

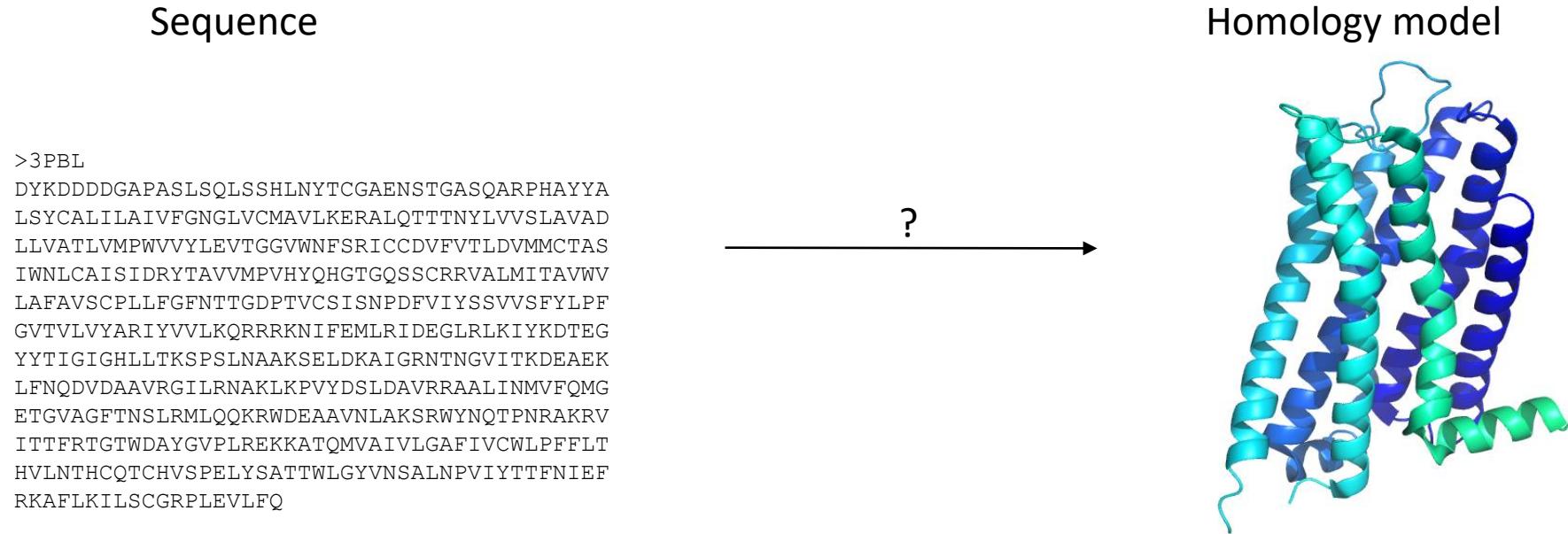
RosettaCM: multi-template comparative modeling



Eli McDonald
Meiler Lab
E-Mail: eli.f.mcdonald@vanderbilt.edu

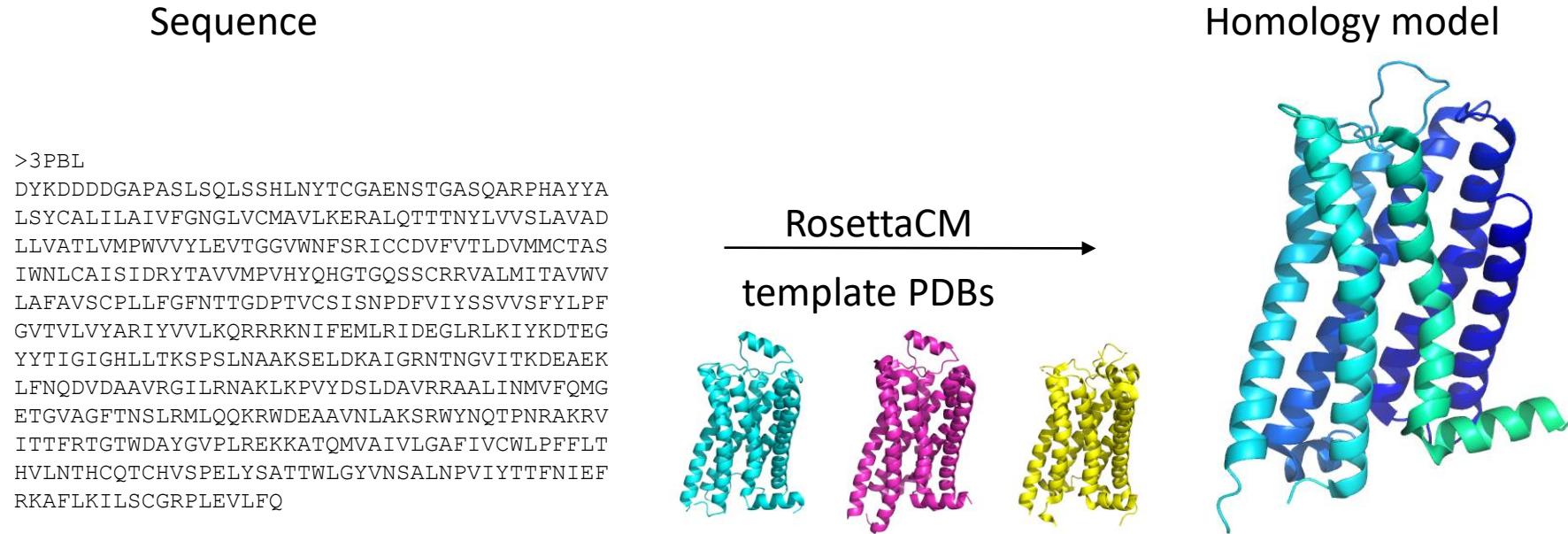
Why comparative modeling with RosettaCM?

- Deep learning-based methods offer less flexibility – you get what you get and you don't throw a fit
- CM is compatible with NMR, EPR, and electron Density constraints data
- Modeling mutations better with biophysical techniques than MSA based techniques



Why comparative modeling with RosettaCM?

- Deep learning-based methods offer less flexibility – you get what you get and you don't throw a fit
- CM is compatible with NMR, EPR, and electron Density constraints data
- Modeling mutations better with biophysical techniques than MSA based techniques



Single template versus multiple template modeling

- Single Template Modeling:
 - Single template as input
 - Uses sequence and template derived fragments
 - Used when available templates have very high identity (>60%)
- Multiple Template Modeling:
 - Multiple templates as input
 - Combine sections of multiple threaded models and sequence derived fragments
 - Used when available templates have low identity (30-50%)
- *Nomenclature Note*
 - Comparative Modeling = Homology Modeling in the land of Rosetta

General workflow for RosettaCM

1. Identification template sequences
2. Preparation of sequence alignments
3. Threading
4. Hybridize
5. Relaxation
6. Scoring and Selection

In the tutorial: Comparative modeling of the Dopamine D3 receptor

Target sequence of dopamine D3 receptor

(PDB: Dopamine D3 receptor 3pbl)

Find this file at */rosetta_cm/demo/input_files/3pbl.fasta*

>3pbl

```
YALSYCALILAIVFGNGLVCM AVLKERALQTTNYLVVSLAVADLLVATL VMPWVVYLEVTGGVWNFSRICCDVF  
VTLDVMMCTASIWNLC AISIDRYTAVVMPVHYQHGTGQSSCRRVALMITAVWVL AFAVSCPLLFGFNTTGDPTVC  
SISNPDFVIYSSVVSFYLPFGVTVLVYARIYVVLKQRRRKAAAAAAAAGVPLREKKATQMVAIVLGAFIVCWLPF  
FLTHVLNTHCQTCHVSP ELYSATTWLGYVNSALNPVIYTFNIEFRKAFLKILSC
```

The screenshot shows the NCBI Protein database search page. At the top, there's a navigation bar with links for 'NCBI', 'Resources', 'How To', 'My NCBI', and 'Sign In'. Below the navigation is a search bar with dropdown menus for 'Search' (set to 'Protein'), 'Limits', 'Advanced search', and 'Help'. A large search input field contains the sequence 'YALSYCALILAIVFGNGLVCM AVLKERALQTTNYLVVSLAVADLLVATL VMPWVVYLEVTGGVWNFSRICCDVF...'. To the right of the input field are 'Search' and 'Clear' buttons. Below the search bar, there's a large protein sequence visualization with various amino acid residues highlighted in different colors. To the right of the sequence, a dark blue sidebar contains the word 'Protein' and a descriptive text about the database.

Protein
Translations of Life

Search: Protein Limits Advanced search Help

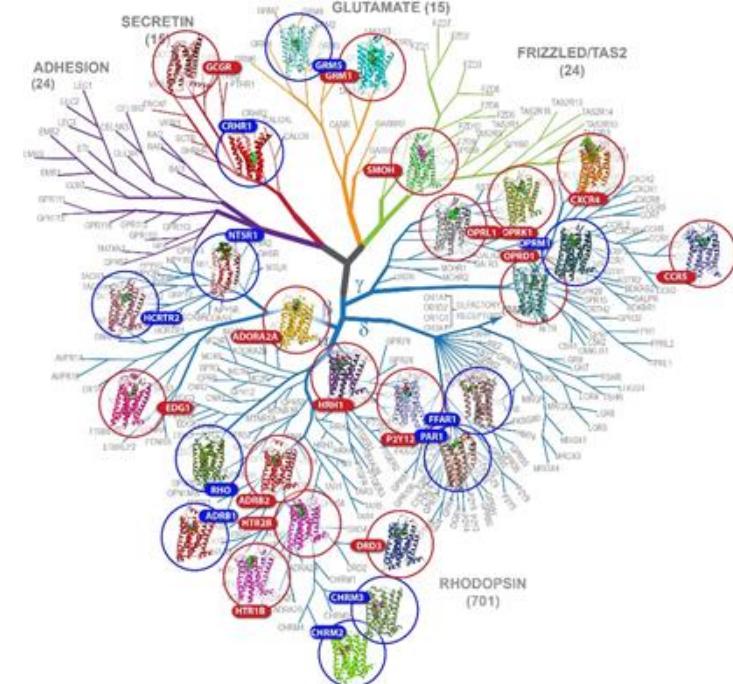
YALSYCALILAIVFGNGLVCM AVLKERALQTTNYLVVSLAVADLLVATL VMPWVVYLEVTGGVWNFSRICCDVF...
Search Clear

The Protein database is a collection of sequences from several sources, including translations from annotated coding regions in GenBank, RefSeq and TPA, as well as records from SwissProt, PIR, PRF, and PDB. Protein sequences are the fundamental determinants of biological structure and function.

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

Template identification for dopamine D3 receptor

- PDB ID: 3pb1
- Class A G-protein coupled receptor (GPCR)
- No high identity templates
- 7 transmembrane helices
- 3 extracellular loops, 3 intracellular loops
- Highly conserved GPCR residues



GPCR phylogenetic tree with crystal structures (2014). Taken from <https://katritch.usc.edu/research.html>

Template identification for dopamine D3 receptor

- Similarity of Sequences :
 - compare proteins based on amino acid sequences (BLASTP using PDB as search database)
 - suitable templates have ideally >30% sequence identity to the target
- Fold Recognition:
 - using predicted secondary structure information to detect proteins with similar 3D characteristics (DALI, PHYRE)

Template identification for Dopamine D3 receptor

- It is advisable to use multiple templates due to the low sequence identity in available templates

Template	PDB ID	% Seq id
β2-adrenoceptor	3SN6	36
5-HT1B receptor	4IAR	32
β2-adrenoceptor	3D4S	34
5-HT2B receptor	5TVN	32
M1 receptor	5CXV	32
H1 receptor	3RZE	31
M4 receptor	5DSG	29
A2A receptor	2YDO	28
A1 receptor	5N2S	27

Template identification for dopamine D3 receptor

Human 5HT-1B receptor (PDB: 4iar)

Human beta1-adrenoceptor (PDB: 4bvn)

Human B2-adrenergic receptor (PDB: 2rh1)

Human M4 muscarinic acetylcholine receptor (PDB: 5dsg)

Human M1 muscarinic acetylcholine receptor (PDB: 5cxv)

Find these files at */rosetta_cm/template_pdbs/original_files/*

The screenshot shows the RCSB PDB homepage. At the top, there's a navigation bar with links for "Contact Us | Print", "PDB ID or Text" search, "Advanced Search", and "Customize This Page". The main header reads "A Resource for Studying Biological Macromolecules" and "An Information Portal to Biological Macromolecular Structures". A message indicates "As of Tuesday Feb 22, 2011 at 4 PM PST there are 71415 Structures" with links to "Search" and "PDB Statistics". On the left, there's a sidebar with sections for "MyPDB" (Login to your Account, Register a New Account), "Home" (News & Publications, Usage/Reference Policies, Deposition Policies, Website FAQ, Contact Us, About Us, Careers, External Links, Sitemap, New Website Features), "Deposition" (All Deposit Services, Electron Microscopy, X-ray | NMR, Validation Server, BioSync Beamline, Related Tools), and "Search" (Advanced Search). The central content area features "Featured Molecules" with a "Structural View of Biology" section showing a tree-like molecular structure and a "Molecule of the Month: Integrin" section with a detailed molecular model. To the right, there are sections for "New Features" (Transporter Classification Database Browser, Latest features released, Website Release Archive), "RCSB PDB News" (Weekly | Quarterly | Yearly), and "Structural Neighbors" (a visualization of structural neighbors). The bottom right corner displays the URL <http://www.rcsb.org>.

Multiple sequence alignment

CLUSTAL O(1.2.4) multiple sequence alignment

```
5cxv      -----KGPWQVAFIGITTGLLSLATVTGNLLVLISFKVNTELKTVNNYFLLSLACADL
5dsg      GPSSHNRYETVEMVFIATVTGSLSLTVVGNILVMLSIVNRLQTVNNYFLFSLACADL
3pbl      -----
4iar      YIYQDSISLPWKV-LLVMLLALITLATTLSNAFVIATVYRTRKLHTPANYLIASLAVAL
2rhl      -----
4bvn      -----DEVVVV-GMGIVMSLIVLAIIVFGNVLVITAIAKFERLQTVNYFITSLACADL
                   -----LSQQWEA-GMSLLMALVVLLIVAGNVLVIAAIGSTQRLQTLTNLFITSLACADL
```

Find this file at */demo/alignment_files/3pbl_alignments.txt*

The screenshot shows the Clustal Omega web interface. At the top, there's a navigation bar with links for 'Input form', 'Web services', 'Help & Documentation', 'Share', and 'Feedback'. Below the navigation, a breadcrumb trail indicates the current location: 'Tools > Multiple Sequence Alignment > Clustal Omega'. The main title 'Multiple Sequence Alignment' is displayed. A descriptive text explains that Clustal Omega is a new multiple sequence alignment program using seeded guide trees and HMM profile-profile techniques to generate alignments between three or more sequences. It also notes that for two sequences, the pairwise alignment tools should be used instead. There are three input methods: a text area for pasting sequences, a file upload button, and a browse button. Below this is a parameter section for 'OUTPUT FORMAT' set to 'Clustal W|GAP numbers'. A note states that default settings are suitable for most users. At the bottom, there's a 'STEP 3 - Submit your job' button.

<http://www.ebi.ac.uk/Tools/msa/clustalo/>

Adjusting multiple sequence alignments

- Experimental expectations:
 - Highly conserved residues
 - Secondary structure elements

Raw ClustalO alignment:

3pbl	-----YALSICALILAI	VPGNGIVCM	AVLK	ERALQT	TTN	LVVSLA	VADL			
5cxv	-----KGPWQVAFIGITG	LLSLATVTG	NLLVLISFKVN	TELKT	VNNYF	LLSLAC	ADL			
5dsg	GPSSHNR	YETVEMVFIATVTG	SLVTVVG	NILVMLS	IKVN	RQLQT	VNNYFL	FSLAC	ADL	
4iar	YIYQDSI	SLPWKV	-LLVMLLALITLATL	SNAFVIATVYR	TRKLHT	PANYLIASLA	VTD	DL		
2rh1	-----DEVWVV	-GMGIVMSL	IIVLAI	IVFGNV	LVITAIAK	FERLQT	VTNYF	ITS	SLAC	ADL
4bvn	-----LSQQWEA	-GMSL	LLMALV	VLLIVAGN	VNLVIAAIGST	QRLQT	TNL	FITS	SLAC	ADL

Adjusted alignment:

3pbl	-----YALSICALILAI	VPGNGIVCM	AVLK	-RALQT	TTN	LVVSLA	VADL			
5cxv	-----KGPWQVAFIGITG	L	L	SLATVTG	N	LLV	SLAC	ADL		
5dsg	GPSSHNR	YETVEMVFIATVTG	SLVTVVG	N	ILVMLS	IKVN	-RQLQT	VNNYFL	FSLAC	ADL
4iar	-YIYQDSI	SLPWKV	LLVMLLALITLATL	SNAFVIATVYR	-RKLHT	PANYLIASLA	VTD	DL		
2rh1	-----DEVWVV	-GMGIVMSL	IIVLAI	IVFGNV	LVITAIAK	-ERLQT	VTNYF	ITS	SLAC	ADL
4bvn	-----LSQQWEA	-GMSL	LLMALV	VLLIVAGN	VNLVIAAIGST	QRLQT	-TNL	FITS	SLAC	ADL

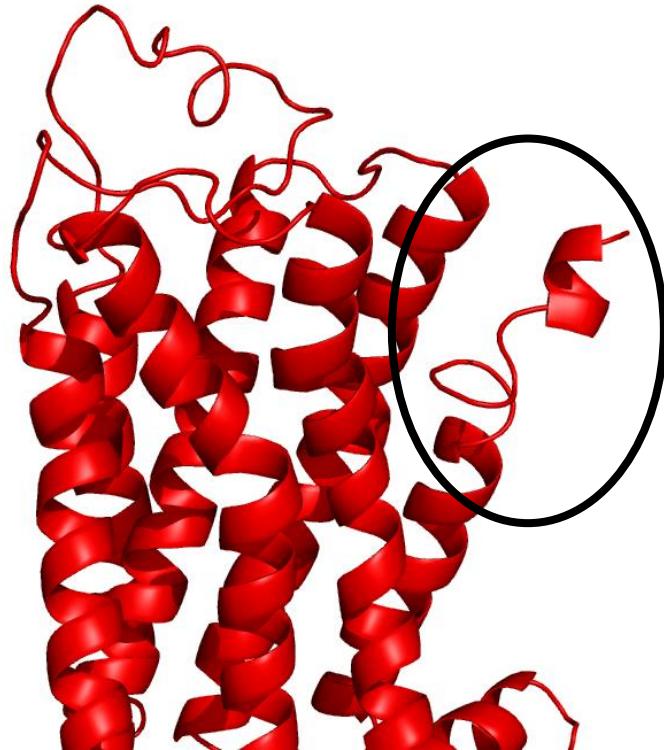
helix regions

highly conserved residues

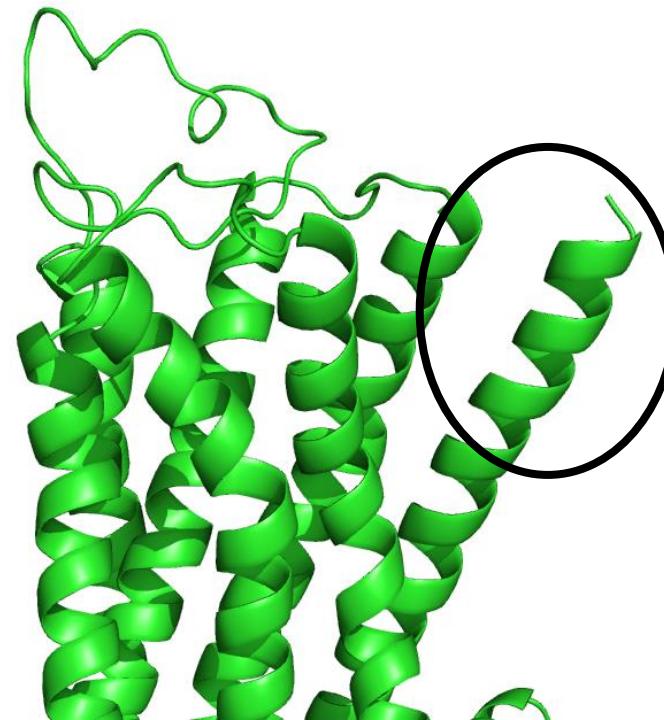
Alignment issues to be resolved

predicted membrane spanning region from OCTOPUS

Adjusted multiple sequence alignments result in improved modeling performance



Example model using
raw alignment

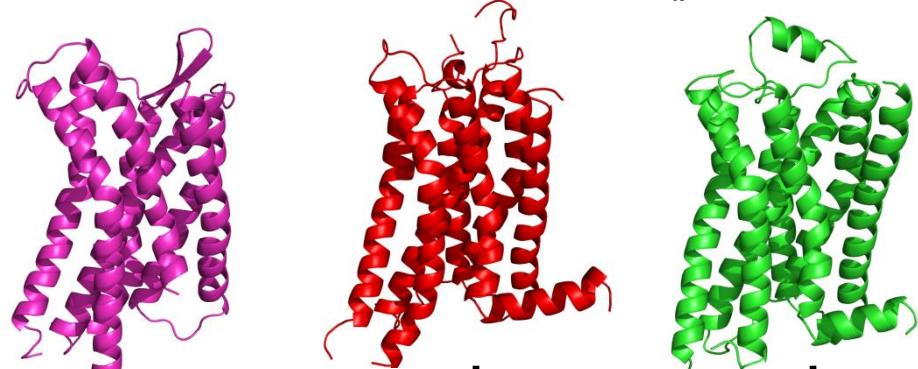


Example model using
adjusted alignment

Partial Thread

-----PWQFSM--LAAYMFLIMLGFPINFLTLVTVQHKKLRTPLNYILLNLAVADLFM
ANFNKIFL-----PTIYSIIFLTGIVGNGLVILVMGYQKKLRSMTDKYRLHLSVADLLF
---DEVVWVVGGMGIVMS---LIVLAIIVFGNVLVITAIAKFERLQTVTNYFITSACADLVM
-----IMGSSVYITVELATAVT.ATT.GNVT.VCWAVWT.NSNLQNVTNYFVVSLAAADIAV

alignment

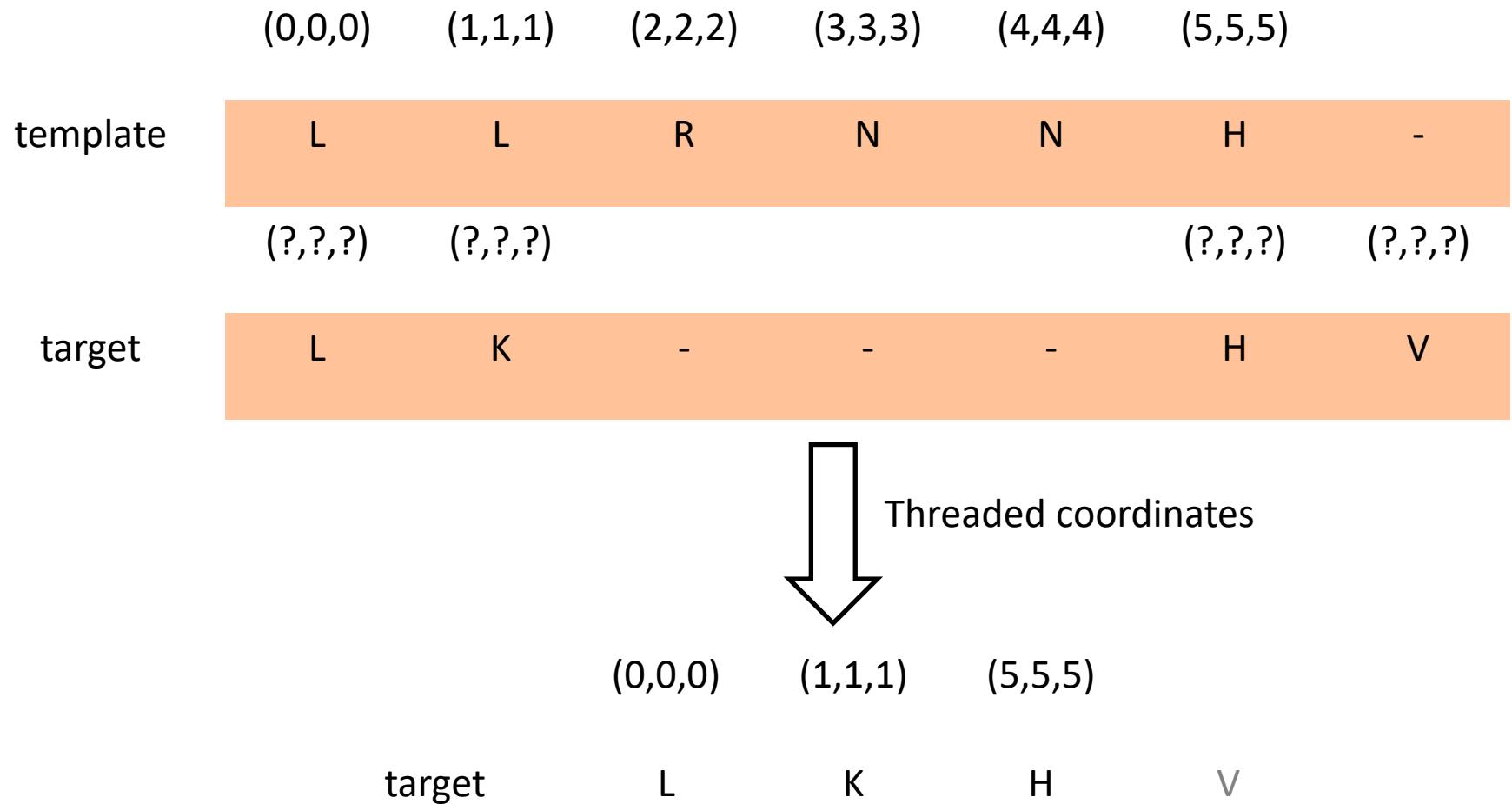


template pdbs



threaded pdbs

Partial Thread



Partial thread only excepts alignments in grishin format

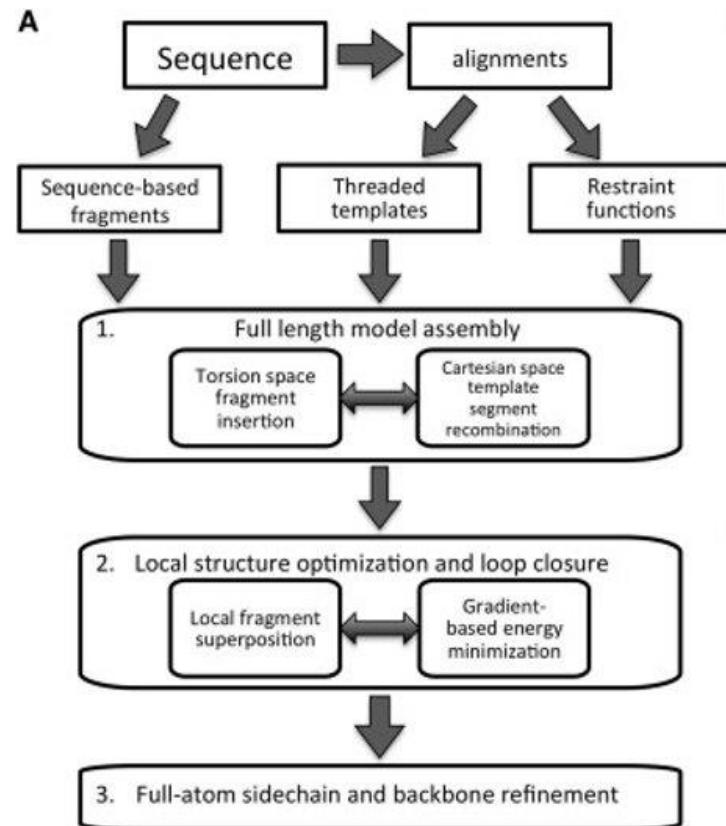
- ClustalO format:
 - All sequences in one file
 - Sequences broken up over several lines
- Grishin format:
 - One file per alignment pair
 - Sequences continuous over one line each
 - Contains header information
 - Due to complicated format, we have provided a script for conversion
`make_alignment_files.sh` for your use back home

Find converted Grishin alignment files at */rosetta_cm/demo/alignment_files/*

(2rh1.aln 4bvn.aln 4iar.aln 5c xv.aln 5dsg.aln)

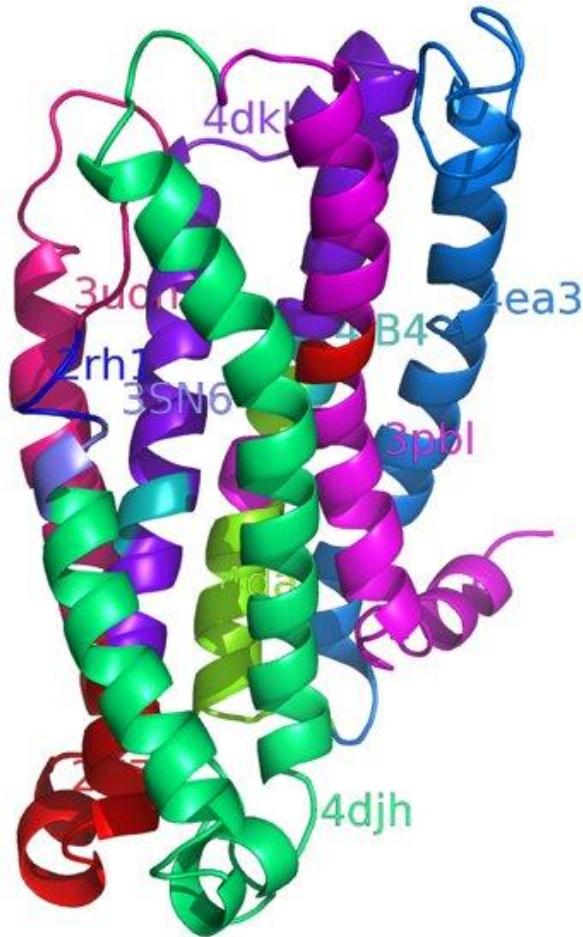
Hybridize protocol contains three stages

1. Generate initial models from template alignments
2. Explore deviations from templates and close loops in 2 steps :
 - MC: Randomly select de novo or template-based fragment and substitute into current conformation
 - Cartesian space full-backbone minimization
3. Full atom backbone and side chain refinement and final relax



Song, Y.; et al. *Structure*, 2013

Final models contain template information from multiple templates



Input files for RosettaCM

Bare minimum:

- Partial-threaded structures
- Mover definition and options

Specific to membrane proteins (not needed if modeling soluble proteins):

- Membrane spanning regions (span file)
- Membrane weight patches

Optional files based on available information:

- Constraint information (eg. atom pair connectivity)
- Disulfide Connectivity

Membrane spanning regions

Find this file at */rosetta_cm/demo/input_files/3pbl.span*

TOPCONS



Results

- Submitted: 2022-11-28 22:51:01 CET
- Status: **Finished**
- Wait time in the queue: 1 sec
- Running Time: 0 sec

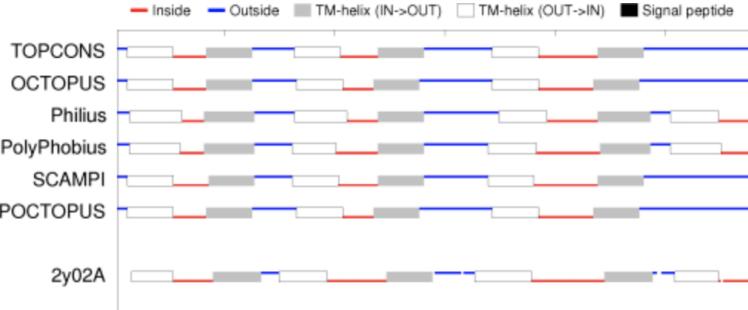
Results of your prediction with jobid: **rst_f926tcvb** (jobname: **3pbl**)

Zipped folder of your result can be found in [rst_f926tcvb.zip](#)

Dumped prediction in one text file can be found in [query.result.txt](#)

The sequence(s) you submitted can be found in [query.raw.fa](#)

Predicted topologies and predicted ΔG values:



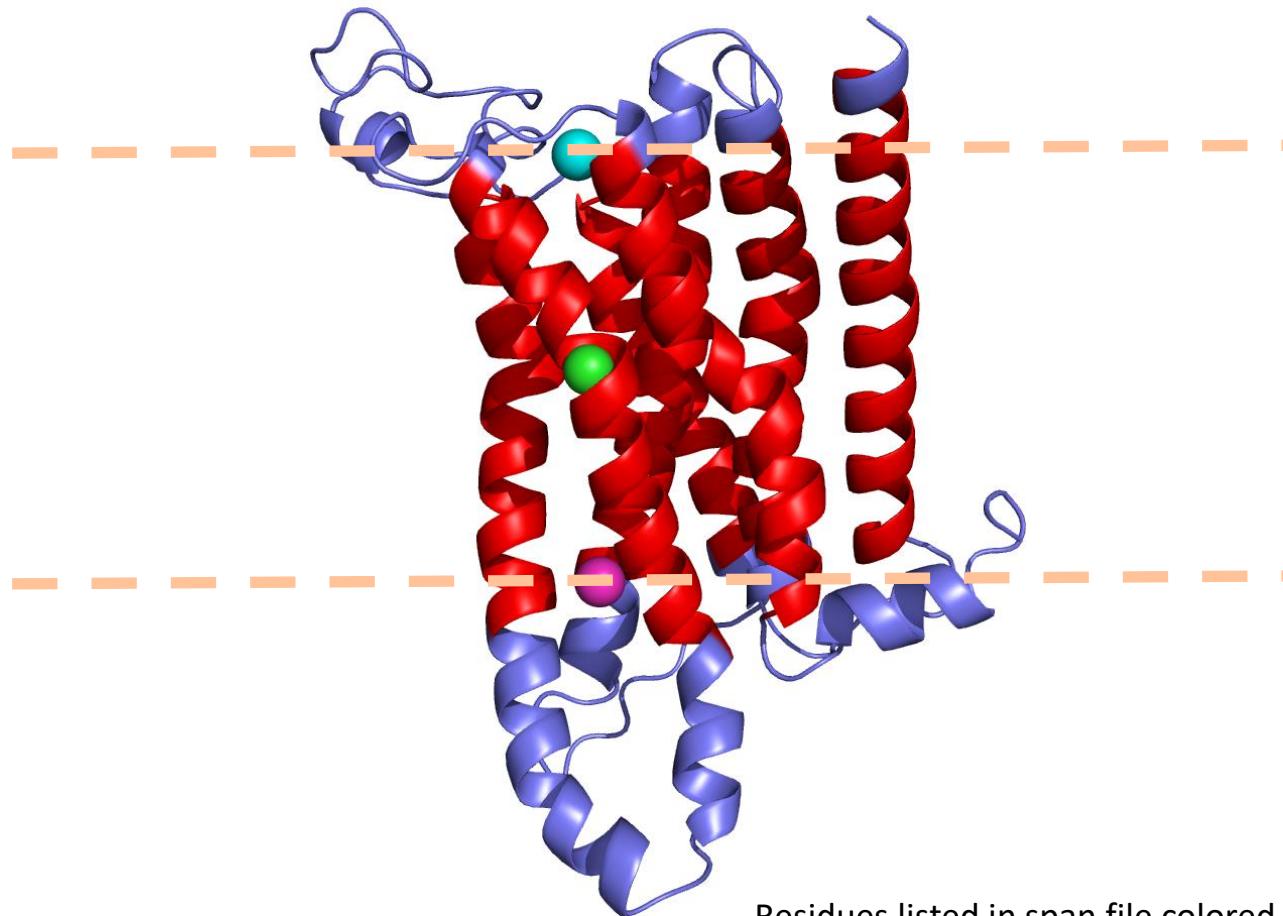
Legend: Inside (red), Outside (blue), TM-helix (IN->OUT) (grey), TM-helix (OUT->IN) (white), Signal peptide (black)

Method	Top Consensus	OCTOPUS	Philius	PolyPhobius	SCAMPI	SPOCTOPUS	2y02A
TOPCONS	Blue	Blue	Blue	Blue	Blue	Blue	Blue
OCTOPUS	Red	Grey	Red	Red	Red	Red	Red
Philius	Blue	Grey	Blue	Blue	Blue	Blue	Blue
PolyPhobius	Red	Grey	Red	Red	Red	Red	Red
SCAMPI	Blue	Blue	Blue	Blue	Blue	Blue	Blue
SPOCTOPUS	Blue	Grey	Blue	Blue	Blue	Blue	Blue
2y02A	Blue	Grey	Blue	Blue	Blue	Blue	Blue

<https://topcons.cbr.su.se/pred/>

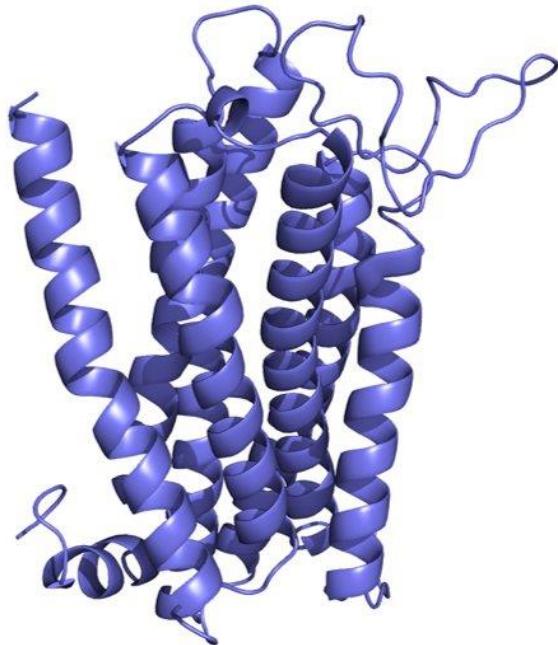
query.result.txt

The span file helps RosettaMembrane to define transmembrane regions

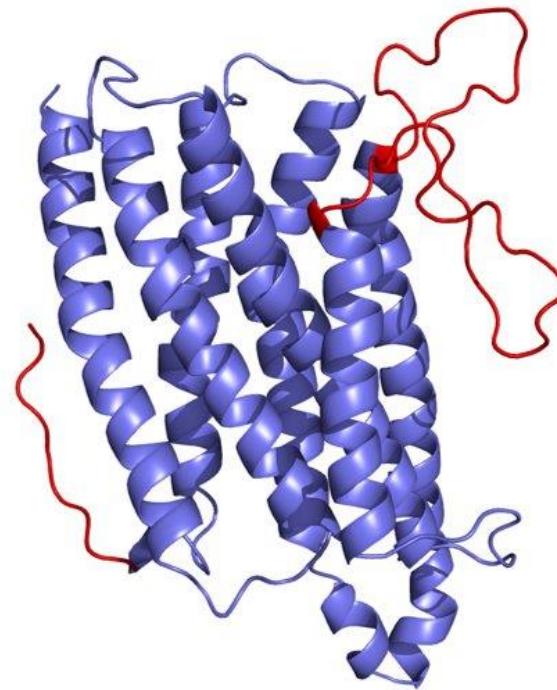


Why are membrane scoring terms important?

With membrane penalties/weights



Without membrane penalties/weights

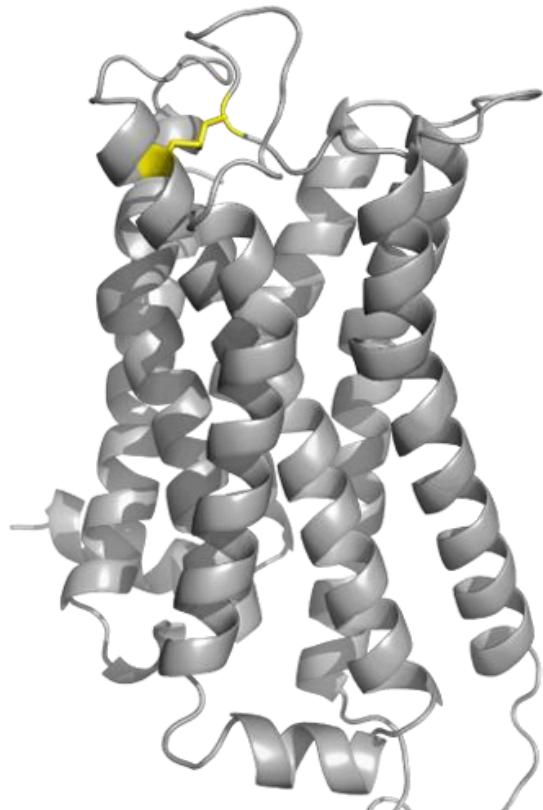


Disulfide constraints

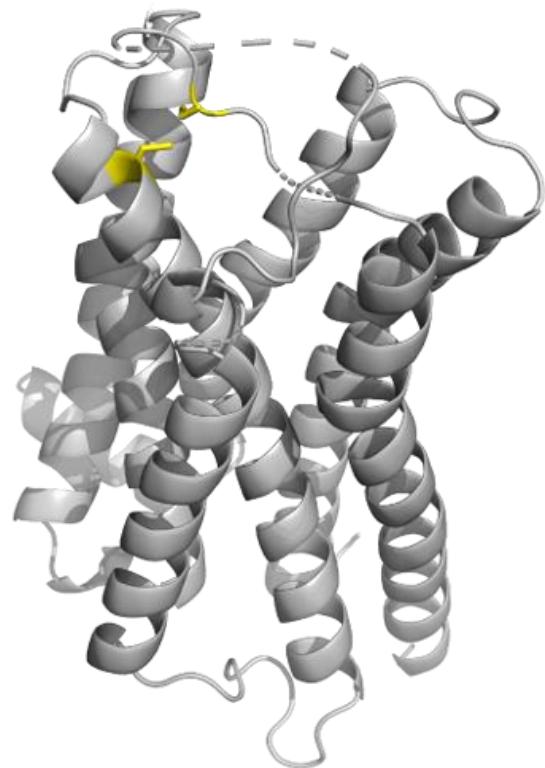
Find this file at */rosetta_cm/demo/input_files/3pbl.disulfide*

72 150

3pbl crystal structure



3pb1 thread into 2rh1



RosettaCM XML

/rosetta_cm/demo/input_files/rosetta_cm.xml

```
<SCOREFXNS>
    <ScoreFunction name="stage1" weights="input_files/stage1_membrane.wts" symmetric="0">
        <Reweight scoretype="atom_pair_constraint" weight="1"/>
    </ScoreFunction>
    <ScoreFunction name="stage2" weights="input_files/stage2_membrane.wts" symmetric="0">
        <Reweight scoretype="atom_pair_constraint" weight="0.5"/>
    </ScoreFunction>
    <ScoreFunction name="fullatom" weights="input_files/stage3_rlx_membrane.wts"
symmetric="0">
        <Reweight scoretype="atom_pair_constraint" weight="0.5"/>
    </ScoreFunction>
    <ScoreFunction name="membrane" weights="membrane_highres_Menv_smooth" symmetric="0">
        <Reweight scoretype="cart_bonded" weight="0.5"/>
        <Reweight scoretype="pro_close" weight="0"/>
    </ScoreFunction>
</SCOREFXNS>
```

*Find all .wts files in */rosetta_cm/demo/input_files*



RosettaCM XML

/rosetta_cm/demo/input_files/rosetta_cm.xml

```
<MOVERS>
  <Hybridize name="hybridize" stage1_scorefxn="stage1" stage2_scorefxn="stage2"
    fa_scorefxn="fullatom" batch="1" stage1_increase_cycles="1.0" stage2_increase_cycles="1.0"
    linmin_only="1" realign_domains="0" disulf_file="input_files/3pbl.disulfide"
    fa_cst_file="fullatom.cst">
    <Template pdb="threaded_pdbs/4iar_out.pdb" cst_file="AUTO" weight="1.000" />
    <Template pdb="threaded_pdbs/4bvn_out.pdb" cst_file="AUTO" weight="1.000" />
    <Template pdb="threaded_pdbs/2rh1_out.pdb" cst_file="AUTO" weight="1.000" />
    <Template pdb="threaded_pdbs/5dsg_out.pdb" cst_file="AUTO" weight="1.000" />
    <Template pdb="threaded_pdbs/5cxv_out.pdb" cst_file="AUTO" weight="1.000" />
  </Hybridize>
  <ClearConstraintsMover name="clearconstraints"/>
  <FastRelax name="relax" scorefxn="membrane" repeats="1" dualspace="1" bondangle="1"/>
</MOVERS>
<OUTPUT scorefxn="membrane"/>
```

RosettaCM Options

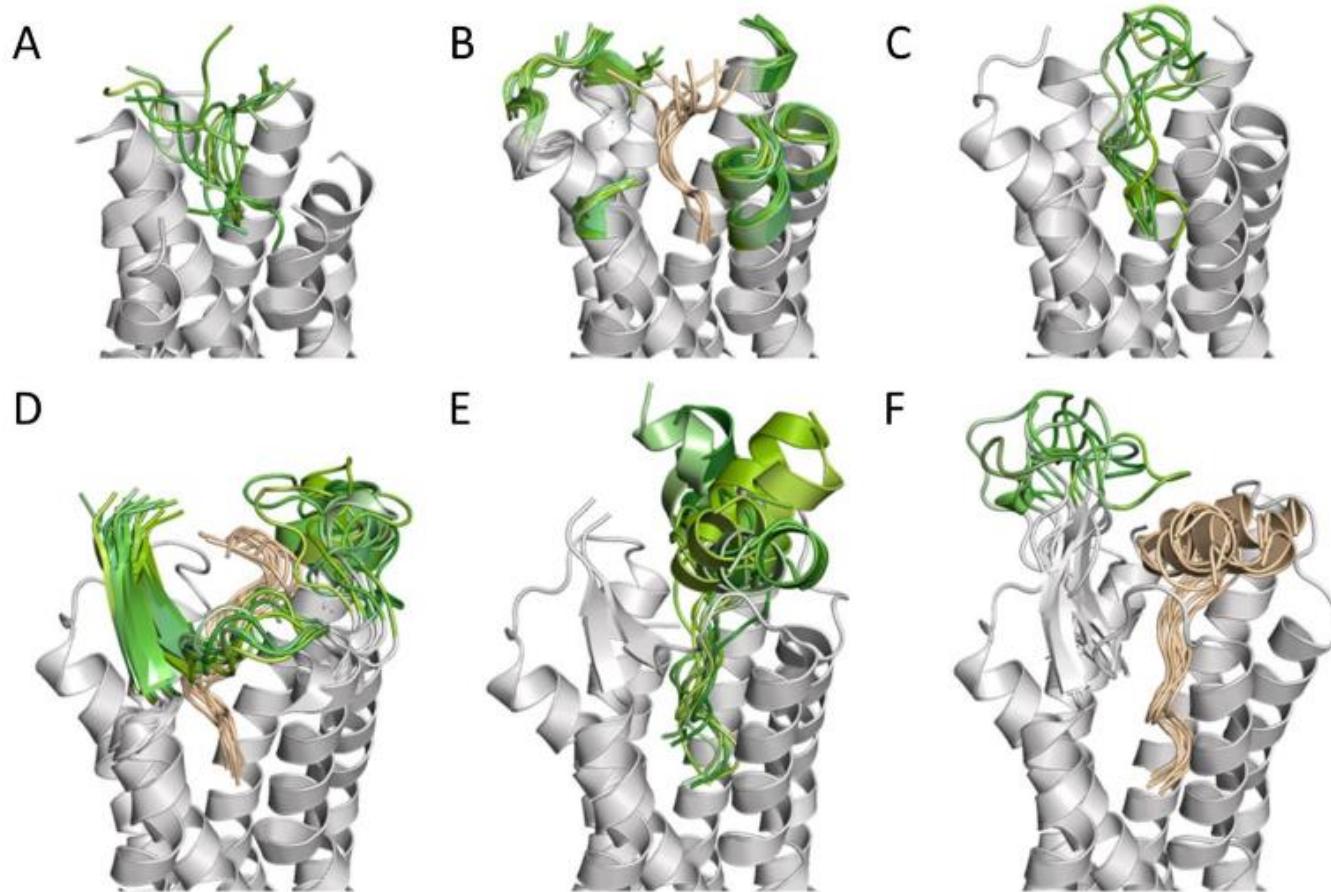
/rosetta_cm/3_hybridize/rosetta_cm.options

```
# i/o
-in:file:fasta input_files/3pb1.fasta                                ##### your target
sequence
-parser:protocol input_files/rosetta_cm.xml
-out:path:all output_files/

#Initialize membrane                                                 ##### only if modeling a membrane
protein
-in:file:spanfile input_files/3pb1.span
-membrane:no_interpolate_Mpair
-membrane:Menv_penalties
-rg_reweight .1
-restore_talaris_behavior

# relax options
-relax:minimize_bond_angles
-relax:minimize_bond_lengths
-relax:jump_move true
-default_max_cycles 200
-relax:min_type lbfgs_armijo_nonmonotone
-score:weights input_files/stage3_rlx_membrane.wts                  ##### use ref2015_cart if soluble
protein
-use_bicubic_interpolation
-hybridize:stage1_probability 1.0
-sdg_upper_bound 15
```

Consecutive modeling of the Ghrelin/GHSR complex



Bender, B.J.; et al. *Structure*, 2019

Tutorial

Comparative modeling of D3 receptor with five class A GPCR templates

Four steps:

1. Setup
2. Threading
3. RosettaCM hybridize
4. Final model selection

References

- RosettaCM documentation
https://www.rosettacommons.org/docs/latest/application_documentation/structure_prediction/RosettaCM
- RosettaCM: Multi-template
Yifan Song, et al. (2013). High-Resolution Comparative Modeling with RosettaCM. *Structure*, 21(10), 1735-1742.