

Comparative modelling of Protein Structure

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 intricacies, 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)

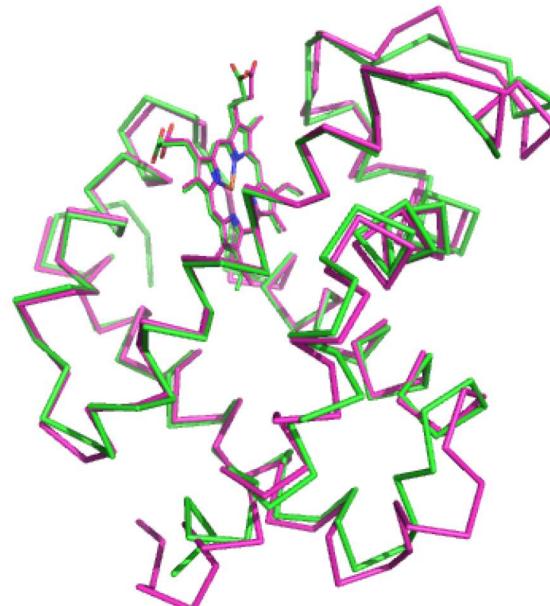
Relation between sequence and structure similarity

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

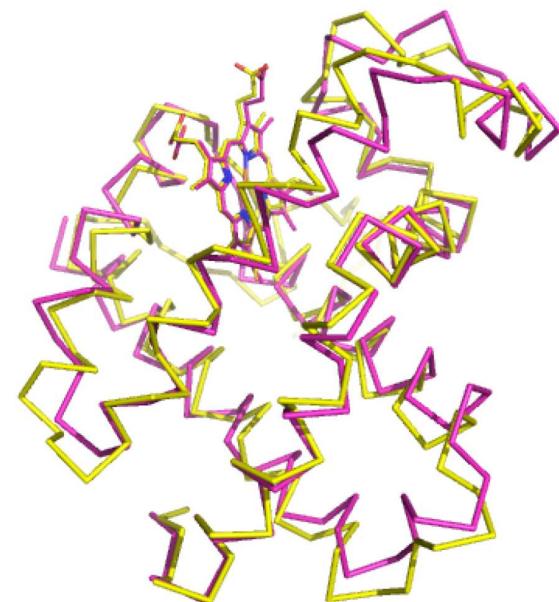
MYOGLOBIN



Human x Horse
88% identity

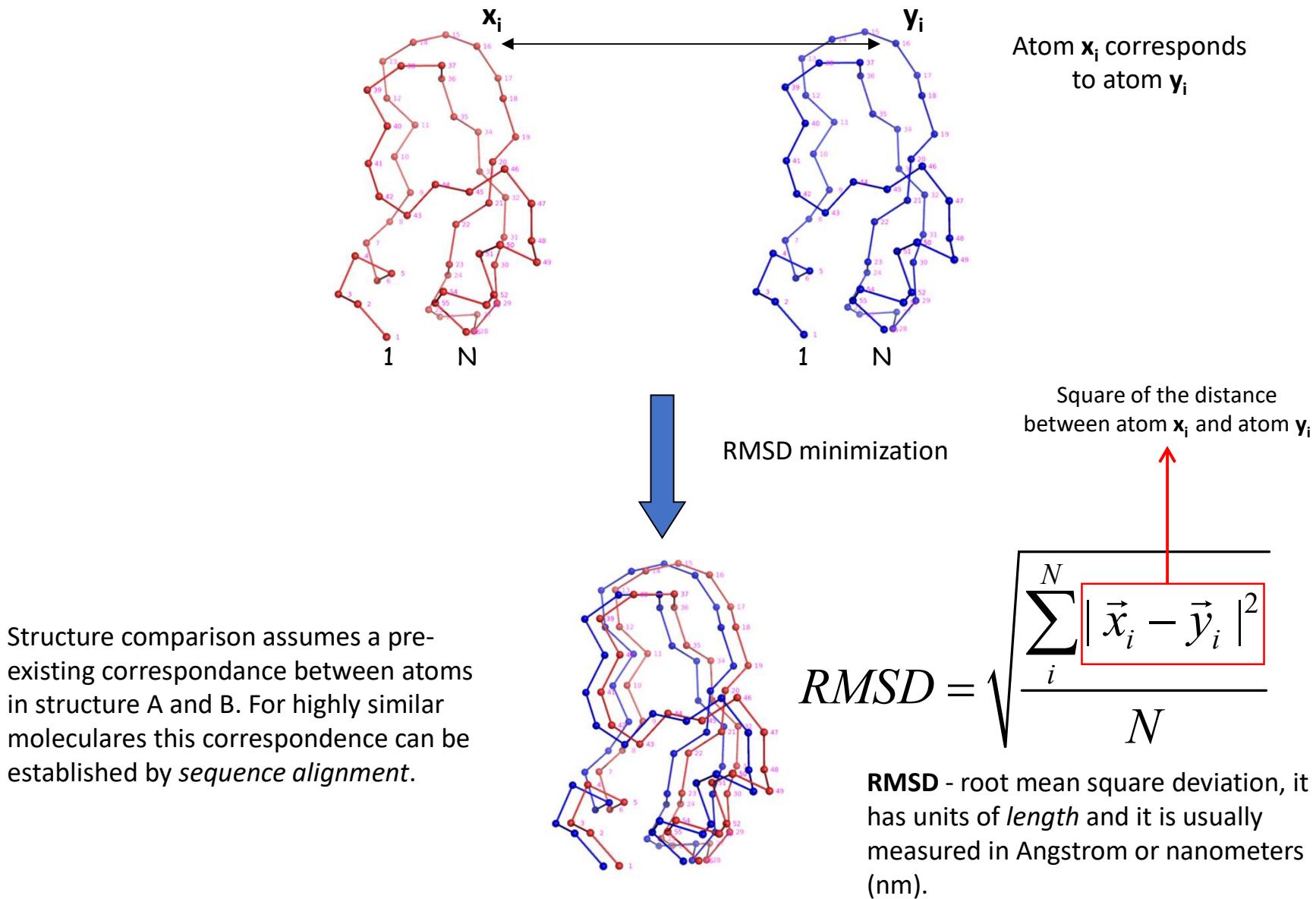


Human x Tuna
45% identity



Myoglobin x Cytoglobin
29% identity

Measuring Structural Similarity



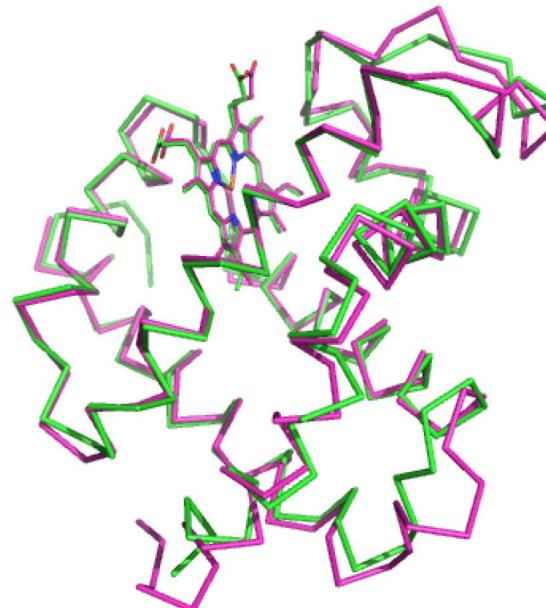
Relation between sequence and structure similarity

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

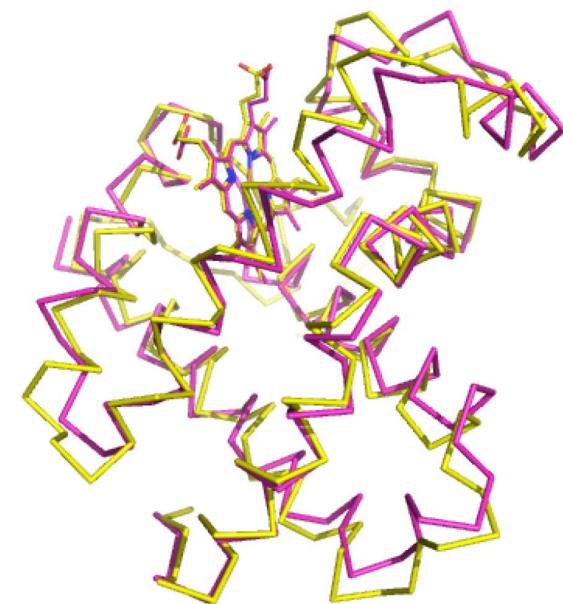
MYOGLOBIN



Human x Horse
88% identity
RMSD = 1.16



Human x Tuna
45% identity
RMSD = 1.72



Myoglobin x Cytoglobin
29% identity
RMSD = 2.53

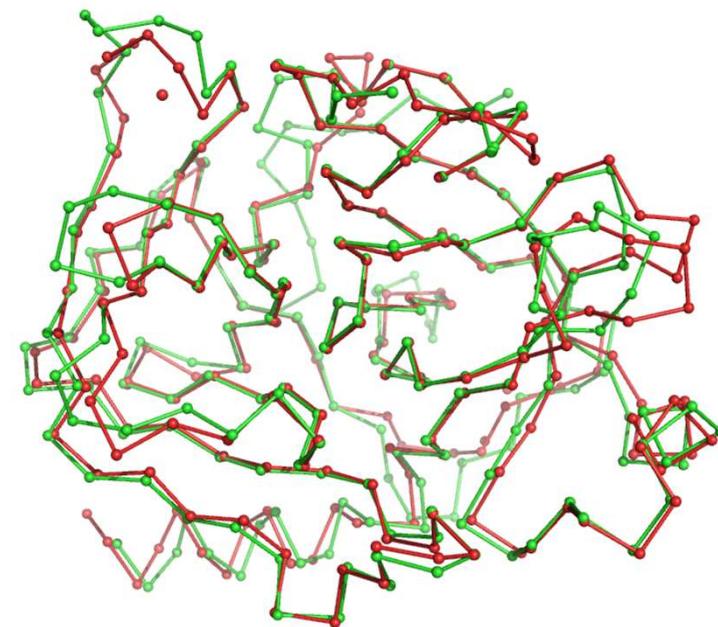
Sequence similarity implies structural similarity



Human x Horse Myoglobin

88% identity

RMSD = 1.16

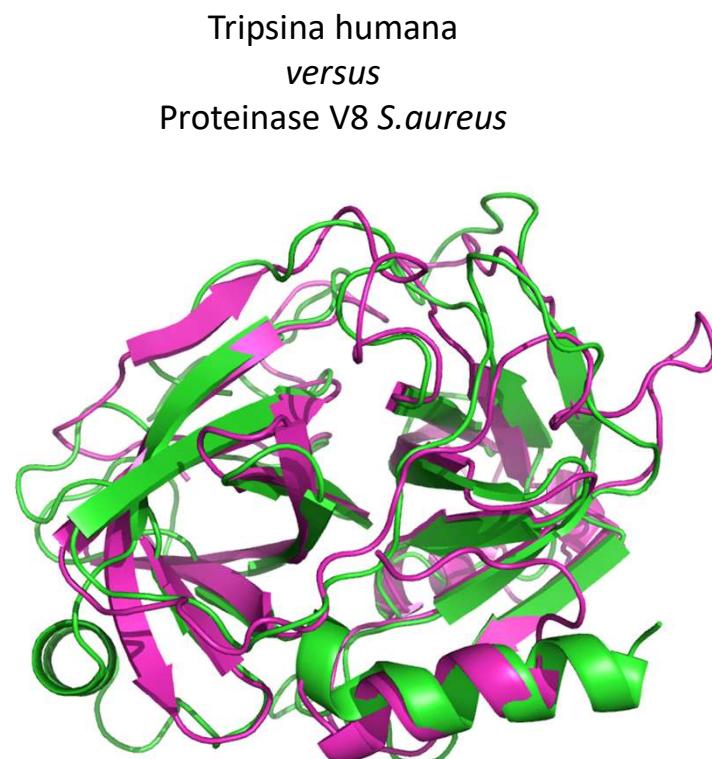


Human x Bovine Trypsin

74% identity

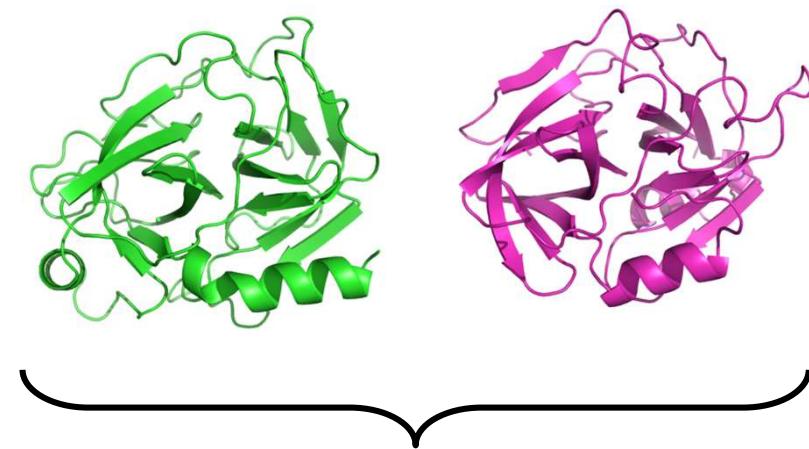
RMSD = 0.76

Structural similarity does not imply sequence similarity



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

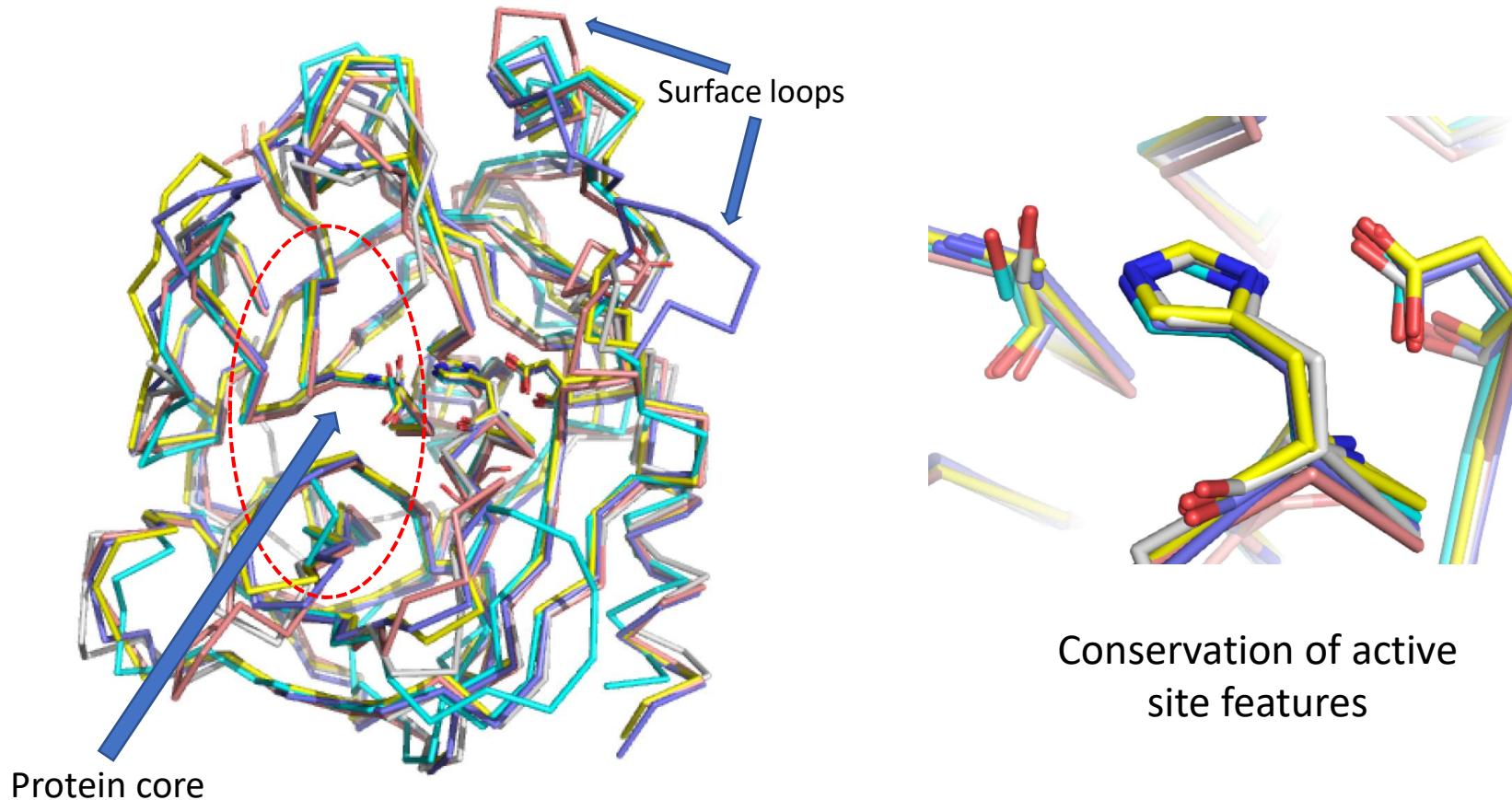
Conservation of structural features

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.

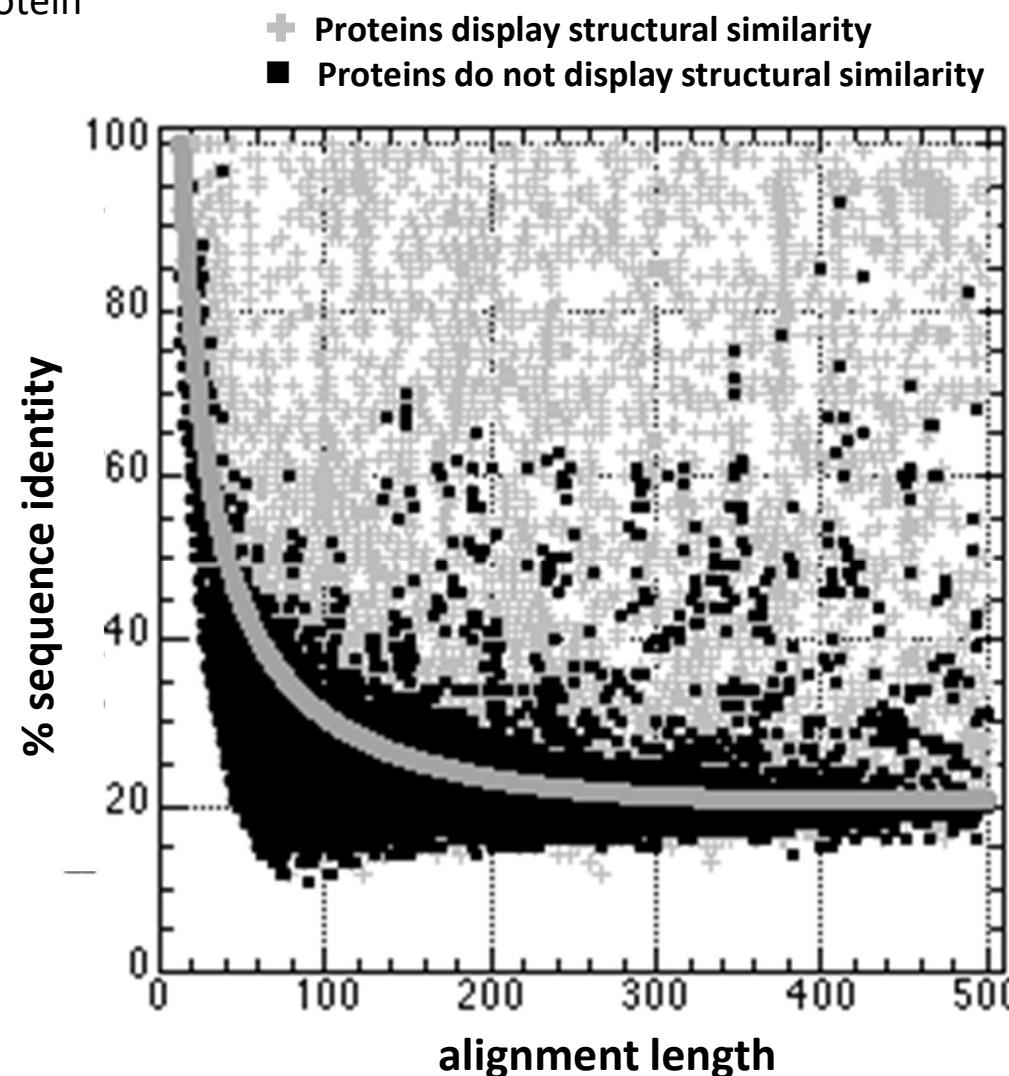
Conservation of structural features

Example: conservation of structural features across the serine protease family

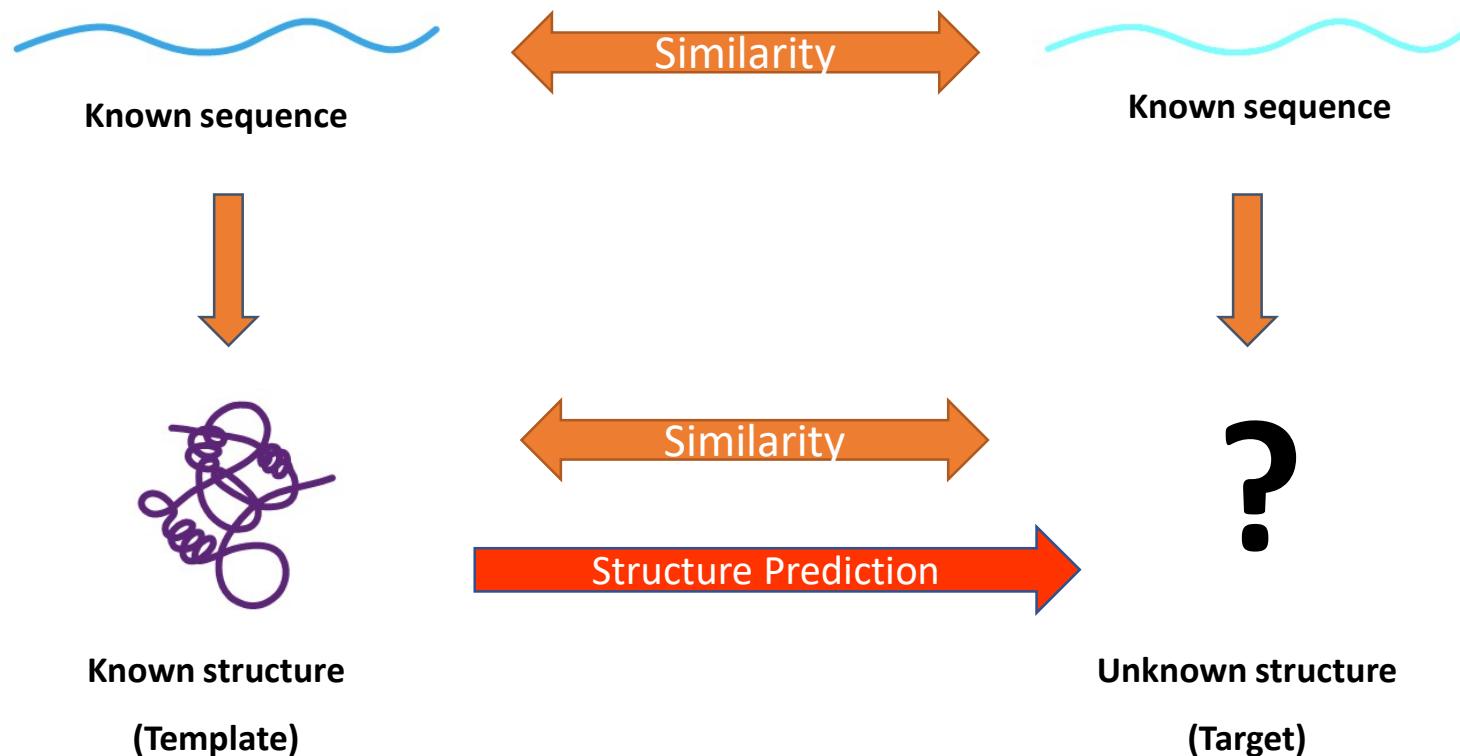


Threshold for structural homology

Each  or  in the graph
represents a protein-protein
comparison



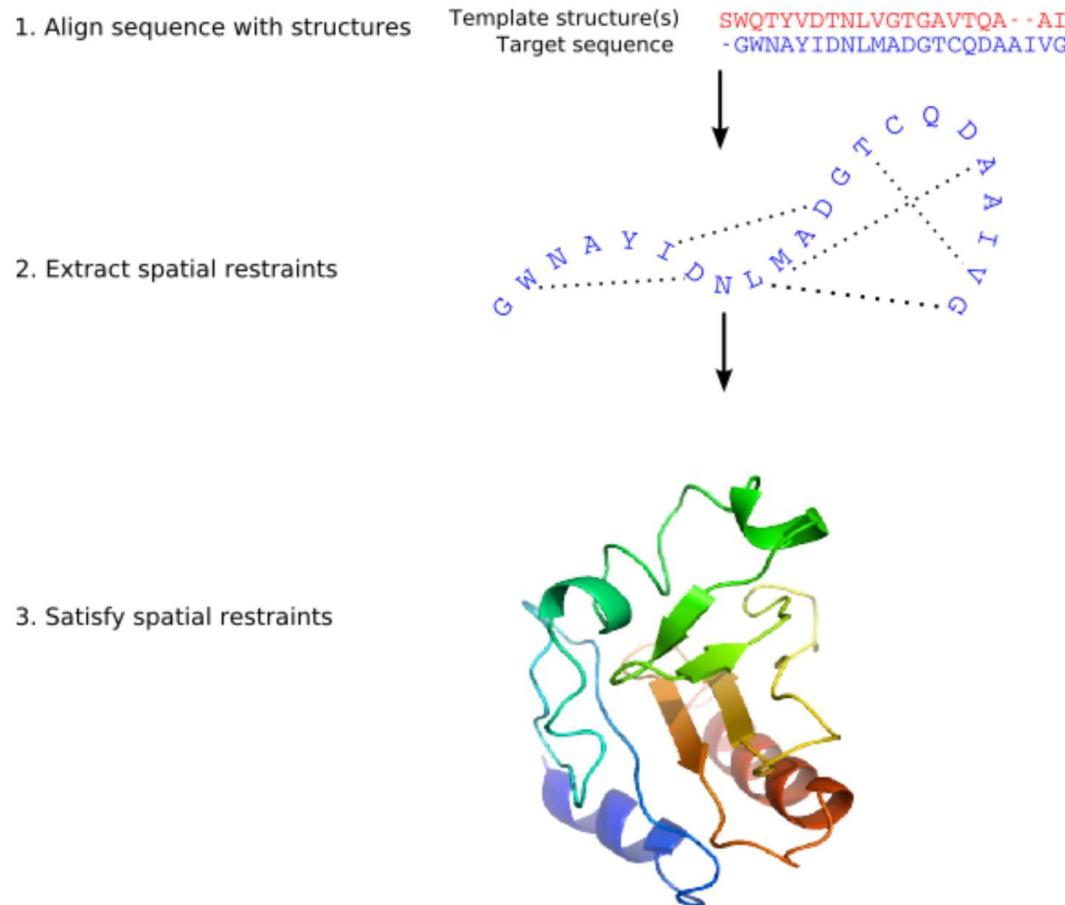
Comparative modelling of protein structures

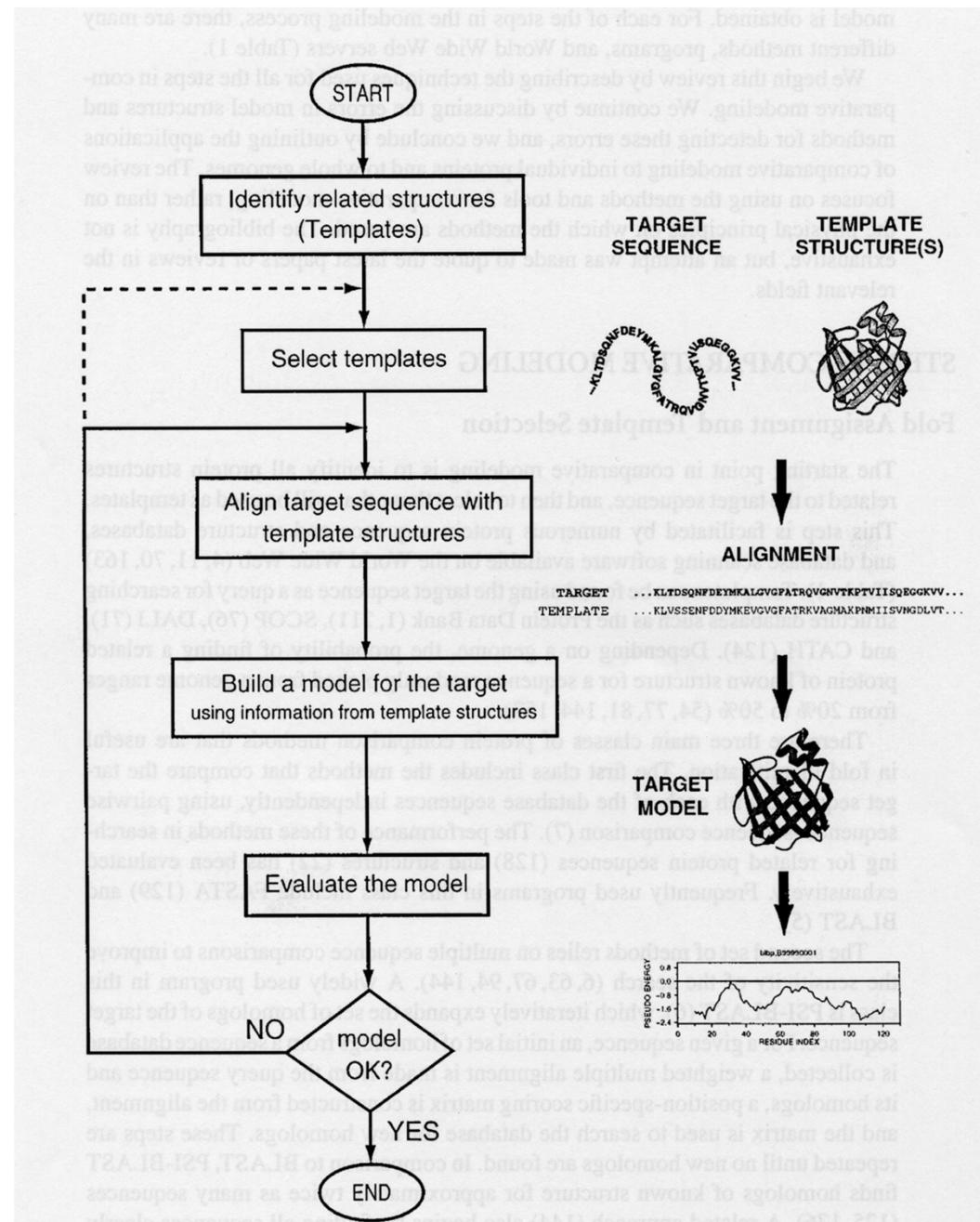


Similar sequences imply similar structures, and so:

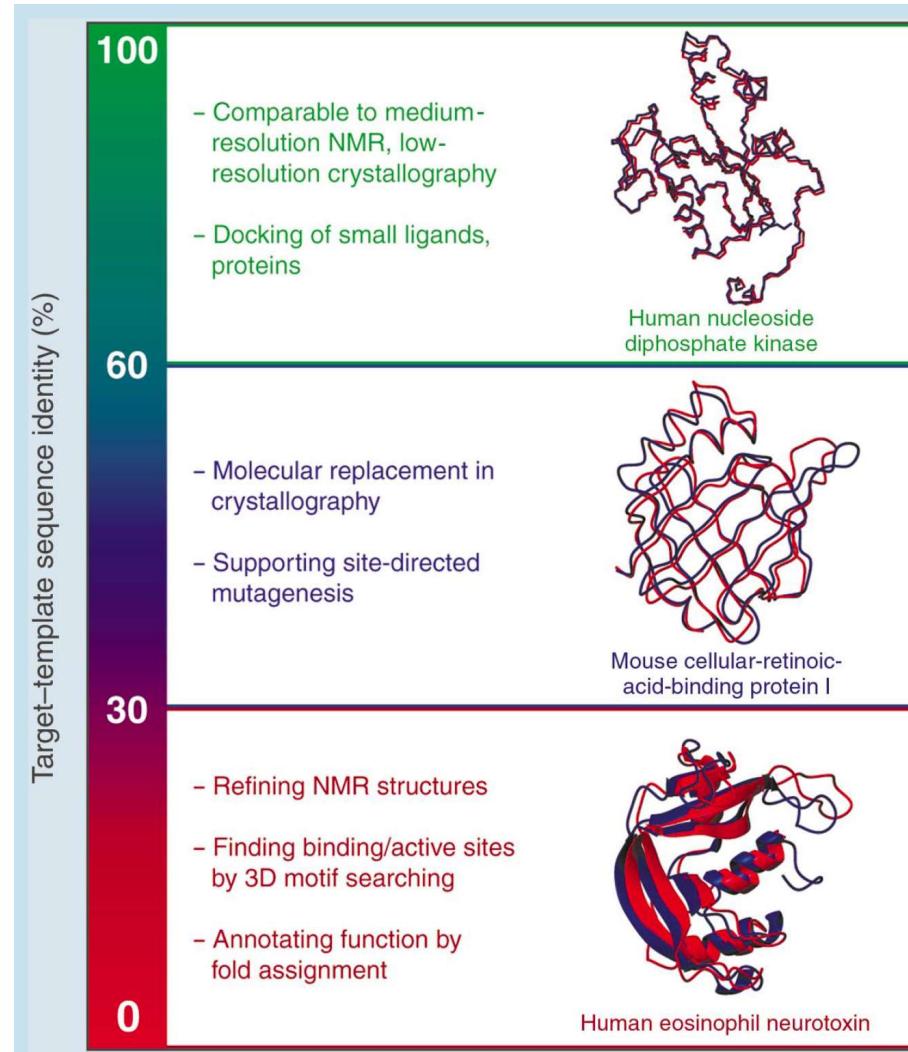
The unknown structure of a target protein can be modelled on the known structure of another protein (template) displaying sufficient sequence similarity with the target.

Modelling by satisfaction of spatial restraints (MODELLER)





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 some 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://raptorg.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 produce extensive databases of publicly 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, go 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 against the Uniprot PDB subset (sequences of proteins of known structure):

The screenshot shows the UniProt BLAST interface. At the top, there's a navigation bar with the UniProt logo, a dropdown menu for 'UniProtKB', and links for 'BLAST', 'Align', 'Retrieve/ID mapping', and 'Peptide search'. Below this is a large orange header 'BLAST'.

The main area contains a 'How to use this tool' section with a brief description of BLAST's purpose. To the right of the description are three numbered steps: 1. E, 2. C, and 3. C.

Below the description is a text input field containing a protein sequence:

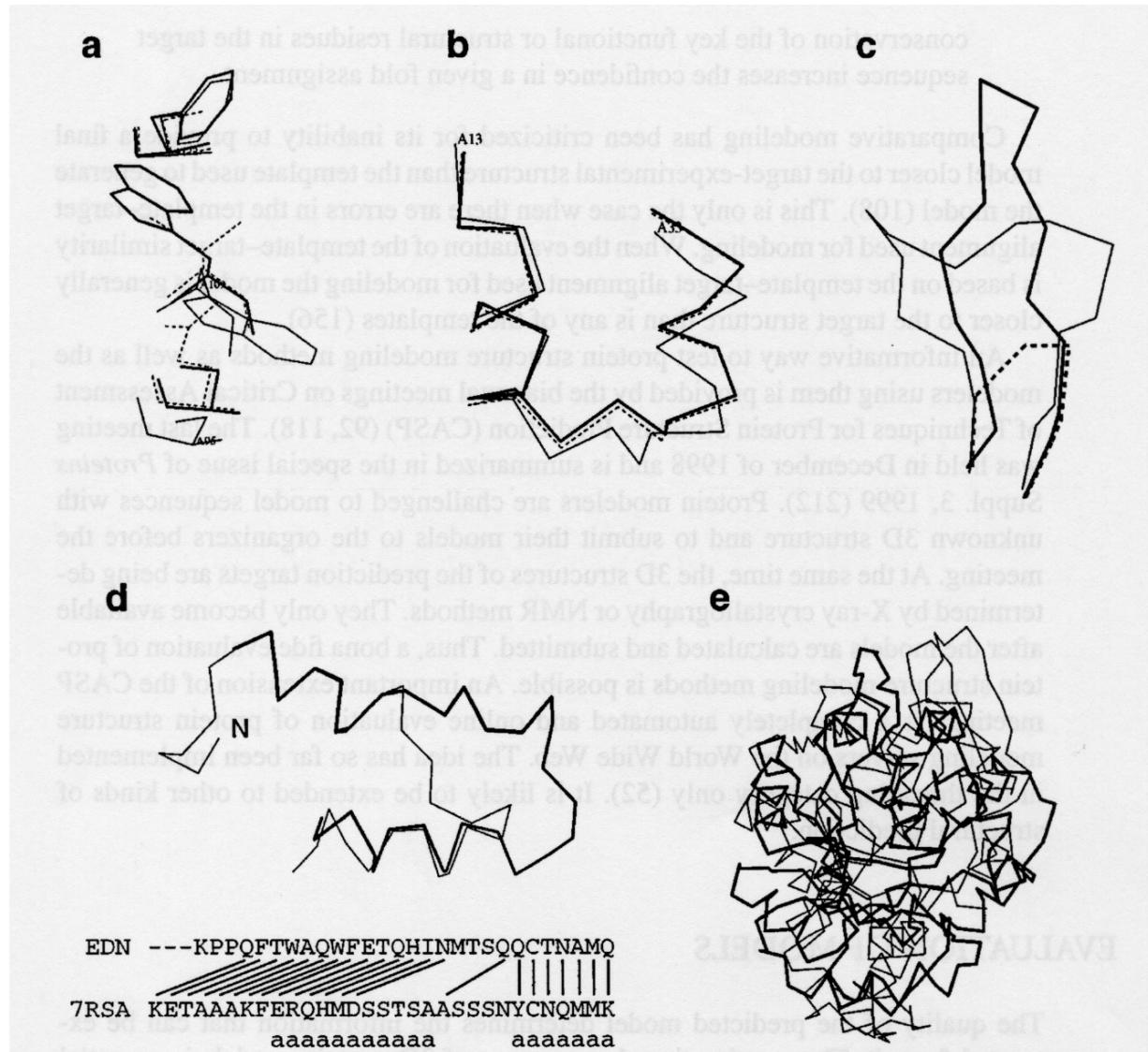
```
>sp|P02144|MYG_HUMAN Myoglobin OS=Homo sapiens OX=9606 GN=MB PE=1 SV=2
MGLSDGEWQLVNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFHLKSEDEMKA
DLKKHGATVLTALGGILKKKGHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKH
PGDFGADAQGAMNKALELFRKDMASNYKELGFQG
```

At the bottom, there are search parameters:

- Target database: ...with 3D structure (PDB) (highlighted with a red box)
- E-Threshold: 10
- Matrix: Auto
- Filtering: None
- Gapped: yes
- Hits: 250

There's also a checkbox for 'Run BLAST in a separate window.' and two buttons at the bottom: 'Clear' and 'Run BLAST'.

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. This 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 are various publicly available quality assessment servers for protein structures, including:
 - **QMEAN**: <https://swissmodel.expasy.org/qmean/>
 - **Molprobity**: <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

The screenshot shows a web browser window for the QMEAN tool on the Swiss-Model website. The title bar reads "QMEAN". The address bar shows the URL "swissmodel.expasy.org...". The toolbar includes standard icons for back, forward, search, and other functions. Below the toolbar, the bookmarks bar lists "Chrome", "Ualg", "Tools", "Code Tools", "Resources", "Covid-19", and "Other bookmarks". The main navigation menu at the top of the page includes "SWISS-MODEL", "Modelling", "Repository", "Tools" (which is selected), "Documentation", "Log in", and "Create Account". The "Tools" section contains a sub-menu with "QMEAN", "CAMEO evaluation", "Help", and "Examples". A sub-menu for "Untitled (created 2020-03-18 23:16:37)" is also visible. The main content area is titled "QMEAN". It features a green button labeled "+Select Coordinate File...". Below this, a note says "For optimum performance, please add the SEQRES of your model here". There are three radio buttons for "Method": "QMEAN" (selected), "QMEANDisCo", and "QMEANBrane". There are two input fields: "Project Name (Optional)" and "Email (Optional)". A large green "Submit" button is located at the bottom right.

QMEAN quality assessment

QMEAN | Untitled Project

swissmodel.expsay.o...

Downloads

Quality for 4ekz.cif

Predicted Local Similarity to Target

Residue Number

Local Quality Estimate

Comparison with Non-redundant Set of PDB Structures

Normalized QMEAN4

Protein Size (Residues)

QMEAN4 Value: 0.33

⚠ During preprocessing, 94 atoms were removed and 87 residues were removed.

removing atoms with zero occupancy
--> removed 7 atoms with zero occupancy

Toggle Selected

PV

This screenshot shows the QMEAN web interface for protein quality assessment. The main panel displays the results for the structure 4ekz.cif. It features a ribbon model of the protein, a line graph showing the predicted local similarity to the target residue by residue, and a scatter plot comparing the structure against a non-redundant set of PDB structures. The QMEAN4 value is 0.33. A message at the bottom notes that 94 atoms and 87 residues were removed during preprocessing. On the right, there is a larger, more detailed 3D ribbon model of the protein. The browser's toolbar and bookmarks bar are visible at the top.

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

The screenshot shows a web browser window for the SWISS-MODEL Repository at swissmodel.expasy.org/repository/. The page features a header with the BIOZENTRUM logo, UniProtKB AC or Entry Name search fields, and sections for Fetch by UniProtKB AC or Entry Name, experimental structures, and free text search. Below the search bar, it displays statistics: 1,683,091 models from SWISS-MODEL for UniProtKB targets and 149,863 structures from PDB with mapping to UniProtKB. It also mentions reference proteomes for various organisms based on UniProtKB release 2019_10.

The SWISS-MODEL Repository is a database of annotated 3D protein structure models generated by the SWISS-MODEL homology-modelling pipeline.

Bienert S, Waterhouse A, de Beer TA, Tauriello G, Studer G, Bordoli L, Schwede T (2017). The SWISS-MODEL Repository - new features and functionality *Nucleic Acids Res.* 45(D1):D313-D319. [\[doi\]](#)

The aim of the SWISS-MODEL Repository is to provide access to an up-to-date collection of annotated 3D protein models generated by automated homology modelling for relevant model organisms and experimental structure information for all sequences in UniProtKB. Regular updates ensure that target coverage is complete, that models are built using the most recent sequence and template structure databases, and that improvements in the underlying modelling pipeline are fully utilised. It also allows users to assess the quality of the models using the latest QMEAN results. If a sequence has not been modelled, the user can build models interactively via the SWISS-MODEL workspace.

Currently the repository contains 1,683,091 models from SWISS-MODEL for UniProtKB targets as well as 149,863 structures from PDB with mapping to UniProtKB.

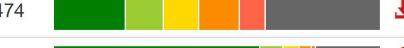
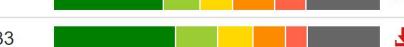
We currently provide models for the **reference proteomes** of the following model organisms, based on UniProtKB release 2019_10. If you want to download a large number of models, please contact us.

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

SWISS MODEL repository

SWISS-MODEL Repository

swissmodel.expasy.org/repository/

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

The screenshot shows a web browser window for the SWISS-MODEL Repository. The URL in the address bar is swissmodel.expasy.org/repository/. The page header includes the BIOZENTRUM logo and the SWISS-MODEL logo. A navigation menu at the top right offers links to Modelling, Repository, Tools, Documentation, Log in, and Create Account. Below the header, a search bar contains the text "O09185". An arrow from the left side of the image points to this search bar. The main content area displays search results for the UniProt code O09185, including homology models (F1P6T8, Q83XK2_ECOLX, B4IFM4, W9KYS2_FUSOX) and experimental structures (AOA0E0UR70, ULA1_HUMAN, P04439, PSA2_YEAST). It also provides a free text search option.

The SWISS-MODEL Repository is a database of annotated 3D protein structure models generated by the SWISS-MODEL homology-modelling pipeline.

Biernert S, Waterhouse A, de Beer TA, Tauriello G, Studer G, Bordoli L, Schwede T (2017). The SWISS-MODEL Repository - new features and functionality *Nucleic Acids Res.* 45(D1):D313-D319. [\[doi\]](#)

The aim of the SWISS-MODEL Repository is to provide access to an up-to-date collection of annotated 3D protein models generated by automated homology modelling for relevant model organisms and experimental structure information for all sequences in UniProtKB. Regular updates ensure that target coverage is complete, that models are built using the most recent sequence and template structure databases, and that improvements in the underlying modelling pipeline are fully utilised. It also allows users to assess the quality of the models using the latest QMEAN results. If a sequence has not been modelled, the user can build models interactively via the SWISS-MODEL workspace.

Currently the repository contains 1,683,091 models from SWISS-MODEL for UniProtKB targets as well as 149,863 structures from PDB with mapping to UniProtKB.

We currently provide models for the **reference proteomes** of the following model organisms, based on UniProtKB release 2019_10. If you want to download a large number of models, please contact us.

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

SWISS MODEL repository

swissmodel.expasy.org/repository/uniprot/O09185

BIOZENTRUM University of Basel The Center for Molecular Life Sciences SWISS-MODEL

O09185 (P53_CRIGR) *Cricetulus griseus* (Chinese hamster) (*Cricetulus barabensis* griseus)
Cellular tumor antigen p53 ★ UniProtKB[®] InterPro[®] STRING[®]

393 aa; Sequence (Fasta)

009185
homo-4-mer; 94-356L
homo-4-mer; 94-356L
monomer; 1-37
monomer; 2-56
monomer; 1-54

50 100 150 200 250 300 350

4mzr.1.B Cellular tumor antigen p53

Seq Identity 82.63%
Seq Similarity 0.56
4 x ZINC ION
SMTL Version 2019-12-06
Download Model

Model Quality Estimate

QMEAN	-2.18
C β	0.09
All Atom	-0.70
solvation	-0.84
torsion	-1.86

Sequence Features

Metal binding Site Natural variant
DNA binding InterPro

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