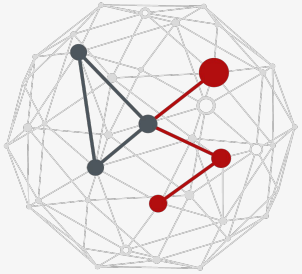


DATA SCIENCE
UNIVERSITY OF PADOVA

COMPARATIVE MODELLING

Master of Science in Data Science

Damiano Piovesan

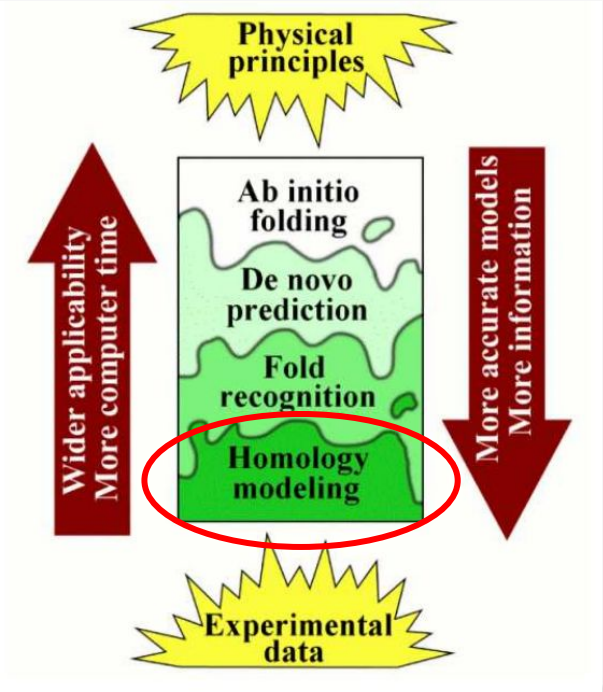
Structure prediction - why?

The **protein structure** allows a molecular and mechanistic understanding of the **protein function**

- Identification of **active sites** and the positions of **key residues**
- Prediction of protein-protein and protein-ligand **interactions**, which are mostly determined by steric (shape) and chemical (e.g. charge) complementarity
- Filling the **sequence / structure gap**. Million sequences are known, while the PDB contain only ca. 200K structures

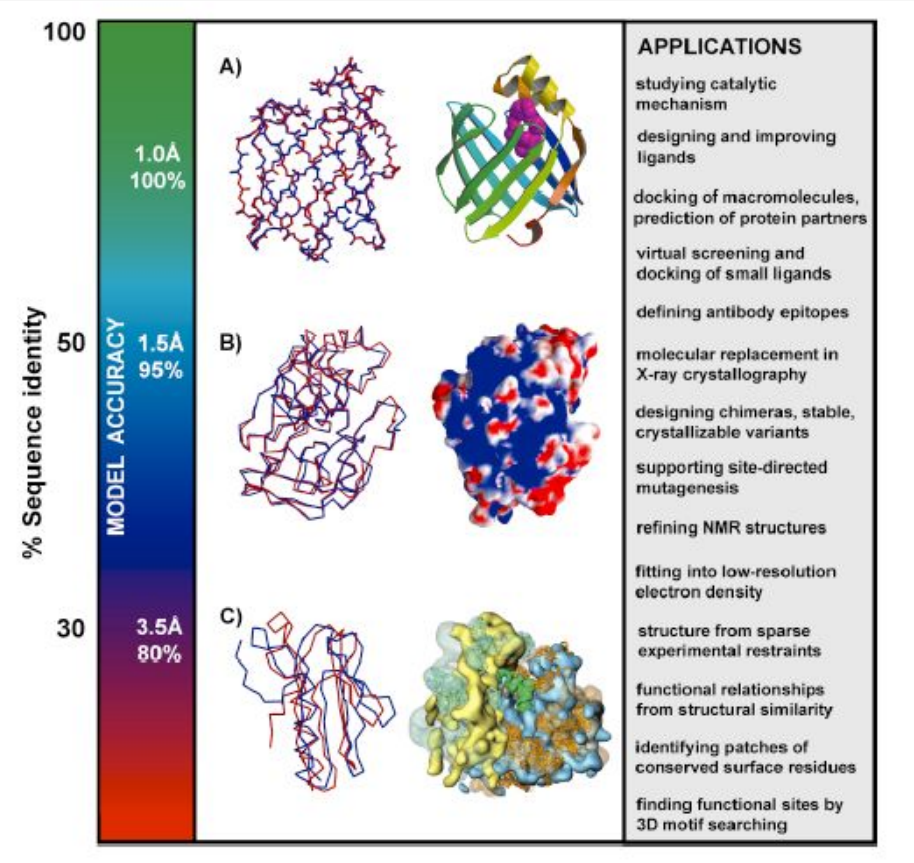
However

- Many proteins have the sequence not similar enough to build an *in silico* model by homology
- Many folds are not represented in the PDB

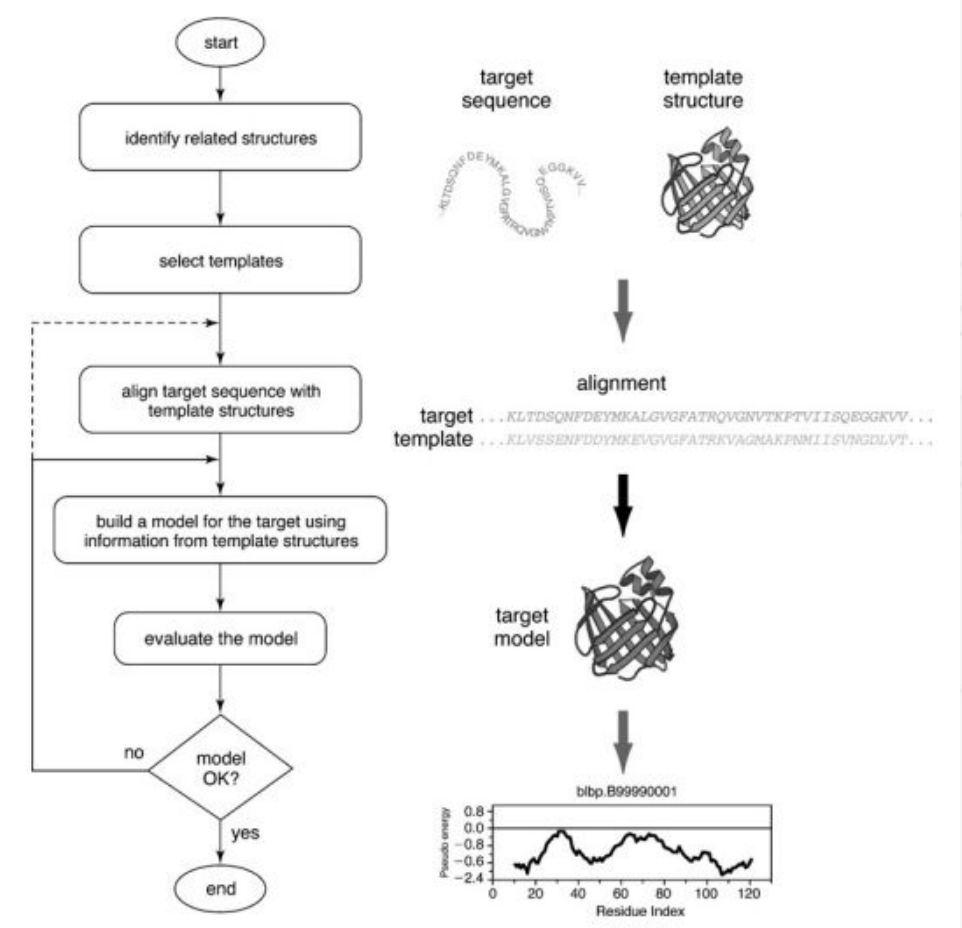


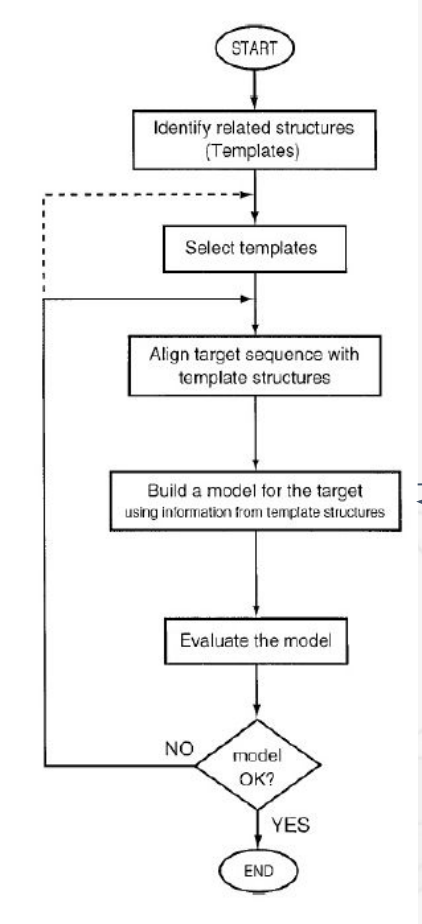
- **De novo prediction / Ab initio**
  - Secondary structure prediction; conformation of short fragments (Rosetta); molecular dynamics; Monte Carlo; quantum mechanics (unfeasible)
  - Heavy computation
- **Fold recognition**
  - Try to fit with known folds
  - The fold space is not completely known (50% success)
- **Homology modelling**
  - Similar sequences have similar structures (>30% sequence identity)
  - 40% of genes are not homologous to known structures

Model accuracy



Comparative modelling





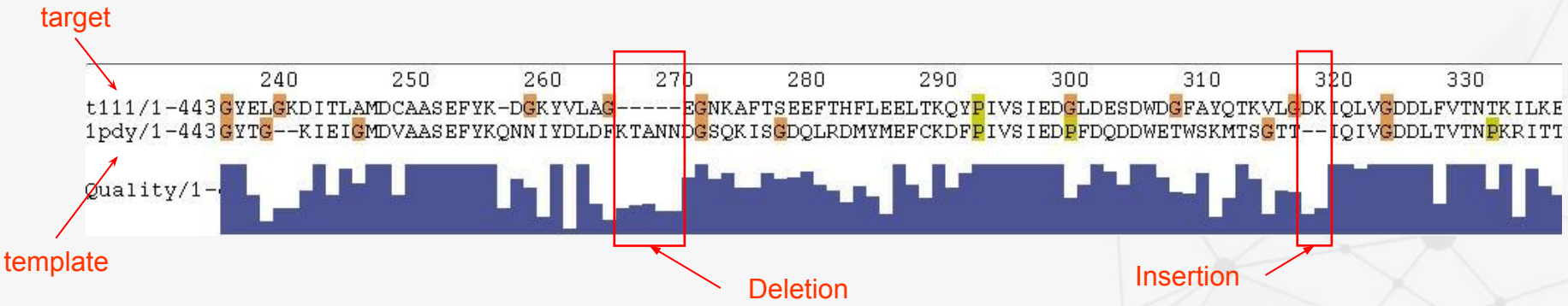
Raw model

Loop modeling

Side chain placement

Refinement

Alignment



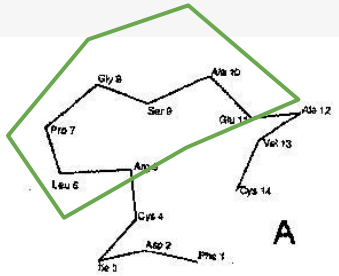
1. **Database search** - Find homologous sequences with known structure. Generally a heuristic algorithm **BLAST / PSI-BLAST**

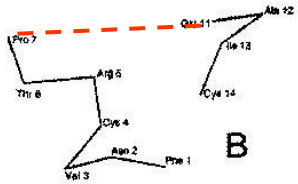
2. **Pairwise alignment** - Assign equivalent positions between **target** and **template**. Determine insertion and deletion. Optimal alignment with **Smith-Waterman** (local) or **Needleman-Wunsch** (global) algorithms

Improve the sequence alignment

- Errors in the alignment cannot be corrected in the following steps!
- Often the best sequence alignment is non optimal for the structure

How do you model this part?





Better sequence alignment
Worse structure model

Worse sequence alignment
Better structure model

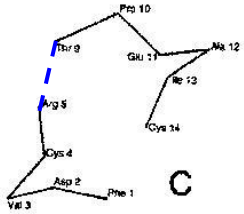


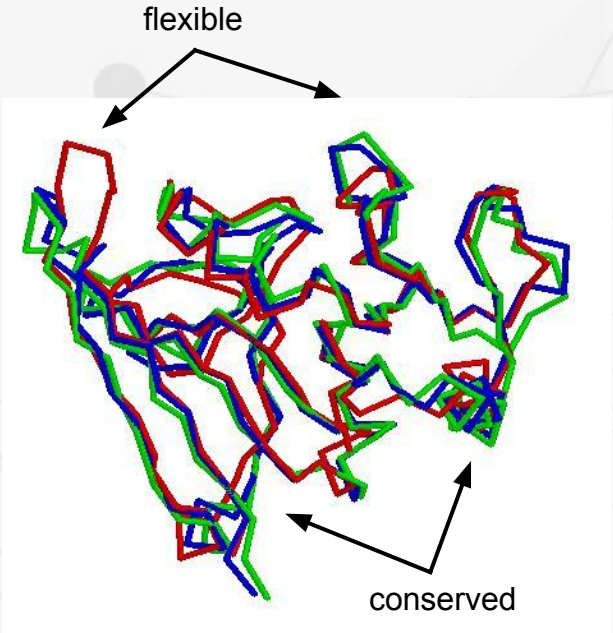
Figure 8. The C_α traces resulting from the alignment in figure 7. A: template. B: middle alignment from figure 7. C: bottom alignment from figure 7.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	
A	PHE	ASP	ILE	CYS	ARG	LEU	PRO	GLY	SER	ALA	GLU	ALA	VAL	CYS	(green in fig 8)
B	PHE	ASN	VAL	CYS	ARG	THR	PRO	---	---	---	GLU	ALA	ILE	CYS	(red in fig 8)
C	PHE	ASN	VAL	CYS	ARG	---	---	---	THR	PRO	GLU	ALA	ILE	CYS	(blue in fig 8)

Figure 7. Example of sequence alignment in an area where a deletion needs to be modelled.

Building the raw model

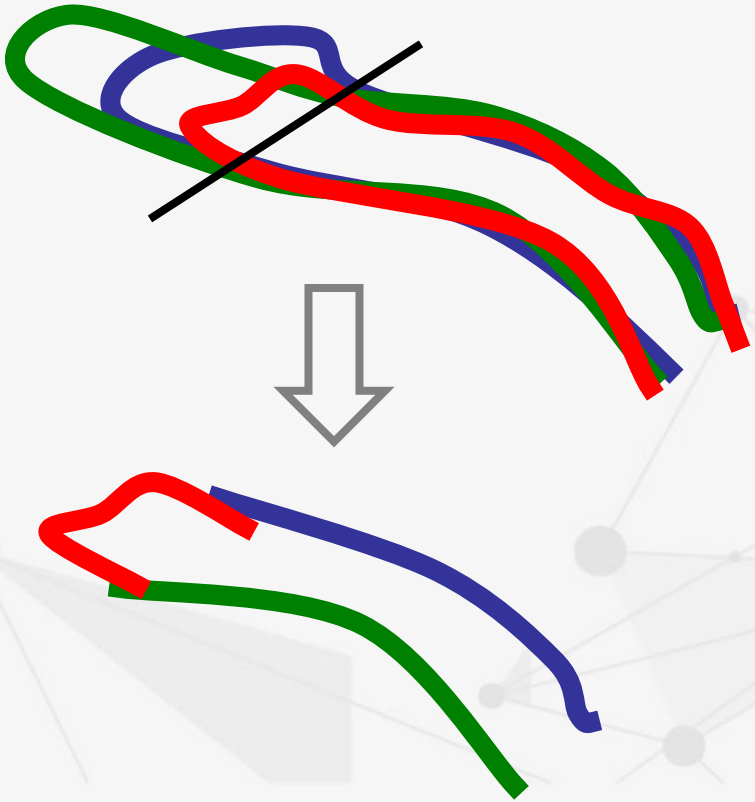
- 3D coordinates of the template residues can be directly used
- The **variable regions** of the structures (generally loops) and in particular position near indels have to be **predicted**



Two principal methods are used for the construction of the raw model

- Fragment-based
- Restraint-based

Fragment-based building



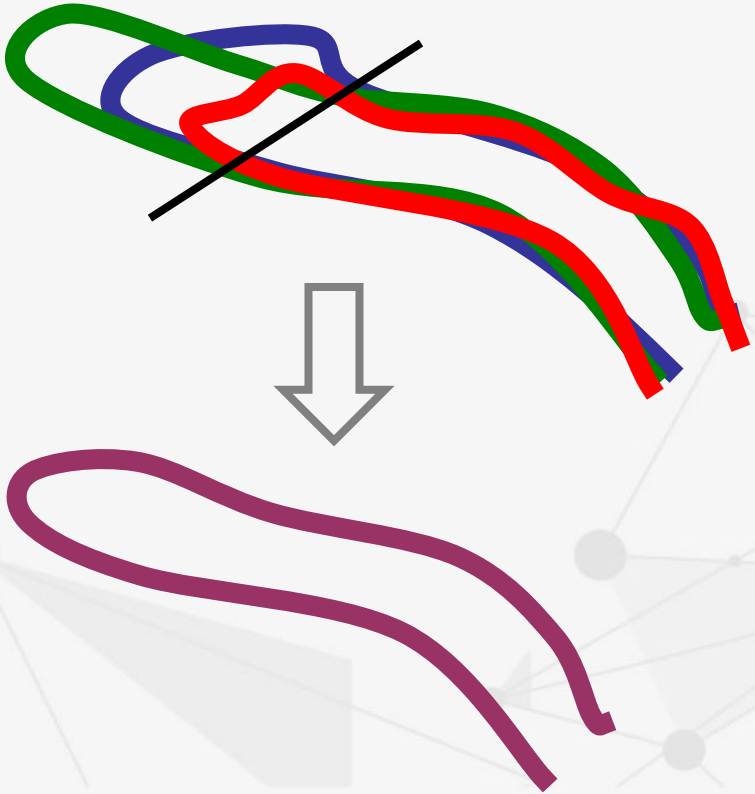
Idea → Copy “useful” coordinates of fragments

- Build a “sharp” set
- Keeps the geometry, eg the active site

Software

- 3D-JIGSAW (Bates et al.)
- COMPOSER (Blundell et al.)
- HOMER (Tosatto et al.)

Restraint-based building



Idea → Use the template to derive restrictions at the atomic positions. Optimize the structure based on the restrictions

- “Spread” errors on the whole structure, but minimize it globally
- Does not ensure the local geometry, eg. in the active site

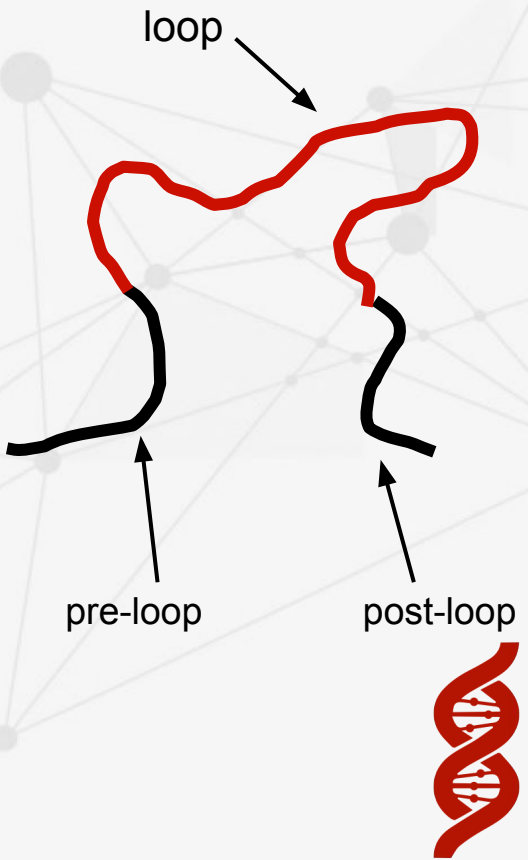
Software

- MODELLER (Šali et al.)

Loop modeling

- Entire fragments of backbone can be missing in the raw-model
  - Not conserved in the protein family
  - Insertion
  - Deletion

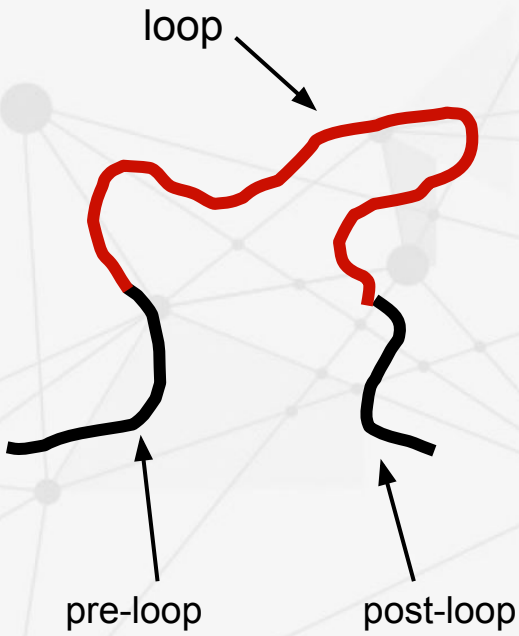
- Problem description
  - Identify the conformation of the fragment (loop, *k* residues) that can connect the pre- to the post-loop
  - Φ and ψ are the only free parameters



Loop modeling

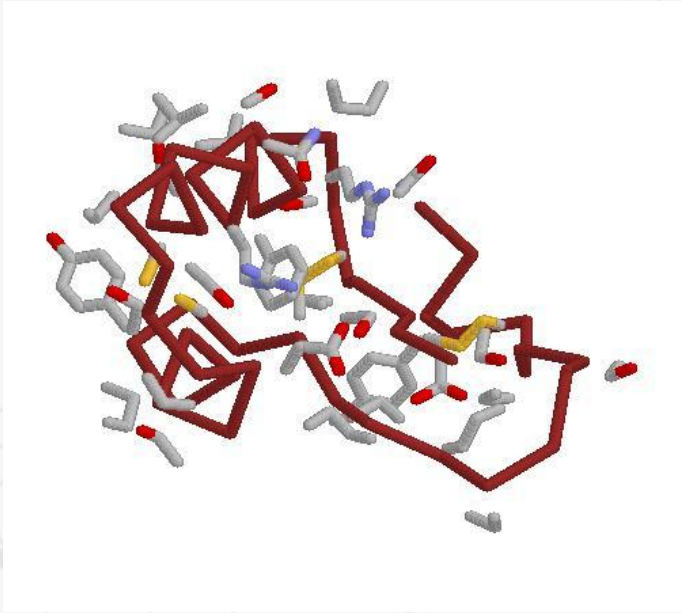
- Database methods
  - Extract loop fragments from PDB
  - Choose the fragment that fits better, based on geometric constraints
  - Not all possible conformations are available in PDB

- Ab initio methods
  - Identify best conformations based on the geometric constraints (torsion angles)
  - Select the “best” fragment
  - Problem: computing time



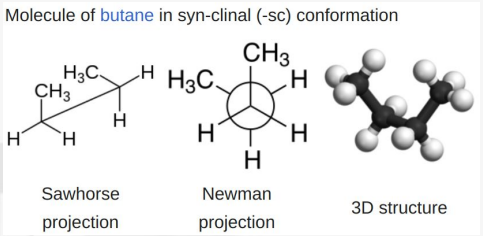
Side chains

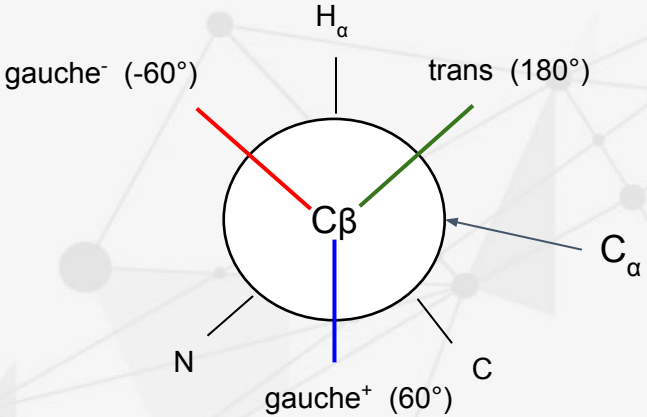
- **Amino acid differences** are not managed when applying the coordinates of the template to the sequence of the target (dimension and position of the side chains)
- Assuming **50% sequences identity**, half side-chains are replaced
- The **RMSD** change is relatively low, but the conformation of important residues (eg. active site) may change
- Effective methods exist to solve this problem, eg. **SCWRL**

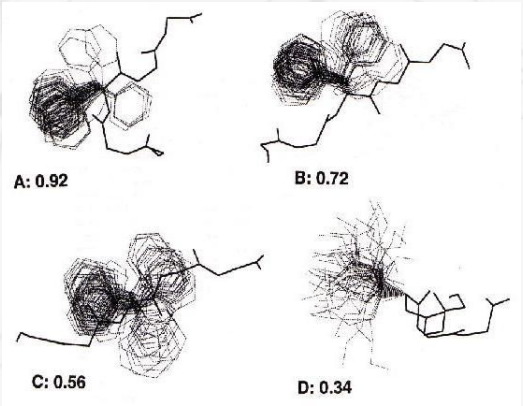


Side chains

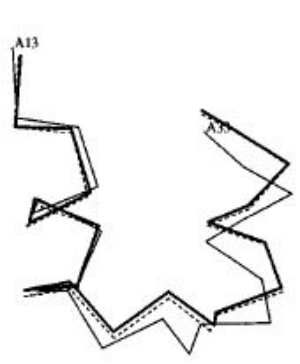
- Rotamers → 3 preferred positions for each torsion angles χ
- The propensity of a rotamer depends on the backbone torsion angles (ϕ, ψ) and the type of amino acid
- Interdependence, domino effect
- Where possible, it is better to maintain the conformation of the template side chains



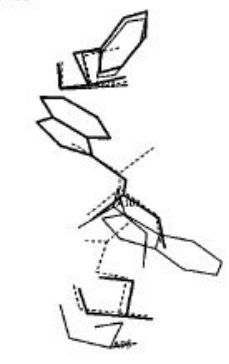




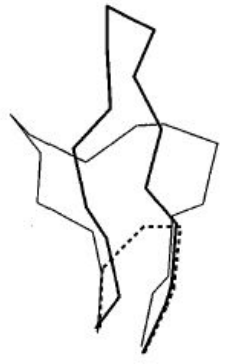
Typical errors



Shift

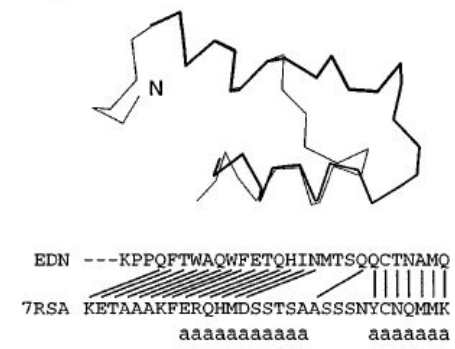


Side chains



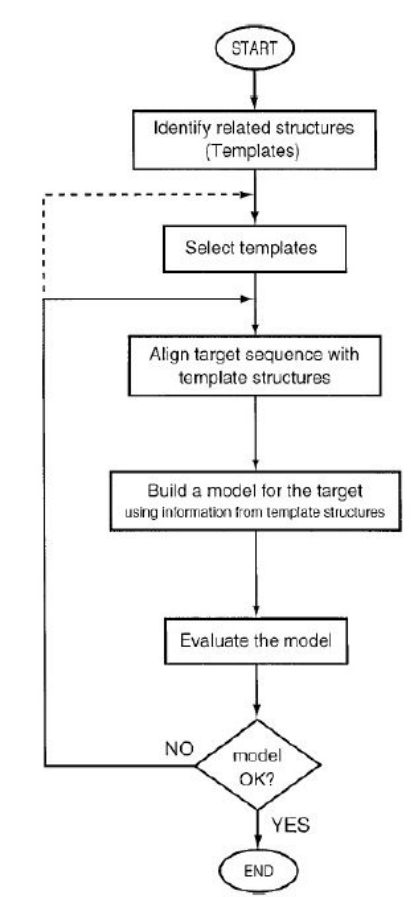
Loops

Wrong template



Wrong alignment

Model assessment

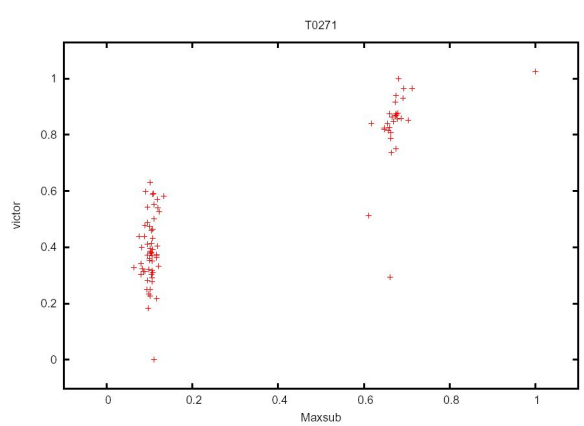


- How do we solve these “problems” without worsening the model?
- Models from **alternative alignments** considering **energy profiles**
  - It is necessary to know which part of the alignment to change
- **CASP** suggested the “*Don’t touch it*” philosophy for a long time
  - It is better to avoid local modifications of the structure
  - Changed over the last few years

Model assessment

- Quality parameters
  - Steric hindrance and **clashes**
  - Deviation from the **geometry** of standard parameters
  - Frequency profiles or energy (**statistical potentials**)

- Software
  - *PROCHECK*
  - *VERIFY-3D*
  - *FRST*
  - *QMEAN*



DIPARTIMENTO MATEMATICA

SWISS-MODEL

Modelling Repository Tools Documentation Log in Create Account

Welcome to SWISS-MODEL

SWISS-MODEL is a fully automated protein structure homology-modelling server, accessible via the ExPASy web server, or from the program DeepView (Swiss Pdb-Viewer). The purpose of this server is to make protein modelling accessible to all life science researchers worldwide.

Start Modelling

Protein Structure Bioinformatics Group
c/o Prof. Torsten Schwede
Swiss Institute of Bioinformatics
Biozentrum, University of Basel
Klingelbergstrasse 50/70
CH-4056 Basel / Switzerland
help-swissmodel@unibas.ch

BIOZENTRUM
The Center for Molecular Life Sciences

20 YEARS SIB Swiss Institute of Bioinformatics

When you publish or report results using SWISS-MODEL, please cite the relevant publications:

- Biasini, M., Bienert, S., Waterhouse, A., Arnold, K., Studer, G., Schmidt, T., Kiefer, F., Cassarino, T.G., Bertoni, M., Bordoli, L., Schwede, T. SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. Nucleic Acids Res. 42, W252-W258 (2014). doi
- Bienert, S., Waterhouse, A., de Beer, T.A., Tauriello, G., Studer, G., Bordoli, L., Schwede, T. The SWISS-MODEL Repository - new features and functionality. Nucleic Acids Res. 45, D313-D319 (2017). doi
- Guex, N., Peitsch, M.C., Schwede, T. Automated comparative protein structure modeling with SWISS-MODEL and Swiss-PdbViewer: A historical perspective. Electrophoresis 30, S162-S173 (2009). doi
- Benkert, P., Biasini, M., Schwede, T. Toward the estimation of the absolute quality of individual protein structure models. Bioinformatics 27, 343-350 (2011). doi
- Bertoni, M., Kiefer, F., Biasini, M., Bordoli, L., Schwede, T. Modeling protein quaternary structure of homo- and hetero-oligomers beyond binary interactions by homology. Scientific Reports 7 (2017). doi