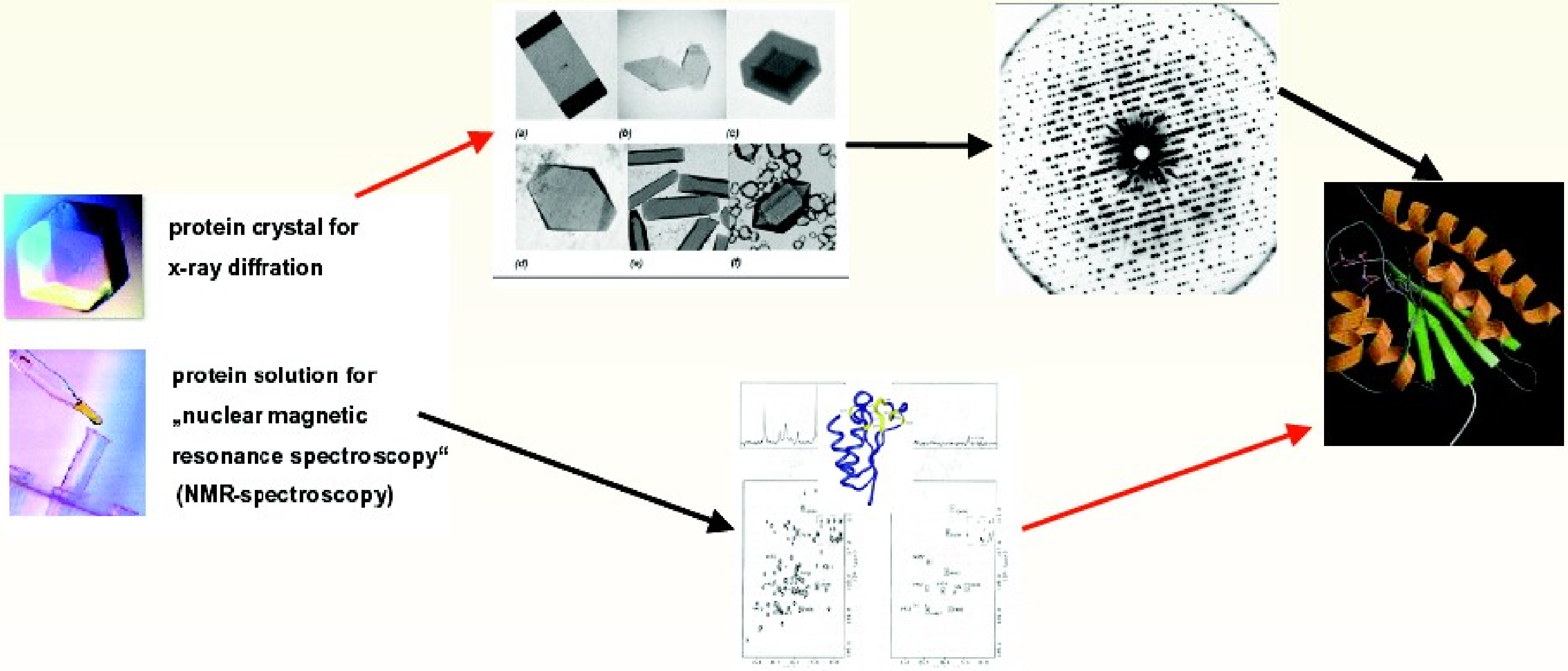
Structure determination is an expensive and time consuming procedure



protein crystal for
x-ray diffration

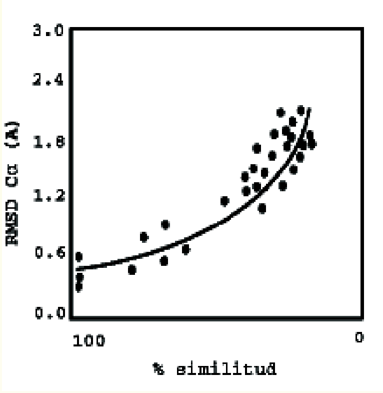
protein solution for
„nuclear magnetic
resonance spectroscopy“
(NMR-spectroscopy)

X-ray -> christalization
NMR -> spectra asignment
EM -> low reolution

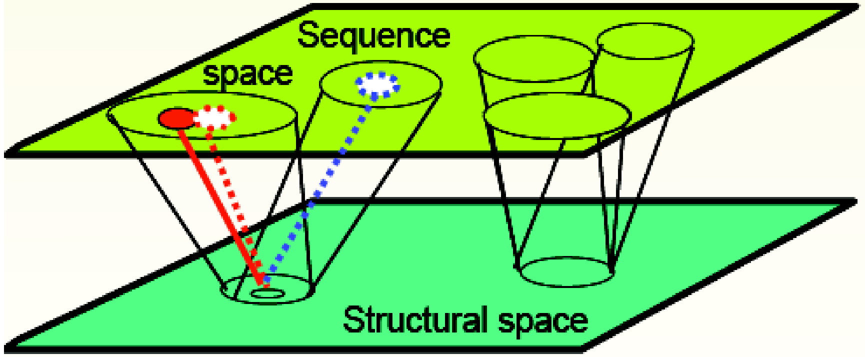
Domain classification

An Introduction to the Protein Structure Databases

Sequence similarity vs structural similarity

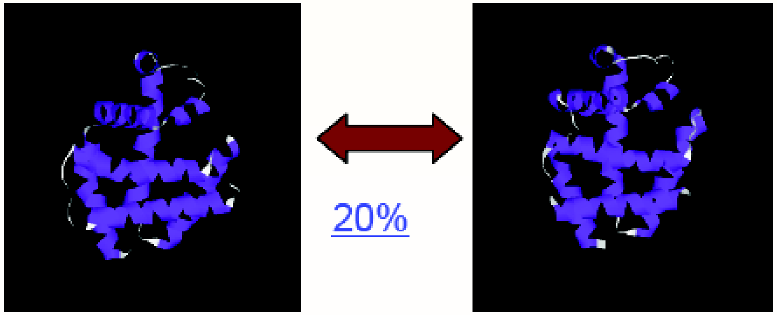


Similar sequences tend to fold into similar structures



Different sequences might fold into similar structures
Convergence

These two principles allow to build protein models

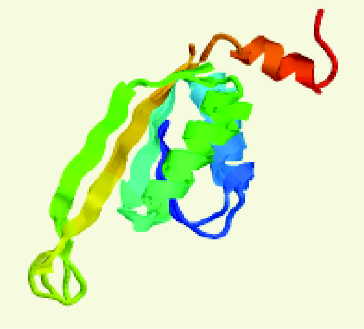


The challenge: to solve all the structural space

Modelling

Simple structure-based modelling

To build a 3D model two things are necessary:



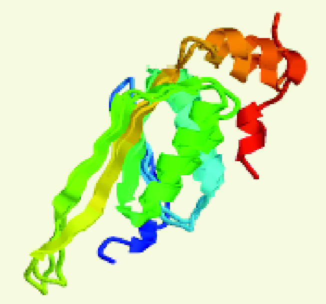
A protein with known structure (the template)

The more similar the template to the sequence of target, the better the model.

+

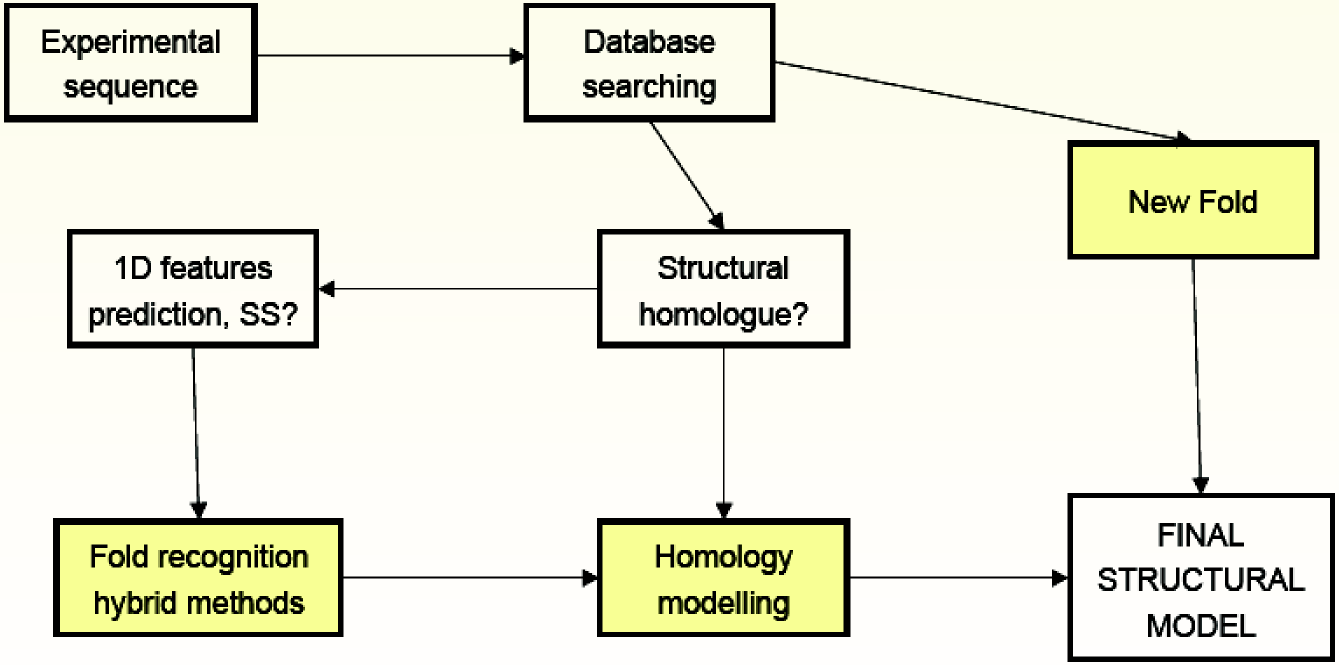
```
TARGET   FIALGQILSESQYQDSSIDKIEDRDIIVRYGCEPTADKFVPIEIFQILRRVKEMDEFLKRVS
         |||.||||.|||..|||||||||||..|||.|||||.|||||.||||.||||||||.|||||
TEMPLATE FIAVGQILPESQWNDSSIDKIEDRDFLVRYACEPTA                KEMDEYLKRVS
```

An alignment between the target sequence and the template protein



= Model structure

Simple structure modelling flowchart



Experimental sequence

Database searching

New Fold

1D features prediction, SS?

Structural homologue?

Fold recognition hybrid methods

Homology modelling

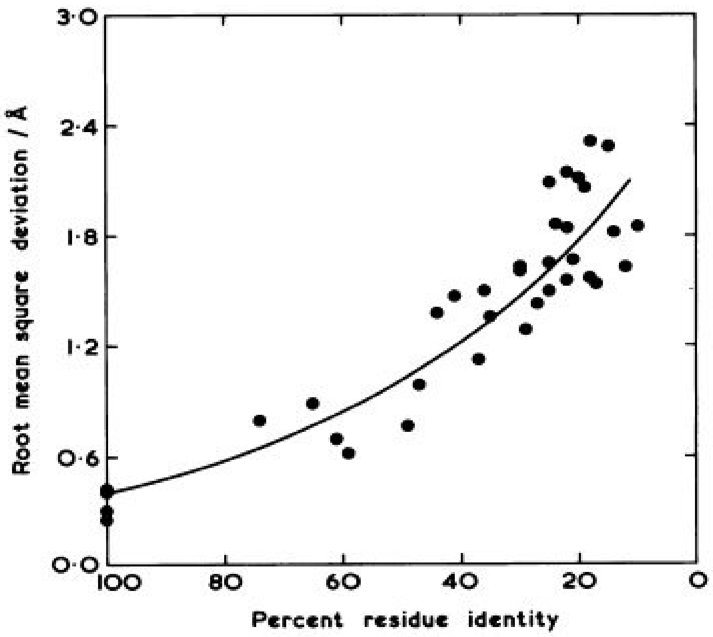
FINAL STRUCTURAL MODEL

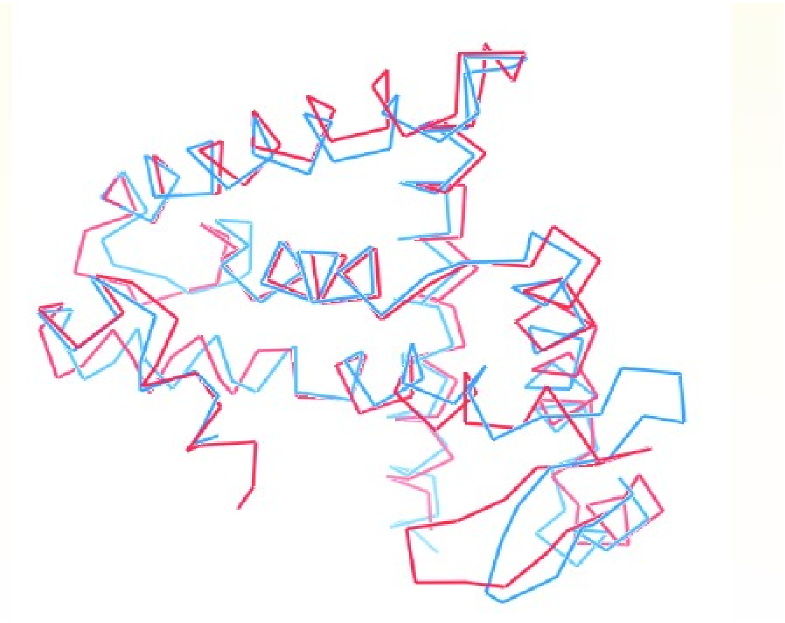
Modelling

RMSD:
Two structures being compared can be experimental or predicted models

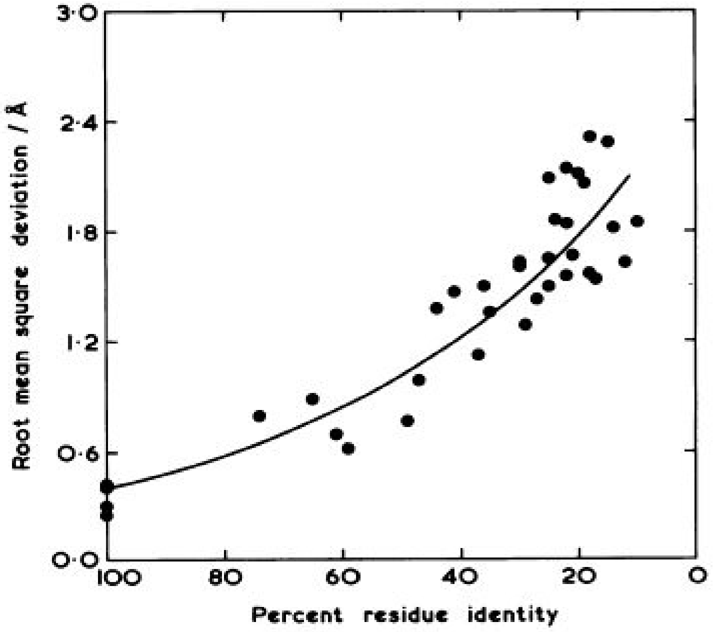
The root mean square deviation (RMSD) is the measure of the average distance between the backbones of superimposed proteins.

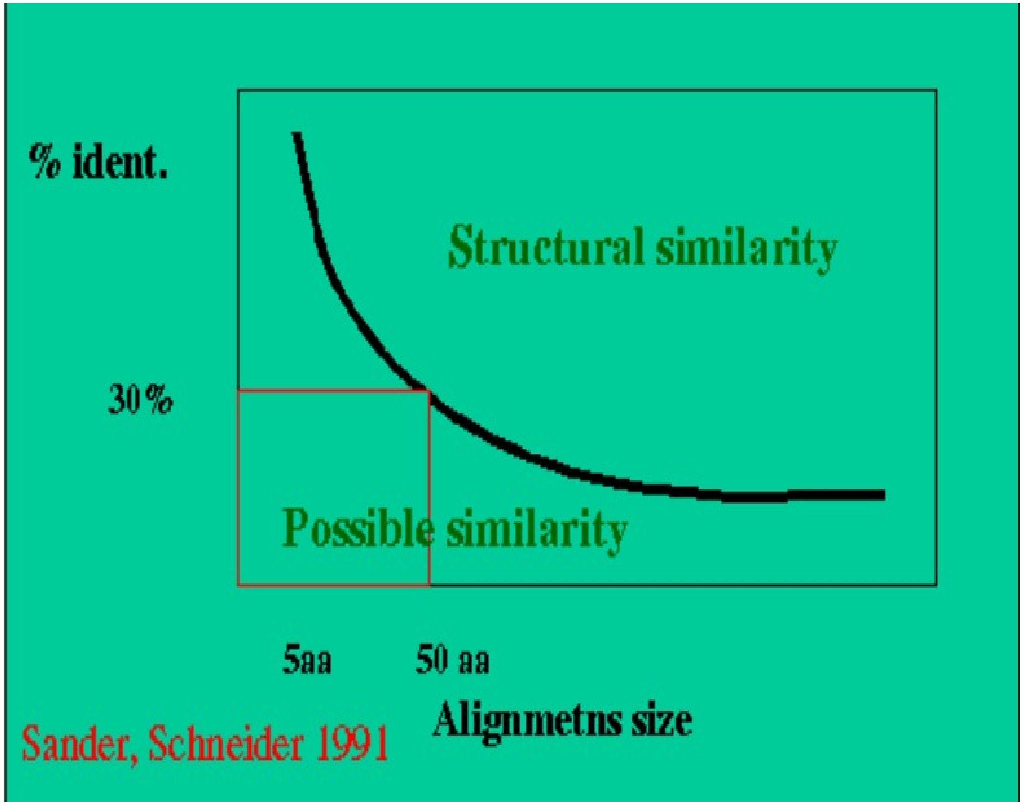
$$rmsd = \sqrt{\frac{\sum_{i=1}^{n}\left(x_{1,i} - x_{2,i}\right)^{\overset{=0.8}{2}}}{n}}$$





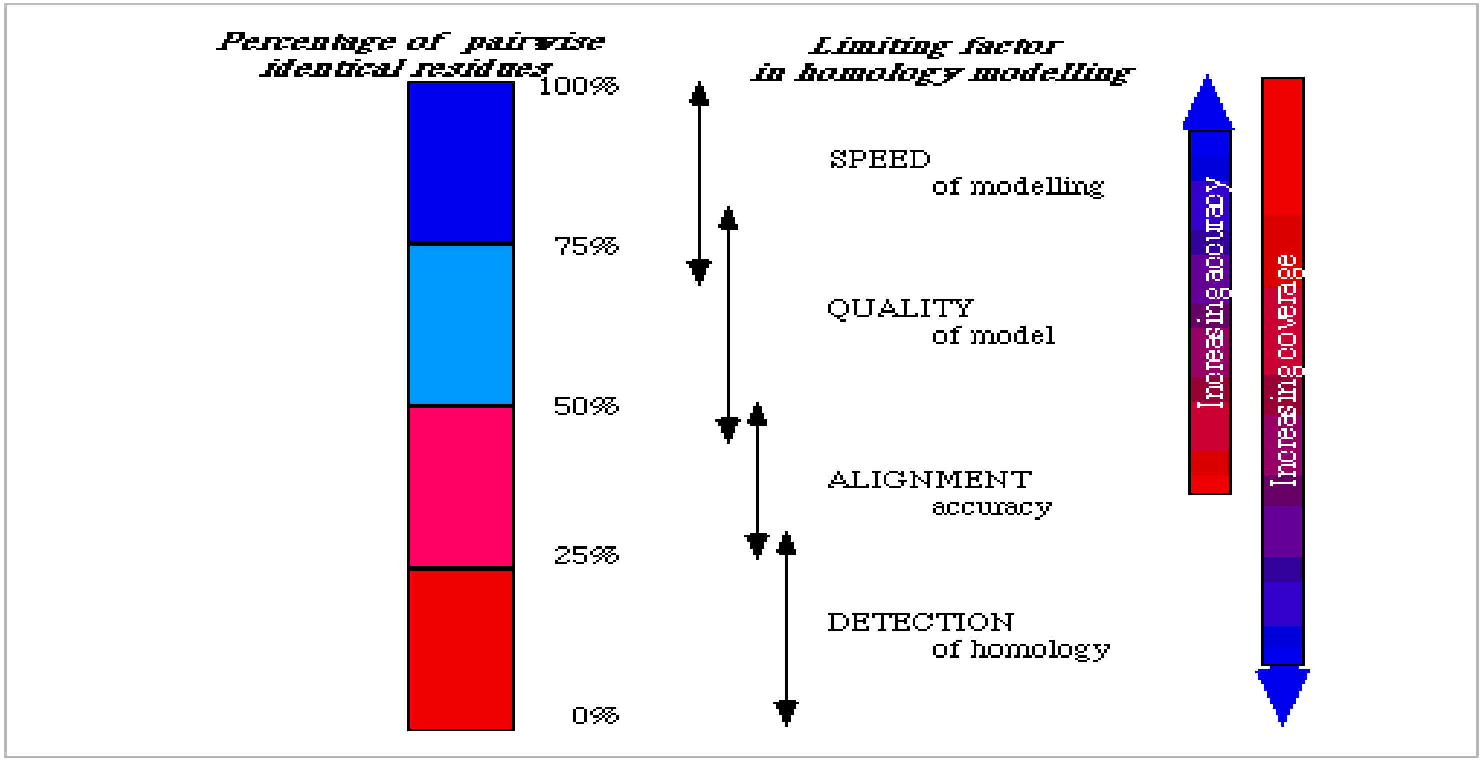
Modelling





Sander, Schneider 1991

Modelling



Accuracy of homology modelling is proportional to the level of pairwise sequence identity between the protein of unknown structure and its target of known structure. For high levels of identity, CPU time is the major constraint, for lower levels, loop regions become a problem (and thus the quality of the model). Below 40-50% sequence identity errors in the sequence alignment become fatal. Below 25-30% sequence identity, **fold recognition** (**threading**) techniques have to replace (or complement) the sequence alignment procedure.

Modelling

Detecting Templates with BLAST

The first step in predicting the structure of a protein is a pairwise BLAST search of the protein structural database (PDB).

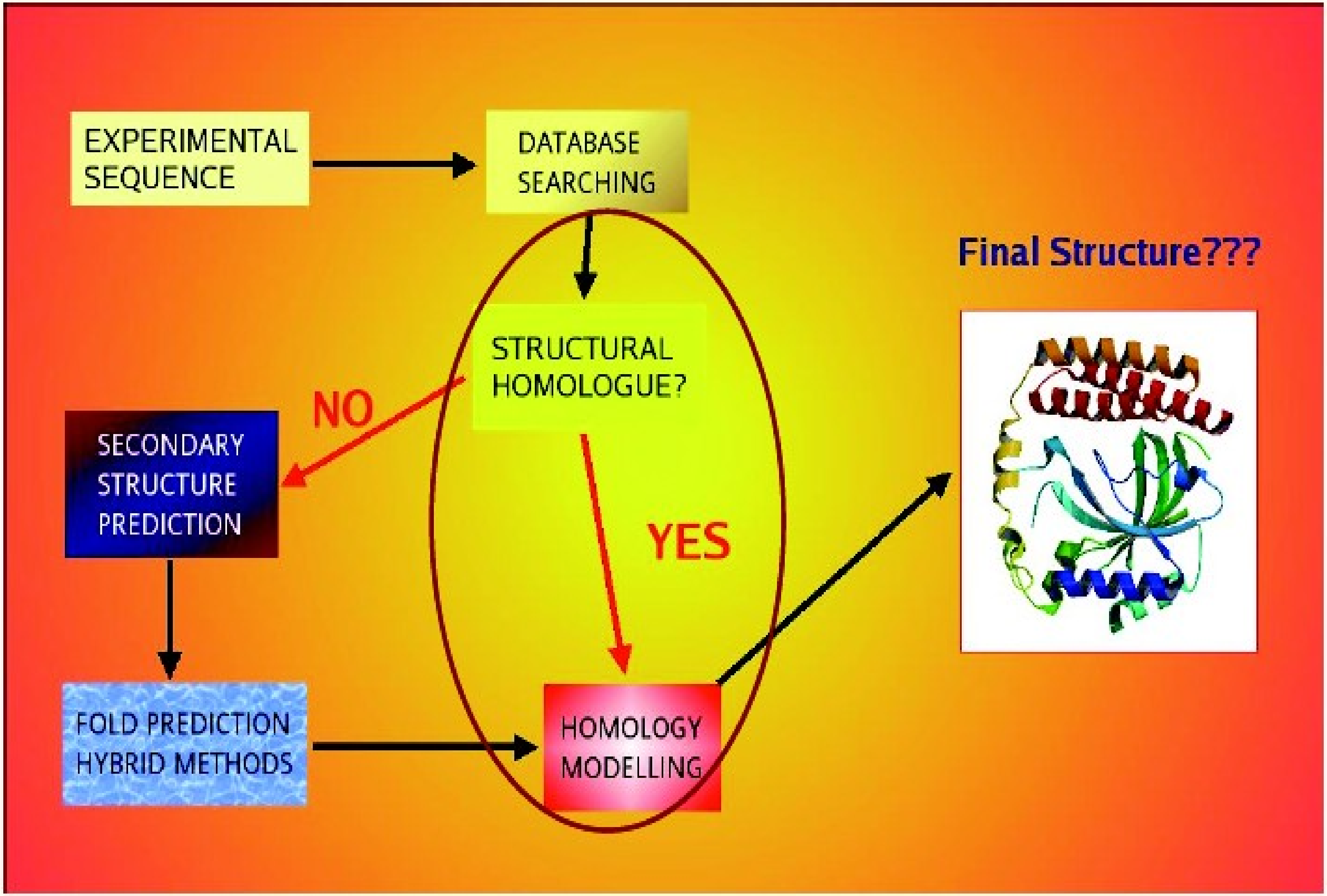
This is important for two reasons:

- The BLAST search will find any highly similar structures and in this case the alignment and modelling will usually be fairly easy.

- In addition a BLAST search can often give clues to the structure of the entire protein - is it likely to have a signal peptide, trans-membrane helices? How many domains might it have?

It is important to consider splitting the target into domains at this point. We can predict the structures of separate domains we are not good at orienting the domains.

Simplified structure prediction flow chart



EXPERIMENTAL SEQUENCE

DATABASE SEARCHING

STRUCTURAL HOMOLOGUE?

NO

YES

SECONDARY STRUCTURE PREDICTION

FOLD PREDICTION HYBRID METHODS

HOMOLOGY MODELLING

Final Structure???

Modelling

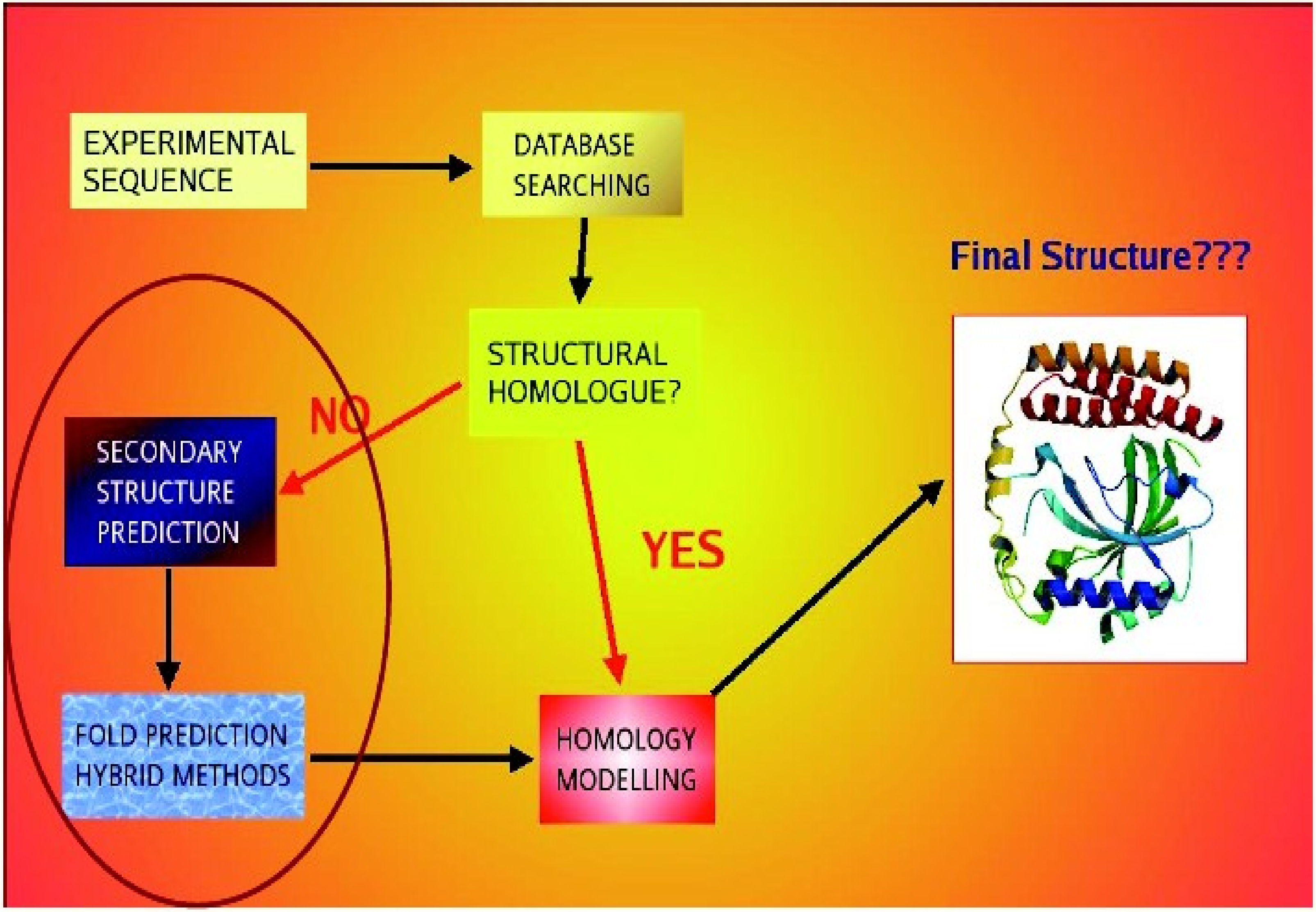
Homology Modeling

Lateral chains(rotamers)

Loops modeling

Evaluation of the model

Simplified structure prediction flow chart



If no similar PDB template exists ...

If no template can be found with BLAST, the structure prediction process is more complicated.

As a start we can use more complex sequence search methods (PSIBLAST, FFAS, HMMs).

If these methods fail we will have to use fold recognition methods to find a remotely related template.

If fold recognition methods fail to find a template, the protein can still be modelled ab initio.

However, with each step the reliability of the prediction decreases.

Fold Recognition

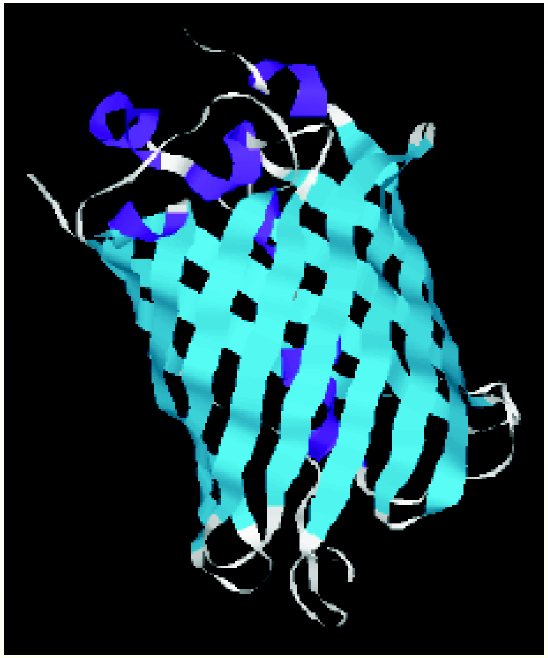
When fold recognition methods were first developed it was thought that they could detect analogous. proteins – those that were structurally similar but with no evolutionary relationship.

In fact most of these predictions were later shown to be homologous (have an evolutionary relationship) once advanced sequence comparison methods, such as PSI-BLAST, were developed.

Fold recognition methods are used because they allow you to find more distant templates.

Fold recognition techniques find templates that sequence based methods cannot because they use structural information as well as sequence similarity to evaluate templates.

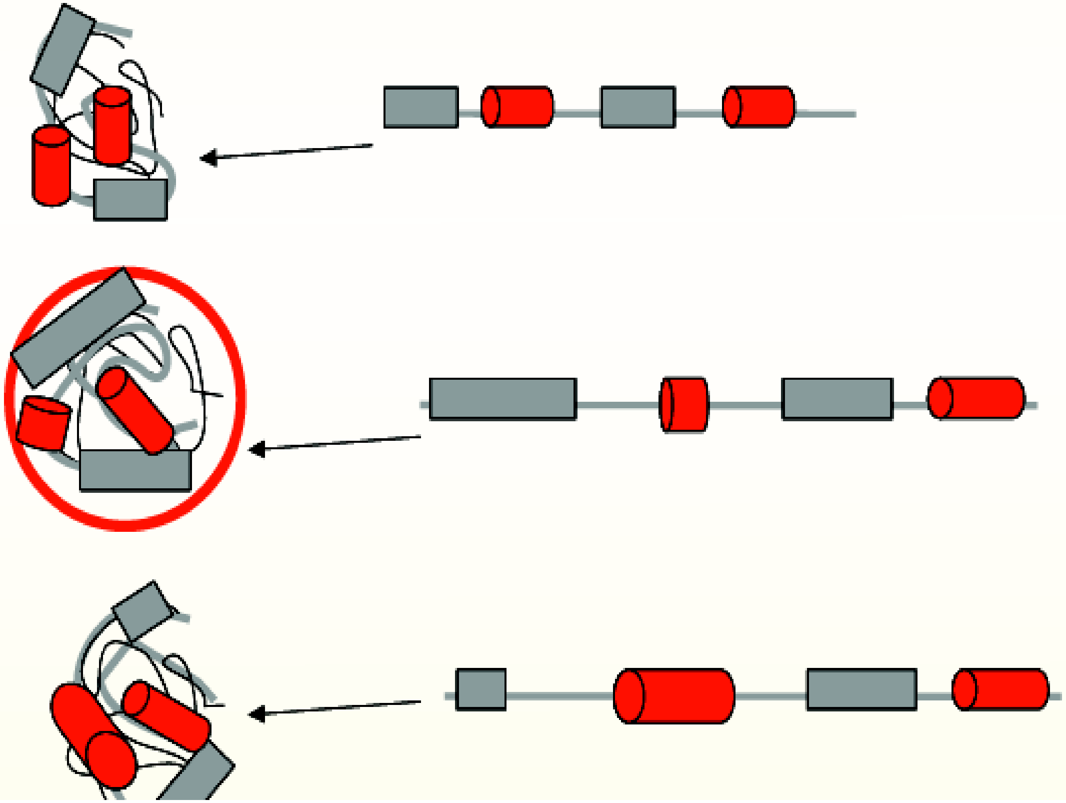
In fold recognition we are asking : "starting from all known protein structures, can I fit my sequence onto them?"



Fold recognition methods have built in fold libraries

Fold recognition methods work by superimposing the target onto a database of known 3D structures (folds) and evaluating the sequence-fold alignments.

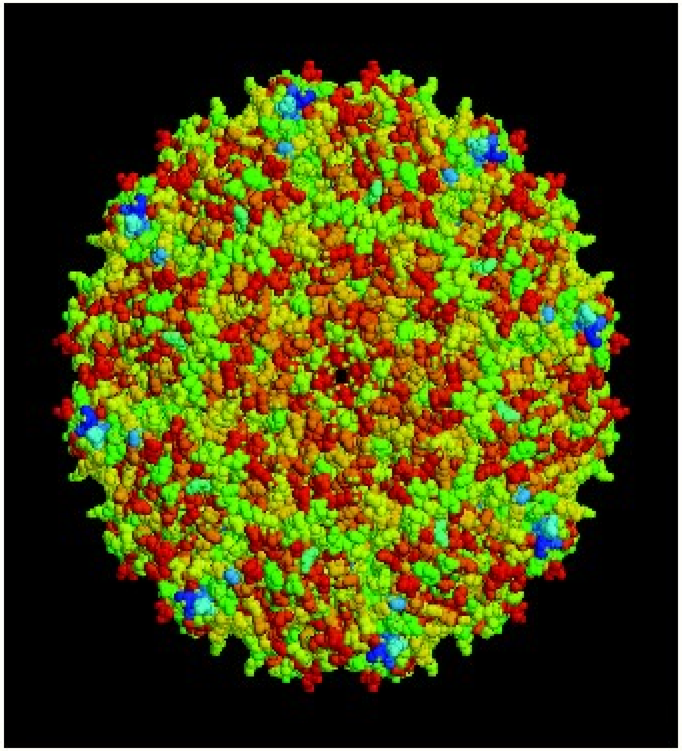
Each method has its own non-redundant database of folds to save calculation time.



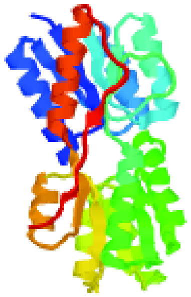
Scoring Functions

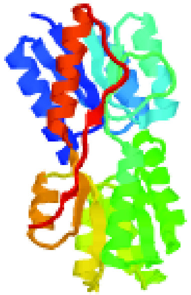
Scoring functions for evaluating the sequence-structure fit :

- Similarity between the known and predicted residue environments
- Coincidence of real and predicted secondary structure/accessibility
- Solvation energy
- Pair potentials
- Evolutionary information (from aligned structures and sequences)



Using Structural Environments Scores





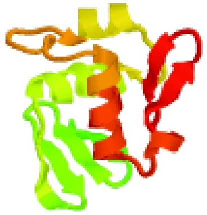
Scoring matrices are pre-generated for the probabilities of finding each of the twenty amino acids in each of the environment classes. These probabilities are calculated from databases of known structures.

Using these probabilities a 3D profile is created for each fold in the fold library. This 3D matrix defines the probability of finding each amino acid in each environment class.

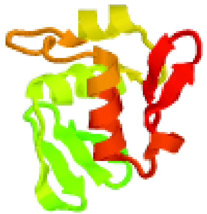
When the target sequence is aligned with the fold, a score can be calculated from the pre-generated 3D profile for each of the positions in the alignment.

The final environment score will be the sum of the probabilities for each residue.

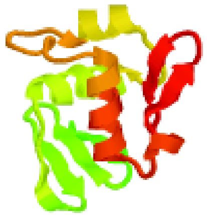
Solvation Energy



Solvation potential is a term used to describe the preference of an amino acid for a specific level of residue burial.

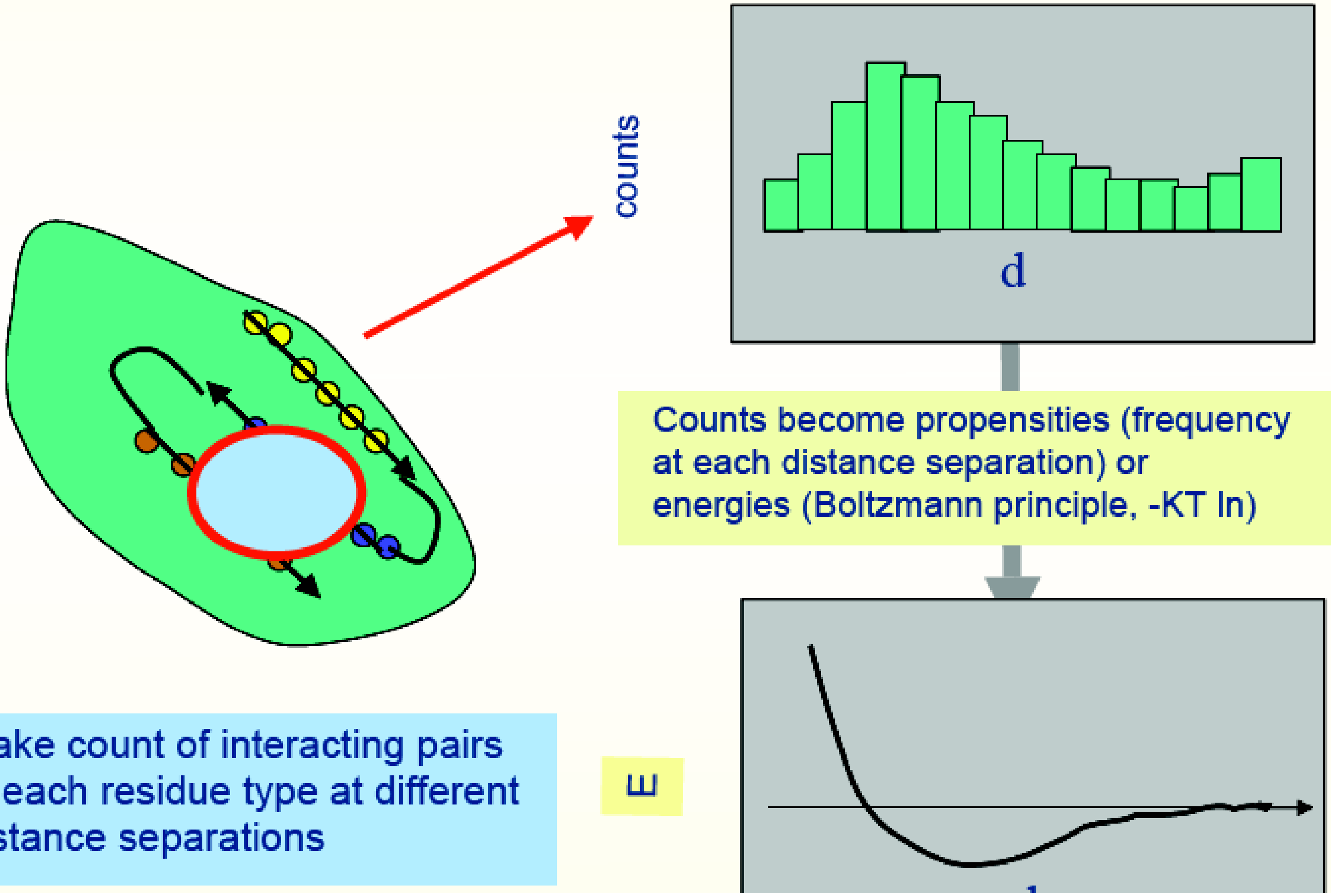


It is derived by comparing the frequency of occurrence of each amino acid at a specific degree of residue burial to the frequency of occurrence of all other amino acid types with this degree of burial.



The degree of burial of a residue is defined as the ratio between its solvent accessible surface area and its overall surface area.

Pair or contact potentials -
the tendency of residues to be in contact

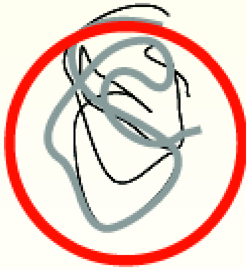


Counts become propensities (frequency at each distance separation) or energies (Boltzmann principle, -KT ln)

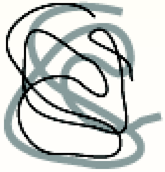
Make count of interacting pairs of each residue type at different distance separations

Fold Recognition Servers

3D-PSSM - www.sbg.bio.ic.ac.uk/~3dpssm/
Based on sequence profiles, solvatation potentials and secondary structure.

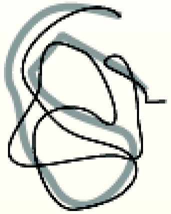


mGenTHREADER - www.psipred.net/
Combines profiles and sequence-structure alignments. A neural network-based jury system calculates the final score based on solvation and pair potentials.



RAPTOR - software.bioinformatics.uwaterloo.ca/~raptor/
Best-scoring server in CAFASP3 competition in 2002. ACE server (based on Raptor) best FR server in CASP6. You have to ask to use it first ...

SPARKS - http://sparks.informatics.iupui.edu/
Top servers in CASP 6. Sequence, secondary structure Profiles And Residue-level Knowledge-based Score for fold recognition



Consensus Fold Recognition

No one method can hope to correctly identify every fold.

Often the best predictions are when server predictions agree.

Human experts have recognised this, human experts usually use several different fold recognition methods and predict folds after evaluating all the results (not just the top hits) from a range of methods.

So why not produce an algorithm that mimics the human experts?

The first consensus server, Pcons, sent the target sequence to six publicly available fold recognition web servers.

Predictions were structurally superimposed and evaluated for their similarity. The best model was predicted from similarity to other predicted models.

Consensus Fold Recognition Servers

3D Jury - http://bioinfo.pl/meta/
3D Jury is a consensus predictor that utilizes the results of fold recognition servers, such as FFAS, 3D-PSSM, FUGUE and mGenTHREADER, and uses a jury system to select alignments and templates. Models are built with Modeller.

GeneSilico - http://genesilico.pl/meta/
A gateway to various methods for protein structure prediction. Domains are identified by HmmPfam, and there are several methods for secondary and tertiary structure (FR) prediction. Consensus predictions are made with the Pcons consensus server and you can also send a subset of alignments to the FRankenstein server.

Pcons - www.sbc.su.se/~arne/pcons/
Pcons was the first consensus server for fold recognition. It has been relaunched recently.

Ab initio and *de novo* protein modelling

- Rather than using previously solved structures, *ab initio* methods build 3D models from physical principles such as energy functions, or try to mimic protein folding.

- *Ab initio* methods work best on very small proteins. They require vast computational resources and the physical basis of protein structural stability and the necessary energy functions are not fully understood.

- *De novo* modelling is a form of *ab initio* modelling that uses protein fragments to build models that are evaluated by physical principles and statistical properties. *De novo* modelling has had some notable successes. Again this works best on smaller proteins.

De Novo Servers

ROBETTA - http://robetta.bakerlab.org/
ROBETTA makes both ab initio and template-based predictions. It detects fragments with BLAST, FFAS03, or 3DJury and uses fragment insertion and assembly.

FoldPRO -
http://www.igb.uci.edu/?page=tools&subPage=psss
A server that attempts to assemble fragments in a similar way to Robetta.

I-TASSER - http://zhang.bioinformatics.ku.edu/I-TASSER/
A de novo server developed by the successful Zhang group in Kansas – predicts small proteins, predictions take about a week.