

# ROSETTACM: MULTI-TEMPLATE COMPARATIVE MODELING

---

Oanh Vu

[oanh.t.vu.2@vanderbilt.edu](mailto:oanh.t.vu.2@vanderbilt.edu)

Georg Kuenze

[georg.kuenze@vanderbilt.edu](mailto:georg.kuenze@vanderbilt.edu)

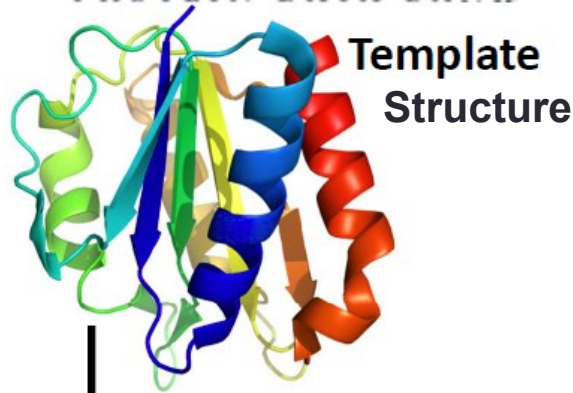
Brian Bender

[brian.j.bender@vanderbilt.edu](mailto:brian.j.bender@vanderbilt.edu)

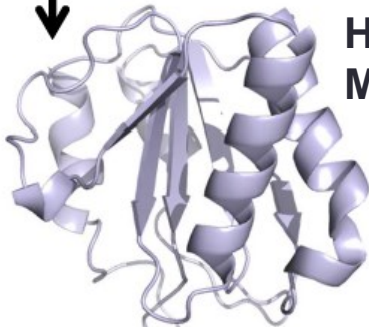
# Introduction: RosettaCM Homology Modeling

## Target Sequence

MKIVYWSGTGNTERMAIAKGIIESGKDVNTI  
NVSDVNIDELLNEDIIGCSAMGDEVLEESEF  
EPFIEEISTKISGKIALFGSYGWDGKWMRDF  
EERMNGYGCVVETIVQNEPDEAEQDCIEFG  
KKIANI



## Homology Model



## • 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%)

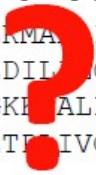
## \*Nomenclature Note\*

- Comparative Modeling = Homology Modeling in the land of Rosetta

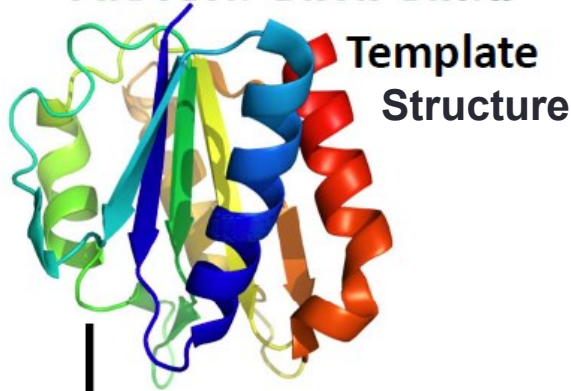
# Identifying Template Structures

## Target Sequence

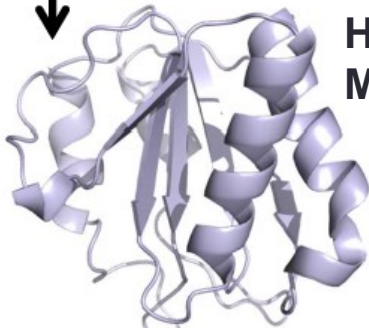
MKIVYWSGTGNTERMAIAKGIIESGKDVNTI  
NVSDVNIDELLNEDILGCSAMGDEVLEESEF  
EPFIEEISTKISGKIALFGSYGWDGKWMRDF  
EERMNGYGCVVVETIVQNEPDEAEQDCIEFG  
KKIANI



RCSB **PDB**  
PROTEIN DATA BANK



Template  
Structure



Homology  
Model

- **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**)

# Practice Target: 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

# Low Identity Templates

- 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

# Comparative Modeling Protocol

- **Step 1:** Align target sequence to template sequences
- **Step 2:** Partial-thread the target sequence onto template structures
- **Step 3:** Combine pieces from different templates using RosettaCM Hybridize

```
-----PWQFSM--LAAYMFLLIMLGFPINFLTLYVTVQHKKLRTPNLYILLNLAVADLFM  
ANFNKIFL-----PTIYSIIFLTGIVGNGLVILVMGYQKKLRSMTDKYRLHLSVADLLF  
---DEVWVVGMGIVMS---LIVLAIVFGNVLVITAIKFERLQTVTNYFITSACADLVM  
-----IMGSSVYITVELAIAVLAILGNVLVCWAVWLNSNLQNVVTNYFVVSLAAADIAV
```

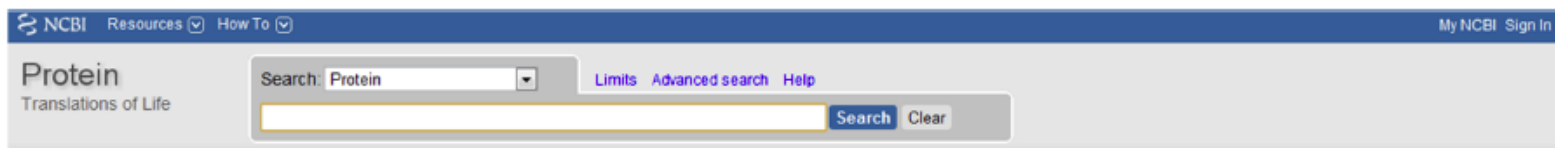
# Target Sequence

(PDB: Dopamine D3 receptor 3pbl)

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

>3pbl

```
YALSYCALILAIIVFGNGLVCM AVLKERALQTTTNYLVVSLAVADLLVATL  
VMPWVVYLEVTGGVWNFSRICCDVFVTL DVM MCTASIWNLC AISDR  
YTAVVMPVHYQHGTGQSSCRRVALMITAVWVLAFAVSCPLLFGFNTT  
GDPTVCSISNPDFVIYSSVVSFYLPFGVTVLVYARIYVVLKQRRRRKAAA  
AAAAAGVPLREKKATQMVAIVLGAFIVCWLPFFLTHVLNTHCQTCHVS  
PELYSATTWLG YVNSALNPVIYTTFNIEFRKAFLKILSC
```



The screenshot shows the top of the NCBI Protein database search page. The header includes the NCBI logo, 'Resources' and 'How To' dropdown menus, and a 'My NCBI Sign In' link. Below the header, the 'Protein' section is titled 'Translations of Life'. A search bar contains the text 'Protein' with a dropdown arrow. To the right of the search bar are links for 'Limits', 'Advanced search', and 'Help'. Below the search bar is a large empty text input field, followed by 'Search' and 'Clear' buttons.



## 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 PDBs

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 interface. At the top, the RCSB PDB logo is on the left, and a banner on the right states 'A MEMBER OF THE PDB' and 'An Information Portal to Biological Macromolecular Structures'. Below the banner, a status line reads: 'As of Tuesday Feb 22, 2011 at 4 PM PST there are 71415 Structures | PDB Statistics'. The main navigation bar includes 'Contact Us | Print', a search box for 'PDB ID or Text', and a link to 'Advanced Search'. The left sidebar contains links for 'MyPDB' (Login, Register), 'Home' (News, Policies, FAQ, etc.), and 'Deposition' (All Deposit Services, etc.). The main content area is titled 'A Resource for Studying Biological Macromolecules' and contains introductory text about the PDB archive and its resources. The right sidebar features a 'Customize This Page' button, 'New Features' (Transporter Classification Database Browser), 'RCSB PDB News' (Weekly, Quarterly, Yearly), and 'Structural Neighbors' (2011-02-22). At the bottom, there is a 'Featured Molecules' section with a 'List View of Archive By: Title | Date | Category' and a 'Structural View of Biology' section with icons for various biological categories.



# Multiple Sequence Alignment

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

CLUSTAL O(1.2.4) multiple sequence alignment

```
5cxv      -----KGPWQVAFIGITTGLLSLATVTGNLLVLISFKVNTELKTVNNYFLLSLACADL
5dsg      GPSSHNRYETVEMVFIATVTGSLSLVTVVGNIIVMLSIKVNRLQTVNNYFLFSLACADL
3pbl      -----YALSYCALILAIIVFGNGLVCMAVLKERALQTTTNYLVVSLAVADL
4iar      YIYQDSISLPWKV-LLVMLLALITLATTLSNAFVIATVYRTRKLHTPANYLIASLAVTDL
2rh1      -----DEVWVV-GMGIVMSLIVLAIVFGNVLVITAIKFERLQTVTNYFITSACADL
4bvn      -----LSQQWEA-GMSLLMALVLLIVAGNVLVIAAIGSTQRLQTLTNLFITSACADL
```

## Clustal Omega

[Input form](#) | [Web services](#) | [Help & Documentation](#)

[Share](#) [Feedback](#)

[Tools](#) > [Multiple Sequence Alignment](#) > Clustal Omega

### Multiple Sequence Alignment

Clustal Omega is a new multiple sequence alignment program that uses seeded guide trees and HMM profile-profile techniques to generate alignments between **three or more** sequences. For the alignment of two sequences please instead use our [pairwise sequence alignment tools](#).

#### STEP 1 - Enter your input sequences

Enter or paste a set of **PROTEIN** sequences in any supported format:

Or, upload a file:  No file selected.

#### STEP 2 - Set your parameters

**OUTPUT FORMAT** **Clustal w/o numbers**

The default settings will fulfill the needs of most users and, for that reason, are not visible.

(Click here, if you want to view or change the default settings.)

#### STEP 3 - Submit your job

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

# Adjusting multiple sequence alignment

## Experimental expectations:

- Highly conserved residues
- Secondary structure elements

## Raw ClustalO alignment:

```
3pbl      -----YALSICALILAIVFGNGLVCMVVLKERALQTTNLYLVVSLAVADL
5cxv      -----KGPWQVAFIGITTGLLSLATVTGNLLVLISFKVNTLKTNNYFLLSLACADL
5dsg      GPSSHNRYETVEMVFIATVTGSLSLVTTVGNNILVMSIKVNRQLQTVNNYFLFSLACADL
4iar      YIQDSISLPWKVLLVMLLALITLATTLNFAFVIATVYRTRKLHTPANYLIASLAVTDL
2rh1      -----DEVVVVGMGIVMSLIVLAIVFGNVLVITAIKFERLQTVTNFYFITSACADL
4bvn      -----LSQQWEAGMSLLMALVLLIVAGNVLVIAAIGSTQRLQTLTNLFITSACADL
```

## Adjusted alignment:

```
3pbl      -----YALSICALILAIVFGNGLVCMVVLKE-RALQT-TNLYLVVSLAVADL
5cxv      -----KGPWQVAFIGITTGLLSLATVTGNLLVLISFKVN-TELKT-VNNYFLLSLACADL
5dsg      GPSSHNRYETVEMVFIATVTGSLSLVTTVGNNILVMSIKVN-RQLQTVNNYFLFSLACADL
4iar      YIQDSISLPWKVLLVMLLALITLATTLNFAFVIATVYRTRKLHT-PANYLIASLAVTDL
2rh1      -----DEVVVVGMGIVMSLIVLAIVFGNVLVITAIKFERLQTVTNFYFITSACADL
4bvn      -----LSQQWEAGMSLLMALVLLIVAGNVLVIAAIGST-QRLQTLTNLFITSACADL
```

helix regions

highly conserved residues

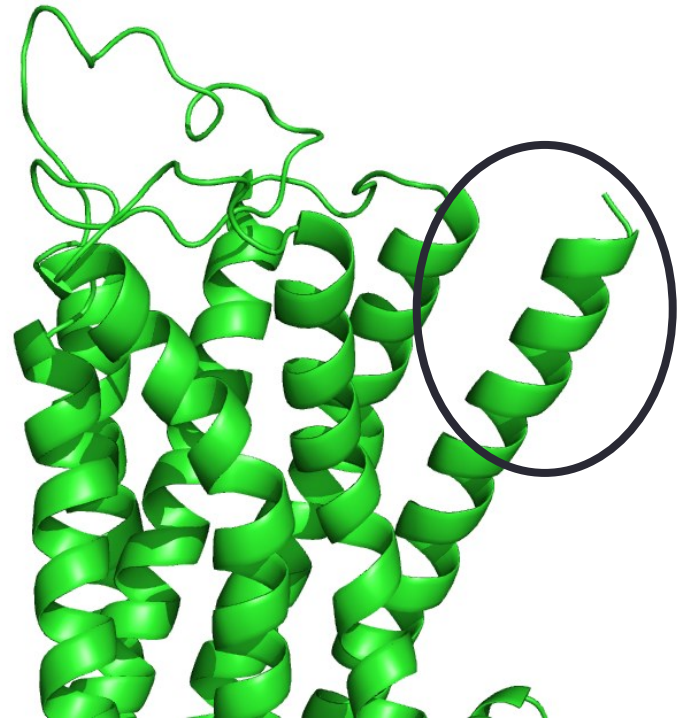
Alignment issues to be resolved

predicted membrane spanning region from OCTOPUS

# Removing helix gaps



Example model using  
raw alignment



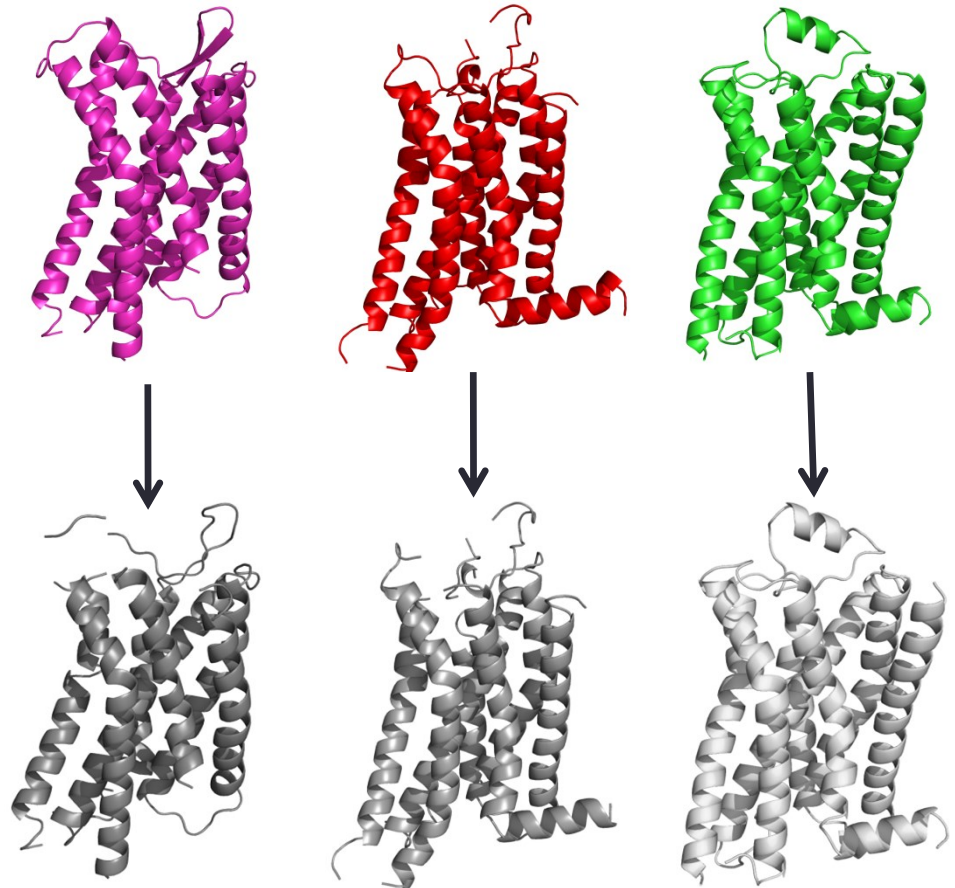
Example model using  
adjusted alignment

# Comparative Modeling Protocol

- **Step 1:** Align target sequence to template sequences
- **Step 2:** Partial-thread the target sequence onto template structures
- **Step 3:** Combine pieces from different templates using RosettaCM Hybridize

```
-----PWQFSM--LAAYMFLIMLGFPINFLTLYVTVQHKKLRTPLNYILLNLAVADLFM  
ANFNKIFL-----PTIYSIIFLTGIVGNGLVILVMGYQKKLRSMTDKYRLHLSVADLLF  
---DEVVVVGMGIVMS---LIVLAIVFGNVLVITAIKFERLQTVTNYFITSACADLVM  
-----IMGSSVYITVELAIAVLAILGNVLVCWAVWLNSNLQNVTNYFVVSLAAADIAV
```

+



# Threading

Template: 

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

Target: 

L	K	-	-	-	H	V
---	---	---	---	---	---	---

*Thread  
Coordinates*



Target: 

(0,0,0)	(1,1,1)	(5,5,5)	
L	K	H	V

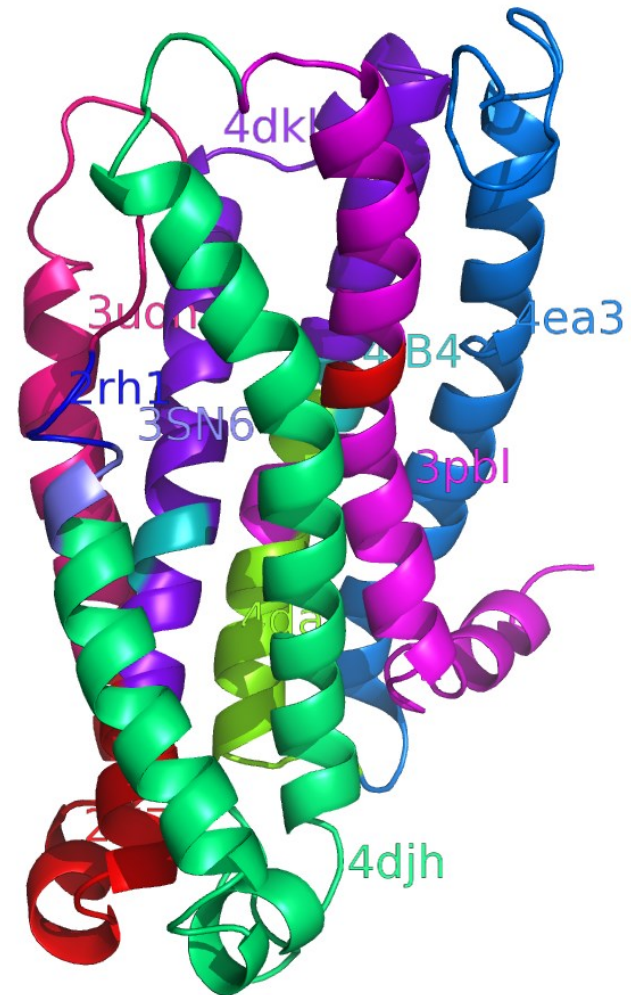
# Grishin Format Alignments Needed for Rosetta Threading

- 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)

# Comparative Modeling Protocol

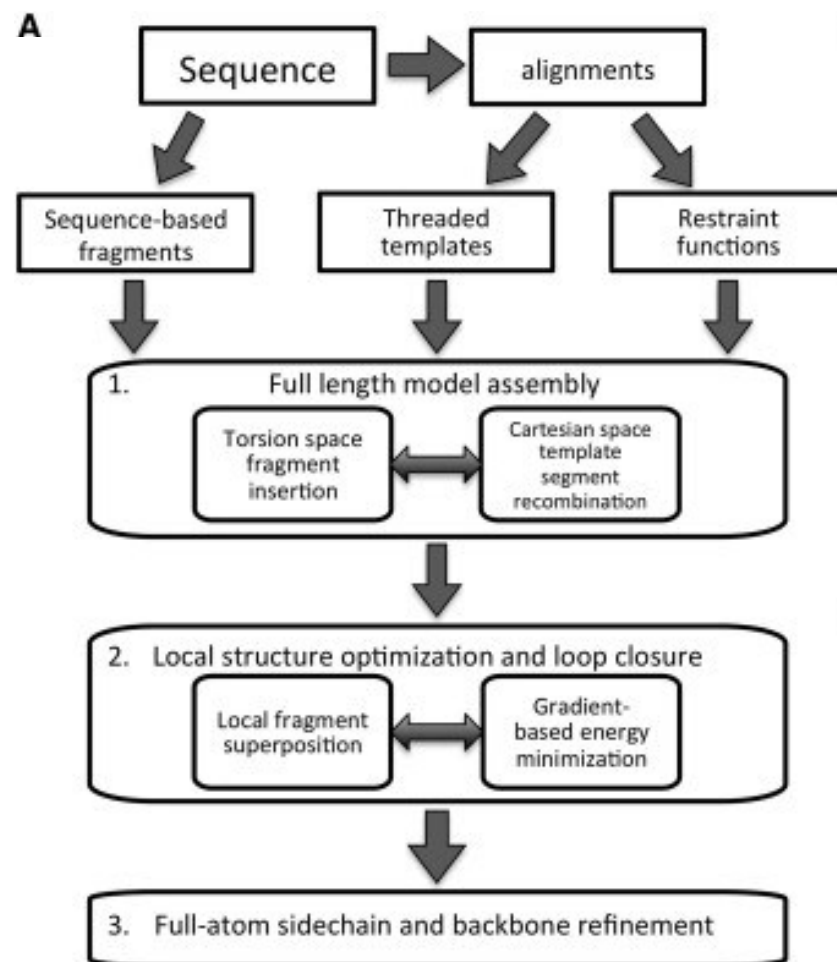
- **Step 1:** Align target sequence to template sequences
- **Step 2:** Partial-thread the target sequence onto template structures
- **Step 3:** Combine pieces from different templates using RosettaCM Hybridize





# RosettaCM: 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, et al. 2013



# Input Files for RosettaCM

Bare minimum:

- Partial-threaded structures
- Mover definition and options

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

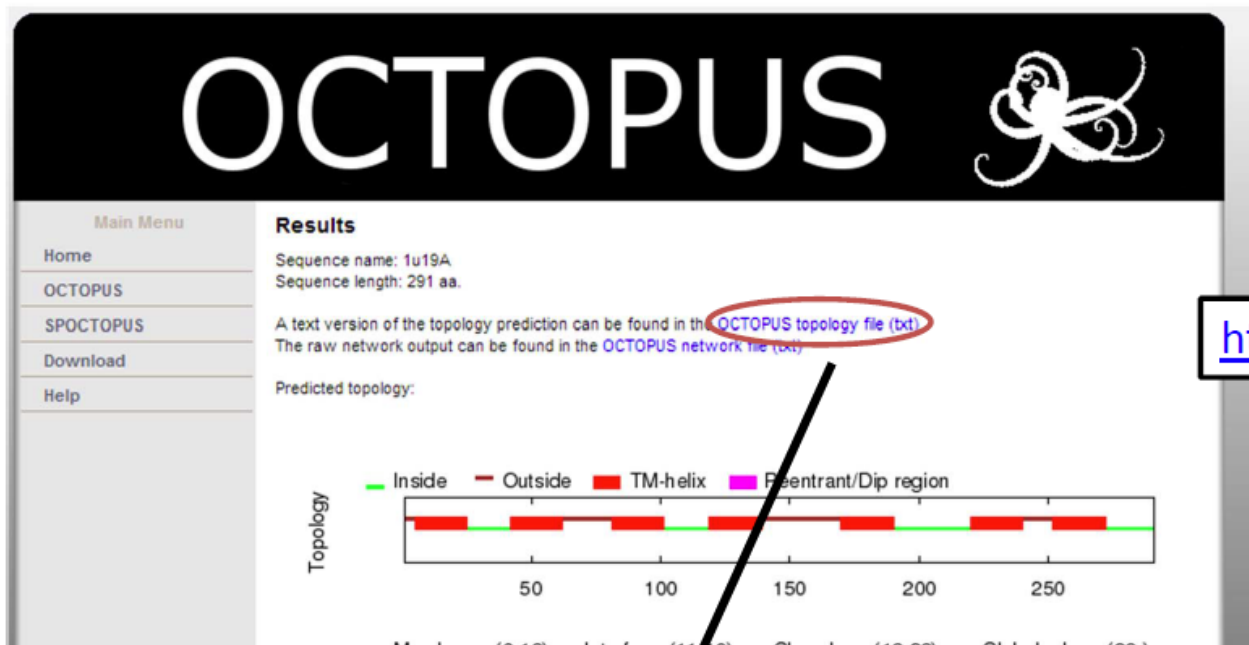
- Membrane spanning regions (span file)
- Membrane weight patches

Optional files based on available information:

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

# Membrane spanning regions

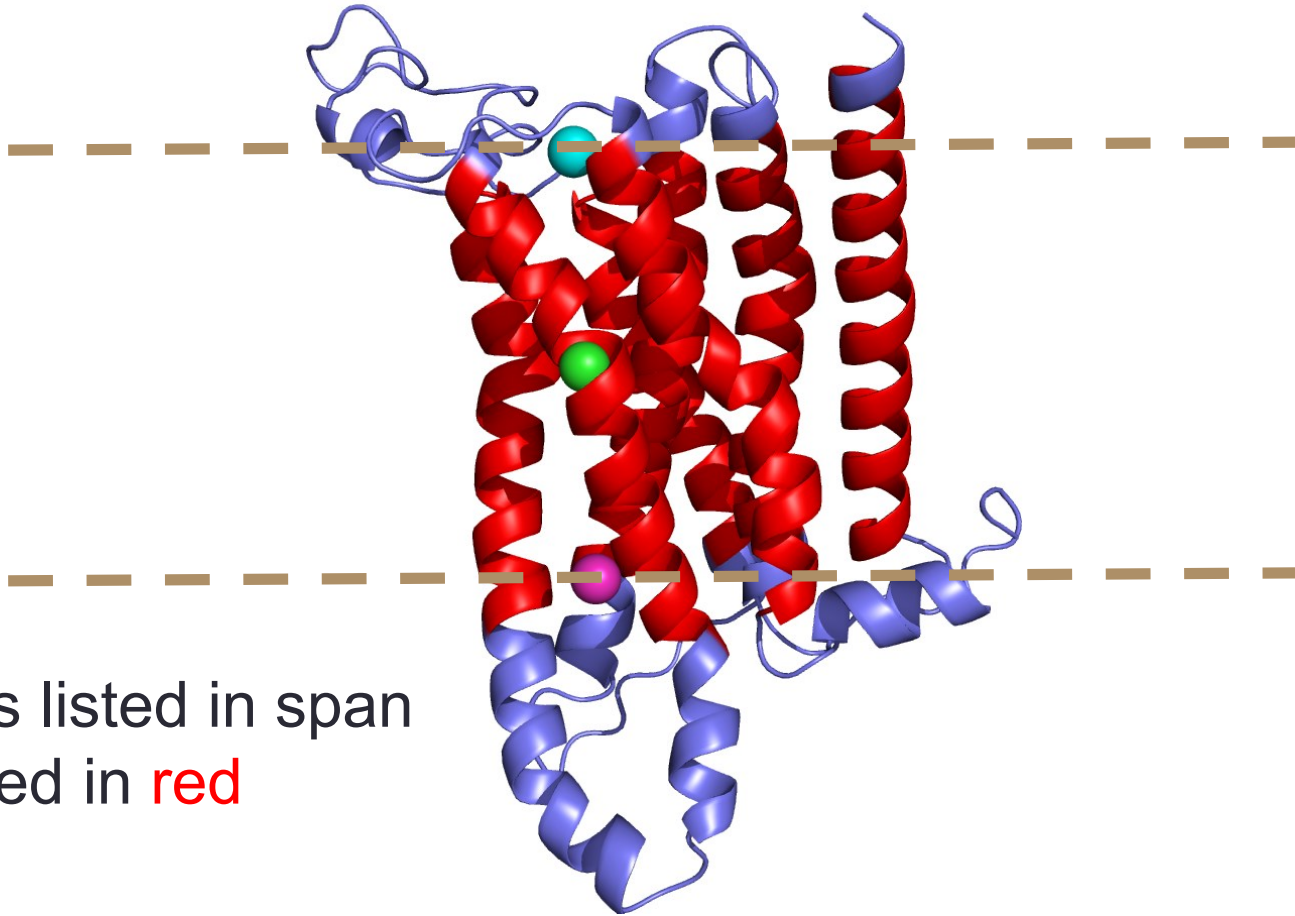
Find this file at `/rosetta_cm/demo/input_files/3pbl.span`



<http://octopus.cbr.su.se/>

`octopus2span.pl 3pbl.octopus`

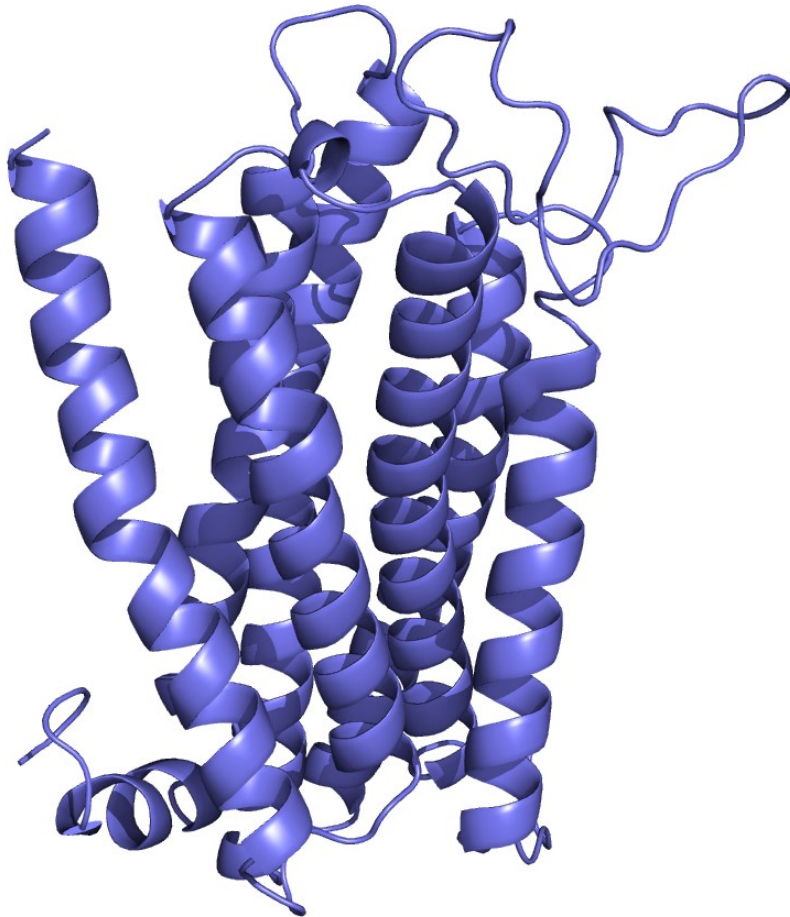
# Rosetta Membrane



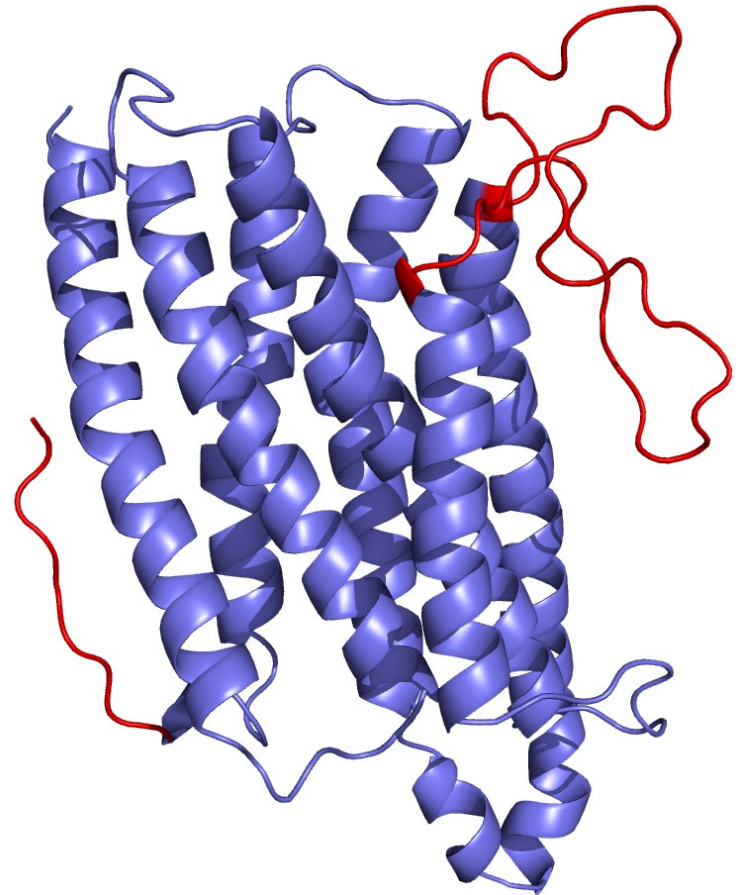
Residues listed in span  
file colored in red

# Why use membrane scoring terms?

**With** membrane penalties/weights



**Without** membrane penalties/weights

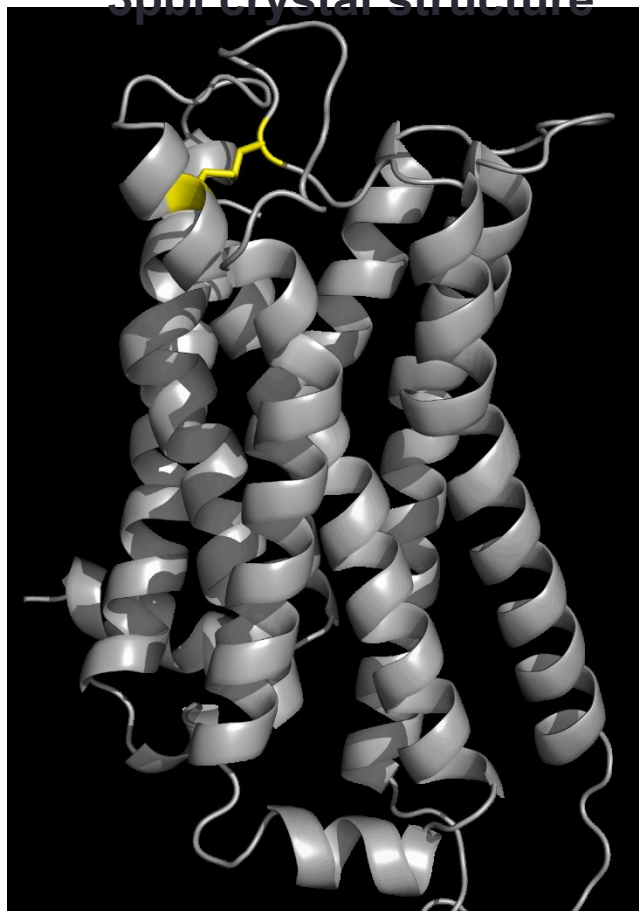


# Disulfide constraints

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_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**  
-parser:protocol input\_files/rosetta\_cm.xml  
-out:path:all output\_files/

#### your target sequence

## #Initialize membrane

-in:file:spanfile **input\_files/3pbl.span**  
-membrane:no\_interpolate\_Mpair  
-membrane:Menv\_penalties  
-rg\_reweight .1  
-restore\_talaris\_behavior

#### only if modeling a membrane protein

## # 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\_bicubic\_interpolation  
-hybridize:stage1\_probability 1.0  
-sog\_upper\_bound 15

#### use ref2015\_cart if soluble protein



# Tutorial

Comparative modeling of D3 receptor with five class A GPCR templates

Four stages:

- I. Setup
- II. Threading
- III. RosettaCM hybridize
- IV. Final model selection

# References

- **Rosetta User Guide & Documentation**

<https://www.rosettacommons.org/docs/latest/Home>

- **Membrane Proteins Documentation**

[https://www.rosettacommons.org/docs/latest/application\\_documentation/Application%20Documentation#Membrane-Proteins](https://www.rosettacommons.org/docs/latest/application_documentation/Application%20Documentation#Membrane-Proteins)

- **RosettaCM: Multi-template**

Yifan Song, et al. (2013). High-Resolution Comparative Modeling with RosettaCM. Structure, 21(10), 1735-1742.