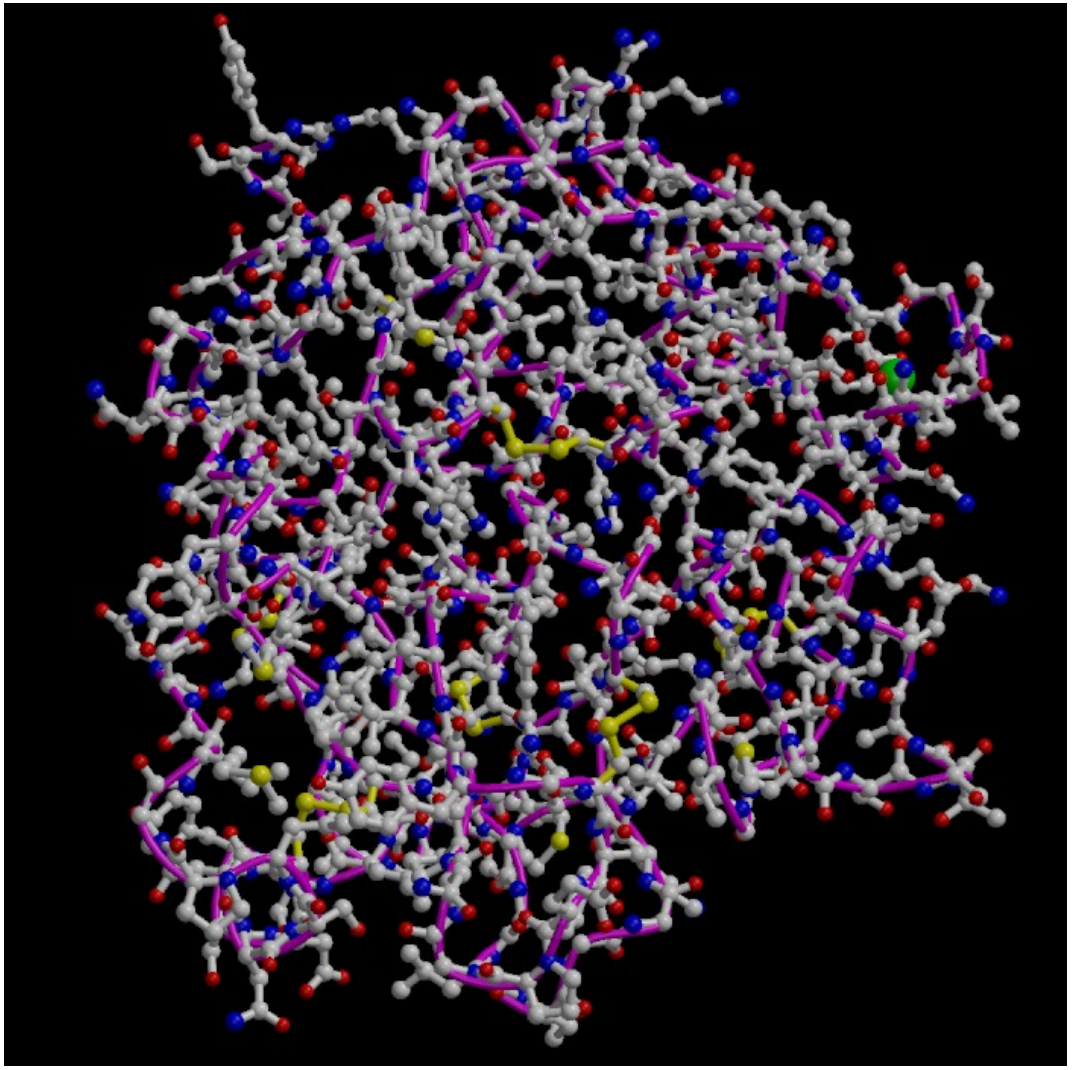
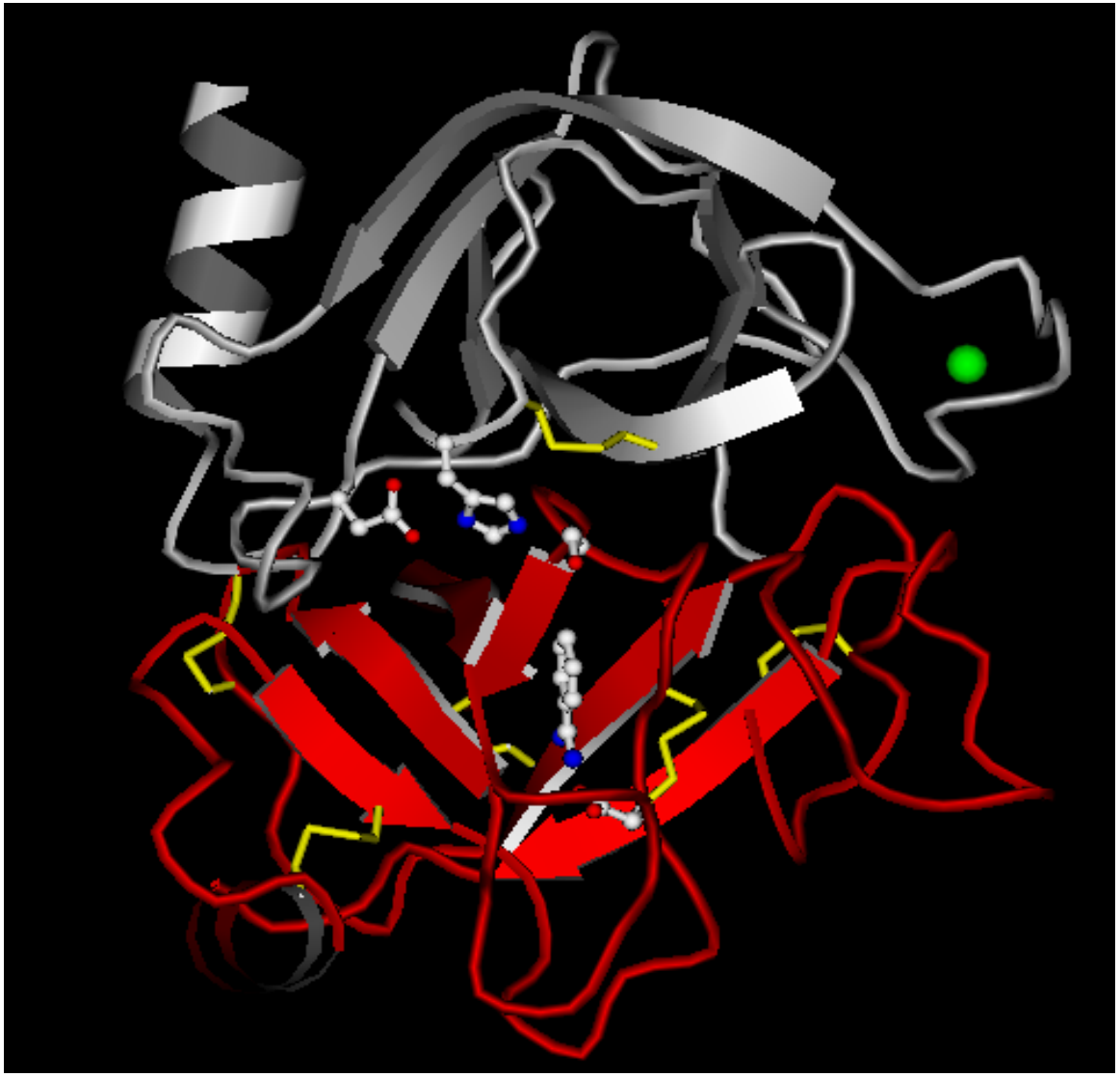
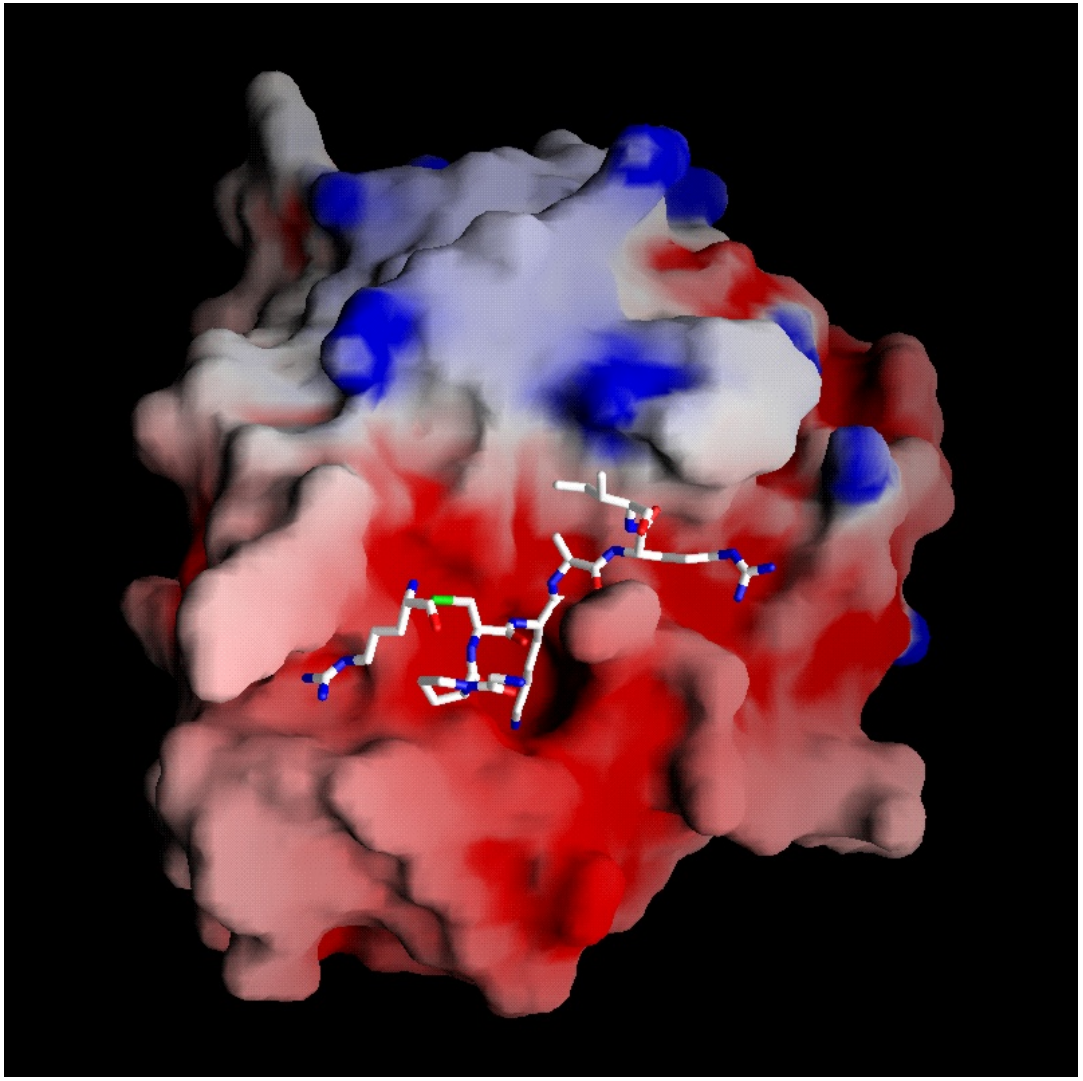
Structual informatics



Structual informatics



Structual informatics



Structual informatics

Primary structure

The primary structure of a protein is its linear sequence of amino acids and the location of any disulfide (-S-S-) bridges.

Secondary structure

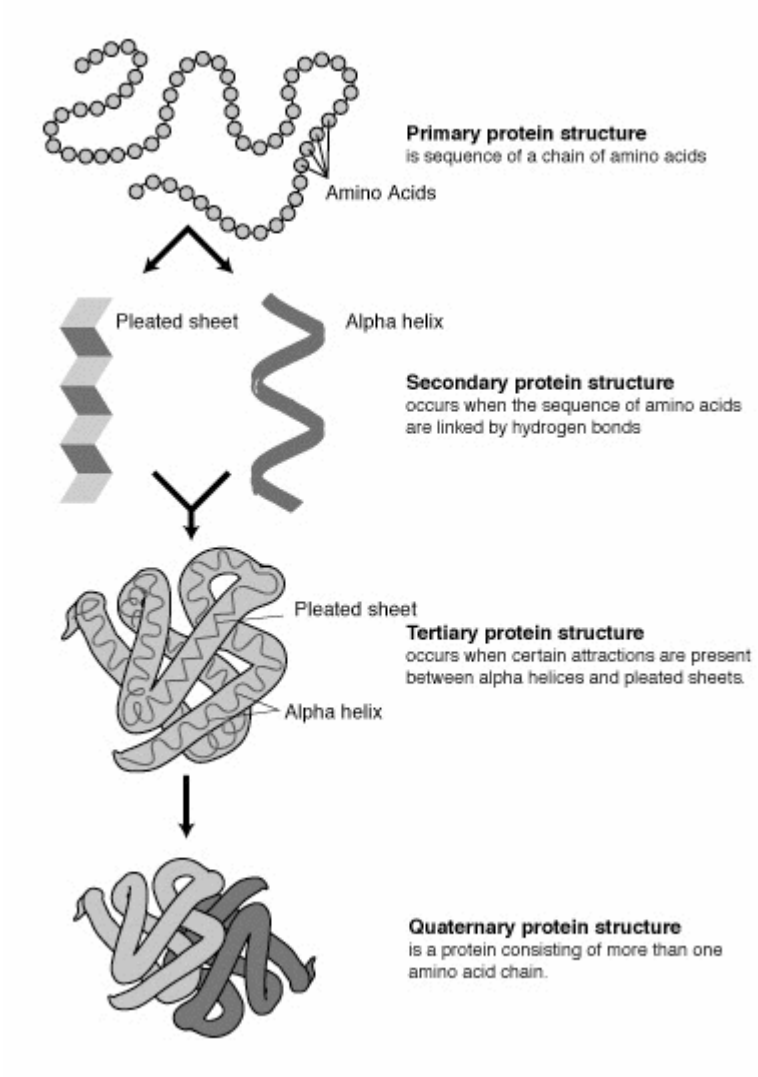
Most proteins contain one or more stretches of amino acids that take on a characteristic structure in 3-D space. The most common of these are the **alpha helix** and the **beta conformation**.

Tertiary structure

Tertiary structure refers to the three-dimensional structure of the entire polypeptide chain.

Quartenary structure

Complexes of 2 or more polypeptide chains held together by noncovalent forces but in precise ratios and with a precise 3-D configuration.



Structual informatics

Primary structure

Amino acid

-Building blocks of proteins

-20 different amino acids

-Common structure – side chains vary

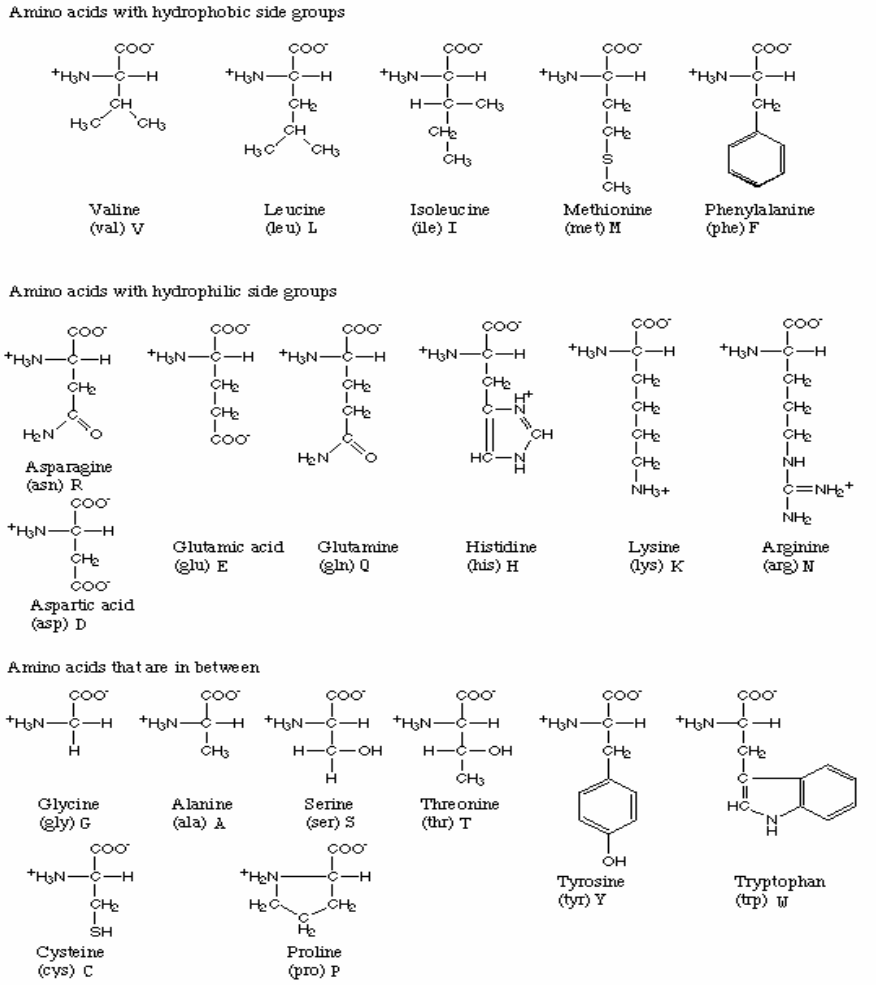
Many ways to classify amino acids

Hydrophilic – hydrophobic

Aromatic – aliphatic

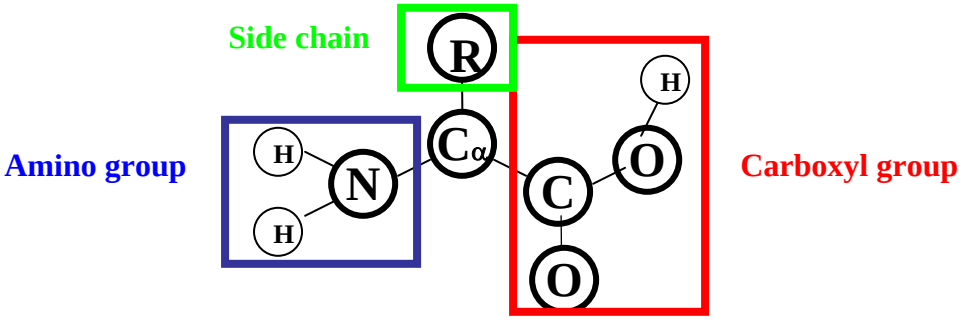
Charged – uncharged

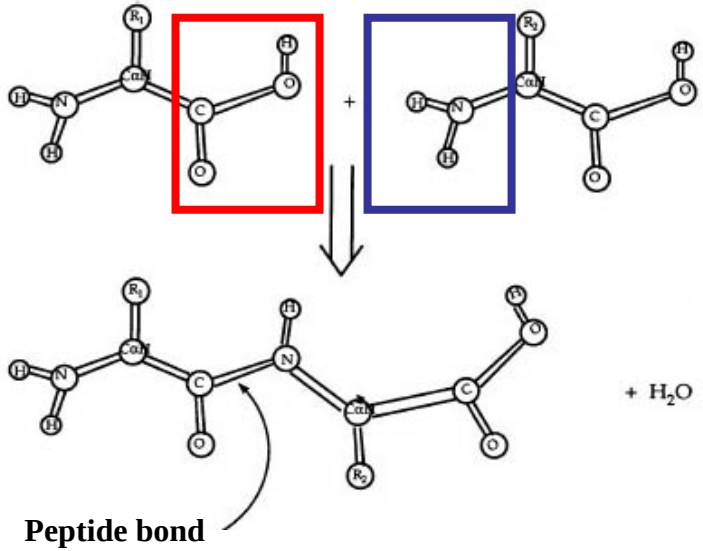
Small – large

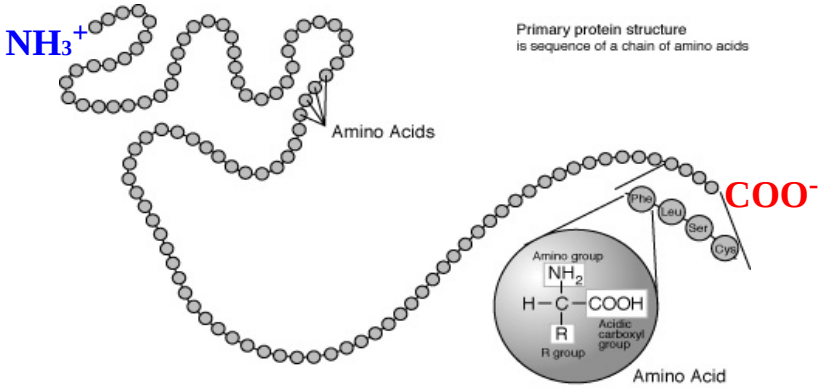


Structual informatics

Primary structure

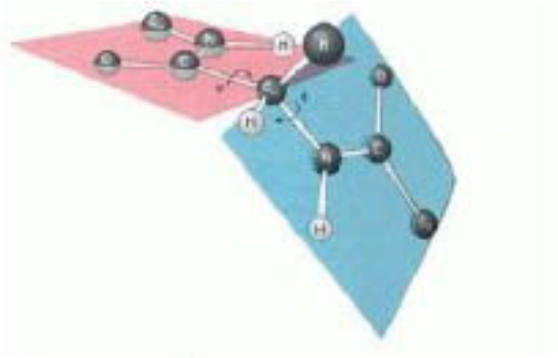






Structual informatics

Primary structure



Amino acids are planar

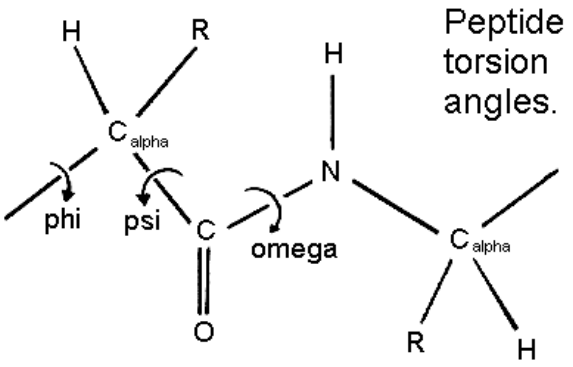
-Due to double bond character

-Gives no rotation over the Omega bond

Torsion angles

-Psi and phi rotation determines the folding of the protein

-Restriction of rotation due to steic hindarence, side-chain atoms from one amino acid come to close to side-chain atoms of another



Structual informatics

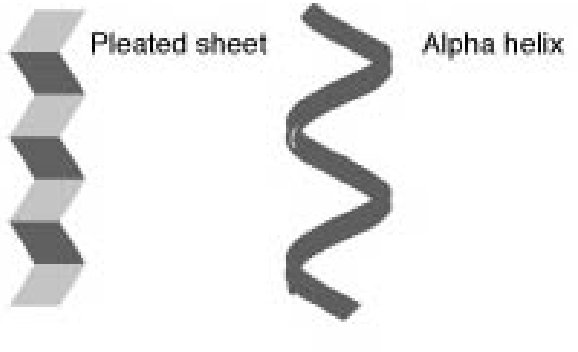
Secondary structure

Secondary structure

-Local arrangements of amino acids

-Formed by intramolecular hydrogen bonds between backbone atoms

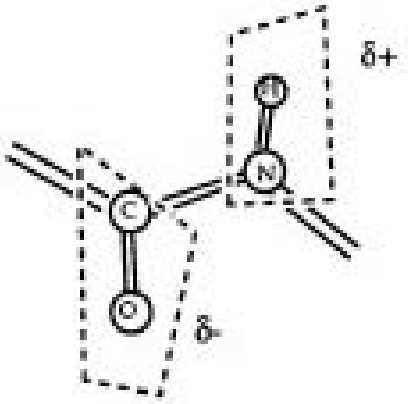
-α-helices and β-sheet are the most common secondary structure elements



Dipole

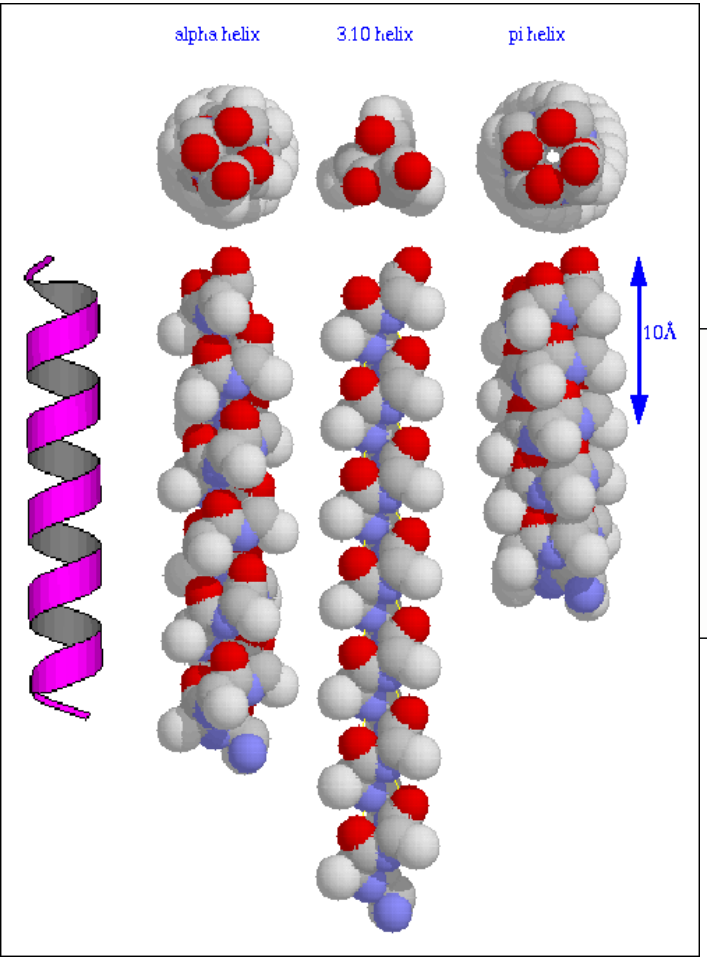
-The backbone atoms in the peptide chain have a dipole

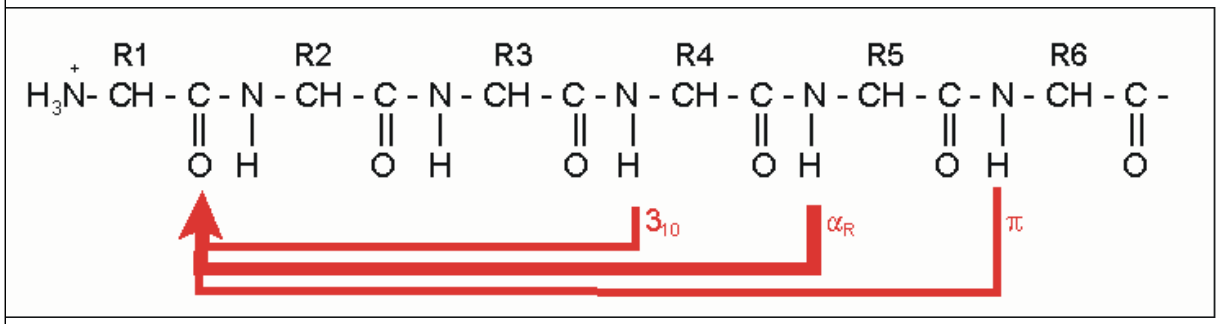
-Can make hydrogen bonds (forms secondary structure elements)



Structual informatics

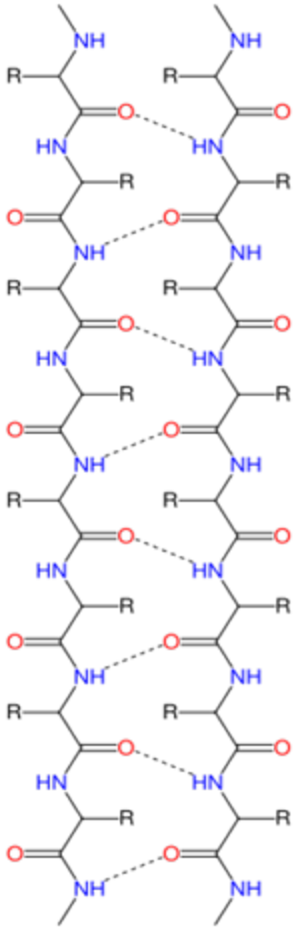
Secondary structure -helix



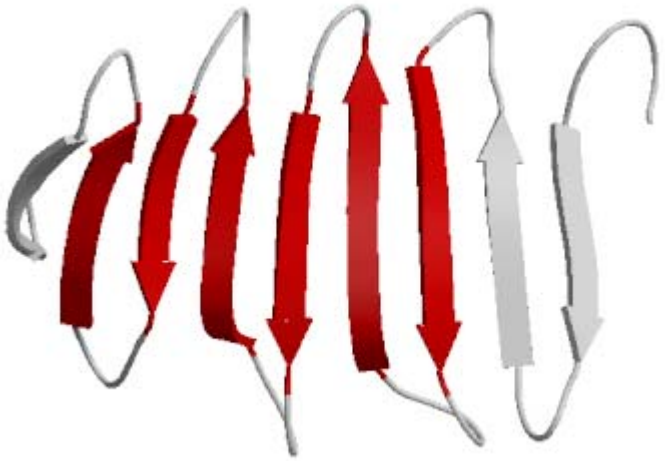


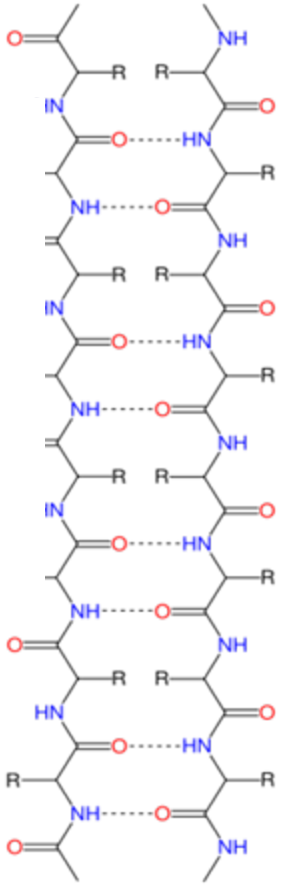
Structual informatics

Secondary structure - sheet



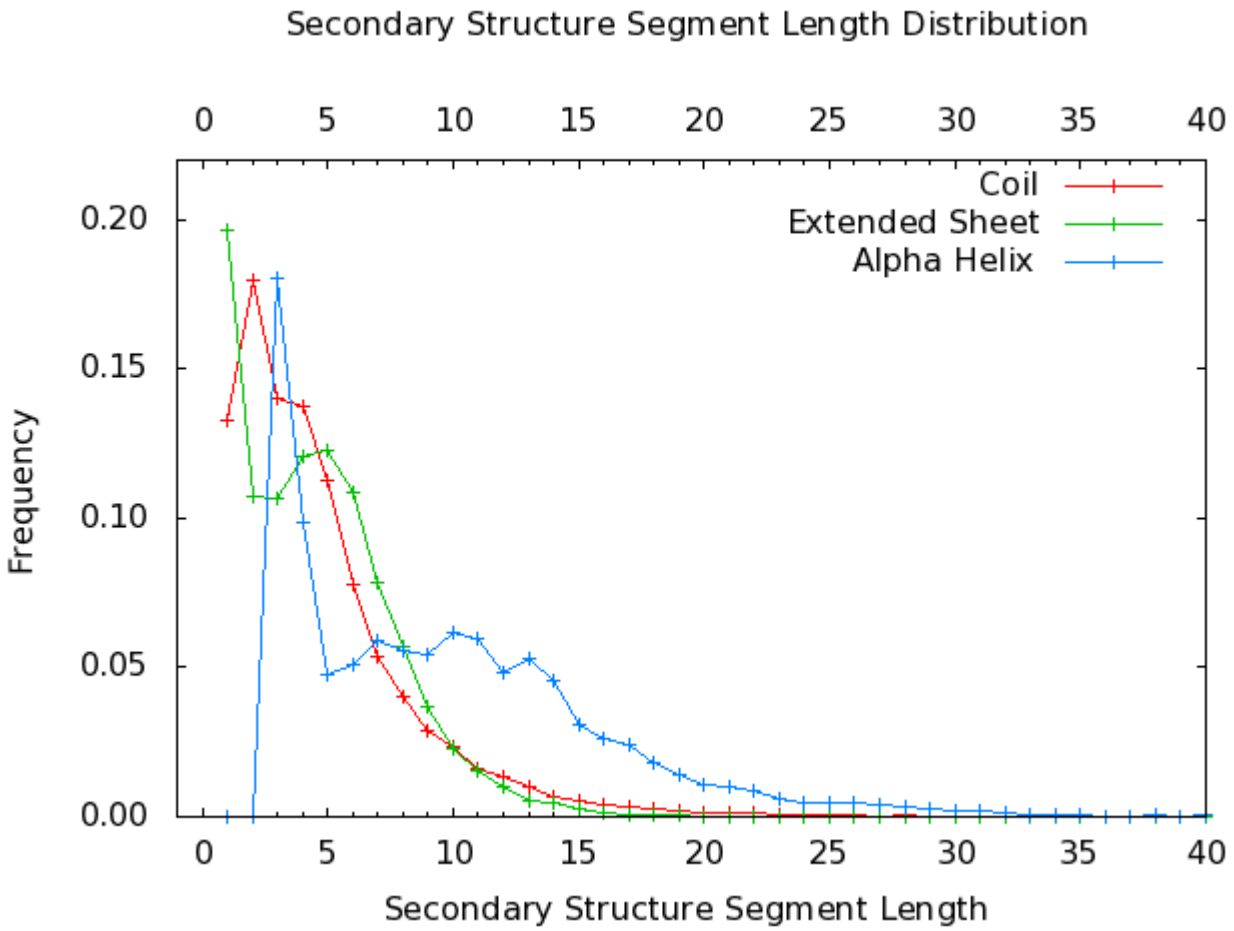






Structual informatics

Secondary structure



Structual informatics

Secondary structure prediction

Prediction of secondary structure elements (α-helices, β-sheet, coils) of *globular* proteins

Given a sequence (primary structure):

QHTAWCLTSEQHTAAVIWDCETPGKQNGAYQEDCA

Prediction of secondary structure:



QHTAWCLTSEQHTAAVIWDCETPGKQNGAYQEDCA
HHHHHHCCEEEEEEEEEECCHHHHHHHCCCCCCC

H = α-helices

E = β-sheet/strands

C = coils

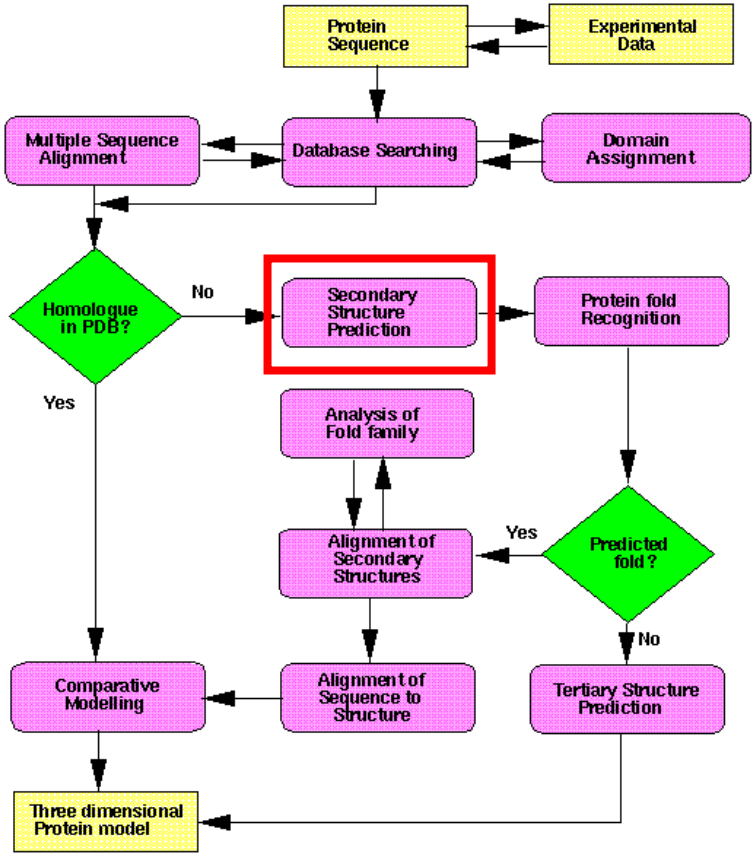
Structual informatics

Why do we want to predict secondary structure of proteins?

- 20-40% of all new proteins will not have homologous structures in protein databases
- Prediction of secondary structure is a step towards 3D structure prediction
- The type and order of secondary structure elements can give important information for the structural classification of the protein
- Structure is closely related to function – we can get functional information, eg. activity and specificity
- Improving an alignment

Structual informatics

Secondary structure prediction is often a part of a larger sequence analysis



Structual informatics

How can we do secondary structure prediction?

Free tools available on the internet:

- Secondary structure prediction

Predictprotein

Psipred

PSA

- Prediction of transmembrane helices

Tmpred

TMHMM

- Coils

COILS

Structual informatics

Secondary structure prediction methods

- Statistical methods

  From PDB database, calculate the propensity for a given amino acid to adopt a certain secondary structure type

- Nearest neighbours

  Predict secondary structure of the central residue of a given segment from homologous segments (neighbors).

- Neural network

  Machine learning approach - use known information to train the computers to recognize patterns in known secondary structure

Structual informatics

Secondary structure prediction methods

Statistical methods: Chou-Fasman

Propensity is the tendency for each amino acid to participate in α-helices, β-sheet and turns based on observed frequencies in a set of sample proteins with known structure

Name	P(H)	P(E)	P(turn)	f(i)	f(i+1)	f(i+2)	f(i+3)
Alanine	142	83	66	0.06	0.076	0.035	0.058
Arginine	98	93	95	0.07	0.106	0.099	0.085
Aspartic Acid	101	54	146	0.147	0.11	0.179	0.081
Asparagine	67	89	156	0.161	0.083	0.191	0.091
Cysteine	70	119	119	0.149	0.05	0.117	0.128
Glutamic Acid	151	37	74	0.056	0.06	0.077	0.064
Glutamine	111	110	98	0.074	0.098	0.037	0.098
Glycine	57	75	156	0.102	0.085	0.19	0.152
Histidine	100	87	95	0.14	0.047	0.093	0.054
Isoleucine	108	160	47	0.043	0.034	0.013	0.056
Leucine	121	130	59	0.061	0.025	0.036	0.07
Lysine	114	74	101	0.055	0.115	0.072	0.095
Methionine	145	105	60	0.068	0.082	0.014	0.055
Phenylalanine	113	138	60	0.059	0.041	0.065	0.065
Proline	57	55	152	0.102	0.301	0.034	0.068
Serine	77	75	143	0.12	0.139	0.125	0.106
Threonine	83	119	96	0.086	0.108	0.065	0.079
Tryptophan	108	137	96	0.077	0.013	0.064	0.167
Tyrosine	69	147	114	0.082	0.065	0.114	0.125
Valine	106	170	50	0.062	0.048	0.028	0.053

Structual informatics

Secondary structure prediction methods

Statistical methods: Chou-Fasman

How does it work?

1) Score each residue in your sequence

2) Identify α-helix and β-sheet regions by scanning a window of 6 residues where 4 amino acids have P(H) > 100 for helices and P(E) > 100 for strands

3) Extend the regions in both directions until 4 amino acids in a window have scores < 100

4) If structures overlap, average values for P(H) and P(E) are compared, and secondary structure is assign based on best scores

5) Turns are identified as tetrapeptides, where the score for each amino acid residues is derived from the sum of the following three residues

Structual informatics

Secondary structure prediction methods

Statistical methods: Chou-Fasman

1) Score each residue in your sequence

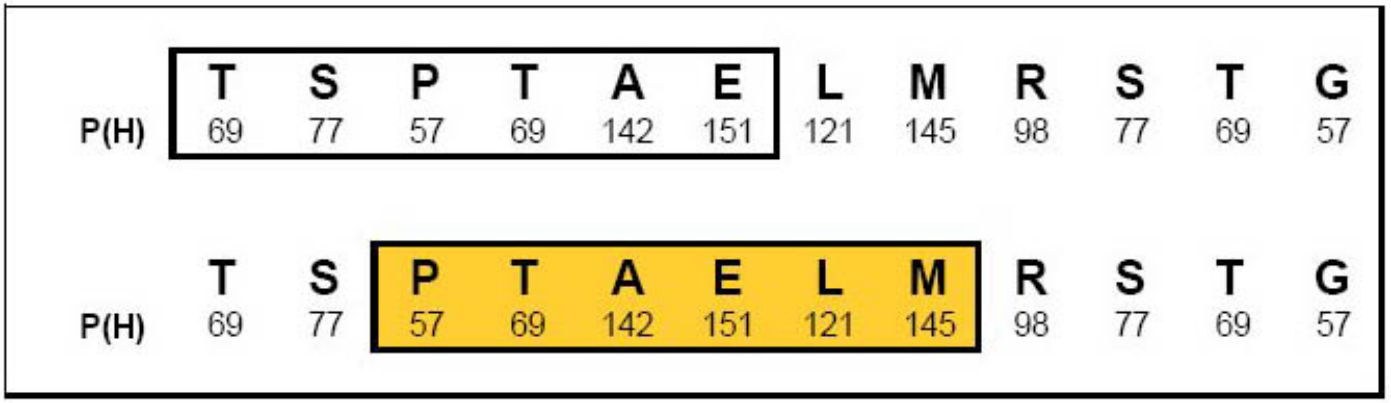
	T	S	P	T	A	E	L	M	R	S	T	G
P(H)	69	77	57	69	142	151	121	145	98	77	69	57
P(E)	147	75	55	147	83	37	130	105	93	75	147	75
P(turn)	114	143	152	114	66	74	59	60	95	143	114	156

Structual informatics

Secondary structure prediction methods

Statistical methods: Chou-Fasman

2) Identify α-helix and β-sheet regions by scanning a window of 6 residues where 4 amino acids have P(H) > 100 for helices and P(E) > 100 for strands



Structual informatics

Secondary structure prediction methods

Statistical methods: Chou-Fasman

3) Extend the regions in both directions until 4 amino acids in a window have scores < 100



Repeat step 1 – 3 for the entire peptide

Structual informatics

Secondary structure prediction methods

Statistical methods: Chou-Fasman

4) If structures overlap, average values for P(H) and P(E) are compared, and secondary structure is assign based on best scores

	T	S	P	T	A	E	L	M	T	S	T	G
P(H)	69	77	57	69	142	151	121	145	69	77	69	57
P(E)	147	75	55	147	83	37	130	105	147	75	147	75

Structual informatics

Secondary structure prediction methods

Statistical methods: Chou-Fasman

5) Turns are identified as tetrapeptides, where the score for each amino acid residues is derived from the sum of the following three residues

Position in turn P(t) = f(i) + f(i+1) + f(i+2) + f(i+3)

X	X	X	A	R	N	D	X	X	X

P(t) = 0.06 0.106 0.191 0.081

Predict turn in the four residues if:

P(t) > 0.000075

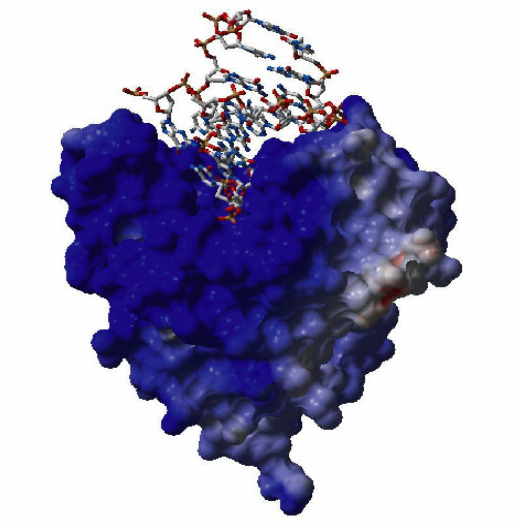
Average P(turn) > 1.00

ΣP(a-helix) < ΣP(turn) > ΣP(b-sheet)

		f(i)	f(i+1)	f(i+2)	f(i+3)
Alanine	(**A**)	0.06	0.076	0.035	0.058
Arginine	(**R**)	0.07	0.106	0.099	0.085
Aspartic acid	(**D**)	0.147	0.11	0.179	0.081
Aspargine	(**N**)	0.161	0.083	0.191	0.091

Structual informatics

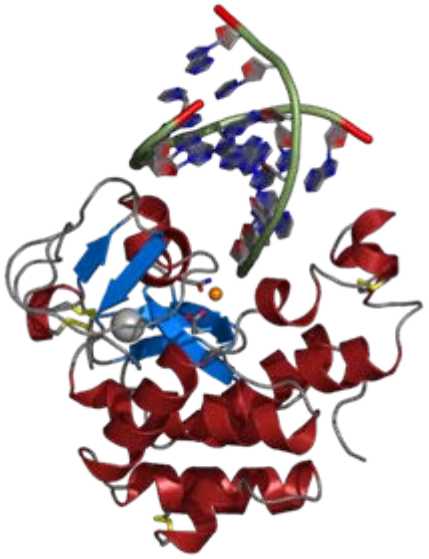
Tertiary structure



-Folding of the total peptide chain

-Secondary structure elements linked by loops and turns

-Describes the 3D-arrangement of all amino acid residues in the polypeptide

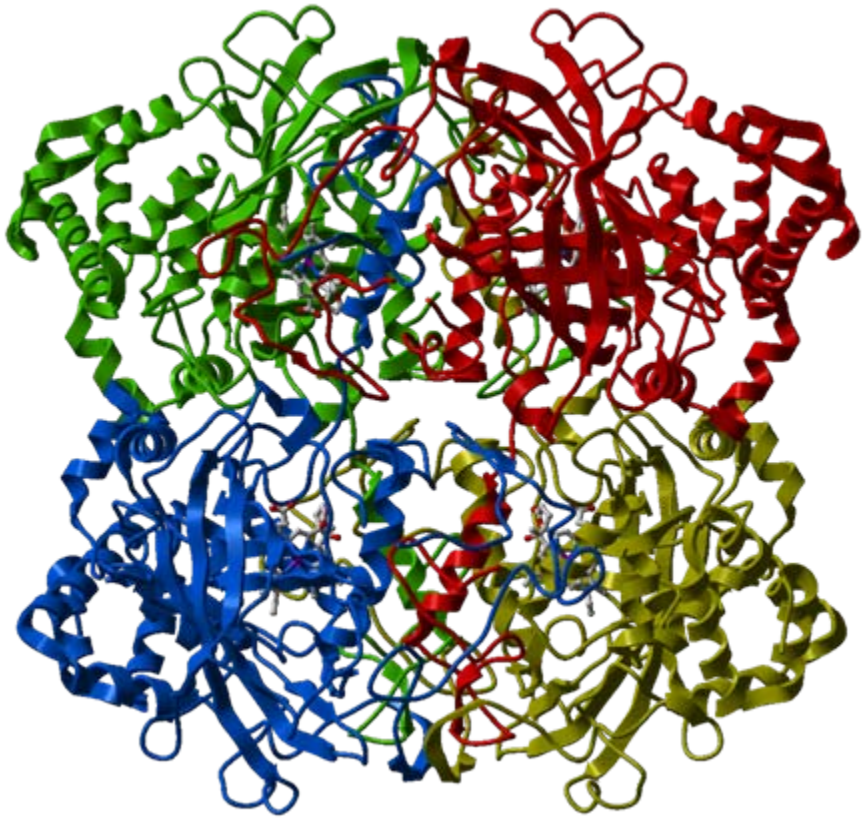


Structual informatics

Quarternary structure

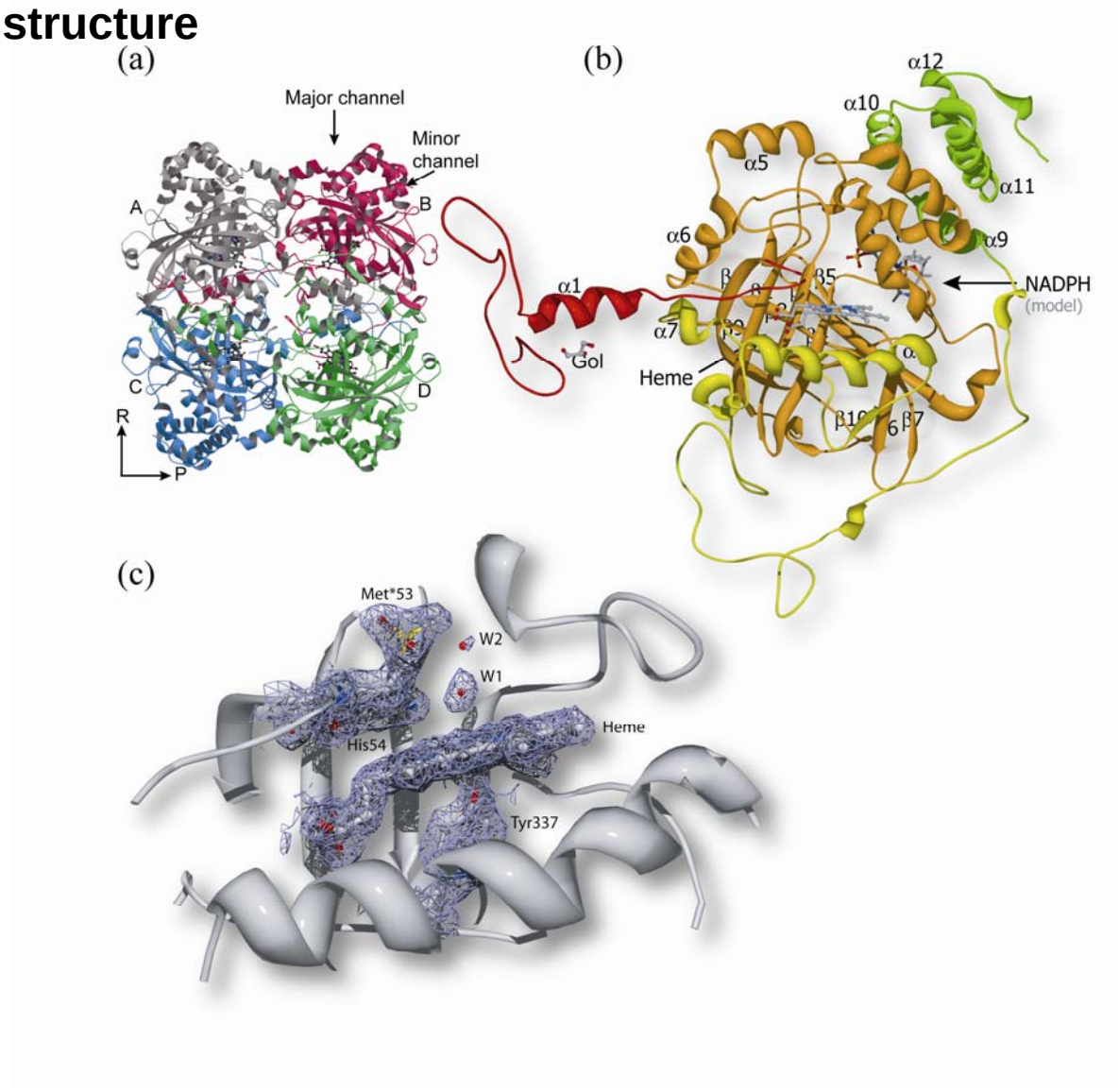
-Association of *two or more* polypeptide chains to form a native protein structure

-The polypeptides are held together by non-covalent bonds



Structual informatics

Quarternary structure



Structual informatics

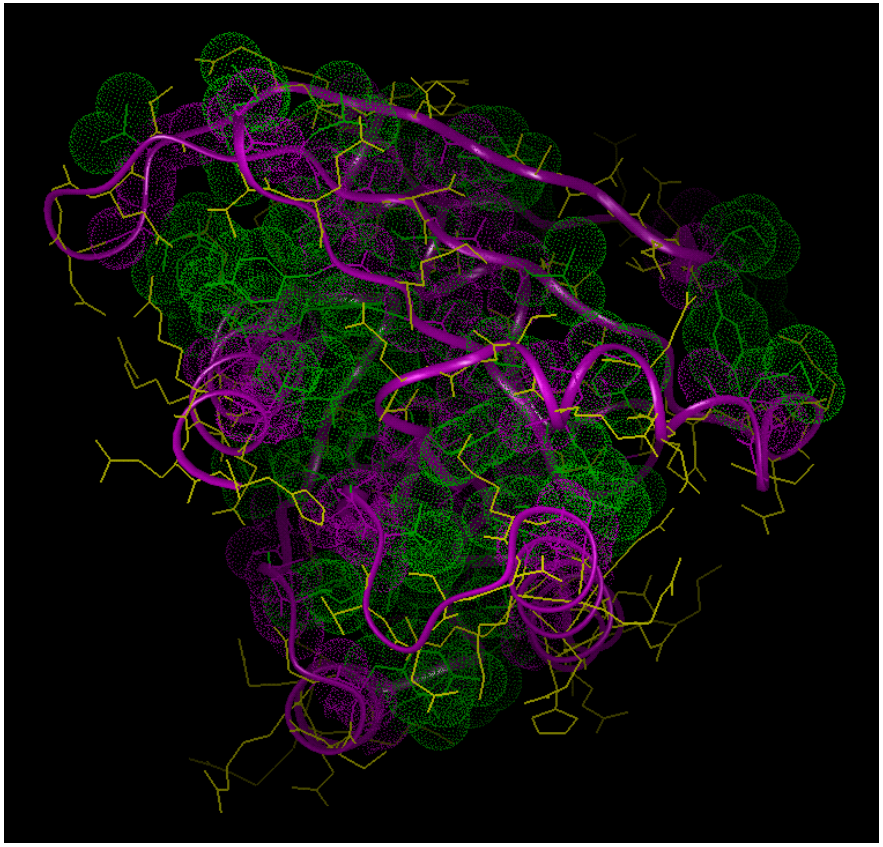
Hydrofobicity effect

-Thought to be the main factor for stabilization of the protein structure

-General rule: hydrophobic core, hydrophilic surface

Green = Hydrophobic amino acids

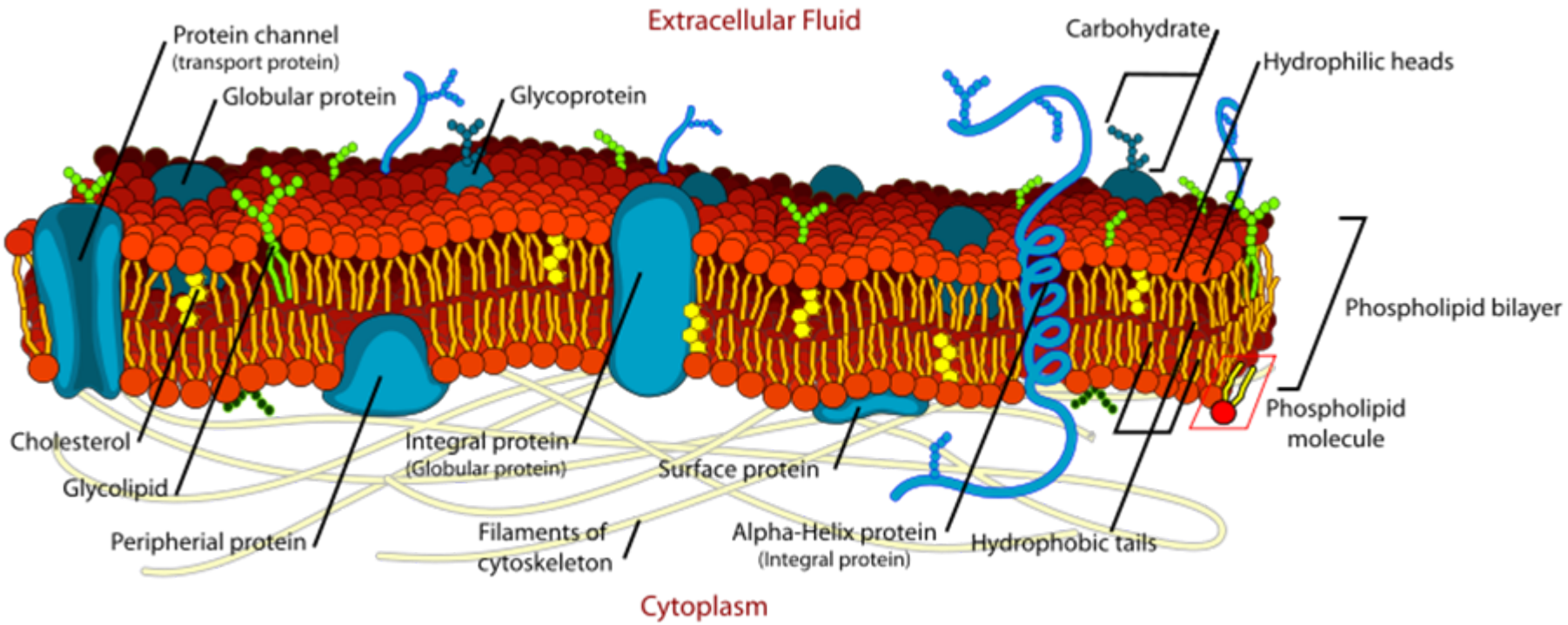
Purple = Hydrophilic amino acids



Structual informatics



Structual informatics



Structual informatics

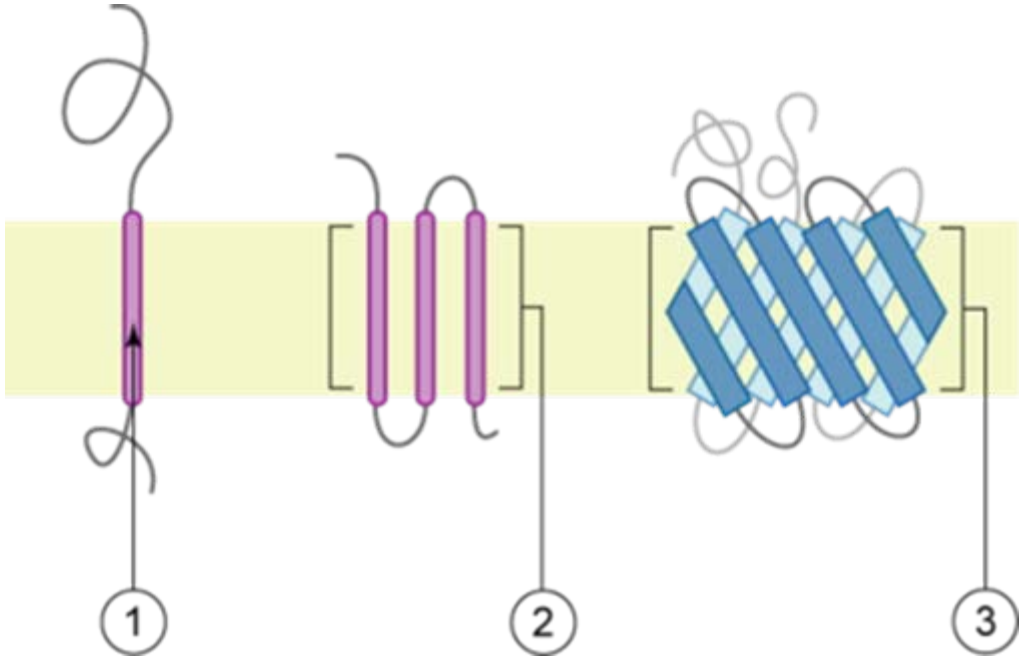
- An **Integral Membrane Protein** (**IMP**) is a protein molecule (or assembly of proteins) that is permanently attached to the biological membrane.

- IMPs can be divided into two groups:
  - Transmembrane proteins
  - Integral monotopic proteins

- A **transmembrane protein** is a protein that spans the entire biological membrane. Transmembrane proteins aggregate and precipitate in water.

- **Peripheral membrane proteins** are proteins that adhere only temporarily to the biological membrane with which they are associated.

Structual informatics

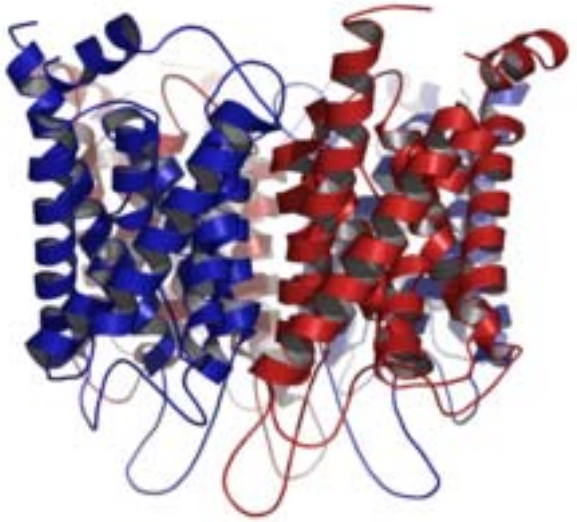


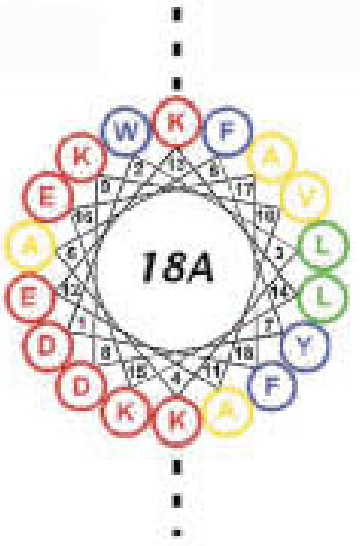
Schematic representation of transmembrane proteins:

1. single transmembrane α-helix (bitopic membrane protein)
2. polytopic α-helical protein
3. transmembrane β barrel

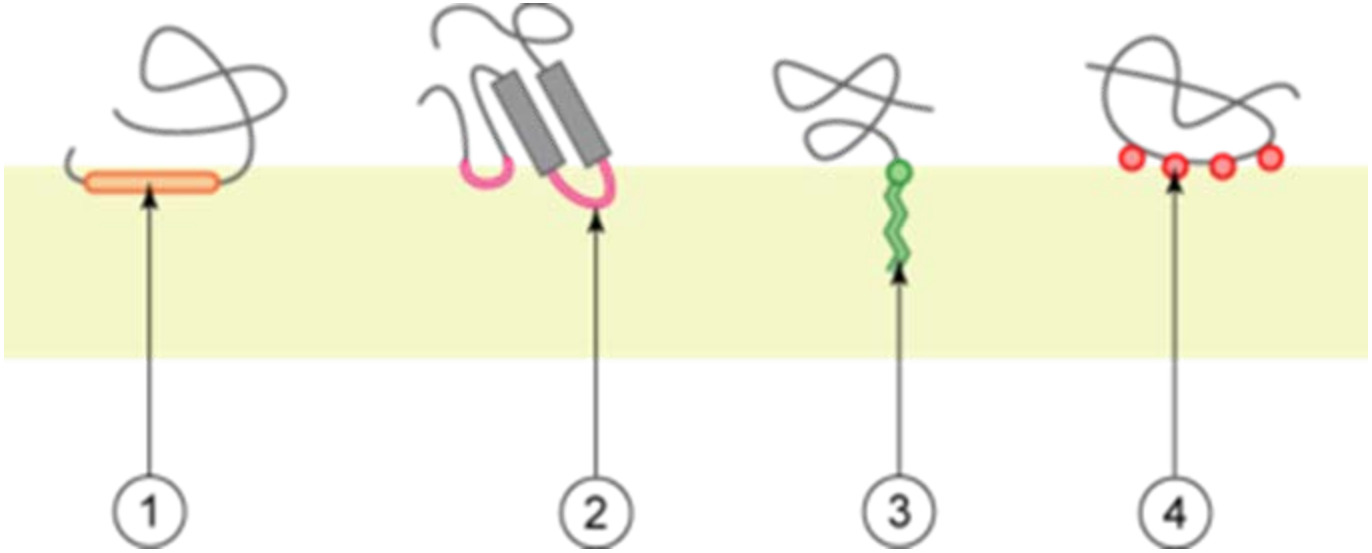
Structual informatics

Transmembrane segments have a hydrophobic side on the surface facing the hydrophobic membrane and a hydrophilic side in the core





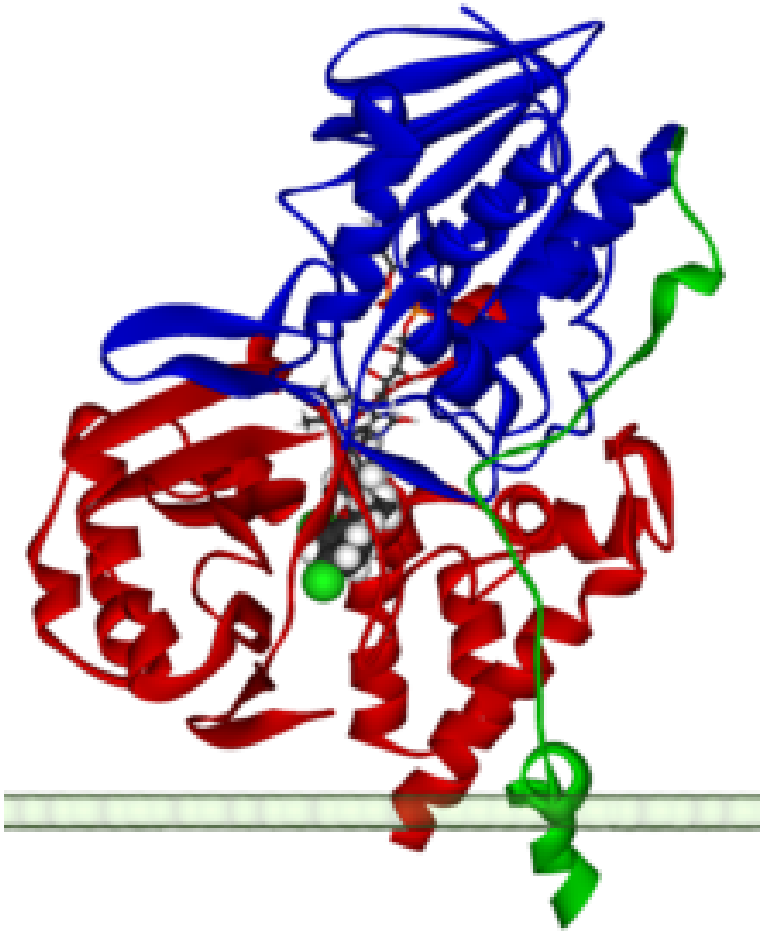
Structual informatics



Schematic representation of the different types of interaction between membrane proteins and the cell membrane:

1. Interaction by an amphipathic α-helix parallel to the membrane plane
2. Interaction by a hydrophic loop
3. Interaction by a covalently bound membrane lipid (*lipidation*)
4. Electrostatic or ionic interactions with membrane lipids (*e.g.* through a calcium ion)

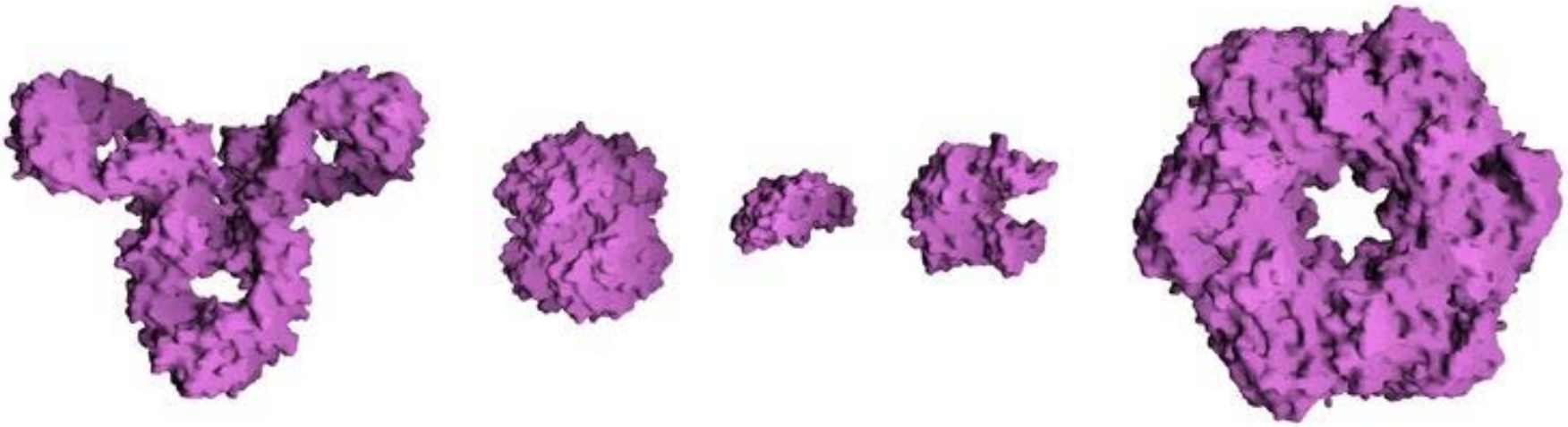
Structual informatics



- **Monoamine oxidases** (singular abbreviation **MAO**) are enzymes that catalyze the oxidation of monoamines. They are found bound to the outer membrane of mitochondria in most cell types in the body.

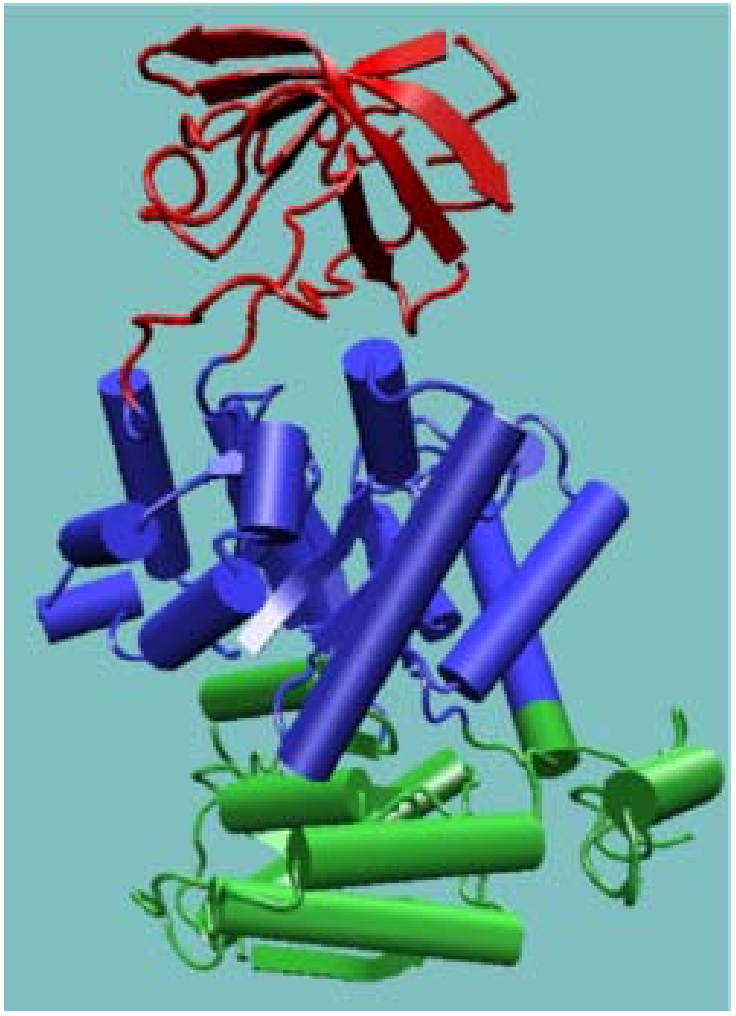
Structual informatics

Classification of structures



Molecular surface of several proteins showing their comparative sizes. From left to right are: Antibody (IgG), Hemoglobin, Insulin (hormone), Adenylate kinase (enzyme), and Glutamine synthetase (enzyme).

Structual informatics



Structual informatics

Protein domains

There does not exist a general accepted definition of what a domain is, but some properties are:

- A domain is part of a polypeptide chain of a protein or the whole chain
- It does not need to be a contiguous region of the polypeptide chain
- It can fold independently to its stable fold
- It has its own function
- It contains at least one hydrophobic core
- It is local compact

Structual informatics

Classification of structures

- **SCOP:** http://scop.mrc-lmb.cam.ac.uk/scop/
- **CATH:** http://www.biochem.ucl.ac.uk/bsm/cath/
- **CE:** http://cl.sdsc.edu/ce.html
- **Dali Domain Dictionary:** http://columba.ebi.ac.uk:8765/holm/ddd2.cgi
- **FSSP:** http://www2.ebi.ac.uk/dali/fssp/
- **HOMSTRAD:** http://www-cryst.bioc.cam.ac.uk/~homstrad/
- **HSSP:** http://swift.embl-heidelberg.de/hssp/

Structual informatics

CATH - Protein Structure Classification

- ***CATH has six levels***
  - Class (C)
  - Architecture (A)
  - Topology (T)
  - Homologous superfamily (H)
  - Sequence family (S)
  - PDB entry
- Version 3.1.0 released January 2007

Structual informatics

CATH - Protein Structure Classification

Class, C-level

Class is determined according to the secondary structure composition and packing within the structure.

- Mainly-alpha
- Mainly-beta
- Alpha-beta (alpha/beta and alpha + beta)
- Protein domains with low secondary structure content

Structual informatics

CATH - Protein Structure Classification

Architecture, A-level

- The architecture describes the arrangement of secondary structures as determined by the orientations of the secondary structures, independent of topology (i.e., ignores the connectivity between the secondary structures).
- Assigned manually, e.g., barrel, 3-layer

Structual informatics

CATH - Protein Structure Classification

Topology (fold family), T-level

- Structures are grouped into fold families at this level depending on both the overall shape and connectivity of the secondary structures.
- This is done using the structure comparison algorithm SSAP (Taylor & Orengo (1989)). Parameters for clustering domains into the same fold family have been determined by empirical trials throughout the databank (Orengo et al. (1992), Orengo et al. (1993)).
- Structures which have a SSAP score of 70 and where at least 60% of the larger protein matches the smaller protein are assigned to the same T level or fold family.

Structual informatics

CATH - Protein Structure Classification

Homologous superfamily, H-level

- This level groups together protein domains which are thought to share a common ancestor and can therefore be described as homologous.

- Similarities are identified first by sequence comparisons and subsequently by structure comparison using SSAP.

- Structures are clustered into the same homologous superfamily if they satisfy one of the following criteria:
  - Sequence identity >= 35%, overlap >= 60% of larger structure equivalent to smaller.
  - SSAP score >= 80.0, sequence identity >= 20%, 60% of larger structure equivalent to smaller.
  - SSAP score >= 70.0, 60% of larger structure equivalent to smaller, and domains which have related functions, which is informed by the literature and Pfam protein family database, (Bateman *et al.,* 2004).
  - Significant similarity from HMM-sequence searches and HMM-HMM comparisons using SAM (Hughey &Krogh, 1996), HMMER (http://hmmer.wustl.edu) and PRC (http://supfam.org/PRC).

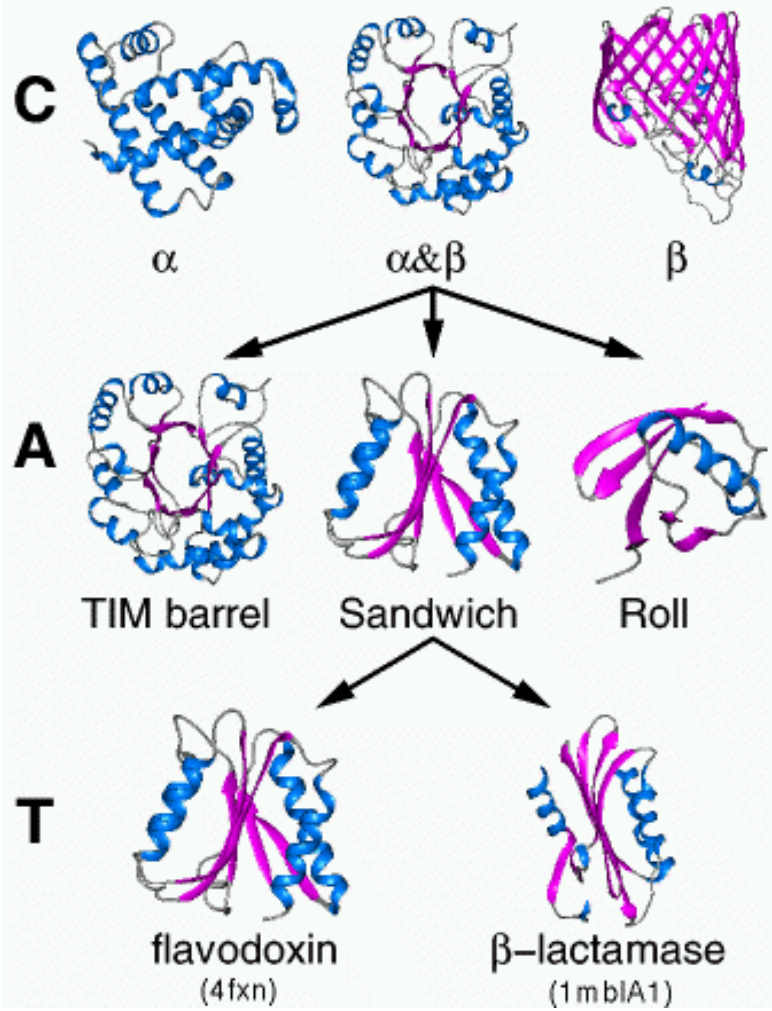
Structual informatics

CATH - Protein Structure Classification

Sequence Family Levels: (S,O,L,I, D)

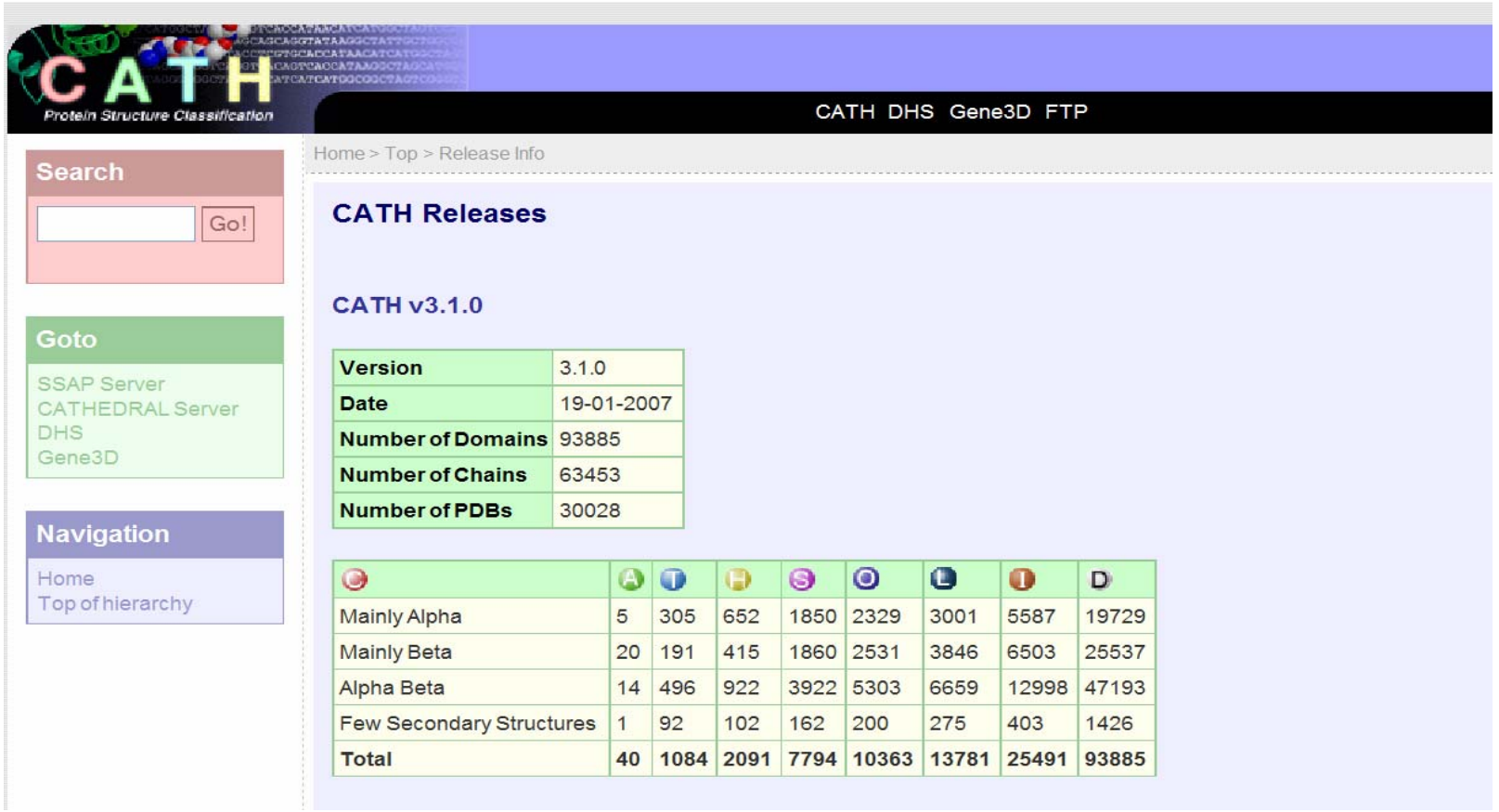
Structures within each H-level are further clustered on sequence identity. Domains clustered in the same sequence families have sequence identities >35% (with at least 60% of the larger domain equivalent to the smaller), indicating highly similar structures and functions.

Structual informatics



Structual informatics

CATH - Statistics



Version	3.1.0
Date	19-01-2007
Number of Domains	93885
Number of Chains	63453
Number of PDBs	30028

C	A	T	H	S	O	L	I	D
Mainly Alpha	5	305	652	1850	2329	3001	5587	19729
Mainly Beta	20	191	415	1860	2531	3846	6503	25537
Alpha Beta	14	496	922	3922	5303	6659	12998	47193
Few Secondary Structures	1	92	102	162	200	275	403	1426
Total	**40**	**1084**	**2091**	**7794**	**10363**	**13781**	**25491**	**93885**

http://cathwww.biochem.ucl.ac.uk/cgi-bin/cath/GotoCath.pl?link=class.html

Structual informatics

Classification by SCOP

- Done manually
- Seven levels
  - Class
  - Fold
  - Superfamily
  - Family
  - Protein domain
  - Species
  - PDB entry
- Version 1.71 released in October 2006

Structual informatics

Classification by SCOP

Class

- All alpha
- All beta
- Alpha/beta
- Alpha+beta
- Multi-domain proteins (folds consisting of two or more domains belonging to different classes)
- Membrane and cell surface proteins and peptides
- Small proteins
- Coiled coil proteins

Structual informatics

Classification by SCOP

Fold

- Major structural similarity
- 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
- 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

Structual informatics

Classification by SCOP

Superfamily

- Probable common evolutionary origin
- Proteins that have low sequence identities, but whose structural and functional features suggest that a common evolutionary origin is probable are placed together in superfamilies.
- For example, actin, the ATPase domain of the heat shock protein, and hexakinase together form a superfamily.

Structual informatics

Classification by SCOP

Family

- Clear evolutionarily relationship
- Proteins clustered together into families are clearly evolutionarily related.
- Generally, this means that pairwise residue identities between the proteins are 30% and greater.
- However, in some cases similar functions and structures provide definitive evidence of common descent in the absence of high sequence identity
- For example, many globins form a family though some members have sequence identities of only 15%.

Structual informatics

Scop Classification Statistics

SCOP: Structural Classification of Proteins. **1.71** release
27599 PDB Entries (18 Jan 2005). 75930 Domains. 1 Literature Reference
(excluding nucleic acids and theoretical models)

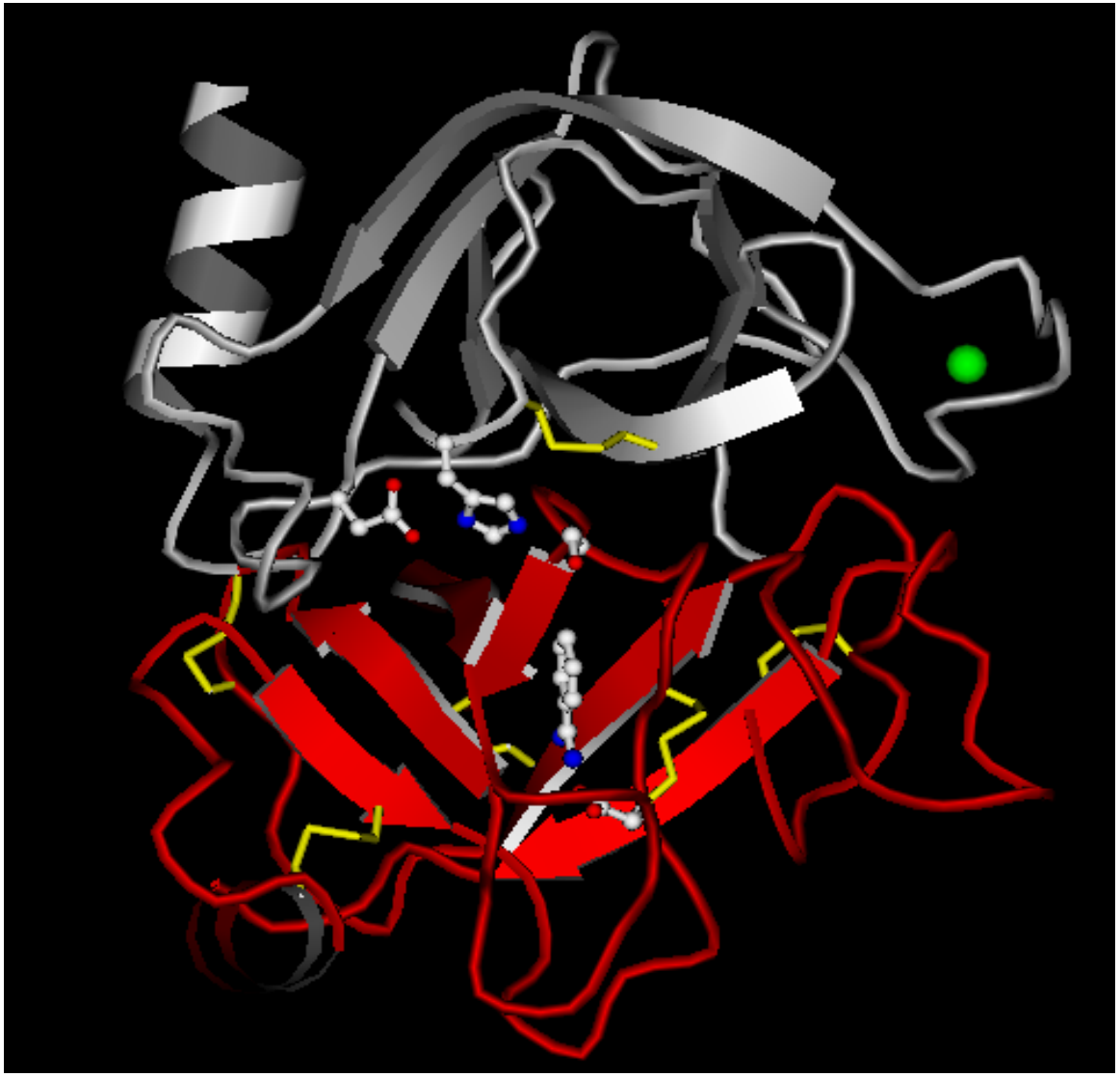
Class	Number of folds	Number of superfamilies	Number of families
All alpha proteins	226	392	645
All beta proteins	149	300	594
Alpha and beta proteins (a/b)	134	221	661
Alpha and beta proteins (a+b)	286	424	753
Multi-domain proteins	48	48	64
Membrane and cell surface proteins	49	90	101
Small proteins	79	114	186
Total	971	1589	3004

Structual informatics

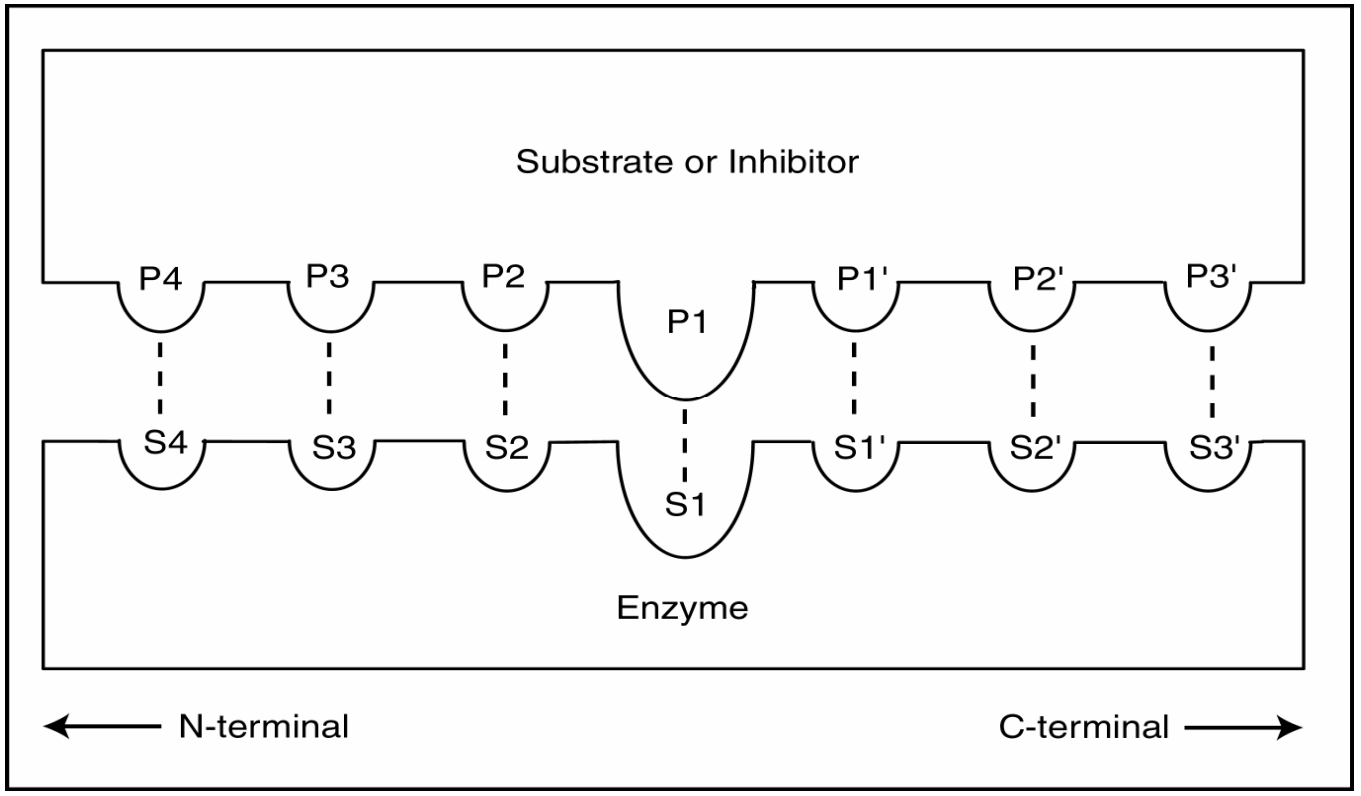
Comparing protein structures: Why?

- Understanding the relationship between structure and function
- Evolutionary relationships - how proteins evolve new functions.
- Useful to describe protein fold space
  - Which of the possible folds exist in nature
  - How many different folds exist. If there exist a finite number of folds, the structure prediction problem becomes easy
- Useful to classify a new structure as being of a known/new fold
- Classification makes protein 3D structure data more accessible and understandable

Structual informatics

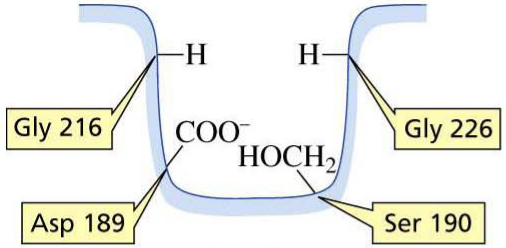


Structual informatics

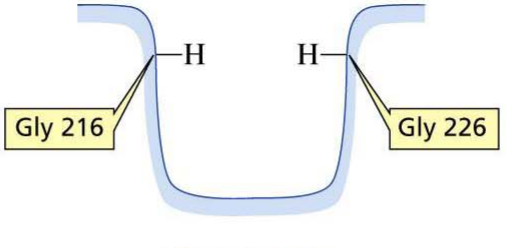


Nomenclature of Schechter & Berger (1967).

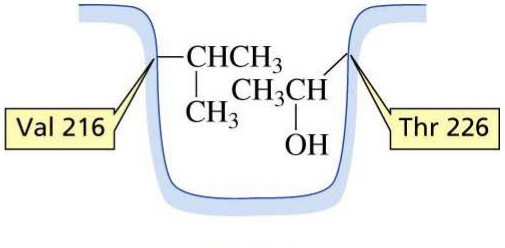
Structual informatics



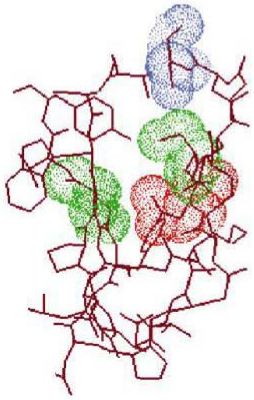
trypsin

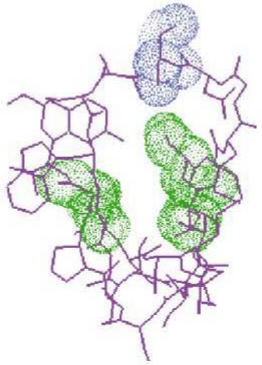


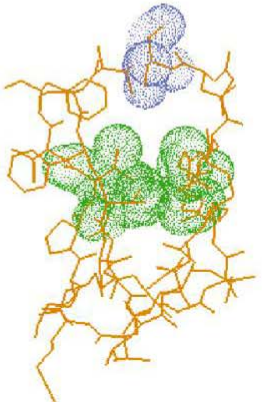
chymotrypsin



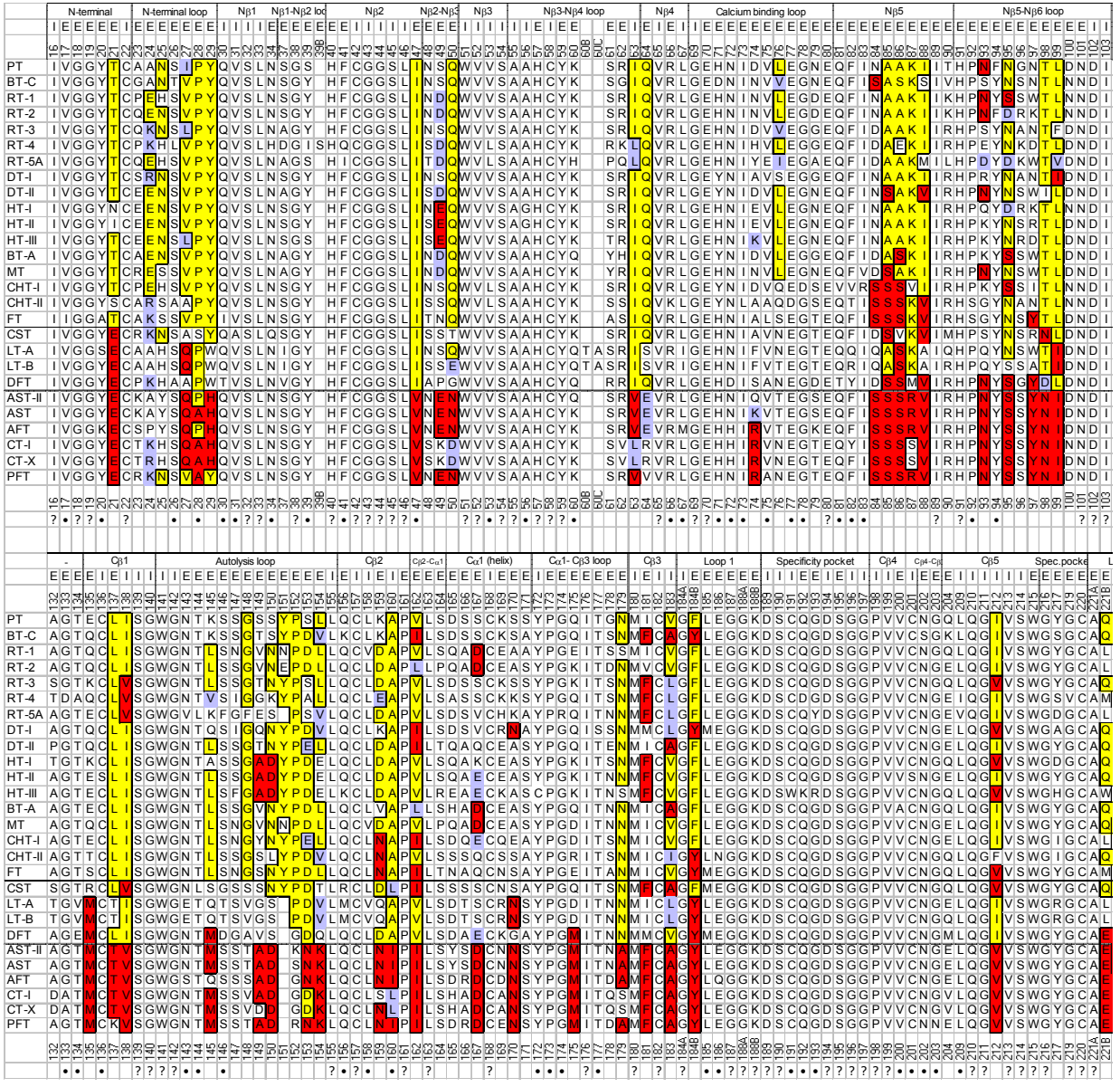
elastase



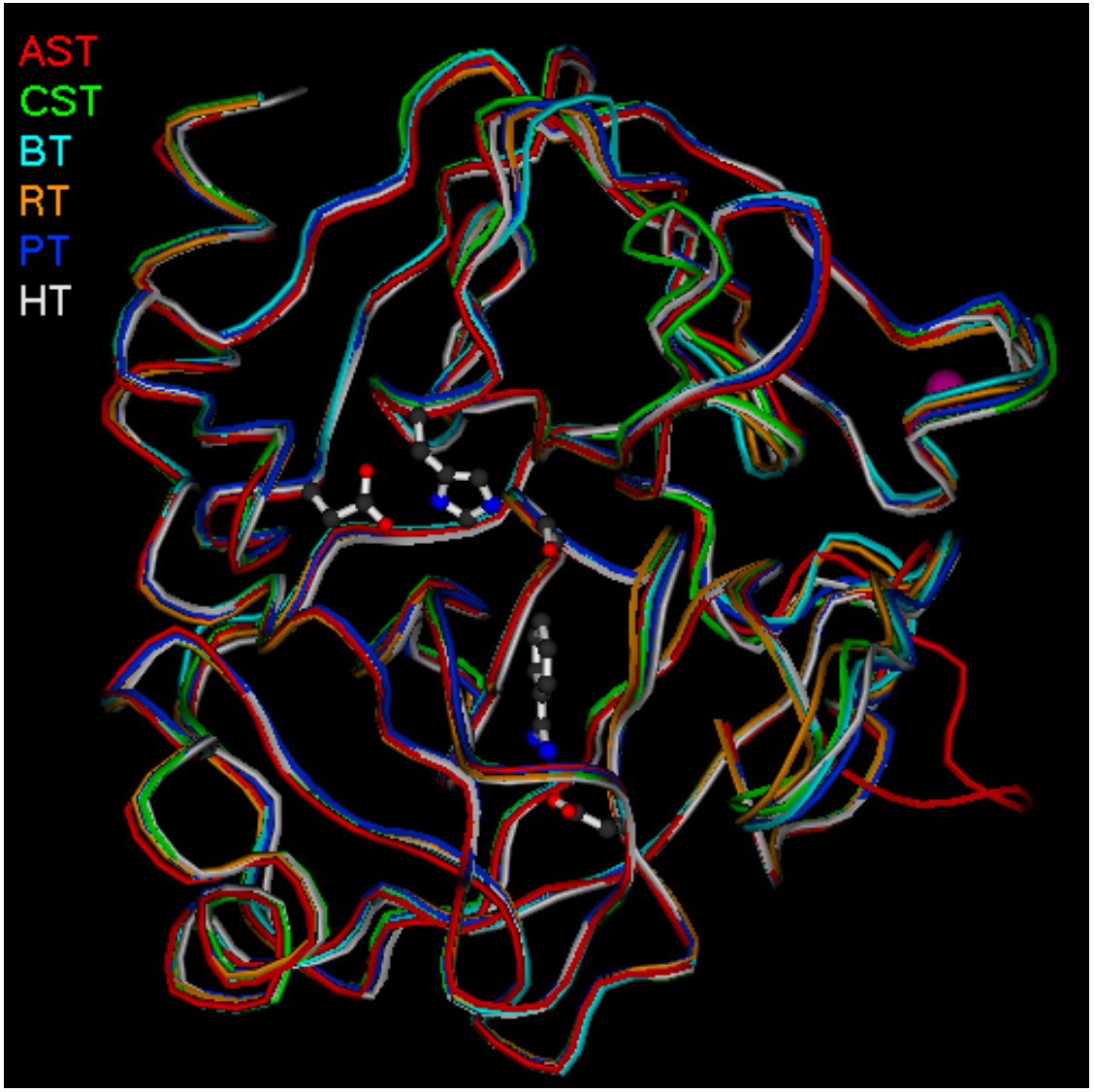




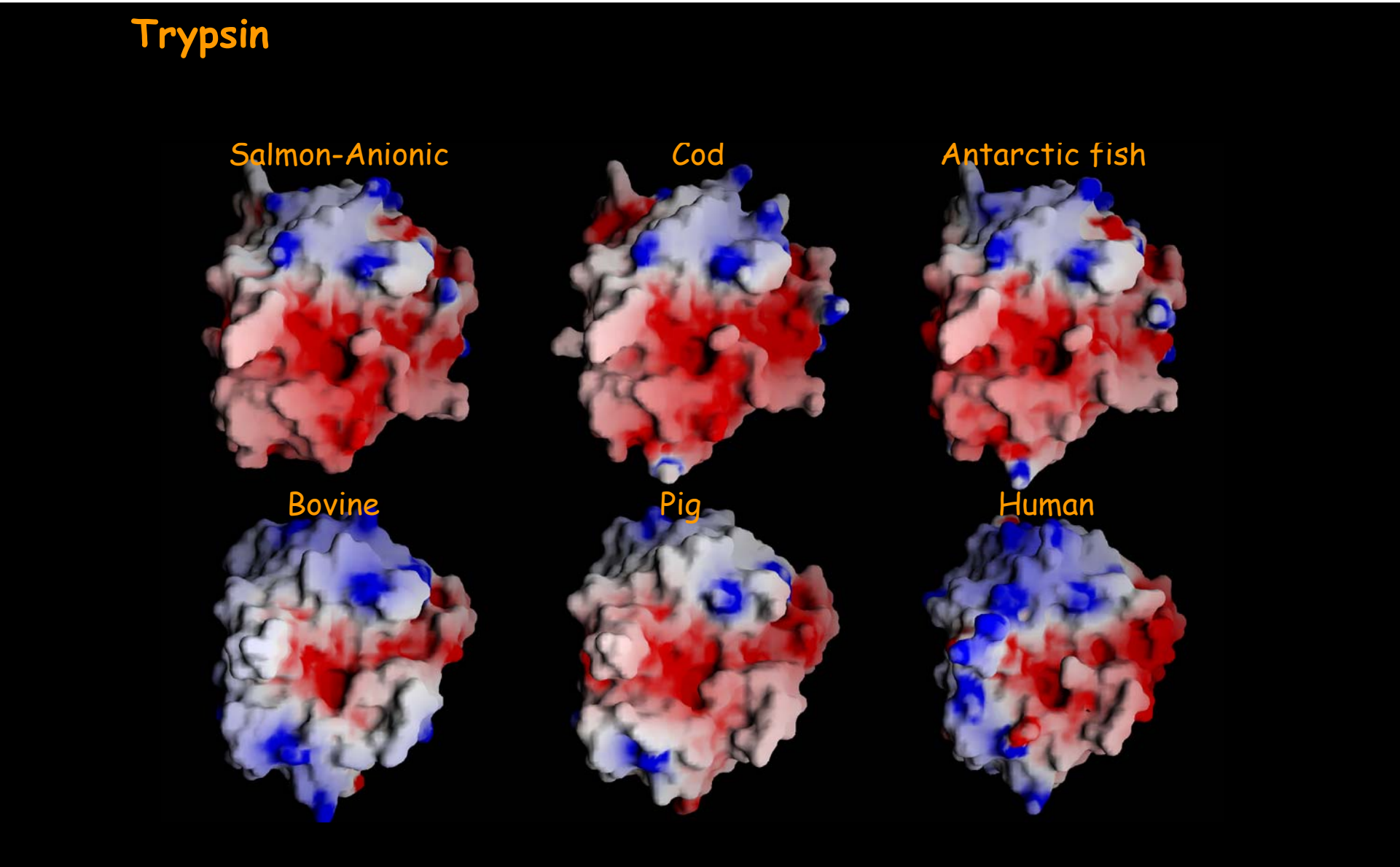
Structual informatics



Structual informatics

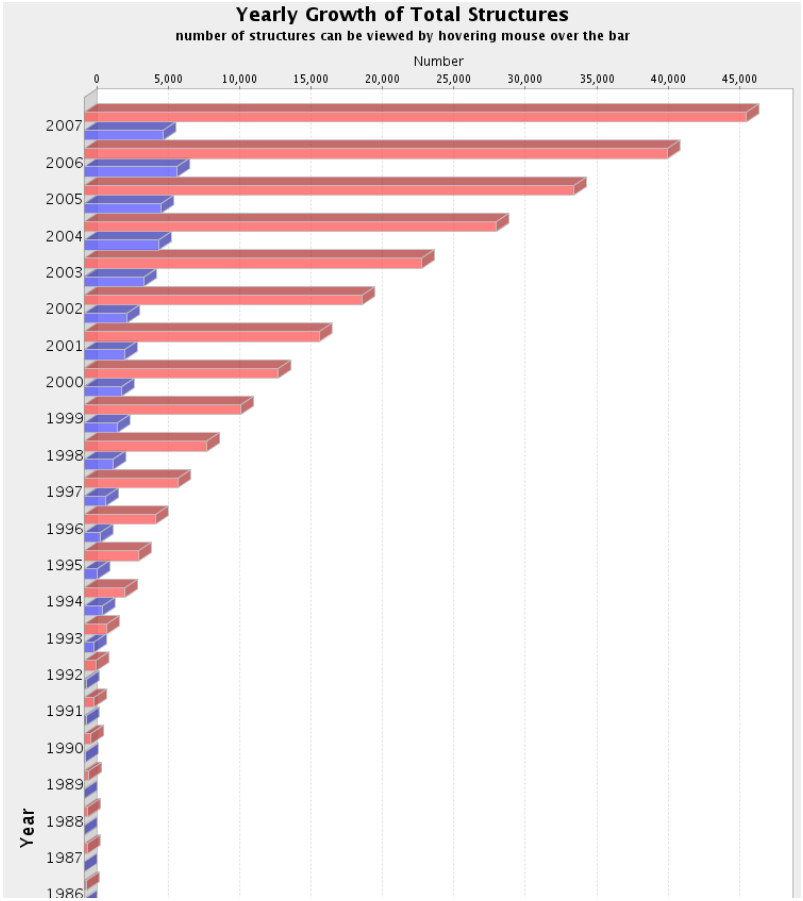


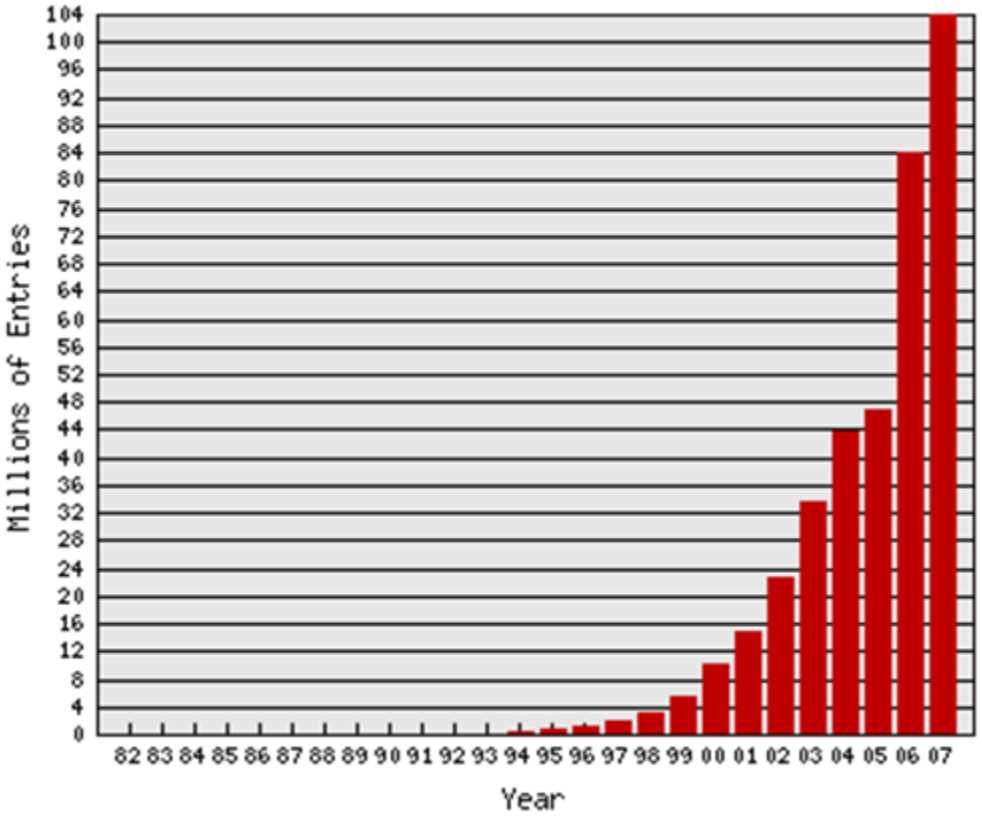
Structual informatics



Structual informatics

Sequence-Structure gap





Structual informatics

Protein structure prediction

3-D Structure of natural proteins of are guided by two sets of principles:

- Laws of Physics
  - Chemical bonds, H-bonds, electonic interactions
- Theory of Evolution
  - Families of similar proteins share similar sequences, structure, and often function

Structual informatics

Two general methods to model structures

Each based on the principles governing 3-D structure

- **Comparative/homology** modeling – based on similarities from evolution
- **De novo** or ***ab initio*** – based on physics…energies

Structual informatics

Steps in comparative modeling (CM)

1. Find suitable template(s)
2. Build alignment between target and template(s)
3. Build model(s)
   - Replace sidechains
   - Resolve conflicts in the structure
   - Model loops (regions without an alignment)
4. Evaluate and select model(s)

Structual informatics

Steps in comparative modeling (CM)

1. Finding template(s)

   – 3-D structures are more conserved that primary sequences through evolution

     - Sequence-comparison
       – E.g., PSI-BLAST
     - Threading (sequence-structure comparison)
       – Often capable of finding more distant relationships between proteins than sequence comparison alone
       – Requires a database of templates and an energy function
       – Each sequence-structure comparison is assessed via the energy function

Structual informatics

Steps in comparative modeling (CM)

2. Build alignment between target and template(s)

    - multiple alignment of close to fairly distant homologues

Structual informatics

Steps in comparative modeling (CM)

3. Build model(s)

   – After finding similar sequences, a structure is modeled using that related sequence.

   – Core regions are kept intact (helices, sheets)

   – Loops and side chains using various methods to minimize energies in the presence of the core regions.

   – Three main methods
      1. Loops and side chains based on alignment to related sequences in other proteins
      2. Approx. position of atoms is calculated from approx. position of conserved atoms in template.
      3. Distance Geometry or spatial restraints are optimized

Structual informatics

Steps in comparative modeling (CM)

4. Evaluate and select model(s)

   - Accuracy associated with the percent of the sequence that matches the template
   - Most errors in side chains or in loops
     - These are mostly insignificant as these parts don't effect function.
   - Other errors include *small* shifts and distortions in other parts of the protein
   - Error increases *rapidly* below 30% sequence similarity

Structual informatics

De novo (ab initio) prediction

- Comparative modeling and threading rely on some sort of detectable similarity between the target and template
  - Sometimes, we cannot detect any relationship
  - CM & Threading cannot predict entirely new folds
- De novo prediction is not limited to known folds
  - The only assumption is that the protein adopts the global free-energy minimum

Structual informatics

De novo (ab initio) prediction

- De novo requires
  - Search procedure
  - Free-energy function
- Different techniques may make certain simplifications
  - E.g., reduced-atom representations

Structual informatics

De novo (ab initio) prediction

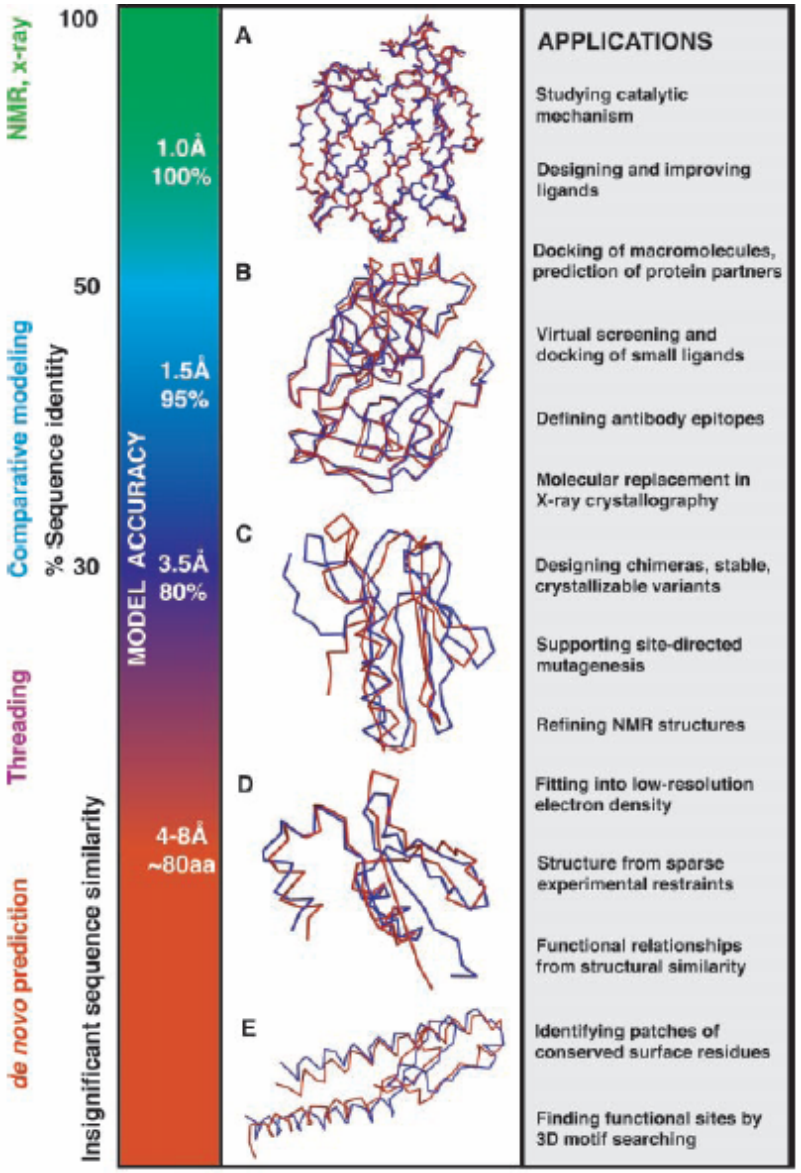
- Accuracy is much lower than comparative methods (with >30% sequence similarity)
- Basic structure can still be determined reasonably well with 150 amino acids or less
- Accuracy too *low* if high-resolution info is needed.
- Low-resolution info is useful in finding structural and functional relationships no apparent by comparison of AA sequence.

Structual informatics

Accuracy

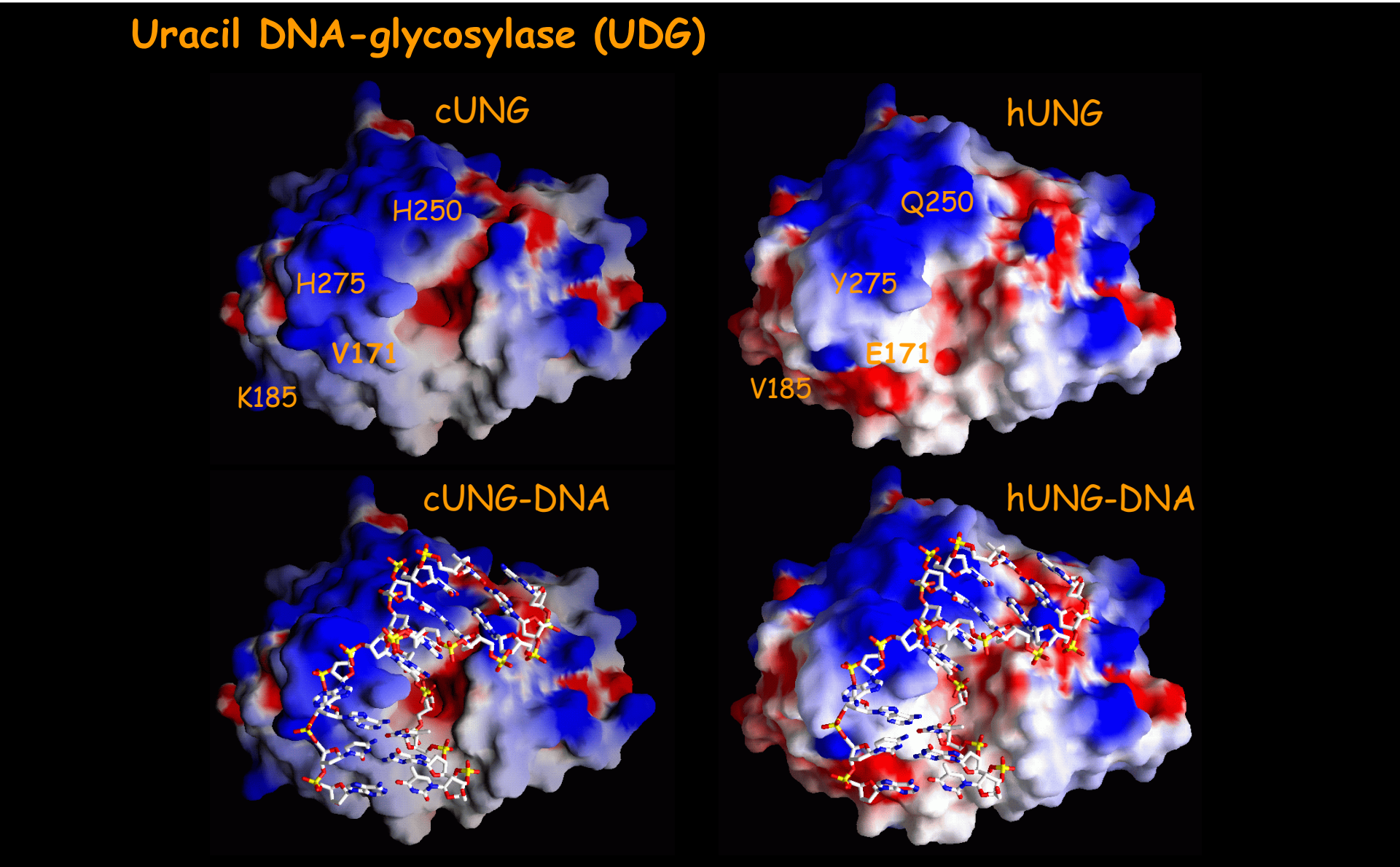
- Comparative more accurate…*but* requires other structures to be known
- De novo can be used for any protein but requires more computing power and won't produce highly accurate models.

Structual informatics



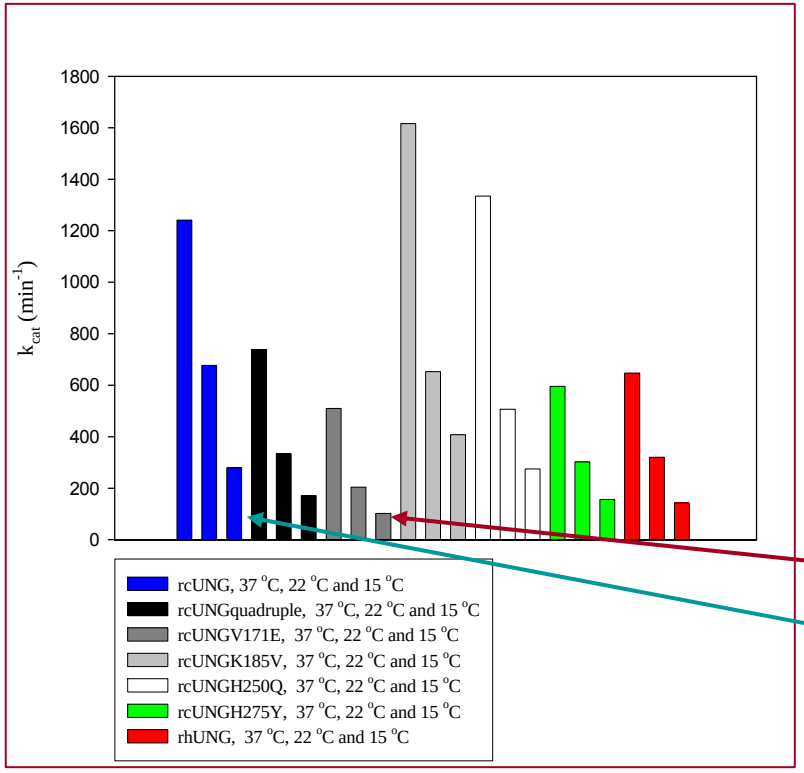
Structual informatics

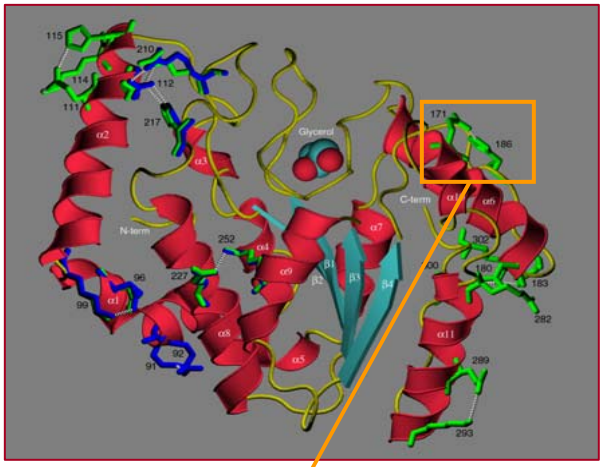
Structual informatics

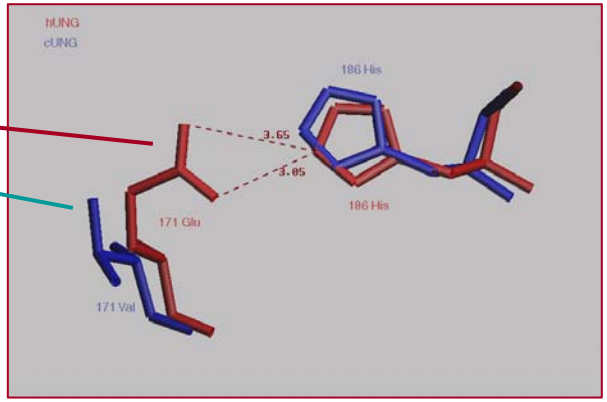


Structual informatics

Uracil DNA-glycosylase (UDG)







Structual informatics

Uracil DNA-glycosylase (UDG)

