1222·2022 800 ANNI

UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO MATEMATICA

DATA SCIENCE
UNIVERSITY OF PADOVA

# COMPARATIVE MODELLING

Master of Science in Data Science

Damiano Piovesan

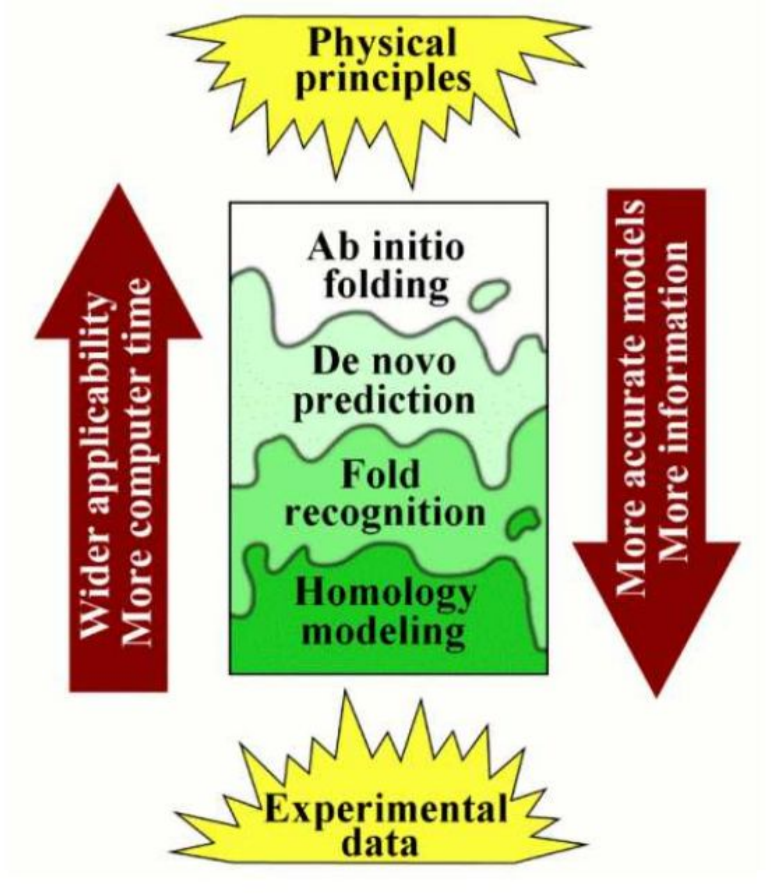
# Motivation

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

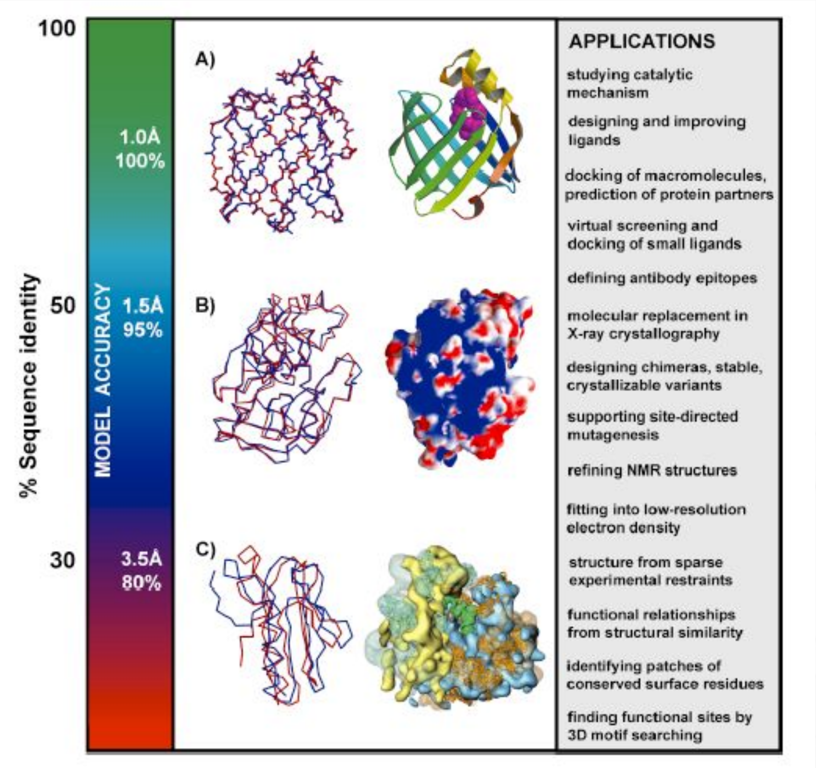
- Identification of **active sites** and the positions of **key residues**
- Prediction of protein-protein and protein-ligand **interactions**, which are mostly determined by steric (shape) and chemical (e.g. charge) **complementarity**
- Filling the **sequence / structure gap**. Million sequences are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

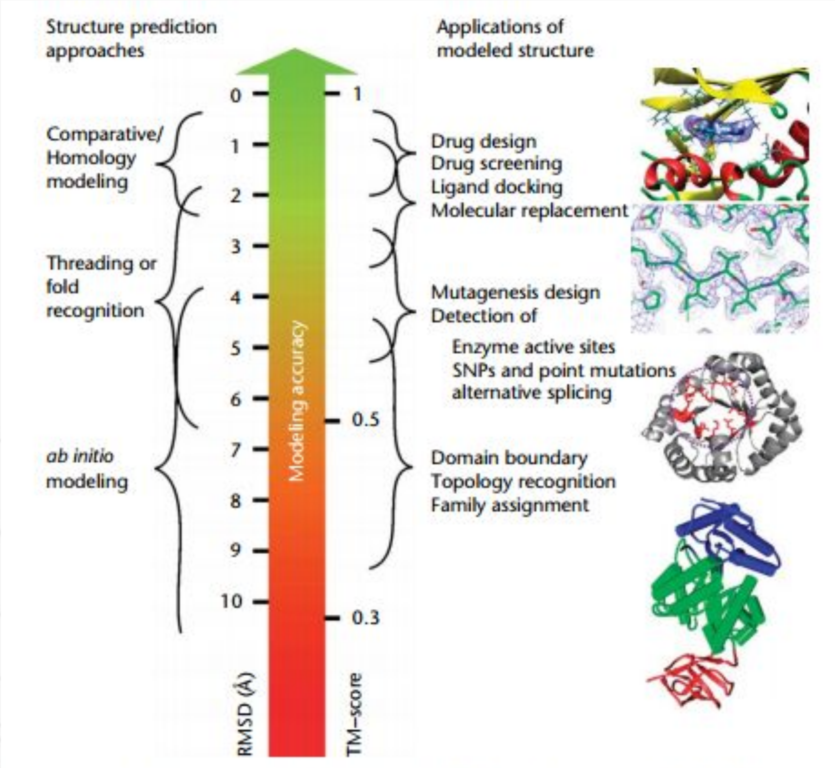
However

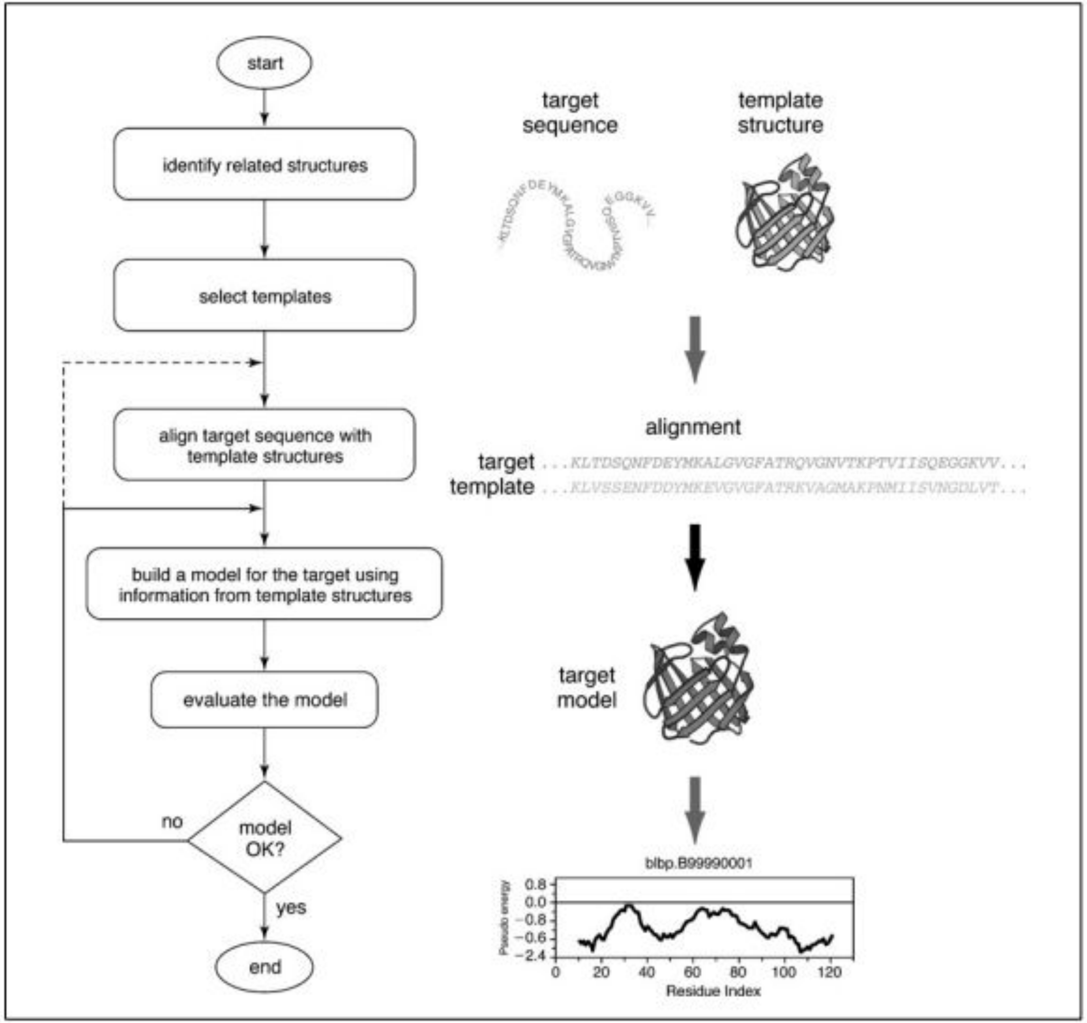
- Many proteins have the sequence not similar enough to build an *in silico* model by homology
- Many folds are not represented in the PDB

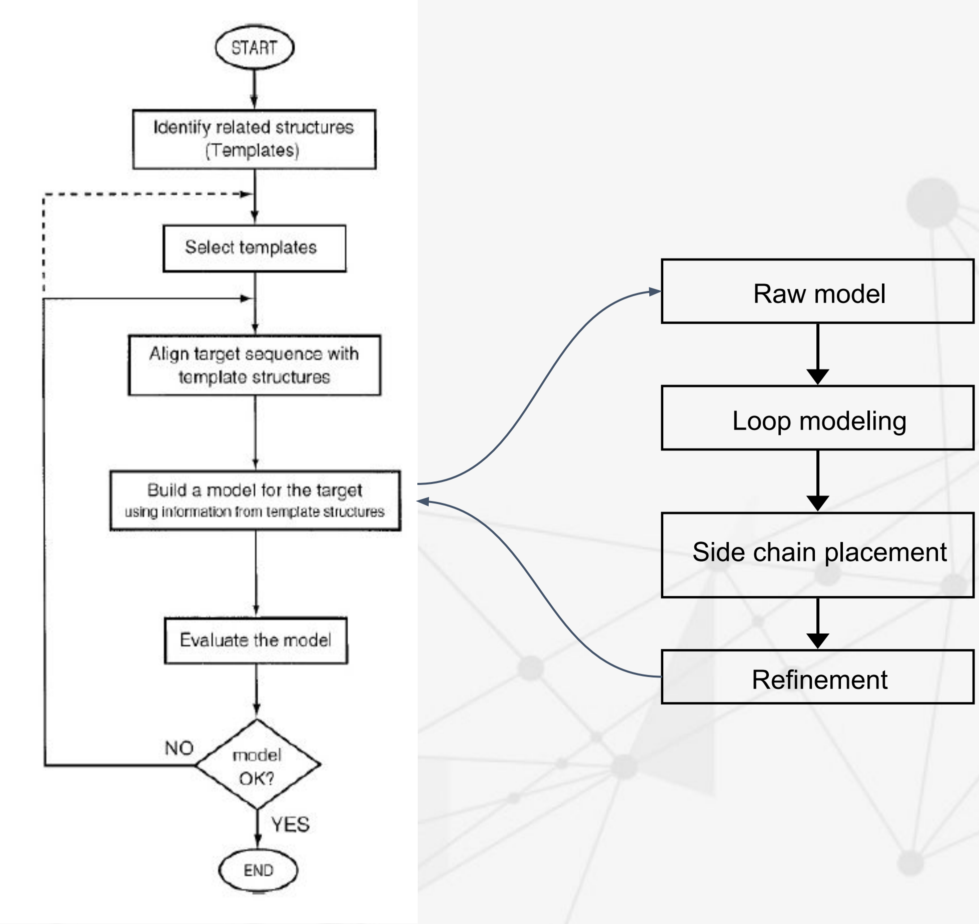


- **De novo prediction / Ab initio**
  - Secondary structure prediction; conformation of short fragments (Rosetta); molecular dynamics; Monte Carlo; quantum mechanics (unfeasible)
  - Tough computation
- **Fold recognition**
  - Try to fit with known folds
  - The fold space is not completely known (50%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


100
50
30
% Sequence identity
MODEL ACCURACY
1.0Å
100%
1.5Å
95%
3.5Å
80%
A)
B)
C)
APPLICATIONS
studying catalytic mechanism
designing and improving ligands
docking of macromolecules, prediction of protein partners
virtual screening and docking of small ligands
defining antibody epitopes
molecular replacement in X-ray crystallography
designing chimeras, stable, crystallizable variants
supporting site-directed mutagenesis
refining NMR structures
fitting into low-resolution electron density
structure from sparse experimental restraints
functional relationships from structural similarity
identifying patches of conserved surface residues
finding functional sites by 3D motif searching


Structure prediction approaches
Comparative/ Homology modeling
Threading or fold recognition
ab initio modeling
0
1
2
3
4
5
6
7
8
9
10
RMSD (Å)
Modeling accuracy
1
0.5
0.3
TM-score
Applications of modeled structure
Drug design
Drug screening
Ligand docking
Molecular replacement
Mutagenesis design
Detection of
Enzyme active sites
SNPs and point mutations
alternative splicing
Domain boundary
Topology recognition
Family assignment


start
identify related structures
select templates
align target sequence with template structures
build a model for the target using information from template structures
evaluate the model
no
model OK?
yes
end
target sequence
template structure
...KLTDSQNFDEYMKALGVGFATRQVGNVTKPTVIISQEGGKVV...
alignment
target ...KLTDSQNFDEYMKALGVGFATRQVGNVTKPTVIISQEGGKVV...
template ...KLVSSENFDDYMKEVGVGFATRKVAGMAKPNMIISVNGDLVT...
target model
blbp.B99990001
Pseudo energy
0.8
0.0
−0.8
−0.6
−2.4
0
20
40
60
80
100
120
Residue Index


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
NO
model OK?
YES
END

Raw model

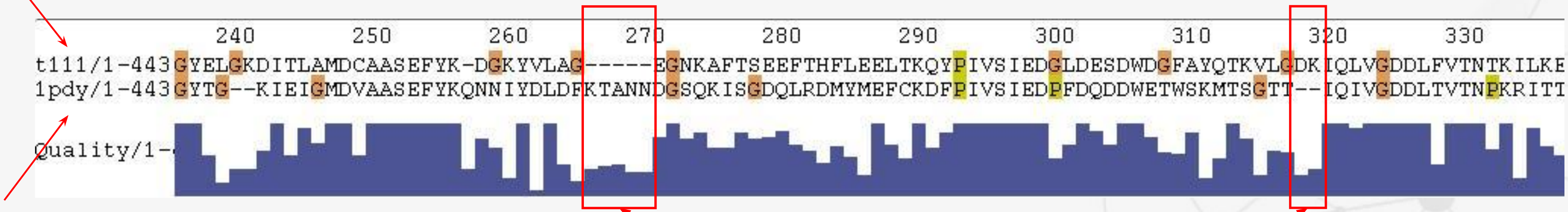
Loop modeling

Side chain placement

Refinement

# Alignment

target



```
              240       250       260       270       280       290       300       310       320       330
t111/1-443GYELGKDITLAMDCAASEFYK-DGKYVLAG-----EGNKAFTSEEFTHFLEELTKQYPIVSIEDGLDESDWDGFAYQTKVLGDKIQLVGDDLFVTNTKILKE
1pdy/1-443GYTG--KIEIGMDVAASEFYKQNNIYDLDFKTANNDGSQKISGDQLRDMYMEFCKDFPIVSIEDPFDQDDWETWSKMTSGTT--IQIVGDDLTVTNPKRITT

Quality/1-
```

template

Deletion

Insertion

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

- 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 you model this?

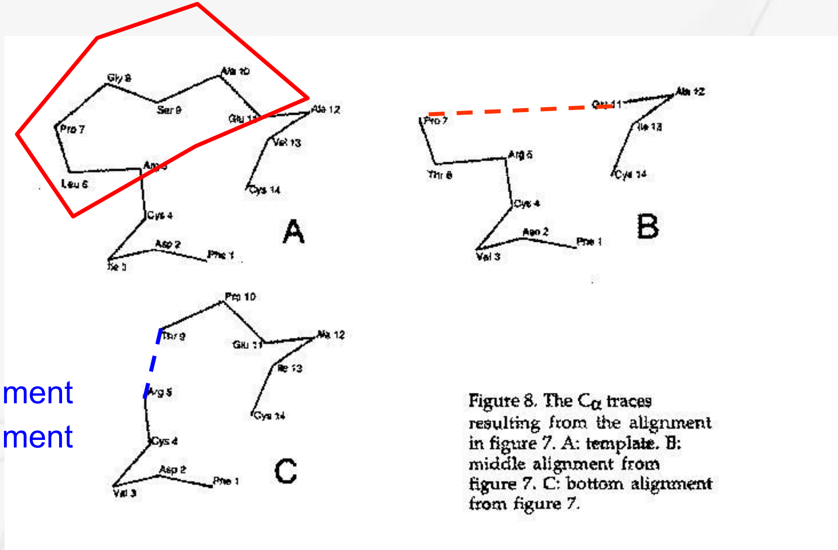
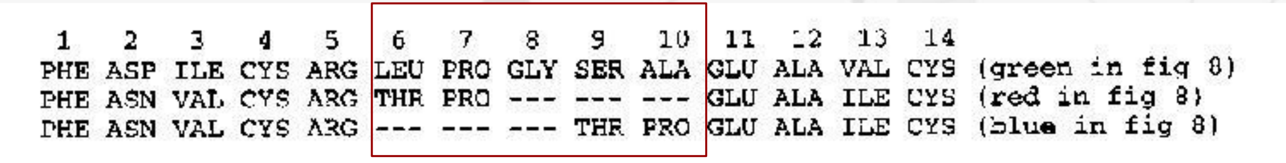


Figure 8. The Cα traces resulting from the alignment in figure 7. A: template. B: middle alignment from figure 7. C: bottom alignment from figure 7.

Worse sequence alignment
Better structure alignment

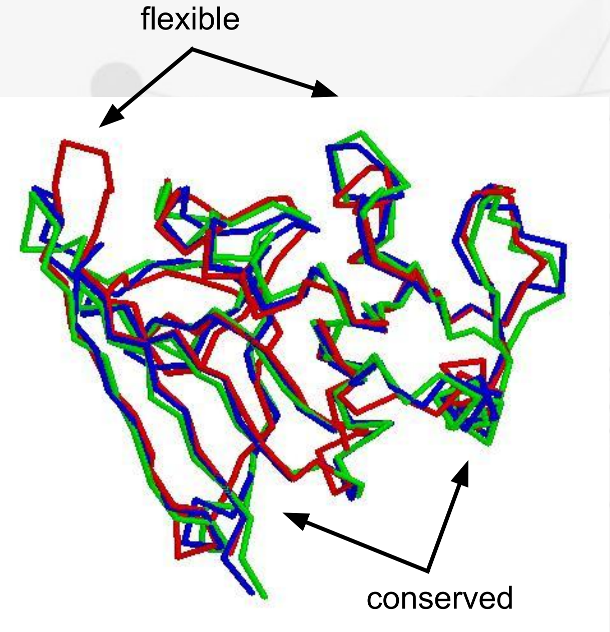


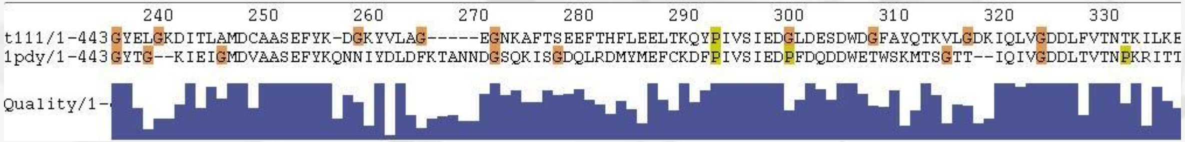
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | PHE | ASP | ILE | CYS | ARG | LEU | PRO | GLY | SER | ALA | GLU | ALA | VAL | CYS | (green in fig 8) |
| B | PHE | ASN | VAL | CYS | ARG | THR | PRO | --- | --- |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

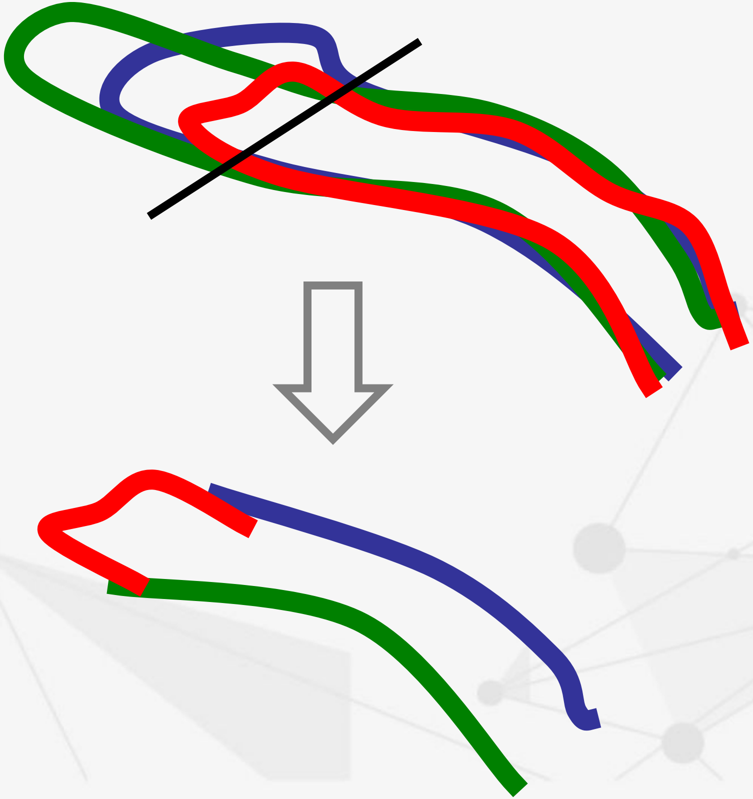




```
             240       250       260       270       280       290       300       310       320       330
t111/1-443GYELGKDITLAMDCAASEFYK-DGKYVLAG-----EGNKAFTSEEFTHFLEELTKQYPIVSIEDGLDESDWDGFAYQTKVLGDKIQLVGDDLFVTNTKILKE
1pdy/1-443GYTG--KIEIGMDVAASEFYKQNNIYDLDFKTANNDGSQKISGDQLRDMYMEFCKDFPIVSIEDPFDQDDWETWSKMTSGTT--IQIVGDDLTVTNPKRITI

Quality/1-
```

# Fragment-based building



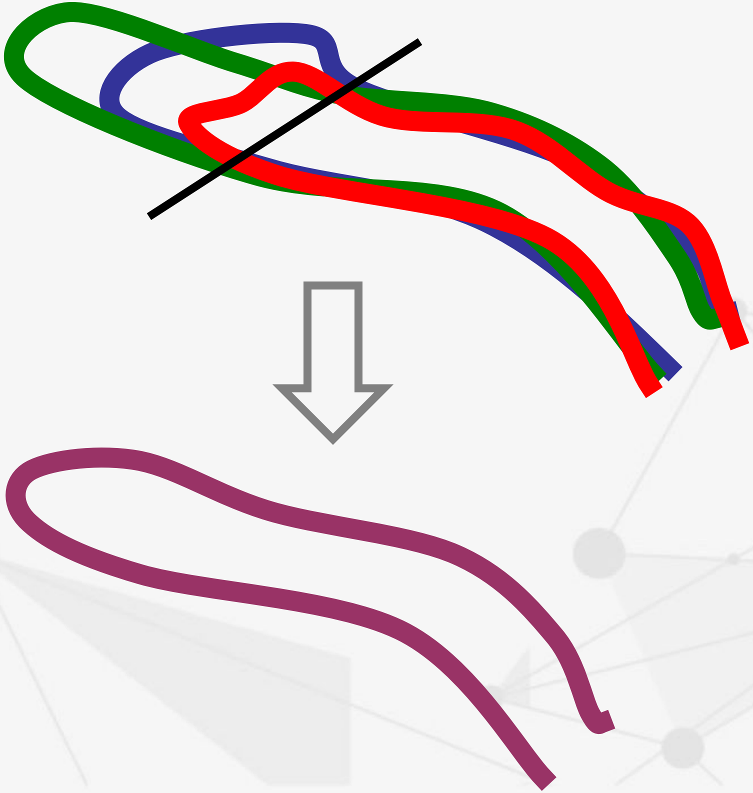
Idea → Copy “useful” coordinates of fragments

- 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

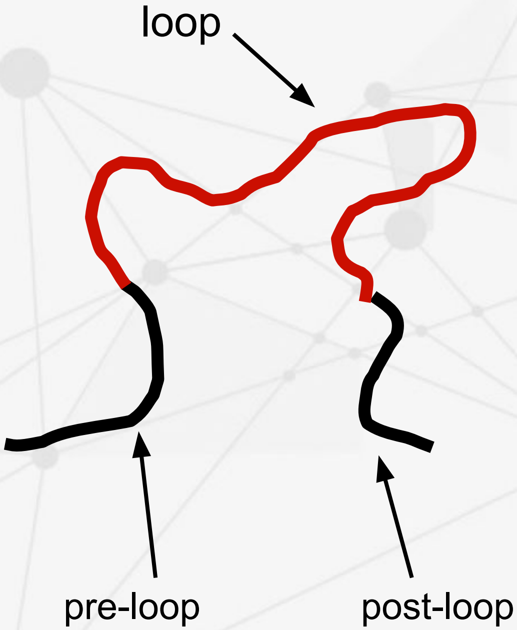
Software

- 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

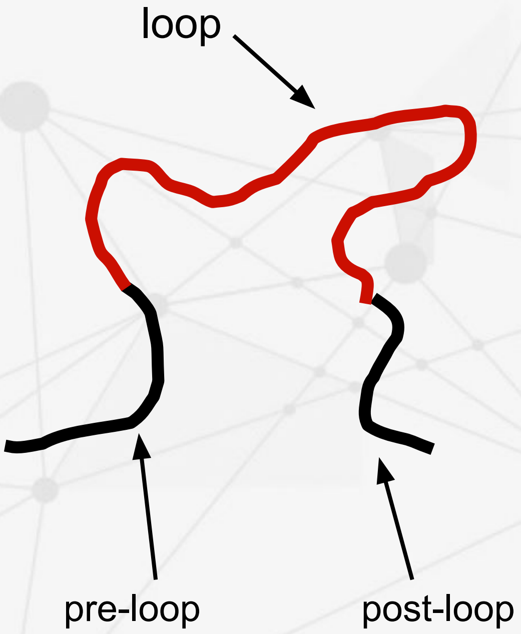
- Problem description
  - Identify the conformation of the fragment (loop, k residues) that can connect the pre- to the post-loop
  - Φ and ψ are the only free parameters



# Loop modeling

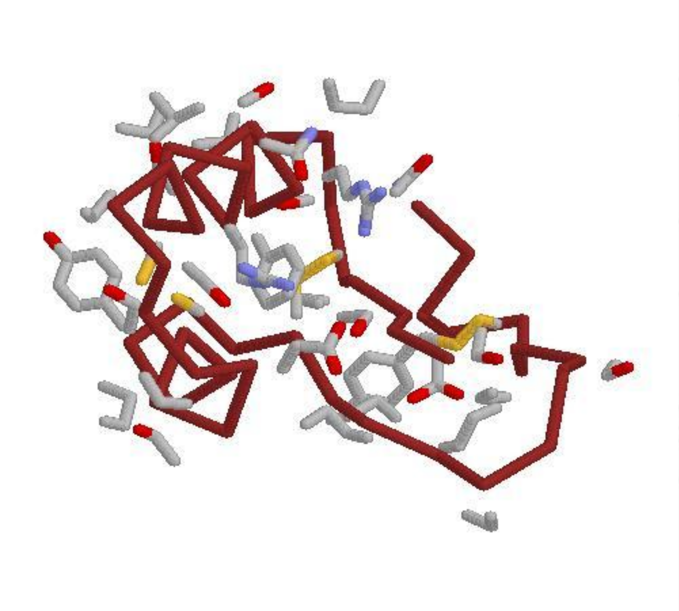
- Database methods
  - Extract loop fragments from PDB
  - Choose the fragment that fits better, based on geometric constraints
  - Not all possible conformations are available in PDB

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



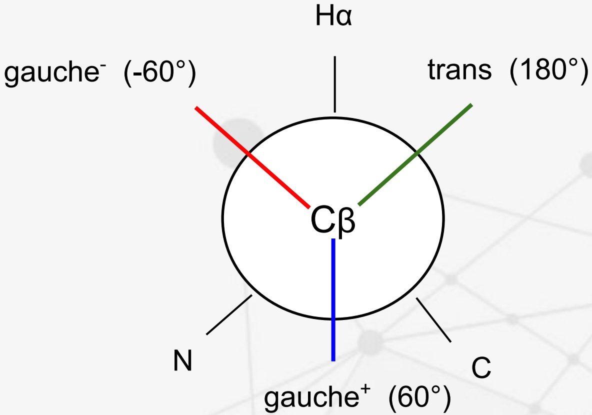
# Side chains

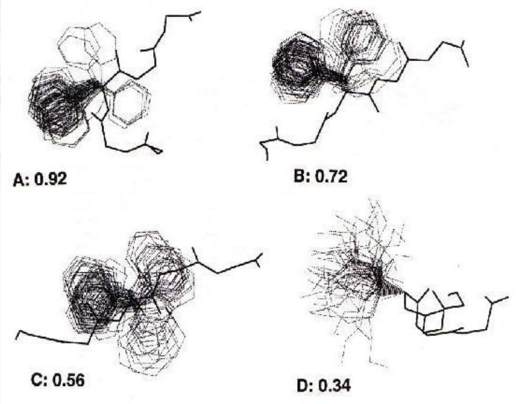
- **Amino acid differences** are not managed when applying the coordinates of the template to the sequence of the target (dimension and position of the side chains)
- Assuming **50% sequences identity**, half side-chains are replaced
- The **RMSD** change is relatively low, but the conformation of important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



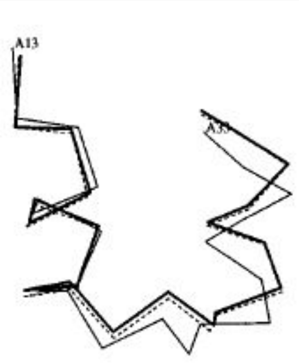
# Side chains

- Rotamers
  - 3 preferred positions for each torsion angles $\chi$
- 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

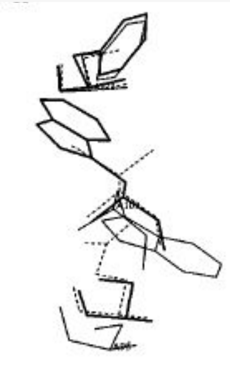




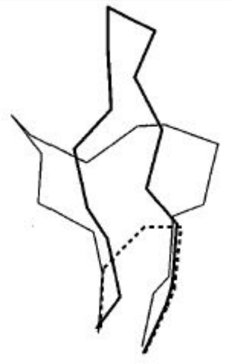
# Typical errors



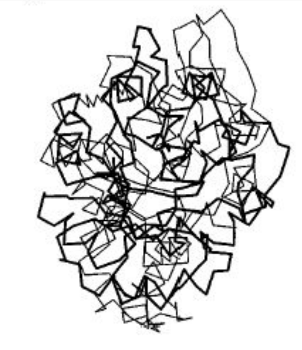
Shift



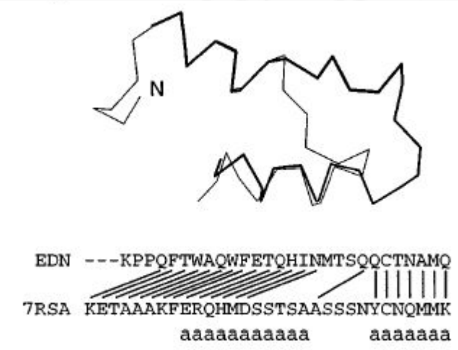
Side chains



Loops



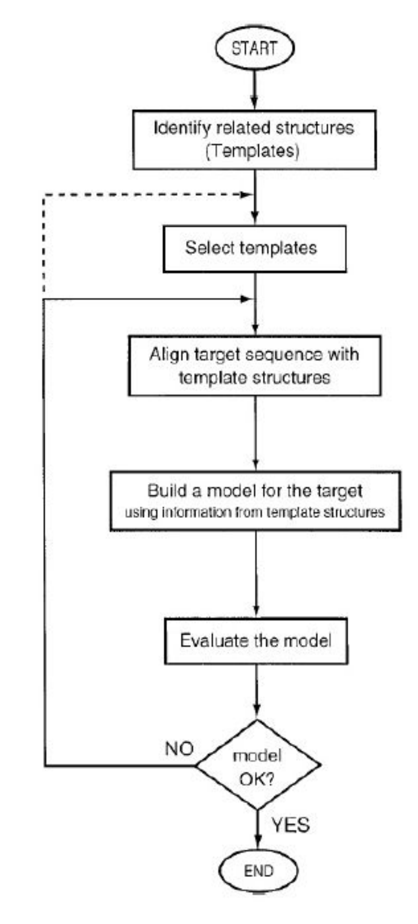
Wrong template



```
EDN  ---KPPQFTWAQWFETQHINMTSQQCTNAMQ
7RSA KETAAAKFERQHMDSSTSAASSSNYCNQMMK
             aaaaaaaaaaa     aaaaaaa
```

Wrong alignment

# Model assessment


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

- How to solve these “problems” without worsening the model?

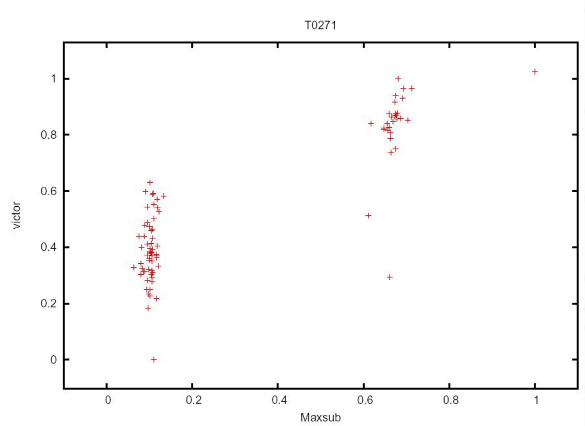
- Models from **alternative alignments** considering **energy profiles** can improve
  - It is necessary to know which part of the alignment to change

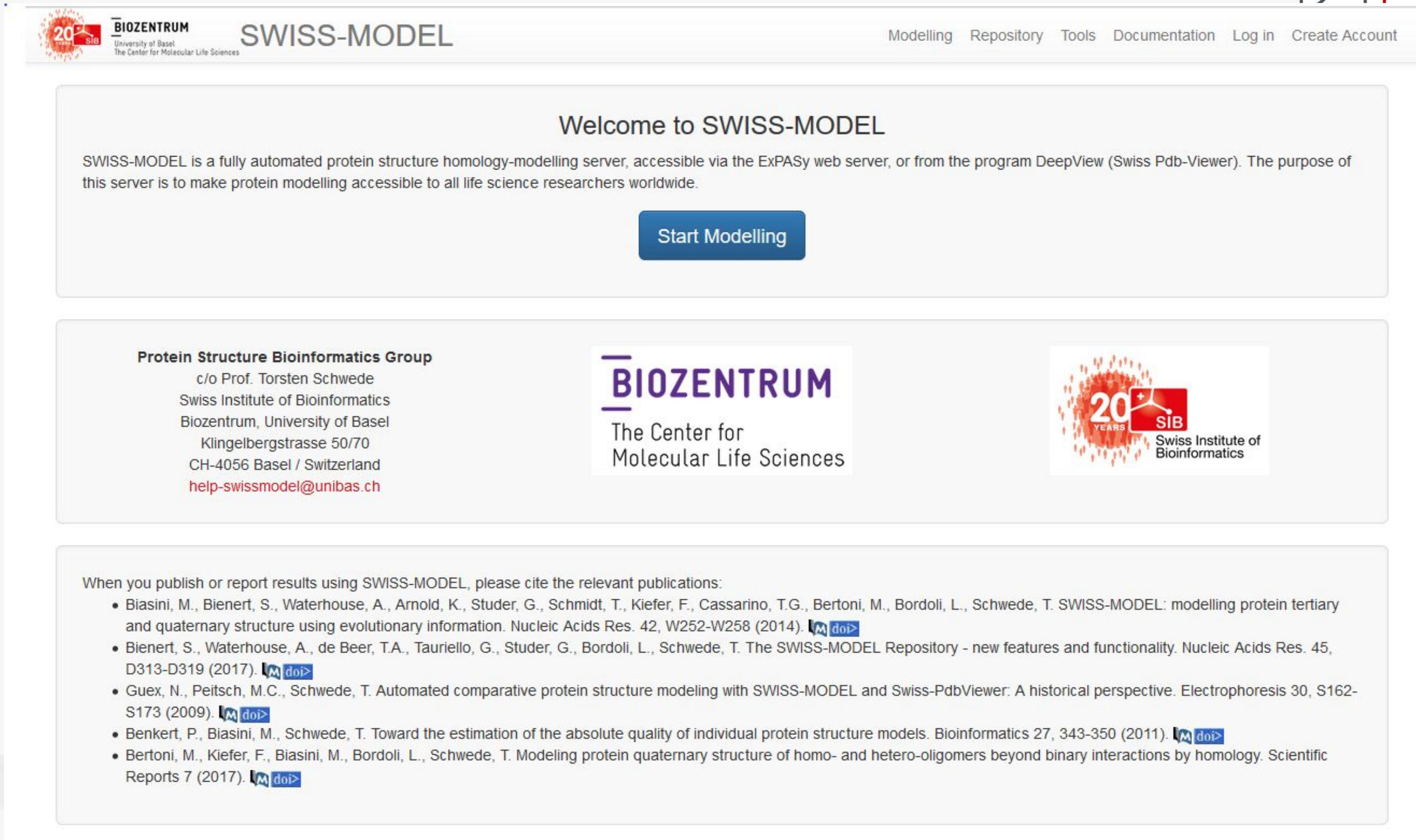
- **CASP** suggested the “*Don’t touch it*” philosophy for a long time
  - It is better to avoid local modifications of the structure
  - Changed over the last few years

# 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*





## 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.

**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

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>