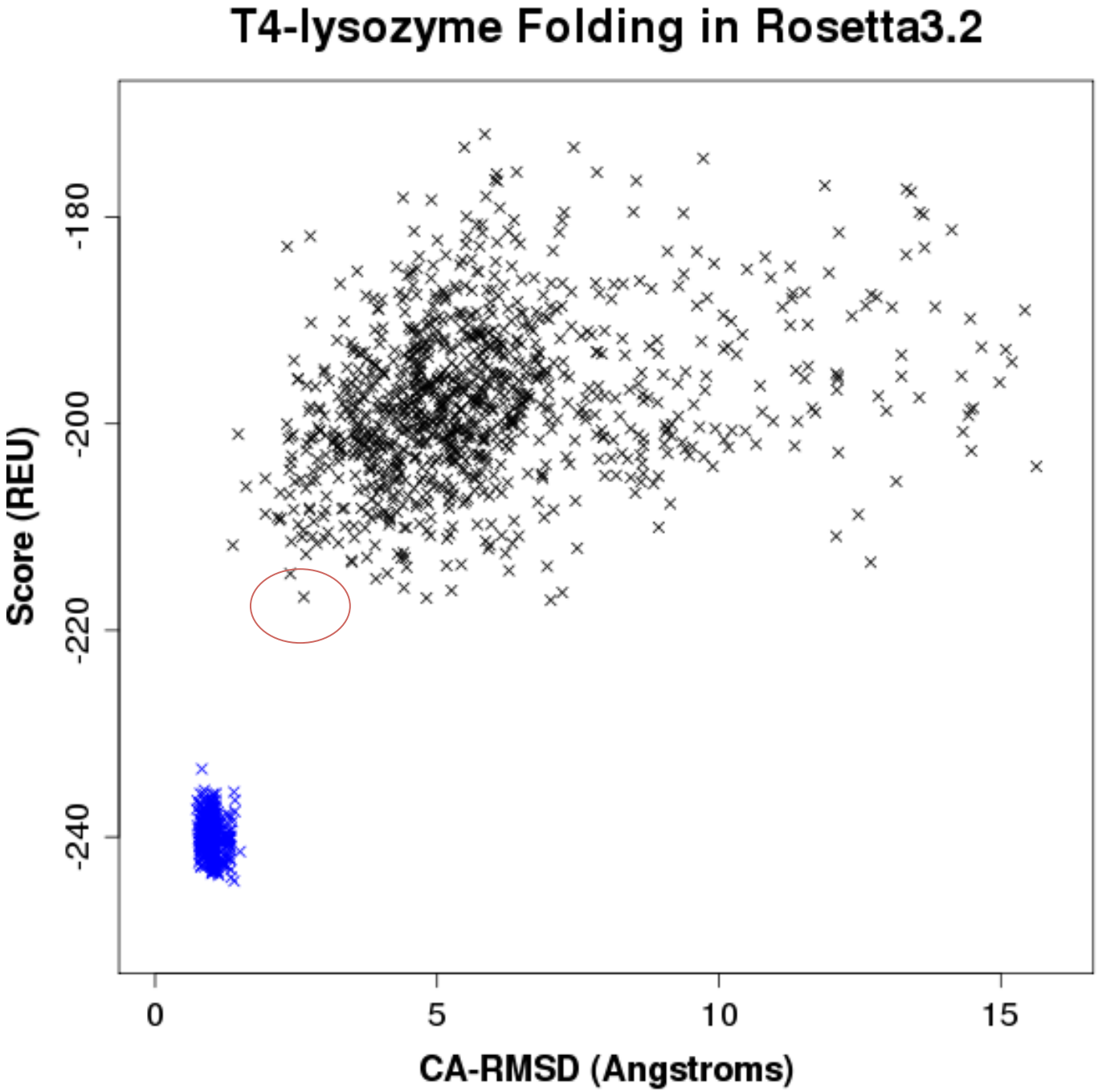
What's Actually Happening? Folding of Ubiquitin



Assessing Model Quality: Score vs. RMSD

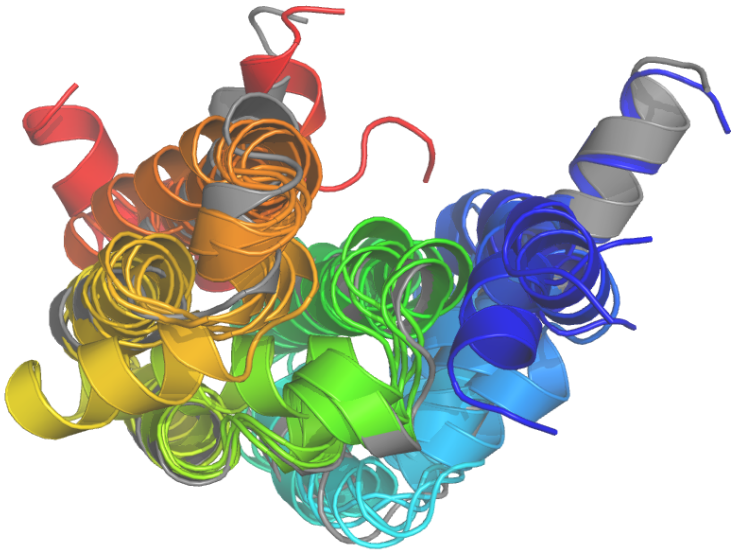
- Determine how well Rosetta energy correlates with model quality (RMSD, MaxSub, etc.)
- If you don't have a native structure (e.g., crystal structure) or a homolog that you'd like to compare the structure to, assume the lowest-scoring model is the native.
- Plot score vs. RMSD. Do you see "clusters" or populations of models? How does score relate to RMSD?
- Can also cluster (will be covered in another tutorial)


T4-Lysozyme Folding in Rosetta3.2: Compare to Lowest-Scoring
Score (REU)
-180
-190
-200
-210
-220
-230
-240
0
5
10
15
RMSD (Angstroms)

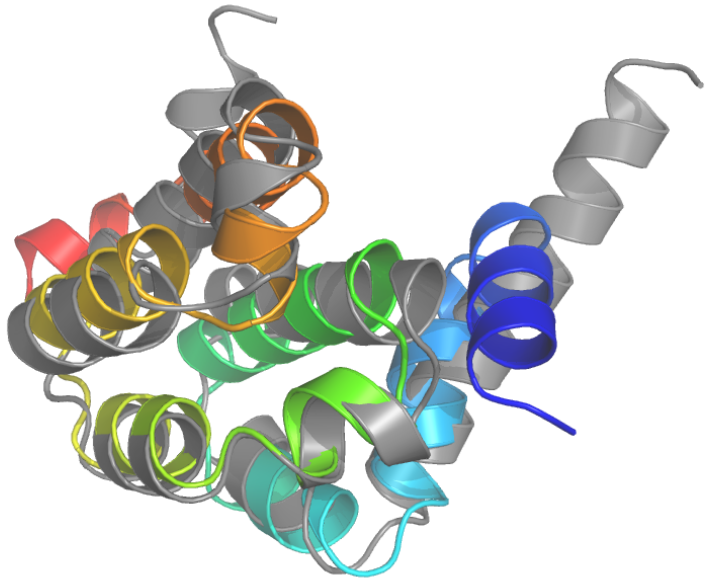

T4-lysozyme Folding in Rosetta3.2
Score (REU)
-180
-200
-220
-240
0
5
10
15
CA-RMSD (Angstroms)

Looking at Models in PyMol

```
/Applications/MacPyMOL.app/Contents/MacOS/MacPyMOL
2LZM_.pdb S_00000175_3*.pdb S_00000129_1*.pdb
S_00000026_2*.pdb S_00000168_2*.pdb S_00000028_4*.pdb
```

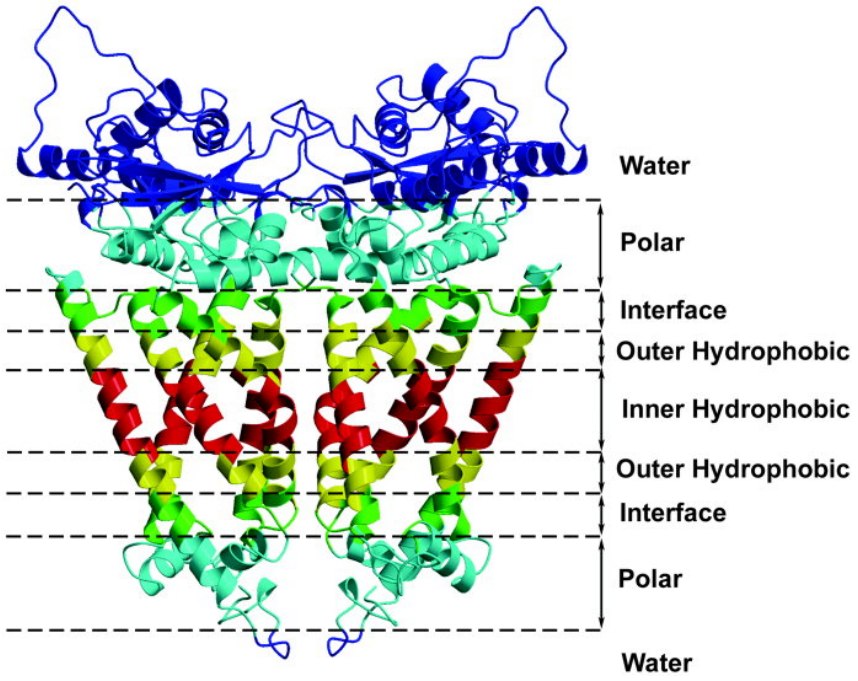


Top 5 scoring



Best-scoring

Folding Membrane Proteins



- The steps to follow are basically the same (including making fragments) with a couple extra steps. Data analysis pretty much the same.
- RosettaMembrane divides up the membrane into hydrophobic, hydrophilic, and soluble layers
- Membrane protein-specific scoring functions have been derived and are used in the MembraneAbinitio application

Figure from Yarov-Yarovoy, Schonbrun, and Baker 2006.

Input Files

Spanfile - *.span

--transmembrane topology prediction file generated using octopus2span.pl script

--Input OCTOPUS topology file is generated at http://octopus.cbr.su.se/ using protein sequence as input.

Lipopholicity prediction file - *.lips4

--Generate using run_lips.pl script

--Need input FASTA file, spanfile, blaspgp and nr (NCBI) database to run

Fragment generation

--Advised to use SAM but not JUFO or PSIPRED, which predict TMH regions poorly

Example Inputs and Command Line

comment

 # TMHs

 # residues

TM spans (2X)

spanfile

```
TM region prediction for
BRD4 predicted using OCTOPUS
4 123
antiparallel
n2c
  6    26     6    26
 31    51    31    51
 58    78    58    78
 97   117    97   117
```

lips4 file

```
Lipid exposed data:
resnum mean-lipo lipophil entropy
     6  -1.000    3.004   1.211
     9  -1.000    2.268   2.137
    10  -1.000    4.862   1.095
    13  -1.000    1.304   1.552
    16  -1.000    3.328   2.025
```

MembraneAbinitio Options

```
-in:file:native <input native PDB>
-in:file:fasta <primary sequence in FASTA format>
-in:file:frag3 <3mer fragment file>
-in:file:frag9 <9mer fragment file>
-in:file:spanfile <spanfile>  # newly created spanfile
-in:file:lipofile <lipophilicity lips4 file>   # newly created lipo file
```

```
-run:reinitialize_mover_for_each_job
-score:find_neighbors_3dgrid
-abinitio:membrane # specify membrane abinitio protocol
-membrane:no_interpolate_Mpair # membrane scoring specification
-membrane:Menv_penalties # turn on membrane penalty scores
-rg_reweight 0.01 # radius of gyration weight not so important for MPs
-stage2_patch <score_membrane_s2.wts_patch> # weights for scores
-stage3a_patch <score_membrane_s3a.wts_patch> # weights for scores
-stage3b_patch <score_membrane_s3b.wts_patch> # weights for scores
-stage4_patch <score_membrane_s4.wts_patch> # weights for scores
-evaluation::gdtmm # output global distance test info
```

```
-out:nstruct 1   # minimum of 1000 recommended
-out:file:scorefile <path to scorefile>
-out:file:silent <path to silent output file>
-out:file:silent_struct_type binary
```

`$ROSETTA_BIN/membrane_abinitio2.$ROSETTA_SUFFIX`
`   @BRD4_mem_abrlx.options –database $ROSETTA_DATABASE >& logfile &`

Folding with Restraints

- Basically the same as normal *de novo* folding protocol except add a few flags to options file

```
-fold_cst # use FoldConstraints protocol
        -force_minimize # minimize in FoldConstraints protocol
-constraints
        -cst_file ./2LZM_dist_w1.cst # path to your cst file
        -cst_weight 4 # factor by which total cst score multiplied by
        -epr_distance # Use RosettaEPR knowledge-based potential
```

Constraint info					Constraint Function info				
<cst type>	<atom1>	<res1>	<atom2>	<res2>	<cst_func>	<RosettaEPR>	<Dcb>	<weight>	<bin>
AtomPair	CB	32	CB	36	SPLINE	EPR_DISTANCE	16.0	1.0	0.5
AtomPair	CB	59	CB	74	SPLINE	EPR_DISTANCE	19.0	1.0	0.5
AtomPair	CB	62	CB	71	SPLINE	EPR_DISTANCE	19.0	1.0	0.5
AtomPair	CB	62	CB	74	SPLINE	EPR_DISTANCE	25.0	1.0	0.5
AtomPair	CB	63	CB	74	SPLINE	EPR_DISTANCE	14.0	1.0	0.5
AtomPair	CB	66	CB	74	SPLINE	EPR_DISTANCE	23.0	1.0	0.5
AtomPair	CB	83	CB	90	SPLINE	EPR_DISTANCE	13.0	1.0	0.5

What is with this Constraints File?

- There are *constraint* types and *function* types
  - **Constraint types:** AtomPair, Angle, Dihedral, etc.
  - **Function types:** Bounded, Spline, Harmonic, Gaussian, etc.
- Each constraint you define is scored individually, and the total constraint score is the sum of all individual scores
- Each constraint can have its own constraint type and function type.
  - In some cases, like when using Spline function, each constraint can have its own weight
- How you define the constraint and how it's scored depends on the constraint type; this is same with function type.

Analysis After Folding with Restraints

- Can often filter by constraint score so that only look at models that satisfy experimental data the best
- Can plot score vs. RMSD, constraint score vs. RMSD, total score vs. constraint score, etc. to get idea of correlation of constraint score with total energy of model
- Can see how many violations your model has, how big the violations are, etc.

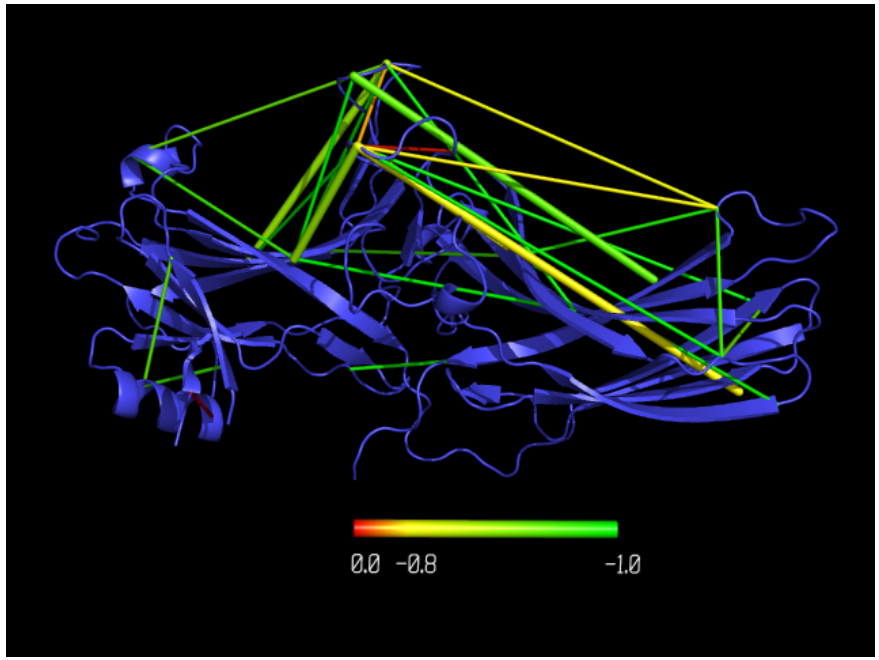


Figure courtesy of Nathan Alexander

A Few More Things to Keep in Mind

- Rosetta3.2 *de novo* folding performs best with small proteins (< 100 residues)
- Folding larger, more complex proteins probably requires more restraints
- *Can* fold membrane proteins with experimental restraints (EPR, NMR, etc.).
  - Exact protocol seems to depend on system and problem being addressed
- More folding capabilities in more recent versions that have not been released (more to come!)

Useful Links and Papers

- **Rosetta User's Guide:**
  - http://www.rosettacommons.org/manuals/archive/rosetta3.2_user_guide/
- ***De novo* folding**
  - http://www.rosettacommons.org/manuals/archive/rosetta3.2_user_guide/abinitio.html
  - Rohl C, *Methods Enzymol.,* 2004.
- **Membrane protein folding**
  - http://www.rosettacommons.org/manuals/archive/rosetta3.2_user_guide/membrane_abinitio.html
  - Yarov-Yarovoy V, Schonbrun J, Baker D, *Proteins*, 2006; Barth P, Schonbrun J, Baker D, *PNAS*, 2007; Barth P, Wallner B, Baker D, *PNAS*, 2009.
- **Using constraints/restraints in Rosetta 3.2**
  - http://www.rosettacommons.org/manuals/archive/rosetta3.2_user_guide/constraints.html
  - Rohl, C, *Methods Enzymol*., 2005; Raman *et al*., *Science*, 2010; Hirst *et al*., *J. Struct. Biol.*, 2011.