Protein Structure Primer

Shoba Ranganathan
Bioinformatics Centre
National University of Singapore
(shoba@bic.nus.edu.sg)

Why protein structure?

- In the factory of the living cell, proteins are the workers, performing a variety of tasks
  - *Each protein adopts a particular folding pattern that determines its function*
    - The 3D structure of a protein brings into close proximity residues that are far apart in the amino acid sequence

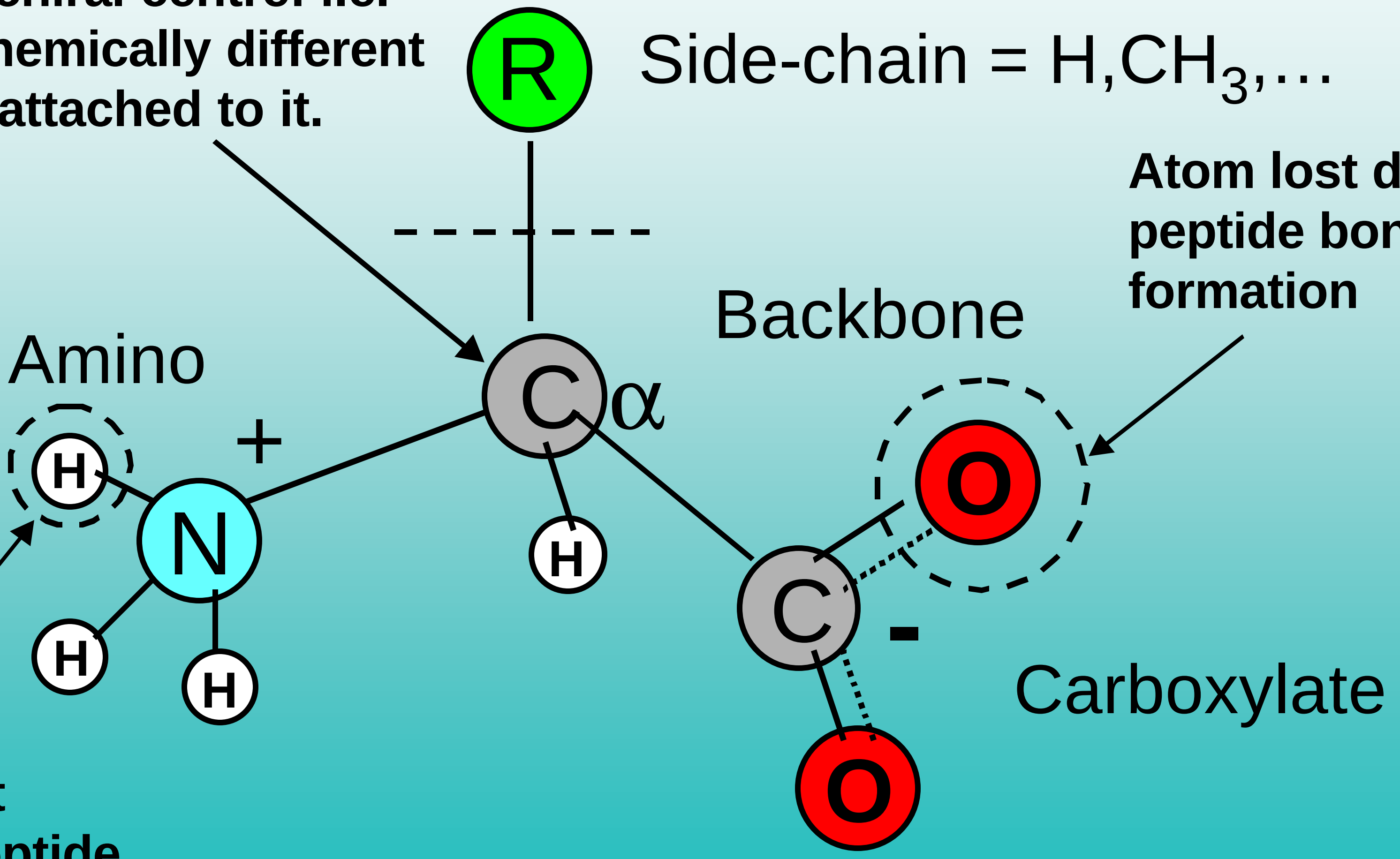
How does a protein fold?

- Most newly synthesized proteins fold without assistance!
  - *Ribonuclease A: denatured protein could refold and recover its activity (C. Anfinsen -1966)*
    - "Structure implies function"
      - *The amino acid sequence encodes the protein's structural information*

The basics

- Proteins are linear heteropolymers: one or more polypeptide chains
  - *Repeat units: 20 amino acid residues*
    - Range from a few 10s-1000s
      - *Three-dimensional shapes (“folds”) adopted vary enormously*
        - Experimental methods: X-ray crystallography, electron microscopy and NMR (nuclear magnetic resonance)

The (L-)amino acid



Ca is a chiral centre: i.e. has 4 chemically different groups attached to it.

Side-chain = H,CH$_3$,…

Cα

Amino

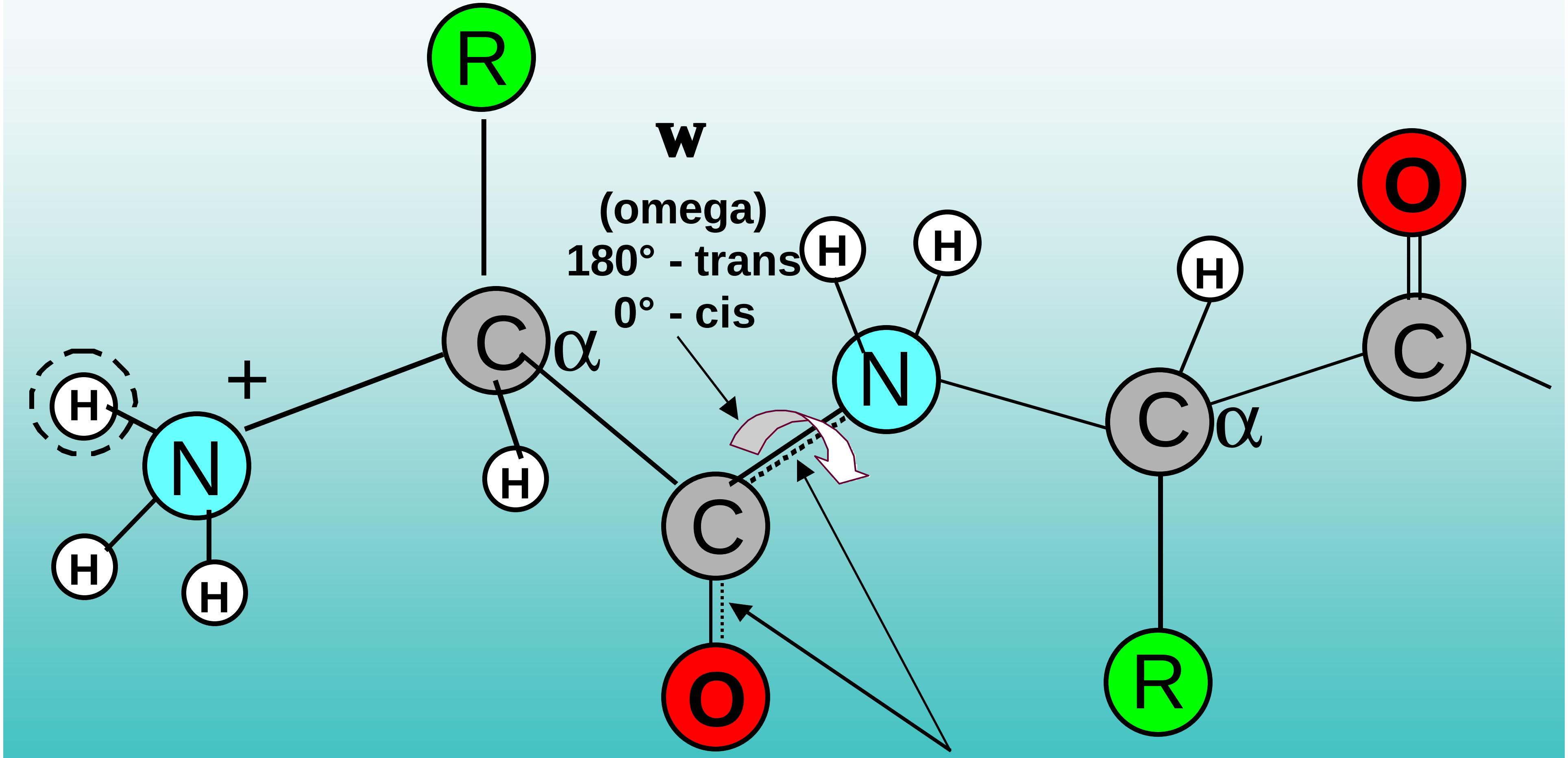
Backbone

Carboxylate

Atom lost during peptide bond formation

Atom lost during peptide bond formation

Formation of polypeptide chain



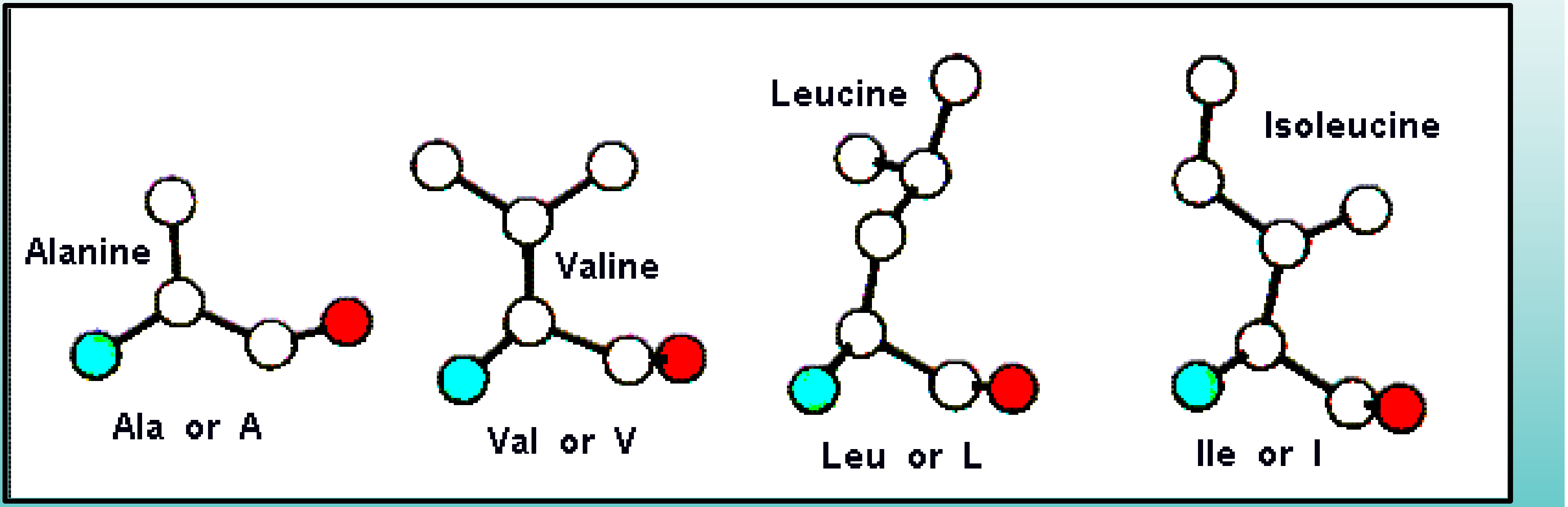
w

(omega)
180° - trans
0° - cis

Partial double bond character

Aliphatic residues

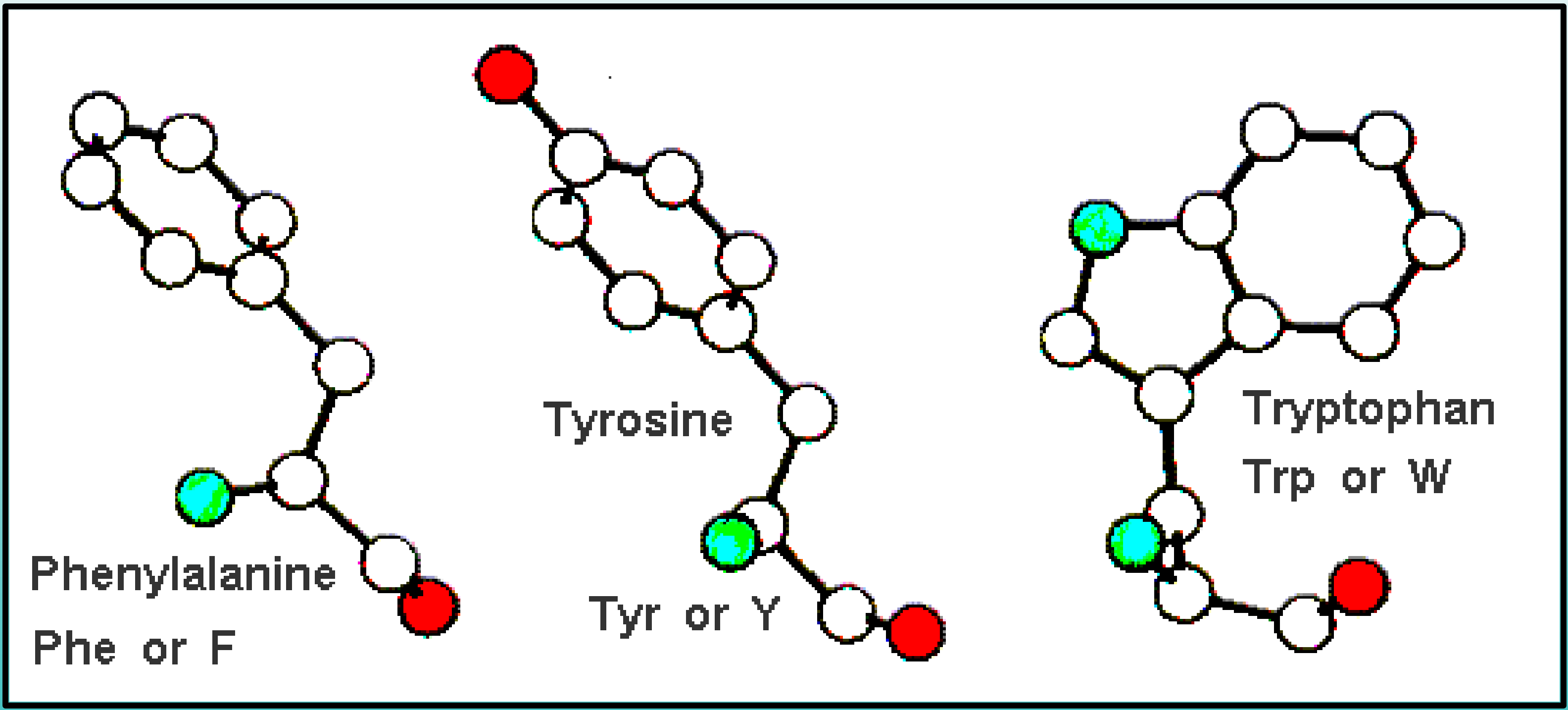
Hydrocarbon sidechains.



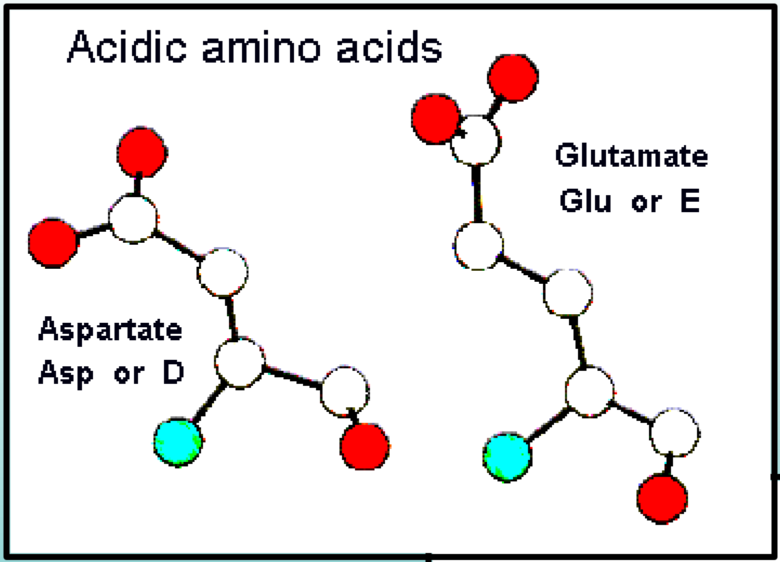
Only heavy atoms are usually shown (i.e. no hydrogens)

Also, the residue lacks the one oxygen atom in the carboxylate group.

Aromatic residues



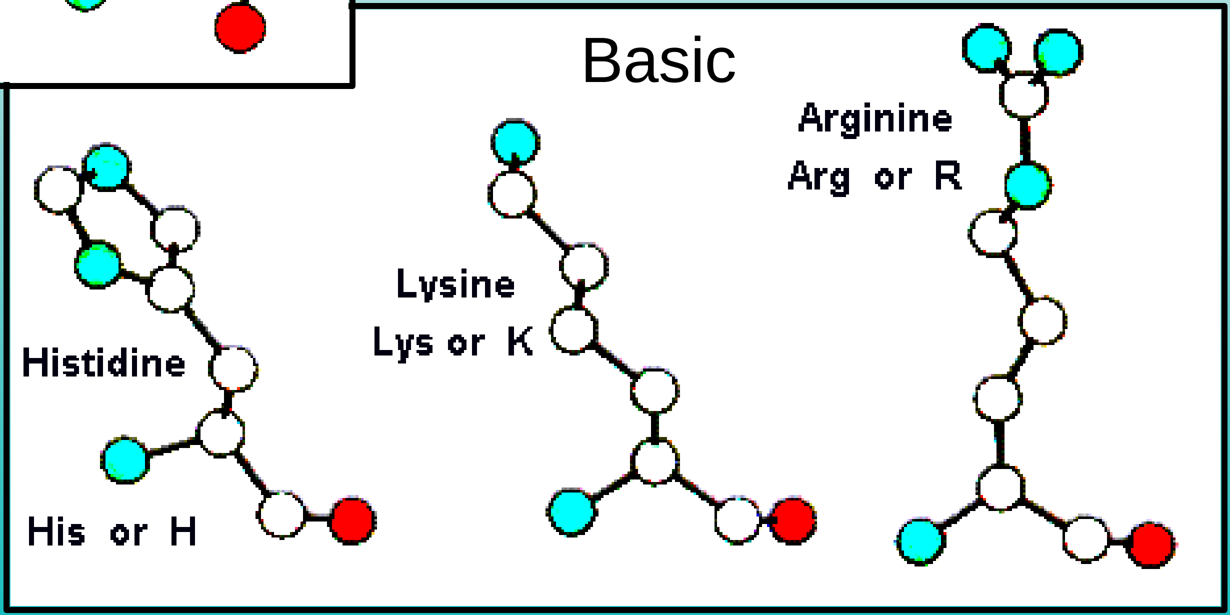
Charged residues



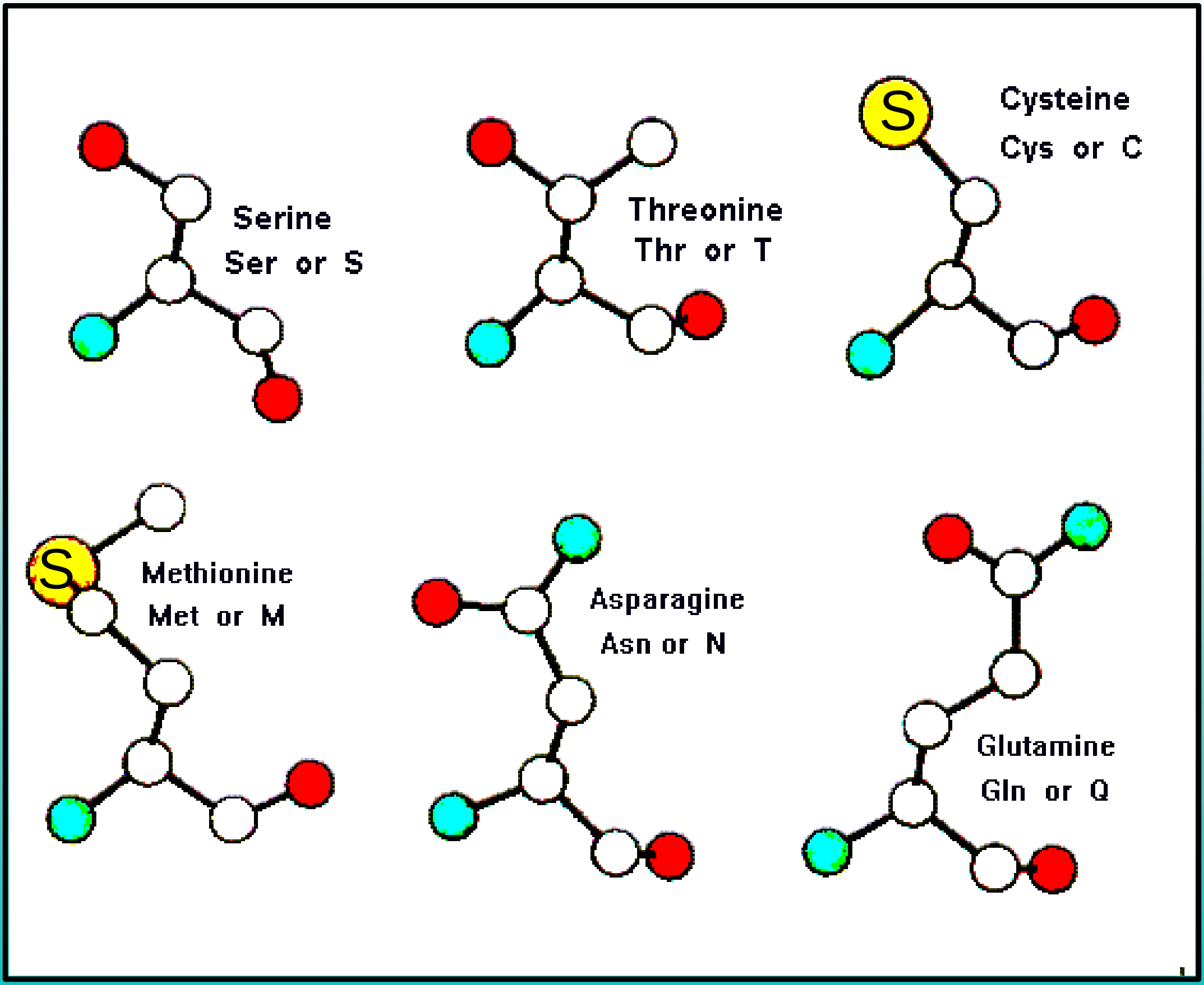
These contain side-chains that are charged under physiological conditions, i.e. pH 7.0:

•acidic – negative charge and

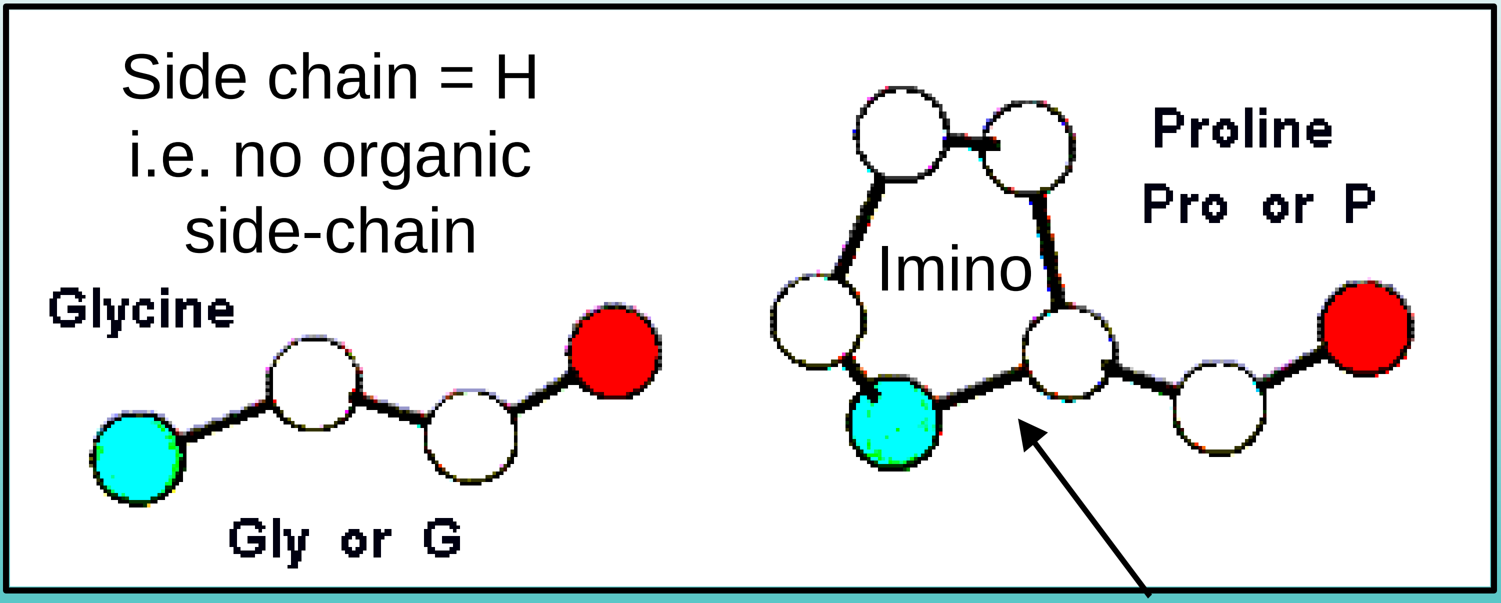
•basic – positive charge



Polar residues



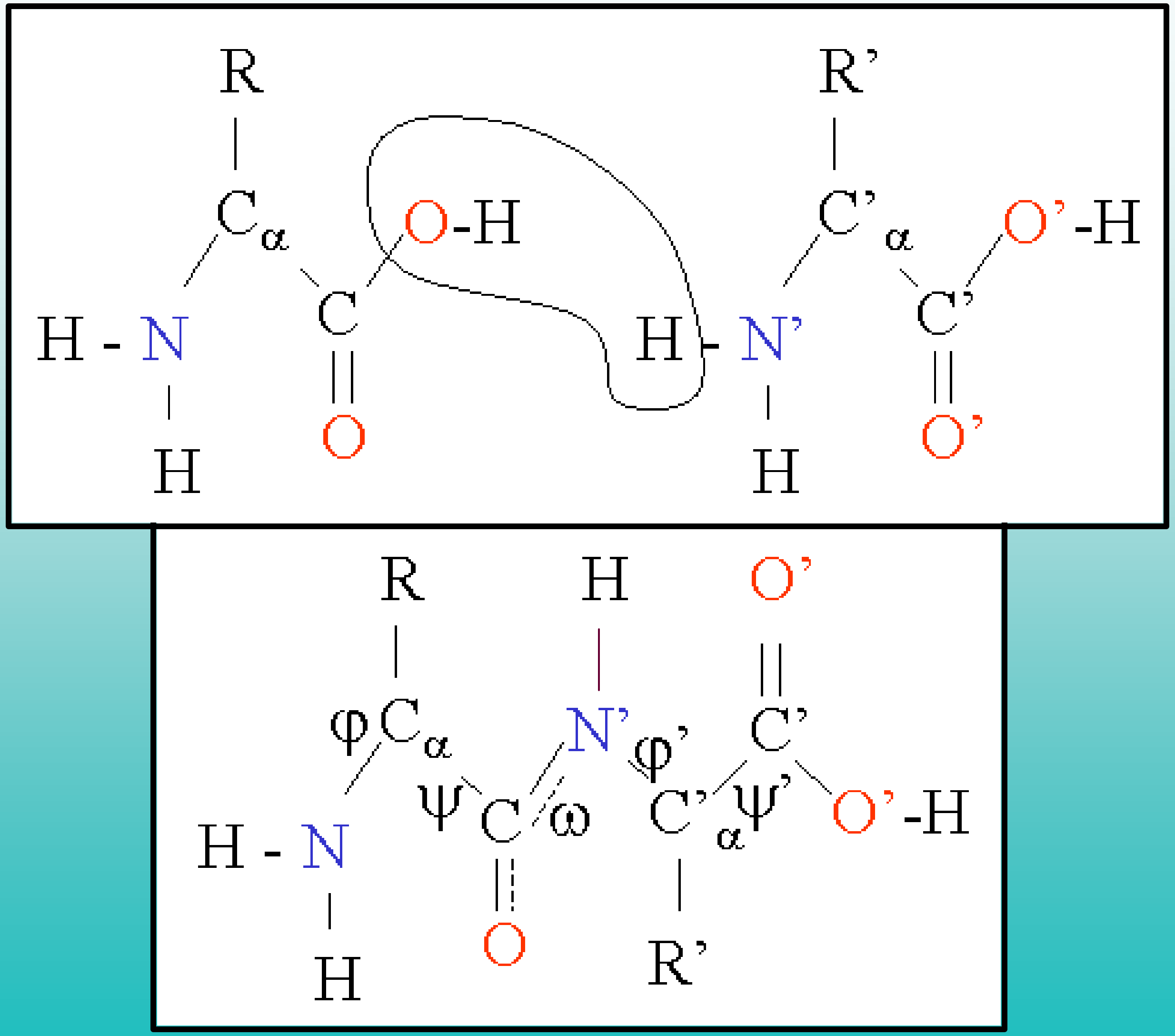
The odd couple



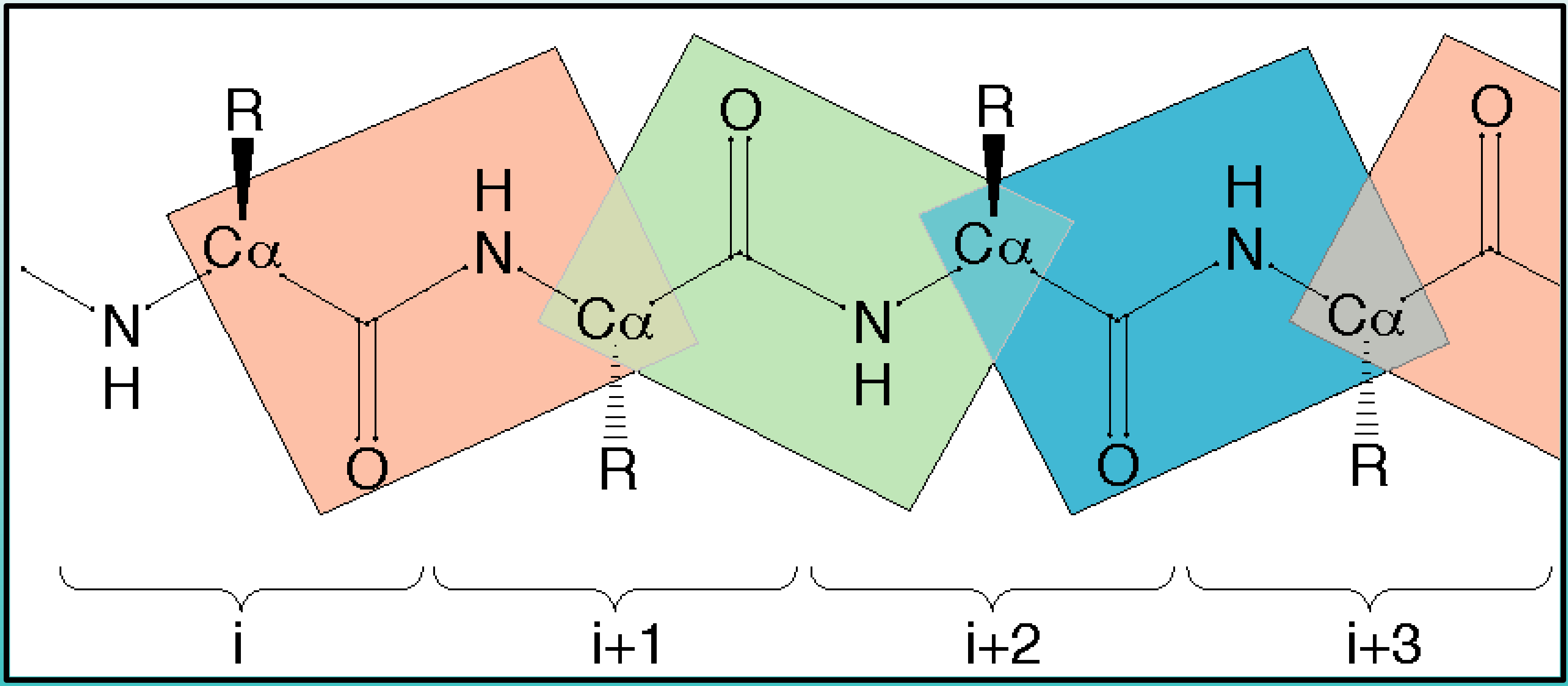
Side chain = H
i.e. no organic
side-chain

Can form cis-peptide bonds

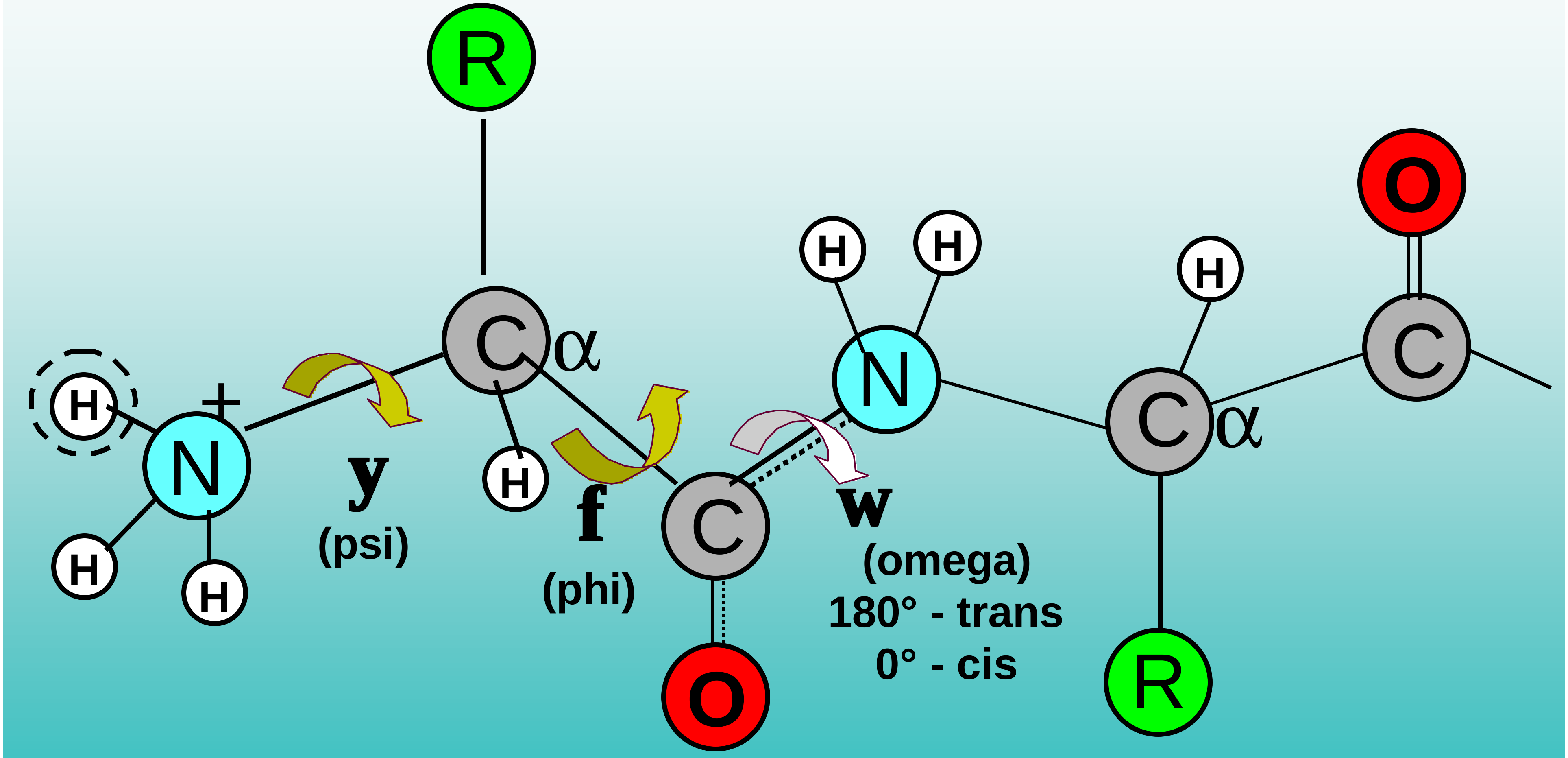
The peptide bond



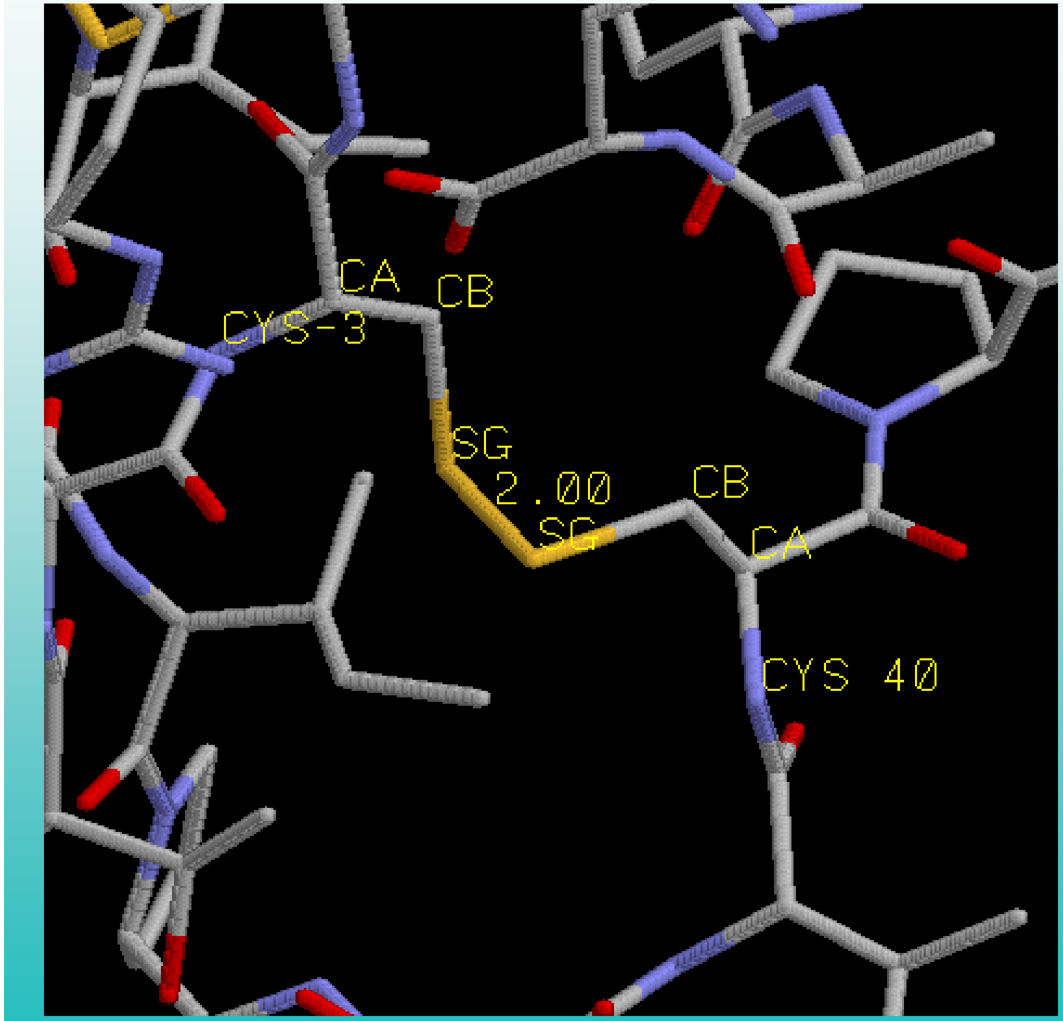
Coplanar atoms



Backbone Torsion Angles



The disulfide bond



- = “disulfide bridge”
- Only in extracellular proteins
- Formed by oxidation of the SH (thiol) group of cysteine residues
- Covalent bond between the Sγ (or ‘SG’) atoms of two cysteine residues

Structural information

- Protein Data Bank: maintained by the Research Collaboratory for Structural Bioinformatics
  - http://www.rcsb.org/pdb/
  - > 10,000 structures of proteins
  - Also contains structures of DNA, carbohydrates and protein-DNA complexes.
- Structures are principally determined by X-ray crystallography. Other methods are electron microscopy and NMR. Some structures are also theoretically predicted.

The PDB data

- Text files
- Each entry is identified by a unique 4-letter code: say 1emg
- 1emg entry
  - Header information
  - Atomic coordinates in Å (1 Ångstrom = 1.0e-10 m)

PDB Header details

- identifies the molecule, any modifications, date of release of PDB entry

```
HEADER    GREENFLUORESCENT PROTEIN                        12-NOV-98   1EMG
TITLE     GREEN FLUORESCENT PROTEIN (65-67 REPLACED BY CRO, S65T
TITLE    2 SUBSTITUTION, Q80R)
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: GREEN FLUORESCENT PROTEIN;
COMPND   3 CHAIN: A;
COMPND   4 ENGINEERED: YES;
COMPND   5 MUTATION: 65 - 67 REPLACED BY CRO, S65T SUBSTITUTION, Q80R
COMPND   6 SUBSTITUTION;
COMPND   7 BIOLOGICAL_UNIT: MONOMER
```

- organism, keywords, method
- Authors, reference, resolution if X-ray structure
- Sequence, x-reference to sequence databases

The data itself

- Coordinates for each heavy (non-hydrogen) atom from the first residue to the last

```
ATOM      1  N   SER A   2      29.089   9.397  51.904  1.00 81.75
ATOM      2  CA  SER A   2      27.883  10.162  52.185  1.00 79.71
ATOM      3  C   SER A   2      26.659   9.634  51.463  1.00 82.64
ATOM      4  O   SER A   2      26.718   8.686  50.686  1.00 81.02
ATOM      5  CB  SER A   2      28.039  11.660  51.932  1.00 75.59
ATOM      6  OG  SER A   2      27.582  12.038  50.639  1.00 43.28
-------
ATOM   1737  CD1 ILE A 229      39.535  21.584  52.346  1.00 41.62
TER    1738      ILE A 229
```

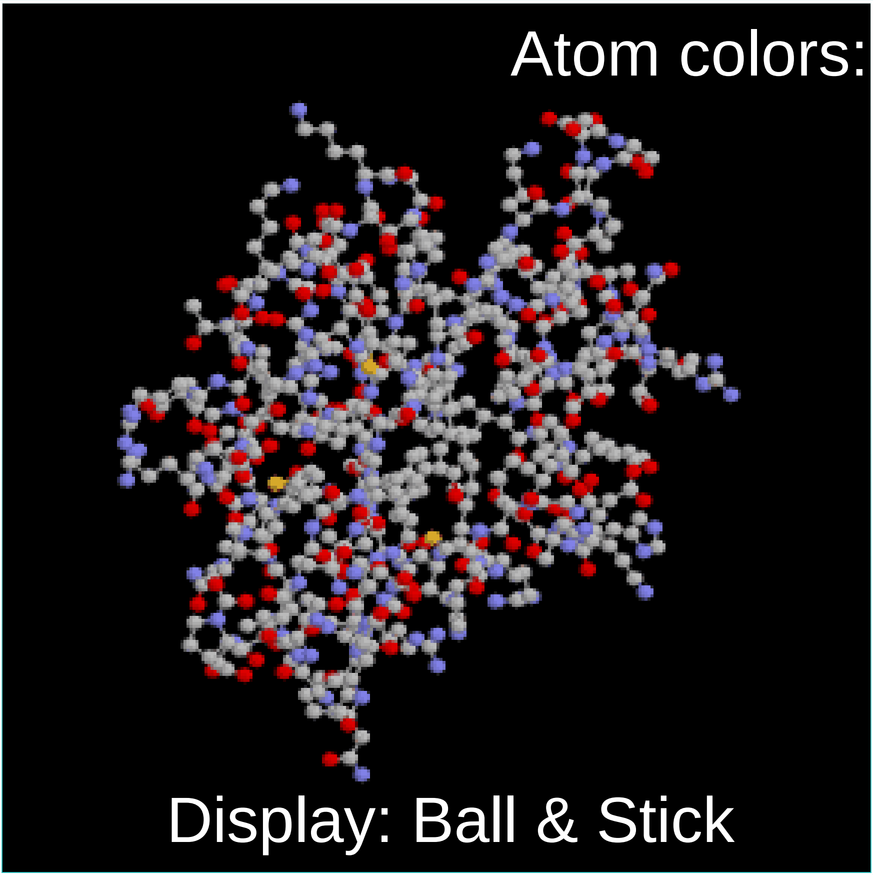
- Any ligands (starting with HETATM) follow the biomacromolecule
- O atoms of water molecules at the end

Visualizing PDB information

- RASMOL: most popular, available for all platforms
  http://www.bernstein-plus-sons.com/software/rasmol
- Swiss PDB Viewer: from Swiss-Prot
  http://expasy.nhri.org.tw/spdbv/
- Chemscape Chime Plug-in: for PC and Mac
  http://www.mdli.com/download/chimedown.html

RASMOL views - SH2 domain

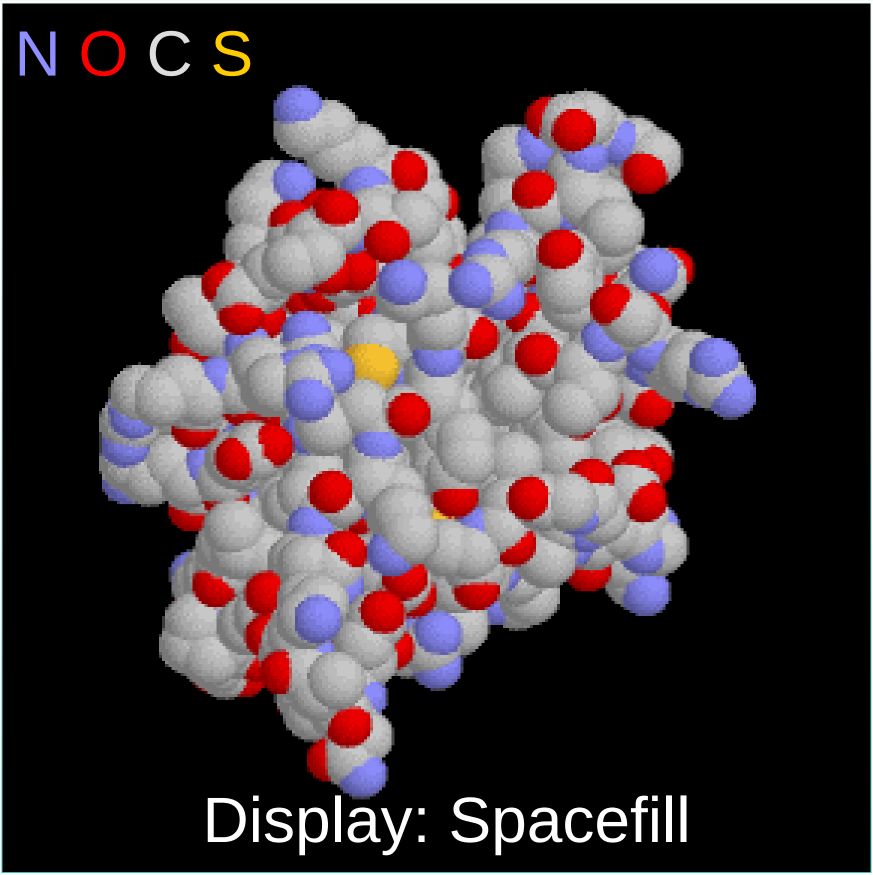
All-atom model



Atom colors:

Display: Ball & Stick

Space-filling model



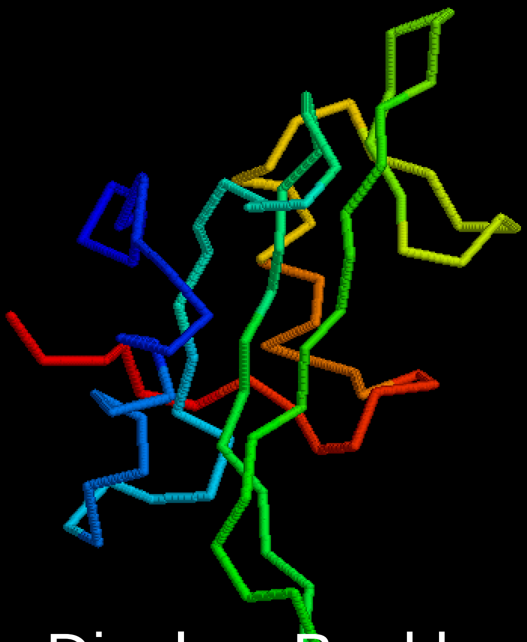
N O C S

Display: Spacefill

RASMOL views – 1sha

Cα Trace

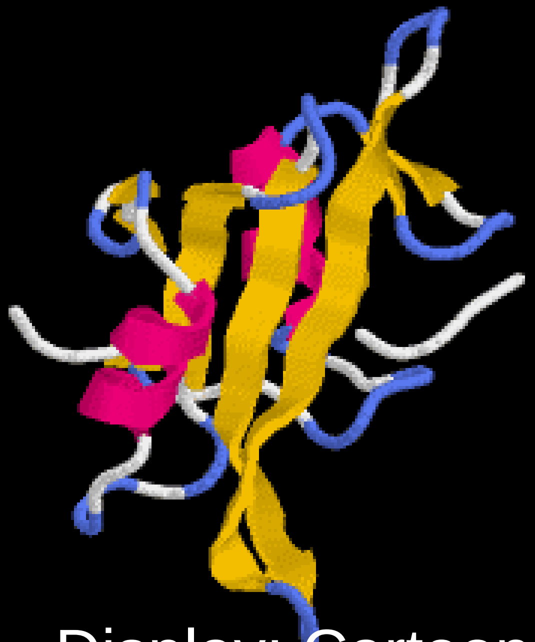
Rainbow coloring: N to C



Display: Backbone
Colours: Group

Ribbon

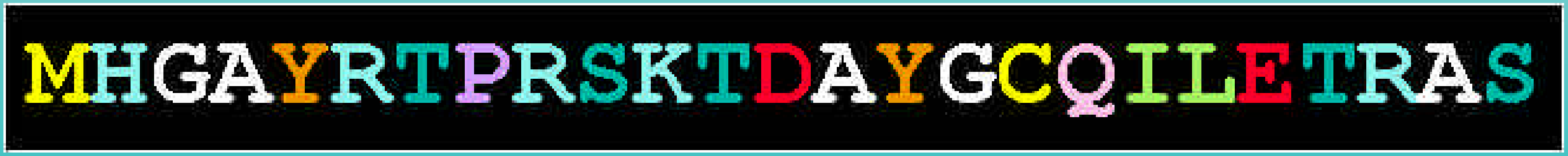
Coloring: by structural units



Display: Cartoons
Colours: Structure

Levels of protein structure: 0,1

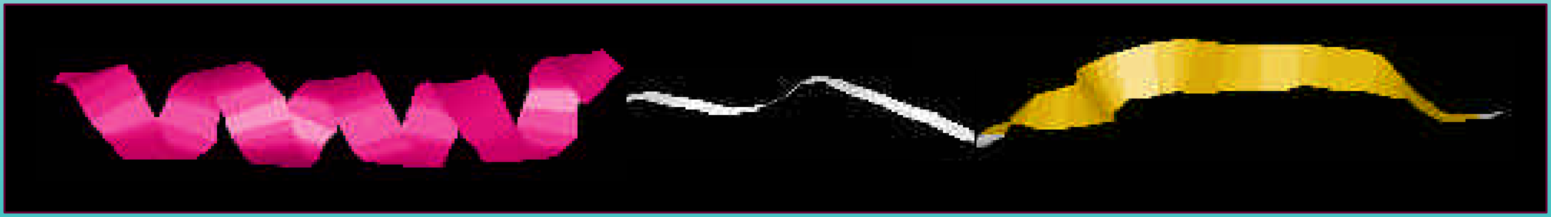
- Zeroth: amino acid composition – no structural information
- Primary
  - This is simply the order of covalent linkages along the polypeptide chain, i.e. the sequence itself



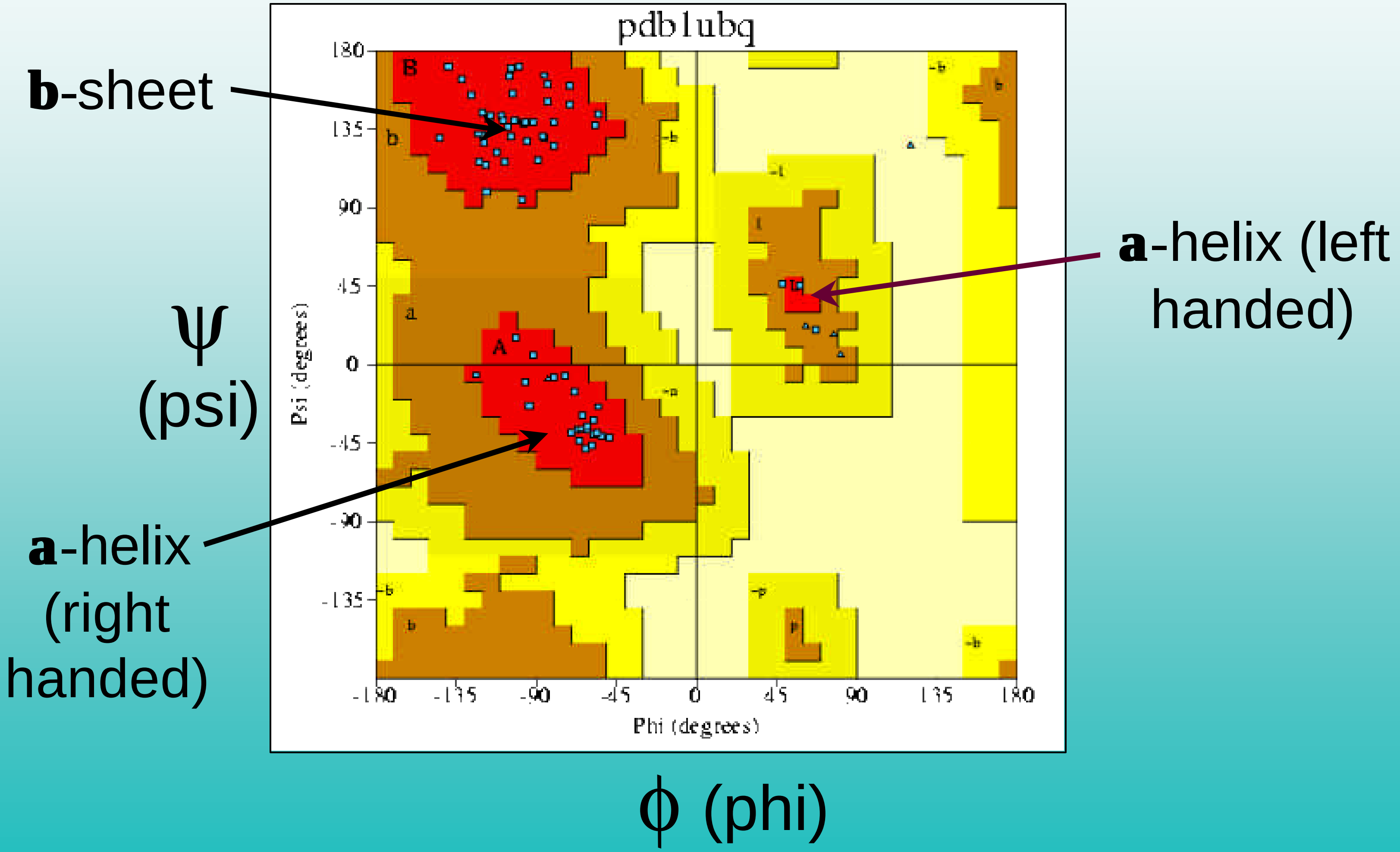
MHGAYRTPRSKTDAYGCQILETRAS

Levels of protein structure: 2

- Secondary
  - Local organization of the protein backbone: **a**-helix, **b**-strand (which assemble into **b**-sheets), turn and interconnecting loop



Ramachandran / phi-psi plot



b-sheet

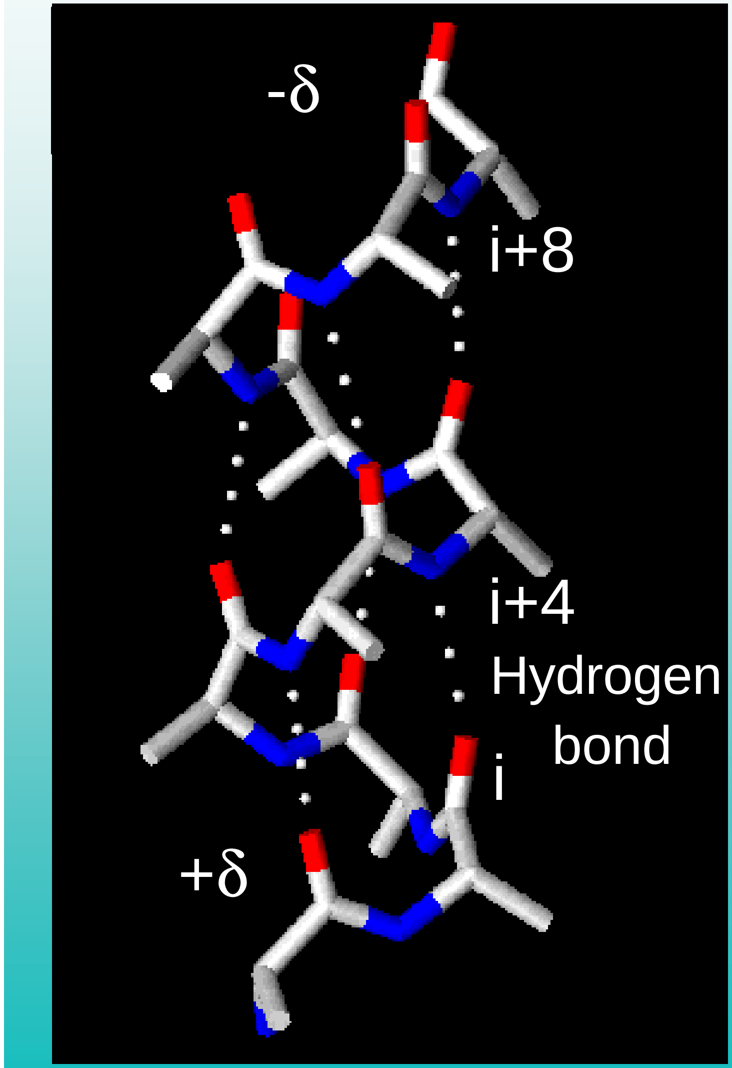
ψ (psi)

a-helix (right handed)

a-helix (left handed)

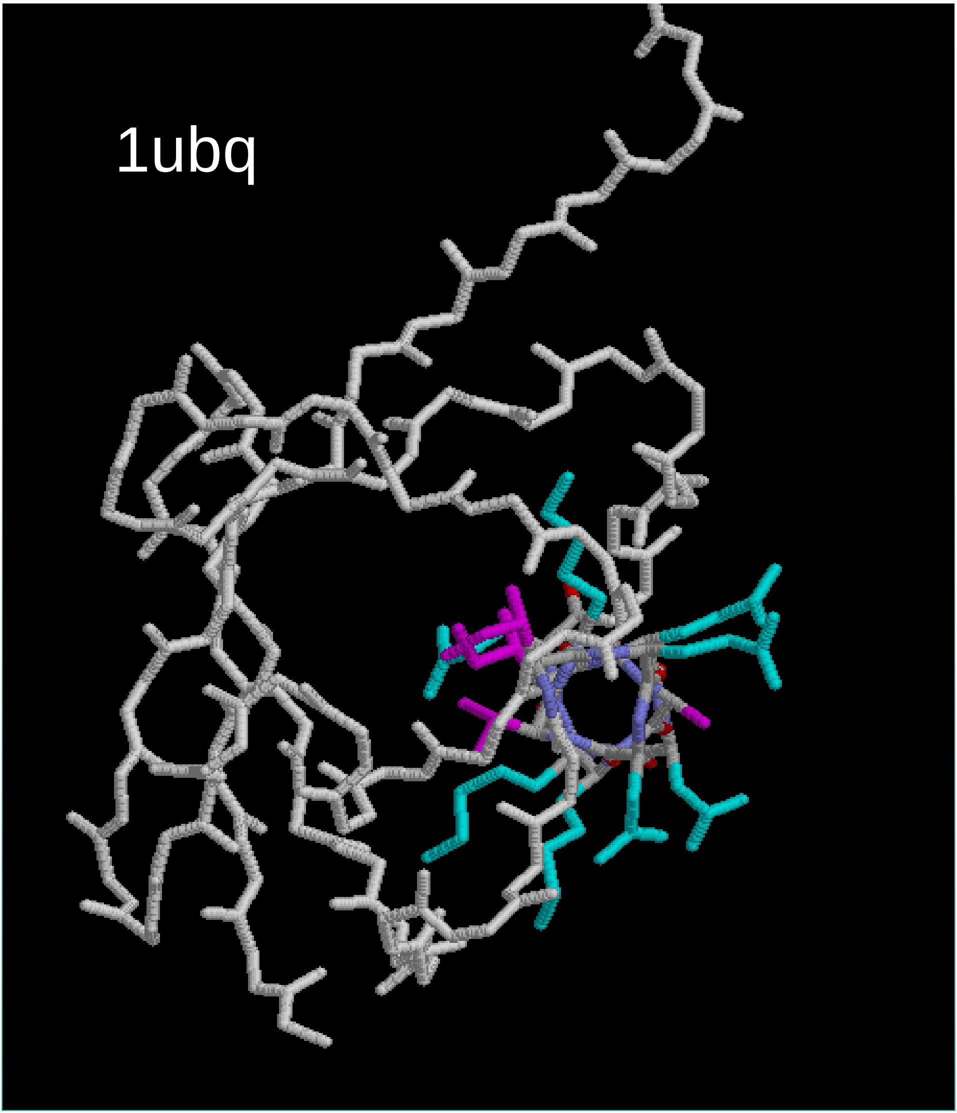
ϕ (phi)

The **a**-helix



- First structure to be predicted (Pauling, Corey, Branson: 1951) and experimentally solved (Kendrew *et al.* 1958) – myoglobin
- Turn: 3.6 residues
- Pitch: 5.4 Å/turn
- Rise: 1.5 Å/residue
- Dipole: start +ve and end –ve
- One of the most closely packed arrangement of residues

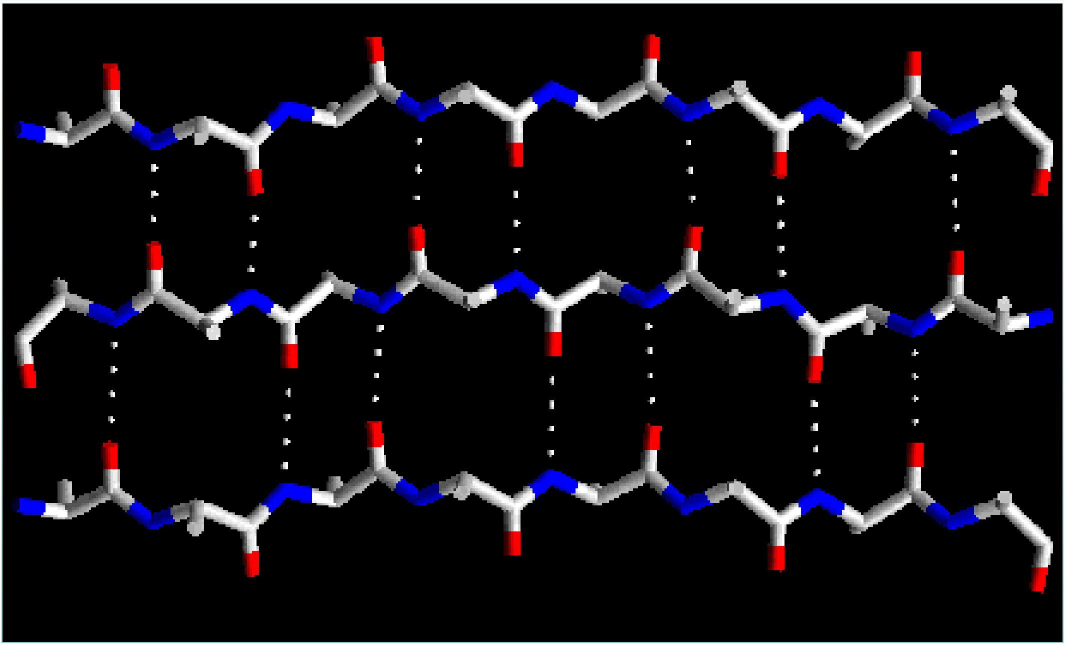
Properties of the **a**-helix



- Side-chains project outwards: proline only fits the start
- Amphipathicity if solvent exposed: hydrophilic residues in cyan; hydrophobic resides in magenta

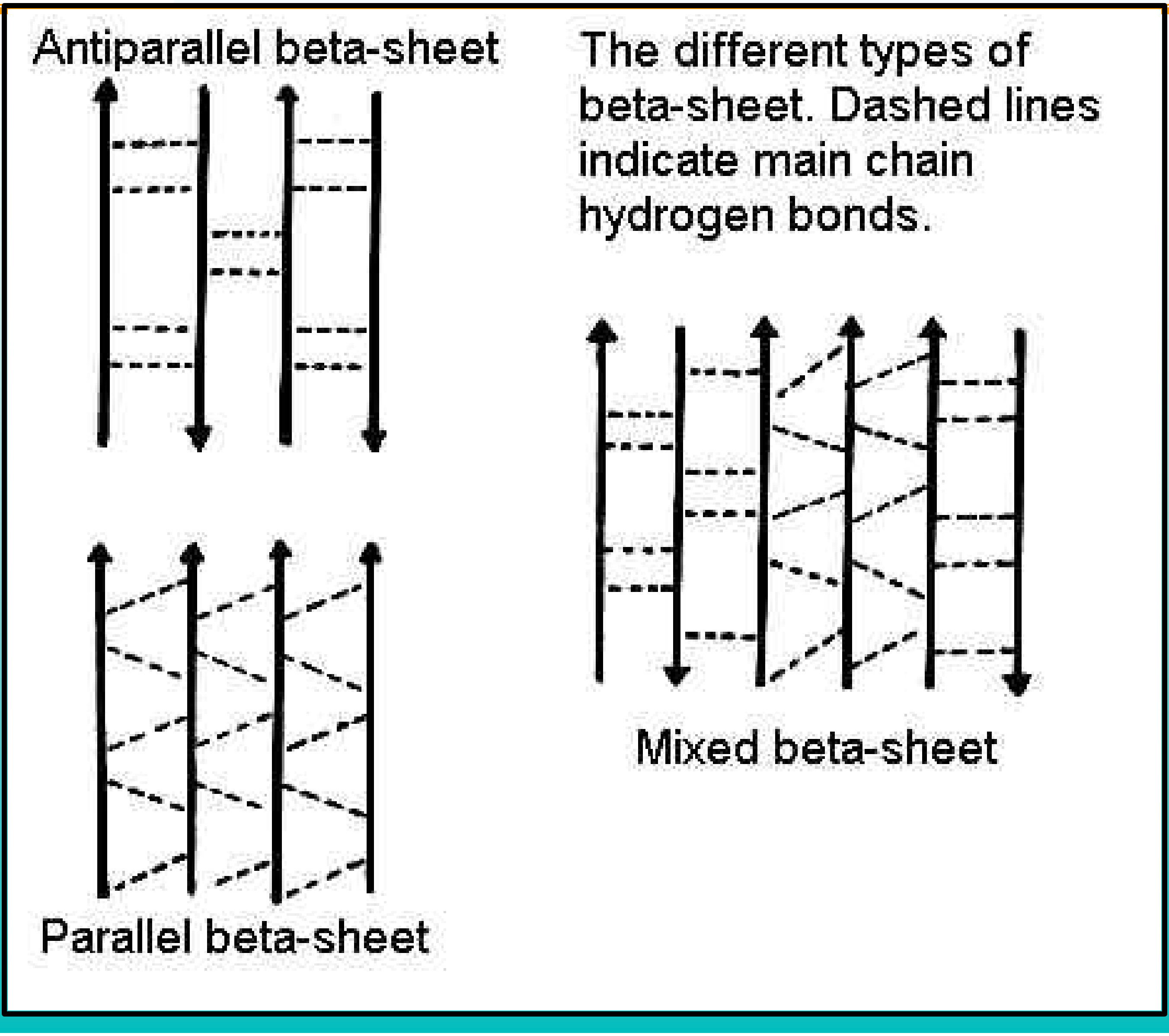
The **b**-sheet

- Side-chains project alternately up or down



- Amphipathicity if solvent exposed: hydrophilic residues on one face; hydrophobic ones on the other
- Backbone almost fully extended: thus one of the lost loosely packed arrangements of residues.

Topologies of **b**-sheets



Antiparallel beta-sheet

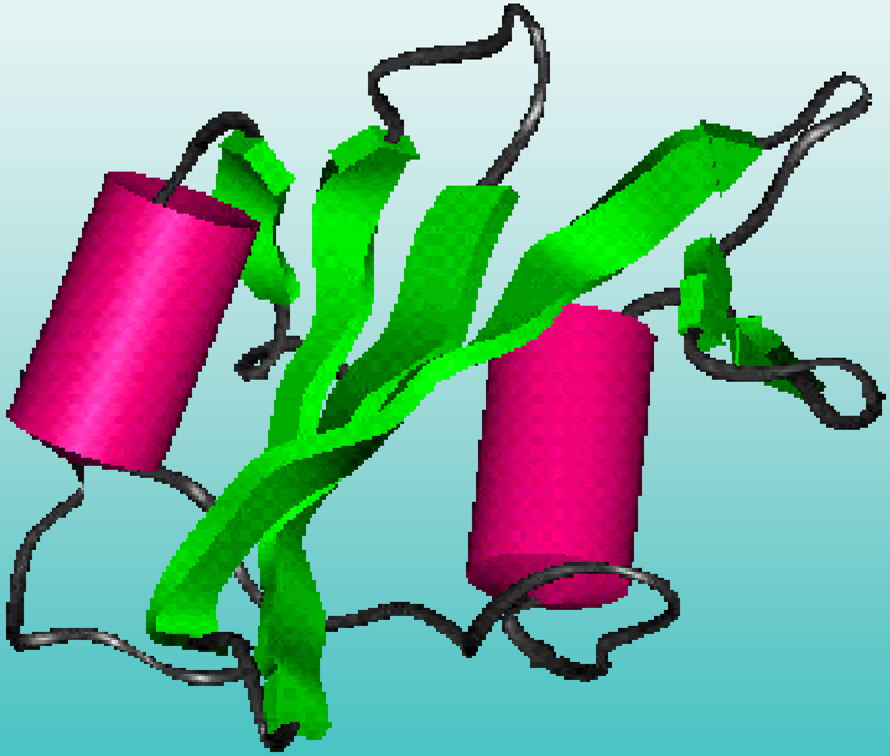
Parallel beta-sheet

Mixed beta-sheet

The different types of beta-sheet. Dashed lines indicate main chain hydrogen bonds.

bioinformatics centre NATIONAL UNIVERSITY OF SINGAPORE

Levels of protein structure: 3

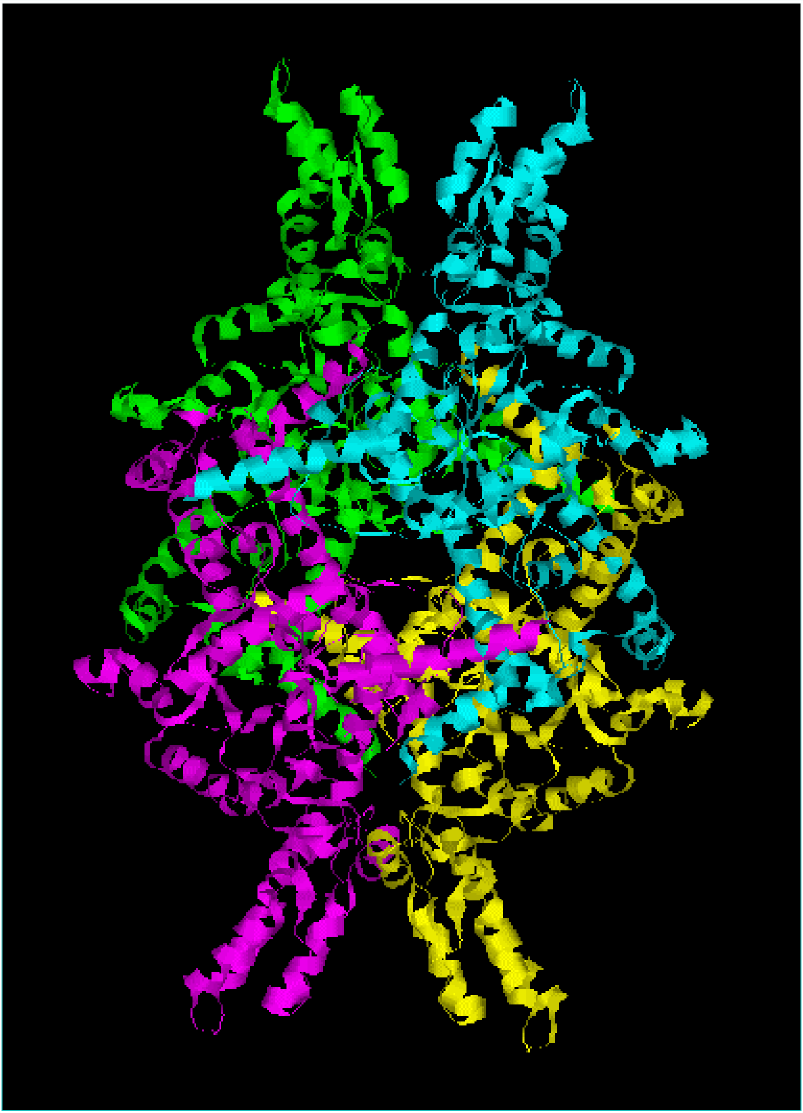


- Tertiary
  - packing of secondary structure elements into a compact spatial unit
  - “Fold” or domain – this is the level to which structure prediction is currently possible

Driving forces in protein folding

- Stabilization by forming hydrogen bonds
- Exposing hydrophilic residues (with charged and polar side-chains) and burying hydrophobic residues (with aliphatic and aromatic side-chains)
- For small proteins (usually > 75 residues)
  - Formation of disulfide bridges
  - Interactions with metal ions

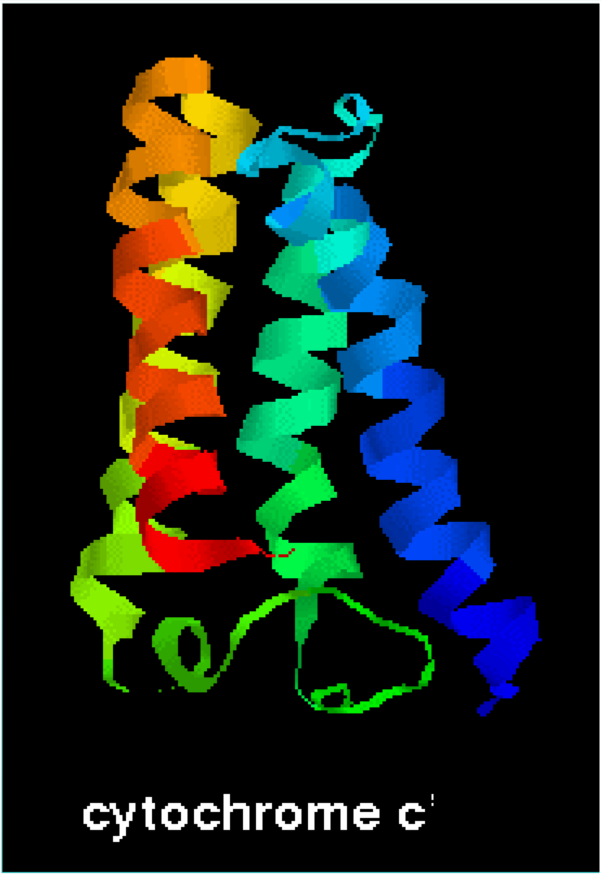
Levels of protein structure: 4



- Quaternary
  - Assembly of homo- or heteromeric protein chains
  - Usually the functional unit of a protein, especially for enzymes

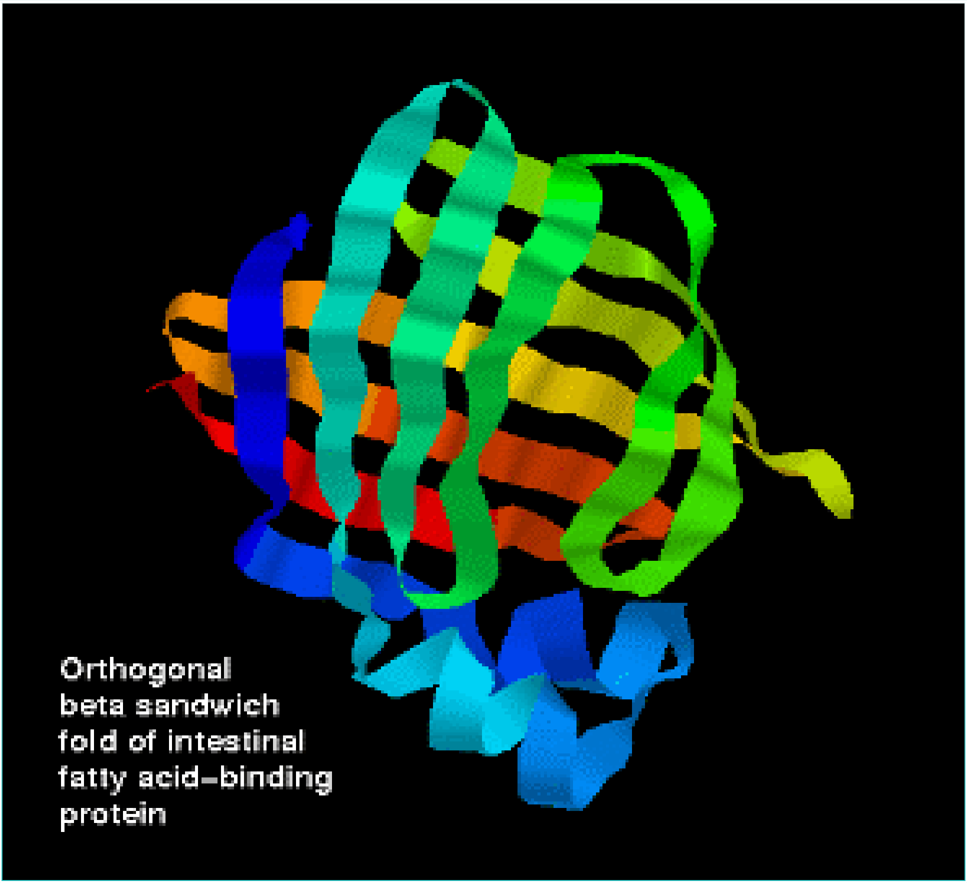
Structural classes: 1

All-**a** (helical)



cytochrome c'

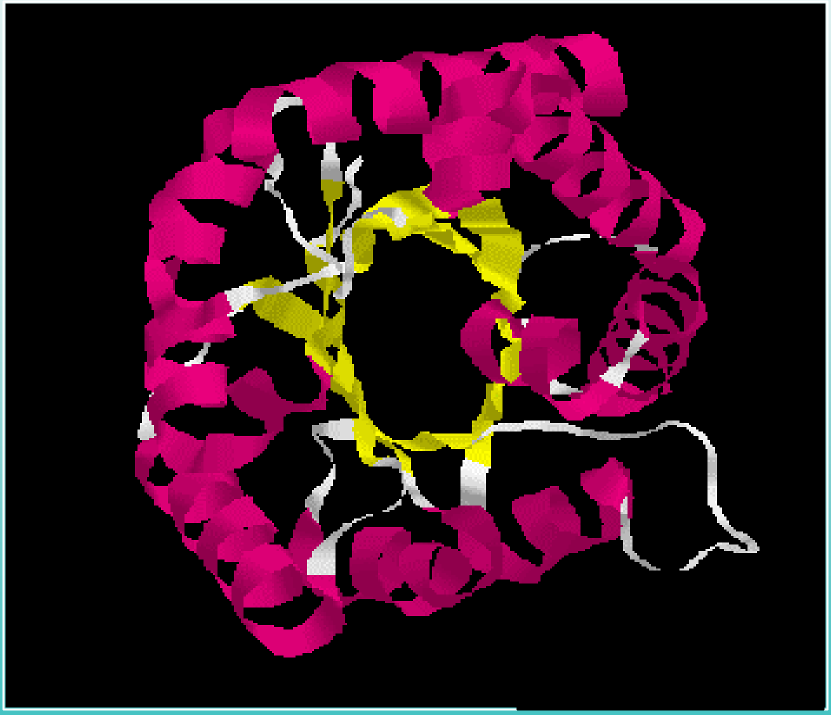
All-**b** (sheet)



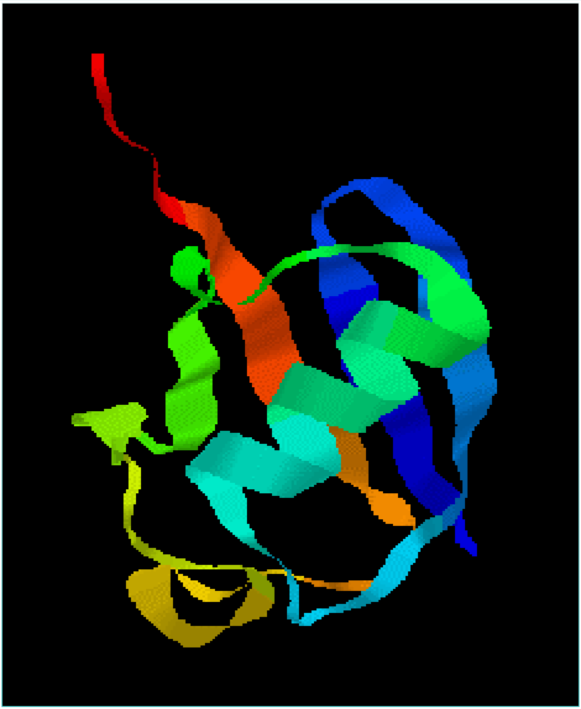
Orthogonal beta sandwich fold of intestinal fatty acid-binding protein

Structural classes: 2

a/b (parallel **b**-sheet)



a+b (antiparallel **b**-sheet)



Most popular class!

Domain: a LEGO piece

- A domain is a compact folding unit of protein structure, usually associated with a function.
    - It is usually a “fold” - in the case of monomeric soluble proteins.
        - Comprises normally only one protein chain: rare examples involving 2 chains are known.
            - Domains can be shared between different proteins.

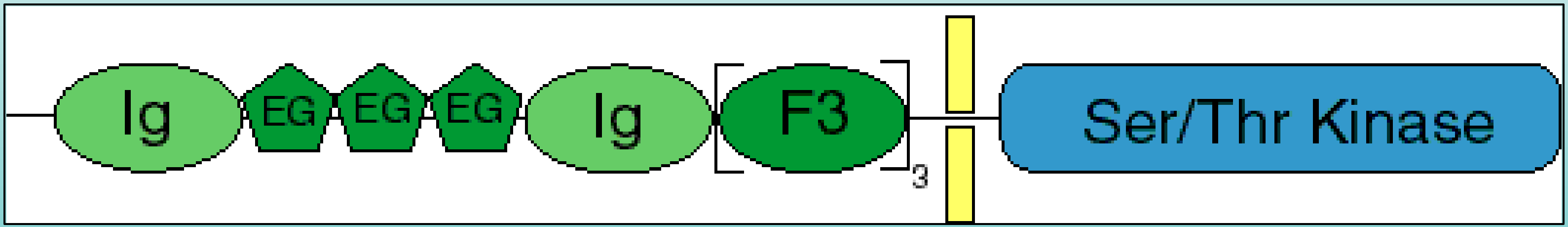
A multi-domain protein

L-lactate dehydrogenase (LDH)

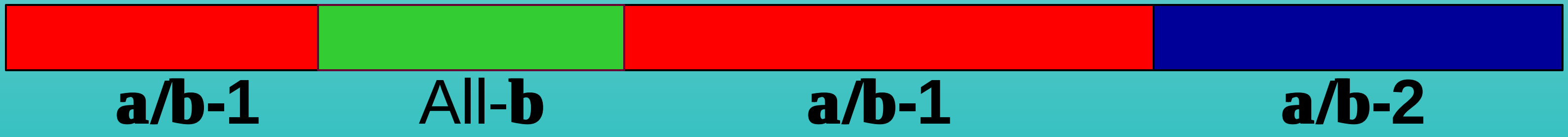
- Essential enzyme in anaerobic glycolysis
  - Catalyses the reversible conversion of pyruvate to L-lactate - oxamate is an inhibitor
    - Nicotinamide adenine dinucleotide (NAD) is the cofactor for the reaction, with a proton from a His residue of the proteinIt is usually a “fold” - in the case of monomeric soluble proteins.

Protein architectures

- Beads-on-a-string: sequential location: tyrosine-protein kinase receptor TIE-1 (immunoglobulin, EGF, fibronectin type-3 and protein kinase)

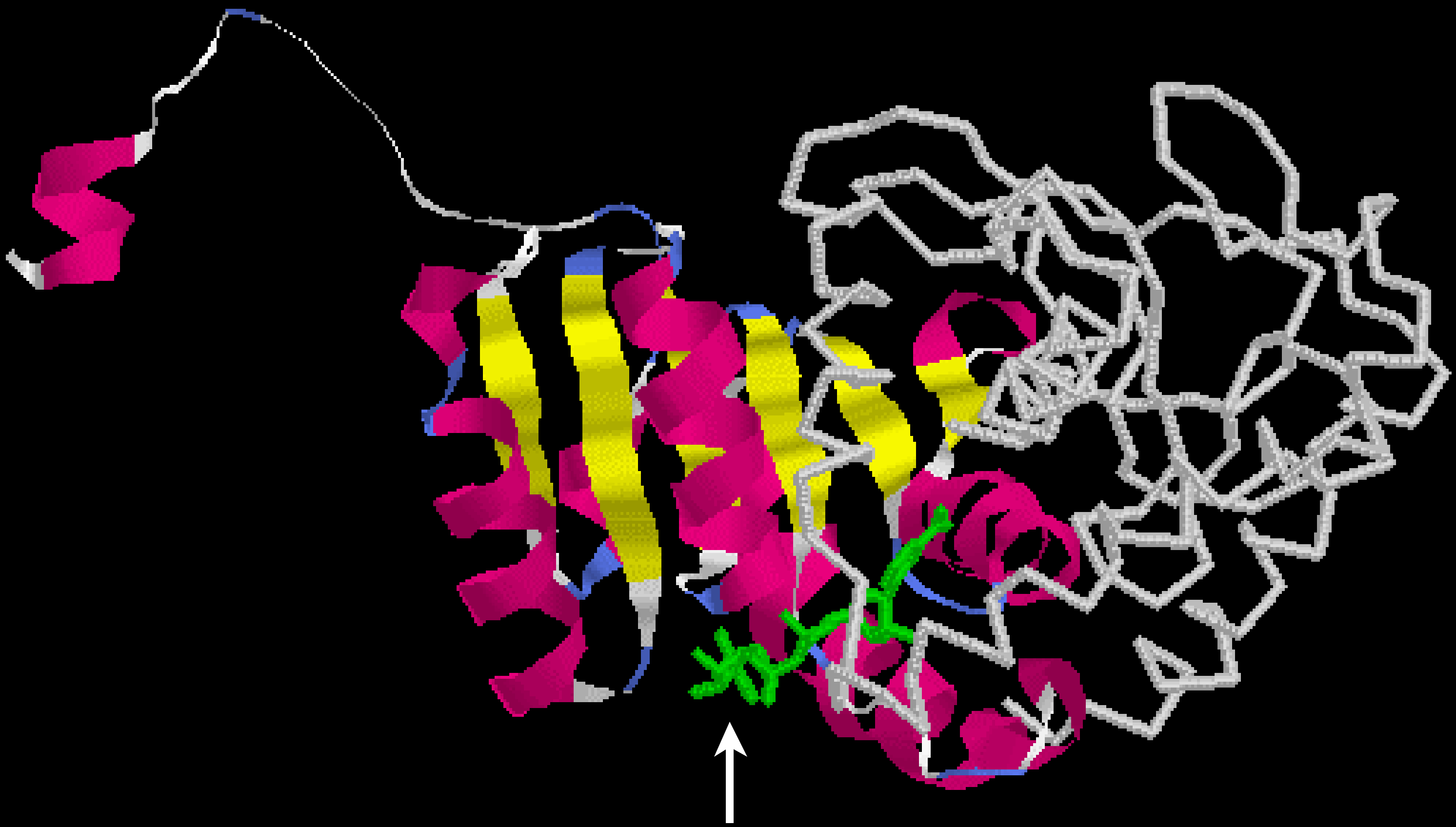


- Domain insertions: “plugged-in” - pyruvate kinase (1pkn): 3 domains - split domain 1



LDH – domain structure

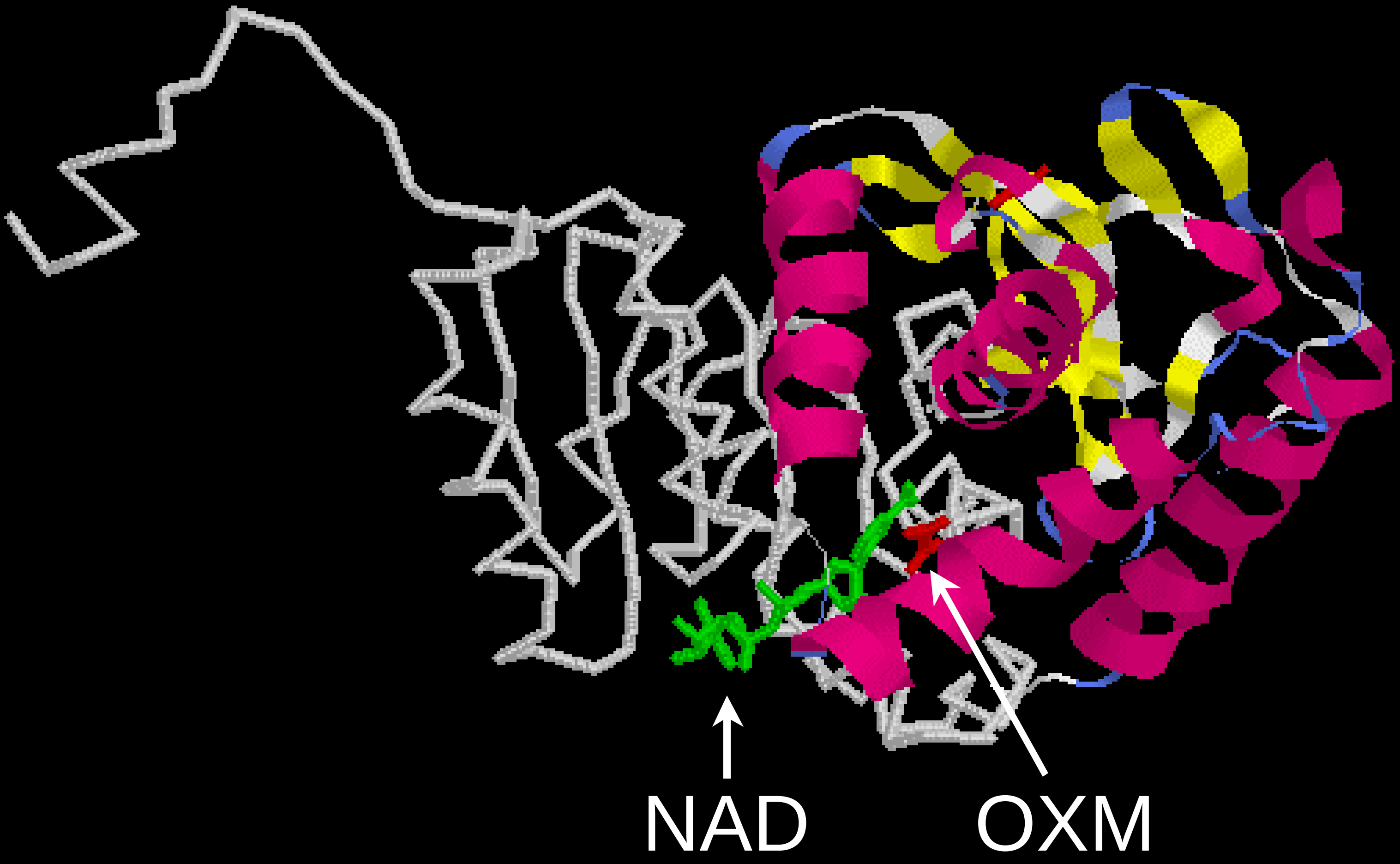
Domain 1: Rossman-fold (α/β)



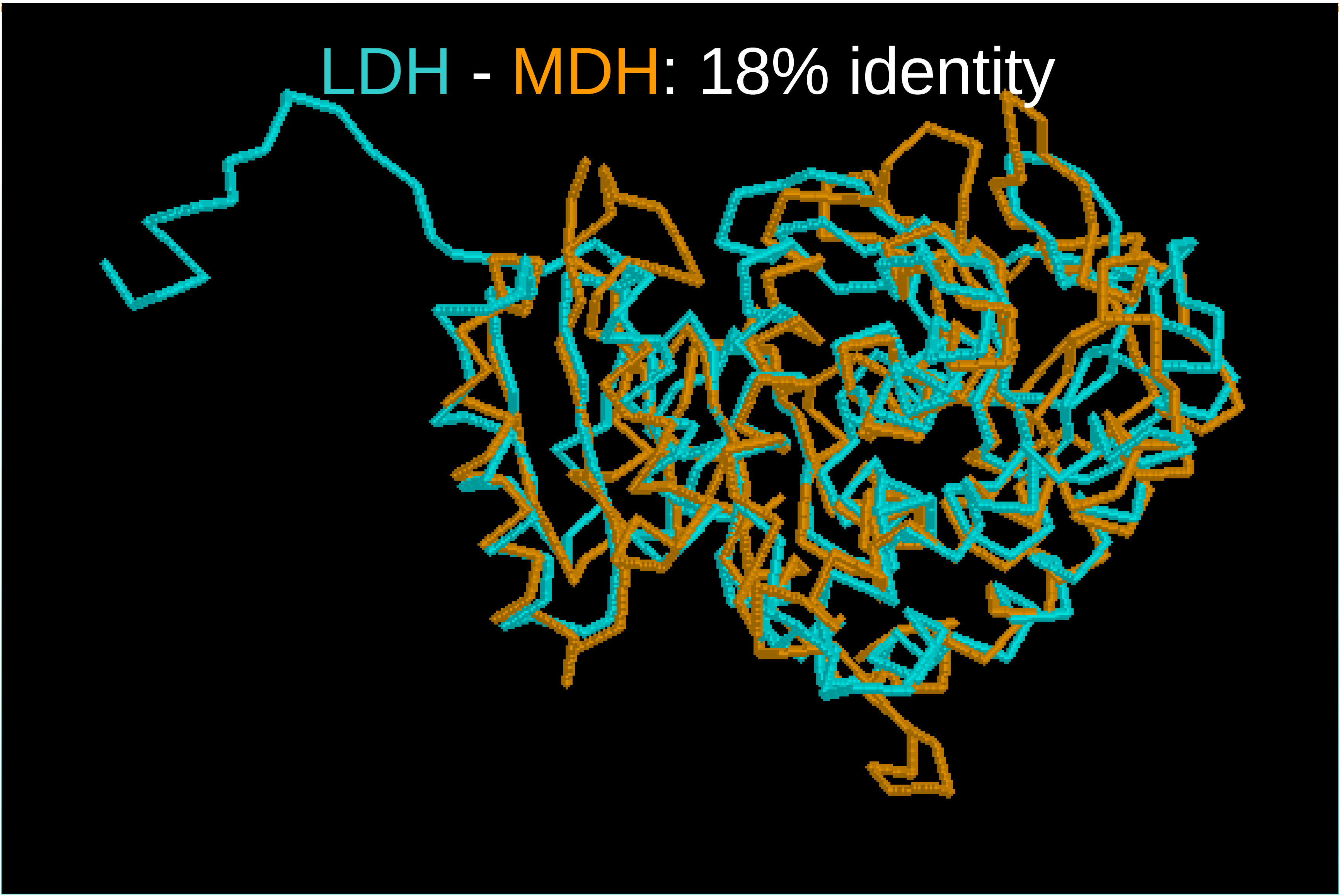
NAD

LDH – domain structure

Domain 2: substrate-binding (α+β):

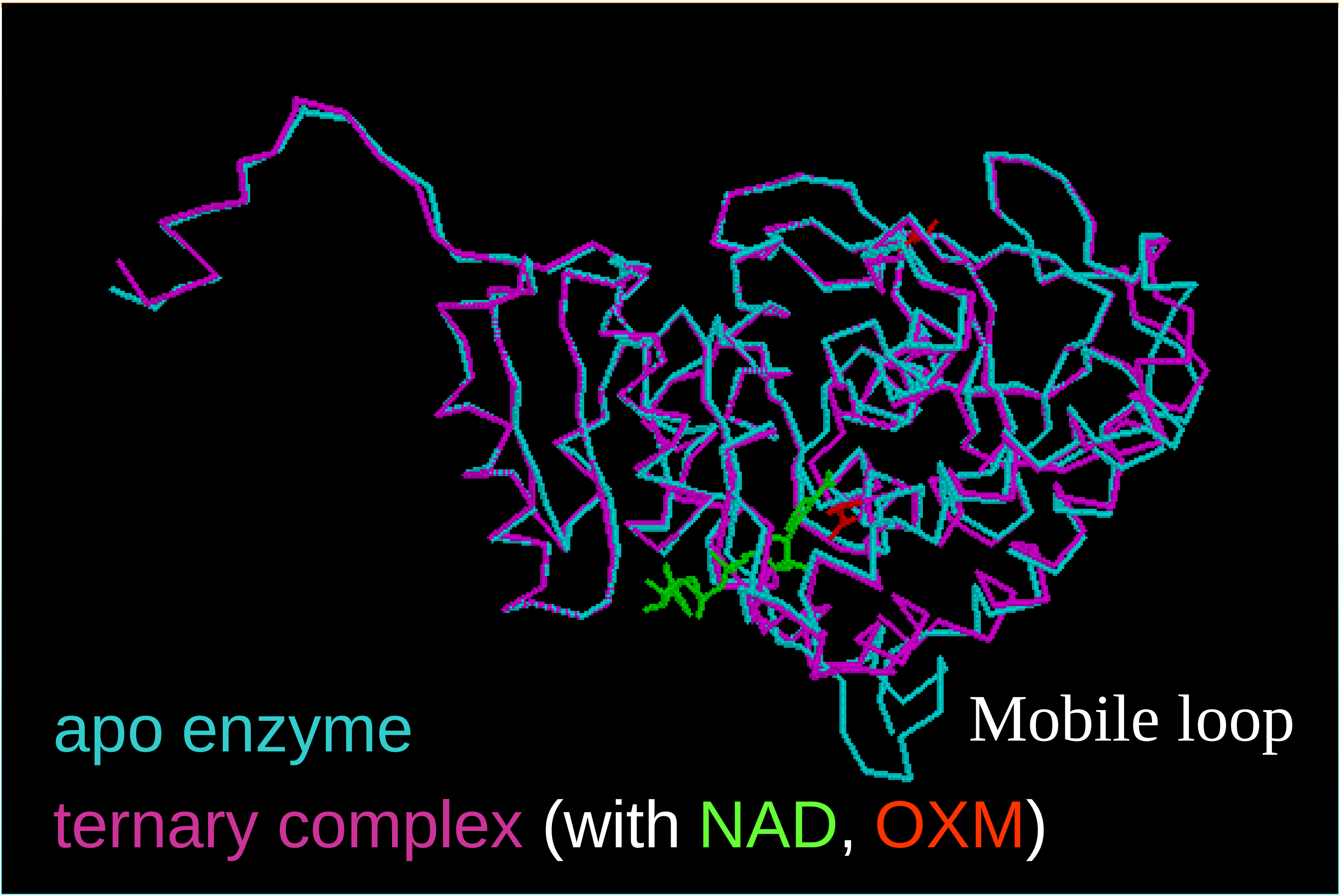


Lactate & Malate dehydrogenases



LDH - MDH: 18% identity

Structural changes in LDH



Mobile loop

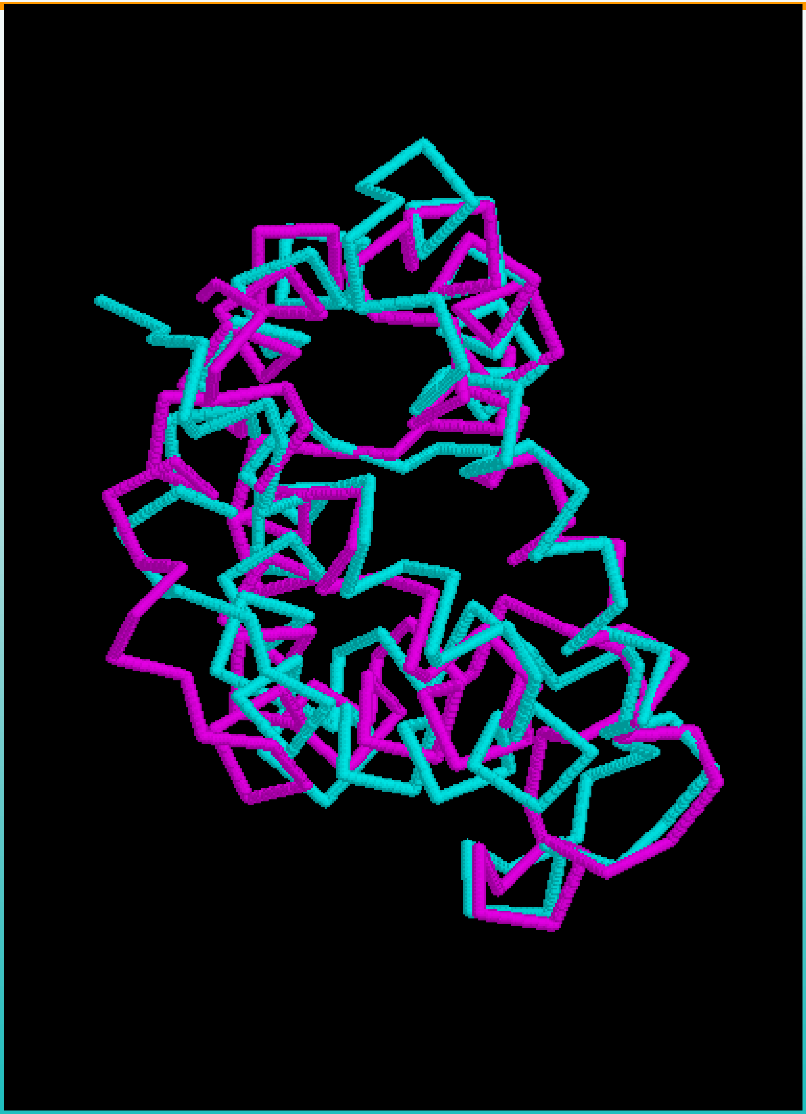
apo enzyme

ternary complex (with NAD, OXM)

bioinformatics centre NATIONAL UNIVERSITY OF SINGAPORE

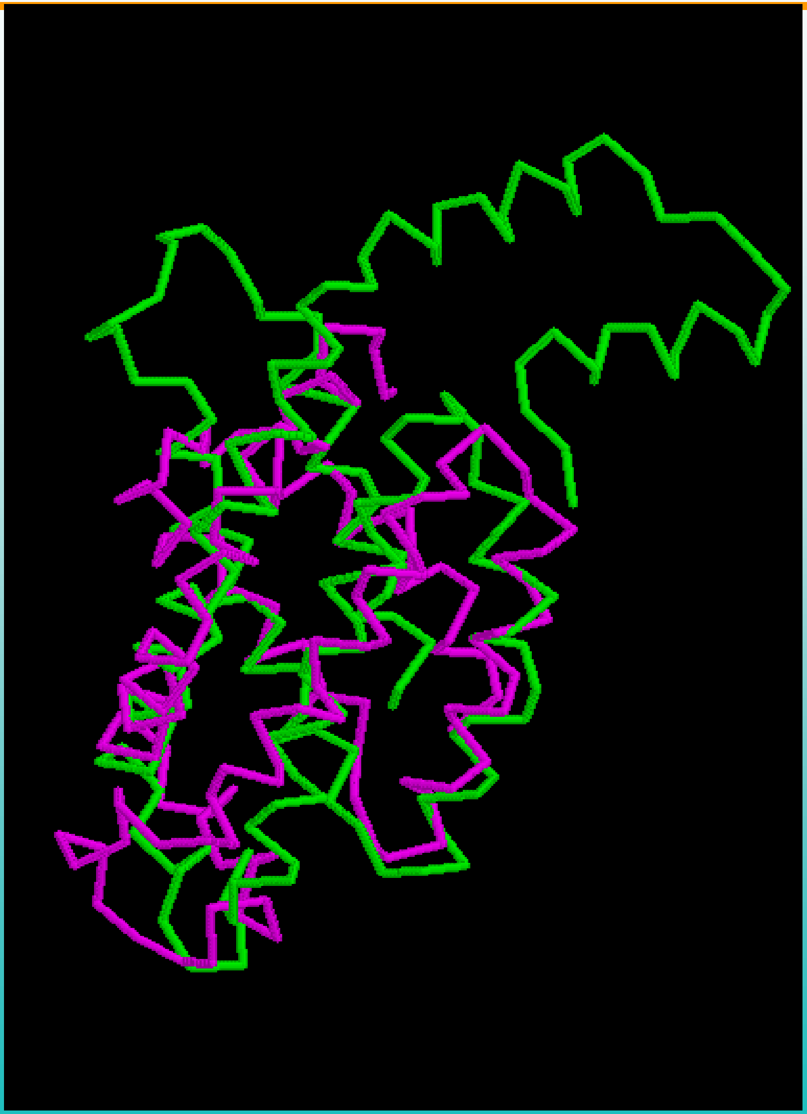
Homologous folds

- Hemoglobin and erythrocruorin: 31% sequence identity
- Normally at least 25% sequence identity
- Identical or closely related functions



Analogous folds

- Hemoglobin and phycocyanin: 9% sequence identity
- Structural architechture quite similar
- Functions not conserved.



Structure comparison facts

- Proteins adopt a limited number of topologies.
    - Homologous sequences show very similar structures: variations in non-conserved regions.
        - In the absence of sequence homology, some folds are preferred by vastly different sequences.

Structure comparison facts

- The “active site” (a collection of functionally critical residues) is remarkably conserved, even when the protein fold is different.
  - Structural models (especially those based on homology) provide insights into possible function for new proteins.
    - Implications for
      - protein engineering
      - ligand/drug design,
      - function assignment for genomic data.