Universitas Esa Unggul

Smart, Creative and Entrepreneurial

IBT 432 Aplikasi Bioinformatika

Protein modelling III: Homology modelling

Riza Arief Putranto

Pertemuan X – 1 Desember 2018

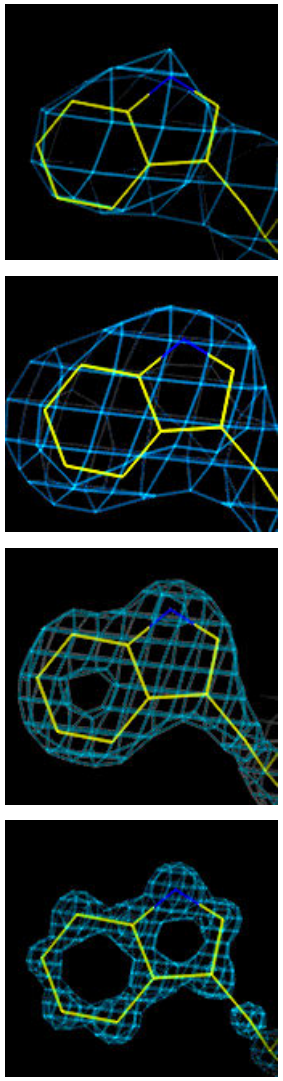
Prodi Bioteknologi

Fakultas Ilmu-Ilmu Kesehatan

Rencana Perkuliahan

1. ~~Kontrak belajar dan pengenalan bioinformatika aplikatif~~
2. ~~Database sekuen dan analisis genomika~~
3. ~~Anotasi sekuen ke genom - Praktik~~
4. ~~Analisis komparasi genomika I~~
5. ~~Analisis komparasi genomika II~~
6. ~~Analisis komparasi genomika III~~
7. ~~Analisis komparasi genomika – Praktik~~
8. ~~Protein modelling I~~
9. ~~Protein modelling II~~
10. Protein modelling III
11. Protein modelling - Praktik
12. Visualisasi protein modelling
13. Visualisasi protein modelling - Praktik
14. Presentasi mahasiswa

Remember the resolution importance



4 Å

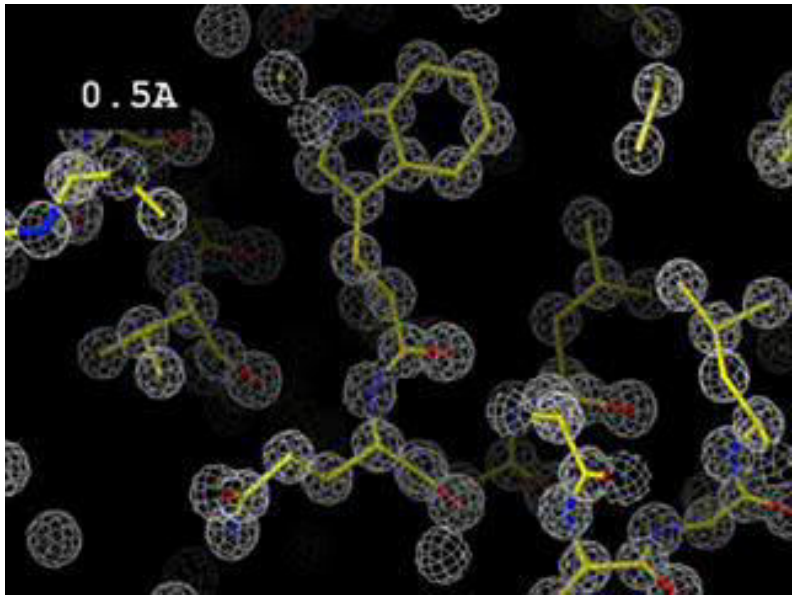
3 Å

2 Å

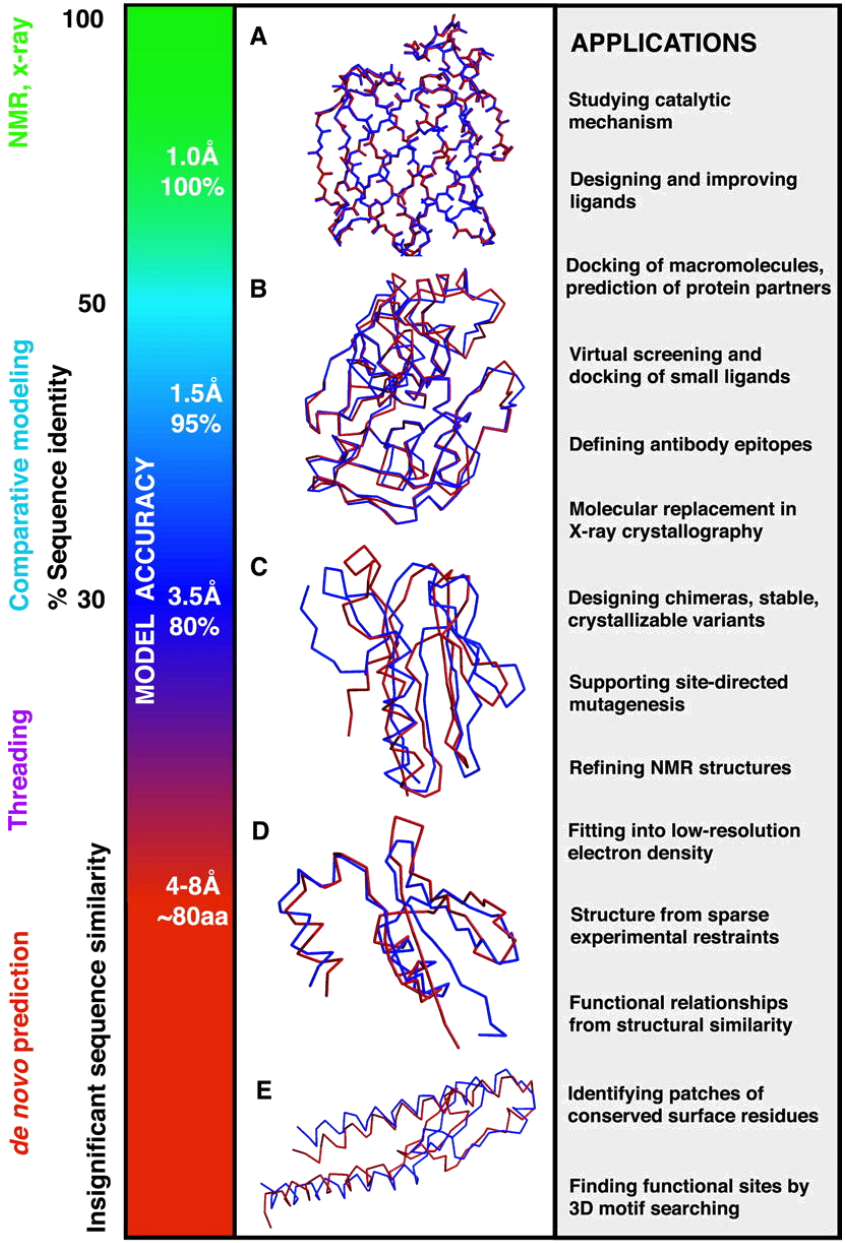
1 Å

low

high



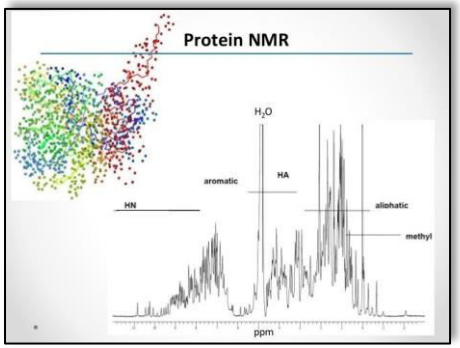
Remember the resolution importance



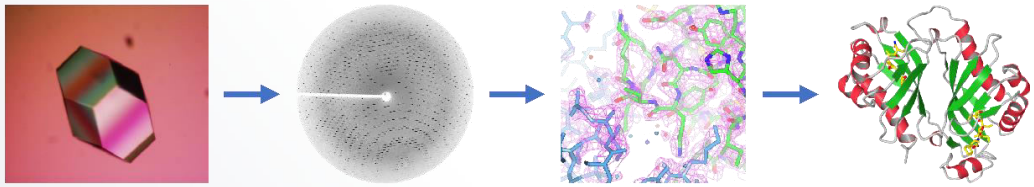
Protein structure gap

Experimental protein structure solution (eg. by **NMR** or **X-Ray crystallography**) is **labor** intensive and **expensive**.

For the majority of proteins in any given proteome, experimental structures are not available.



1. Is it possible to **predict** 3-dimensional protein structures **computationally**?
2. Which computational methods are **feasible** and applicable in a life science research context?



Homology modeling

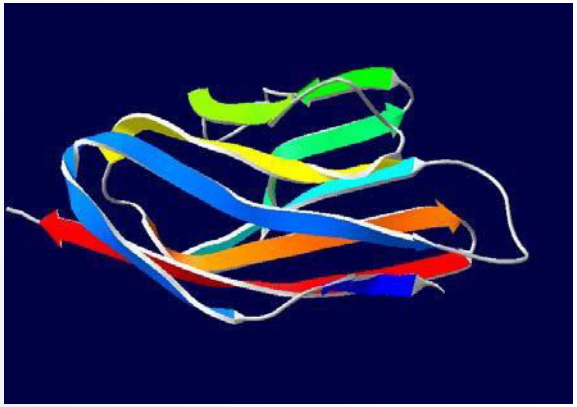
“The biology perspective”

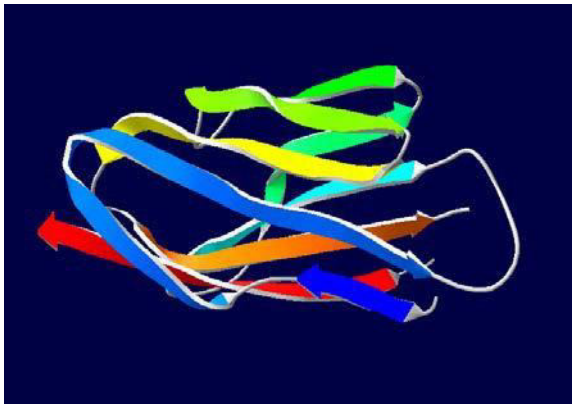
Homologous proteins have **evolved** by molecular evolution from a common ancestor over millions of years. If we can establish **homology to a known protein**, we can **predict aspects of structure and function** of a protein by similarity - **Charles Darwin**

Darwin's evolution of protein structures

Protein structure is better conserved than sequence

Similar Sequence = Similar Structure

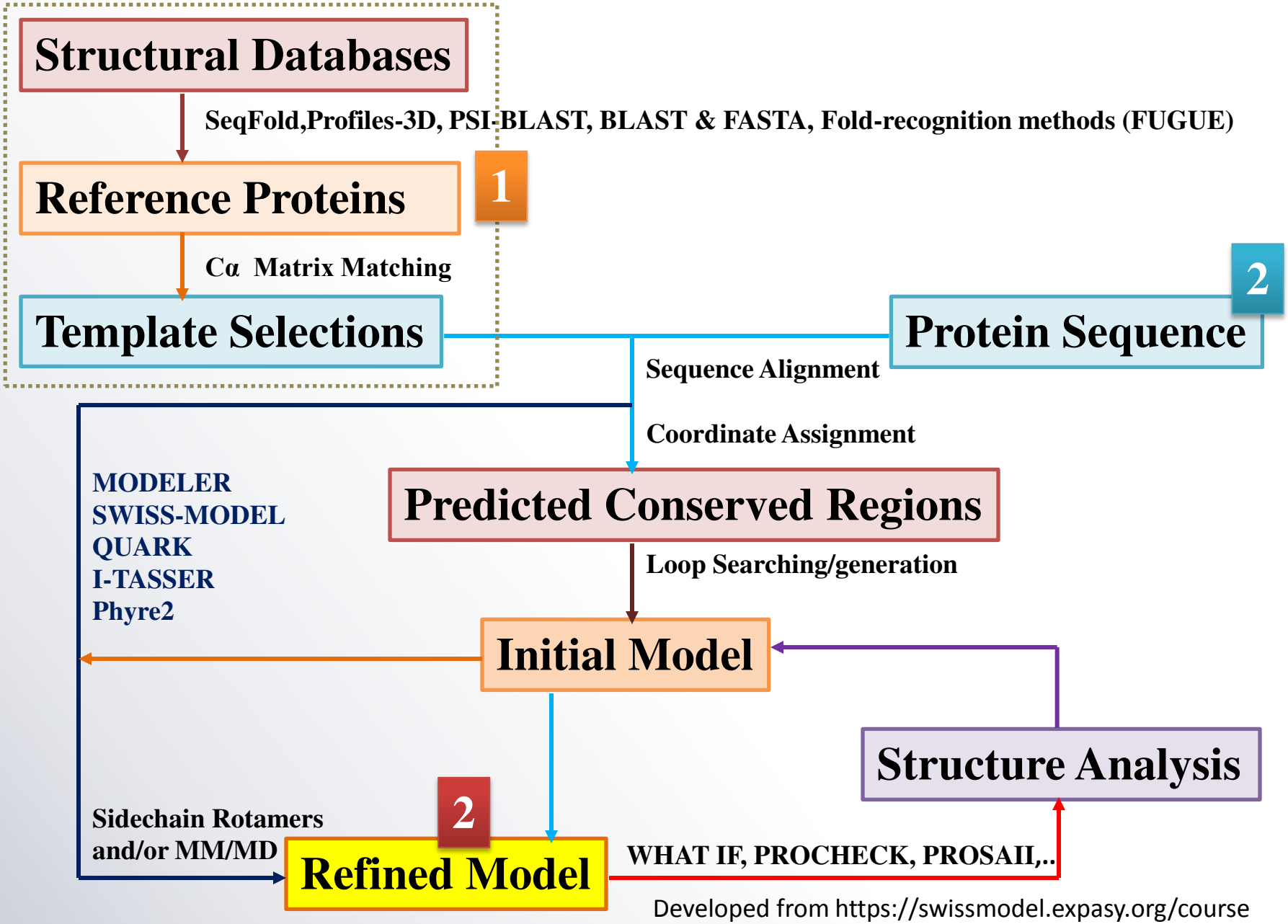




Homology modeling = Comparative protein modeling

Idea: **Using experimental 3D-structures of related family members (templates) to calculate a model for a new sequence (target).**

The principal of homology modeling



Developed from https://swissmodel.expasy.org/course

Protein modeling tools

SWISS-MODEL-Hhpred

Server for homology detection and structure prediction by **HMM-HMM comparison**.

I-Tasser

I-TASSER is a server for protein structure and function predictions. 3D models are built based on multiple-threading alignments by **LOMETS** and **iterative TASSER assembly simulations**.

QUARK

QUARK is a computer algorithm for **ab initio protein structure** prediction and protein peptide folding, which aims to construct the correct protein 3D model from amino acid sequence only. QUARK models are built from small fragments (1-20 residues long) by replica-exchange Monte Carlo simulation under the guide of **an atomic-level knowledge-based force field**.

M4T

Comparative Modelling using a combination of **multiple templates** and **iterative optimization** of alternative alignments.

Modeller

Software for homology or comparative modeling of protein three-dimensional structures. MODELLER implements comparative protein structure modeling by **satisfaction** of **spatial restraints**.

ModWeb

A web server for automated comparative modeling that relies on **PSI-BLAST**, **IMPALA** and **MODELLER**.

Phyre2

A fold recognition server for predicting the structure and/or function of your protein sequence.

SWISS-MODEL

- **SWISS-MODEL** is a **web-based integrated service** dedicated to protein structure homology modelling.
- Building a homology model comprises **four main steps**: (1) identification of structural **template(s)**, (2) **alignment** of target sequence and template structure(s), (3) **model-building**, and (4) **model quality evaluation**.
- Modelling modes:
  - ✓ **Automated -** **requires the amino acid sequence or the UniProtKB accession code**
  - ✓ **Alignment -** **if the template protein is known**
  - ✓ **Project -** **visual inspection and manual manipulation**

Swiss Institute of Bioinformatics

SWISS-MODEL Template Library (SMTL)

- SMTL **aggregates information** of experimental structures from the **Protein Data Bank** and augments it with derived information. When a new structure is released by the PDB, the coordinates and accompanying information are processed and imported into the template library.
- The current SMTL contains:
  - ✓ **548.438 chains**
  - ✓ **94.303 unique SEQRES sequences**
    (primary sequence of the polymeric molecules present)
  - ✓ 223.434 biounits

RCSB PDB

SIB
Swiss Institute of Bioinformatics

SWISS-MODEL Template Library (SMTL)

Enter PDB ID or SMTL ID: Submit

SMTL Update: 2018-04-12

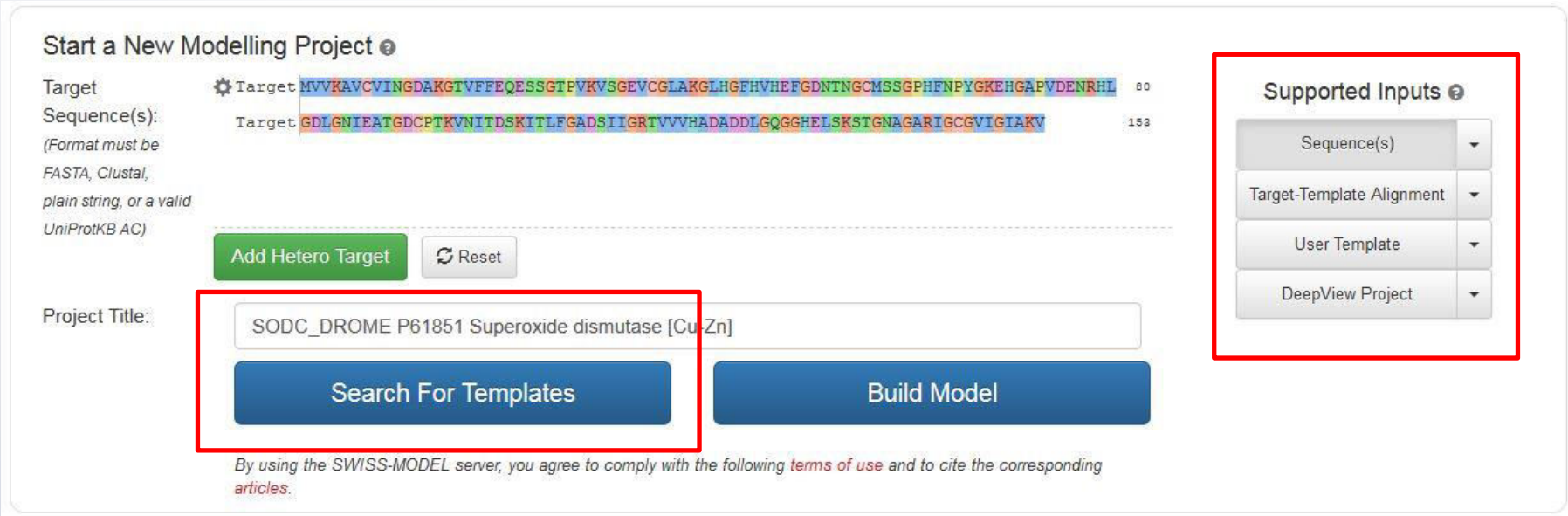
- New Entries

2009 H1N1 PA Endonuclease in complex with RO-7	5vpt
2009 H1N1 PA Endonuclease in complex with RO-7 and Magnesium	5vrj
A New Class of Beta-glucosidase inhibitor	5ost
ARABIDOPSIS THALIANA GSTU23, GSH bound	6ep7
ARABIDOPSIS THALIANA GSTU23, reduced	6ep6
AlfA from B. subtilis plasmid pLS32 filament structure at 3.4 A	6f95
Alpha-1,6-mannosyl-glycoprotein 2-beta-N-acetylglucosaminyltransferase with Bound Acceptor	5vcs
Alpha-1,6-mannosyl-glycoprotein 2-beta-N-acetylglucosaminyltransferase with bound UDP and Manganese	5vcm
Alpha-1,6-mannosyl-glycoprotein 2-beta-N-acetylglucosaminyltransferase with bound uranium dioxide	5vcr
Aminoglycoside Phosphotransferase (2")-Ia S376N mutant in complex with GMPPNP and Magnesium	6ch4
Aminoglycoside Phosphotransferase (2")-Ia in complex with GMPPNP, Magnesium, and Amikacin	6cgd
Aminoglycoside Phosphotransferase (2")-Ia in complex with GMPPNP, Magnesium, and Arbekacin	6cgg
Aminoglycoside Phosphotransferase (2")-Ia in complex with GMPPNP, Magnesium, and Dibekacin	6cav
Aminoglycoside Phosphotransferase (2")-Ia in complex with GMPPNP, Magnesium, and Lividomycin moieties	6cey

SIB
Swiss Institute of Bioinformatics

Chose your modelling mode

- You can either **paste the protein sequence** or **provide the UniprotKB** of your target sequence in the input form.
- To search for available template structures, click on the "**Search for Templates**" button



Swiss Institute of Bioinformatics

Your template results

- Build model by **selecting your template(s)**
- Chose the **best sequence similarity** (above 30% prefered)

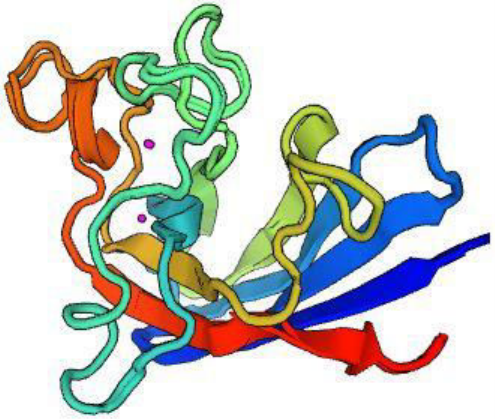
Template Results

Templates | Quaternary Structure | Sequence Similarity | Alignment of Selected Templates | More

	Name	Title	Coverage	Identity	Method	Oligo State	Ligands
☑ ✓	3l9e.1.A	Superoxide dismutase [Cu-Zn]		68.63	X-ray, 2.0Å	homo-dimer ✓	2 x ZN
☑	3l9y.1.A	Superoxide dismutase [Cu-Zn]		67.97	X-ray, 1.8Å	homo-dimer ✓	2 x ZN, 2 x CU
☑	3l9e.1.A	Superoxide dismutase [Cu-Zn]		67.97	X-ray, 2.0Å	homo-dimer ✓	2 x ZN
☑	3l9y.1.A	Superoxide dismutase [Cu-Zn]		67.32	X-ray, 1.8Å	homo-dimer ✓	2 x ZN, 2 x CU
☐	2zky.1.A	Superoxide dismutase [Cu-Zn]		61.84	X-ray, 2.4Å	homo-dimer ✓	2 x ZN
☐	4b3e.1.A	SUPEROXIDE DISMUTASE [CU-ZN]		61.84	X-ray, 2.1Å	homo-dimer ✓	2 x ZN, 2 x CU
☐	1n18.1.A	Superoxide dismutase [Cu-Zn]		61.84	X-ray, 2.0Å	homo-dimer ✓	2 x ZN, 2 x CU1
☐	2zky.1.A	Superoxide dismutase [Cu-Zn]		61.18	X-ray, 2.4Å	homo-dimer ✓	2 x ZN
☐	3ltv.1.A	Superoxide dismutase [Cu-Zn],Superoxide dismutase [Cu-Zn]		61.33	X-ray, 2.5Å	homo-dimer ✓	2 x ZN
☐	4b3e.1.A	SUPEROXIDE DISMUTASE [CU-ZN]		61.18	X-ray, 2.1Å	homo-dimer ✓	2 x ZN, 2 x CU
☐	1n19.1.A	Superoxide Dismutase [Cu-Zn]		61.18	X-ray, 1.9Å	homo-dimer ✓	2 x ZN, 2 x CU1
☐	2zkx.1.B	Superoxide dismutase [Cu-Zn]		61.18	X-ray, 2.7Å	homo-dimer ✓	2 x ZN, 2 x CU1
☐	2zkw.1.A	Superoxide dismutase [Cu-Zn]		61.18	X-ray, 1.9Å	homo-dimer ✓	2 x ZN, 2 x CU1
☐	3gtt.1.A	Superoxide dismutase [Cu-Zn]		60.67	X-ray, 2.4Å	homo-dimer ✓	2 x ZN
☐	3ltv.1.A	Superoxide dismutase [Cu-Zn],Superoxide dismutase [Cu-Zn]		60.67	X-ray, 2.5Å	homo-dimer ✓	2 x ZN

Build Models 4

Clear Selection



PV | Cartoon

3l9e.1.A

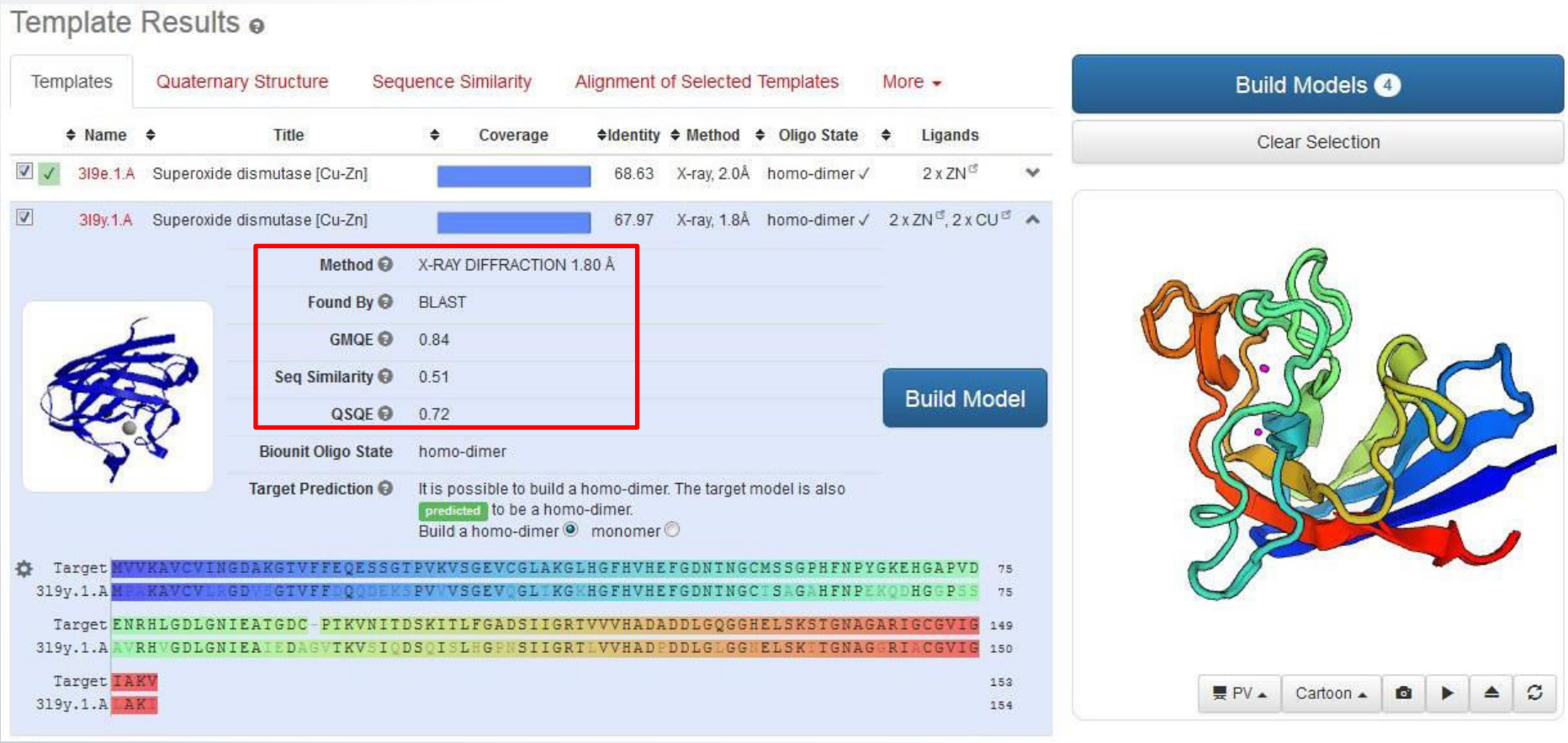
3l9y.1.A

3l9e.1.A

3l9y.1.A

Model results & evaluation

- **GMQE (Global Model Quality Estimation)** combines properties from the target–template alignment and the template search method. GMQE score is expressed as a **number between 0 and 1**, reflecting the **expected accuracy of a model built** with that alignment and template and the coverage of the target.

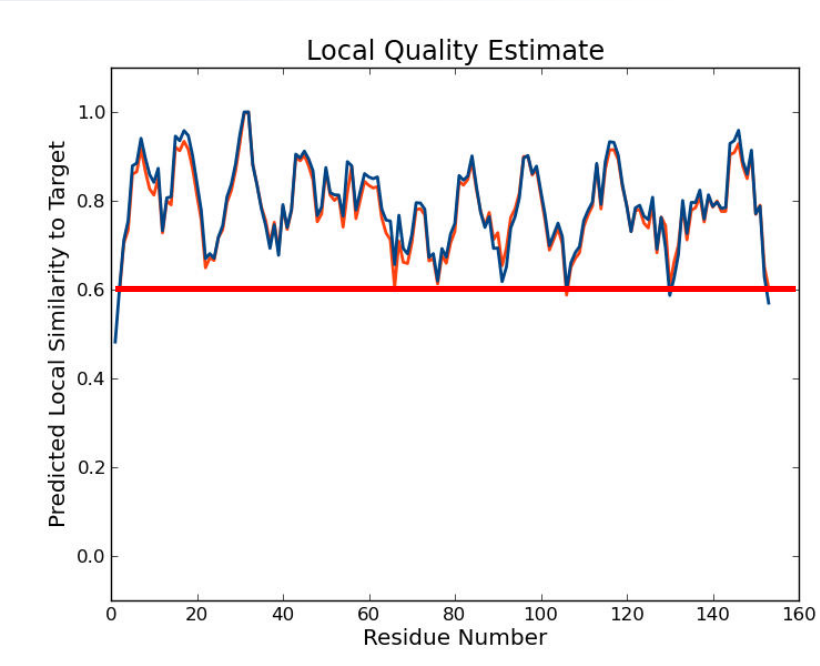


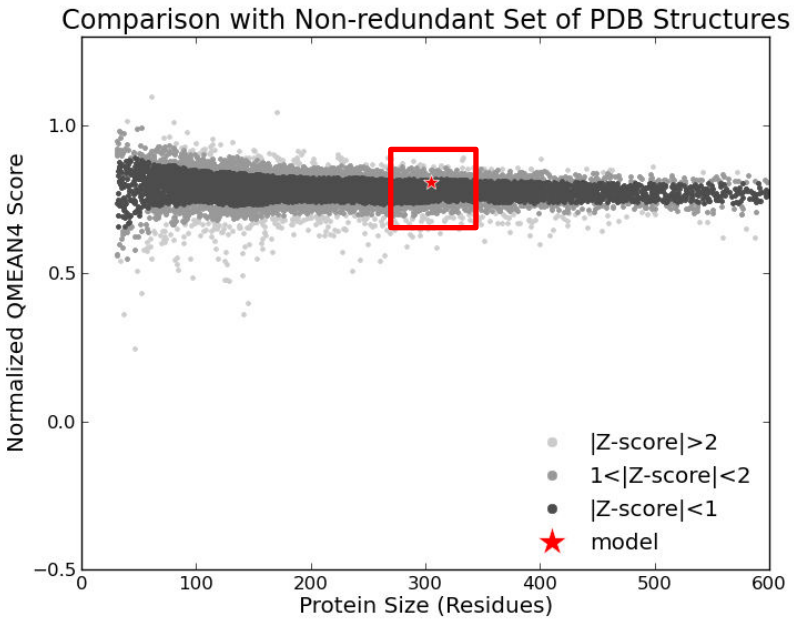
Model results & evaluation

- **QMEAN**, a **composite estimator** based on different geometrical properties and provides both **global** (i.e. for the entire structure) and **local** (i.e. per residue) absolute quality estimates on the basis of **one single model**.
- QMEAN Z-scores **around zero is good**, but of **-4.0 or below** are an indication of models with **low quality**

Model results & evaluation

- The "**Local Quality**" estimates, for each residue of the model (reported on the x-axis), the expected similarity to the native structure (y-axis). Typically, residues showing **a score below 0.6 are expected to be of low quality**.
- The "**Comparison**" plot models quality scores of **individual models** related to scores obtained for **experimental structures** of similar size. **Query inside normal distribution** of existing model is **great**.



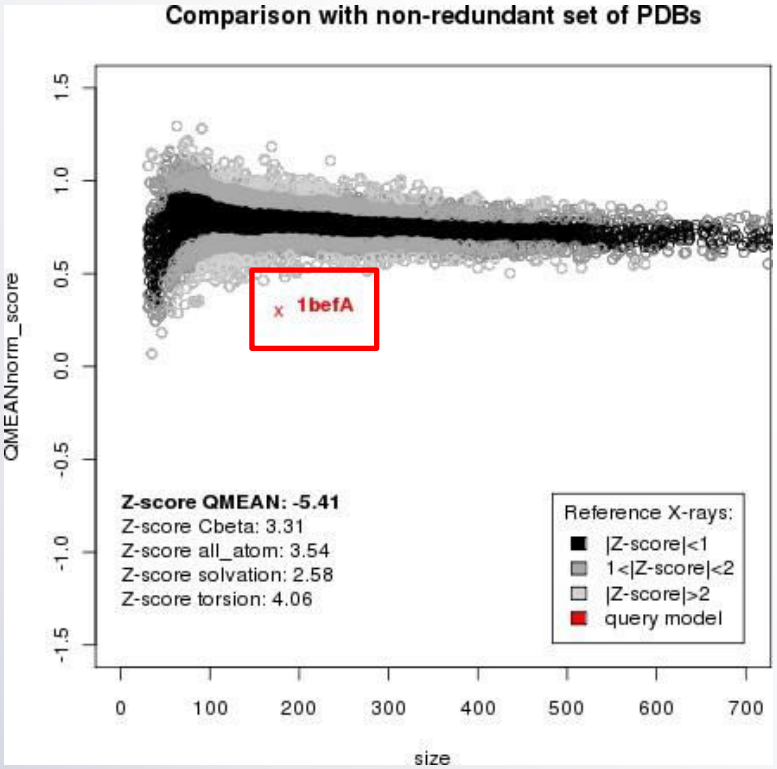


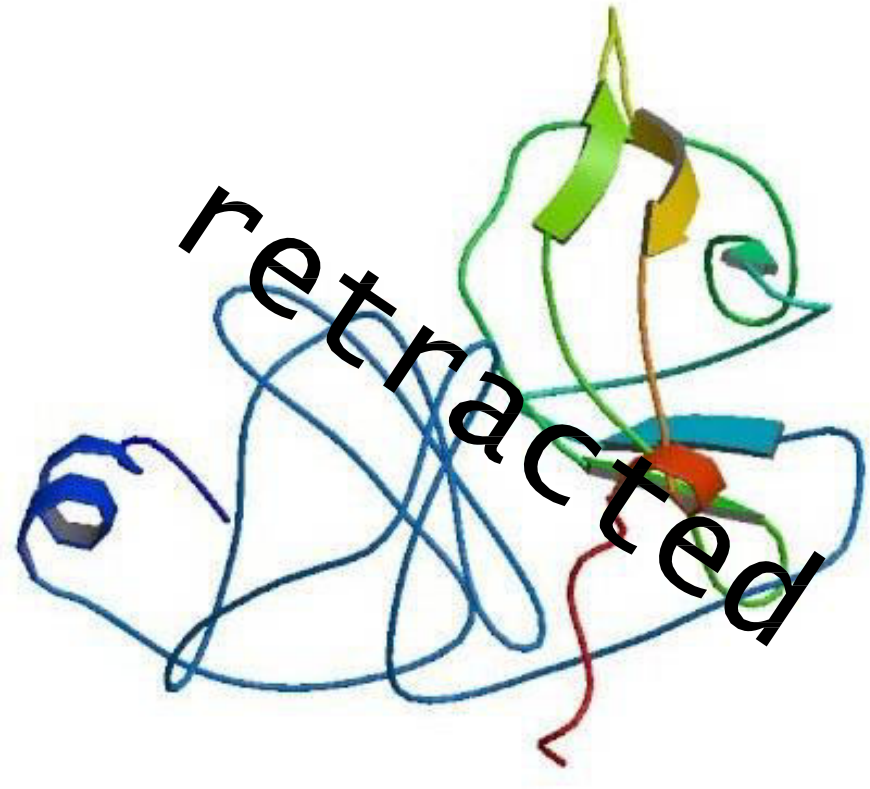
Model conclusion

Sometimes in modelling, you can **fail**. It is real!

If your model is **so far away from normal distribution of references.**

REDO your modelling, compare it with other methods

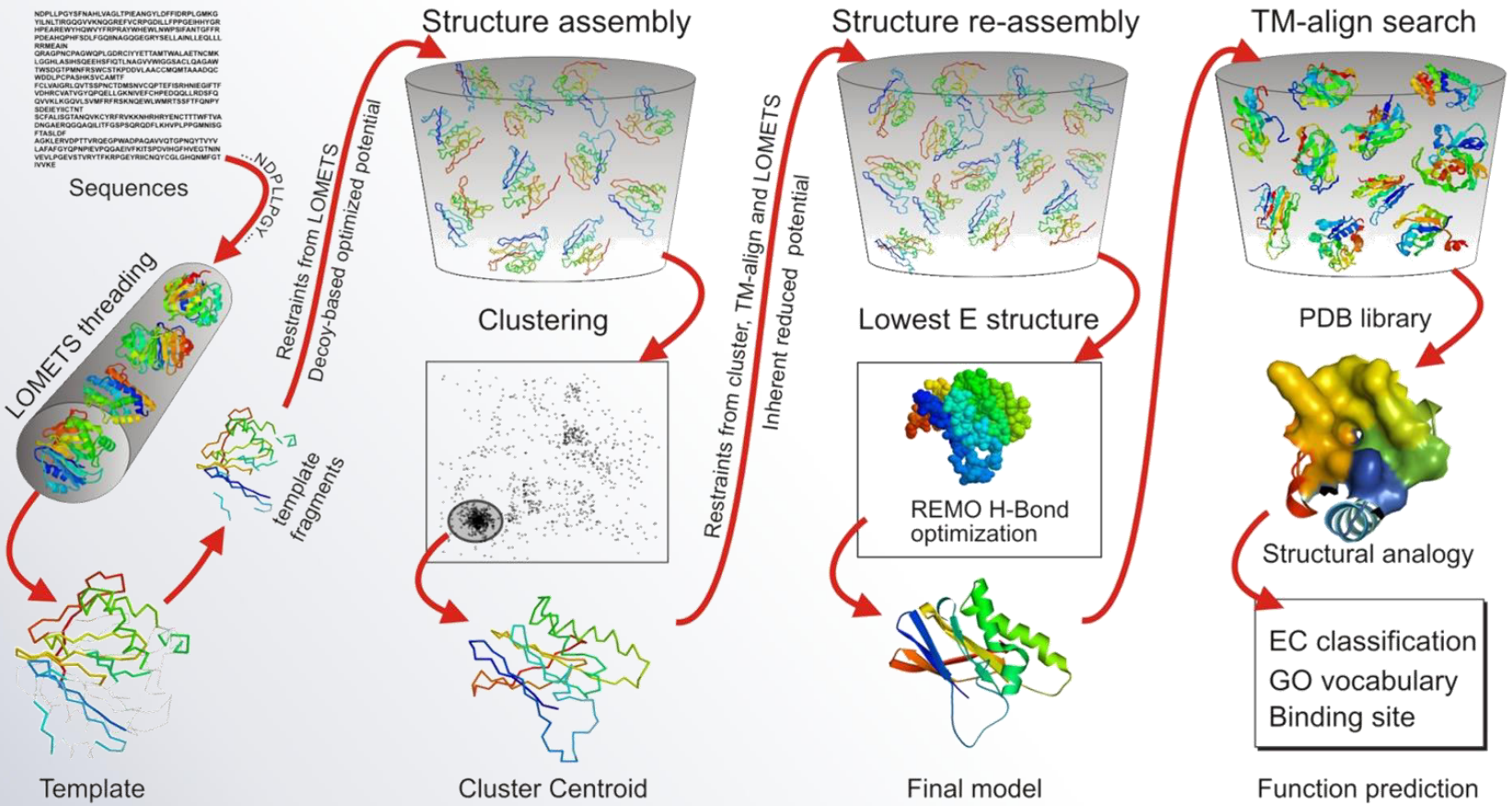




The case of Dengue Virus NS3 Serine Protease

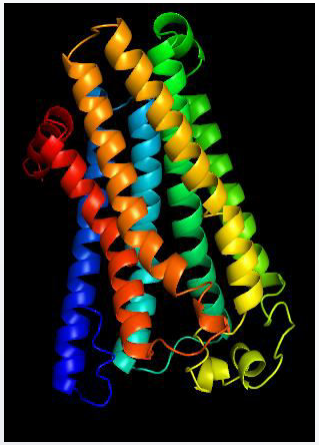
I-TASSER

- **I-TASSER (Iterative Threading ASSEmbly Refinement)**, a hierarchical approach to protein structure and function prediction.
- **It forms structural templates from the PDB by multiple threading approach LOMETS**, with full-length atomic models constructed by **iterative template fragment assembly simulations**.



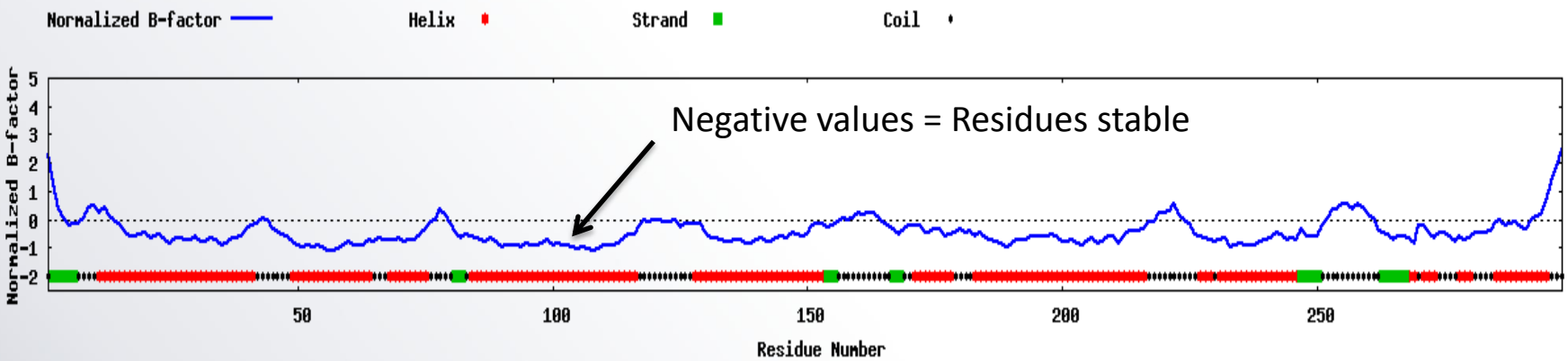
I-TASSER parameters of a good result

A. Visualization of tertiary structure



- Download Model 1
- C-score=-0.08
- Estimated TM-score = 0.70±0.12
- Estimated RMS = 6.3±3.9Å

- **C-score** is typically **in the range of [-5,2]**, where a C-score of higher value signifies a model with a high confidence and vice-versa.
- **TM-score >0.5** indicates a model of correct topology and **TM-score<0.17** means a random similarity.

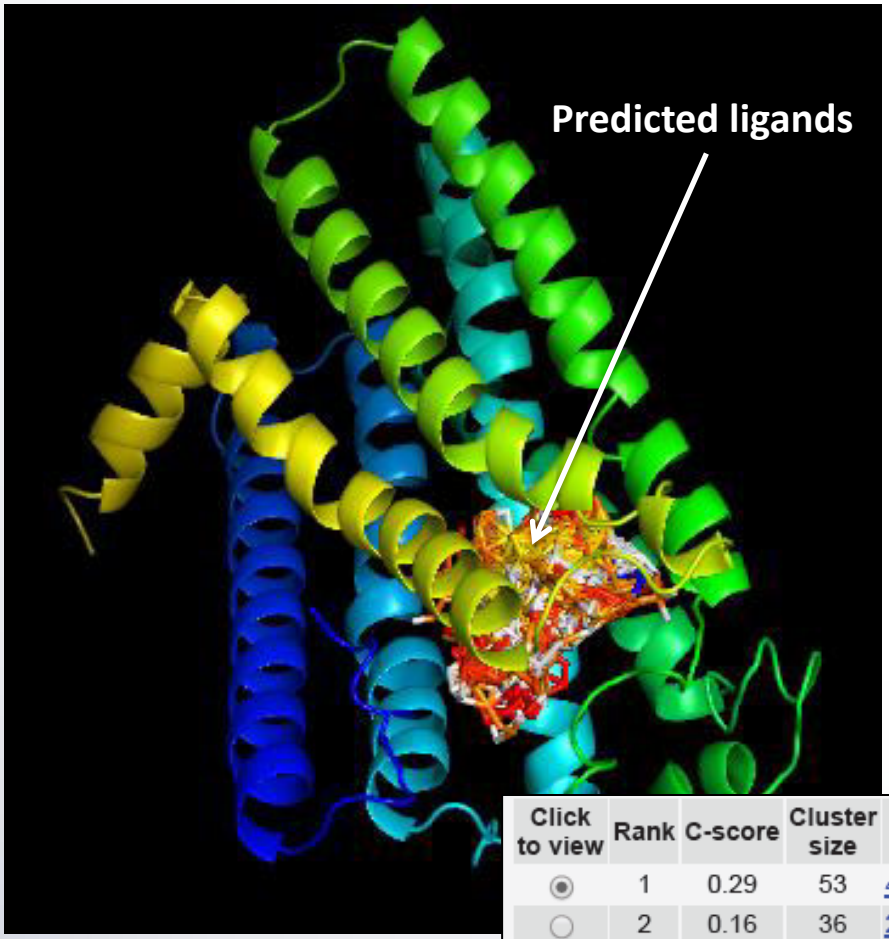


B-factor, a value to indicate the extent of the inherent thermal mobility of residues/atoms in proteins

The normalized B-factor (**B-factor profile, BFP**), predicted using a combination of both **template-based assignment** and **profile-based prediction.**

I-TASSER parameters of a good result

B. Prediction of ligand binding sites



a) **C-score** is the confidence score of the prediction. C-score ranges [0-1], where a higher score indicates a more reliable prediction.

b) **Cluster size** is the total number of templates in a cluster.

c) **Lig Name** is name of possible binding ligand. Click the name to view its information in the BioLiP database.

d) **Rep** is a single complex structure with the most representative ligand in the cluster, i.e., the one listed in the Lig Name column.

e) **Mult** is the complex structures with all potential binding ligands in the cluster.

Click to view	Rank	C-score	Cluster size	PDB Hit	Lig Name	Download Complex	Ligand Binding Site Residues
◉	1	0.29	53	4amiA	G90	Rep, Mult	90,91,94,95,98,180,181,182,192,196,242,24
○	2	0.16	36	2y04A	Y01	Rep, Mult	51,55,58,96,128,132,135,139
○	3	0.08	24	4ea3A	0NN	Rep, Mult	71,91,94,95,98,99,192,196,242,246,249,266
○	4	0.05	12	3dqbA	PEPTIDE	Rep, Mult	48,112,115,116,118,211,215,218,221,284,28
○	5	0.04	18	2ycwA	2CV	Rep, Mult	208,211,212,225,228,229,235

I-TASSER parameters of a good result

C. Prediction of protein function using Gene Ontology

Consensus prediction of GO terms

Molecular Function	GO:0003796	GO:0004940	GO:0004941	GO:0004995	
GO-Score	0.58	0.48	0.35	0.33	
Biological Process	GO:0071875	GO:0019835	GO:0009253	GO:0042742	GO:0016998
GO-Score	0.66	0.58	0.58	0.58	0.58

- **CscoreGO** is a combined measure for evaluating global and local similarity between query and template protein. It's **range is [0-1]** and higher values indicate more confident predictions.
- **TM-score** is a measure of global structural similarity between query and template protein.

GO:0003796 F JSON

lysozyme activity

Molecular Function

Definition (GO:0003796 GONUTS page)

Catalysis of the hydrolysis of the beta-(1->4) linkages between N-acetylmuramic acid and N-acetyl-D-glucosamine residues in a peptidoglycan. PMID:22748813

GO:0071875 P JSON

adrenergic receptor signaling pathway

Biological Process

Definition (GO:0071875 GONUTS page)

A series of molecular signals generated as a consequence of an adrenergic receptor binding to one of its physiological ligands.

QUARK

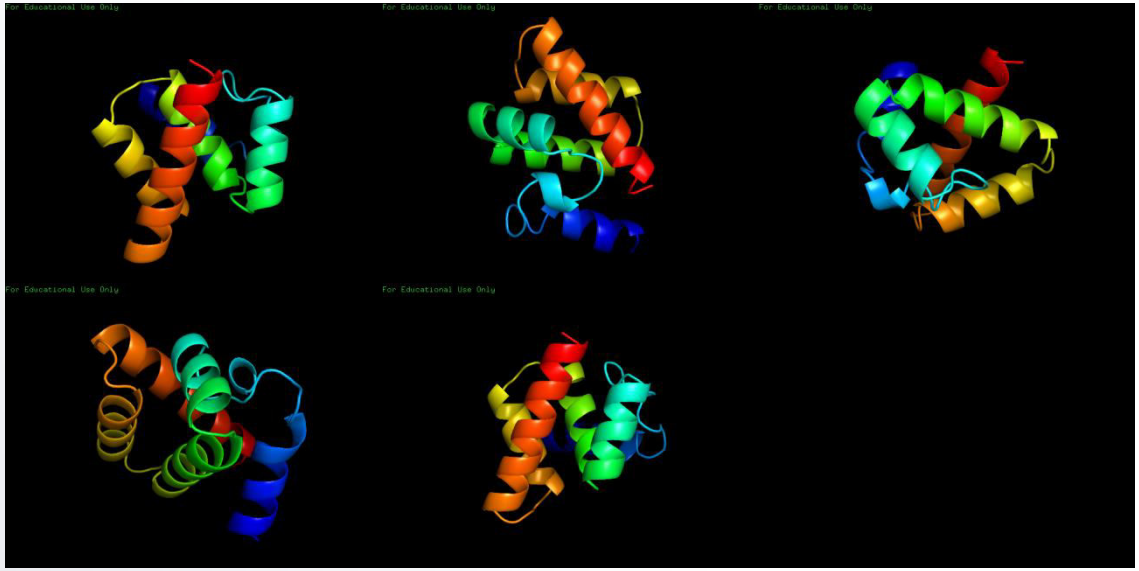
- **QUARK**, a computer algorithm for **ab initio protein structure prediction** and **protein peptide folding**, which aims to construct the correct protein 3D model from amino acid sequence only.
- QUARK models, built from small fragments (1-20 residues long) by **replica-exchange Monte Carlo simulation** under the guide of an atomic-level knowledge-based force field.
- QUARK was ranked as the No 1 server in Free-modeling (FM) in CASP9 and CASP10 experiments.

All you need is amino acid sequences…

```
>1ci4A
TTSQKHRDFVAEPGEKPVGSLAGIGEVLGKKLEERGFDKAYVVLGQFLVLKKDEDLFRE
WLKDTCGANAKQSRDCFGCLREWCDAFL
```

QUARK results

A. Prediction of five ab initio proteins from query



Confidential score 1-9, close to 10 perfect

B. Predicted secondary structure

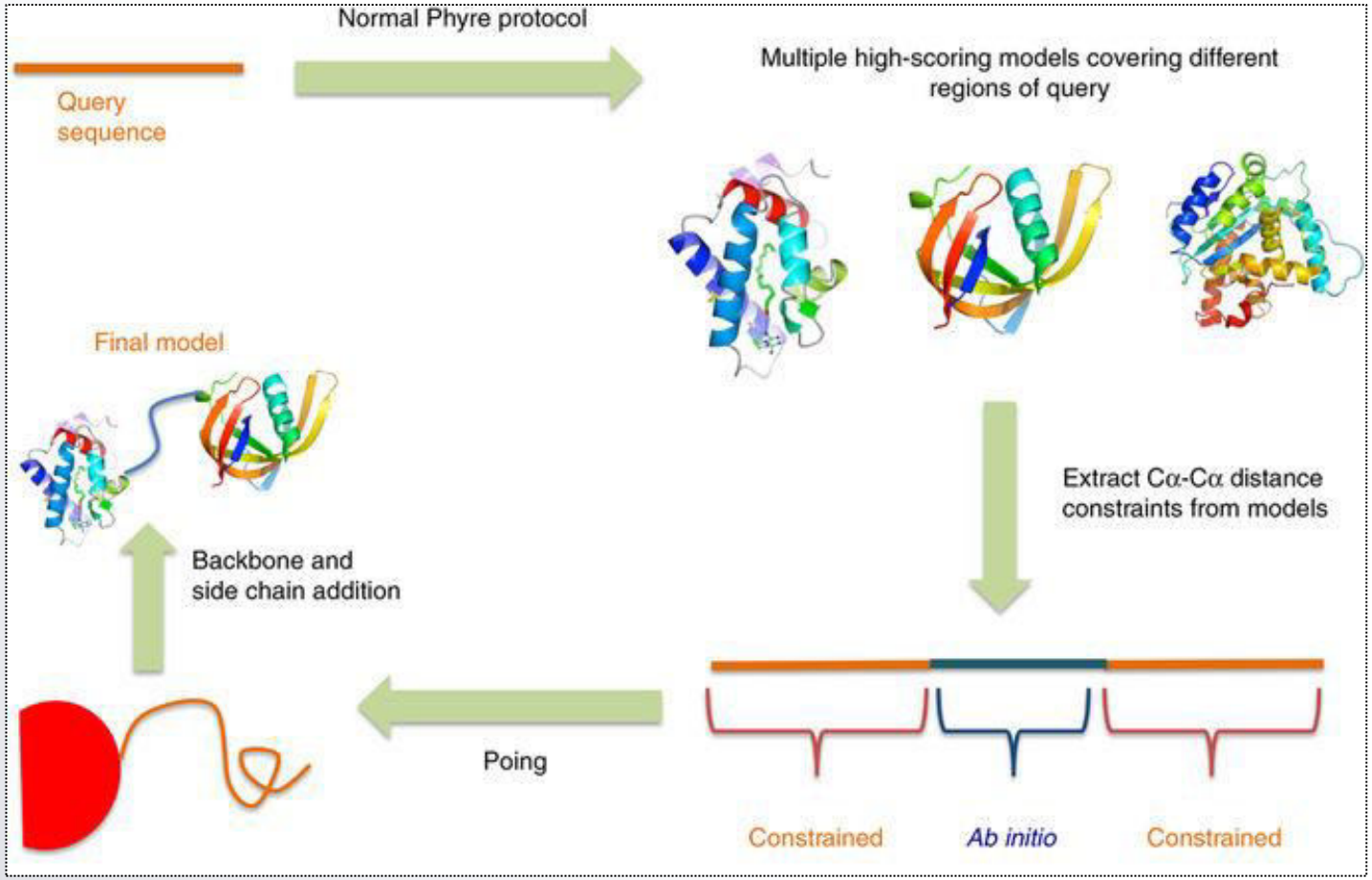
```
                          20                  40                  60                  80
                           |                   |                   |                   |
Sequence    TTSQKHRDFVAEPGEKPVGSLAGIGEVLGKKLEERGFDKAYVVLGQFLVLKKDEDLFREWLKDTCGANAKQSRDCFGCLREWCDAFL
Prediction  CCCHHHHHHHCCCCCCCCCCCCCHHHHHHHHHCCCHHHHHHHHHHHHCCCHHHHHHHHHHHCCCHHHHHHHHHHHHHHHHHHC
Conf.Score  988899999879999987447898899999999979659999999999958889999999999968899999999999999859
            H:Helix; S:Strand; C:Coil
```

C. Predicted solvent acessibility

```
                          20                  40                  60                  80
                           |                   |                   |                   |
Sequence    TTSQKHRDFVAEPGEKPVGSLAGIGEVLGKKLEERGFDKAYVVLGQFLVLKKDEDLFREWLKDTCGANAKQSRDCFGCLREWCDAFL
Prediction  553330221123223321120110032002102421132002000200113232310220022102031310310010022003324
            Values range from 0 (buried residue) to 9 (highly exposed residue)
```

Phyre2

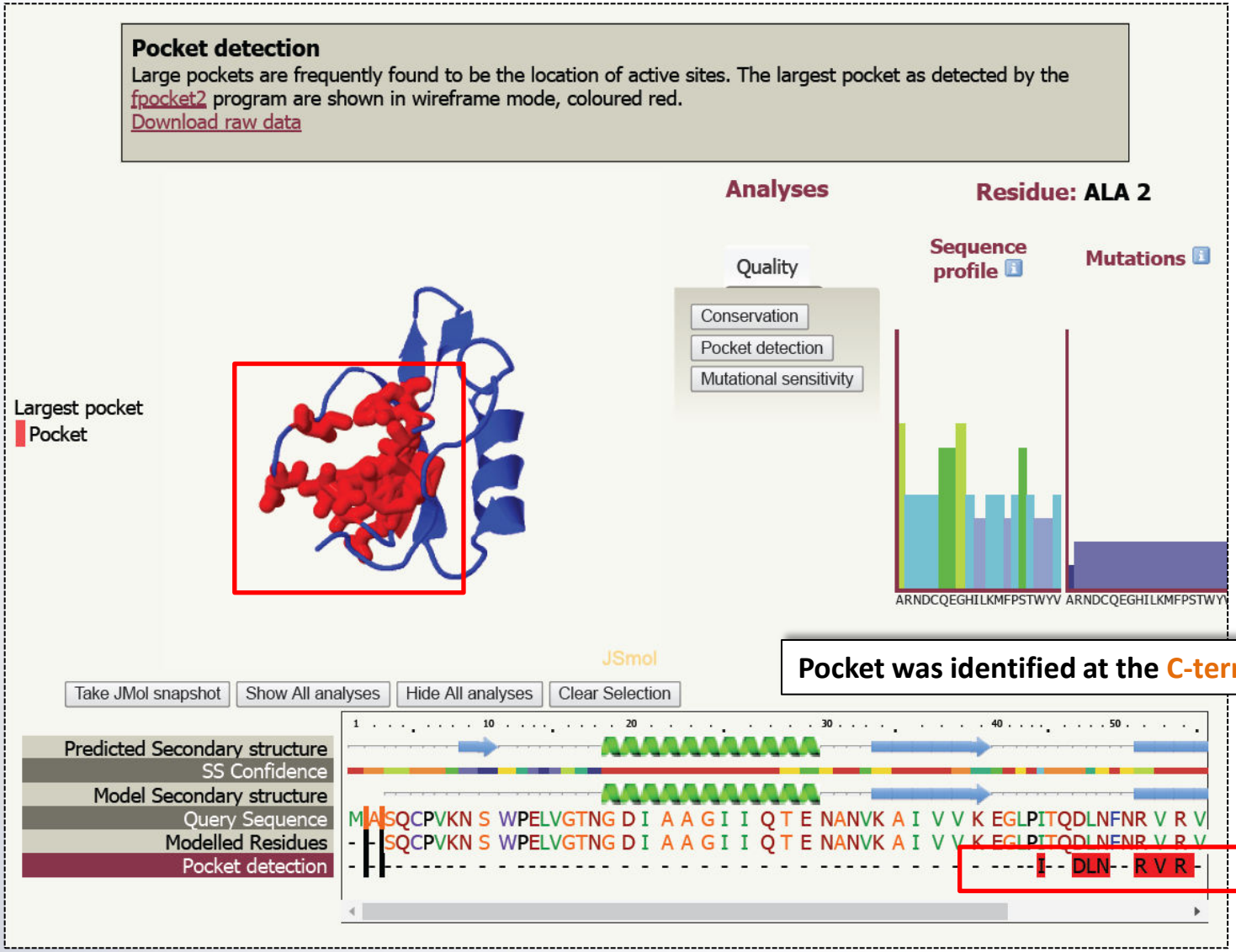
- **Protein Homology/analogY Recognition Engine V 2.0** (Phyre2) provides an intensive mode to create a complete full-length model of a sequence through a **combination of multiple template modeling** and simplified **ab initio folding simulationists** with a simple and intuitive interface.



Eliminate **constraints** from several models and create final model using P**oing** (protein-folding simulator)

Phyre2 results

A. Identification of pocket using Phyre2 investigators



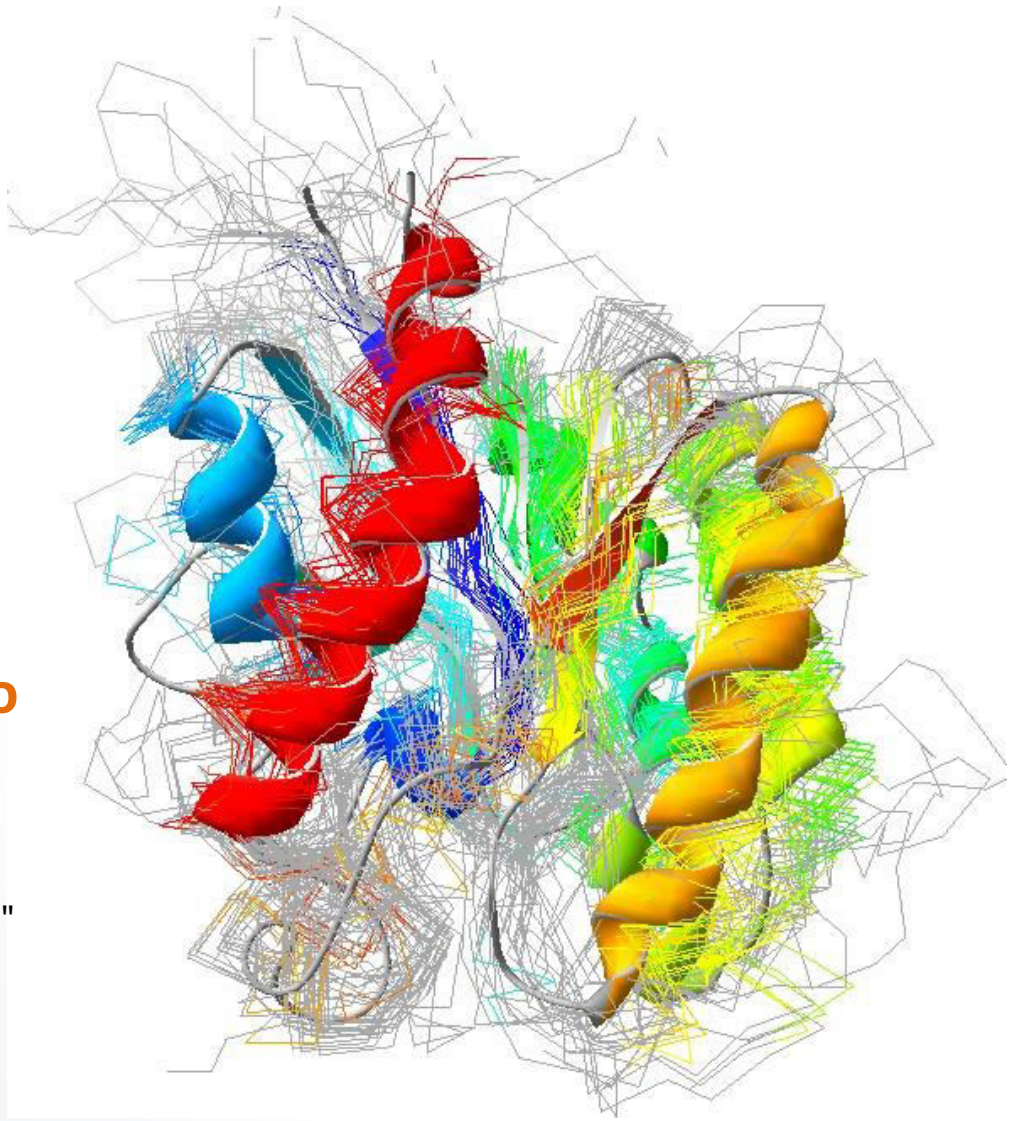
Pocket was identified at the C-terminal sites

Paradox: modeling is not a real protein

Ceci n'est pas une protéine.

"... a model must be **wrong**, in some respects --- else it would be the thing itself. The trick is **to see** ... **where it is right**."

Henry A. Bent
"Uses (and Abuses) of Models in Teaching Chemistry," J. Chem. Ed. 1984 61, 774.



It was still the tenth course, don’t get dizzy yet