# Structure of biomolecules
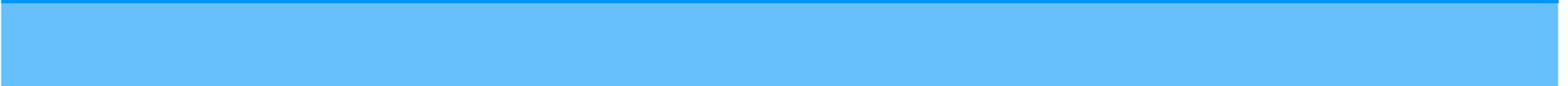
Outline

- Proteins

  - primary structure

  - secondary structure

  - tertiary structure

  - motifs and folds

  - quaternary structure

- Nucleic acids

  - Main types of structures

- Structural data formats

  - PDB format

- Primary structural databases

Hierarchy of protein structure

Primary structure	Secondary structure	Tertiary structure	Quaternary structure
Amino acid residues	α Helix	Polypeptide chain	Assembled subunits
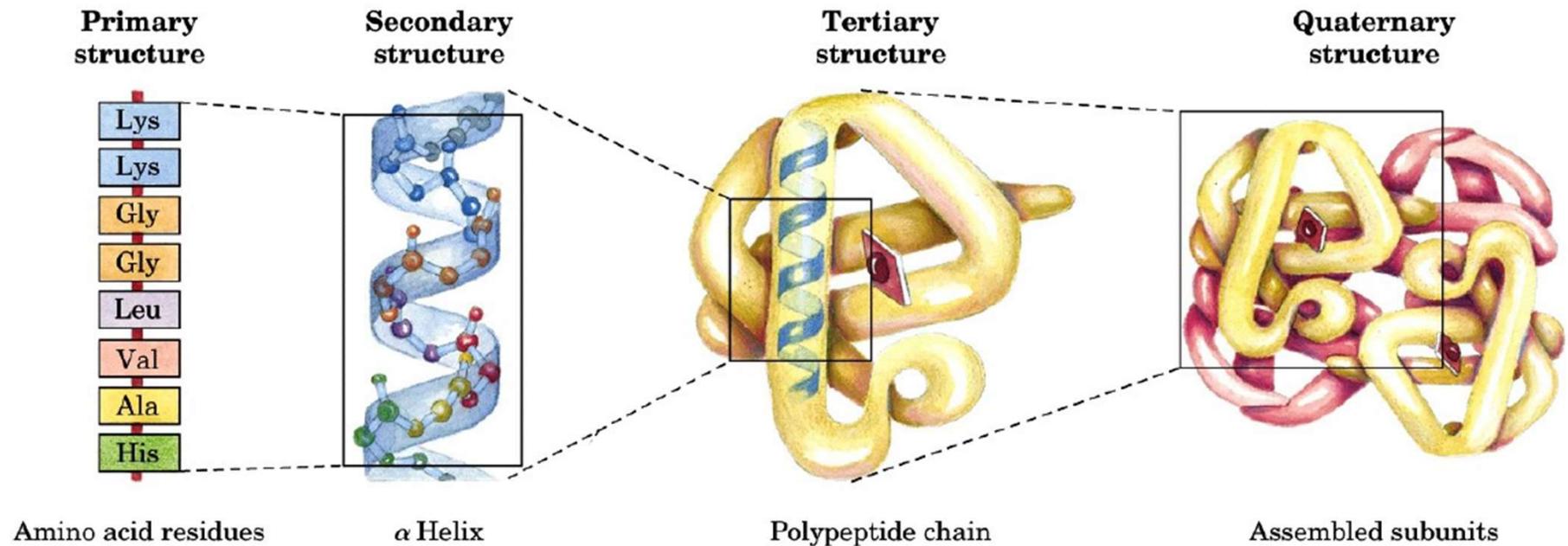
Amino acids

❑ 20 **L**-amino acids (natural)

❑ Side chains

▪ charged, polar, hydrophobic

Side chain → R · H

Amino group → H_2N · COOH

Chiral centre (Cα)

Acid group
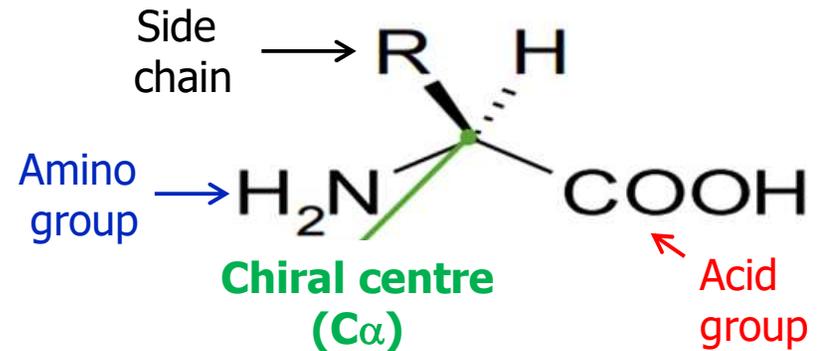
A. Amino Acids with Electrically Charged Side Chains

Positive

| Arginine (Arg) R | Histidine (His) H | Lysine (Lys) K |

Negative

| Aspartic Acid (Asp) D | Glutamic Acid (Glu) E |

B. Amino Acids with Polar Uncharged Side Chains

| Serine (Ser) S | Threonine (Thr) T | Asparagine (Asn) N | Glutamine (Gln) Q |

C. Special Cases

| Cysteine (Cys) C | Glycine (Gly) G | Proline (Pro) P |

D. Amino Acids with Hydrophobic Side Chains

| Alanine (Ala) A | Valine (Val) V | Isoleucine (Ile) I | Leucine (Leu) L |

| Methionine (Met) M | Phenylalanine (Phe) F | Tyrosine (Tyr) Y | Tryptophan (Trp) W |
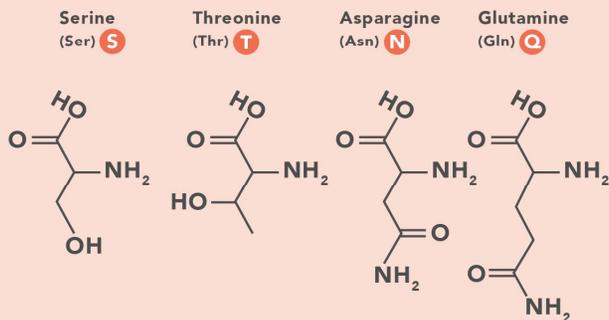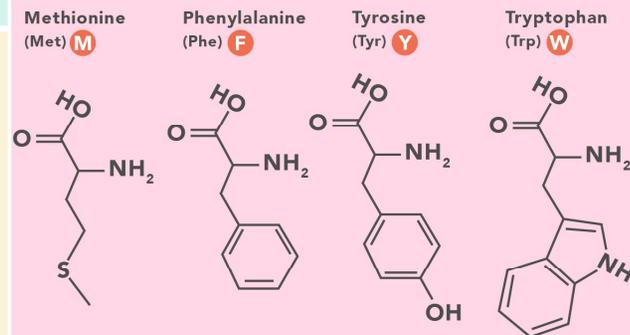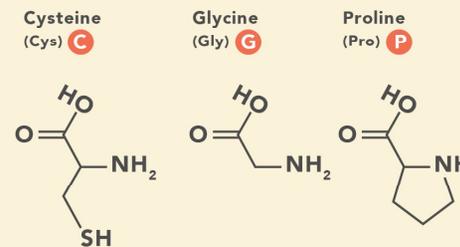
Primary structure

❑ Linear chain of amino acid residues

- MSLGAKPFGEKKFIEIKGRRMAYIDEGTGDPILFQHGNPTSSYLWRNIM

N-terminus   C-terminus

❑ Protein backbone

- covalent bonds

- from N-terminus to C-terminus

❑ Peptide bond

- partial double bond character

→ planar geometry
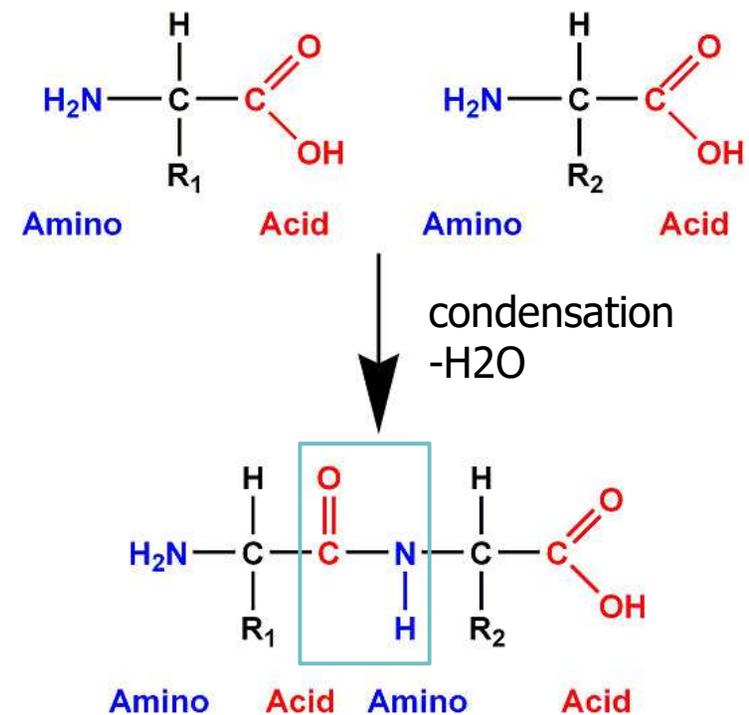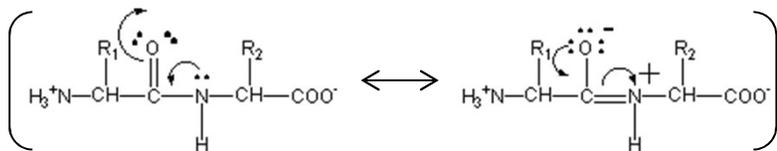
Geometry of protein backbone

❑ Conformation of the peptide chain

   ▪ defined by Φ (fi) and Ψ (psi) angles

❑ Ramachandran (Φ, Ψ) plot

   → the majority of proteins follow this distribution
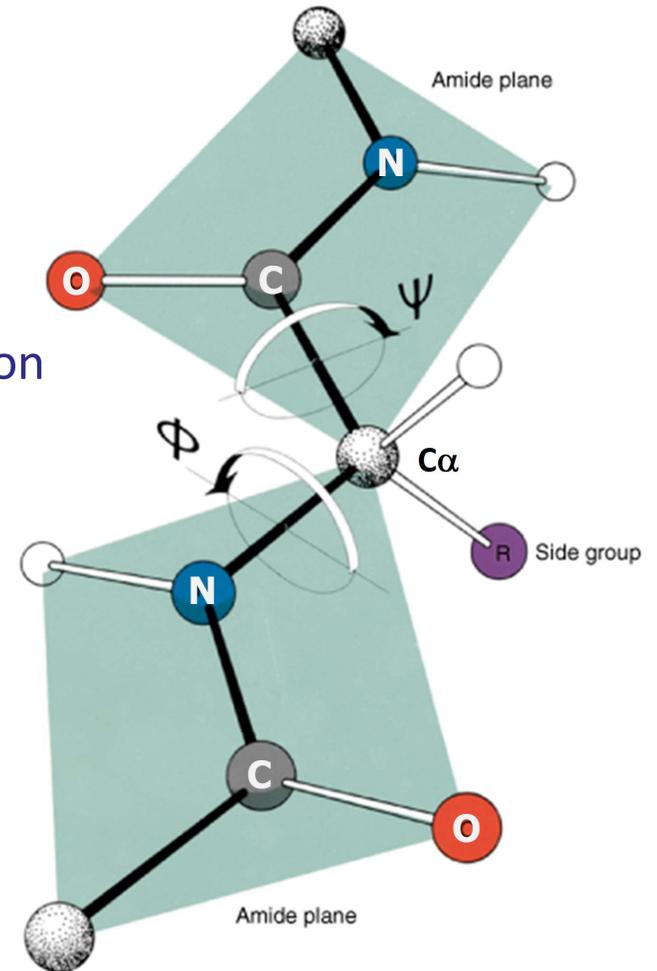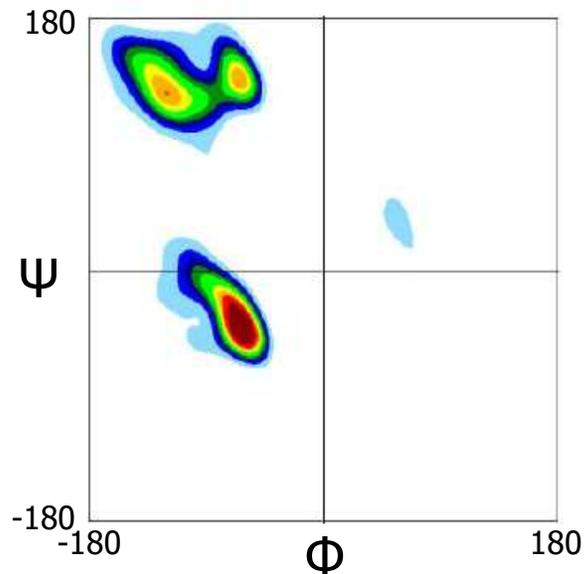
Secondary structure

- ❑ **Local** three-dimensional **structure** of polypeptide chain

- ❑ Governed by **hydrogen bonding** between backbone atoms

- ❑ Types of structures

  - ▪ helices
  - ▪ β-structures   } regular
  - ▪ loops and coils - irregular

Secondary structure

❑ DSSP (hydrogen bond estimation algorithm)

- The most common method for assigning secondary structure

- Starts by identifying the intra-backbone hydrogen bonds

  (between NH ····· O=C)

- Hydrogen bond exists if $E \leq$ -0.5 kcal/mol

- The type of repetition will assign the residue to one of 7 types

  (3 major types: helices, strands and loops)

$$E = 0.084 \left\{ \frac{1}{r_{ON}} + \frac{1}{r_{CH}} - \frac{1}{r_{OH}} - \frac{1}{r_{CN}} \right\} \cdot 332 \, \text{kcal/mol}$$

Helices

❑ Types of helices

- 3.6_{13} helix (α-helix) – common

- 3_{10} helix – less frequent, end of α-helices

- 4.1_{16} helix (π-helix) – rare

→ represented by helical

cartoons or cylinders

❑ Hydrogen bonding

- within a single chain

3_{10}-helix

Ψ

π-helix

α-helix

Φ
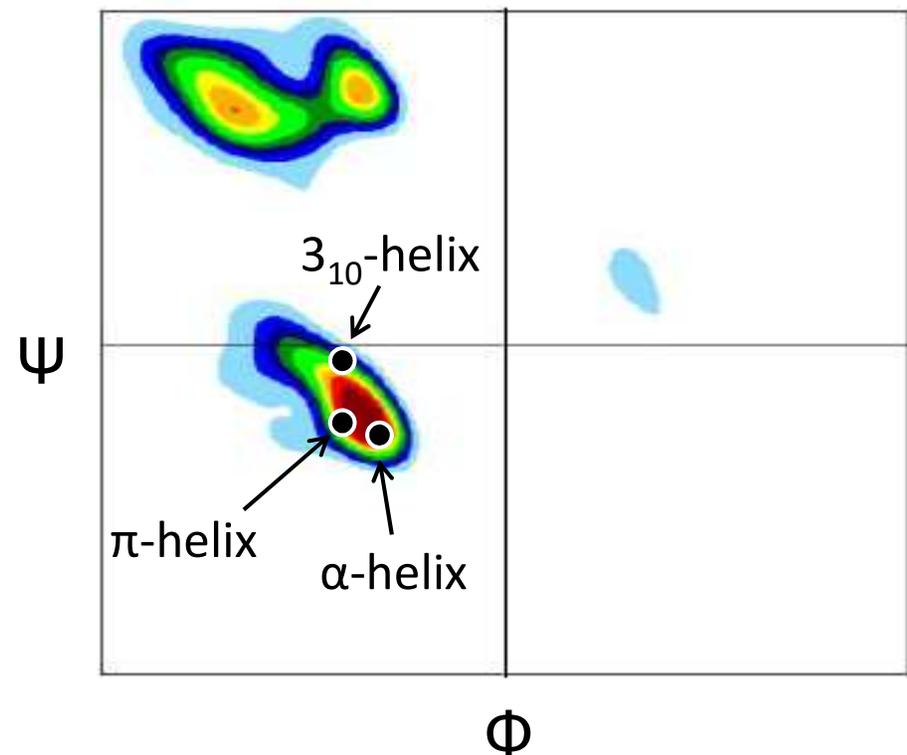
Helices

Type	3_{10}	α	π
Residues per turn	3.0	3.6	4.1
Atoms in H-bonded ring	10	13	16
Hydrogen bonding	$n - n + 3$	$n - n + 4$	$n - n + 5$
Angle between neighboring residues	120	100	88
Helical rise per amino acid residue (Å)	2.0	1.5	1.15
ϕ (°)	−75	−60	−75
ψ (°)	−5	−45	−40

3_{10} helix α helix π helix
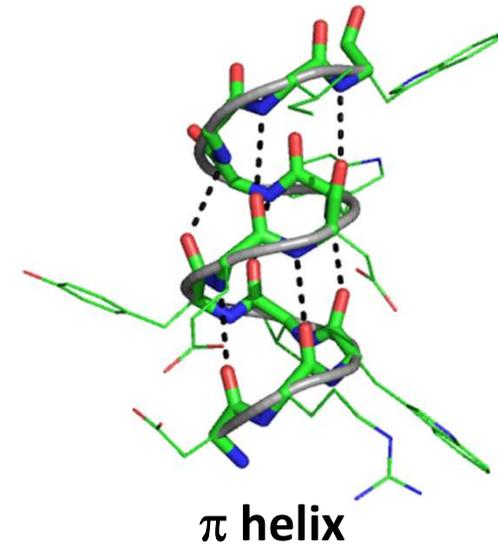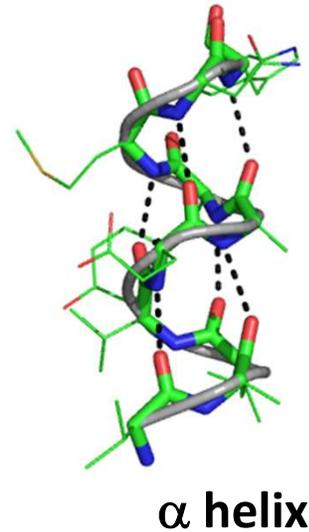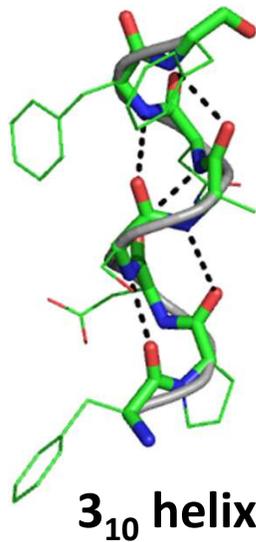
β-structures

- ❑ Types of typical β-structures
  - β-sheets
  - β-turns
  - β-bulge
  - polyproline helices

- ❑ Hydrogen bonding
  - between adjacent chains

polyproline helices

β-sheets

Ψ

Φ
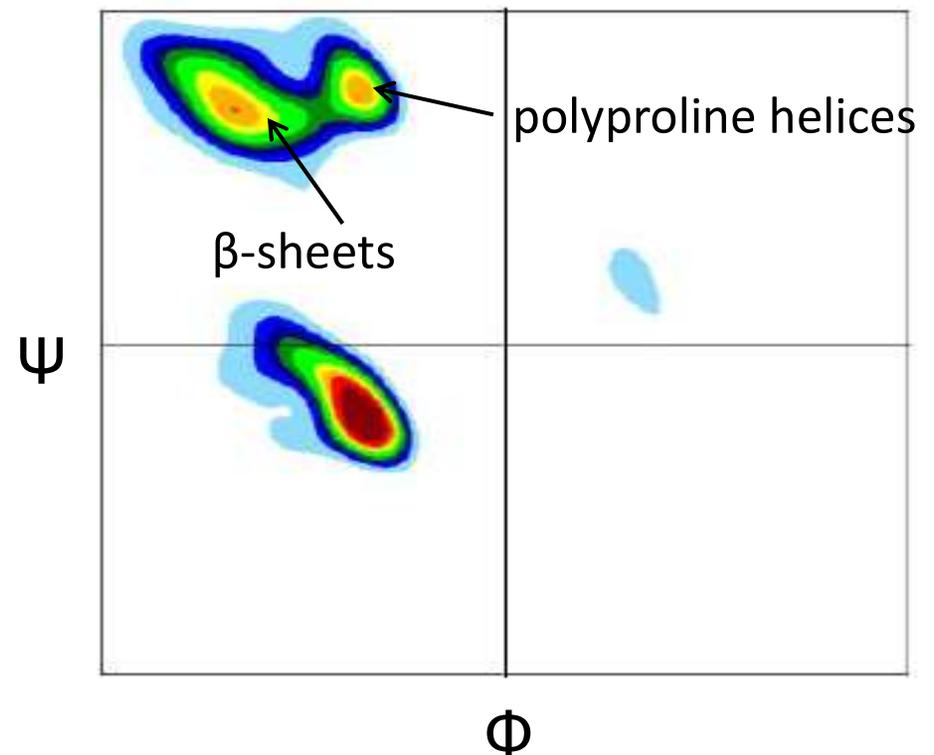
β-structures

- ❑ Types of β-sheets
  - ▪ parallel
  - ▪ antiparallel (stronger)
  - ▪ mixed
 - → represented by arrows
    indicating the sequence
    direction
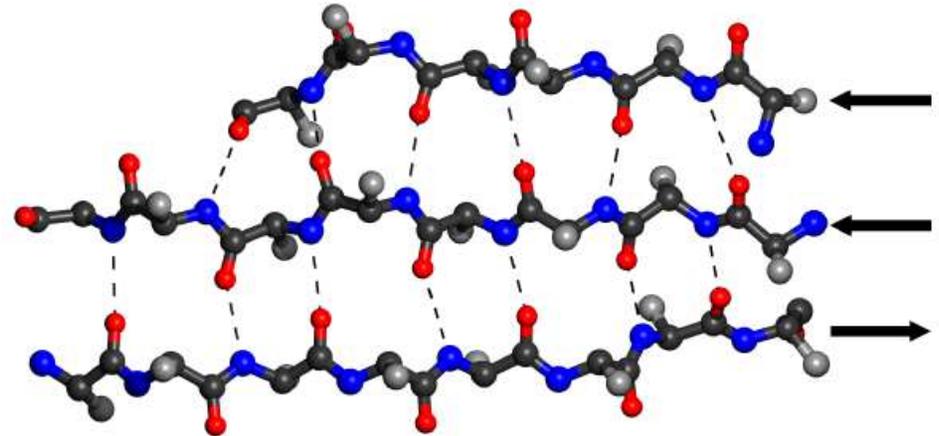
- ❑ Side-chains
  - ▪ towards the sides of

the sheets
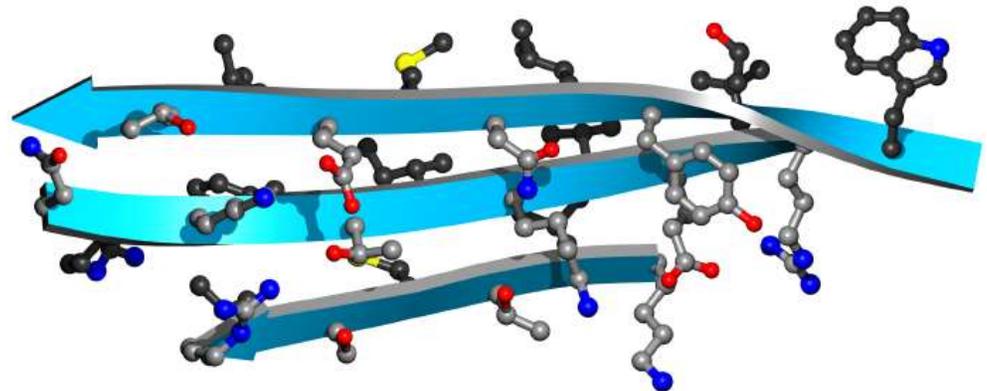
β-structures

❑ **β-turns**

- short structures (4-5 residues)

- connects two β-strands

- ideally H-bond between backbone of n and $n+3$ residues

- often includes glycine or proline on specific positions
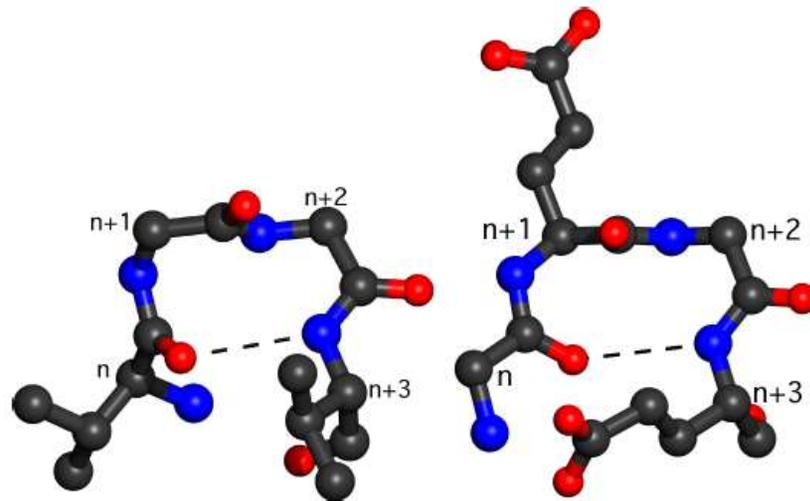
# β-structures

- β-bulge
  - frequently occurs in antiparallel β-sheets
  - disrupts ideal H-bonding pattern
  - increases twists of a sheet
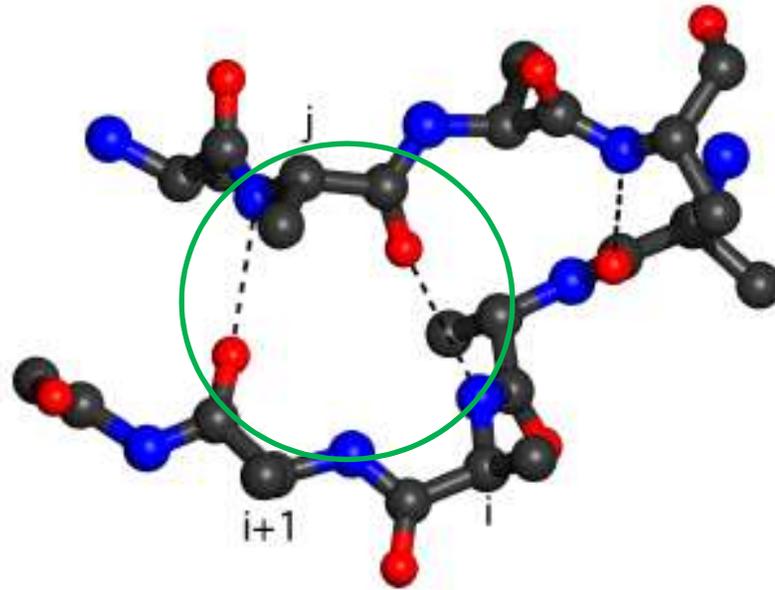
β-structures

□ Polyproline helices

- typical in collagen and other strong fibers

- left handed triple-stranded helix (unlike all other helices)

- composed of three chains of repetitive sequence (Proline-Hydroxyprolin-Glycine)$_n$
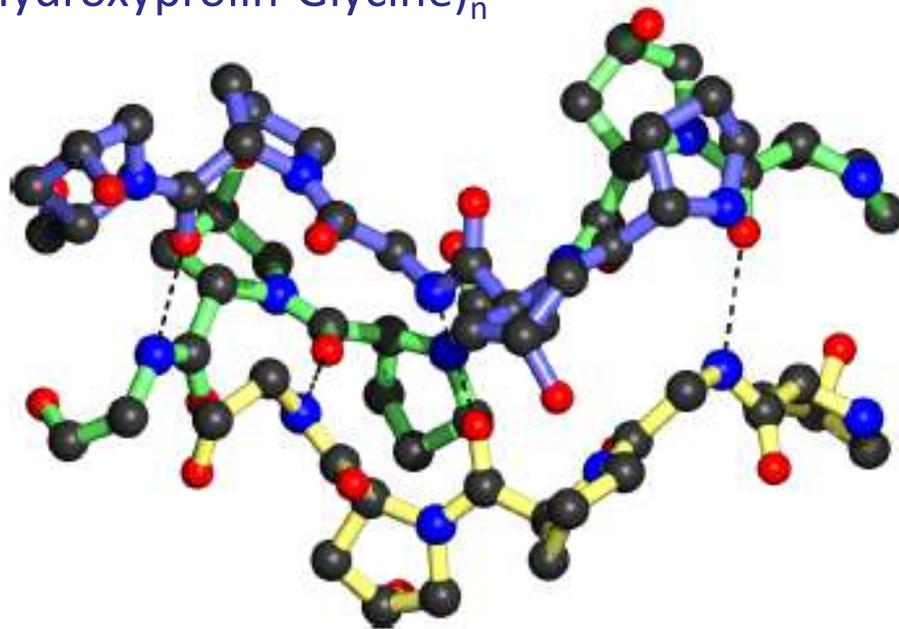
Tertiary structure

❑ Global three dimensional structure of protein
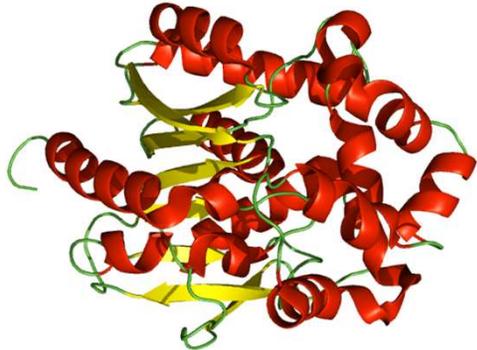
❑ Governed mainly by hydrophobic interactions involving side chains of amino acid residues
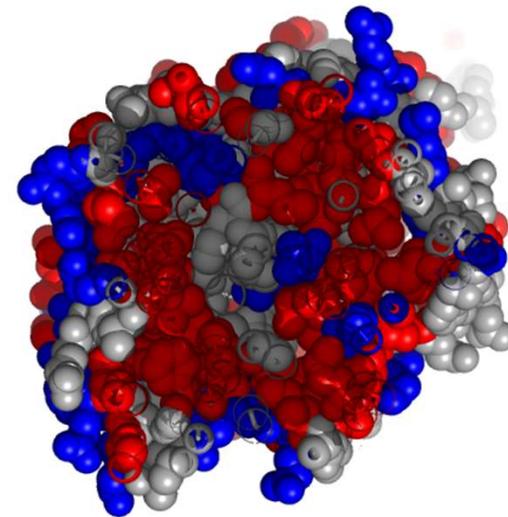
Tertiary structure

- ❑ Supersecondary structures (Motifs)

  - ▪ small substructures formed by several secondary structures

- ❑ Domain

  - ▪ structurally (functionally) independent regions

  - ▪ compact parts of structure – around single hydrophobic core

  - ▪ formed in separate folding unit

- ❑ Fold

  - ▪ general architecture of protein

  - ▪ type of protein structure

Protein motifs

- β-harpin
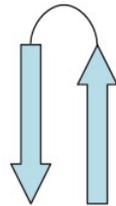
- β-meander
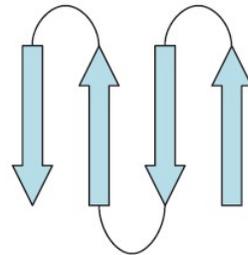
- Greek key
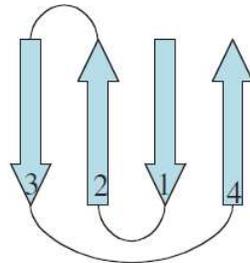
- Jellyroll
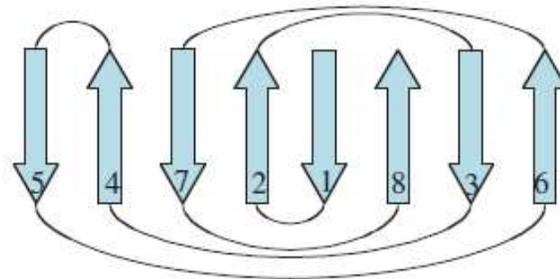
Protein motifs

□ Helix-turn-helix
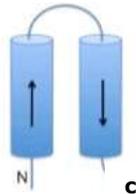
□ Helical bundle
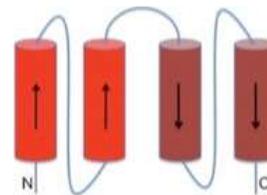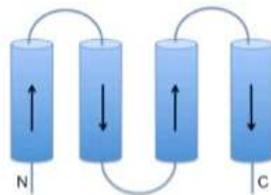
□ βαβ unit
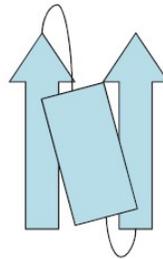
Protein domains

❑ Parts of tertiary structure

- separate folding
- Independent structure
- Usually up to 200 residues

Pyruvate kinase in glycolysis

pyruvate kinase

ADP → ATP

PEP → pyruvate

Nucleotide binding

Substrate binding

Regulatory domain
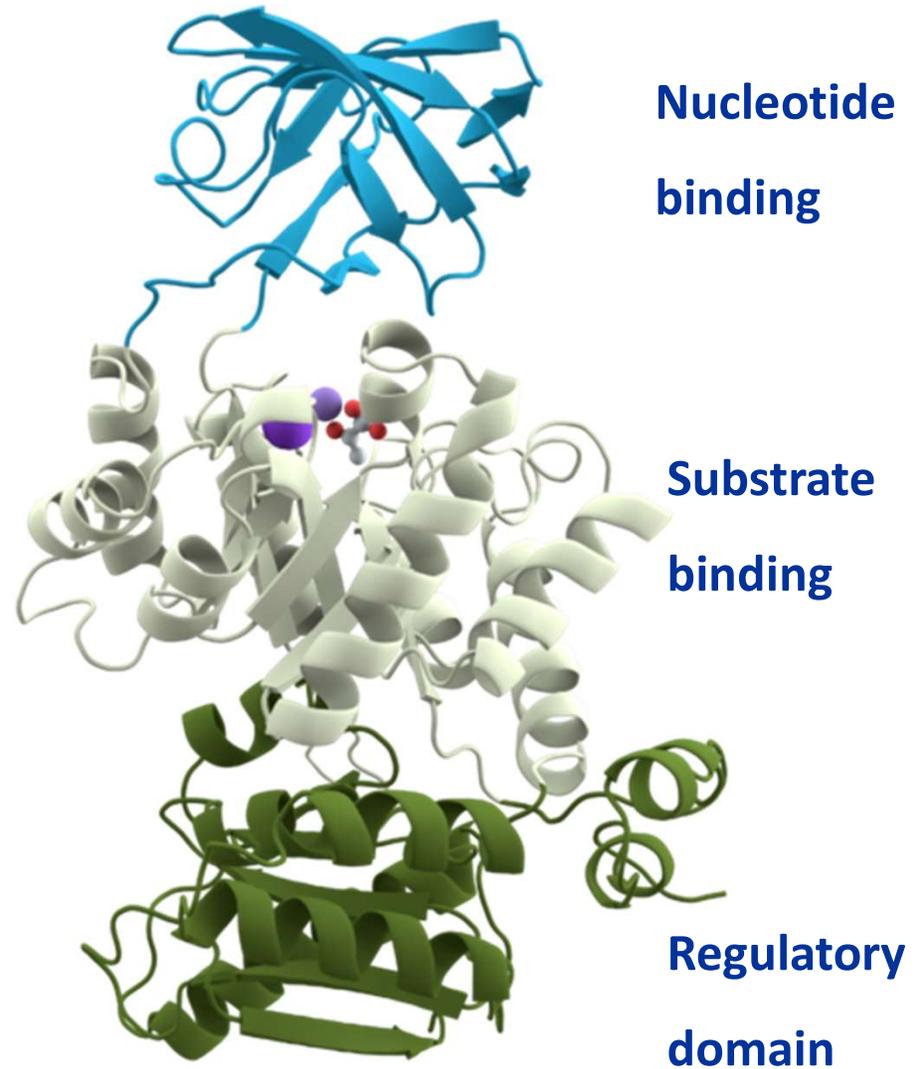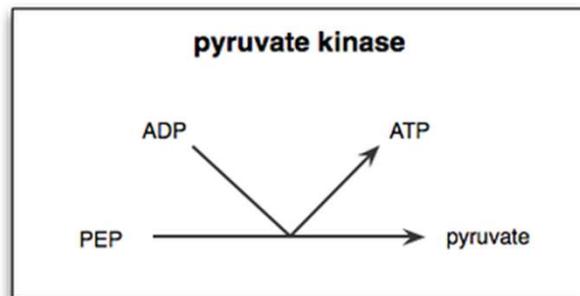
Protein folds

❑ Some folds are very common, some are rare

❑ Classification of folds

   ▪ biochemical

      ▪ globular, membrane, fibrous proteins,

        intrinsically disordered

   ▪ structural

      ▪ all-α, all-β, α/β and α+β proteins

❑ Number of folds

   ▪ currently: 1,195 (SCOP) vs 1,373 (CATH)

   ▪ theoretical maximum: 10,000
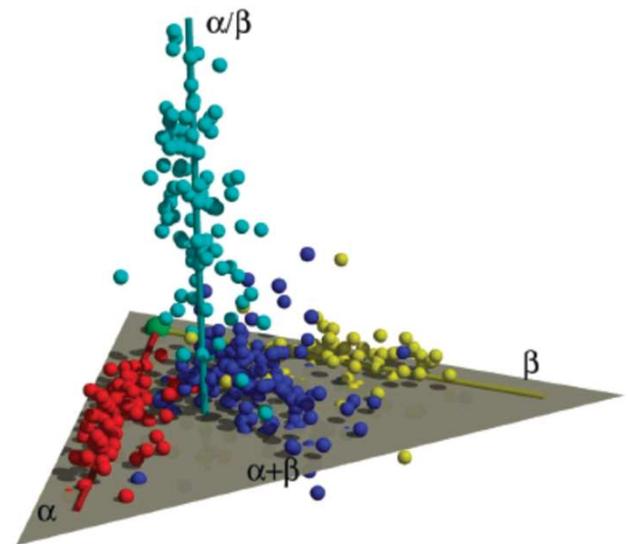
Biochemical classification of folds

❑ **Globular proteins**
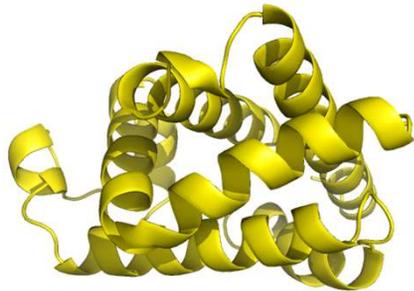
❑ **Membrane proteins**
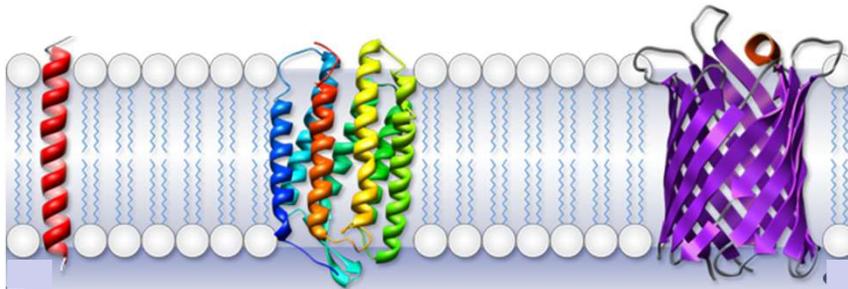
❑ **Fibrous proteins**
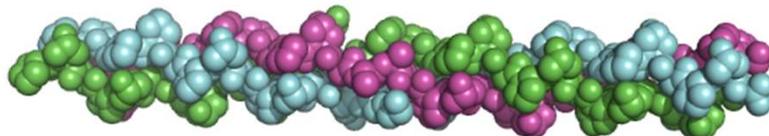
# Structural classification of folds

❑ All-α (entirely α-helices)
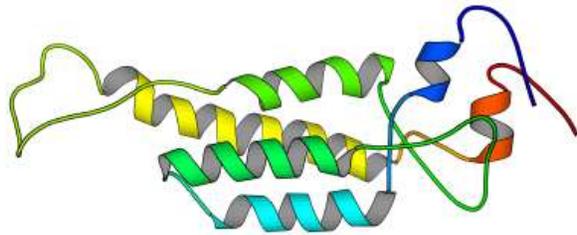

up-and-down bundle
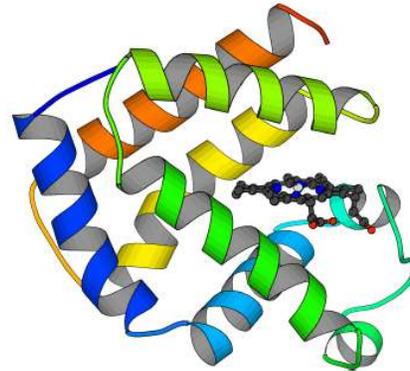
globin-like

❑ All-β (entirely β-strands)
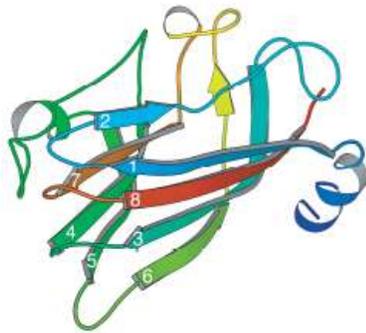
jellyroll
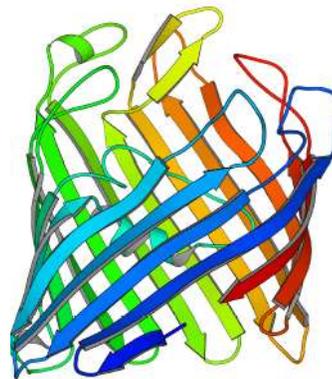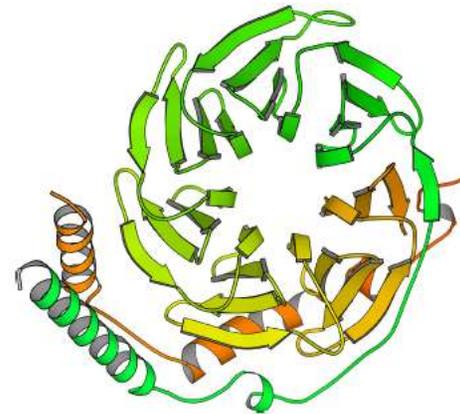
β barrel

β propeller

Structural classification of folds

❑ α/β (alternating α-helices and β-strands)
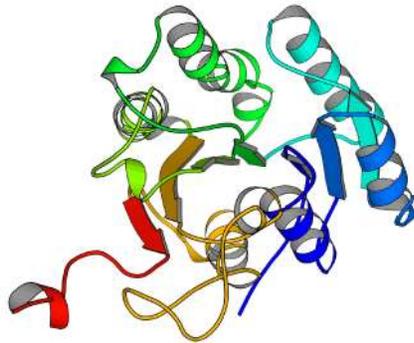
Rossmann
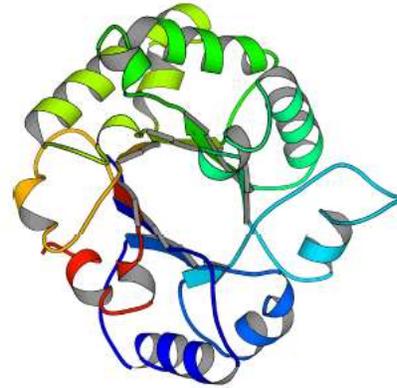
TIM barrel

❑ α+β (α-helices and β-strands occur separately)
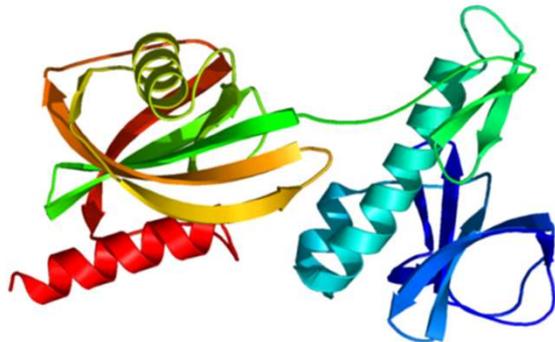
β-Grasp (ubiquitin-like)

Quaternary structure

❑ Association of several protein chains

(monomers/subunits) into oligomers (multimers)

- homomeric protein – from identical monomers

- heteromeric protein – from different types of monomers
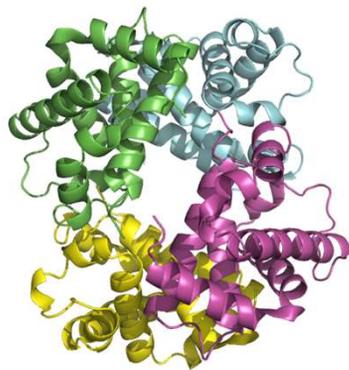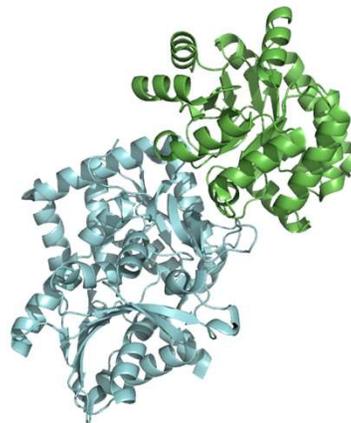
| homotetramer | heterodimer | heterotetramer |
| hemoglobin | tryptophan synthase | immunoglobin |
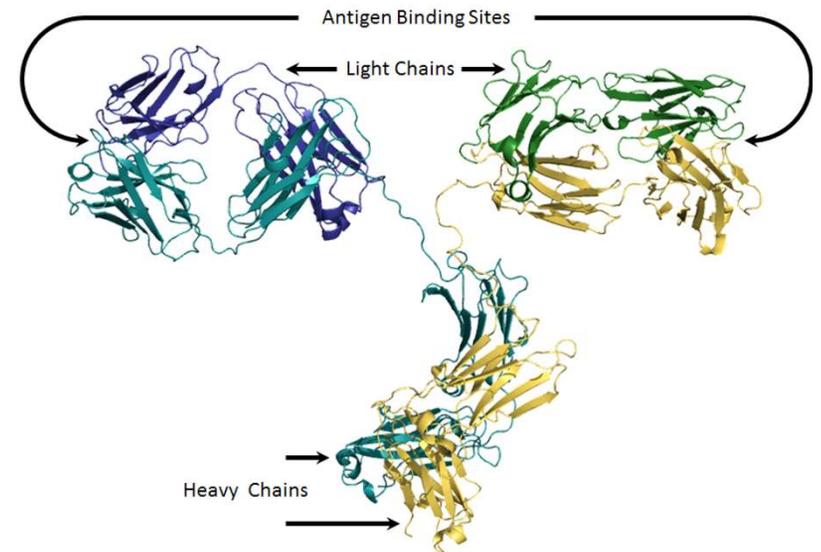
Nucleic acids
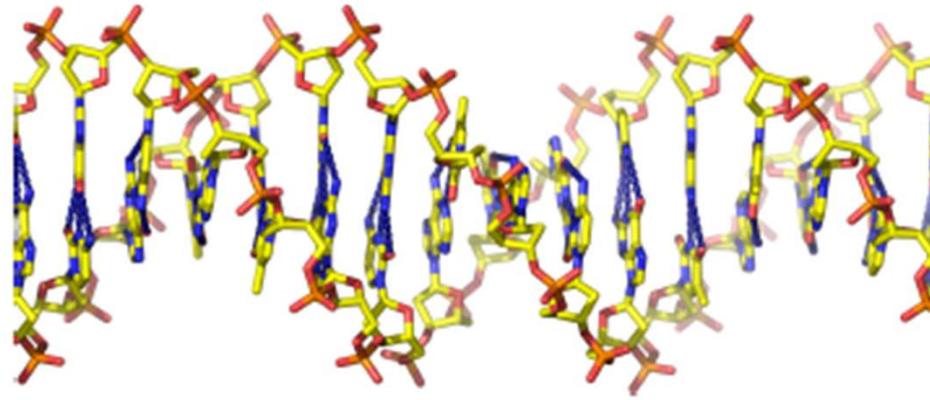
Nucleotides

❑ **Composition**

(a) Nucleotide

Phosphate group is bonded to 5′ carbon of sugar

Nitrogenous base

Phosphate group

5-carbon sugar

Nitrogenous base is bonded to 1′ carbon of sugar

(b) Sugars

Ribose in RNA

Deoxyribose in DNA

(c) Nitrogenous bases

Cytosine (C) Uracil (U) in RNA Thymine (T) in DNA

Pyrimidines

Guanine (G) Adenine (A)

Purines are larger than pyrimidines

Purines
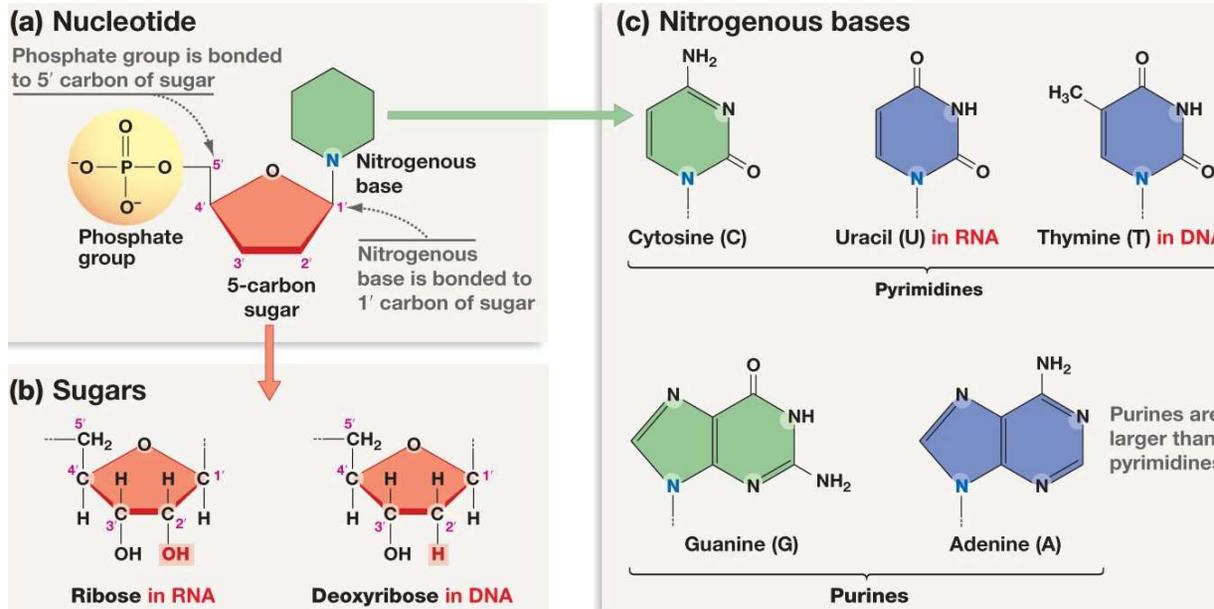
❑ Heterocyclic base

❑ Pentose sugar

❑ Phosphate

❑ DNA bases: **A**, **G**; **C**, **T**

❑ RNA bases: **A**, **G**; **C**, **U**

❑ **Rotation about glycosidic bond**

anti   syn   anti   syn
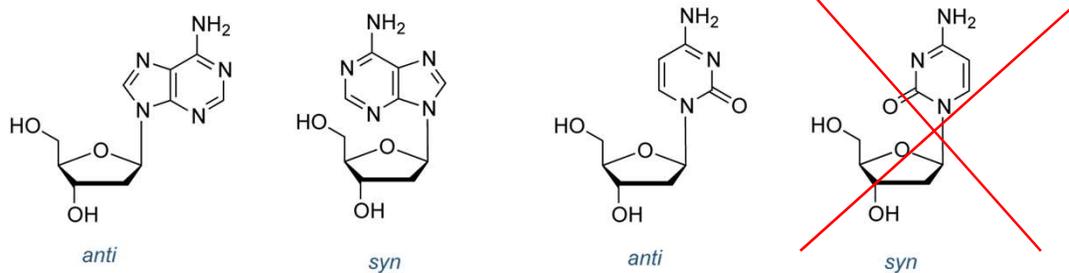
In DNA the *anti* conformation is dominant with rare exceptions

Primary structure

❑ **Linear chain of nucleotides**

   ▪ CGCGAATTCGCG

❑ **Sugar-phosphate backbone**

   ▪ covalent character

   ▪ phosphodiester bond

   ▪ from 5'-end to 3'-end

Primary structure

❑ **Linear chain of nucleotides**

- CGCGAATTCGCG

❑ **Sugar-phosphate backbone**

- covalent character

- phosphodiester bond

- from 5'-end to 3'-end

5' end

G

C

A

Me

T

oligonucleotides dGCAT
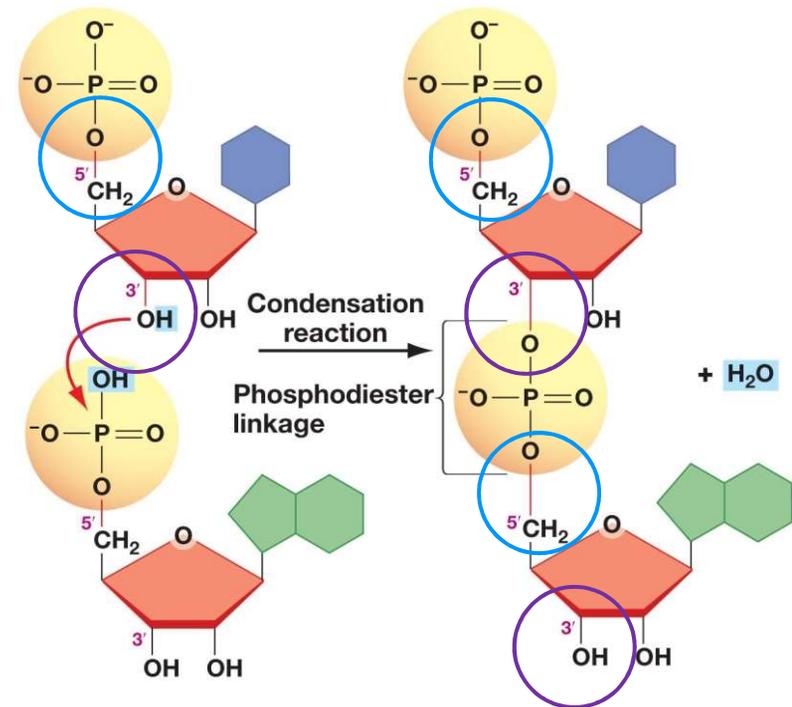
3' end
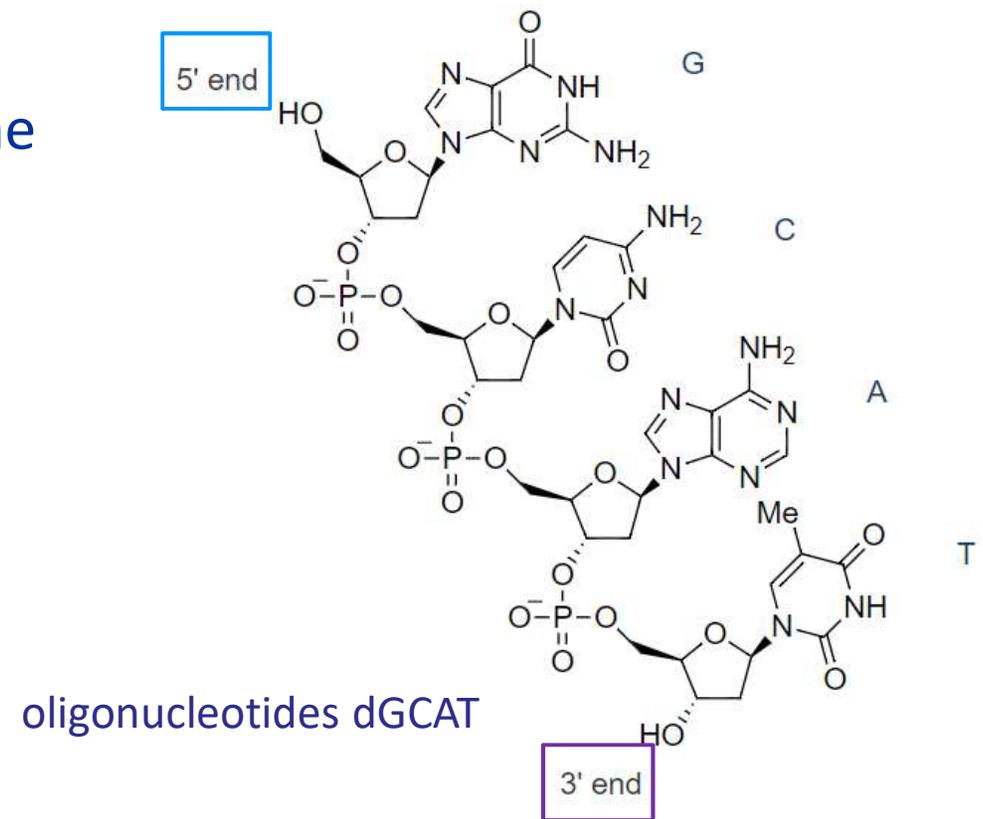
Sugar-phosphate backbone

❑ **Very flexible backbone**

   ▪ six torsion angles
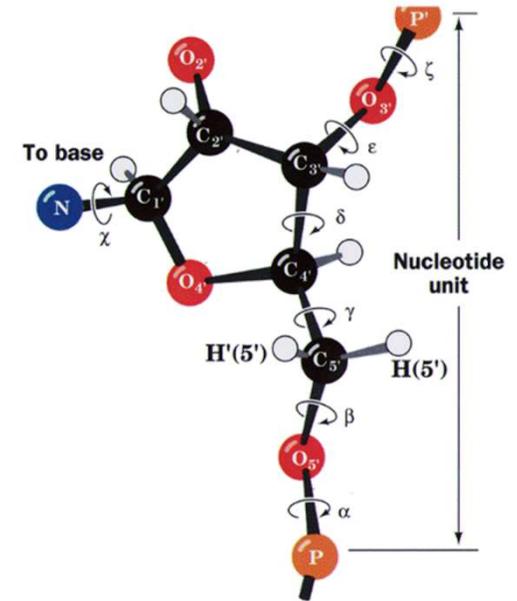
❑ Ribose is not planar → sugar puckering

   ▪ two main types of conformation

C3'-endo (A-Conformation)

C2'-endo (B-Conformation)

2'-deoxyribose (in DNA)
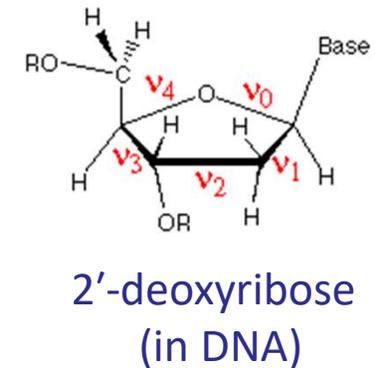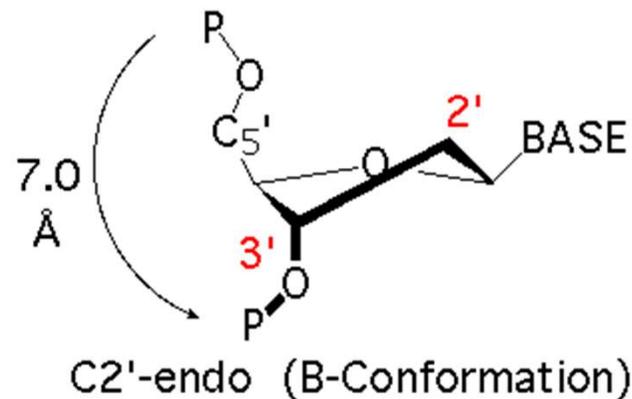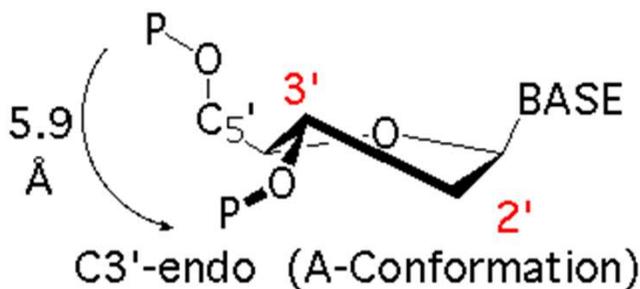
Secondary structure

❑ Governed by hydrogen bonding between bases

→ base pairs
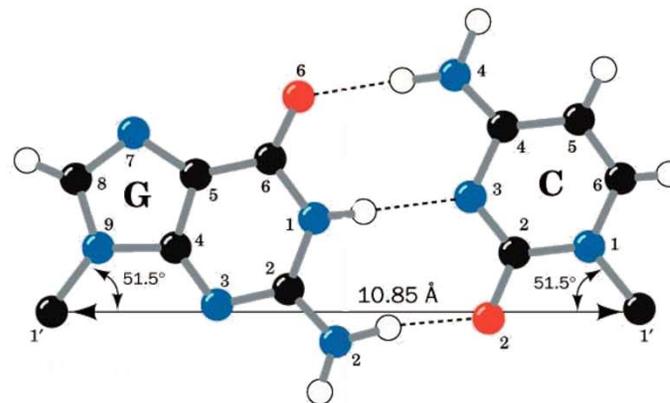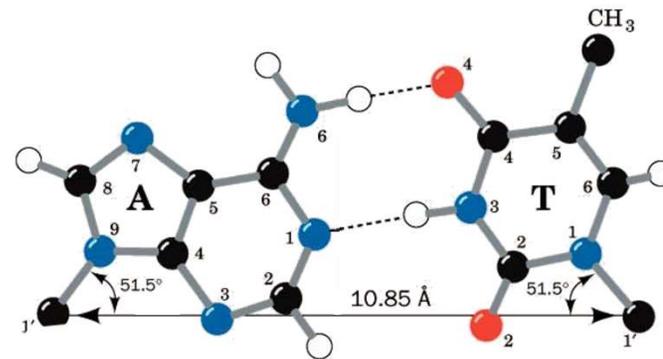
❑ DNA base pairs:

- Adenine, Thymine

- Cytosine, Guanine

❑ RNA base pairs:

- Adenine, Uracil

- Cytosine, Guanine

Secondary structure

- ❑ Leontis /Westhof classification

   - ▪ three base-paring edges

        - ▪ Watson-Crick (WC)

        - ▪ Hoogsteen (H)

        - ▪ sugar (S)

    - ▪ 12 types of base-paring
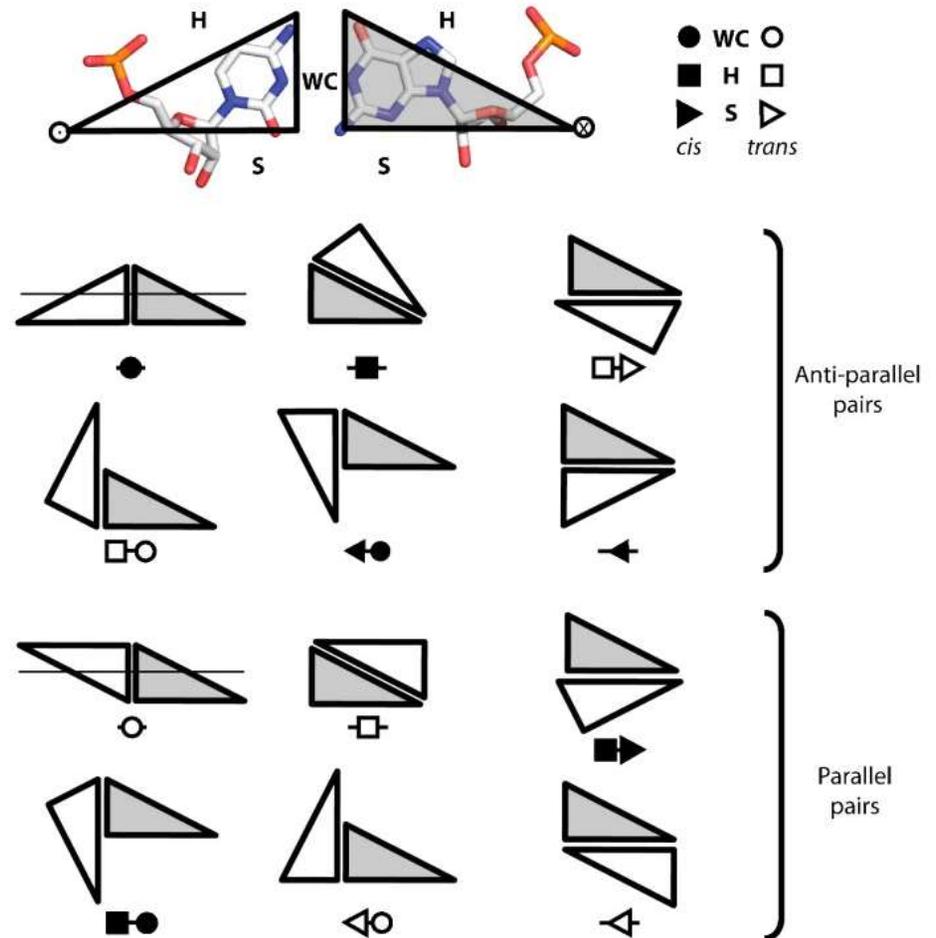
Secondary structure of DNA

Type	A-DNA	B-DNA	Z-DNA
Helix sense	Right	Right	Left
Bases per turn	11	10.5	12
Helical rise per nucleotide (Å)	2.6	3.4	3.7
Sugar pucker	C3'-endo	C2'-endo	C2'-endo C3'-endo

A-DNA
(rare)

B-DNA
(predominant)

Z-DNA

Higher structures of DNA

- ❑ Tertiary structure - mainly canonical B-DNA

- ❑ Quaternary structures - with support of proteins

Nucleosome core particle
10 nm
DNA
10-nm fiber
100 nm
30 nm
30-nm fiber
100 nm

2x histon H2A
2x histon H2B
2x histon H3
2x histon H4

histon H1

DNA
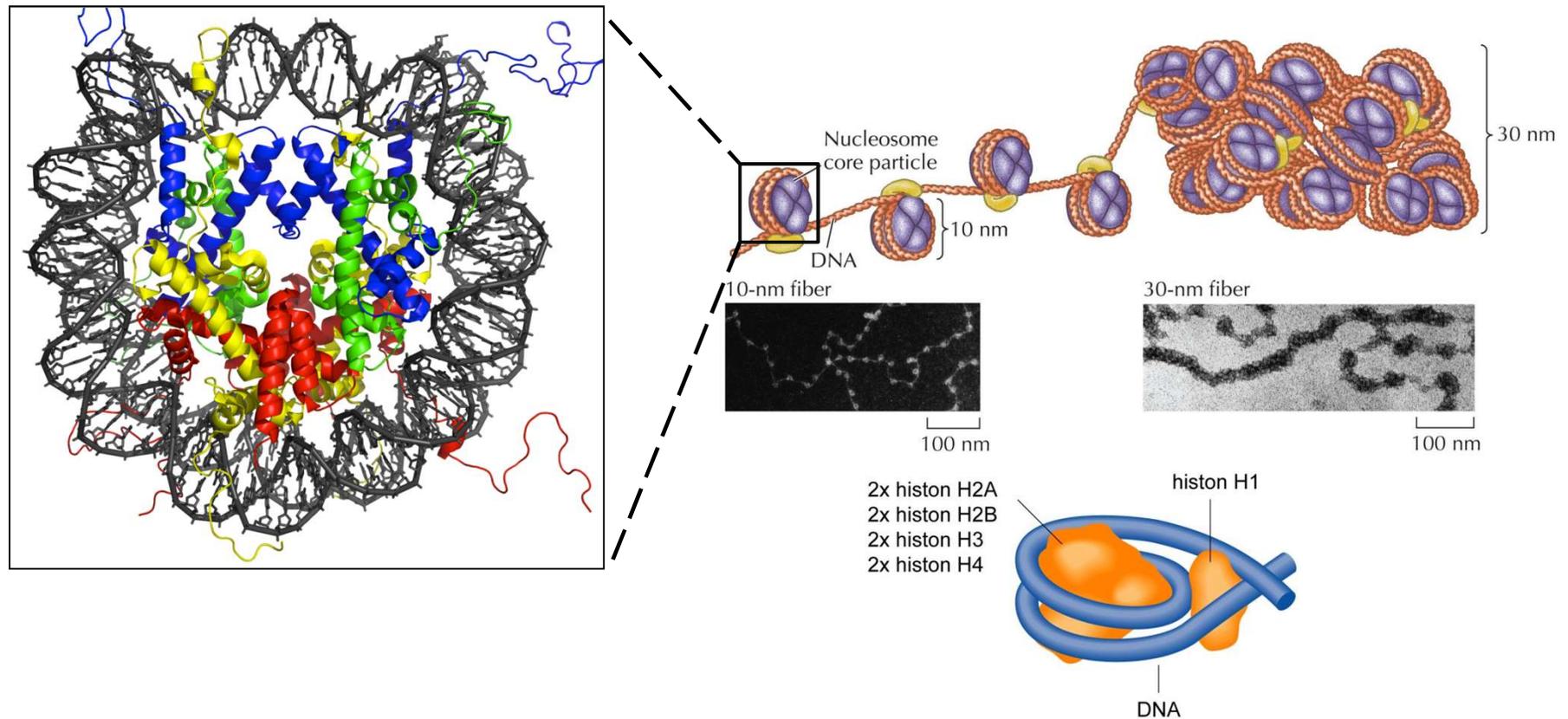
Higher structures of DNA

- ❏ Tertiary structure - mainly canonical B-DNA

- ❏ Quaternary structures - with support of proteins
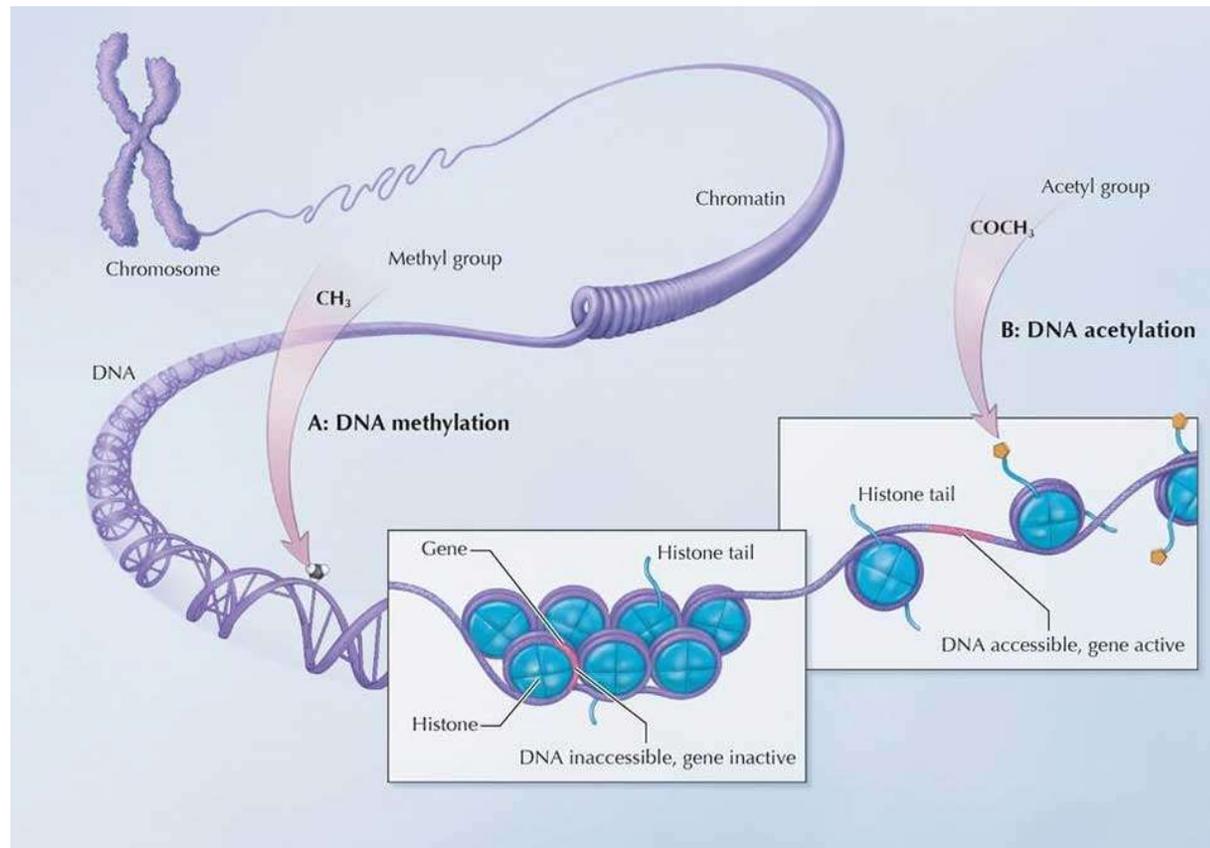
Secondary structures of RNA

❑ A-RNA helix – similar to A-DNA
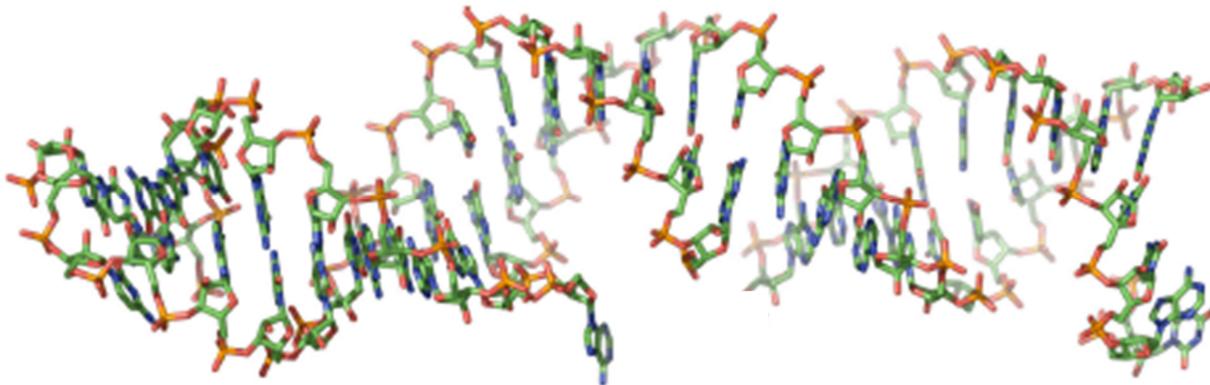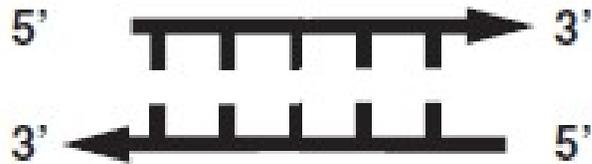
Secondary structures of RNA

❑ Junctions

- regions connecting two or more stems

- two-stem, three-stem and four-stem junction
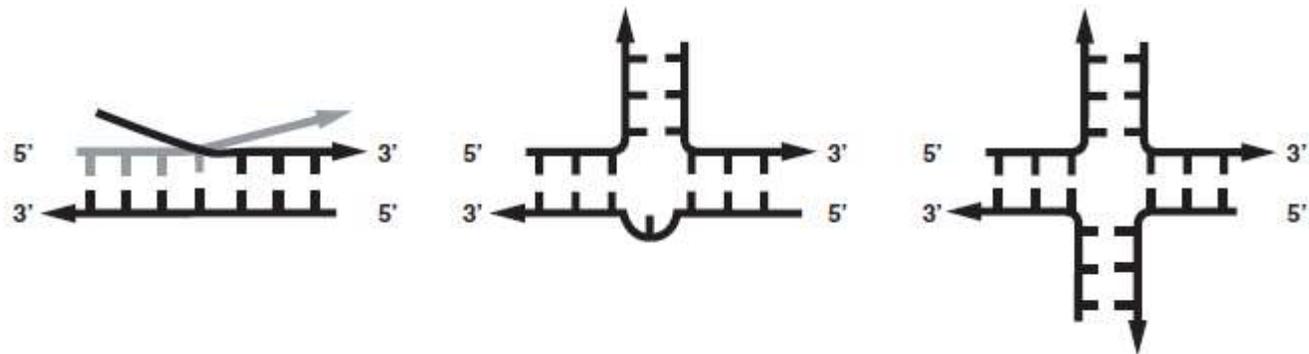
Secondary structures of RNA

❑ Internal loops

- often serve as binding sites for proteins

- many subtypes - e.g.: bulge, K-turn
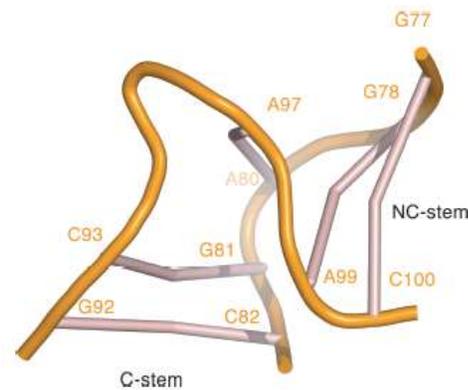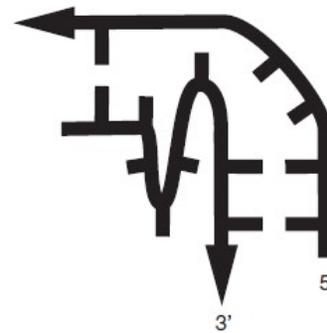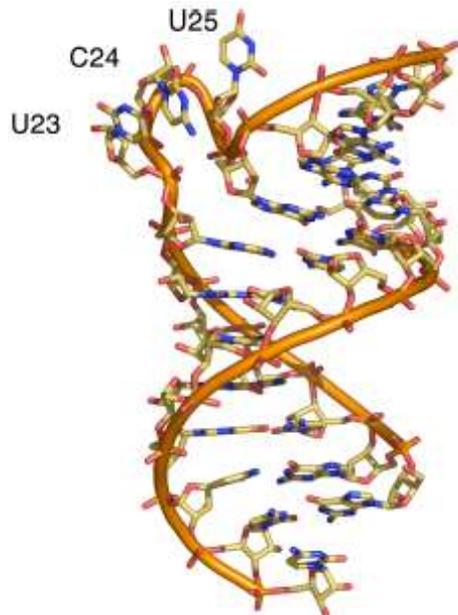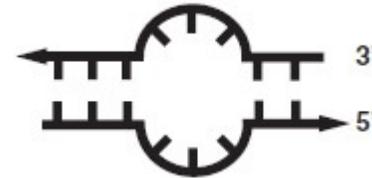
Secondary structures of RNA

❑ Harpin loops

- ▪ sequence inversely self-complementary

- ▪ GGCUGGCUGUUCGCCAGCC
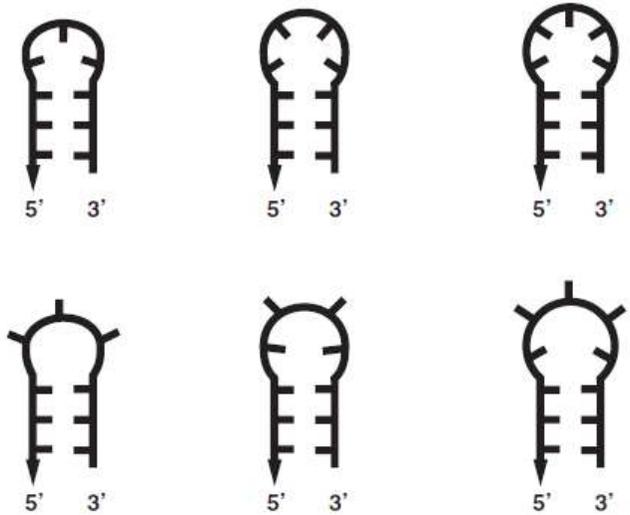
- ▪ many subtypes - e.g.: GNRA, ANYA, UNCG tetraloops
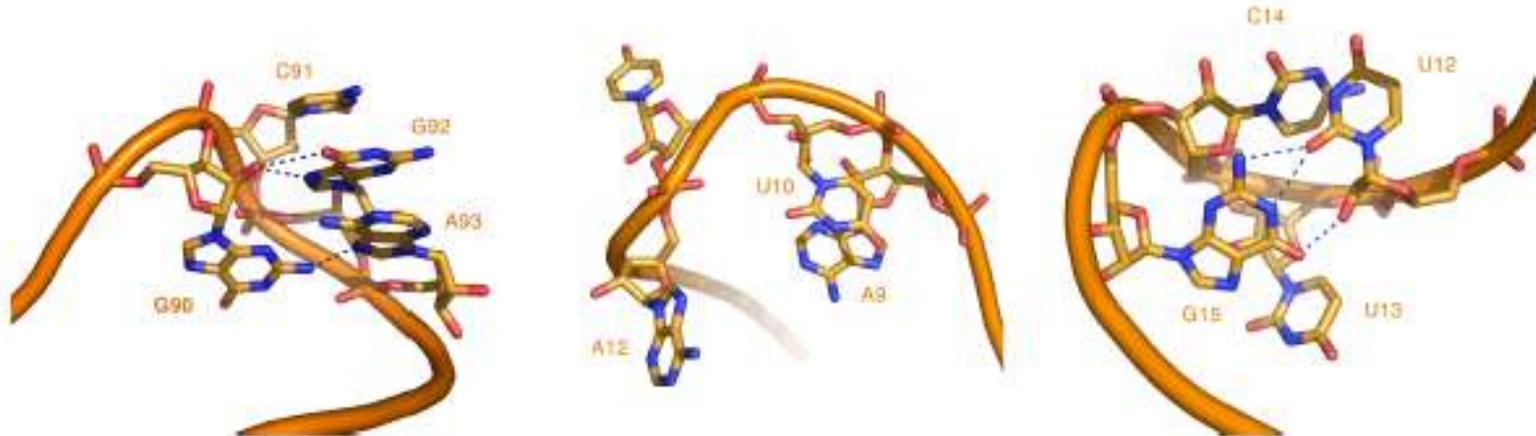
Secondary structures of RNA

❑ Very complex – stem-loop structure
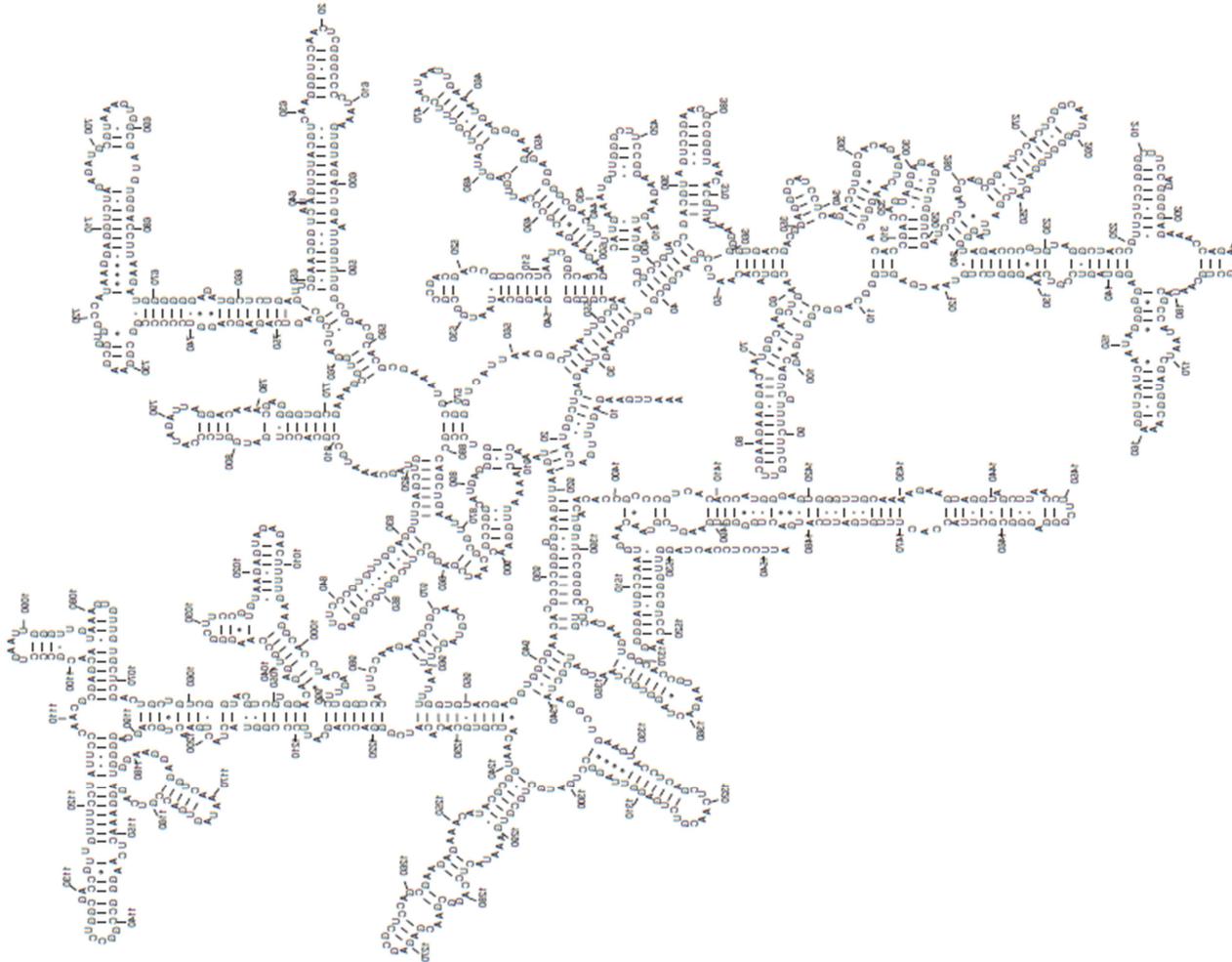
Tertiary structures of RNA

❑ Supersecondary structures (Motifs)

- **pseudoknot**

  - two stem-loop structure
  - loop of the first stem forms part of the second stem
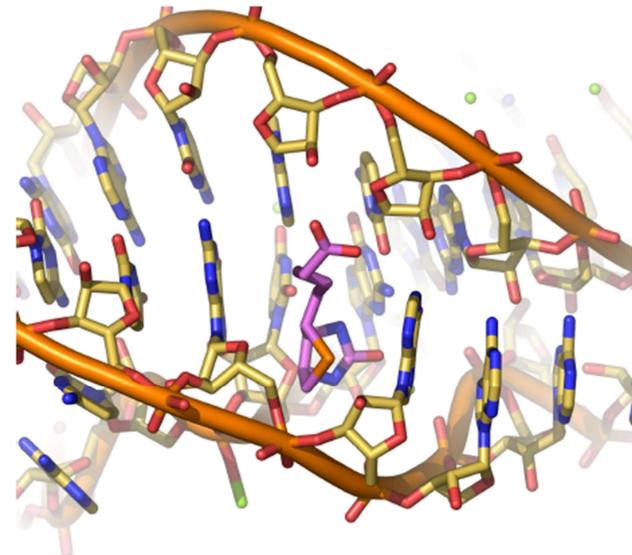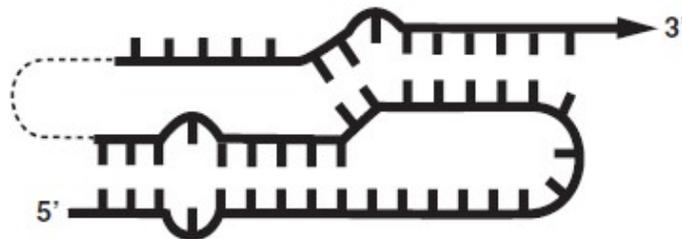
Tertiary structures of RNA

❑ Supersecondary structures (Motifs)

- **A-minor motifs**

  - four subtypes
  - adenines interact with RNA minor grooves
  - involved in the packing of RNA double helices
  - possibly the most important feature forming the tertiary structure of RNAs
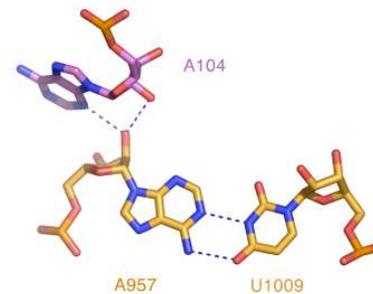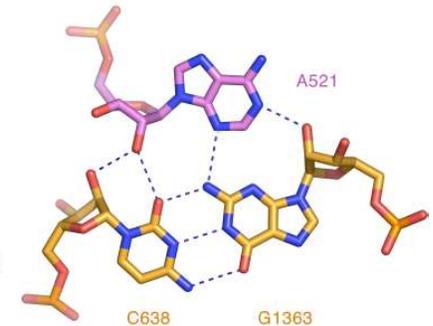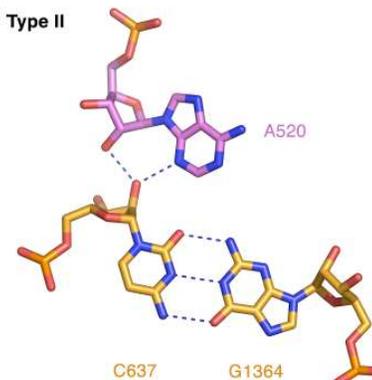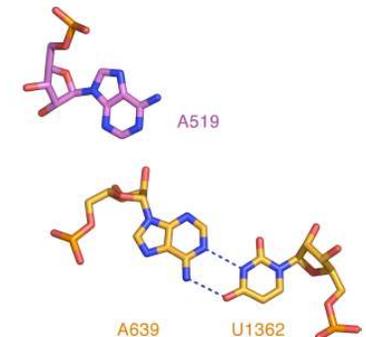
Tertiary structures of RNA

A-RNA dodecamer
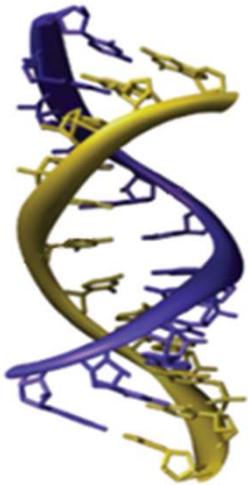
Phenylalanine transfer RNA
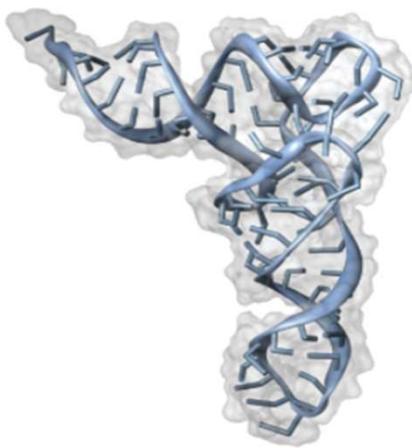
Group I intron ribozyme

Hammerhead ribozyme
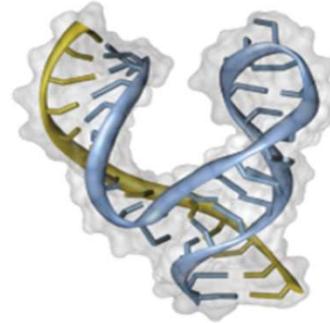
Guanine riboswitch
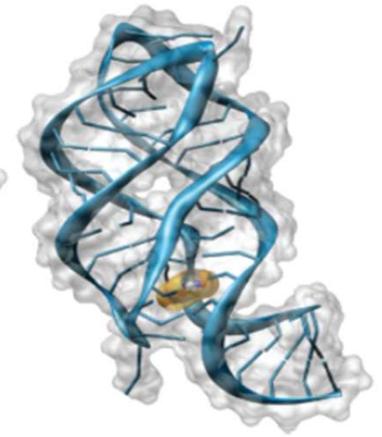
Quaternary structure of RNA

❑ **Association of several chains of RNA**

- ▪ frequently joined with proteins

- ▪ Eukaryotic ribosome - ~ 6800 nt, 79 proteins

https://www.youtube.com/watch?v=Jml8CFBWcDs
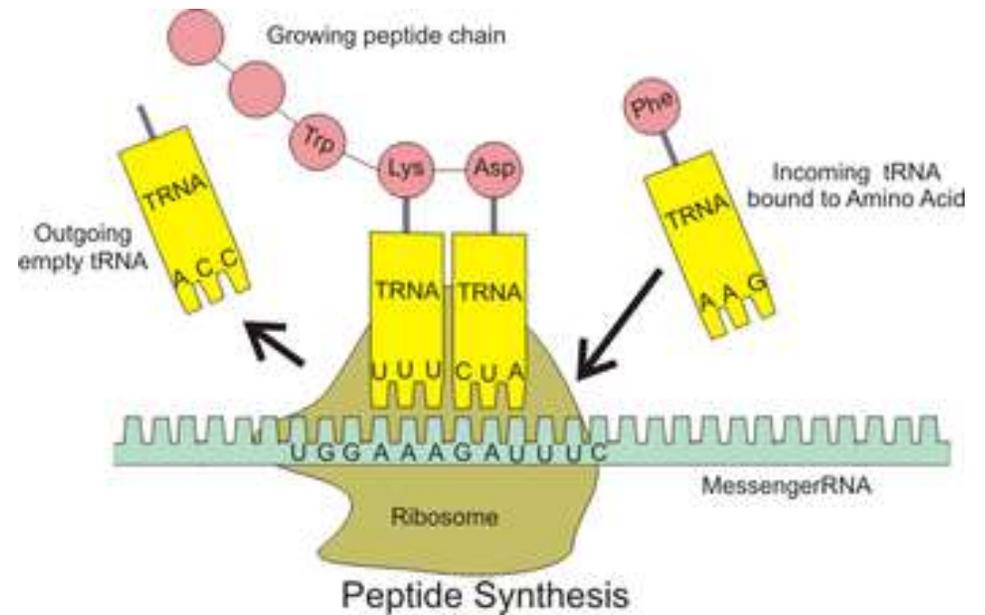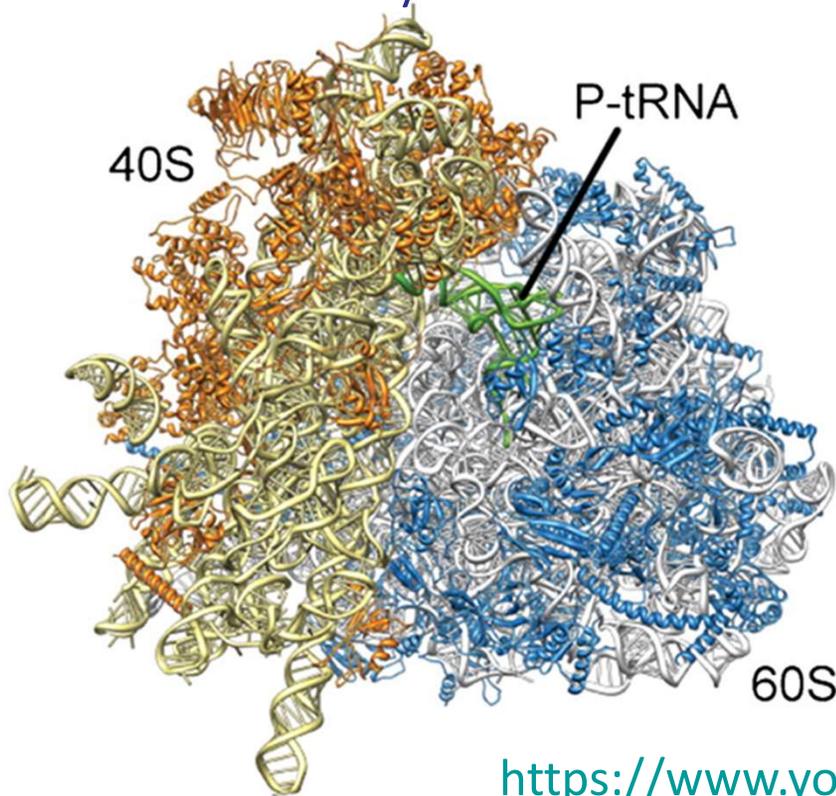
Data formats

❑ different formats are used to represent primary

macromolecular structure data

- PDB

- mmCIF

- PDBML

- …

PDB format

- designed in the early 1970s – first entries of PDB database

- rigid structure of 80 characters per line, including spaces

- still the most widely supported format

- https://www.wwpdb.org/documentation/file-format-content/format33/v3.3.html
- https://www.cgl.ucsf.edu/chimera/docs/UsersGuide/tutorials/pdbintro.html

PDB format

structure annotation

```
HEADER     LYASE (CARBON-CARBON)                    03-JUL-95   1DNP
TITLE      STRUCTURE OF DEOXYRIBODIPYRIMIDINE PHOTOLYASE
... ...
SOURCE     2 ORGANISM_SCIENTIFIC: ESCHERICHIA COLI
KEYWDS      DNA REPAIR, ELECTRON TRANSFER, EXCITATION ENERGY TRANSFER,
KEYWDS     2 LYASE, CARBON-CARBON
... ...
ATOM      21   ND1 HIS A   3     55.365  27.866  62.971  1.00 11.07      N
ATOM      22   CD2 HIS A   3     57.200  28.354  61.894  1.00 13.12      C
ATOM      23   CE1 HIS A   3     56.124  26.783  62.981  1.00 13.03      C
ATOM      24   NE2 HIS A   3     57.243  27.052  62.334  1.00  8.19      N
ATOM      25   N   LEU A   4     55.580  32.694  59.656  1.00 12.61      N
ATOM      26   CA  LEU A   4     54.799  33.803  59.113  1.00 11.56      C
ATOM      27   C   LEU A   4     53.552  33.269  58.374  1.00  7.76      C
ATOM      28   O   LEU A   4     53.650  32.363  57.532  1.00  6.99      O
ATOM      29   CB  LEU A   4     55.656  34.683  58.174  1.00  9.03      C
ATOM      30   CG  LEU A   4     54.946  35.887  57.518  1.00  2.00      C
ATOM      31   CD1 LEU A   4     54.623  36.920  58.550  1.00  6.21      C
... ...
HETATM  7641  AN7 FAD B 472     27.855  78.556  29.073  1.00  4.55      N
HETATM  7642  AC5 FAD B 472     28.524  78.026  27.955  1.00  2.00      C
HETATM  7643  AC6 FAD B 472     29.848  77.609  27.724  1.00  3.40      C
HETATM  7644  AN6 FAD B 472     30.787  77.757  28.664  1.00  6.22      N
```

amino acid field

cofactor filed

atom number / atom name / residue name / polypeptide chain identifier / residue number / x, y, z coordinates / occupancy / temperature factor / atom type

- https://www.wwpdb.org/documentation/file-format-content/format33/v3.3.html
- https://www.cgl.ucsf.edu/chimera/docs/UsersGuide/tutorials/pdbintro.html

PDB format

- atomic coordinates

- chemical and biological features

- experimental details of the structure determination

- structural features

  - secondary structure assignments

  - hydrogen bonding

  - biological assemblies

  - active sites

  - ...

- https://www.wwpdb.org/documentation/file-format-content/format33/v3.3.html
- https://www.cgl.ucsf.edu/chimera/docs/UsersGuide/tutorials/pdbintro.html

Primary structural databases

- Worldwide Protein Data Bank (wwPDB)

  http://www.wwpdb.org/

- RCSB Protein Data Bank (RCSB PDB):

  http://pdb.rcsb.org

- Nucleic Acid Database (NDB)

  http://ndbserver.rutgers.edu/

- Biological Magnetic Resonance Data Bank (BMRB)

  http://www.bmrb.wisc.edu/

- Electron Microscopy Data Bank (EMDB)

  http://www.emdatabank.org/

- Cambridge Structural Database (CSD)

  http://www.ccdc.cam.ac.uk/products/csd/

References

- Gu, J. & Bourne, P. E. (2009). **Structural Bioinformatics, 2nd Edition**, Wiley-Blackwell, Hoboken.

- Liljas, A. *et al*. (2009). **Textbook Of Structural Biology**, World Scientific Publishing Company, Singapore.

- Schwede, T. & Peitsch, M. C. (2008). **Computational Structural Biology: Methods and Applications**, World Scientific Publishing Company, Singapore.

- Schaeffer, R.D & Daggett, V. (2011). Protein folds and protein folding. *Protein Engineering, Design & Selection* **24**:11–19.