Lecture 8: Protein structure analysis

Torgeir R. Hvidsten

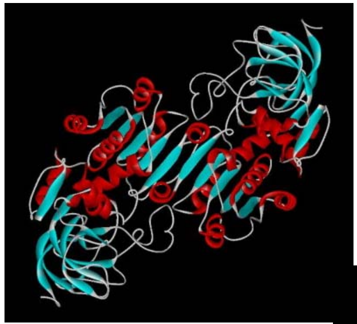
Professor
Norwegian University of Life Sciences

Guest lecturer
Umeå Plant Science Centre
Computational Life Science Cluster (CLiC)

Proteins play key roles in a living system

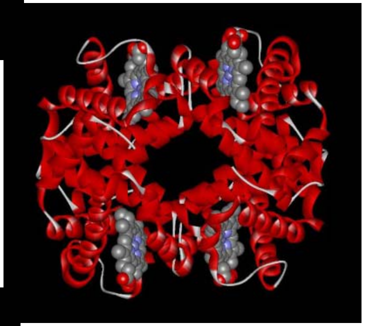
Three examples of protein functions

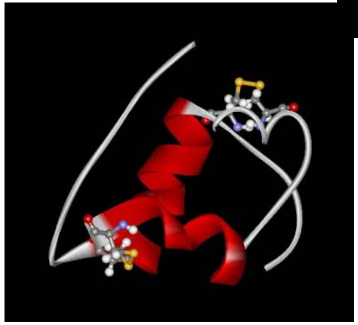
- Catalysis:
  Almost all chemical reactions in a living cell are catalyzed by protein enzymes
- Transport:
  Some proteins transports various substances, such as oxygen, ions, and so on
- Information transfer:
  For example, hormones



Alcohol dehydrogenase oxidizes alcohols to aldehydes or ketones

Haemoglobin carries oxygen

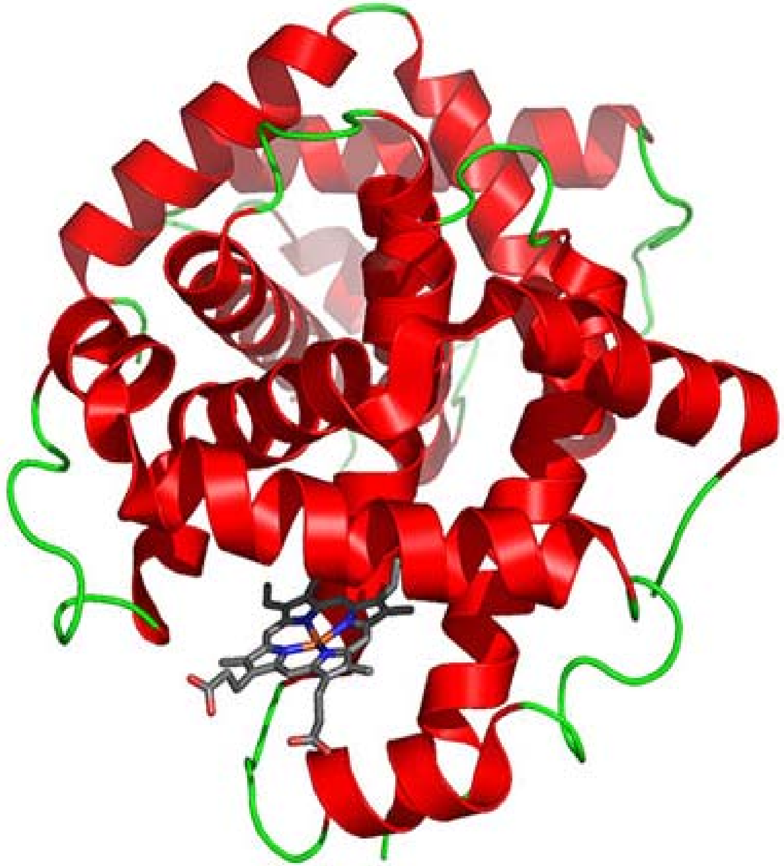




Insulin controls the amount of sugar in the blood

Structure - function

The 3D shape (and chemical properties) of proteins determine their function

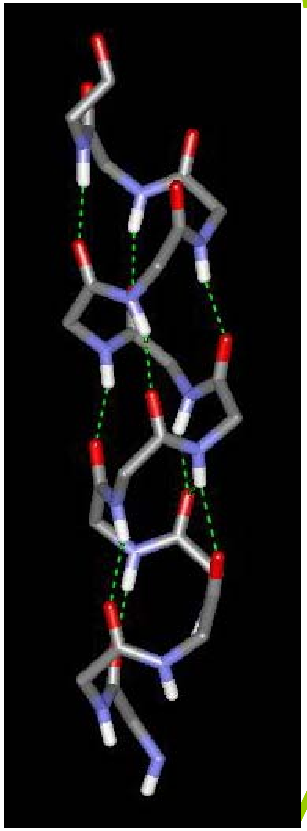


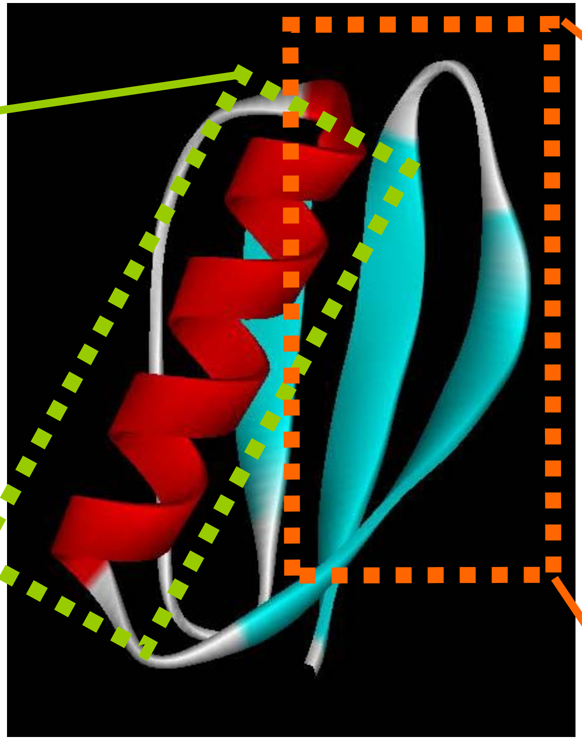
Hemoglobin

Hammer

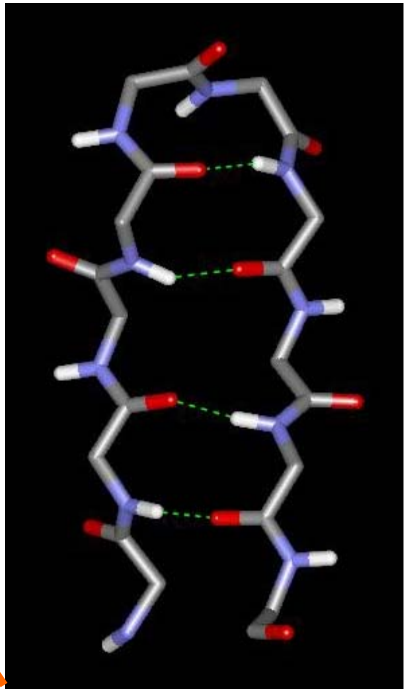
Basic structural units of proteins: Secondary structure

α-helix



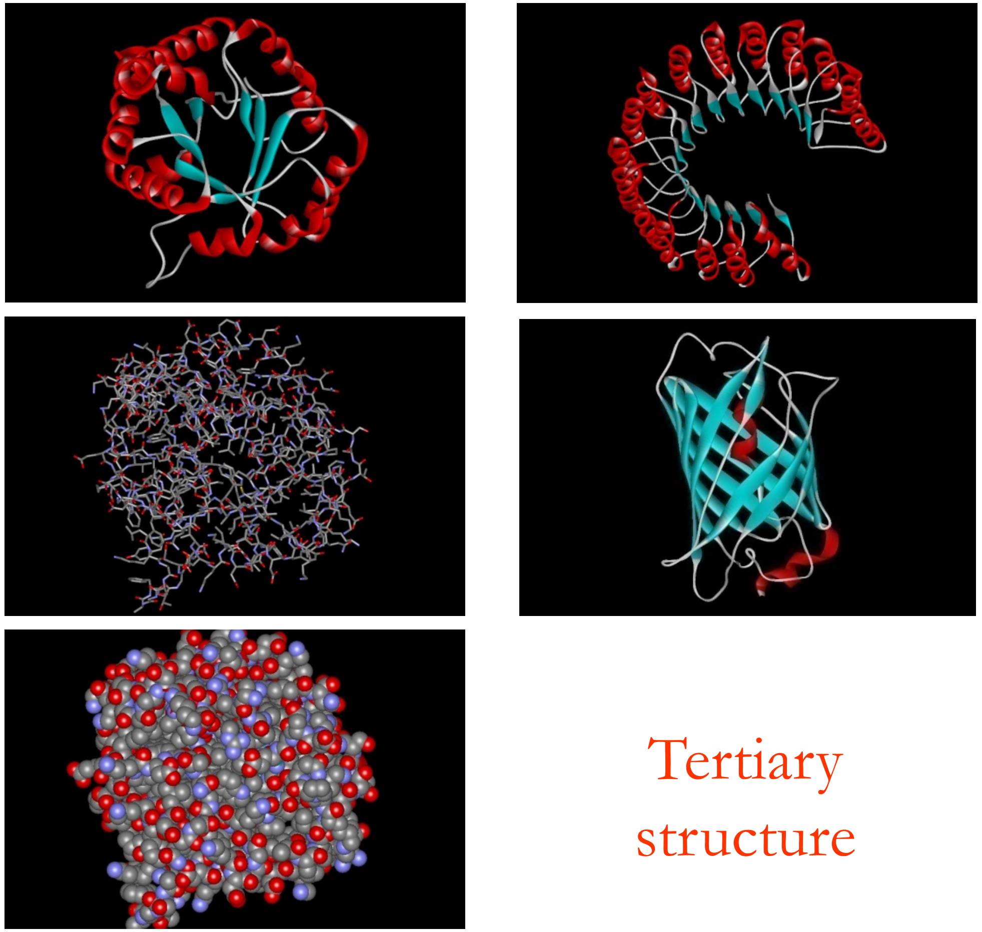


β-sheet

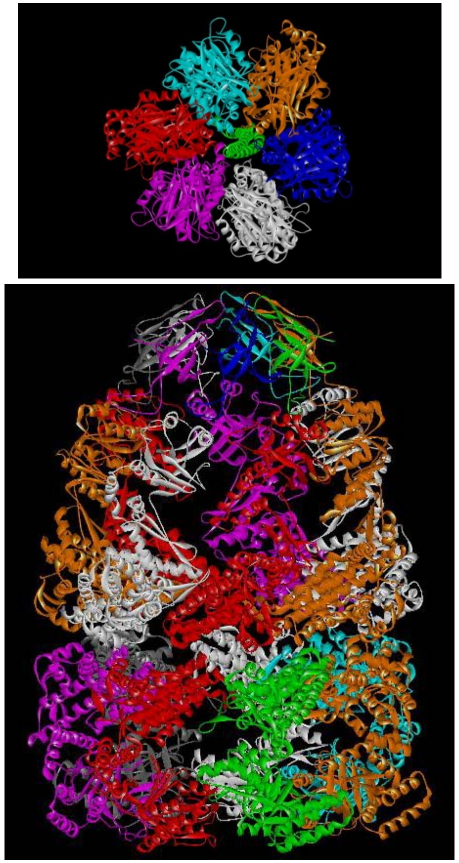


Secondary structures, α-helix and β-sheet, have regular hydrogen-bonding patterns

Three-dimensional structure of proteins



Tertiary structure



Quaternary structure

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 more than one polypeptide chains)

Domains: recurrent units of proteins

- The same or similar domains are found in different proteins

- Each domain has a well determined compact structure and performs a specific function

- Proteins evolve through the duplication and domain shuffling

Protein domains can be defined based on:

- Geometry: group of residues with a high contact density, number of contacts within domains is higher than the number of contacts between domains

- Kinetics: domain as an independently folding unit

- Physics: domain as a rigid body linked to other domains by flexible linkers

- Genetics: minimal fragment of gene that is capable of performing a specific function

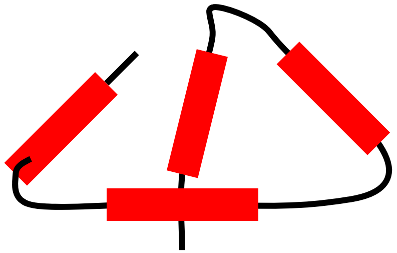
Protein folds

- One domain → one fold

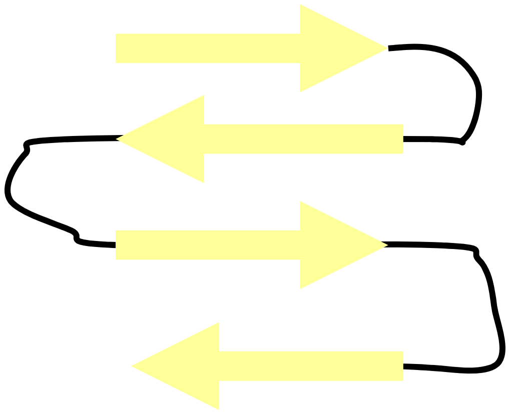
- Fold definition: two folds are similar if they have a similar topology: arrangement/orientation of secondary structure elements (architecture) and connectivity
  - topology = architecture + connectivity
- Fold classification: structural similarity between folds is found using structure-structure comparison algorithms

Domain/fold classification

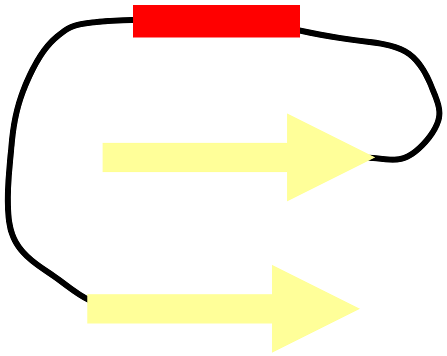
- Class α: a bundle of α helices connected by loops on the surface of protein
- Class β: antiparallel β sheets
- Class α/β: mainly parallel β sheets with intervening α helices
- Class α+β: mainly segregated α helices and antiparallel β sheets
- Multidomain proteins: comprise domains representing more than one of the above four classes
- Membrane and cell-surface proteins: α helices (hydrophobic) with a particular length range, traversing a membrane



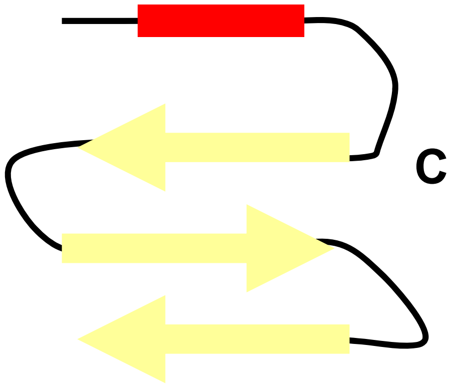
Class α



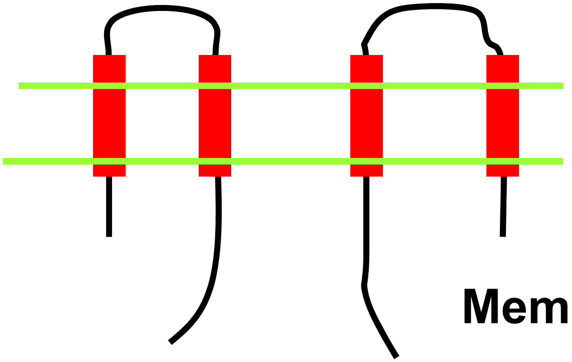
Class β



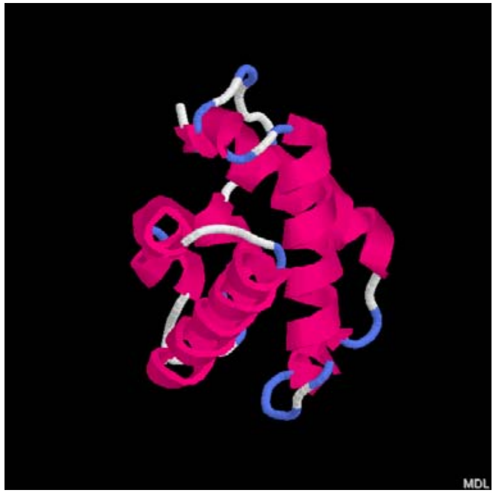
Class α/β



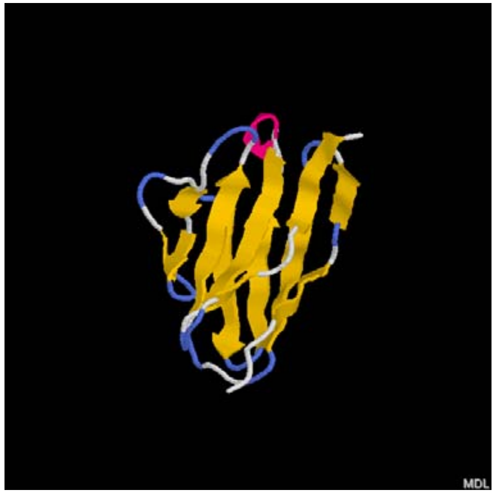
Class α+β



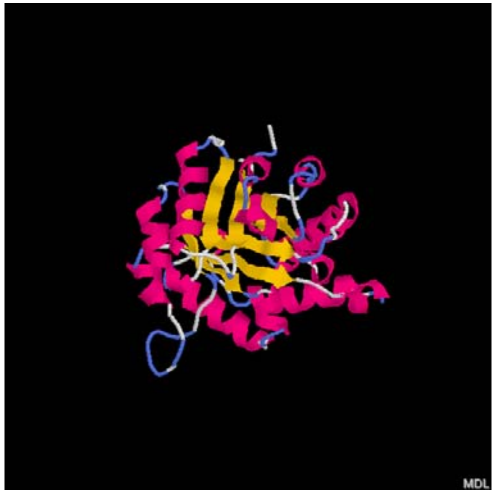
Membrane proteins



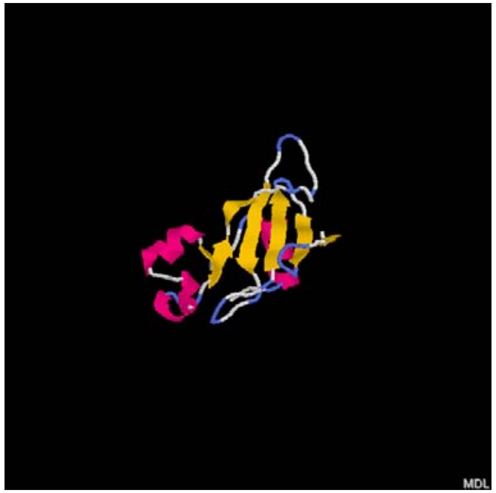
Class α



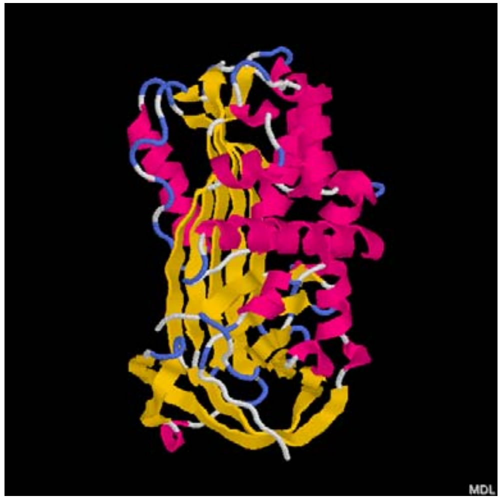
Class β



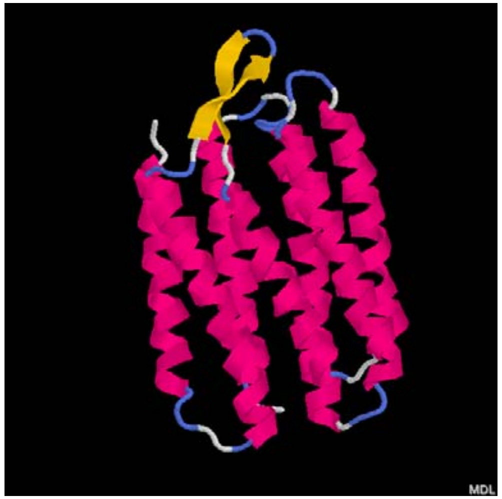
Class α/β



Class α+β



Multi-domain

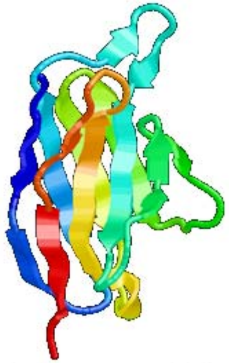


Membrain-bound

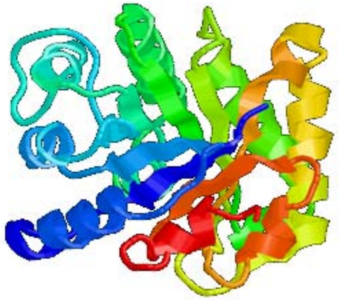
Structural classification of proteins (SCOP)

- The SCOP database aims to provide a detailed and comprehensive description of the structural and evolutionary relationships between all proteins whose structure is known.
- Created by manual inspection and aided by automated methods
- Consists of four hierarchical categories:
  - Class, Fold, Superfamily and Family.

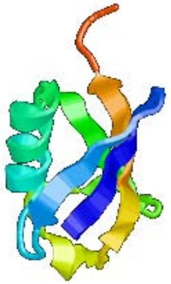
SCOP



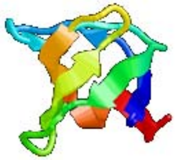
2IMM - Immunoglobulin like



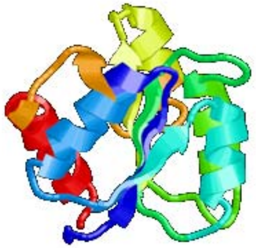
1TPH - beta/alpha barrel



1UBI - beta grasp

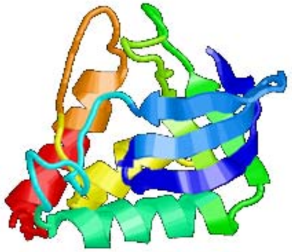


1CKA - SH3-like barrel

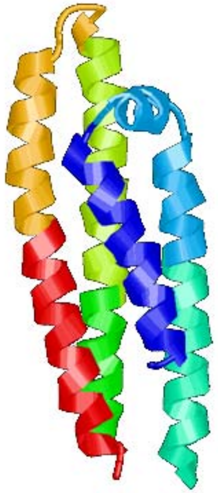


3CHY - Flavodoxin like

1FXD - Ferredoxin like



1SNC - OB-fold



1NFN - 4-helix bundle

The eight most frequent SCOP folds

Why study structure?

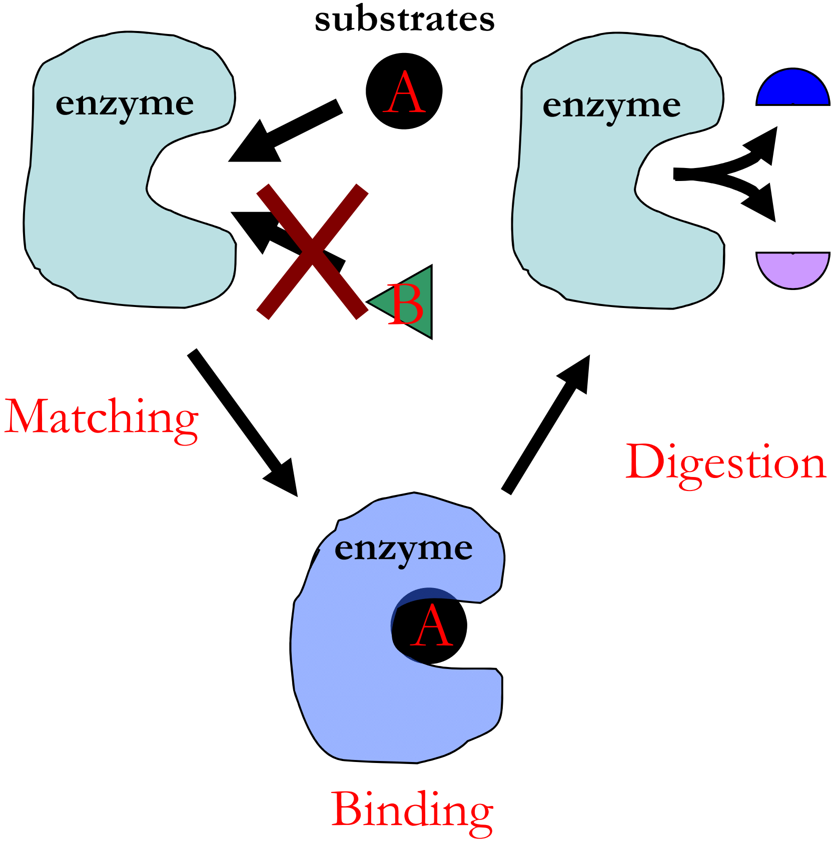
- A full understanding of a molecular system comes from careful examination of the sequence-structure-function triad
- Below 30% protein sequence identity detection of a homologous relationship is not guaranteed by sequence alone
- Structure is much more conserved than sequence

➢ However:

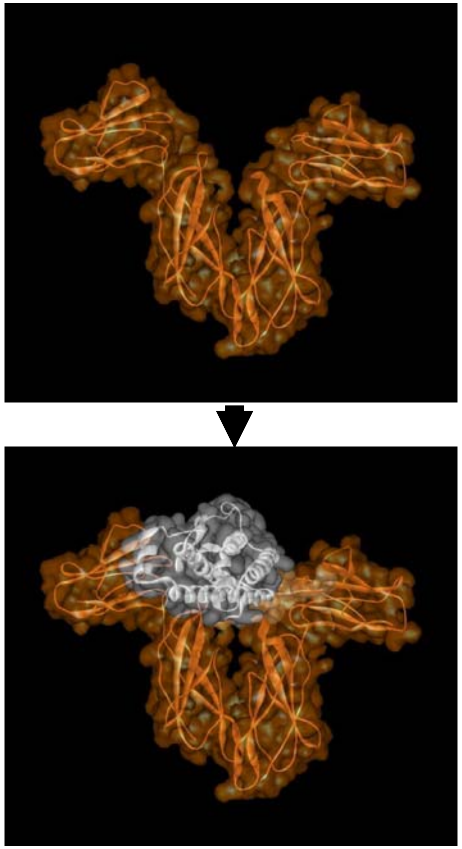
- A non-redundant set of sequences is different than a non-redundant set of structures is different than a non-redundant set of functions

The structure-function relationship

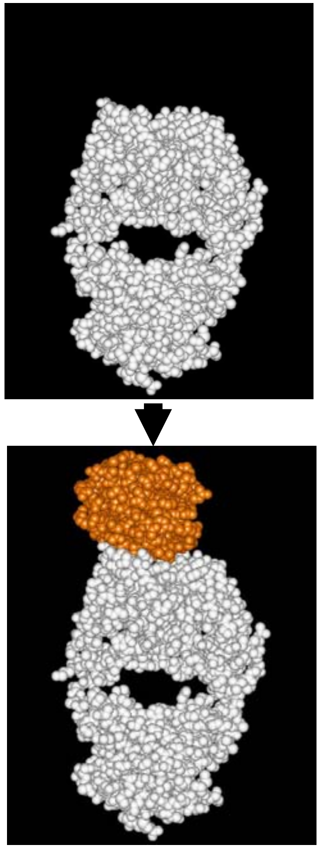
Example of enzyme reaction



Hormone receptor



Antibody

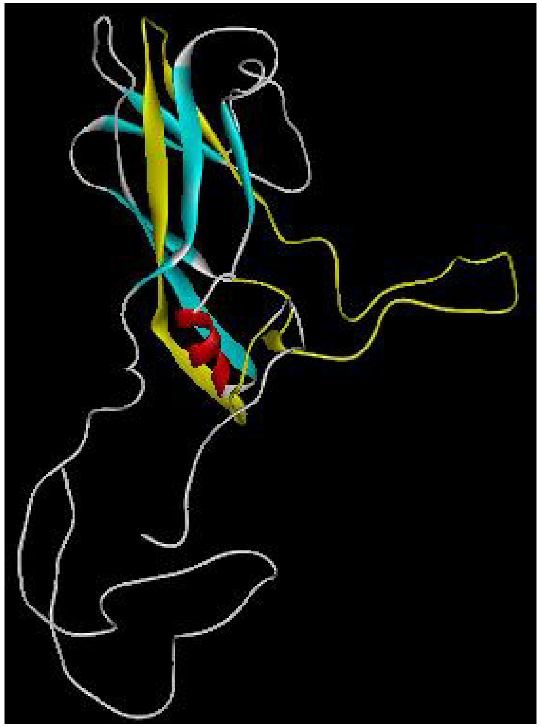


Structure-function relationships

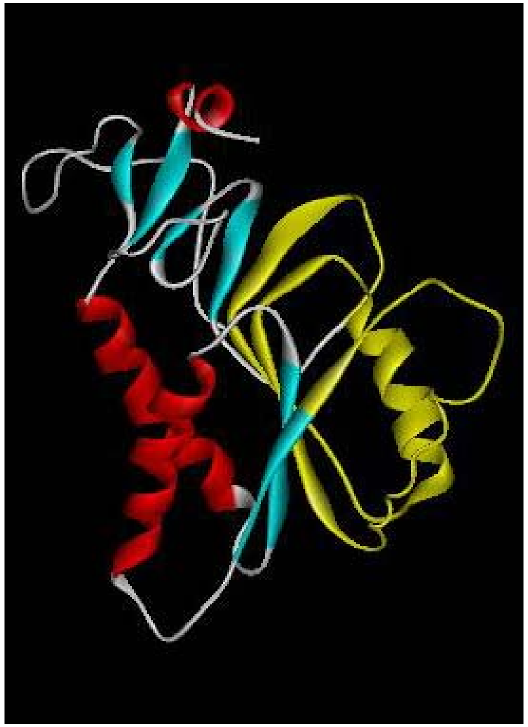
- The golden rule is there are no golden rules – George Bernard Shaw
  - Complication comes from one structure - multiple functions
  - Some folds are promiscuous and adopt many different functions - superfolds
- Above 40% sequence identity, sequences tend to have the same structure and function – but there are exceptions
- Structure and function tend to diverge at ~ 25% sequence identity
- The twilight zone: 20-40% sequence identity
- The structure-function relationship is even more complex than the relationship between sequence and structure (and not as well understood)

Similar sequences – different structures

1PIV:1
Viral Capsid Protein

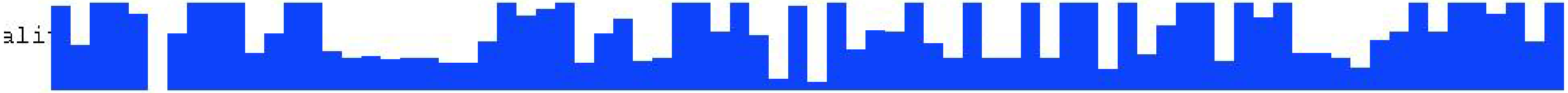


1HMP:A
Glycosyltransferase

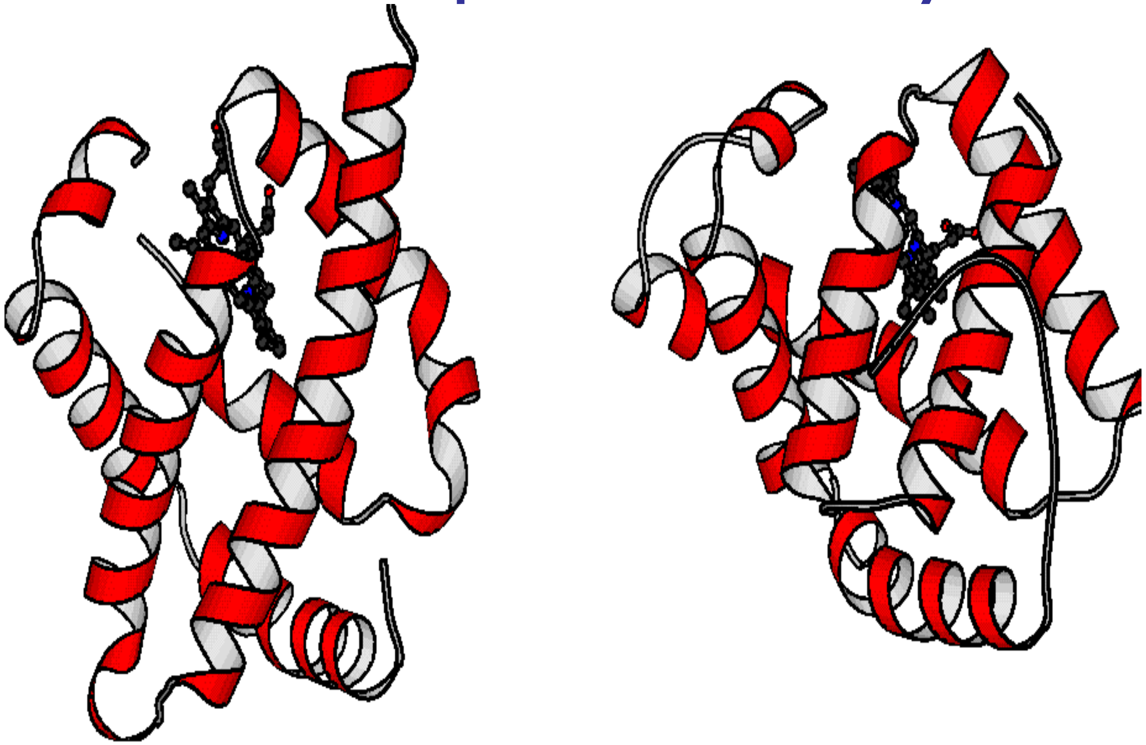


```
             10        20        30        40        50        60        70
-78 VPLKTDANDQ-IGDSL------YSAMTVDDFGVLAVRVVNDHNPTKVT--SKVRIYMKPKHVRV---WCPRPPRAVPY
-78 IRLKSYCNDQSTGDIKVIGGDDLSTLTGKN--VLIVEDIIDTGKTMQTLLSLVRQY-NPKMVKVASLLVKRTPRSVGY
```

44%

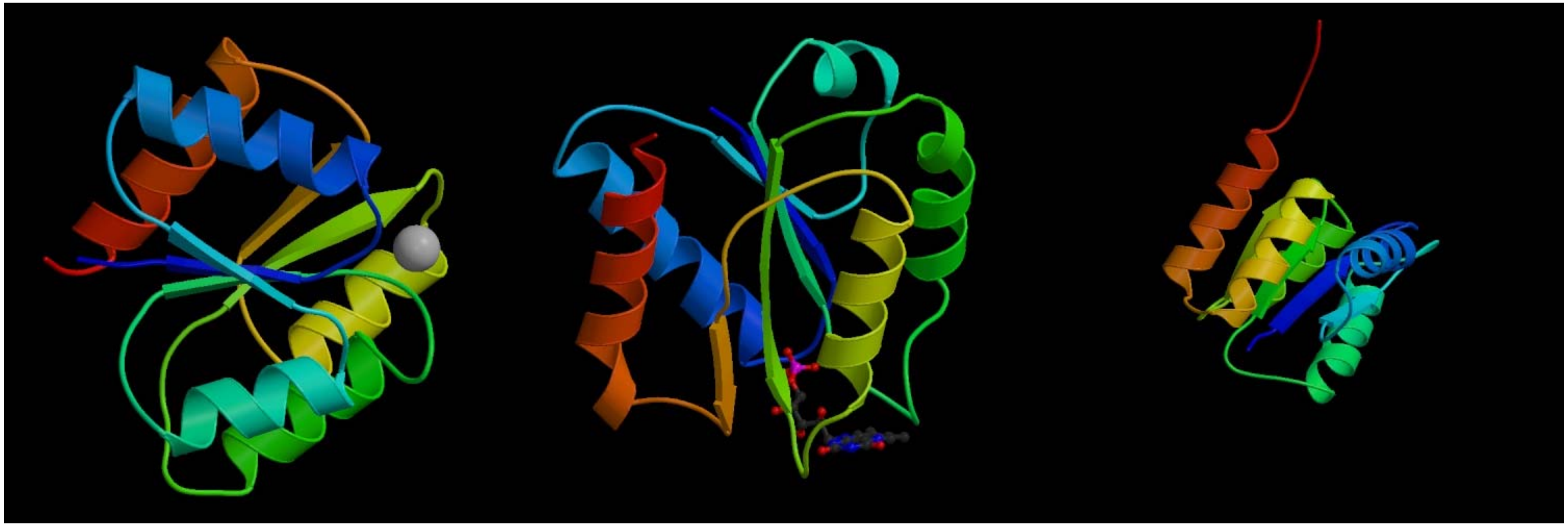


Same structure and function – low sequence identity



The globin fold is resilient to amino acid changes. *V. stercoraria* (bacterial) hemoglobin (left) and *P. marinus* (eukaryotic) hemoglobin (right) share just 8% sequence identity, but their overall fold and function is identical.

Similar structure - different function



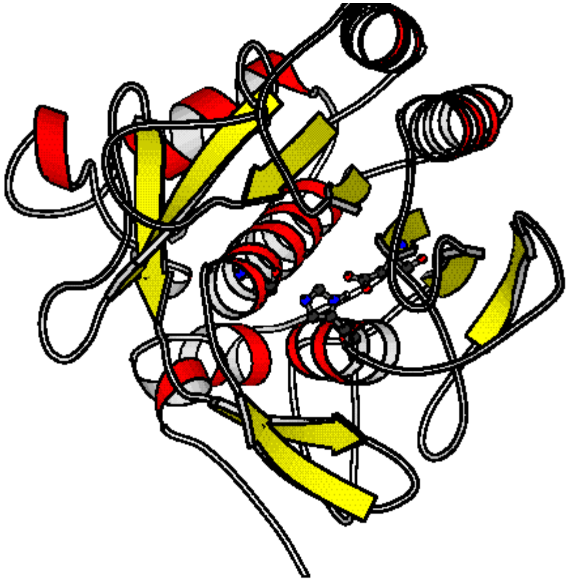
1ymv
CheY
Signal Transduction

1fla
Flavodoxin
Electron Transport

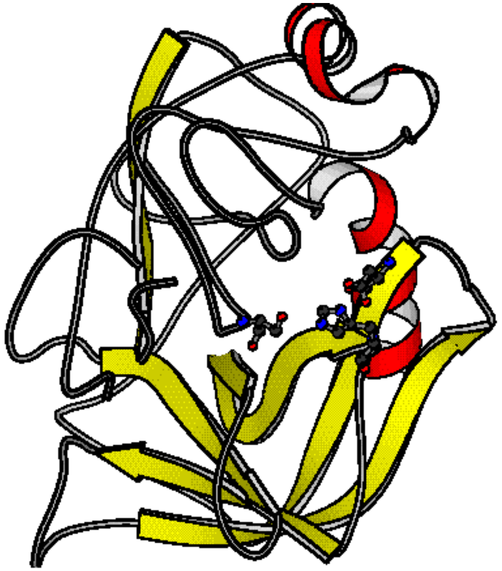
1pdo
Mannose Transporter

Less than 15% sequence identity

Convergent evolution



a. Subtilisin EC 3.4.21.62

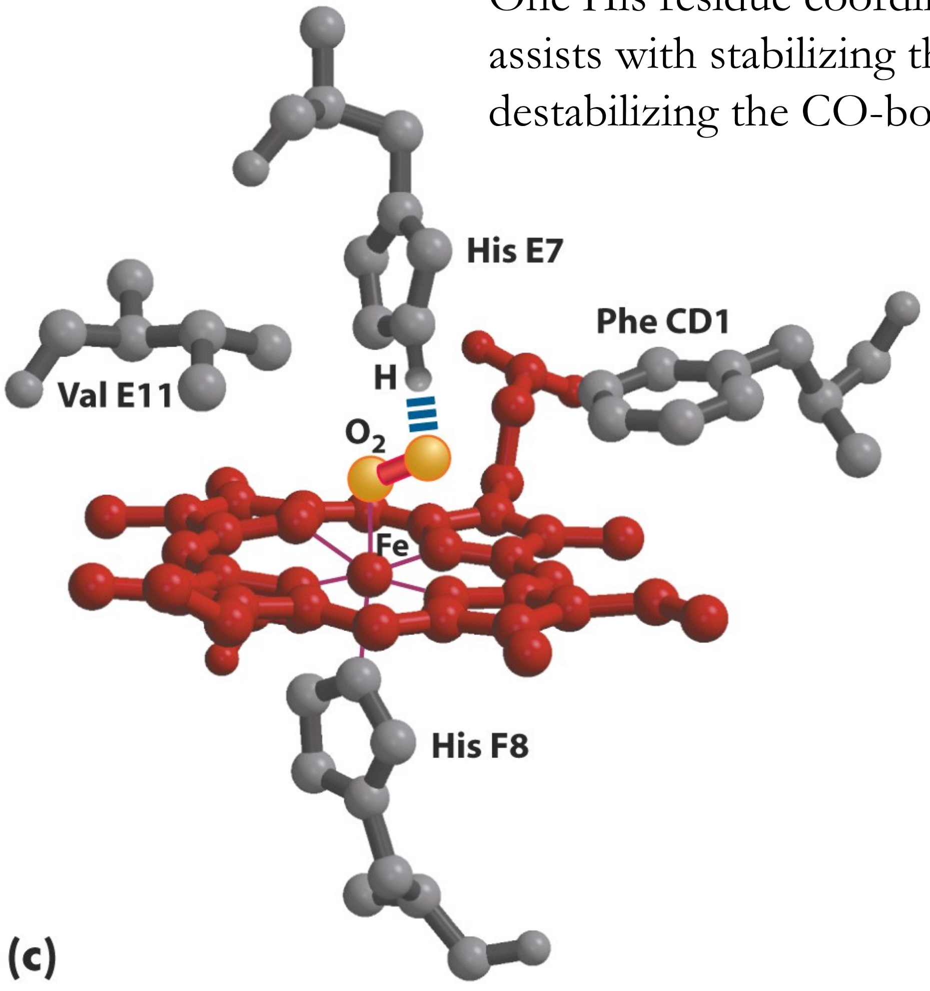


b. Chymotrypsin EC 3.4.21.1

Subtilisin and chymotrypsin are both serine endopeptidases. They share no sequence identity, and their folds are unrelated. However, they have an identical, three-dimensionally conserved Ser-His-Asp catalytic triad, which catalyses peptide bond hydrolysis. These two enzymes are a classic example of convergent evolution.

Functional sites: Oxygen-binding site

One His residue coordinates the iron. The second one assists with stabilizing the O2-bound form and also destabilizing the CO-bound form.



Computational function prediction methods

Major challenges

- The multifunctional nature of proteins
  - → proteins have multiple domains hosting different function
  - → some domain host several functions
- The functional sites in proteins may be
  - – better conserved than global sequence
    - → low sequence similarity between functionally similar proteins
  - – better conserved than global fold
    - → the same function may be hosted by different folds
- … but in some cases functional sites may be
  - – less conserved than global sequence
    - → highly similar sequences do not have the same function
  - – less conserved than global fold
    - → the same fold may host different functions

Computational function prediction methods

➢ Sequence-based

- Sequence alignment: Transfer function information from a known protein with high sequence similarity to the target
- Sequence-motifs: Extract function-specific sequence profiles from conserved sites and use these to assign functional classes to targets

➢ Structure-based

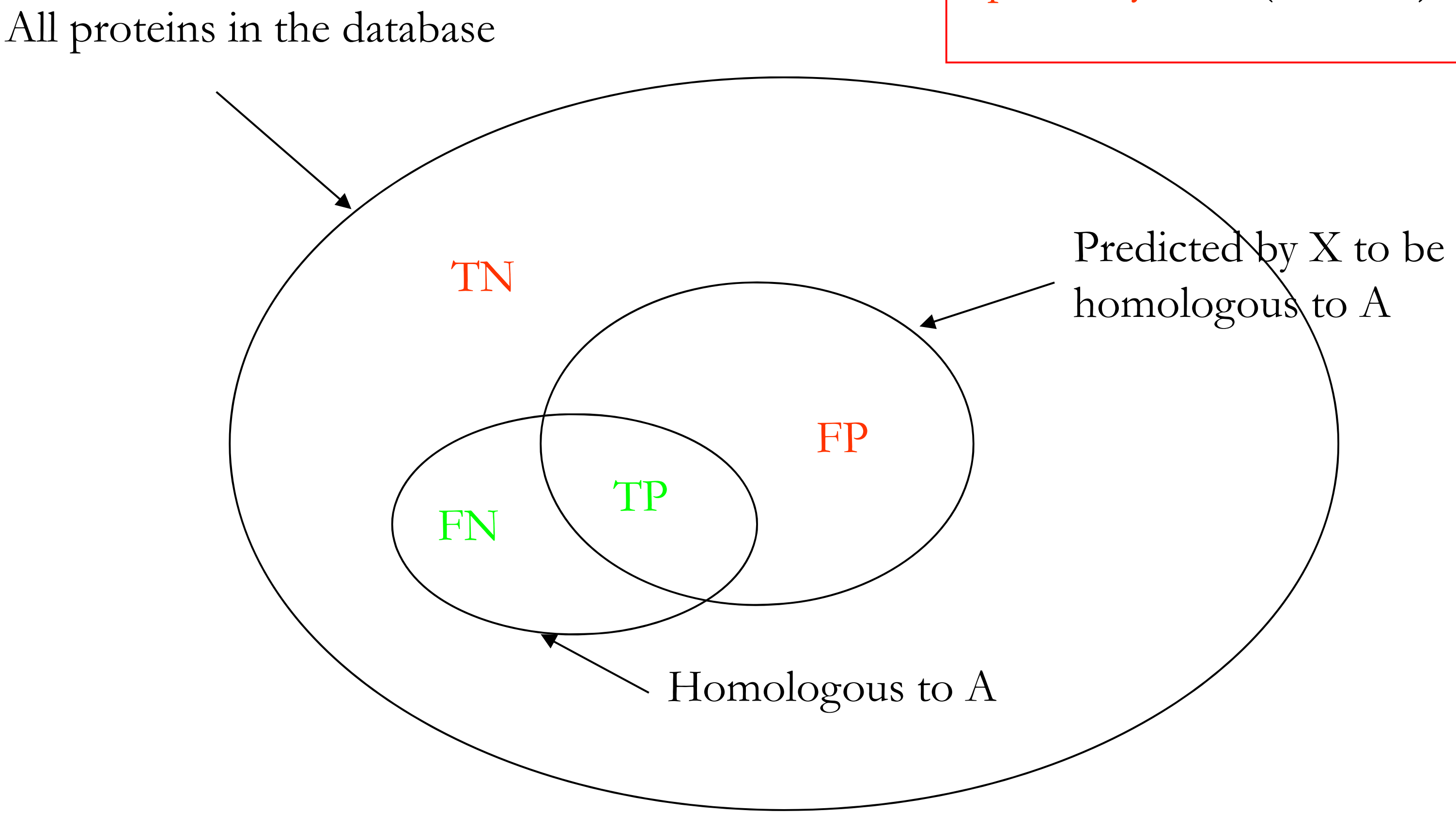
- Structure alignment: Transfer function information from a known protein with high structure similarity to the target
- Structure-motif: Use 3D templates of functional sites, scan the target structure and assign function

Power of computational methods

You want to find homologous proteins to a specific protein A using some computational method X:

Sensitivity: TP/(TP+FN)

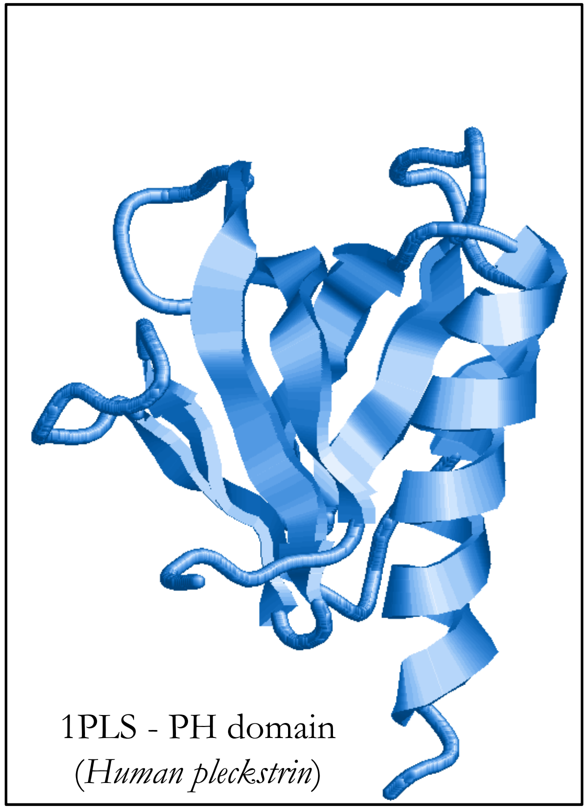
Specificity: TN/(TN+FP)


All proteins in the database
TN
Predicted by X to be homologous to A
FP
TP
FN
Homologous to A

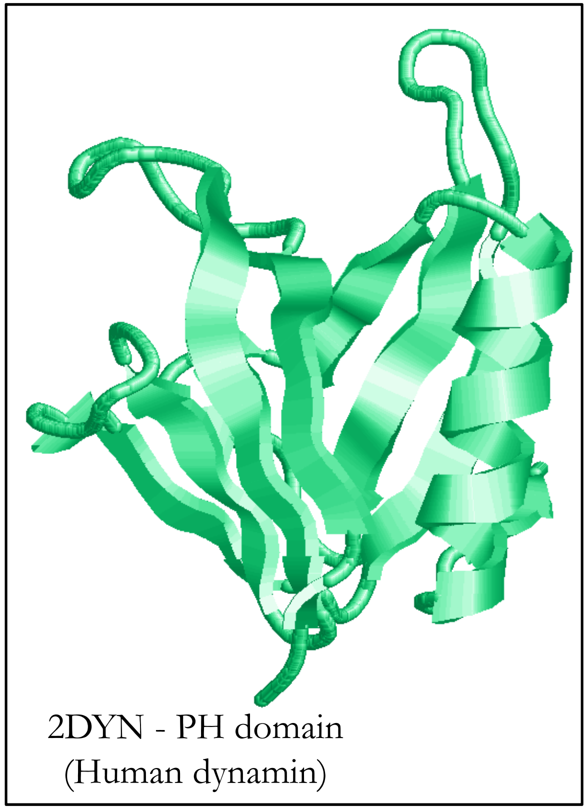
Example method: Global structure similarity

1PLS/2DYN:

23% sequence identity



1PLS - PH domain
(*Human pleckstrin*)



2DYN - PH domain
(Human dynamin)

Example method: Global structure similarity

Dali
http://ekhidna.biocenter.helsinki.fi/dali_server/

CBM Severo Ochoa
CSIC

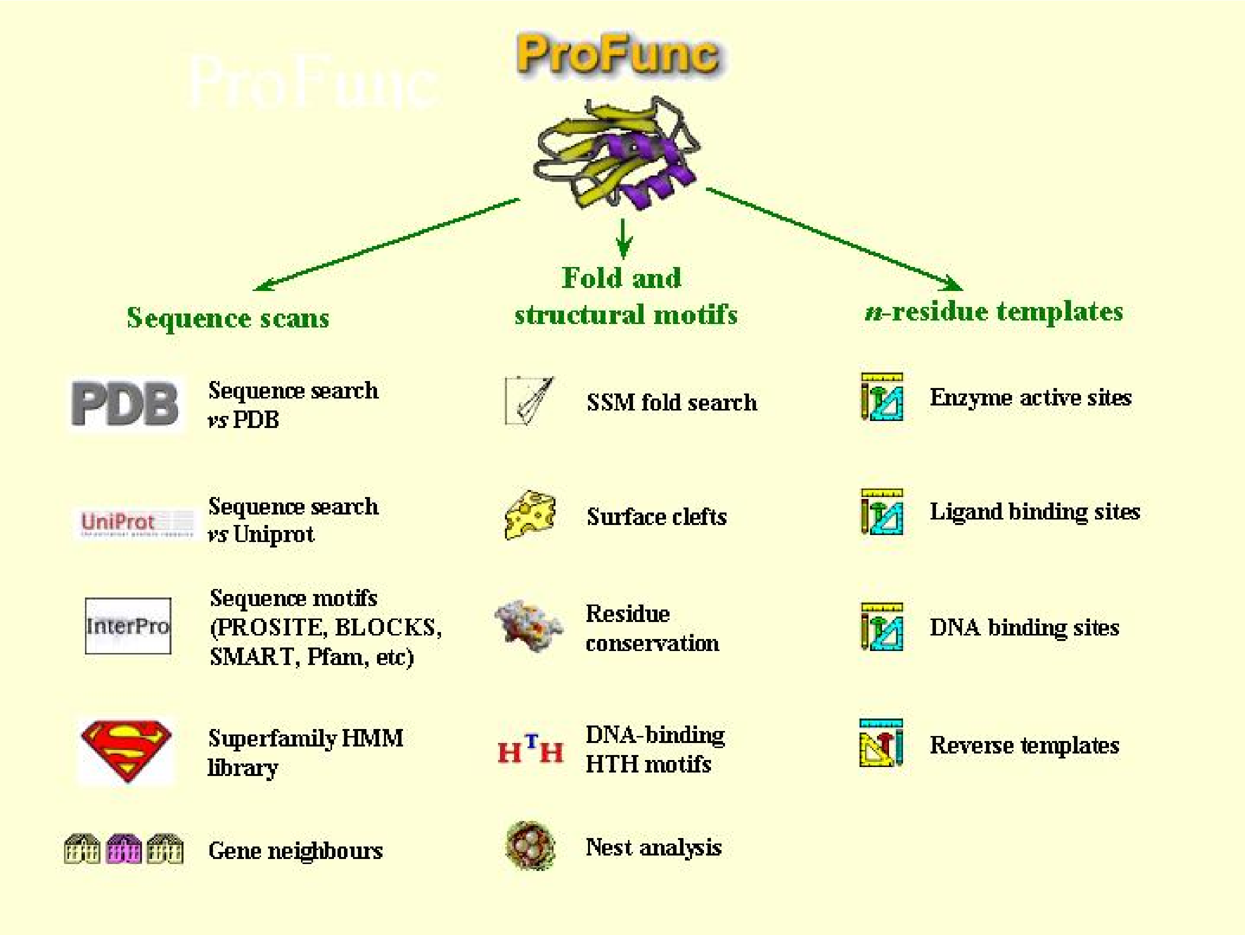
MAMMOTH-mult
Multiple Protein Structure Alignment Server

http://ub.cbm.uam.es/mammoth/pair/index3.php

Structural similarity between Calmodulin and Acetylcholinesterase

Example method: ProFunc

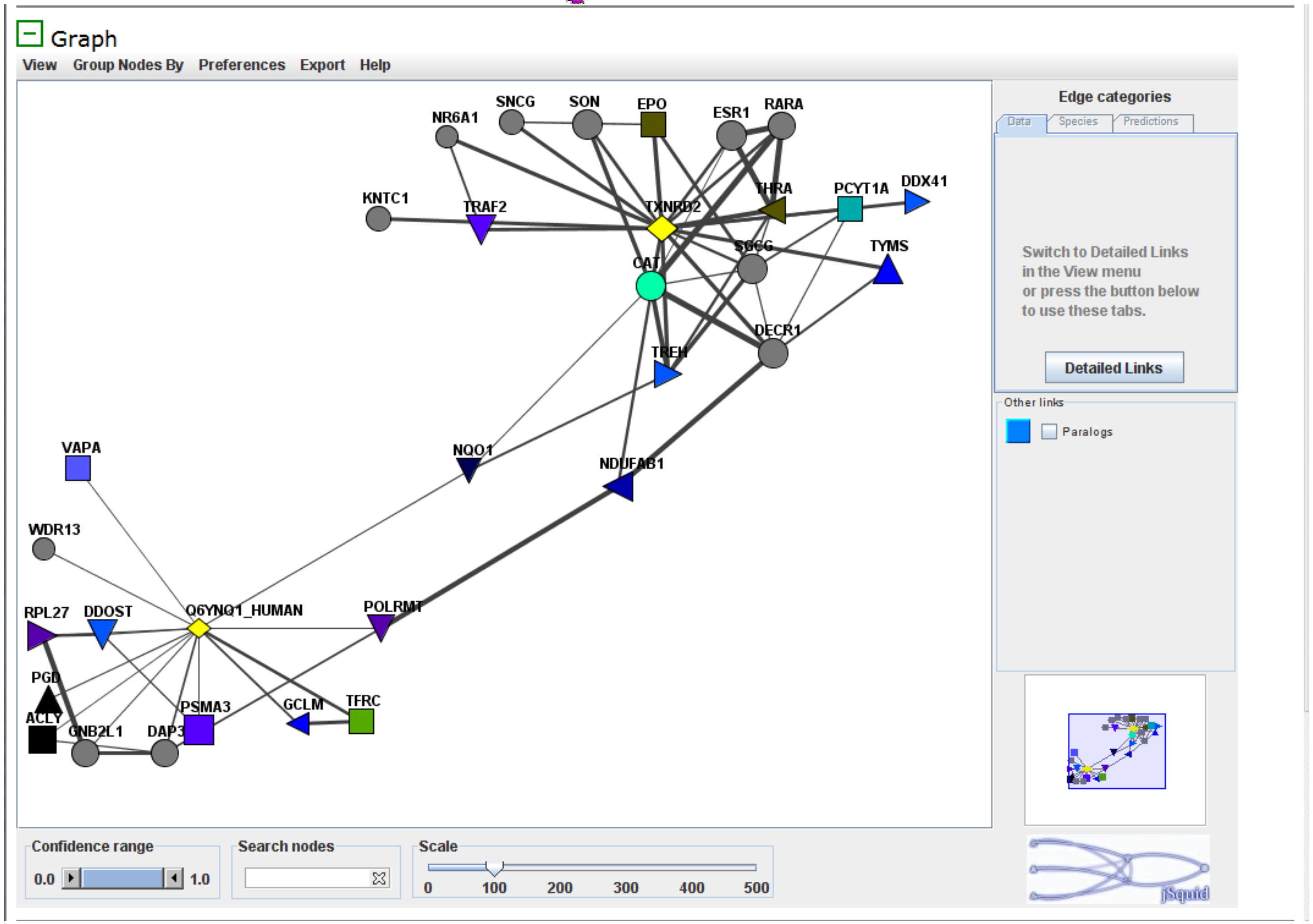
Successful function prediction methods are typcially meta-servers that combine many methods



http://www.ebi.ac.uk/thornton-srv/databases/ProFunc/

Example method:

FunCoup *networks of functional coupling*



Switch to Detailed Links in the View menu or press the button below to use these tabs.

http://funcoup.sbc.su.se/

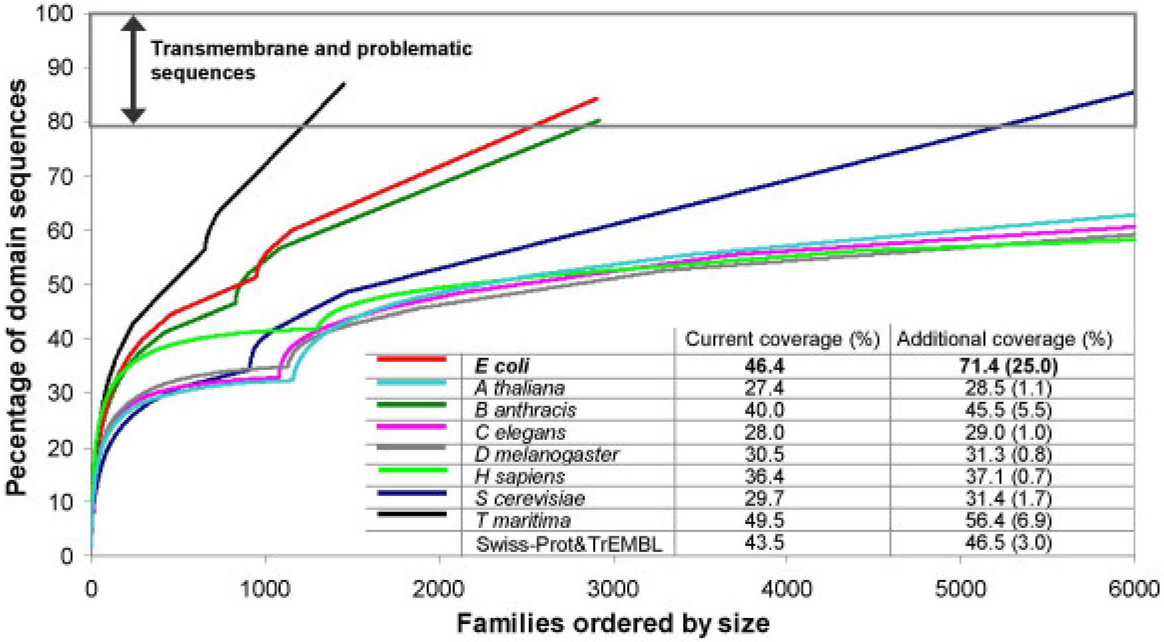
Structural Genomics

- The biggest limitation for predicting function from structure is the low availability of structure information

- Solution: Structural genomics
  - Solve experimentally the structure for a representative set of all protein sequences, e.g., one or a few proteins from each fold
  - Predict the structure for the remaining sequences using homology modeling, i.e., transfer structure from a structurally solved homology
  - Predict function from structure

- Structure prediction methods are better at predicting the core of proteins than the loops

Structural Genomics



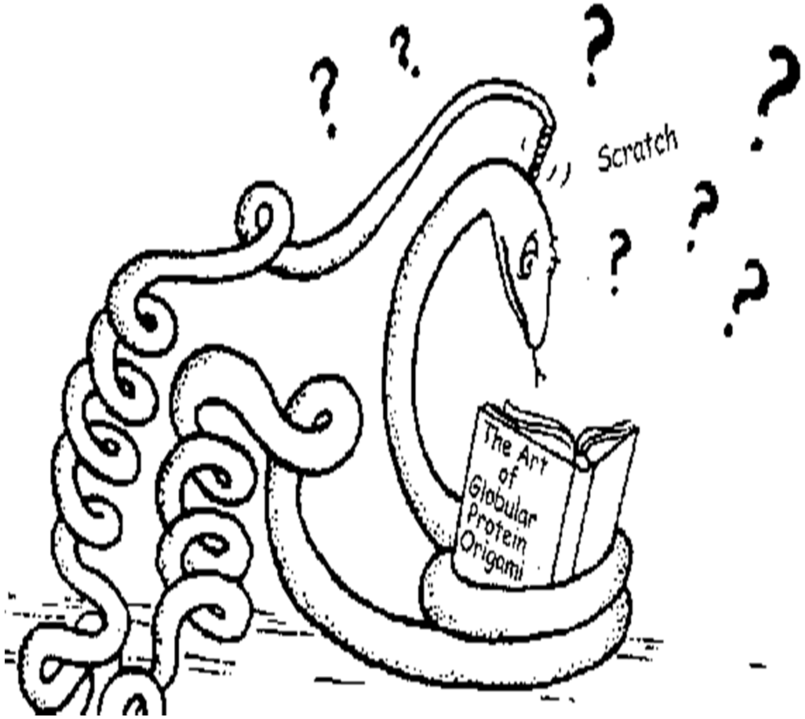
	Current coverage (%)	Additional coverage (%)
E coli	**46.4**	**71.4 (25.0)**
A thaliana	27.4	28.5 (1.1)
B anthracis	40.0	45.5 (5.5)
C elegans	28.0	29.0 (1.0)
D melanogaster	30.5	31.3 (0.8)
H sapiens	36.4	37.1 (0.7)
S cerevisiae	29.7	31.4 (1.7)
T maritima	49.5	56.4 (6.9)
Swiss-Prot&TrEMBL	43.5	46.5 (3.0)

Marsden, Lewis and Orengo. Towards a comprehensive structural coverage of completed genomes: a structural genomics viewpoint. BMC Bioinformatics8: 86, 2007.

A domain sequence is structurally annotated if it can be assigned to a CATH or Pfam-A_struc family through the use of hidden Markov model searches

The protein folding problem

Anfinsen's thermodynamic hypothesis (1973):
Protein folding is a strictly physical process that solely depends on the protein sequence



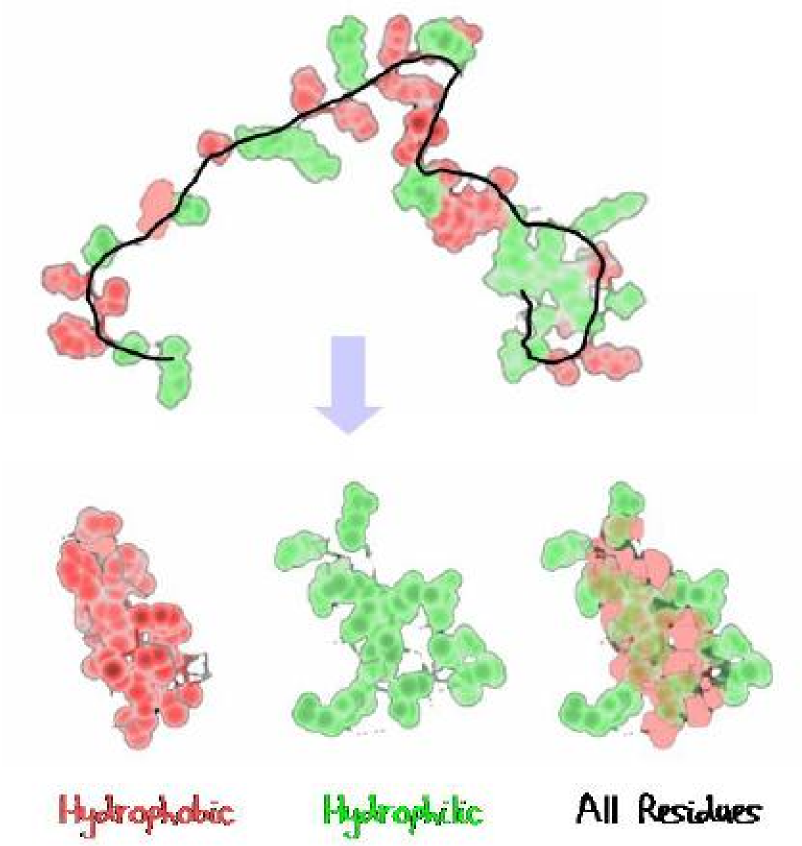
The folding problem:
discover nature’s algorithm for specifying 3D structure of proteins from their amino acid sequences

Hydrophobic interactions (I)

- Atomic charges dictate how folds occur
- Groups of C-H atoms have little charge
  - Called hydrophobic or non-polar
- Hydrophobic groups pack together
  - To avoid contact with solvent (aqueous solution)
  - To minimise energy
- Hydrophobic and hydrophilic regions are the main driving force behind the folding process

Hydrophobic interactions (II)

- Hydrophobicity vs. hydrophilicity
- Van der Waals interaction
- Electrostatic interaction
- Hydrogen bonds
- Disulfide bonds



Folding is directed mainly by internal residues

- Mutations that change surface residues are accepted more frequently and are less likely to affect protein conformations than are changes of internal residues
- This is consistent with the idea of hydrophobic force-driven folding

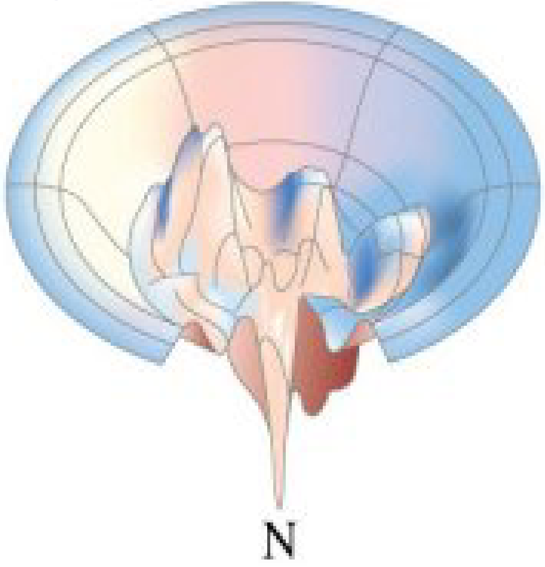
Molten globule

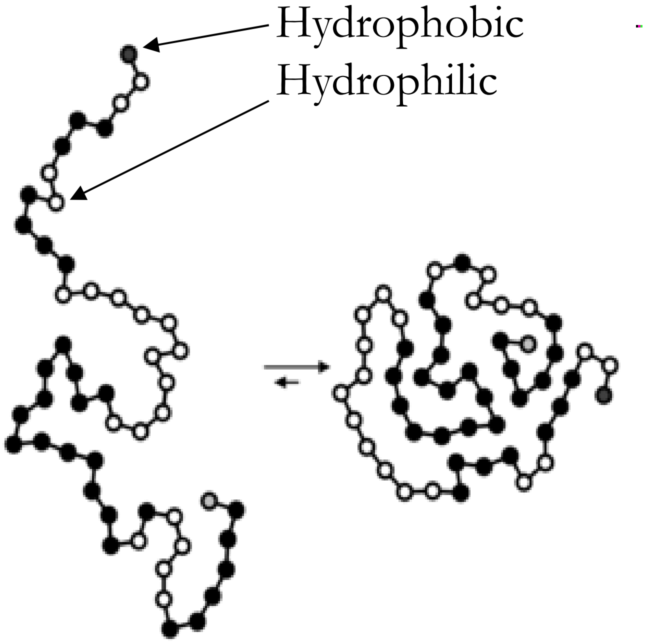
- Phase 1: Much of the secondary structure that is present in a native proteins forms within a few milliseconds
- Phase 2: Hydrophobic collapse into the Molten globule
  - Slightly larger (5-15% in radius) than the native conformation
  - Significant amount of secondary structure formed
  - Side chains are still not ordered/packed
  - Structure fluctuation is much larger - not very thermodynamically stable

Computational folding methods

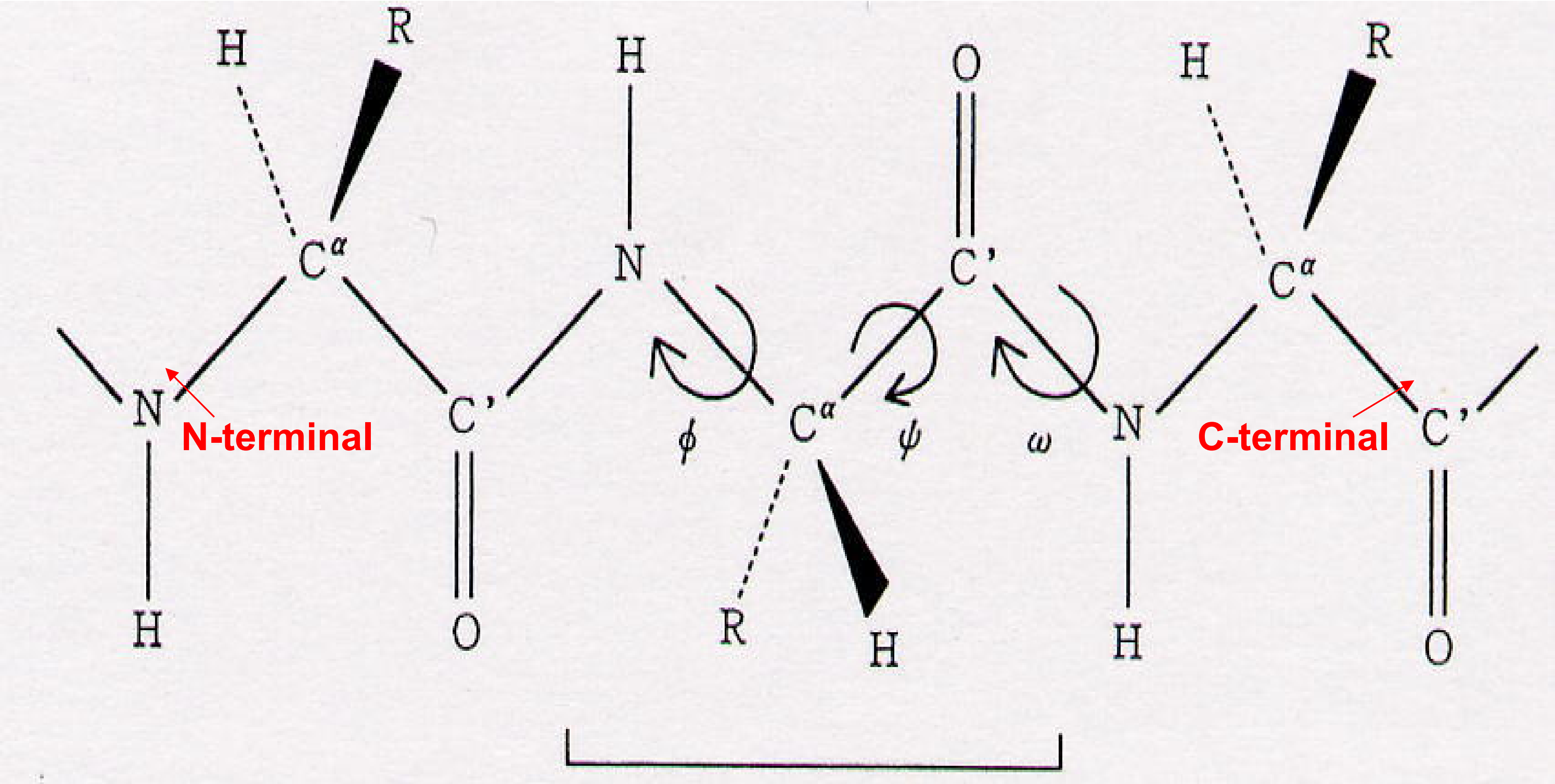
- No effective folding machine exists that is based on physical principles and energy minimization alone

- Current computational methods rely on known protein structures – machine learning approach:
  - Template-based modeling
  - Template-free modeling





Structure represented by angels



Protein folding

- Levinthal’s paradox
  - If for each residue there are only two degrees of freedom (ψ,φ)
  - Assume each can have only 3 stable values
  - This leads to 3^{2n} possible conformations
  - If a protein can explore 10^{13} conformation per second (10 per picosecond)
  - Still requires an astronomical amount of time to fold a protein
- Conclusion: proteins must fold in a way that does not randomly explore each possible conformations!

Structure prediction

- Protein structure prediction is the "holy grail" of bioinformatics

- Since structure is so important for function, solving the structure prediction problem should allow protein design, design of inhibitors, etc

- Huge amounts of genome data - what are the functions of all of these proteins?

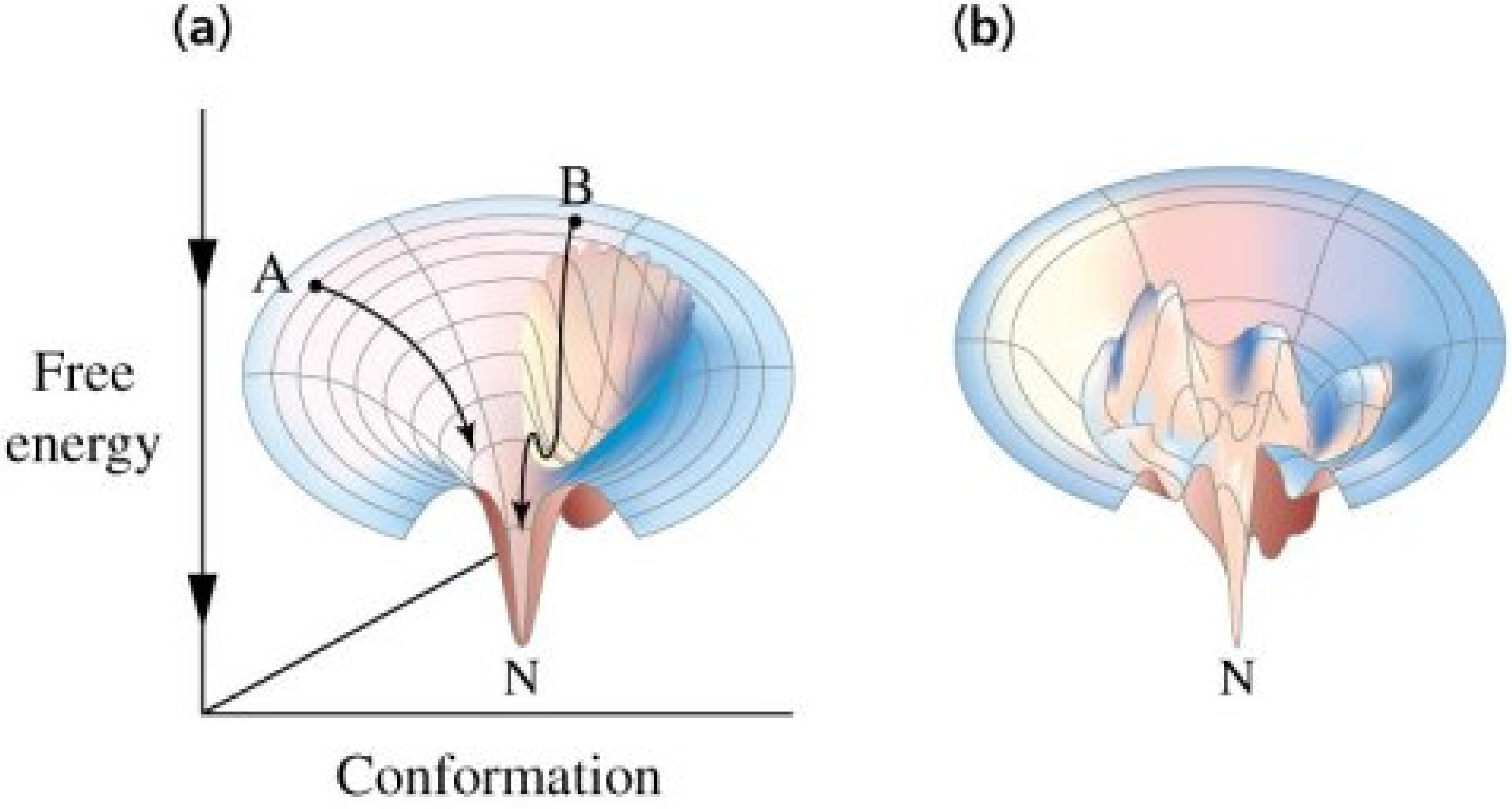
Assumptions

- ➢ Assumption 1: All the information about the structure of a protein is contained in its sequence of amino acids
- ➢ Assumption 2: The structure that a (globular) protein folds into is the structure with the lowest free energy
- ➢ Finding native-like conformations require:

  - A scoring function (potential)

  - A search strategy.

The free energy surface of a protein



Physics-based protein simulation

- All atom quantum mechanics (QM) calculation is not feasible
- QM can be applied to a small set of atoms
  - Modeling of an active site
  - Can get total energies (binding vs. non-binding, pK_a etc.), wave function (charge distribution)
  - QM/MM simulations (i.e. remaining atoms are treated with Molecular Mechanics)

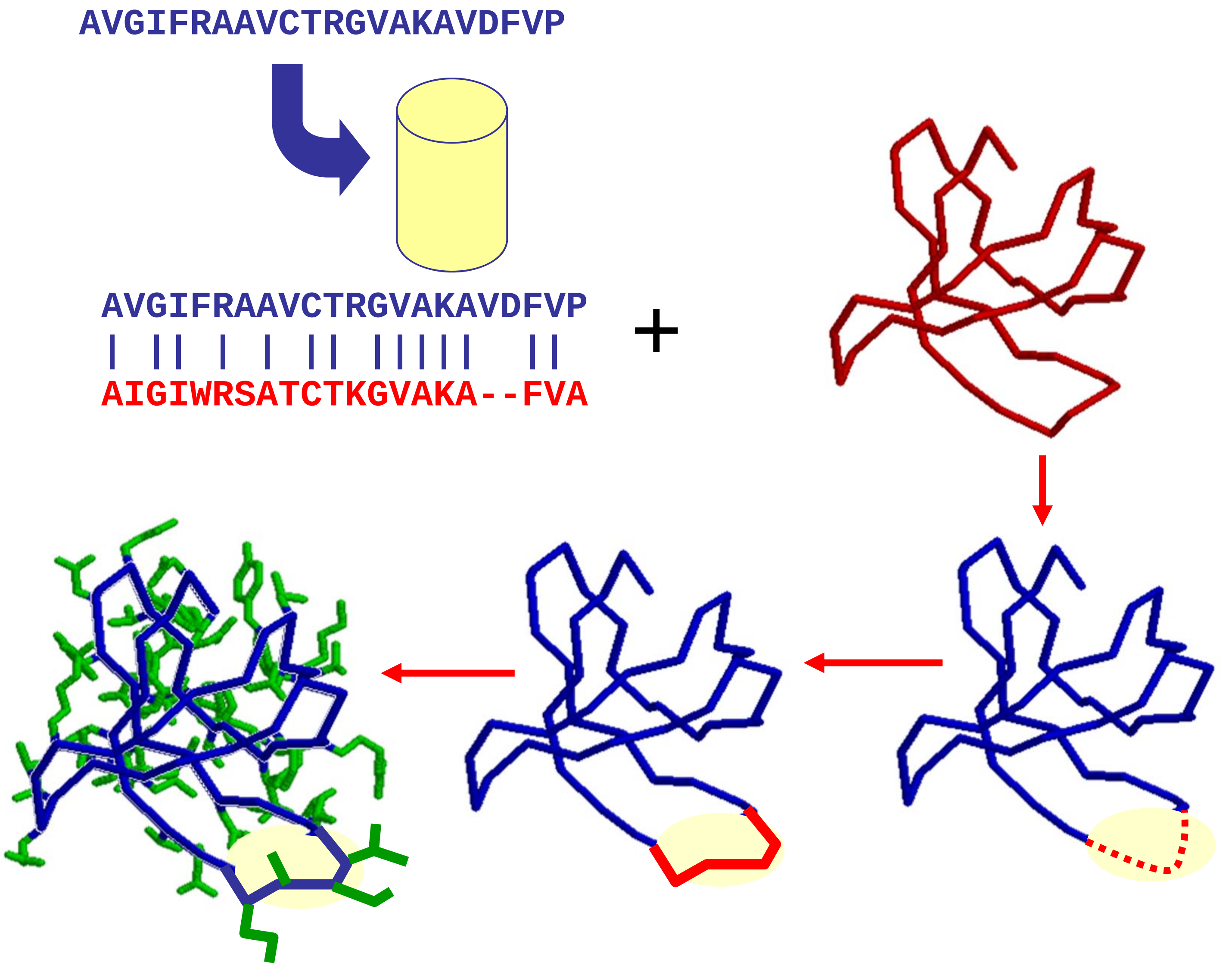
Problems

- Is the energy function correct?
  - Precise enough to discriminate non-native structure.
  - Yet simple enough for computers to carry out efficiently.
- Is the conformational search good enough to cover the global minimum?
- Protein folding without any prior knowledge about protein structure is a difficult task.
- Protein structure prediction is often quoted as an "NP complete problem", i.e. the complexity of the problem grows exponentially as the number of residues increases

Flavors of “knowledge-based” structure prediction

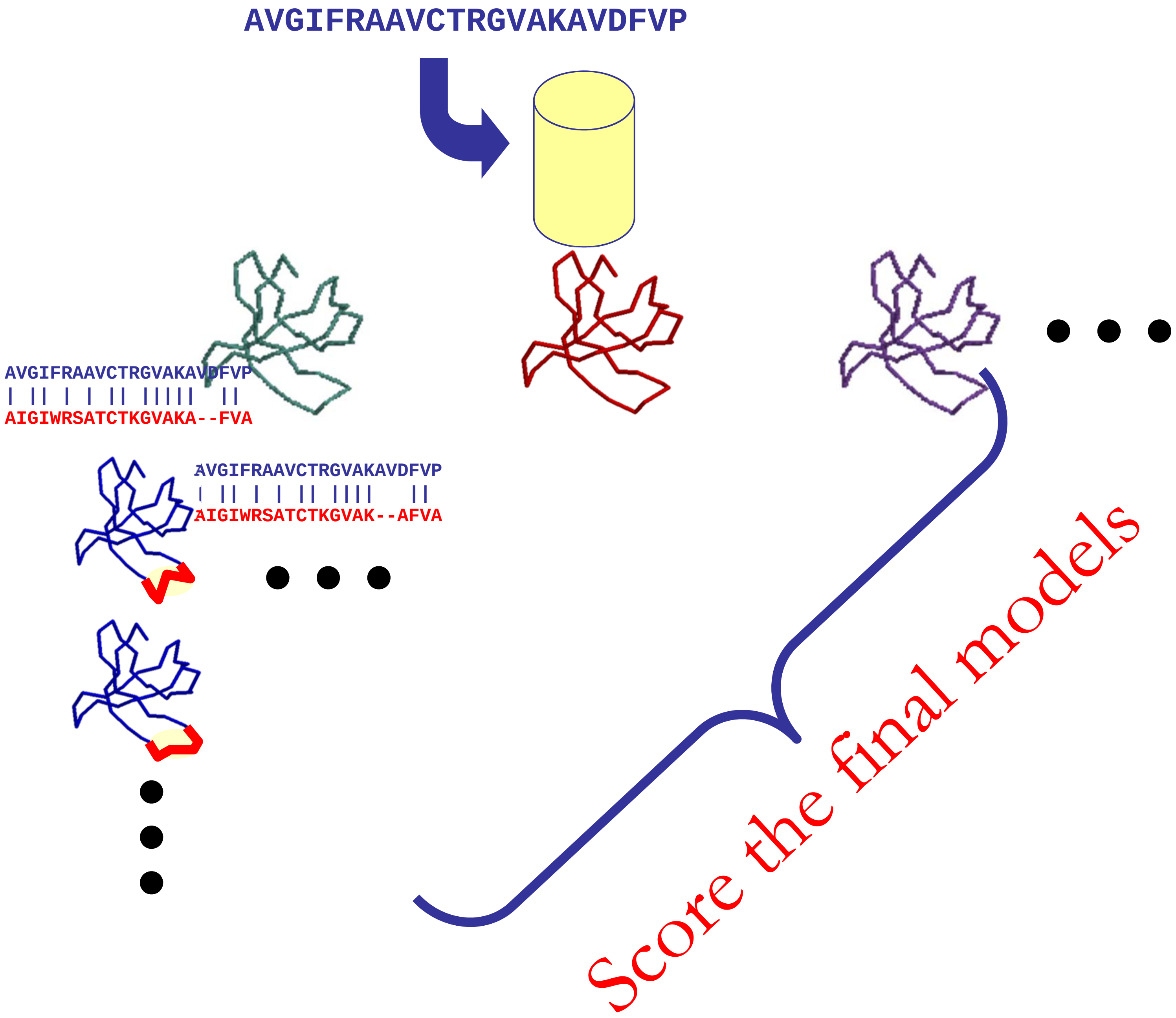
- Experimental data
  - X-ray crystallography
  - NMR spectroscopy
- Computational methods
  - Homology/comparative modeling
  - Fold recognition (threading)
  - Ab initio (de novo, new folds) methods (Ab initio: “from the beginning”.

Comparative modeling



```
AVGIFRAAVCTRGVAKAVDFVP

AVGIFRAAVCTRGVAKAVDFVP
| || | | || |||||  ||
AIGIWRSATCTKGVAKA--FVA
```



AVGIFRAAVCTRGVAKAVDFVP

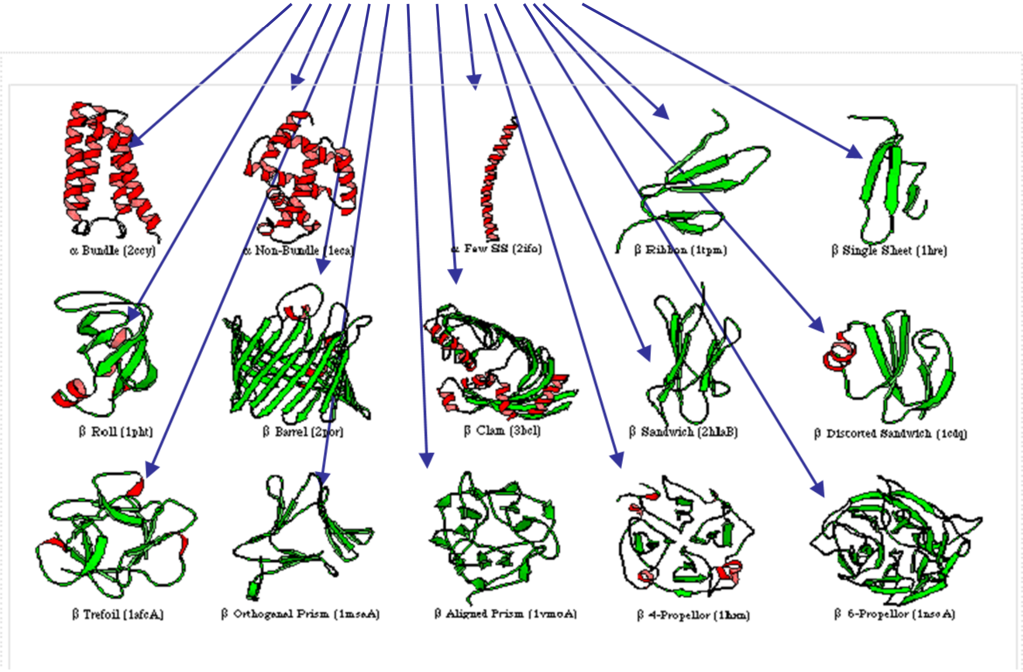
```
AVGIFRAAVCTRGVAKAVDFVP
| || | | || |||||  ||
AIGIWRSATCTKGVAKA--FVA
```

```
AVGIFRAAVCTRGVAKAVDFVP
| || | | || ||||   ||
AIGIWRSATCTKGVAK--AFVA
```

Score the final models

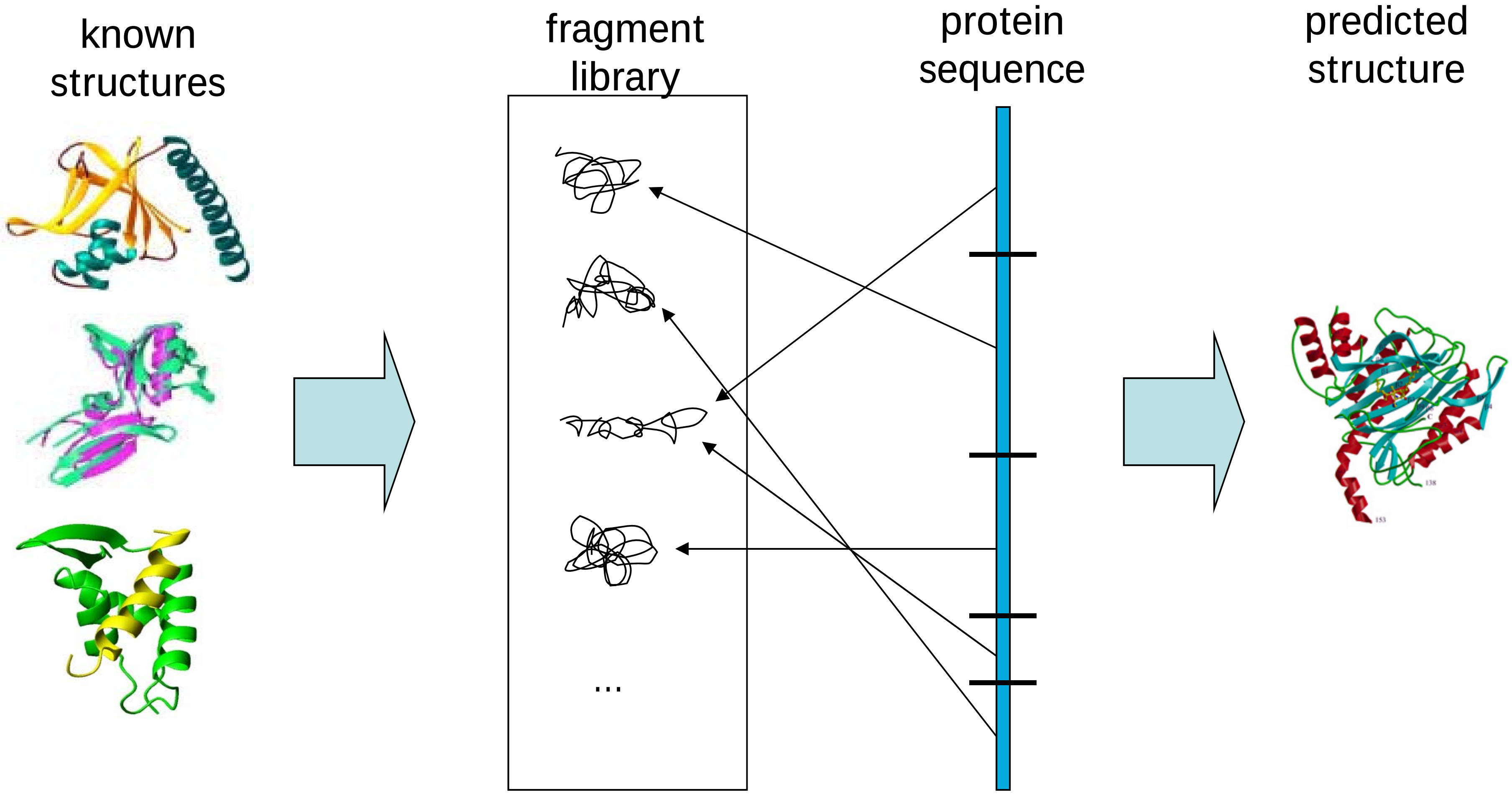
Fold recognition

```
AVGIFRAAVCTRGVAKAVDFVPVESMETTMRSPV
FTDNSSPPAVPQSFQVAHLHAPTGSGKSTKVPAA
YAAQGYKVLVLNPSVAATLGFGAYMSKAHGIDPN
IRTGVRTITTGAPVTYSTYGKFLADGGCSGGAYD
IIICDECHSTDSTTILGIGTVLDQAETAGARLVV
LATATPPGSVTVPHPNIEEVALSNTGEIP
```



Score and select model

Fragment assembly



known structures

fragment library

protein sequence

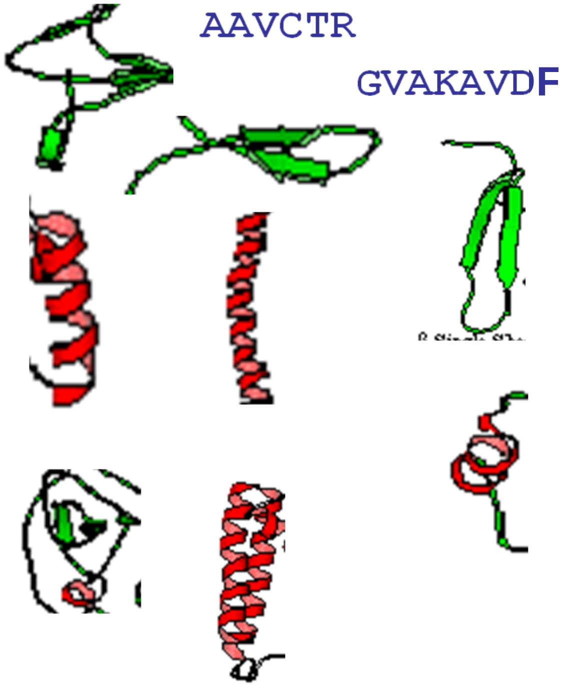
predicted structure

...

New fold/*ab initio* prediction

AVGIFRAAVCTRGVAKAVDFVP…

AVGIFR



AAVCTR

GVAKAVD**F**

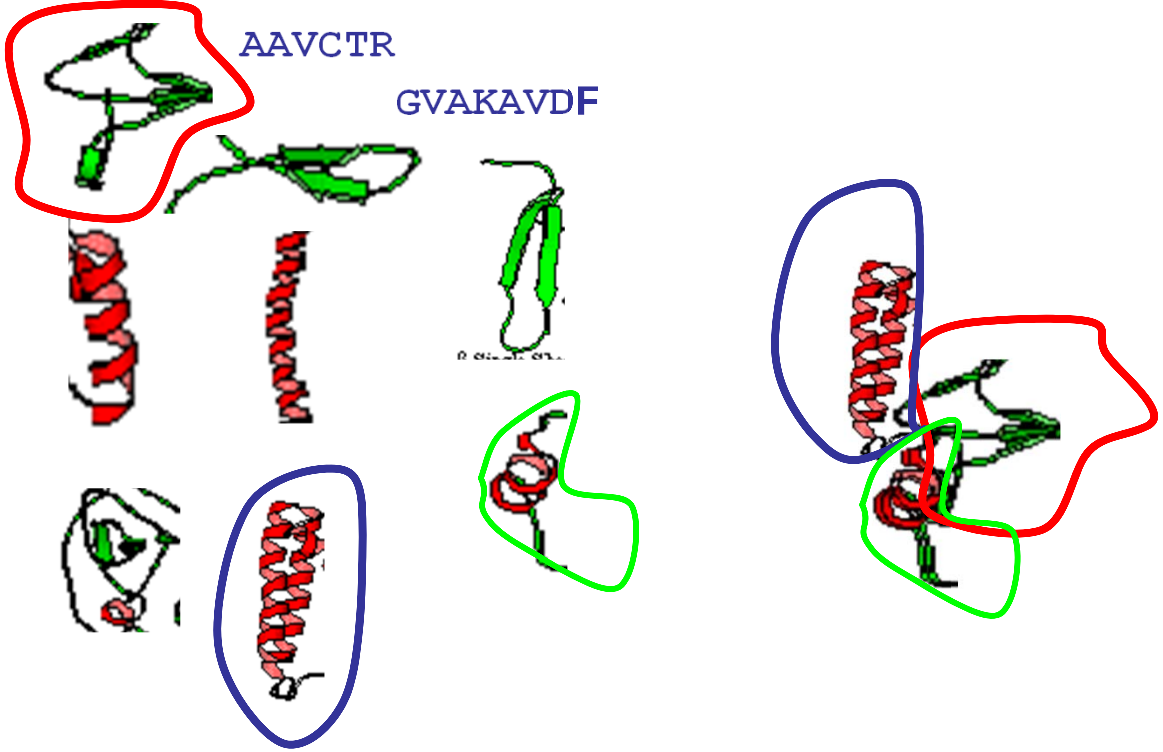
New fold/*ab initio* prediction

AVGIFRAAVCTRGVAKAVDFVP…

AVGIFR

AAVCTR

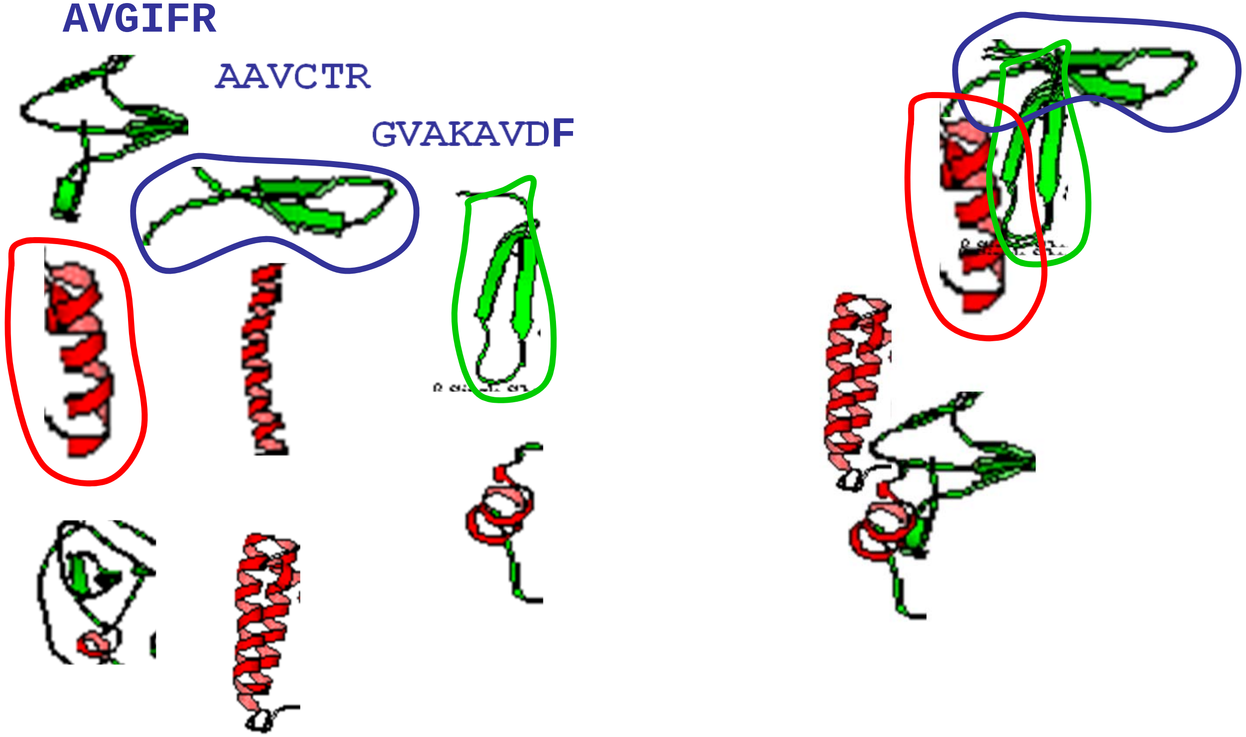
GVAKAVD**F**



New fold/*ab initio* prediction

AVGIFRAAVCTRGVAKAVDFVP…

AVGIFR



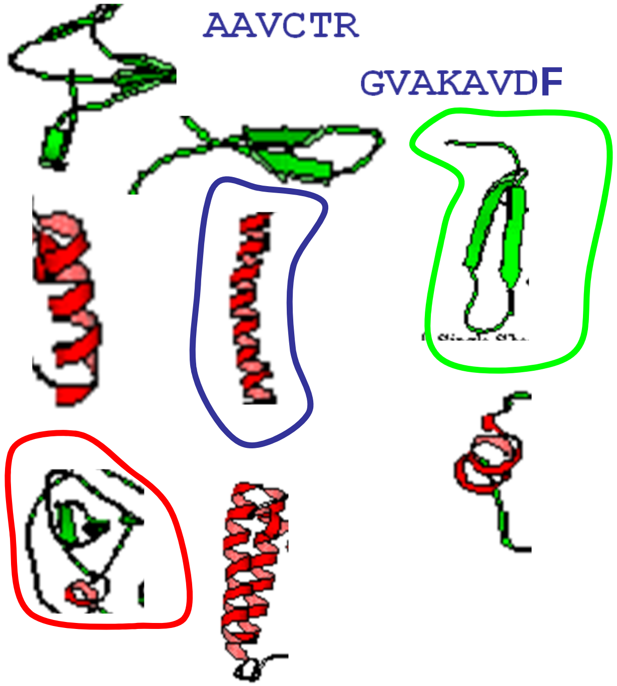
AAVCTR

GVAKAVD**F**

New fold/*ab initio* prediction

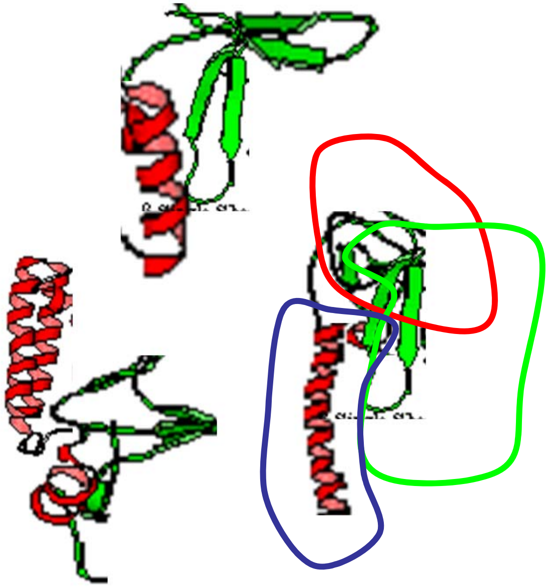
AVGIFRAAVCTRGVAKAVDFVP…

AVGIFR



AAVCTR

GVAKAVDF



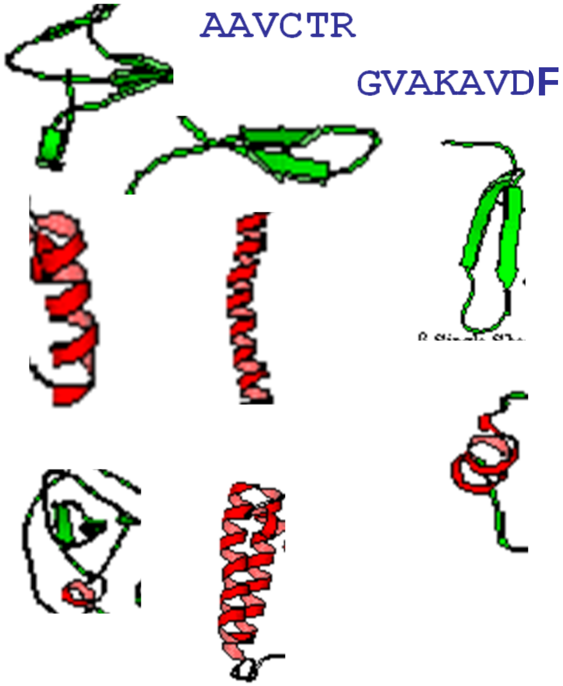
New fold/*ab initio* prediction

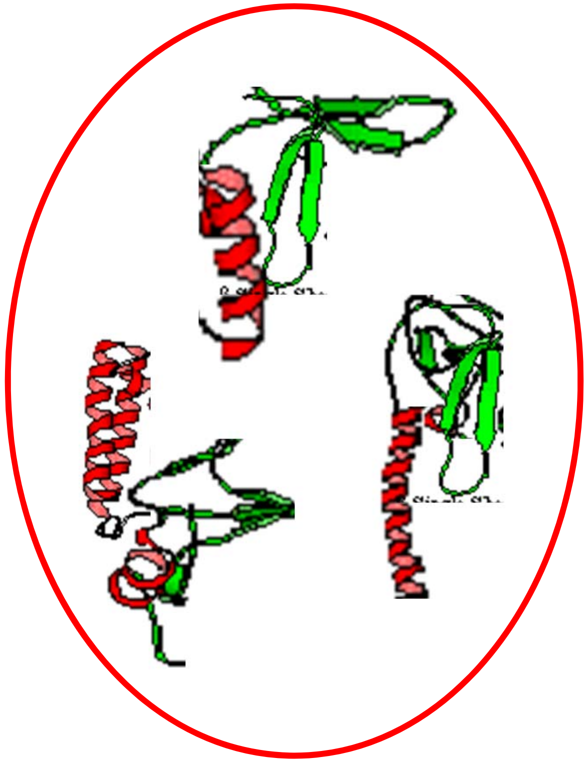
AVGIFRAAVCTRGVAKAVDFVP…

AVGIFR

AAVCTR

GVAKAVD**F**





Score and select model

CASP: Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction

http://www.predictioncenter.org/

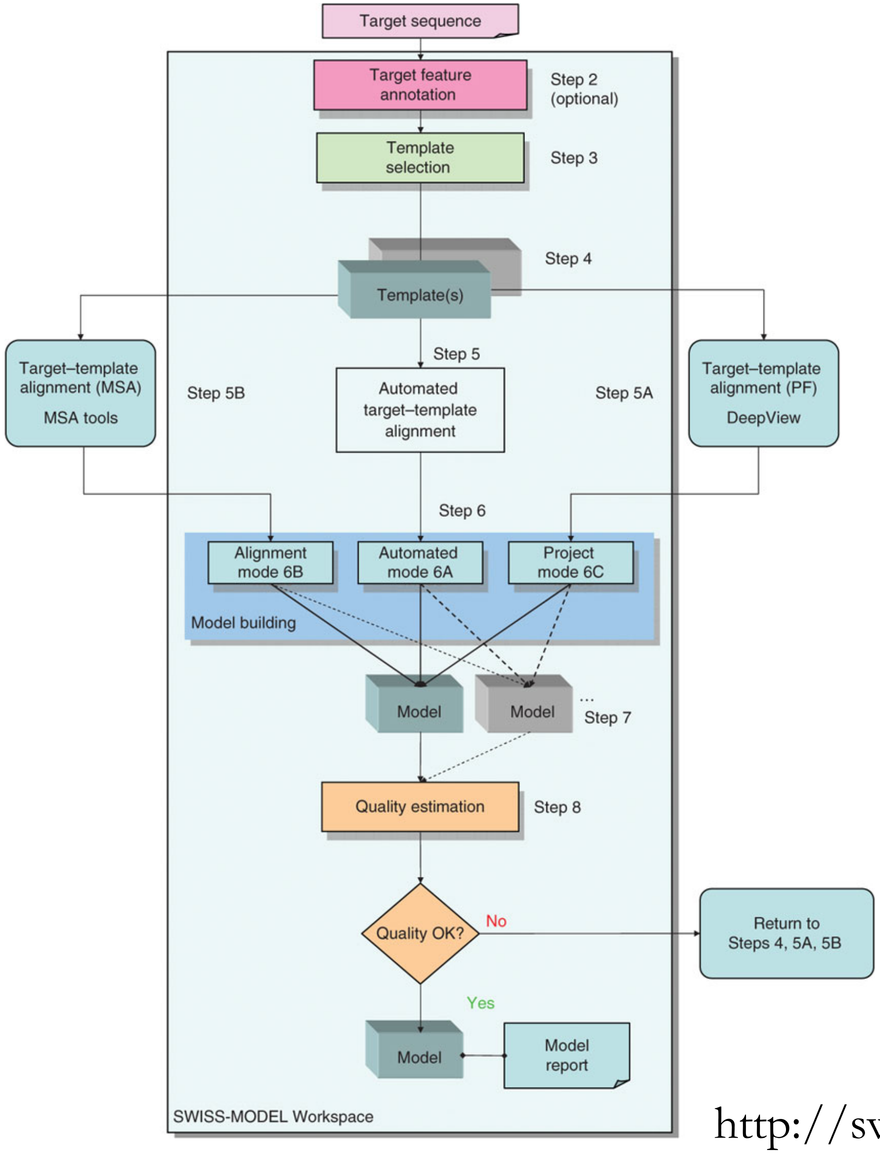
- Aim: obtain an in-depth and objective assessment of our current abilities and inabilities in the area of protein structure prediction
- Participants will predict the structure of a set of sequences soon to be known structures
- These will be true predictions, not ‘post-dictions’ made on already known structures.

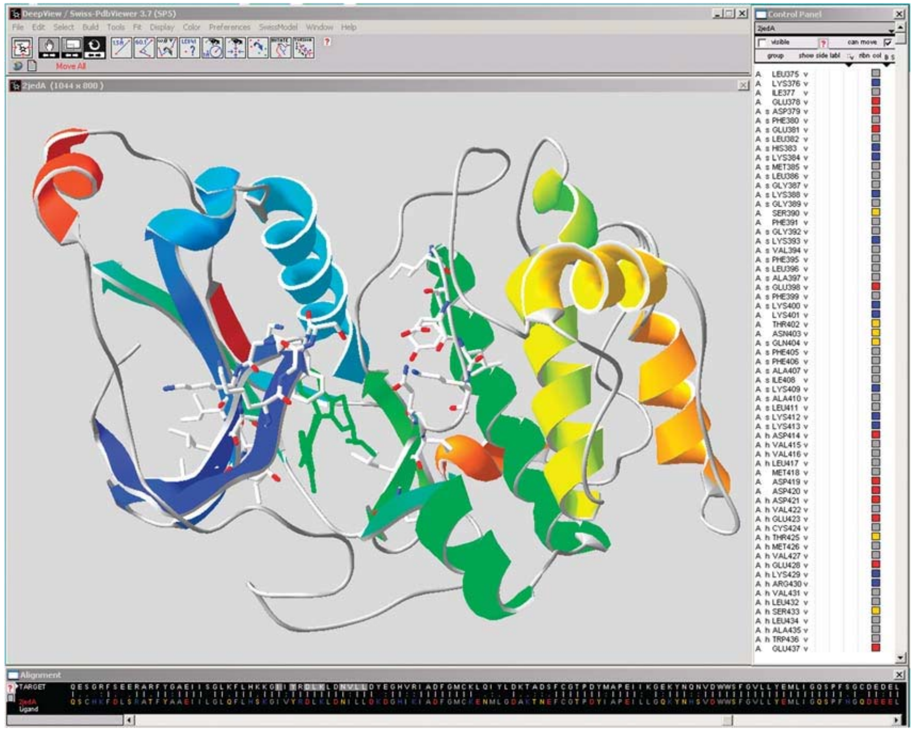
Meta-methods

- Meta-methods combine predictions from individual methods
  - E.g. 3D-Jury: http://bioinfo.pl/Meta/
- Range from methods that select the best prediction to methods that improve and combine other predictions
- Often include methods for all flavors of protein structure prediction

SWISS-MODEL

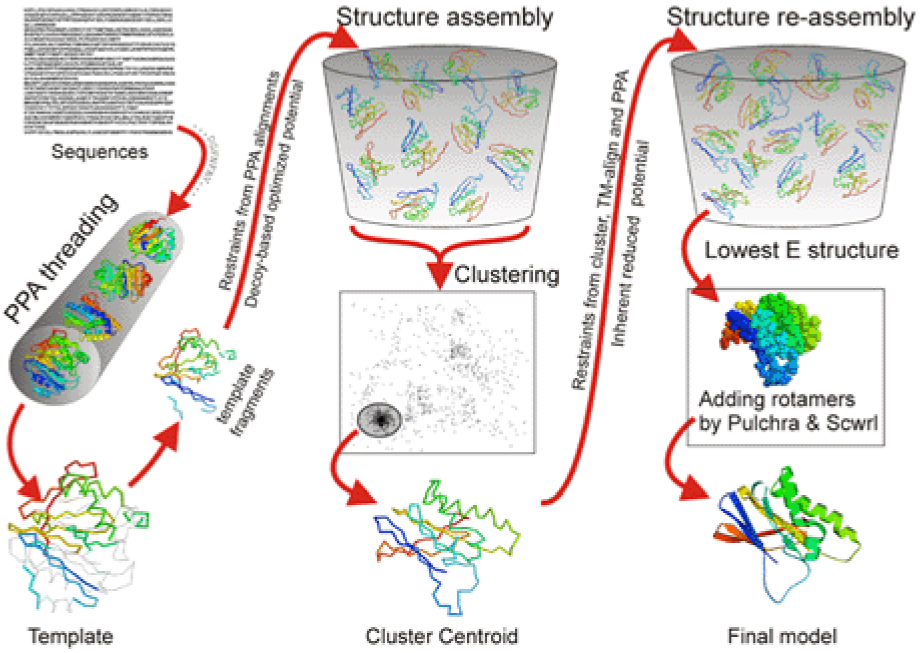
15 years




DeepView / Swiss-PdbViewer 3.7 (SP5)
File Edit Select Build Tools Fit Display Color Preferences SwissModel Window Help
Move All
2jedA (1044 x 800)
Control Panel
2jedA
visible
can move
group
show side labl
rbn col
Alignment
TARGET
2jedA
ligand

http://swissmodel.expasy.org//SWISS-MODEL.html

I-TASSER

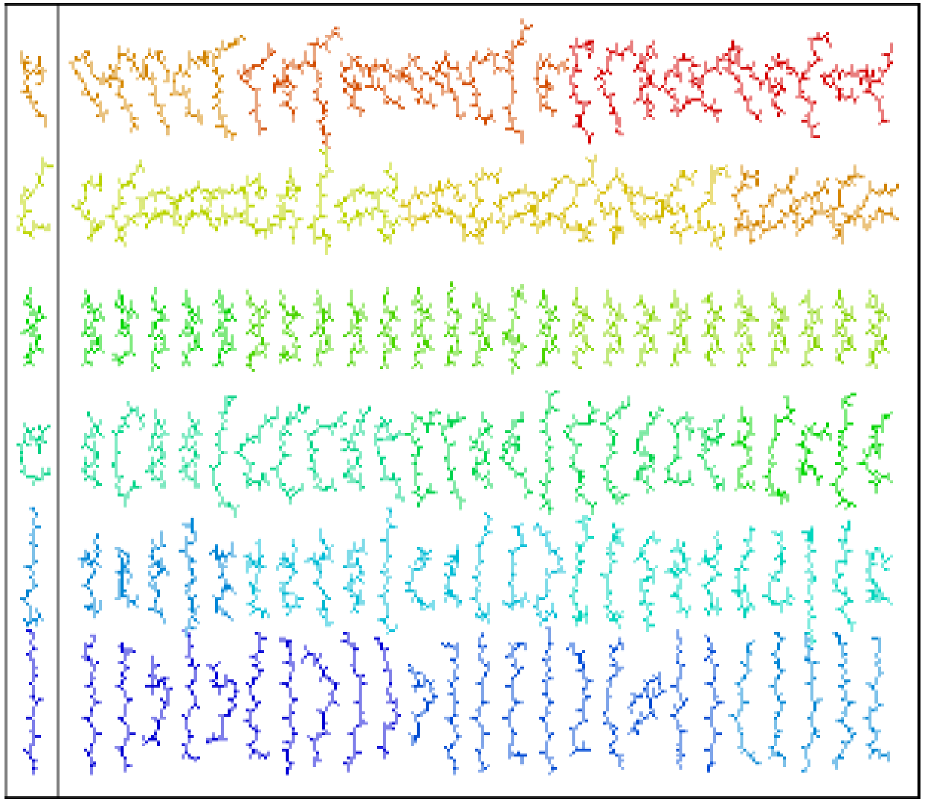

Sequences
PPA threading
Template
template fragments
Restraints from PPA alignments
Decoy-based optimized potential
Structure assembly
Clustering
Cluster Centroid
Restraints from cluster, TM-align and PPA
Inherent reduced potential
Structure re-assembly
Lowest E structure
Adding rotamers by Pulchra & Scwrl
Final model

http://zhang.bioinformatics.ku.edu/I-TASSER/

Rosetta/Robetta

- Decoys are assembled from fragments
- Lowest energy model from a set of generated decoys is selected as the prediction
- Monte Carlo simulated annealing
- Physical energy function with elements of a statistical potential

Fragment library



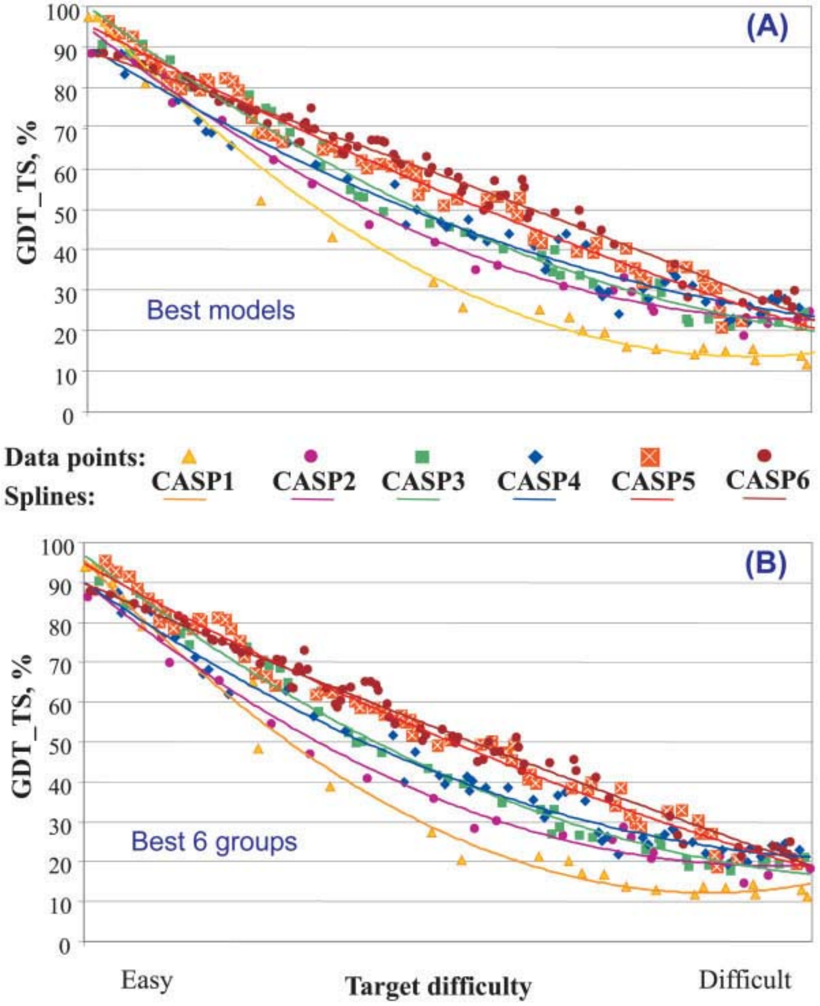
http://robetta.bakerlab.org/

CASP: progress

- Most progress in the fold prediction category and for servers over humans

- GDT_TS = (GDT_P1 + GDT_P2 + GDT_P4 + GDT_P8)/4,

  where GDT_Pn denotes percent of residues under distance cutoff <= nÅ



Kryshtafovych,Venclovas, Fidelis and Moult. Progress Over the First Decade of CASP Experiments. PROTEINS: Structure, Function, and Bioinformatics Suppl 7:225–236, 2005.