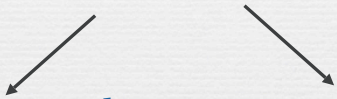
Structure prediction, fold recognition and homology modelling

Marjolein Thunnissen
Lund September 2009

Steps in protein modelling

Similarity search (BLAST)
Multiple alignment



3-D structure known No Structure known
Comparative Modelling Secondary structure prediction
Fold recognition
Ab initio (Rosetta)

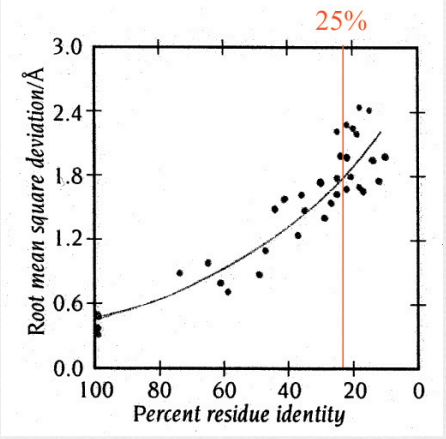
Structure Prediction

1) **Prediction of secondary structure.**
a) Method of Chou and Fasman
b) Neural networks
c) hydrophobicity plots
2) **Prediction of tertiary structure.**
a) Ab initio structure prediction
b) Threading
- 1D-3D profiles
- Knowledge based potentials
c) Homology modelling

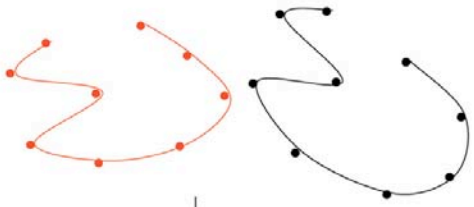
How does sequence identity correlate with structural similarity

Analysis by Chotia and Lesk (89)

- 100% sequence identity: rmsd = experimental error
- <25% (twilight zone), structures might be similar but can also be different
- Rigid body movements make rmsd bigger



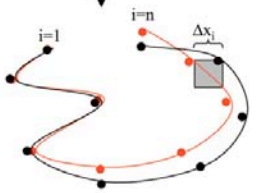
Root mean square deviation (rmsd)



$$rmsd = \sqrt{\frac{\sum_{i=1}^{n} (\Delta \mathbf{x}_i)^2}{n}}$$

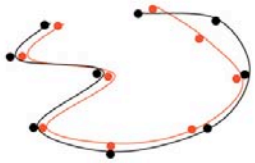
for atom pairs i=1 to n
$\Delta \mathbf{x}_i$ distance between two compared atoms in space

find starting superposition



rmsd=1.6 Å

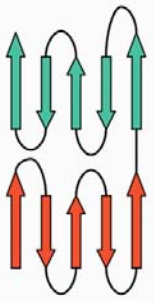
cyclic minimisation of rmsd



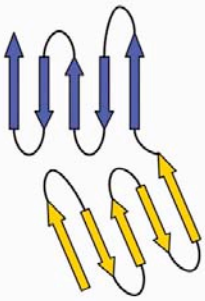
rmsd=1.2 Å

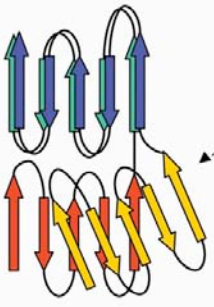
Rigid body movements increase the *rmsd*

Protein 1
(2 domains)

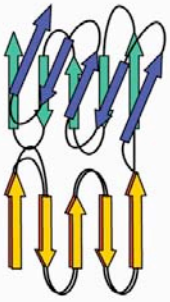


Protein 2
(2 domains)

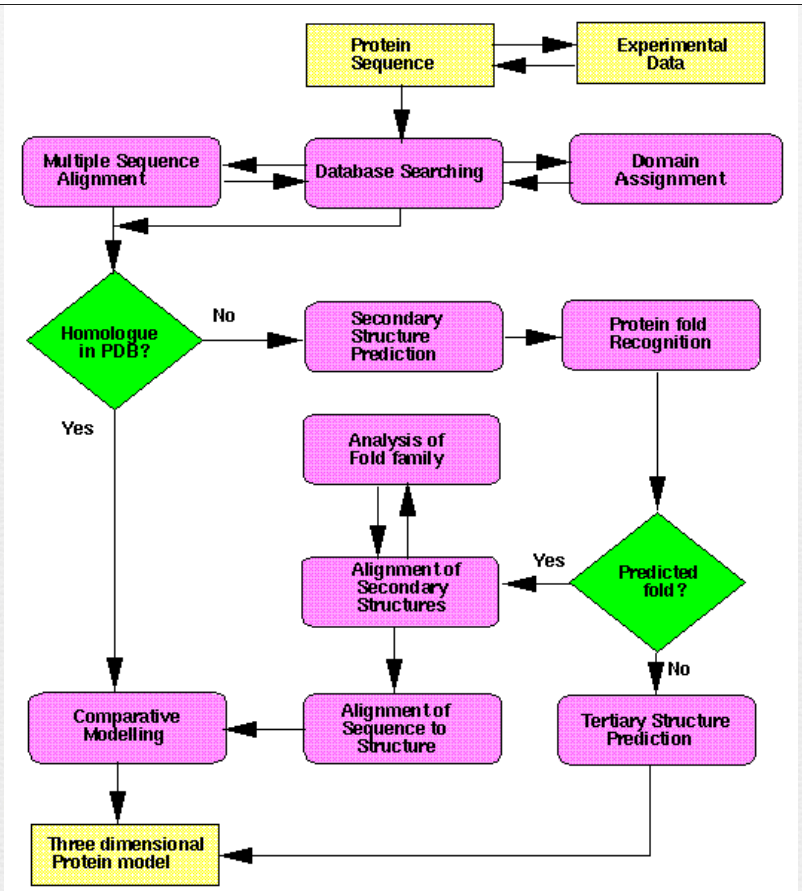




domain I: rmsd (low)
domain II: rmsd (high)



domain I: rmsd (high)
domain II: rmsd (low)



Secondary structure prediction:

Take the sequence and, using rules derived from known structures, predict the secondary structure that is most likely to be adopted by each residue

Why secondary structure prediction ?

- A major part of the general folding prediction problem.
- The first method of obtaining some structural information from a newly determined sequence. Rules governing α-helix and β-sheet structures provide guidelines for selecting specific mutations.
- Assignment of sec. str. can help to confirm structural and functional relationship between proteins when sequences homology is weak (used in threading experiments).
- Important in establishing alignments during model building by homology; the first step in attempts to generate 3D models

Some interesting facts 2nd structure predictions

- based on primary sequence only
- accuracy 64% -75%
- higher accuracy for α-helices than β– strands
- accuracy is dependent on protein family
- predictions of engineered proteins are less accurate

Methods:

- Statistical methods based on studies of databases of known protein structures from which structural propensities for all amino acids are calculated. However, these methods do not take into account physico-chemical knowledge about proteins.
- Physico-chemical methods (helical wheels, hydrophobicity profiles etc.).
- Hybrid methods combines the first two.

Structural Propensities

- Due to the size, shape and charge of the side chain, each amino acid may "fit" better in one type of secondary structure than another.
- Classic example: The rigidity and main chain angle of proline cannot be accommodated in an α-helical structure.

Examples of statistical methods:

- Two classical methods that use the statistical approach (previously determined propensities):
  - Chou-Fasman
  - Garnier-Osguthorpe-Robson

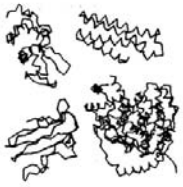
Chou-Fasman method

Has been one of the most popular methods.
Based on calculation of the propensity of each residue to form α-helix or β-strand.
Uses table of conformational parameters (propensities) determined primarily from measurements of secondary structure by CD spectroscopy.
Can result in ambiguity if a region has high propensities for both helix and sheet.

The Chou & Fassman method for secondary structure prediction

Chou & Fassman (1974) *Biochemistry* **13,** 211-221, 222-245

1. Probabilities for all amino acids to be either in α-helix, β-sheet or coil.
   20 * 3 probabilities
2. Set knowledge-based rules to apply probability tables for prediction



```
protein 1   AGPDFVILKRW...
            CCHHHHCCHHH...
protein 2   PGSAALHVIYI...
            CCBBBBCCHHH...
                 ⋮
protein 15  MVEDEHVILTAGF...
            CBBBBBBBBBHHH...
```

Amino acid	No. of residues $\#_{total}$	$\#_{helix}$	$\#_{sheet}$	$\#_{coil}$
Ala	228	119	38	71
Arg	78	22	12	44
.....	...	...	...	...
Val	181	74	51	56

Databank of 15 structures

Assignment of secondary structures to sequences

List of observations

From observations to probabilities

Aminoacid	observations $\#_{total}$	$\#_{helix}$	$\#_{sheet}$	$\#_{coil}$	frequency f_{helix}	f_{sheet}	f_{coil}	probability P_{helix}	P_{sheet}	P_{coil}
Ala	228	119	38	71	0.52	0.97	0.31	1.45	0.97	0.66
Arg	78	22	12	44	0.28	0.15	0.56	0.79	0.90	1.20
.....	...	...	...	...	...	...	...	...	...	...
Val	181	74	51	56	0.41	0.28	0.31	1.14	1.65	0.66
					average <f> 0.36	0.17	0.47	1.0	1.0	1.0

119 / 228 = 0.52

$\#^{Ala}_{helix} / \#^{Ala}_{total} = f^{Ala}_{helix}$

0.52 / 0.36 = 1.45

$f^{Ala}_{helix} / < f_{helix} > = P^{Ala}_{helix}$

Chou-Fasman propensities (partial table)

Chou-Fasman Parameters

Residue	Pα		Residue	Pβ		Residue	Pt
Glu	1.51	Hα	Val	1.70	Hβ	Asn	1.56
Met	1.45		Ile	1.60		Gly	1.56
Ala	1.42		Tyr	1.47		Pro	1.52
Leu	1.21		Phe	1.38		Asp	1.46
Lys	1.16	hα	Trp	1.37	hβ	Ser	1.43
Phe	1.13		Leu	1.30		Cys	1.19
Gln	1.11		Cys	1.19		Tyr	1.14
Trp	1.08		Thr	1.19		Lys	1.01
Ile	1.08		Gln	1.10		Gln	0.98
Val	1.06		Met	1.05		Thr	0.96
Asp	1.01	Iα	Arg	0.93	iβ	Trp	0.96
His	1.00		Asn	0.89		Arg	0.95
Arg	0.98	iα	His	0.87		His	0.95
Thr	0.83		Ala	0.83		Glu	0.74
Ser	0.77		Ser	0.75	bβ	Ala	0.66
Cys	0.70		Gly	0.75		Met	0.60
Tyr	0.69	bα	Lys	0.74		Phe	0.60
Asn	0.67		Pro	0.55	Bβ	Leu	0.59
Pro	0.57	Bα	Asp	0.54		Val	0.50
Gly	0.57		Glu	0.37		Ile	0.47

Chou-Fasman method:

- Calculation rules are somewhat arbitrary
- Example: Method for helix
- Search for nucleating region where 4 out of 6 a.a. have Pα> 1.03
- Extend until 4 consecutive a.a. have an average Pα < 1.00
- If region is at least 6 a.a. long, has an average Pα > 1.03, and average Pα > average Pβ, consider region to be helix

Garnier-Osguthorpe-Robson (GOR):

Build on Chou-Fasman *Pij* values

- Probability of an amino-acid to be in a specific structural elements depends on aminoacid type of residue itself and neighboring atoms
- evaluate each residue PLUS adjacent 8 N-terminal and 8 carboxyl-terminal residues
- sliding window of 17
- underpredicts β-strand regions
- GOR III method accuracy ~64%

Accuracy of predictions

- Both methods are only about 55-65% accurate
- A major reason is that while they consider the local context of each sequence element, they do not consider the global context of the sequence - the type of protein
  - The same amino acids may adopt a different configuration in a cytoplasmic protein than in a membrane protein

"Adaptive" methods

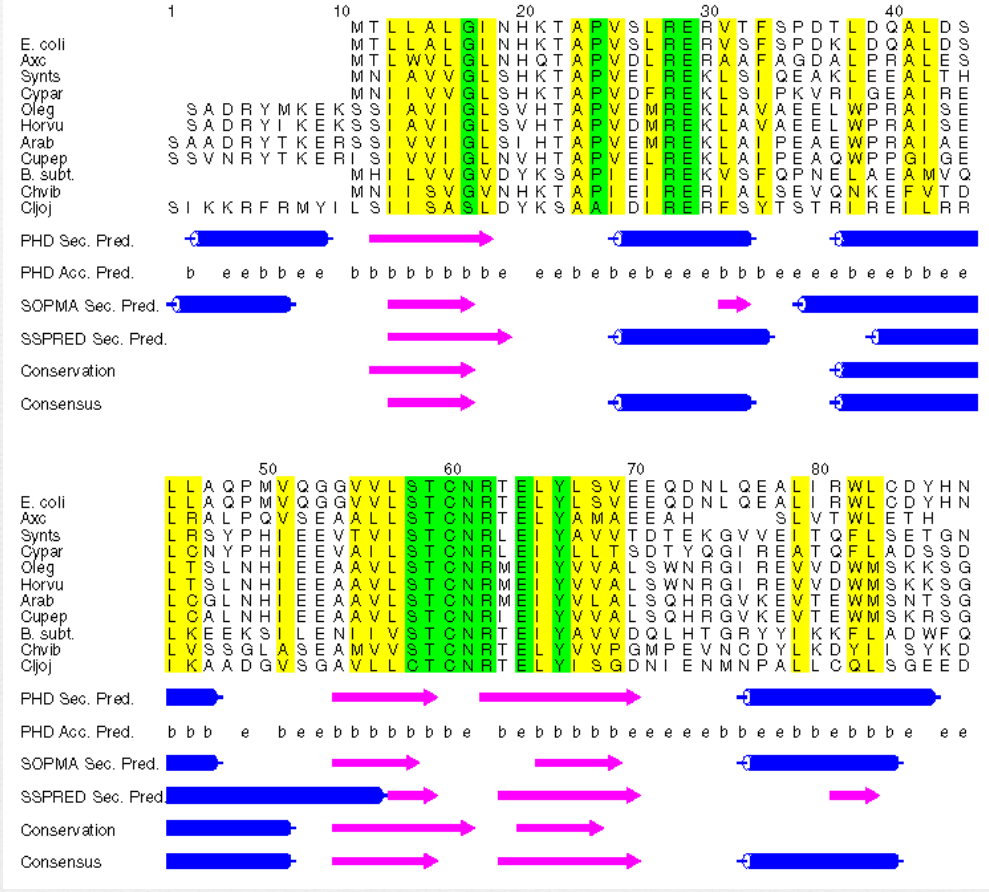
- Neural network methods - train network using sets of known proteins then use to predict for query sequence
  - **Nnpredict**
- Homology-based methods - predict structure using rules derived from proteins homologous to query sequence (multiple seq. alignment):
  - **SOPM**
  - **PHD**

Information from multiple sequence alignment increases accuracy:

- Positions of insertions and deletions suggest regions with surface loops.
- Conserved Gly or Pro suggest a β-turn.
- Hydrophobic residues conserved at i, i+2, i+4, and separated by unconserved hydrophilic residues suggest a surface β-strand.
- A short run of hydrophobic a.a. (4 residues) may suggest a buried β-strand, a longer stretch (20 residues) - a membrane spanning helix.
- Pairs of conserved hydrophobic a.a. separated by pairs of unconserved or hydrophilic residues suggest a helix with one face packed against the protein core. Likewise an i, i+3, i+4, i+7 pattern of conserved hydrophobic residues.

Guidelines:

- Since no method is the best, it is sensible to try different methods and compare the results - region where the methods agree are likely to be correctly predicted.
- Build a consensus !



Hydrophobicity profiles

- Tendency of a residue to occur at the surface or the interior can be described by a partition coefficient between these two phases (hydrophobicity scale).

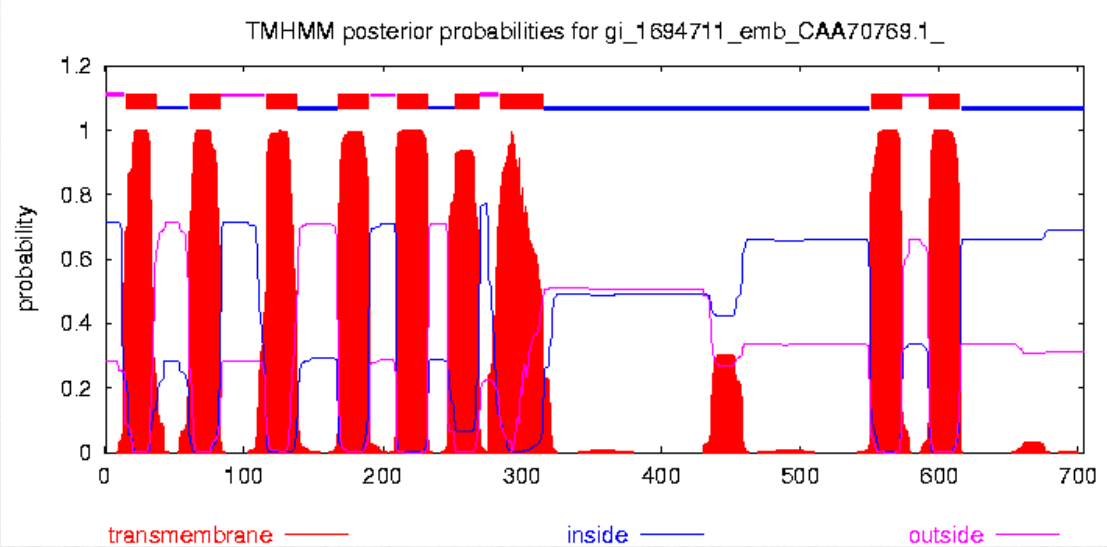
- The profile is computed by averaging the hydrophobicity within a moving window.

- The window size depends on the size of the structural element needed to be resolved. For secondary structure elements the window must be larger than a single turn (more than 4 residues), but smaller than a large segment (a helix, less than approx. 12 residues). For a membrane spanning region the window should have the size of the expected segment (approx.. 20 residues). Small windows are noisy (too many details).

Hydrophobicity profiles

- Can be used to predict turns, exterior and interior regions of a molecule.
- Can be applied to distinguish between membrane and soluble proteins.
- Mostly used to identify transmembrane helices in membrane proteins.

Prediction of transmembrane helices for Arabidopsis ferric reductase (FRO1) by a TMHMM (v. 2.0) at: http://www.cbs.dtu.dk/services/TMHMM-2.0/



Tertiary structure prediction

- Ab initio prediction of protein 3D structures is still problematic at present (Rosetta). However, proteins often adopt similar folds despite no significant sequence or functional similarity. Nature has apparently restricted the number of protein folds.

How to recognise a fold?

Even with no homologue of known 3D structure, it may be possible to find a suitable fold for your protein among known 3D structures using fold recognition methods.

Methods of protein fold recognition attempt to detect similarities between protein 3D structure that are not accompanied by significant sequence similarity - find a fold that is compatable with a particular sequence, or rather than predicting how a sequence will fold, predict how well a fold will fit a sequence.

Fold recognition on the Internet:

- Guide to predicting protein 3D structure (highly recommended): http://www.sbg.bio.ic.ac.uk/people/rob/CCP11BBS. The guide is from 1996, theory and flowchart information is still applicable today.
- JPRED (secondary structure prediction) http://www.compbio.dundee.ac.uk/www-jpred/
- Phyre http://www.sbg.bio.ic.ac.uk/~phyre/
- TOPITS/predictprotein (http://www.predictprotein.org)
- UCLA-DOE Structure Prediction Server (http://fold.doe-mbi.ucla.edu/) Only verification.

Threading or fold recognition

Task: detection of remote homologues behind the twilight zone

Method

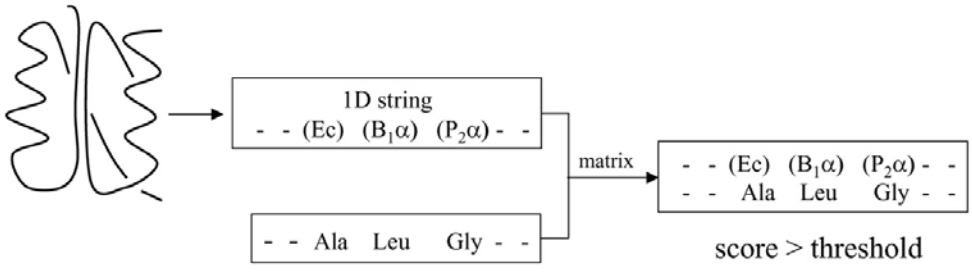
1) Database of folds
2) Dock side chains according to query sequence onto a structural scaffold from a fold library
3) Validate structure/sequence match by energy calculation
   -1D-3D profiles
   -knowledge based potentials
   atomic force fields

Applications: Threading: Fold database + sequence
Reverse Threading: Sequence database + fold

1D - 3D Profiles

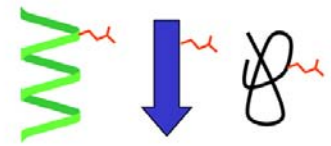
Bowie, Lüthy & Eisenberg, *Science* **253**, 164 (1991)

- compare sequence (1D) with structure (3D)
- convert 3D-structure into a 1D-string of environment classes, compare 1D-sequence with 1D-string using a scoring matrix
- if (score > threshold) than the sequence will fold into this structure



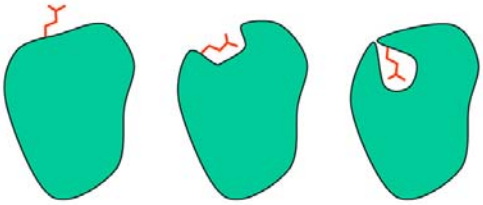
Three properties define the class of a given residue (χ)

1. property: secondary structure



α-helix β-sheet coil

2. property: buried surface area

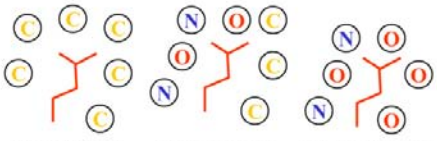


A< 40 Å² exposed (e)

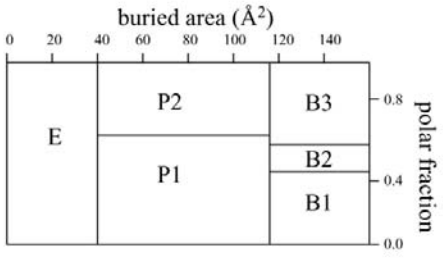
40 Å² < A < 114 Å² partial (p)

A>114 Å² buried (b)

3. property: fraction of polar atoms

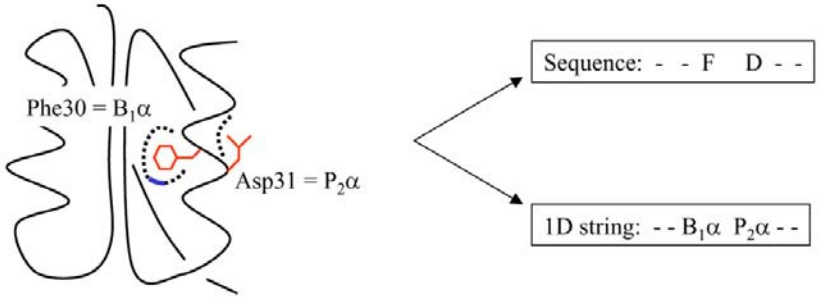


0-45% polar 45-58% polar 58-100% polar

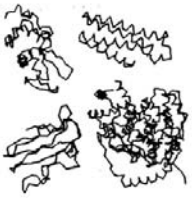


- Combination of properties 2 & 3 gives 6 classes (E, P1, P2, B1, B2, B3)
- Combination of 6 classes with property 1 gives 18 classes (Eα, Eβ, Ec, P1α, P1β, ... B3β, B3c)

Any protein structure can be converted into a 1D-string and a 1D-sequence



The 1D-sequence/1D-string scoring matrix



struture 1: - $B_1\alpha$ $P_2\alpha$ Ec $B_2\alpha$ -
- A P G E -

struture 2: - $B_2\alpha$ $P_2\beta$ $B_2\alpha$ Ec -
- Y S V V -

⋮

struture 16: - $B_2\beta$ $B_2\alpha$ Ec $B_2\alpha$ -
- L L D E -

find homologous sequences

- $B_1\alpha$ $P_2\alpha$ Ec $B_2\alpha$ -
- A P G E -
- A S G D -
- A S G E -

- $B_2\alpha$ $P_2\beta$ $B_2\alpha$ Ec -
- Y V V D -
- F V L G -
- F V I - -

⋮

Database of 16 structures

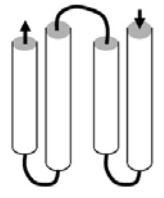
Database of 16 string/sequence pairs

Database of 16 string/alignment pairs

observations
frequencies
probabilities

Environment class	W	F	Y	L	I	V	M	A	G	P	C	T	S	Q	N	E	D	H	K	R
$B_1\alpha$	1.00	1.32	0.16	1.27	1.17	0.66	1.28	-0.66	-2.53	-1.16	-0.73	-1.29	-2.73	-1.06	-1.93	-1.74	-1.97	-0.34	-1.82	-1.67
$B_1\beta$	1.17	0.86	0.07	1.13	1.47	1.09	0.55	-0.79	-2.02	-0.94	-0.22	-1.12	-2.01	-1.67	-1.42	-1.93	-2.56	-1.91	-2.66	-1.16
B_1	1.05	1.45	0.17	1.10	1.11	1.02	0.98	-0.91	-1.92	0.26	-1.22	-1.53	-2.81	-1.17	-2.42	-2.52	-1.76	-1.12	-2.39	-2.18
$B_2\alpha$	0.50	0.90	0.55	1.01	0.63	0.68	1.13	-0.69	-1.49	-2.21	-0.10	-1.50	-1.47	-0.23	-0.61	-0.71	-1.62	0.23	-0.78	0.06
$B_2\beta$	0.01	1.18	1.06	0.78	1.31	1.08	0.64	-1.55	-2.26	-0.49	-0.87	-2.27	-1.77	-1.22	-2.07	-1.07	-1.41	-0.77	-1.14	-0.20
B_2	1.02	1.05	1.12	0.84	0.81	0.60	0.90	-0.66	-1.66	0.19	-0.05	-0.76	-1.17	-0.76	-0.66	-1.35	-1.28	0.46	-2.54	-0.80
$B_3\alpha$	0.92	-0.03	0.56	0.15	0.04	-0.02	0.89	-0.57	-1.86	-0.58	-1.56	-0.57	-0.96	0.22	-0.06	0.08	-0.50	0.73	0.43	0.96
$B_3\beta$	0.75	0.81	1.30	0.18	0.54	0.56	-0.57	-0.93	-1.93	-0.34	-0.54	-0.44	-0.74	0.21	-0.24	-0.14	-0.86	0.82	-0.53	0.13
B_3	1.07	0.70	1.13	0.35	-0.17	-0.03	0.23	-0.86	-0.88	-0.13	-1.20	-0.53	-0.54	0.05	0.04	-0.38	-1.05	1.01	0.10	0.68
$P_1\alpha$	-1.35	-0.82	-0.59	-0.52	-0.24	0.10	-0.03	0.73	-0.49	-0.25	0.85	0.31	0.34	-0.14	-0.54	-0.17	-0.25	-0.53	-0.21	-0.28
$P_1\beta$	0.36	-0.48	0.17	-1.03	0.20	0.48	-0.27	0.84	-0.82	-0.55	1.49	0.93	0.33	-0.27	-1.32	-0.73	-1.07	-0.42	-1.21	-0.77
P_1	-1.26	-1.20	-1.21	-0.52	-0.23	-0.01	-1.19	0.46	-0.24	0.66	1.35	0.56	0.49	-0.52	-0.13	-0.81	0.38	-1.12	-0.74	-1.29
$P_2\alpha$	-1.14	-1.43	-0.79	-0.35	-0.54	-0.48	-0.45	0.06	-0.50	-0.28	-0.93	-0.05	-0.18	0.55	-0.05	0.56	0.28	0.06	0.81	0.50
$P_2\beta$	-0.79	-0.54	-0.84	-1.20	-0.93	0.13	-0.72	-0.55	-0.96	-1.29	-0.57	0.84	0.59	-0.06	-0.16	0.32	0.19	0.87	0.59	0.10
P_2	-0.82	-0.96	-0.51	-0.70	-1.09	-0.88	-0.89	-0.15	-0.40	0.44	-0.60	0.06	0.26	0.27	0.50	0.27	0.49	0.13	0.44	0.30
E α	-1.35	-2.20	-2.10	-1.58	-2.78	-1.10	-2.72	0.46	0.68	0.04	-0.44	-0.17	0.15	0.36	0.28	0.59	0.44	-0.19	0.13	-0.34
E β	0.84	-0.90	0.30	-1.66	-1.47	-1.74	-0.68	0.06	1.46	-0.96	-0.24	0.14	0.65	-0.19	-0.06	-0.16	-0.78	-0.83	-0.52	-0.49
E	-2.14	-1.90	-0.94	-1.19	-1.61	-0.91	-1.67	0.12	1.13	0.20	-0.46	0.12	0.32	-0.03	0.41	0.00	0.22	-0.25	-0.14	-0.32

Any structure can be converted into a 1D-3D profile

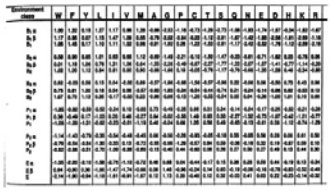


myoglobin structure

Ala1 Asn2 Glu3 Ile4 ....

Ec B_2c Eα $P_2\alpha$ ....

scoring table



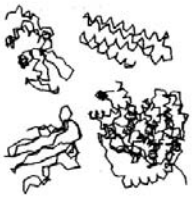
Residue	Environment class	(100 * S_i) A	C	D	...	V	W	Y	Gap penalty Open	Ext
Ala1	Ec	12	-46	22	...	-91	-214	-94	2	0.02
Asn2	B_2c	-66	-5	-128	...	60	102	112	2	0.02
Glu3	Eα	46	-44	44	...	-110	-135	-210	200	200
Ile4	$P_2\alpha$	6	-93	28	...	-48	-114	-79	200	200
....										

sequence/string pair of myoglobin

1D-3D profile of myoglobin

Applications for 1D-3D profiles

1. **Threading:** What is the fold for a give sequence ?
   - convert all structures into 1D-3D profiles
   - compare the query sequence with all 1D-3D profiles
     calculate the Z-score for every sequence/profile pair



Database of unique folds

Fold 1	A	C	D	....	V	W	Y
Ec	12	-46	22	...	-91	-214	-94
B_2c	-66	-5	-128	...	60	102	112
Eα	46	-44	44	...	-110	-135	-210
$P_2α$	6	-93	28	...	-48	-114	-79

S = 22 - 66 - 210 - 48 = -302

novel sequence DAYV....

Fold 2	A	C	D	....	V	W	Y
Ec	12	-46	22	...	-91	-214	-94
Eα	46	-44	44	...	-110	-135	-210
$P_2α$	6	-93	28	...	-48	-114	-79
$B_2α$	-69	-10	-162	...	68	50	85

S = 22 + 46 - 79 + 68 = 57

Database of 1D-3D profiles

2. **Inverse threading:** Which sequence matches a given fold?
   - convert the query fold into a 1D-3D profile
   - compare the query profile against all protein sequences

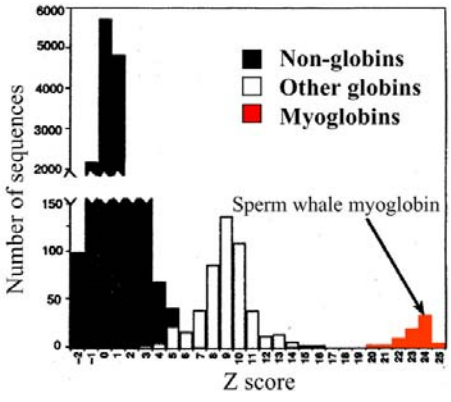
Inverse threading example: Which sequences fit the sperm whale myoglobin fold?

Input:
sperm whale myoglobin 1D-3D profile
protein sequence database (n: #entries)

$$\langle S \rangle = \sum_{i=1}^{n} S_i / n$$ average S-score

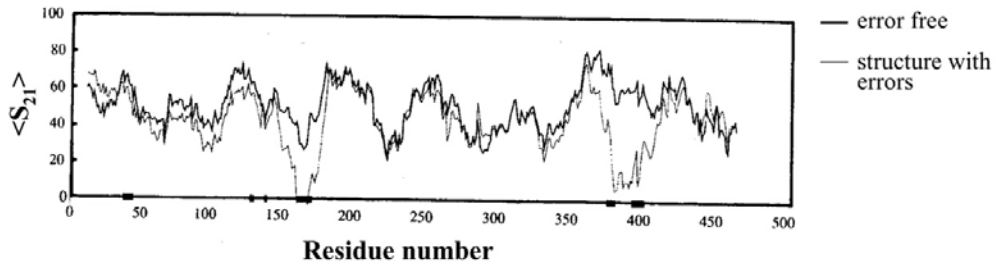
$$\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\langle S \rangle - S_i)^2}$$ standard deviation

$$Z_i = (S_i - \langle S \rangle) / \sigma$$ Z-score



Example 2: Verification of experimentally determined protein structures

$\langle S_{21} \rangle$ Average S-score for a window of 21 residues



$\langle S_{21} \rangle$ high	residues are in their preferred environment
$\langle S_{21} \rangle$ low	residues are in a non-natural environment (Glu in hydrophobic core)

2D to 3D: Comparative (homology) modelling

- A prediction of 3D structure is most successful when a structures of one or more homologues are known.
- Homologous proteins always contain a core region where the general fold of the chain is very similar.

However:

Even in core regions side-chain conformations may vary.

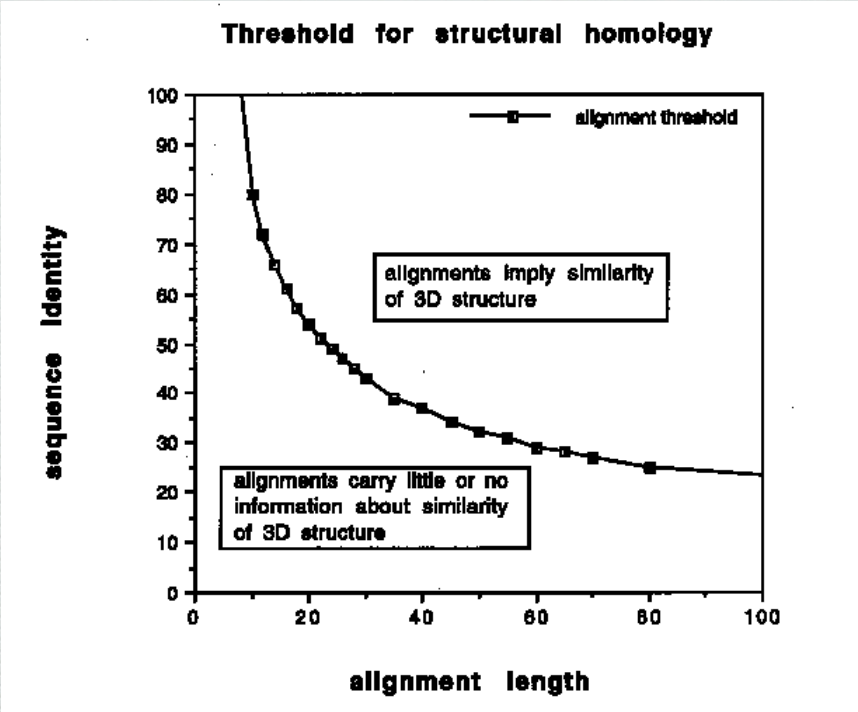
Modelling:

The modelling process can be subdivided into 9 stages:

1. template recognition;
2. alignment;
3. alignment correction;
4. backbone generation;
5. generation of canonical loops (data based);
6. side chain generation plus optimisation;
7. *ab initio* loop building (energy based);
8. overall model optimisation (energy minimisation);
9. model verification with optional repeat of previous steps.

What can be modelled ?

Homology threshold for structurally reliable alignments as a function of alignment length.



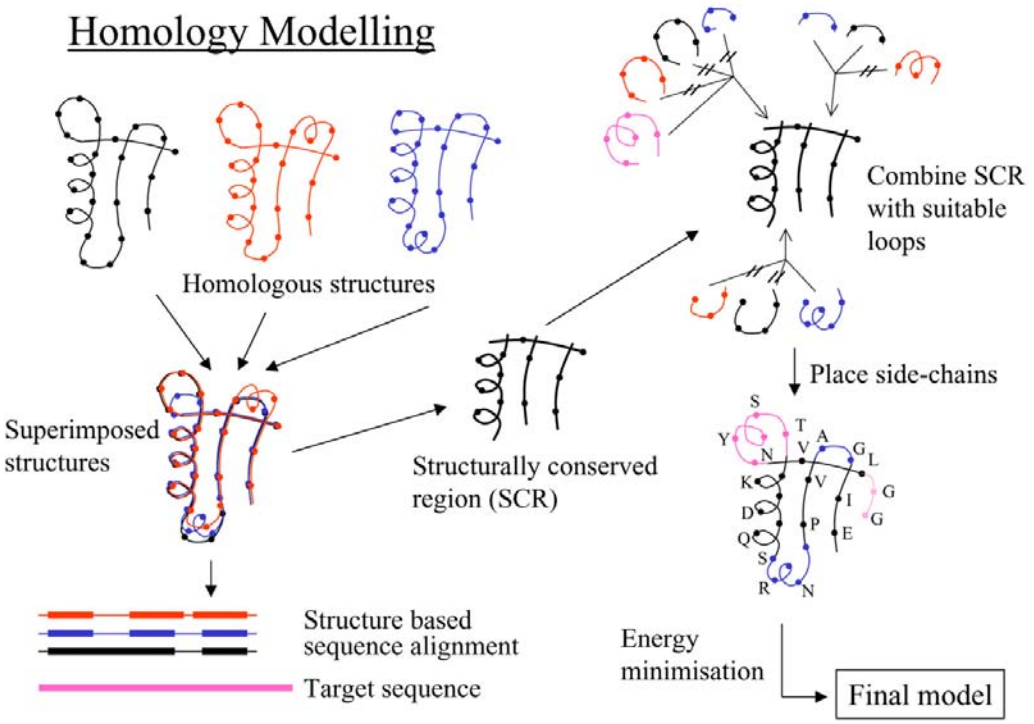
Some general rules:

- Sequence alignments, particularly those involving proteins having low percent sequence identities can be inaccurate. Thus, a model built using the alignment will be wrong in some places. Look over the alignment carefully before building a model.
- The quality of protein models built using homology to a template protein structure is normally determined by the RMS errors in models of proteins of which the structure is known.
- Visual checking of the model is important: check the Ramachandran plot and the energy of your model in SwissPdbviewer, hydrophobic residues should be buried, polar and charged exposed, charged residues avoid having hydrophobic neighbours (Asp-Leu), might help to build a model from another homologous sequence and compare the results, check against secondary structure prediction,

More rules:

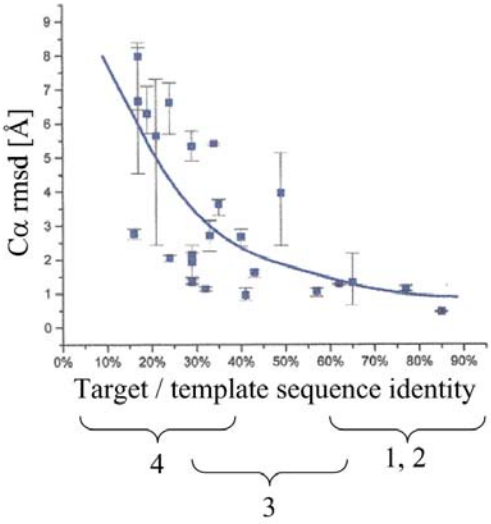
- The observed residue burial or exposure should be compared to residue burial or exposure in the model.
- The conservation of residue properties in experimental structure and model.
- Whether or not the side chains on the core beta-strands pointed in towards the barrel or out towards the helices
- The hydrogen bonding pattern of the beta-strands and helices should be checked.

Homology Modelling



Accuracy of homology modelling

Critical assessment of structure prediction (CASP) *Proteins* Supp. **3**, (1999)



Errors in homology modelling

1. side-chain packing
2. main-chain distortions of fragments (loops, helices, etc.)
3. main-chain distortions of newly modelled fragments
4. miss-alignments

Some results:

- 63% of sequences sharing 40-49% identity with template yield a model deviating by less than 3 Å from the control structure.
- The number increases to 79% for seq. Identities ranging from 50 to 59%.
- Below 30 % the accuracy rapidly degrades.
- The most reliable part of the model is the portion it shares with the template, while loop and other non-

Some links

- List with several different software packages:

  http://en.wikipedia.org/wiki/Protein_structure_prediction_software
- 3D JIGSAW
- http://www.bmm.icnet.uk/servers/3djigsaw/
- SwissModel
- http://www.expasy.org/swissmod/SWISS-MODEL.html

CASP does checks on reliability of the different software packages.

Model Validation

- Does it look like a protein
  - Is chemistry ok (eg. Hydrophobic in core)
  - Geometry ok (eg Phi-Psi angles, bond lengths)
  - Amino acid environment A "correct" model can be completely wrong
  - Atom packing correct
- Accuracy (if we know the answer)
  - Rmsd
  - Fraction of correct no of modelled residues
- Use validation programs

Evaluation of model accuracy

- Many programs come from validation programs for experimentally determined structures
- Check for proper protein stereochemistry
  - Procheck (http://biotech.embl-ebi.ac.uk:8400/)
    - Ramachandran plot, bond length etc
  - Whatif/Whatcheck (http://swift.cmbi.kun.nl/swift/whatcheck/)
    - Packing quality
  - MolProbity (http://molprobity.biochem.duke.edu/)
- Check sequence vs structure
  - Verify3D (http://www.doe-mbi.ucla.edu/Services/Verify_3D/)