

Practical courses 4 and 5

3D structure prediction

This practical course focuses on the modelling of protein structures by comparative modelling and by fold recognition. You will create different models for the acyl carrier protein of *Rhodospirillum rubrum*. The quality of your models will be evaluated with the global Qmean score and Procheck. These tools are available here:

Qmean: <https://swissmodel.expasy.org/qmean/>

Procheck, via "PDBSum Generate":

<https://www.ebi.ac.uk/thornton-srv/databases/pdbsum/Generate.html>

What is the Qmean score based on and how to interpret its value?

QMEAN is a comprehensive scoring function for model quality assessment. The ranking of the models is based on the QMEAN which reflects the predicted global model reliability ranging from 0 to 1. 1 is the best score. QMEAN, which stands for Qualitative Model Energy ANalysis, is a composite scoring function describing the major geometrical aspects of protein structures. Five different structural descriptors are used. The local geometry is analyzed by a new kind of torsion angle potential over three consecutive amino acids. A secondary structure-specific distance-dependent pairwise residue-level potential is used to assess long-range interactions. A solvation potential describes the burial status of the residues. Two simple terms describing the agreement of predicted and calculated secondary structure and solvent accessibility, respectively, are also included.

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2703985/>

https://www.researchgate.net/publication/5912145_QMEAN_A_comprehensive_scoring_function_for_model_quality_assessment

1. Comparative modelling of the 3D structure of the acyl carrier protein of *Rhodospirillum rubrum*: manual approach

A] Perform a Blast (<http://www.ncbi.nlm.nih.gov/blast/>; use "Protein blast") on the ACP sequence to identify a template to model the structure of this protein (choose the appropriate database to scan with BLAST).

From uniprot, we obtain the ACP fasta format file :

UniProtKB - B6IN76 (ACP_RHOCS)

Display

Entry

Publications

Feature viewer

Feature table

None

Function

Names & Taxonomy

Subcellular location

BLAST

Align

Format

Add to basket

History

Protein | Acyl

Gene | acpP

Organism | Rhodospirillum rubrum

Status | R

View format

Text

FASTA (canonical)

XML

RDF/XML

GFF

ATCC 51521 / SW

Protein inferred from homology¹

Function¹

Carrier of the growing fatty acid chain in fatty acid biosynthesis. UniRule annotation

Pathway: fatty acid biosynthesis

```
>sp|B6IN76|ACP_RHOCS Acyl carrier protein OS=Rhodospirillum rubrum (strain ATCC 51521 / SW) OX=414684 GN=acpP PE=3 SV=1
MSDTAERVKKIVIEHLGVEESKVTESASFIDDLGADSLDTVELVMAFEEFFGIEIPDDAA
EKILTVKDAIDFINQKTAA
```

blastn blastp **blastx** tblastn tblastx

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [Clear](#) [Query subrange](#)

MSDTAERVKKIVIEHLGVEESKYTESASFIDDLGADSLDTVELVNAFEEEFGEIPDDAA
EKILTVKDAIDFINQKTAA

From

To

Or, upload file Aucun fichier sélectionné.

Job Title

splB6IN76[ACP_RHOCS Acyl carrier protein OS=Rhodospirillum...

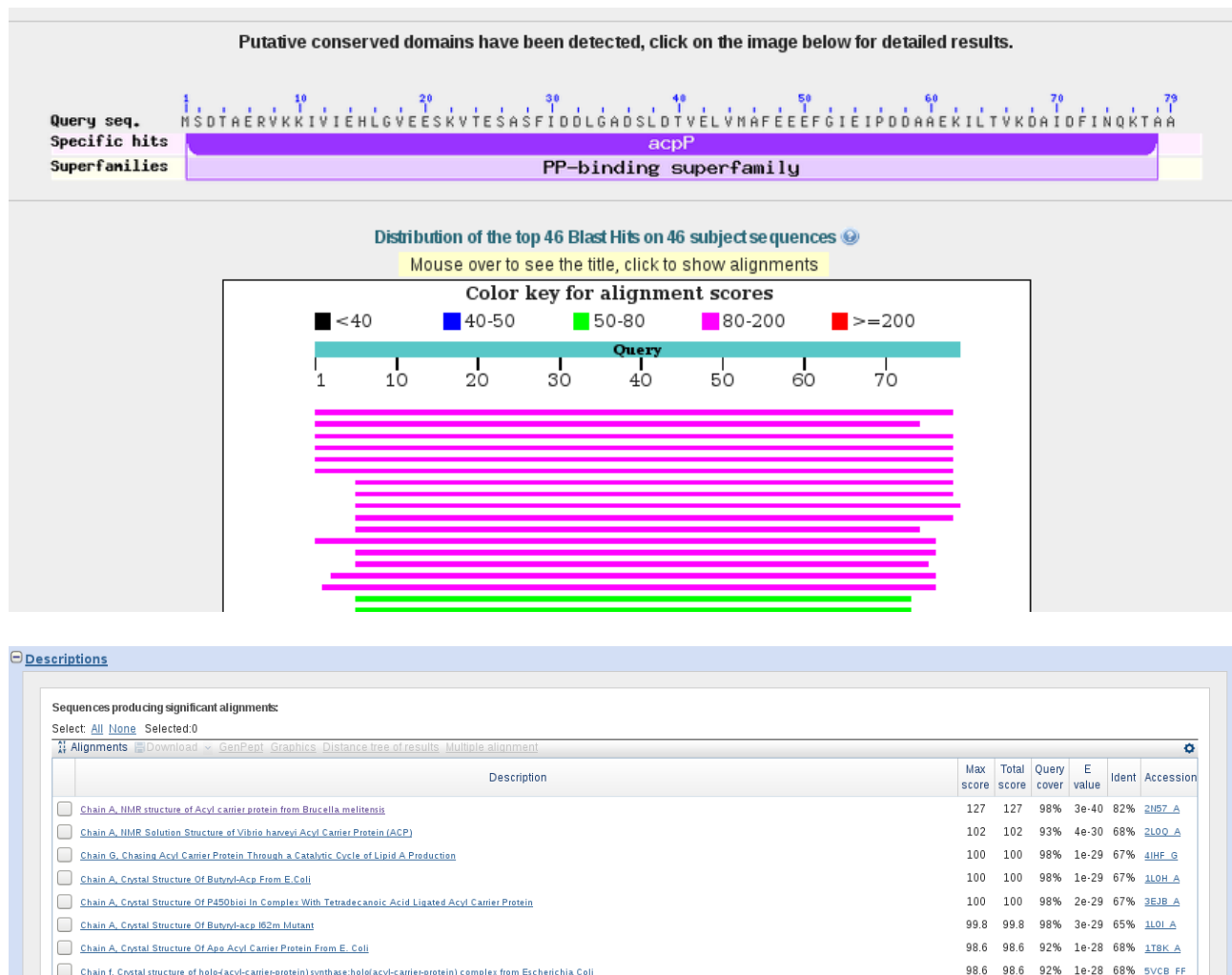
Enter a descriptive title for your BLAST search

☐ Align two or more sequences

Choose Search Set

Database Protein Data Bank proteins(pdb)

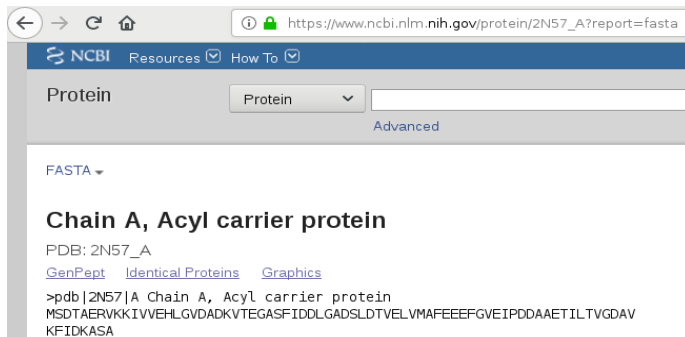
Result of a Protein BLAST



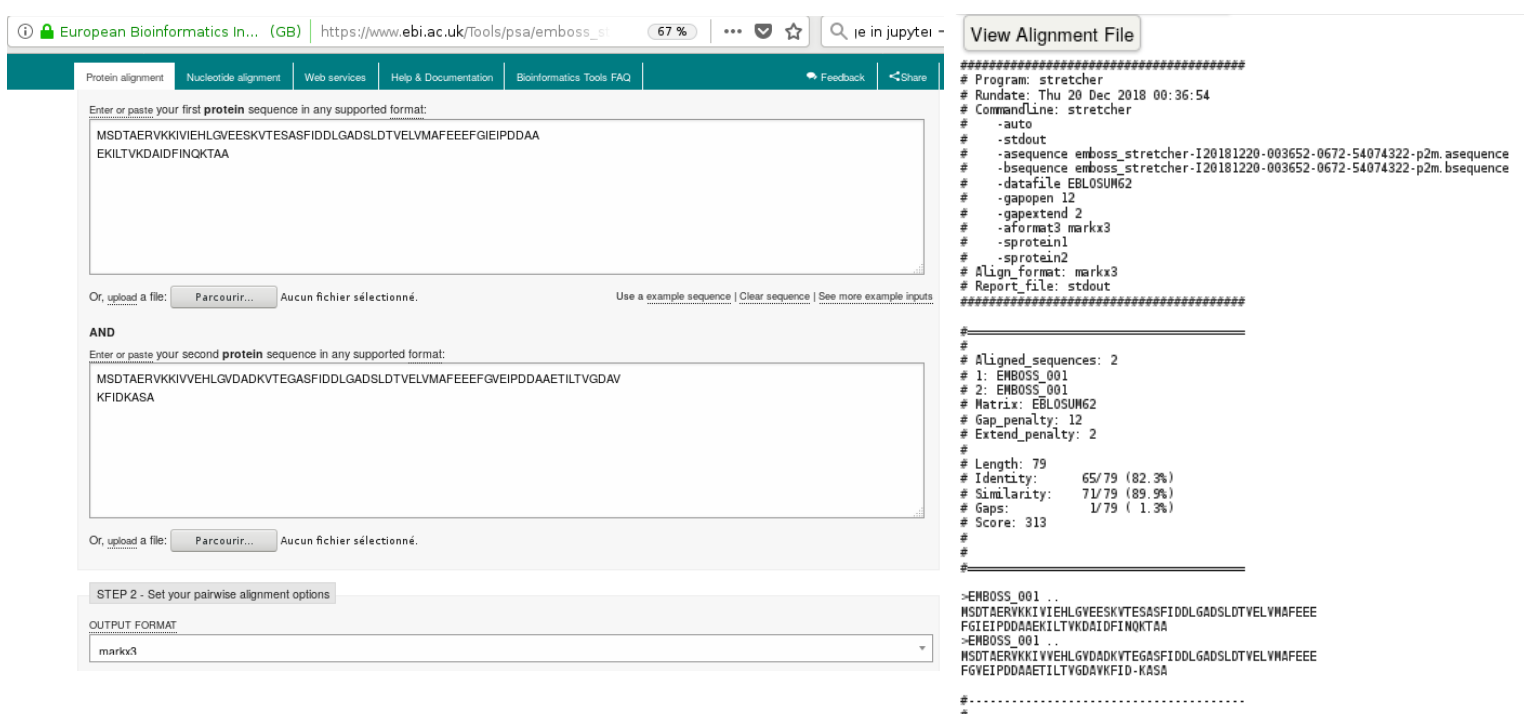
B] Select a template among the hits that have been identified by Blast. For that purpose, take into account the percentage of sequence identity, the percentage of query cover and the

quality of the structure of the template (see the tools used in TP1). Download the sequence of the template from the PDB website.

We selected the first template because of its best results for the identity percentage and for the query cover.



C] Perform a sequence alignment between the template and ACP. For that purpose, use the the Stretcher global alignment program (https://www.ebi.ac.uk/Tools/psa/emboss_stretcher/, choose the "Markx3" output format in the "More options" menu).



European Bioinformatics Institute (GB) | https://www.ebi.ac.uk/Tools/psa/emboss_stretcher/ 67 %

Protein alignment | Nucleotide alignment | Web services | Help & Documentation | Bioinformatics Tools FAQ | Feedback | Share

Enter or paste your first **protein** sequence in any supported format:

```
MSDTAERVKKIVVEHLGVDADKVTGASFIIDL GADSLDTVELVMAFEFEFGVEIPDDAAETILTVGDAVKFIDKASA
```

Or, upload a file: Aucun fichier sélectionné. Use a example sequence | Clear sequence | See more example inputs

AND

Enter or paste your second **protein** sequence in any supported format:

```
MSDTAERVKKIVVEHLGVDADKVTGASFIIDL GADSLDTVELVMAFEFEFGVEIPDDAAETILTVGDAVKFIDKASA
```

Or, upload a file: Aucun fichier sélectionné.

STEP 2 - Set your pairwise alignment options

OUTPUT FORMAT:

View Alignment File

```
#####
# Program: stretcher
# Runday: Thu 20 Dec 2018 00:36:54
# Commandline: stretcher
# -auto
# -stdout
# -asequence emboss_stretcher-I20181220-003652-0672-54074322-p2m.asequence
# -bsequence emboss_stretcher-I20181220-003652-0672-54074322-p2m.bsequence
# -datafile EBL0SUN62
# -gapopen 12
# -gapextend 2
# -aformat3 markx3
# -sprotein1
# -sprotein2
# Align_format: markx3
# Report_file: stdout
#####

#
#
# Aligned sequences: 2
# 1: EMB0SS_001
# 2: EMB0SS_001
# Matrix: EBL0SUN62
# Gap_penalty: 12
# Extend_penalty: 2
#
# Length: 79
# Identity: 65/79 (82.3%)
# Similarity: 71/79 (89.9%)
# Gaps: 1/79 (1.3%)
# Score: 313
#
#
#-----
>EMB0SS_001 ..
MSDTAERVKKIVVEHLGVDADKVTGASFIIDL GADSLDTVELVMAFEFE
FGIEIPDDAAEKILTVKDAIDFINKTAA
>EMB0SS_001 ..
MSDTAERVKKIVVEHLGVDADKVTGASFIIDL GADSLDTVELVMAFEFE
FGVEIPDDAAETILTVGDAYKFIID-KASA
#-----
#-----
```

D] Submit the sequence alignment obtained in the section 1.C. to the Modeller server (<https://toolkit.tuebingen.mpg.de/#/tools/modeller>). The license key to use Modeller is "MODELIRANJE". The sequence alignment must be provided in PIR format. There is on the virtual university a document that explains how to convert the Markx3 format obtained from the sequence alignment into a PIR format that must be submitted to Modeller. Save the PDB file of the model.

ID

Date

Tool

5342200

MODL

Input

3D-Structure

```

>P1;ACP
sequence:ACP:::
MSDTAERVKKIVIEHLGVEESKVTESASFIDDLGADSLDTVELVMAFEEE
FGIEIPDDAAEKILTVKDAIDFINQKTAA*
>P1;2N57
structure:2N57:1 :A:78 :A:::
MSDTAERVKKIVIEHLGVDADKVTESASFIDDLGADSLDTVELVMAFEEE
FGVEIPDDAAETILTVGDAVKFID-KASA*

```

Paste Example

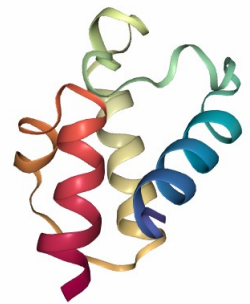
Upload File

Valid PIR alignment.

MODELLER-key is stored in your profile.

5342200_1

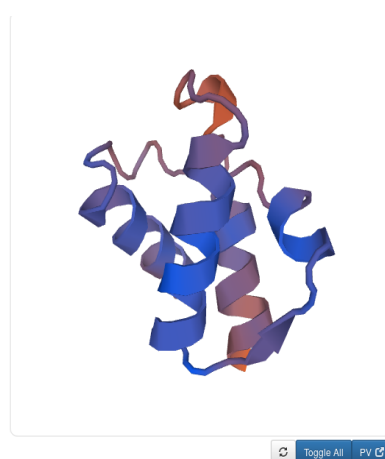
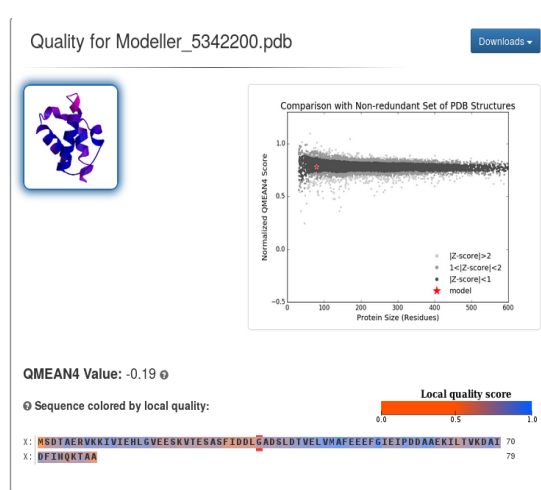
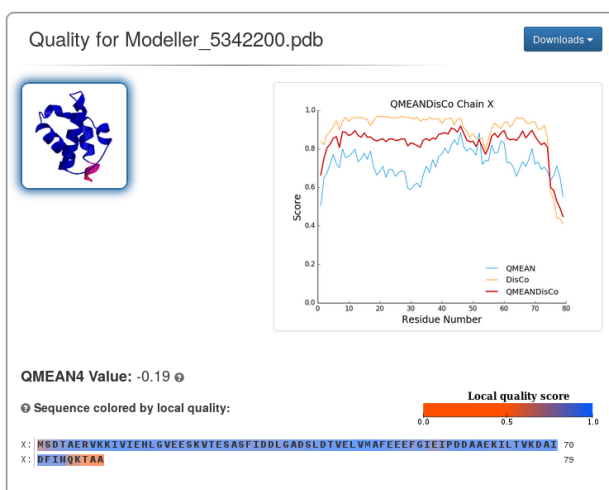
Resubmit Job



We obtain the pdf file.

E] Analyze the quality of this model.

To analyse the quality of a model we can use Qmean (swiss-model) or Procheck.



Qmean= -0.19

The quality of the model based on 6 parameters is good because the Qmean is close to 0. When we calculate the qmean for each aa (between 0 and 1) it must be the closest to 1. When we calculate the global Qmean for all the sequence, it's like a z score and it must be the lowest as possible because it means we obtain a protein score close to the score obtained by high quality proteins.

PROCHECK

Notice: Undefined offset: 1 In /var/www/hlib/work.php on line 166

Notice: Undefined offset: 2 In /var/www/hlib/work.php on line 167

[WHATCHECK](#) • [ERRAT](#) • [Verify3D](#) • [PROVE](#) • [CRYST](#) • [pdbU](#) • [SAVES](#) [?](#)

[New Job](#)

PROCHECK run: **Modeller_5342200.pdb**

PROCHECK Completed

<http://servicesn.mbi.ucla.edu/PROCHECK/?job=30439> | [View Structure](#)

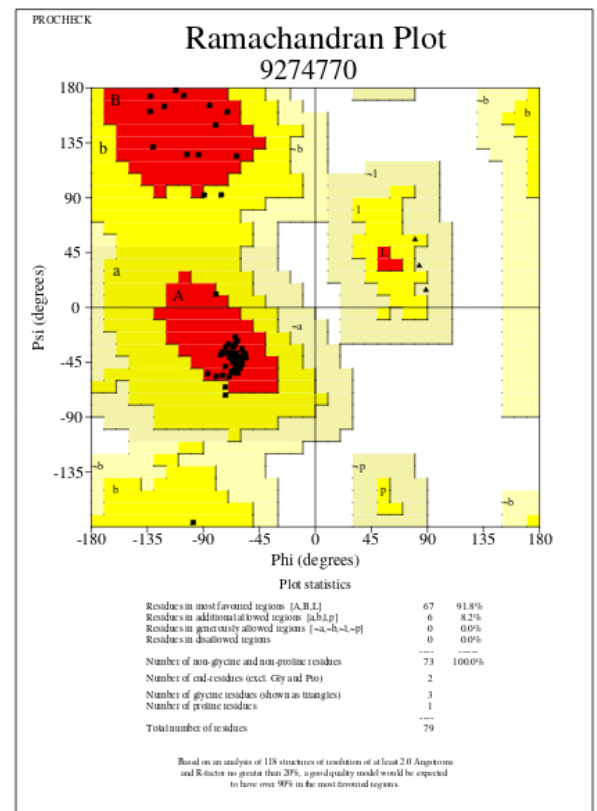
Started: Dec 19th, 2018 at 5:28 PM

Finished: Dec 19th, 2018 at 5:28 PM

Out of 8 evaluations

- **Errors: 0**
- **Warning: 3**
- **Pass: 5**

1. [Main Ramachandran plot](#)
2. [All-residue Ramachandran plots](#)
3. [All-residue chi1-chi2 plots](#)
4. [Main-chain parameters](#)
5. [Side-chain parameters](#)
6. [Residue properties plot](#)
7. [Main-chain bond lengths](#)
8. [Main-chain bond angles](#)
9. [RMS distances from planarity](#)
10. [Distorted geometry](#)
11. [Program Output](#)
12. [Results Summary](#)



We can see that the amino acids (dots) are in the possible domains, it means that the model is good concerning the angular torsion.

2. Comparative modelling of the 3D structure of the acyl carrier protein of *Rhodospirillum centenum*: semi-automatic approach

The HHPred server combined to Modeller (<https://toolkit.tuebingen.mpg.de/#/tools/hhpred>) will be used.

A) Submit the ACP sequence to the HHPred server. Describe the first step performed by HHPred. Several templates are proposed. Compare these templates to those identified with Blast in section 1.B. (sequence identity, quality of the template structure, ...).

Hitlist

Show entries

Search:

Nr	Hit	Name	Probability	E-value	SS	Cols	Target Length
<input type="checkbox"/> 1	6GCS_Q	NUAM protein (E.C.1.6.99.3); NADH dehydrogenase; Complex I, NADH dehydrogenase, Mitochondrion; HET: NDP, SF4, ZMP, CDL, FMN, 3PE; 4.32A (Yarrowia lipolytica)	98.8	1.8e-9	8.9	77	132
<input type="checkbox"/> 2	2CGQ_A	ACYL CARRIER PROTEIN ACPA; RV0033, ACYL CARRIER PROTEIN, PROTEIN; 1.83A (MYCOBACTERIUM TUBERCULOSIS) SCOP: 1.1.1.1, a.28.1.0	98.76	2.3e-9	8.4	78	113
<input type="checkbox"/> 3	6G2J_T	NADH-ubiquinone oxidoreductase chain 3 (E.C.1.6.5.3); Complex I, mitochondria, proton pump; HET: ADP, PC1, AME, AYA, CDL, SF4, 3PE, FMN, FME, EH2, 2MR, NDP; 3.3A (Mus musculus)	98.75	4.8e-9	9.6	78	156
<input type="checkbox"/> 4	3TEJ_A	Enterobactin synthase component F (E.C.2.7.7.-); NONRIBOSOMAL PEPTIDE, THIOESTERASE, CARRIER DOMAIN; HET: UF0; 1.9A (Escherichia coli)	98.68	4.1e-9	8.9	72	329
<input type="checkbox"/> 5	5ISW_A	Gramicidin S synthase 1 (E.C.5.1.1.11); Epimerization domain, NRPS, gramicidin S; HET: GOL; 1.75A (Brevibacillus brevis)	98.67	3.4e-9	8.7	74	573
<input type="checkbox"/> 6	2JGP_A	TYROCIDINE SYNTHETASE 3; MULTIFUNCTIONAL ENZYME, ANTIBIOTIC BIOSYNTHESIS, CONDENSATION; HET: SO4, DIO; 1.85A (BREVIBACILLUS BREVIS)	98.64	5.2e-9	8.7	75	520

Putative uncharacterized protein; helical bundle, acyl

HHpred is a method for sequence database searching and structure prediction that is as easy to use as BLAST or PSI-BLAST and that is at the same time much more sensitive in finding remote homologs.

Here, we see that we obtain different results compared with BLAST because in this case, we rely on the sequence homology (derived from a common ancestor) and not on the sequence identity. Thus, the probability above corresponds to the probability to have homology between 2 sequences.

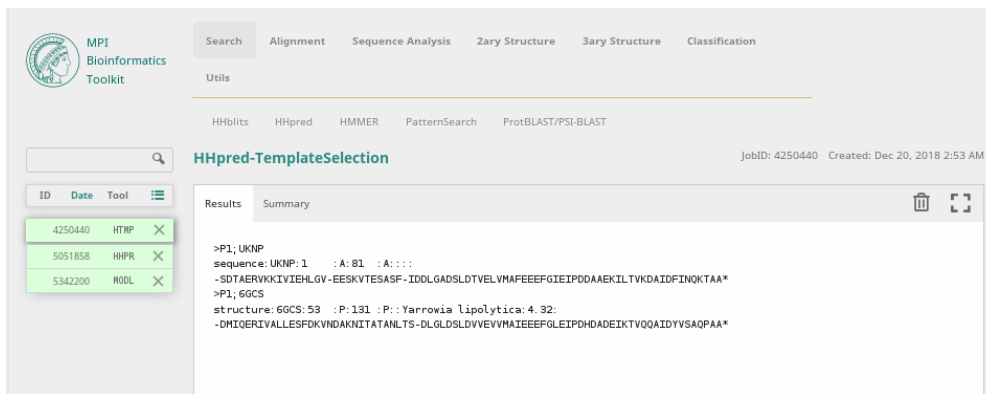
So the first step here is to determine homology and align the sequences using this probability. We determine a profile processing multiple alignments with all the sequences belonging to a family and choosing the more conserved amino acids. And we compare our sequence with this profile.

Compared with BLAST, we can see that we don't have good sequence identities because homology is more important.

B] Select one template and click on "Model using selection". HHpred will align the 2 sequences. Then click on "Forward to Modeller", use the Modeller-key "MODELIRANJE" and click on "Submit job".

Save the PDB file of the model (you will find a "Download PDB file" tab).

We selected here the first template.



MPI Bioinformatics Toolkit

Search Alignment Sequence Analysis 2ary Structure 3ary Structure Classification

Utils

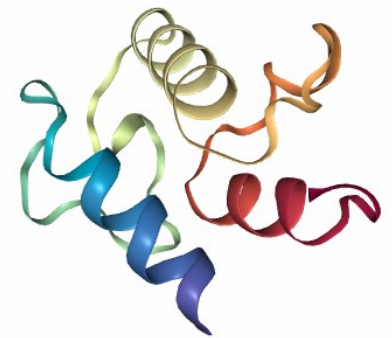
HHblits HHpred HMMER PatternSearch ProtBLAST/PSI-BLAST

JobID: 4250440 Created: Dec 20, 2018 2:53 AM

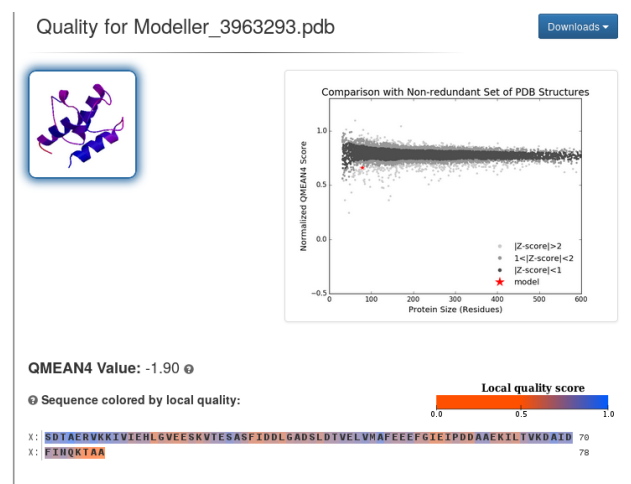
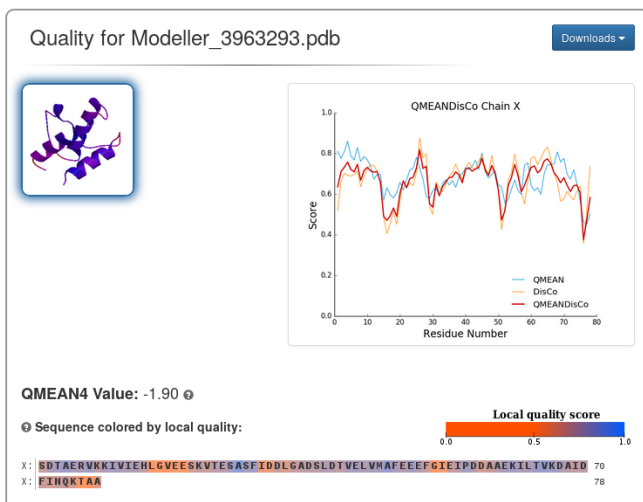
HHpred-TemplateSelection

Results Summary

>P1; UKNP
sequence: UKNP:1 :A: 81 :A: : :
-SDTAERVKKIVIEHLGV-EESKVTESASF-IDDLGADSLDTVELVMAFEFEFGIEIPDDAAEKILTVKDAIDFINKTAA*
>P1; 6GCS
structure: 6GCS:53 :P:131 :P::Yarrowia lipolytica:4.32:
-DMIQERIVALLSFQKVDANKITATANLTS-DLGLDSLQVVEVYMAIEEFGLIEIPDHDAEIKTVQQAIDYVSAQPA*
*



C] Analyze the quality of this model.



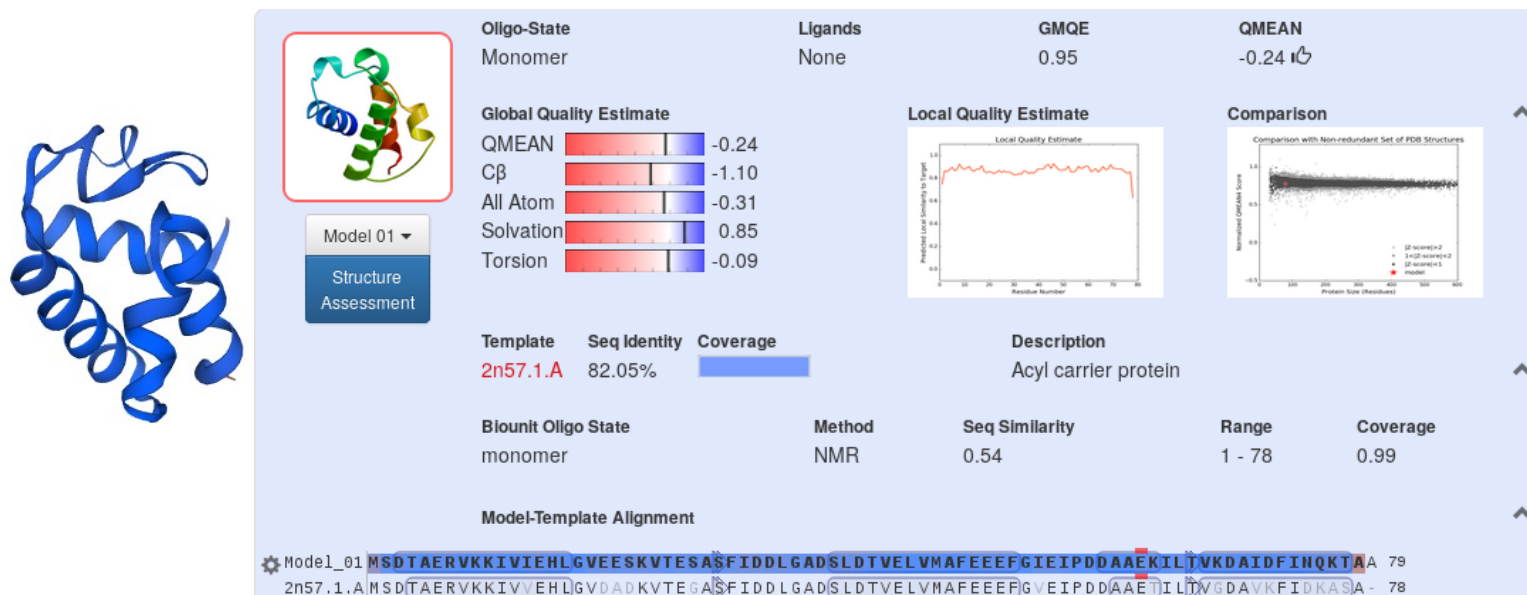
The quality of the model is bad, QMEAN = -1,9

3. Comparative modelling of the 3D structure of the acyl carrier protein of *Rhodospirillum centenum*: automatic approach

Use the SwissModel server (<https://swissmodel.expasy.org>) to build a model of ACP. Use the "Build Model" tab.

Identify the template that has been used. Save the PDB file of the model.

Analyze the quality of this model.



QMEAN= -0,24

We obtain a different Qmean compared with the previous result because the alignment and/or the manner to model can explain this difference. But the quality of the model stay relatively good.

4. Modelling of the 3D structure of the acyl carrier protein of *Rhodospirillum centenum*: by a fold recognition approach

A] Submit the sequence of ACP to Sparks-x (<http://sparks-lab.org/yueyang/server/SPARKS-X/>). What is the best template according to Sparks-x? Download the model obtained with this template (click on "MOD1" to download the model).

Analyze the quality of this model.

SPARKS-X result for unknown

Your input sequence:
>sp|B6IN76|ACP_RHOC5 Acyl carrier protein OS=Rhodospirillum centenum (strain ATCC 51521 / SW) OX=414684 GN=acpP PE=3 SV=1
MSDTAERVKKIVIEHLGVEESKVTESASFIDDLGADSLDTVELVMAFEFEFGIEIPDDAA
EKILTVKDAIDFINQKTA

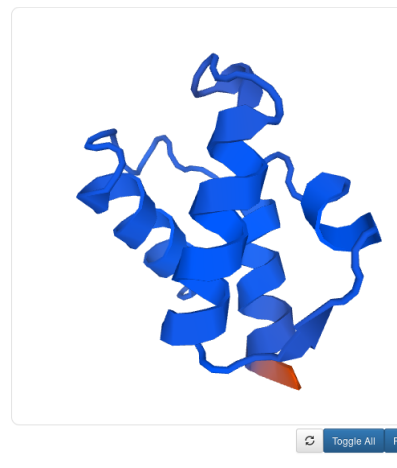
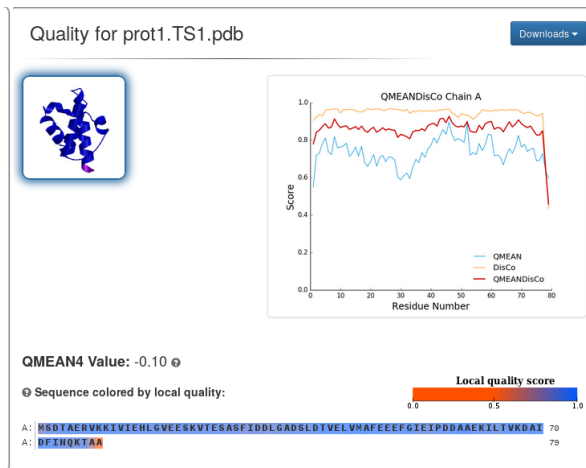
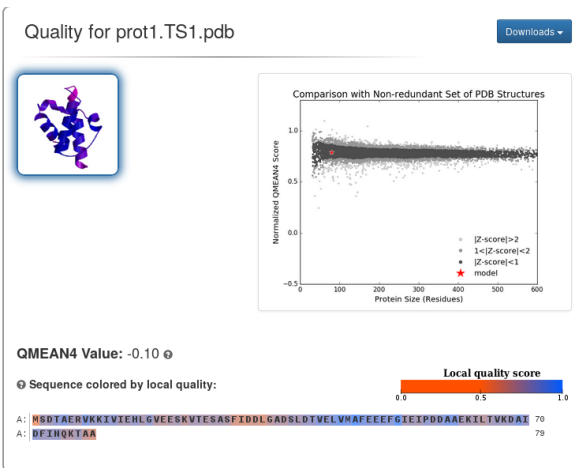
TOP 10 MATCHES (Zscore>6 means significant matching)

RANK TEMPLATE ZSCORE SEQID ALIGNMENT MODEL

1 2n57A 24.71 82.1% [ALT1] [MOD1]
2 3gz1A 24.53 48.1% [ALT2] [MOD2]
3 1x3oA 23.89 54.4% [ALT3] [MOD3]
4 3ejbA 23.71 65.4% [ALT4] [MOD4]
5 2ehsA 23.27 62.7% [ALT5] [MOD5]
6 2qnwA 23.13 49.4% [ALT6] [MOD6]
7 2m5rA 22.61 45.6% [ALT7] [MOD7]
8 2n50A 22.10 40.5% [ALT8] [MOD8]
9 2cnrA 21.65 38.0% [ALT9] [MOD9]
10 4dxeH 21.54 60.8% [ALT10] [MOD10]

Thu Dec 20 12:20:35 2018

Please contact Yuedong_Yang if any questions/comments.



Qmean= - 0.10

We can see here a good quality of the model because the Qmean is near 0 which means a low z score. Moreover, the Qmean for each aa is close to 1.

B] Submit the sequence of ACP to genTHREADER (<http://bioinf.cs.ucl.ac.uk/psipred/>, choose only GenTHREADER-Rapid Fold Recognition). Select the first template proposed and build a model. Download the PDB file of the model.

Choose Prediction Methods

☐ PSIPRED v3.3 (Predict Secondary Structure)

☐ pGenTHREADER (Profile Based Fold Recognition)

☐ BioSerf v2.0 (Automated Homology Modelling)

☐ FFPred 3 (Eukaryotic Function Prediction)

☐ MEMPACK (SVM Prediction of TM Topology and Helix Packing)

☐ DomSerf v2.0 (Automated Domain Modelling by Homology)

☐ DISOPRED3 (Disorder Prediction)

☐ MEMSAT3 & MEMSAT-SVM (Membrane Helix Prediction)

☐ DomPred (Protein Domain Prediction)

☒ GenTHREADER (Rapid Fold Recognition)

☐ pDomTHREADER (Fold Domain Recognition)

[Help...](#)

Input Sequence (Single sequence or Multiple Sequence alignments; as raw sequence or fasta format)

>sp|B6N76|ACP_RHOCS Acyl carrier protein OS=Rhodospirillum rubrum (strain ATCC 51521 / SW) OX=414684 GN=acpP PE=3 SV=1
MSDTAERVKKIVIEHLGVEESKVTESASFIDDLGADSLDTVELVMAFEFFGIEIPDDAAEKILTVKDAIDFTHQKTA

[Help...](#)

If you wish to test these services follow this link to retrieve a [test fasta sequence](#).

Submission Details

Email Address for job completion alert (optional)

[Help...](#)

Password (only required for licenced commercial e-mail addresses)

[Help...](#)

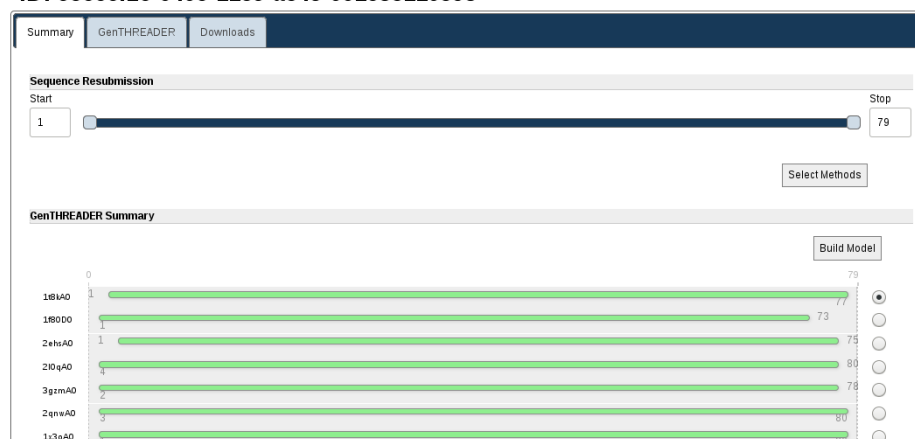
Short identifier for submission

PSIPRED

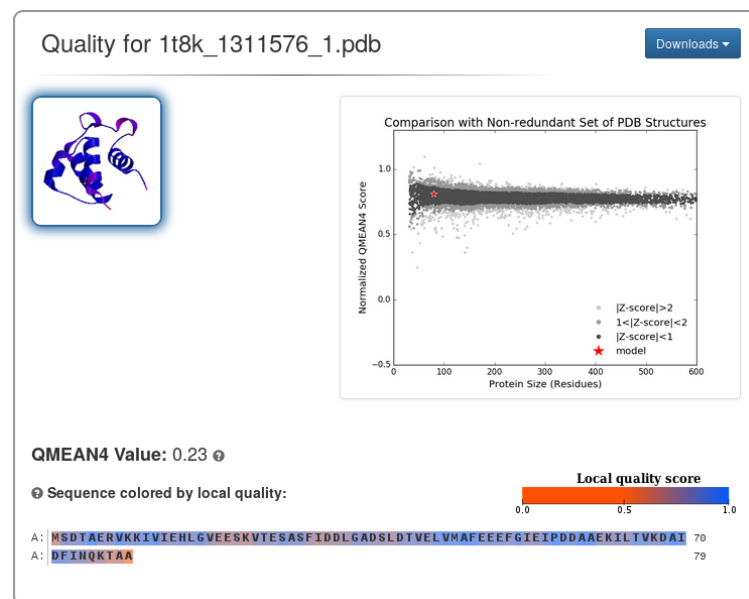
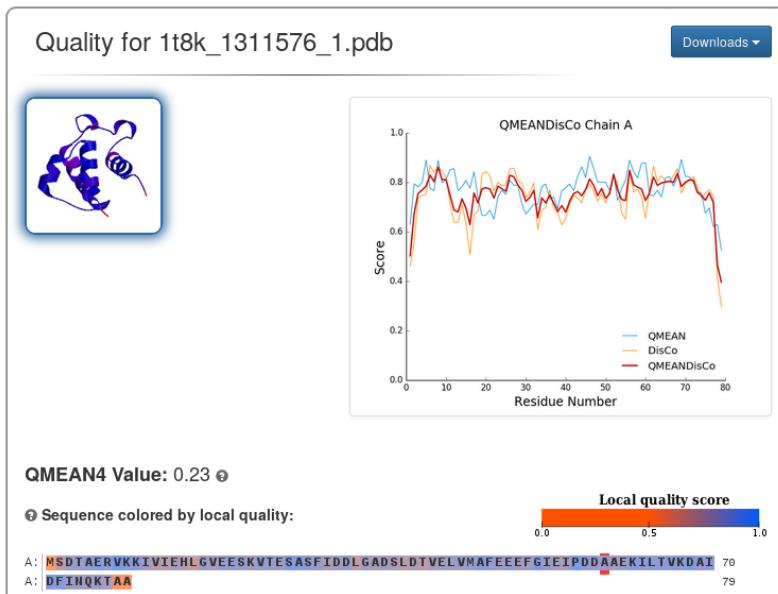
[Help...](#)

Sequence analysis results for job: PSIPRED

ID: 95666f26-040c-11e9-ae45-00163e110593



C] Analyze the quality of this model.

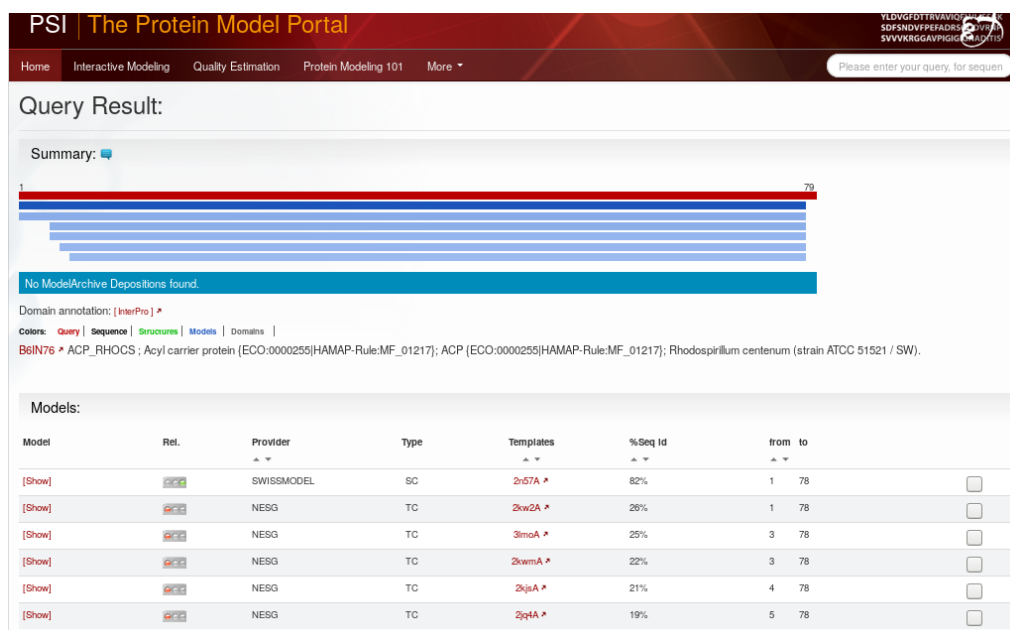


Qmean=0.23 , it's a good model.

5. Protein Model Portal


The Protein Model Portal (<http://www.proteinmodelportal.org>) is a website that collects models that are already available for a protein.

A] Search the Protein Mode Portal for existing models for ACP (Uniprot code B6IN76). How many models are there?



There are 6 models of proteins corresponding to ACP.

B] Select all the models and click on "Start analysis of structural variability" to compare the structure of the different models. Analyze the results.
 C] Download the PDB file of model 5. For that purpose, click on "Show", then on "Model provided by NESG", then on "download: model". Analyze the quality of this model.



Northeast
Structural
Genomics
Consortium

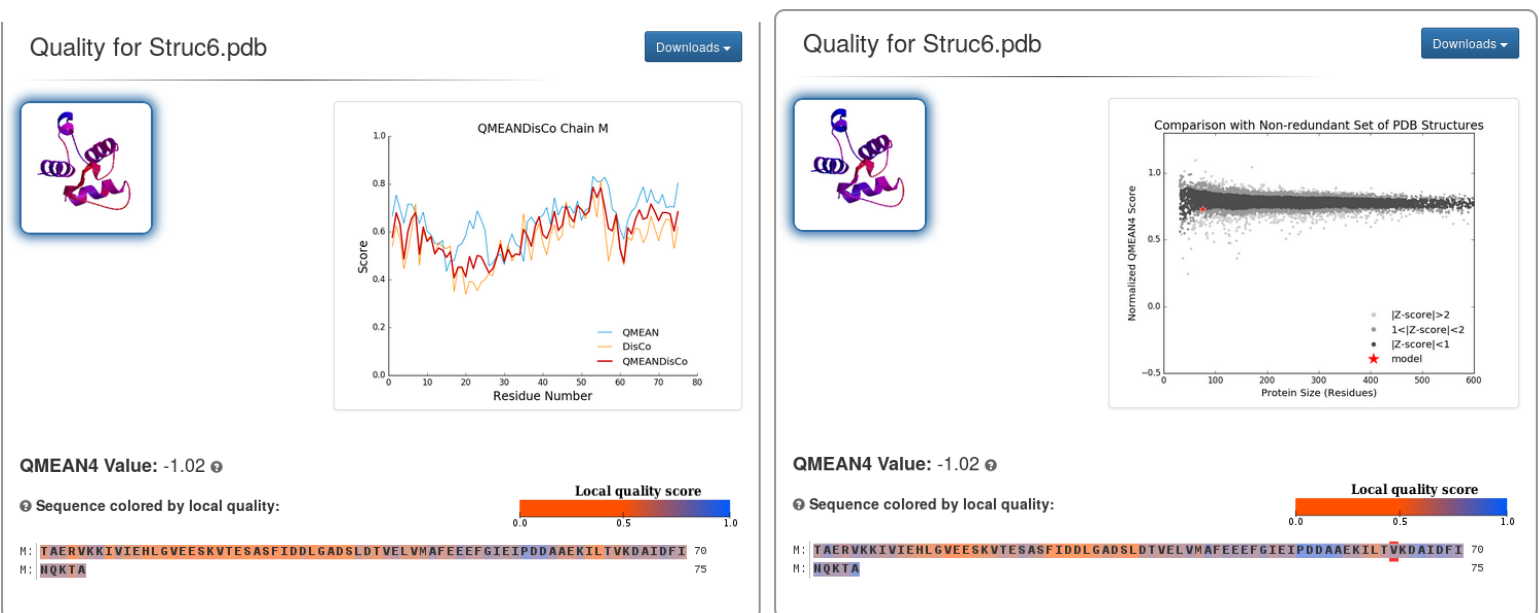
[NESG DATABASES](#)
[DATABASE HELP](#)
[NESG DATABASE](#)
[PDB-60 DATABASE](#)
[MURRAY LAB](#)
[HOME](#)
[LAB MEMBERS](#)
[RESEARCH](#)
[DATA/SOFTWARE](#)
[CONTACT US](#)
[PHARMACOLOGY](#)
[HOME](#)
[NESG](#)
[HOME](#)
[WEB RESOURCES](#)
[ARABIDOPSIS 2010](#)
[NATIONAL SCIENCE FOUNDATION](#)

NESG MODELS DATABASE

The similarity search retrieved the following modeled proteins from the NESG database

Target ID	Template	Model Quality(pG)	Target/Template Sequence Identity	E Value	Target Coverage	Species	Target Description
ACI98973.1	2KJS	0.91	21%	3.4e-32	95%	Rhodospirillum centenum	acyl carrier protein Full=Acyl carrier protein; Short=ACPacyl carrier protein

Then, click on « download model » in the bottom right corner and open it in the QMEAN website.



Qmean= -1.02

This model is not very good regarding the z score between 1 and 2 and the qmean per residu not close to 1.

6. Comparison of the models

You have now 6 models for ACP: 3 models obtained by comparative modelling, 2 models obtained by fold recognition and 1 models from the Protein Model Portal.

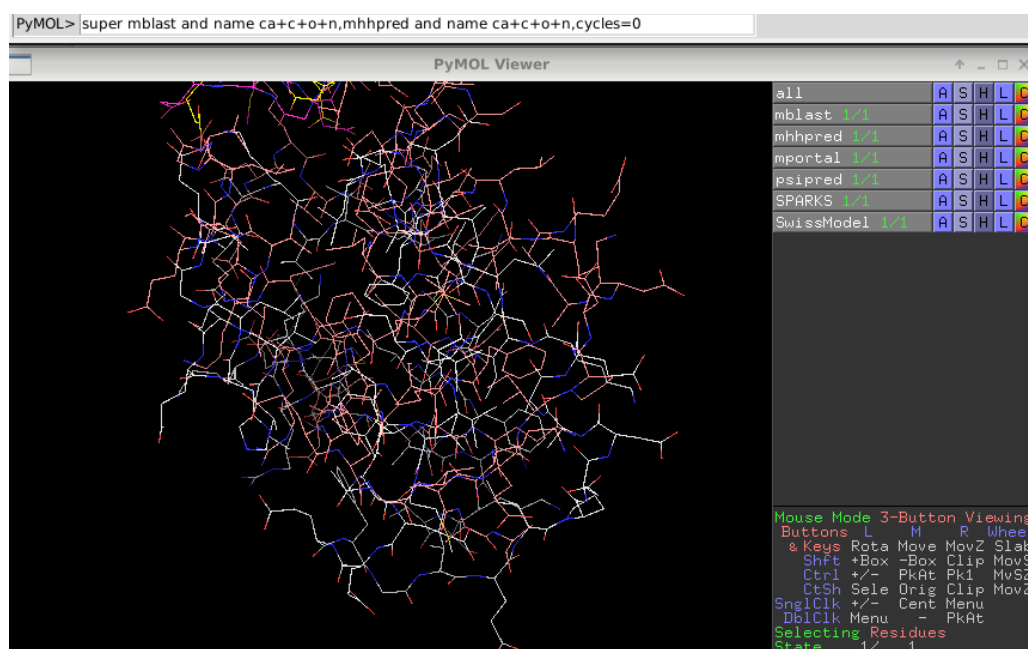
A] Compare the quality of these models.

	Couleur Pymol	Qmean	Type of technic
Comparative modeling			
Manual_BLAST	Vert	-0.19	Manuel
HHpred	Bleu ciel	-1.9	Semi automatique
Swiss	Mauve	-0.24	Automatique
Fold recognition			
Sparks-x	Jaune	-0.10	Automatique
Psipred	blanc	0.23	Automatique
Portal			
5model	Rose pale	-1.02	Automatique

B] Open the 6 PDB files of the models in a same window in Pymol.

Compute the rmsd between all pairs of structure. For that purpose, you can use the "super" command in pymol. To superimpose the main chain atoms of struc1 to those of struc2 and to compute the rmsd, for instance, use the command "super struc2 and name ca+c+o+n,struct1 and name ca+c+o+n,cycles=0". Read the documentation about the "super" command. Note that Python can be used to write scripts in Pymol or to define new functions. Writing a script could be a good idea to perform this repetitive task. You will find a lot of information about this on internet (Pymol tutorials, existing scripts, ...).

Group the models according to their structural similarity.



Comparison		RMSD
Manual-BLAST	HHpred	2.311
	Swiss	0.323
	Sparks-x	0.482
	Psipred	1,765
	5model	2,433
HHpred	Swiss	2,326
	Sparks-x	2.55
	Psipred	2.423
	5model	3.498

Swiss	Sparks-x	0.234
	Psipred	2.004
	5model	2.337
Sparks-x	Psipred	1.511
	5model	2.295
Psipred	5model	2.570

We realize that the models with the lowest RMSD with each other have the best values of qmean (Manual-BLAST, Sparks-x and Swiss). There is a link between the quality and the value of RMSD associated to this models.

C] Use the "super" command, to superimpose all the models on one of the 6 models and to identify the regions where the structural differences are the largest.

PDBefold→ to identify regions which are differente because we have the value of RMSD per aa between two different models. Thus we can identify this particular regions with high RMSD.

We can also observe the superimposition on Pymol, When we do it manually, we realize that the biggest differences are in the loops. This is totally normal considering that the loops are the most variable part and thus, the most difficult to model.



Manual_BLAST vs HHPRED

When superimposing, we compare the structures of the models to identify regions that are similar in the different models. In general, the secondary structures are well superimposed but loops (regions more variable) aren't so well.

It is also important to evaluate the quality of the models by the Qmean and to put it in relation with the superimposition. Can localize regions with higher energy on the models to see if it corresponds to loops for example.