Structural Bioinformatics
Lecture 7

Homology-based modelling of protein 3D structure

UNIVERSITÄT
DES
SAARLANDES

Common uses of comparative protein structure models

- Designing (site-directed) mutants to test hypotheses about function
- Identifying active and binding sites
- Searching for ligands of a given binding site
- Designing and improving ligands of a given binding site
- Modeling substrate specificity
- Predicting antigenic epitopes
- Protein-protein docking simulations
- Inferring function from calculated electrostatic potential around the protein
- Molecular replacement in X-ray structure refinement
- Refining models against NMR dipolar coupling data
- Testing a given sequence - structure alignment
- Rationalizing known experimental observations
- Planning new experiments

Modelling protein 3D structure

Protein sequence

ca. 180,000 structures of proteins and their complexes

Sequence similarity search

NCBI BLAST

HHblits

RCSB PDB PROTEIN DATA BANK

3D structures of proteins with a significant sequence similarity?

Yes → **Template-based modelling; homology modelling**

No → **Template-free modelling; ab initio/de novo modeling**

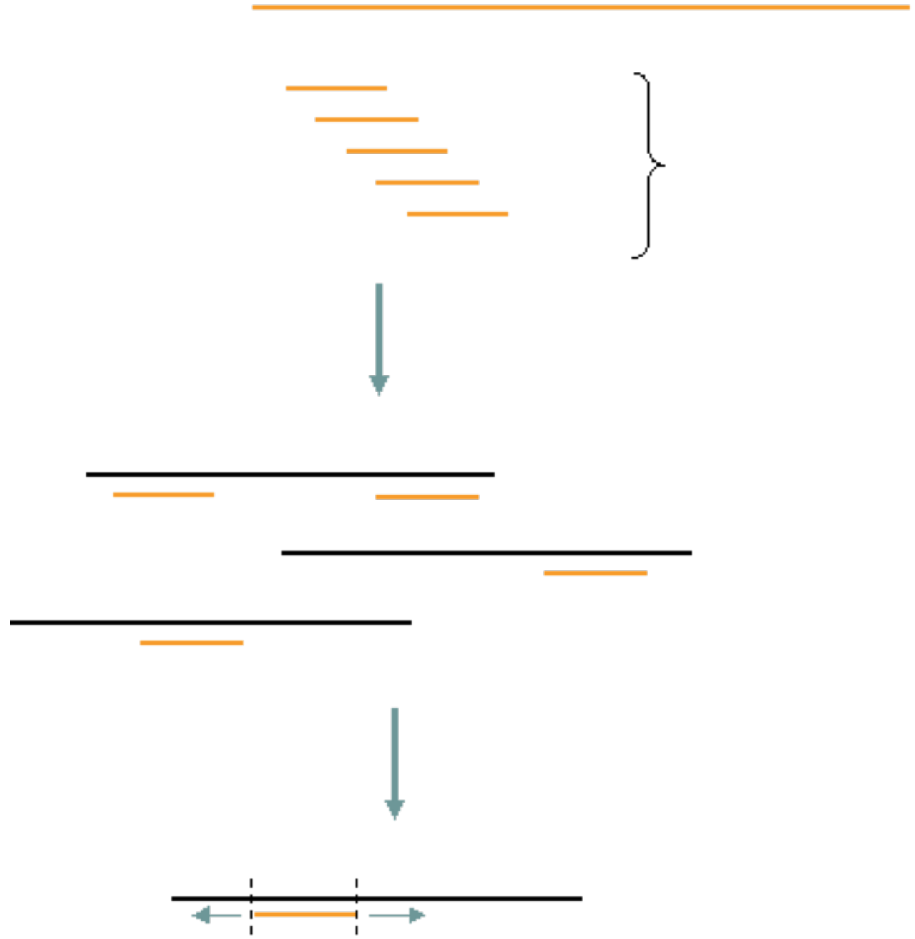
Template search

Sequence similarity search methods

BLAST → PSI-BLAST → Sequence vs. HMM → **HMM vs. HMM**

BLAST

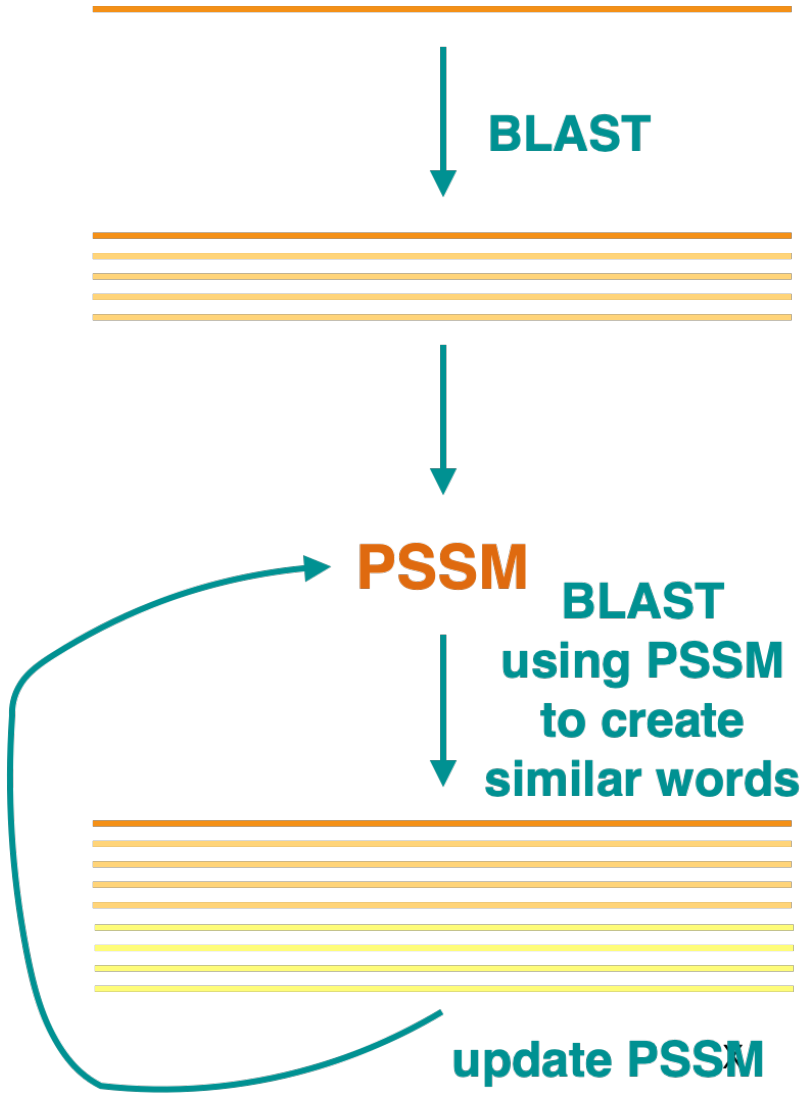
- **Query**: your sequence
- **Database**: all sequences to be searched
- Query is split into words
- Database is searched for words that are highly similar
  - Database records that contain more than one similar word at a small enough distance are selected
- Word alignment is extended until its score drops below a cutoff



PSI-BLAST

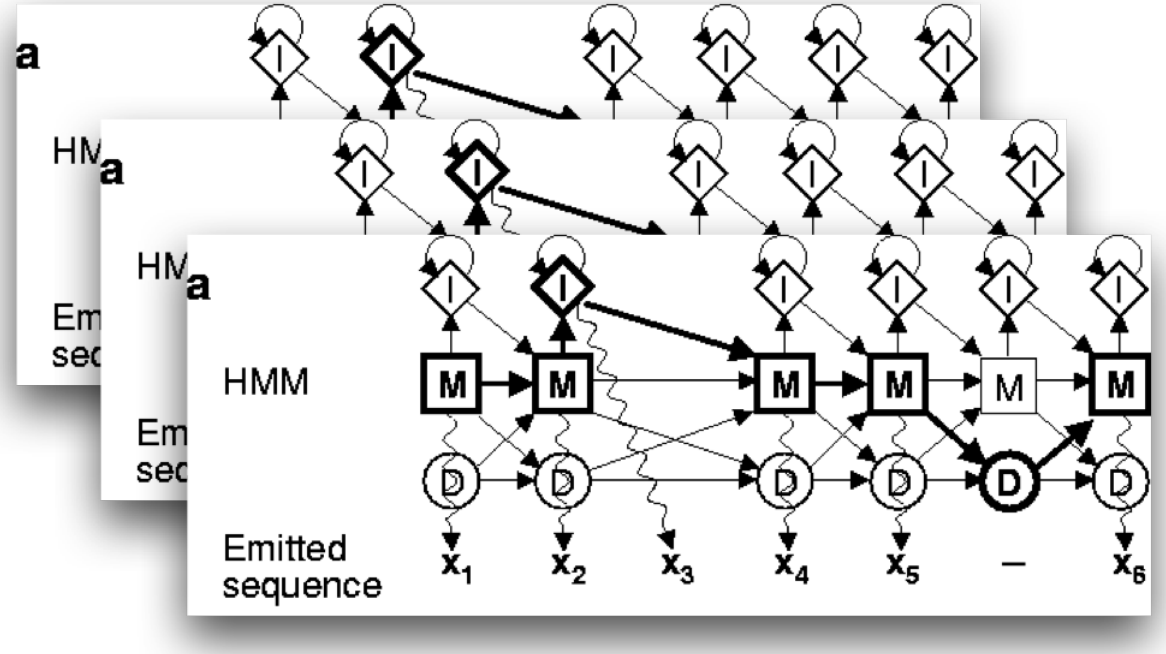
- **Query**: your sequence
- **Database**: all sequences to be searched

- Run BLAST
- Multiple alignment of hits => **position-specific scoring matrix (PSSM)**
  - sort of a alignment profile
- PSSM is used in place of a substitution matrix



HMM-based similarity search (e.g. HMMer)

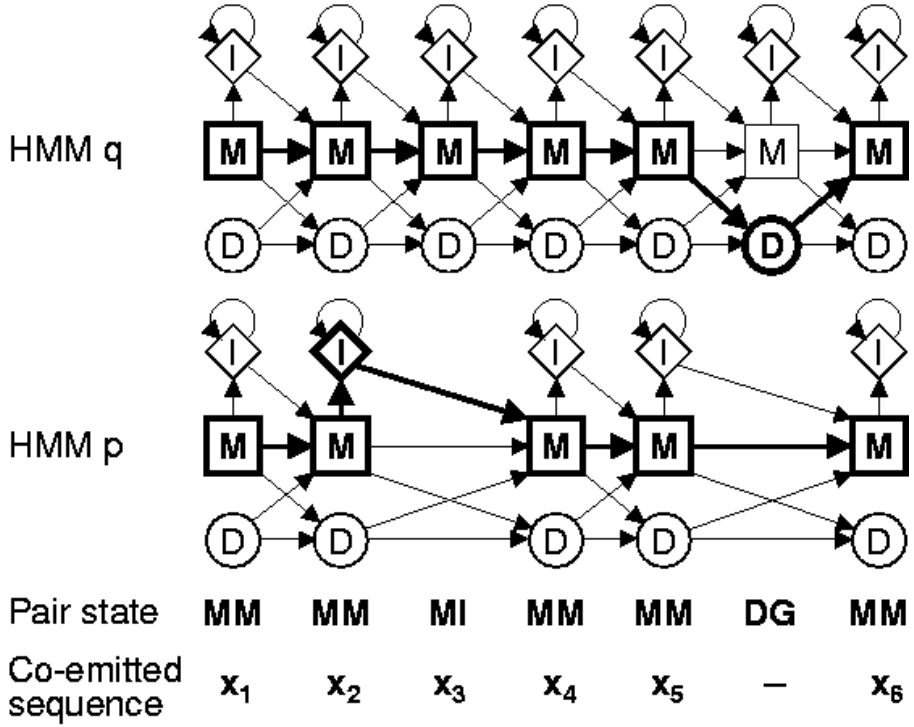
- **Query**: your sequence
- **Database**: set of HMMs



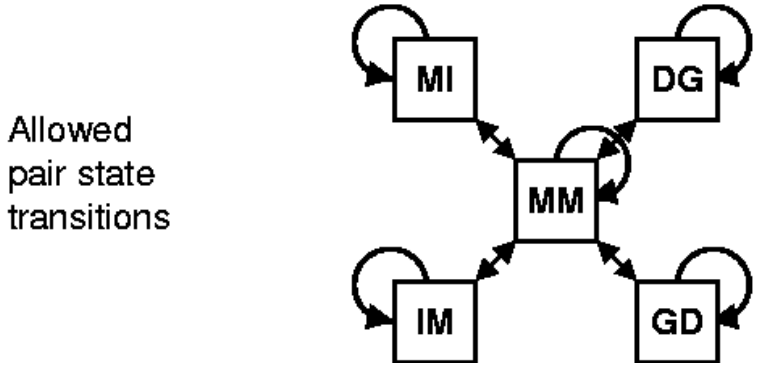
- Given a set of models $\{M_1, \ldots, M_K\}$ and a database of sequences $\{x_1, \ldots, x_N\}$, group them to models
- Compare to the score of the model of a random sequence $P(x|R) = \prod_i q_{x_i}$
- Slow, because all HMMs have to be tested, each $O(n^2)$

HMM-HMM profile comparison (HHblits)

- **Query**: your sequence or sequence alignment
  - sequence turned into an HMM using PSI-BLAST
- **Database**: set of HMMs

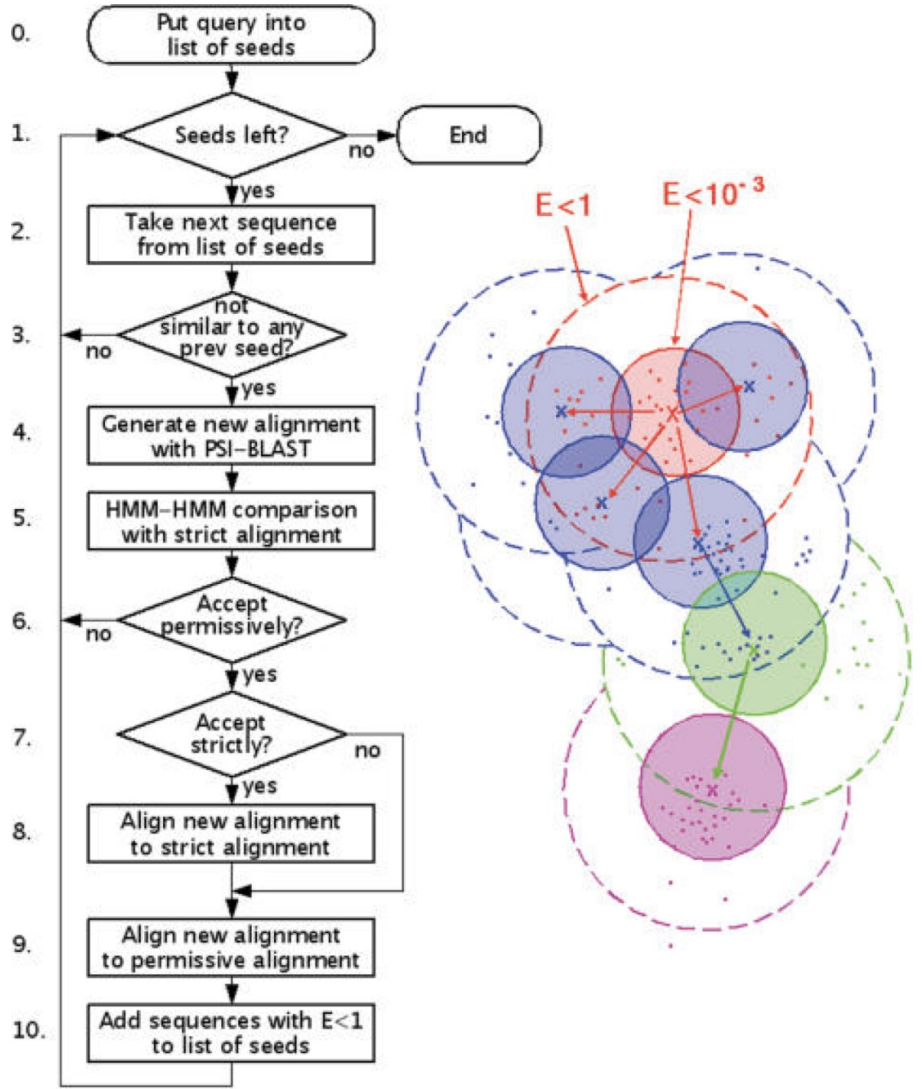


- Create an HMM for the query sequence (PSI-BLAST)
- Compare to each HMM in the database
- Also slow



Even more sensitive HMM-HMM profile search: Hhsenser (Söding, 2006)

- Very slow! — each pair of HMMs has to be compared to each other
- Strict and permissive alignments: different cutoffs

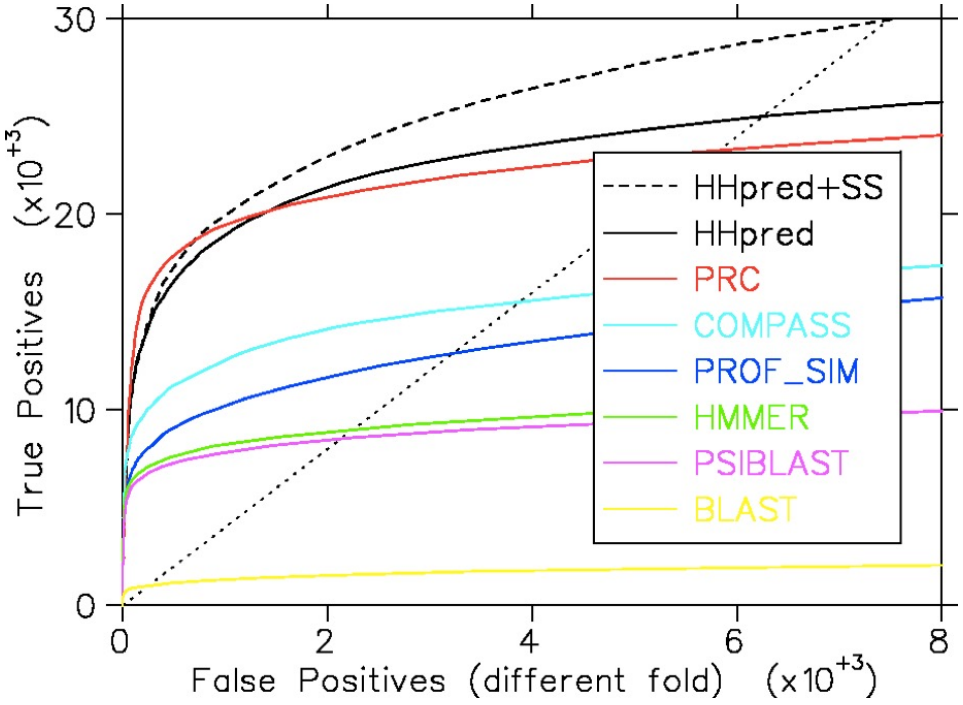


HHpred: implements HMM-HMM profile search against PDB

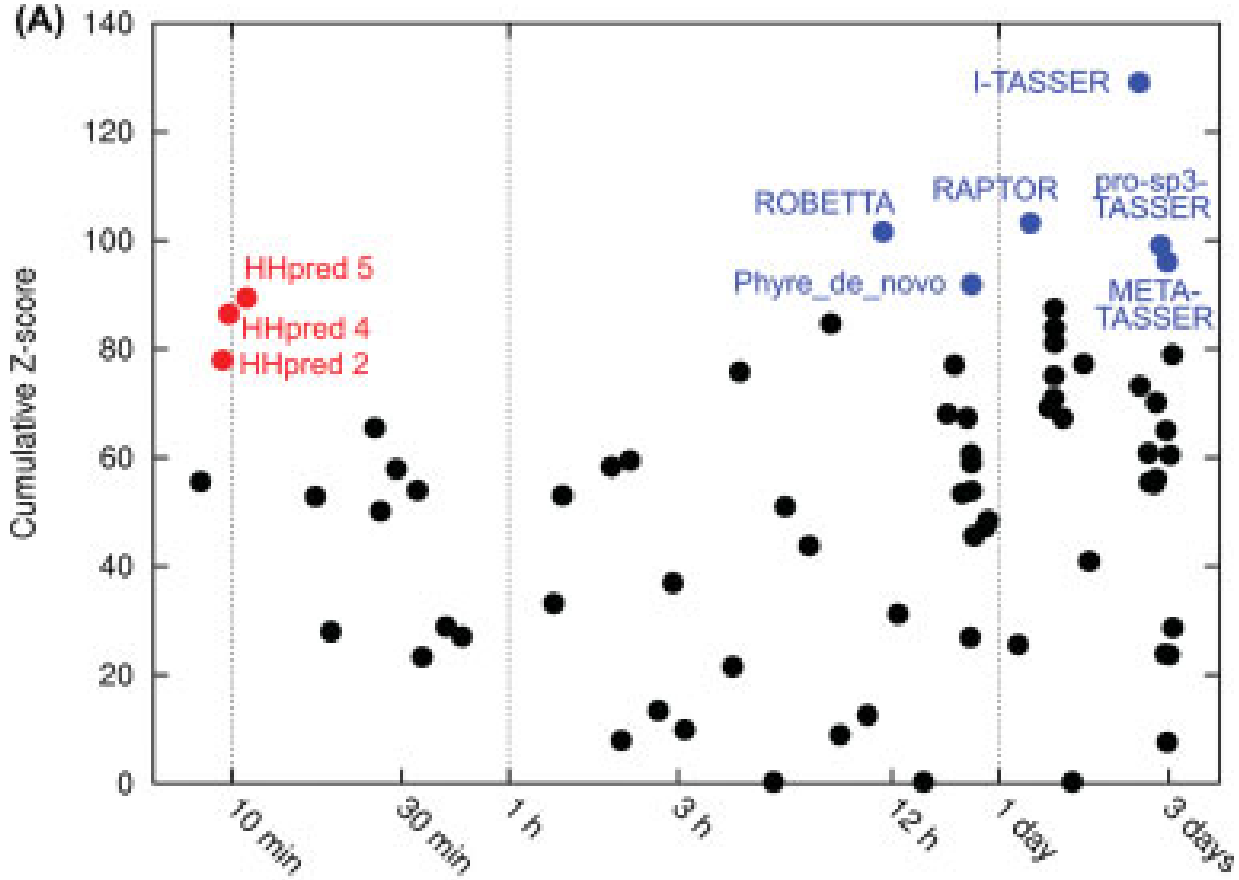
- Implements HMM-HMM profile sequence search
- Can use single sequence or an alignment as input
- Comes with a pre-calculated collection of HMMs corresponding to every 3D structure in PDB
- Is available for download or online as a part of Tübingen Bioinformatics Toolkit (https://toolkit.tuebingen.mpg.de/)

HHpred: workflow

- Search for homologs with PSI-BLAST
- HMM construction
- HMM-HMM comparison (HHsearch)
- (Optional: secondary structure comparison)
- Run modelling tool (MODELLER)



HHpred: results

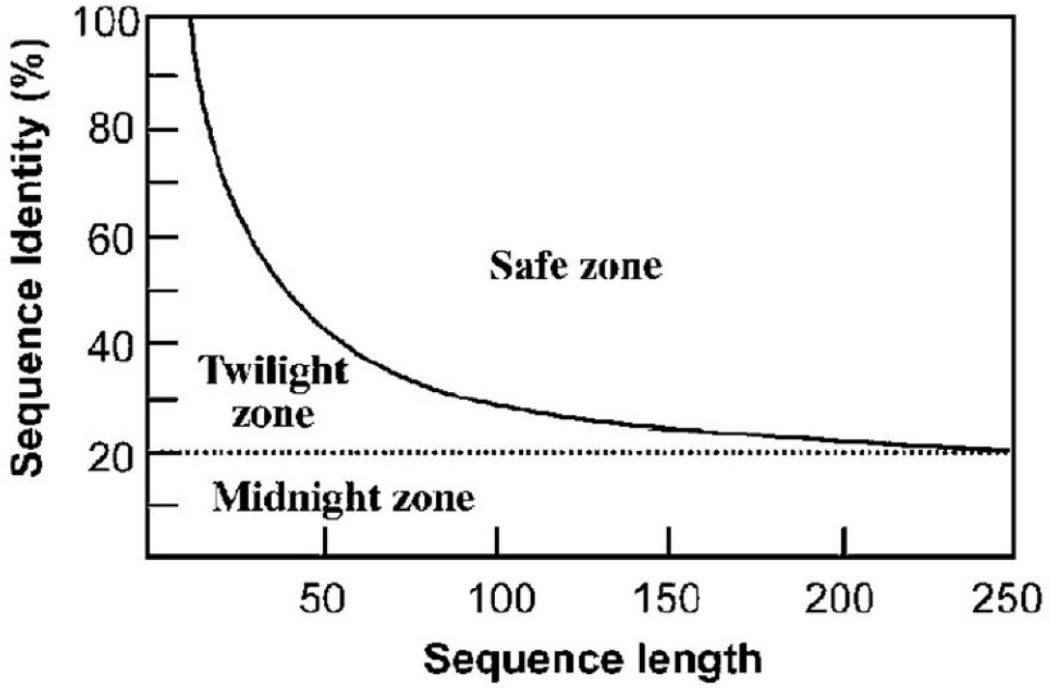


Hildebrandt et al., 2009

- In a blind experiment: predict a newly-resolved 3D structure of a protein without knowing it (CASP)

Template-based modelling a.k.a. homology modelling

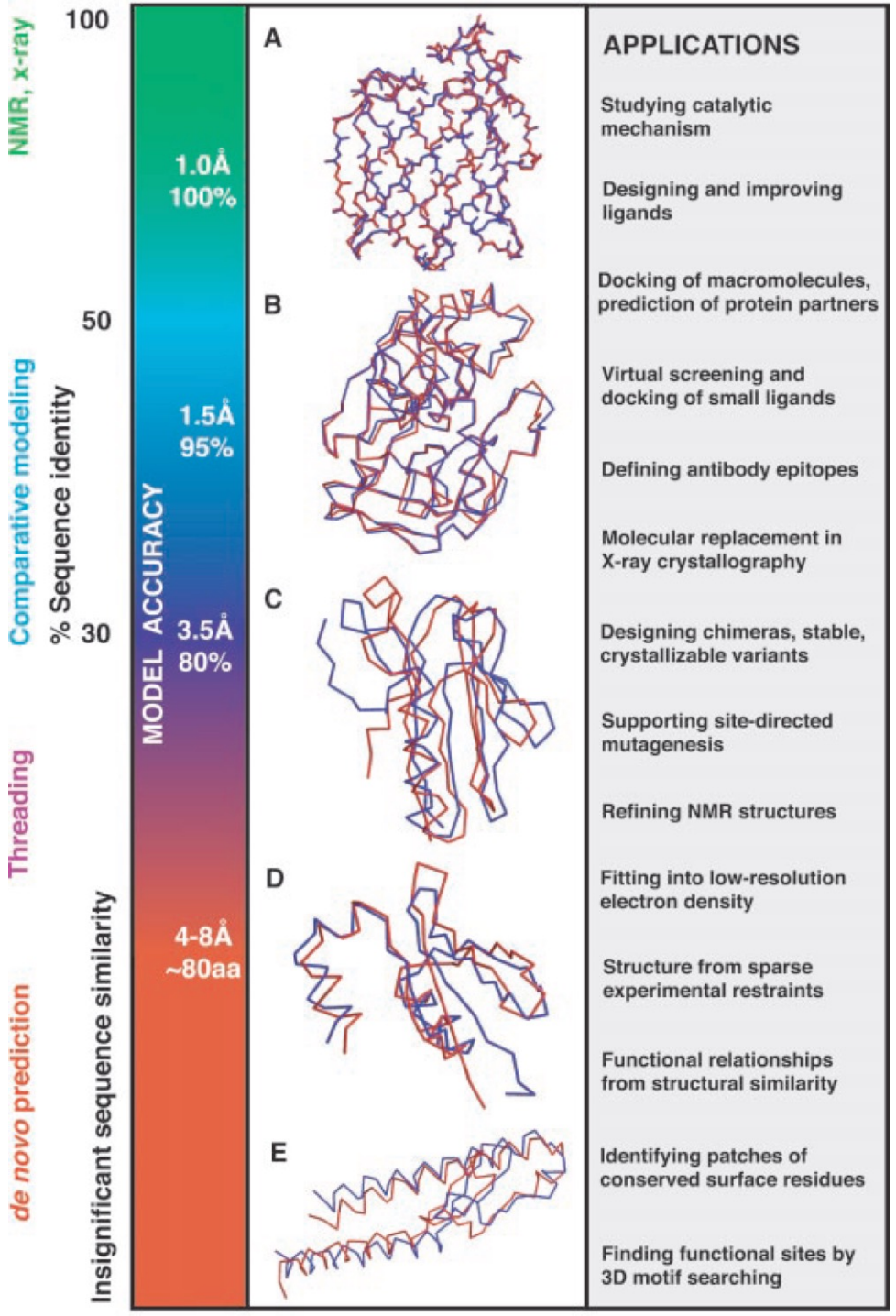
What can be modelled?



- >75%: easy to build a good model automatically
- 50%-75%: alignment correction might be needed
- 25%-50%: alignment has to be constructed very carefully

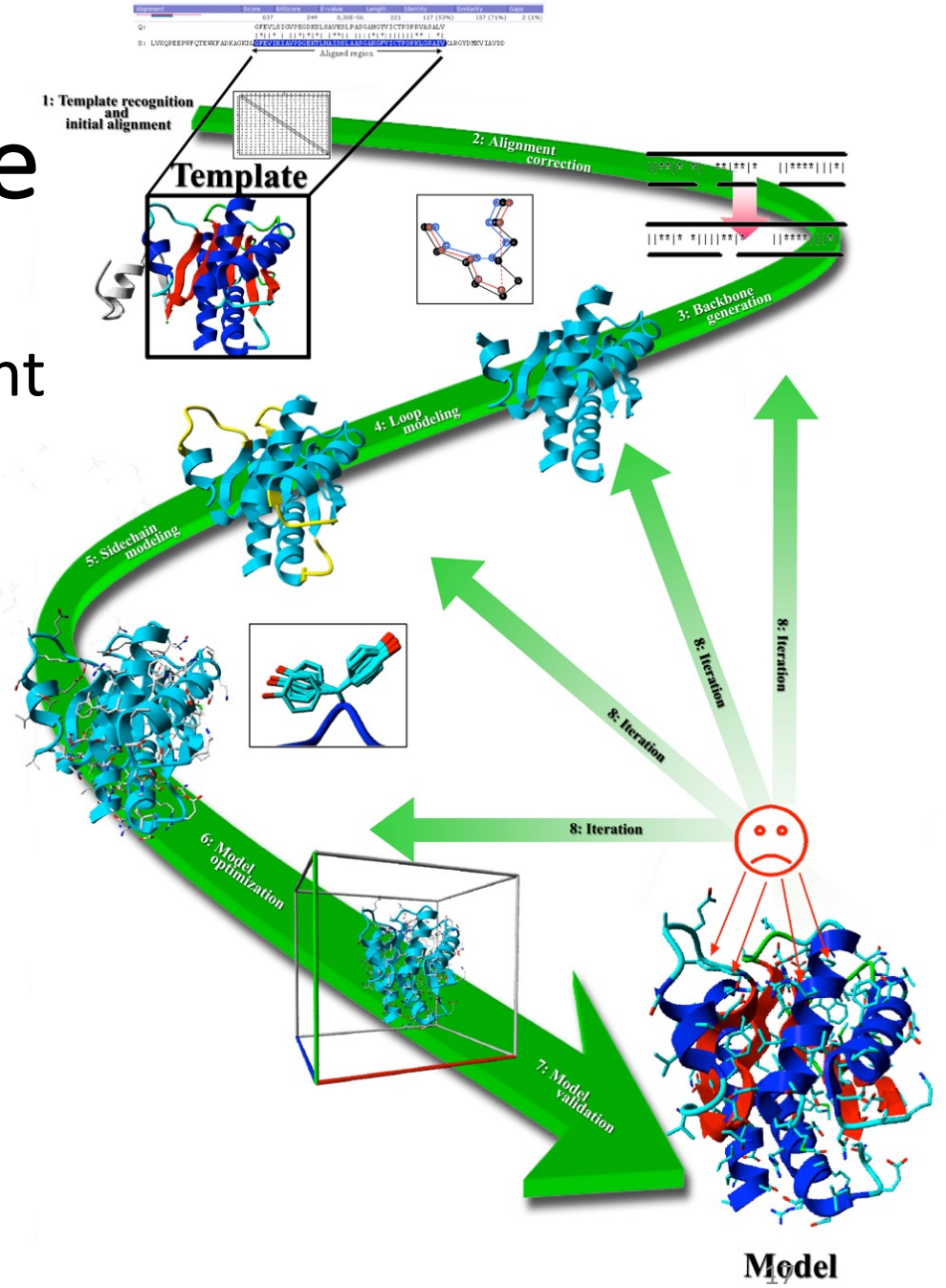
What can be modelled?

- >75%: easy to build a good model automatically
- 50%-75%: alignment correction might be needed
- 25%-50%: alignment has to be constructed very carefully



Homology modelling pipeline

1. Template recognition and initial alignment
2. Alignment correction
3. Backbone generation
4. Loop modeling
5. Side chain modelling
6. Model optimization
7. Model validation
8. Iteration



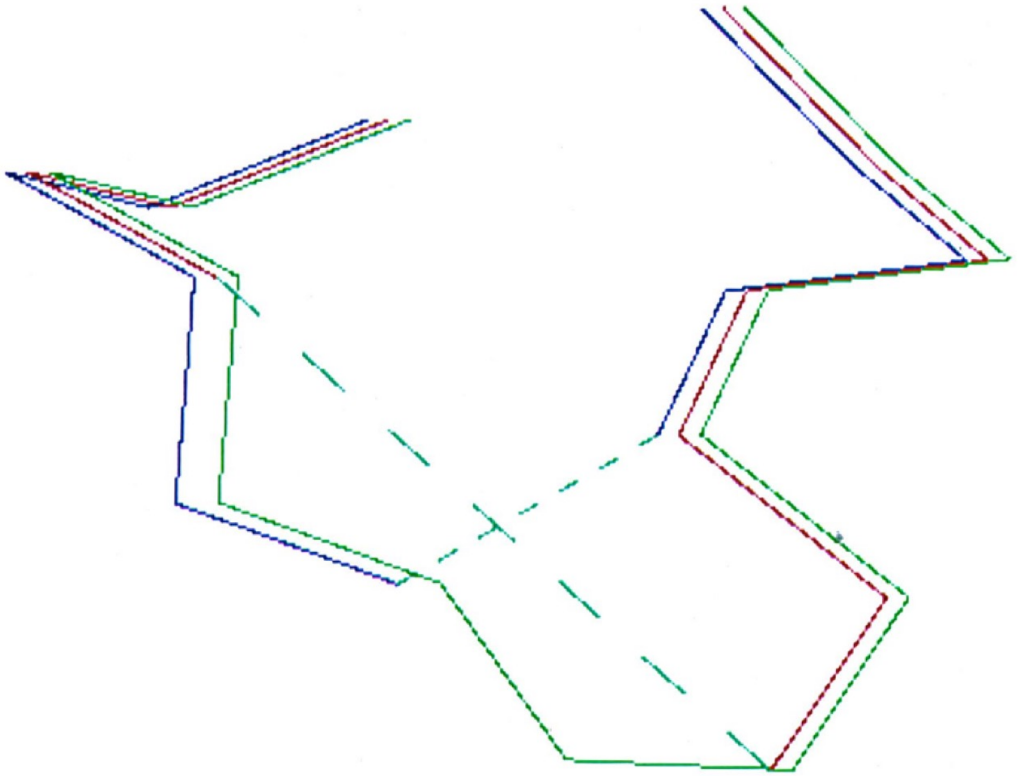
1. Template recognition and alignment

- Sequence similarity search against PDB
  - BLAST, HMM
- Not always the hit with the lowest E-value is the best option:
  - Conformational state (active/inactive)
  - Bound ligands (substrate/product, cofactors, inhibitors)
  - Multimeric state
- BLAST is not good for constructing alignments: use a global alignment tool with relaxed terminal penalty

2. Alignment correction

- Why?

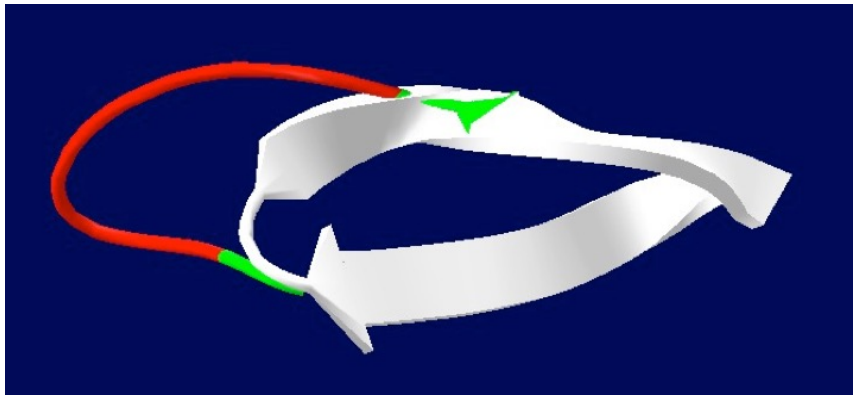
- Multiple alignments
- Gap correction



3. Backbone generation

- Coordinates of the backbone atoms (N, Cα, C, O) are copied from the template to the target
- Often coordinates of Cβ can be also copied
- Errors in PDB files are possible =>
  - re-refined structures
  - RECOORD project (500+ NMR structures)
  - PDBREPORT database (WHAT_CHECK / WHAT_IF program suite, *temporarily<?> unavailable*)
  - multiple templates

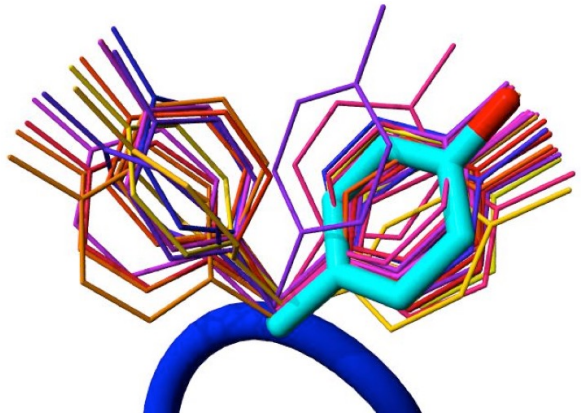
4. Loop closure



- Deletions and insertions in the target sequence
- Deletions create holes, insertions create atoms that cannot be accounted for using the template => loops have to be modelled de novo
- Insertions and deletions rarely occur within regular secondary structure elements
- Even without insertions/deletions, loops tend to assume more variable conformations => in might make sense to re-model all loops
  - Knowledge-based: endpoints and length of the loop in the target match a loop in PDB
  - Hybrid: loop is divided into smaller segments, then as above
  - Energy-based ab initio prediction of the fold of the loop

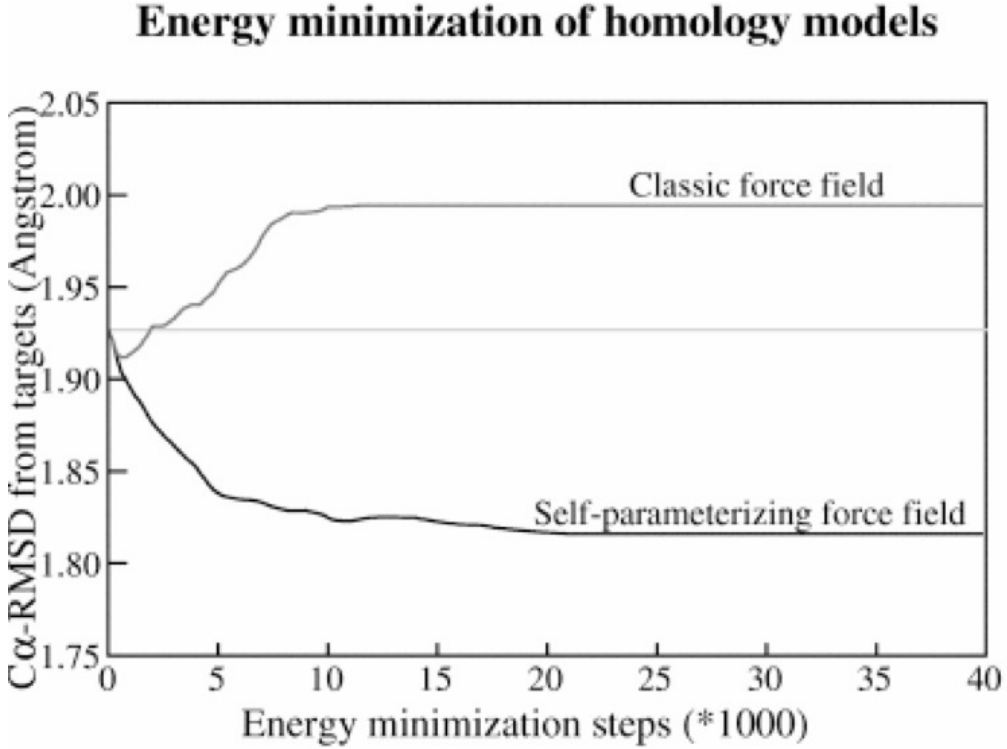
5. Side chain modelling

- Conserved residues often assume same conformation if the template-target identity is high enough (>35%) => network of contacts
- **Rotamers**: side-chain conformations
- Selected from a **rotamer library**: collection of all side-chain conformations observed in PDB, together with their frequencies
- **Backbone-dependent** rotamer libraries: bins for ϕ/ψ angles, or short stretches
- Side chain positions in the core of the protein are much more stable than on the surface:
  - more flexible in experimentally resolved structures
  - have higher energy
  - may have a functional reason to move around



6. Model optimization

- Position of side chains and the backbone are mutually dependent
- Optimization through energy minimization of all atoms (side chain rotamers can be remodelled between the steps)
- **Classical force fields** promote accumulation of small errors
- **Quantum force fields**: still slow and approximate
- **Self-parametrizing force fields**: randomly change a parameter, see if model gets better. Also slow
- **Molecular dynamics simulations**

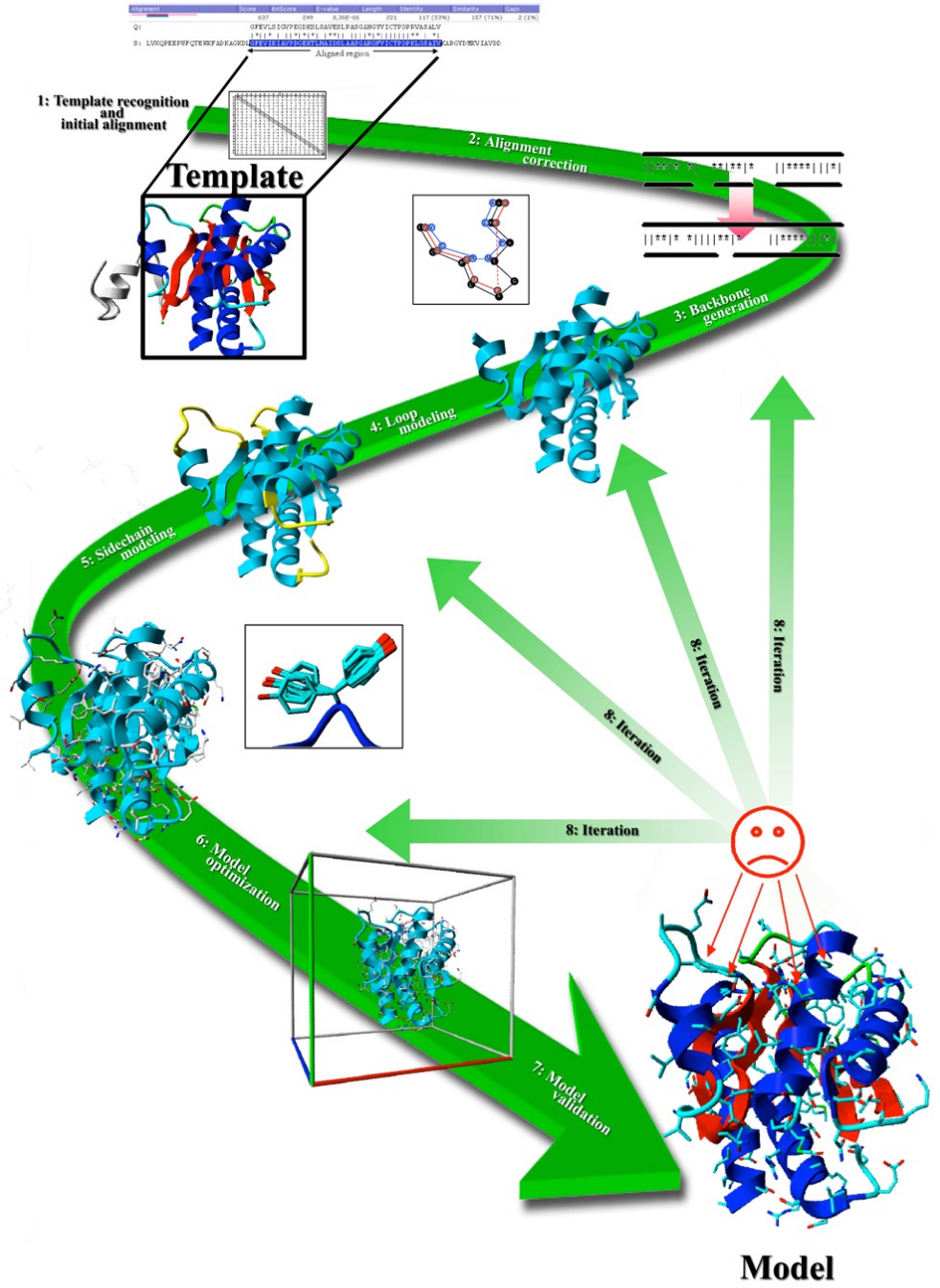


7. Model validation

- Model quality depends on
  - % identity between the template and the target sequences
  - # error in the template structure
- **Normality checks**:
  - bond **lengths**, **angles**, etc.
    - dihedral angles: allowed areas of the Ramachandran plot
  - **inside/outside** distribution of **polar and apolar** residues can identify completely misfolded proteins
  - Radial distributions of certain atom types are characteristic and can be extracted from experimentally resolved structures
  - Atomic contacts

8. Iteration

To any of the steps, depending where the problem is discovered



MODELLER (Andrej Šali lab, UCSF)

J. Mol. Biol. (1993) **234**, 779–815

Comparative Protein Modelling by Satisfaction of Spatial Restraints

Andrej Šali† and Tom L. Blundell

ICRF Unit of Structural Molecular Biology
Department of Crystallography
Birkbeck College, London WC1E 7HX, England

Protein Science (2000), *9*:1753–1773. Cambridge University Press. Printed in the USA.

Modeling of loops in protein structures

ANDRÁS FISER, RICHARD KINH GIAN DO, AND ANDREJ ŠALI

Laboratories of Molecular Biophysics, Pels Family Center for Biochemistry and Structural Biology, The Rockefeller University, 1230 York Ave., New York, New York 10021

(REC

Protein Science (1994), *3*:1582–1596. Cambridge University Press. Printed in the USA.

Derivation of rules for comparative protein modeling from a database of protein structure alignments

ANDREJ ŠALI[1] AND JOHN P. OVERINGTON[2]

[1] Department of Chemistry, Harvard University, Cambridge, Massachusetts 02138
[2] Pfizer Central Research, Ramsgate Road, Sandwich, Kent CT13 9NJ, United Kingdom

(RECEIVED March 8, 1994; ACCEPTED May 16, 1994)

Statistical potential for assessment and prediction of protein structures

MIN-YI SHEN AND ANDREJ SALI

Department of Biopharmaceutical Sciences, Department of Pharmaceutical Chemistry, and California Institute for Quantitative Biomedical Research, University of California at San Francisco, San Francisco, California 94158, USA

(RECEIVED June 28, 2006; FINAL REVISION August 18, 2006; ACCEPTED August 18, 2006)

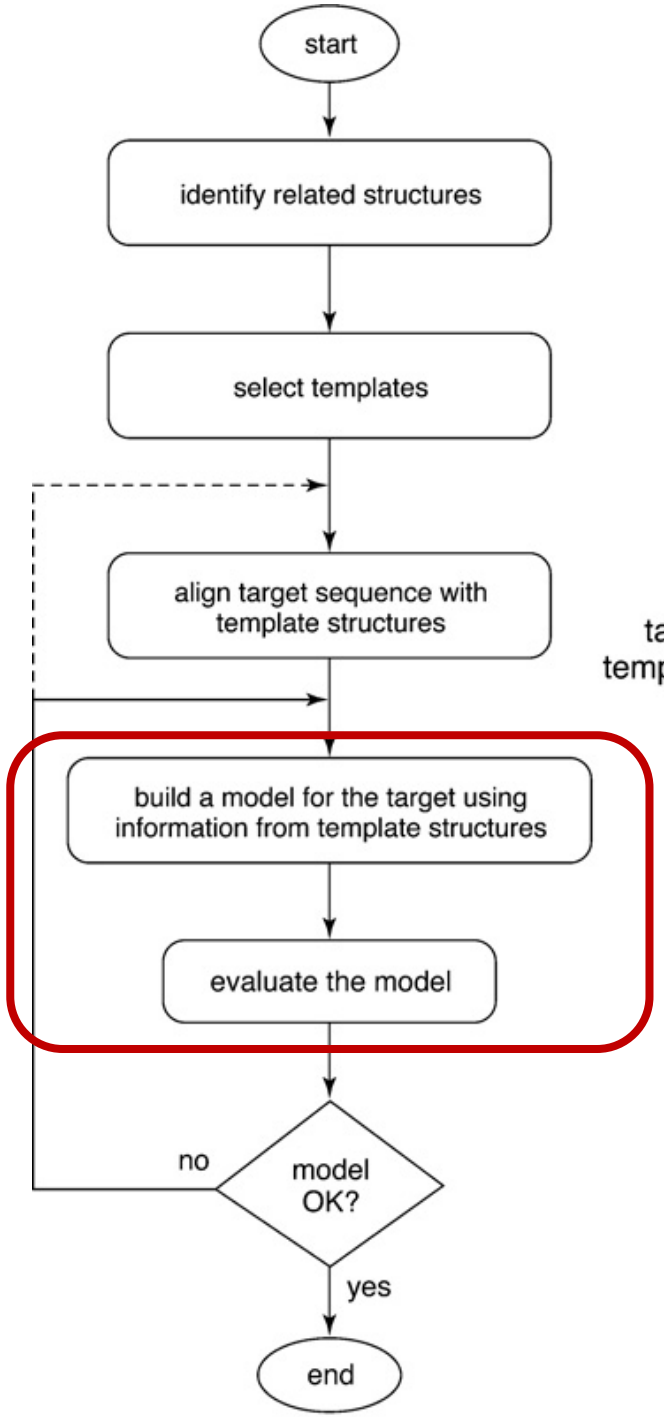
Modeling Structure from Sequence

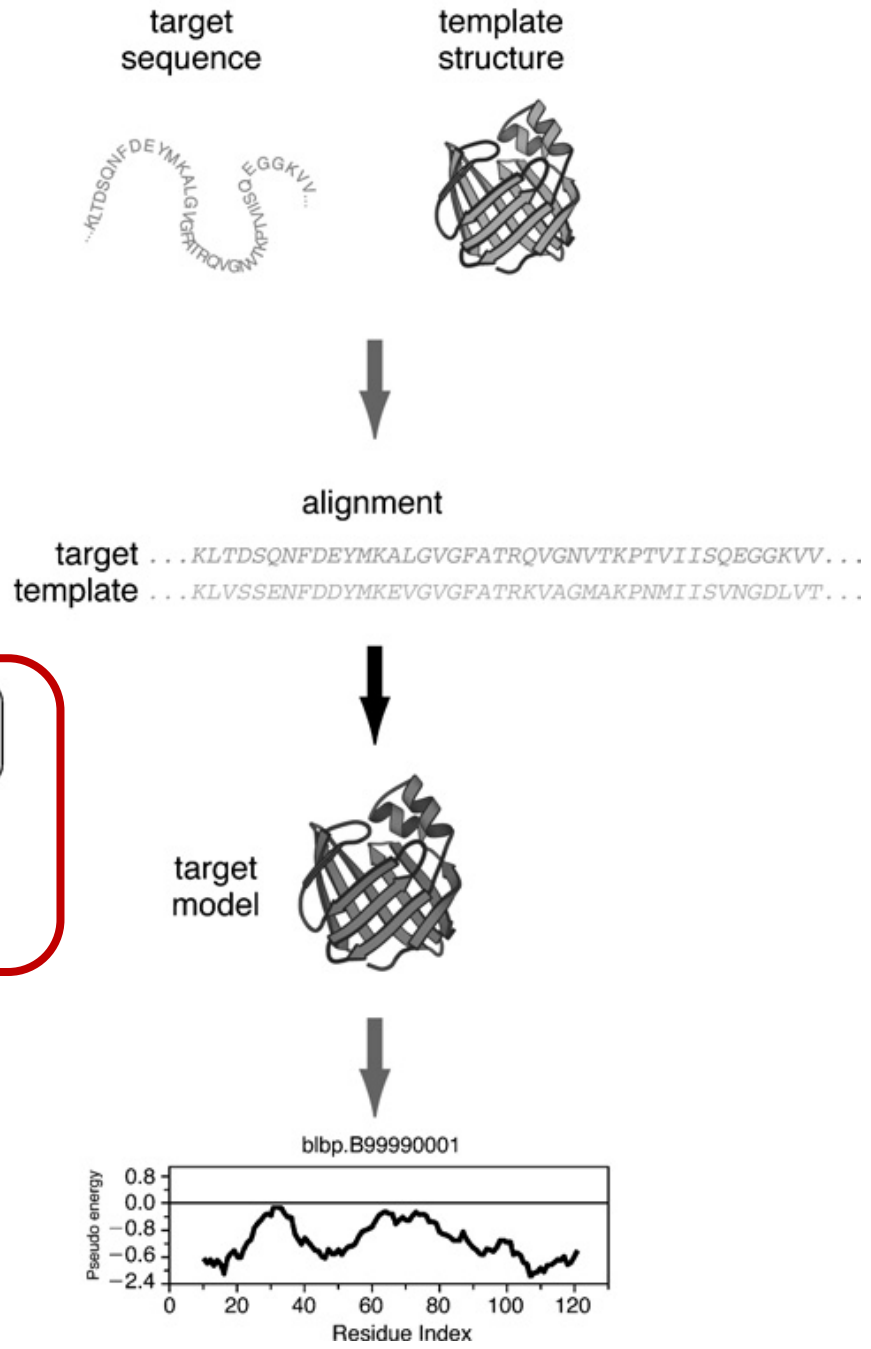
5.6.1

Supplement 15

Contributed by Narayanan Eswar, Ben Webb, Marc A. Marti-Renom, M.S. Madhusudhan, David Eramian, Min-yi Shen, Ursula Pieper, and Andrej Sali

Current Protocols in Bioinformatics (2006) 5.6.1-5.6.30


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
model OK?
no
yes
end


target sequence
template structure
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

General idea

- Extract spatial restraints from the known 3D structures
  - template and others
- Build a model for the 3D structure of the query sequence so that it satisfies these restraints
- **Probabilistic model of restraints**:
  - given a set of structures, how probable are the observed values of restraints — **optimization problem**

What are spatial restraints?

- amino acid residues type
- main chain dihedrals
- residue secondary structure class
- main-chain conformation class
- residue solvent accessibility
- template/target sequence identity
- resolution of the template structure
- distance from alignment gaps
- B factor
- etc.

Spatial restraints expressed as probability density function (PDF)

- Form a multi-dimensional table *W* with observed values for a list of features
- PDF $p(x \mid a, b, c, \ldots) \approx W(x, a, b, c, \ldots)$
  - *a, b, c*: features
  - $\approx f(x, a, b, c, \ldots, \mathbf{q})$, where *f* is an analytic function with a parameter **q** defined to minimise the root mean square deviation of *f* from *W* (for some features, e.g. bond length)
  - Features with floating point values are split into classes or binned
- Using PDF to calculate probability of a certain range of values:
  $P(x_1 \leq x < x_2) = \int_{x_1}^{x_2} p(x)dx$

Statistical potential (knowledge-based potential)

- Energy function *F*(*r*) derived from an analysis of known protein structures
  - Or a probability density function $P(r) = \frac{1}{Z} e^{-\frac{F(r)}{kT}}$
  - Needs a reference state $\Delta F(r) = -kT \ln \frac{P(r)}{Q_R(r)}$
- E.g. **potential of mean force** (Sippl, 1990) describing pairwise distances between all pairs of amino acids:

$$\Delta F_T = \sum F(r_{ij}|a_i, a_j) = -kT \sum \ln \frac{P(r_{ij}|a_i, a_j)}{Q_R(r_{ij}|a_i, a_j)}$$

- $\Delta F_T < 0$: pairwise distances in the sample more probable than in the reference state

Restraints for the whole model

- Target function: $P = \prod p(f_i) \rightarrow \max \Leftrightarrow F = -\ln P \rightarrow \min$, where f_i is the *i*-th restraint
- Optimization starts from local restraints, other restraints are added sequentially: from neighbouring residues to long-range interactions
- Hence MODELLER performs maximum likelihood optimization given a **statistical potential** derived from a databased of known 3D structures
- Same or other statistical potentials can be used to assess the model (e.g. DOPE score)

E.g. modelling loops

- Position (backbone) atoms uniformly
- Randomize by shifting every coordinate by ±5Å
- Optimize using an energy function that includes:
  - bond lengths
  - bond angles
  - dihedral angles
  - improper dihedral angles (between atoms that are not bound covalently)
  - torsion angle conditional on residue
  - side chain torsion angles
  - non-bonded interactions conditional on atom type, distance, and separation in sequence

Using MODELLER

- Stand-alone tool
- https://salilab.org/modeller/
- License free for academic users
- Python-like scripting language

Advanced modelling with MODELLER

- Multiple templates
- Multiple chains
- Hydrogens
- Ligands and water atoms

MODBASE

- A database for models of all proteins in UniProt at https://modbase.compbio.ucsf.edu
  - One template per model
  - Multiple templates are chosen
  - Models are subsequently evaluated
- Models are score based on their **statistical energy potential**, **sequence similarity** of the template and **structural compactness**

SWISS-MODEL
(Torsten Schwede lab, Uni Basel)

Nicolas Guex
Manuel C. Peitsch

SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling

ProMod and Swiss-Model: Internet-based tools for automated comparative protein modelling

M. C. Peitsch

Nucleic Acids Research, 2003, Vol. 31, No. 13 3381–3385
DOI: 10.1093/nar/gkg520

SWISS-MODEL: an automated protein homology-modeling server

Torsten Schwede[1,2,*], Jürgen Kopp[1,2], Nicolas Guex[3] and Manuel C. Peitsch[2,4]

[1]Biozentrum der Universität Basel, Klingelbergstr. 50-70, CH 4056 Basel, Switzerland, [2]Swiss Institute o
Bioinformatics, Basel, Switzerland, [3]GlaxoSmithKline, Research Triangl
4002 Basel, Switzerland

228 Plan-les-Ouates/Geneva, Switzerland

COMPUTER CORNER

TIBS 24 – *SEPTEMBER 1999*

Protein modelling for all

additional model building and analysis algorithms. The current Macintosh and PC versions of the Swiss-PdbViewer rely on Apple's QuickDraw3D for some as-

BIOINFORMATICS ORIGINAL PAPER

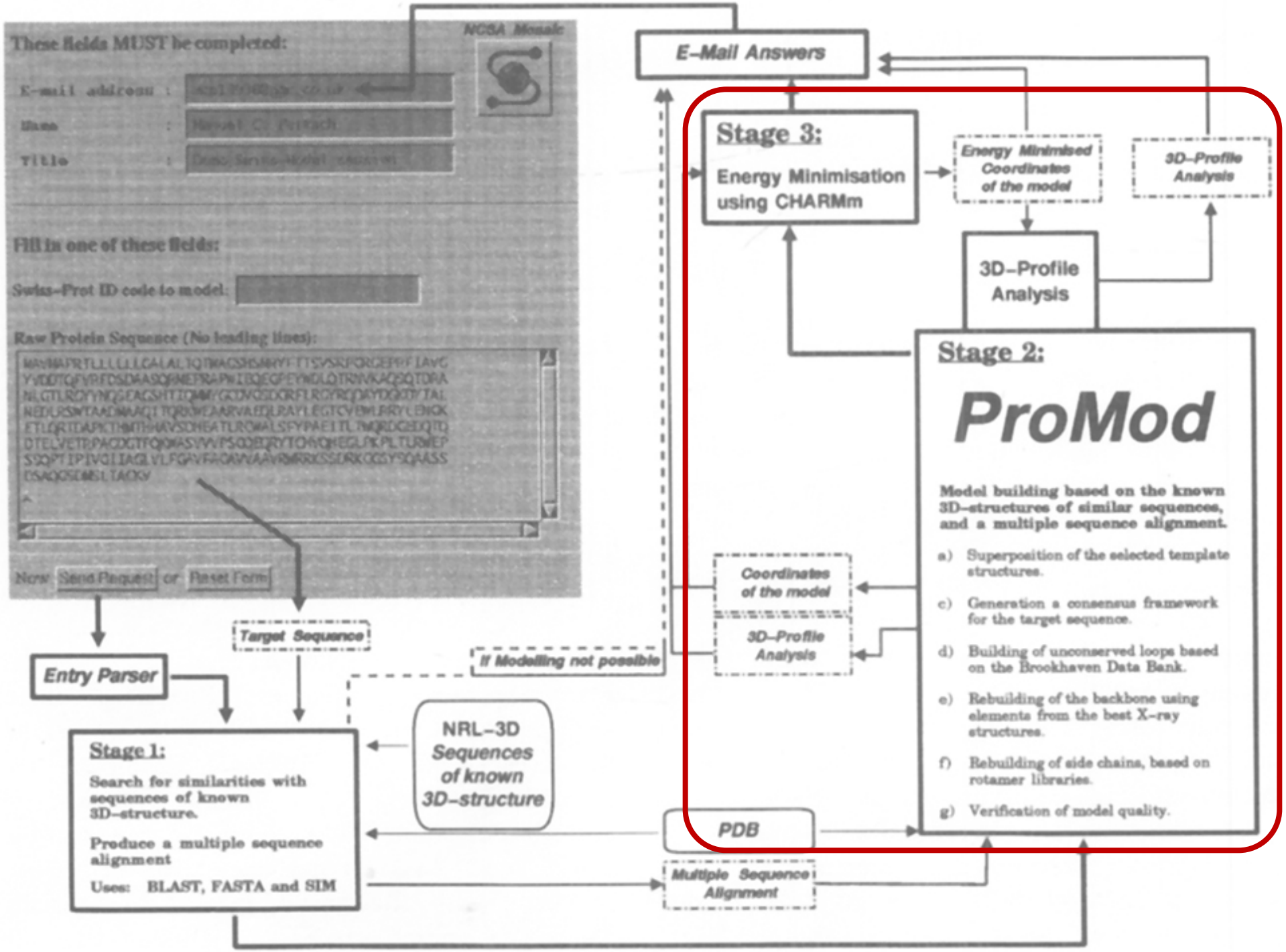
Vol. 22 no. 2 2006, pages 195–201
doi:10.1093/bioinformatics/bti770

Structural bioinformatics

The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling

Konstantin Arnold[1,2], Lorenza Bordoli[1,2], Jürgen Kopp[1,2] and Torsten Schwede[1,2,*]

[1]Biozentrum Basel, University of Basel, Switzerland and [2]Swiss Institute of Bioinformatics, Basel, Switzerland


These fields MUST be completed:
NCSA Mosaic
E-mail address :
Name :
Title :
Fill in one of these fields:
Swiss-Prot ID code to model:
Raw Protein Sequence (No leading lines):
Now Send Request or Reset Form
E-Mail Answers
Stage 3:
Energy Minimisation using CHARMm
Energy Minimised Coordinates of the model
3D-Profile Analysis
3D-Profile Analysis
Stage 2:
ProMod
Model building based on the known 3D-structures of similar sequences, and a multiple sequence alignment.
a) Superposition of the selected template structures.
c) Generation a consensus framework for the target sequence.
d) Building of unconserved loops based on the Brookhaven Data Bank.
e) Rebuilding of the backbone using elements from the best X-ray structures.
f) Rebuilding of side chains, based on rotamer libraries.
g) Verification of model quality.
Coordinates of the model
3D-Profile Analysis
If Modelling not possible
Target Sequence
Entry Parser
Stage 1:
Search for similarities with sequences of known 3D-structure.
Produce a multiple sequence alignment
Uses: BLAST, FASTA and SIM
NRL-3D Sequences of known 3D-structure
PDB
Multiple Sequence Alignment

Template selection

- Multiple templates
- Sequence- and structure-based alignment
- Weights for templates determined at each position based on local target-template sequence similarity

- (Original algorithm, not in the current webserver)

Target structure modelling

- Weighted average position for each **main-chain** atom in the target sequence computed
- **Side-chain** atoms:
  - conserved atoms modelled by weighted average positions
  - non-conserved atoms modelled based on a backbone-dependent rotamer library and a scoring function
    - van der Waals exclusion: **unfavourable interactions**
    - H-bonds, disulfide bonds: **favourable interactions**

Loop modelling

- **Before** adding side chains
- Ensemble of loops compatible with neighbouring stems (C_α atoms of 4 residues before and after the loop) constructed
  - Loop: length + stem conformation in experimentally resolved structures — **not necessarily loops in the sense of secondary structure**!
- Best loop conformation selected based on:
  - steric (van der Waals) clashes
  - favourable interactions
  - energy

Energy of side-chain conformations

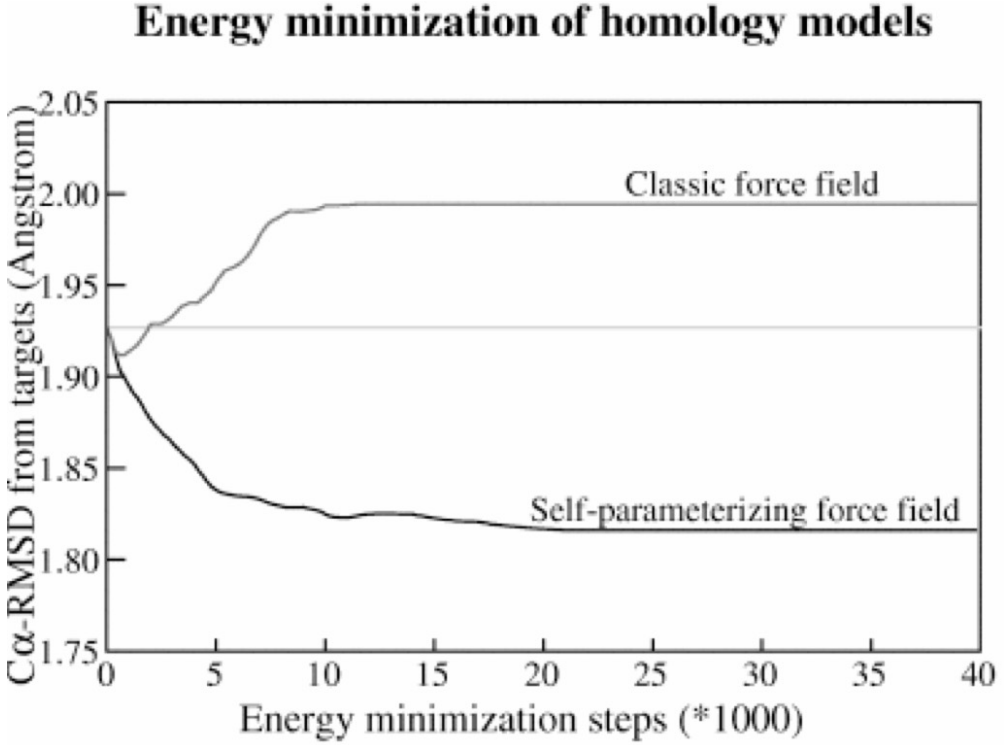
- SCWRL4 (Krivov et al., PROTEINS, 2009)
- Backbone-dependent rotamer library
- Energy function depends on:
  - rotamer frequency
  - steric repulsion (**increases energy**) and H-bond energy (**decreases energy**)
  - self- and pairwise rotamer energy terms
- Efficient way to solve combinatorial packing problem

Model optimization

- Steepest descent using GROMOS96 force field for structure regularisation
- Short molecular dynamics runs for structure optimization
- Empirical force fields to detect structural errors

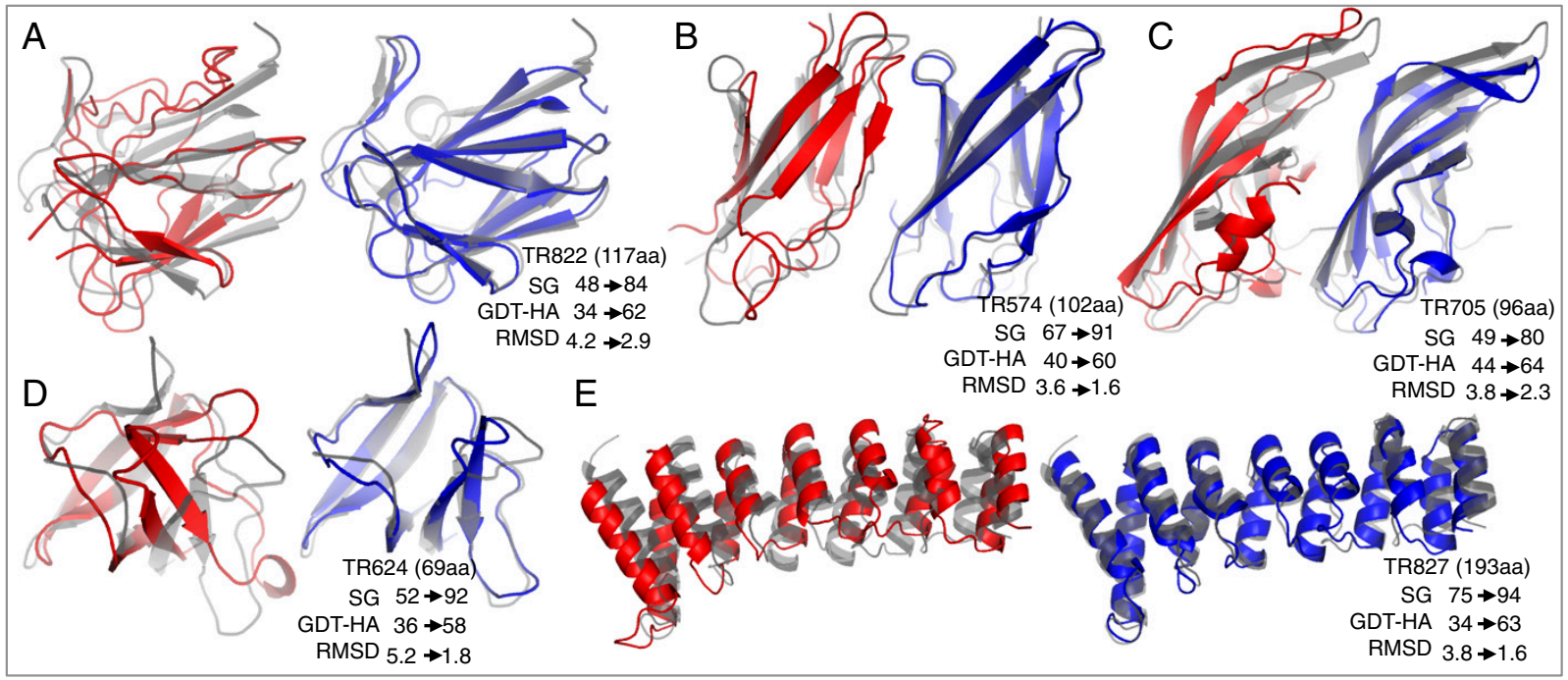
Applying molecular dynamics for model optimization

- **Molecular dynamics**: modelling protein dynamics using the Newtonian laws of motion
- Usually does not work well in model evaluation
  - force fields not accurate
  - energy landscape with a lot of local minima
- => Self-parametrizing force fields



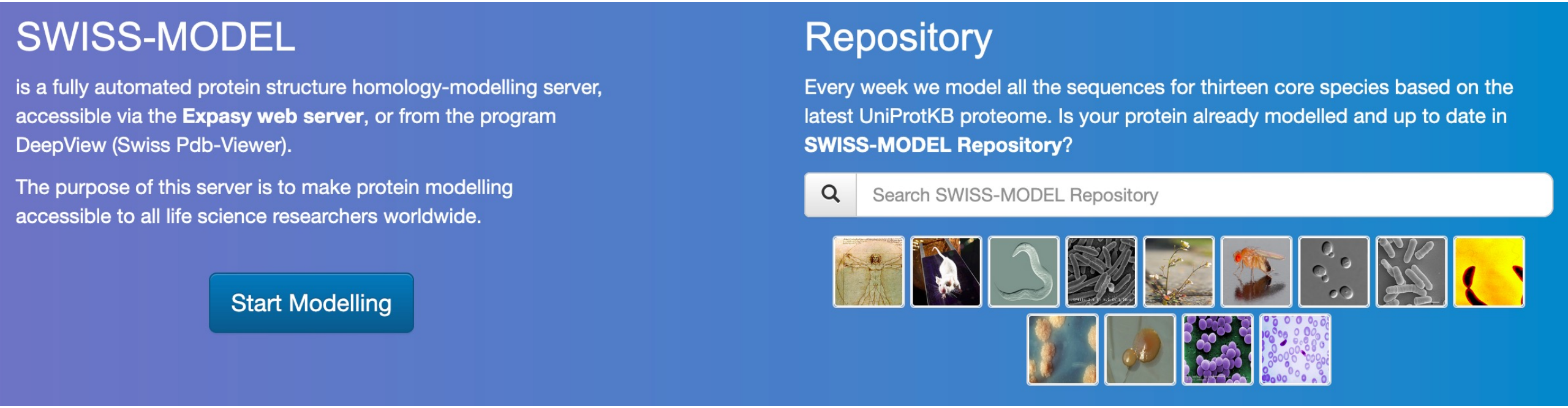
Applying molecular dynamics for model optimization

- Two-step protocols (Park et al., PNAS, 2018)
  - Introduce (significant) structural variation that is not accessible by a standard MD protocol
  - Evolve towards the lowest all-atom energy



Using SWISS-MODEL

- https://swissmodel.expasy.org/



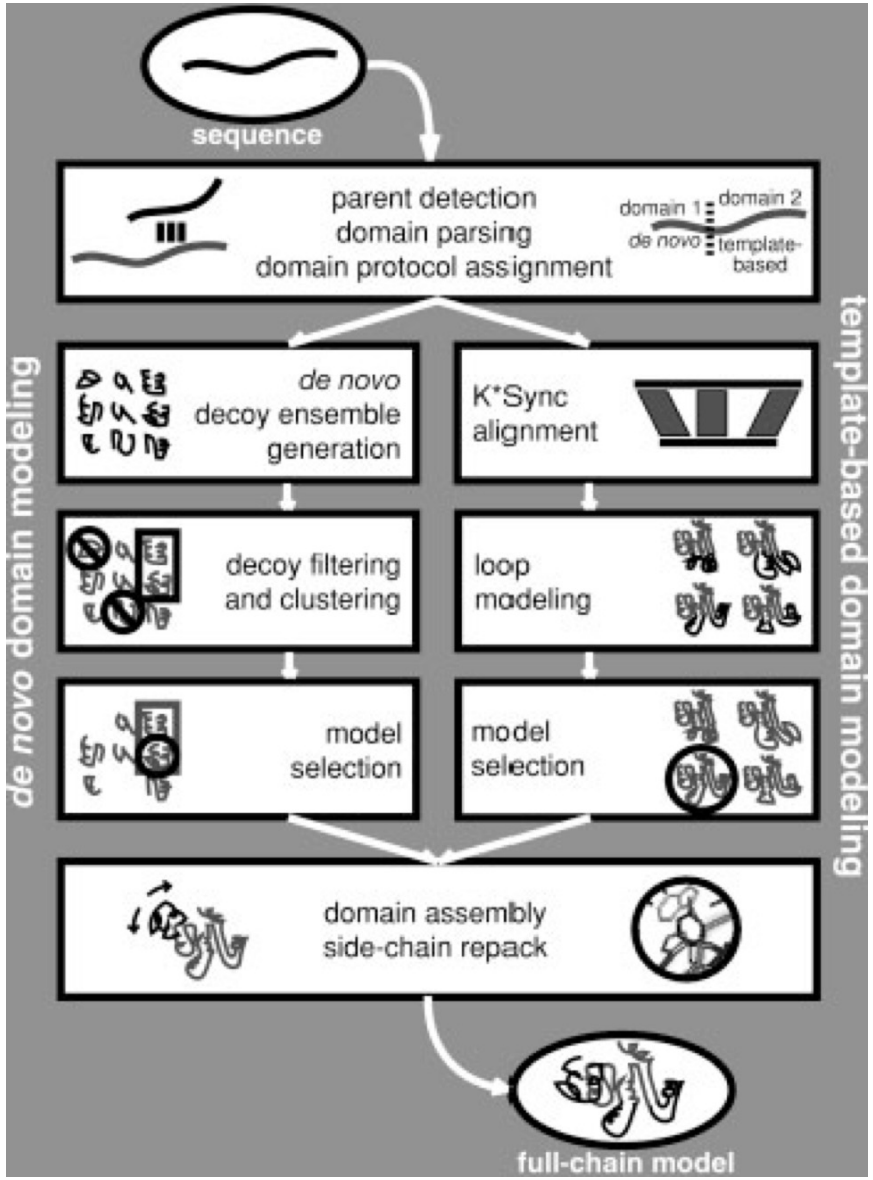
ROBETTA
(David Baker lab, U Washington)

ROBETTA

- A part of the Rosetta suite (*de novo* protein structure prediction, lecture 9)
- Automated server
  - template-based and *de novo* modelling

ROBETTA

1. Target split into domains; a template for each domain
2. Sequence profiles for both template and target
3. Backbone of aligned positions modelled
4. Loops modelled from fragment library
5. Side chains packed using a backbone-dependent rotamer library
6. Fragments are used to orient domains relative to each other
7. Whole model optimized with own potential



Concluding remarks

The key to success: a clever choice of the template

- Combine several templates
- Use different templates for different parts of the structure
- Use different sequence-to-structure alignment methods

Model assessment

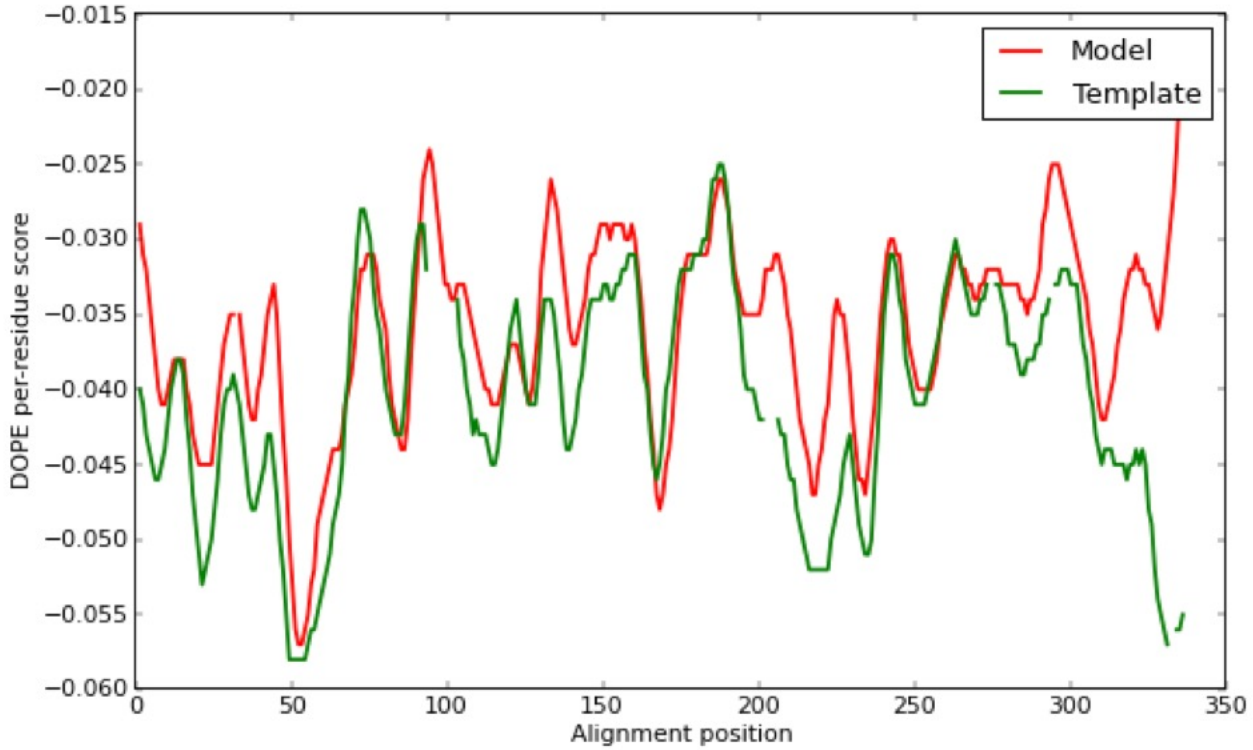
MODELLER assessment tools

- Quality of alignment (`alignment.check()`, but manual inspection important!)
- (Pseudo-)Energy potentials:
  - DOPE (Discrete Optimized Protein Energy)
  - GA341
  - …possibility to plug in any potential (`selection.assess()`)

DOPE (Discrete Optimized Protein Energy, Shen and Šali, *Protein Sci*, 2006)

- `model.assess_dope()`
- **Statistical potential** based on a non-redundant set of 1472 X-ray protein 3D structures
- Pairwise distances between residues (cf. potential of mean force)
- **Reference state**: sphere of an appropriate radius depending on protein size
- Evaluates whether the set of distances in the model are more probable than random
- Can be calculated for individual residues or smoothed over a window

DOPE profile



GA341 (Melo *et al.*, *Protein Sci*, 2002)

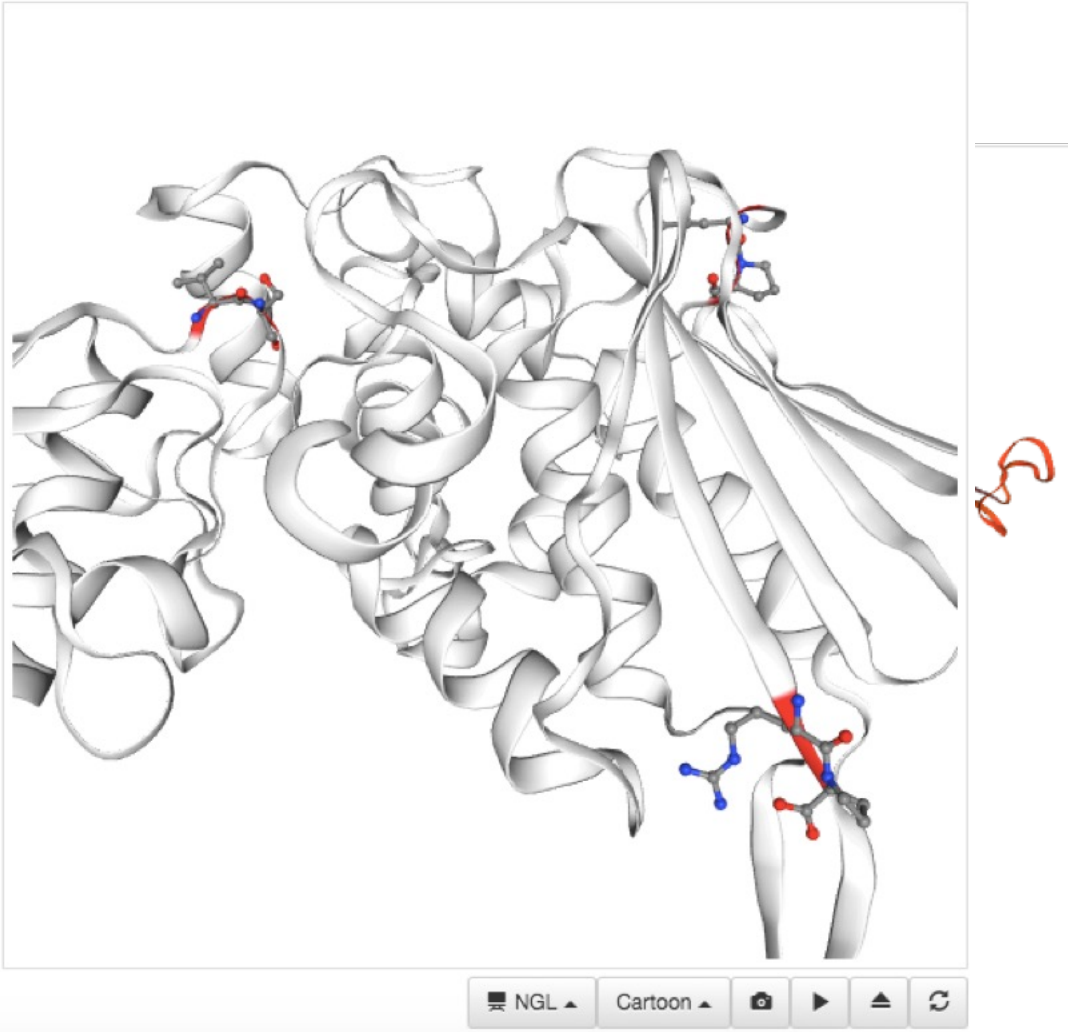
- **4** statistical potentials combined:
  - distance-dependent
  - contact (same as distance-dependent, but 0 above a contact threshold)
  - dihedral angle
  - accessible surface

SWISS-MODEL assessment tools

MolProbity Results

	MolProbity Score	2.21	
☐	Clash Score	3.74	(A46 PHE-A91 HIS)
	Ramachandran Favoured	87.89%	
☐	Ramachandran Outliers	4.47%	A105 VAL, A225 GLY, A178 HIS, A291 ARG, A79 ARG, A24 PHE, A80 PRO, A305 ALA, A52 ASP, A319 ASN, A350 LEU, A41 ALA, A26 ASN, A91 HIS, A193 ASN, A214 PRO, A345 LEU
☐	Rotamer Outliers	3.85%	A30 LEU, A279 VAL, A106 SER, A306 ASP, A203 LEU, A275 ILE, A28 ASP, A21 LYS, A40 THR, A214 PRO, A101 VAL, A337 LEU
☐	C-Beta Deviations	5	A26 ASN, A193 ASN, A111 ASP, A293 TYR, A339 SER
	Bad Bonds	0 / 3002	
☐	Bad Angles	36 / 4078	(A91 HIS-A92 PRO), (A345 LEU-A346 PRO), (A148 THR-A149 PRO), (A182 LEU-A183 PRO), (A160 LEU-A161 PRO), A37 PHE, A64 ASP, (A258 THR-A259 ARG), (A104 GLY-A105 VAL), A41 ALA, A239 ILE, A91 HIS, A189 MET, A214 PRO, A315 HIS, A123 ASP, (A79 ARG-A80 PRO), A178 HIS, A350 LEU, A18 HIS, A40 THR, A238 ASP, A60 HIS, A165 THR, A146 ASP, A144 HIS, (A177 GLN-A178 HIS), A28 ASP, A117 TRP
☑	Cis Non-Proline	1 / 362	(A105 VAL-A106 SER)
☑	Twisted Prolines	2 / 19	(A213 LEU-A214 PRO), (A381 ARG-A382 PRO)

Results obtained using MolProbity in Phenix version 1.13, release 2998



QMEAN: Qualitative Model Energy ANalysis

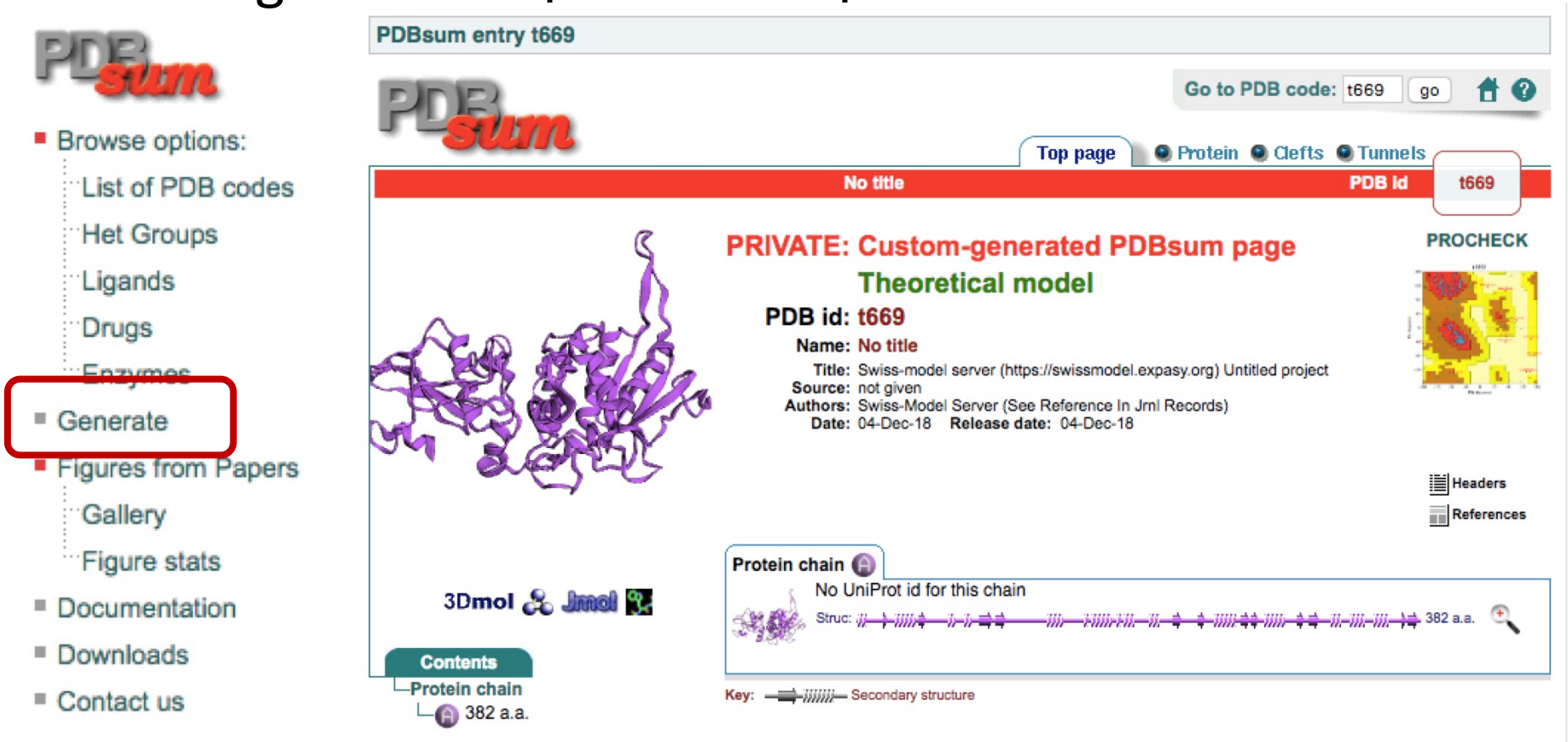
- Linear combination of six structural descriptors:
  1. Distances between Cβ atoms, secondary structure- and residue type-specific
  2. Same for all atoms of a residue
  3. Solvation (relative solvent accessibility), residue type-specific
  4. Torsion angles, reside type-specific
  5. Agreement between secondary structure predicted from sequence and observed in the model
  6. Same for predicted / observed solvent accessibility
- **Statistical potentials** (expected distributions of values) calculated based on a non-redundant subset of high-quality 3D protein structures from PDB
- Normalized to the segment [0, 1]

GMQE: Global Model Quality Estimation

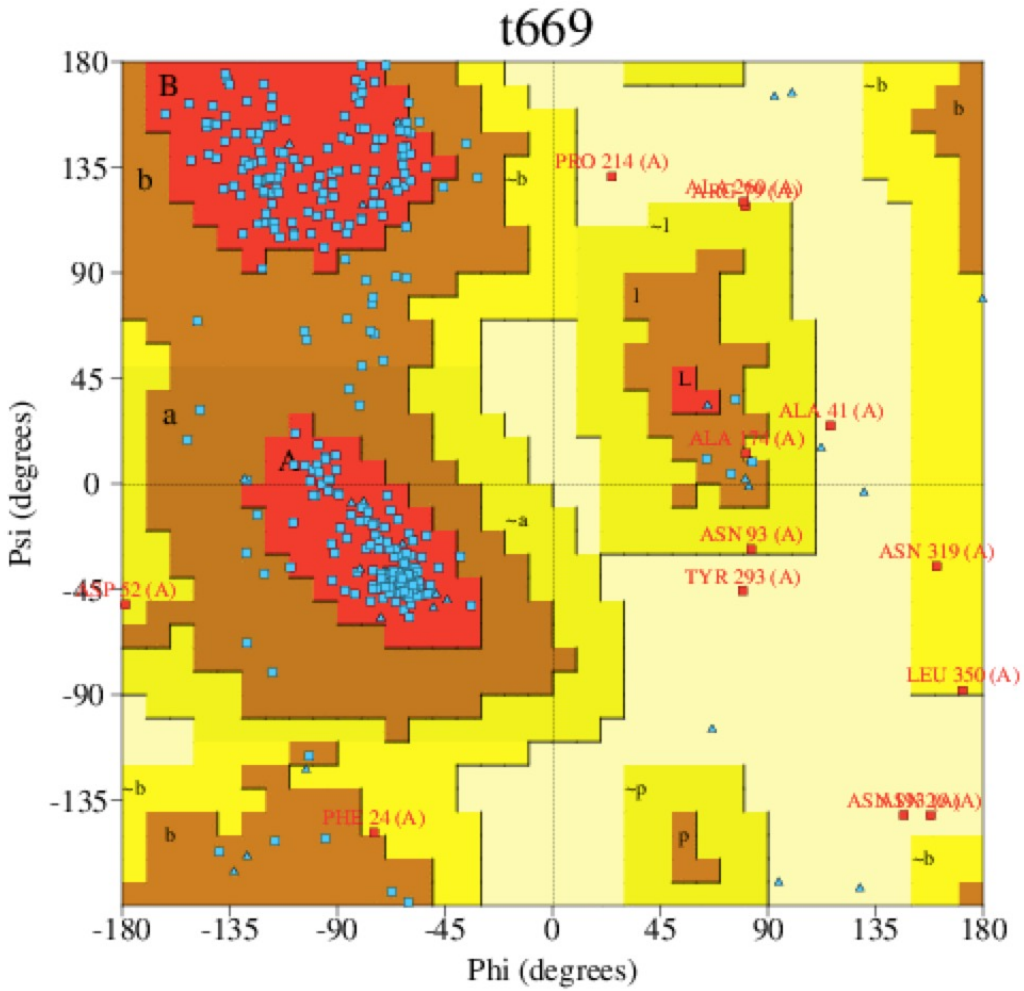
- Combines properties from the target-template alignment and the template search method (BLAST or HHblits)
  1. sequence identity
  2. sequence similarity
  3. alignment score
  4. agreement between predicted secondary structure of target and template
  5. agreement between predicted solvent accessibility between target and template
  6. all normalized by alignment length
- Scaled to the segment [0, 1]
- **QSQE**: a version for protein complexes

External assessment tools: PDBsum

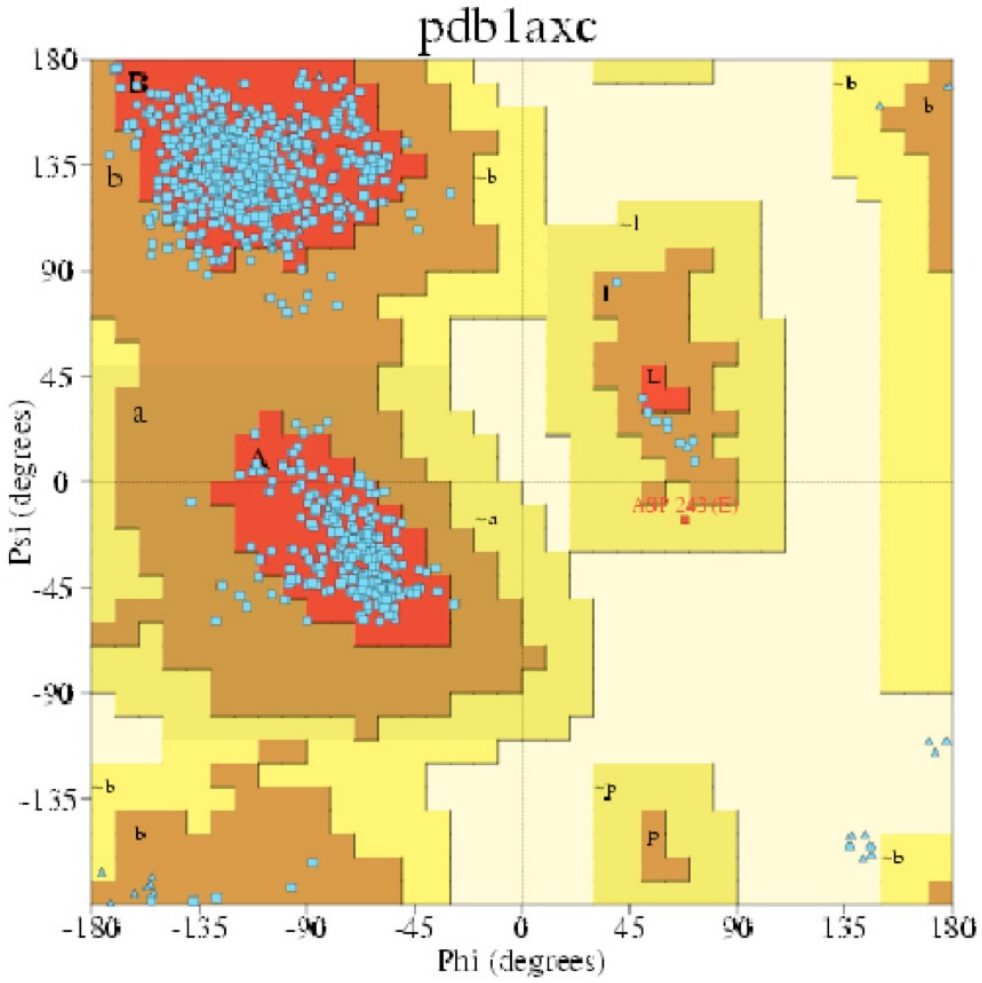
- http://www.ebi.ac.uk/pdbsum
- Designed for experimental protein 3D structures



Theoretical model



Naturally occurring protein



This is a good model!

Summary and possible exam questions

- Sequence similarity search methods: sensitivity and runtime
- Seven steps of homology modelling:
  1. Template recognition and initial alignment
  2. Alignment correction
  3. Backbone generation
  4. Loop modeling
  5. Side chain modelling
  6. Model optimization
  7. Model validation
- Model assessment approaches and tools