

RosettaCM: multi-template comparative modeling



VANDERBILT
UNIVERSITY

Hope Woods

Meiler Lab

E-Mail: hope.woods@vanderbilt.edu

Why comparative modeling with RosettaCM?

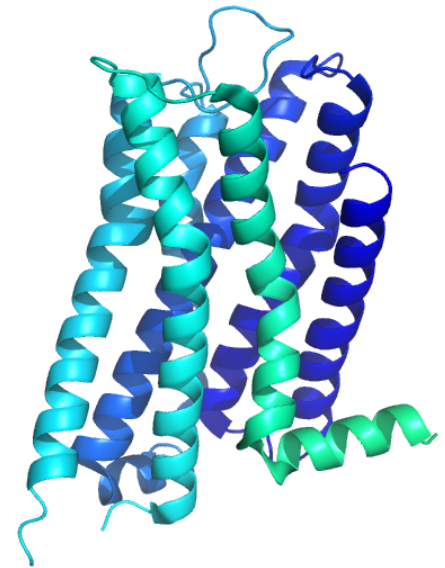
- Ab initio protein folding shows best results <120 amino acid length
- Structural information for a protein is available through crystal structures of related proteins

Sequence

```
>3PBL
DYKDDDDGAPASLSQLSSHLNYTCGAENSTGASQARPHAYYA
LSYCALILAIVFGNGLVCMAYLKERALQTTTNYLVVSLAVAD
LLVATLVMPWVVYLEVTGGVWNFSRICCDVFVTLDDMMCTAS
IWNLCASIDRYTAVVMPVHYQHGTGQSSCRRVALMITAVWV
LAFVAVSCPLLFGFNTTGDPTVCSISNPDFVIYSSVVSFYLPF
GVTVLVYARIYVVLKQRRRKNI FEMLRIDEGLRLKIYKDTEG
YYTIGIGHLLTKSPSLNAAKSELDKAIGRNTNGVITKDEAEK
LFNQDVDAAVRGILRNAKLKPVYDSLDAVRRALINMVFQMG
ETGVAGFTNSLRMLQQKRWDEAAVNLA KSRWYNQTPNRAKRV
ITTFRTGTWDAYGVPLREKKATQMV AIVLGAFIVCWLPFFLT
HVLNTHCQTCHVSPPELYSATTWLG YVNSALNPVIYTTFNIEF
RKAFLKILSCGRPLEVL FQ
```

?

Homology model

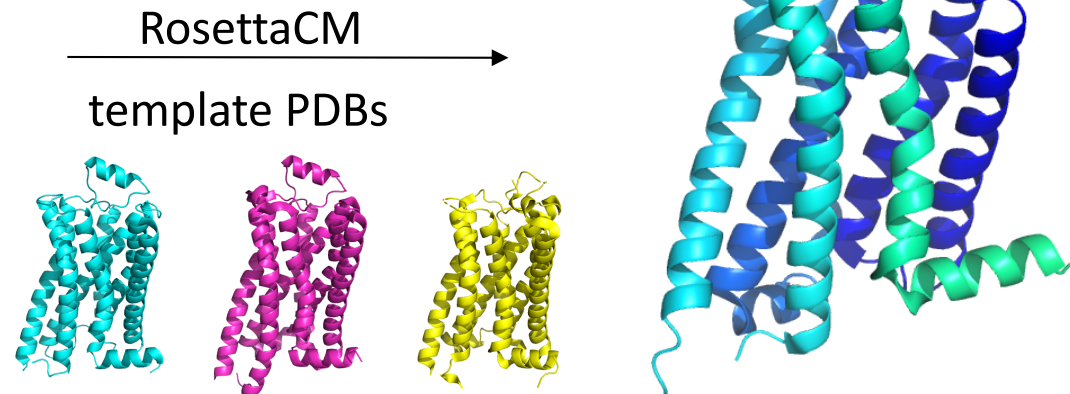


Why comparative modeling with RosettaCM?

- Ab initio protein folding shows best results <120 amino acid length
- Structural information for a protein is available through crystal structures of related proteins

Sequence

```
>3PBL
DYKDDDDGAPASLSQLSSHLNYTCGAENSTGASQARPHAYYA
LSYCALILAIVFGNGLVCMAYLKERALQTTTNYLVVSLAVAD
LLVATLVMPWVVYLEVTGGVWNFSRICCDVFVTLDDMMCTAS
IWNLCASIDRYTAVVMPVHYQHGTGQSSCRRVALMITAVWV
LAFVAVSCPLLFGFNTTGDPTVCSISNPDFVIYSSVVSFYLPF
GVTVLVYARIYVVLKQRRRKNI FEMLRIDEGLRLKIYKDTEG
YYTIGIGHLLTKSPSLNAAKSELDKAIGRNTNGVITKDEAEK
LFNQDVDAAVRGILRNAKLKPVYDSLDAVRRALINMVFQMG
ETGVAGFTNSLRMLQQKRWDEAAVNLA KSRWYNQTPNRAKRV
ITTFRTGTWDAYGVPLREKKATQMVAIVLGAFIVCWLPFFLT
HVLNTHCQTCHVSPPELYSATTWLGYNALNPVIYTTFNIEF
RKAFLKILSCGRPLEVLFQ
```



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



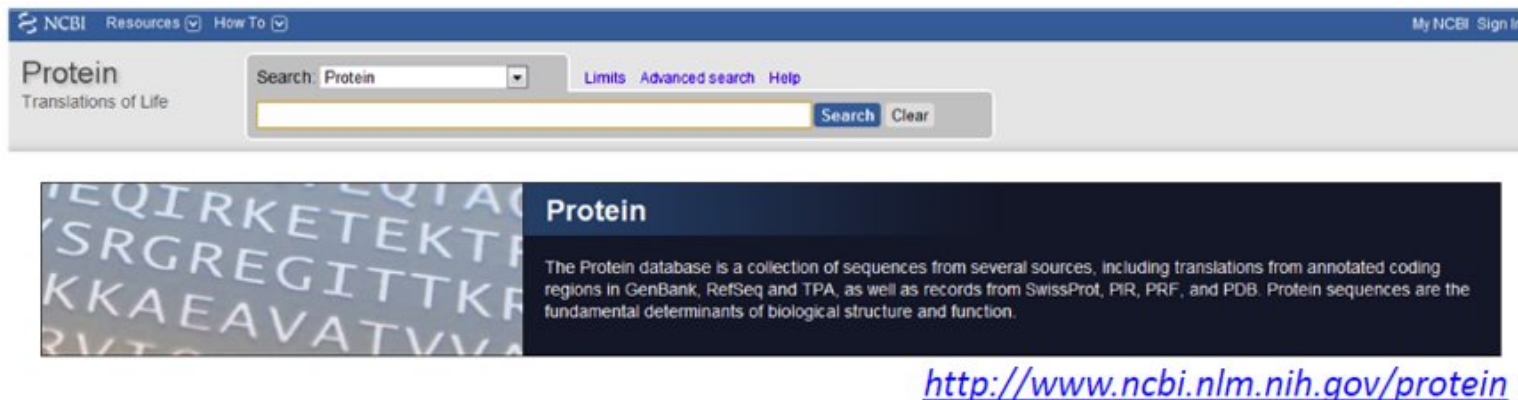
Target sequence of dopamine D3 receptor

(PDB: Dopamine D3 receptor 3pbl)

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

>3pbl

```
YALSYCALILAIIVFGNGLVCMAYLKERALQTTTNYLVVSLAVADLLVATLVMPWVVYLEVTGGVWNFSRICCDVF  
VTLDVMMCTASIWNLCASIDRYTAVVMPVHYQHGTGQSSCRRVALMITAVVVLAFVSCPLLFGFNTTGDPTVC  
SISNPDFVIYSSVVSFYLPFGVTVLVYARIYVVLKQRRRKAAGVPLREKKATQMVAIVLGAFIVCWLPF  
FLTHVLNTHCQTCHVSPELYSATTWLGYVNSALNPVIYTTFNIEFRKAFLKILSC
```



NCBI Resources How To My NCBI Sign In

Protein
Translations of Life

Search: Protein Limits Advanced search Help

Search Clear

Protein

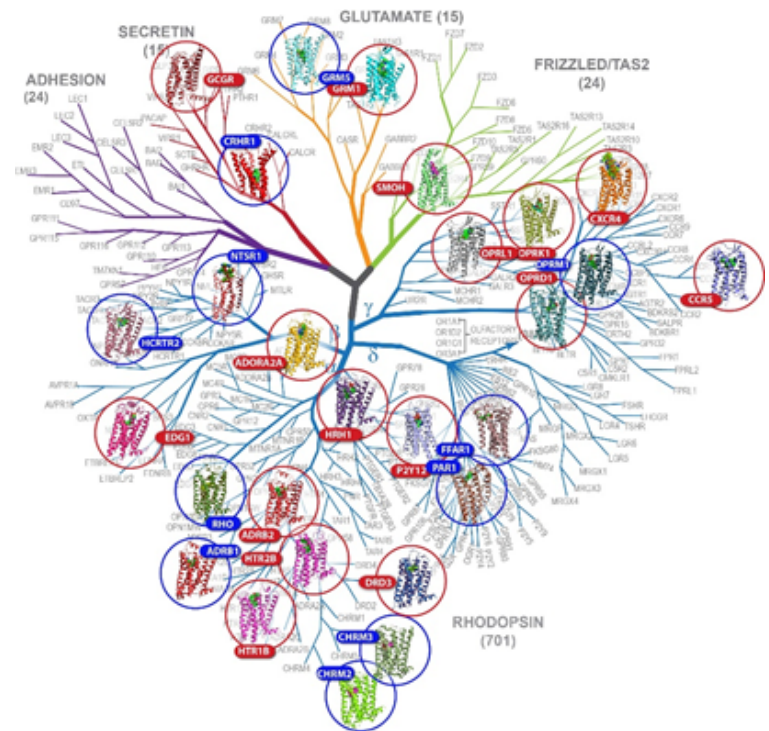
The Protein database is a collection of sequences from several sources, including translations from annotated coding regions in GenBank, RefSeq and TPA, as well as records from SwissProt, PIR, PRF, and PDB. Protein sequences are the fundamental determinants of biological structure and function.

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



Template identification for dopamine D3 receptor

- PDB ID: 3pbl
- 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 |
|------------------------|--------|----------|
| β 2-adrenoceptor | 3SN6 | 36 |
| 5-HT1B receptor | 4IAR | 32 |
| β 2-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_pdb/original_files/*

The screenshot shows the RCSB PDB website homepage. At the top, the logo for RCSB PDB (Protein Data Bank) is displayed, along with the text "A MEMBER OF THE PDB" and "An Information Portal to Biological Macromolecular Structures". Below this, a status bar indicates "As of Tuesday Feb 22, 2011 at 4 PM PST there are 71415 Structures" and provides links to "PDB Statistics". The main navigation bar includes "Contact Us | Print" and a search bar with "PDB ID or Text" and "PDB ID lookup or Text search of the complete structure file" options. The left sidebar contains links for "MyPDB", "Home", "Deposition", and "Search". The main content area features a "Resource for Studying Biological Macromolecules" section, a "Featured Molecules" section with a "Molecule of the Month: Integrin", and a "Structural View of Biology" section. The right sidebar includes a "Customize This Page" section, a "New Features" section, and a "RCSB PDB News" section. A "Structural Neighbors" section at the bottom right shows a 3D protein structure and a link to explore structural neighbors.

<http://www.rcsb.org>



Multiple sequence alignment

CLUSTAL O(1.2.4) multiple sequence alignment

```
5cxv      -----KGPWQVAFIGITTGLLSLATVTGNLLVLISFKVNTTELKTVNNYFLLSLACADL
5dsg      GPSSHNRYETVEMVFIATVTGSLSLVTVVGNILVMLS IKVNRQLQTVNNYFLFSLACADL
3pbl      -----YALSYCALILAI VFGNGLVCM AVLKERALQTTTNYLVVSLAVADL
4iar      YIYQDSISLPWKV-LLVMLLALITLATTLSNAFVIATVYRTRKLHTPANYLIASLAVTDL
2rh1      -----DEVWVV-GMGIVMSLIVLAIVFGNVLVITAIKFERLQTVTNYFITSLACADL
4bvn      -----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 teal header with the text "Clustal Omega". Below the header, there are navigation links: "Input form", "Web services", and "Help & Documentation". To the right of these links are "Share" and "Feedback" buttons. Below the navigation bar, there's a breadcrumb trail: "Tools > Multiple Sequence Alignment > Clustal Omega". The main heading is "Multiple Sequence Alignment". Below this, a paragraph describes Clustal Omega as a new multiple sequence alignment program that uses seeded guide trees and HMM profile-profile techniques to generate alignments between three or more sequences. It also mentions that for two sequences, pairwise sequence alignment tools should be used. The interface is divided into three steps: "STEP 1 - Enter your input sequences", "STEP 2 - Set your parameters", and "STEP 3 - Submit your job". In STEP 1, there's a text area for entering or pasting a set of PROTEIN sequences in any supported format. Below this text area, there's a link to "Or, upload a file" and a "Browse..." button, with the text "No file selected" next to it. In STEP 2, there's a section for "OUTPUT FORMAT" with a dropdown menu currently set to "Clustal w/o numbers". Below this, there's a note stating "The default settings will fulfil the needs of most users and, for that reason, are not visible." and a "More options..." button with a link to "(Click here, if you want to view or change the default settings.)".

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



Adjusting multiple sequence alignments

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

Raw ClustalO alignment:

```
3pbl  -----YALSICALILAIVFGNGLVCMVLKERALQTTTNYLVVSLAVADL
5cxv  -----KGPWQVAFIGITTGLLSLATVTGNLLVLISFKVNTLKTNNYFLLSLACADL
5dsg  GPSSHNRYETVEMVFIATVTGSLSLVTVVGNIIVMLSIKVNRLQTVNNYFLFSLACADL
4iar  YIIQDSISLPWKV-LLVMLLALITLATTLSNAFVIATVYTRKLHTPANYLIASLAVTDL
2rhl  -----DEVWVV-GMGIVMSLIVLAIVFGNVLVITAIKFERLQTVTNYFITSACADL
4bvn  -----LSQQWEA-GMSLLMALVVLLIVAGNVLVIAAIGSTQRLQTLTNLFITSACADL
```

Adjusted alignment:

```
3pbl  -----YALSICALILAIVFGNGLVCMVLKE-RALQT-TTNYLVVSLAVADL
5cxv  -----KGPWQVAFIGITTGLLSLATVTGNLLVLISFKVN-TELKT-VNNYFLLSLACADL
5dsg  GPSSHNRYETVEMVFIATVTGSLSLVTVVGNIIVMLSIKVN-RQLQT-VNNYFLFSLACADL
4iar  -YIIQDSISLPWKVLLVMLLALITLATTLSNAFVIATVYTRKLHT-PANYLIASLAVTDL
2rhl  -----DEVWVVGMGIVMSLIVLAIVFGNVLVITAIKFERLQTVTNYFITSACADL
4bvn  -----LSQQWEAGMSLLMALVVLLIVAGNVLVIAAIGST-QRLQTLTNLFITSACADL
```

helix regions

highly conserved residues

Alignment issues to be resolved

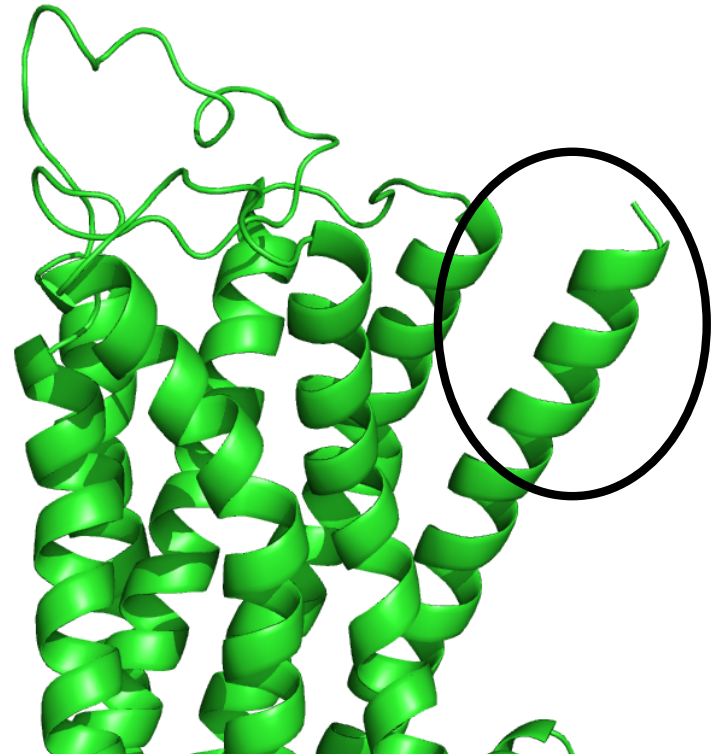
predicted membrane spanning region from OCTOPUS



Adjusted multiple sequence alignments result in improved modeling performance



Example model using
raw alignment

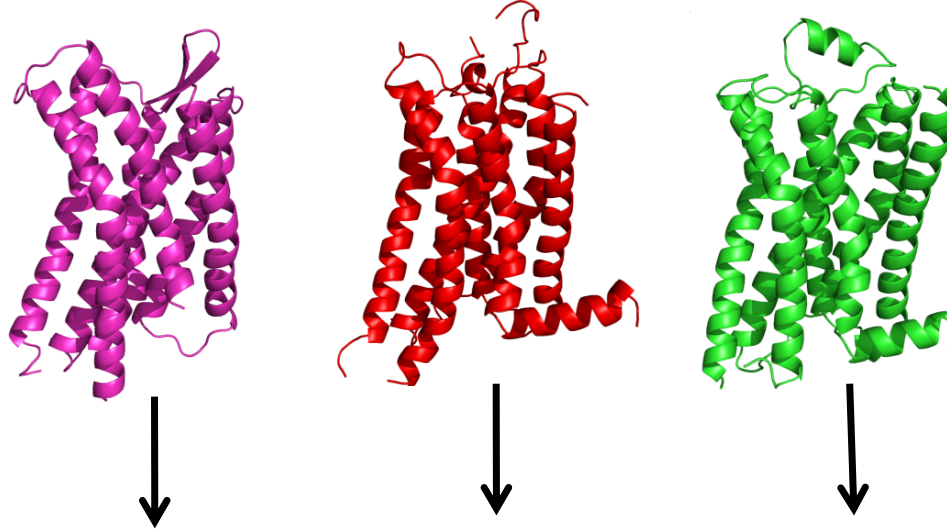


Example model using
raw alignment

Partial Thread

```
-----PWQFSM--LAAYMFLLIMLGFPINFLTLYVTVQHKKLRTPNLNYILLNLAVADLFM  
ANFNKIFL-----PTIYSIIFLTGIVGNGLVILVMGYQKKLRSMTDKYRLHLSVADLLF  
---DEVVVVGMGIVMS---LIVLAIVFGNVLVITAIKFERLQTVTNFYFITSACADLVM  
-----IMGSSVYITVELATAVIATTGNVIVCWAVWI.NSNLQNVTNFYFVVSIAAADIAV
```

alignment



template pdbs

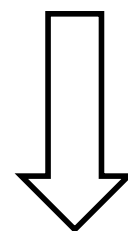


threaded pdbs



Partial Thread

| | (0,0,0) | (1,1,1) | (2,2,2) | (3,3,3) | (4,4,4) | (5,5,5) | |
|----------|---------|---------|---------|---------|---------|---------|---------|
| template | L | L | R | N | N | H | - |
| | (?,?,?) | (?,?,?) | | | | (?,?,?) | (?,?,?) |
| target | L | K | - | - | - | H | V |



Threaded coordinates

| | (0,0,0) | (1,1,1) | (5,5,5) | |
|--------|---------|---------|---------|---|
| target | L | K | H | V |



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

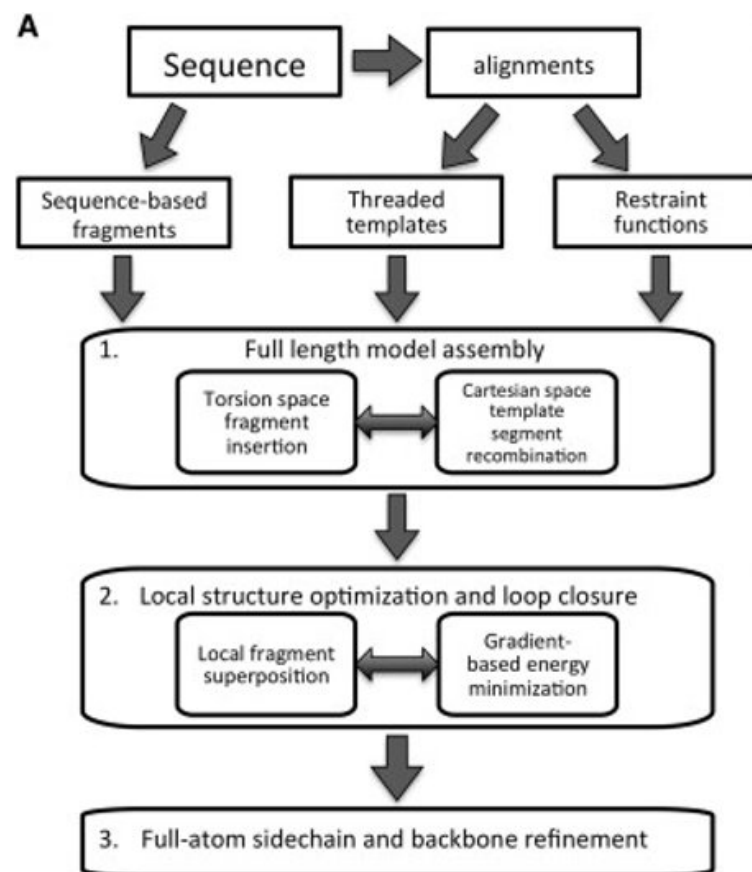
Find converted Grishin alignment files at */rosetta_cm/demo/alignment_files/*

(2rh1.aln 4bvn.aln 4iar.aln 5cxv.aln 5dsg.aln)



Hybridize protocol contains three stages

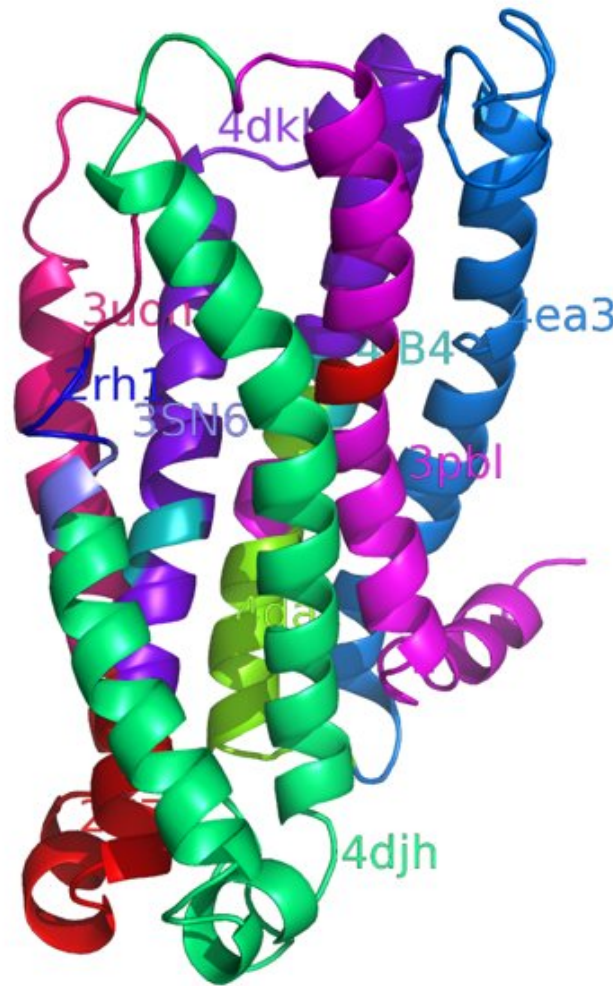
1. Generate initial models from template alignments
2. Explore deviations from templates and close loops in 2 steps :
 - MC: Randomly select de novo or template-based fragment and substitute into current conformation
 - Cartesian space full-backbone minimization
3. Full atom backbone and side chain refinement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

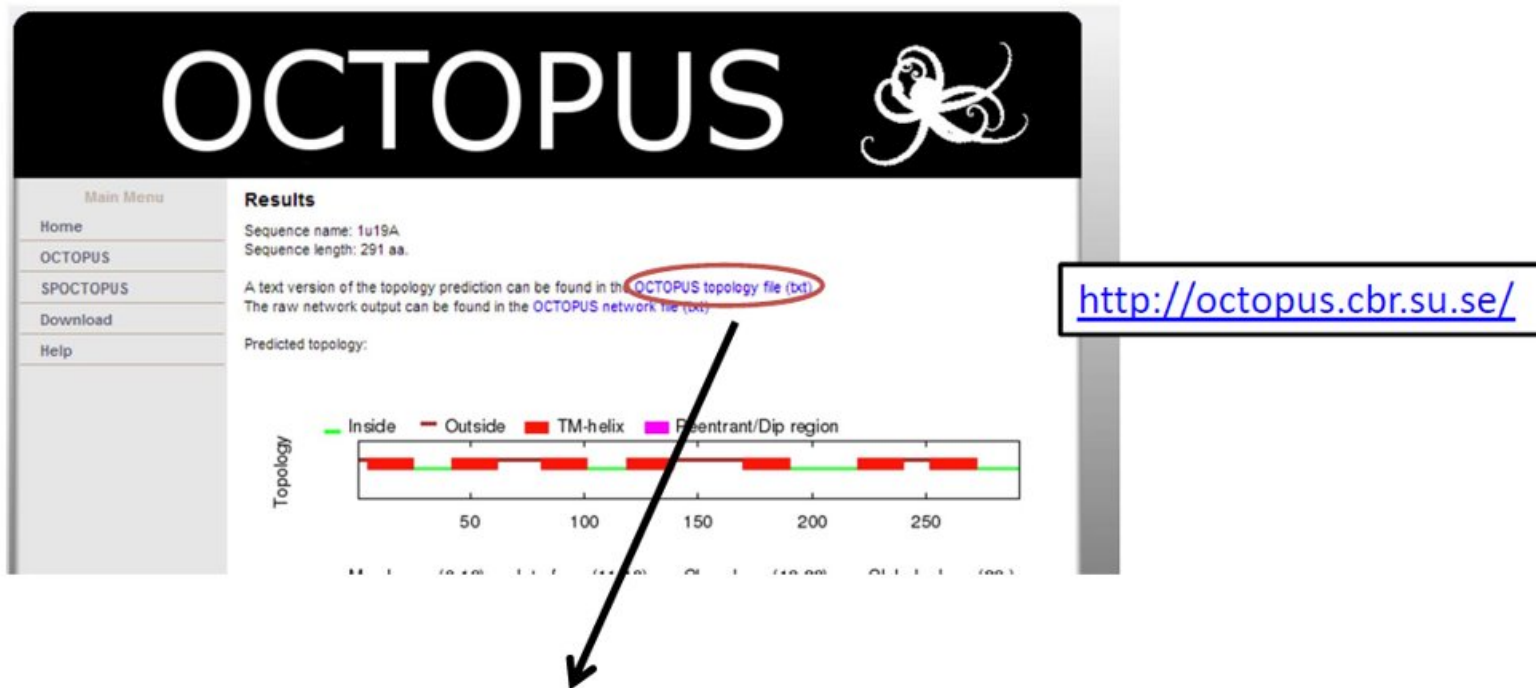
Optional files based on available information:

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



Membrane spanning regions

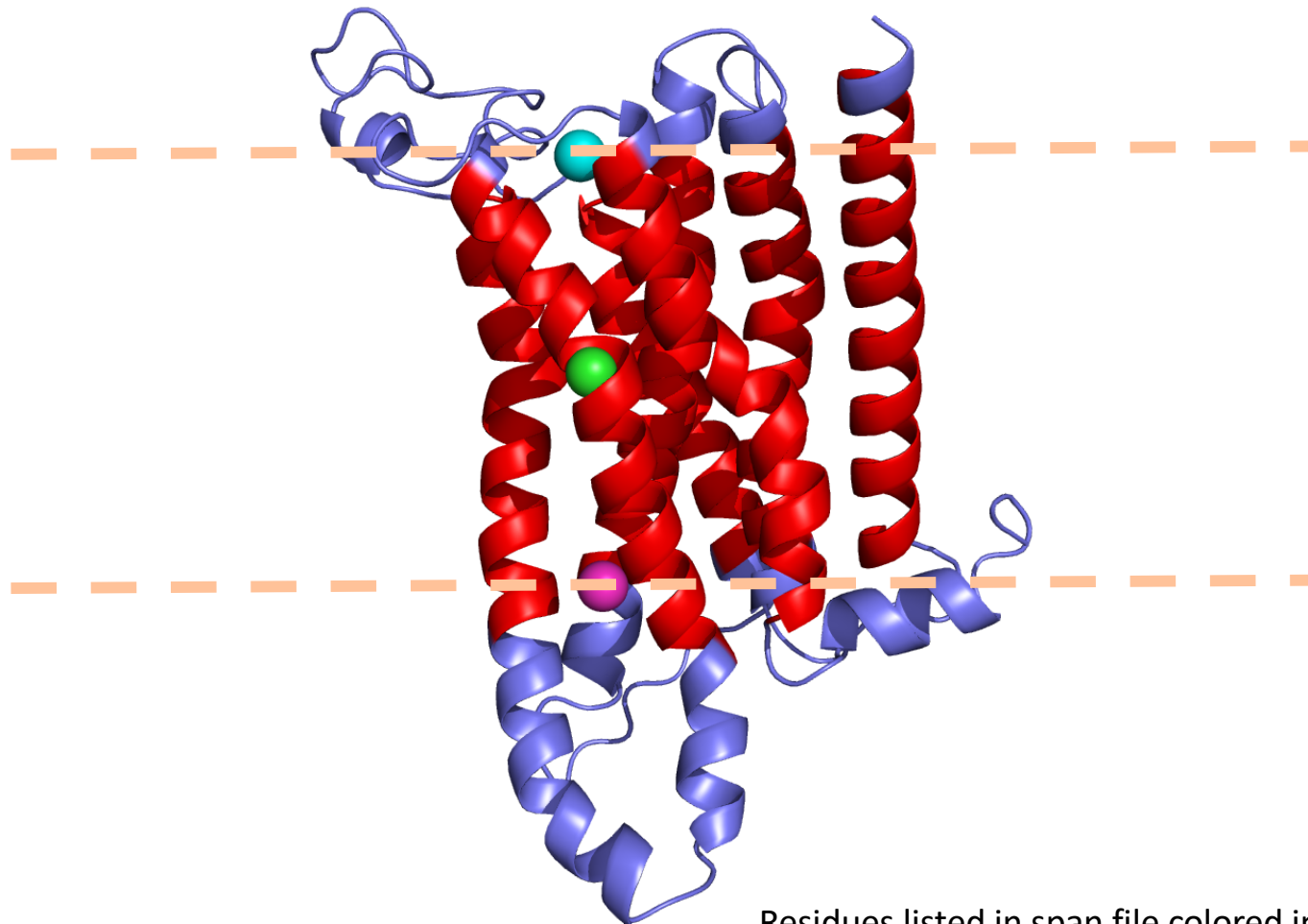
Find this file at `/rosetta_cm/demo/input_files/3pbl.span`



```
octopus2span.pl 3pbl.octopus
```

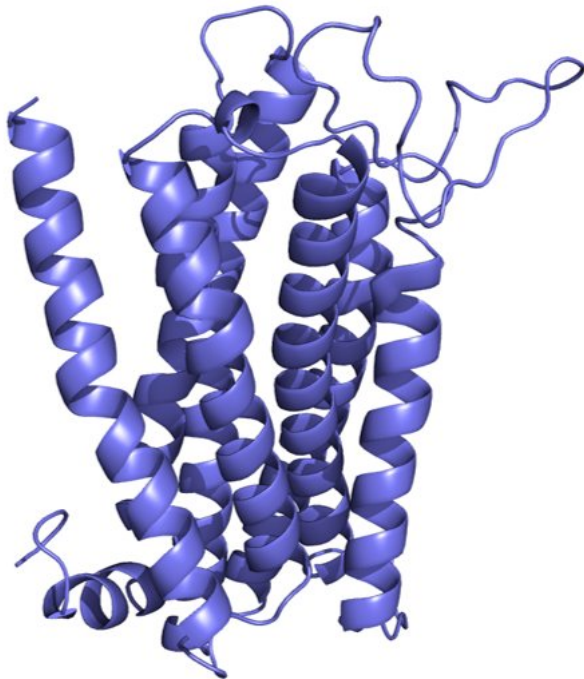


The span file helps RosettaMembrane to define transmembrane regions

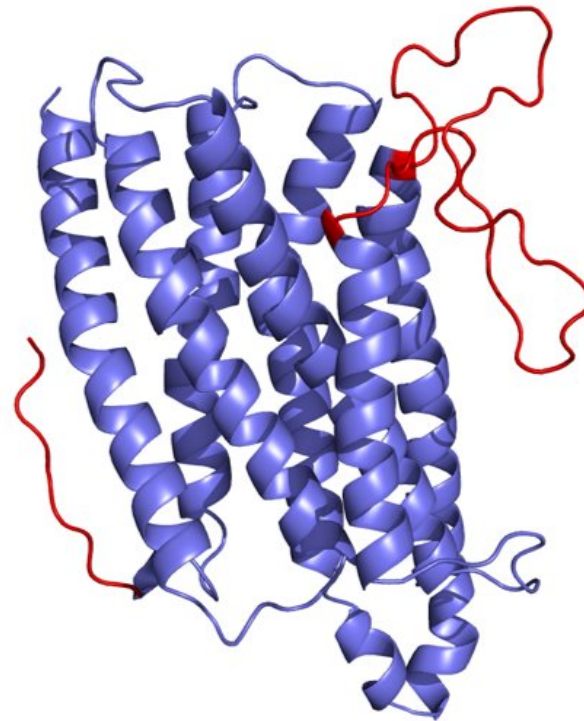


Why are membrane scoring terms important?

With membrane penalties/weights



Without membrane penalties/weights

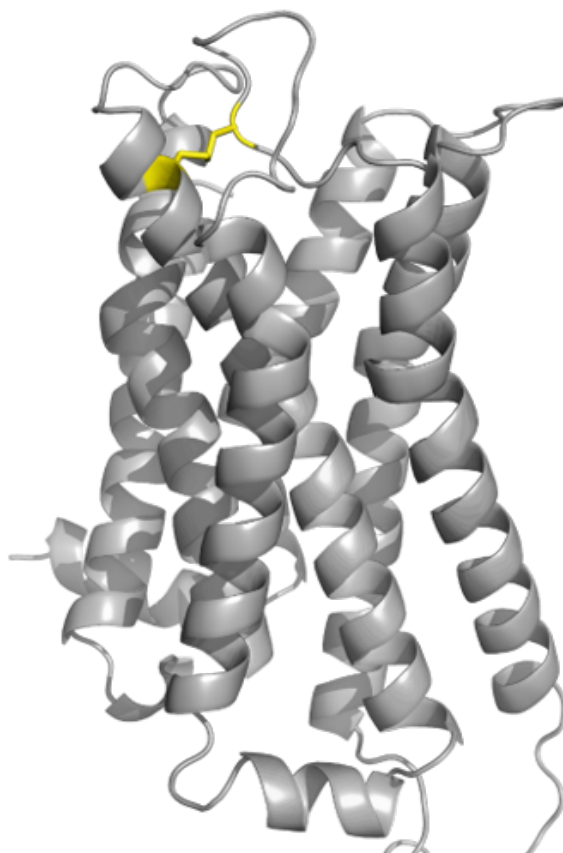


Disulfide constraints

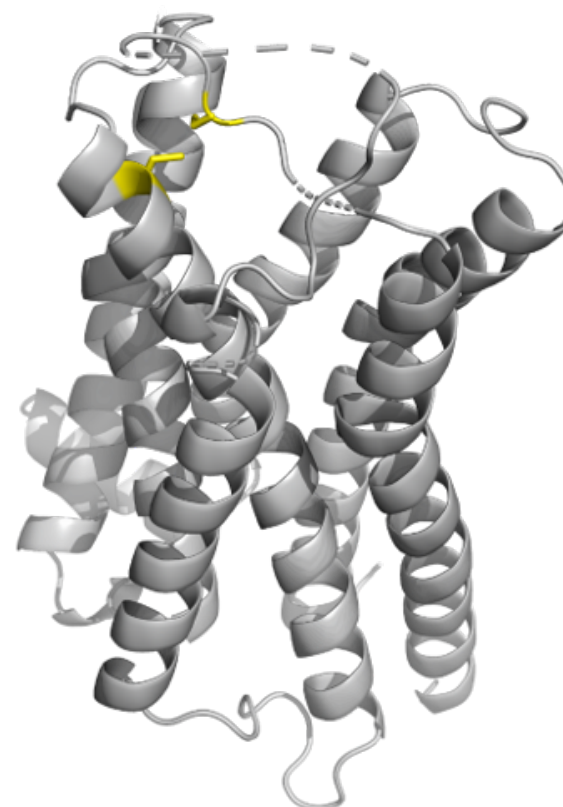
Find this file at */rosetta_cm/demo/input_files/3pbl.disulfide*

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/3pb1.disulfide"
fa_cst_file="fullatom.cst">
    <Template pdb="threaded_pdb/4iar_out.pdb" cst_file="AUTO" weight="1.000" />
    <Template pdb="threaded_pdb/4bvn_out.pdb" cst_file="AUTO" weight="1.000" />
    <Template pdb="threaded_pdb/2rh1_out.pdb" cst_file="AUTO" weight="1.000" />
    <Template pdb="threaded_pdb/5dsg_out.pdb" cst_file="AUTO" weight="1.000" />
    <Template pdb="threaded_pdb/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

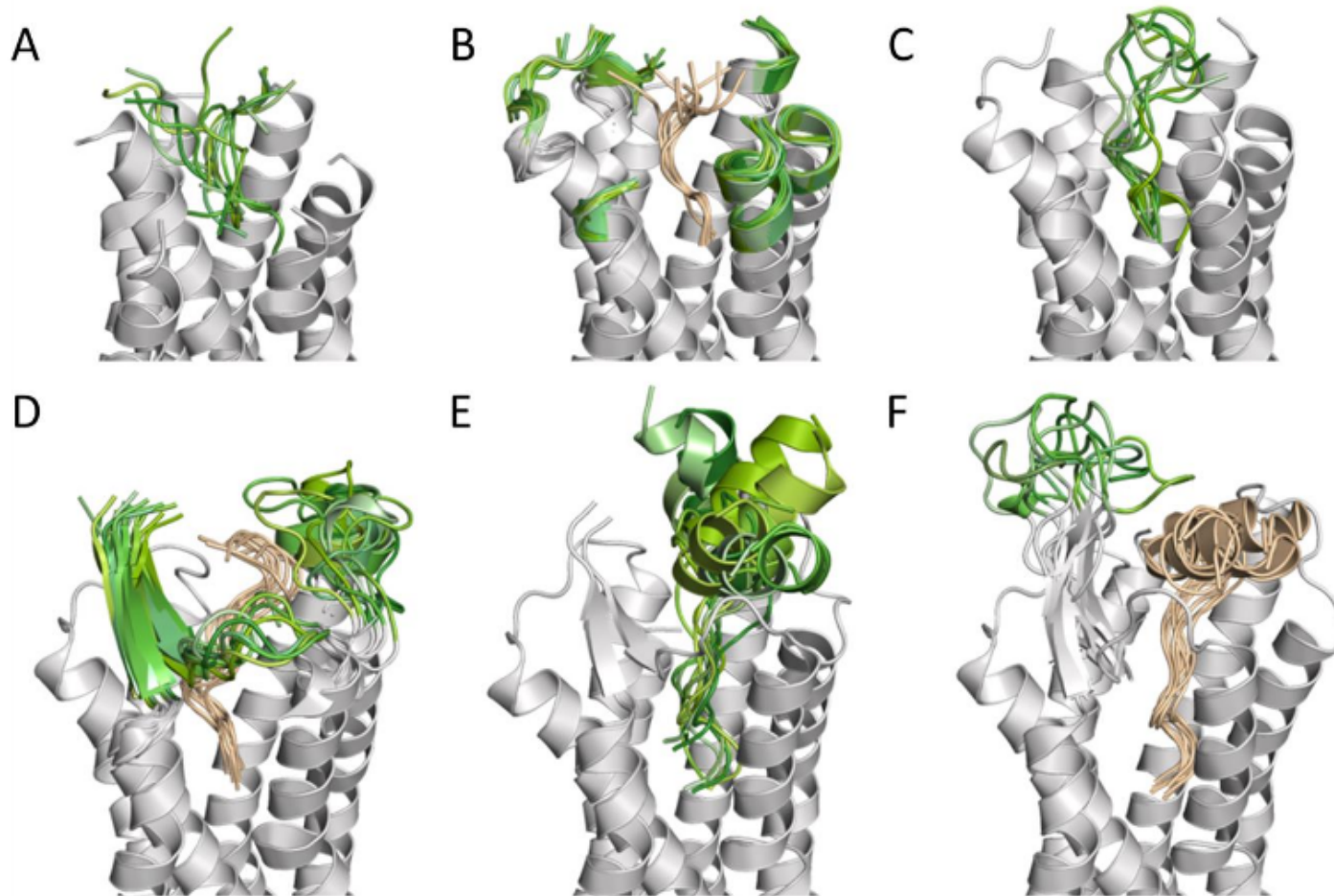
```
# i/o
-in:file:fasta input_files/3pbl.fasta          ##### your target
sequence
-parser:protocol input_files/rosetta_cm.xml
-out:path:all output_files/

#Initialize membrane                        ##### only if modeling a membrane
protein
-in:file:spanfile input_files/3pbl.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
-scg_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.

