

De novo Folding with RosettaMembrane

Rosetta Workshop
April 17, 2014

Yan Xia
yan.xia@vanderbilt.edu
Meiler Laboratory, Vanderbilt University
<http://www.meilerlab.org>

Tutorial outline

- Intro to *de novo* protein folding
- How is it done in Rosetta
- *De novo* folding of membrane proteins
 - Necessary input files
 - Running *de novo* folding in membrane mode
 - Running folding with structural restraints
 - Extracting data and analyzing output
- Useful references and websites

Intro to *de novo* protein folding

Template modeling is tempting, but not always feasible.

There is about 1000 distinct integral membrane protein families out there.

100 protein families will need to be structurally characterized in order to cover the human alpha-helical proteome to 58% of all sequences (Pieper et al, 2013).

For proteins without a template, *de novo* folding is one alternative.

Intro to *de novo* protein folding

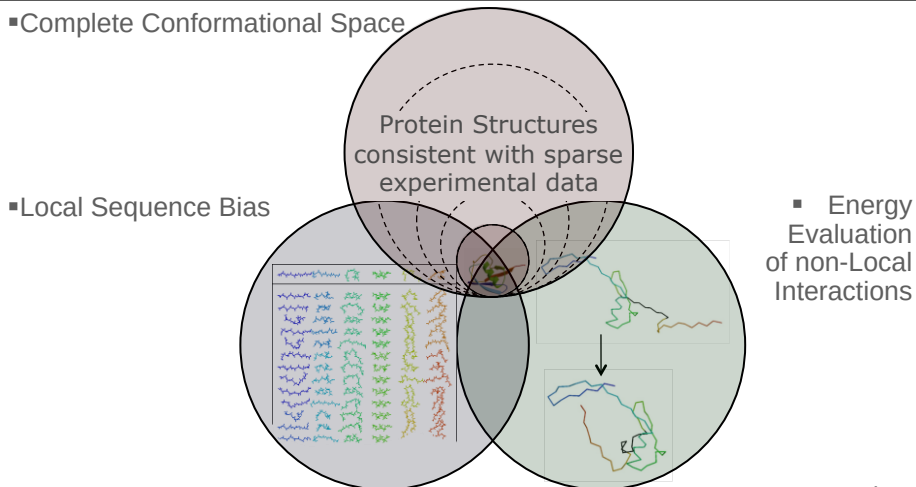
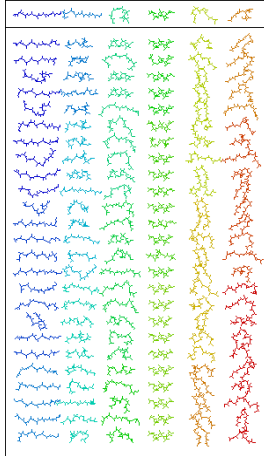


Figure from Jens Meiler

Intro to *de novo* protein folding

MDIGPANEIQEL

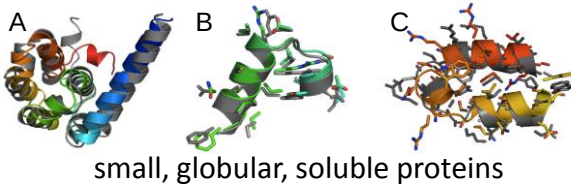


- We are folding *from the primary sequence* using secondary structure prediction and peptide fragments from the PDB
- Use the fragments to change the geometry of the protein and score to keep good fragment insertions

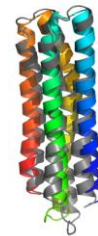
5
Figure courtesy of Jens Meiler

What can Rosetta actually fold?

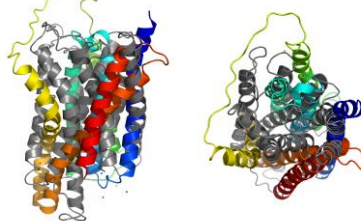
T4-lysozyme C-terminal domain



**V-type Na⁺
ATP synthase
subunit**



rhodopsin

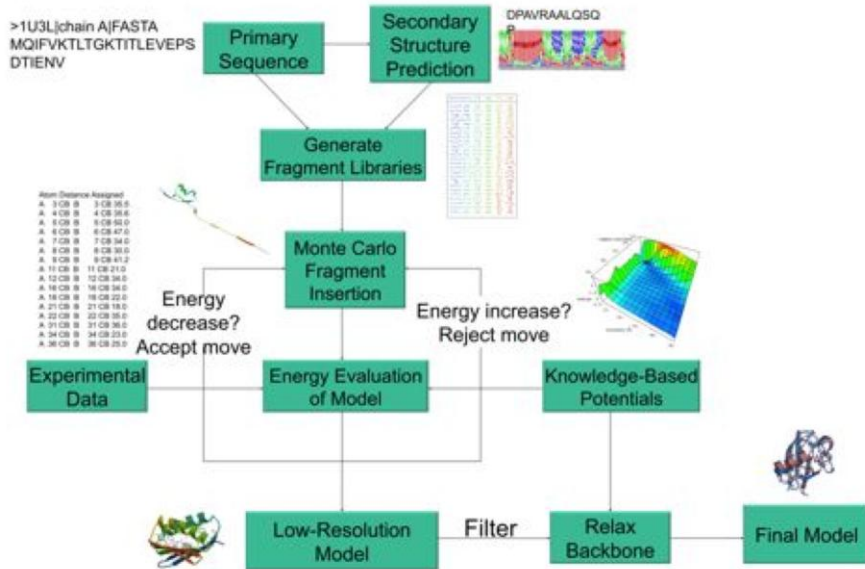


small, simple
membrane proteins

...but not large, complex
proteins

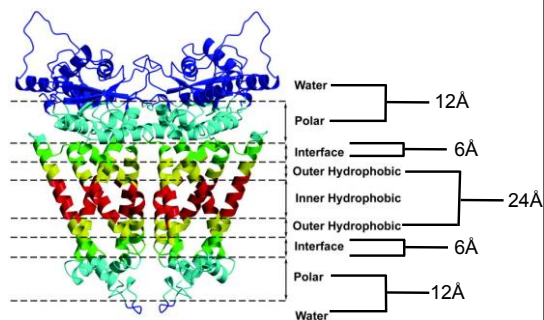
6

Rosetta *de novo* folding protocol



How does Rosetta place the protein in the membrane?

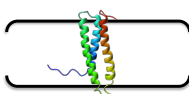
Membrane center and normal defined by input TM span information



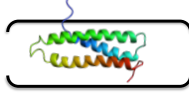
Yarov-Yarovoy, Schonbrun, and Baker 2006.

RosettaMembrane penalties

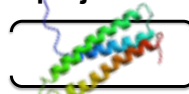
non-helix



termini



projection



span

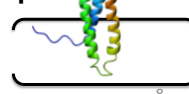
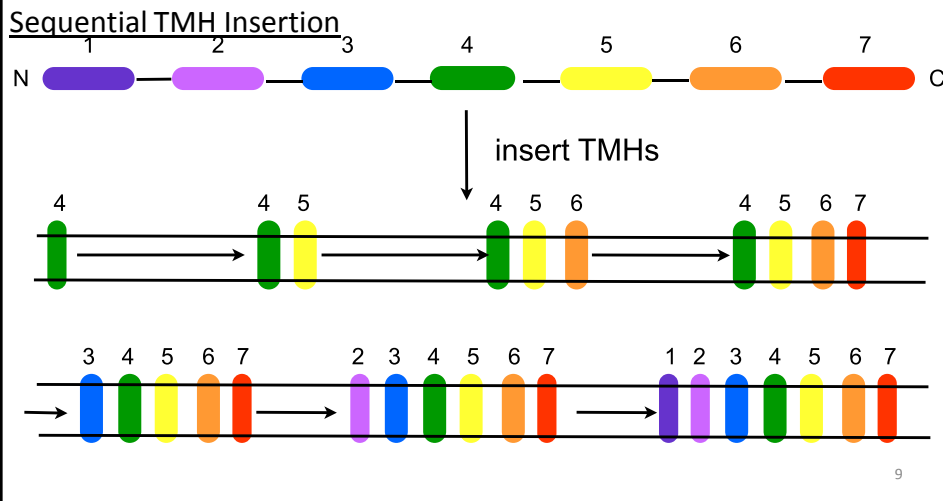


Figure from Elizabeth D. Nguyen

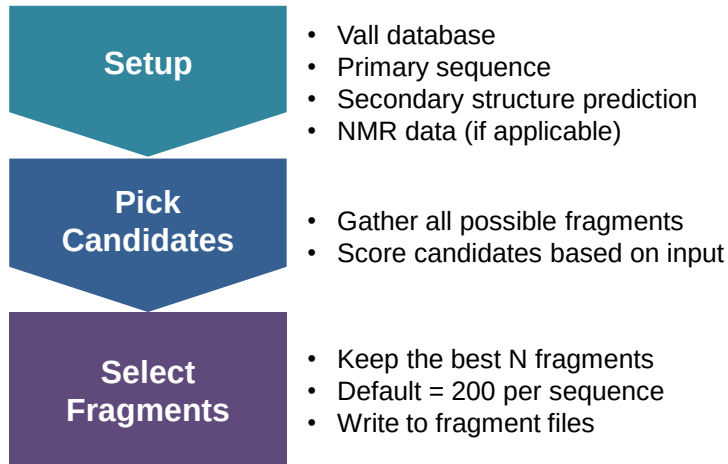
RosettaMembrane folds proteins differently than soluble proteins



Step 1: Necessary input files for the abinitio folding protocol

- FASTA file of your protein sequence
- “Clean” PDB file of native structure (optional)
- **Fragment library files**
- Options file
- Topology broker setup file
- Rigid definition file
- Secondary structure prediction file

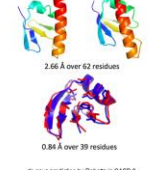
Fragment file generation



A note about making fragments in the tutorial

- In Step 1.c in the tutorial, you are told how to make fragments for folding
- There are 2 ways to make fragments: Using Robetta (5.1), if you are not-for-profit user, or `make_fragments.pl` (5.2)
- You can CHOOSE which one you would like to use. Keep in mind that the Robetta server may become too busy, resulting in long wait times.
- The fragment files provided for you were made using Robetta.

Making fragments with Robetta



Model 1 **Target – T0513**

2.66 Å over 62 residues

0.86 Å over 39 residues

de novo prediction by Robetta in CASP-6

ROBETTA BETA
Full-chain Protein Structure Prediction Server

Registration: [Register] [Update] [Login]

Documentation: [Docs] [FAQs]

Services: [Download] [Submit]

Interface: [Alpha3 Scanning] [Query] [Submit]

DNA Interface: [Residue Scanning] [Query] [Submit]

Related Sites: Rosetta@home, RosettaAntibody Server, RosettaDesign Server, RosettaDock Server, Rosetta Commons, Foldit, Rosetta@home

www.robetta.org

➔

ROBETTA BETA
Full-chain Protein Structure Prediction Server

Structure Prediction: [Query] [Submit] | Fragment Libraries: [Query] [Submit] | Alanine Scanning: [Query] [Submit] | DNA Interface Scan: [Register] [Update] [Docs] [FAQs] [Login]

Submit a job to the Fragment Server

*Please submit one job at a time

• Identifier must be at least 5 alphanumeric characters

Required

Registered Username: _____ or Registered Email Address: _____

Target Name:

Paste Fasta

> 2L2M Sequence
 PTQKASRLPNQDVDAVRGILNNAKLKIPVYDSDLVARRCCLMMVFMQGTGV
 AGFTNLSRLMLQQKRRVDEAAVNLAKSRVYWNQTNRAKRVITFTCTWDVAYNL

or Upload Fasta: no file selected

Optional

Identifier:

Exclude Homologues:

Rosetta NMR (link looks better from input format)
Chains to Shift: no file selected
NOE Constraints: no file selected
Disorder Constraints: no file selected

ROBETTA BETA
Full-chain Protein Structure Prediction Server

Structure Prediction: [Query] [Submit] | Fragment Libraries: [Query] [Submit] | Alanine Scanning: [Query] [Submit] | DNA Interface Scan: [Register] [Update] [Docs] [FAQs] [Login]

Fragment Server Queue

6 jobs in Queue

Username: _____ Target: _____ Host: _____ (Stop) (Wait) (Cancel)

Pages: 1 2 3 4 5 6

ID	Status	Date (YYYY)	Username	Length	Target	Host
15152	Complete	10/11/11 15:48:49 AM	af	225	dhg232031918103.asterlab.c...	Itan.s.x
15151	Complete	10/11/11 04:44:42 AM	jamrzed	2	1G2D.R4d	Itan.s.x
15150	Complete	10/11/11 05:15:40 AM	jamrzed	25	1G2D.R4d	Itan.s.x
15149	Complete	10/11/11 05:36:01 AM	awsewer	38	2QV4	Fasta@P.asterlab.c...
15148	Complete	10/11/11 05:17:48 AM	awsewer	41	1KTV	Fasta@P.asterlab.c...
15147	Complete	10/11/11 05:11:05 AM	awsewer	41	1KTV	Fasta@P.asterlab.c...
15146	Complete	10/11/11 05:31:07 AM	jamrzed	25	1G2D.Fast	Itan.s.x
15145	Complete	10/11/11 05:30:01 AM	jamrzed	25	1G2D.Asd	Itan.s.x
15144	Complete	10/11/11 07:51:58 AM	jamrzed	24	1G2D.Asd	Itan.s.x
15143	Complete	10/11/11 07:12:35 AM	awsewer	24	1G2D.D	Itan.s.x
15142	Complete	10/11/11 05:02:04 AM	gty	441	1P4N	fasta@cd.asterlab.c...
15141	Complete	10/11/11 05:03:47 PM	msd	58	GB1	1G2.15D.s.x
15140	Complete	10/11/11 05:07:56 AM	gty	176	6	126.127.s.x
15139	Complete	10/08/11 05:35:47 AM	gty	120	New beta 2.aster	126.124.s.x
15138	Complete	10/08/11 05:30:51 AM	awsewer	208	1E05	126.124.s.x

➔

Making fragments with make_fragments.pl

- If you are an industry user, you will not have access to Robetta
- You can use the `make_fragments.pl` script, which is discussed in the tutorial
- This script will need to be modified to work with your own environment

```
#####
# USER CONFIGURATION -- ONLY CHANGE THIS SECTION.
#                               YOU MUST EDIT THESE PATHS BEFORE USING
#####

# change the following paths to point to the locations of your copies of these
# files, databases or directories.

# BLAST install path (Requires non-blast+ NCBI version)
# recommended: blast-2.2.17
# ftp://ftp.ncbi.nih.gov/blast/executables/release/2.2.17/
my $BLAST_DIR = "/sb/meiler/Linux2/x86_64/blast/blast-2.2.18"
; # "/my_src_dir/blast-2.2.17_x64/blast-2.2.17";
my $BLAST_NUM_CPUS = 1; # number of processors to use (blastpgp -a option)

# NR blast database filename
# Create blast db files using 'formatdb -o T -i nr'
# ftp://ftp.ncbi.nih.gov/blast/db/FASTA/nr.gz
my $NR = "/sb/meiler/scripts/sequence_analysis/db/nr"; #"/my_database_dir/blast/nr";

# RCSB PDB files for homolog detection and date filtering when benchmarking
my $PDB_SEQRES = "/work/rosetta/databases/pdb/pdb_seqres.txt";
my $PDB_ENTRIES_IDX = "/work/rosetta/databases/pdb/entries.idx";

# DEFAULT PDB STRUCTURE, PSSM, and SPARTA (chemical shift) DATABASE (VALL)
# 'vall_files' option overrides this
my $VALL = "/blue/meilerlab/apps/rosetta/rosetta-3.4/rosetta_tools/fragment_tools/vall.jul19.2011.gz"
; #"/my_database_dir/vall.apr24.2008.extended";

# pdb2vall.py script for adding specific PDBs to the vall (-add_pdb_to_vall)
my $PDB2VALL = "/blue/meilerlab/apps/rosetta/rosetta-3.4/rosetta_tools/fragment_tools/pdb2vall/pdb2vall.py";
my $PDB2VALL_IGNORE_ERRORS = 1;

# ROSETTA FRAGMENT PICKER
my $FRAGMENT_PICKER =
"/blue/meilerlab/apps/rosetta/rosetta-3.4/rosetta_source/bin/fragment_picker.default.linuxgccrelease"
; #"/my_src_dir/mini/bin/picker.linuxgccrelease";
# fragment_picker must be built with extras=boost_thread for multiple threads
my $FRAGMENT_PICKER_NUM_CPUS = 1; # number of processors to use

my $ROSETTA_DATABASE = "/blue/meilerlab/apps/rosetta/rosetta-3.4/rosetta_database"
; #"/my_database_dir/rosetta_database";
```

15

Step 1: Necessary input files for the abinitio folding protocol

- FASTA file of your protein sequence
- “Clean” PDB file of native structure (optional)
- **Fragment library files**
- **Options file**
- Topology broker setup file
- Rigid definition file
- Secondary structure prediction file

16

Step 2: RosettaMembrane Input Files

```
-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
-membrane:Membed_init # initialize membrane embedding
-rg_reweight 0.01 # radius of gyration weight not so important for MPs
-abinitio:stage2_patch <score_membrane_s2.wts_patch> # weights for scores
-abinitio:stage3a_patch <score_membrane_s3a.wts_patch> # weights for scores
-abinitio:stage3b_patch <score_membrane_s3b.wts_patch> # weights for scores
-abinitio:stage4_patch <score_membrane_s4.wts_patch> # weights for scores

-evaluation::gdtmm # output global distance test info
-evaluation:rmsd <native file> <score column suffix> <rigid file>
-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
```

17

Step 2: RosettaMembrane 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.

Lipophilicity prediction file - *.lips4

--Generate using run_lips.pl script

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

Rigid file specifying TM regions

-- So that we compute RMSD in the TM regions

18

Step 2: Example span and lipophilicity files

comment — TM region prediction for BRD4
predicted using OCTOPUS

TMHs — 4 123 — **# residues**
antiparallel
n2c

TM spans (2X) —

spanfile

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

19

Step 3: Running the membrane protein folding job

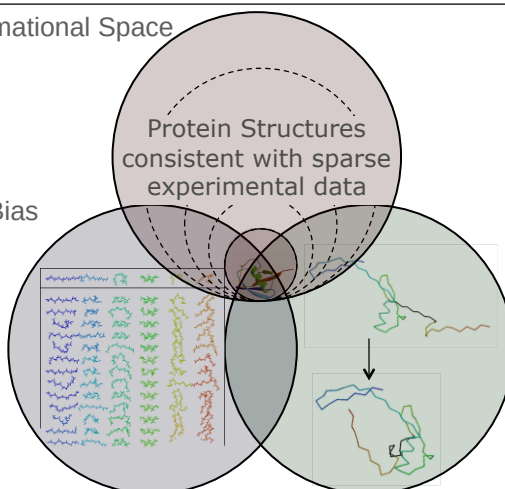
```
~/rosetta_workshop/Rosetta/main/source/bin/membrane_abinitio2.default.linuxgccrelease\
-database
~/rosetta_workshop/Rosetta/main/database
@BRD4_mem_abrlx.options
```

20

Why (Not) combine experimental restraints with Rosetta?

Complete Conformational Space

Local Sequence Bias



Energy Evaluation of non-Local Interactions

Figure from Jens Meiler

Step 4: Folding with restraints

- Similar for soluble and membrane proteins
- 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
-cst_fa_file ./2LZM_dist_w1.cst # path to your cst file
-cst_fa_weight 4 # factor by which 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	

Step 4: Folding with restraints

- Similar for soluble and membrane proteins
- 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
-cst_fa_file ./2LZM_dist_w1.cst # path to your cst file
-cst_fa_weight 4 # factor by which 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

23

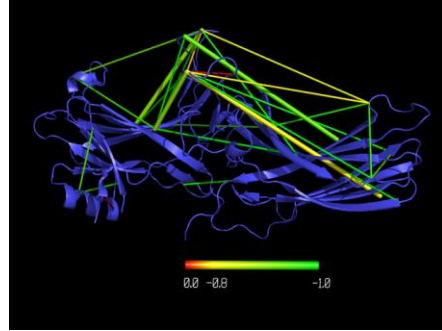
What is the constraints file?

- 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.

24

Step 4: 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.



25
Figure courtesy of Nathan Alexander

Step 5: 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)

26

A few more things to keep in mind

- Rosetta3.5 *de novo* folding performs best with small proteins (80-100 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!)

27

Useful links and papers

- **Rosetta User's Guide:**
 - http://www.rosettacommons.org/manuals/archive/rosetta3.5_user_guide/
- ***De novo* folding**
 - https://www.rosettacommons.org/manuals/archive/rosetta3.5_user_guide/d/d26/_abinitio_relax.html
 - Rohl C, *Methods Enzymol.*, 2004.
- **Membrane protein folding**
 - https://www.rosettacommons.org/manuals/archive/rosetta3.5_user_guide/d/d38/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 restraints in Rosetta 3.5**
 - https://www.rosettacommons.org/manuals/archive/rosetta3.5_user_guide/d/d50/constraint_file.html
 - Rohl, C, *Methods Enzymol.*, 2005; Raman *et al.*, *Science*, 2010; Hirst *et al.*, *J. Struct. Biol.*, 2011.

28