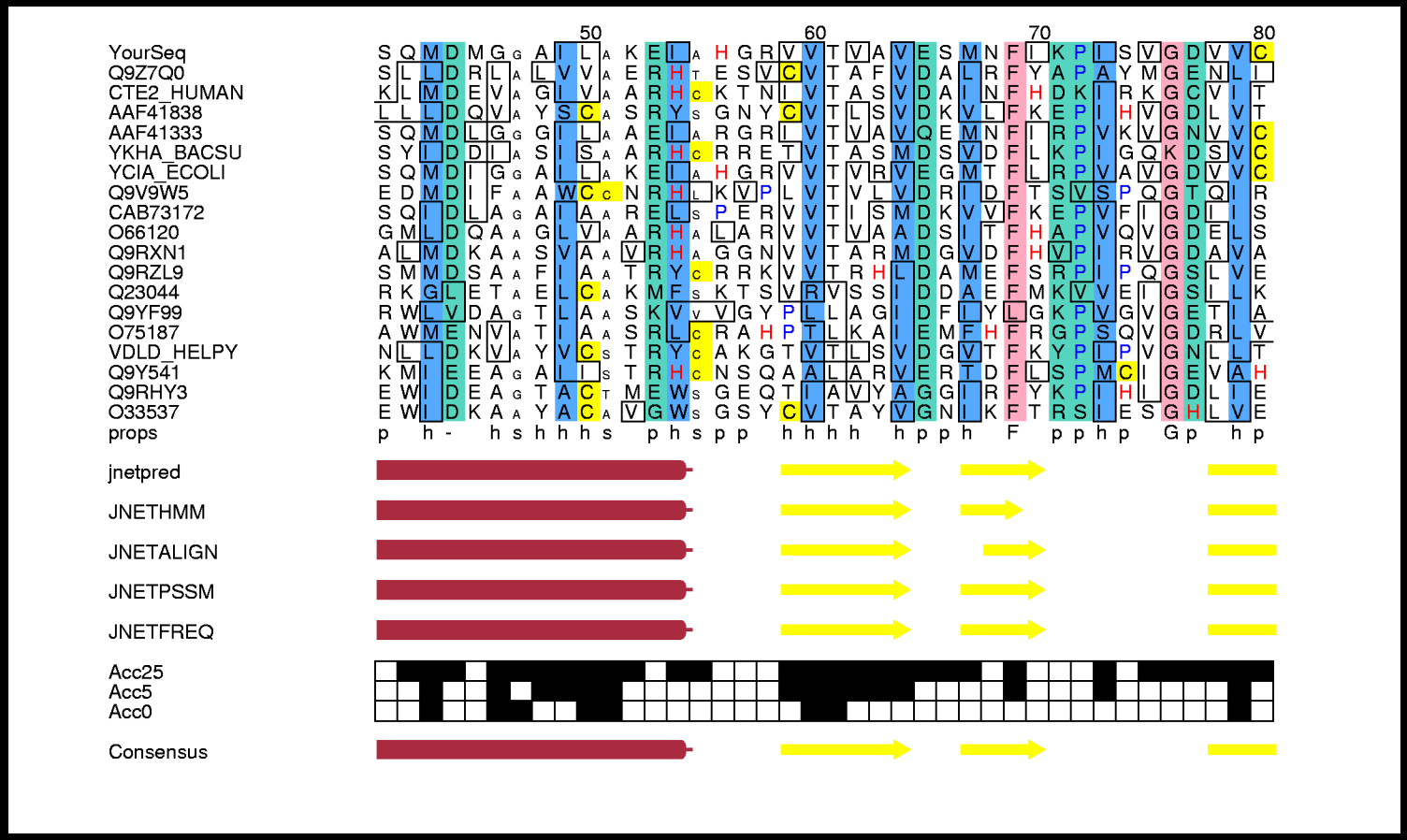
Protein Tools

Part 1

Jaroslaw Pillardy
Computational Biology Service Unit
Cornell Theory Center

Prediction of Secondary Structures



Two simple questions

What is secondary structure?

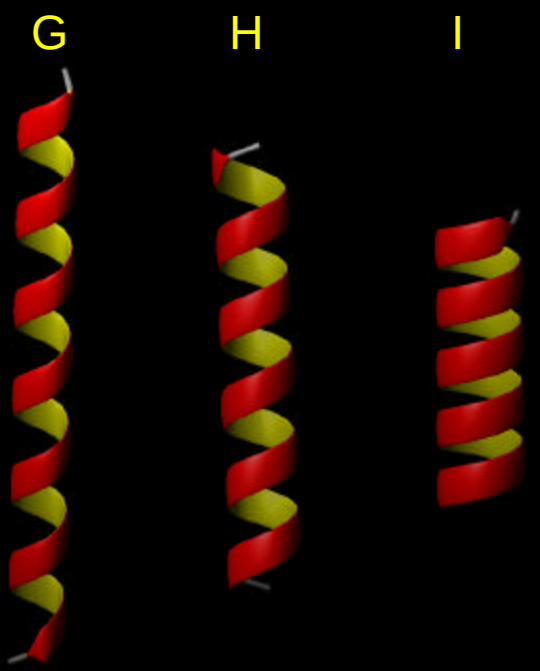
Protein folding unit which is brought about by linking the carbonyl (CO) and imide (NH) groups of the backbone together by means of hydrogen bonds.

What is secondary structure for?

Many proteins usually fold into globular forms in solution, the inside of a globule being highly hydrophobic. Packing backbone (that has highly polar [= hydrophilic] NH and CO groups) into such a hydrophobic environment would have been impossible without screening NH and CO groups by bonding them together with a set of H-bonds in secondary structure segments.

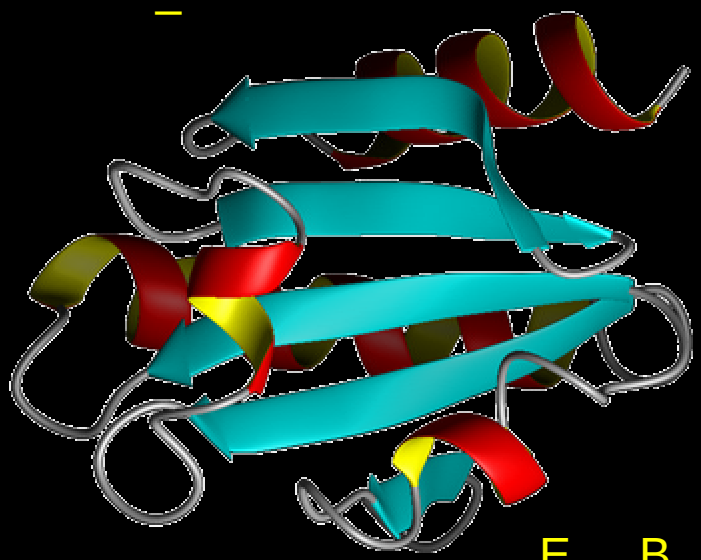
Dictionary of **S**econdary **S**tructure assignment of **P**roteins (DSSP)

H	– α helix	H
I	– π helix	H
G	– 3_{10} helix	H
E	– extended strand	E
B	– bridge	E
T	– turn	–
S	– bend	–
	– other	–



32% H
21% E
47% –

33% T



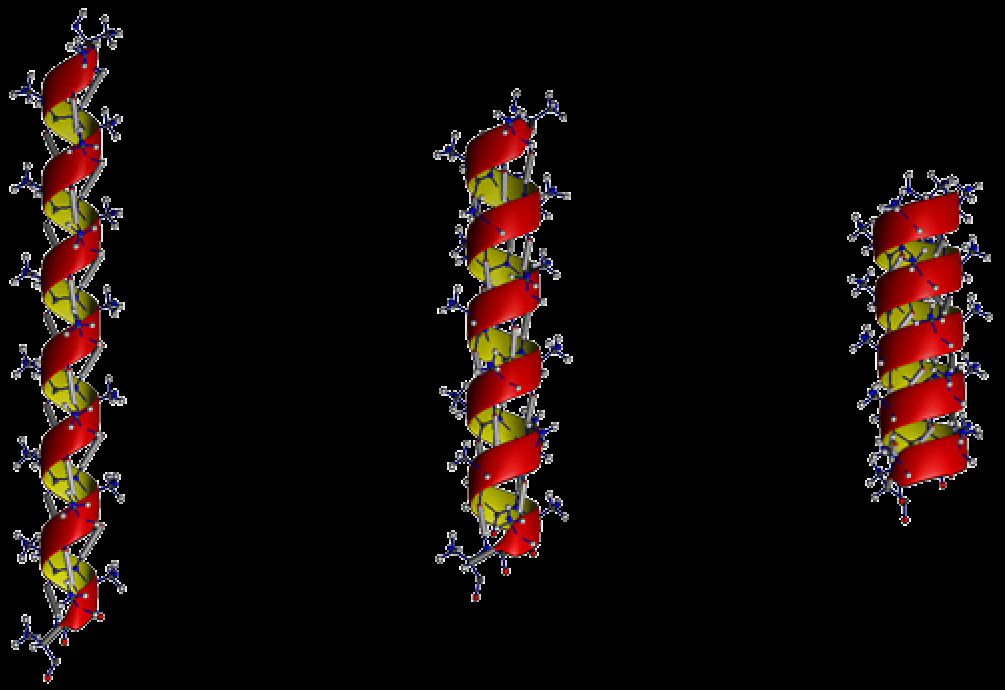
Helices

α helix (H)
3.6 residues per turn
13 atoms in H-bond ring

π helix (I)
4.4 residues per turn
16 atoms in H-bond ring

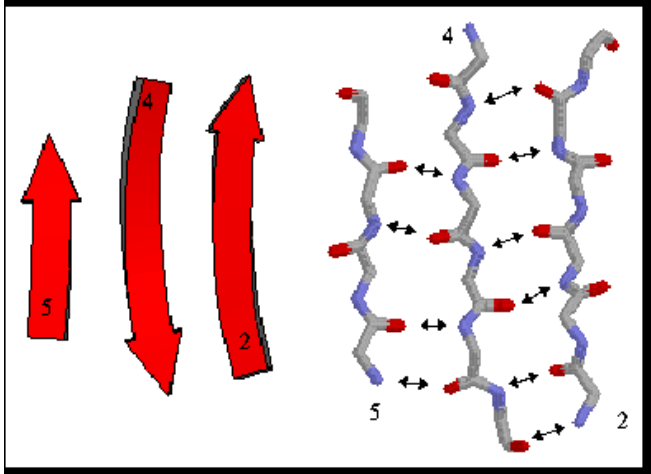
3_{10} helix (G)
3.0 residues per turn
10 atoms in H-bond ring

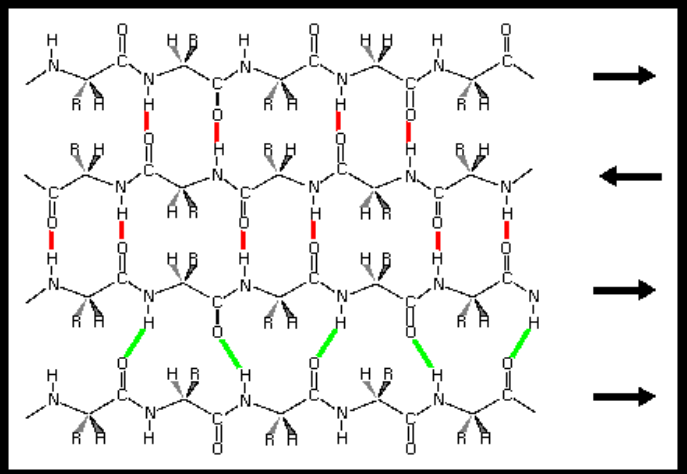
G H I



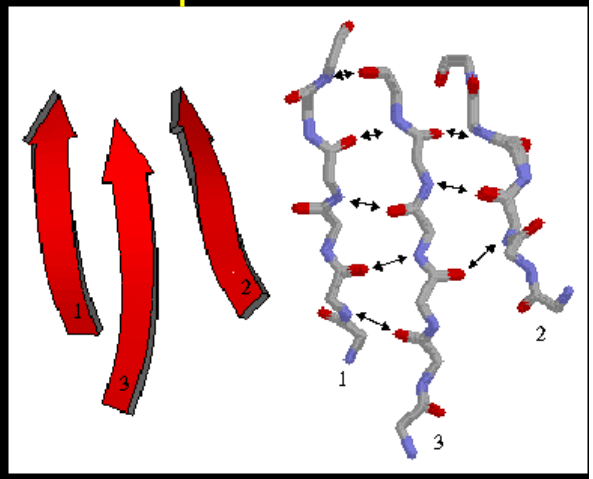
Strands (sheets)

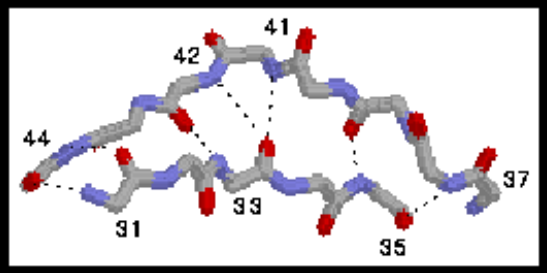
antiparallel





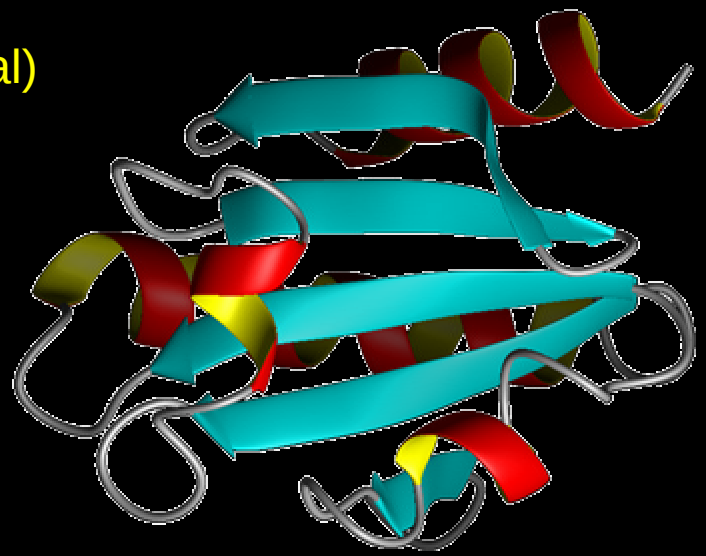
parallel





Connections:

- hairpin
- crossover (helical)

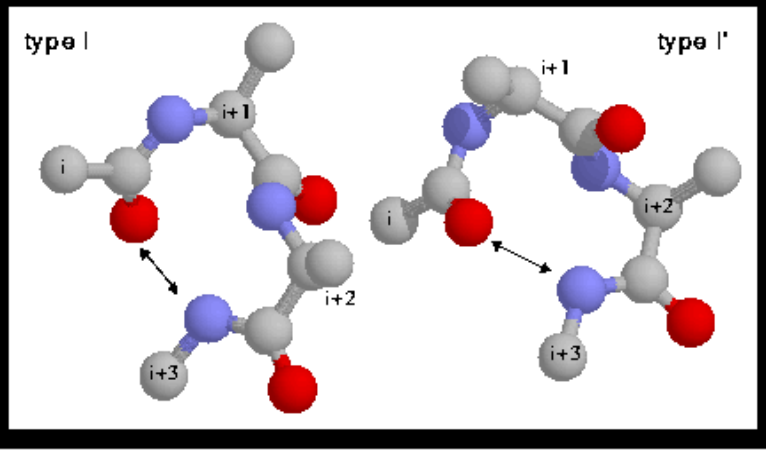


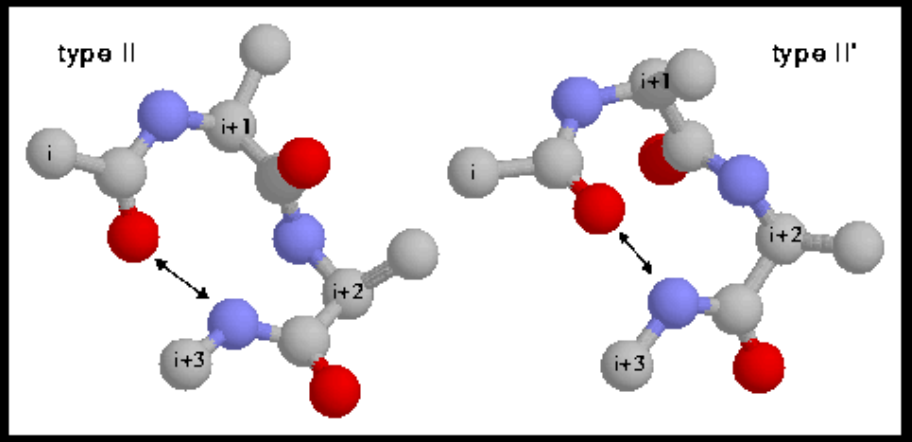
Most sheets observed in globular proteins are twisted (0 to 30 degrees per residue). This twist is always of the same handedness.

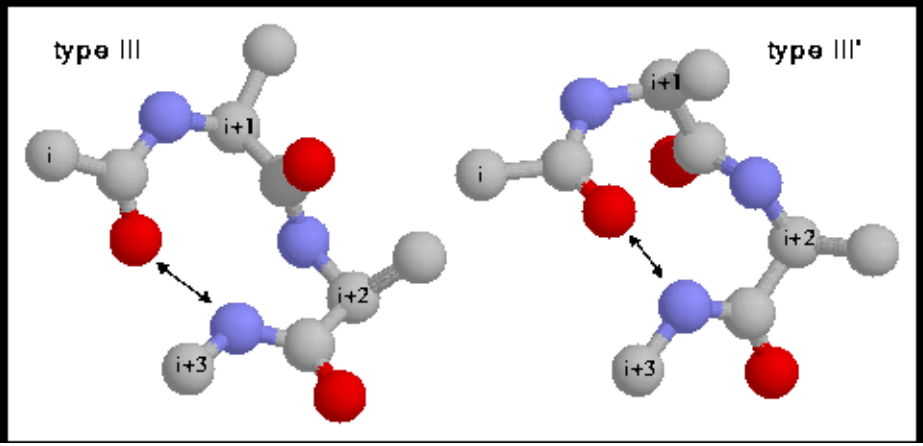
- angle of strand crossings: left-handed
- progressive twist of the hydrogen-bonding direction: right-handed.

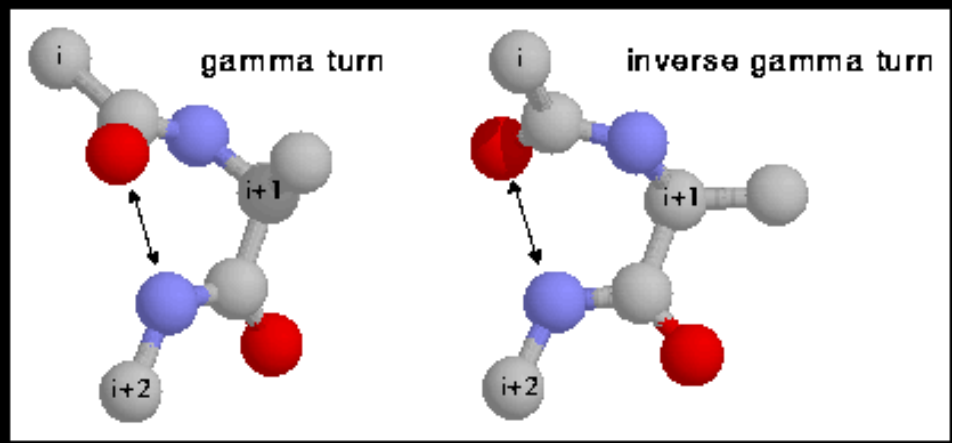
Turns

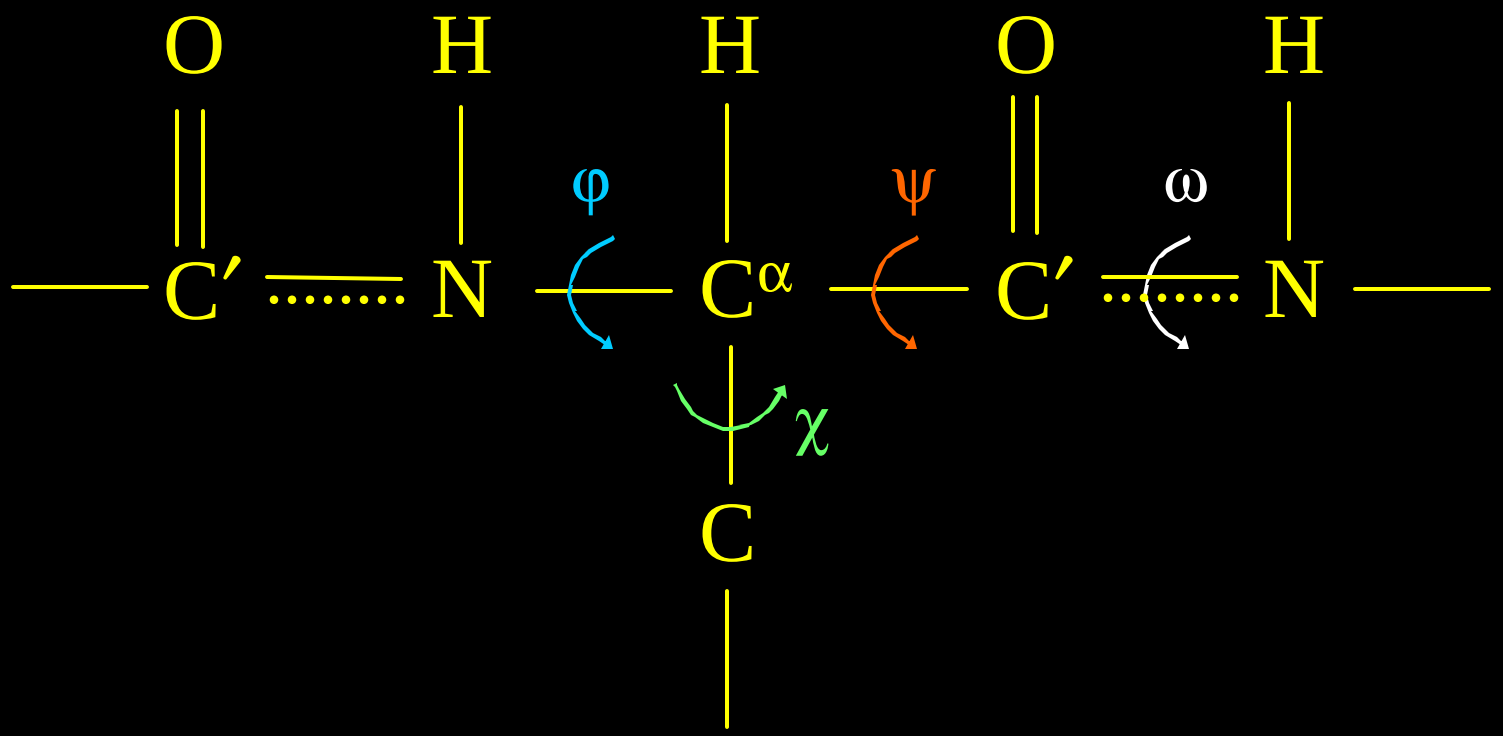
Turns are the third of the three "classical" secondary structures with approximately one-third of all residues in globular proteins are contained in turns that serve to reverse the direction of the polypeptide chain. Turns are located primarily on the protein surface and accordingly contain polar and charged residues. Antibody recognition, phosphorylation, glycosylation, hydroxylation, and intron/exon splicing are found frequently at or adjacent to turns.





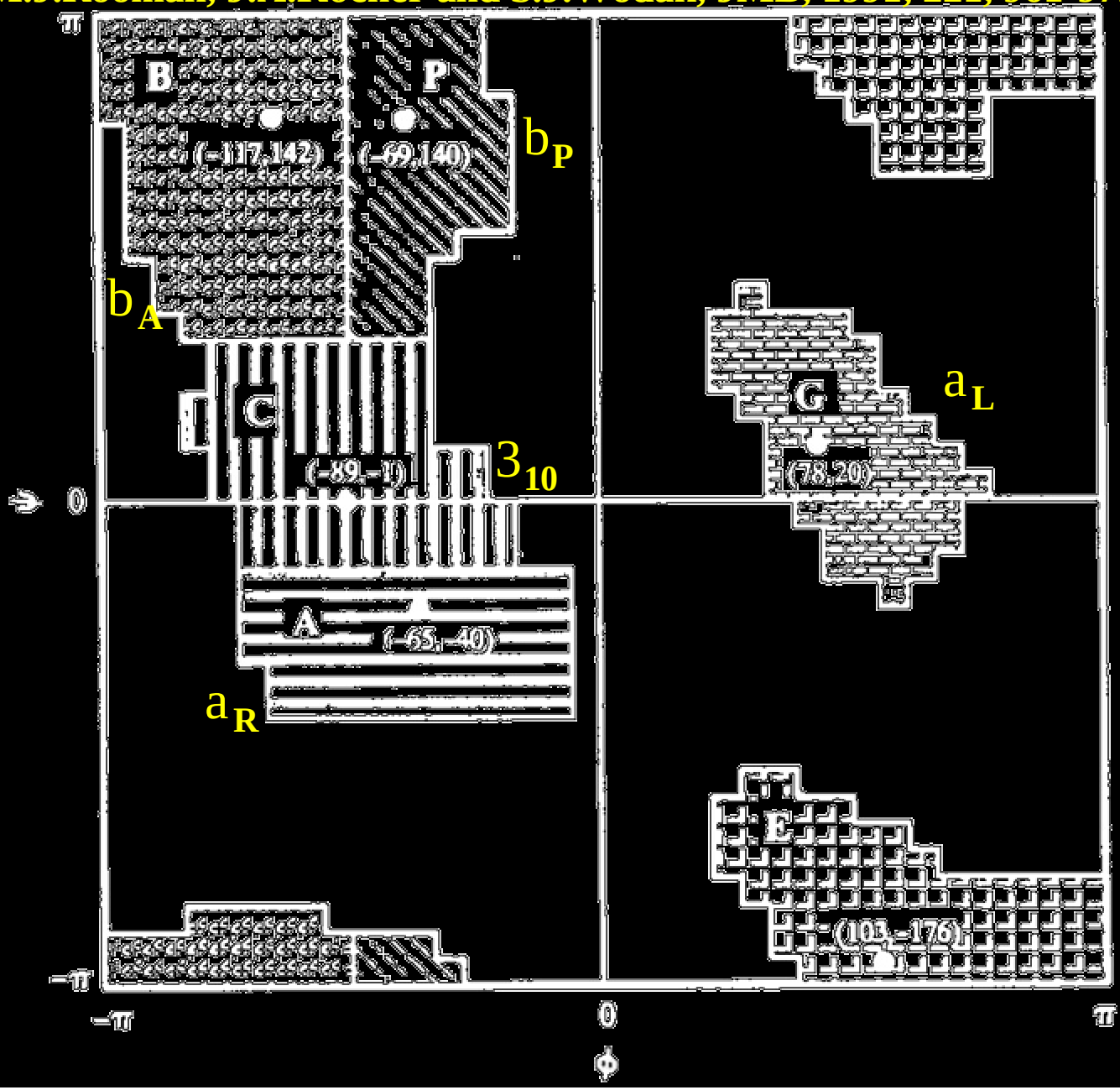





O
H
H
O
H
φ
ψ
ω
C′
N
Cα
C′
N
χ
C

Classifying backbone conformations into 8 types

(M.J.Rooman, J.A.Kocher and S.J.Wodak, JMB, 1991, 221, 961-979)



Secondary structure prediction: what is it useful for?

- Verifying model or alignment
- Finding domains
- Looking for possible function assignments
- Improving alignments

Secondary structure prediction: prediction accuracy

Per-residue prediction accuracy:

$$Q_3 = 100(c_H + c_E + c_-)/N$$

Normalized per-residue prediction accuracy:

$$norm\ Q_3 = 100(Q_3 - Q_3^{RAN})/(Q_3^{HM} - Q_3^{RAN})$$

Secondary structure prediction: first generation methods

The oldest group of methods: *C+F, Lim, GORI*.

Based on single-residue statistics: tried to correlate amino acid composition of fragments with secondary structure.

Normalized three-state-per-residue accuracy is 25% - 40%

Secondary structure prediction: second generation methods

Examples: *Schneider*, *ALB, GORIII, COMBINE, S83* .

Based on segment statistics: typically 11-21 adjacent residues are taken into account in order to predict the secondary structure state of a residue. Physicochemical properties of residues are often considered. Many advanced recognition and machine learning algorithms are used (neural networks, graph theory, multivariate statistics, expert rules and more).

Normalized three-state-per-residue accuracy is 40% - 60%, β-strands predicted with (non-normalized) accuracy 28%-48%, predicted segments often too short.

Secondary structure prediction: third generation methods

Examples: *NSSP, LPAG, PHD*.

Multiple sequence alignment is used and the secondary structure prediction with methods used in the second generation methods is carried out.

Normalized three-state-per-residue accuracy is 70%, β-strands are predicted with the same accuracy as α-helices, and segment lengths are usually predicted correctly.

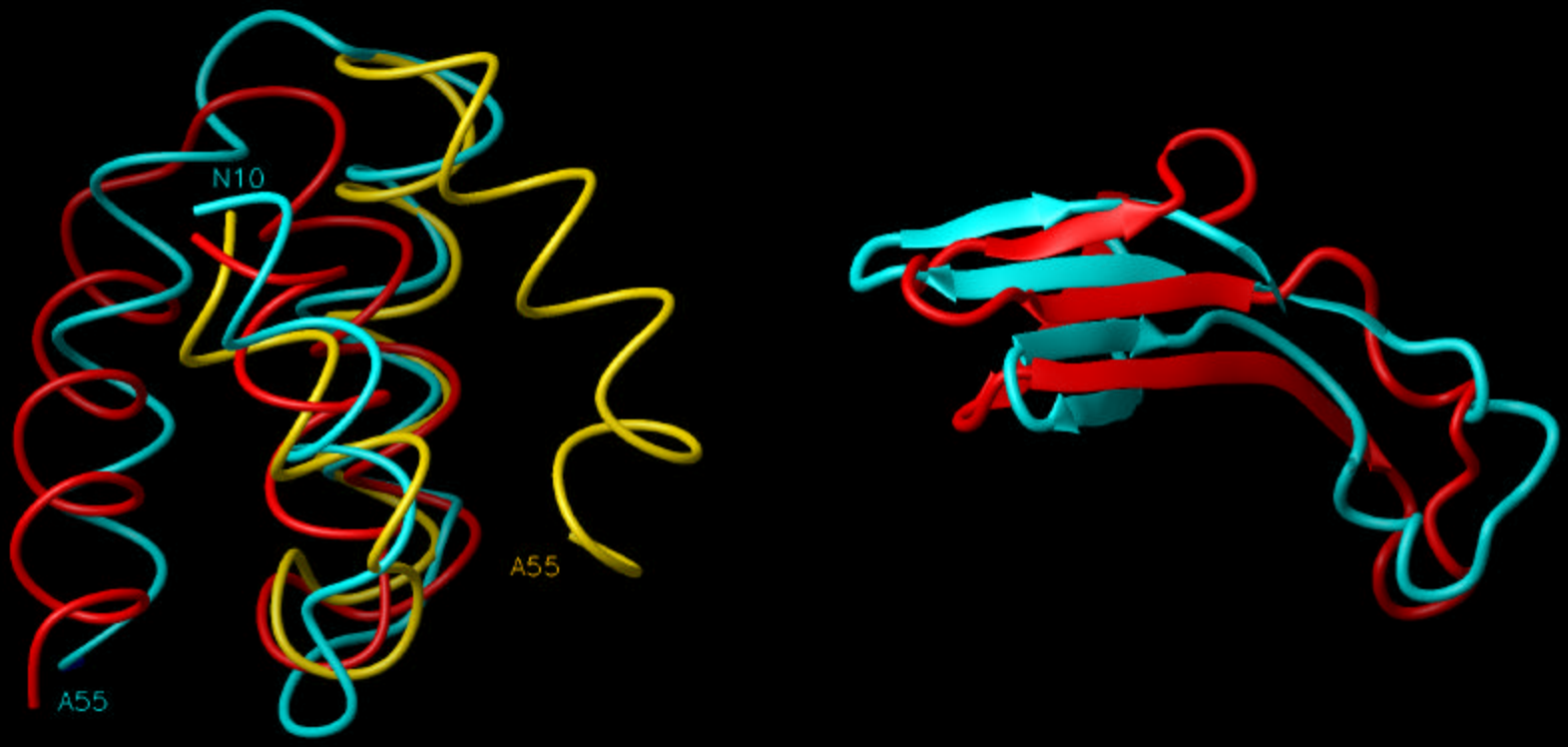
Secondary structure prediction: third-and-half generation methods

Examples: *PSI-PRED, JNET*.

Multiple sequence alignment with PSI-BLAST (until convergence) is used and the secondary structure prediction with methods used in the second generation methods is carried out.

Three-state-per-residue accuracy is about 75% (at least 2%-5% better than 3rd generation), β-strands are predicted with the same accuracy as α-helices, and segment lengths are usually predicted correctly.

Protein Structure Comparison



Identical proteins, different conformations: RMSD

RMSD may be calculated over many different features, the most popular being positions of atoms and torsional angles. Calculating RMSD is easy when alignment is known.

Position-RMSD:

- The most important are C^{α} atoms from protein backbone since they define the fold (C^{α}RMSD).
- Usually quite good measure if structures are close, but it makes little sense when they are very distant.
  Example: 1fsd. Example: 1igd.
- Easily fooled by misplacing (similar) domains.
  Example: 1fcc .
- Size-dependent.
- Should be calculated in two ways simultaneously: global (whole sequence) and local (the biggest segment below certain RMSD threshold).
  Example: 1e68.

Identical proteins, different conformations: RMSD

RMSD may be calculated over many different features, the most popular being positions of atoms and torsional angles. Calculating RMSD is easy when alignment is known.

Angle-RMSD:

- Gives very good indication about secondary structure agreement.
- Very unintuitive measure: change of one angle contributes little to angle-RMSD but structures are quite different.
- Very resistant to misplacing (similar) domains (see above).
- Size-dependent.

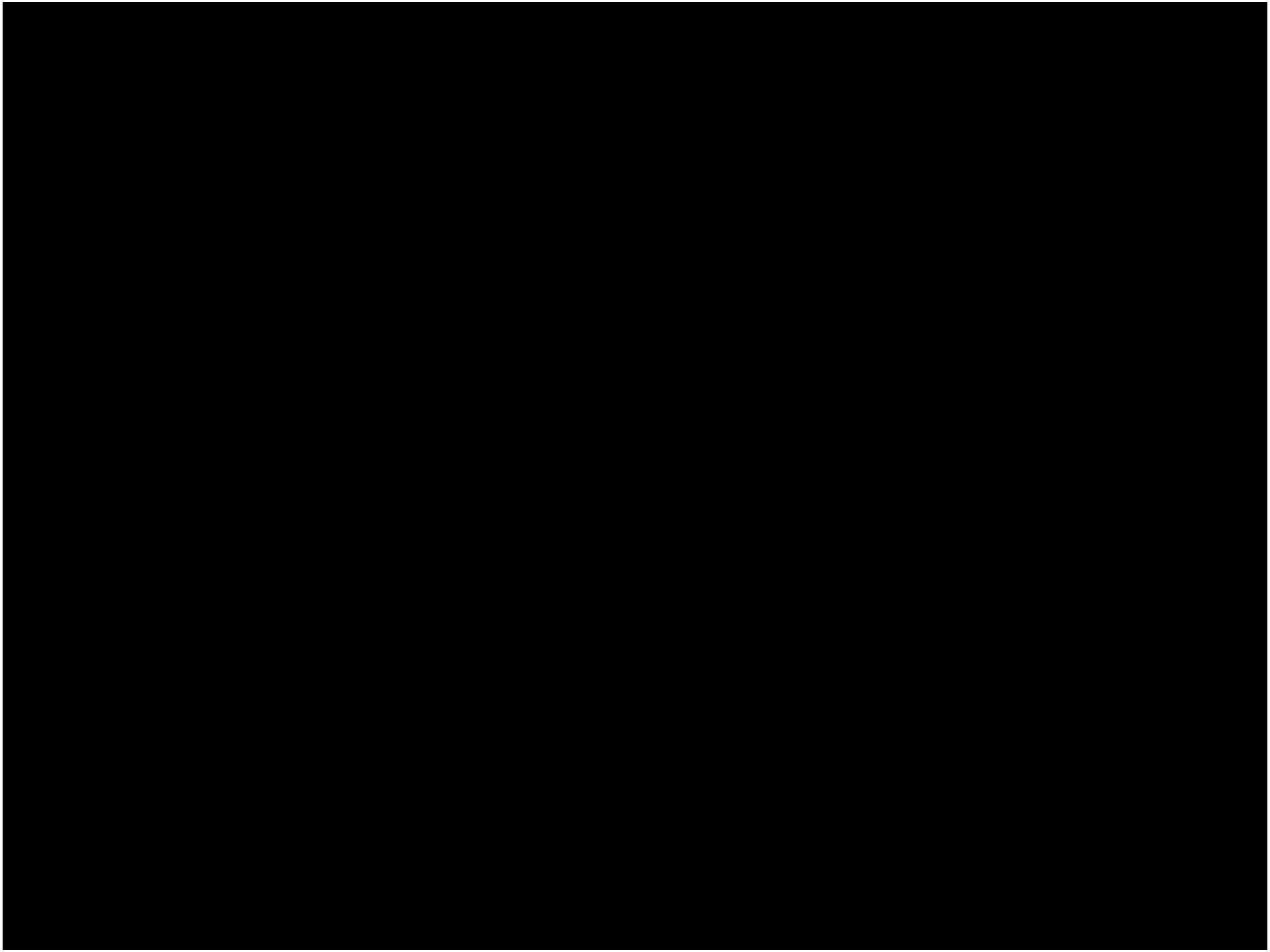
Different proteins: RMSD and alignment

RMSD is calculated usually as position-RMSD, the really difficult part is finding optimal alignment.

The problem: find optimal sequence alignment, such that the structural difference (RMSD) is as small as possible. It is similar to classical sequence-sequence alignment, but scoring function depends on superposition of 3D structures associated with sequences.

Different algorithms for alignment search are used:

- Monte Carlo (DALI in *FSSP*)
- Combinatorial extension (CE)
- Double dynamic programming (SAP in *CATH*)



CASP4 Target T0102, RMSD=4.2Å,
native=blue, calculated=red
AS48, Bacteriocin AS-48, E. faecalis,
70 amino acids, cyclical,
PDB code 1E68