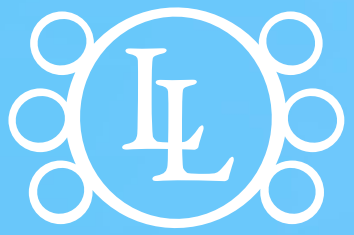


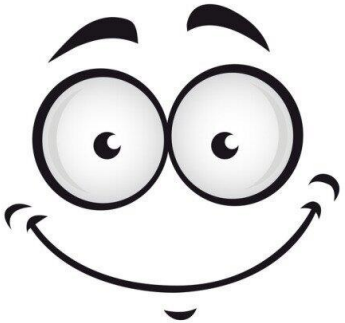
Structure of biomolecules

Outline

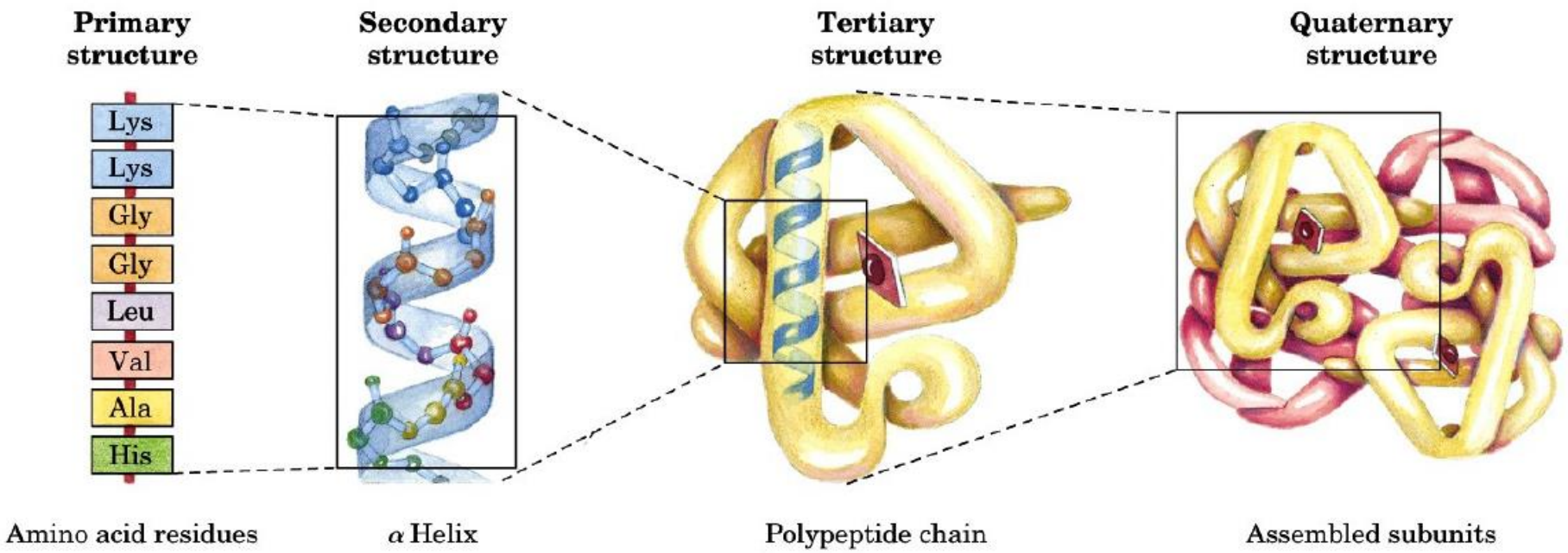
- Proteins
  - Primary structure
  - Secondary structure
  - Tertiary structure
  - Motifs and folds
  - Quaternary structure
- Nucleic acids
  - Main types of structures
- Primary structural databases
- Structural data formats
  - PDB and mmCIF formats

Protein structure

Structure of proteins are…

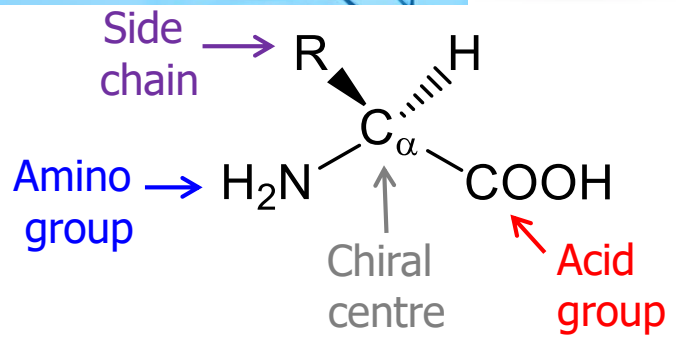


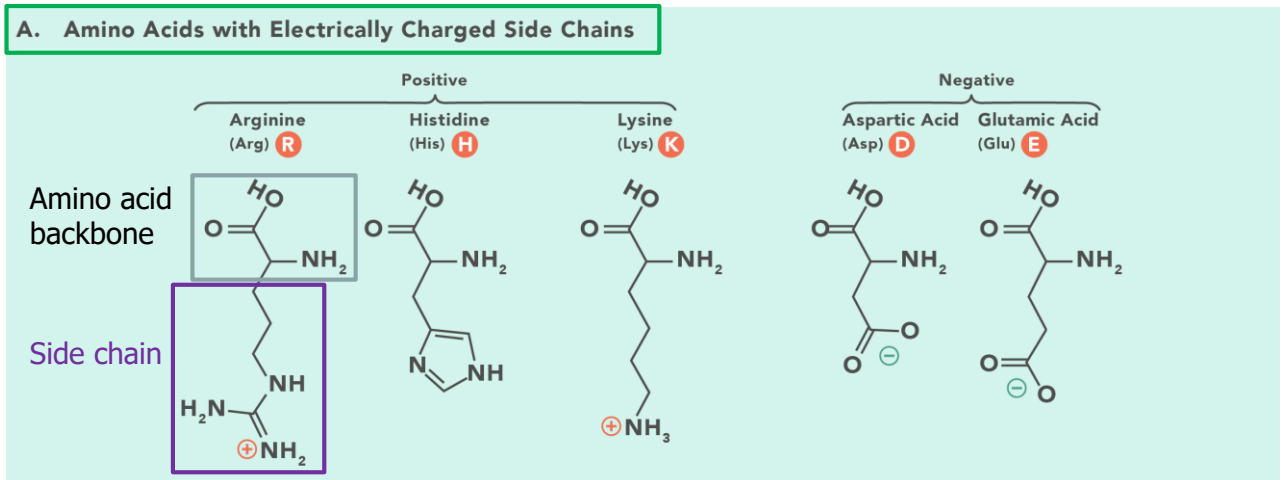
Hierarchy of protein structure

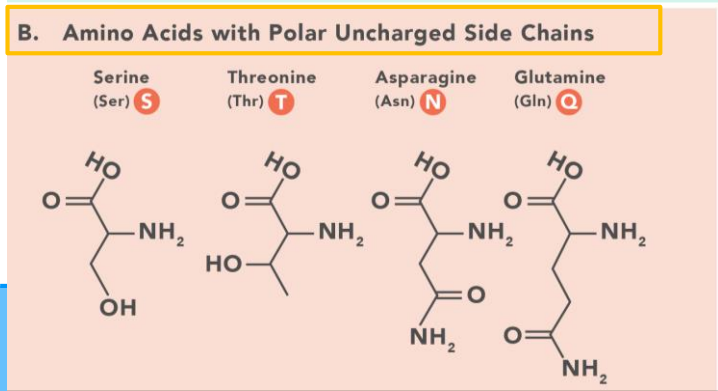


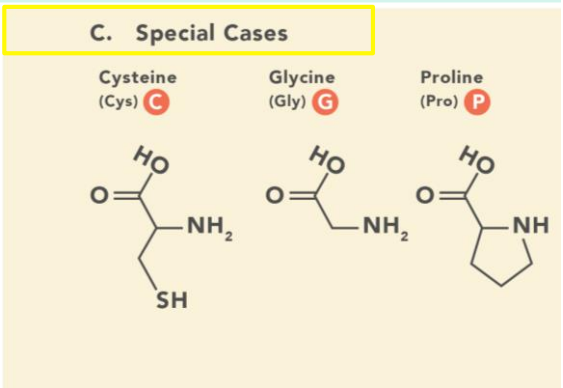
Amino acids

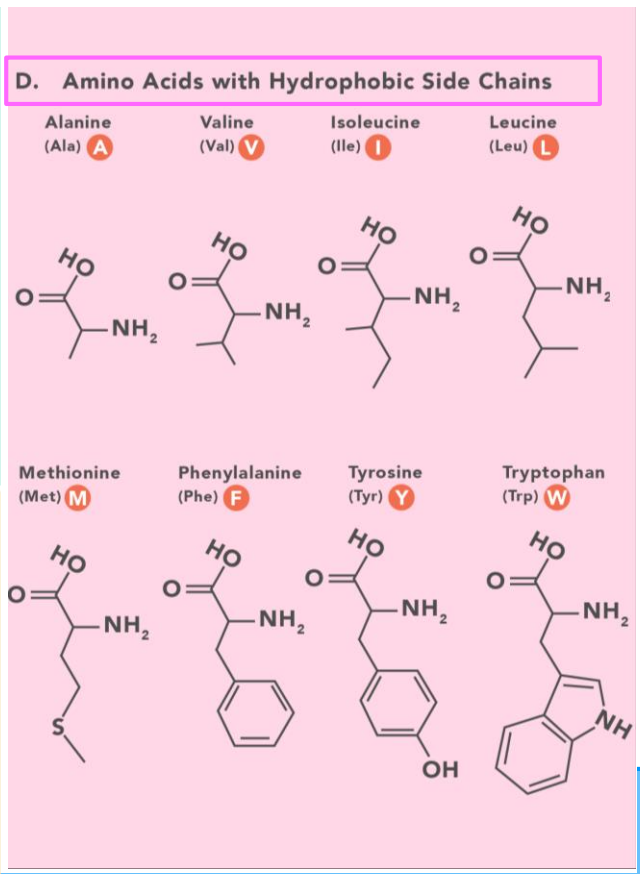
- 20 L-amino acids (natural)
- Side chains
  - Charged, polar, hydrophobic











Primary structure

- Linear chain of amino acid residues

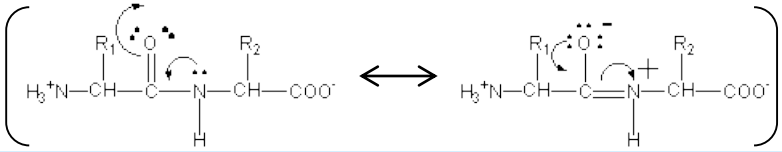
MSLGAKPFGEKKFIEIKGRRMAYIDEGTGDPILFQHGNPTSSYLWRNIM

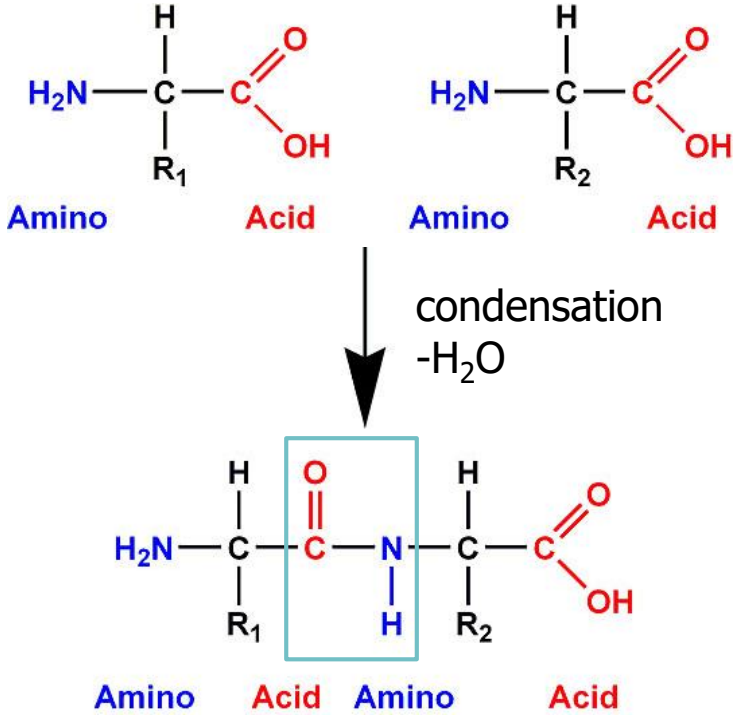
N-terminus

C-terminus

- Protein backbone
  - From N-terminus to C-terminus
  - Connected by covalent bonds
- Peptide bond (amide bond)
  - Partial double bond character

→ Planar geometry



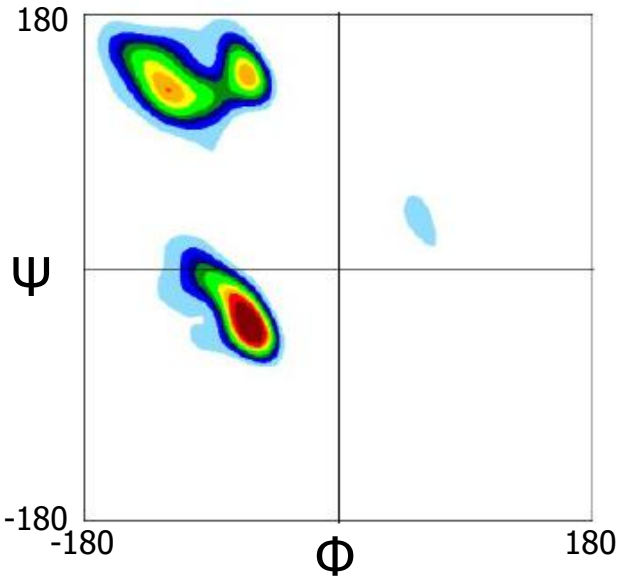


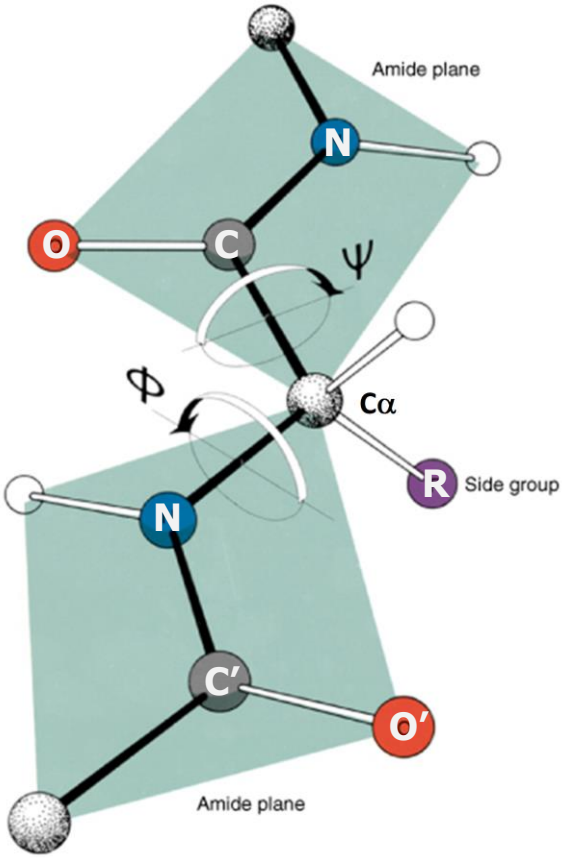
Geometry of protein backbone

- Conformation of the peptide chain
  - Defined by Φ (phi) and Ψ (psi) dihedral angle

- Ramachandran plot (Φ, Ψ)

→ The majority of proteins follow this distribution





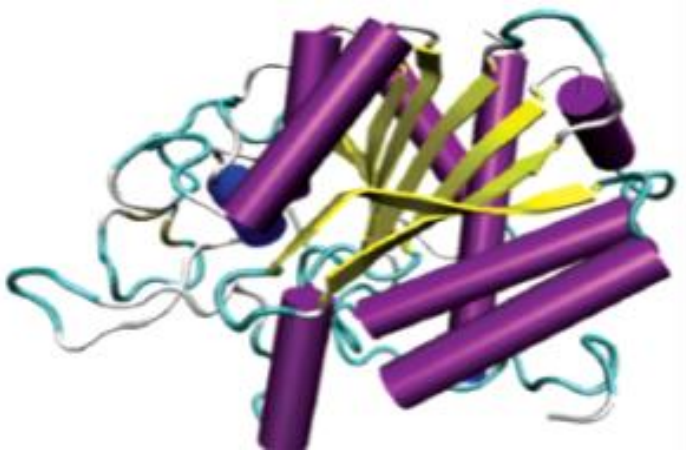
φ (phi) = dihedral angle {C' − N − C_α − C}
ψ (psi) = dihedral angle {N − C_α − C − N}

Secondary structure

- Local three-dimensional structure of polypeptide chain
- Governed by hydrogen bonding between backbone atoms
- Types of structures
  - Helices
  - β-Structures
  - Loops and coils

Helices, β-Structures: Regular patterns

Loops and coils: - Irregular patterns



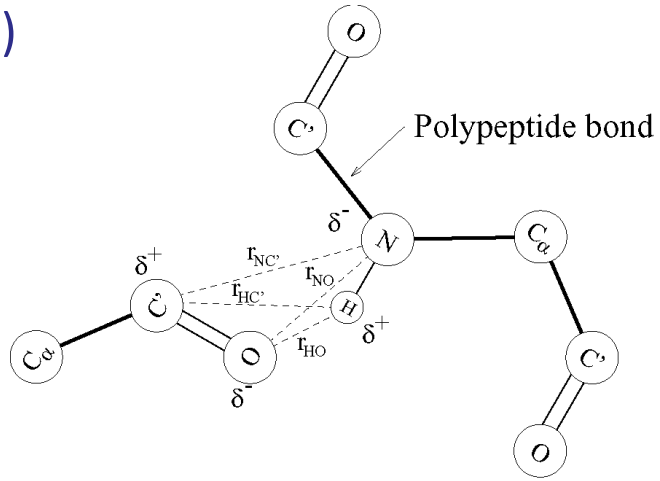
Secondary structure

- 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

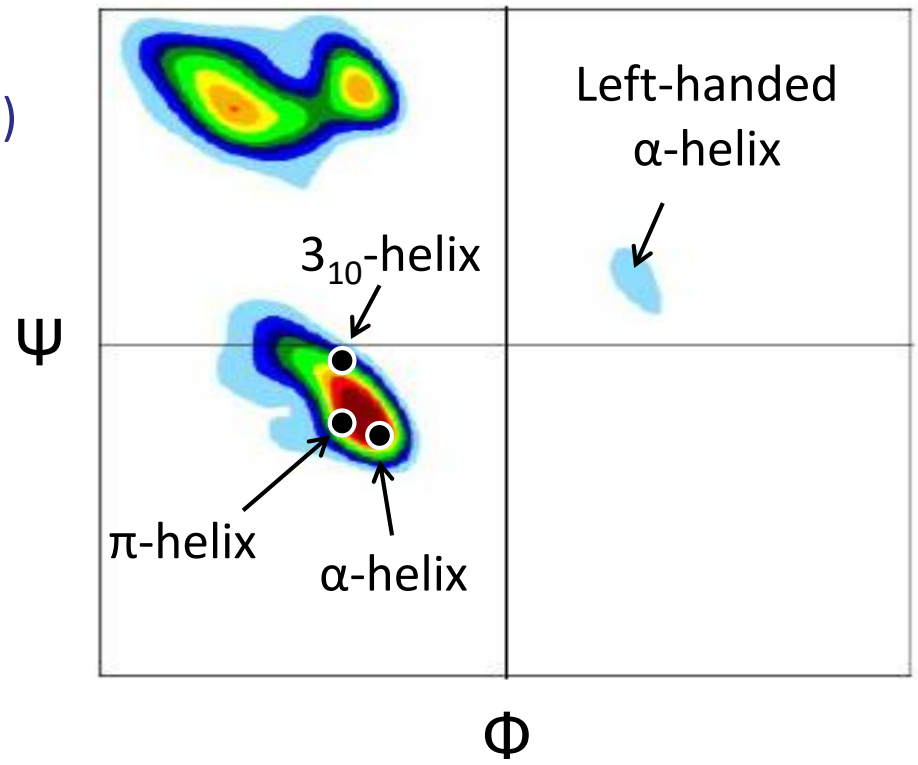
    (3 major types: helices, strands and loops)

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



Helices

- Types of helices
  - 3.6_{13} helix (α-helix) – most common
  - 3_{10} helix – less frequent, end of α-helices
  - 4.1_{16} helix (π-helix) (rare)
  - Left-handed helix (very rare)
  - → Represented by helical cartoons or cylinders

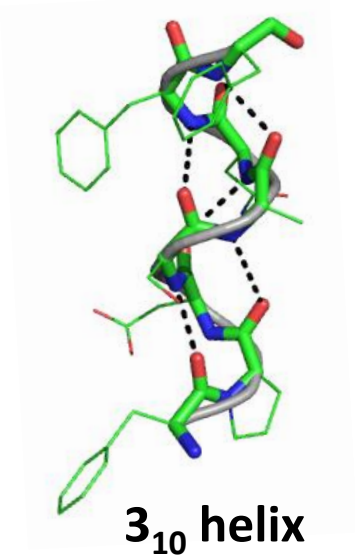


- Right-handed (mostly)
- Hydrogen bonding
  - Within a single chain

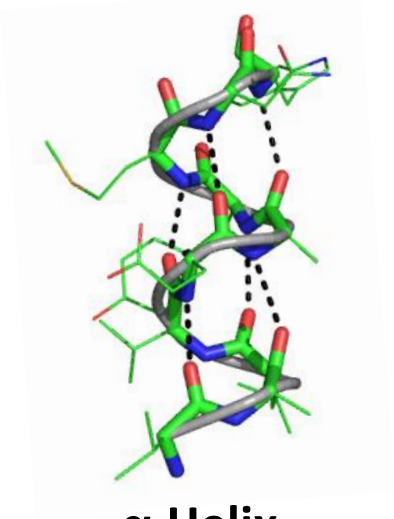
Helices

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

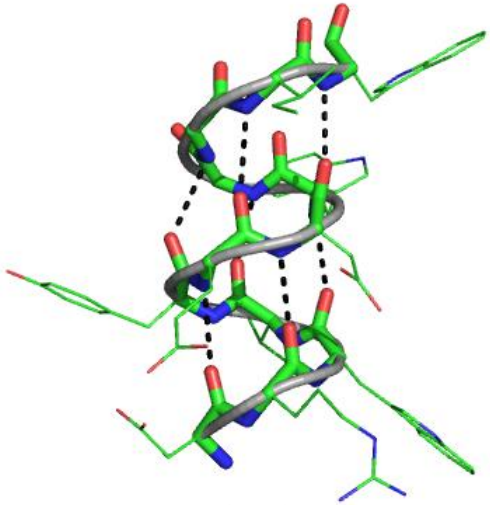
H-bonds



3_{10} helix



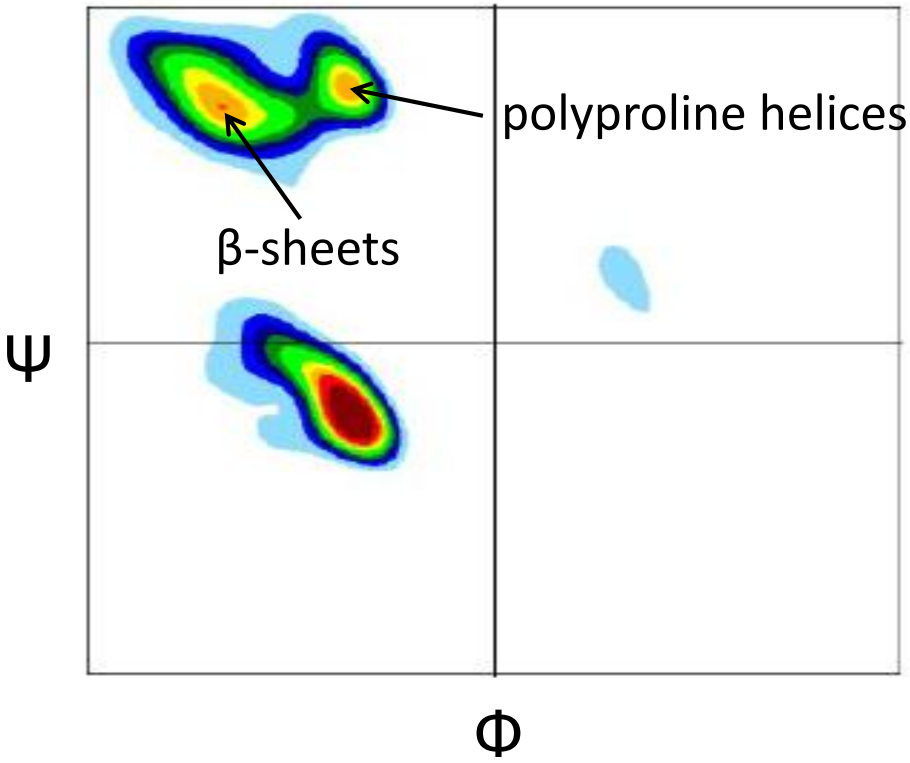
α Helix



π helix

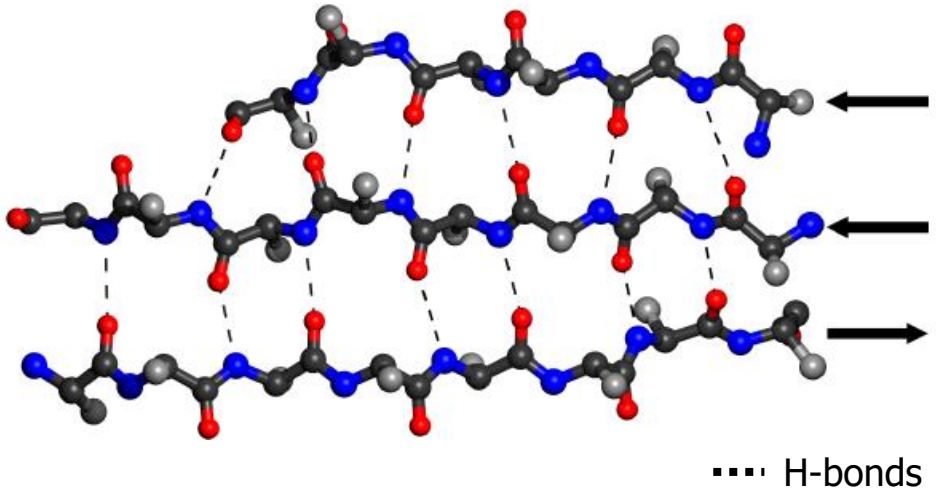
β-structures

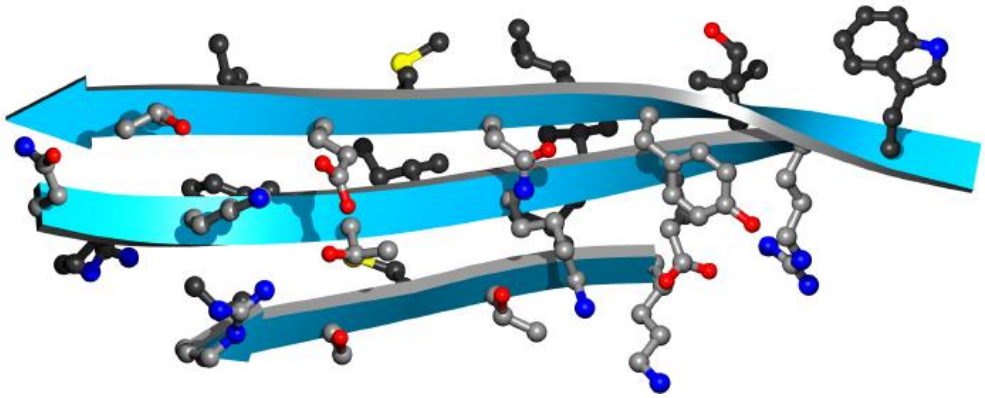
- Types of typical β-structures
  - β-sheets
  - β-turns
  - β-bulge
  - Polyproline helices
- Hydrogen bonding
  - Between adjacent chains



β-structures

- Types of β-sheets
  - Parallel
  - Antiparallel (stronger)
  - Mixed
  - → Represented by ribbons with arrows indicating the sequence direction

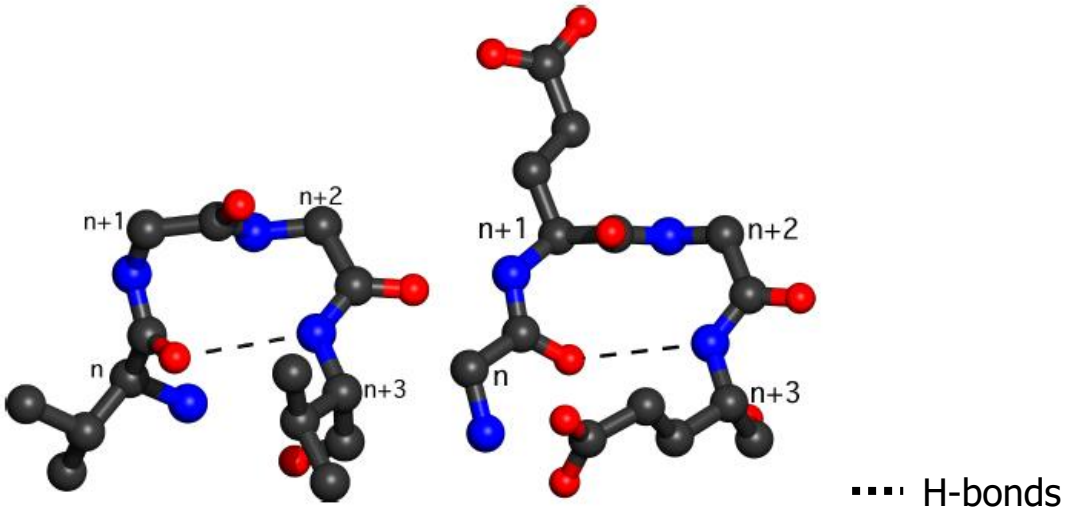




- Side-chains
  - Towards the sides of the sheets

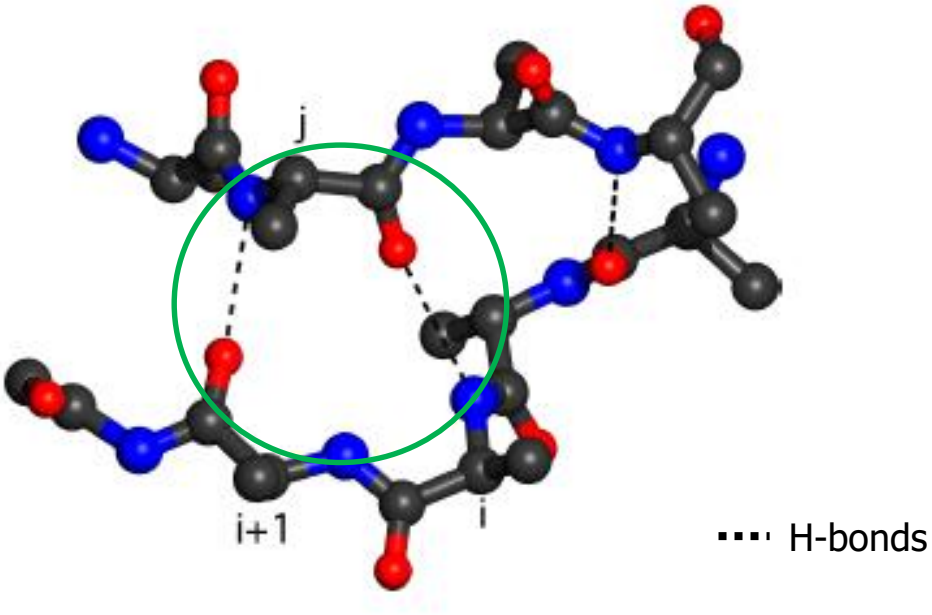
β-structures

- 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



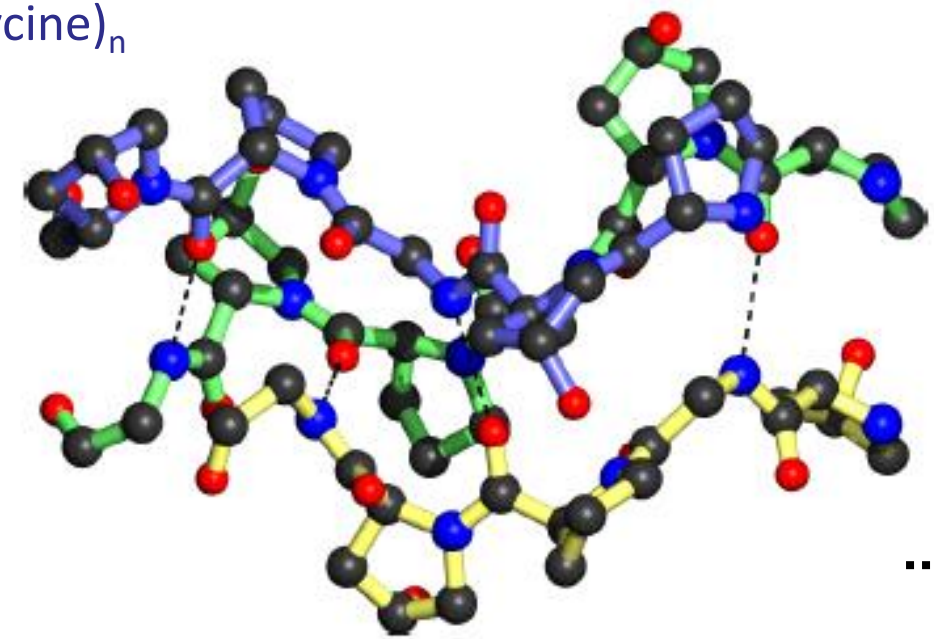
β-structures

- β-bulge
  - Frequently occurs in antiparallel β-sheets
  - Disrupts ideal H-bonding pattern
  - Increases twists of a sheet



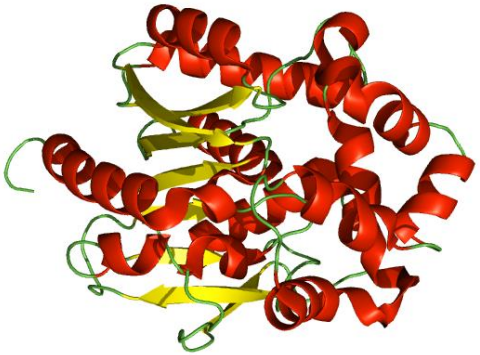
β-structures

- Polyproline helices
  - Typical in collagen and other strong fibers
  - Left-handed triple-stranded helix (unlike most of other helices)
  - Composed of three chains of repetitive sequence (Proline-Hydroxyprolin-Glycine)$_n$

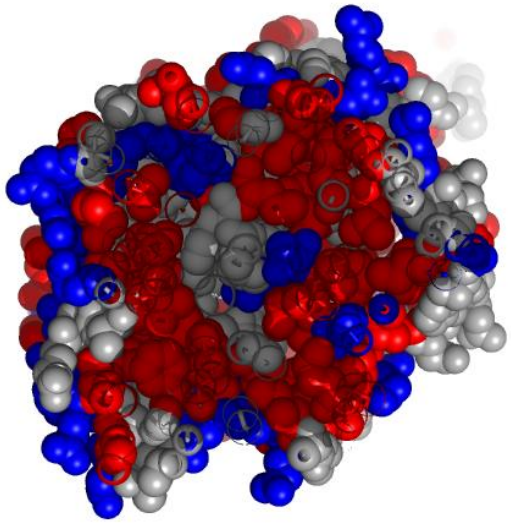


Tertiary structure

- Global three-dimensional structure of protein



- Governed mainly by hydrophobic interactions involving side chains of amino acid residues



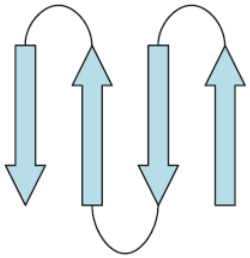
Tertiary structure

- 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 independently)
- Fold
  - General architecture of protein
  - Type of protein structure

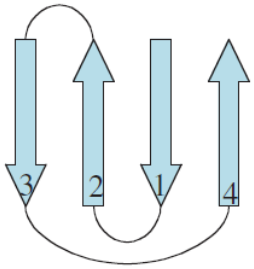
Protein motifs

- β-harpin

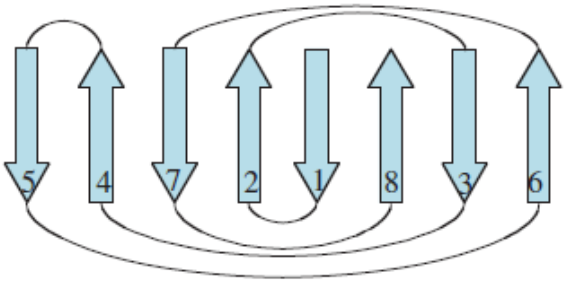
- β-meander



- Greek key

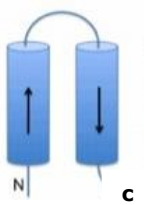


- Jellyroll

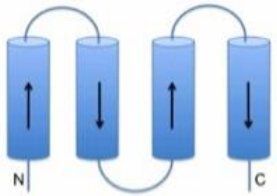


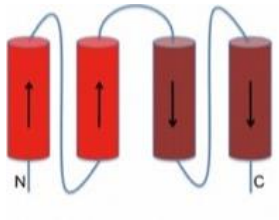
Protein motifs

- Helix-turn-helix

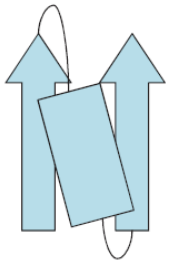


- Helix bundle



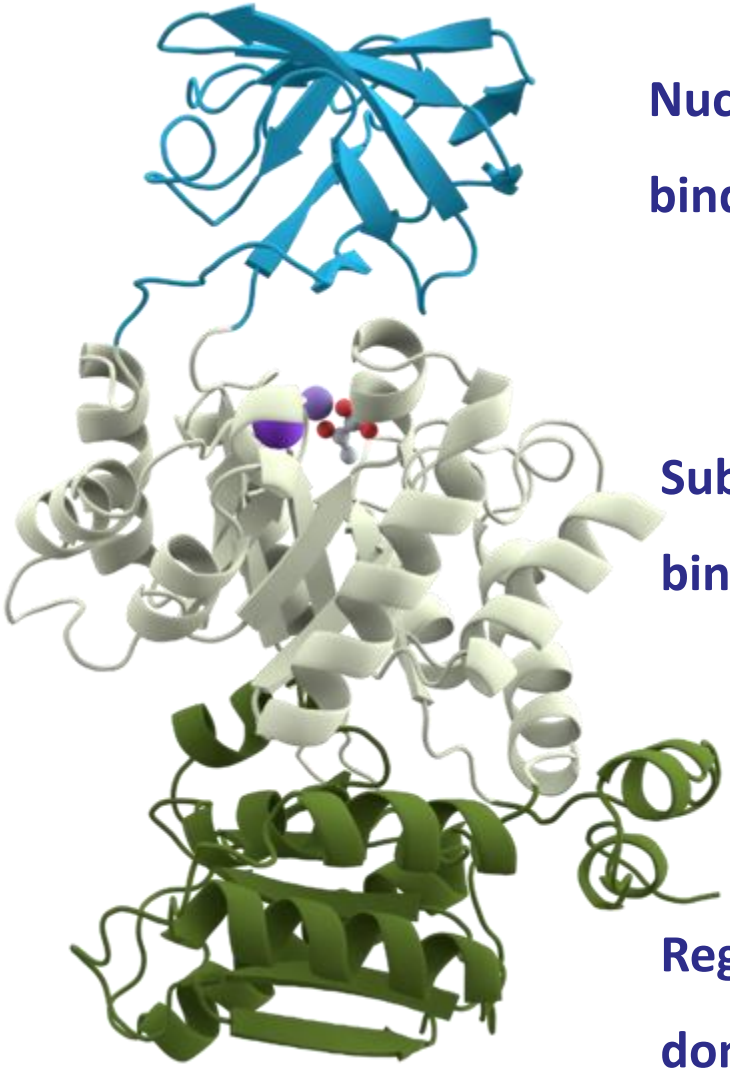


- βαβ unit

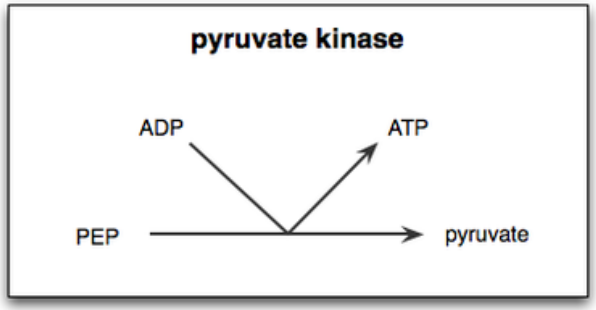


Protein domains

- Parts of tertiary structure
  - Separate folding
  - Independent structure
  - Usually up to 200 residues

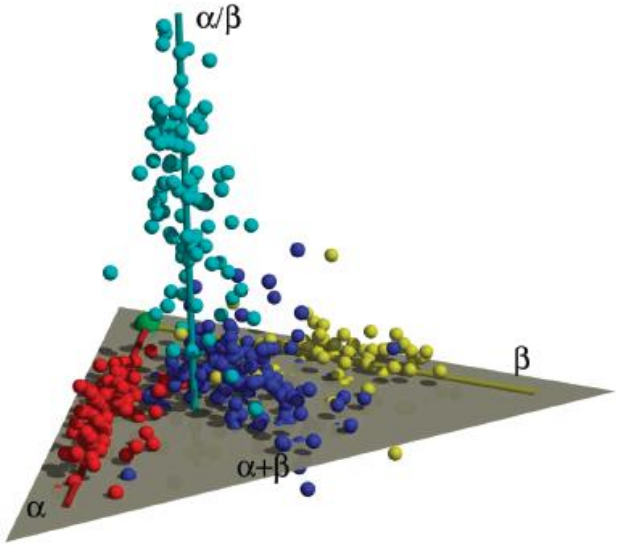


Pyruvate kinase in glycolysis



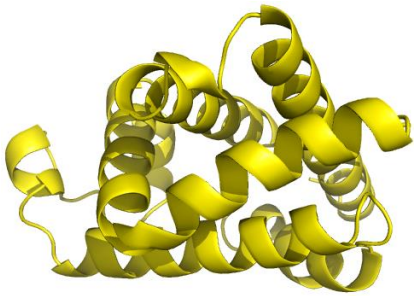
Protein folds

- 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

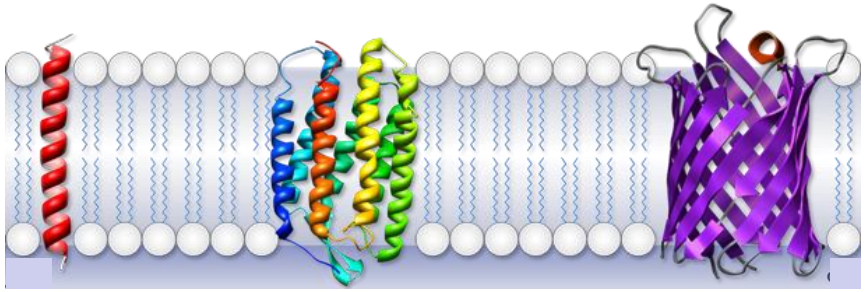


Biochemical classification of folds

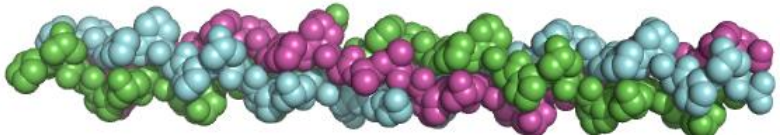
- Globular proteins



- Membrane proteins

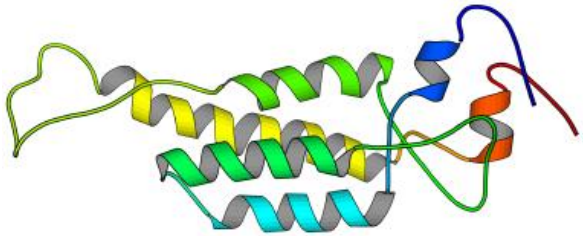


- Fibrous proteins

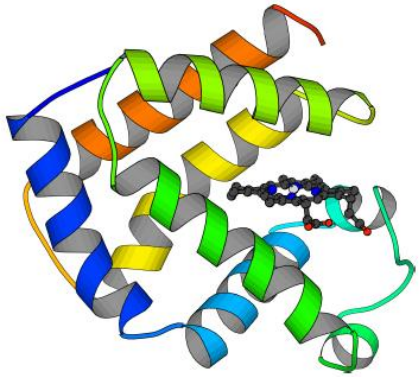


Structural classification of folds

- All-α (entirely α-helices)

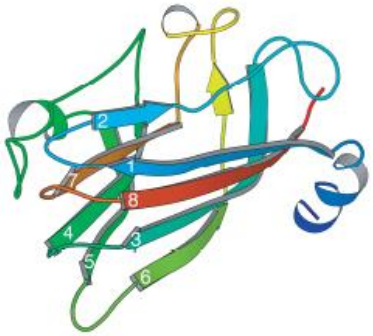


Up-and-down bundle

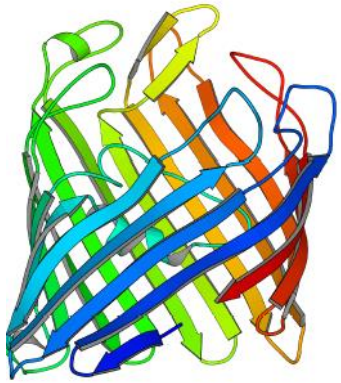


Globin-like

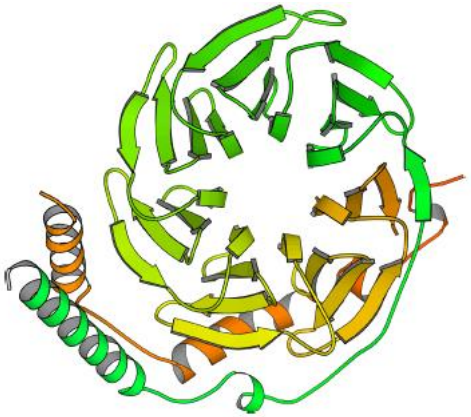
- All-β (entirely β-strands)



Jellyroll



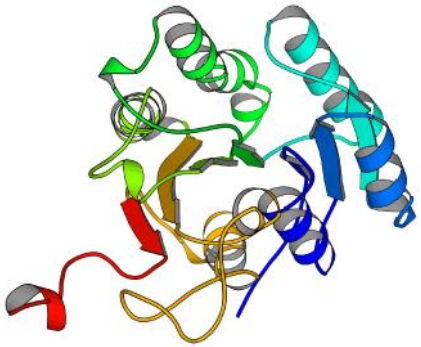
β barrel



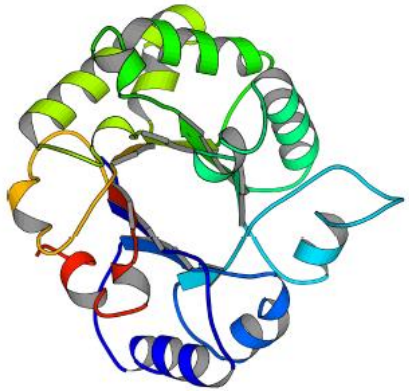
β propeller

Structural classification of folds

- α/β (sequence alternates between α-helices and β-strands)

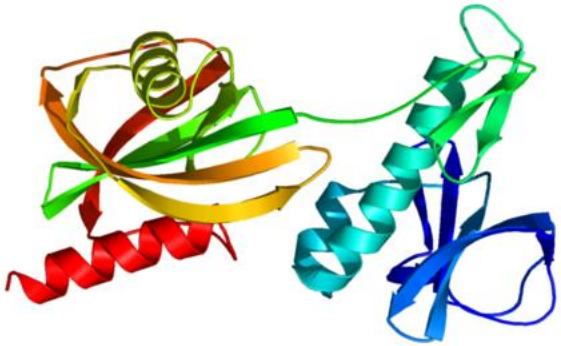


Rossmann



TIM barrel

- α+β (α-helices and β-strands occur separately in sequence)



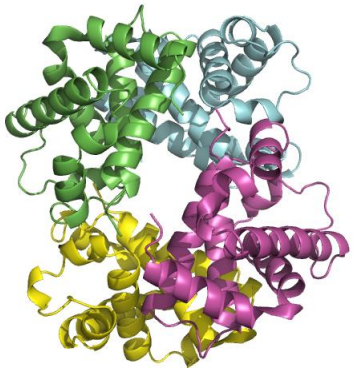
β-Grasp (ubiquitin-like)

Quaternary structure

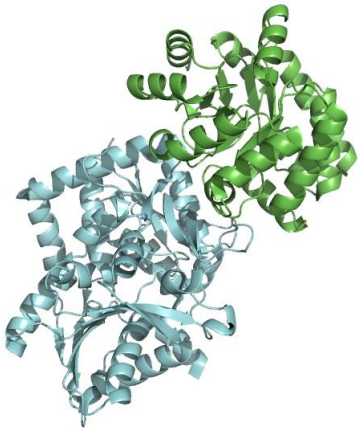
- Association of several protein chains

(monomers/subunits) into oligomers (multimers)

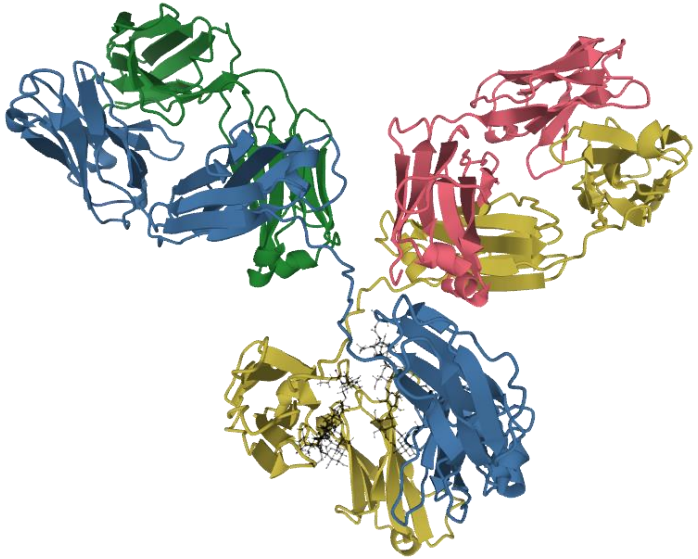
  - Homomeric protein – from identical monomers
  - Heteromeric protein – from different types of monomers



Homotetramer
hemoglobin

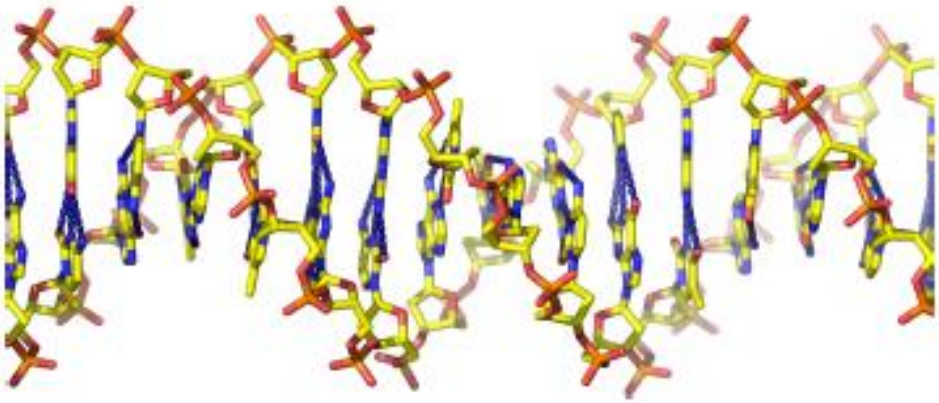


Heterodimer
tryptophan synthase

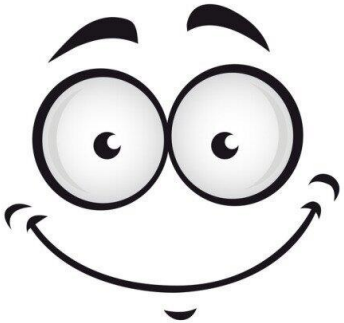


Heterotetramer
immunoglobin

Nucleic acids



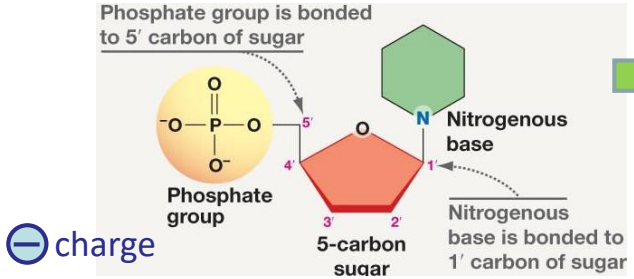
Structure of nucleic acids…



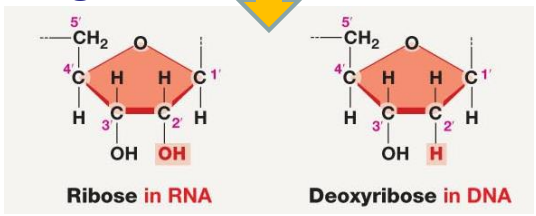
Nucleotides

❑ Composition

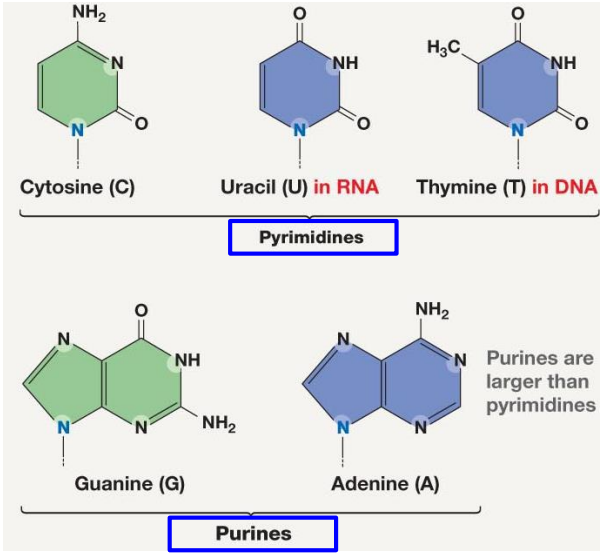
Nucleotide



Sugar



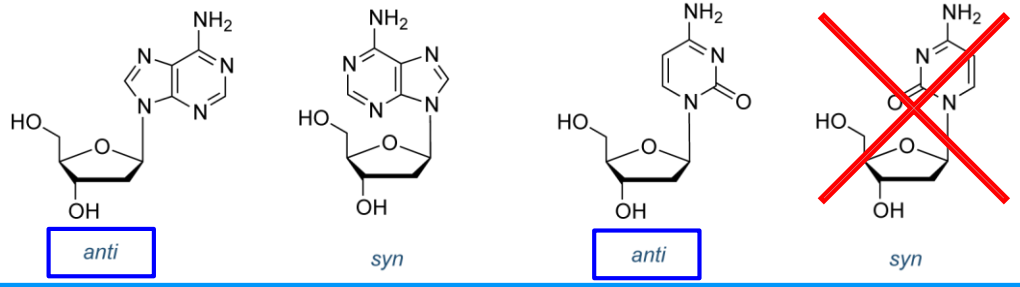
Nitrogenous base



- ❑ Phosphate
- ❑ Pentose sugar
- ❑ Heterocyclic base

- ❑ DNA bases: **A**, **G**; **T**, **C**
- ❑ RNA bases: **A**, **G**; **U**, **C**

❑ Rotation about glycosidic bond



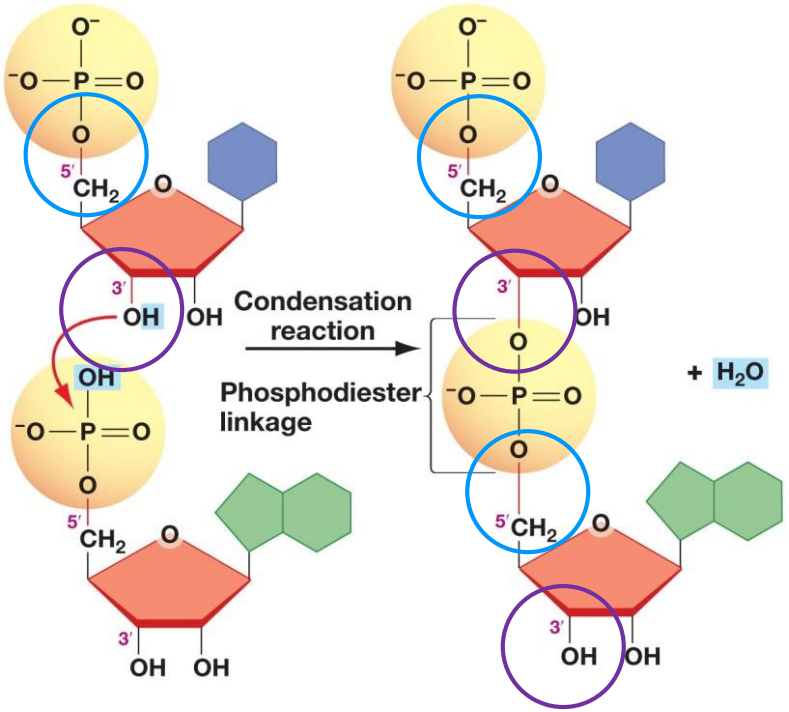
The *anti* conformation is dominant in DNA with rare exceptions

Primary structure

- Linear chain of nucleotides (oligonucleotides or polynucleotides)

  CGCGAATTCGCG

- Sugar-phosphate backbone
  - Covalent character
  - Phosphodiester bond
  - From 5’-end to 3’-end

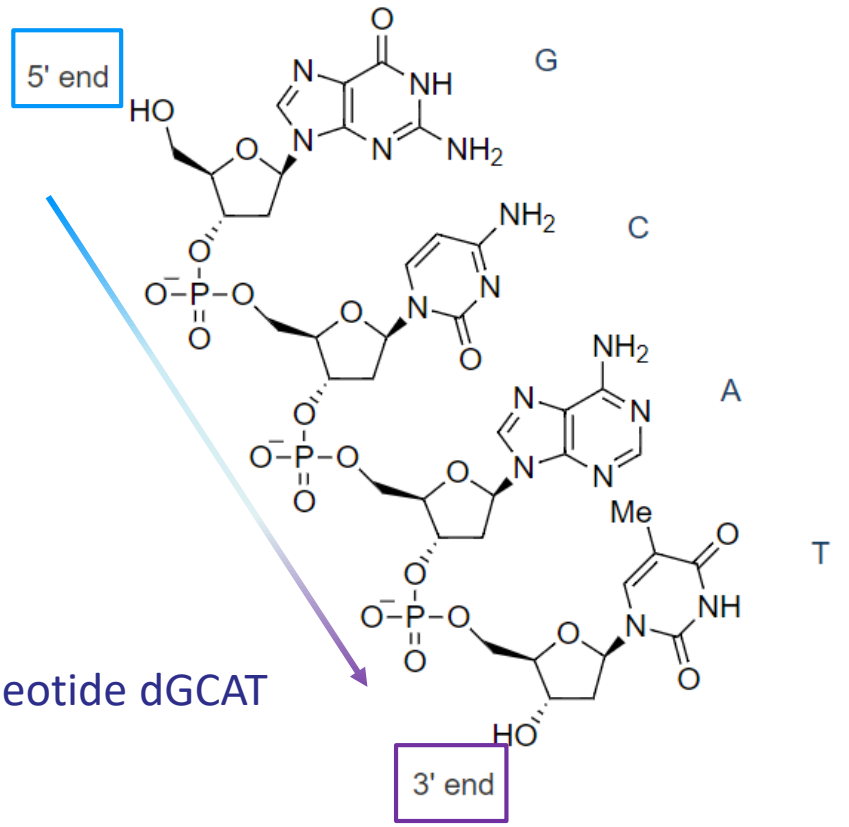


Primary structure

- Linear chain of nucleotides (oligonucleotides or polynucleotides)

  CGCGAATTCGCG

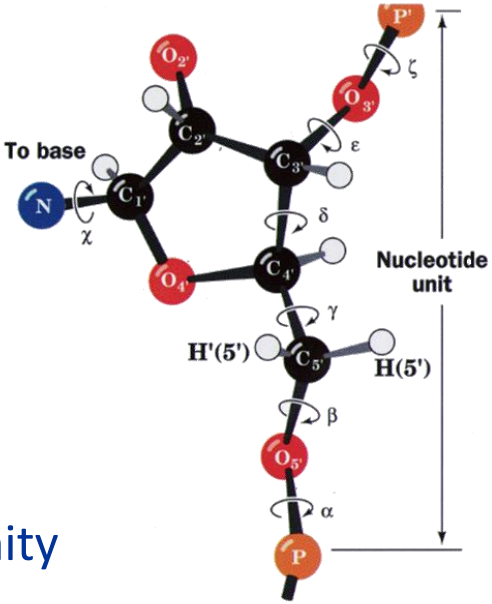
- Sugar-phosphate backbone
  - Covalent character
  - Phosphodiester bond
  - From 5’-end to 3’-end



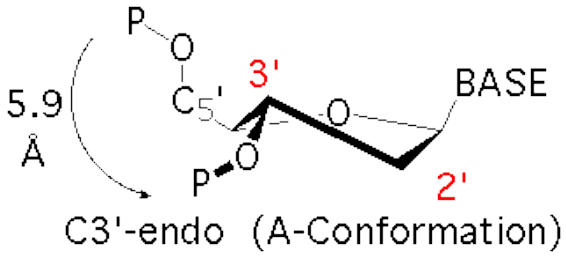
oligonucleotide dGCAT

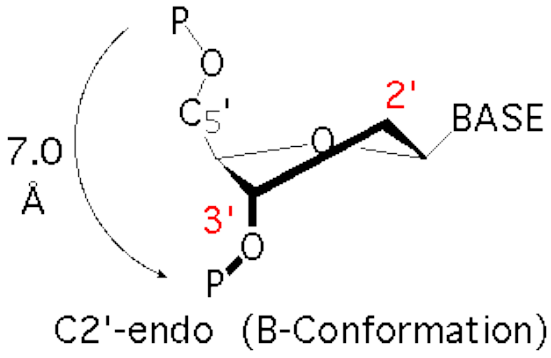
Sugar-phosphate backbone

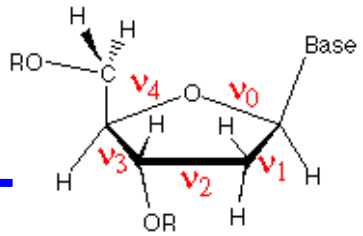
- Very flexible backbone
  - Six torsion angles



- Ribose is not planar → sugar puckering
  - Denotes the phosphate-phosphate proximity
  - Two main types of conformation





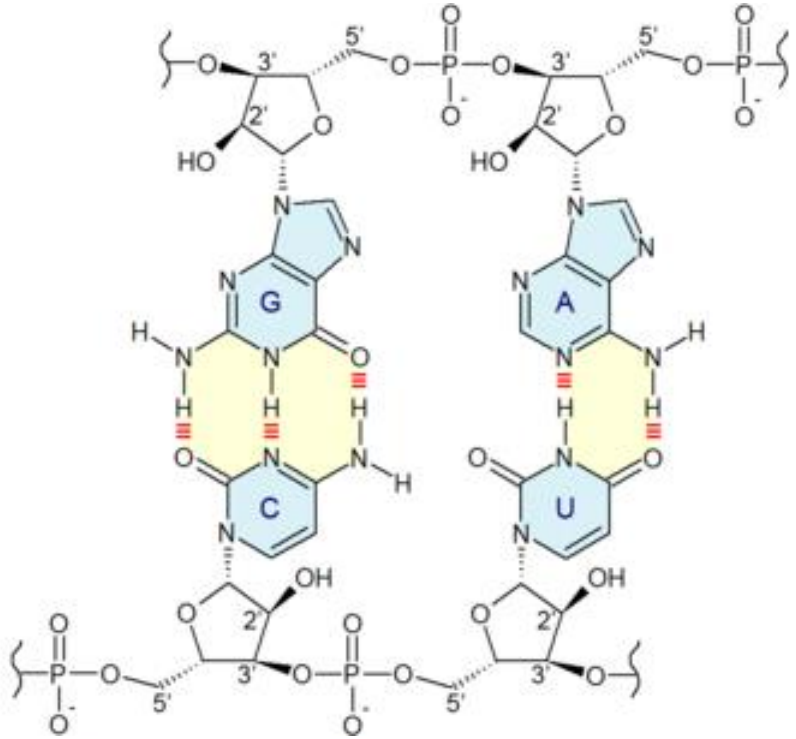


2′-deoxyribose (in DNA)

Secondary structure

- Local interactions between nucleotide bases

→Base pairs



≡ H-bonds

- DNA base pairs:

  Adenine - Thymine

  Cytosine - Guanine

- RNA base pairs:

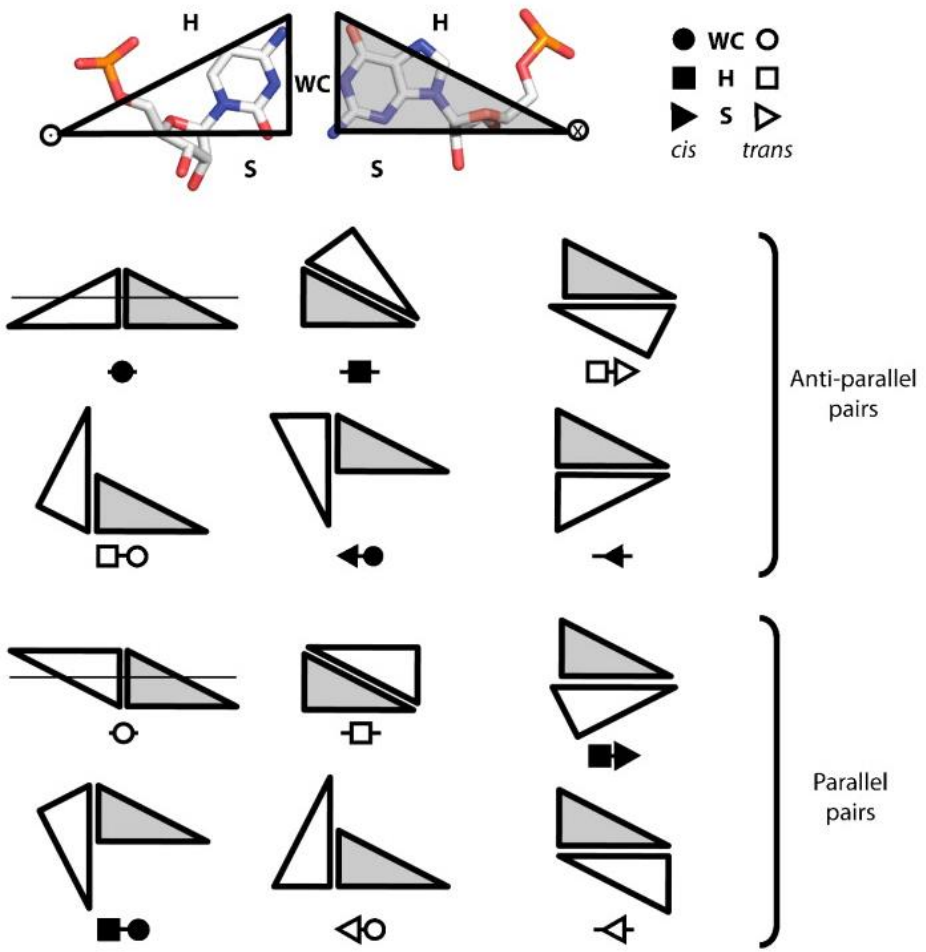
  Adenine - Uracil

  Cytosine - Guanine

- Complementarity due to hydrogen bonds

Secondary structure

- Leontis /Westhof classification
  - Three base-paring edges
    - Watson-Crick (WC)
    - Hoogsteen (H)
    - Sugar (S)
  - 12 types of base-paring



Tertiary structure of DNA

- Overall three-dimensional arrangement and folding
- Three types: A-DNA, B-DNA, Z-DNA
- B-DNA is the most common

  (described by Watson & Crick)

A-DNA
(rare)

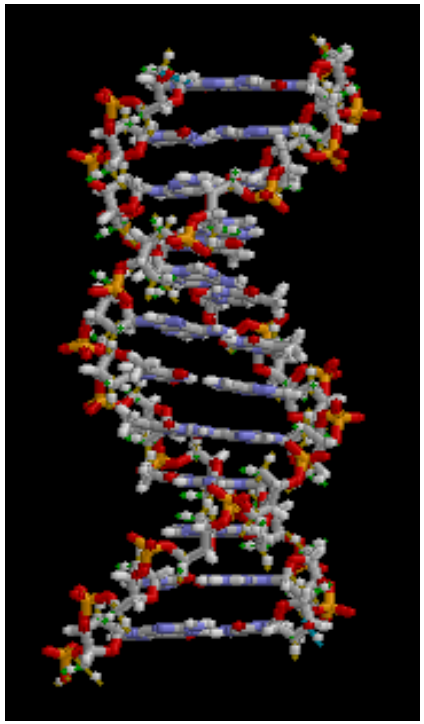
B-DNA
(predominant)

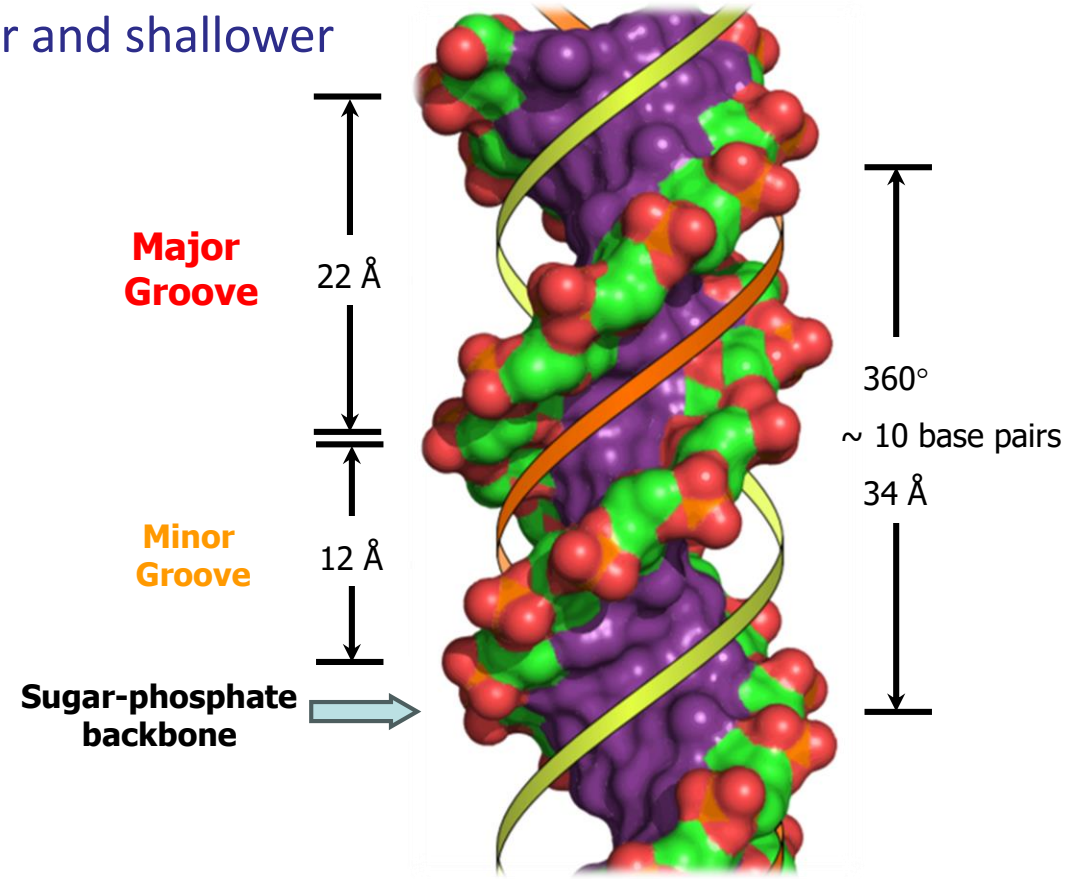
Z-DNA

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

Tertiary structure of DNA

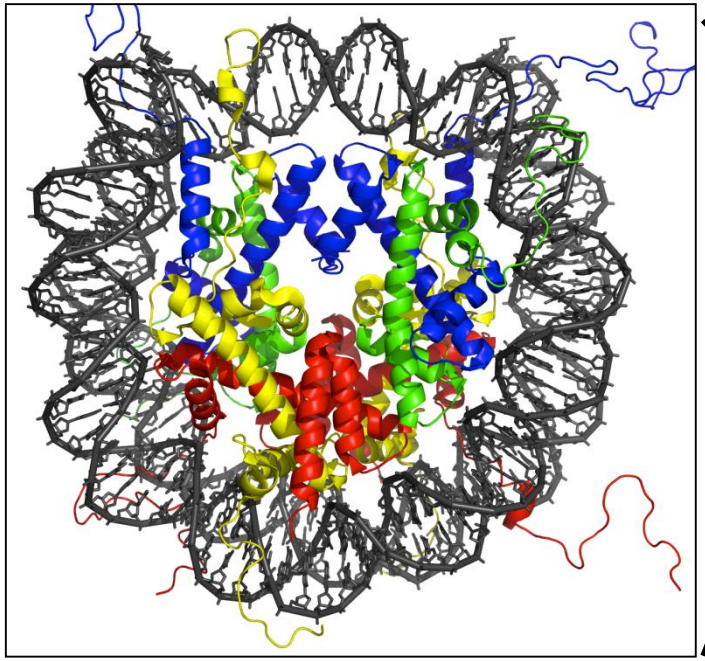
- Grooves: crucial for DNA-protein interactions
  - Major groove: wide and deep – where most proteins interact
  - Minor groove: narrower and shallower

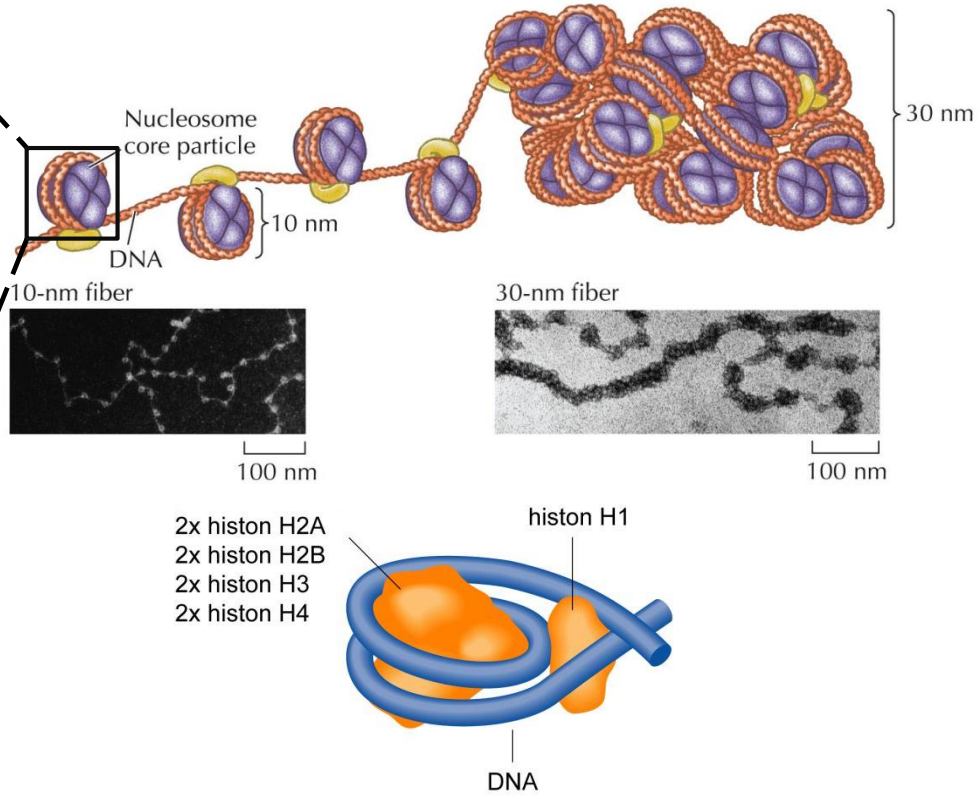




Higher structures of DNA

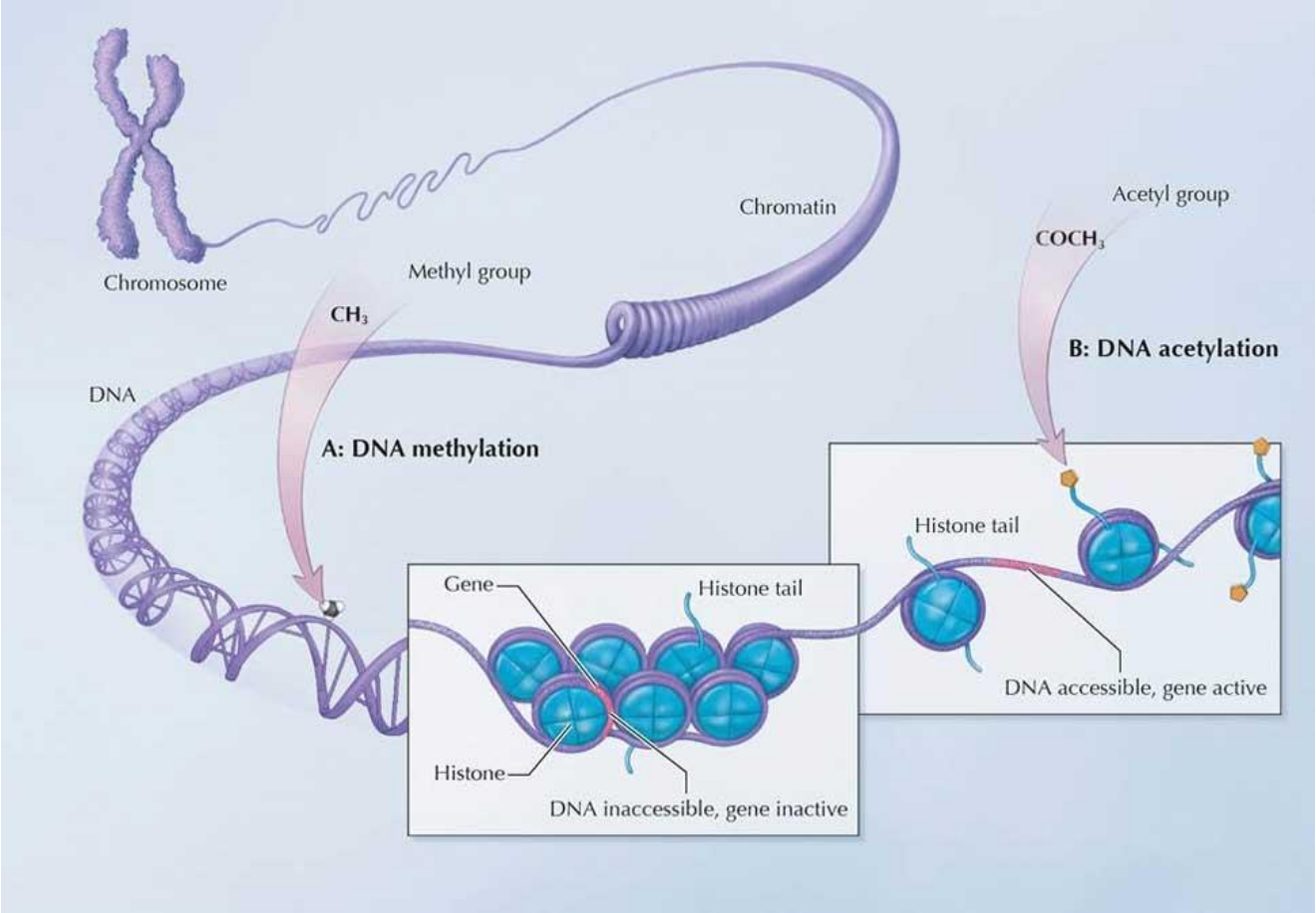
- Quaternary structures - with support of proteins





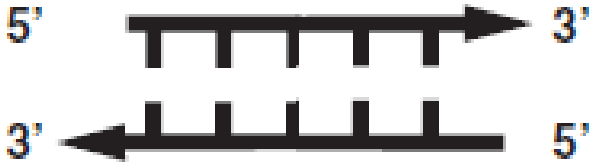
Higher structures of DNA

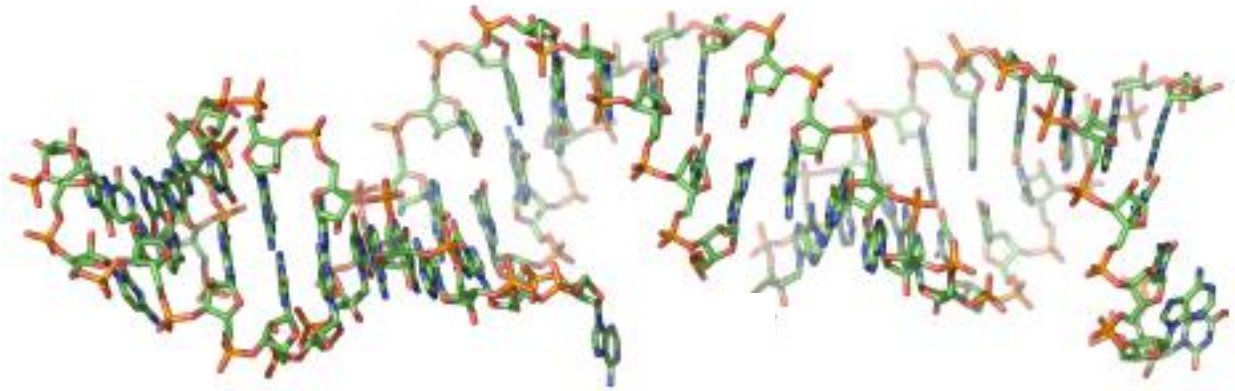
- Quaternary structures - with support of proteins



Secondary structures of RNA

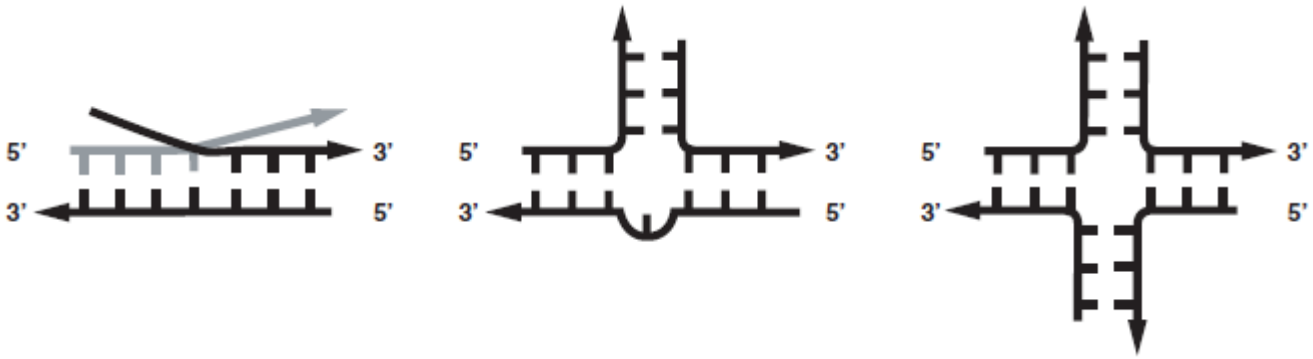
- Most common form: A-RNA helix (similar to A-DNA)





Secondary structures of RNA

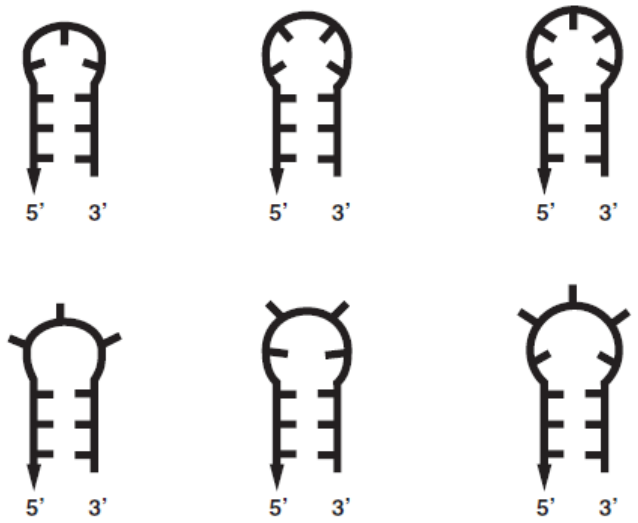
- Junctions
  - Regions connecting two or more stems
  - Two-stem, three-stem and four-stem junction



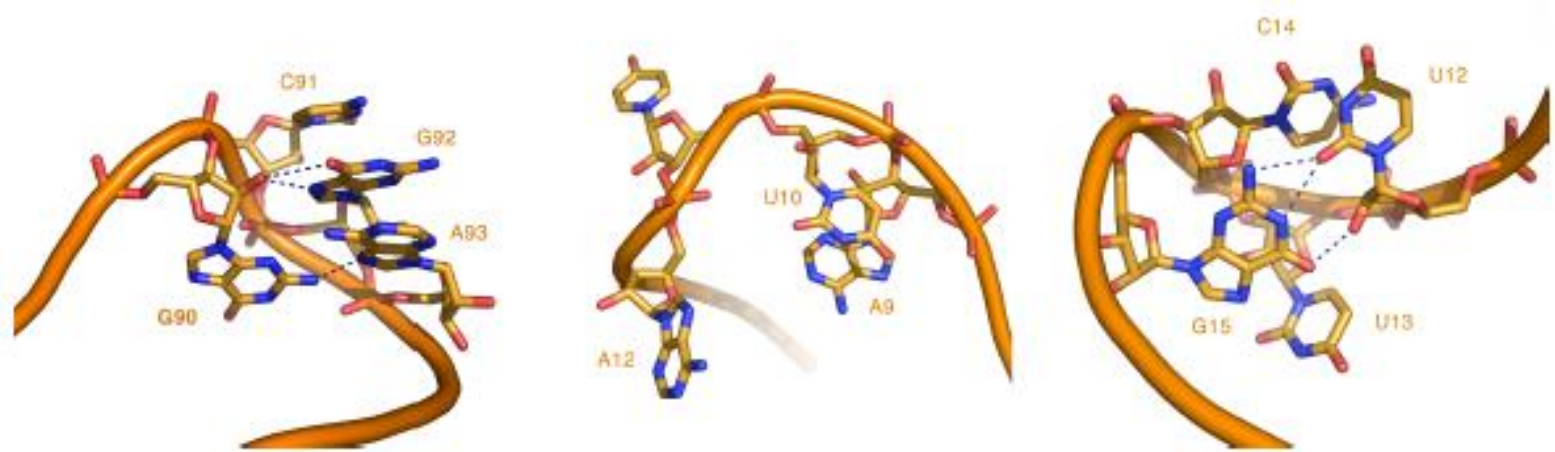
Secondary structures of RNA

- Harpin loops
  - Sequence inversely self-complementary

    GGCUGGCUGUUCGCCAGCC

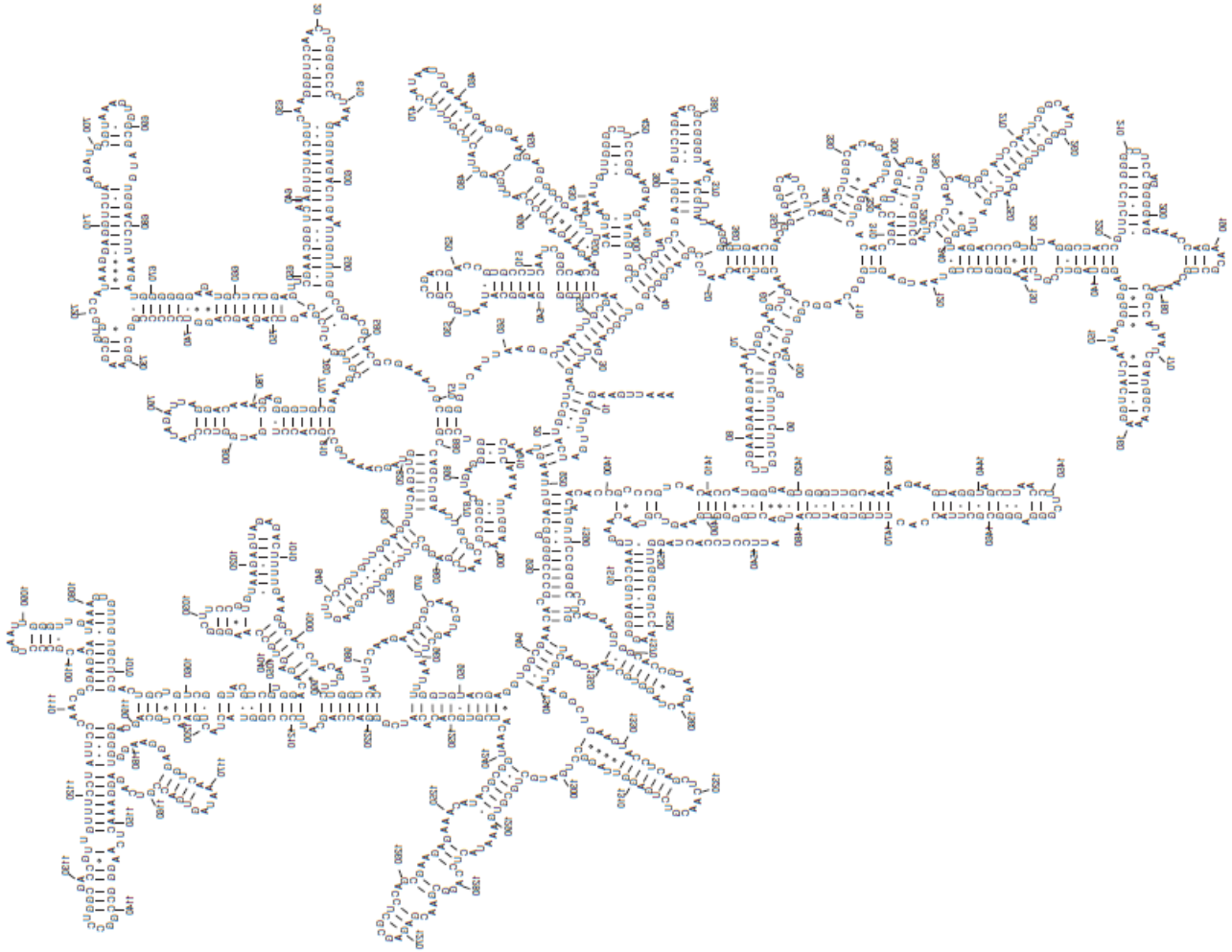


  - Many subtypes - e.g.: GNRA, ANYA, UNCG tetraloops



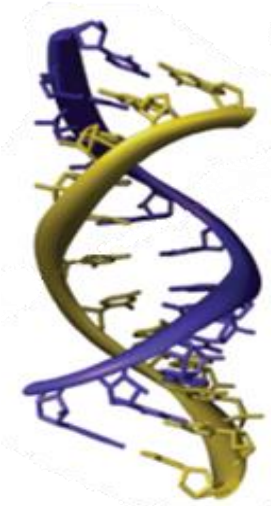
Secondary structures of RNA

- Very complex – stem-loop structure

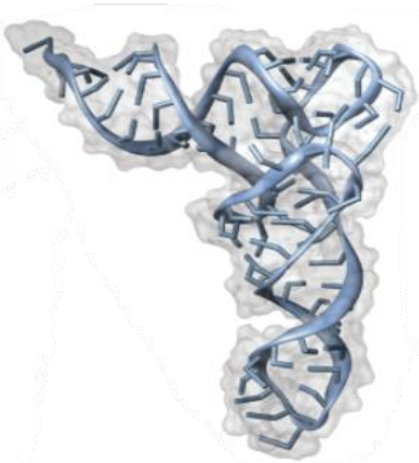


Tertiary structures of RNA

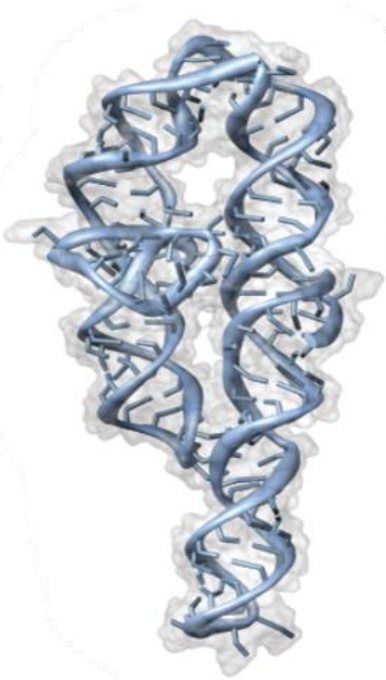
A-RNA dodecamer



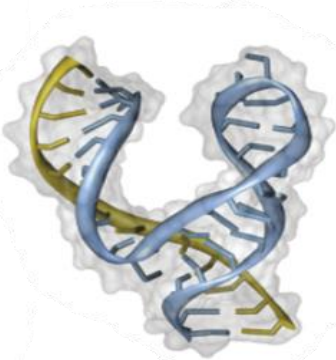
Phenylalanine transfer RNA



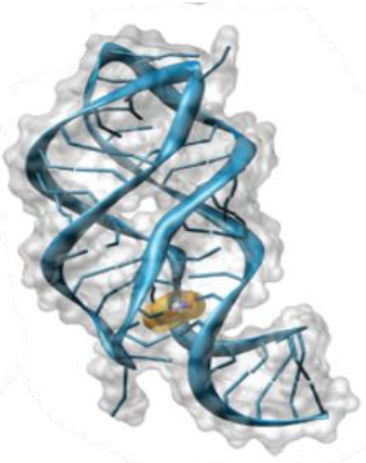
Group I intron ribozyme



Hammerhead ribozyme

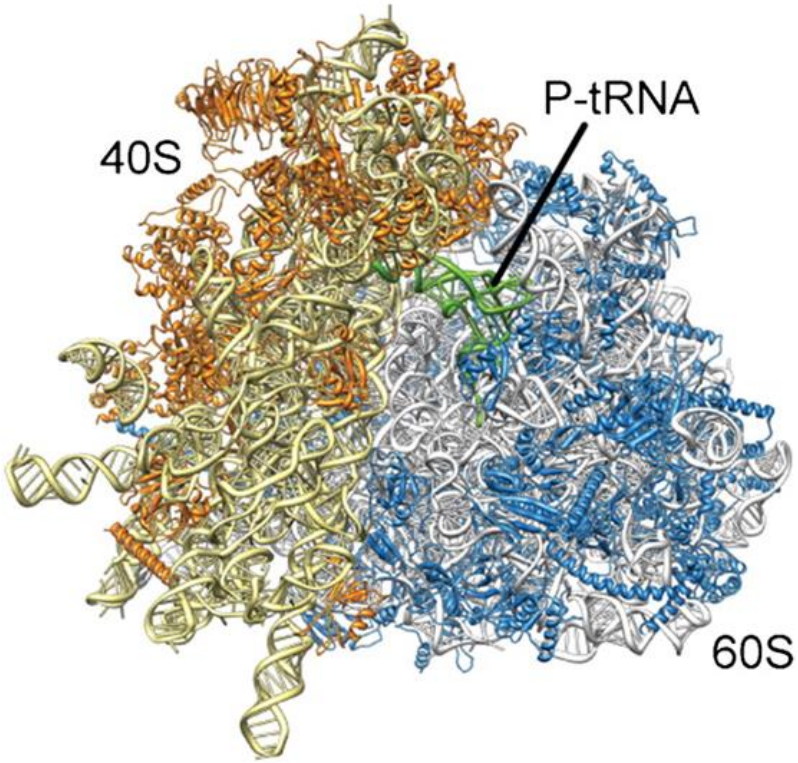


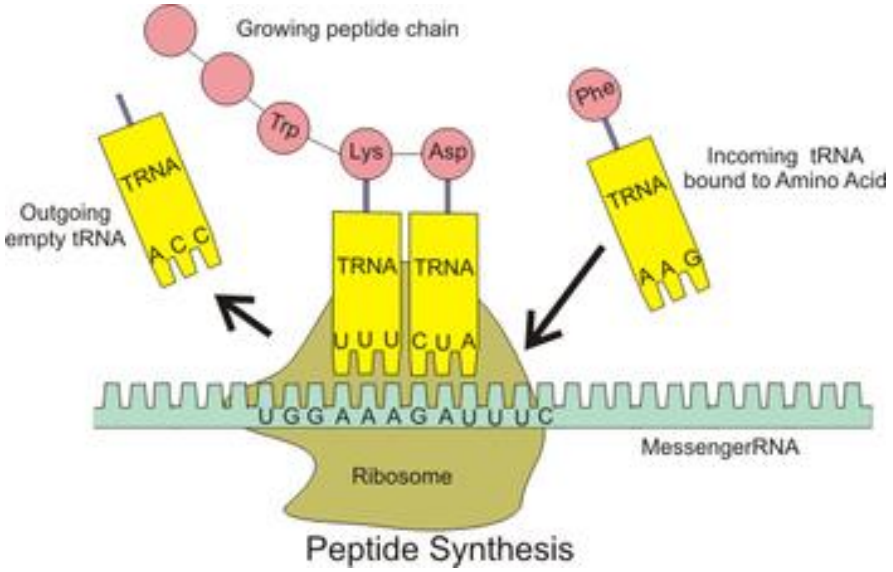
Guanine riboswitch



Quaternary structure of RNA

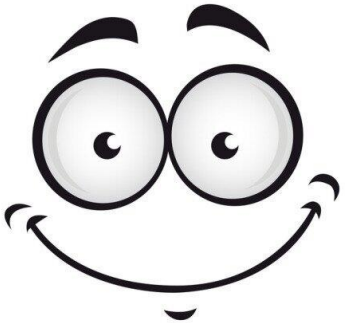
- Association of several chains of RNA
  - Frequently joined with proteins
  - Eukaryotic ribosome - ~ 6800 nt, 79 proteins





Ribosome in action:
https://www.youtube.com/watch?v=Jml8CFBWcDs

Structural databases?



Primary structural databases

- Worldwide Protein Data Bank (wwPDB)

  http://www.wwpdb.org/
- RCSB Protein Data Bank (RCSB PDB)

  http://pdb.rcsb.org
- Nucleic Acid Knowledgebase (Nucleic Acid Database)

  https://www.nakb.org/
- Biological Magnetic Resonance Data Bank (BMRB)

  https://bmrb.io/
- Electron Microscopy Data Bank (EMDB)

  http://www.emdatabank.org/
- Cambridge Structural Database (CSD)

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

…More details in lesson 4!

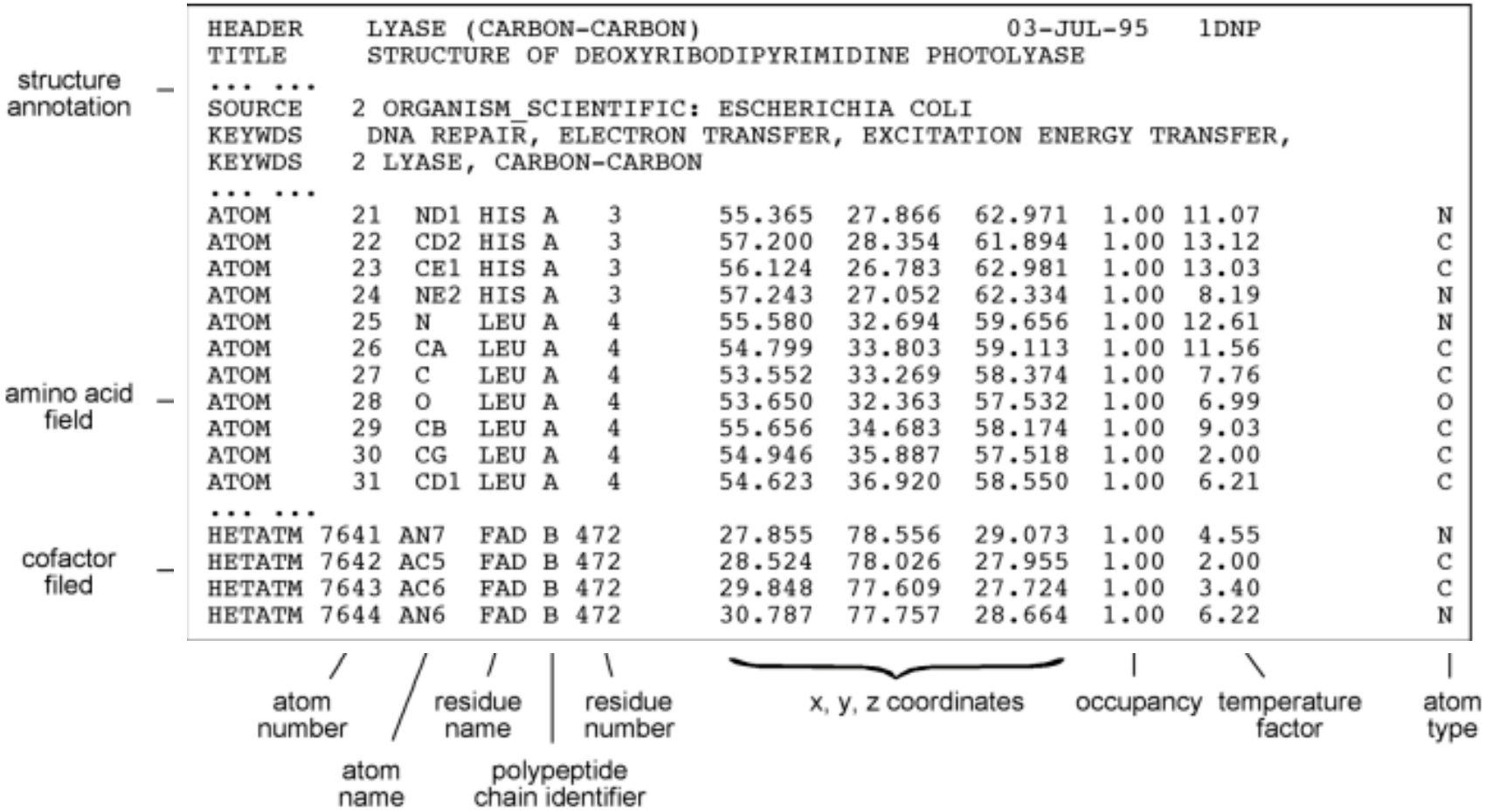
Structural data formats

- Different file formats used to represent 3D structure data
  - PDB
  - mmCIF
  - PDBML
  - MOL2
  - …
- The spatial 3D coordinates and other information are recorded for each atom

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

PDB format


structure annotation
amino acid field
cofactor filed
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
atom number
atom name
residue name
residue number
polypeptide chain identifier
x, y, z coordinates
occupancy
temperature factor
atom type

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

PDB format

- Advantages
  - Widely used → supported by majority of tools
  - Easy to read and easy to use
  - Can be manually edited

→ Suitable for accessing individual entries

PDB format

- Disadvantages
  - Potential inconsistency between individual PDB entries as well as

PDB records within one entry

Ex: different residue numbering in SEQRES and ATOM sections

→ Not suitable for computer extraction of information

Primary sequence

Atoms and residues in the file

```
SEQRES    1    396  MET ASP GLU ASN ILE THR ALA ALA PRO ALA ASP PRO ILE
SEQRES    2    396  LEU GLY LEU ALA ASP LEU PHE ARG ALA ASP GLU ARG PRO
. . .
. . .
ATOM      1  N   MET     5      41.402  11.897  15.262  1.00 48.61
ATOM      2  CA  MET     5      40.919  13.262  15.600  1.00 47.70
ATOM      9  N   PHE     6      39.627  14.840  14.228  1.00 48.66
ATOM     10  CA  PHE     6      39.199  15.440  12.964  1.00 45.33
. . .
```

PDB format

- Disadvantages
  - Absolute limits on the size of certain items of data

Ex.: max. number of atom records limited to 99,999; max. number of chains limited to 26, etc.

→ Large systems such as the ribosomal subunit must be divided into multiple PDB files

→ Not suitable for analysis and comparison of experimental and structural data across the entire database

mmCIF format

- **M**acro**m**olecular **c**rystallographic **i**nformation **f**ile (mmCIF)
- Developed to handle increasingly complicated structural data
- Each field of information is explicitly assigned by a tag and linked to other fields through a special syntax

PDB HEADER PLANT SEED PROTEIN 11-OCT-91 1CBN

mmCIF

```
_struct.entry_id '1CBN'
_struct.title 'PLANT SEED PROTEIN'
_struct_keywords.entry_id '1CBN'
_struct_keywords.text 'plant seed protein'
_database_2.database_id 'PDB'
_database_2.database_code '1CBN'
_database_PDB_rev.rev_num 1
_database_PDB_rev.date_original '1991-10-11'
```

mmCIF format

- Advantages
  - Easily parsable by computer software
  - Consistency of data across the database

→ Suitable for analysis and comparison of experimental and structural data across the entire database

- Disadvantages
  - Difficult to read
  - Rarely supported by visualization and computational tools

→ Not suitable for accessing individual entries

PDBML format

- **P**rotein **D**ata **B**ank **M**arkup **L**anguage (PDBML)
- E**x**tensible **M**arkup **L**anguage (XML) version of PDB format

```
<?xml version="1.0" encoding="UTF-8" ?>
<PDBx:datablock  datablockName="EXAMPLE"
   xmlns:PDBx="http://deposit.pdb.org/pdbML/pdbx-v1.000.xsd"
   xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
   xsi:schemaLocation="http://deposit.pdb.org/pdbML/pdbx-v1.000.xsd
            pdbx-v1.000.xsd">
   <PDBx:entity_polyCategory>
      <PDBx:entity_poly entity_id="1">
         <PDBx:type>polypeptide(L)</PDBx:type>
         <PDBx:nstd_linkage>no</PDBx:nstd_linkage>
         <PDBx:nstd_monomer>no</PDBx:nstd_monomer>
         <PDBx:pdbx_seq_one_letter_code>
          DIVLTQSPASLSASVGETVTITCRASGNIHNYLAWYQQKQGKSPQLLVYYTTTLADG
          VPSRFSGSGSGTQYSLKINSLQPEDFGSYYCQHFWSTPRTFGGGTKLEIK
         </PDBx:pdbx_seq_one_letter_code>
         <PDBx:pdbx_seq_one_letter_code_can>
          DIVLTQSPASLSASVGETVTITCRASGNIHNYLAWYQQKQGKSPQLLVYYTTTLADG
          VPSRFSGSGSGTQYSLKINSLQPEDFGSYYCQHFWSTPRTFGGGTKLEIK
         </PDBx:pdbx_seq_one_letter_code_can>
      </PDBx:entity_poly>
   </PDBx:entity_polyCategory>
</PDBx:datablock>
```

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.