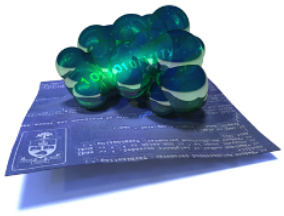
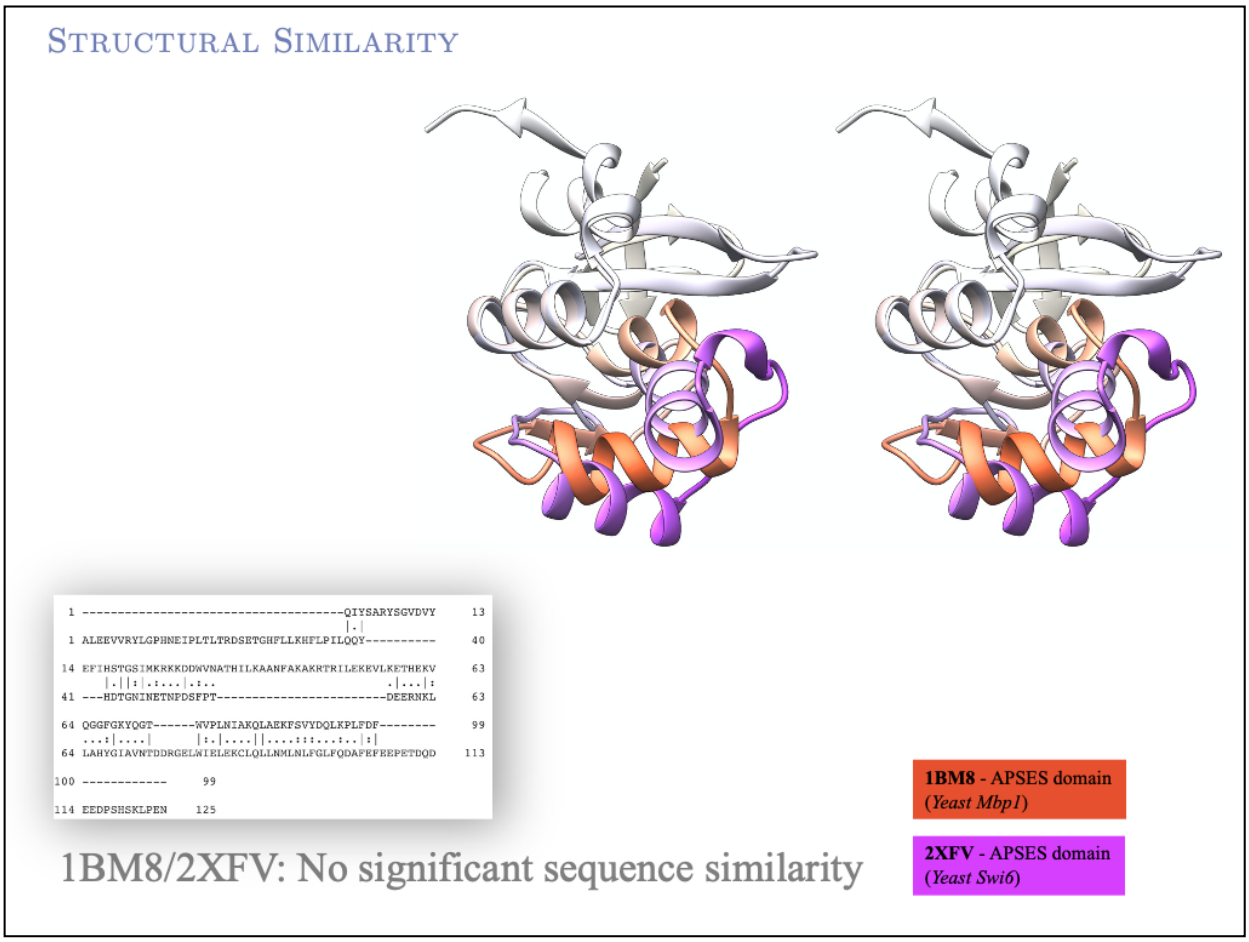
A
Bioinformatics
Course

Homology Modelling



Boris Steipe

department of biochemistry – department of molecular genetics
University of Toronto

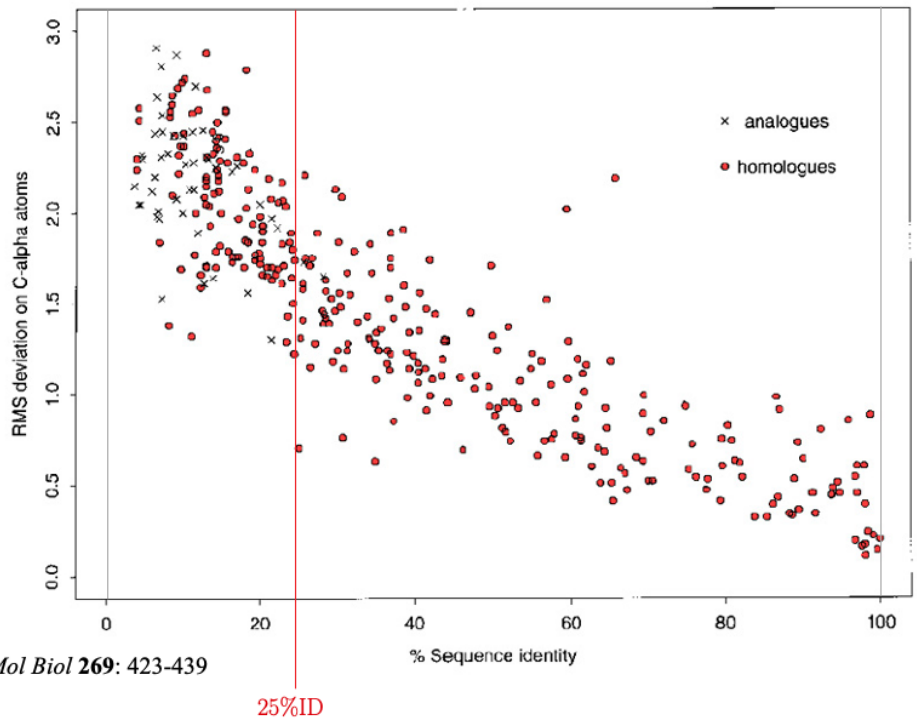


Structures can be similar despite sequences being dissimilar.

In the case of the Mbp1 APSES domain compared to its homologue Swi6, an indel after the DNA binding helix has shifted the orientation of two helices (red and magenta) and destroyed the ability of the Swi6 ancestor to bind DNA, the domain took on different fucntions and has diverged beyond any detectable sequence similarity. Nevertheless, the structures at the termini of the domains can be perfectly superimposed (white).

Homologues proteins retain structural similarity even if they are highly diverged. As a corrollary, once homology can be established between two sequences, based on sequence similarity, it is virtually certain that the structure of one protein can be modelled from knowledge of the structure of the other

Structural Similarity



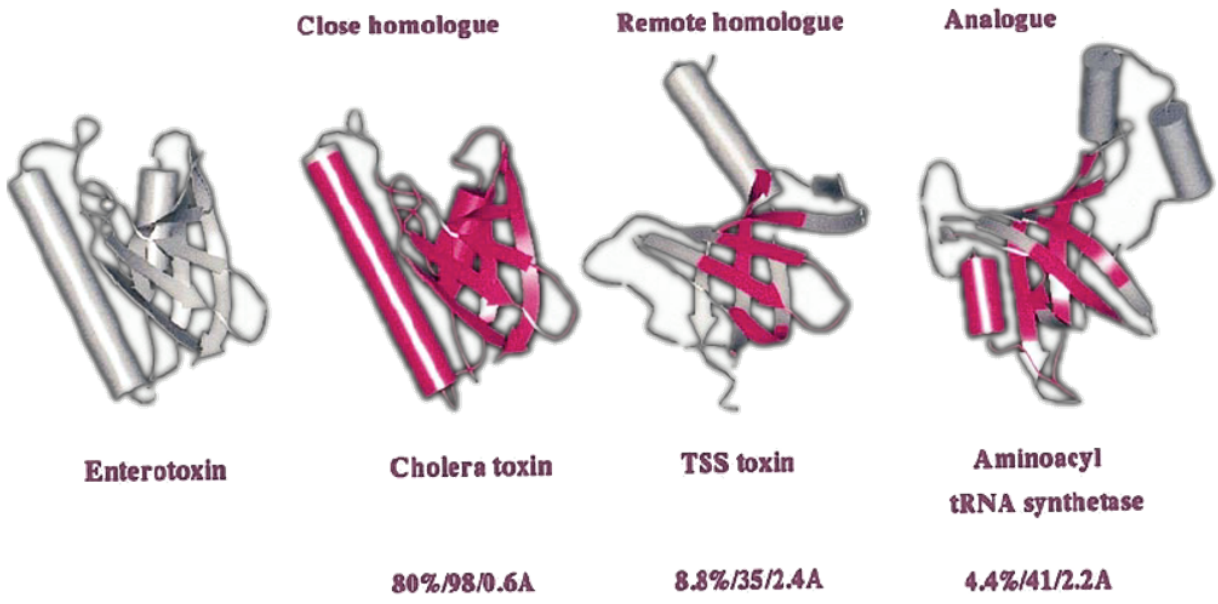
Russell *et al.* (1997) *J Mol Biol* **269**: 423-439

Proteins that diverge in evolution maintain their global fold !

Note the 25% ID level, which we usually take as the cutoff for inferring homology from sequence similarity.

Even at 100% sequence identity (*i.e.* the structure of the *same* protein observed under different conditions) structures can vary by 0.3Å RMSD or more. But up to ~50% identity, structure is conserved at less than 1.0Å RMSD – less than the length of a carbon-hydrogen bond! At the extreme of dissimilarity, even in the complete absence of sequence identity, structural similarity does not drop below 3.0Å for homologous structures. However, at the left hand of this distribution, homologous and analogous folds cannot be distinguished by sequence or structure similarity alone.

Homology and Analogy



Fold space is finite, and structures that have a similarly folded core can have arisen by convergent evolution. These are not homologous folds, they are **analogous.**

This example show close and remote homologues of enterotoxin, and a structurally similar aminoacyl tRNA synthetase structure. Numbers in brackts are %ID / number of matched residues / RMSD of superposition. By these metrics, the tRNA synthetase is a better match to enterotoxin than the TSS toxin. But it is not a homologue.

cf. Russell *et al. (1997)* Recognition of analogous and homologous protein folds: analysis of sequence and structure conservation. *J Mol Biol.* **269**:423-439.

Homology Modelling

How to build a homology model (in three easy steps).

1: Align a **target** sequence with the sequence of a suitable structure **template**.

2: **Replace** the coordinates of sidechains in the template structure file with coordinates for target residue sidechains. This is the **model**.

3: That's all. *(except you can still mess it up...)*

Target: the sequence for which you want to obtain a modelled structure.

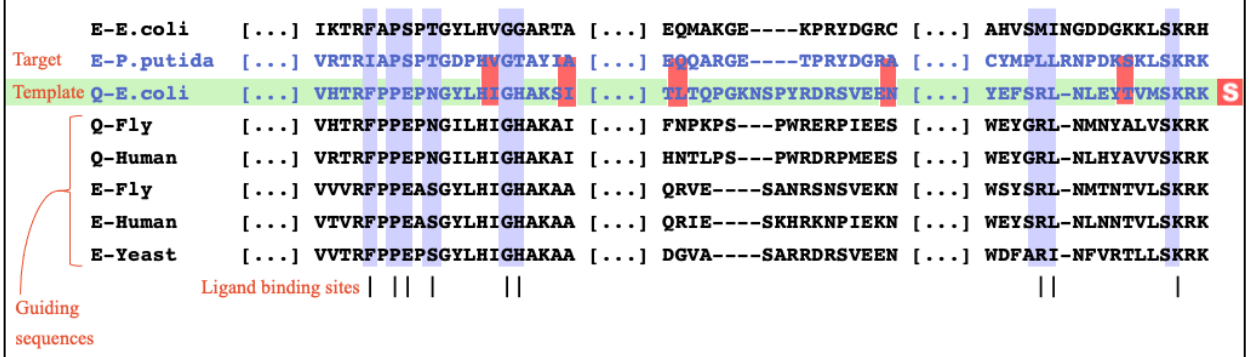
Template: a sequence that is homologous to the target and for which the structure is known.

Model: Structural model of the target, obtained by replacing template sidechans with target sidechains

Sometimes researchers add a third step to homology modelling: energy refinement of the model to adopt more reasonable atom-atom contacts, bond-lengths and angles *etc.* While this practice is popular, it is likely to do more harm than good: usually the RMSD between the model and the true structure gets worse.

Thus energy refinement of homology models is in general an example of Cargo Cult science.

Structure Conservation



QRS *E. coli* vs. ERS *P. putida*: ~ 19% pairwise ID

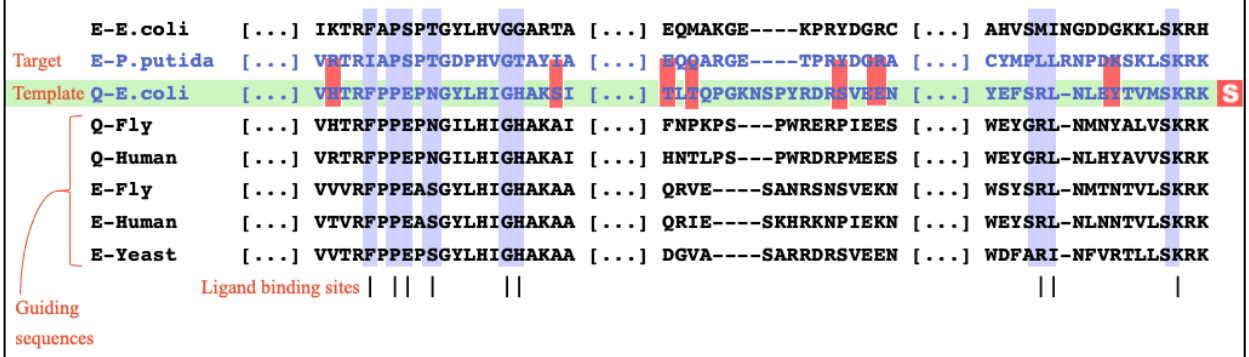
Many regions are expected to be highly conserved in structure.

Some changes are straightforward to model.

Alignment of aminoacyl t-RNA synthetase sequences, and consequences of the types of changes we observe for modeling 3D-structure. In this example, we consider the target of *P. putida* glutamyl tRNA ligase model on the structure of *E. coli* glutaminyl tRNA ligase.

Side chains which can be changed by deleting atoms (I→V, I→A, N→A, T→S), or which merely require changing chemical elements (N→D, even L→Q), are straightforward to model and don't require changing coordinates.

Structure Conservation



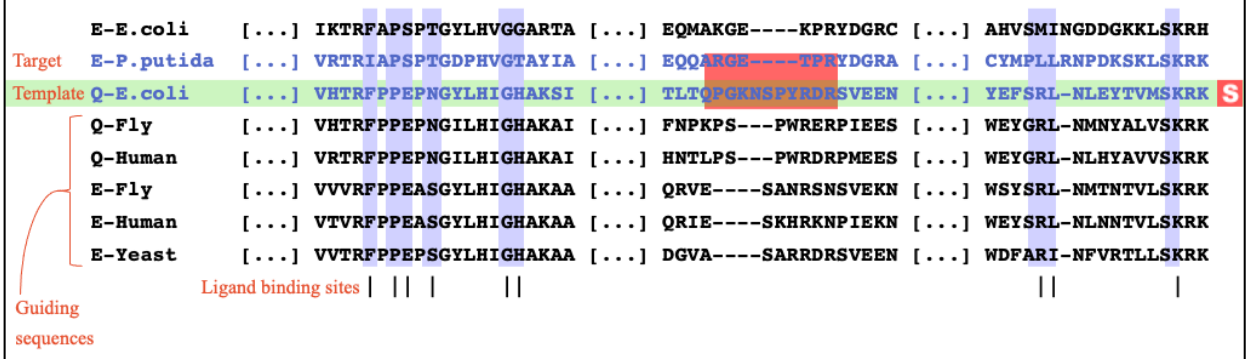
QRS *E. coli* vs. ERS *P. putida*: ~ 19% pairwise ID

How would sidechain rotamers be modeled?

- conserved dihedral angles
- preferred rotamers
- DEE (Dead End Elimination theorem) for global consistency.

For model sidechains that are larger than template sidechains (H→R, S→I, T→Q, V→K ...), we need to decide on the correct geometry. Looking up preferred conformations in "rotamer" dictionaries helps make the decision.

Structure Conservation



How would you (or should you even) model indels?

- Where should the insertion be placed?
- What is the conformation of the new residues?
- Which residues should be deleted?
- How many additional residues need to change conformation?

Indels cannot be modelled with any certainty.

Homology modelling accuracy is determined by sequence alignment accuracy.

The goal for an alignment to be used in homology modelling is a sequence to structure mapping, not sequence to sequence!

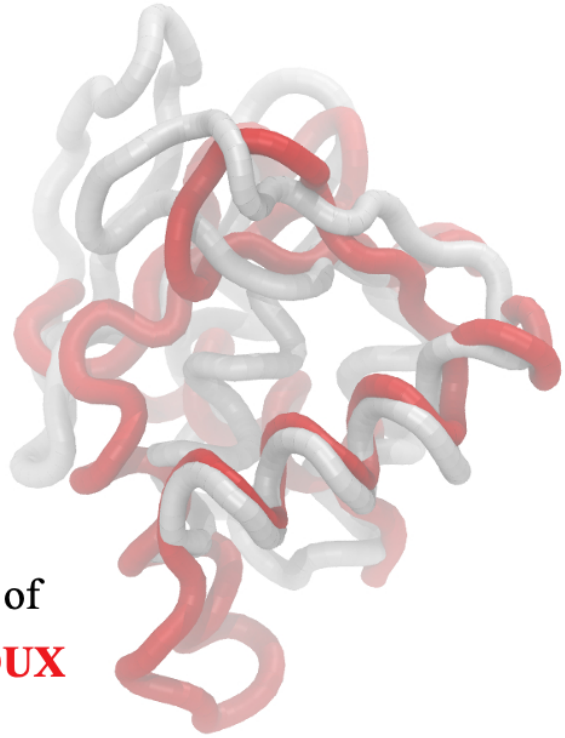
A sequence to structure mapping asks: which residue of the target should replace which residue of the model – this is not necessarily the same as asking which residue of the target is related by evolution to which residue of the model.

Superposition and Alignment

The coordinates of two proteins can be superimposed in space.

An alignment may be derived from a **superposition,** by correlating residues that are close in space.

An **optimal sequence alignment** may lead to a different correlation ...



Superposition of 1BM8 and 1DUX

Alignment is not superposition. These are different procedures and they have different objectives. One can *derive* an alignment from a superposition, by aligning e.g. all residue pairs for which the C^{α} atom distance fall below a threshold, such as 1.9Å, *i.e.* half the average C^{α}–C^{α} separation in proteins.

INDELS

Indels may be structurally accommodated some distance away from the site of their sequence insertion.

Example - a two residue deletion in GFP

Sequence alignment (shows what happened)

```
GFP   DGSVQLADHYQQNTPIGDGPVLLP
      ||...|.:.  ::..:...||.||
RFP   DGGHYLVEF--KSIYMAKKPVQLP
```

Structure superposition (shows how it's accomodated)

```
GFP DGSVQLADHYQQNTPIGDGPVLLP
    ||...|.:.         .||.||
RFP DGGHYLVEFKSIYMAK..KPVQLP
```

Sequence alignment and superposition have different objectives. Alignment – based on an evolutionary model – recovers information on an evolutionary event. Superposition shows how the event has been structurally accommodated.

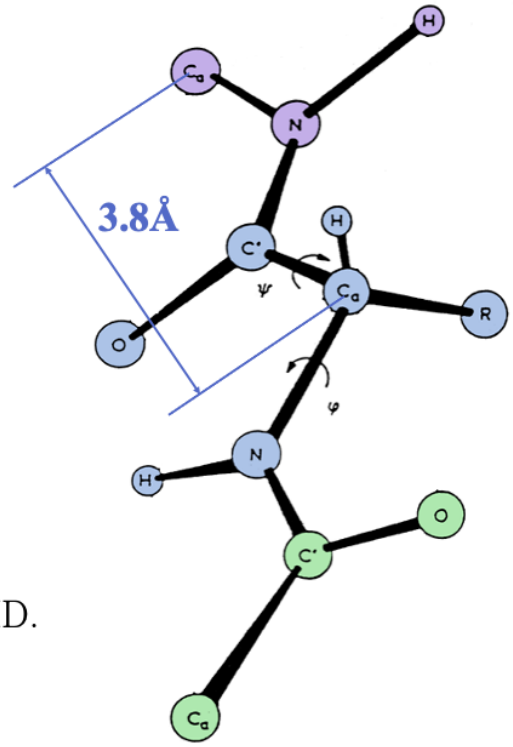
To recover an alignment from a superposition, align each residue with the one it is closest to in the superimposed pair of structures.

Alignment Error and Model Accuracy

A shift in alignment of just *one* residue corresponds to an error in the modelled structure of about 4 Å.

Nothing you can do AFTER the alignment will fix this error. Energy refinements (in general) can't correct errors as large as that.

Remember that RMSDs of homologous proteins are less than 3Å, regardless of %ID.



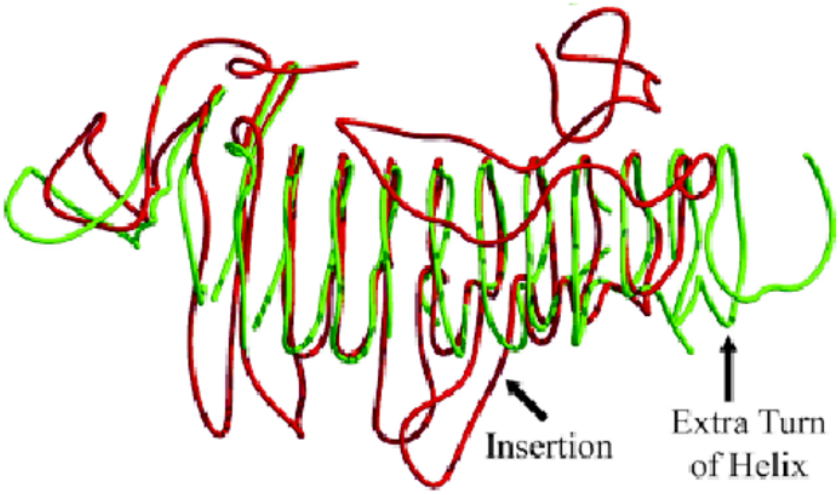
Indels can cause alignment errors. Alignment errors cannot be corrected through modelling.

Alignment is the Most Important Factor for Accuracy

Sequence alignment is the limiting step for homology model accuracy. No amount of energy minimization or other computational procedure will put a misaligned residue in the right place just because it may be more plausible there!
(And sometimes *plausible* still means *wrong*.)

TRUE structure

Homology model



From:

HOMSTRAD @ CASP4:

Williams MG et al. (2001) *Proteins* **Suppl.5**: 92-97

In this example, the alignment "slipped" on an inserted loop in the beta-helix ("Insertion") and added an additional loop at its end. This makes more than half of the model "wrong", and there is *no way* to resolve this error by computational means, short of outright *ab initio* structure prediction.

It's absolutely crucial to get the alignment right.

Alignment is the Most Important Factor for Accuracy

Consequently:
use the best alignment that is available, in situations where there is any doubt (i.e. always in the presence of indels)

Use a carefully computed multiple sequence alignment

Include your target and your selected template sequence

Include many relevant homologues to resolve ambiguities

Manually edit results for plausibility of the target/template alignment, consider template secondary structure
(cf. SAS: https://www.ebi.ac.uk/thornton-srv/databases/sas/ or
PDBSUM: https://www.ebi.ac.uk/pdbsum)

Extract the input pairwise alignment from the multiple alignment

Your alignment should strive for the most plausible sequence-to-structure mapping.

After producing an MSA from carefully selected sequences, extract the pairwise alignment of target and template by copying the two rows, and removing all gap characters that are present in **both** sequences.

Note that PSI-BLAST is probably not the best tool to search for related sequences for homology modeling: the goal is not to have a comprehensive set of homologues, but to include sequences that improve the accuracy of the MSA; that will generally require high levels of sequence similarity. In contrast, the use-case for PSI-BLAST is detecting very distant relationships.

As a rule: **homologous sequences that are not descendants of the Last Common Ancestor of target and template, are unlikely to be useful for the alignment.**

Modelling indels

- Comparisons of alignments and structures demonstrate that *uniform gap penalty assumptions* are ***not biological.***
- Indels are most often observed in loops, less often in secondary structure elements
- When they do not occur in loops, there is frequently a maintenance of helical or strand properties and the indel is structurally accommodated in a position different from where the insertion occurred.
- In that case, the homology model should be based on the superposition, not on the optimal sequence alignment.

Accordingly:

- Use structure-weighted gap penalties
- Use a multiple sequence alignment
- Use manual improvement of alignment

Structure-Aware Gap Penalties

- Should we use **position specific, structure aware** gap penalties for general alignment, or more specifically to align for homology modeling?
- Some MSA algorithms support the use of **secondary structure masks**.
- Alternatively: use a manual **sequence alignment editor** to move gaps out of secondary structure regions, which you can identify from the *template* structure(s).

However: This may be implicitly achieved by modern (consistency based) multiple sequence alignment programs.

How to choose a template:

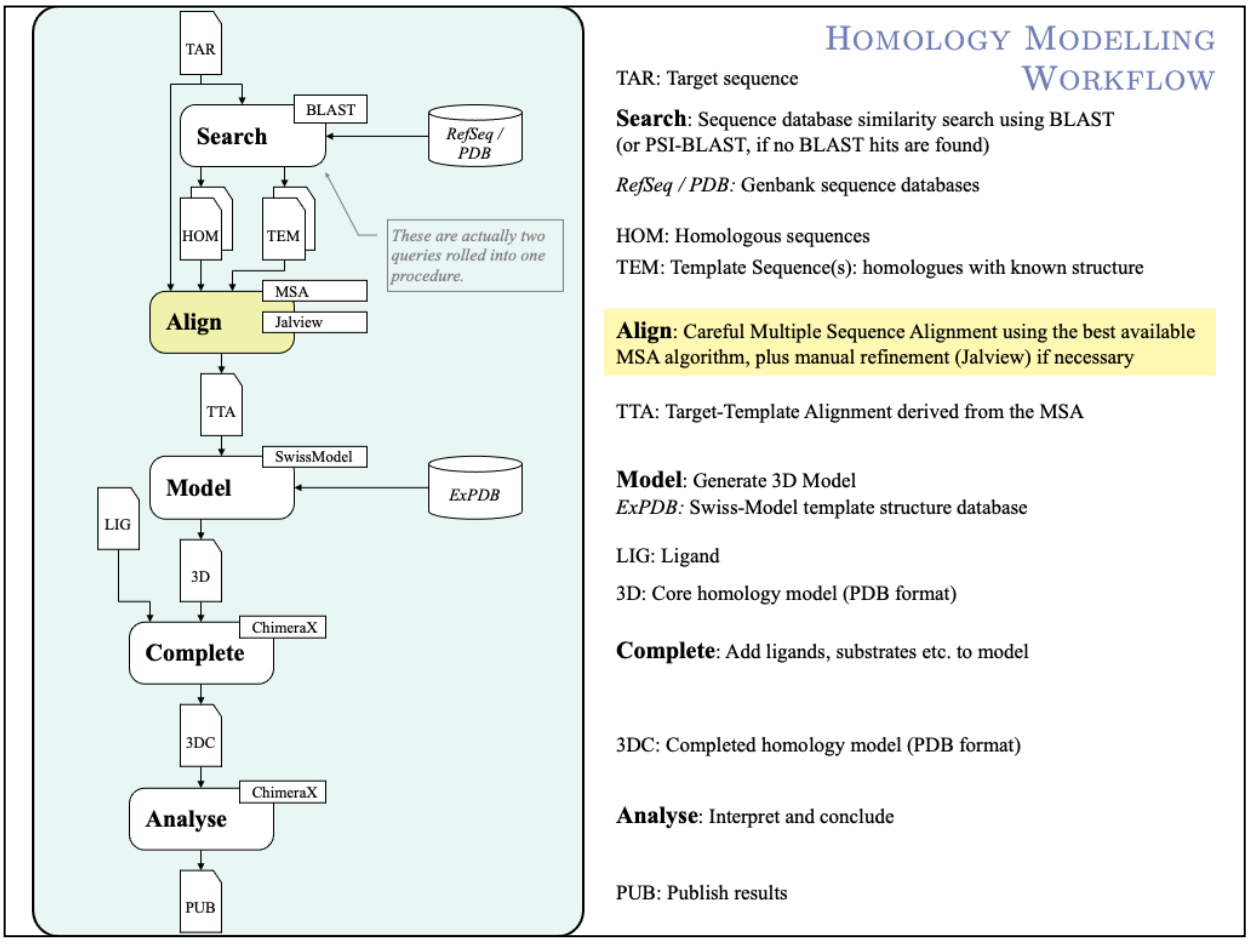
1: Choose the structure with the highest sequence similarity (smallest number of indels!)

2: Choose a structure with bound ligands, cofactors, or other biologically relevant modifications

3: Choose the highest resolution X-ray structure, (except if the resolution is worse than ~3.0Å, then NMR structures might be just as accurate). Don't trust loops to yield correct coordinates.

If 1-3 don't all indicate the same structure, you have to balance the requirements intelligently, based on the purpose of the model. Or create several models and then intelligently cut and paste coordinates.

Include more than one template sequence in the MSA if possible: comparing a template **superposition** with the template **alignment** will be very informative in terms of alignment details.



Workflow for homology modeling sketched in SPN (Structured Process Notation).

Homology Modelling: Online Servers Perform Well

Online servers such as Swiss-Model provide convenient, high-quality homology modelling services for free. Here the APSES domain of *Cryptococcus Neoformans* Mbp1 has been modelled on the orthologous domain of *Magnaporthe oryzae*.

How to assess accuracy:

1: Assume all indels to be wrong.

2: Assume disordered portions of the template have lead to wrongly modelled sections.

3: Structure analysis ("threading", "solvent accessibility", compatibility with ligands) can point out possible alignment errors.

4: Homology modelling programs have tools to assess local reliability of structure.

But: there is no point in "repairing" stereochemistry, you can only review the alignment and try again once you have improved it.

Assume all indels to be wrong: *SwissModel* flags "invented" conformations (loops) by giving them a value of 0.00 in the *occupancy* column of the PDB file.

Solvent exposed sidechain conformations are unreliable because of the lack of packing constraints – except if engaged in conserved ligand interactions.

Ligands may include structurally conserved water molecules – check the template. Water molecules will not have been included by the modeling algorithm.

Homology Modelling: Use of Results

Biochemical inference based on 3D similarity – *i.e.* homology models

- **Bonds**
- **Angles, plain and dihedral**
- **Surfaces, solvent accessibility**

Don't bother!

- Amino acid functions, presence in structure patterns
- **Spatial relationship of residues to active site**
- **Spatial relationship to other residues**
- **Participation in function / mechanism**
- Static and dynamic disorder
- Electrostatics
- *Conservation patterns (structural and functional)*
- *Posttranslational modification sites (but not structural consequences!)*
- *Suitability as drug target*

Just as in structure analysis, treat your model as a spatial map of features and annotations, not necessarily as a represantation of accurate coordinates.

Is it possible to predict function from models? Usually not, however some functions may be incompatible with the model and thus can be excluded from consideration.

Homology Modelling: Cargo Cult

Some types of analysis amount to Cargo Cult:

- Modelling properties that cannot / will not be verified
- Analysing detailed geometry of the model
- Interpreting loop structures near indels
- Inferring relative domain arrangement
- Inferring structures of complexes

You need to consider whether there actually IS a use for the model coordinates – or whether the model is just intended as a pretty picture.

Homology Modelling: Use of Results

Prototype 1: **Analytical**

Explain mechanistic aspects of protein.

(e.g. in terms of)

- residues involved in catalysis
- global properties (like electrostatics)
- shape, relative orientation and distances of domains or subdomains
- flexibility and dynamics - e.g. hypothesizing about the rate limiting step

Homology Modelling: Use of Results

Prototype 2: **Comparative**

Bring conservation patterns into a spatial context in order to infer causality from (database) correlations –

- describe context specific conservation patterns and analyze these according to conserved properties;
- analyze the predicted effect of sequence variation (e.g. for engineering changes, fusing domains or predicting SNP effects);
- distinguish physiological vs. nonphysiological interactions.

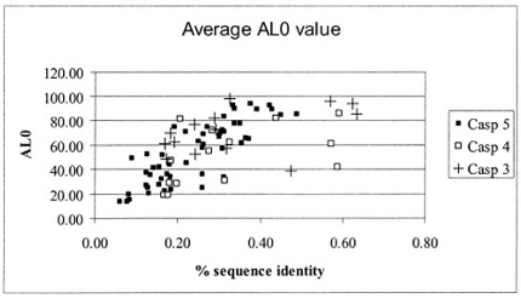
Homology Modelling: Energy Minimization

- Brings protein to lowest energy in about 1-2 minutes CPU time
- Removes atomic overlaps and unnatural strains in the structure
- Stabilizes or reinforces strong hydrogen bonds, breaks weak ones
- Efficient way of "polishing" your protein model
- .... but is it **"true"**???

Energy Minimization Often Makes Models Worse

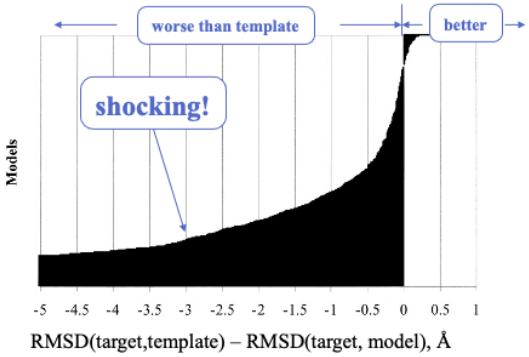
Homology modeling is getting better ...

... because:



Remote sequence similarity detection methods have improved.

... not because:



Coordinate manipulations introduce inaccuracies where there would not need to be any!

Tramontano A & Morea V (2003) Assessment of homology based predictions in CASP5
Proteins **S6:**352-368

Comparison of RMSD(target,template) – RMSD(target, model):

target is the true structure of the target protein that has been independently solved and used for comparison to validate the modelling procedure. If RMSD(target,**template**) is smaller than RMSD(target,**model**), this means that the true structure is more similar to the original template than to the homology model that was produced.

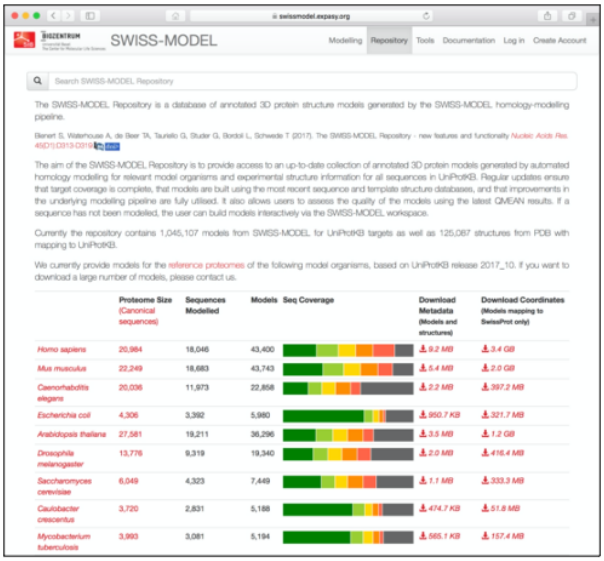
Structure prediction assessments have shown that often the template structure is more similar to the true structure of the target, than the model structure. This is troubling.

Something was done to the template backbone (*i.e.* energy refinement) that actually made the model more wrong than simply keeping the template as-is would have been.

The number of cases were such manipulations **improved** the model – if anything, by a tiny amount – is vanishingly small.

Model Databases may Already Contain the Structure you Need

https://modbase.compbio.ucsf.edu



https://swissmodel.expasy.org/repository

These databases provide high-quality automated models on a genome scale. As you can infer from the coloured bars of the Swiss-Model repository, approximately 50% of a given model organism proteome can be modelled with high confidence based on existing protein structures.

http://steipe.biochemistry.utoronto.ca/abc

B O R I S . S T E I P E @ U T O R O N T O . C A

DEPARTMENT OF BIOCHEMISTRY & DEPARTMENT OF MOLECULAR GENETICS
UNIVERSITY OF TORONTO, CANADA