

De novo Protein folding

Rosetta Workshop

November 2018

Hope Woods

hope.woods@vanderbilt.edu

Meiler Laboratory, Vanderbilt University

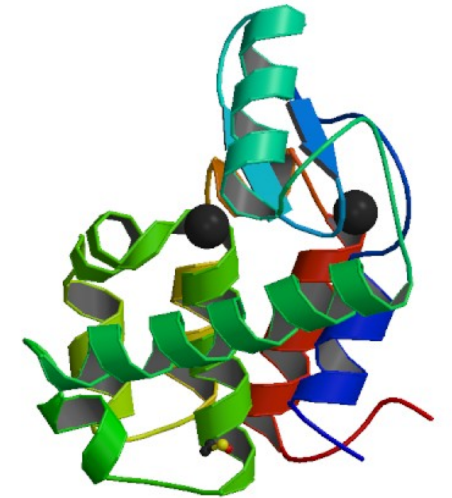
<http://www.meilerlab.org>

De novo Protein folding

- Determination of protein structure from sequence data alone
- No template structures, unlike homology modelling
- Useful for small, simple proteins (<200 residues)

MNIFEMLRIDEGLRLKIYK
DTEGYTIGIGHLLTKSPS
LNASKSELDKAIGRNTNGV
ITKDEAEKLFNQDVDAVR
GILRNAKLKPVYDSLDAVR
RAALINMVFQMGETGVAGF
TNSLRMLQQKRWDEAAVNL
AKSRWYNQTPNRAKRVITT
FRTGTWDAYKNL

Primary Sequence



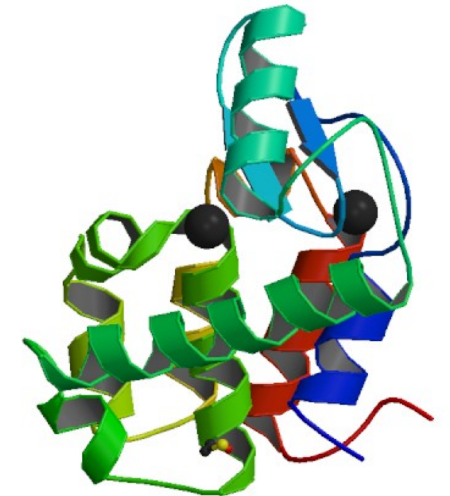
Tertiary Structure

De novo Protein folding

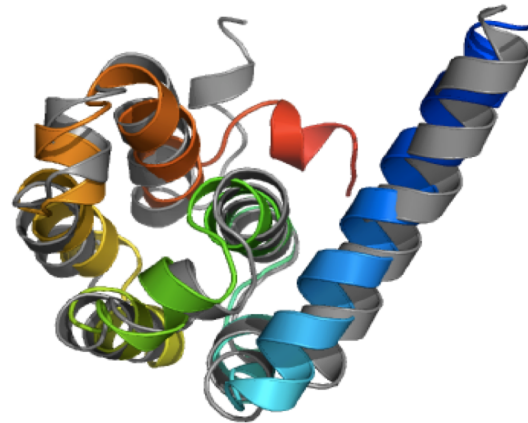
- Determination of protein structure from sequence data alone
- No template structures, unlike homology modelling
- Useful for small, simple proteins (<200 residues)

MNIFEMLRIDEGLRLKIYK
DTEGYTYIGIGHLLTKSPS
LNASKSELDKAIGRNTNGV
ITKDEAEKLFNQDVDAVR
GILRNAKLKPVYDSLDAVR
RAALINMVFQMGETGVAGF
TNSLRMLQQKRWDEAAVNL
AKSRWYNQTPNRAKRVITT
FRTGTWDAYKNL

Primary Sequence

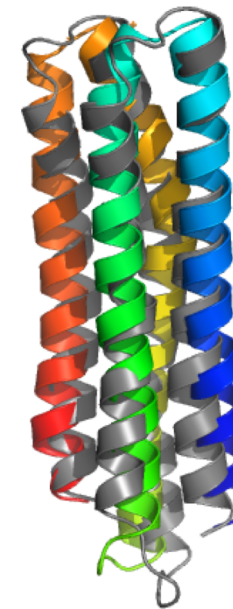


Tertiary Structure



Small, globular,
soluble proteins

T4-lysozyme
C-terminal domain



Small, simple
membrane proteins

V-type Na⁺ ATP
synthase subunit



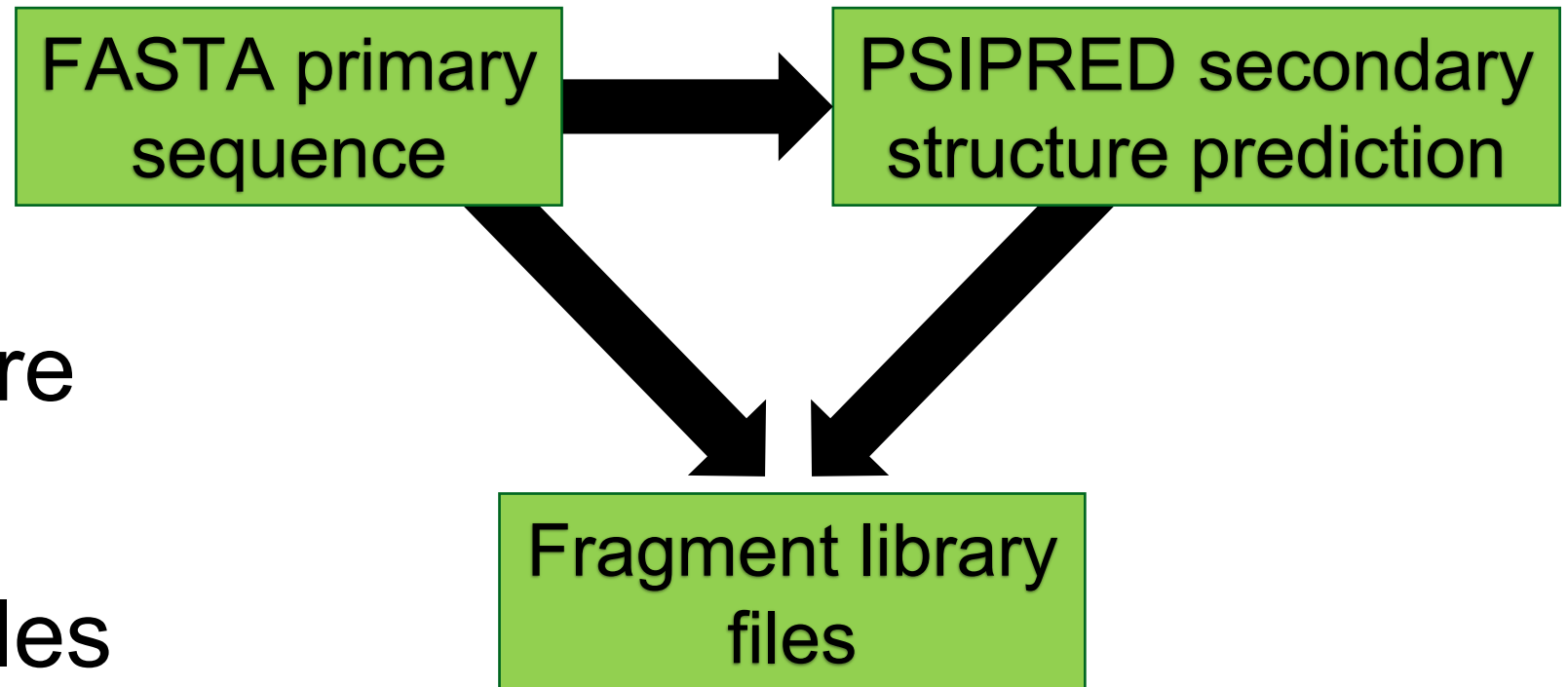
...but not large,
complex proteins

Rhodopsin

Inputs

- **Necessary:**

- FASTA primary sequence
- Secondary structure prediction
- Fragment library files
- Options file

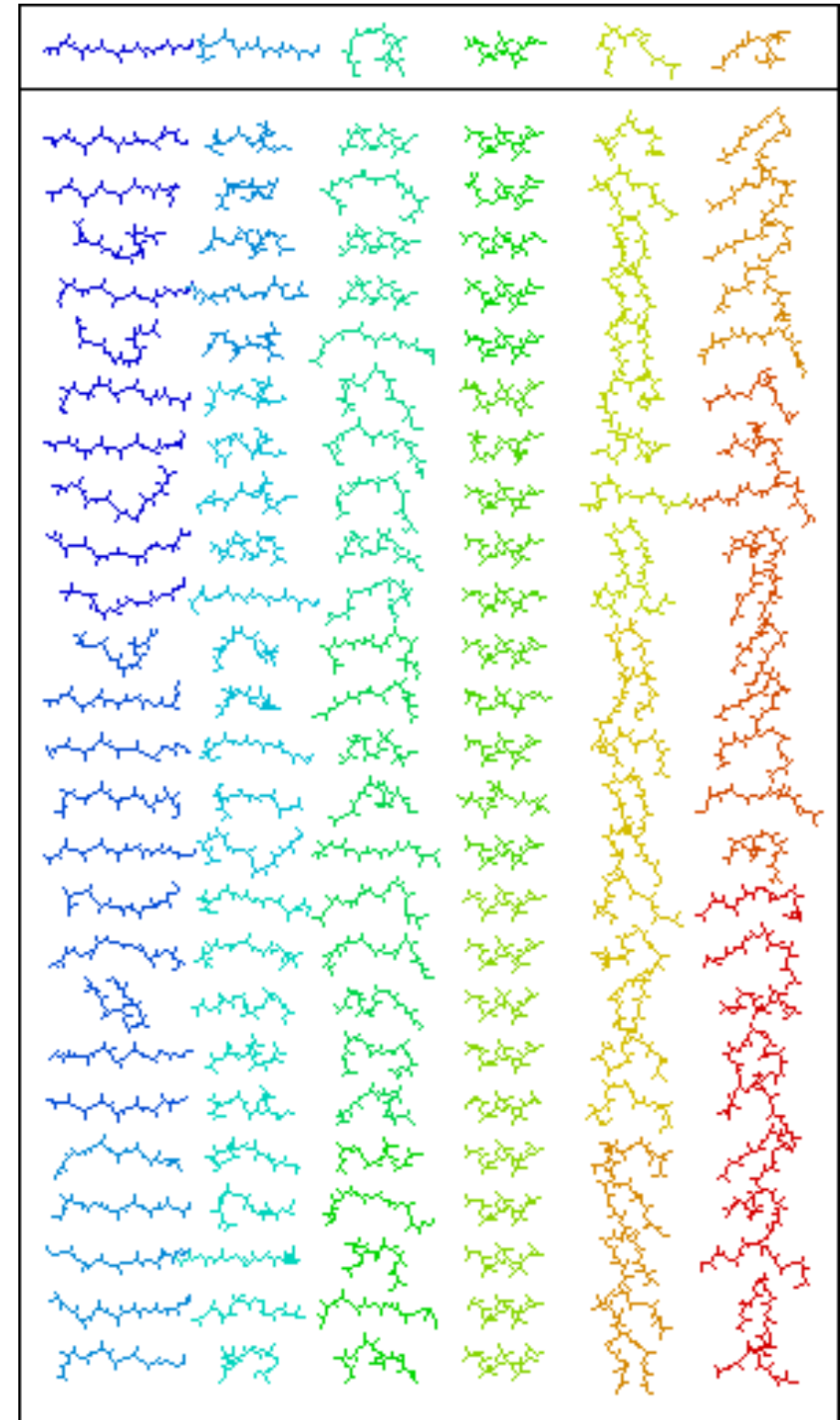


- **Optional:**

- Native structure (PDB)
- Constraints file

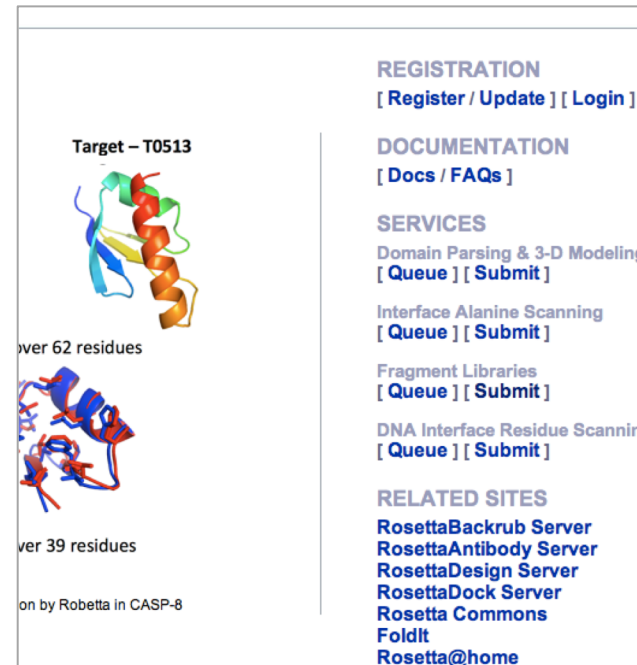
Inputs: fragment library files

- Peptide sequences generated from PDB: either 3 or 9 residues long
- Introduced to backbone to change geometry of protein
- High-quality fragments identified prior to application
- Can be selected using NMR data



Getting fragments: Robetta

- Online server for fragment selection, two simultaneous jobs per user
- Academic users only
- Alternative: perl script that is environment-specific (make_fragments.pl, discussed in the tutorial)



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

The screenshot shows the Robetta homepage with a navigation menu on the right including: 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]), and RELATED SITES (RosettaBackrub Server, RosettaAntibody Server, RosettaDesign Server, RosettaDock Server, Rosetta Commons, Foldit, Rosetta@home). On the left, there are two protein structure visualizations: one labeled 'Target - T0513' and another labeled 'over 62 residues'.

ID	Status	Date (PST)	Username	Length	Target	Host
18162	Complete	02/10/11 10:48:48 AM	vij4	226	anceu	dhcp128038156
18161	Complete	02/10/11 10:44:02 AM	jamsmad	26	1GZLIEnd	titan
18160	Complete	02/10/11 10:14:27 AM	jamsmad	22	1GZLShort	titan
18159	Complete	02/10/11 09:36:01 AM	zwenhor	38	2I2V4	Farid-HP.
18158	Complete	02/10/11 09:15:17 AM	zwenhor	41	1K1V	Farid-HP.
18157	Complete	02/10/11 09:11:09 AM	zwenhor	41	1K1V	Farid-HP.
18156	Complete	02/10/11 08:31:07 AM	jamsmad	26	1GZLFront	titan
18155	Complete	02/10/11 08:30:07 AM	jamsmad	26	1GZLAdd	titan
18154	Complete	02/10/11 07:51:38 AM	jamsmad	24	1GZLAdd	titan
18153	Complete	02/10/11 07:12:33 AM	jamsmad	24	1GZLDel	titan
18152	Complete	02/10/11 04:32:04 AM	Only Dym	441	PAN	wisweb2-out.v
18151	Complete	02/09/11 08:03:47 PM	manuti	56	GB1	142.1
18150	Complete	02/09/11 09:27:59 AM	drx	176	f9	128.2
18149	Complete	02/09/11 08:35:47 AM	gise	126	Nav beta-2 extra	139.1
18148	Complete	02/09/11 08:33:55 AM	zwenhor	208	1EOG	129.1

The screenshot shows the Robetta BETA submission form. It includes fields for 'Registered Username' or 'Registered Email Address' (with example 'stephanie.j.hirst@vanderbi'), 'Target Name' (with example '2LZM_'), and a 'Paste Fasta' field containing a protein sequence. There are also 'Upload Fasta' and 'Submit' buttons. Below the main form, there are optional fields for 'Identifier' (with example '2LZM_'), 'Exclude Homologues' (checkbox), and 'Rosetta NMR' (click links below for input format) with sub-fields for 'Chemical Shifts', 'NOE Constraints', and 'Dipolar Constraints'.

Options for AbInitioRelax

options.txt (1/2)

```
-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      # Discontinues refinement if a filter fails

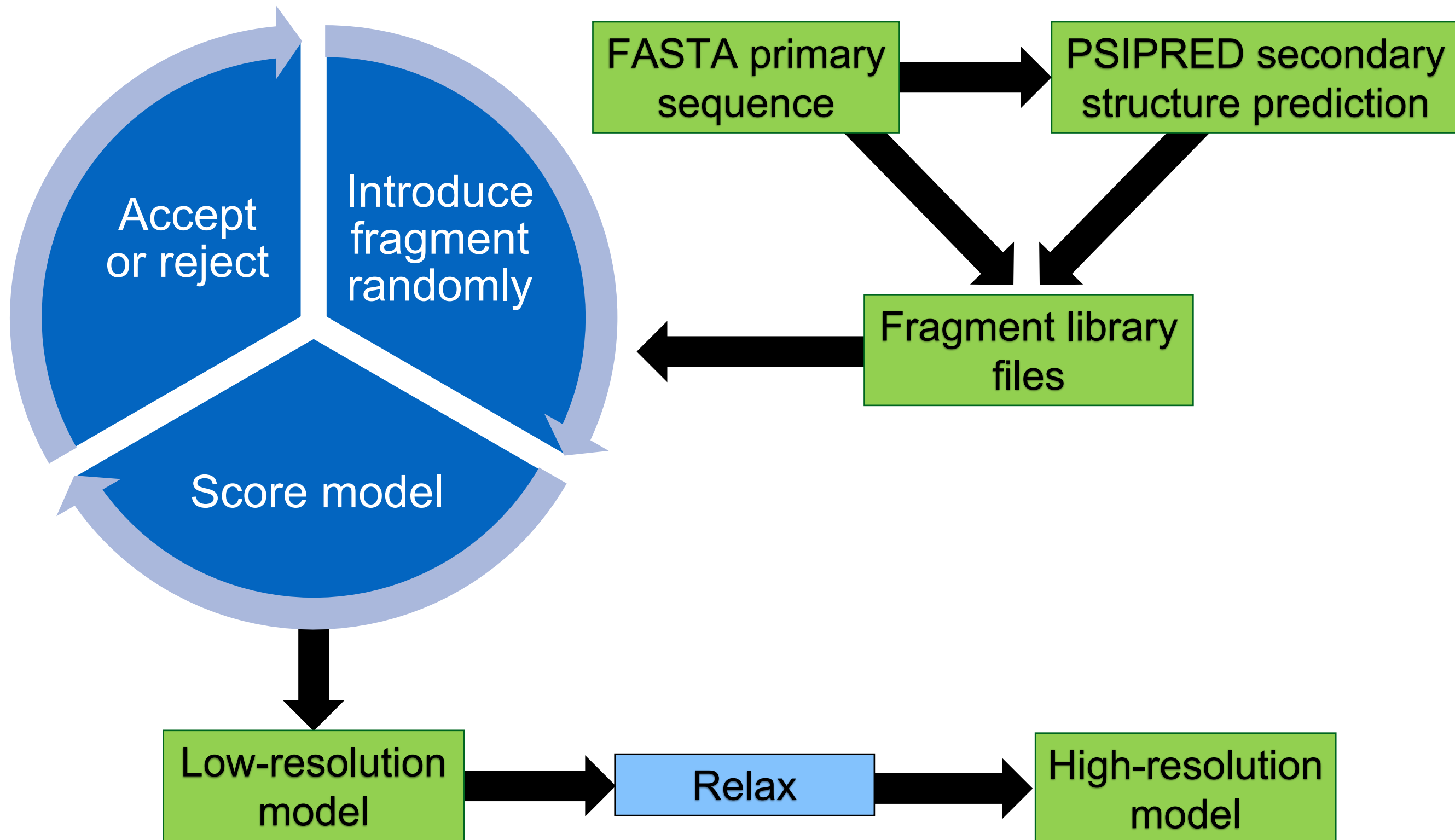
-relax
  -fast                  # Type of relax protocol. This has been shown to be the best deal for speed and
                        robustness.
```

Options for AbInitioRelax

options.txt (2/2)

```
-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_breaker 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
```

Running AbInitioRelax



Adding experimental data

options_cst.txt

```
-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

- Constraint type (AtomPair, Angle, etc.)
- Function type (Gaussian, spline, etc.)

Adding experimental data

GREMLIN.BAKERLAB.org

Powered by OPENSEQ.org

Home

GREMLIN

Complexes

PFAM

New Projects

Contact Us

Submit Sequence for Coevolution Analysis

Note, due to limited resources your submission may take forever to complete (Jobs Running: 47).



Email

(Will only be used to send you notice of Completion/Error).

Job name

Starting alignment

Alignment must be in FASTA format. Alternatively, you may provide a single sequence in plain text or in FASTA format.

No file chosen

NOTE: We treat your input as a "starting alignment" which is then enriched using HHblits. If Jackhmmer is selected only the first (or query) sequence is used. [20 < query_length < 1000] [contact us](#) if you wish to submit a sequence larger than the limit.

NEW: We have a separate [submission page](#) for protein-protein coevolution analysis.

Options

Generate MSA using: **HHblits**

Filter MSA

GREMLIN

E-value: **1E-10**

Iterations: **4**

Coverage: **75**

Remove gaps: **75**

Prior: **Vanilla**

• HHblits/Jackhmmer

- HHblits and Jackhmmer are methods used to generate the Multiple Sequence Alignment (MSA).
- <E-value/Iterations> options are used to control the diversity of MSA generation.
- **To use your own alignment**, without MSA enrichment, set Iterations to 0.

• Filter MSA

- <Coverage> Filter MSA to remove sequences that don't cover at least (%) of query.
- <Remove gaps> After coverage filter, remove positions in the alignment that have (%) gaps.

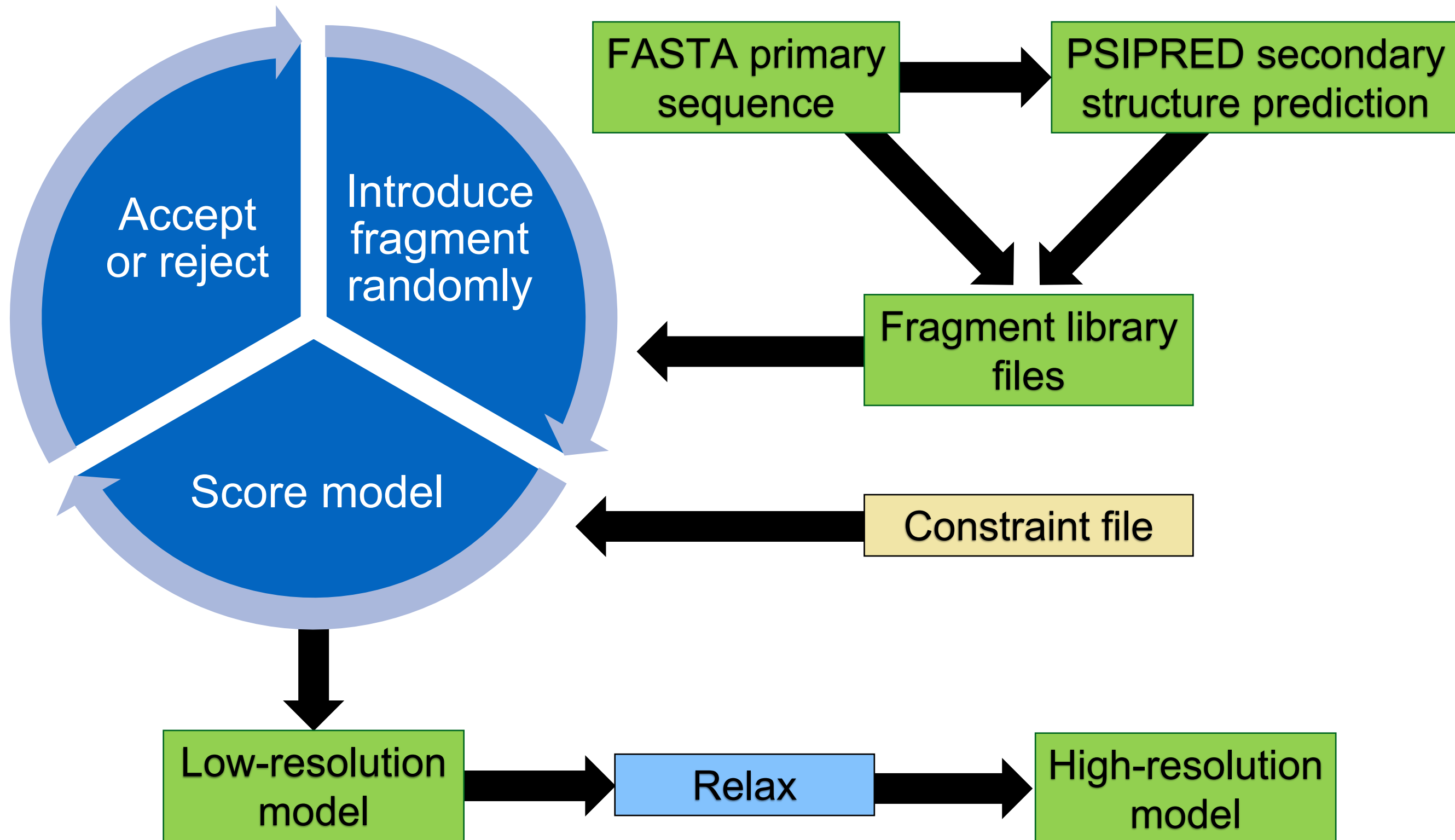
• GREMLIN

- only **Vanilla** works right now, other options are under construction.
- Vanilla - Pure sequence based analysis (no prior).
- SS - Add Secondary Structure prior, based on PSIPRED predictions.
- SVM - Add SVM-CON prior.

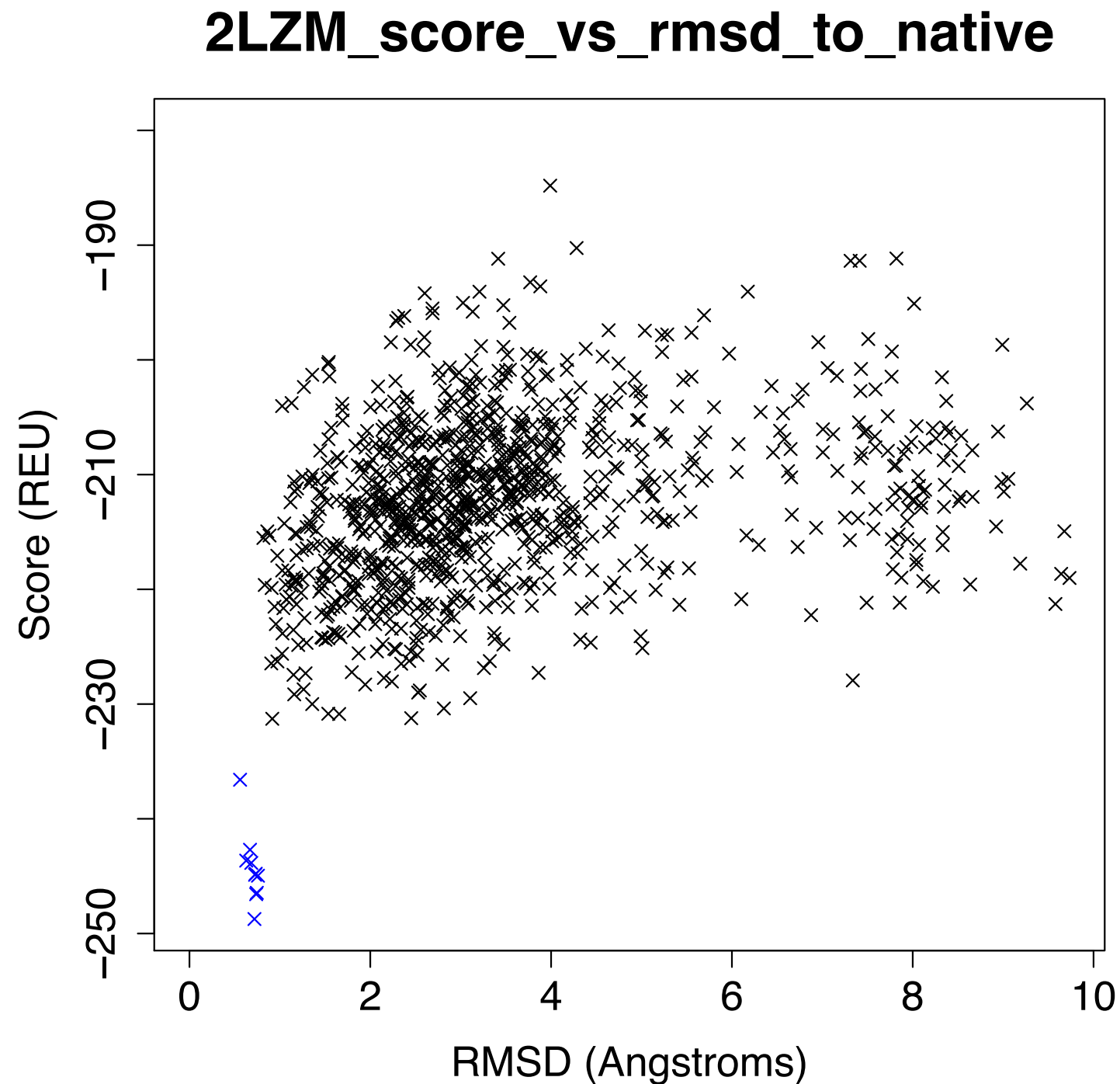
1 + 2 =

☒ List results for others to see.

Running AbInitioRelax

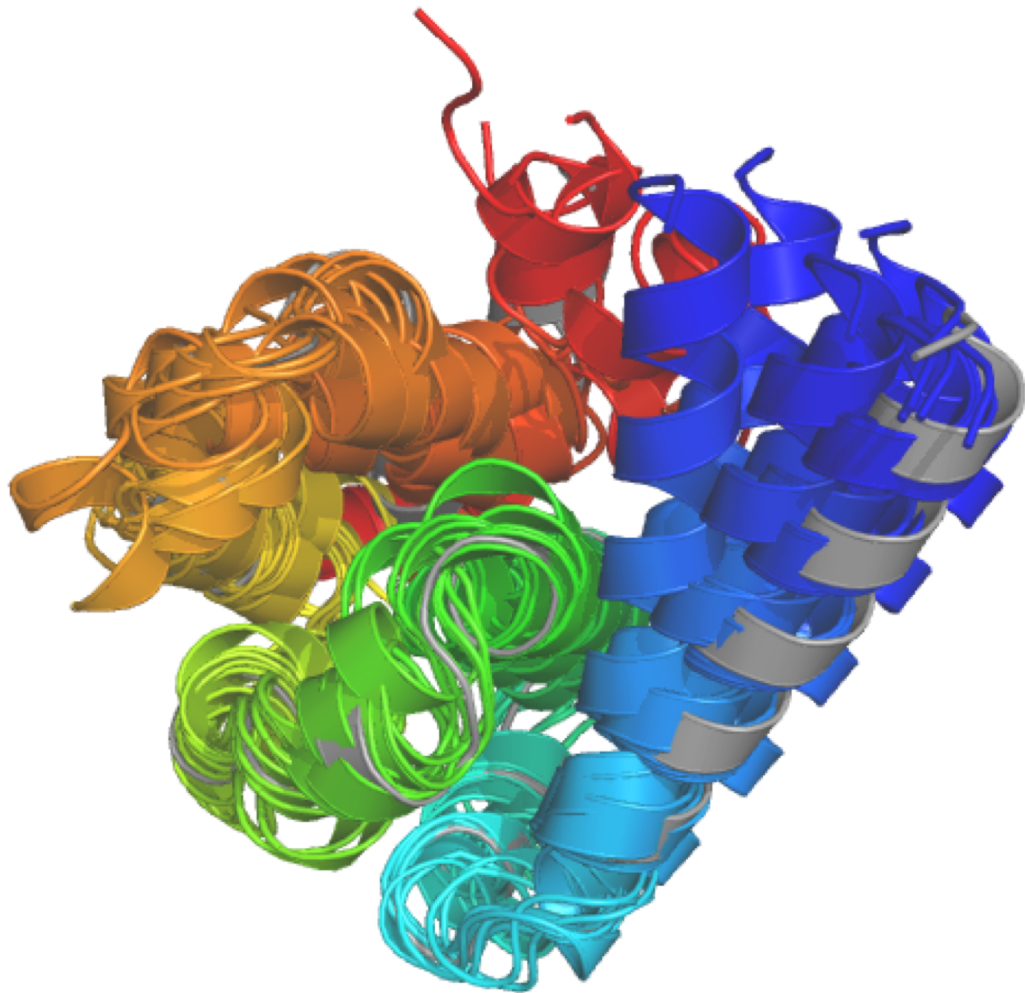


Evaluating model quality

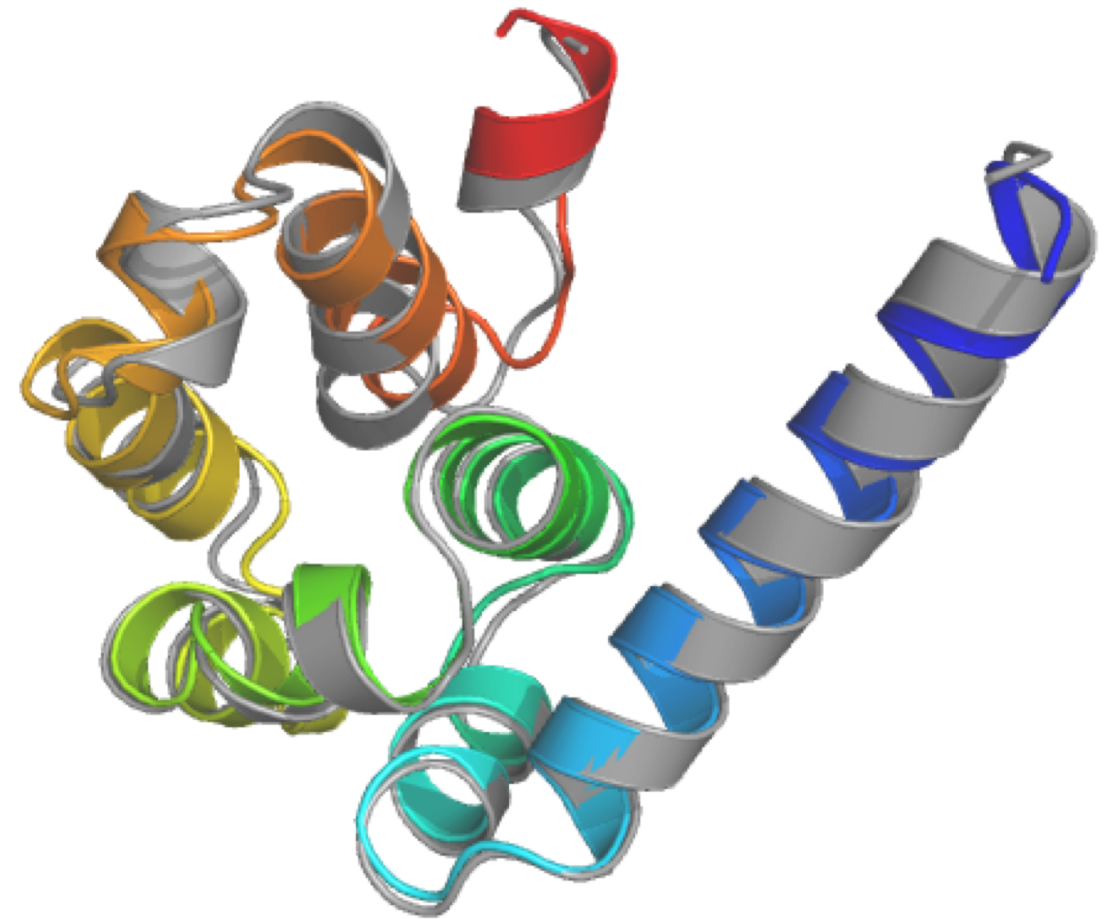


Evaluating model quality

Run: `pymol *.pdb`



Top 10 Scoring



Best Scoring

Conclusions and resources

- Rosetta AbInitio useful for small protein structure prediction
- Experimental data can improve prediction outcomes
- Additional resources:

