

De novo Folding in Rosetta

Rosetta Workshop
November 2013

Jessica A. Finn
jessica.a.finn@vanderbilt.edu
Meiler Laboratory, Vanderbilt University
<http://www.meilerlab.org>

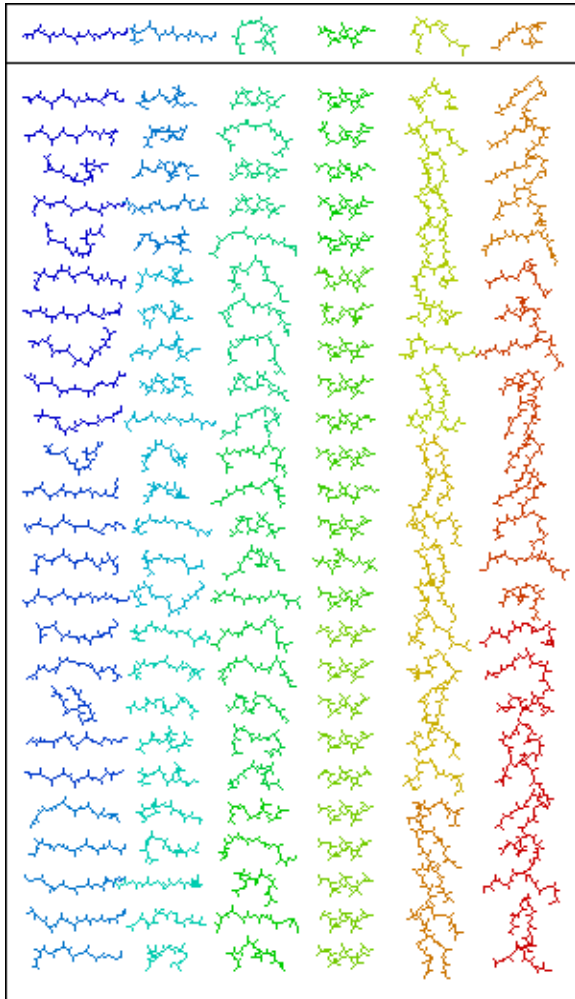
Tutorial outline

- Introduction to *de novo* protein folding
- *De novo* folding of soluble proteins
 - Preparing input files
 - *De novo* folding with restraints
 - Analyzing output files
- Useful references and websites

What is *de novo* protein folding?

- A technique to determine tertiary structure *from a primary sequence* using secondary structure prediction and peptide fragments selected from the PDB
- Differs from homology modeling, which starts with a template structure

What are fragments?



- 3 and 9 amino acid peptides generated from the PDB
- Fragments change the geometry of the protein
- Scoring functions identify and maintain good fragments

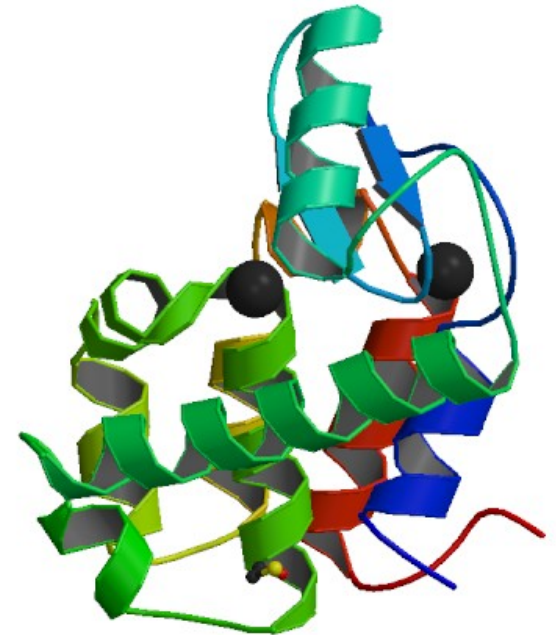
What is *de novo* protein folding?

MNIFEMLRIDEGLRLKI
YKDTEGYTIGIGHLLT
KSPSLNASKSELDKAIG
RNTNGVITKDEAEKLFN
QDVDAAVRGILRNAKLK
PVYDSLDAVRRALINM
VFQMGETGVAGFTNSLR
MLQQKRWDEAAVNLAKS
RWYNQTPNRAKRVITTF
RTGTWDAYKNL

Primary Sequence

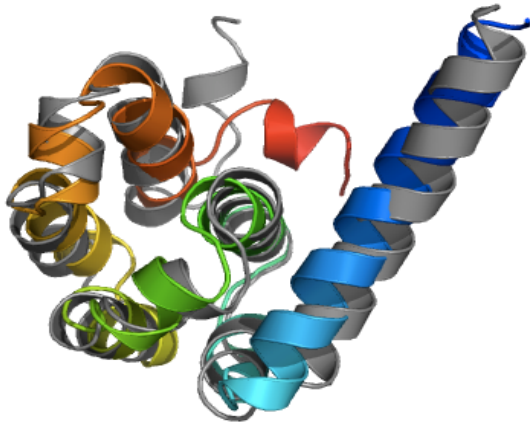
Secondary Structure
Prediction, PDB
Fragments, Rosetta
Scoring Functions

de novo folding



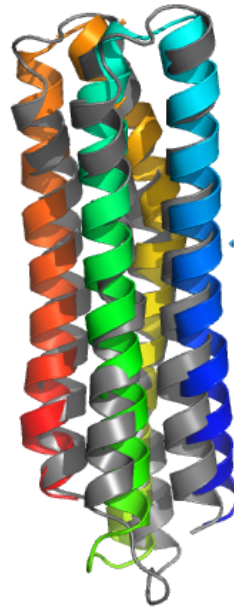
Tertiary Structure

What can Rosetta actually fold?



Small, globular,
soluble proteins

**T4-lysozyme
C-terminal domain**



Small, simple
membrane proteins

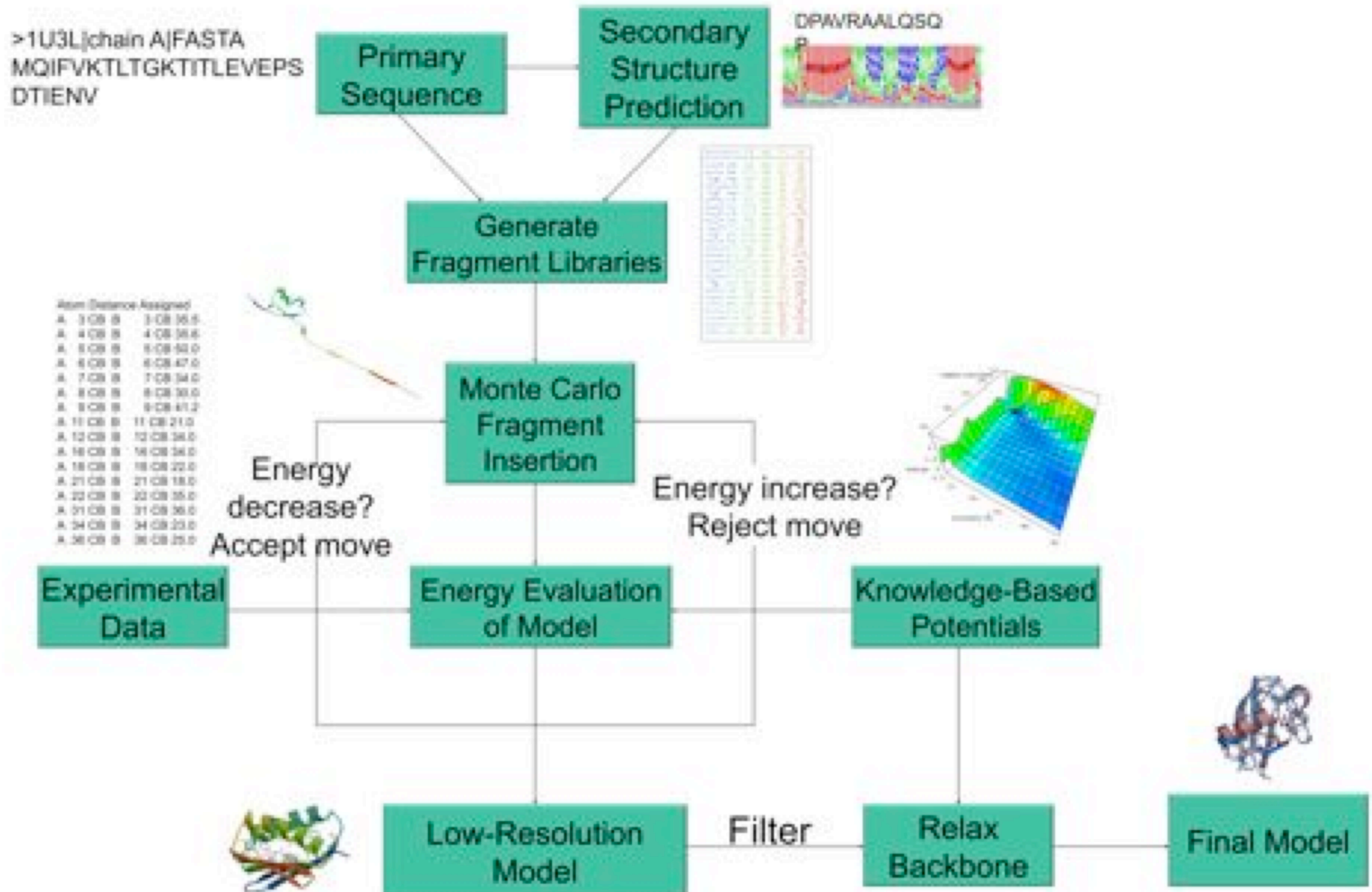
**V-type Na⁺ ATP
synthase subunit**



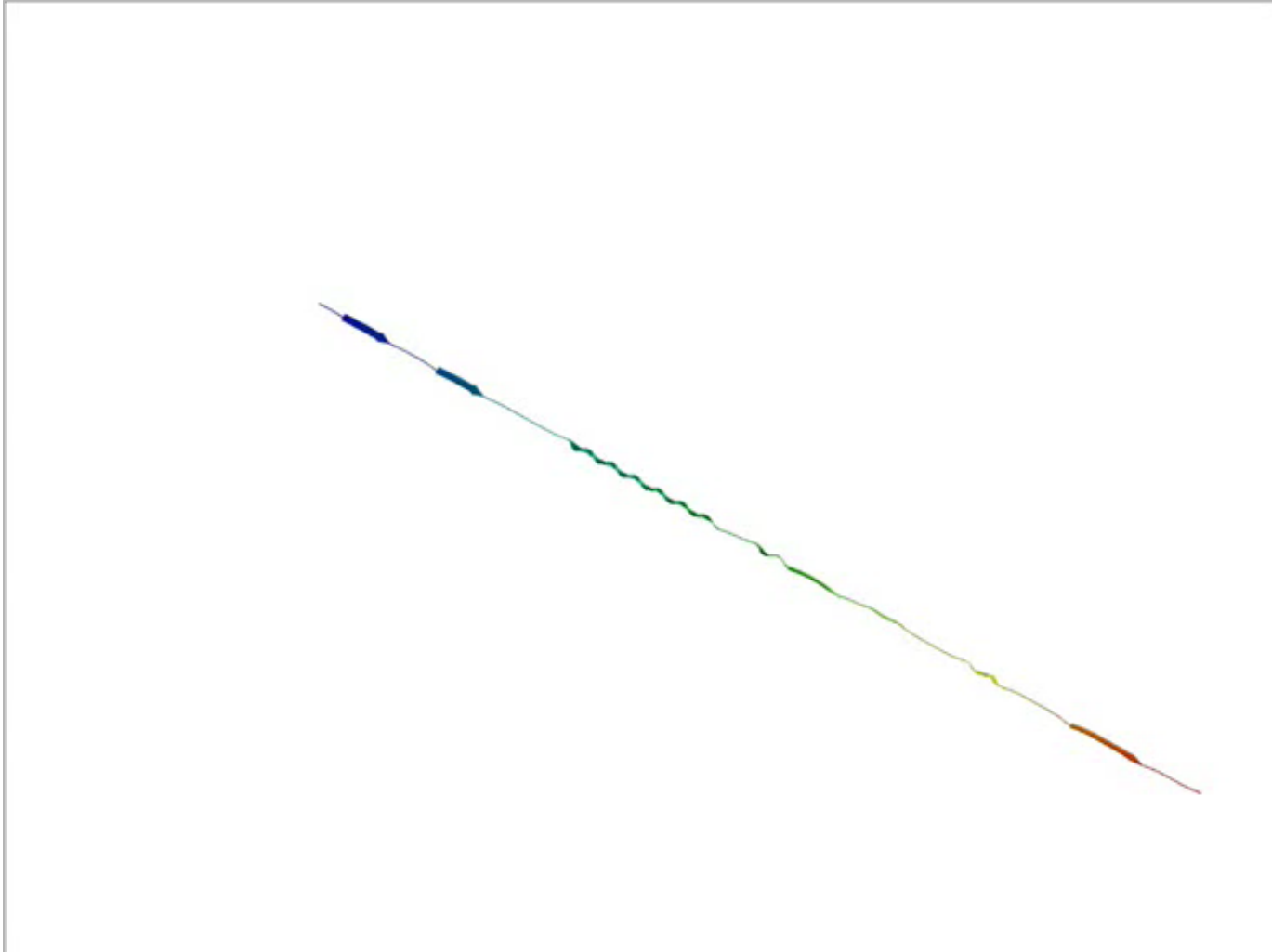
...but not large,
complex proteins

Rhodopsin

Rosetta *de novo* folding protocol



Running a folding job: ubiquitin



Necessary input files for the *de novo* folding protocol

- FASTA file of your protein sequence
- Secondary structure prediction files
- **Fragment library files**
- **Options file**
- Topology broker setup file
- *(Optional) PDB file of native structure*
- ***(Optional) Constraints file***

Fragment file generation

Setup

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

Pick Candidates

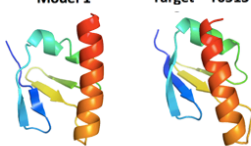
- Gather all possible fragments
- Score candidates based on input

Select Fragments

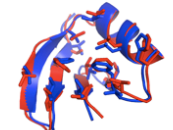
- Keep the best N fragments
- Default = 200 per sequence
- Write to fragment files

Making fragments with Robetta

Model 1 **Target – T0513**



2.66 Å over 62 residues



0.84 Å over 39 residues

de novo prediction by Robetta in CASP-8

REGISTRATION
[Register / Update] [Login]

DOCUMENTATION
[Docs / FAQs]

SERVICES
Domain Parsing & 3-D Modeling
[Queue] [Submit]
Interface Alanine Scanning
[Queue] [Submit]
Fragment Libraries
[Queue] [Submit]
DNA Interface Residue Scanning
[Queue] [Submit]

RELATED SITES
[RosettaBackrub Server](#)
[RosettaAntibody Server](#)
[RosettaDesign Server](#)
[RosettaDock Server](#)
[Rosetta Commons](#)
[FoldIt](#)
[Rosetta@home](#)

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

ROBETTA BETA www.bakerlab.org
Full-chain Protein Structure Prediction Server

Structure Prediction [Queue] [Submit] Fragment Libraries [Queue] [Submit] Alanine Scanning [Queue] [Submit] DNA Interface Scan [Queue] [Submit]
[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:** stephanie.j.hirst@vanderbil

Target Name:
 2LZM_

Paste Fasta
2LZM Sequence
ITKDEAEKLFNQDVAARVILRNALPKPVYDSLDAVRRCALINMVFQMGETGV
AGFTNSLRMLQQRWDEAAVNLAWSRWYQTPNRAKRVITFTRTGTWDAYKNL

or Upload Fasta: no file selected

Optional
Identifier: 2LZM_
Exclude Homologues: ☐

Rosetta NMR (click links below for input format)
Chemical Shifts: no file selected
NOE Constraints: no file selected
Dipolar Constraints: no file selected

ROBETTA BETA www.bakerlab.org
Full-chain Protein Structure Prediction Server

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

Fragment Server Queue
0 Job(s) Queued
Username: Target: Host:

Page 1 2 3 4 5 6

ID	Status	Date (PST)	Username	Length	Target	Host
18182	Complete	02/10/11 10:48:48 AM	vji4	226	anceu	dhcp128038158138.central.x.x
18181	Complete	02/10/11 10:44:02 AM	jamsmad	26	1GZLIEnd	titian.x.x
18180	Complete	02/10/11 10:14:27 AM	jamsmad	22	1GZLShort	titian.x.x
18159	Complete	02/10/11 09:36:01 AM	zwenhor	38	2I2V4	Farid-HP.vsnat.x.x
18158	Complete	02/10/11 09:15:17 AM	zwenhor	41	1K1V	Farid-HP.vsnat.x.x
18157	Complete	02/10/11 09:11:09 AM	zwenhor	41	1K1V	Farid-HP.vsnat.x.x
18156	Complete	02/10/11 09:31:07 AM	jamsmad	26	1GZLFront	titian.x.x
18155	Complete	02/10/11 08:30:07 AM	jamsmad	26	1GZLAdd	titian.x.x
18154	Complete	02/10/11 07:51:36 AM	jamsmad	24	1GZLAdd	titian.x.x
18153	Complete	02/10/11 07:12:33 AM	jamsmad	24	1GZLDel	titian.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	dxr	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	zwenhor	208	1EOG	129.174.x.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

Setting up options for folding

```
-in
  -file
    -native <native PDB file>          # native PDB file (optional)
    -fasta <primary sequence in FASTA format> # or command line: -in:file:fasta
    -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)
  -broker
    -setup <topology broker file>        # topology broker file
-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 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
    -use_filters true      # 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
  -relax
    -fast                  # Type of relax protocol. This has been shown to be the best deal for speed and
robustness.
```

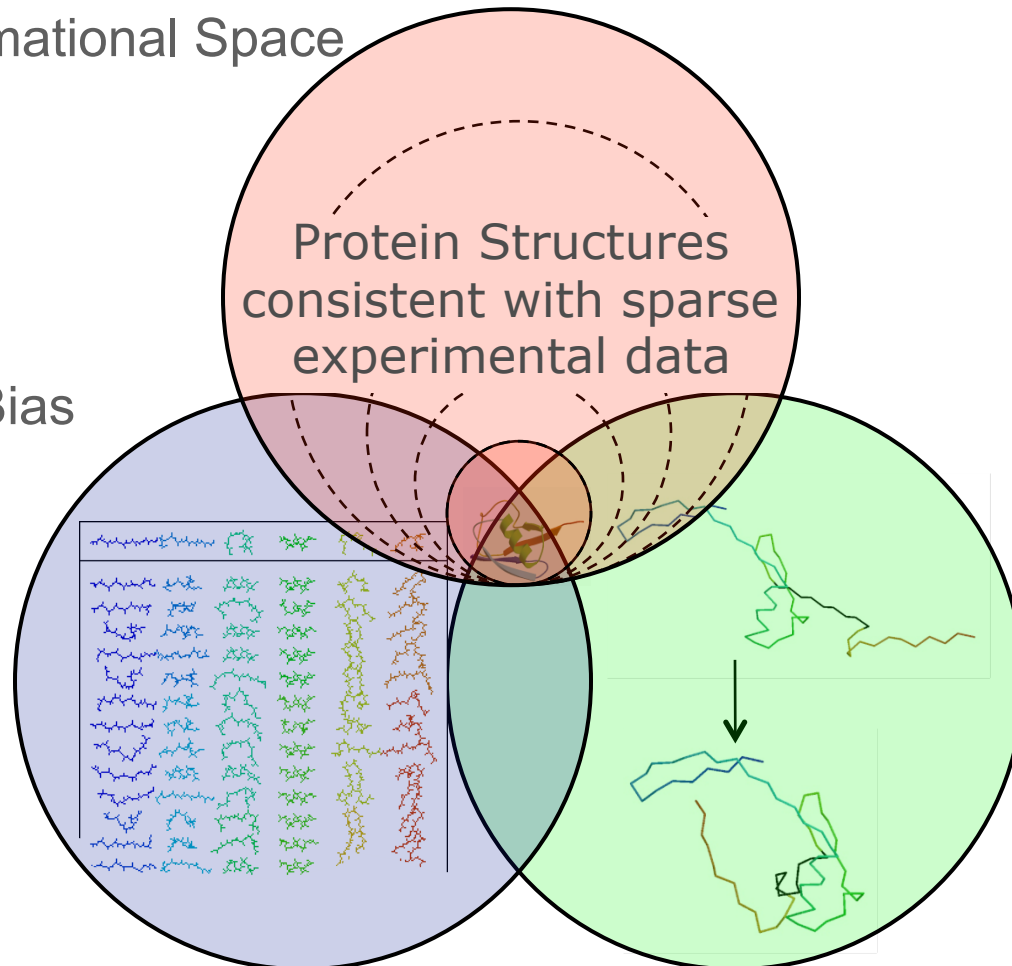
Setting up options for folding cont.

-run	
-reinitialize_mover_for_each_job	# Job distributor generates fresh copy of its mover before each apply (once per job)
-protocol broker	# run the topology_broker protocol for de novo folding
-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
-out	
-output	# use this to tell Rosetta you actually want output
-nstruct 1	# how many structures do you want to generate?
-file	
-silent <silent output file>	# full path to silent file output
-silent_struct_type binary	# we want binary silent files
-scorefile <scorefile>	# full path to scorefile
-overwrite	# overwrite any existing output with the same name you may have generated

Why combine experimental restraints with Rosetta?

- Complete Conformational Space

- Local Sequence Bias



- Energy Evaluation of non-Local Interactions

Adding experimental restraints

```
-fold_cst                # use FoldConstraints protocol
    -force_minimize      # minimize in FoldConstraints protocol
-constraints
    -cst_file <file>     # path to your cst file
    -cst_weight 4        # factor by which to multiply cst score
    -cst_fa_file <file>  # 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

Formatting 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

Initiating *de novo* folding

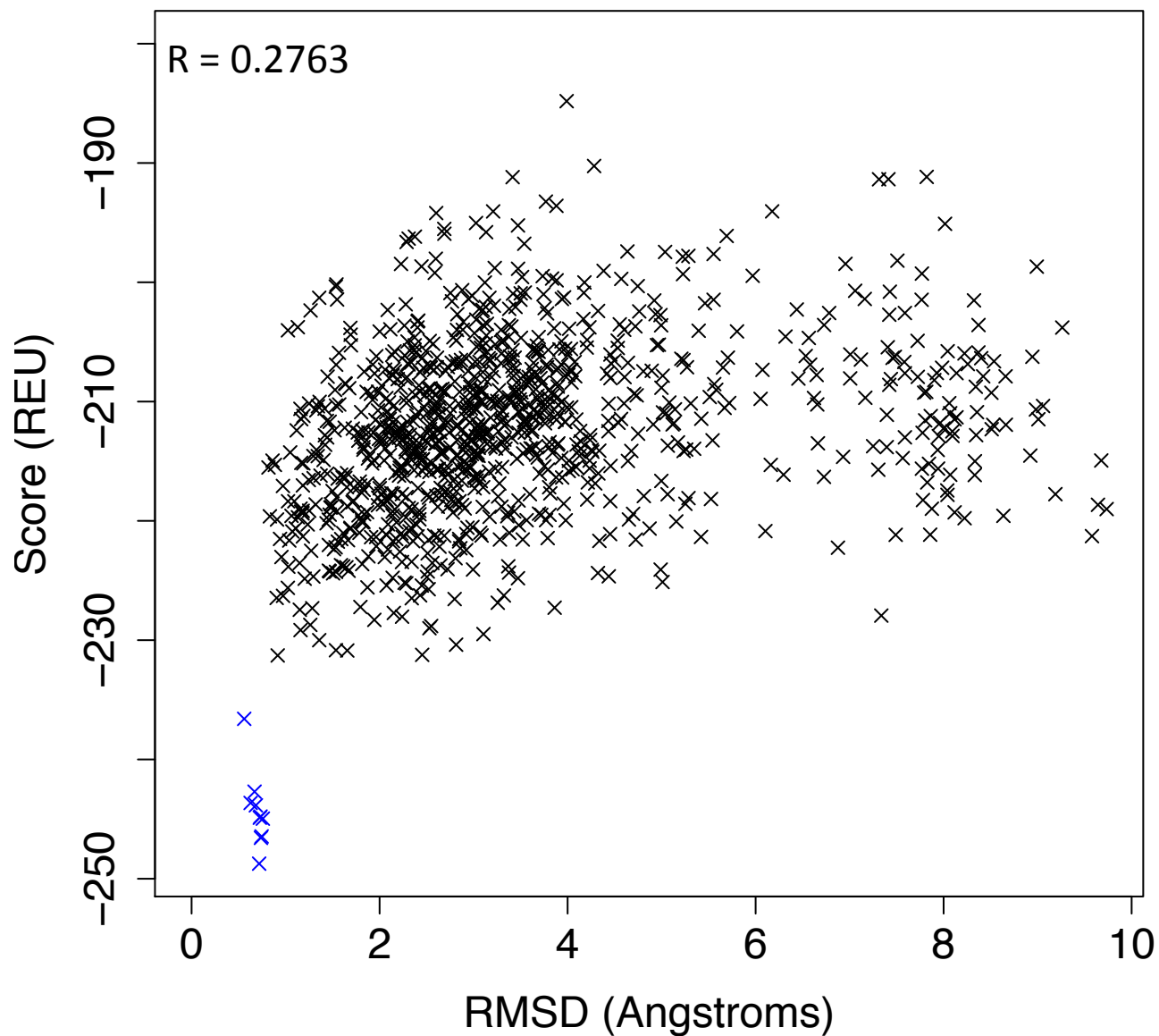
```
minirosetta.default.linuxgccrelease  
@2LZM_broker.options -database  
path_to/Rosetta/database/
```

- MiniRosetta calls many protocols, including abinitio

Assessing model quality: score vs. RMSD

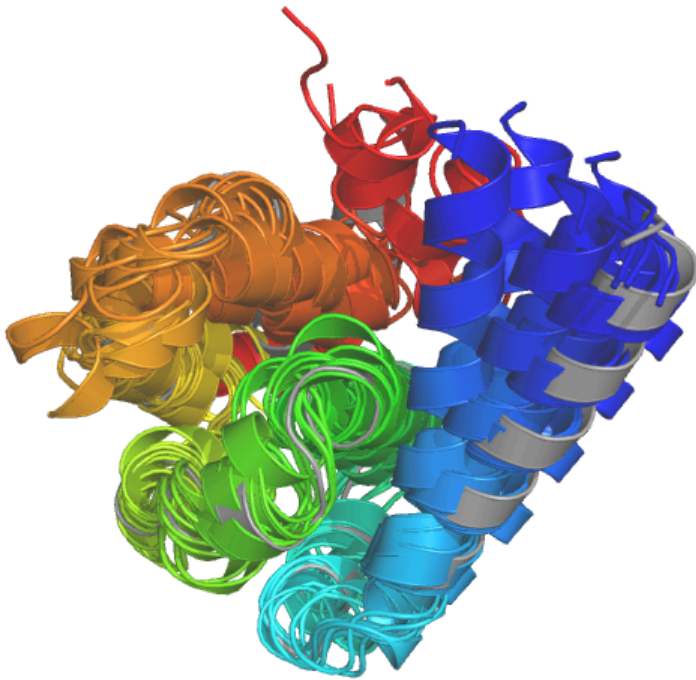
- Determine how well Rosetta energy correlates with model quality (RMSD, MaxSub, etc.)
- However this requires a native structure (e.g., crystal structure) or a homolog to which you'd like to compare the structure

2LZM_score_vs_rmsd_to_native

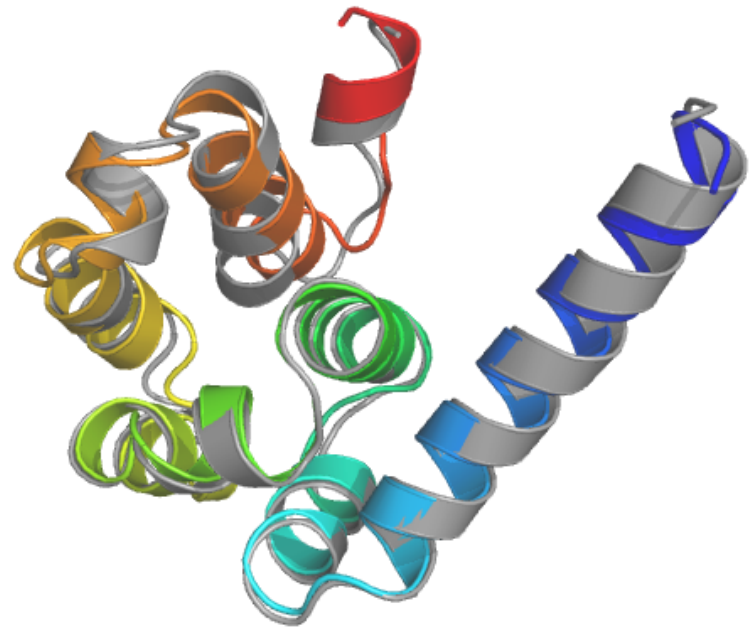


Assessing model quality: PyMol

Run: `pymol *.pdb`



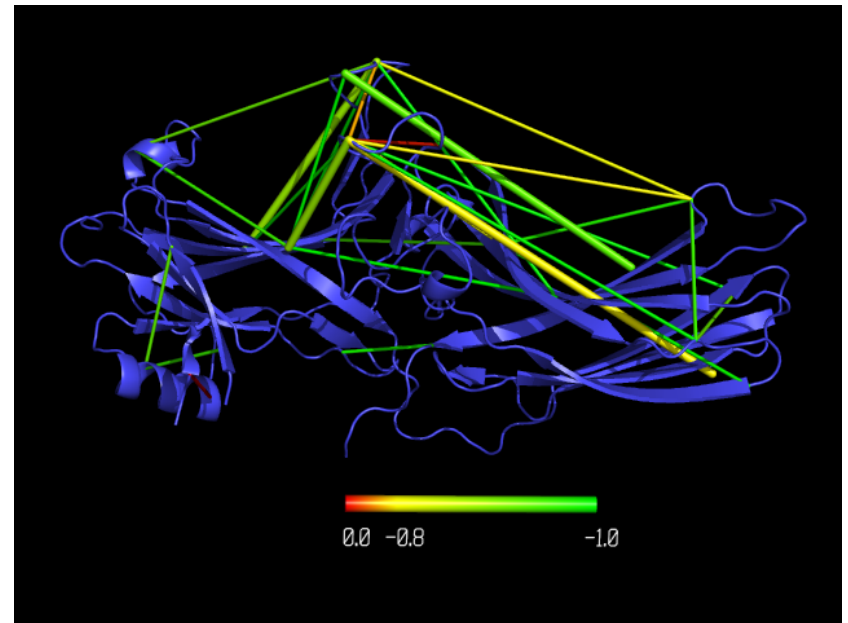
Top 10 Scoring



Best Scoring

Assessing model quality: restraints

- Filter by constraint score to look at models that satisfy experimental data
- Plot constraint score vs. RMSD, total score vs. constraint score, etc. to identify correlation of constraint score with total energy of model
- Identify constraint violations in your model, how big the violations are, etc.



Conclusions

- Rosetta *de novo* folding performs best with small proteins (80-100 residues)
- Folding larger, more complex proteins requires more information
- Experimentally derived knowledge can be provided in the form of restraints to improve model quality

Additional resources

- **Rosetta User's Guide:**
 - http://www.rosettacommons.org/manuals/archive/rosetta3.5_user_guide/
- **Protein Structure Prediction Using Rosetta**
 - Rohl, Strauss, Misura and Baker, *Methods Enzymol.*, 2004
- **Using restraints in Rosetta**
 - Rohl, *Methods Enzymol.*, 2005
 - Raman *et al.*, *Science*, 2010
 - Hirst *et al.*, *J. Struct. Biol.*, 2011.
- **Constraint File Instructions**
 - https://www.rosettacommons.org/manuals/archive/rosetta3.5_user_guide/de/d50/constraint_file.html

Today's Tutorial

- Bacteriophage T4 lysozyme
 - Small, soluble, globular protein
 - PDBID: 2LZM
- Tutorial 1 is a *de novo* folding benchmark experiment
 - *Reminder: If the Robetta queue is long, copy the fragment files from the input_files directory*
- Tutorial 2 adds experimental restraints