Structural Bioinformatics
Lecture 3

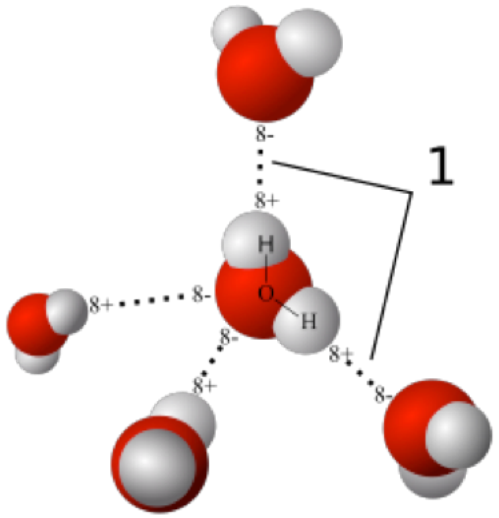
Protein structural organization.
Classification of proteins by structure

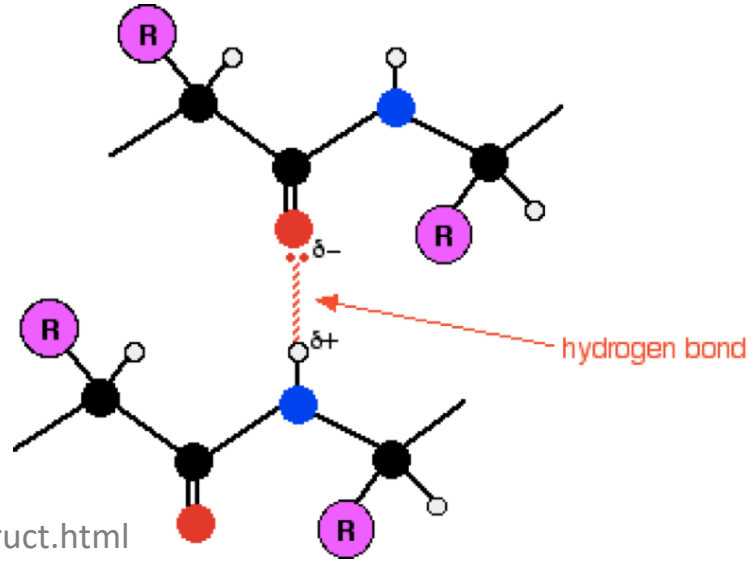
UNIVERSITÄT
DES
SAARLANDES

Four levels of organization of protein 3D structure

Hydrogen bonds in proteins

- Non-covalent interaction between polar molecules
- Electronegative-H ... electronegative
  - H is partially shared between the two molecules
- Hydrogen donor and acceptor
- 5...30 kJ/mol
- **Can form between different atoms in proteins**
  - **Importantly, also between mainchain N and O**

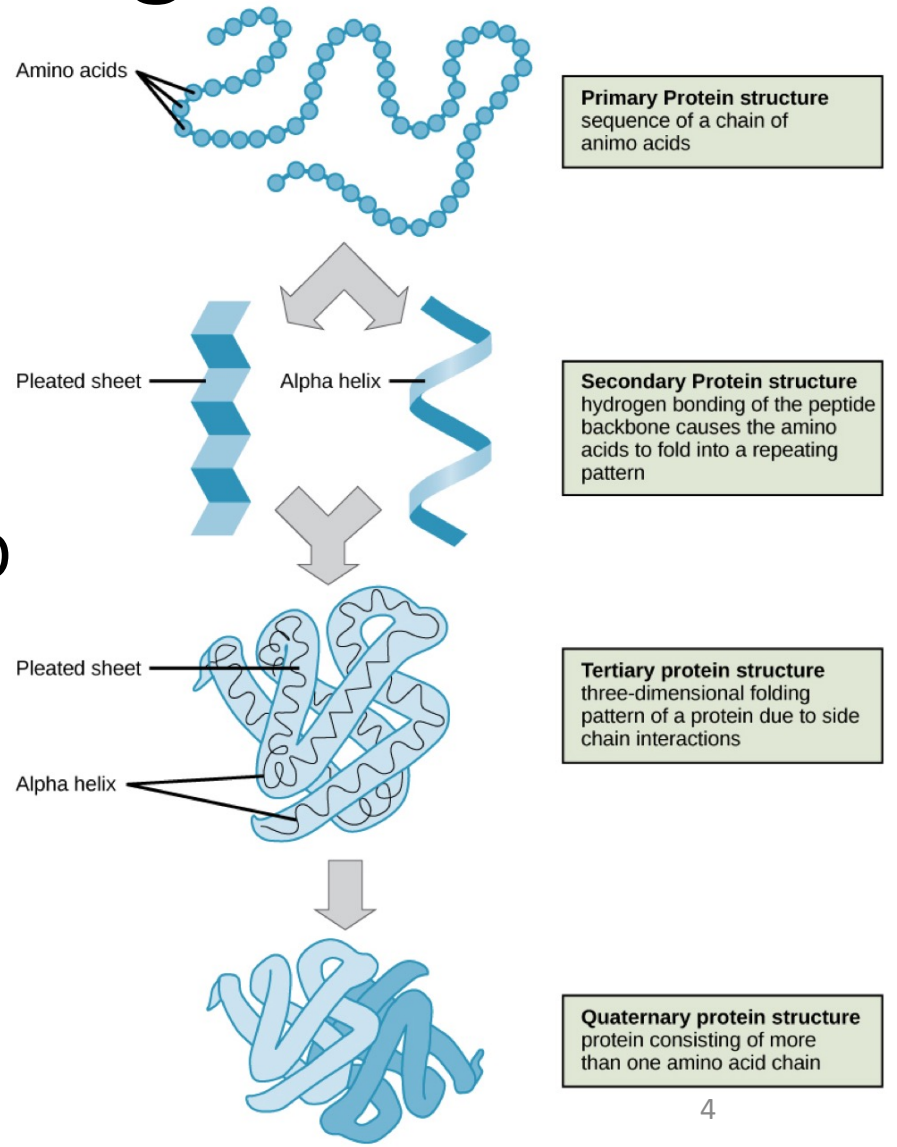




https://www.chemguide.co.uk/organicprops/aminoacids/proteinstruct.html

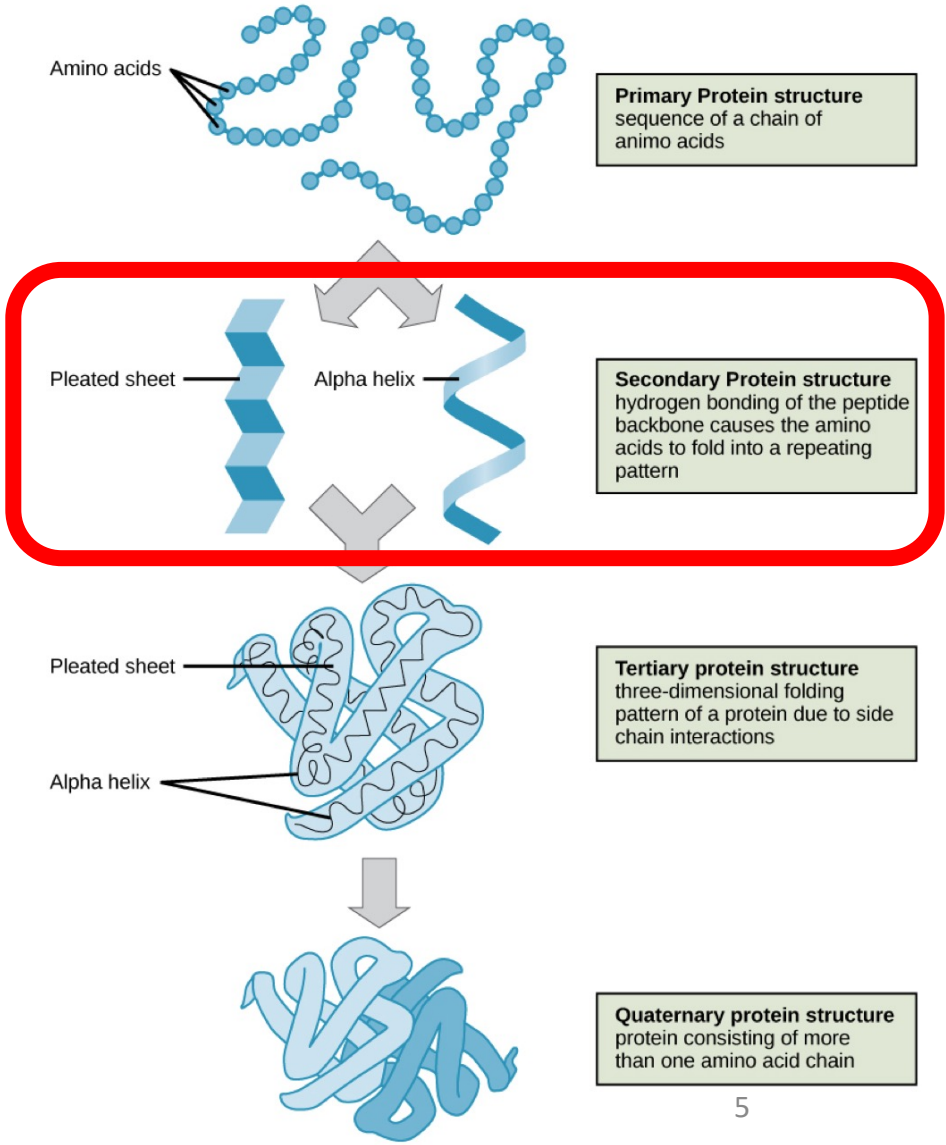
Four levels of protein structure organisation

- **Primary structure** (rarely used): sequence
- **Secondary structure**: local arrangement in 3D
  - α-helices and β-strands (and some other)
- **Tertiary structure**: arrangement of secondary structure elements relative to each other in 3D
- **Quaternary structure**: mutual arrangement of multiple (protein) chains to form a functional complex



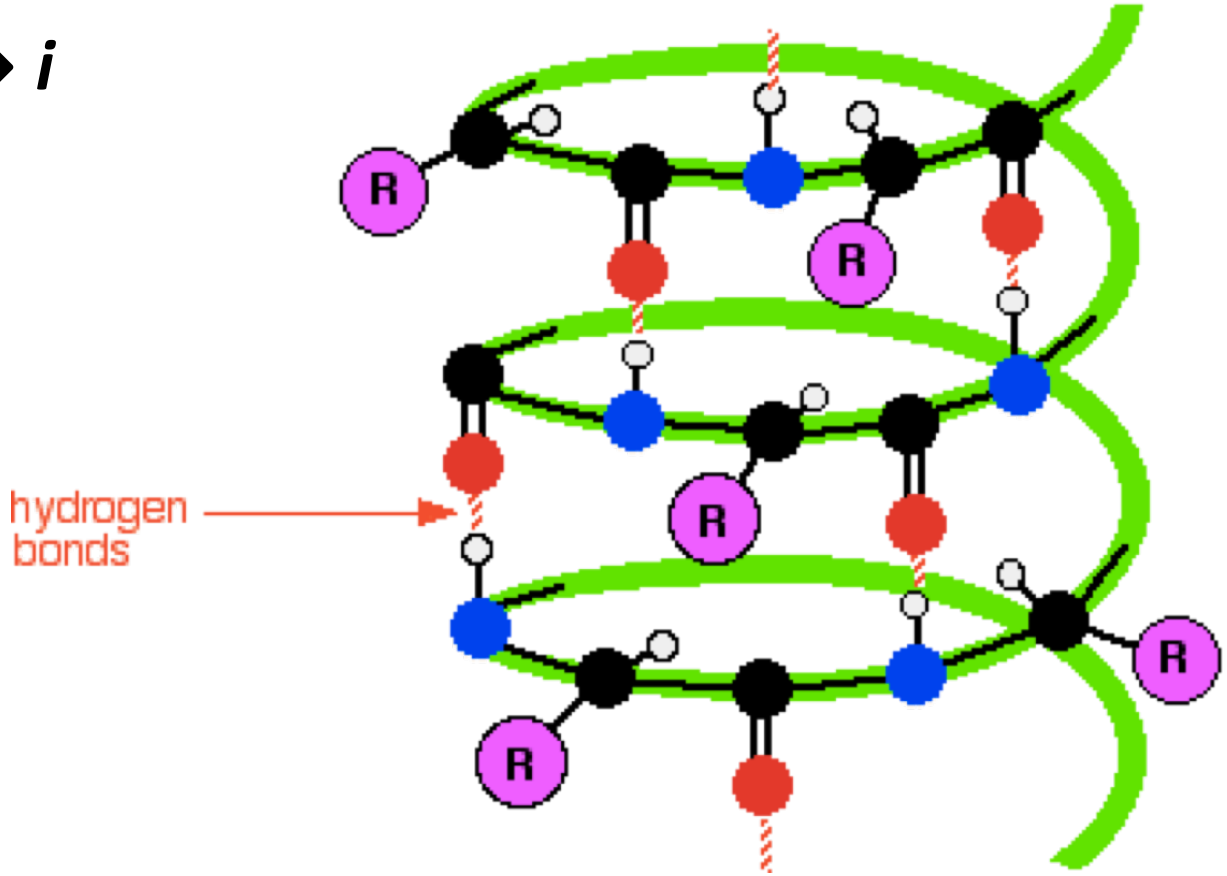
Secondary structure

- General 3D form of **local segments** of the polypeptide chain
- Supported by hydrogen bonds



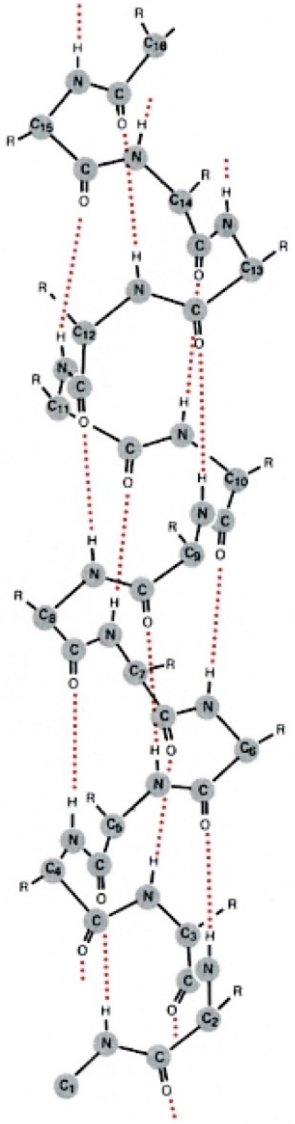
Secondary structure: helices

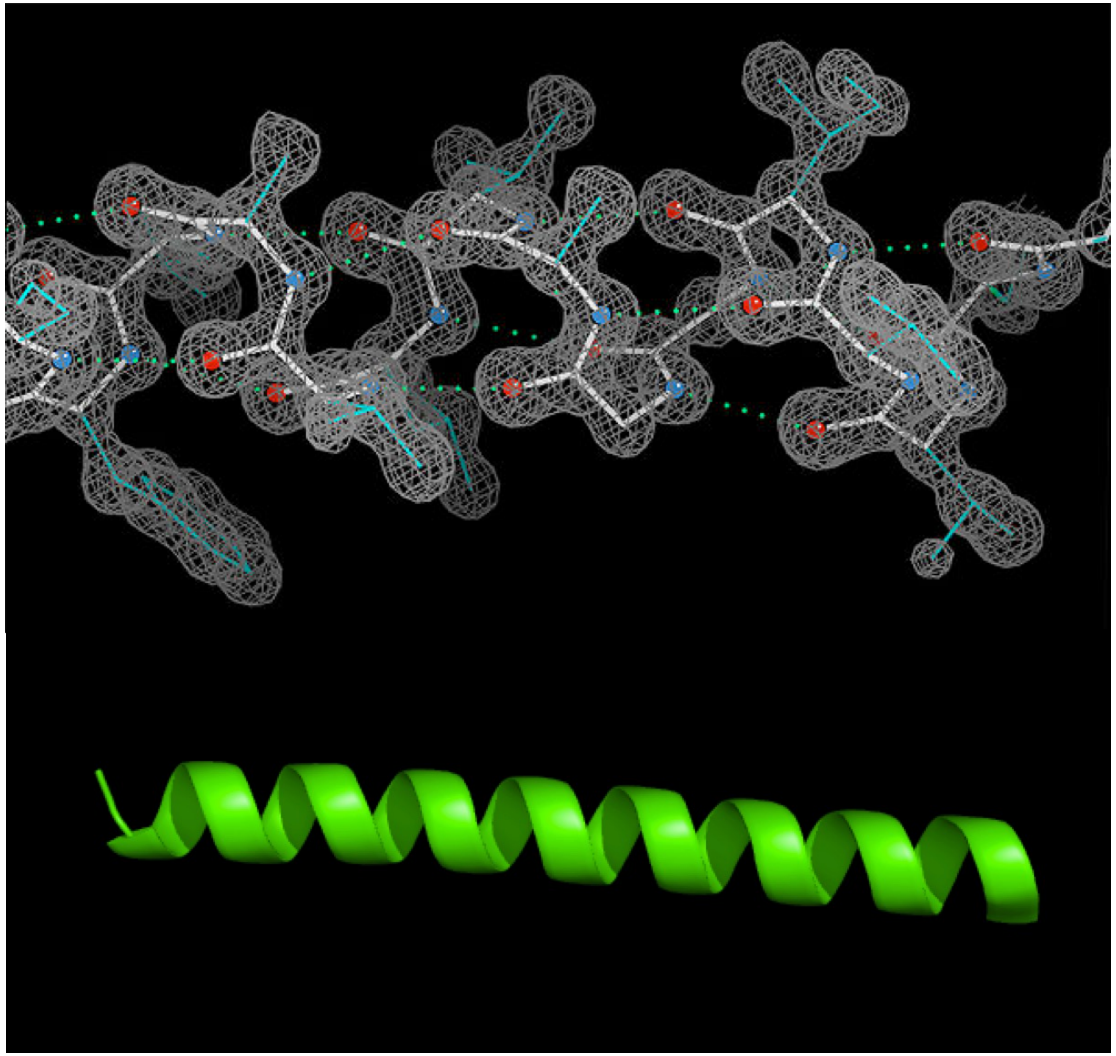
- Regilar H-bonding pattern: ***i* + *N* → *i***
- α-helix: *i* + 4 → *i*
  (3_{10}-helix: *i* + 3 → *i*;
  π-helix: *i* + 5 → *i*)



https://www.chemguide.co.uk/organicprops/aminoacids/proteinstruct.html

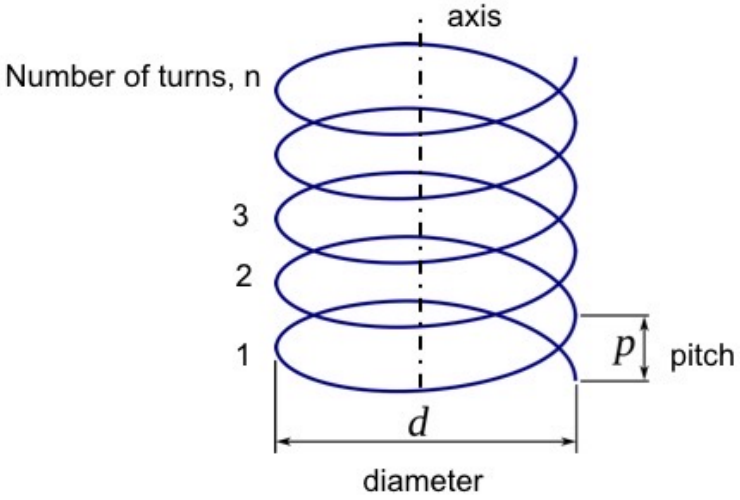
α-helices: ways of representation

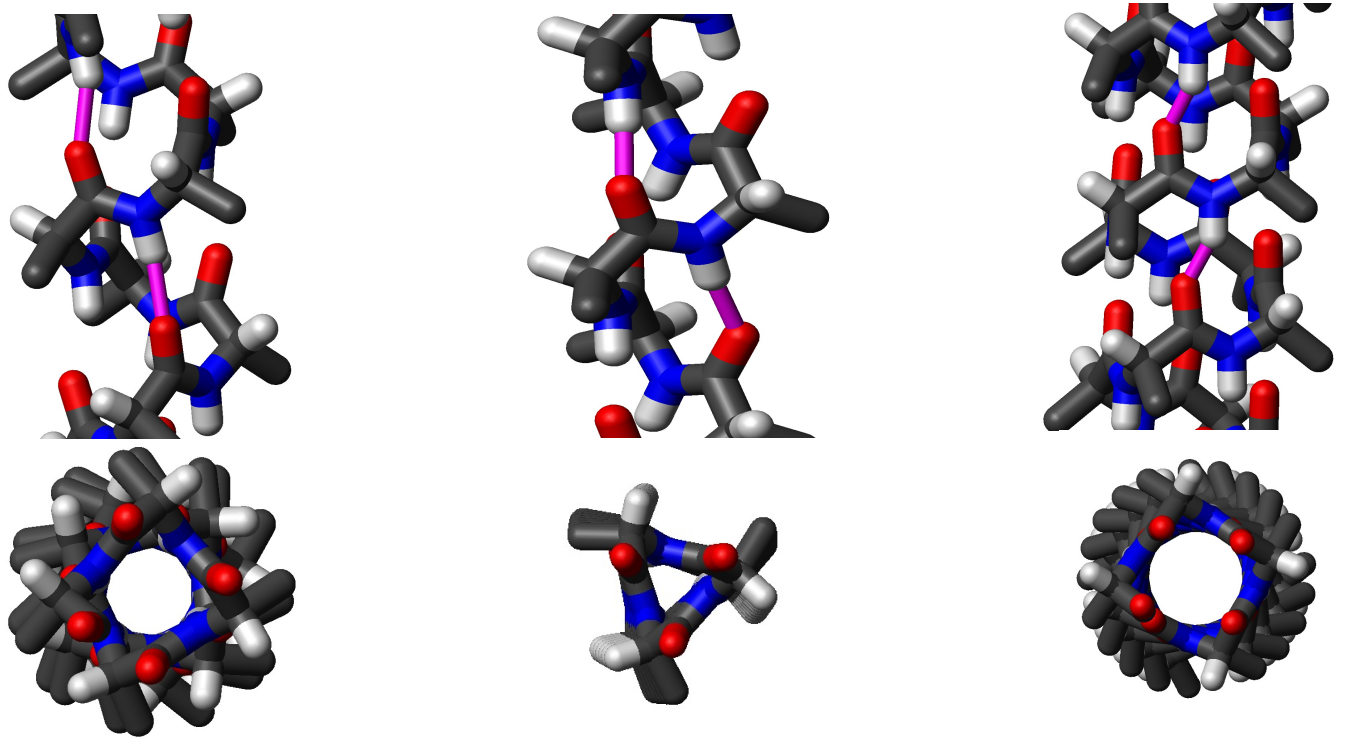




Comparison of helical structures

Geometry attribute	α-helix	3_{10} helix	π-helix
Residues per turn	3.6	3.0	4.4
Translation per residue	1.5 Å (0.15 nm)	2.0 Å (0.20 nm)	1.1 Å (0.11 nm)
Radius of helix	2.3 Å (0.23 nm)	1.9 Å (0.19 nm)	2.8 Å (0.28 nm)
Pitch	5.4 Å (0.54 nm)	6.0 Å (0.60 nm)	4.8 Å (0.48 nm)



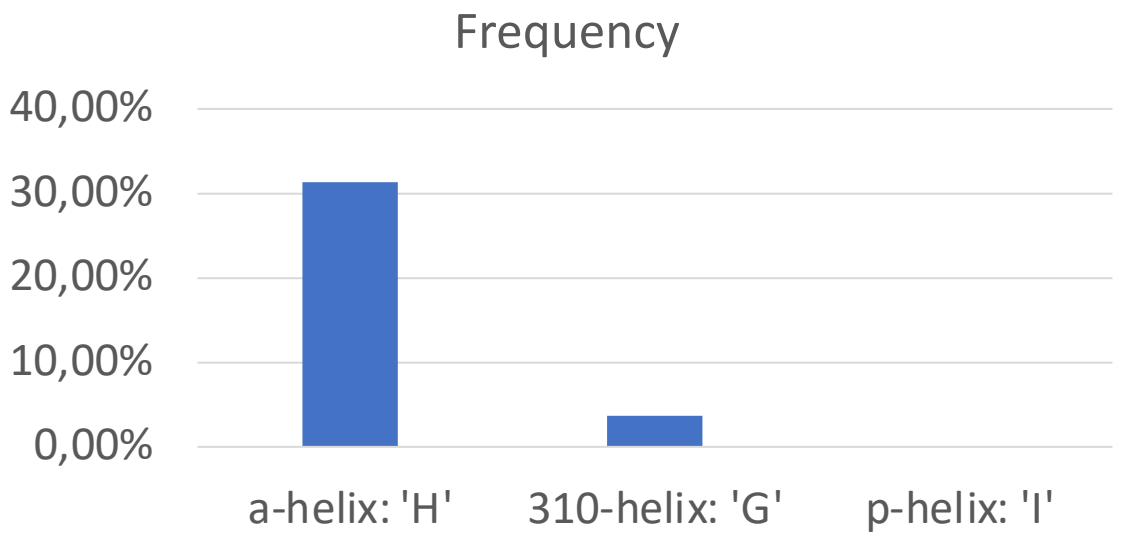


Side view

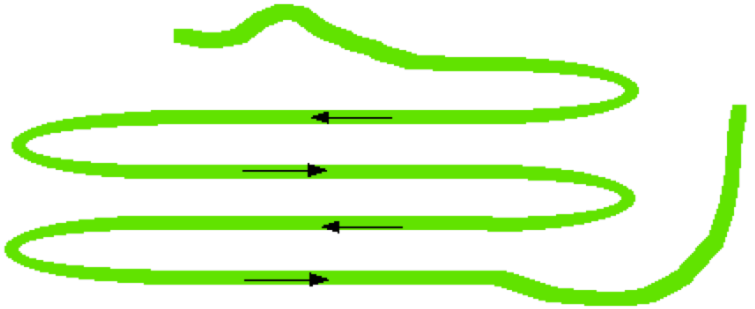
Top view

Comparison of helical structures

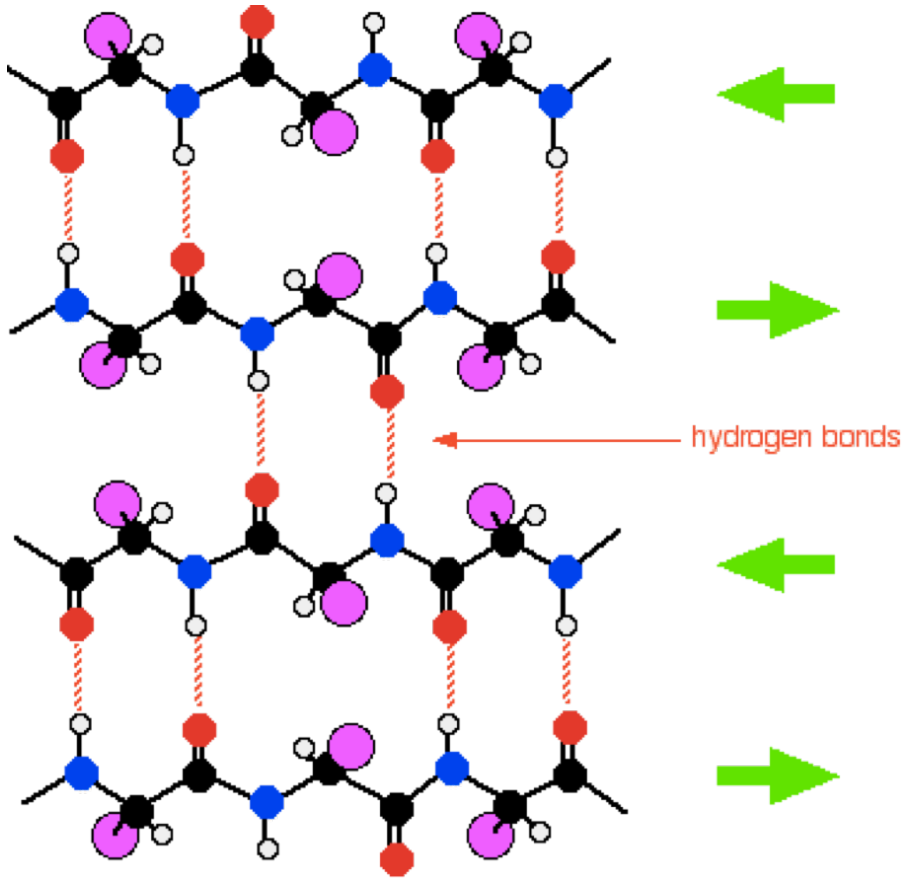
	a-helix: 'H'	3_{10}-helix: 'G'	p-helix: 'I'
Frequency	31.3%	3.7%	0.04%
Average length	11.2 residues	3.4 residues	5.2 residues



Secondary structure: β-sheets

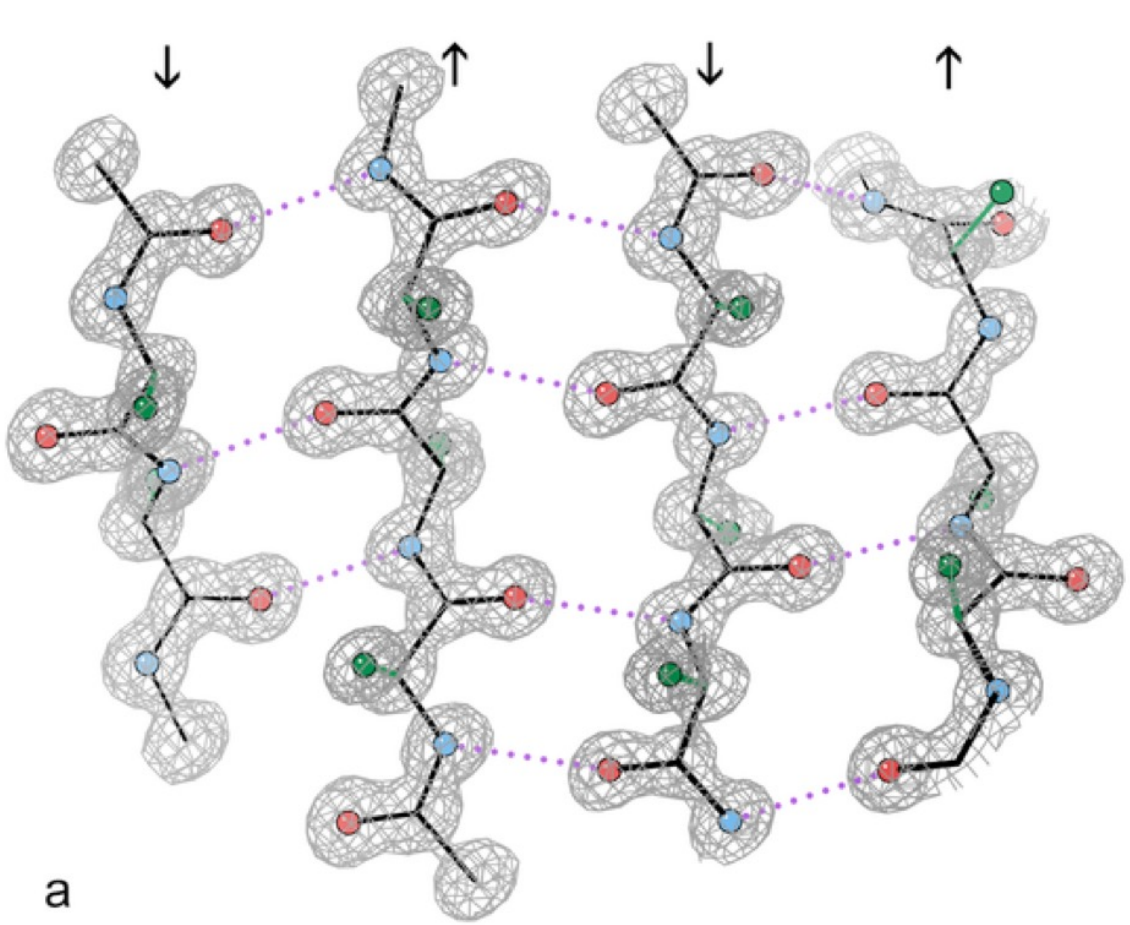


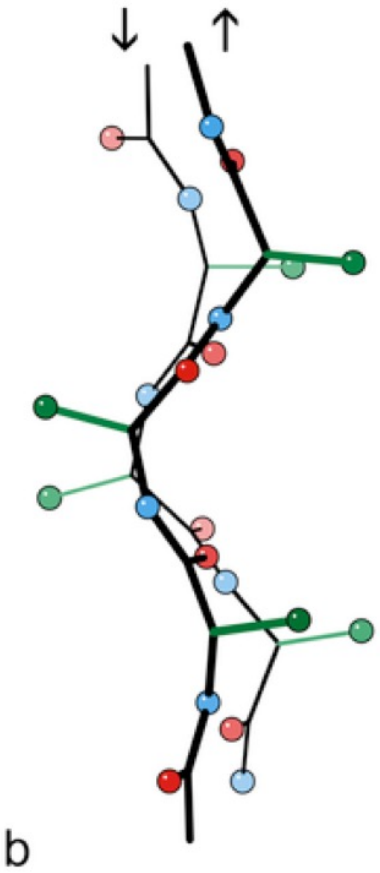
- β-sheet: consists of β-strands
- Parallel and anti-parallel
- Connected by hydrogen bonds



https://www.chemguide.co.uk/organicprops/aminoacids/proteinstruct.html

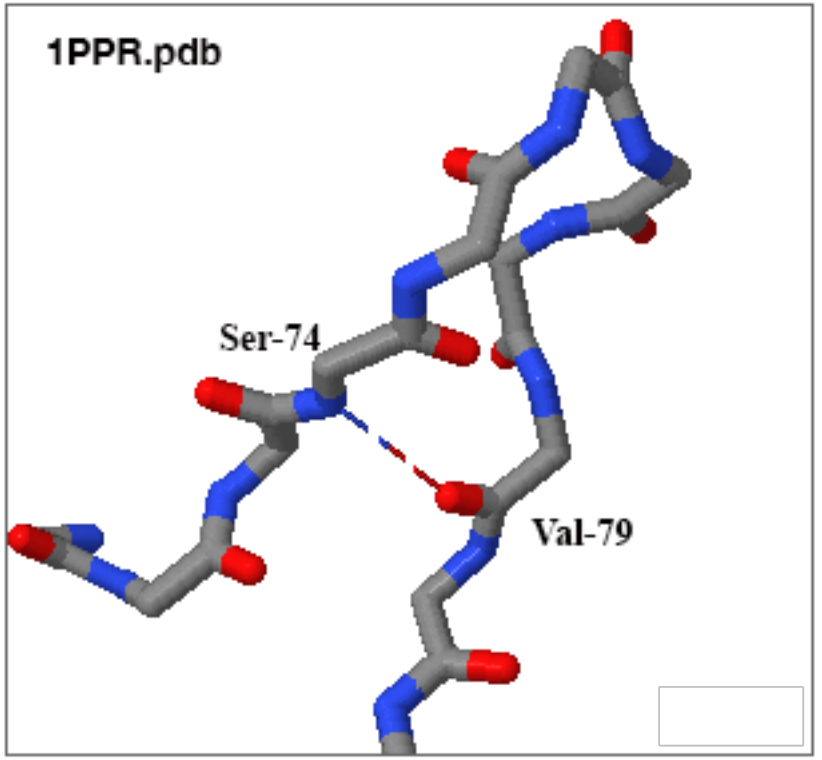
β-sheets: ways of representation





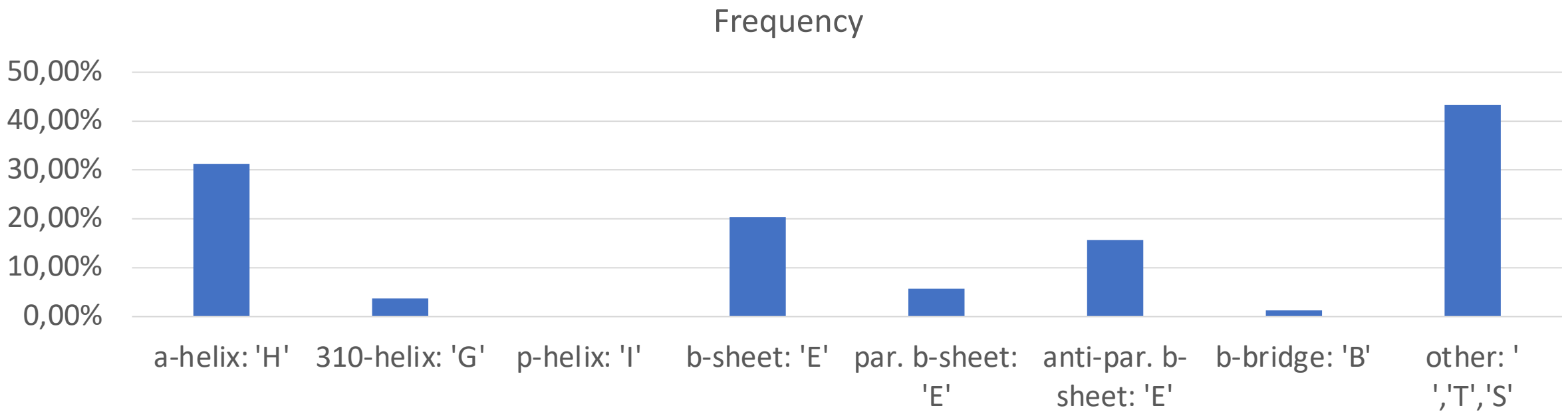
Secondary structure: coils and others

- **Coils**: regions without defined secondary structure (turns, bends, simply unstructured)
- **β-bridges**: two isolated residues bound with a H-bond in the same geometry as in β-sheets

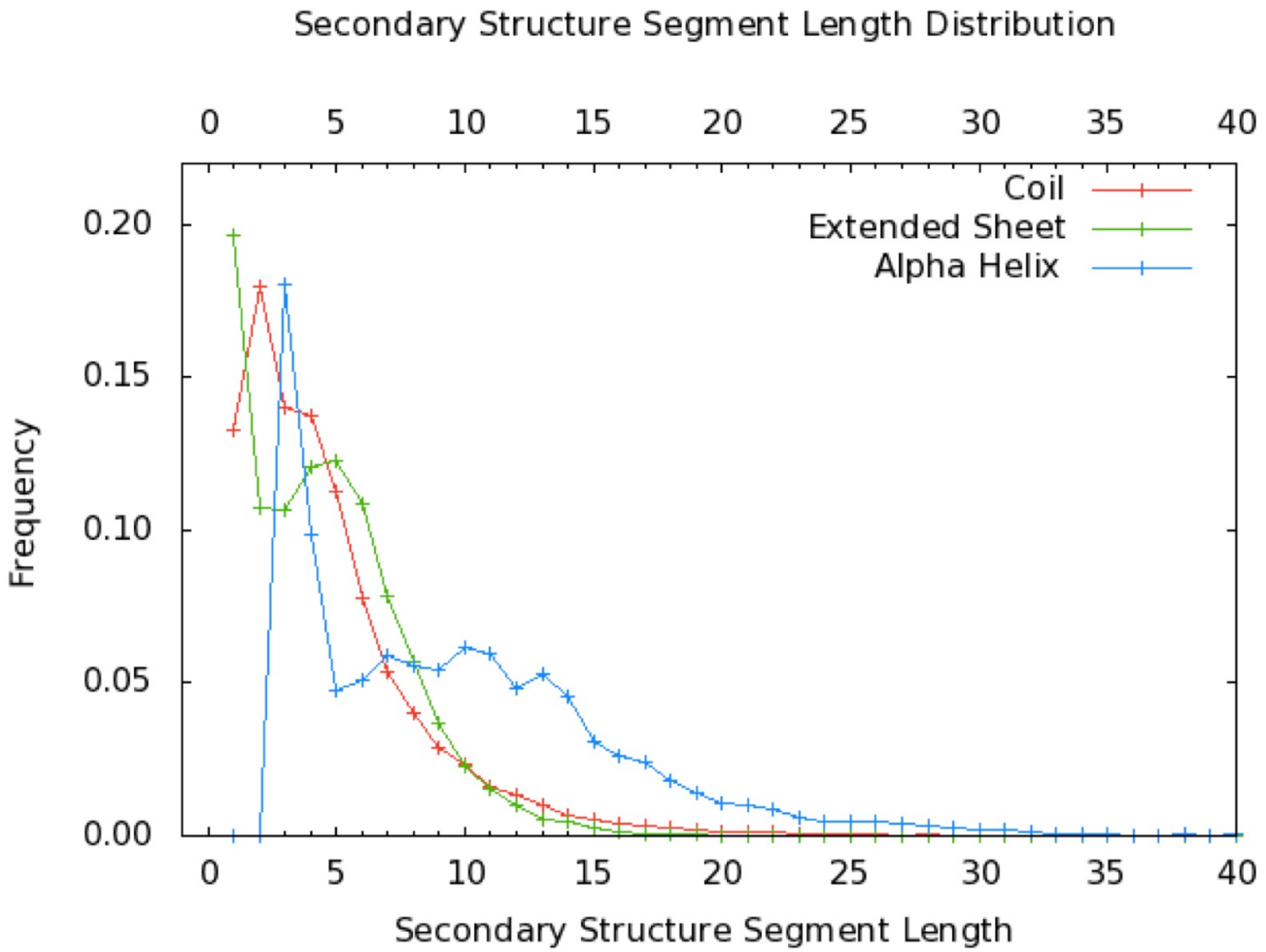


Comparison of secondary structure elements

	a-helix: 'H'	3_{10}-helix: 'G'	p-helix: 'I'	b-sheet: 'E'	par. b-sheet	anti-par. b-sheet	b-bridge: 'B'	other
Frequency	31,30%	3,70%	0,04%	20,40%	5,70%	15,70%	1,30%	43,30%
Average length	11.2 residues	3.4 residues	5.2 residues	4.4 residues	4.0 residues	4.6 residues	1 res. per definition	6.8 residues



Length of secondary structure elements



Secondary structure assignment: DSSP

- H = alpha helix
- B = residue in isolated beta-bridge
- E = extended strand, participates in beta ladder
- G = 3-helix (3/10 helix)
- I = 5 helix (pi helix)
- T = hydrogen bonded turn
- S = bend
- C/- = no secondary structure (coil)

#	RESIDUE	AA	STRUCTURE	BP1	BP2	ACC	N-H-->O	O-->H-N	N-H-->O	O-->H-N	TCO	KAPPA	ALPHA	PHI	PSI	X-CA	Y-CA	Z-CA
1	0	A	M		0	0	87	0, 0.0	2,-0.3	0, 0.0	140,-0.0	0.000	360.0	360.0	360.0	158.1	24.4	9.9	-9.9
2	1	A	V	-	0	0	123	0, 0.0	2,-0.3	0, 0.0	79,-0.1	-0.999	360.0	-139.4	-144.9	137.5	27.3	11.9	-11.4
3	2	A	L	-	0	0	12	-2,-0.3	2,-0.1	77,-0.1	128,-0.0	-0.700	21.5	-123.2	-95.0	149.8	28.8	15.3	-10.4
4	3	A	S	> -	0	0	64	-2,-0.3	4,-2.2	1,-0.1	5,-0.2	-0.395	28.4	-108.3	-82.7	166.7	29.9	17.9	-12.9
5	4	A	E	H > S+	0	0	101	1,-0.2	4,-2.4	2,-0.2	5,-0.2	0.881	121.5	58.3	-62.5	-36.3	33.5	19.2	-12.9
6	5	A	G	H > S+	0	0	46	2,-0.2	4,-1.6	1,-0.2	-1,-0.2	0.907	106.9	46.3	-59.6	-42.0	32.1	22.5	-11.5
7	6	A	E	H > S+	0	0	51	2,-0.2	4,-2.1	1,-0.2	-2,-0.2	0.910	111.4	51.2	-68.7	-40.2	30.7	20.6	-8.5
8	7	A	W	H X S+	0	0	15	-4,-2.2	4,-2.9	1,-0.2	5,-0.2	0.872	107.4	54.7	-63.5	-36.8	34.0	18.7	-8.0
9	8	A	Q	H X S+	0	0	98	-4,-2.4	4,-2.0	2,-0.2	-1,-0.2	0.897	107.3	49.3	-64.8	-38.5	35.9	22.0	-8.1
10	9	A	L	H X S+	0	0	64	-4,-1.6	4,-1.6	2,-0.2	5,-0.2	0.926	113.6	47.0	-64.8	-42.1	33.7	23.4	-5.3
11	10	A	V	H X S+	0	0	0	-4,-2.1	4,-2.1	1,-0.2	-2,-0.2	0.952	115.2	44.2	-63.8	-50.5	34.2	20.3	-3.2

Secondary structure synopsis

H-bond energy

Relative index of the partner

Angle between C=O's

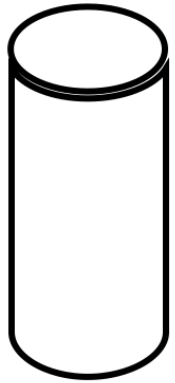
Bend angle

Summary: how SSEs are held together

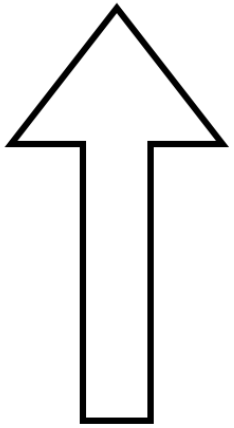
- H-bonds
- Between main chain atoms
- In proteins, 90% of backbone C=O and NH groups participate in H-bonds (62% in intra-backbone)

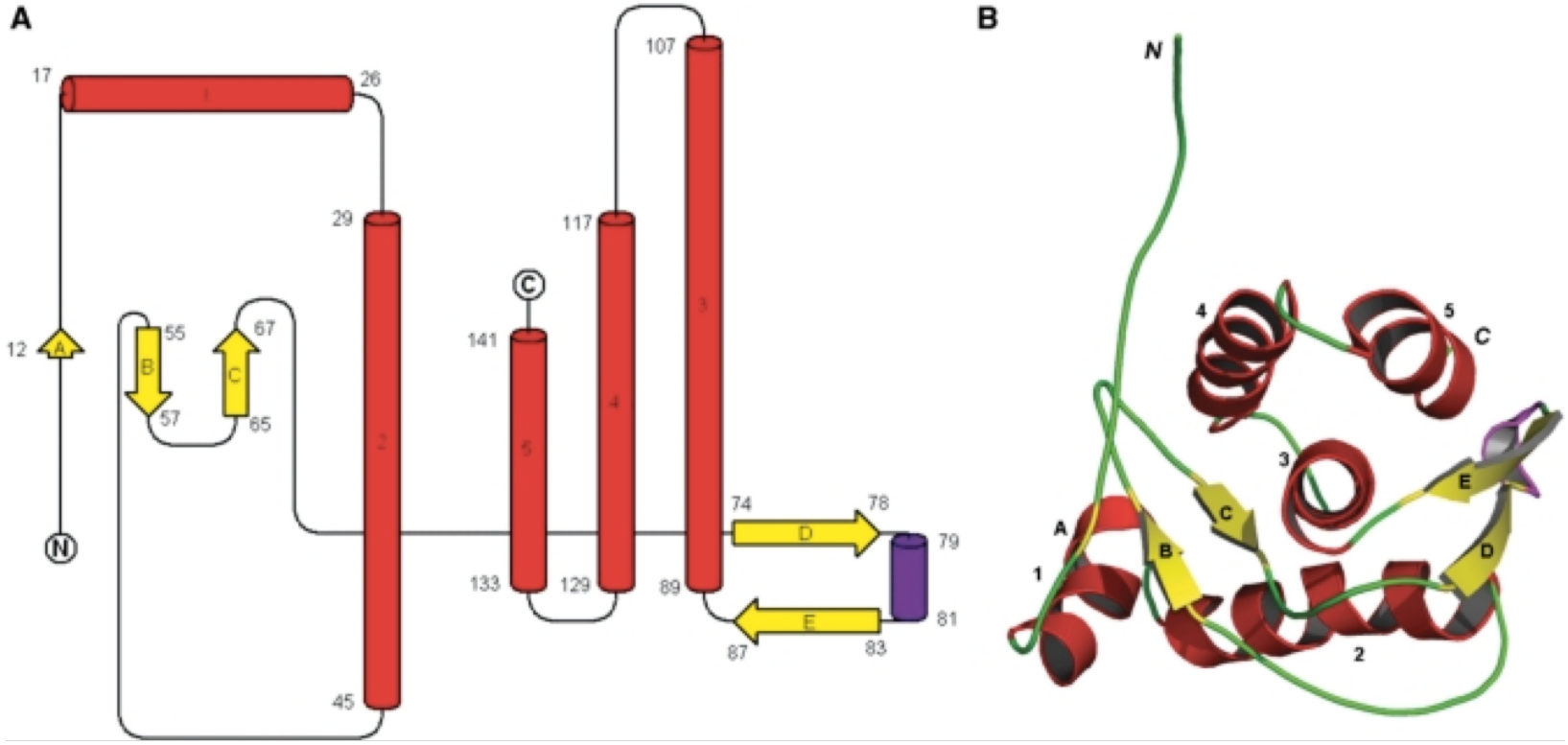
SSE schematic representation: topology diagram

α-helix:



β-strand:

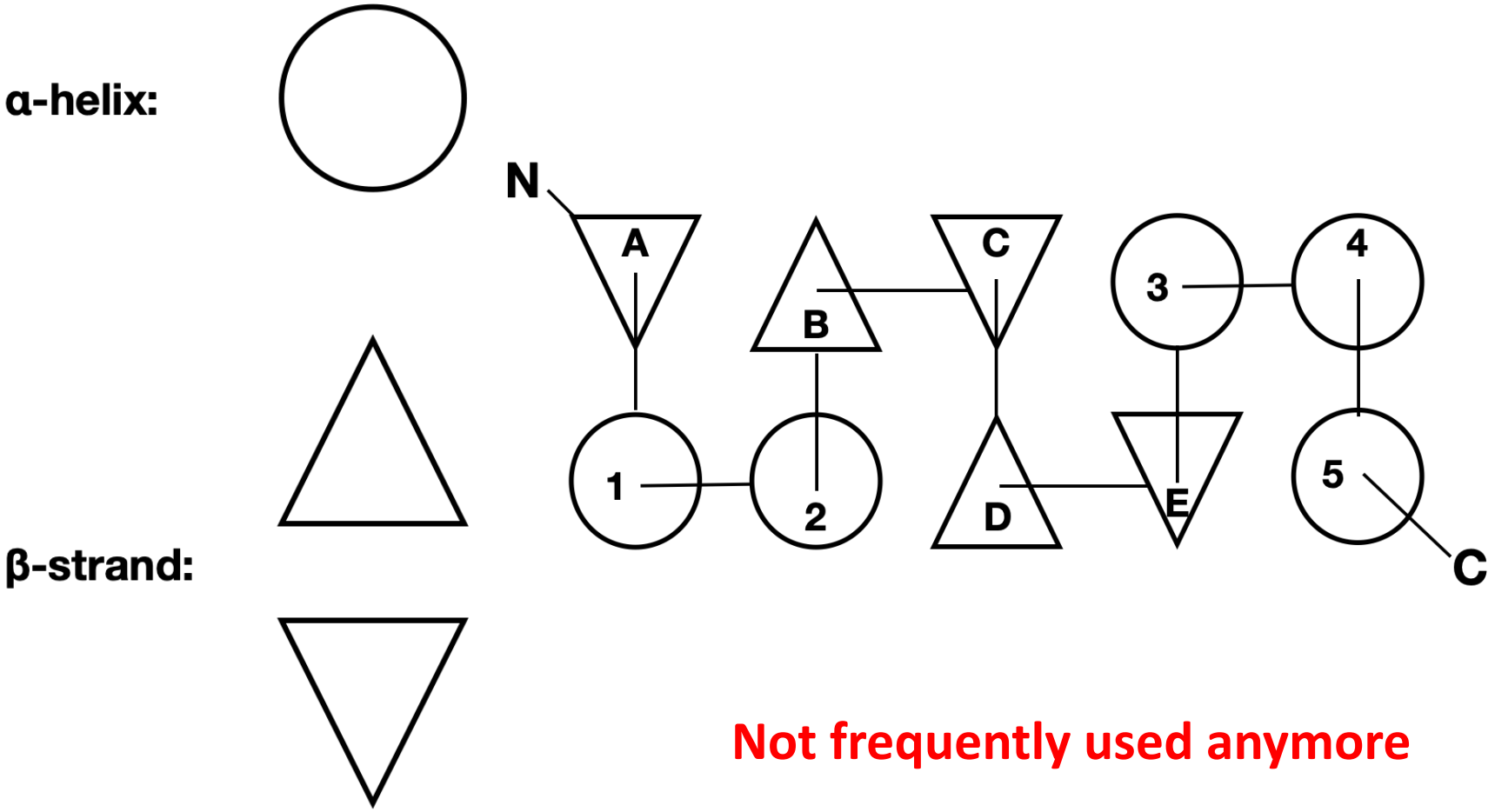


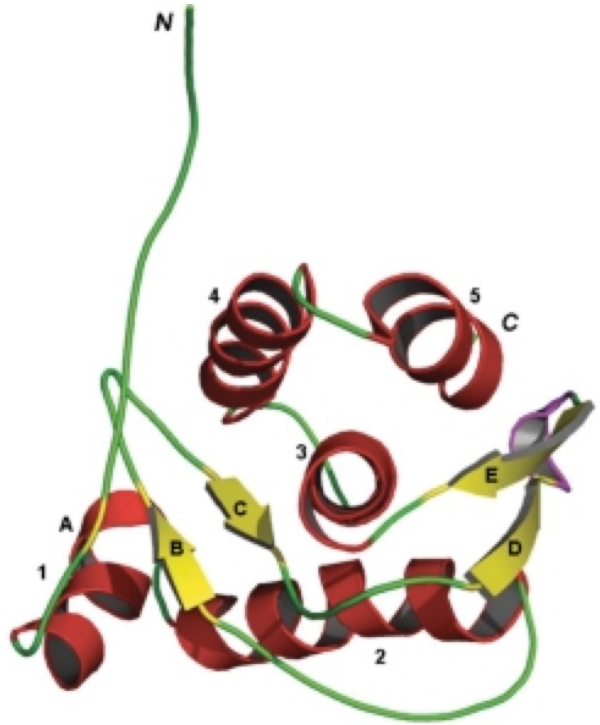


Serfiotis-Mitsa et al., 2010

Can be created using specialized software, e.g.
http://munk.csse.unimelb.edu.au/pro-origami/

SSE schematic representation: TOPS diagram

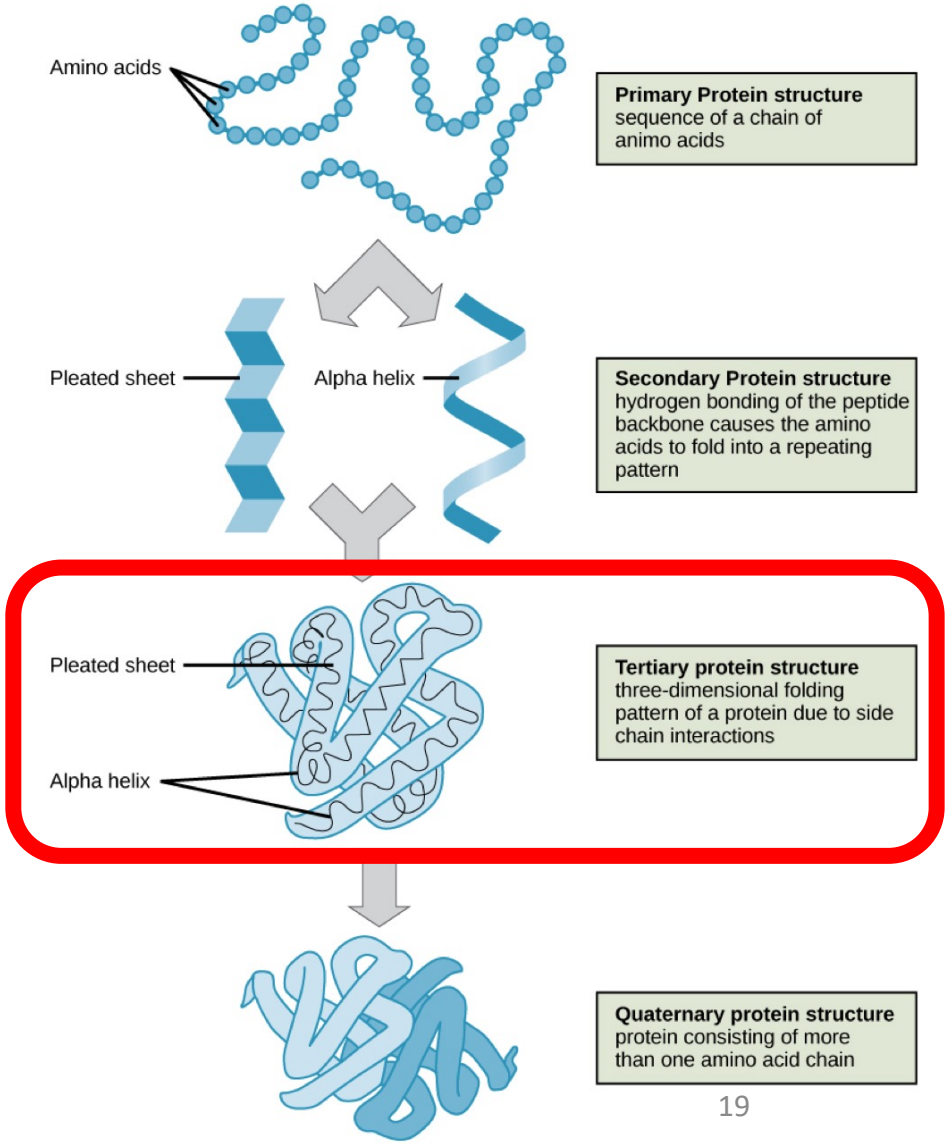




Not frequently used anymore

Tertiary structure

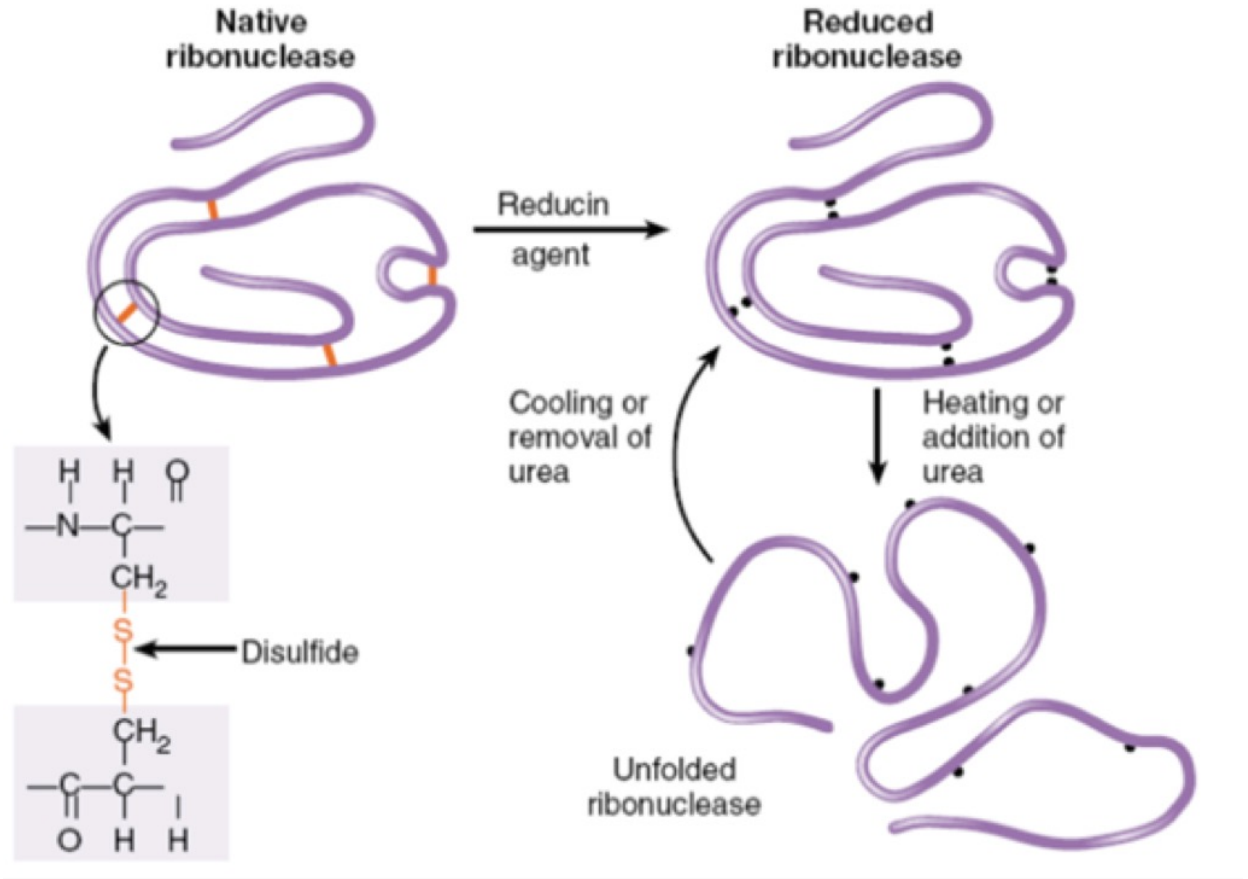
- Mutual arrangement of secondary structure elements of one chain
- Unique shape of a protein
- Carries a function



Anfinsen dogma

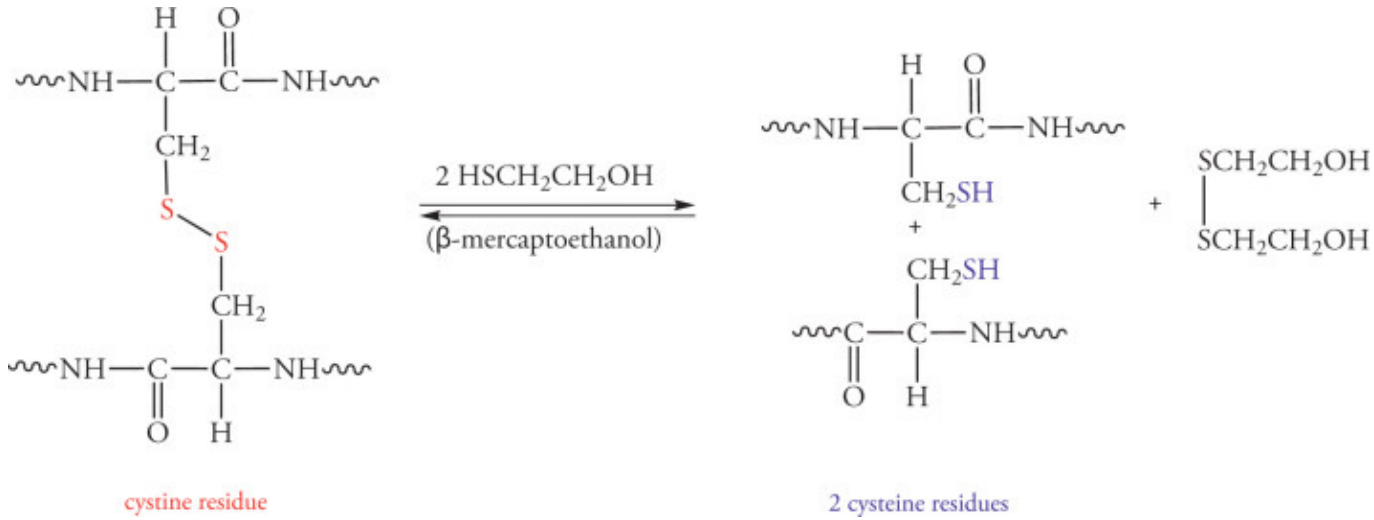
- At least for small globular proteins, at the environmental conditions at which folding occurs, the native structure is a unique, stable and kinetically accessible minimum of the free energy. (*Christian B. Anfinsen*)
  - **Uniqueness**: the sequence does not have any other configuration with a comparable free energy
  - **Stability**: small changes in the surrounding environment cannot give rise to changes in the minimum configuration
  - **Kinetic accessibility**: the path in the free energy surface from the unfolded to the folded state must be reasonably smooth

Anfinsen dogma: folding and unfolding of ribonuclease



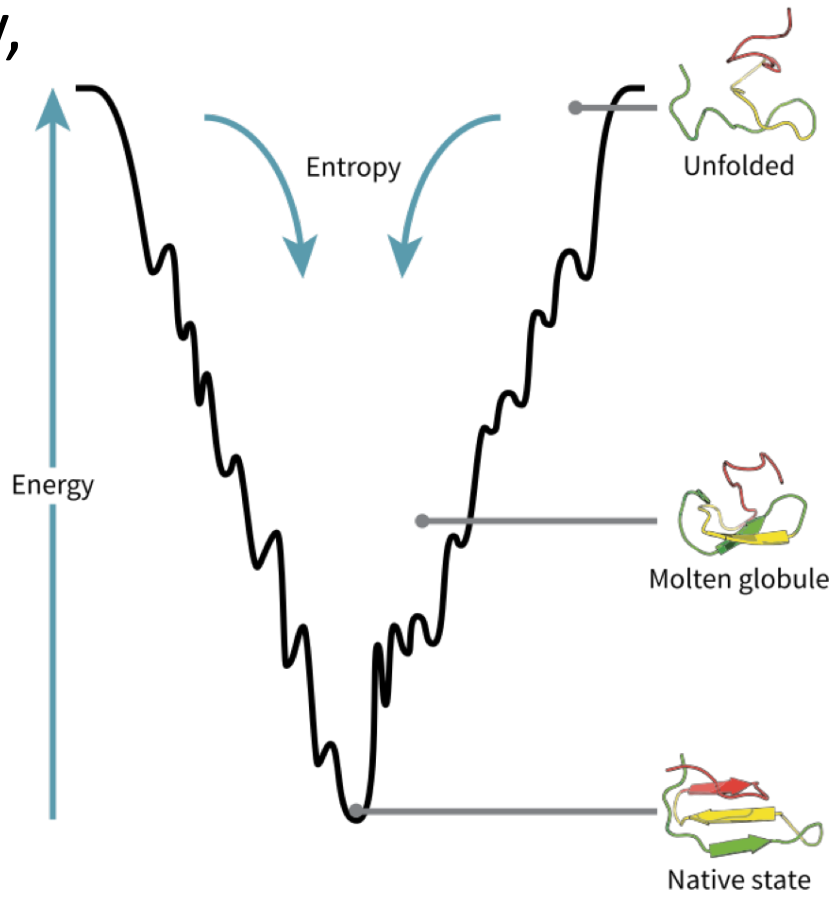
Disulfide bonds

- Can form between two **cysteine** residues, forming **cystine**
- Reversible
- In oxidizing conditions
- Unstable in cytosol (as a rule)
  - => in eukaryotes often found in secreted proteins
- Stabilize protein fold
- **But: most proteins assume a stable structure without disulfide bonds**



Free energy landscape

- Simple case: protein can unfold and refold rapidly, reversibly, two-state mechanism
- Gibbs free energy *G*: the **maximum** amount of work that can be extracted from a thermodynamically closed system at constant temperature and volume
- Each state of the protein is characterised by a certain *G*
- $\Delta G = G_{unfolded} - G_{folded}$
- ΔG is directly related to protein melting temperature



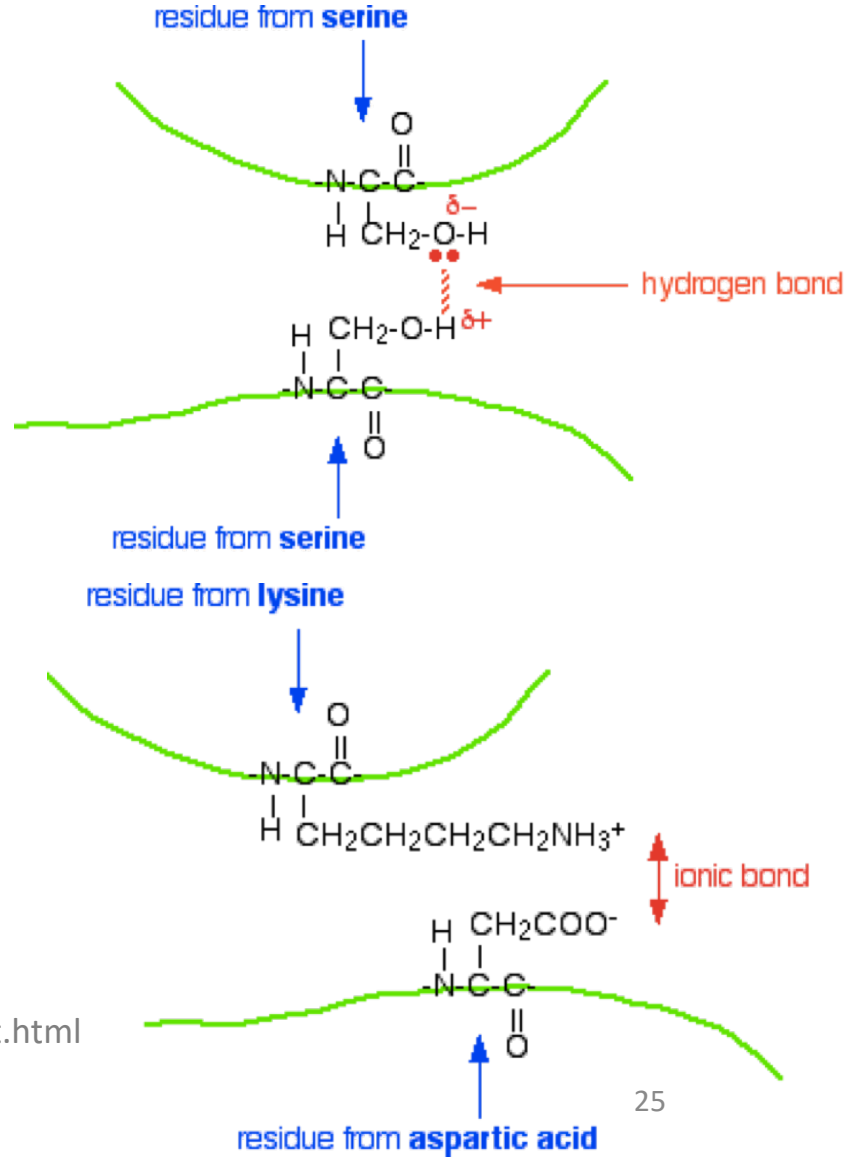
https://commons.wikimedia.org/w/index.php?curid=28353539

Folding to tertiary structure

- Spontaneous
- Assisted by **chaperones**
- ↑temperature; ↑↓pH => “Un-Folding” — **Denaturation**

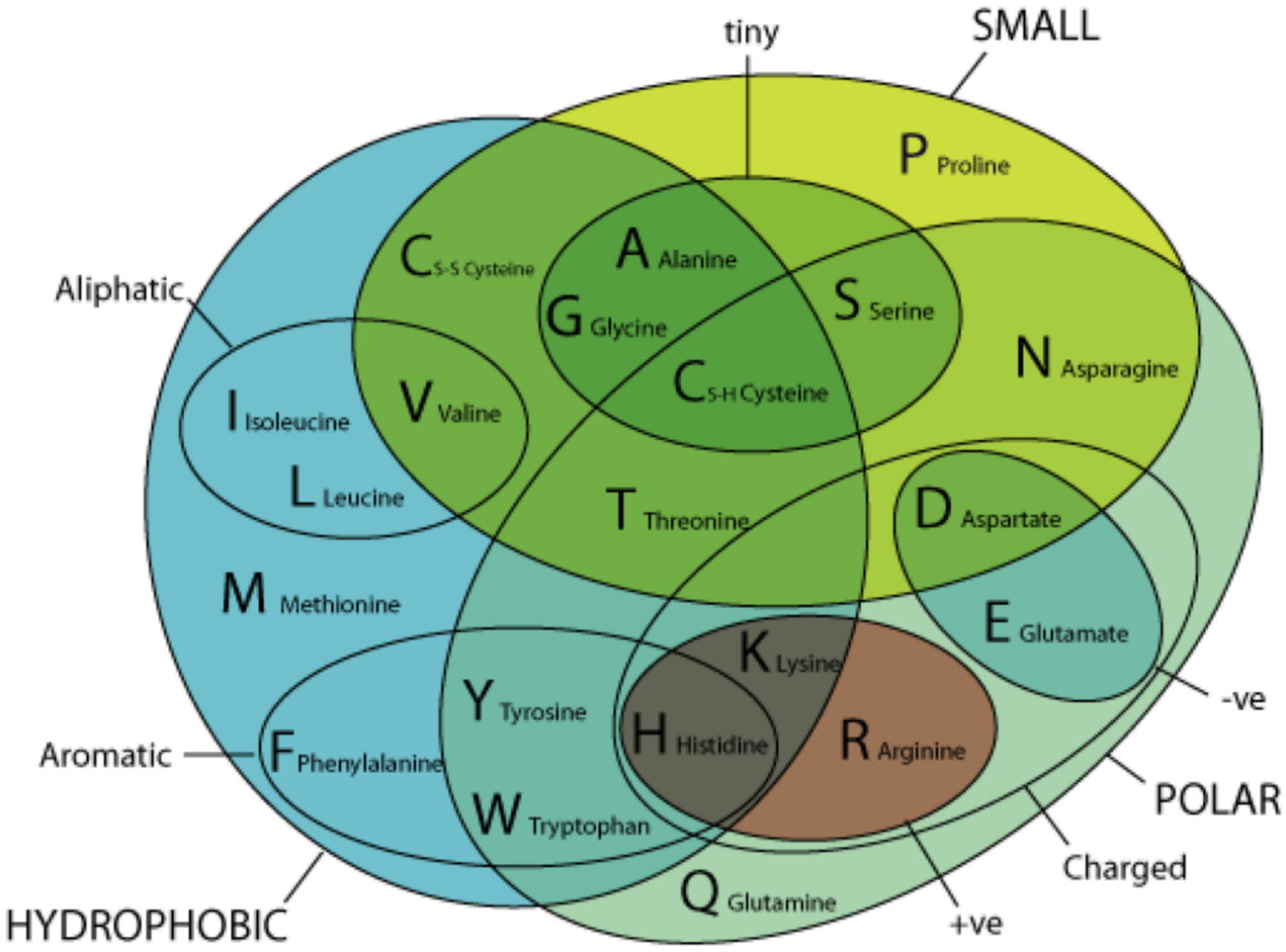
Holding tertiary structure together

- Hydrogen bonds
- Ionic interactions (salt bridges)
- Van der Waals forces
- Disulfide bonds
- **Act on residues (side chains)!**



https://www.chemguide.co.uk/organicprops/aminoacids/proteinstruct.html

Amino acids have different chemical properties



Learn them!

H-bond donors and acceptors in amino acid side chains

- 3 amino acids (arginine, lysine and tryptophan): hydrogen donor atoms
- 2 amino acids (aspartic acid, glutamic acid): hydrogen acceptor atoms
- 6 amino acids (asparagine, glutamine, histidine, serine, threonine and tyrosine): both hydrogen donor and acceptor atoms

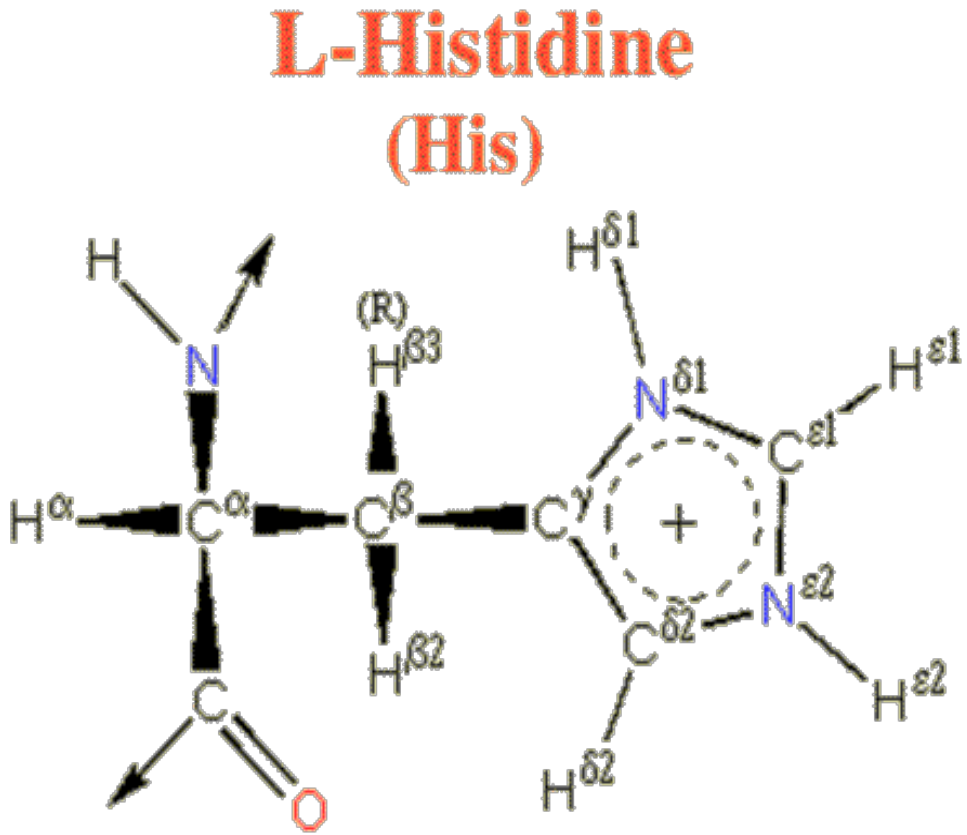
Amino acids	Hydrogen donor atoms	Hydrogen acceptor atoms
Arginine (Arg, R)	NE, NH1 (2), NH2 (2)	
Asparagine (Asn, N)	ND2 (2)	OD1 (2)
Aspartic acid (Asp, D)		OD1 (2), OD2 (2)
Glutamine (Gln, Q)	NE2 (2)	OE1 (2)
Glutamic acid (Glu, E)		OE1 (2), OE2 (2)
Histidine (His, H)	ND1, NE2	ND1, NE2
Lysine (Lys, K)	NZ (3)	
Serine (Ser, S)	OG	OG (2)
Threonine (Thr, T)	OG1	OG1 (2)
Tryptophan (Trp, W)	NE1	
Tyrosine (Tyr, Y)	OH	OH

(number of hydrogens, if >1)

http://www.imgt.org/IMGTeducation/Aide-memoire/_UK/aminoacids/charge/#hydrogen

A detour: atom naming conventions

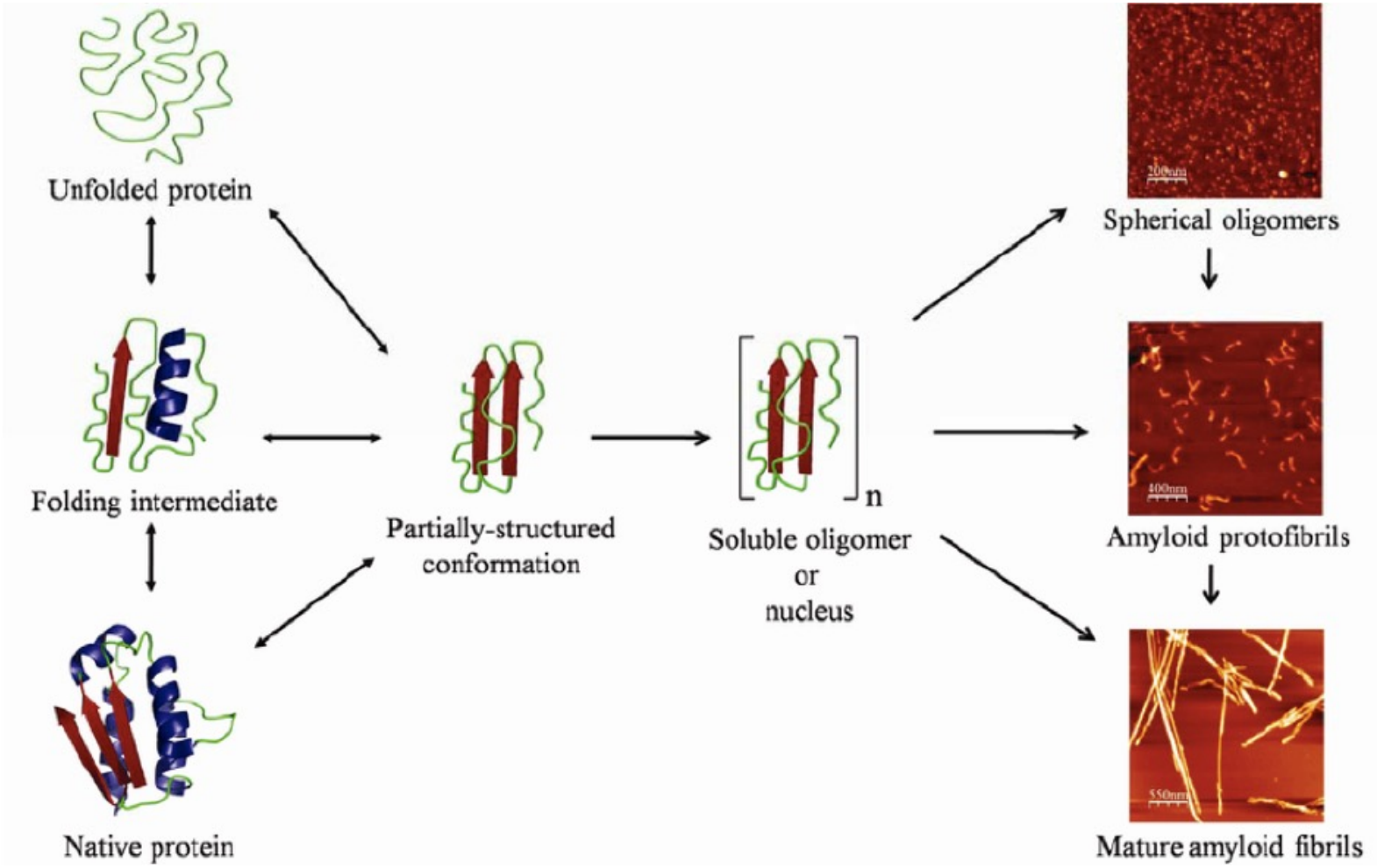
- Main chain: N, Cα, C, O
- Side chain: remoteness from the main chain, Greek letters
  - β, γ, δ, ε, ζ, η
- In PDB file (and often in the literature):
  - α → A, β → B, γ → G, δ → D, ε → E, ζ → Z, η → H
- + numerical branch indicator
- => e.g. in His CD1, ND2



Visit https://bmrb.io/referenc/commonaa.php to inspect the other

Folding defects can cause diseases

- **Alzheimer**: Amyloid fibrils in nerve cells
- "**Mad cow disease**" (bovine spongiform encephalopathy), **Creutzfeld Jakob disease**: **Prions**

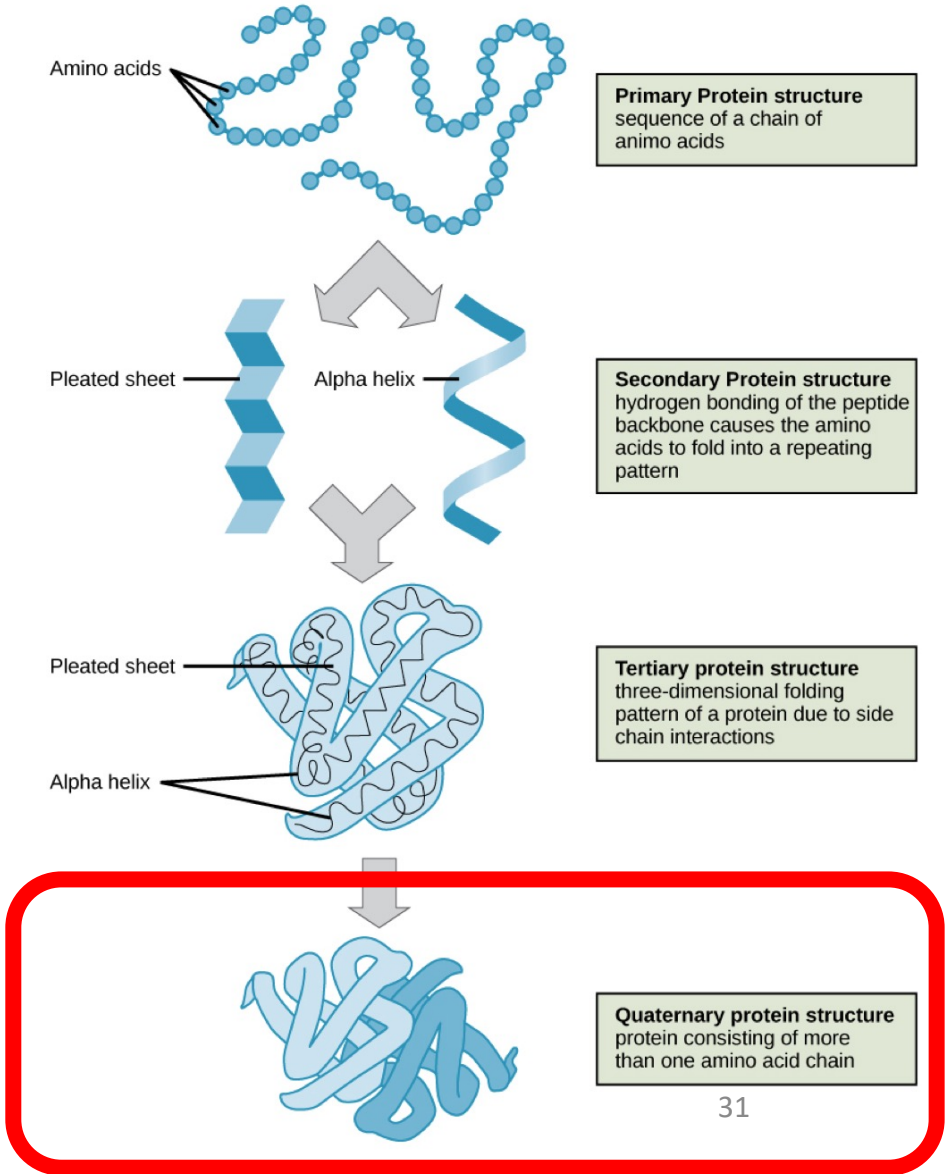


Protein fold

- **Fold**: specific 3D arrangement of the secondary structure elements
- Folds are categorized, compared and stored in databases
- Between 1,000 and 10,000 unique folds
- ~800 have resolved 3D structures so far

Quaternary structure

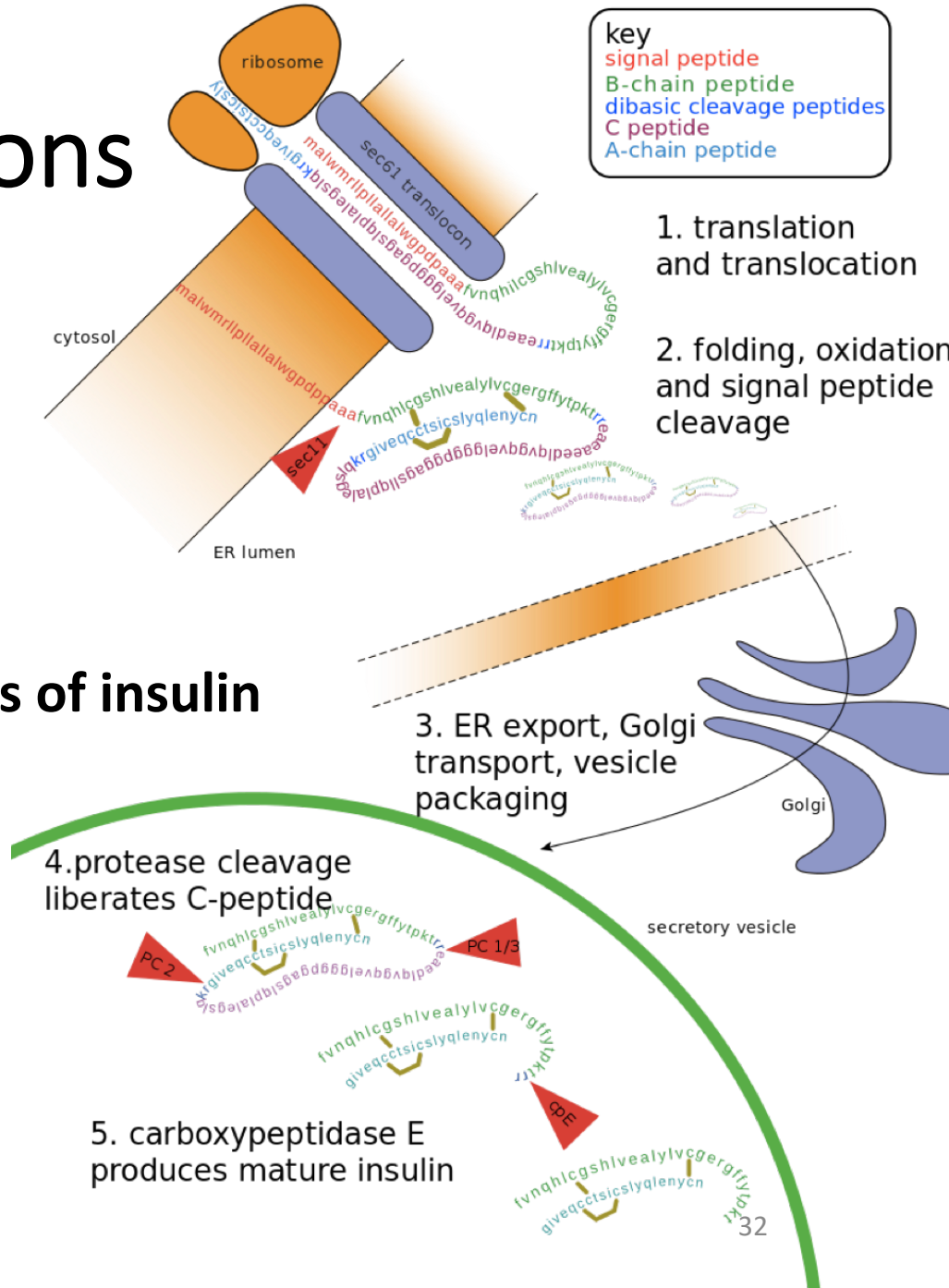
- Spatial arrangement of multiple polypeptide chains
- Necessary for function
  - Active site may be formed by multiple subunits
  - Co-localization of multiple active sites
  - Interaction sites may be shared by multiple subunits
  - Scaffold (e.g. tubulin)



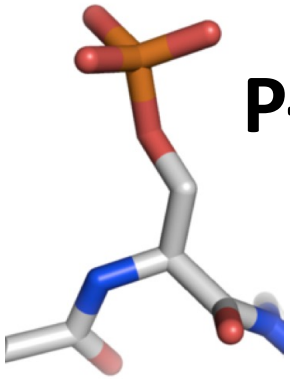
Post-translational modifications

- Covalent and/or enzymatic modification of proteins after their synthesis
  - Cleavage
  - Formation of disulphide bonds
  - Methylation
  - Phosphorylation
  - Glycosylation
  - Lipidation
  - …
- Reversible or irreversible

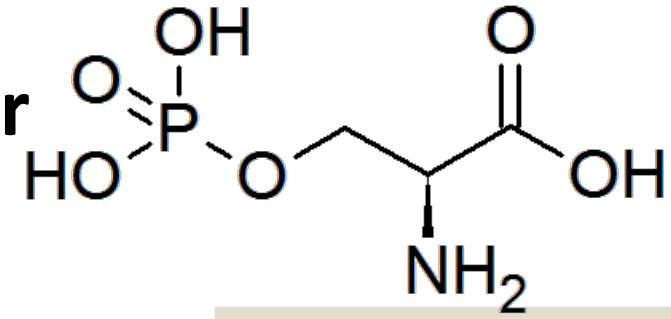
E.g. PTMs of insulin



Phosphorylation



P-Ser

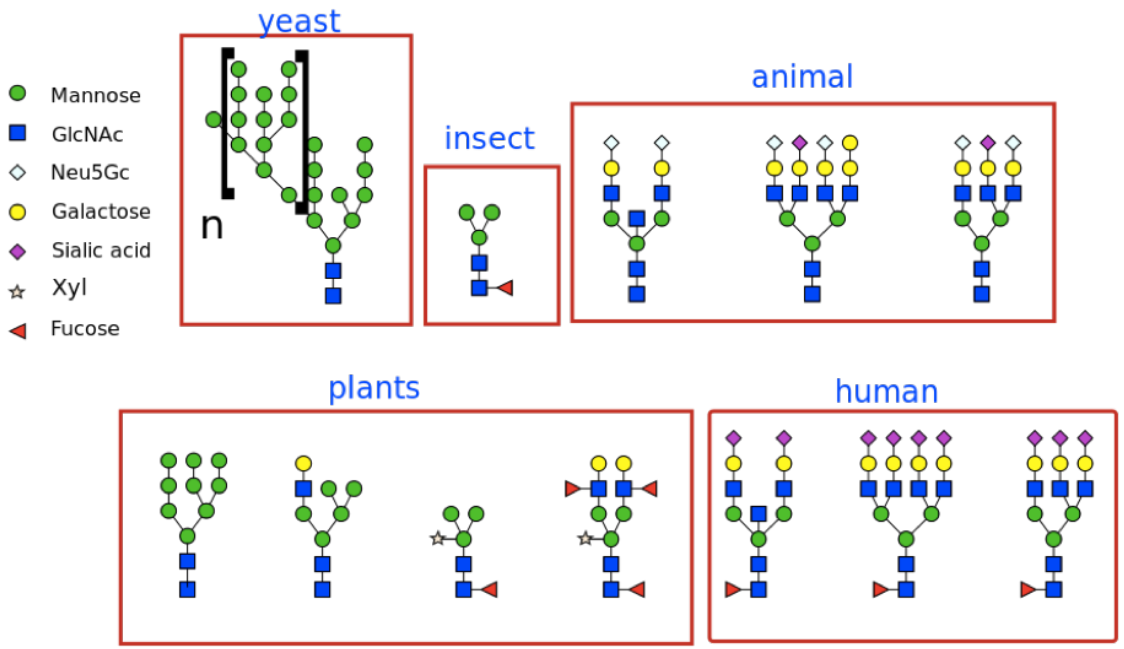


- Reversible
- Regulatory mechanism in prokaryotes and eukaryotes
- Phosphorylated amino acids (anywhere in the protein):
  - Ser > Thr >> Tyr
  - His, Asp (prokaryotes >> eukaryotes)
- Carried out by **kinases**
  - Often a specific motif (sequence of amino acids) is targeted, e.g. several kinase types target a consensus **Ser/Thr-Pro**

Kinase	Full name	Consensus phosphorylation site
PKA	Protein kinase A or cAMP-dependent protein kinase	R-R-X-S/T-Φ
CDK	Cyclin-dependent kinase	S/T-P-X-K/R
ERK2	Extracellular-regulated kinase-2	P-X-S/T-P
CK1*	Casein kinase-1	pS-X-X-S/T
CK2‡	Casein kinase-2	S/T-D/E-X-E/D
GSK3	Glycogen synthase kinase-3	S-X-X-X-pS
CaMK2	Calmodulin-dependent protein kinase-2	R-X-X-S/T
ABL	Abelson murine leukaemia virus tyrosine kinase	I/V/L-Y-X-X-P/F
EGFR	Epidermal growth factor receptor	E-E-E-Y-F
Src	Rous sarcoma virus tyrosine kinase	E-E-I-Y-E/G-X-F
IRK	Insulin receptor tyrosine kinase	Y-M-M-M
PKB/AKT	Protein kinase B	R-X-R-X-X-S/T
PKD	Protein kinase D	L/I-X-R-X-X-S/T
PIM1–3	Proviral integration site kinases 1–3	R-X-R-X-X-S/T

Glycosylation

Examples of glycans

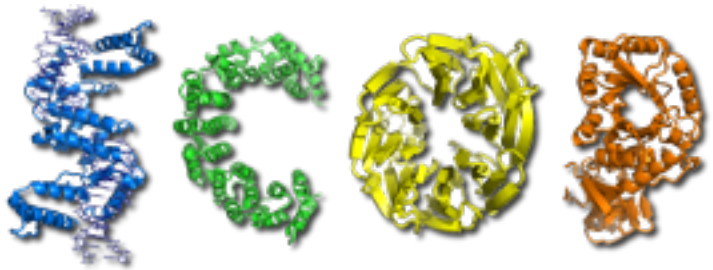


- Attachment of glycans (carbohydrates)
- In eukaryotes, happens in ER and Golgi
- Generally irreversible
- Most frequent types:
  - N-linked glycosylation: attachment to N in Asn, targets a consensus motif **Asn-X-Ser/Thr**
  - O-linked glycosylation: attachment to O in any amino acid; no simple consensus known, depends on the glycan
- Abundant in viral envelope proteins

Protein structure classification

Two major resources

- **SCOP**: Structural Classification of Proteins
- Manually curated => more accurate
- http://scop.mrc-lmb.cam.ac.uk/

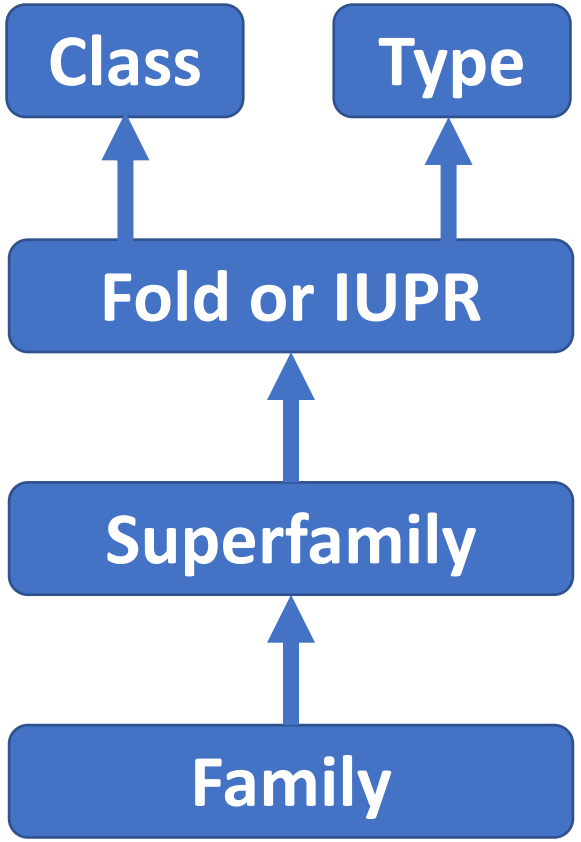


- **CATH**: Protein Structure Classification Database
- Automated => larger
- https://www.cathdb.info/

SCOP: a semi-manual hierarchy of proteins

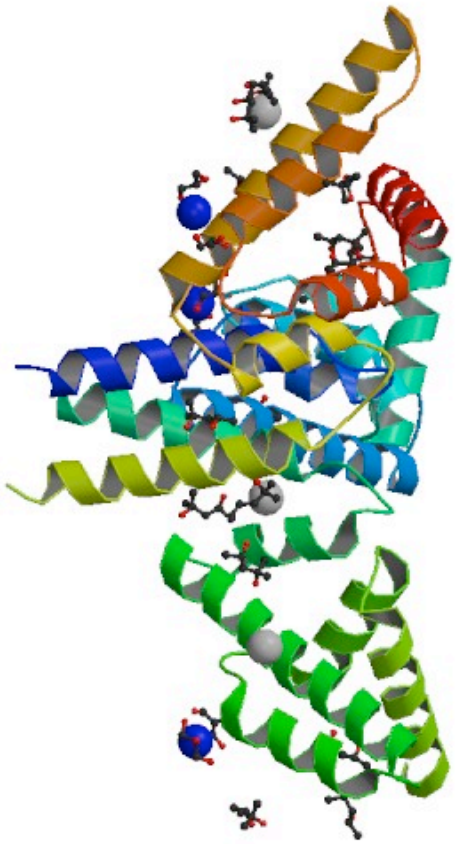
- **Family**: groups closely related proteins with a clear evidence for their evolutionary origin
- **Superfamily**: families, whose structures and functions suggest common evolutionary origin
- **Fold**: same major secondary structure elements arranged similarly; changes at periphery
- **IUPR** (Intrinsically Unstructured Protein Region): superfamilies of proteins or protein regions that do not adopt globular folded structure
- **Class**: folds and IUPRs defined by secondary structural content
  (1) all alpha; (2) all beta; (3) alpha and beta (interspersed); (4) alpha plus beta (segregated); (5) small
- **Type**: (1) globular; (2) membrane; (3) fibrous; (4) intrinsically unstructures

Hierarchical tree (in most cases)



Four major classes in SCOP

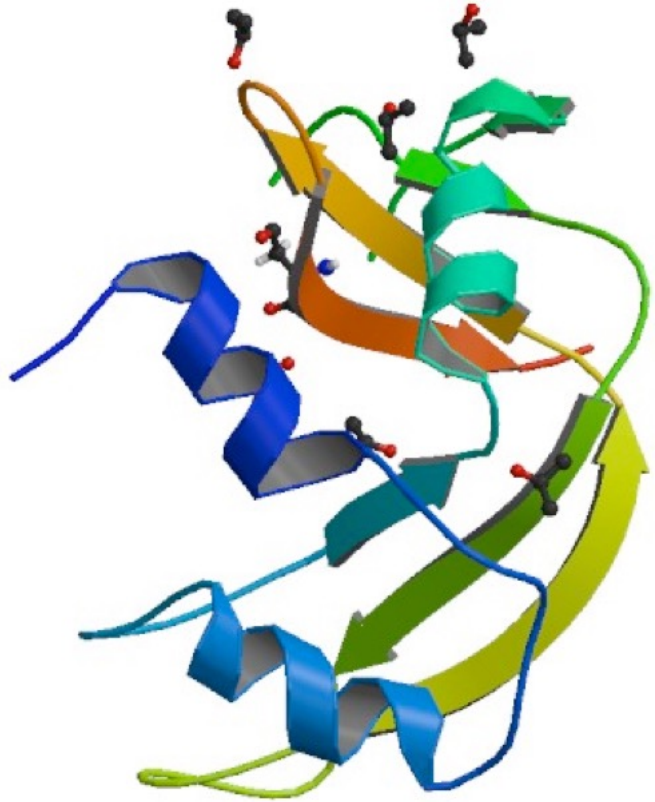
ALL-ALPHA



ALL-BETA

ALPHA / BETA

ALPHA + BETA

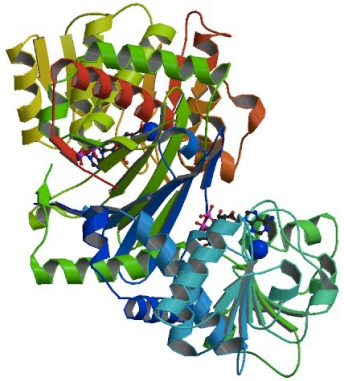


Example of SCOP hierarchy

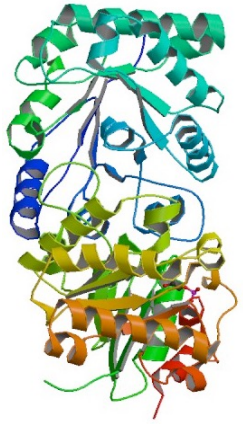
Class: α/β

Folds: TIM-barrel

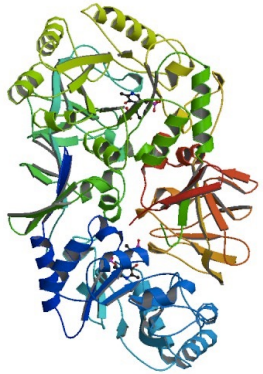
Rossmann fold



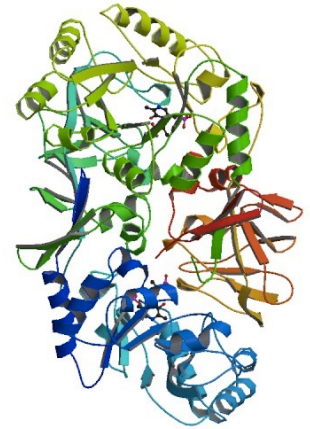
Superfamilies: Thiophosphate isomerase (TIM)



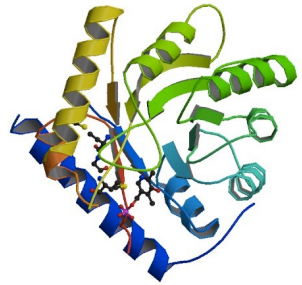
PLP-binding barrel



Families: Alanine racemase-like, N-terminal domain



"Hypothetical" protein ybl036c



Accessing SCOP

- http://scop.mrc-lmb.cam.ac.uk/: big relaunch in 2020, SCOP 2.0
  - Keyword and ID search; sequence search
- Downloadable parsable files and format description: http://scop.mrc-lmb.cam.ac.uk/download
- (BioPython module https://biopython.org/docs/1.75/api/Bio.SCOP.html, **not recommended**: uses legacy SCOP 1.75 format and classification)
- REST interface: http://scop.mrc-lmb.cam.ac.uk/api/

SCOP stats

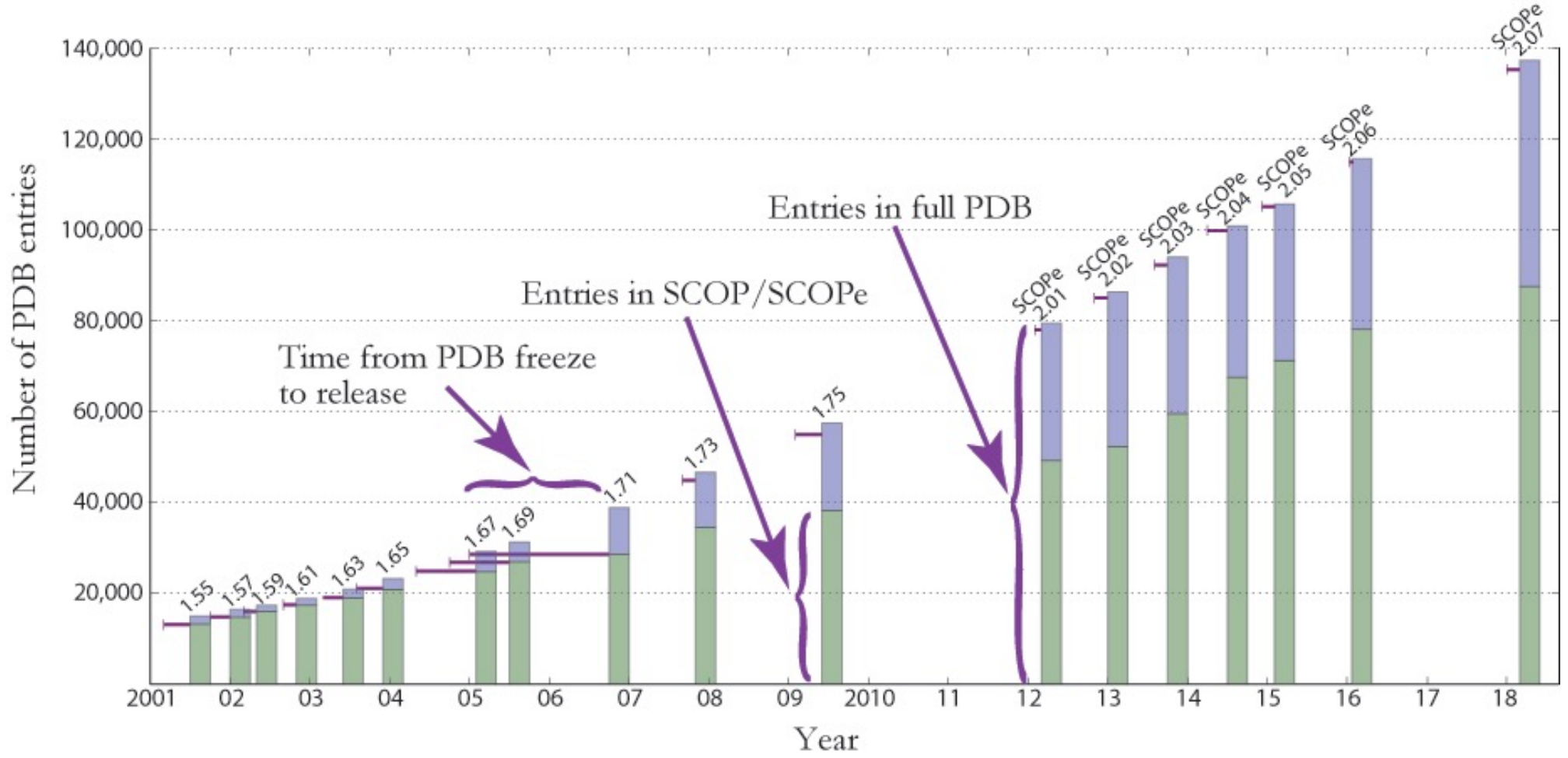
	SCOP2	SCOP 1.75
Number of folds	1487	1195
Number of IUPR	22	n.a
Number of hyperfamilies	18	n.a
Number of superfamilies	2660	1962
Number of families	5562	3902
Number of inter-relationships	60	n.a

61,528 non-redundant domains representing
700,279 protein structures (2020-10-20)

A parallel effort: SCOPe (extended SCOP)

- http://scop.berkeley.edu/
- Regular updates (last update: July 2020)
- Available for browsing and downloads
- Also heavily depends on manual curation
- Uses the old SCOP nomenclature: e.g. “a.1.1.1“
- More classes than in SCOP
- Aims to classify all PDB entries (SCOP does not)

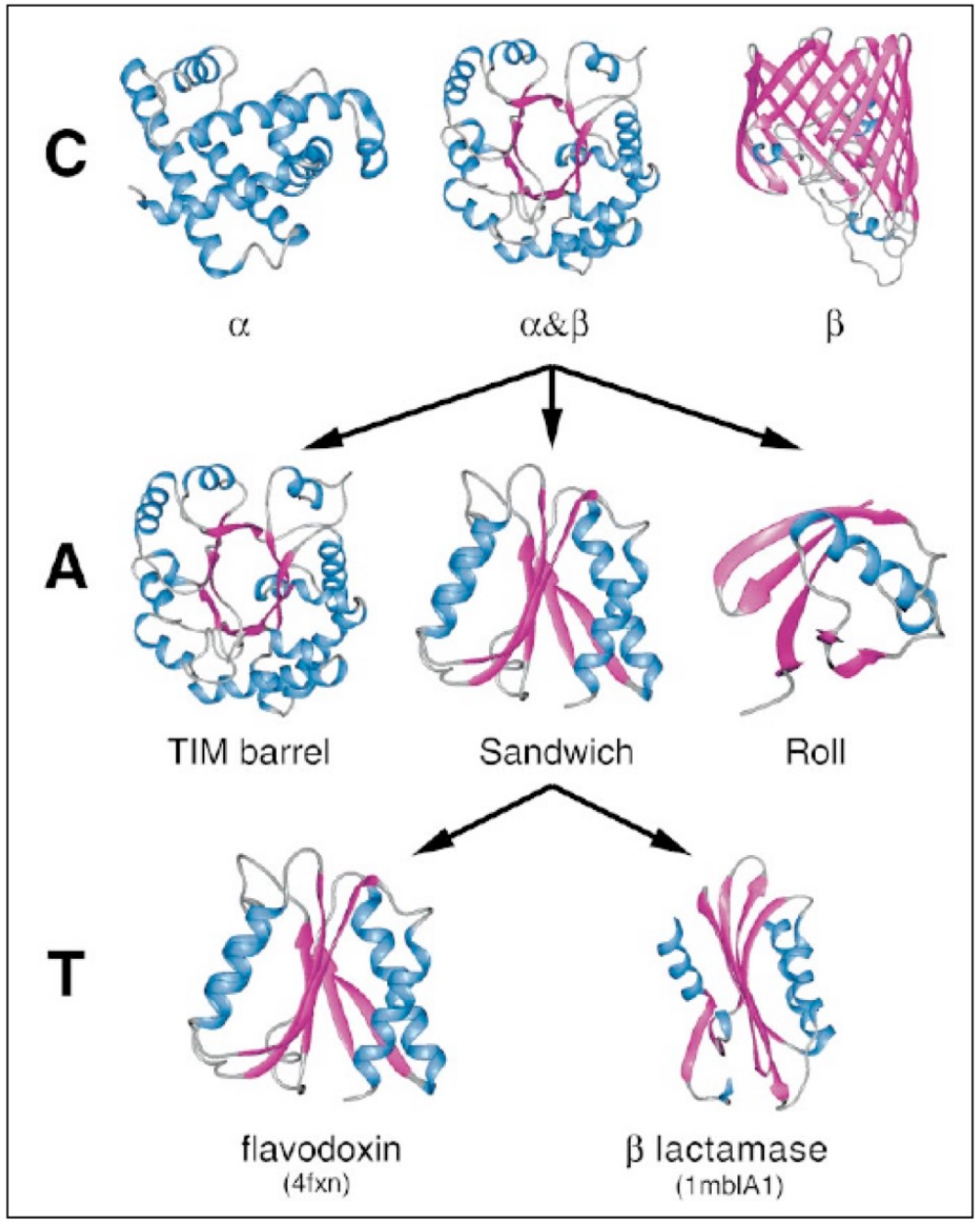
A parallel effort: SCOPe (extended SCOP)



CATH: an automated hierarchy of proteins

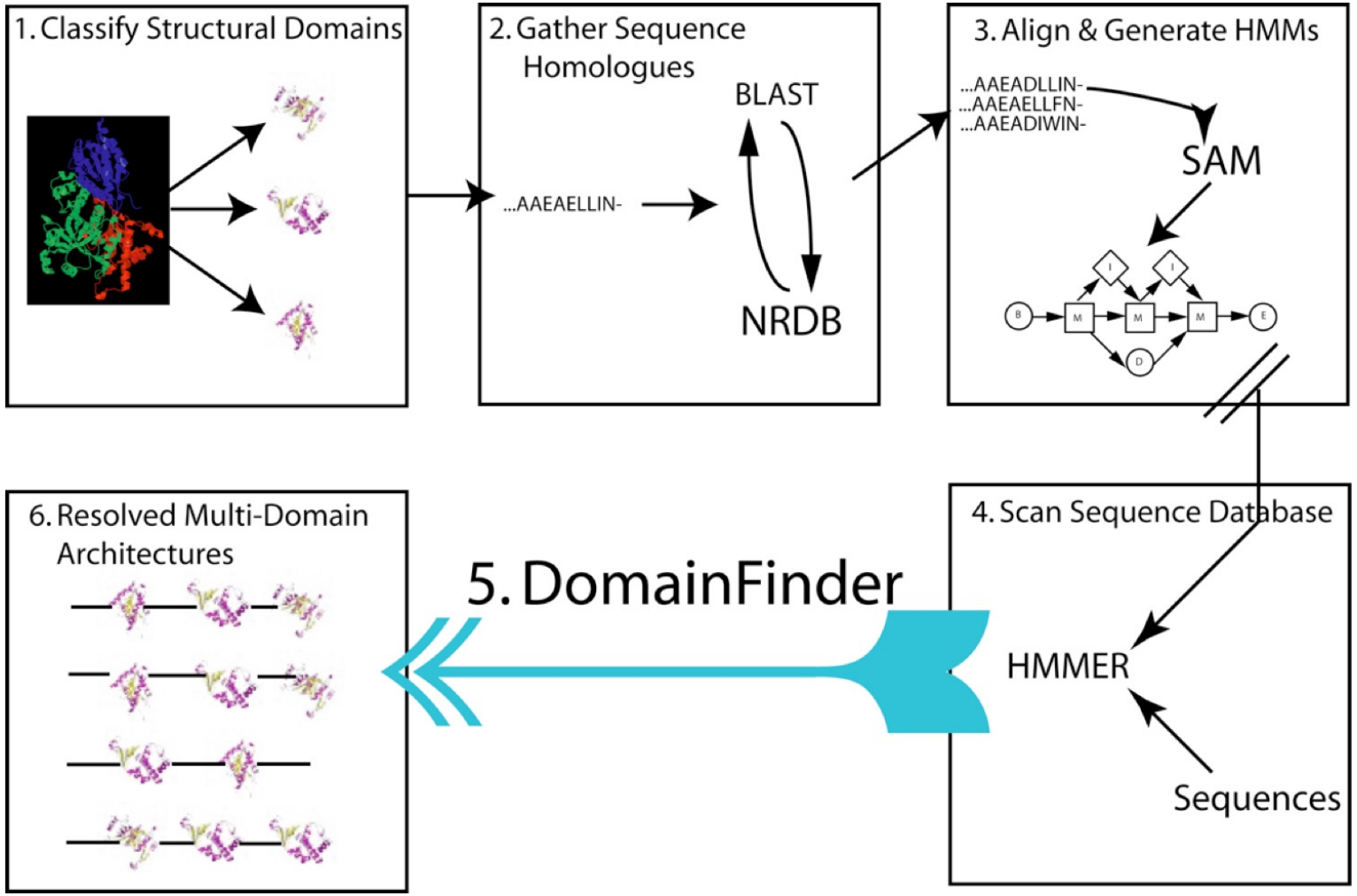
- CATH: class (C), architecture (A), topology (T), homologous superfamily (H)
- C: secondary structure composition and contacts
  - (1) mainly α; (2) mainly β; (3) α-β
- A: overall arrangement of secondary structure, independent of connectivity
- T: similar number and arrangement of the secondary structure elements
- H: structure and function similarity => common ancestry
- S: sequence identity > 35%

CATH categories



Gene3D: an integral part of CATH

- Assigns sequences with no 3D structure to CATH domains
- **82,665,384** sequences in **151,013,797** domains

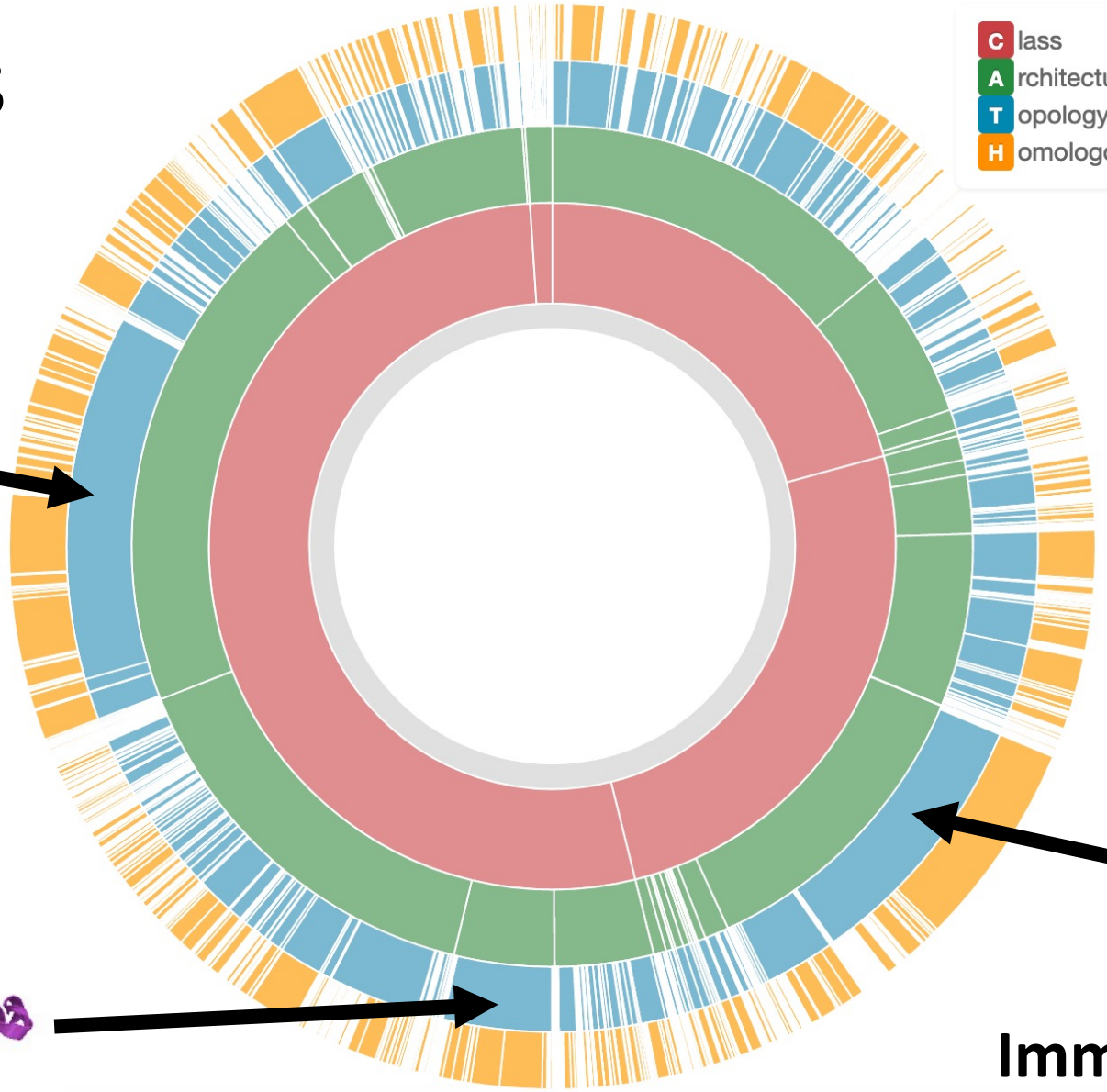


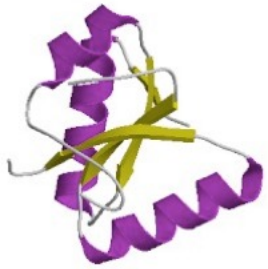
CATH stats

	CATH-Plus 4.3.0	CATH (daily snapshot)
PDB Release	01.07.19	3 months ago
Domains	500238	536769
Superfamilies	5481	6631
Annotated PDBs	131091	165026

Aims to classify all structures in PDB

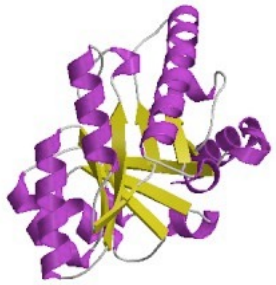
CATH stats





Rossmann fold

TIM barrel

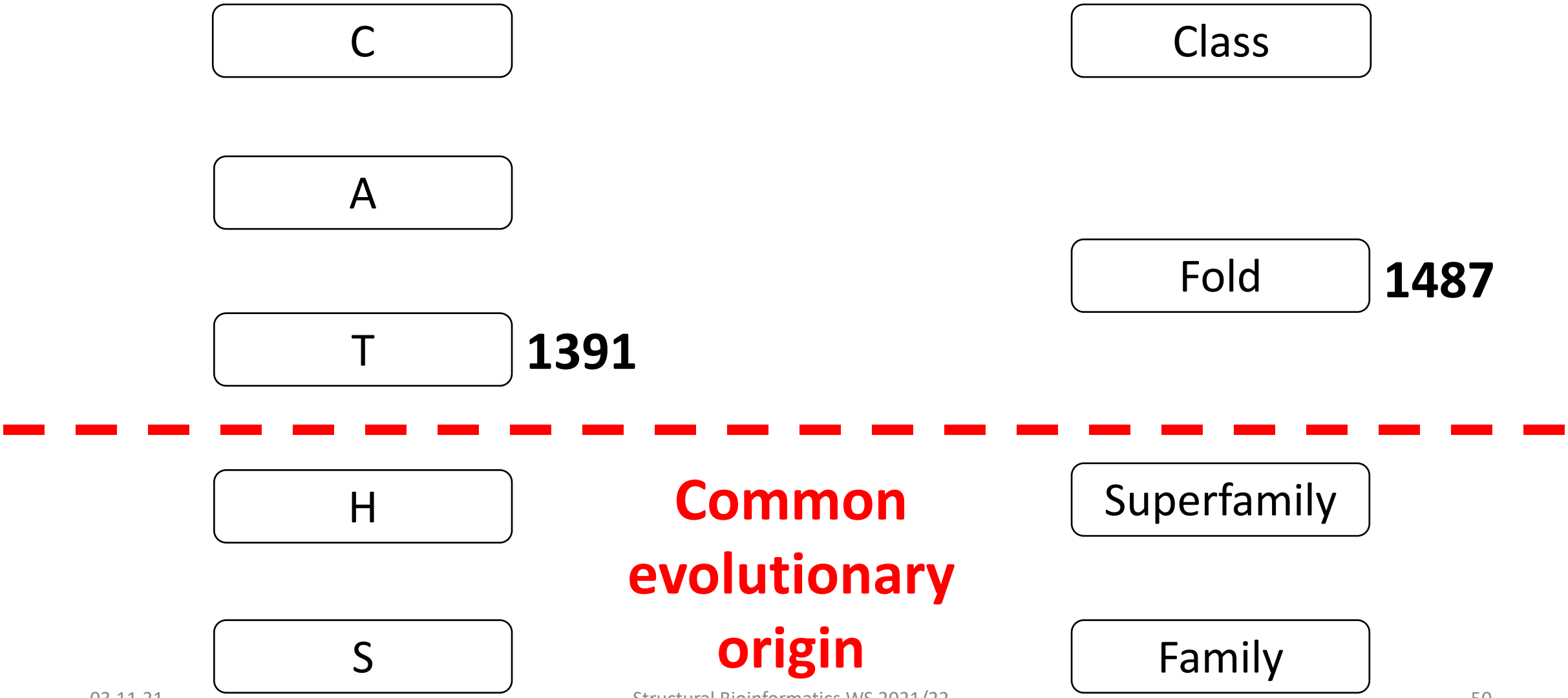


Immunoglobulin fold

Accessing CATH

- Various search options
  - Keyword and ID search; search by sequence; search by structure
- Downloadable files and documentation: http://www.cathdb.info/wiki?id=data:index
- Related databases and resources
  - E.g. FunFams: classification of proteins into functional families with annotated conserved (=> presumably functionally important) positions

CATH vs. SCOP


C
Class
A
Fold
1487
T
1391
H
Superfamily
Common evolutionary origin
S
Family

Summary and possible exam questions

- Name the four levels of organization of protein 3D structure
- What types of chemical bonds maintain protein 3D structures? To what organizational levels do they correspond?
- What are post-translational modifications, and what is their role in protein function
- Levels of SCOP and CATH classification
- At what level of structural similarity do you expect common evolutionary origin?