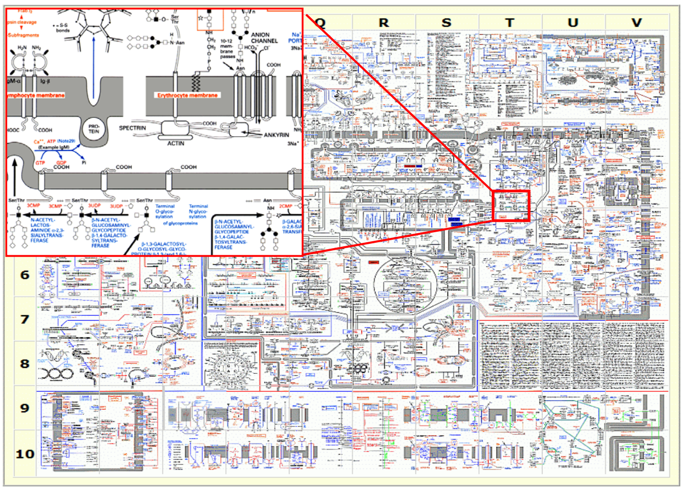
COSC 348: Computing for Bioinformatics

Lecture 24:
Protein structure determination

Lubica Benuskova

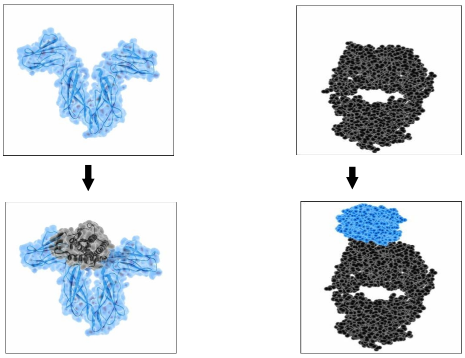
http://www.cs.otago.ac.nz/cosc348/

Life is maintained by protein networks



Protein 3D structure determines its function

- Two examples of structure to function relationship: molecules interact based on their shape and physical properties.

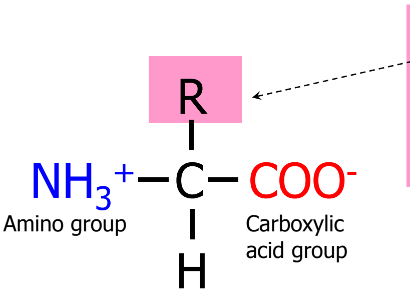


Relation of function to 3D structure

- Similar structure, mutually exclusive function, e.g. Lysozyme versus α-lactalbumin.
- Same function, completely different structures, e.g. Carbonic anhydrases from *M. thermophila* and mouse
- One structure, multiple functions:
- Gal1p – Kinase as well as regulator of Gal-gene expression Gal3p – 70% similar; does not have kinase activity

These are exceptions – most of time 3D structure is a good predictor of protein function, therefore protein structure prediction is a very important part of bioinformatics with important effect on medicine and pharmacology.

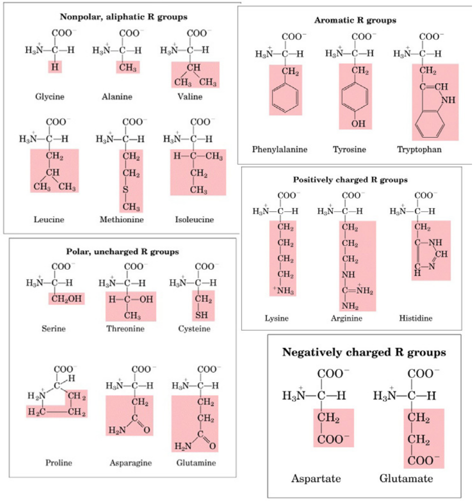
Amino acid: basic unit of a protein



Different side chains or R-groups (residues), determine the properties of 20 amino acids.

C = carbon, O = oxygen, H = hydrogen, N = nitrogen (COHN)

Twenty standard Amino Acids



- Classification of AA based on polarity of their residues
- Side chains (or R-groups) have different properties
  - Polarity (charge)
  - Acidity
  - Hydrophobicity
  - …

Protein primary structure: sequence of AA

- Proteins are linear *polymers* (i.e. "many"-mers) of amino acids
  - Peptide ~ 2-10 amino acids
  - Polypeptide ~ 10-50 amino acids
  - Protein > 50- amino acids
- A given protein has *always* the same amino acid sequence
  - (Protein sequence is determined by the gene DNA sequence)
- A given protein has *always* a unique three- dimensional structure.
  - (Protein structure is determined by protein sequence)

always = biological always (there are always exceptions)

Hierarchical nature of protein structure

Primary structure (Amino acid sequence)

↓

Secondary structure (α-helix, β-sheet)

↓

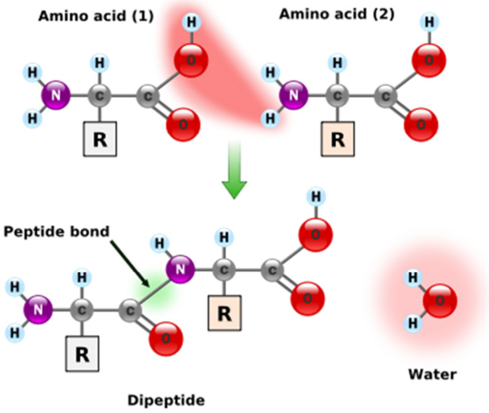
Tertiary structure (Three-dimensional structure formed by assembly of secondary structures)

↓

Quaternary structure (Structure formed by several protein chains)

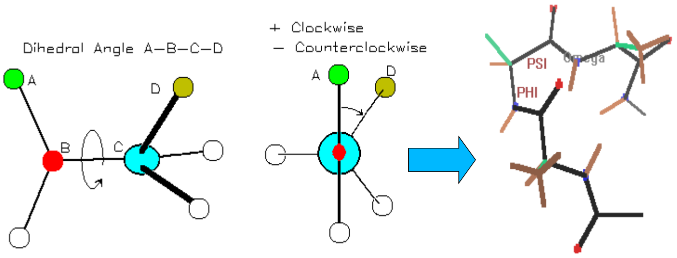
3D

Peptide bond binds amino acids into the primary structure of proteins

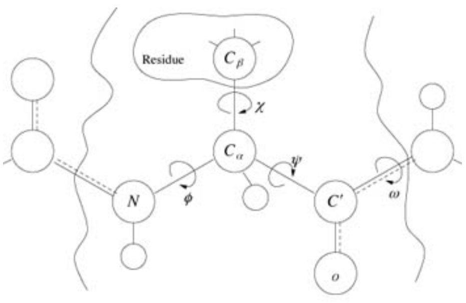


Secondary structure from bond angle

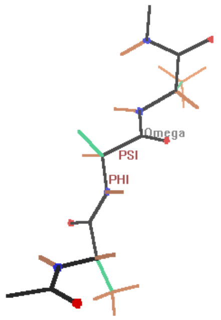
- Peptide bond rotation determines protein folding, i.e. *its secondary structure*
- *Torsion angles*



Torsion angles determine protein folding into the secondary structure

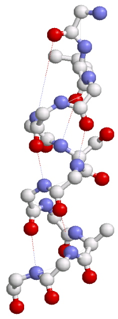


Backbone torsion angles of a protein

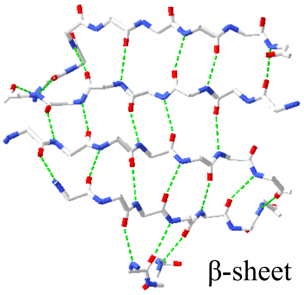


2 basic types of secondary structure

- regular patterns of backbone peptide torsion angles and hydrogen bonds (dotted lines)
- Two types of secondary structure



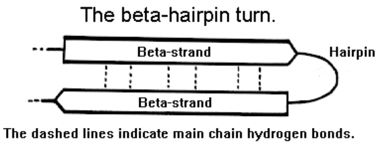
α-helix



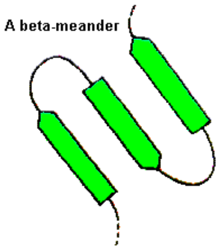
β-sheet

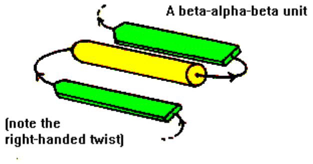
Tertiary structure of proteins

- Tertiary structure emerges as an arrangement of secondary structure elements:

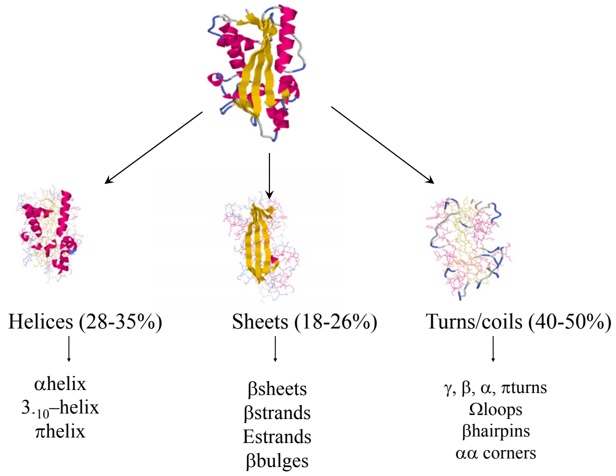


The dashed lines indicate main chain hydrogen bonds.





Secondary structures within the tertiary one

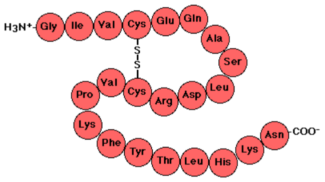


Helices (28-35%)	Sheets (18-26%)	Turns/coils (40-50%)
αhelix 3.$_{10}$–helix πhelix	βsheets βstrands Estrands βbulges	γ, β, α, πturns Ωloops βhairpins αα corners

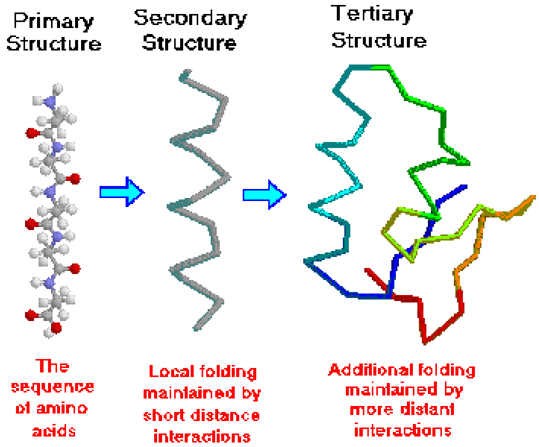
Bonds of tertiary structure

- Noncovalent bonds between atoms:
  - Van der Waals forces (weak electrical attraction between atoms)
  - Hydrophobic (clustering of nonpolar groups)
  - Hydrogen bonding

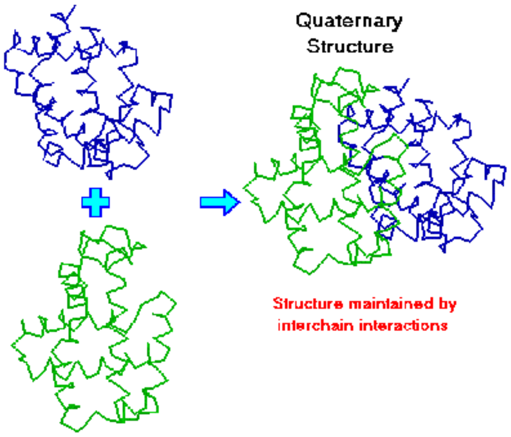
- Covalent bonds between atoms:
  - Disulfide bonds (contribute to the stability of the folded state by linking distant parts of the polypeptide chain)



Protein structure: summary


Primary Structure
Secondary Structure
Tertiary Structure
The sequence of amino acids
Local folding maintained by short distance interactions
Additional folding maintained by more distant interactions

Protein structure of composite proteins



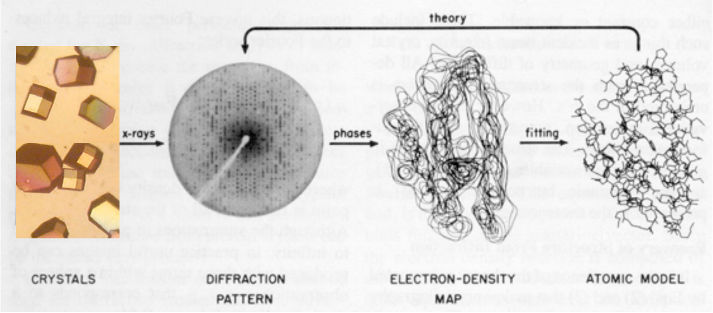
Protein 3D-structure determination

- Experimental physical methods
  - X-Ray crystallography
  - NMR (nuclear magnetic resonance)
  - Cryo-EM ("frozen" electron microscopy)

- Computational methods
  - *Ab initio* modelling (i.e. from the start or begining)
  - Comparative / homology modelling
  - Threading or fold recognition

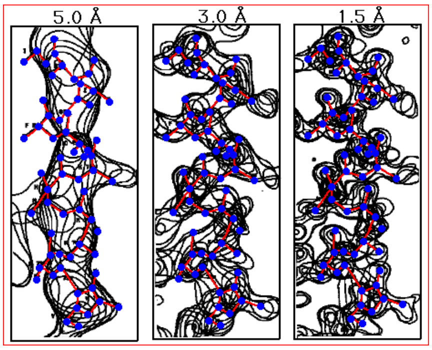
X-ray crystallography

- Crystallize protein
- Collect diffraction patterns under different angles
- Calculate the electron density map and atomic model



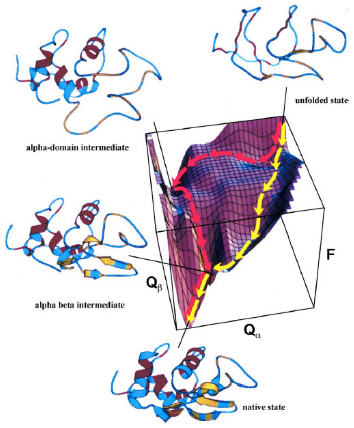
X-ray: electron density maps

- > 60,000 3D structures of proteins and other biomolecules have been determined by X-rays.
- X-ray crystallography is now used routinely to determine how a pharmaceutical drug interacts with its protein target.
- Problems: expensive machine, knowledge of physics and many proteins cannot be prepared as crystals.



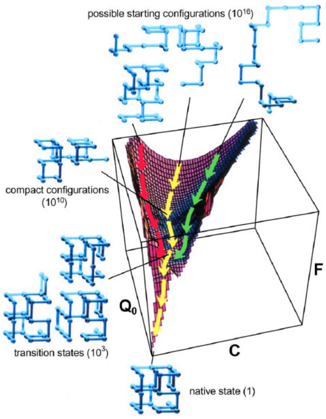
Resolution on the level of atom size 10^{-10} m.

Ab Initio methods



- Based on protein denaturation process:
  - Denatured state = unfolded state
  - Native state = folded state
  - Denaturation by heat, urea, salts
- Depends heavily on the analysis of known protein structures and establishing ***3D*** *structure to energy relationship.*

Ab Initio methods


possible starting configurations (10^{16})
compact configurations (10^{10})
transition states (10^3)
native state (1)
F
Q_0
C

Many possibilities, computationally very demanding to optimise them all.

Gamer community by playing the game **FoldIt** solved the protein structure: http://arstechnica.com/science/news/2010/08/gamers-beat-algorithms-for-finding-protein-structures.ars

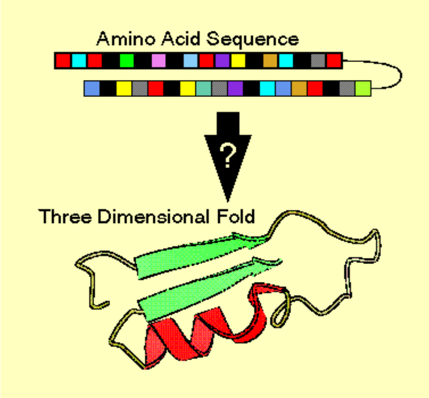
FoldIt

- The FoldIt algorithm can handle huge energy landscape and mappings to different configurations. But often it gets stuck – and that's where the gamers came into play.
- Users were given a set of controls that let them poke and prod the protein's structure in three dimensions; displays provide live feedback on the energy of a configuration.
- Nobody had a biochemical education, there were simple structural tasks to learn on, there were leaderboards, team and individual challenges, user forums, etc.

Rosetta@home

- FoldIt was based on a program called Rosetta@home, which is still available online for anyone to take part.
- By running the Rosetta@home program on your computer while you don't need it you will help to speed up the research. Anyone can thus help the efforts at designing new proteins to fight diseases such as HIV, malaria, cancer, etc.
- 2011 year's issue of Nature magazine had an article by Sievers et al. describing work they are doing with collaborators using Rosetta software (https://www.rosettacommons.org/) to design a new class of drug for Alzheimer's disease.

Computational challenge: predicting 3D protein structure **JUST** from its primary structure



- Basis for this kind of prediction is this:
- 3D structure is much more conserved than AA sequence during evolution.

Comparative/homology modelling

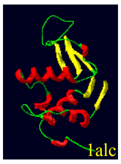
- Predicting protein structure by homology. Homology is the similarity between two sequences owing to common ancestry.

```
KQFTKCELSQNLYDIDGYGRIALPELICTMFHT
SGYDTQAIVENDESTEYGLFQISNALWCKSSQS
PQSRNICDITCDKFLDDDITDDIMCAKKILDIK
GIDYWIAHKALCTEKLEQWLCEKE
```

Homologous

Share Similar Sequence

```
KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFE
SNFNTQATNRNTDGSTDYGILQINSRWWCNDGRTP
GSRNLCNIPCSALLSSDITASVNCAKKIVSDGNGM
NAWVAWRNRCKGTDVQAWIRGCRL
```



Use as a template

Predicted structural motif

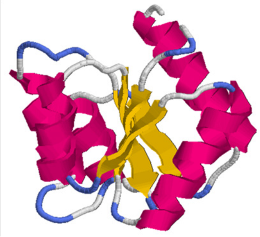
Known structural motif

Basis of comparative modelling

- A **structural alphabet** that is a library of ***small structural motifs*** which approximate every part of the protein 3D structure.
  - Protein block (PB) is a set of 16 short structural motifs;
  - each motif is represented by a vector of 8 torsion angles of 5 or so consecutive amino acids;
  - Motifs are denoted by letter *a, b, …,p.*
- Structural alphabet is derived based on the study of the relation of primary, secondary and tertiary structure by other methods (X-ray).
- Protein block (PB) substitution matrix is then used to evaluate and optimise the homology models.

Generation of PB substitution matrix

- First we encode known 3D structures as 1-dimensional protein block (PB) sequences;
- We do it for every known 3D – 1D relationship (it's not 1-to-1 !!!).
- Calculations of substitution frequency of each combination of PB for the particular primary structure. The 3D structure that scores the highest wins.



3D structure

Sequence of PBs

```
KBCCDDDDFBFKLMMMMMMMMMMNOP
ABDCDDFBFKLMMMMMMNGOIABDCD
DFBDGHILMLMMMMMMMMPMKLMMPC
CDDDDDFBDCFKLMMMMMMNOPABDCD
DDDDFKLMMMMMMMMMMMNO
```

Protein block (PB) substitution matrix

Protein blocks	*a*	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*	*j*	*k*	*l*	*m*	*n*	*o*	*p*
a	2.28															
b	-0.12	2.49														
c	0.54	-0.21	1.69													
d	-0.29	-0.44	0.17	1.35												
e	-1.59	-0.48	-1.10	-0.36	3.05											
f	-0.54	-1.53	-0.39	-0.49	0.75	2.21										
g	0.31	-0.73	0.18	-1.29	1.37	-0.33	3.25									
h	-1.14	0.20	-1.63	-1.20	0.66	-0.34	-0.74	3.07								
i	0.39	0.24	-1.11	-1.12	-1.15	-1.07	-0.19	-0.92	3.37							
j	-1.15	0.32	-1.03	-0.92	-0.76	-0.34	-0.51	1.18	1.54	3.74						
k	-1.75	-0.03	-2.45	-2.63	-0.38	-0.04	-1.39	0.51	-0.15	0.07	2.52					
l	-0.60	0.04	-2.21	-1.56	-1.76	-0.33	-0.74	-0.36	-0.22	-0.12	0.19	2.24				
m	-2.40	-2.98	-2.70	-5.20	-4.75	-2.14	-1.10	-2.93	-3.15	-2.00	-1.02	-0.68	1.06			
n	-1.40	-0.83	-1.68	-3.07	-0.58	-1.99	1.07	-1.07	-0.97	-0.44	-0.56	-0.27	-0.77	3.65		
o	-0.54	-0.55	-0.65	-2.66	-2.48	-1.41	-0.01	0.96	-0.89	-0.48	-1.71	0.06	-1.26	0.26	3.36	
p	-0.36	0.33	-0.01	-2.10	-2.22	-1.91	0.47	-1.81	1.32	0.60	-1.35	-1.23	-1.10	0.36	0.24	2.83

Tyagi M, Venkataraman SG, Srinivasan N, de Brevern AG, Offmann B: A substitution matrix for structural alphabet based on structural alignment of homologous proteins and its applications. *Proteins*, 2006; 65:32–39.

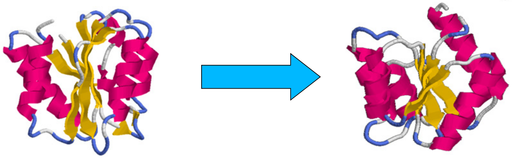
Summary of homology modelling

- First perform the primary structure alignment

```
KBCCDDDDFBFKLMMMMMMMMMMNOPABDCDDFB
FKLMMMMMMNGOIABDCDDFBDGHILMLMMMMMM
MMPMKLMMPCCDDDDFBDCFKLMMMMMMNOPABD
CDDDDDFKLMMMMMMMMMMMNO
```

```
DDDDFKOMMMMMMMMMMMMMMMNOPACDDDDFKL
PCFKLMMMPFBDCDDDDDEHJLPACFKLMMMMMM
MMMMMMGHIACDDEHIAFKLNOMMMMMMMMMMMM
NOPACFBACDDDEHJACKLMMMMMMMMMMMMM
```

- Then re-arrange structural motifs based on the PB sequence alignment and choose the PB alignment with the the highest score.



Protein Block Expert (PBE)

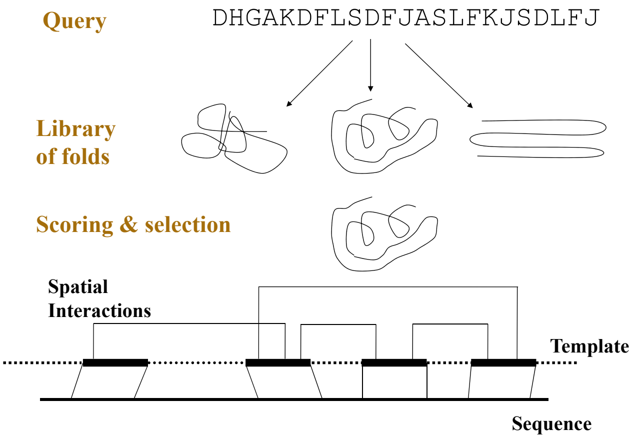
- Protein Block Expert, an online tool:
  - Pair-wise PB sequence alignment to compare two protein structures
  - Mining of structurally similar proteins from databases
  - Provides both local and global alignment algorithms
  - Limitation: for proteins that have no similarity to proteins with known structures the homology search cannot be used

Tyagi M, Sharma P, Swamy CS, Cadet F, Srinivasan N, Brevern AG, Offmann B: Protein Block Expert (PBE): A web-based protein structure analysis server using a structural alphabet. *Nucl Acids Res*. 2006 Jul 1; 34(Web Server issue):W119-23.

Threading or Fold Recognition

- Basis
  - It is estimated there are only around 1000 to 10 000 stable folds in nature
  - Fold recognition is essentially finding the best fit of a sequence to a set of candidate folds – *does not use sequence alignment !*
  - Selects the best sequence-fold mapping using a fitness scoring function
  - NP-complete problem

Threading: Basic Strategy



Query DHGAKDFLSDFJASLFKJSDLFJ

Library of folds

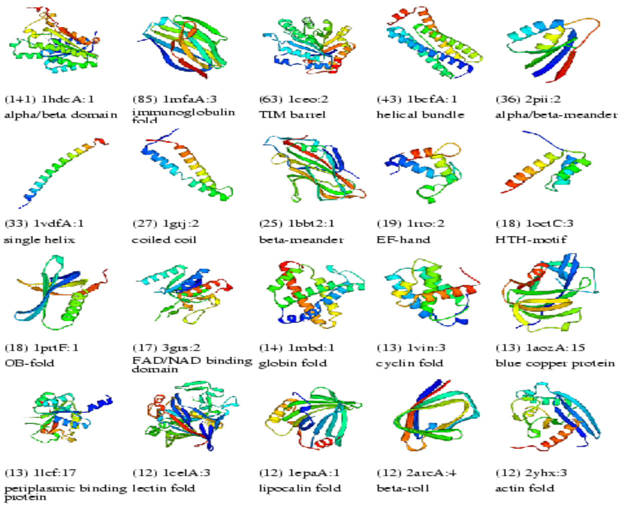
Scoring & selection

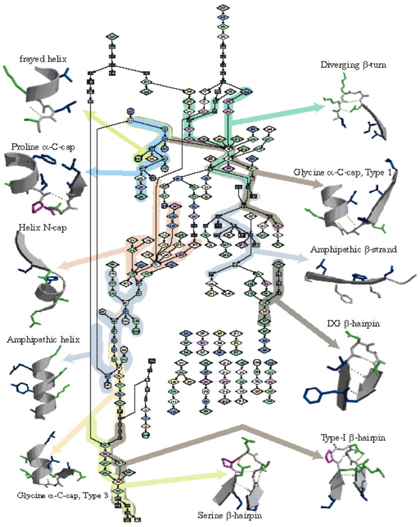
Spatial Interactions

Template

Sequence

Dominant domain fold types





Threading:

Means identification and scoring of structural motifs within a 3D protein structure based on the primary structure – it's like assembling a 3D puzzle

Major protein structure www sources

- > 80,000 entries in the world-wide Protein Databank (PDB)
  - http://www.wwpdb.org/index.html
- Major structural sources of proteins based on folds:
  - SCOP (Structural Classification Of Proteins)
    http://scop.mrc-lmb.cam.ac.uk/scop/
  - CATH (Class, Architecture, Topology, Homology)
    http://www.biochem.ucl.ac.uk/bsm/cath/
  - DALI DOMAIN DICTIONARY
    http://ekhidna.biocenter.helsinki.fi/dali_server/