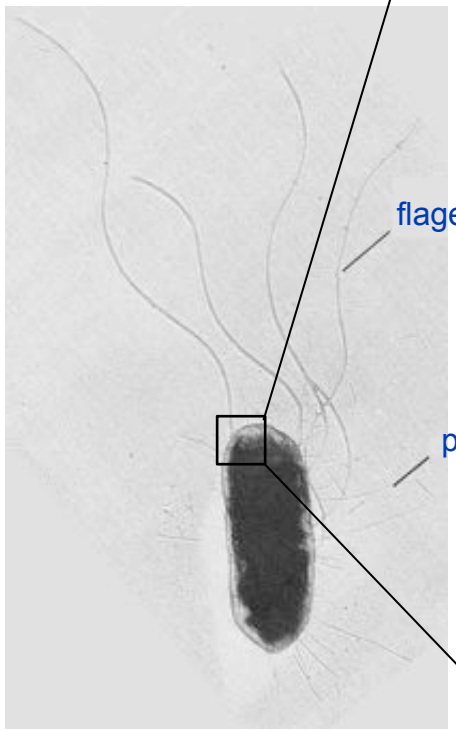
Proteins:
an introduction

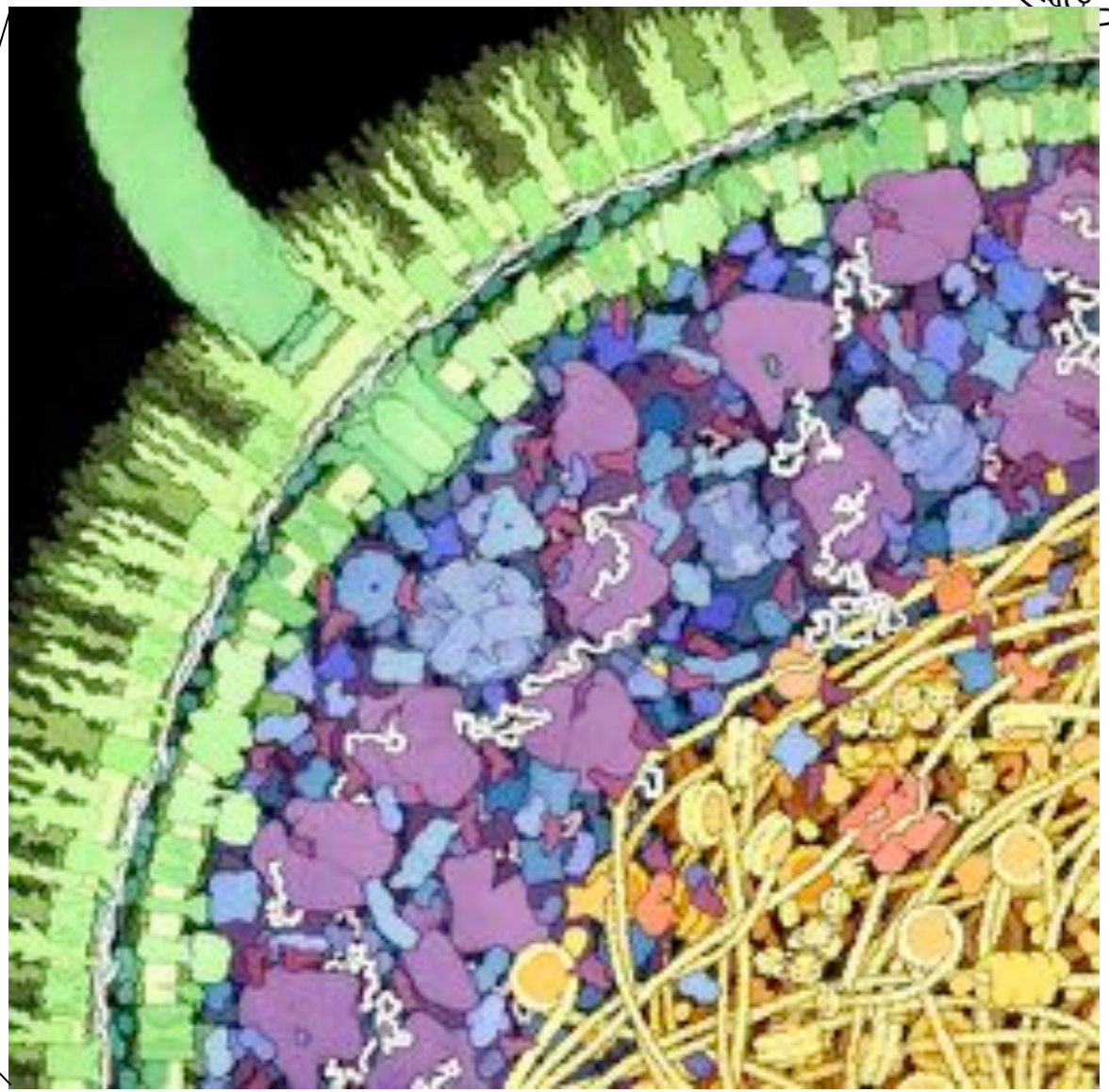
Life is the result of the chemical activity of proteins

(with nucleic acids essentially confined to encoding and producing them)

A cross-section through *Escherichia coli* shows the diversity and density of macromolecules forming a cell



Howard Berg



David Goodsell

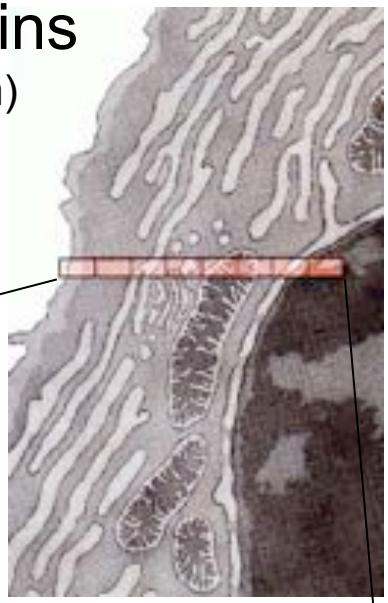
Life is the result of the chemical activity of proteins

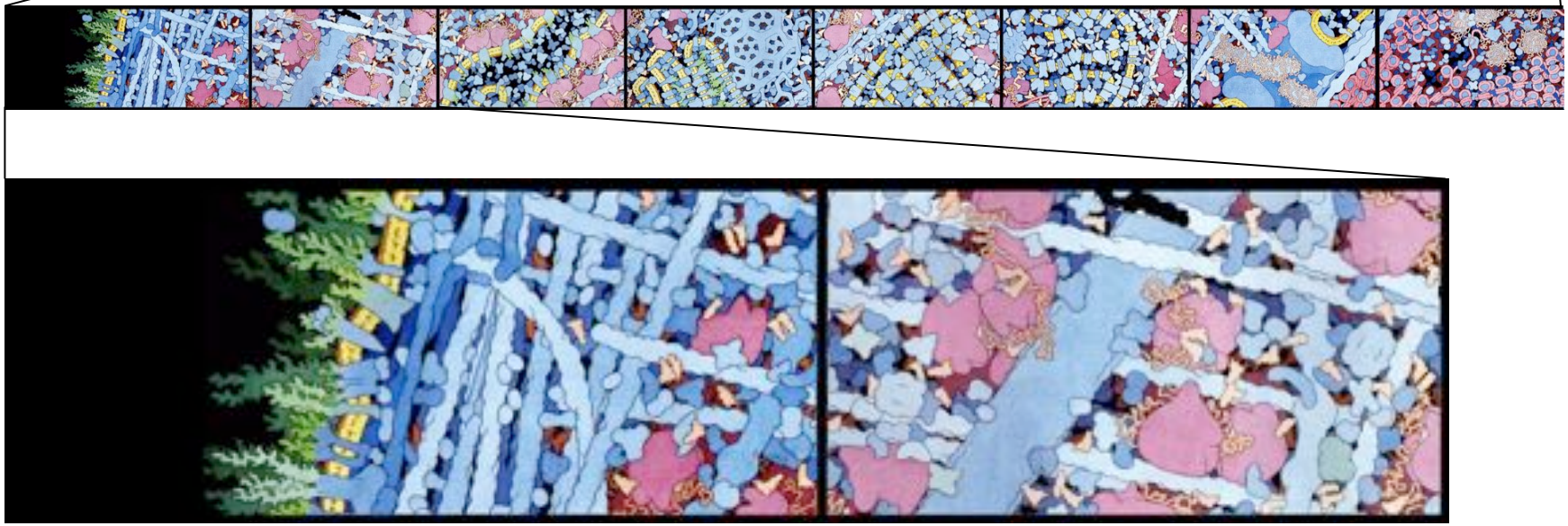
(with nucleic acids essentially confined to encoding and producing them)

A cross-section through a human cell, from the plasma membrane to the nucleus

The enlargement shows cell-surface gycoproteins, the sub-membraneous cytoskeleton (primarily actin filaments), a microtubulus, ribosomes (in pink; they are everywhere except the nucleus)

relative size of *E. coli*





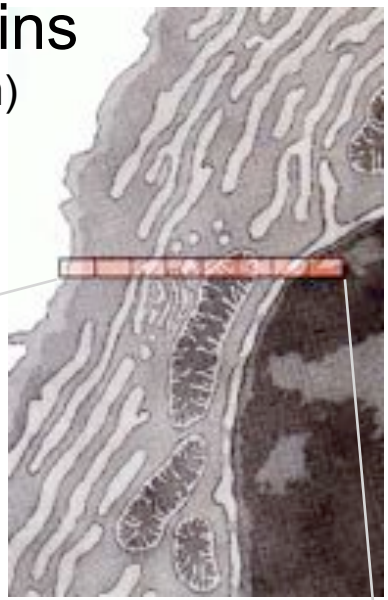
David Goodsell

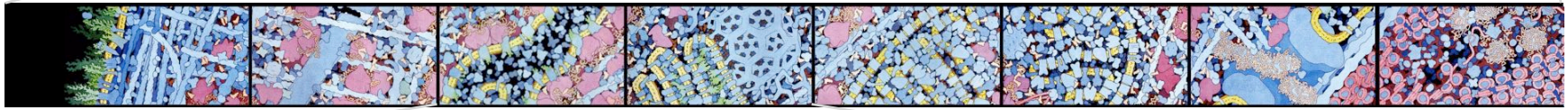
Life is the result of the chemical activity of proteins

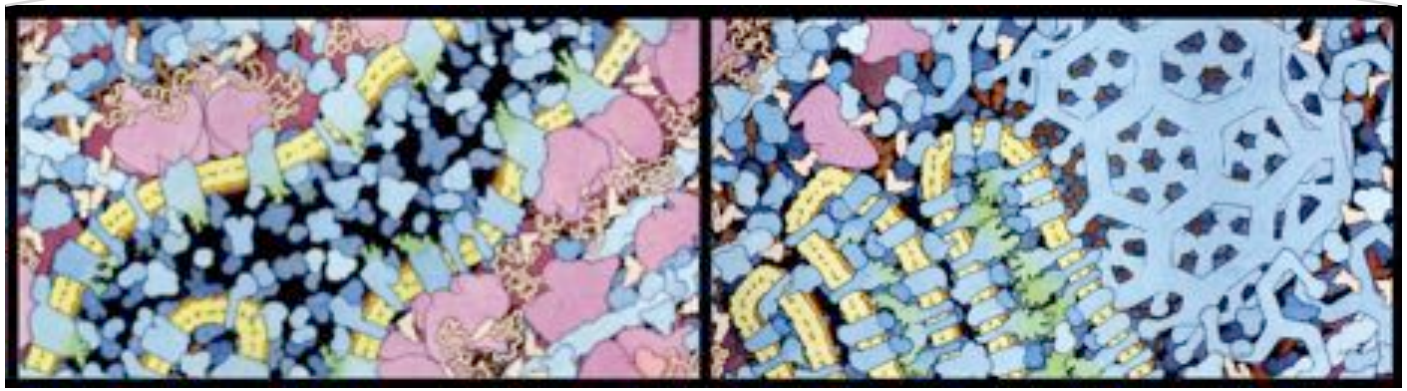
(with nucleic acids essentially confined to encoding and producing them)

A cross-section through a human cell, from the plasma membrane to the nucleus

The enlargement shows the endoplasmatic reticulum (‚rough ER'), Golgi stacks, and a clathrin cage assembling for vesicle budding







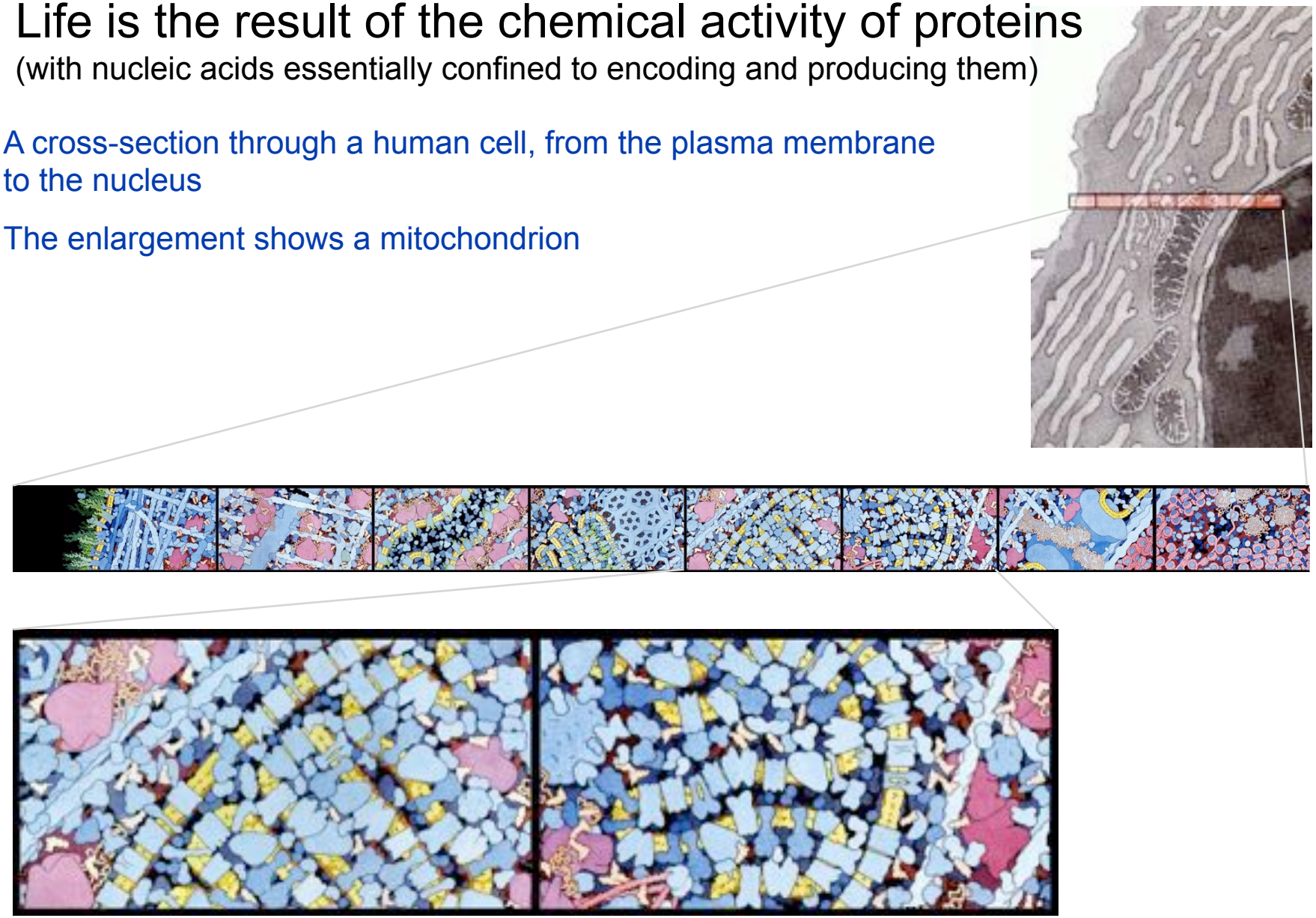
David Goodsell

Life is the result of the chemical activity of proteins

(with nucleic acids essentially confined to encoding and producing them)

A cross-section through a human cell, from the plasma membrane to the nucleus

The enlargement shows a mitochondrion



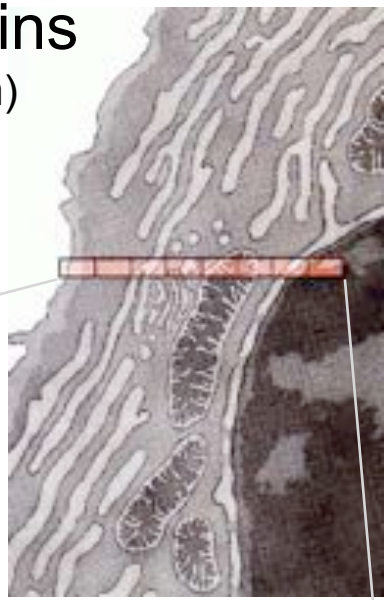
David Goodsell

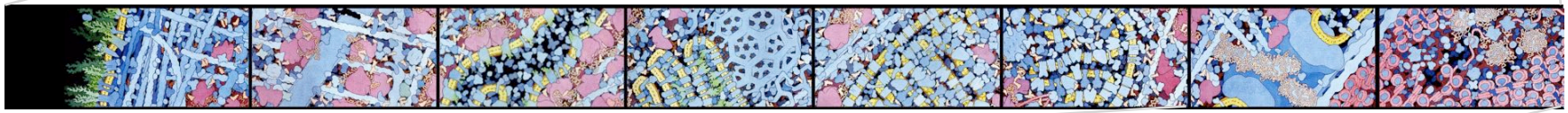
Life is the result of the chemical activity of proteins

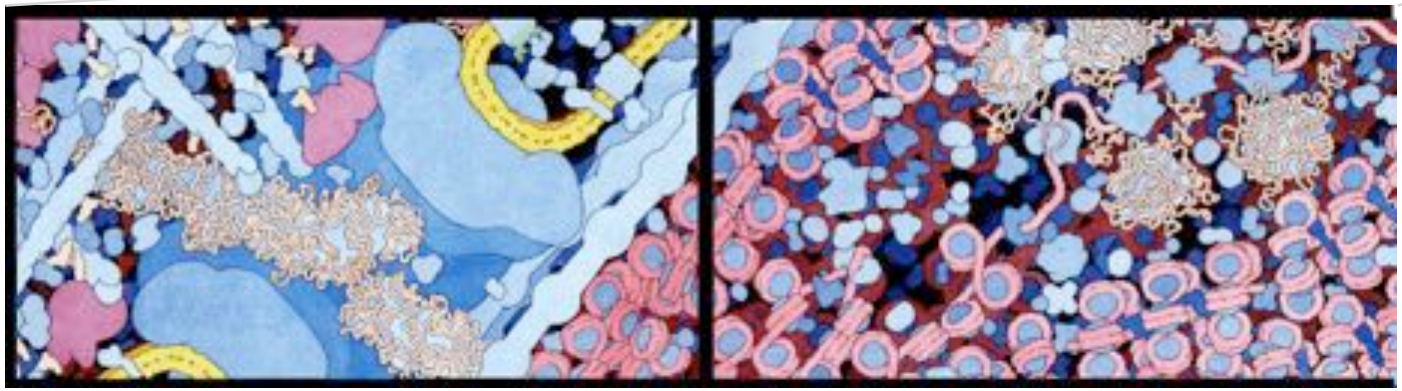
(with nucleic acids essentially confined to encoding and producing them)

A cross-section through a human cell, from the plasma membrane to the nucleus

The enlargement shows a nuclear pore, the pernuclear and intranuclear cytoskeleton, chromatin, spliceosomes and other RNP particles

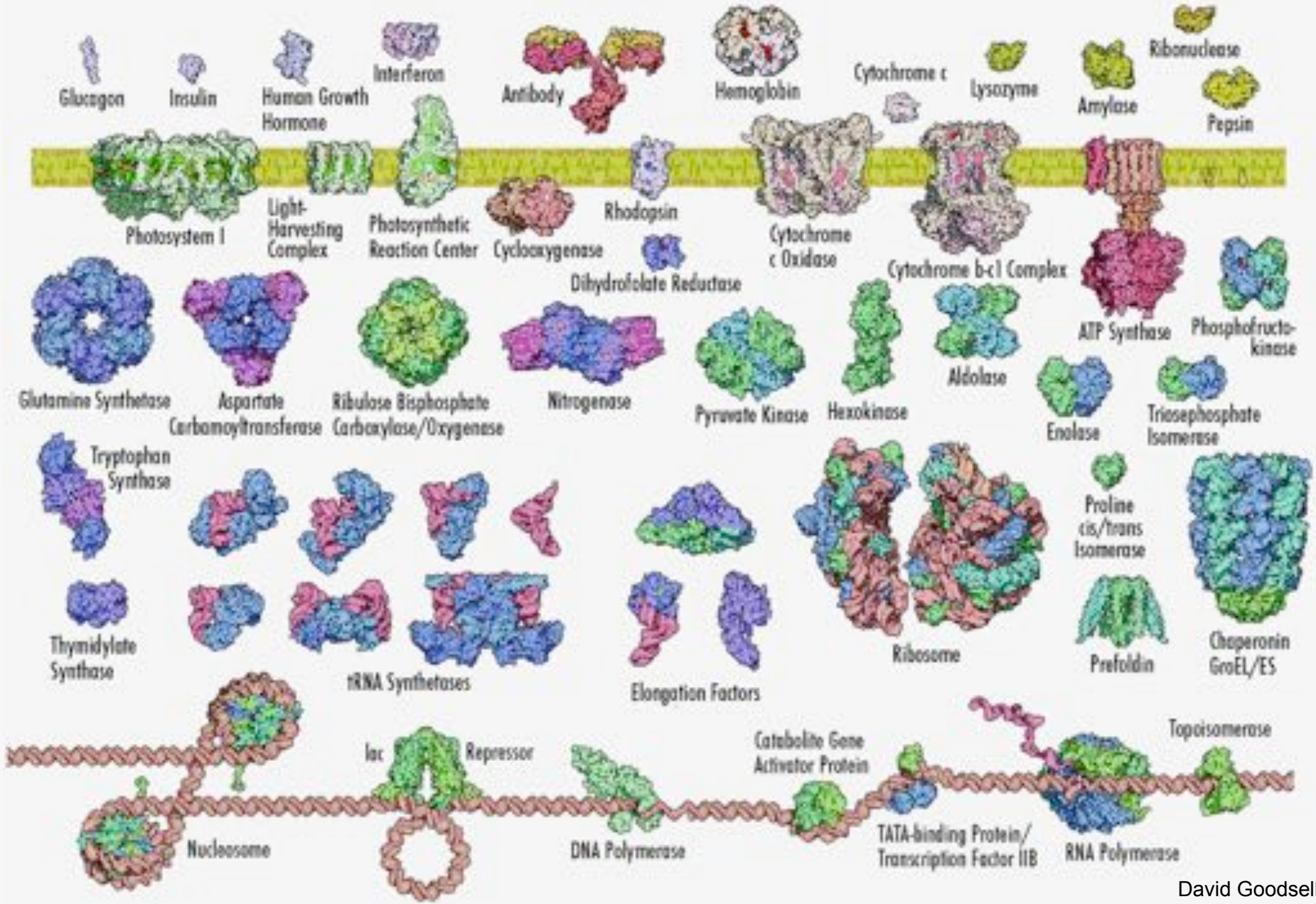






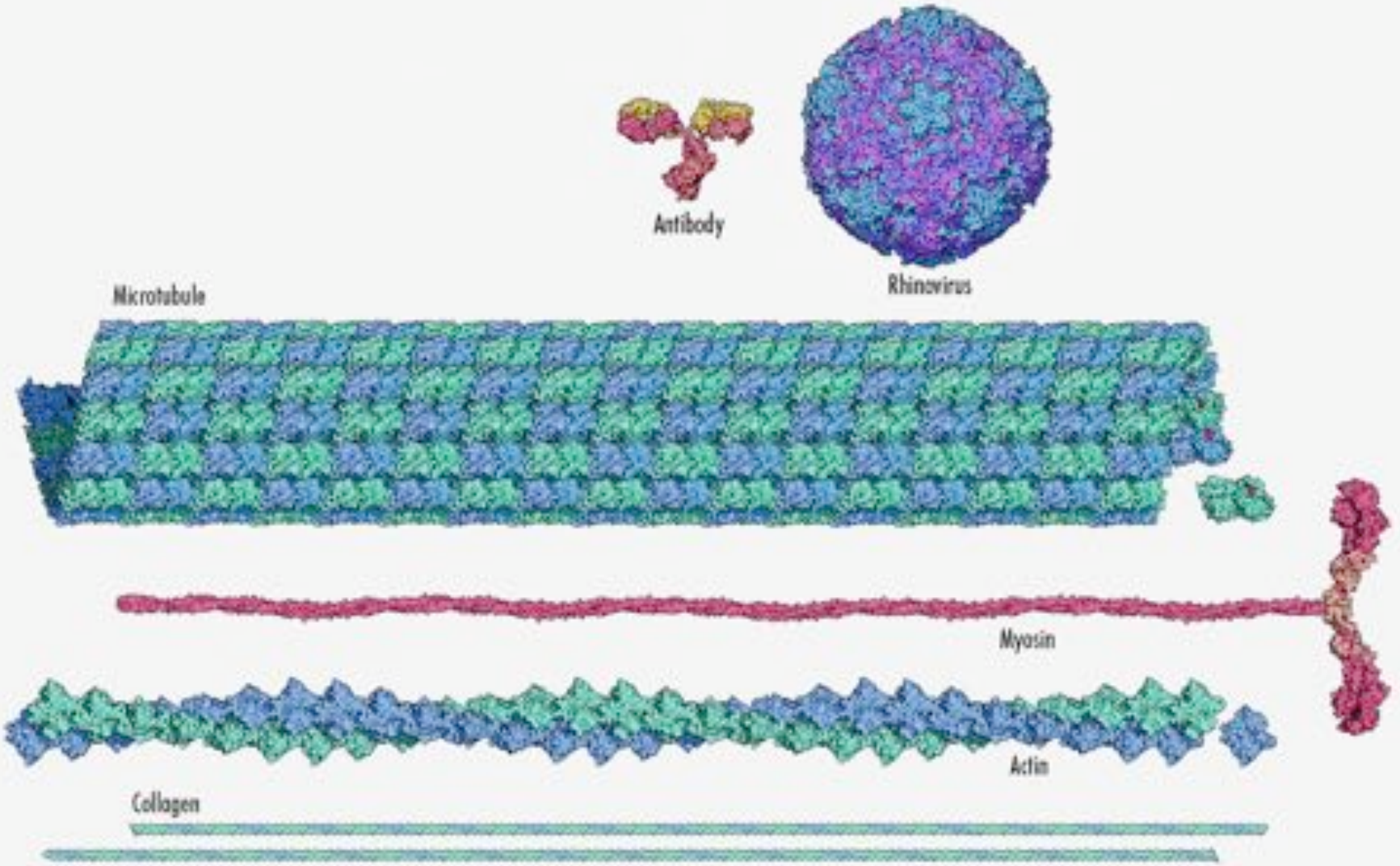
David Goodsell

Proteins are extremely diverse



David Goodsell

Proteins are extremely diverse



David Goodsell

What is a protein?

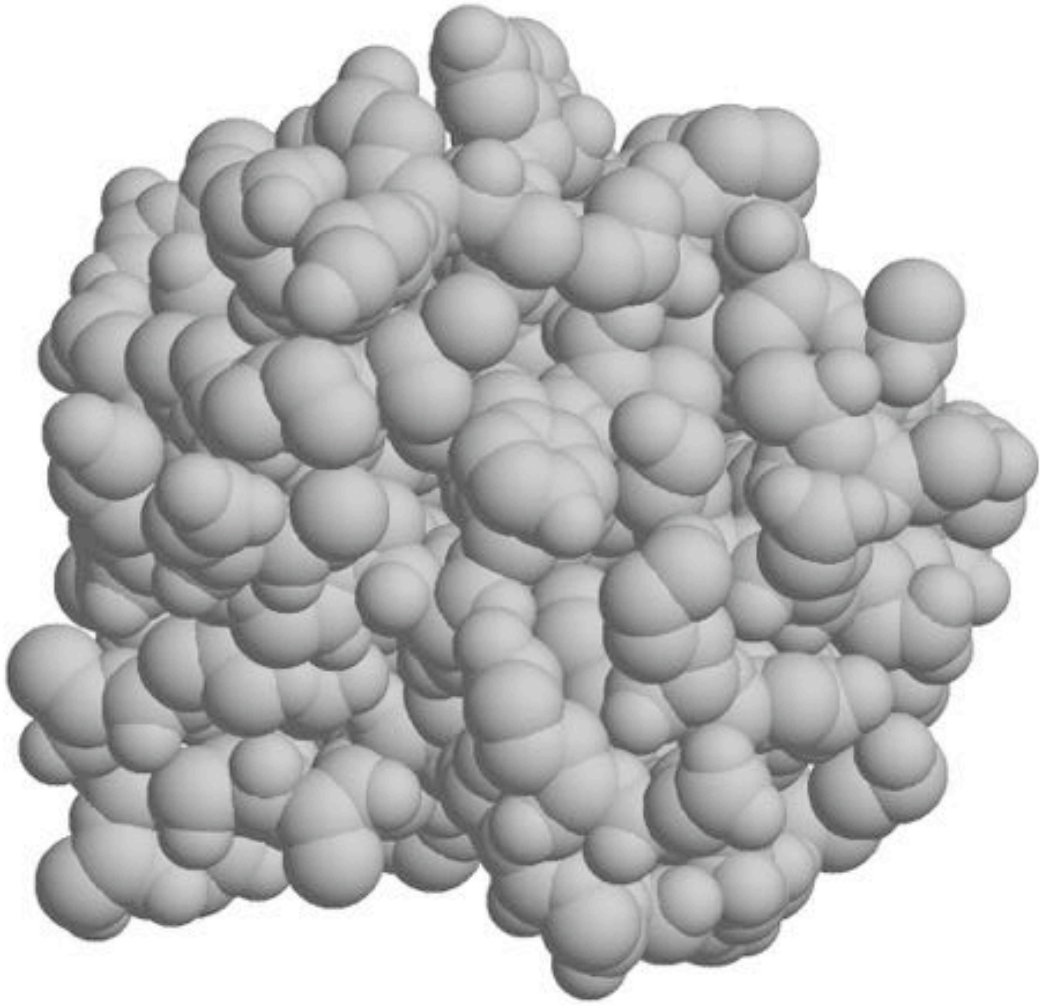
• *protein* - 1844, from Fr. protéine, coined 1830s by Dutch chemist Mulder from Gk. proteios "the first quality," from protos "first." Originally a theoretical substance thought to be essential to life, the modern use is from Ger. Protein, borrowed in Eng. 1907. (Etymology Online)

• Unbranched polymer chain of α-L-amino acids connected by peptide bonds (condensation between the carboxyl- and amino-groups of consecutive monomers)

• "The term protein is usually reserved for those chains with a specific sequence, length, and **folded conformation**." (Creighton, Proteins, W.H.Freeman and Co. 1984)

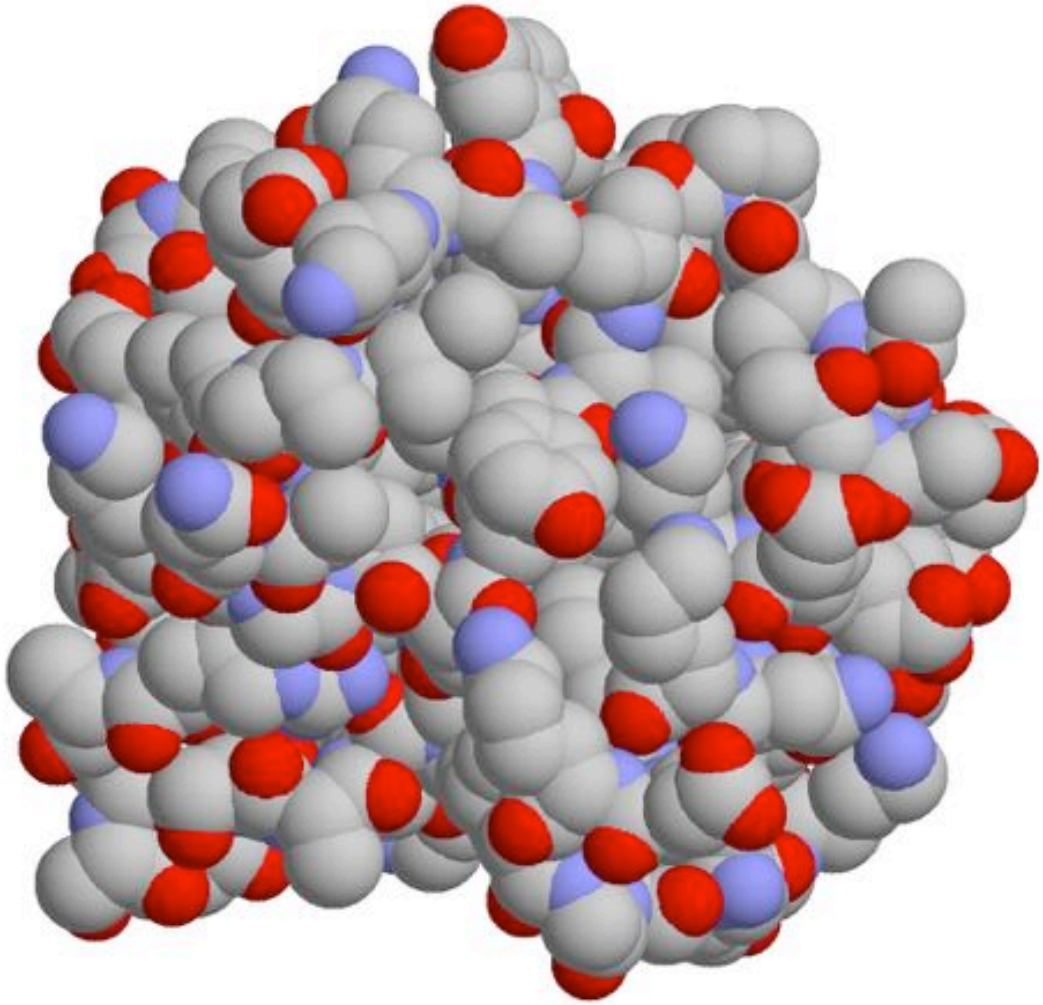
Schematic representations of protein structure

- spacefilling representation



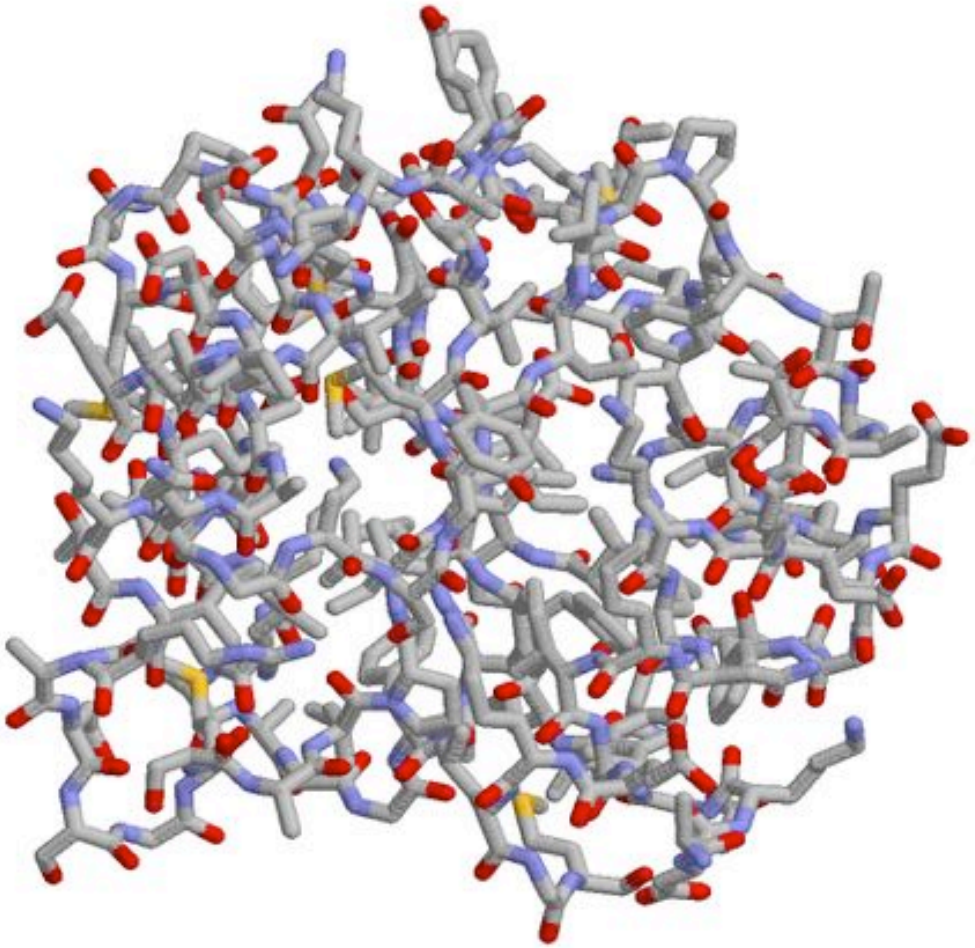
Schematic representations of protein structure

- spacefilling representation with CPK (Corey-Pauling-Koltun) colors
(C - grey, N - blue, O - red, S - yellow, P - purple, halogens - green)



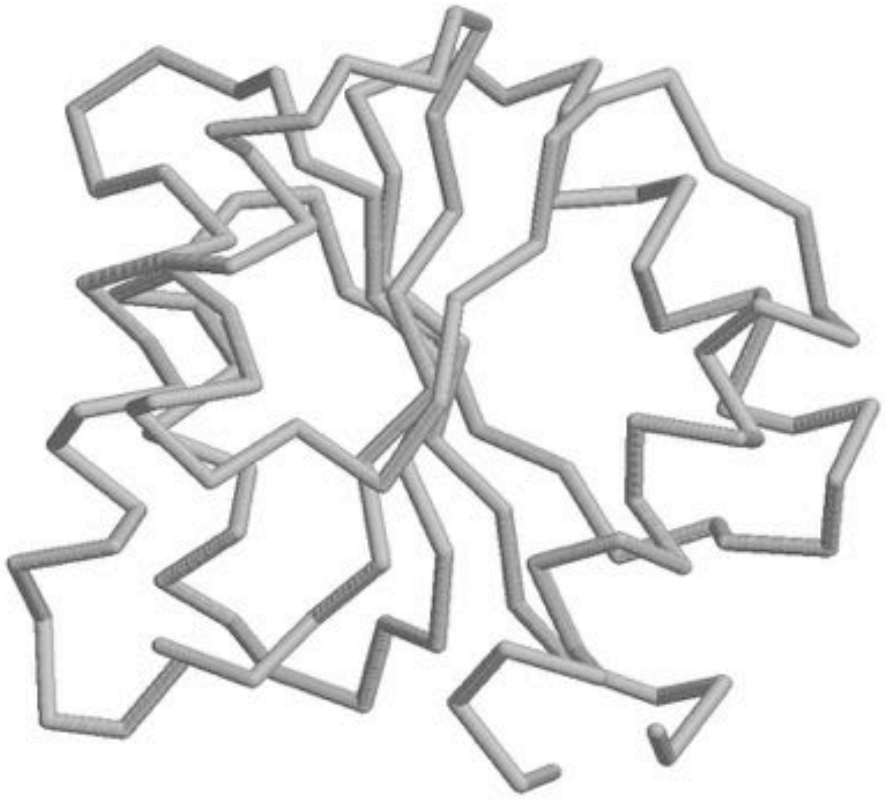
Schematic representations of protein structure

- stick representation with CPK colors



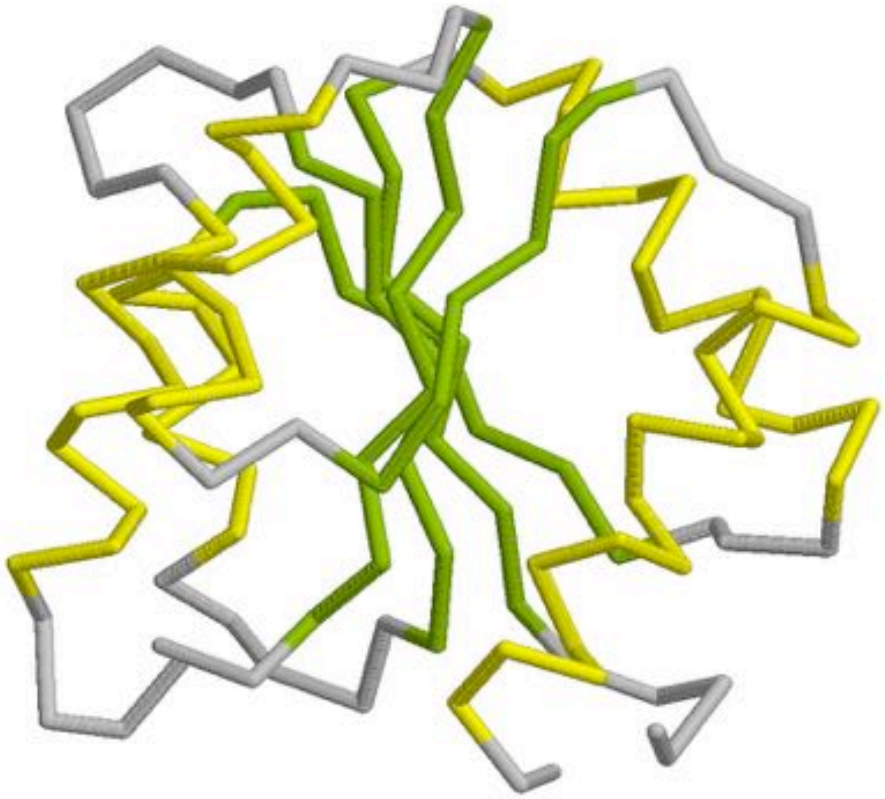
Schematic representations of protein structure

- Cα trace



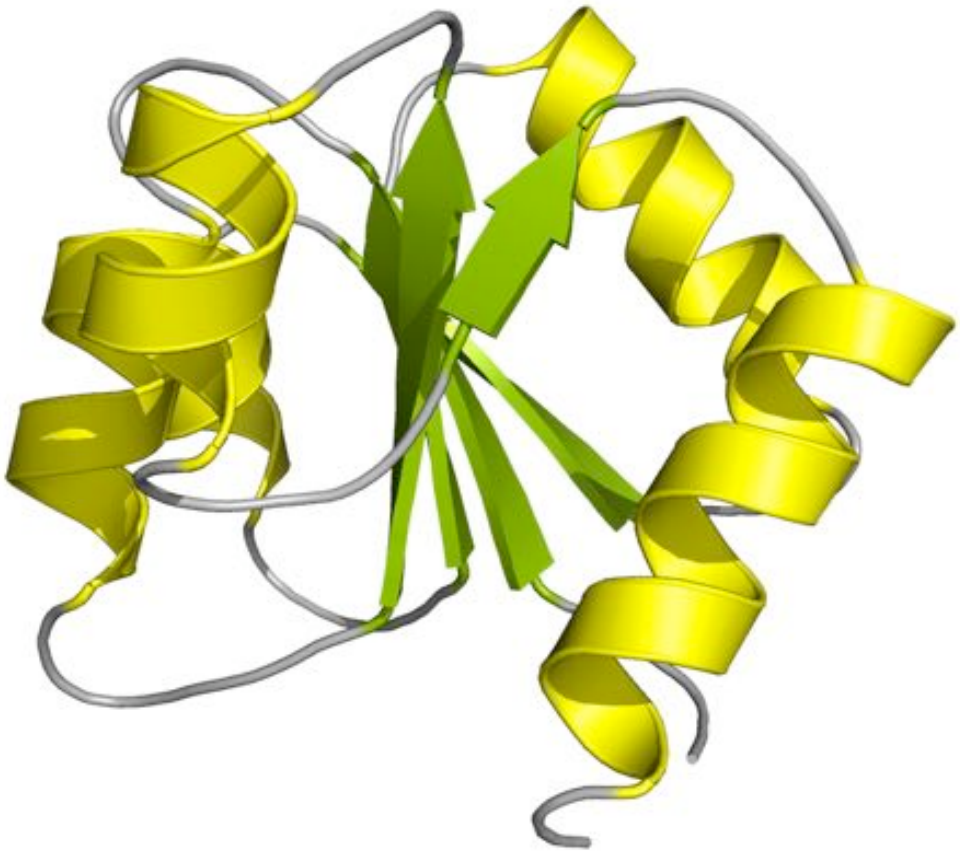
Schematic representations of protein structure

- Cα trace colored by secondary structure (strand - green, helix - yellow)



Schematic representations of protein structure

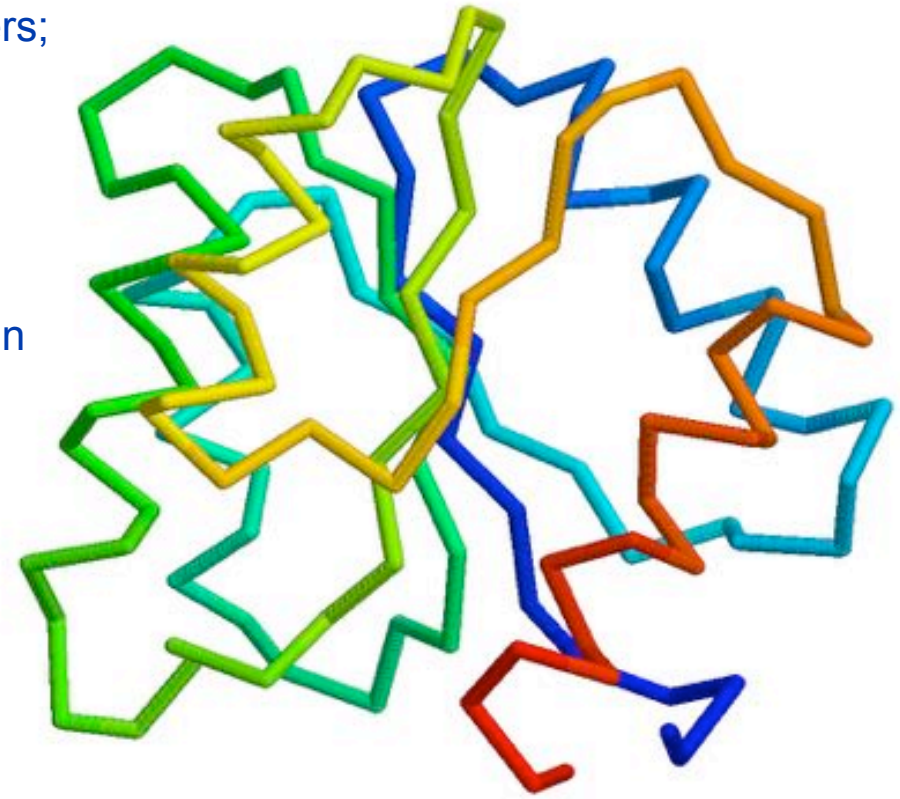
- cartoon representation colored by secondary structure



The path of a polypeptide chain

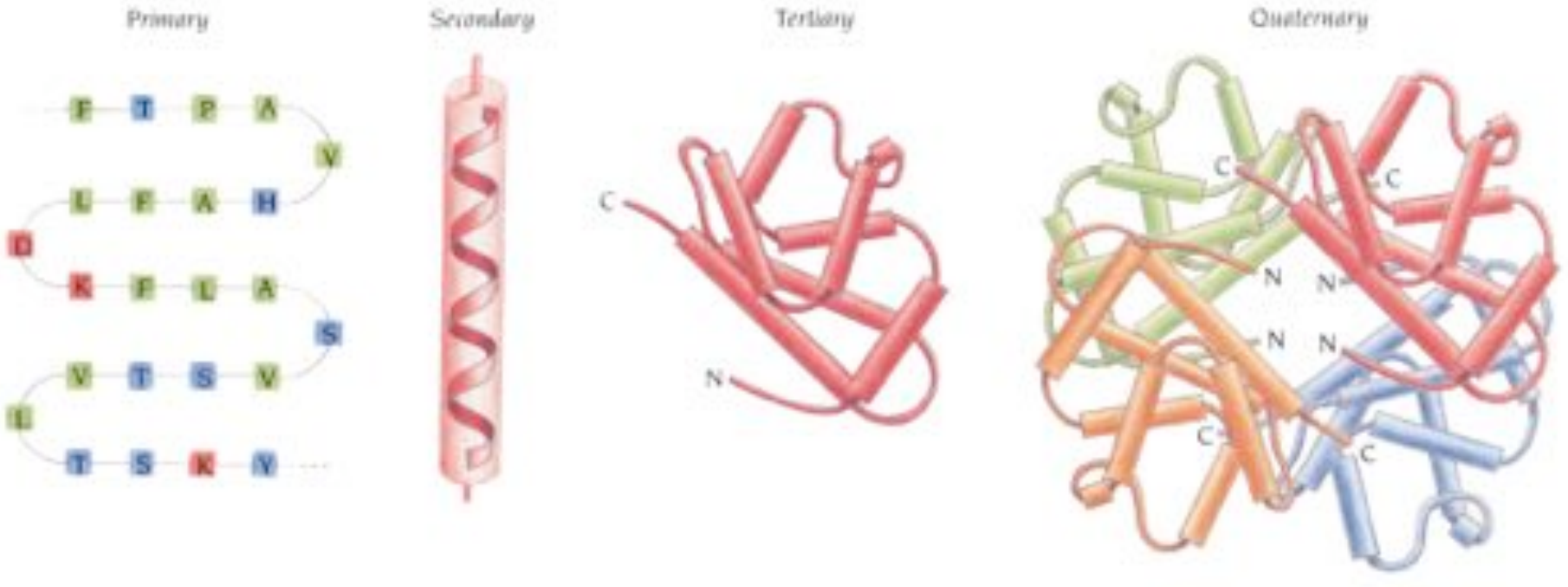
- Cα trace in chainbow coloring
(from blue at the N-terminus to red at the C-terminus)

Despite the convoluted way in which their backbone folds, proteins are unbranched polymers; there are however instances of branching through covalent sidechain-sidechain (e.g. disulfide bonds) and sidechain-mainchain (e.g. ubiquitination) bonds introduced post-translationally



Levels of protein structure

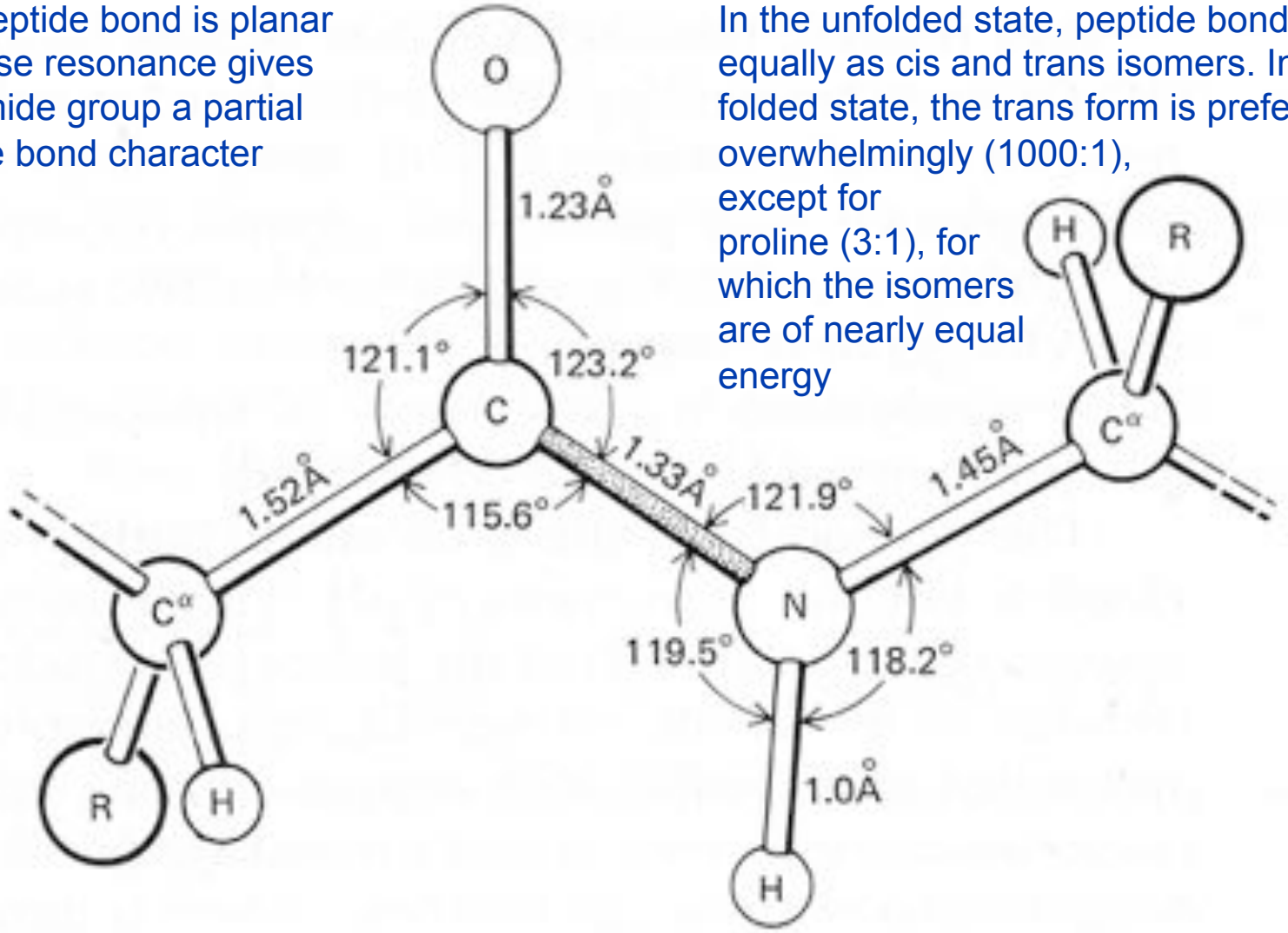
Primary - the succession of amino acids in the polypeptide chain (N to C)
Secondary - local, hydrogen-bonded configurations of the polypeptide chain
Tertiary - topological arrangement of secondary structures in 3D
Quaternary - non-covalent association of tertiary structures into a complex



Branden & Tooze, Introduction to Protein Structure

Primary structure - the peptide bond

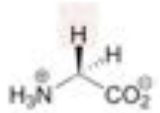
The peptide bond is planar because resonance gives the amide group a partial double bond character

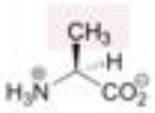


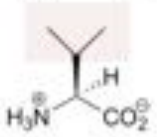
In the unfolded state, peptide bonds occur equally as cis and trans isomers. In the folded state, the trans form is preferred overwhelmingly (1000:1), except for proline (3:1), for which the isomers are of nearly equal energy

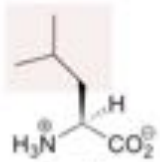
Primary structure - amino acid sidechains (= residues)

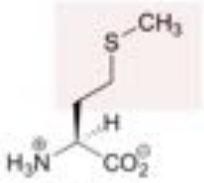
Nonpolar, aliphatic side groups

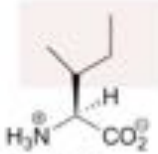

Glycine
Gly, G


Alanine
Ala, A

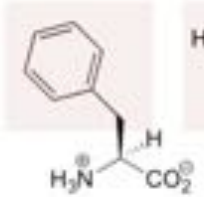

Valine
Val, V

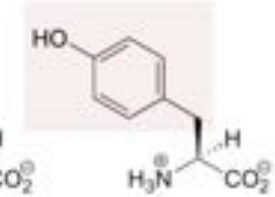

Leucine
Leu, L

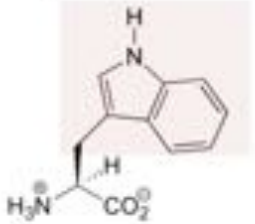

Methionine
Met, M


Isoleucine
Ile, I

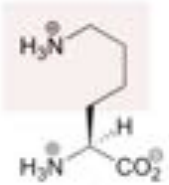
Aromatic side groups

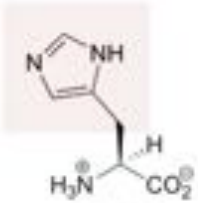

Phenylalanine
Phe, F

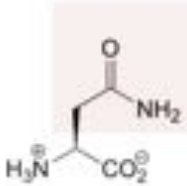

Tyrosine
Tyr, Y


Tryptophan
Trp, W

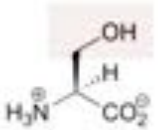
Positively charged side groups

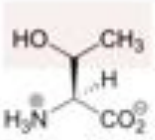

Lysine
Lys, K

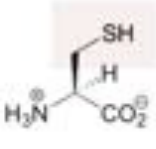

Histidine
His, H

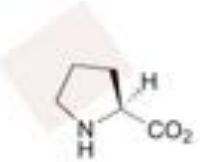

Asparagine
Asn, N

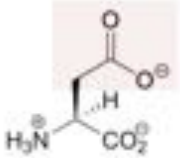
Polar, uncharged side groups

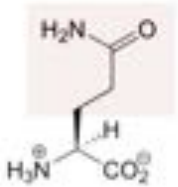

Serine
Ser, S


Threonine
Thr, T

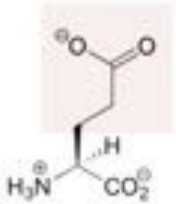

Cysteine
Cys, C

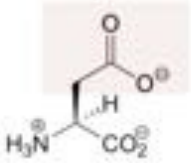

Proline
Pro, P


Aspartate
Asp, D


Glutamine
Gln, Q

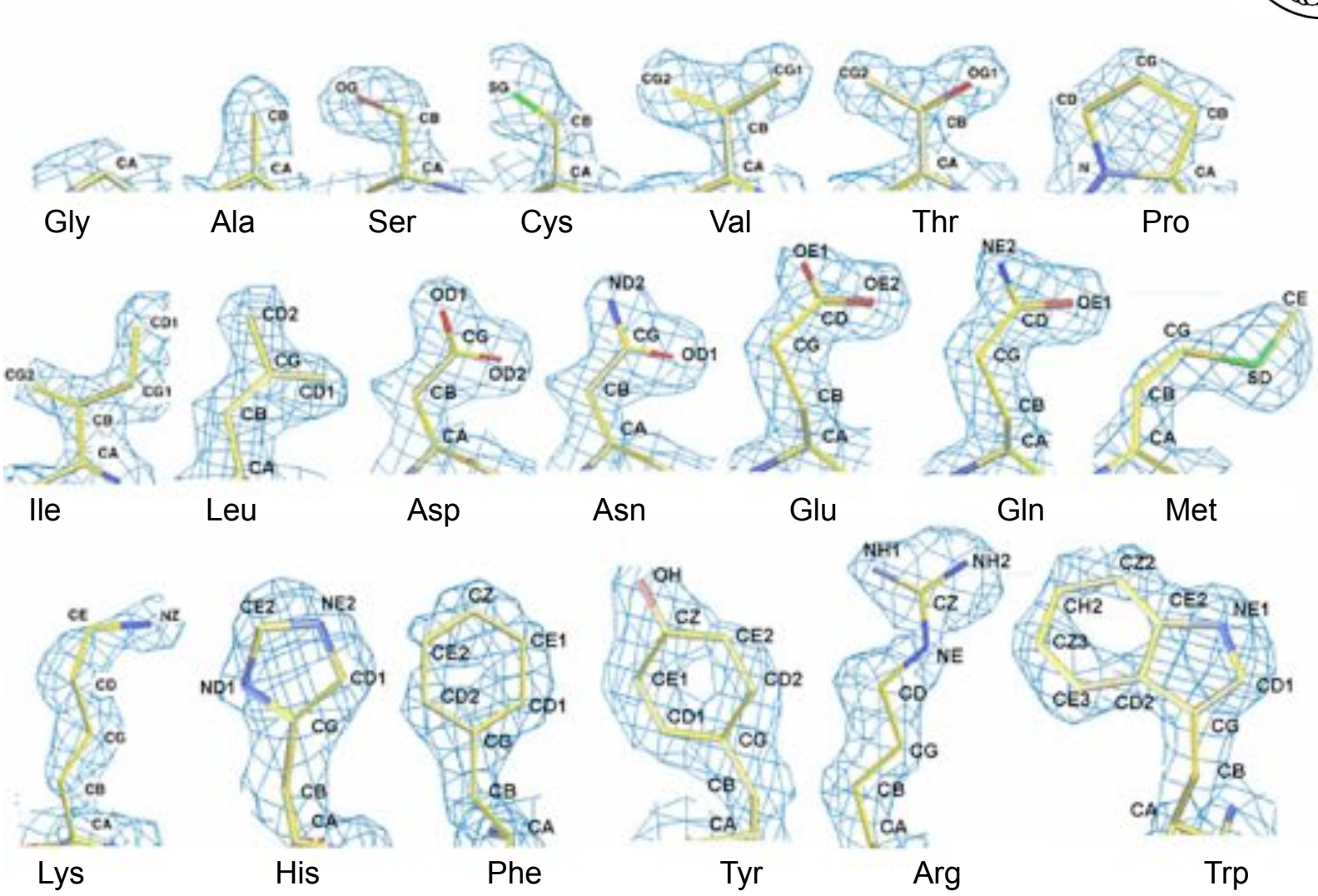
Negatively charged side groups


Glutamate
Glu, E


Aspartate
Asp, D

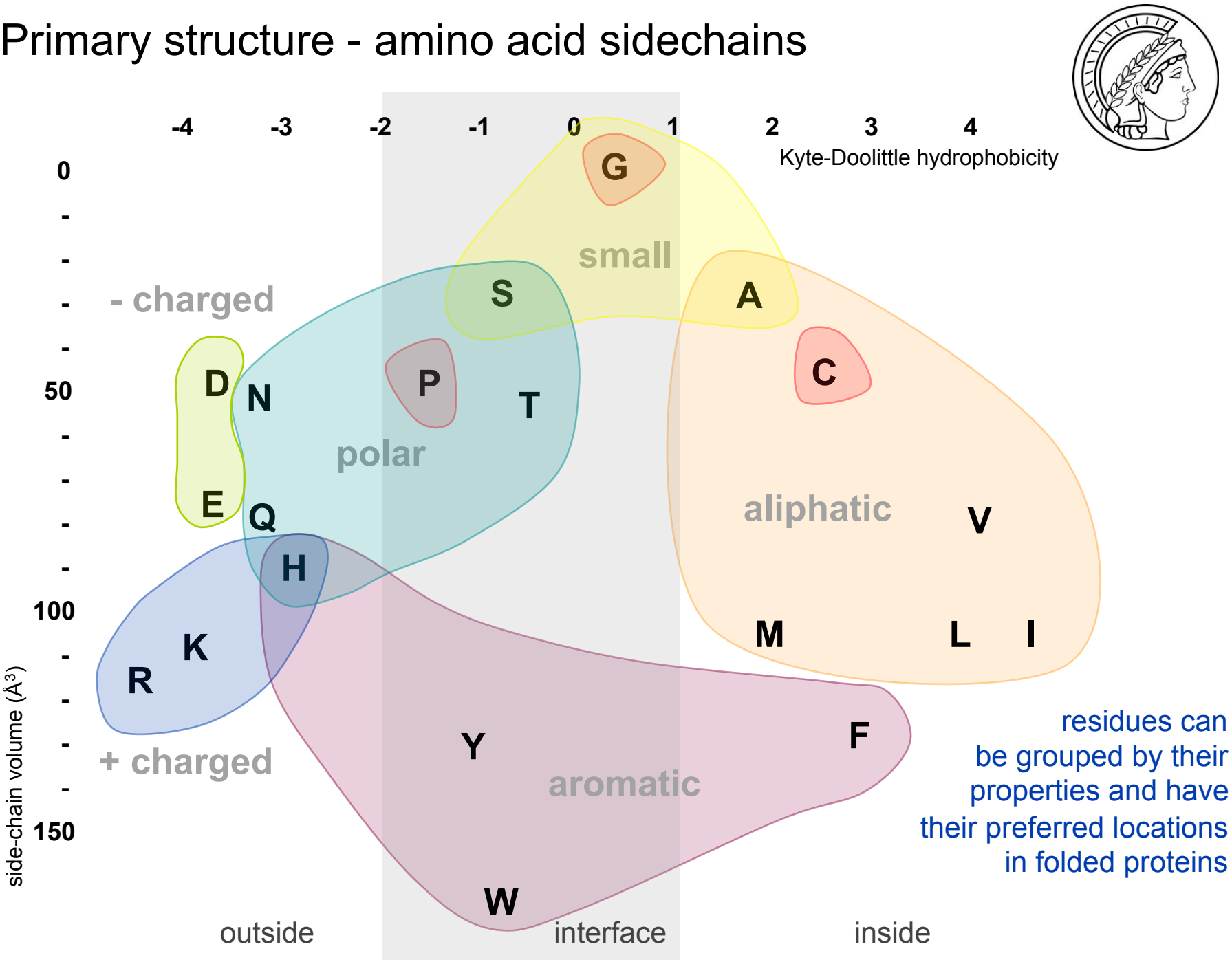
ChemWiki

Primary structure - amino acid sidechains

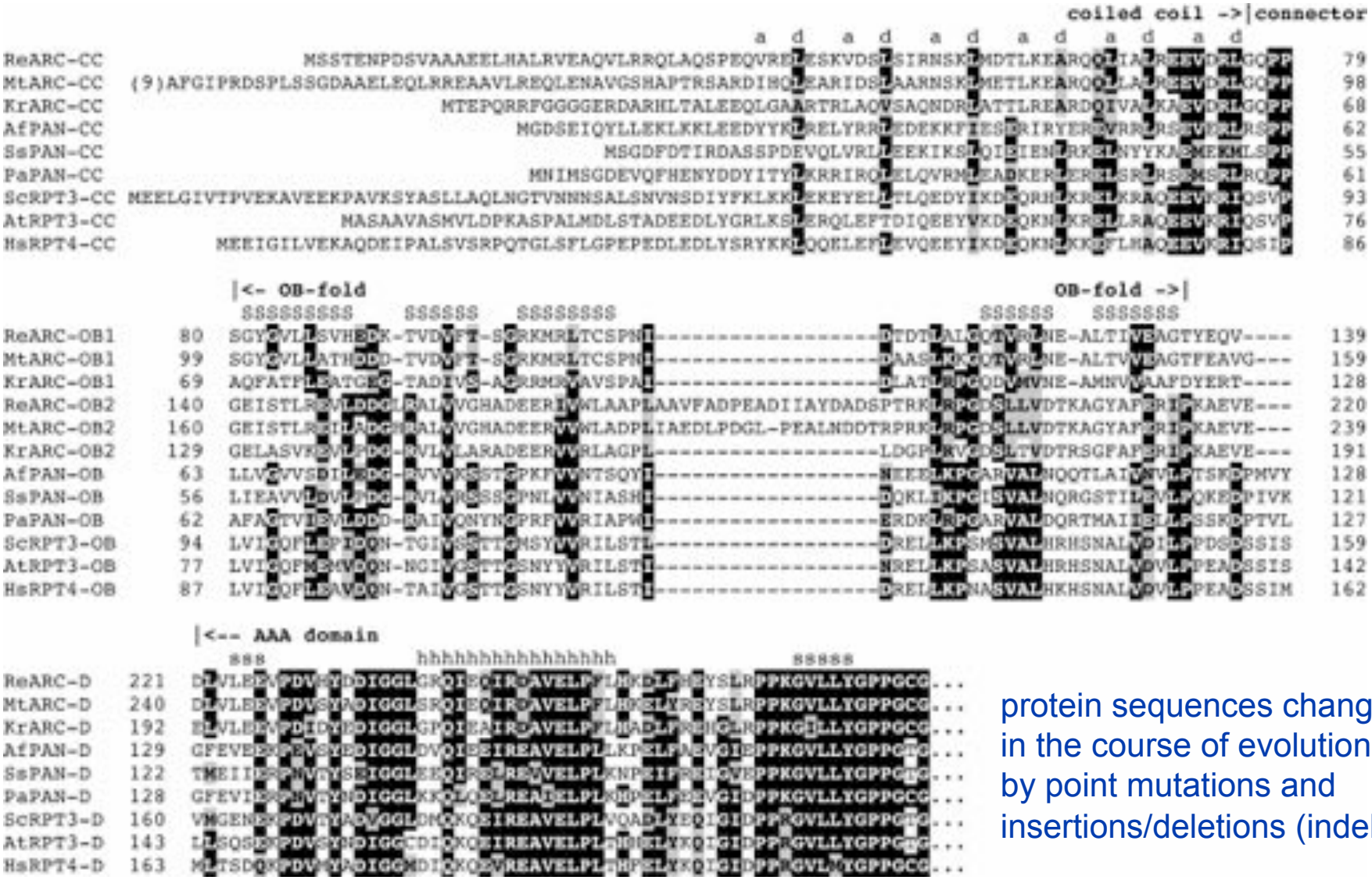


Mike Sawaya

Primary structure - amino acid sidechains

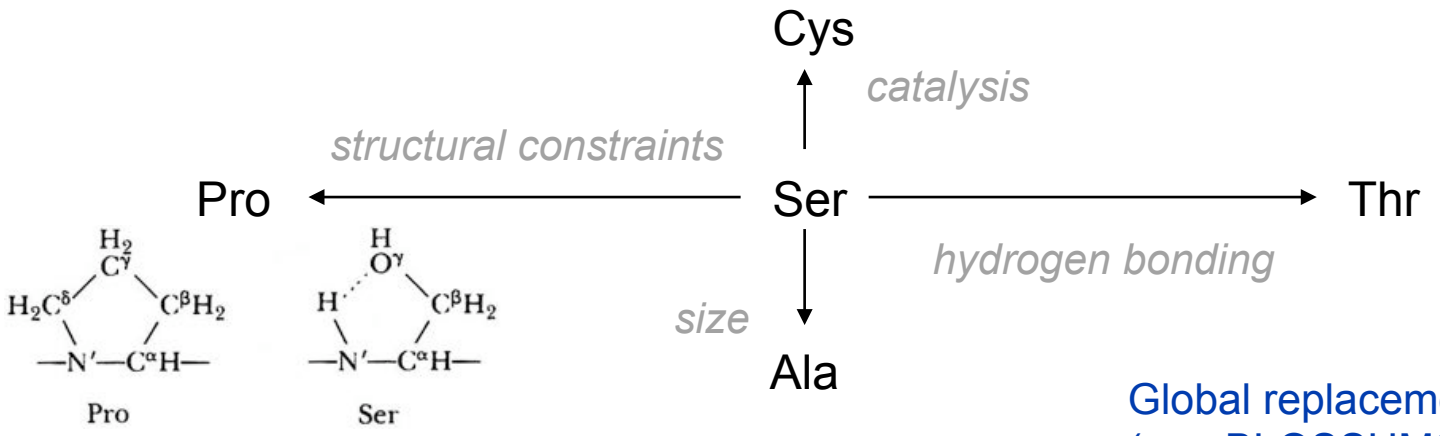


Primary structure - sequences and alignments



protein sequences change in the course of evolution by point mutations and insertions/deletions (indels)

Primary structure - amino acid substitutions



	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9																				C
S	−1	4																			S
T	−1	1	5																		T
P	−3	−1	−1	7																	P
A	0	1	0	−1	4																A
G	−3	0	−2	−2	0	6															G
N	−3	1	0	−2	−2	0	6														N
D	−3	0	−1	−1	−2	−1	1	6													D
E	−4	0	−1	−1	−1	−2	0	2	5												E
Q	−3	0	−1	−1	−1	−2	0	0	2	5											Q
H	−3	−1	−2	−2	−2	−2	1	−1	0	0	8										H
R	−3	−1	−1	−2	−1	−2	0	−2	0	1	0	5									R
K	−3	0	−1	−1	−1	−2	0	−1	1	1	−1	2	5								K
M	−1	−1	−1	−2	−1	−3	−2	−3	−2	0	−2	−1	−1	5							M
I	−1	−2	−1	−3	−1	−4	−3	−3	−3	−3	−3	−3	−3	1	4						I
L	−1	−2	−1	−3	−1	−4	−3	−4	−3	−2	−3	−2	−2	2	2	4					L
V	−1	−2	0	−2	0	−3	−3	−3	−2	−2	−3	−3	−2	1	3	1	4				V
F	−2	−2	−2	−4	−2	−3	−3	−3	−3	−3	−1	−3	−3	0	0	0	−1	6			F
Y	−2	−2	−2	−3	−2	−3	−2	−3	−2	−1	2	−2	−2	−1	−1	−1	−1	3	7		Y
W	−2	−3	−2	−4	−3	−2	−4	−4	−3	−2	−2	−3	−3	−1	−3	−2	−3	1	2	11	W
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

Global replacement matrices (e.g. BLOSSUM62) assign the same value to a substitution irrespective of its location and have global penalties for indels

Position-specific scoring matrices (PSSMs) derive specific substitution values from positional frequencies in multiple alignments, but maintain global penalties for indels

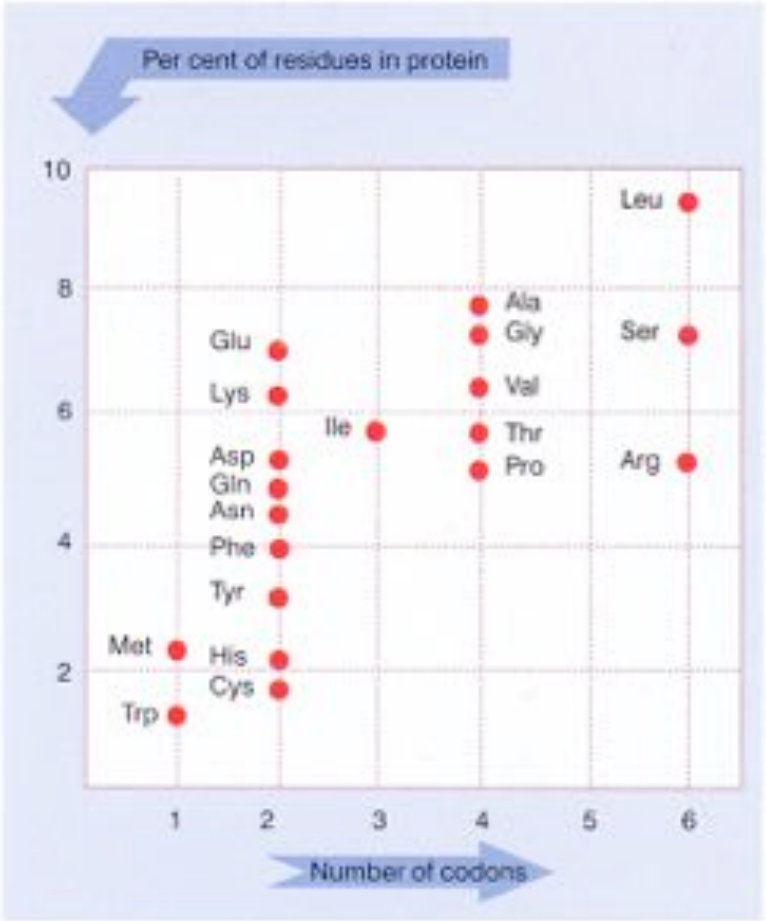
Profile Hidden Markov Models (HMMs) derive position-specific substitution and indel values from multiple alignments

Primary structure - the linearity of the peptide chain originates from the linear nature of the gene

The DNA code

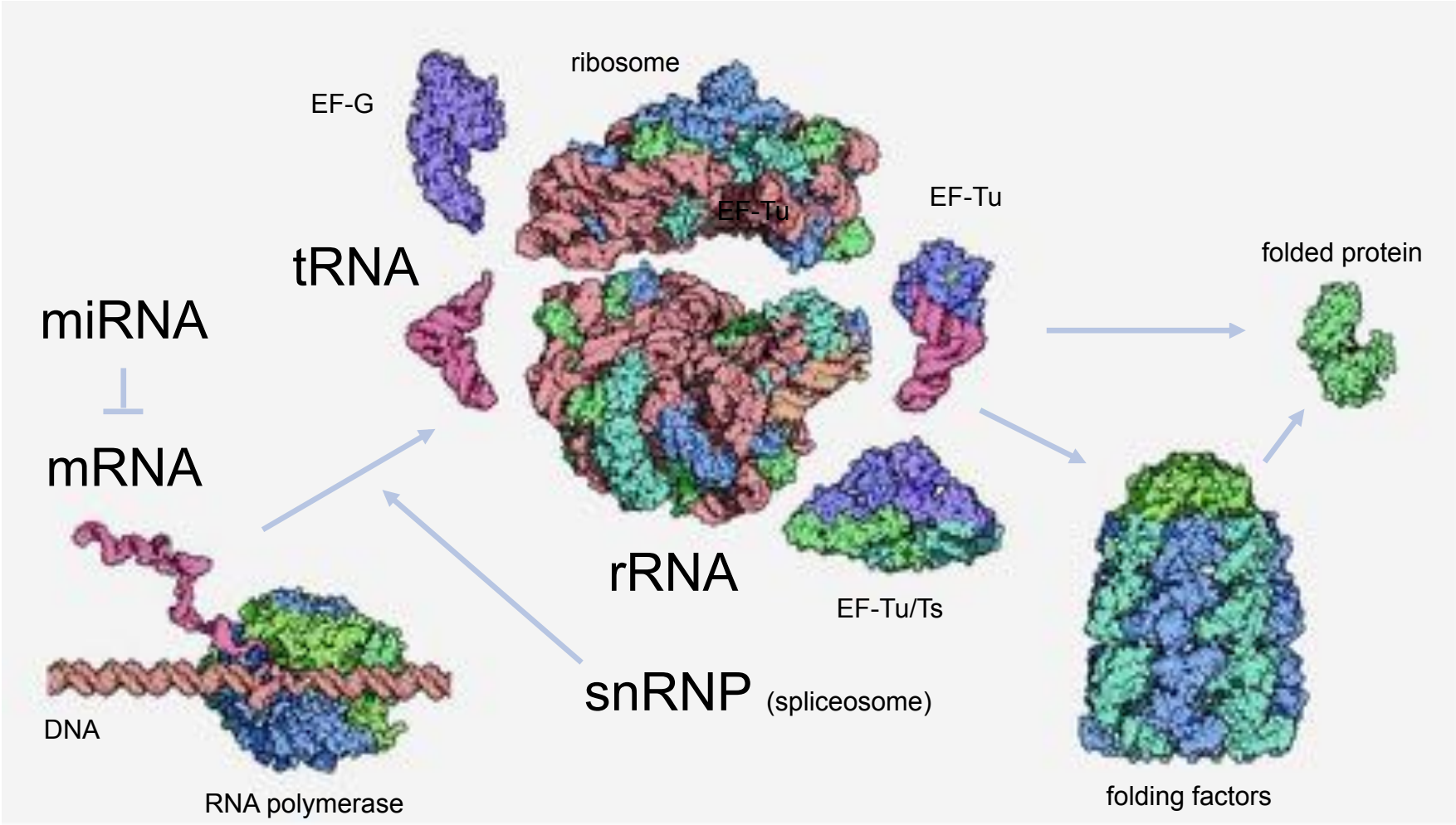
	U	C	A	G
U	UUU Phe UUC UUA Leu UUG	UCU UCC UCA Ser UCG	UAU Tyr UAC UAA STOP UAG	UGU Cys UGC UGA STOP UGG Trp
C	CUU CUC CUA Leu CUG	CCU CCC CCA Pro CCG	CAU His CAC CAA Gln CAG	CGU CGC CGA Arg CGG
A	AUU AUC AUA Ile AUG Met	ACU ACC ACA Thr ACG	AAU Asn AAC AAA Lys AAG	AGU Ser AGC AGA Arg AGG
G	GUU GUC GUA Val GUG	GCU GCC GCA Ala GCG	GAU Asp GAC GAA Glu GAG	GGU GGC GGA Gly GGG

Frequency of residues as a function of their number of codons



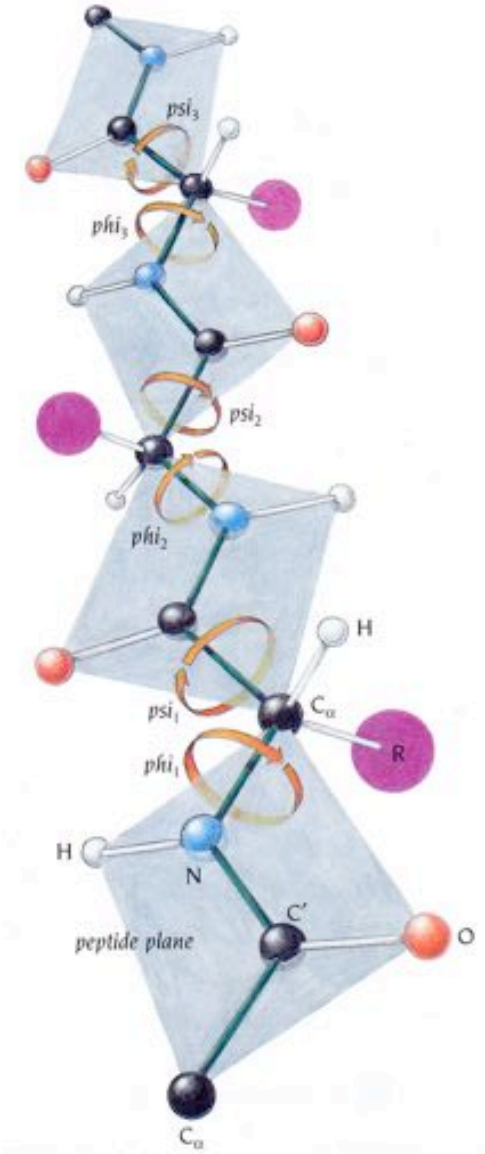
Benjamin Lewin, genes VII

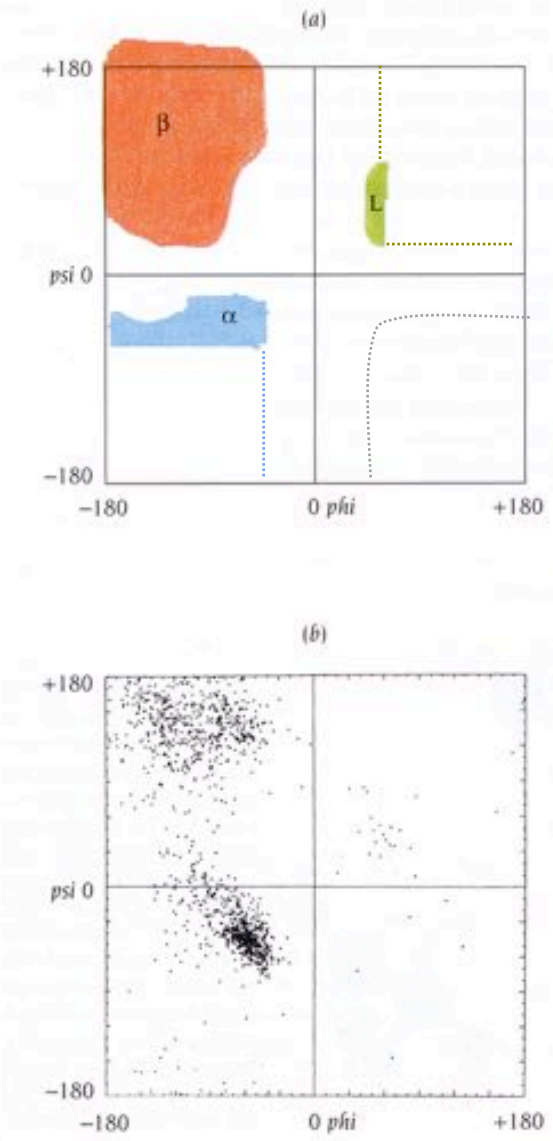
Primary structure - the central dogma



RNA molecules are the key players in converting genetic information into protein
→ remnants of an '**RNA world**' ?

Secondary structure - the Ramachandran plot





Because of its partial double bond nature, the peptide bond is always planar (ω = 180° (trans) or 0° (cis)).

Rotation is only possible around the N-Cα (phi ϕ) and Cα-C' (psi ψ) bonds.

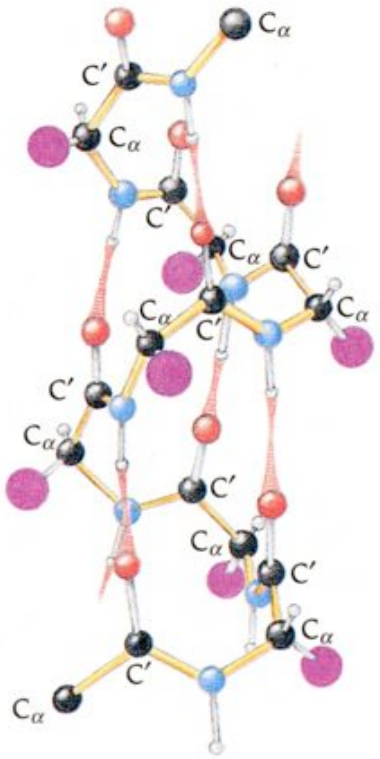
The Ramachandran plot shows the phi and psi angles that can be assumed by the peptide chain.

Because of the bulkiness of the Cβ carbon, angles of phi > 80 and psi < -80 are severely disfavored, except for glycine

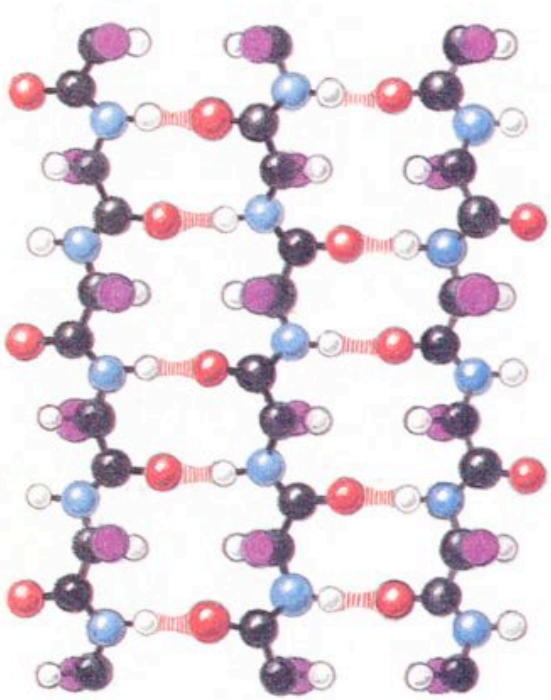
Branden & Tooze, Introduction to Protein Structure

Secondary structure - helices and sheets

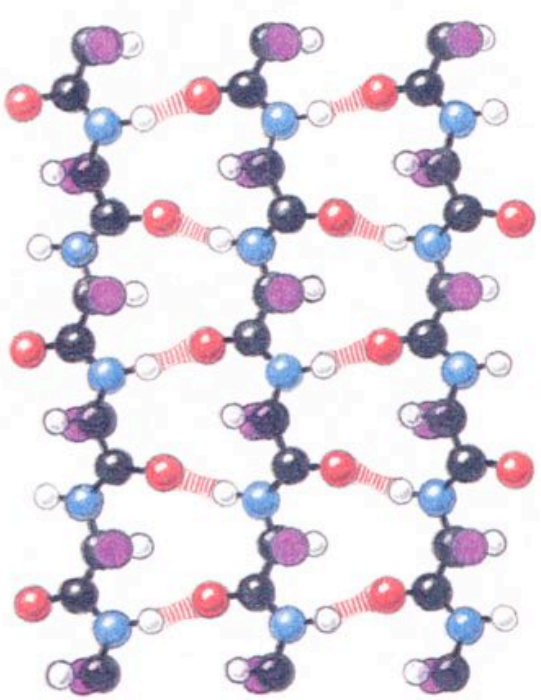
α-helix



antiparallel β-sheet

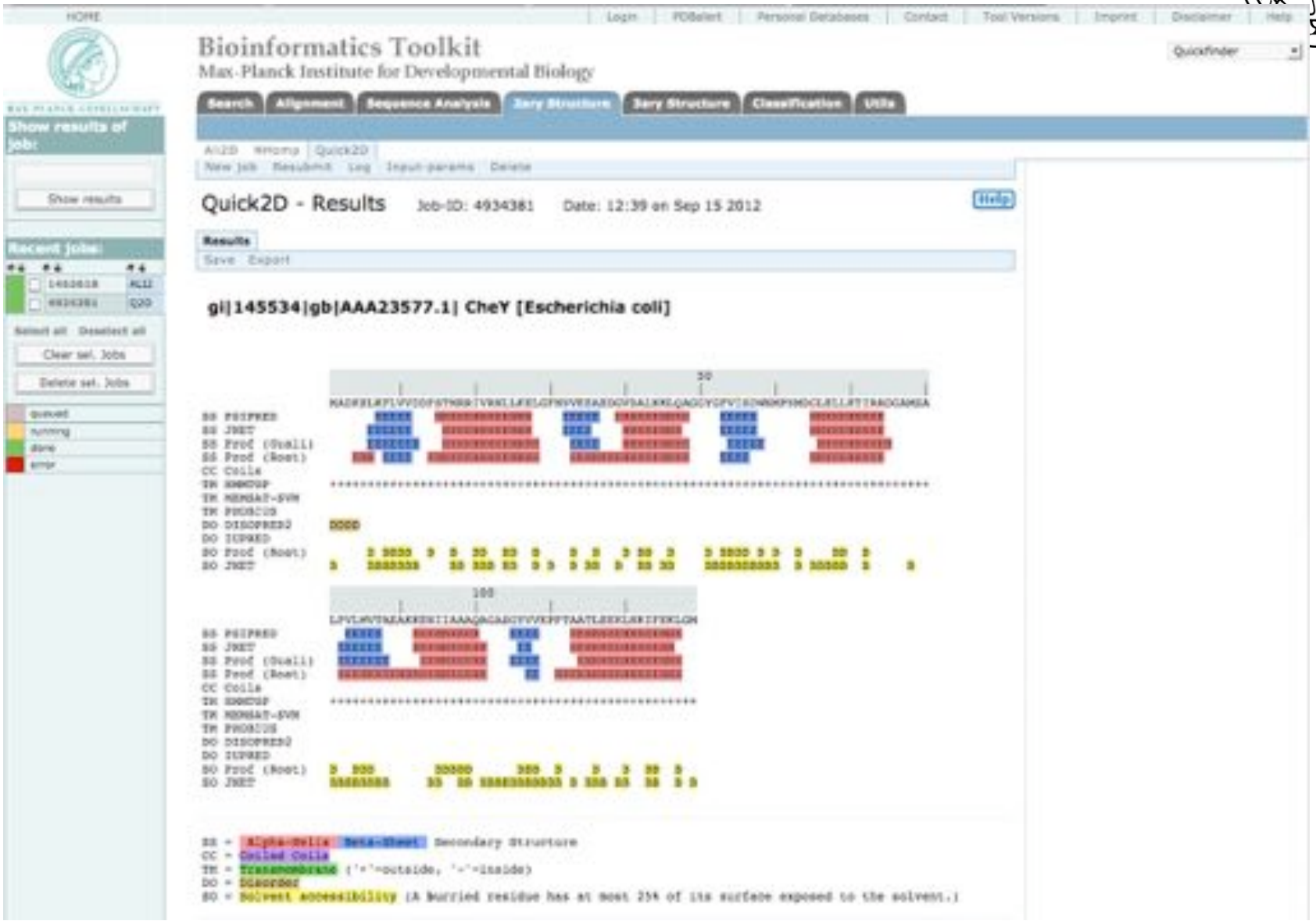


parallel β-sheet

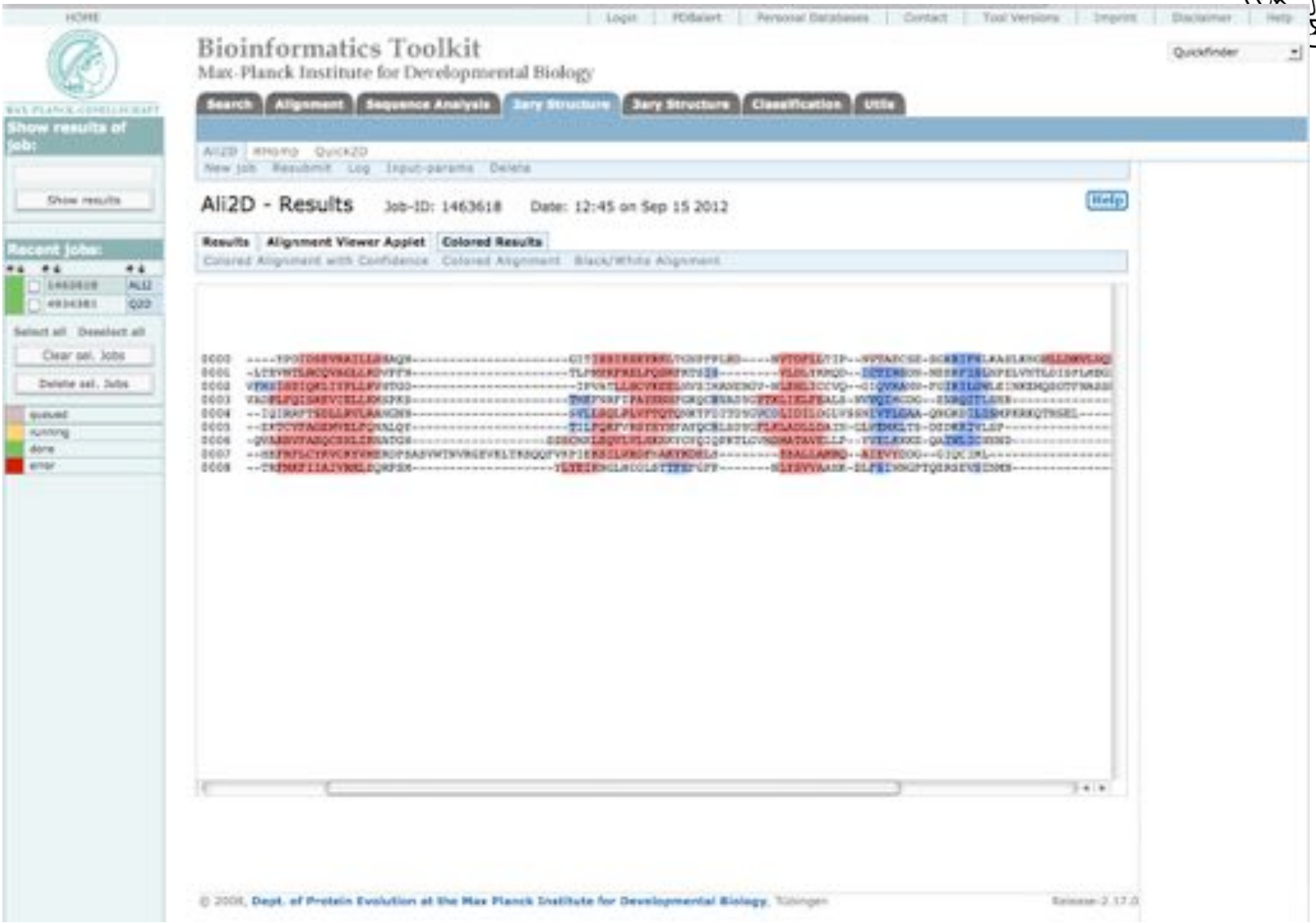


Branden & Tooze, Introduction to Protein Structure

Secondary structure - prediction: Quick2D

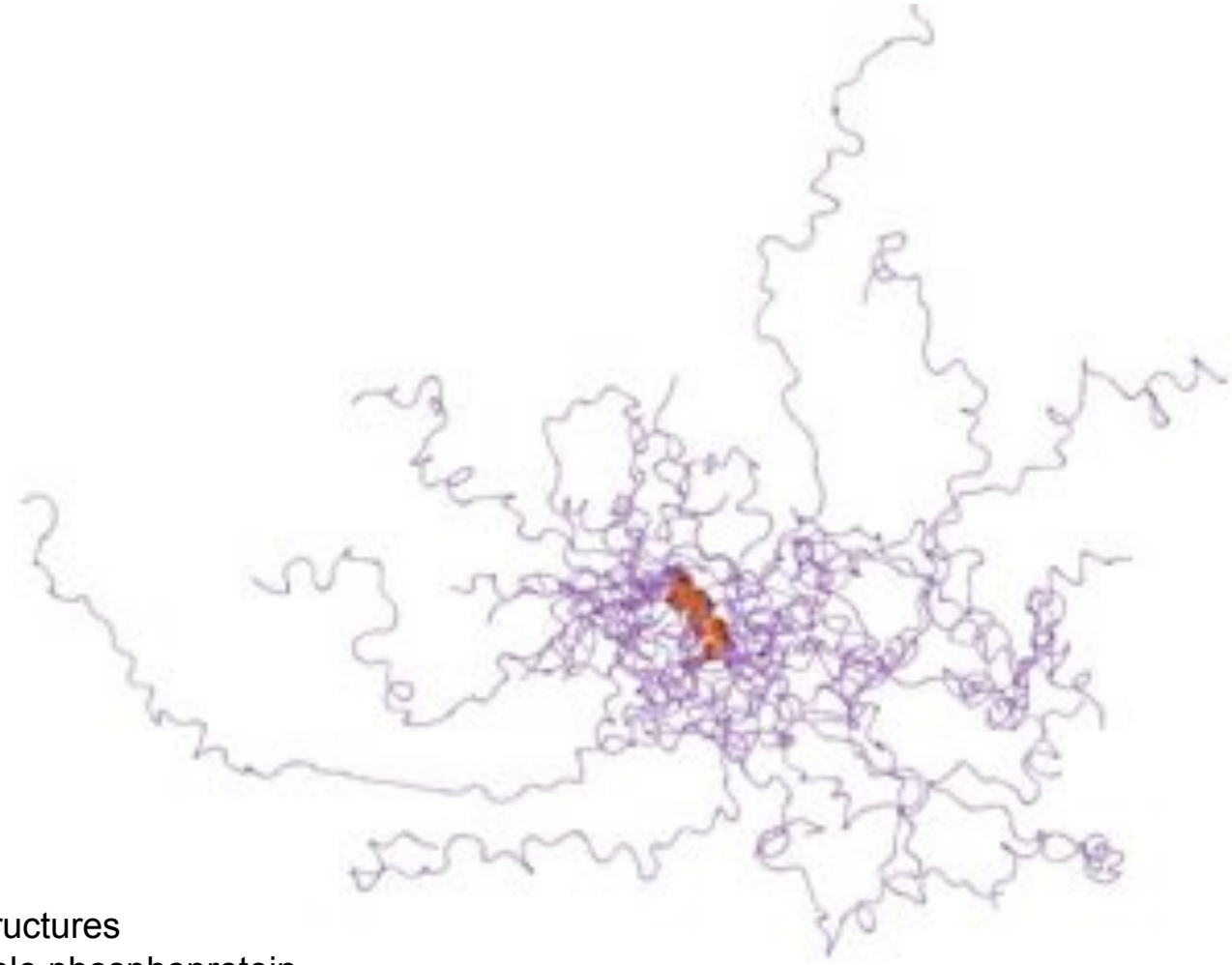


Secondary structure - prediction: Ali2D



Secondary structure - natively unstructured proteins

Sometimes proteins lack secondary structure.

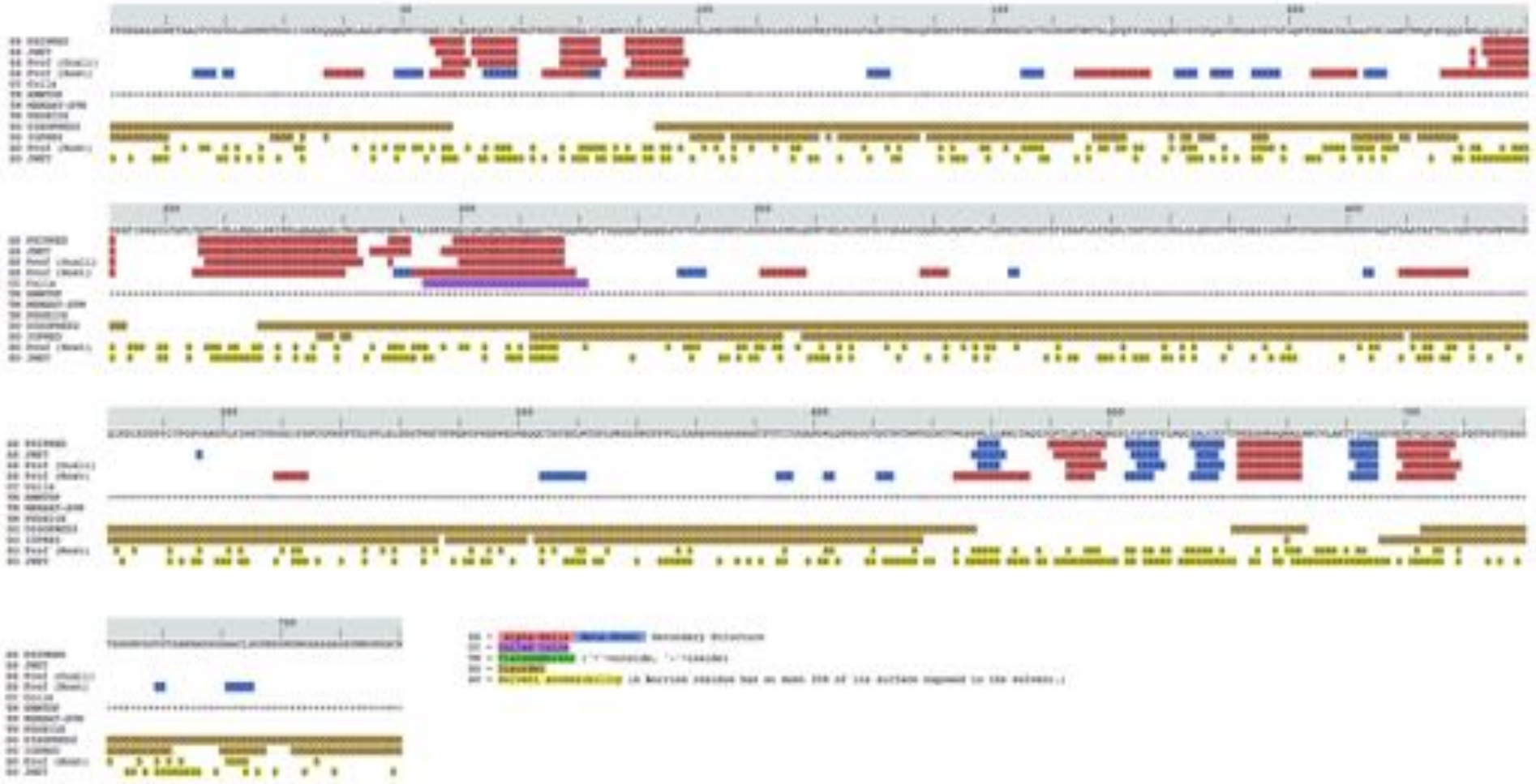


Ensemble of NMR structures
of the Thylakoid soluble phosphoprotein
TSP9, which shows a largely flexible protein chain (PDB 2fft)

Secondary structure - natively unstructured proteins

Sometimes proteins lack secondary structure. Few proteins, however, are entirely unstructured. Usually, proteins contain unstructured segments between folded regions.

Consensus secondary structure prediction for GW182, a protein with large unstructured segments

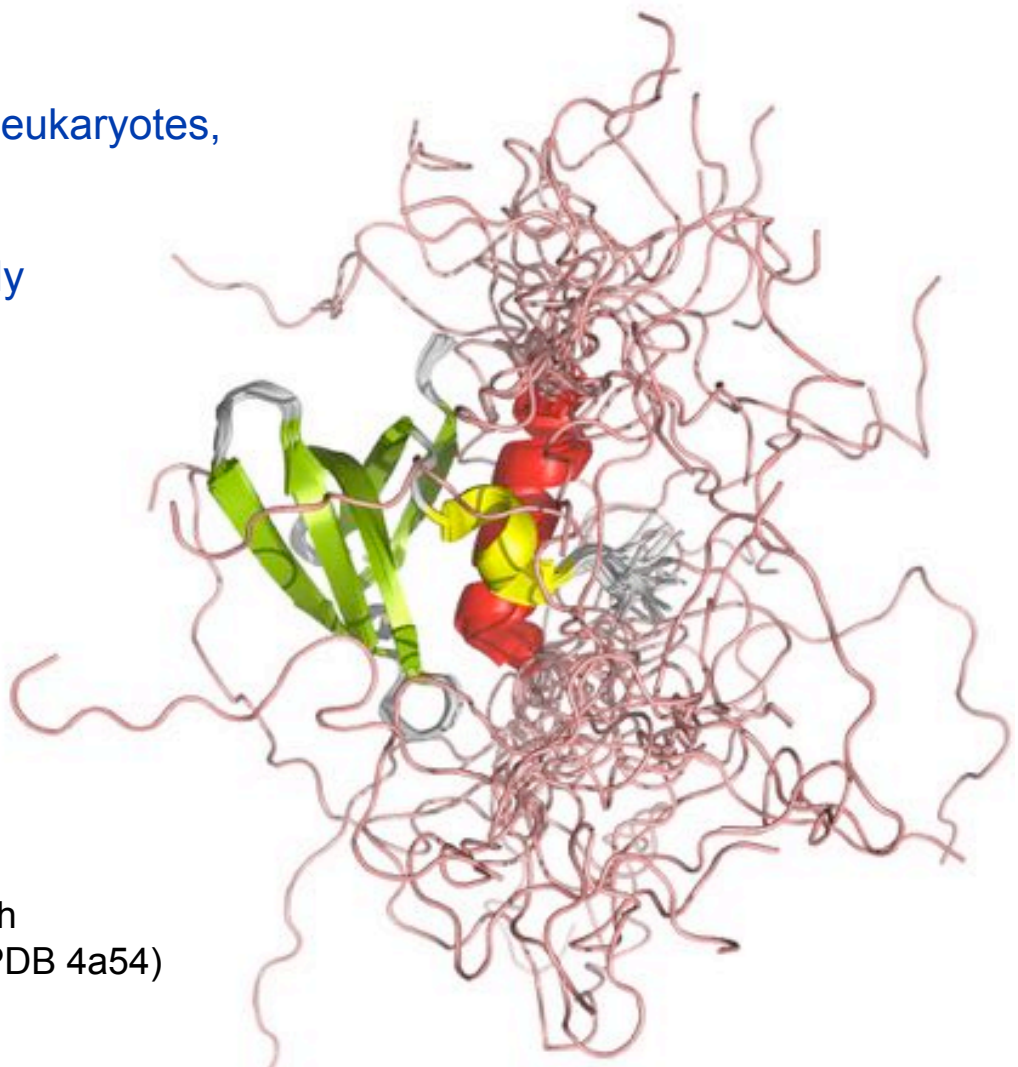


Secondary structure - natively unstructured proteins

Sometimes proteins lack secondary structure. Few proteins, however, are entirely unstructured. Usually, proteins contain unstructured segments between folded regions.

Unstructured segments are common in eukaryotes, but rare in prokaryotes.

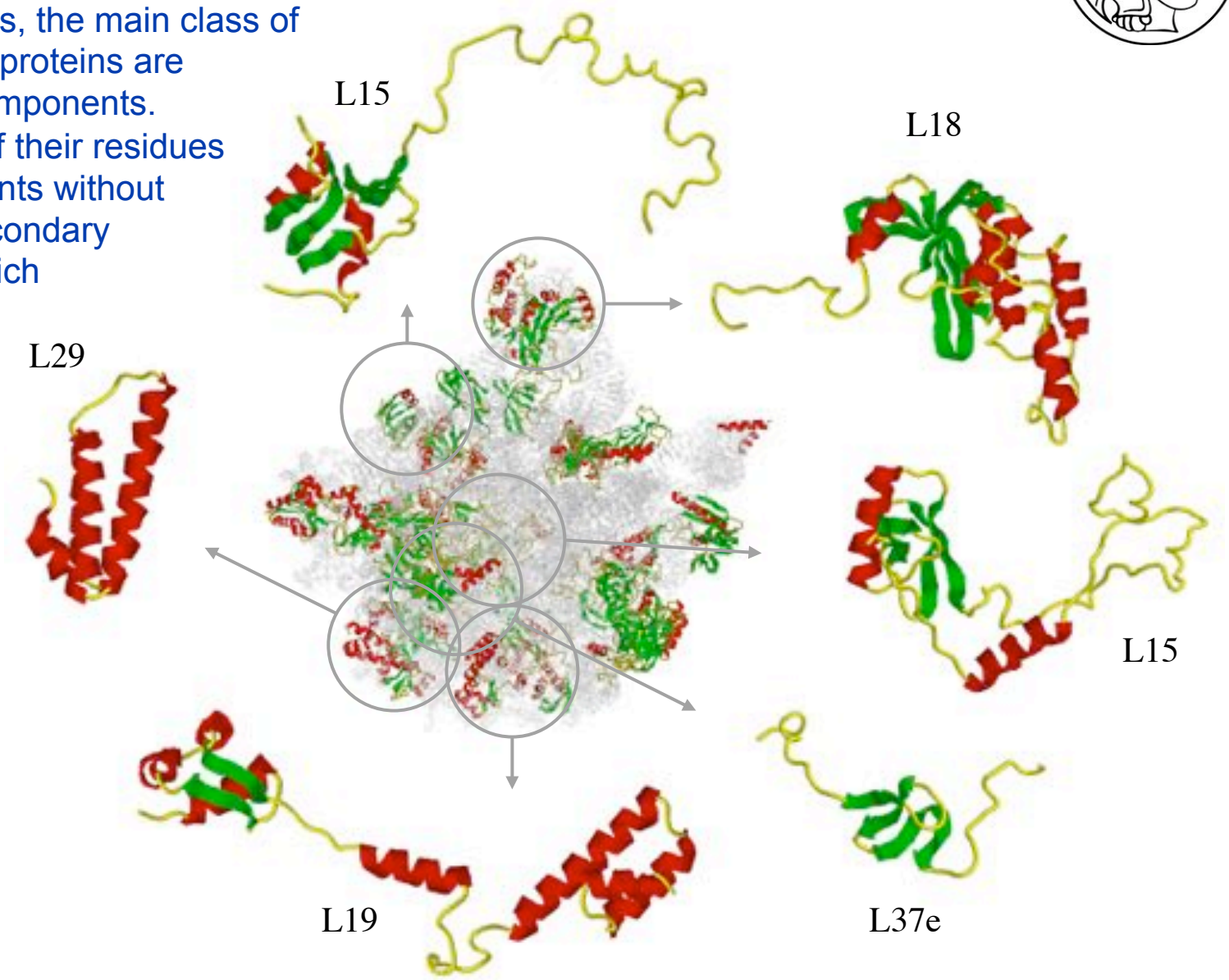
Unstructured segments almost invariably assume a defined structure in complex with a structured partner.



yeast Edc3 LSm domain in complex with
a leucine-rich motif (HLM) from Dcp2 (PDB 4a54)

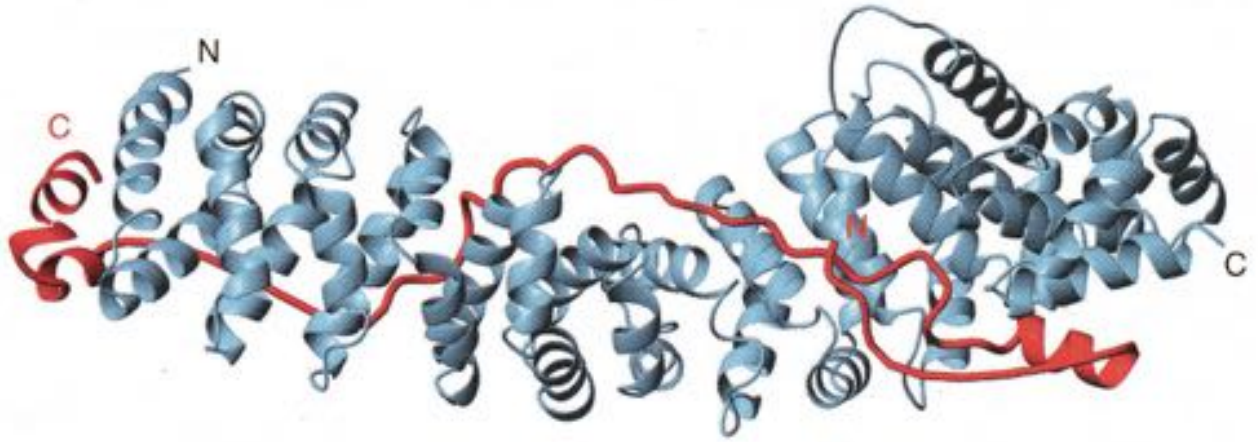
Secondary structure - natively unstructured proteins

In prokaryotes, the main class of unstructured proteins are ribosomal components. About 60% of their residues are in segments without a defined secondary structure, which in complex with the RNA.

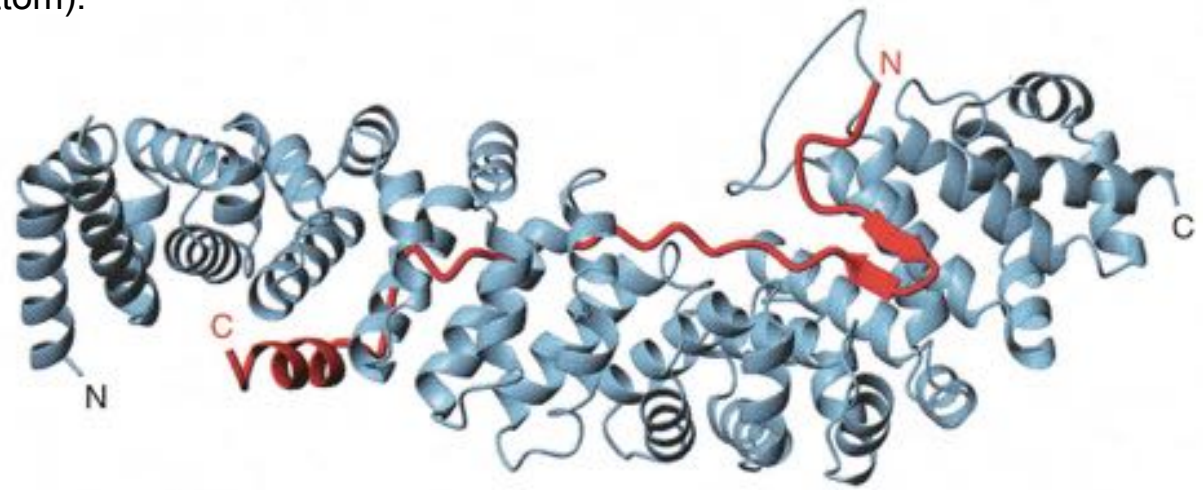


Secondary structure - natively unstructured proteins

In eukaryotes, unstructured segments are the main location for linear motifs, short sequences that mediate the specific association to folded ('scaffold') proteins.

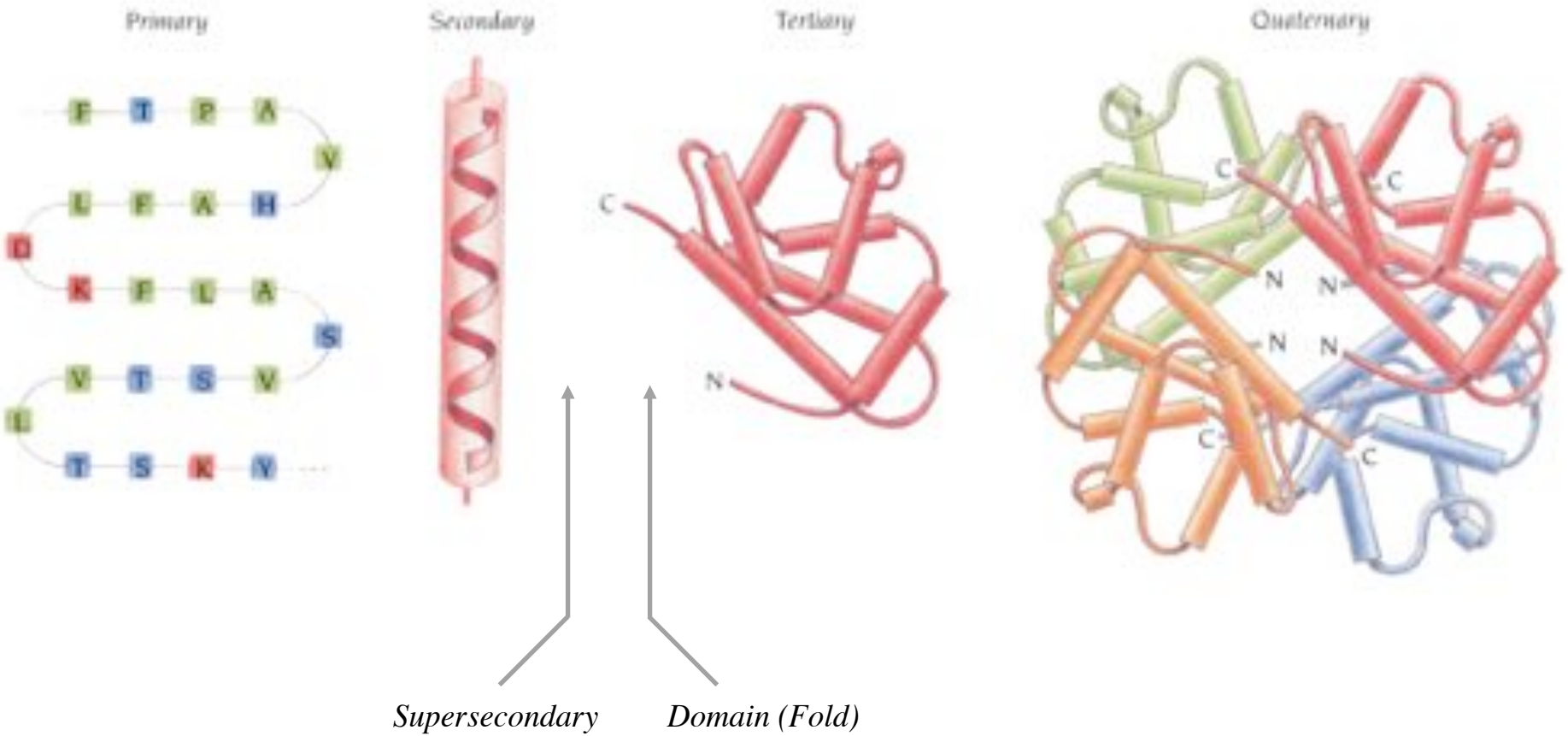


The armadillo repeat domain of β-catenin in complex with E-cadherin (top) and Tcf3 (bottom).



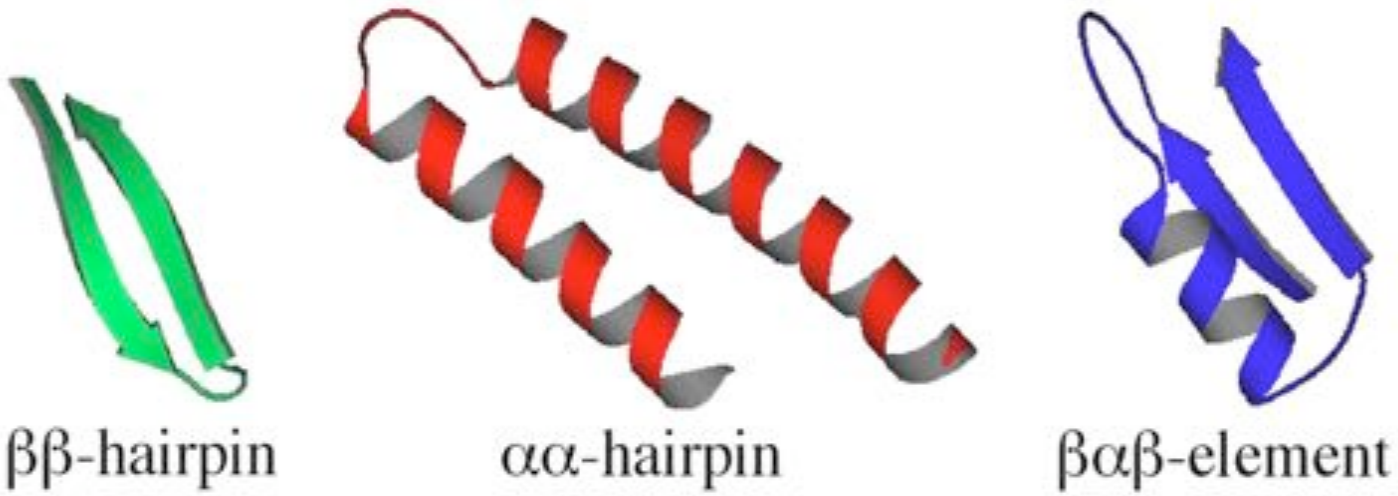
Dyson H.J. and Wright P.E., Curr.Opin.Struct.Biol. 12:54-60 (2002)

From secondary to tertiary structure



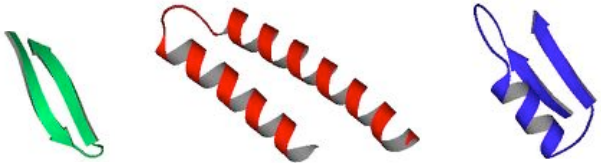
Branden & Tooze, Introduction to Protein Structure

The main supersecondary structure elements



Supersecondary structures are at the core of domains - the hierarchy of fold complexity

supersecondary structures

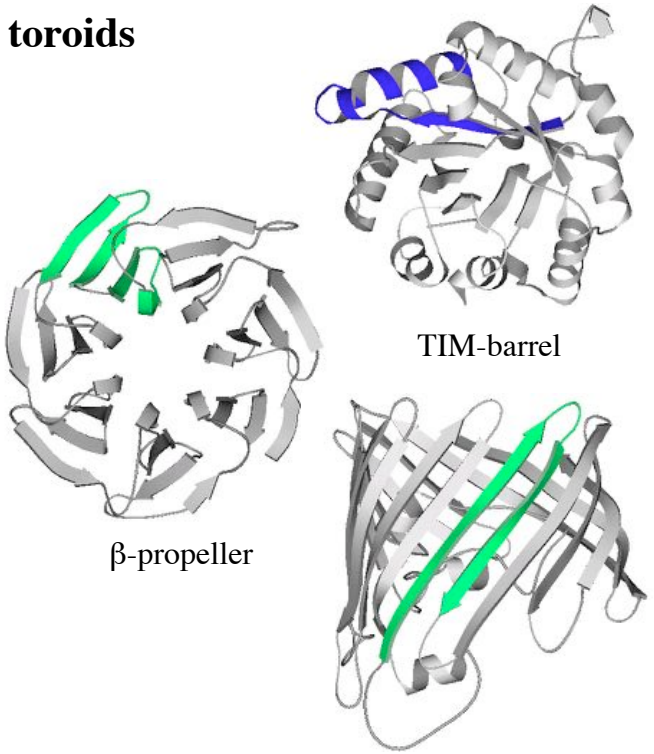


ββ-hairpin αα-hairpin βαβ-element

solenoids

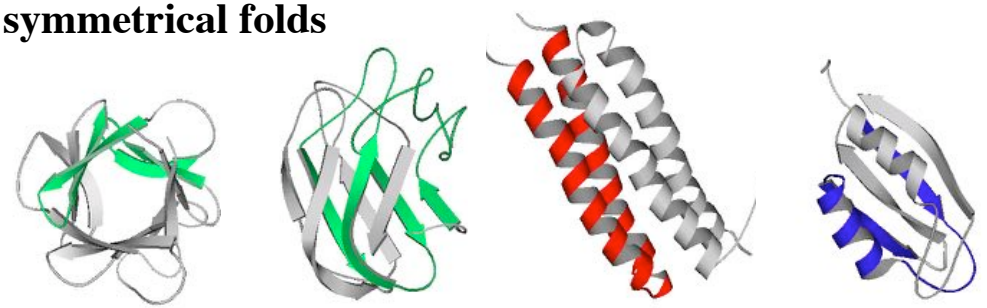
choline-binding domains tetratricopeptide repeats leucine-rich repeats

toroids



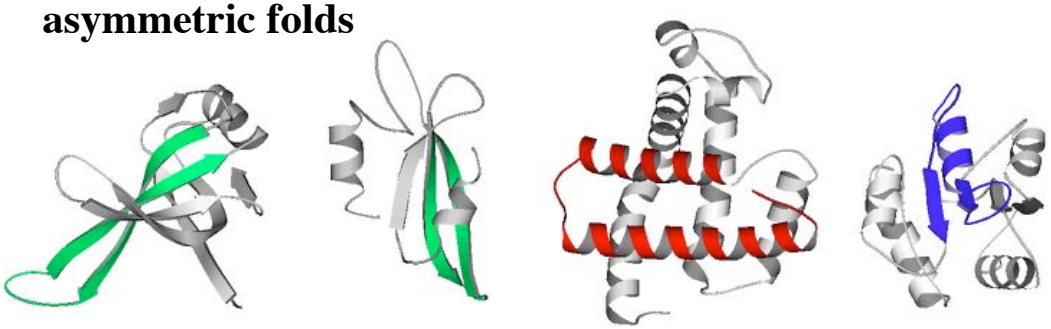
β-propeller TIM-barrel porin

symmetrical folds



β-trefoil jelly-roll 4-helix bundle ferredoxin

asymmetric folds

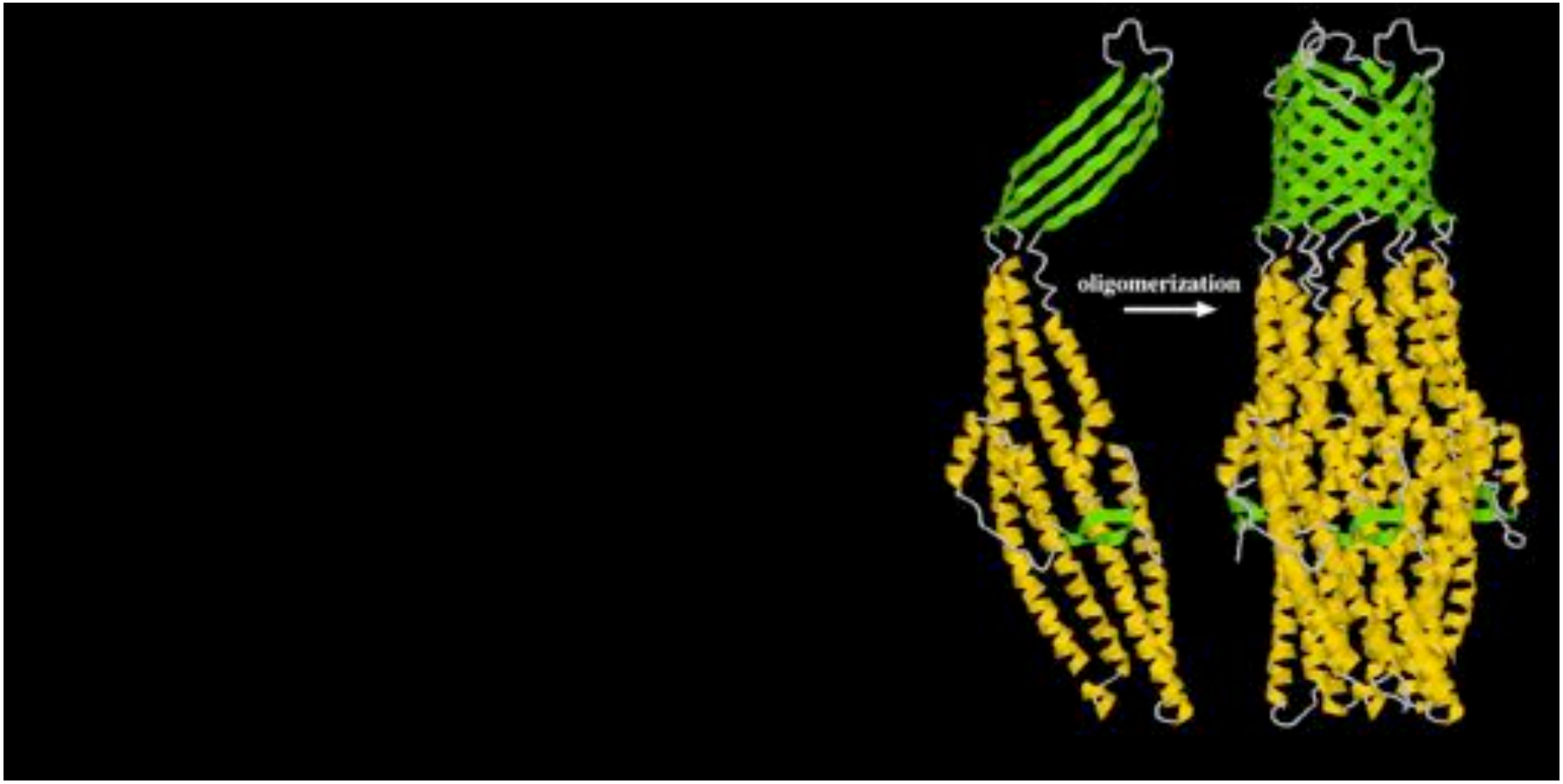


OB-fold UB-roll globin flavodoxin

Deconstructing a protein into supersecondary structures

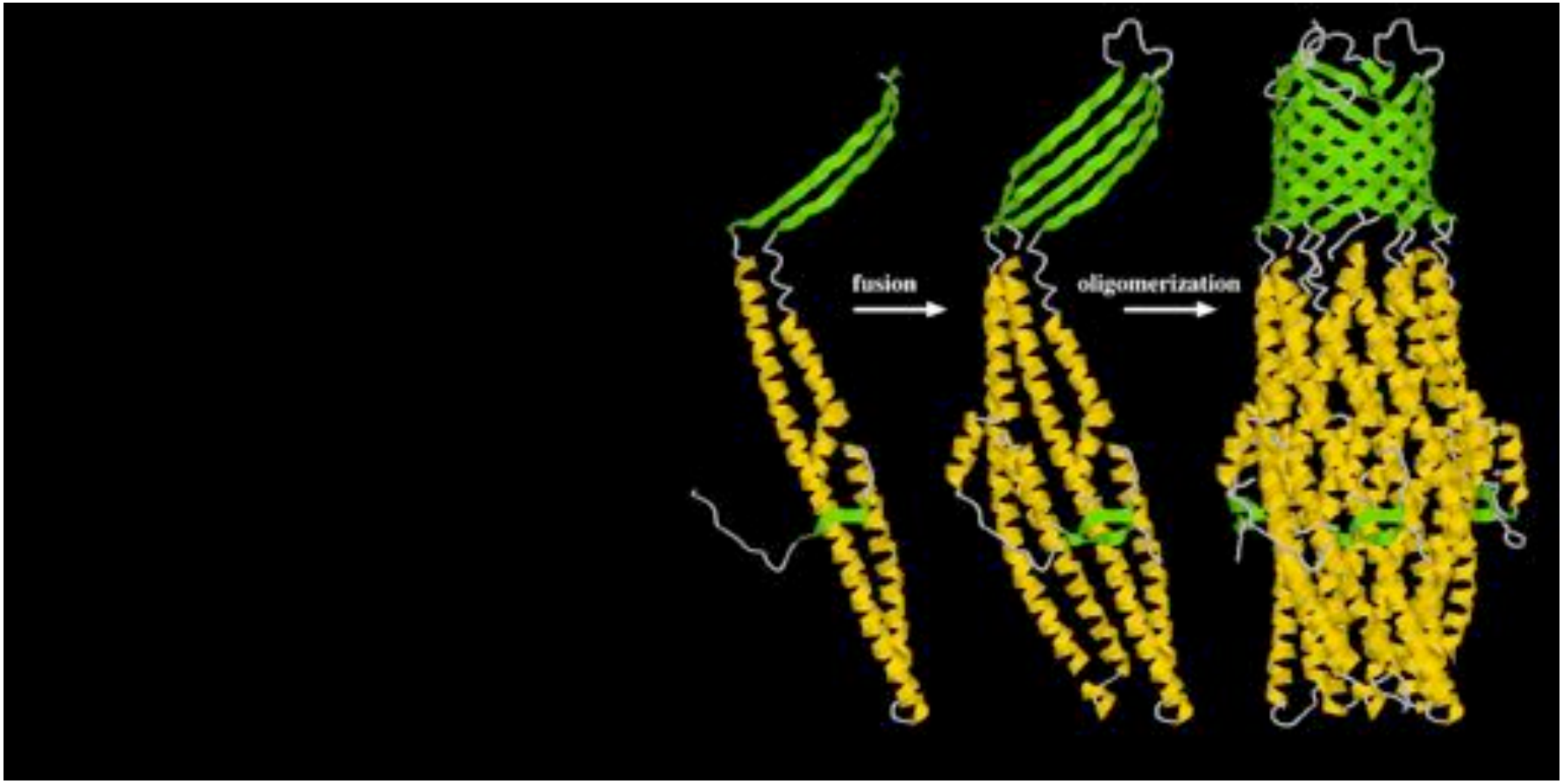
TolC

Deconstructing a protein into supersecondary structures



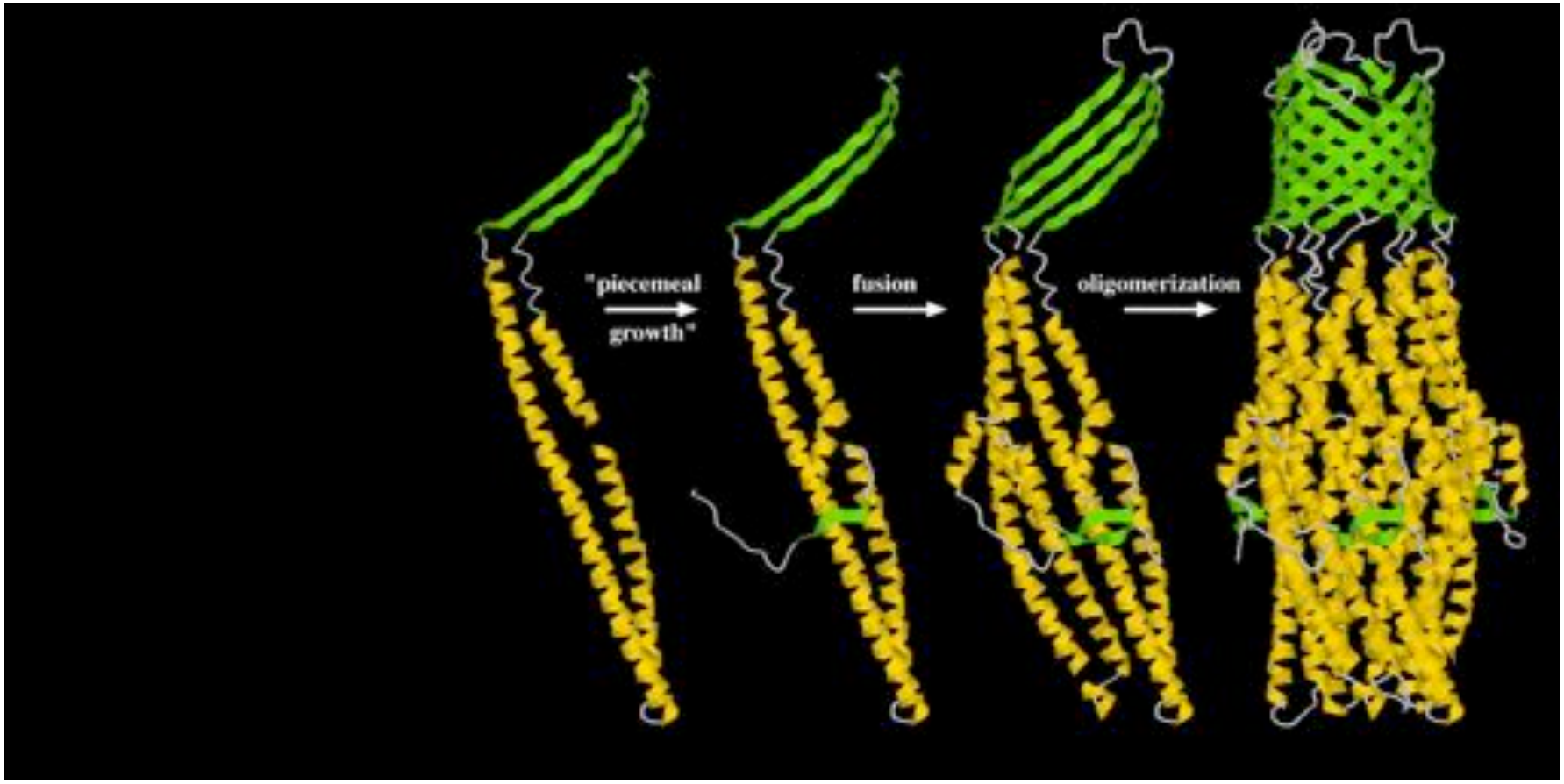
TolC

Deconstructing a protein into supersecondary structures



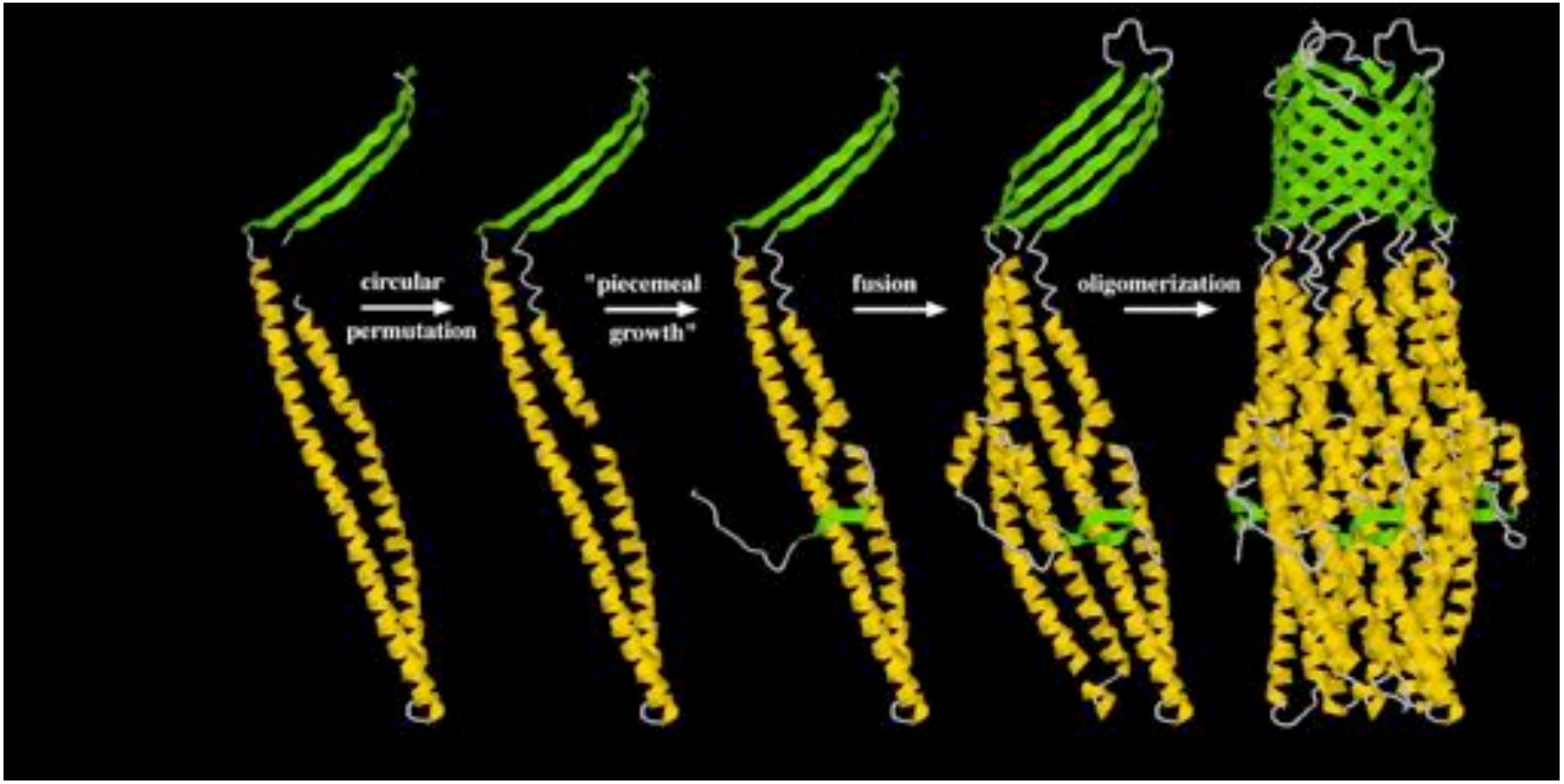
TolC

Deconstructing a protein into supersecondary structures



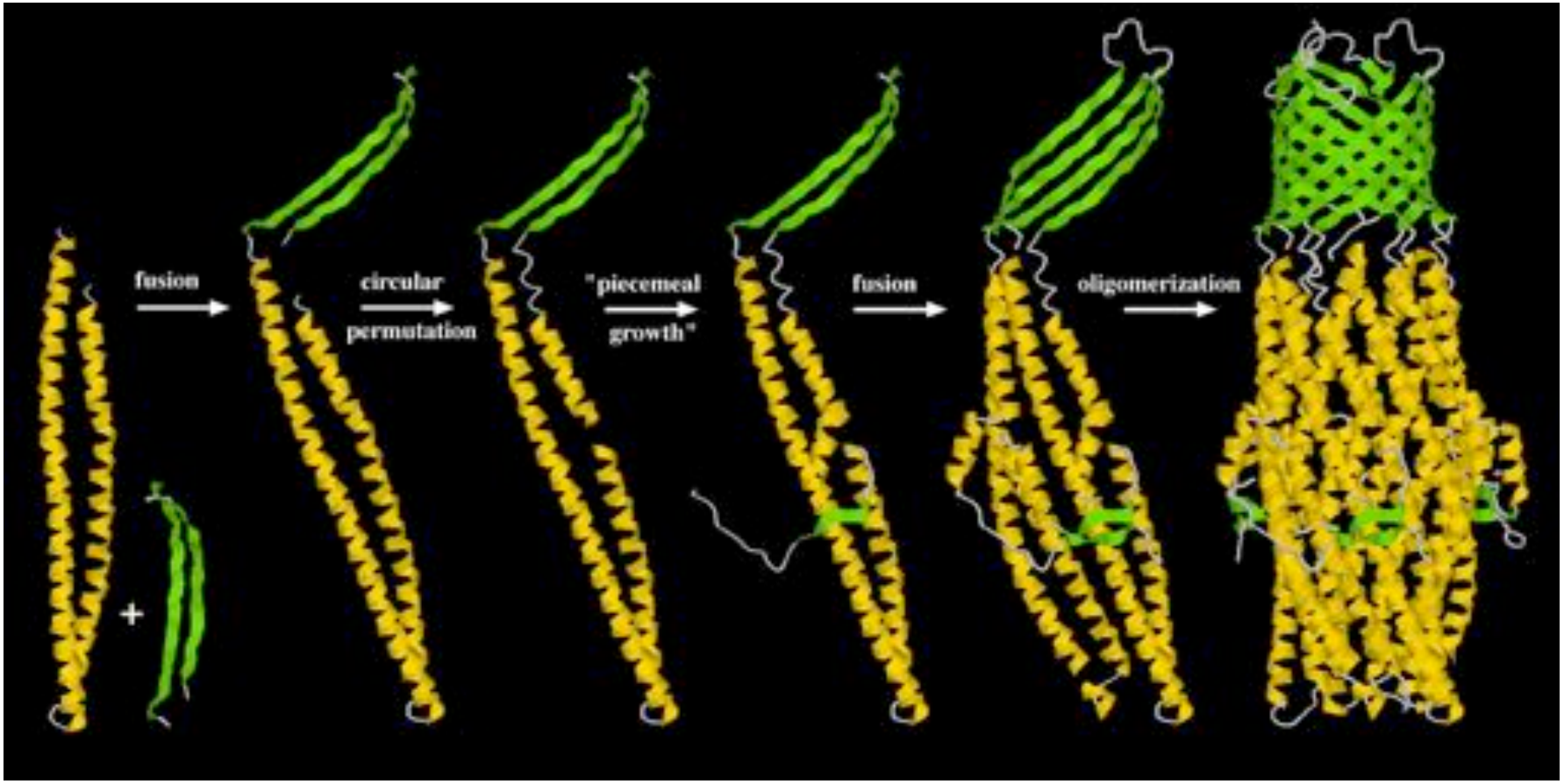
TolC

Deconstructing a protein into supersecondary structures



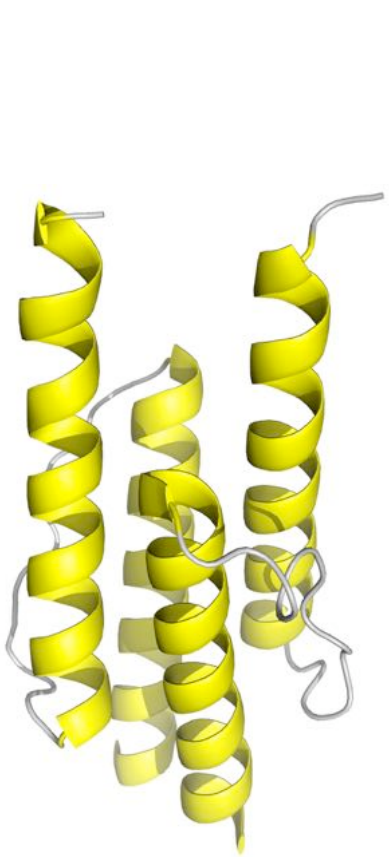
TolC

Deconstructing a protein into supersecondary structures



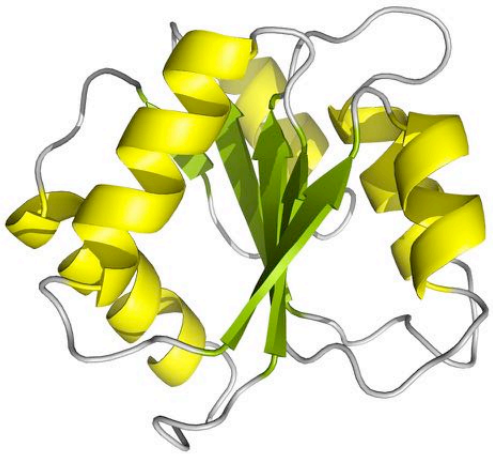
TolC

Tertiary structure - the fold classes

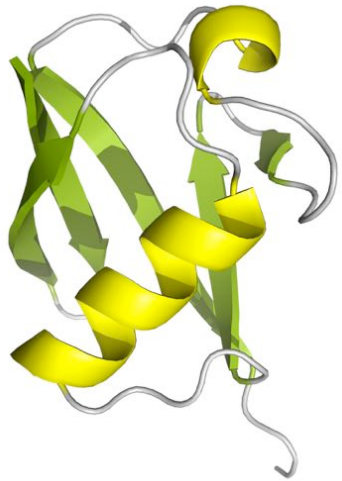


α (formed of α-helices)

β (formed of β-strands)

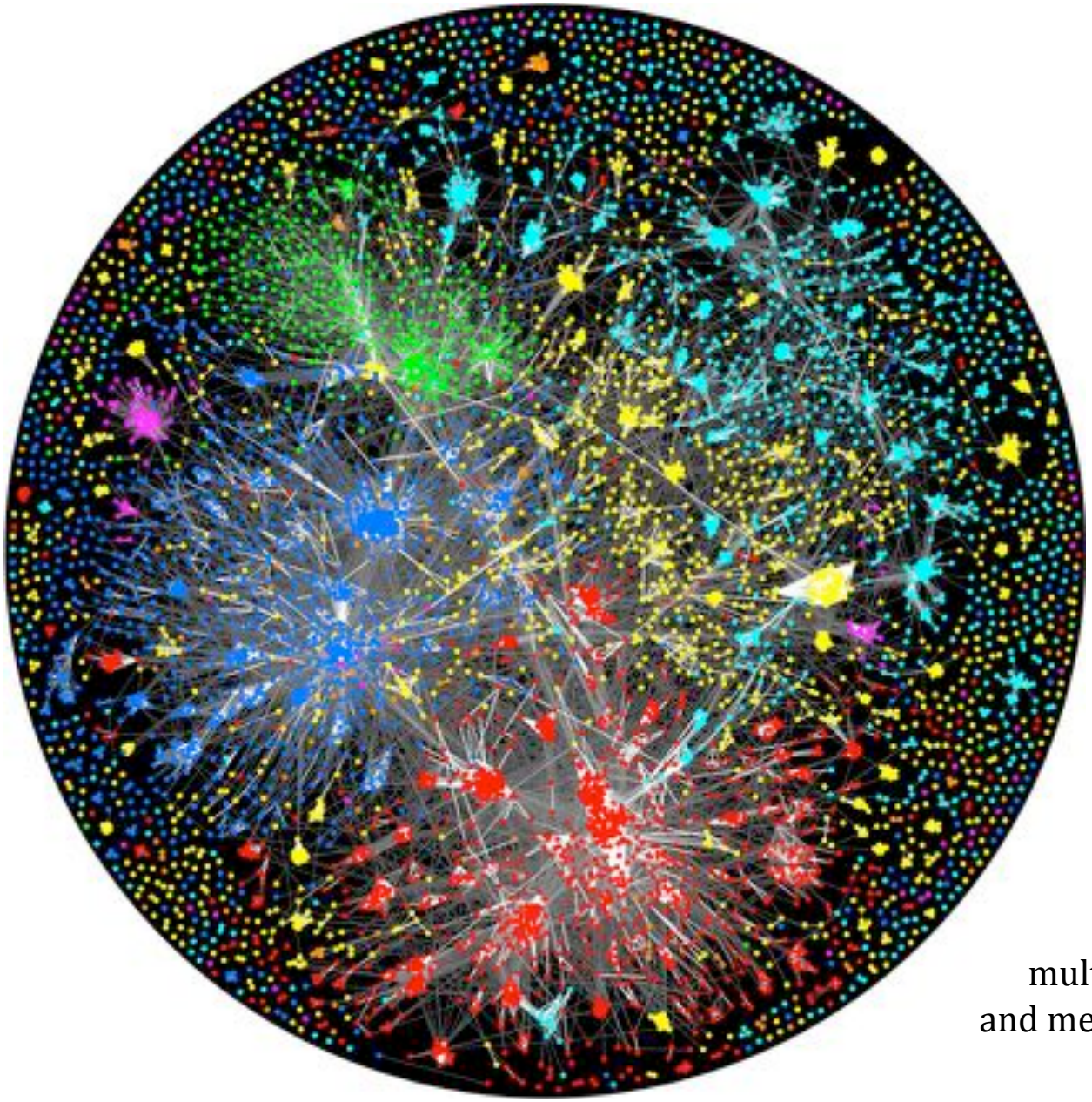


α/β (formed of alternating α-helices and β-strands; strands mainly parallel)



α+β (formed of irregular arrangements of α-helices and β-strands; strands mainly antiparallel)

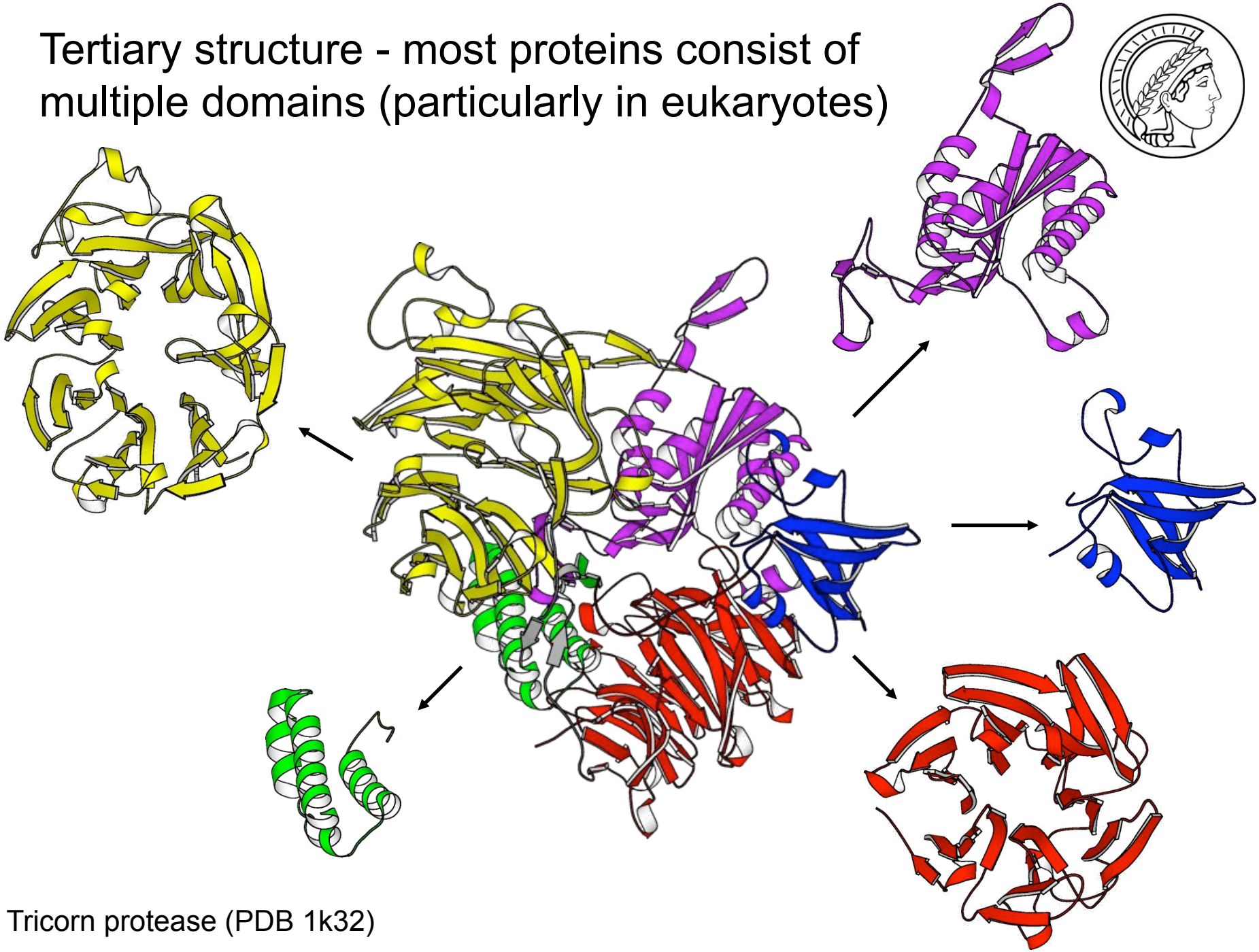
Tertiary structure - the fold classes



Structure classes have preferred residue composition

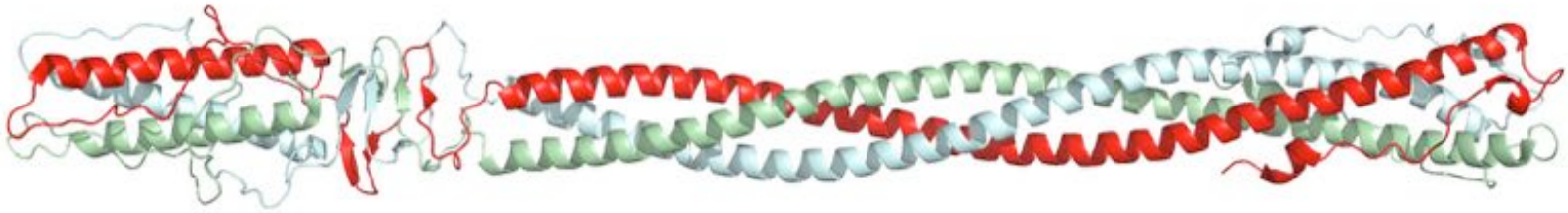
Proteins of known structure, clustered by sequence similarity and colored by structural class: all-α (blue), all-β (cyan), α/β (red), α+β (yellow), small proteins (green), multi-domain proteins (orange), and membrane proteins (magenta).

Tertiary structure - most proteins consist of multiple domains (particularly in eukaryotes)

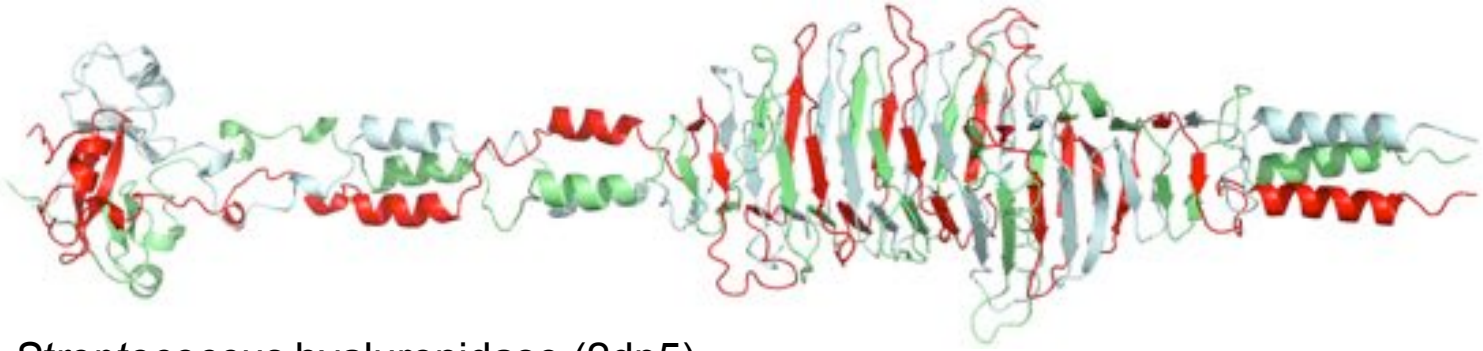


Tricorn protease (PDB 1k32)

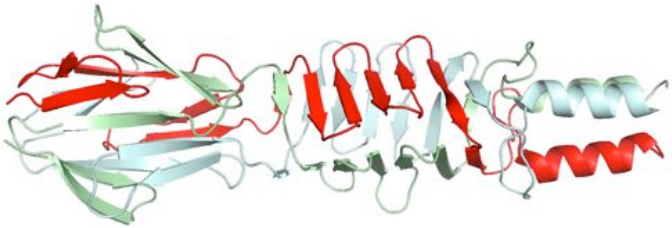
From secondary to quaternary structure - protein fibers



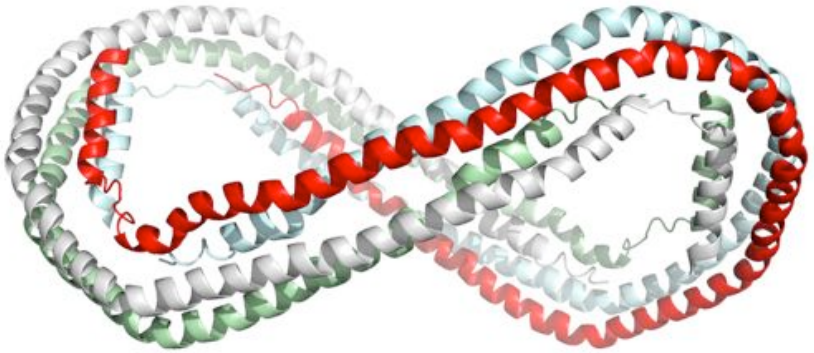
Salmonella phage P22 tail spike protein (2poh)



Streptococcus hyaluronidase (2dp5)

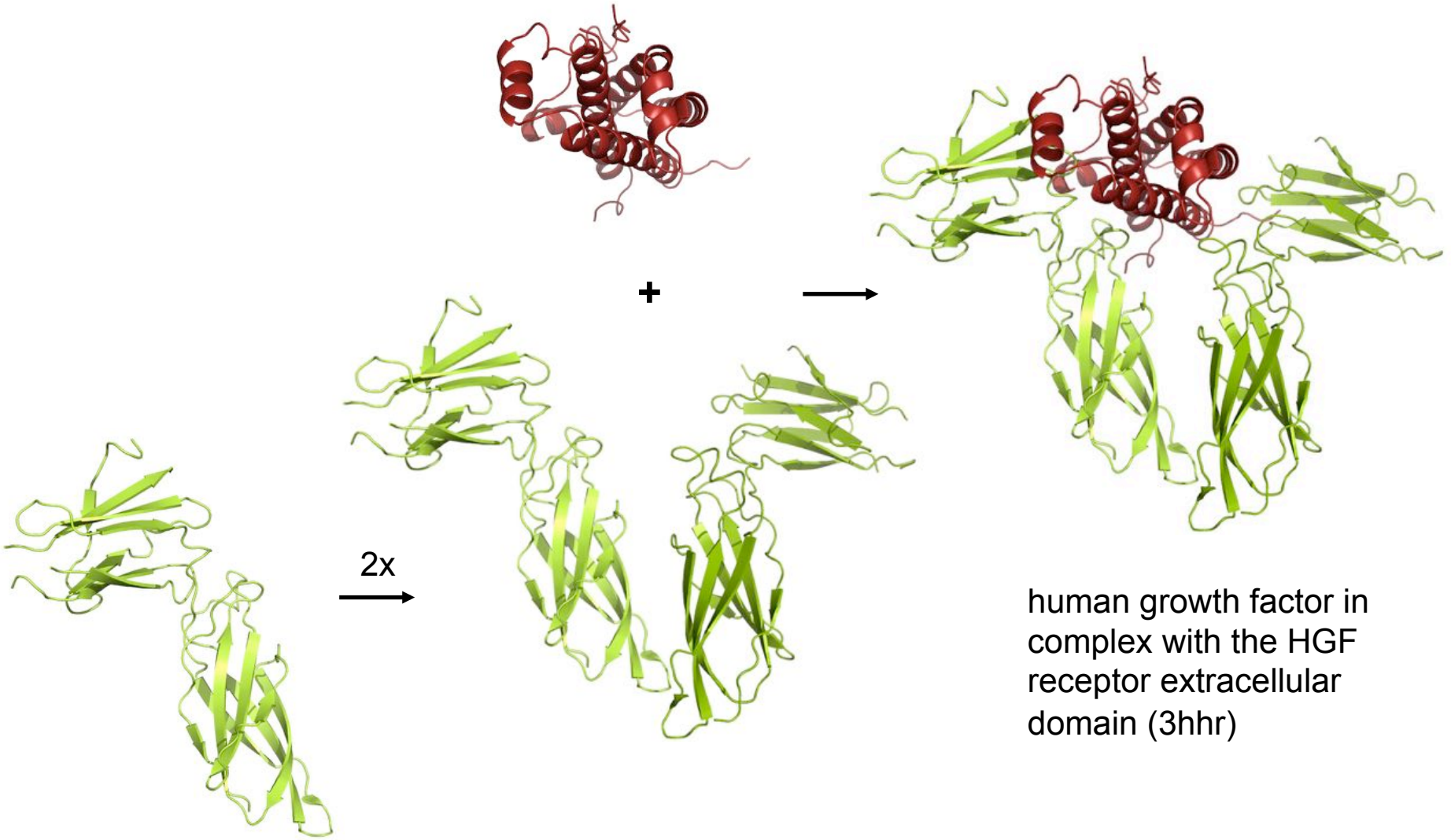


BadA adhesin head segment (3d9x)



apolipoprotein A-I (1av1)

Quaternary structure - homo- and hetero-oligomers



human growth factor in complex with the HGF receptor extracellular domain (3hhr)

Qaternary structure - complex assemblies