Comparative modelling of Protein Stucture

The importance of structure prediction

The vast (and steadily growing) number of proteins of unknown structure puts a heavy demand on ever faster methods for 3D structure determination. Due to their intrincacies, such methos simply cannot cope with the fast pace of sequencing, and the gap widens more and more. This situation is not likely to change, ever.

So what can we do ?

We need to be able *predict* the 3D structure of proteins from their aminoacid sequence. In general terms this is a very hard computational problem, but there are special situations where it is very feasible. That is currently our best hope of coping up.

Prediction of the 3D structure of proteins is thus one of the most fundamental problems of bioinformatics / computational biology.

Protein structure prediction

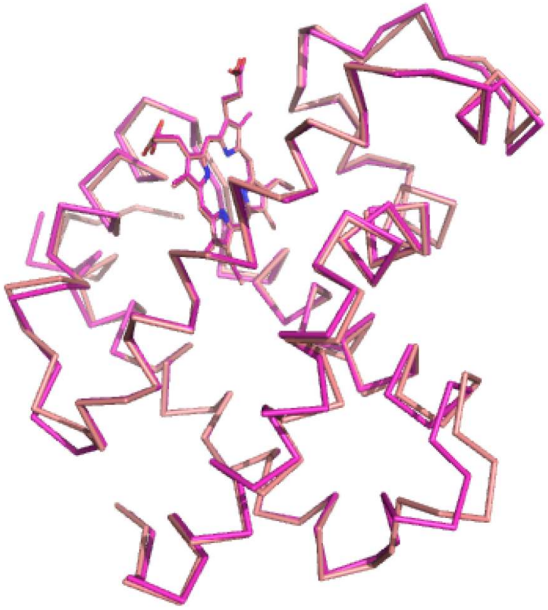
- **Secondary structure prediction**: predicting the location of secondary structural elements (helices, sheets, loops, etc.) within the protein sequence. Mostly done with automated servers that can solve the problem with about 85-90% accuracy.

- **Tertiary structure prediction:** prediction the 3D shape of proteins with atomic detail. This can be done using different approaches:
  - ***Ab initio modelling*:** based on first physical-chemical principles, without relying on previous knowledge or assumptions about the final structure. (*e.g.* protein folding by Molecular Dynamics)
  - ***Knowledge-based modelling*:** as the name implies, this type of modelling relies on our accumulated knowledge on protein structures. (e.g. comparative (or homology) modelling of protein structure)

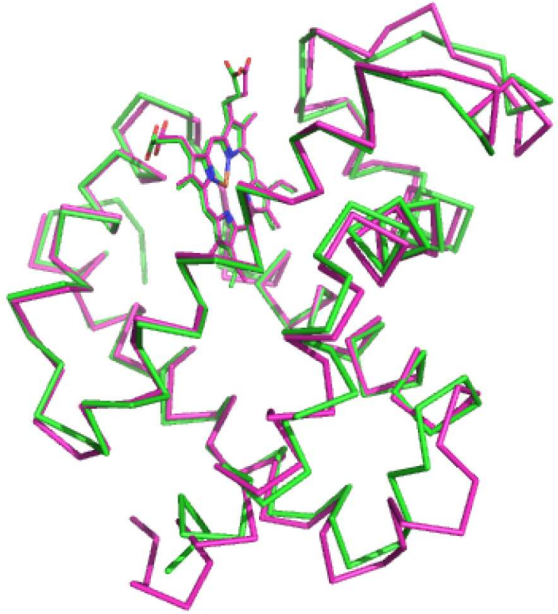
Relation between sequence and structure similarity

Given that protein structure is largely a consequence of aminoacid sequence, it's expectable that *similar sequences give rise to similar structures.*

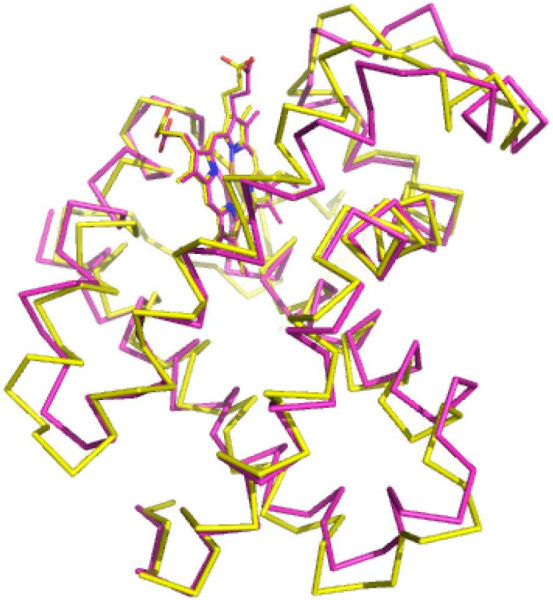
MYOGLOBIN



Human **x** Horse
88% identity

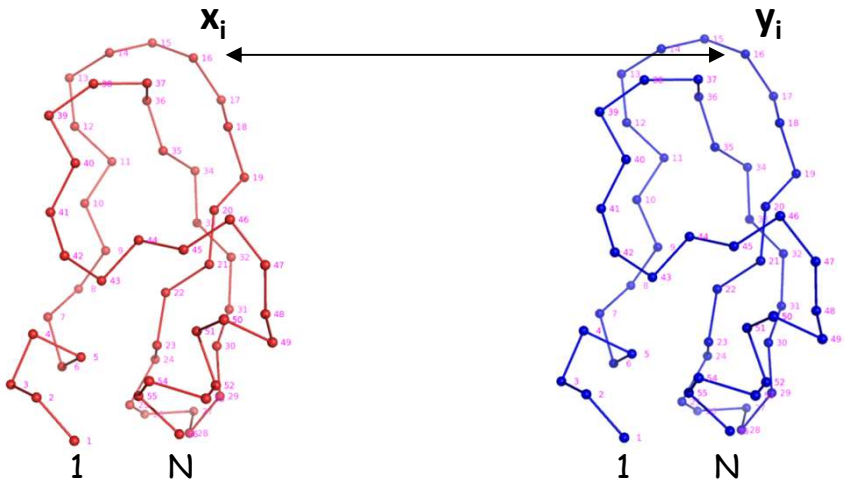


Human **x** Tuna
45% identity



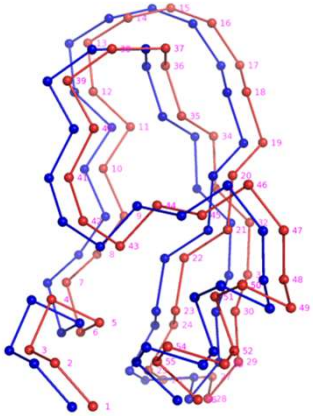
Myoglobin **x** Cytoglobin
29% identity

Measuring Sructural Similarity



Atom $\mathbf{x_i}$ corresponds to atom $\mathbf{y_i}$

RMSD minimization



Square of the distance between atom $\mathbf{x_i}$ and atom $\mathbf{y_i}$

$$RMSD = \sqrt{\frac{\sum_{i}^{N} | \vec{x}_i - \vec{y}_i |^2}{N}}$$

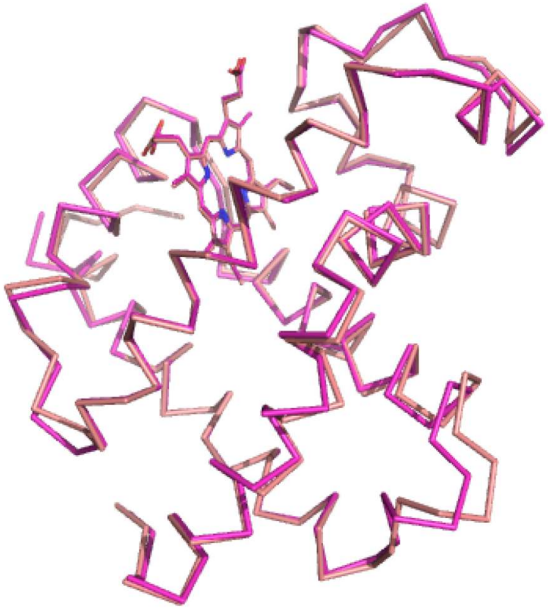
Structure comparison assumes a pre-existing correspondance between atoms in structure A and B. For highly similar moleculares this correspondence can be established by *sequence alignment*.

RMSD - root mean square deviation, it has units of *length* and it is usually measured in Angstrom or nanometers (nm).

Relation between sequence and structure similarity

Given that protein structure is largely a consequence of aminoacid sequence, it's expectable that *similar sequences give rise to similar structures.*

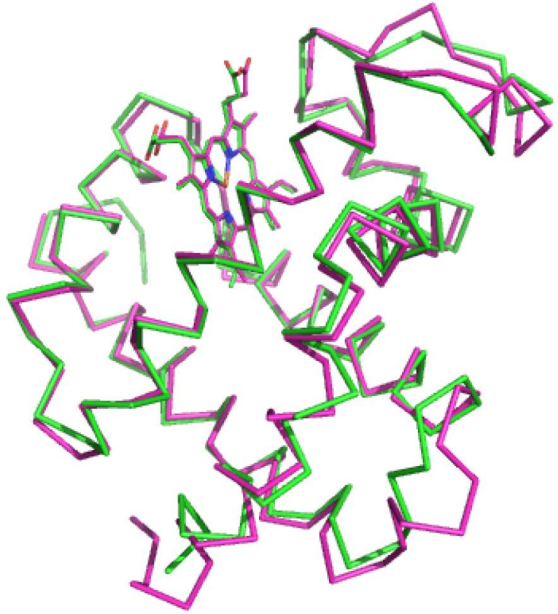
MYOGLOBIN



Human **x** Horse

88% identity

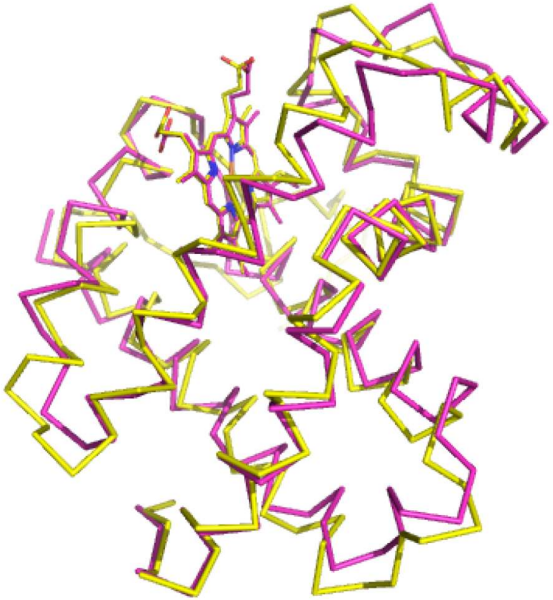
RMSD = 1.16



Human **x** Tuna

45% identity

RMSD = 1.72

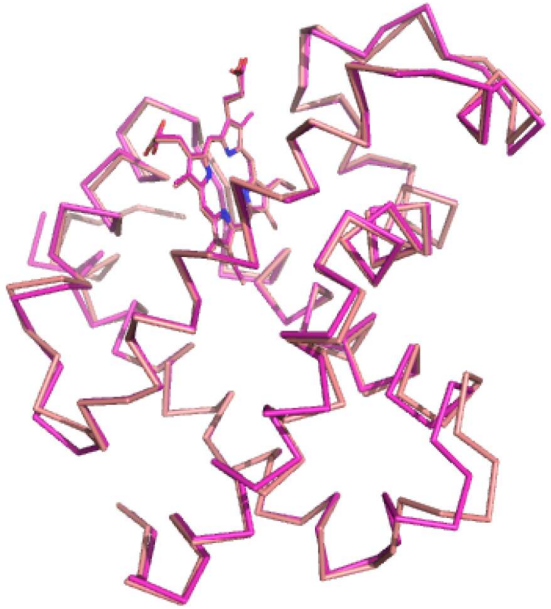


Myoglobin **x** Cytoglobin

29% identity

RMSD = 2.53

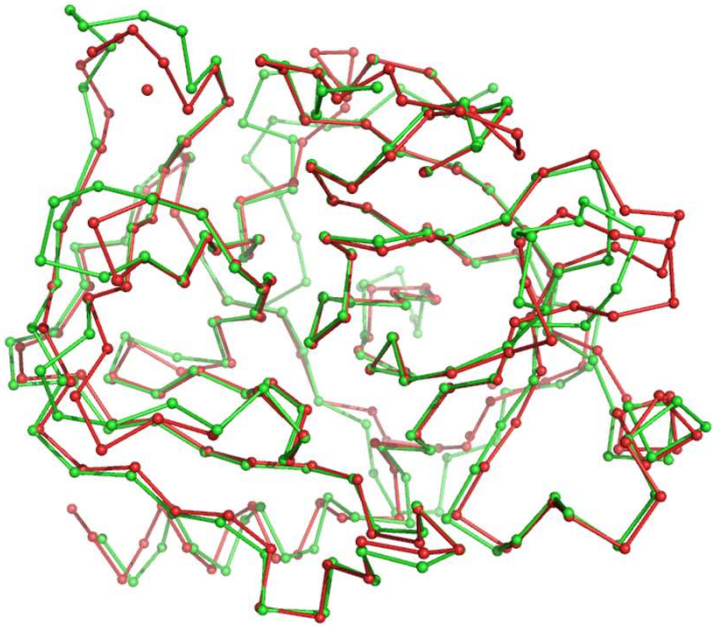
Sequence similarity implies structural similarity



Human **x** Horse Myglobin

88% identity

RMSD = 1.16



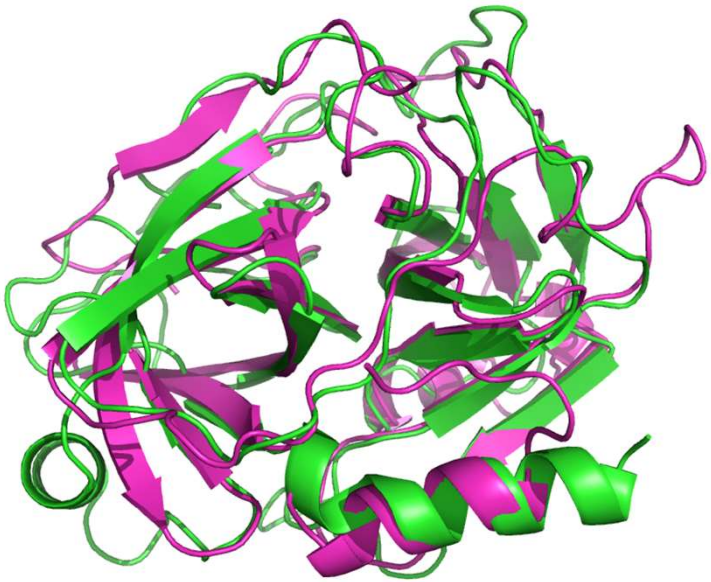
Human **x** Bovine Trypsin

74% identity

RMSD = 0.76

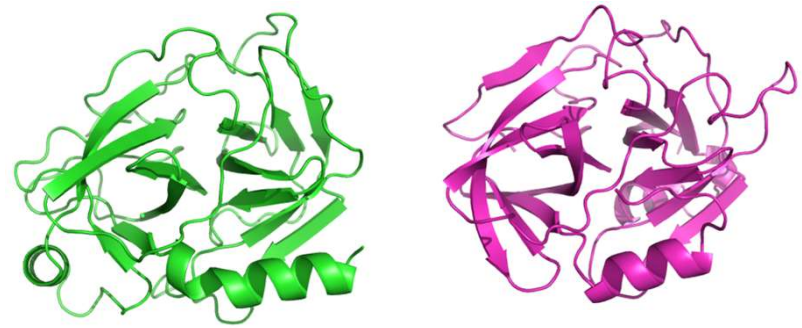
Structural similarity does not imply sequence similarity

Tripsina humana
versus
Proteinase V8 *S.aureus*



RMSD 2.5 Å

19% identidade de sequência



As duas proteínas têm clara semelhança estrutural, mas esta não é detectável por comparação das duas sequências

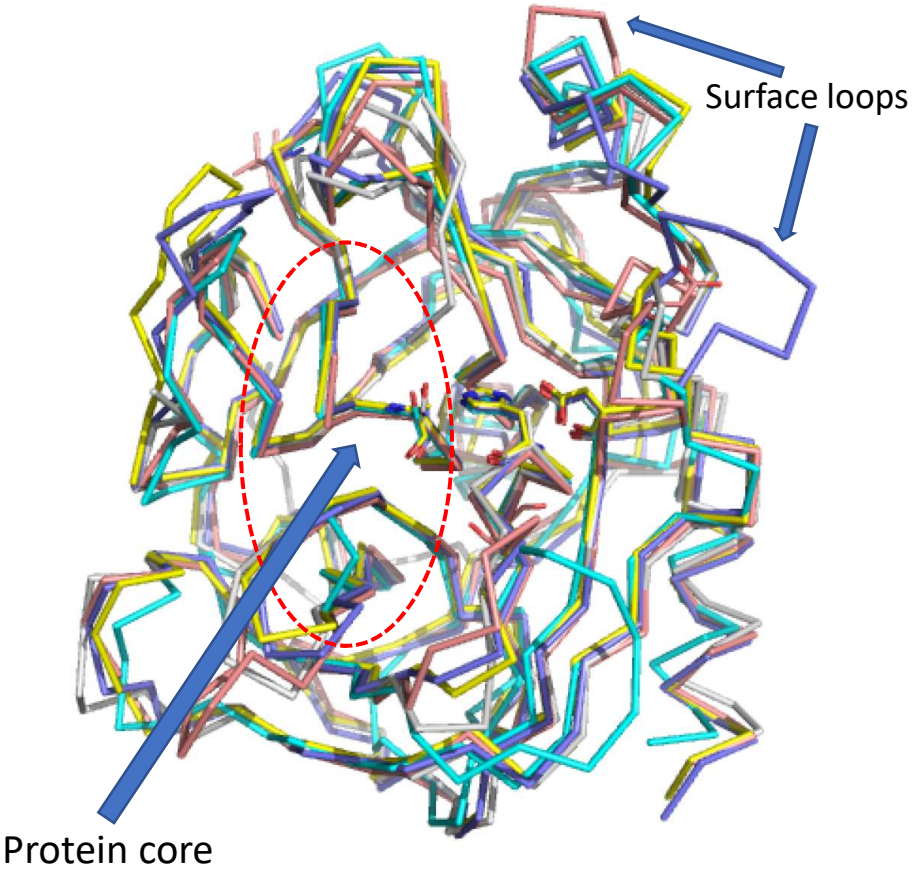
Conservation of structural features

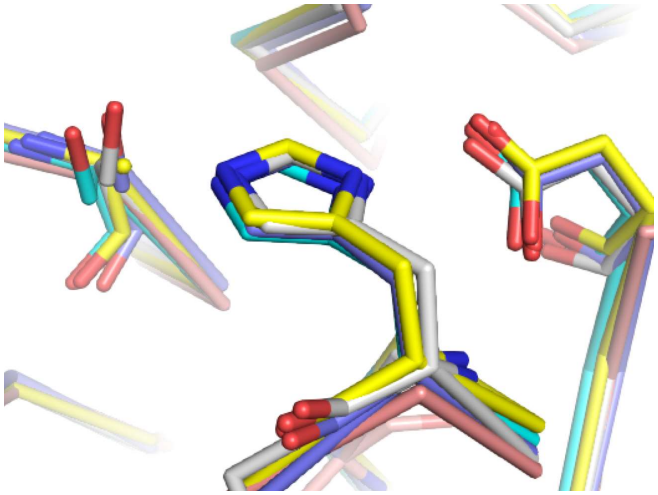
Conservation of structural features is not constant along the protein sequences:

- **Core regions** tend to be more conserved, as mutations there tend to have a greater impact upon the structure.
- **Functional important regions** are associated with function and thus are under strong selective pressure (e.g. enzyme active sites)
- **Surface-exposed loops** tend to be less conserved, provided they aren't related with the protein's function.

Conservation of structural features

Example: conservation of structural features across the serine protease family

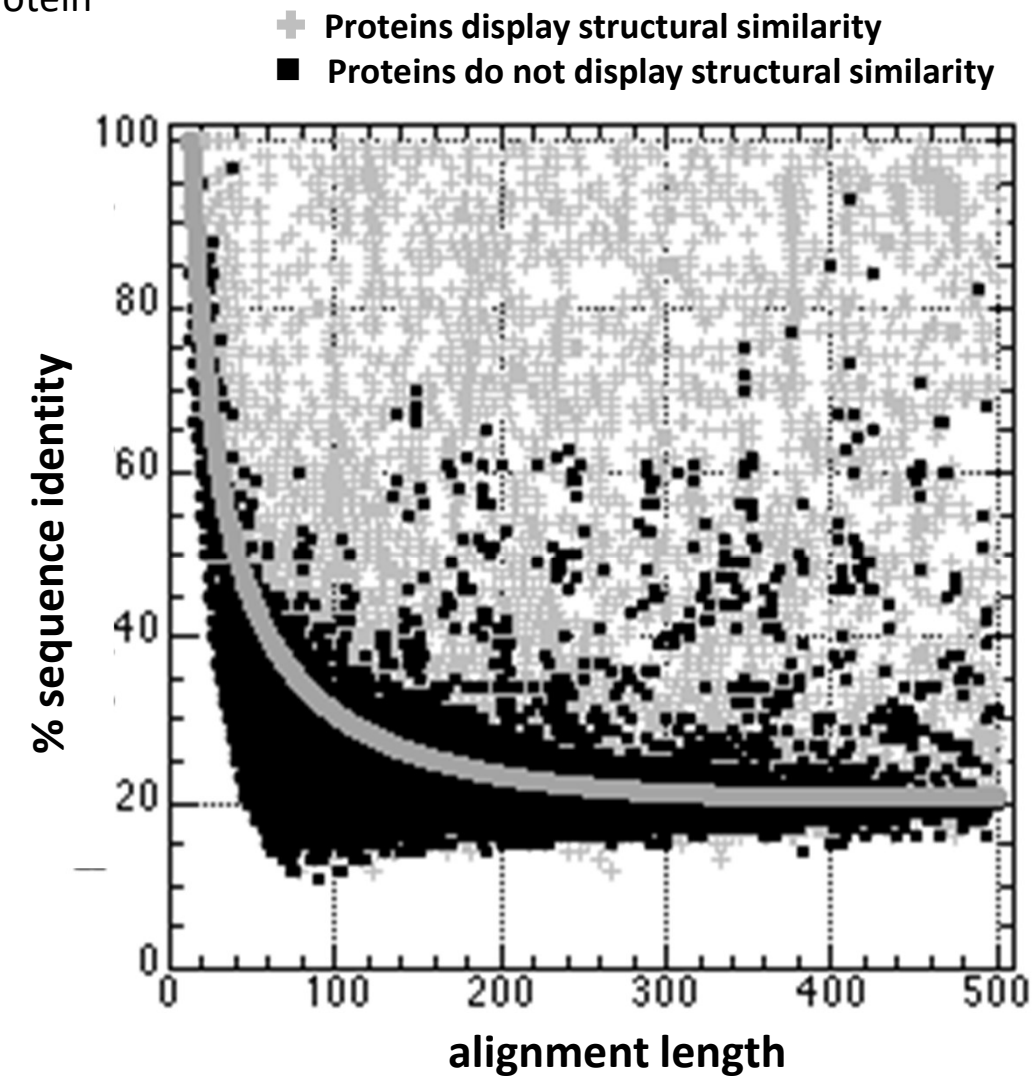




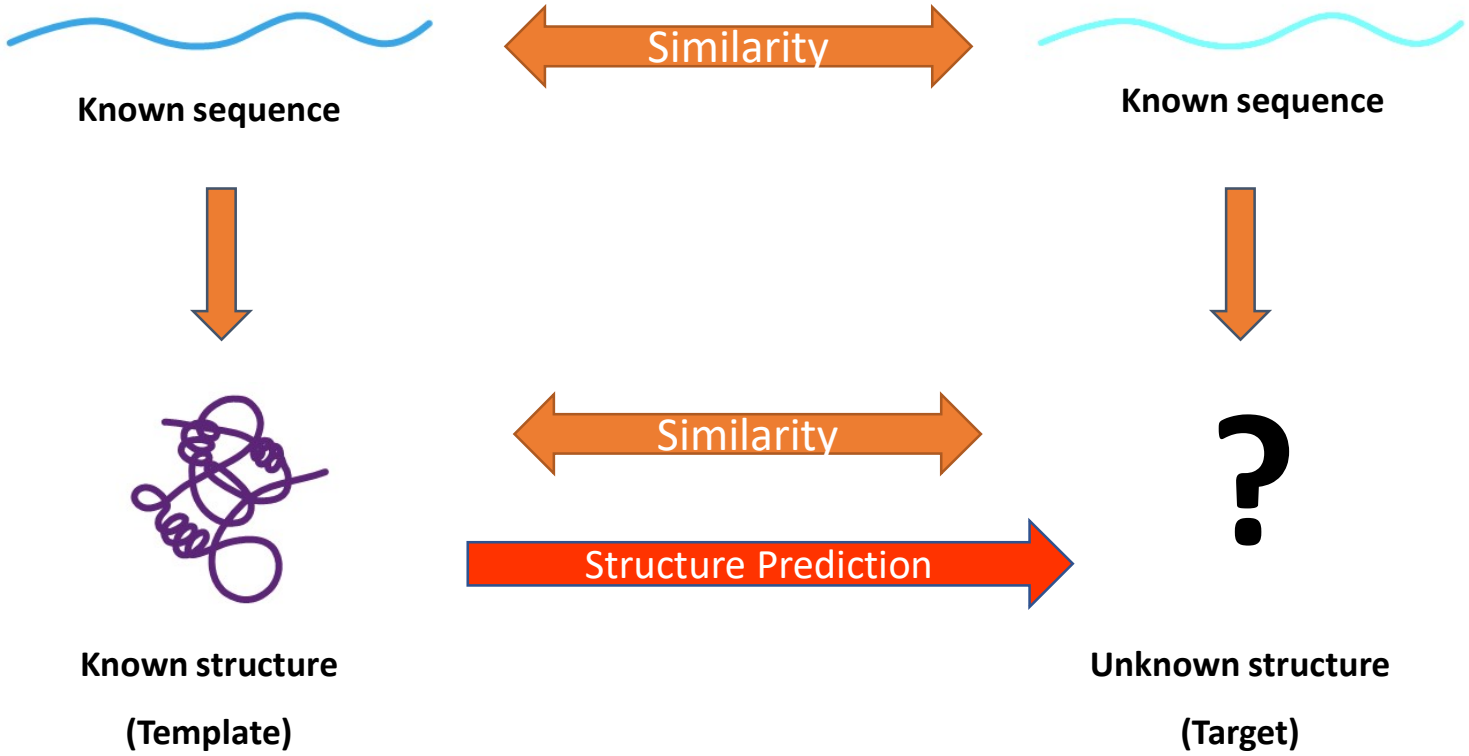
Conservation of active site features

Threshold for structural homology

Each ✚ **or** ■ in the graph representes a protein-protein comparison



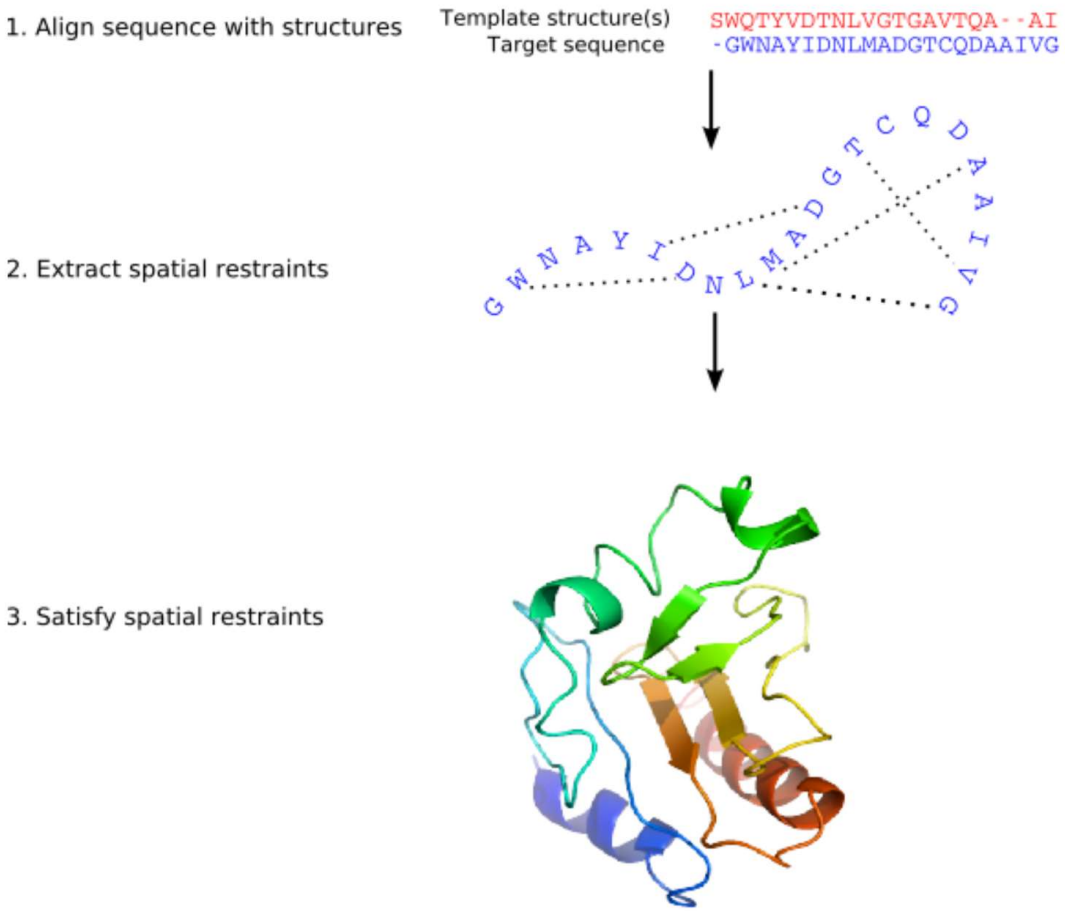
Comparative modelling of protein structures

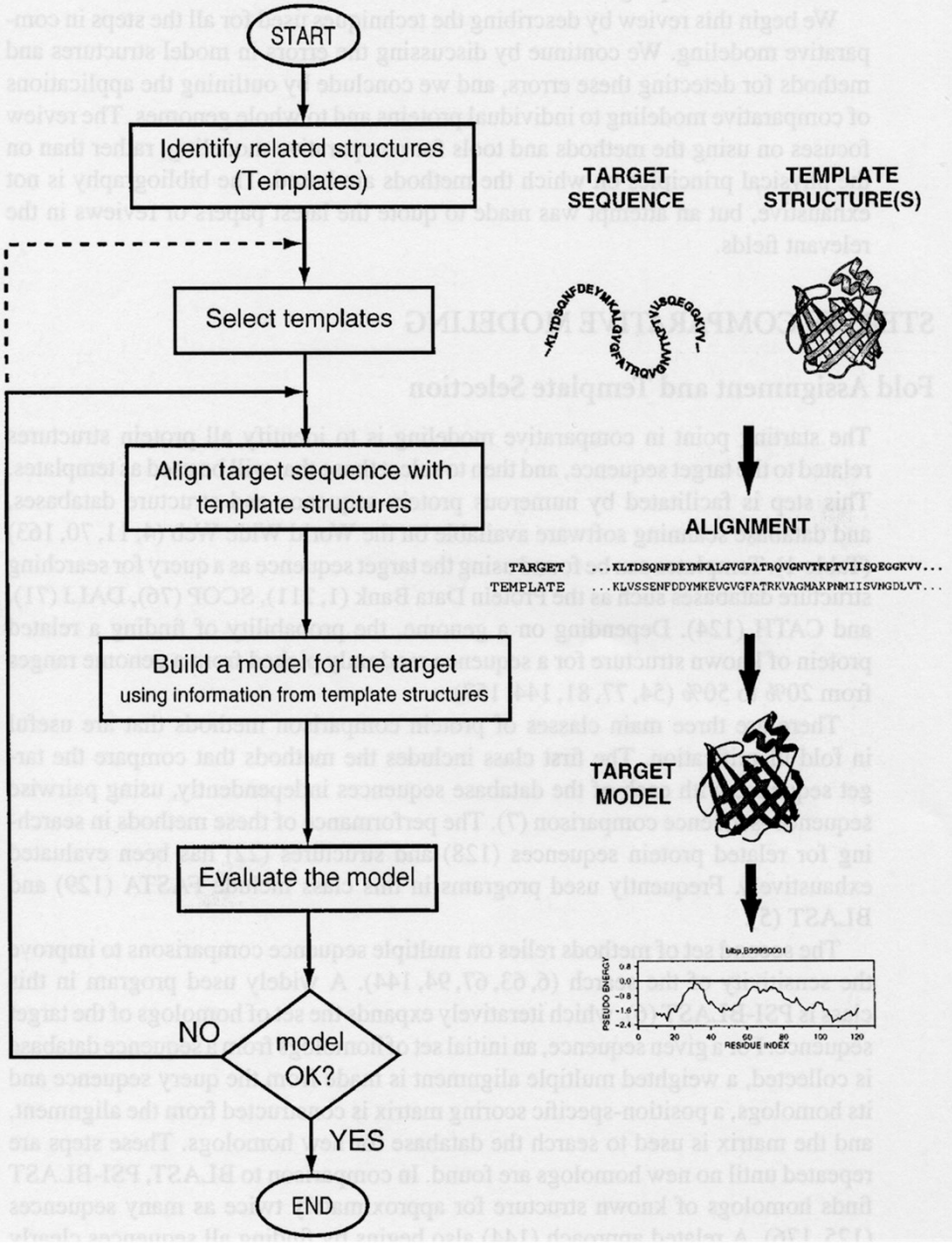


Similar sequences imply similar structures, and so:

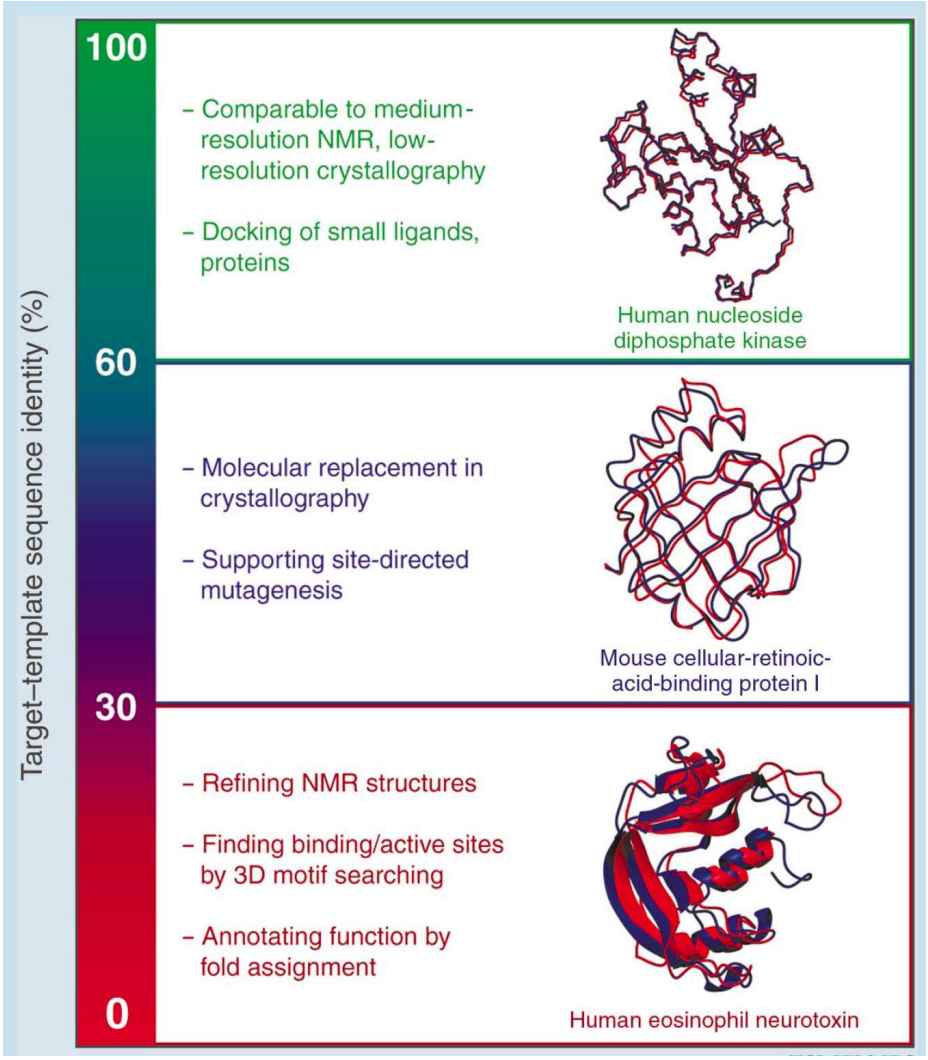
The unknow structure of a target protein can be modelled on the the known structure of another protein (template) displaying sufficient sequence similarity with the target.

Modelling by satisfaction of spatial restraints (MODELLER)




START
Identify related structures (Templates)
Select templates
Align target sequence with template structures
Build a model for the target using information from template structures
Evaluate the model
model OK?
NO
YES
END
TARGET SEQUENCE
TEMPLATE STRUCTURE(S)
ALIGNMENT
TARGET ...KLTDSQNFDEYMKALGVGFATRQVGNVTKPTVIISQEGGKVV...
TEMPLATE ...KLVSSENFDDYMKEVGVGFATRKVAGMAKPNMIISVNGDLVT...
TARGET MODEL
PSEUDO ENERGY
RESIDUE INDEX

Applications of comparative modelling



On-line servers for automated comparative modelling

Homology models can be built by several online servers, using as input the target sequence and optionally (in same cases) the target model or PDB ID(s).

- **Swiss Model -** http://swissmodel.expasy.org/
- **I-Tasser:** http://zhanglab.ccmb.med.umich.edu/I-TASSER
- **ModWeb:** http://salilab.org/modweb
- **Phyre2:** http://www.sbg.bio.ic.ac.uk/phyre2
- **RaptorX:** http://raptorx.uchicago.edu/

The Protein Model Portal (https://www.proteinmodelportal.org/) is portal web site allowing submission of modelling queries to multiple automated sites at once.

Public databases of protein models

Several automated modelling algorithms have been used to produced extensive databases of publically available pre-computed models. They can be used for very easy (high similarity) cases or for having a “quick and dirty” preliminary model.

- **Protein Model Portal -** https://www.proteinmodelportal.org/
- **Swiss Model Repository :** https://swissmodel.expasy.org/repository
- **ModBase:** https://modbase.compbio.ucsf.edu/
- **MobiDb:** http://www.sbg.bio.ic.ac.uk/phyre2

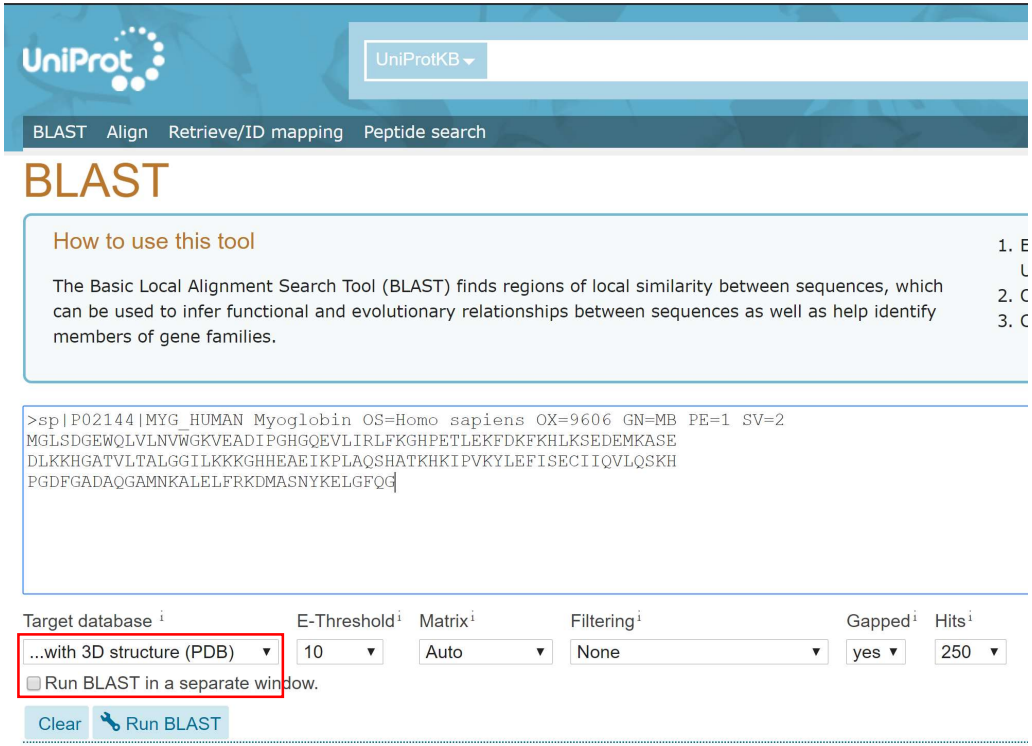
Steps in manual comparative modelling

1. Choice of the template (or templates)
2. Alignment of the target against the template(s)
3. Building and energy optimization of 3D models of the target guided by satisfaction of the spatial restraints imposed by the template.
4. (Optional) Optimization of long surface loops no present on the template (*ab initio* modelling)
5. Quality assessment of the produced models by different methods
6. If models don't meet minimum quality requirements, bo back to 1. (choice of a new template) or 2. (correction of template-target sequence alignment errors)
7. Repeat the process until models have sufficient quality

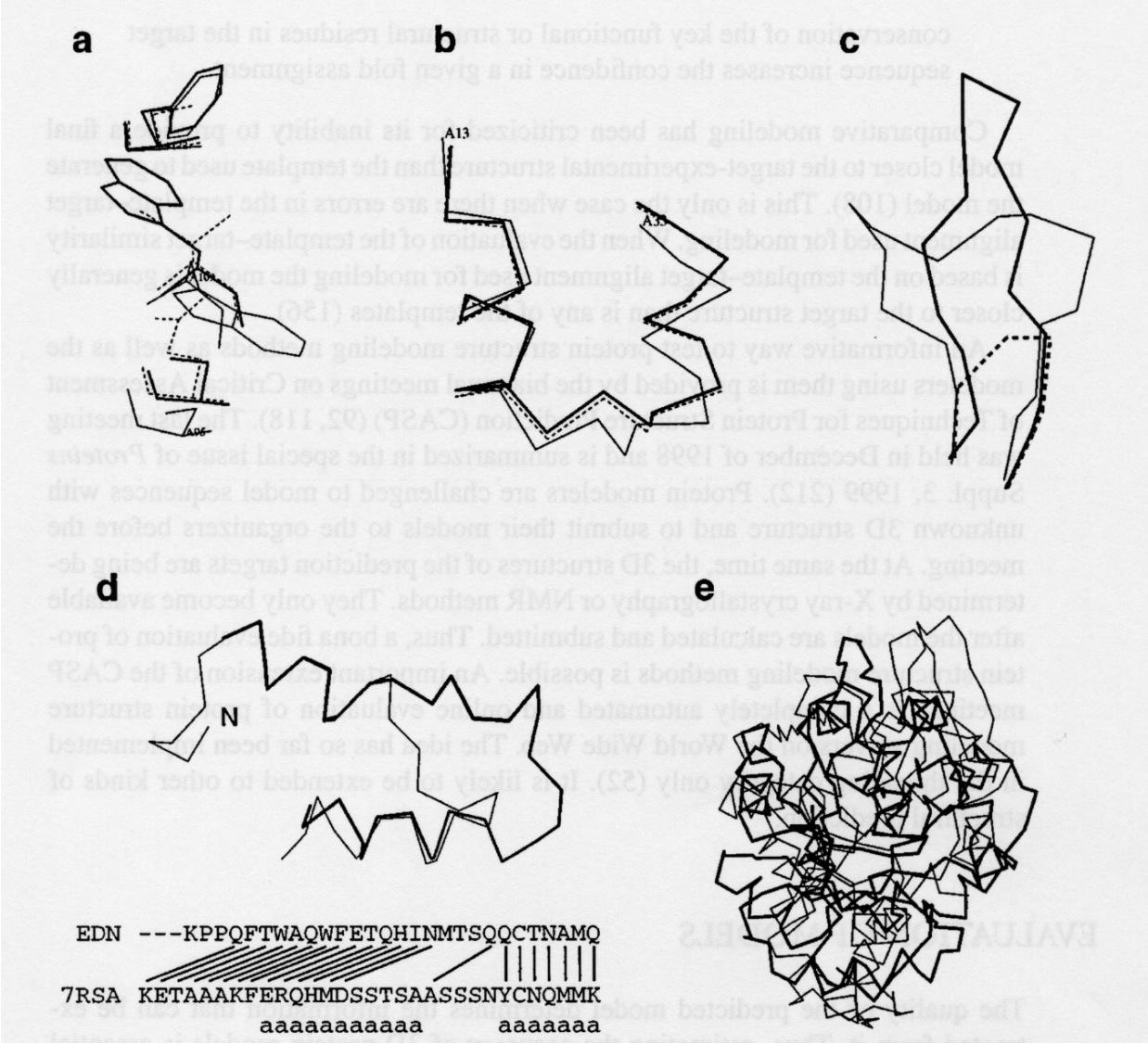
Template search

The template protein must be of known structure and with sufficient similarity with the target sequence

- The template structure must have good resolution and R-value and few or no missing atoms
- Templates are often found by BLASTing the target sequence agains the Uniprot PDB subset (sequences of proteins of known structure:



Errors in comparative modelling



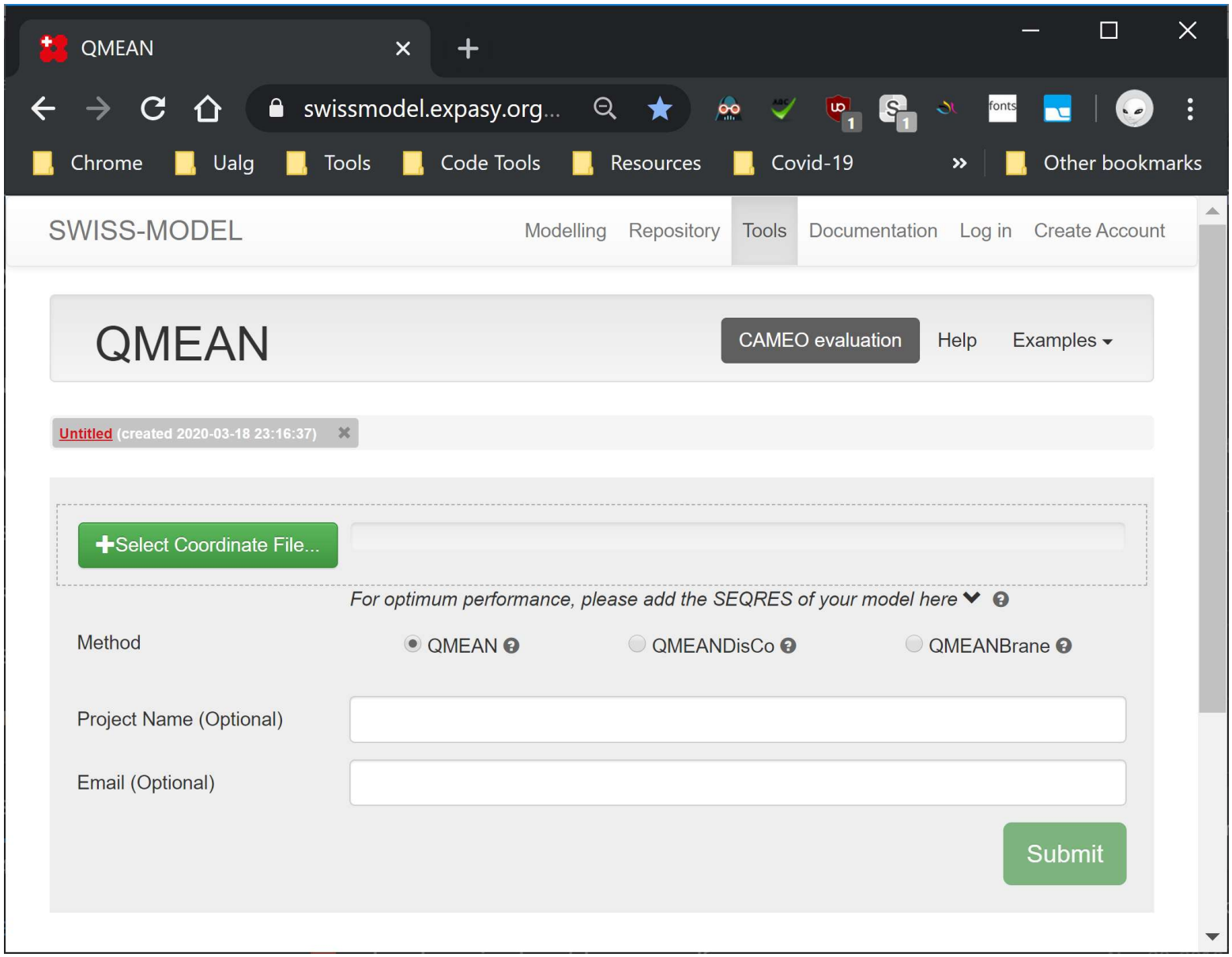
Quality assessment of models

After models are produced by the software, they need to be *evaluated* for their chemical correctness and native-like properties. The can happen at two levels:

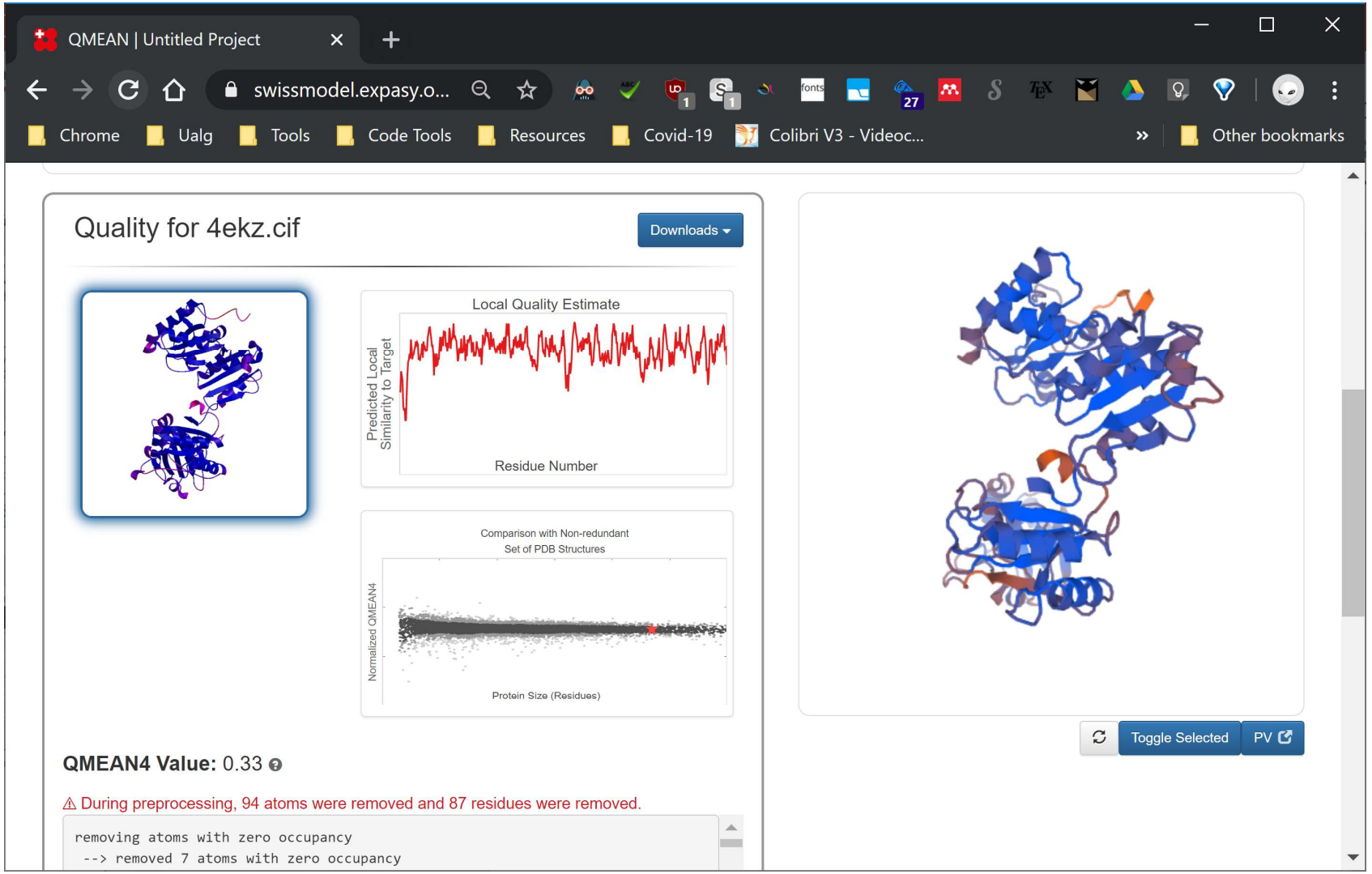
- **Internal –** checks generated by the modelling package itself. In the case of MODELLER, these can be the molecular PDF function, restraint violation lists, and DOPE and GAC stores.

- **External –** There various publicly available quality assessment servers for protein structures, including:
  - **QMEAN:** https://swissmodel.expasy.org/qmean/
  - **Molbrobity: https://molprobity.biochem.duke.edu/**
  - **PDBsum: http://www.ebi.ac.uk/thornton-srv/databases/pdbsum/Generate.html**
  - **Prosa2:** https://prosa.services.came.sbg.ac.at/prosa.php
  - **WHATCHECK:** https://swift.cmbi.umcn.nl/gv/whatcheck/

QMEAN quality assessment



QMEAN quality assessment

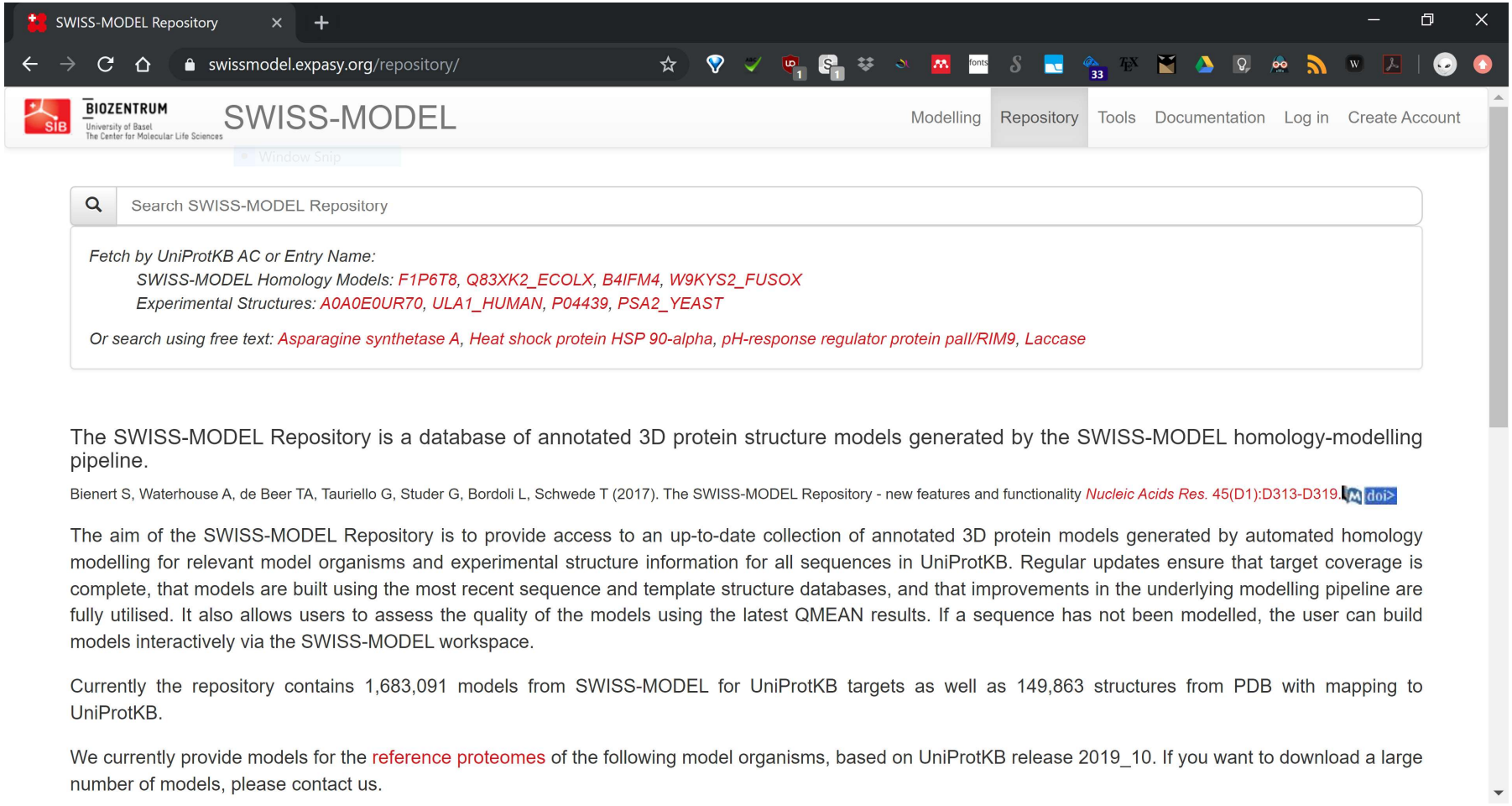


Comparative modelling databases

- Model repositories containing automatically computed models for a large fraction of the known sequences
- Easy quick and dirty way to get a model for a protein of unknown structure.
- Generally “safe” sequence similarities above > 70-75%
- Can be refined or generated for multiple templates
- It is important to look at measures of model quality

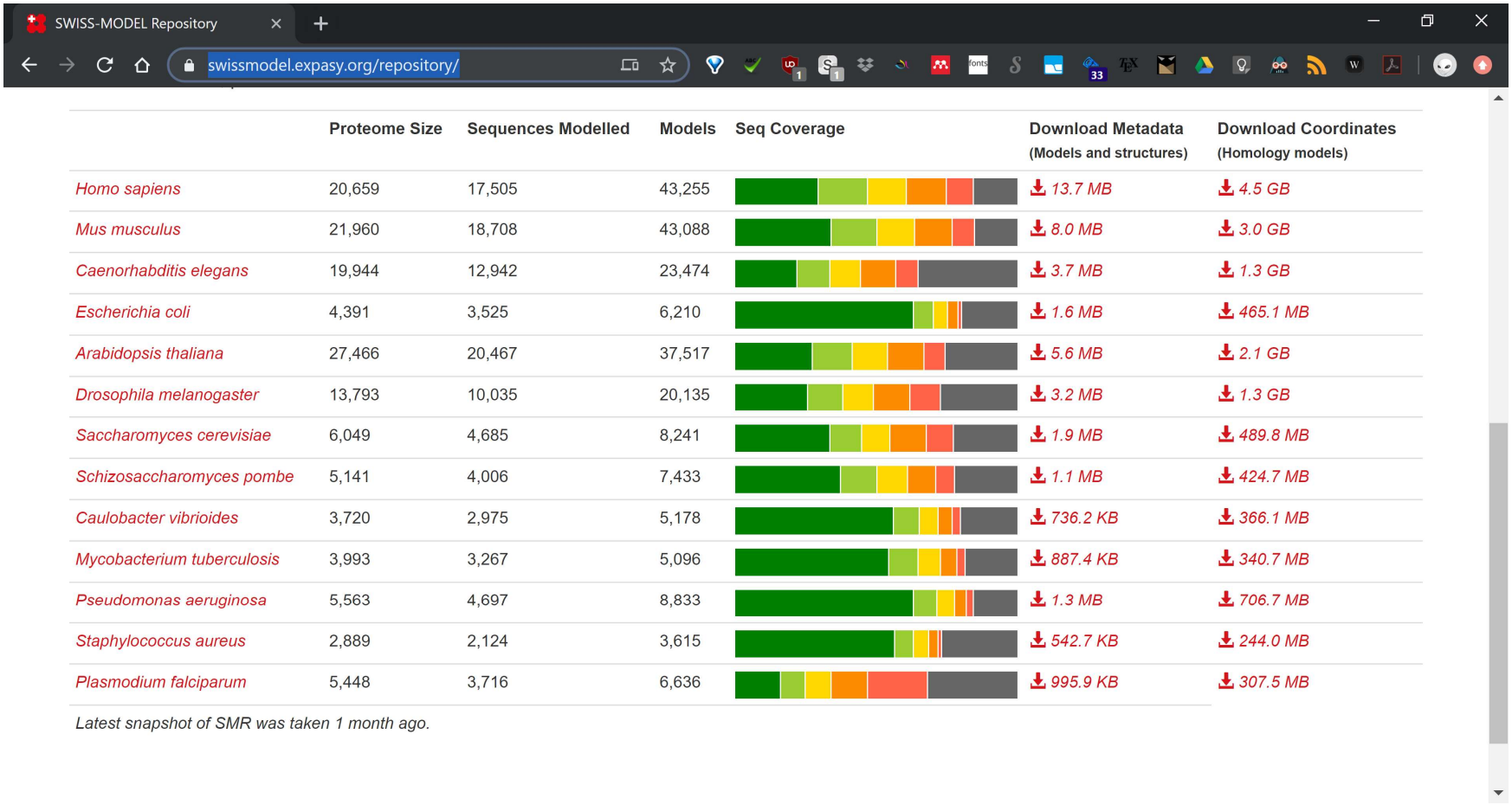
- SWISS Model Repository – https://swissmodel.expasy.org/repository/
- ModBase - https://modbase.compbio.ucsf.edu/i

SWISS MODEL repository



https://swissmodel.expasy.org/repository/

SWISS MODEL repository

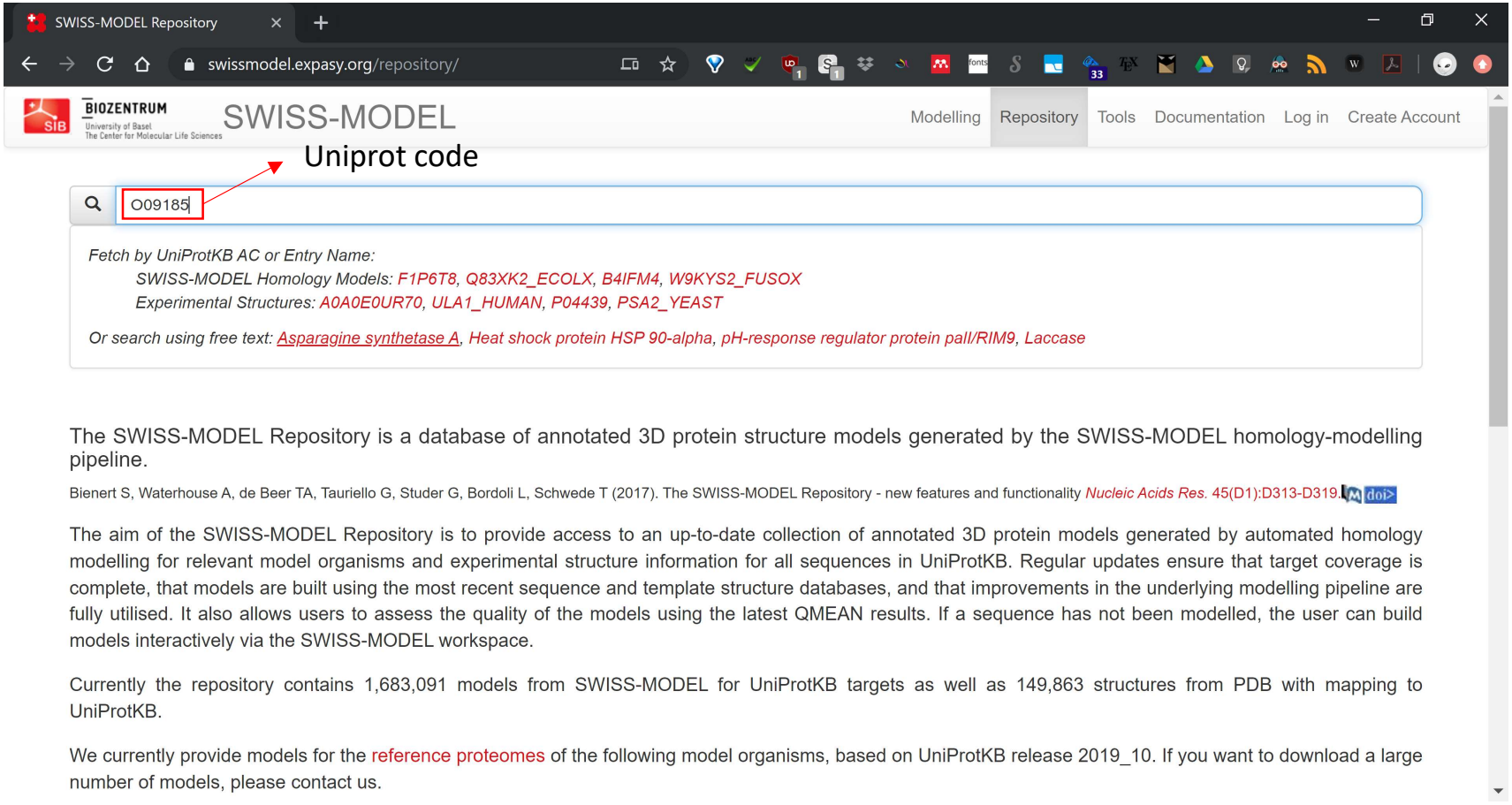


	Proteome Size	Sequences Modelled	Models	Seq Coverage	Download Metadata (Models and structures)	Download Coordinates (Homology models)
Homo sapiens	20,659	17,505	43,255		13.7 MB	4.5 GB
Mus musculus	21,960	18,708	43,088		8.0 MB	3.0 GB
Caenorhabditis elegans	19,944	12,942	23,474		3.7 MB	1.3 GB
Escherichia coli	4,391	3,525	6,210		1.6 MB	465.1 MB
Arabidopsis thaliana	27,466	20,467	37,517		5.6 MB	2.1 GB
Drosophila melanogaster	13,793	10,035	20,135		3.2 MB	1.3 GB
Saccharomyces cerevisiae	6,049	4,685	8,241		1.9 MB	489.8 MB
Schizosaccharomyces pombe	5,141	4,006	7,433		1.1 MB	424.7 MB
Caulobacter vibrioides	3,720	2,975	5,178		736.2 KB	366.1 MB
Mycobacterium tuberculosis	3,993	3,267	5,096		887.4 KB	340.7 MB
Pseudomonas aeruginosa	5,563	4,697	8,833		1.3 MB	706.7 MB
Staphylococcus aureus	2,889	2,124	3,615		542.7 KB	244.0 MB
Plasmodium falciparum	5,448	3,716	6,636		995.9 KB	307.5 MB

Latest snapshot of SMR was taken 1 month ago.

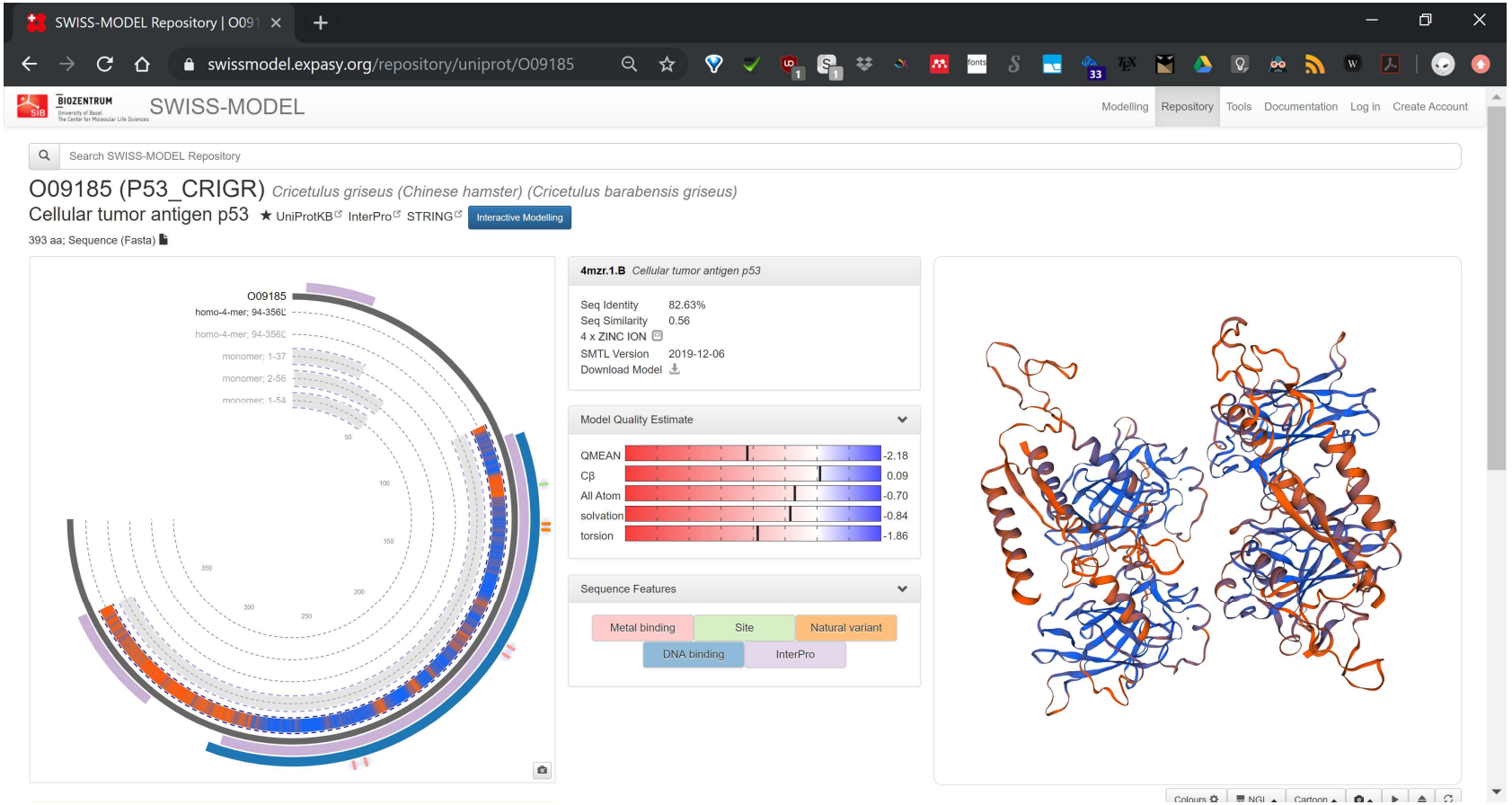
https://swissmodel.expasy.org/repository/

SWISS MODEL repository



https://swissmodel.expasy.org/repository/

SWISS MODEL repository



https://swissmodel.expasy.org/repository/