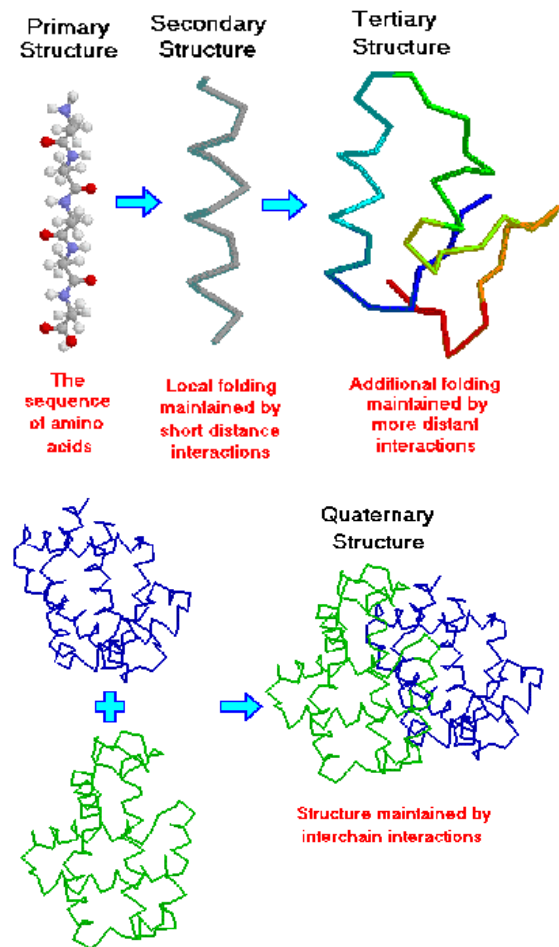


Protein Structure Prediction through web-based tools

- Domain/Motif Databases eg. SMART, Pfam, NCBI's CDD
- Protein databank – Coordinates of solved structures
- Secondary structure prediction tools eg. PHD, PSIPRED, SSPro etc
- Homology modeling – automated and interactive eg. Swiss model, 3D-Jigsaw
- Fold recognition or threading alignments eg. 123D+, Fugue, 3D-PSSM
- Model Evaluation tools eg. Verify 3D
- Structure superposition tools eg. CE

Protein Structure

- Proteins fold in three dimensions
- Protein structure is organized hierarchically from a *primary* to *quaternary* level.
 - **Primary structure** is the sequence of residues in the polypeptide chain.
 - **Secondary structure** is a local regularly occurring structure in proteins and is mainly formed through hydrogen bonds between backbone atoms. So-called random coils, loops or turns don't have a stable secondary structure.
 - **Tertiary structure** describes the packing of alpha-helices, beta-sheets and random coils with respect to each other on the level of one whole polypeptide chain.
 - **Quaternary structure** only exists if there is more than one polypeptide chain present in a complex protein. Then quaternary structure describes the spatial organization of the chains.



Secondary structure

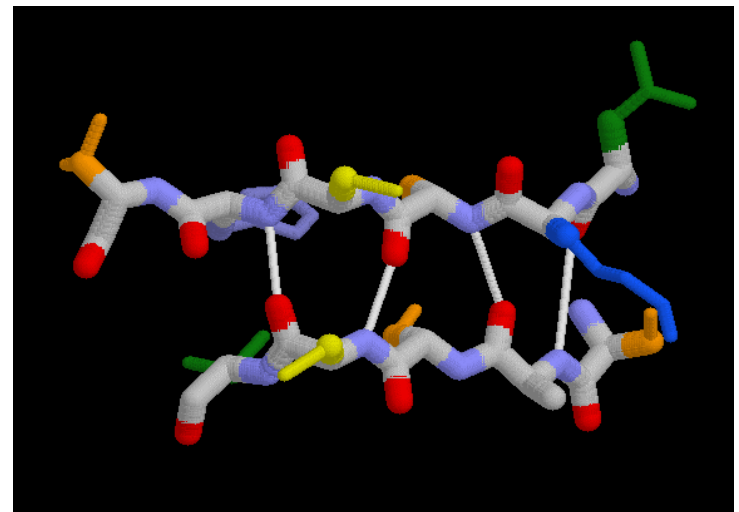
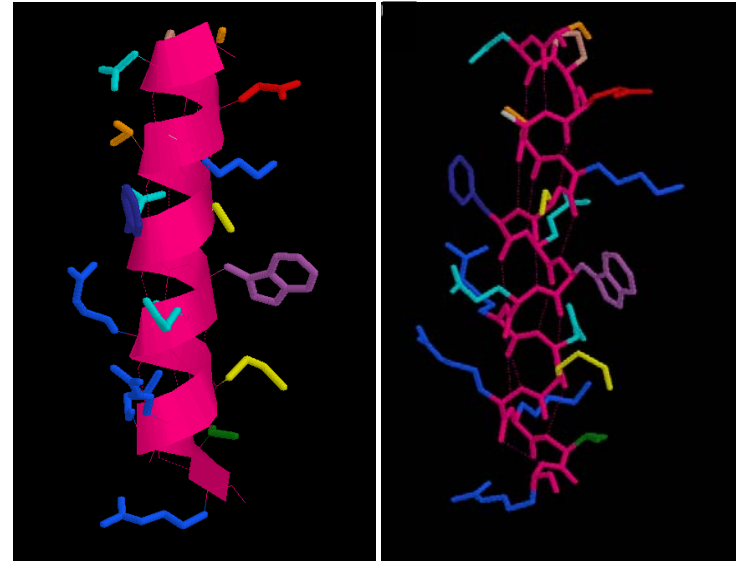
- There are two types of stable secondary structures

Alpha helix

- The backbone of an alpha helix is arranged in a spiral similar to that seen on a cork screw; the side-chains stick more-or-less straight out from the backbone.
- The alpha helix is stabilized by hydrogen bonds between the carbonyl oxygen of one amino acid and the backbone nitrogen of a second amino acid located four positions away.

Beta sheet

- The backbone of an beta sheet is arranged in zigzag or pleated fashion; the side-chains stick from the backbone on each side of the sheet. A minimum of two strands is required to define a beta sheet; many beta sheets have more.
- The beta sheet is stabilized by hydrogen bonds between the carbonyl oxygen of an amino acid in one strand and the backbone nitrogen of a second amino acid in another strand. Beta sheets can be either parallel or anti-parallel.



Why predict the secondary structure of a given amino acid sequence?

- Useful first step in understanding how the amino acid sequence of a protein determines the native state
- Important in establishing alignments during model building by homology
- The assignment of secondary structure can help confirm a structural and functional relationship between proteins when there is a weak sequence relationship.
- To select specific mutants for the design of novel proteins

Some popular programs

	Q ₃	Website
Prof	77.0	http://www.aber.ac.uk/~phiwww/prof/index.html
Psipred	76.6	http://bioinf.cs.ucl.ac.uk/psipred/
SSPro	76.3	http://www.igb.uci.edu/tools/scratch/
Jpred2	75.2	http://www.compbio.dundee.ac.uk/~www-jpred/submit.html
PHD	75.1	http://www.predictprotein.org/

Q₃: Accuracy of the prediction in %

Psipred

- Pioneering program which introduced the use of iterated PSI-BLAST searches in a automated method
- Based on a simplified neural networks algorithm
- High degree of prediction accuracy (up to ~77%)
- Prediction method
 - Generation of a sequence profile
 - Prediction of initial secondary structure
 - Filtering of the predicted structure

Website: <http://bioinf.cs.ucl.ac.uk/psipred/>

Output from Psipred

PSIPRED PREDICTION RESULTS

Key

Conf: Confidence (0=low, 9=high)
 Pred: Predicted secondary structure (H=helix, E=strand, C=coil)
 AA: Target sequence

PSIPRED HFORMAT (PSIPRED V2.3 by David Jones)

Conf: 95231005888845323343133445442133325446773430100057540565478
 Pred: CCCCCECCCCCCCCCHHHHHHHHHHHHHHHHHHHHCCCEEECCCHHHCCCCCEEECC
 AA: RRLSDIVYPNMKSPLAKCIRVGYLLKKTESKSFYTKGYFVLTNNYLHEFKSSDFFLDKSKS
 10 20 30 40 50 60

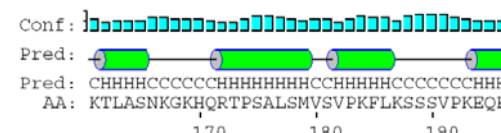
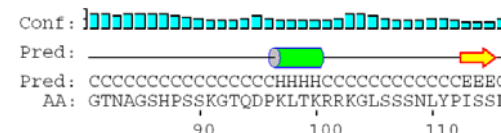
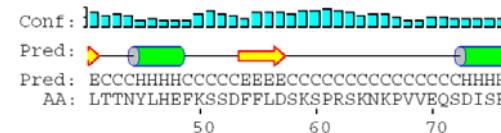
Conf: 885354300110000001466656776431001241000002663100123100023343
 Pred: CCCCCCCCCCHHHHCCCCCCCCCCCCCCCCCCCCCHHHHCCCCCCCCCEEECCCC
 AA: PRSKNKPVEQSDISRNVKDGTDNAGSHPSKGTQDPKLTKRKGLSSSNLYPISSLSLND
 70 80 90 100 110 120

Conf: 312246673278973034567400001234443200035420220443331003555313
 Pred: CCCCCCEEEEEECCCCCCCCCCCCCCCCCCCCCCCCCHHHHCCCCCHHHHHHHHC
 AA: CSLKDSTDSTFVLQGYASYHSPEDTCTKESSTSDLACPTKTLASNKGKHQRTPSALSMV
 130 140 150 160 170 180

Conf: 20355415788410012456777776542024327899851676888788889999999
 Pred: CHHHHCCCCCHHHHHHHHHHHHHHHHHHCCCEEEEEECCCCCHHHHHHHHHHHH
 AA: SVPKFLKSSSVPKEQKKAKEEANINKKSICEKRVEWTFKIFASLEPTPEESKNFKKWVQ
 190 200 210 220 230 240

Conf: 98765301111369
 Pred: HHHHHHCCCCC
 AA: DIKALTSFNSTQER
 250

Reliable



Legend:

= helix	= strand	= coil
= confidence of prediction	= helix	
	= strand	
	= coil	

Pred: predicted secondary structure
 AA: target sequence

Which program to use?

As of yet, there is no clear “best prediction method”

- Apply a few different methods to your sequence
- Incorporate information from homologous proteins
 - Using a prediction method that does it implicitly
 - Manually
- Prediction should be smoothed to remove unlikely predictions e.g. an isolated residue predicted to be in an α -helix.
- Use reliability indices to guide your prediction.

Limitation

- Stated accuracy for a program may not reflect the accuracy of prediction for all sequences
 - The current state of art for prediction ranges from 65% given only a single sequence to 70% when homologues are available; if the structural domain is known accuracy can rise to 80%

Domains, Motifs, Patterns, And Profiles

Motif: the biological object one attempts to model

A functional or structural domain, active site, phosphorylation site etc.

Domain: an independent structural and functional modular unit within a protein

Different regions along a single polypeptide chain that can act as independent units, to the extent that they can be excised from the chain, and still be shown to fold correctly, and often still exhibit biological activity.

Pattern: a qualitative motif description

Based on a regular expression-like syntax

Profile: a quantitative motif description

Assigns a degree of similarity to a potential match i.e. rather than identifying only the “consensus” or most common amino acid at a particular location, we can assign a probability to each amino acid in each position of the domain.

SMART (Simple Modular Architecture Research Tool)

- Web-based resource used for :
 - rapid annotation of protein domains.
 - analysis of domain architectures.
- Relies on hand curated multiple sequence alignments of representative family members from PSI-BLAST
 - constructs alignments, profiles, and Hidden Markov models to search databases iterative for more sequences until no more homologues detected
- Domain termini or boundaries, being the least conserved regions of alignments, are often inaccurate and several profiles represent incomplete portions of domains
- <http://smart.embl-heidelberg.de/>

Prosite Patterns

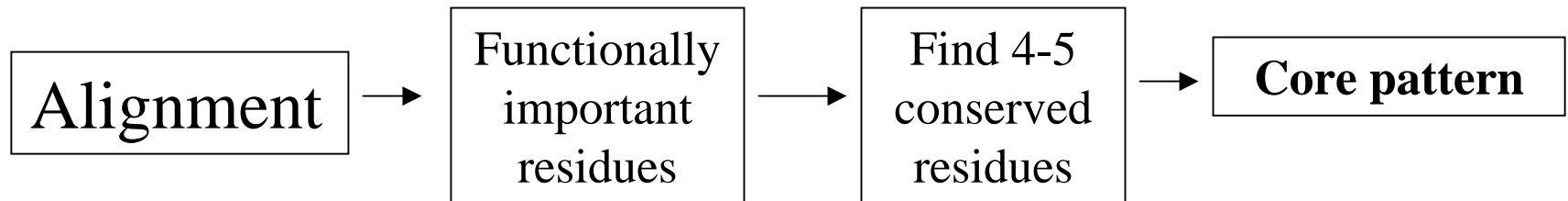
Pattern is given as regular expression:

`[AC]-x-V-x(4)-{ED}`

ala/cys-any-val-any-any-any-any-(any except glu or asp)

Building a pattern:

- from literature
- new patterns

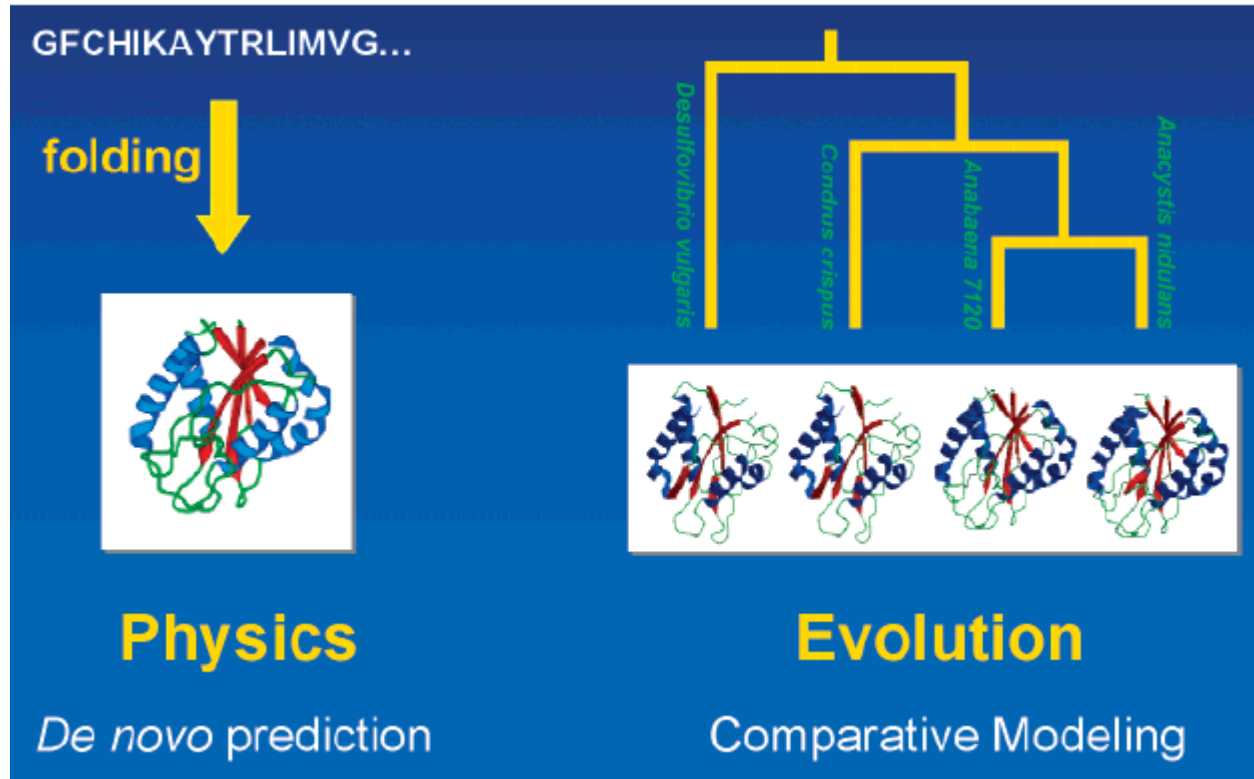


<http://www.expasy.ch/prosite/>

Common Protein Pattern Databases

- Prosite patterns
 - Prosite profiles
 - Pfam
 - SMART
 - Prints
 - TIGRFAMs
 - BLOCKS
 - ProDom
 - PIR-ALN
 - ProtoMap
 - Domo
 - ProClass
- Many different protein signature databases from small patterns to alignments to complex HMMs
 - Have different strengths and weaknesses
 - Have different database formats
- Therefore:** best to combine methods

3-D Structure of Proteins



Laws of Physics

Use first principles to computationally fold proteins

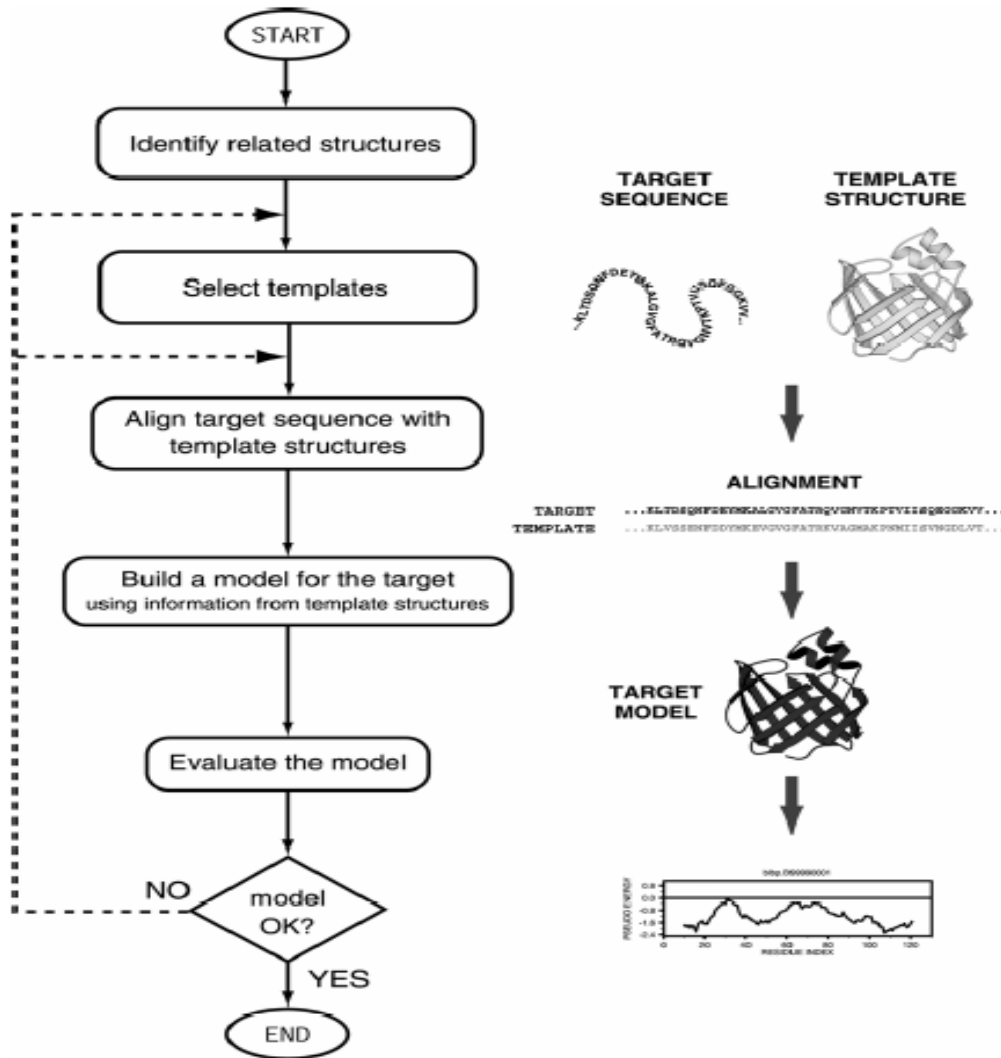
Theory of Evolution

Based on the concept that families of similar proteins share similar sequences, structure, and often function

Comparative modeling

Four Main Steps in this process

- Finding known structures related to the target sequence being modeled (finding templates)
- Aligning sequence with known structures
- Building a model
- Assessing the model



Template identification/Selection

- Sequence comparisons using various sequence alignment programs
- Fold assignment

Fold: *Major structural similarity*

[*Structural Classification of Proteins (SCOP)*]

Proteins are defined as having a common fold if they have the same major secondary structures in the same arrangement and with the same topological connections. Different proteins with the same fold often have peripheral elements of secondary structure and turn regions that differ in size and conformation. In some cases, these differing peripheral regions may comprise half the structure. Proteins placed together in the same fold category may not have a common evolutionary origin: the structural similarities could arise just from the physics and chemistry of proteins favoring certain packing arrangements and chain topologies.

Threading or Fold Recognition

Basis

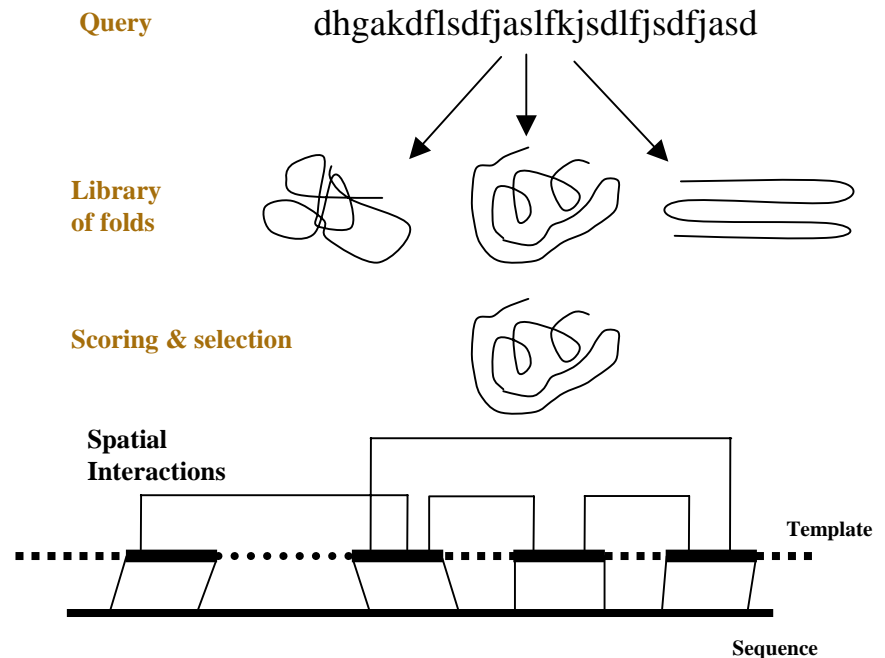
It is estimated there are only around 1000 to 10000 stable folds in nature and irrespective of the amino acid sequence, a protein will adopt one of these folds

The basic idea

Fold recognition is essentially finding the best fit of a sequence to a set of candidate folds i.e. placing a protein sequence onto a structural template “optimally”. The best sequence-fold alignment is selected using a fitness scoring function

Key components

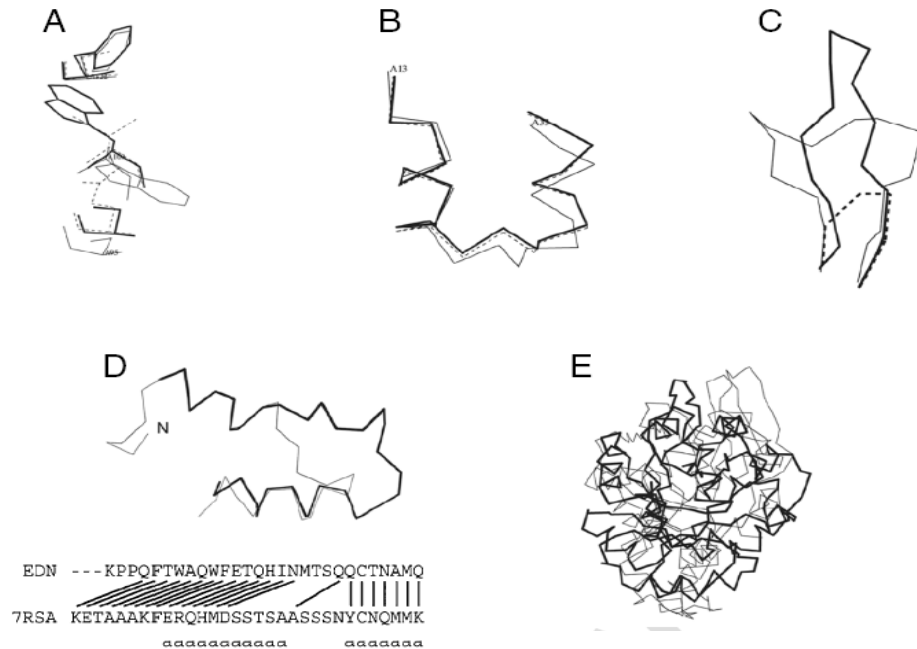
- A structural template database
- An “energy” function for measuring quality of a placement (alignment)
- An algorithm for finding an optimal placement
- A capability for assessing the reliability of prediction



Template and target alignment

- Most crucial step in making the model
- For closely related protein sequences with identity higher than 40%, the alignment is almost always correct. Regions of low local sequence similarity become common when the overall sequence identity is below 40%
- May need to be hand edited based on input from other resources such as secondary structure prediction, solvent accessibility prediction, mutational studies etc.

Model Evaluation- Typical errors



- A. Side-chain packing
- B. Distortions and shifts in correctly aligned regions
- C. Errors in regions without a template
- D. Errors due to misalignments
- E. Errors due to an incorrect template

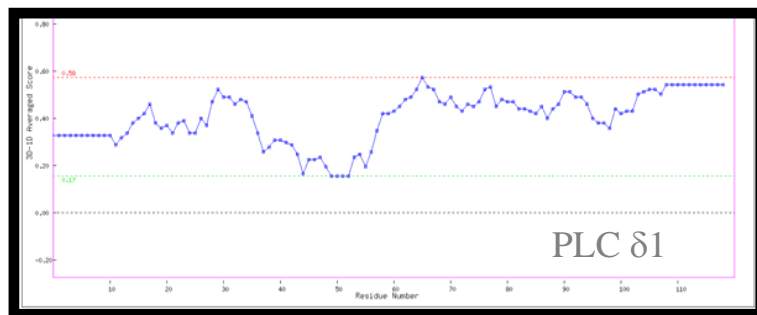
Accuracy associated with percent sequence that matches the template

- Most errors in side chains or in loops
- Other errors include *small* shifts and distortions in other parts of the protein
- Error increases *rapidly* below 30% sequence similarity

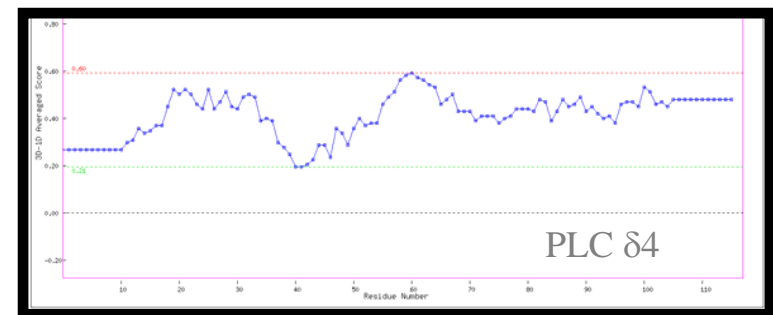
Verify 3D: Model Evaluation

Assesses the compatibility of a protein structure with its sequence

- Can be used to identify the regions that are unreliable or have been modeled improperly and require remodeling.
- Reliable means to assess the quality of a modeled protein- scores are consistently high for high-resolution experimentally determined structures.
- Can discriminate among potential models for a single sequence - homology models constructed based on alignments to templates of decreasing sequence similarity have correspondingly degraded scores



Template



Target

De novo/ Ab initio prediction

- Can be done with any protein...no other members of the family need be known
- Based on assumption – Native State is at the global free energy minimum
- A large scale search is carried out to find this minimum
 - Methods for carrying out search efficiently
 - Free energy function used to calculate energy
- Accuracy is much lower than comparative methods (with >30% sequence similarity)
- Basic structure can still be determined reasonably well with 150 amino acids or less - Low-resolution information may be useful in finding structural and functional relationships not apparent by other methods

Summary

- When a suitable template structure exists in PDB, using homology modeling on target sequence is best for predicting the structure
 - Comparative modeling is more accurate...*but* requires suitable structural templates to be known
 - *De novo/ Ab initio* structure can be used for any protein but requires more computing power and as yet doesn't produce highly accurate models.
- Fold Recognition servers can help find a template when conventional sequence analysis methods fail
- Combining elements from several sources may allow you to construct reasonably accurate models