Structural bioinformatics KFC/STBI

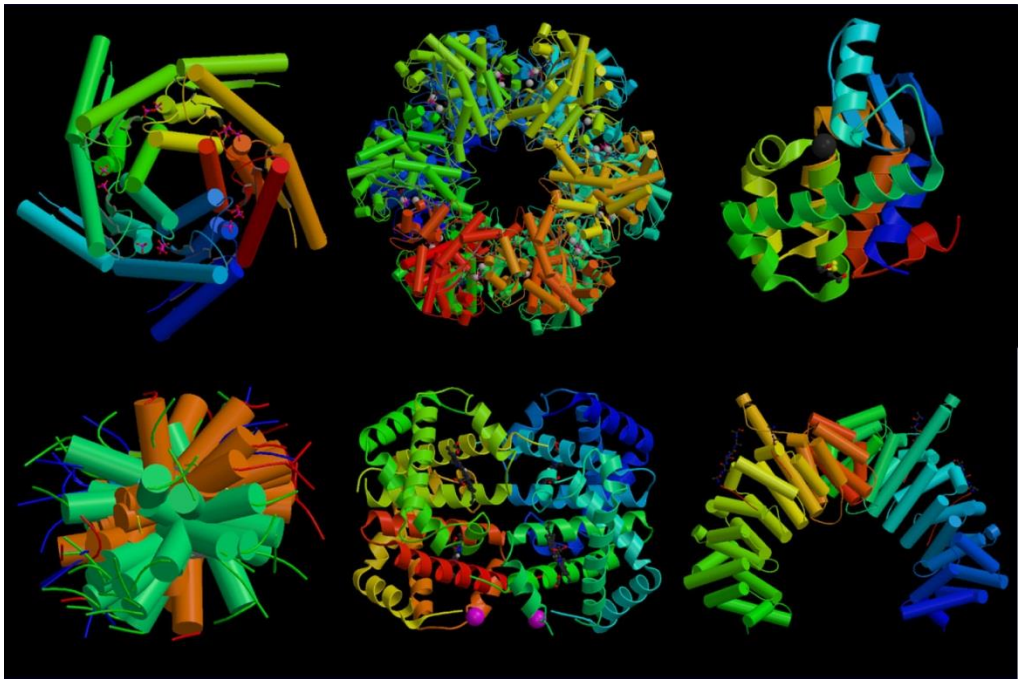
Biomolecules and how to understand their structures

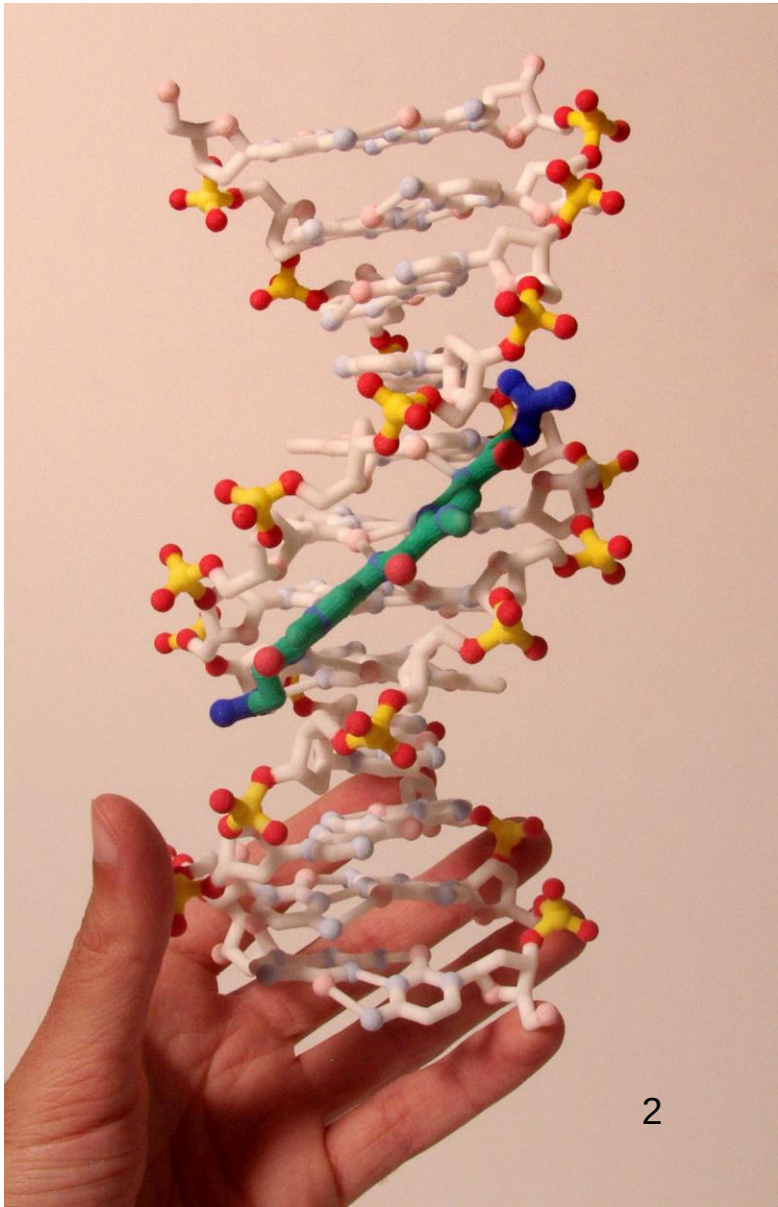
Karel Berka

Miroslav Krepl

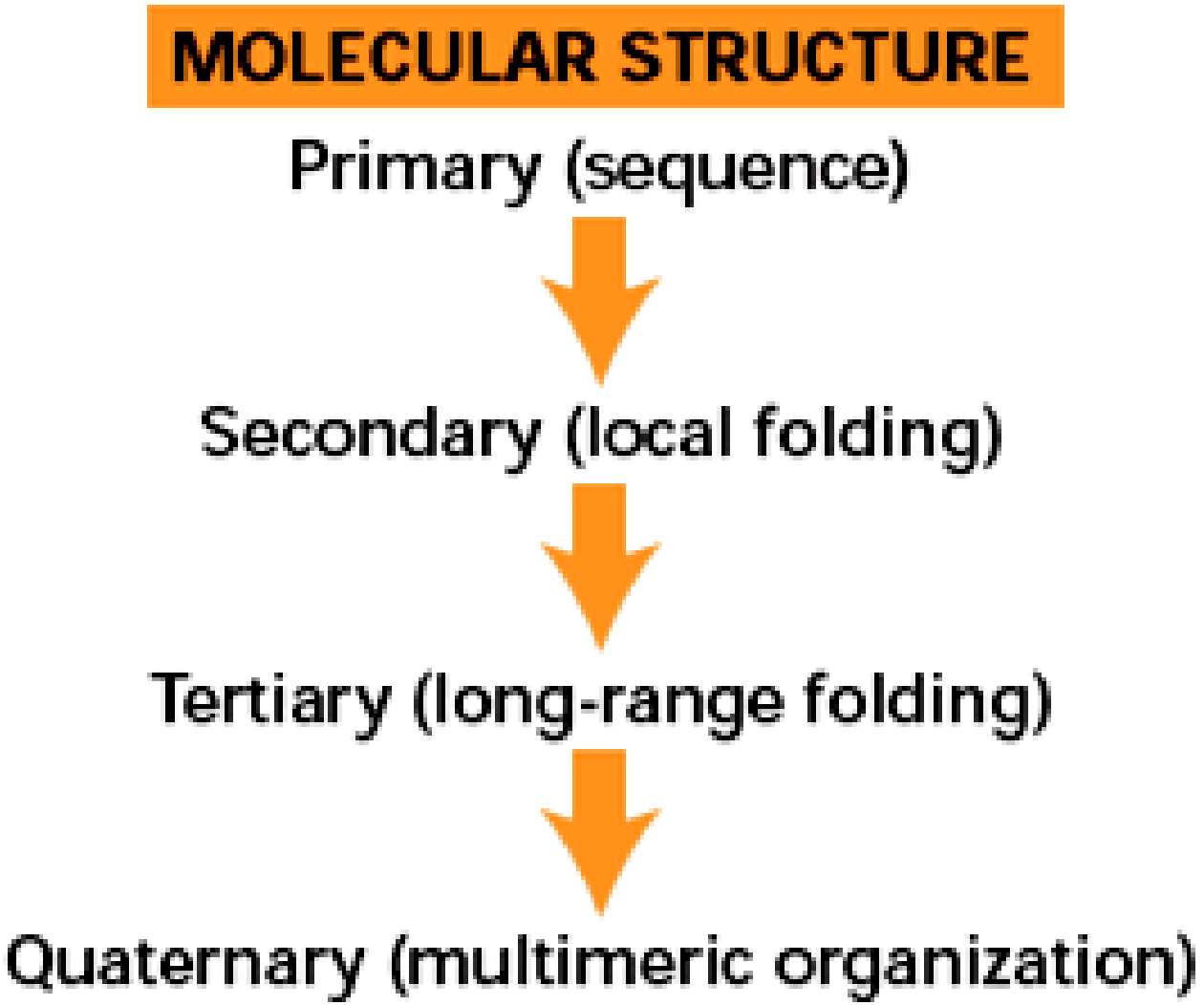
Biomolecules

- proteins
- NA – DNA, RNA
- lipids
- polysaccharides
- Small molecules (hormones, drugs)





Structural Hierarchy

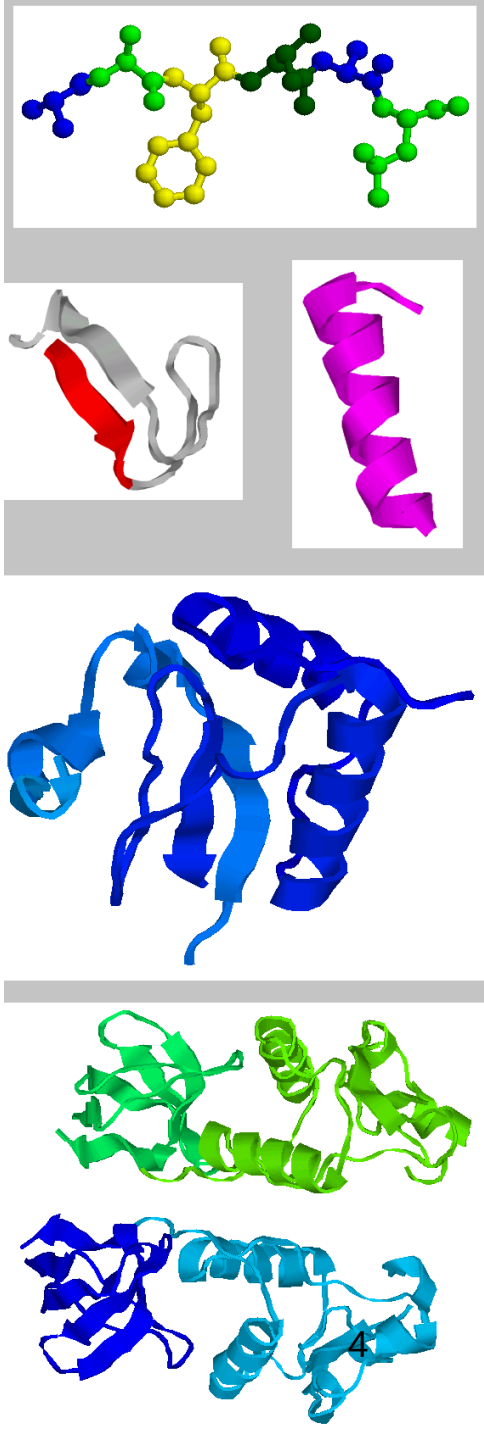


Lodish, Molecular Cell Biology, 5th Ed.

Proteins

- Amino acids
- Backbone and Sidechains

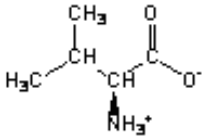
- Primary structure
  - sequence of amino acids
- Secondary structure
  - Local structural patterns
- Tertiary structure
  - Domain Fold
- Quarternary structure
  - Multichain organization



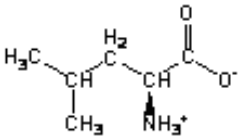
http://cs.wikipedia.org/wiki/Soubor:ProteinStructures.png

Amino acids

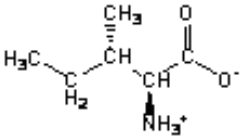
Amino acids with hydrophobic side chains



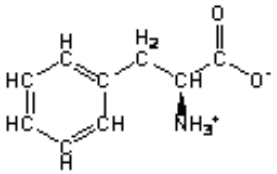
Valine
Val
V



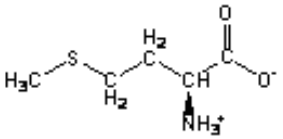
Leucine
Leu
L



Isoleucine
Ile
I

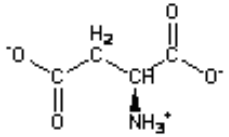


Phenylalanine
Phe
F

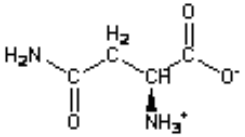


Methionine
Met
M

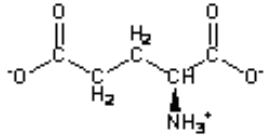
Amino acids with hydrophilic side chains



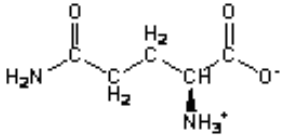
Aspartic acid
Asp
D



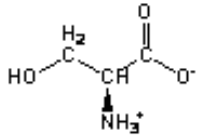
Asparagine
Asn
N



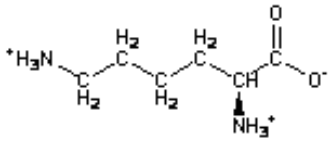
Glutamic acid
Glu
E



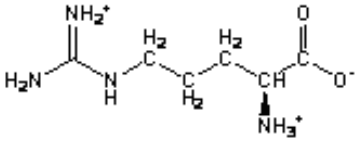
Glutamine
Gln
Q



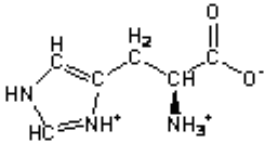
Serine
Ser
S



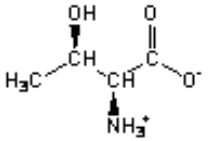
Lysine
Lys
K



Arginine
Arg
R

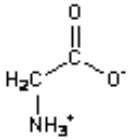


Histidine
His
H

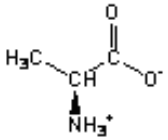


Threonine
Thr
T

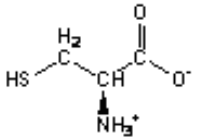
Amino acids with intermediate side chains



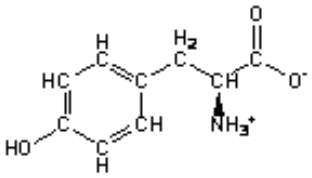
Glycine
Gly
G



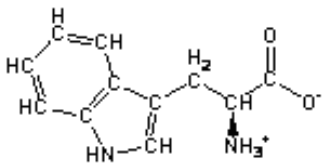
Alanine
Ala
A



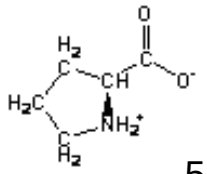
Cysteine
Cys
C



Tyrosine
Tyr
Y

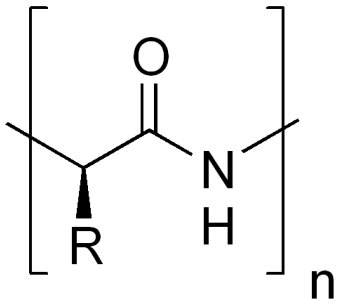


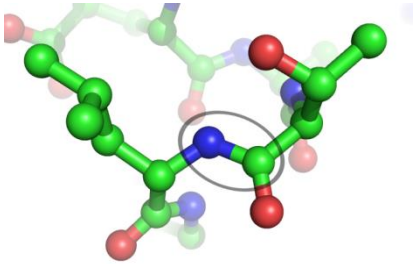
Tryptophan
Trp
W



Proline
Pro
P

Primary Structure of Protein





AMINO ACID			SIDE CHAIN
Aspartic acid	Asp	D	negative
Glutamic acid	Glu	E	negative
Arginine	Arg	R	positive
Lysine	Lys	K	positive
Histidine	His	H	positive
Asparagine	Asn	N	uncharged polar
Glutamine	Gln	Q	uncharged polar
Serine	Ser	S	uncharged polar
Threonine	Thr	T	uncharged polar
Tyrosine	Tyr	Y	uncharged polar

POLAR AMINO ACIDS

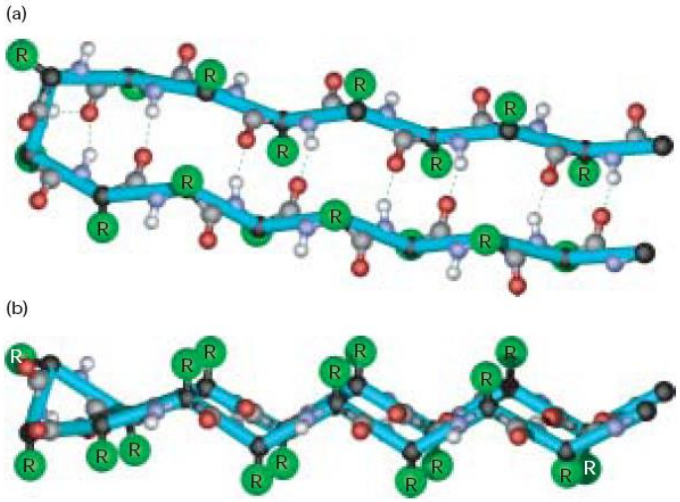
AMINO ACID			SIDE CHAIN
Alanine	Ala	A	nonpolar
Glycine	Gly	G	nonpolar
Valine	Val	V	nonpolar
Leucine	Leu	L	nonpolar
Isoleucine	Ile	I	nonpolar
Proline	Pro	P	nonpolar
Phenylalanine	Phe	F	nonpolar
Methionine	Met	M	nonpolar
Tryptophan	Trp	W	nonpolar
Cysteine	Cys	C	nonpolar

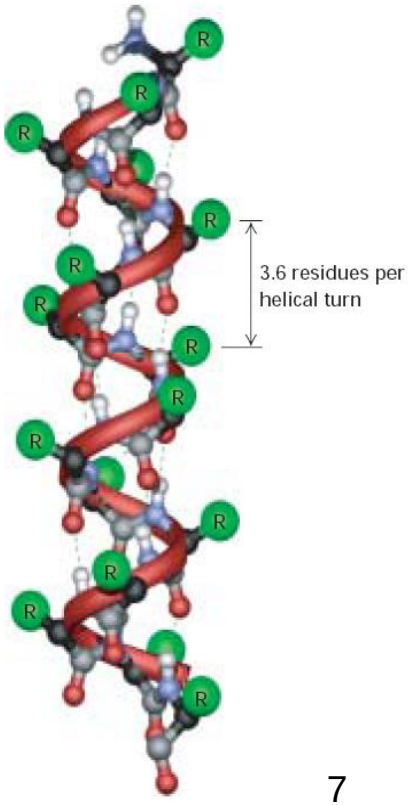
NONPOLAR AMINO ACIDS

Alberts, Molecular Biology of the Cell, 5th Ed.

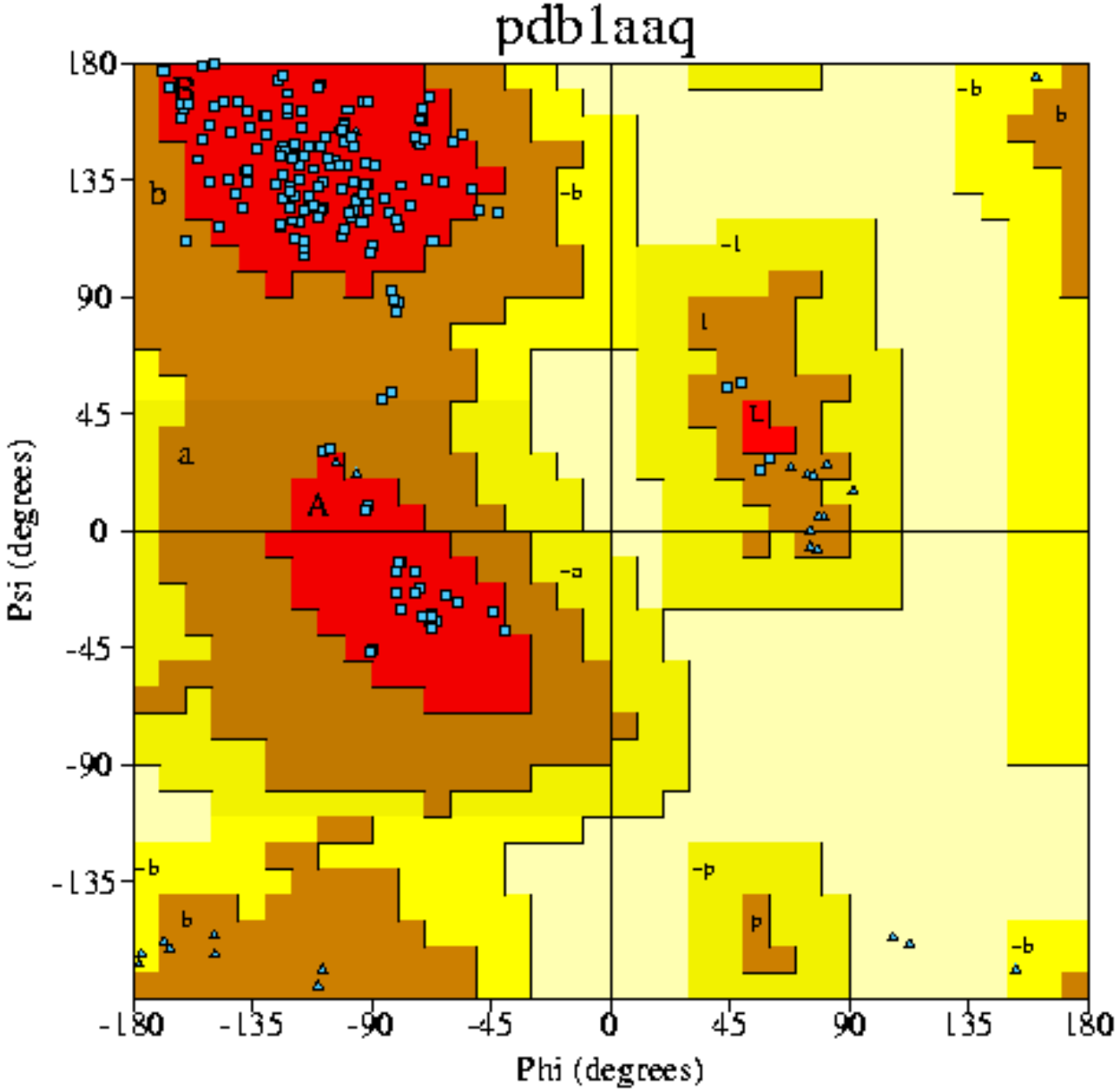
Secondary structure of Proteins

- Local folding
- Secondary structure depends on amino acid sequence
  - **α-helix**
  - **3-10 helix**
  - **β-sheet**
  - **β-turn, loop**





Ramachandran plot



PROCHECK summary for 1aaq

PROCHECK statistics

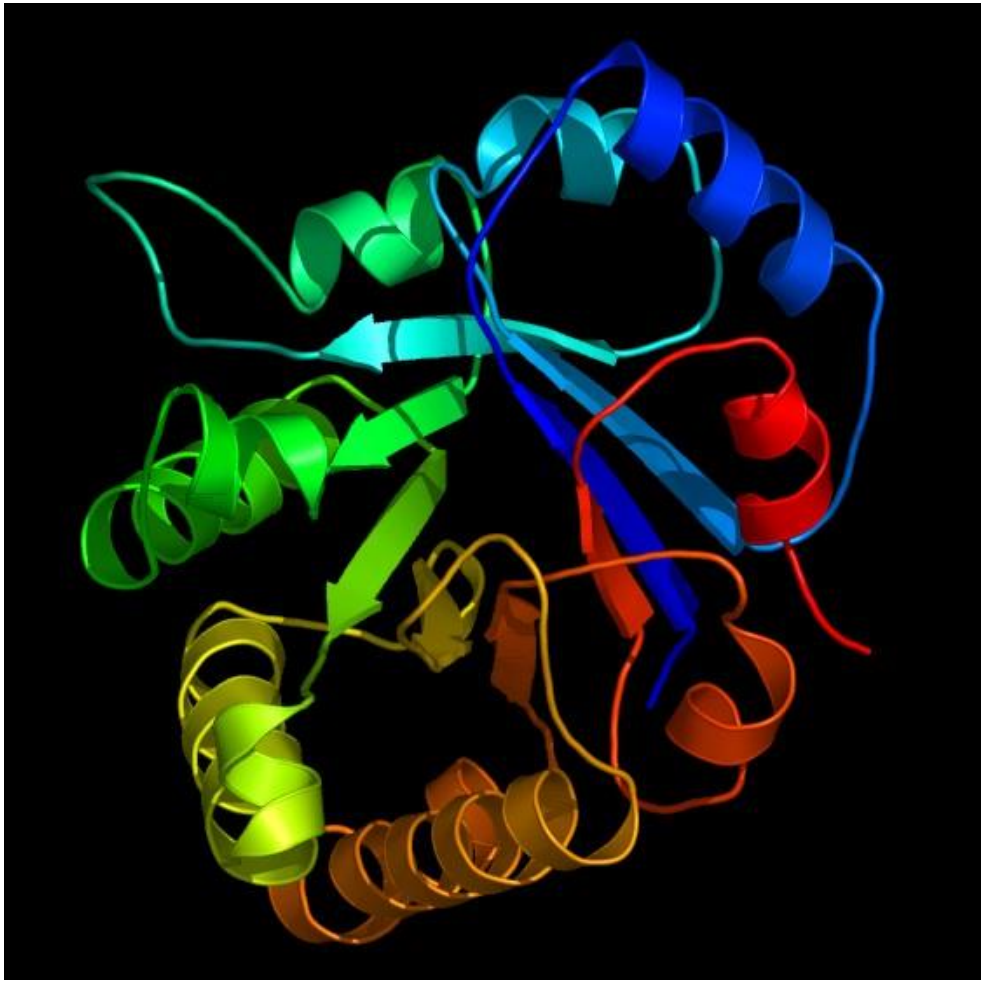
Ramachandran Plot statistics

	No. of residues	%-tage
Most favoured regions [A,B,L]	146	92.4%
Additional allowed regions [a,b,l,p]	12	7.6%
Generously allowed regions [~a,~b,~l,~p]	0	0.0%
Disallowed regions [XX]	0	0.0%
Non-glycine and non-proline residues	158	100.0%

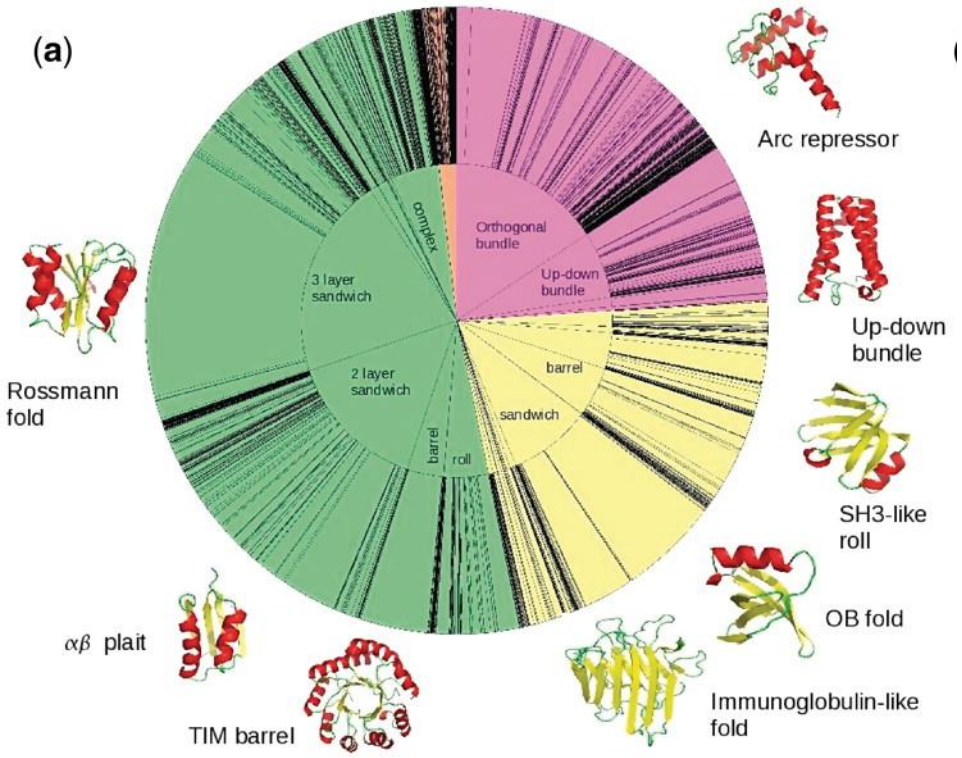
End-residues (excl. Gly and Pro)	2
Glycine residues	26
Proline residues	12
Total number of residues	198

Tertiary Structure

- fold
  - globular
  - membrane
  - Fibrilar
  - IUP
- Necessary for FUNCTION
- domains



‘CATHerine wheels’.



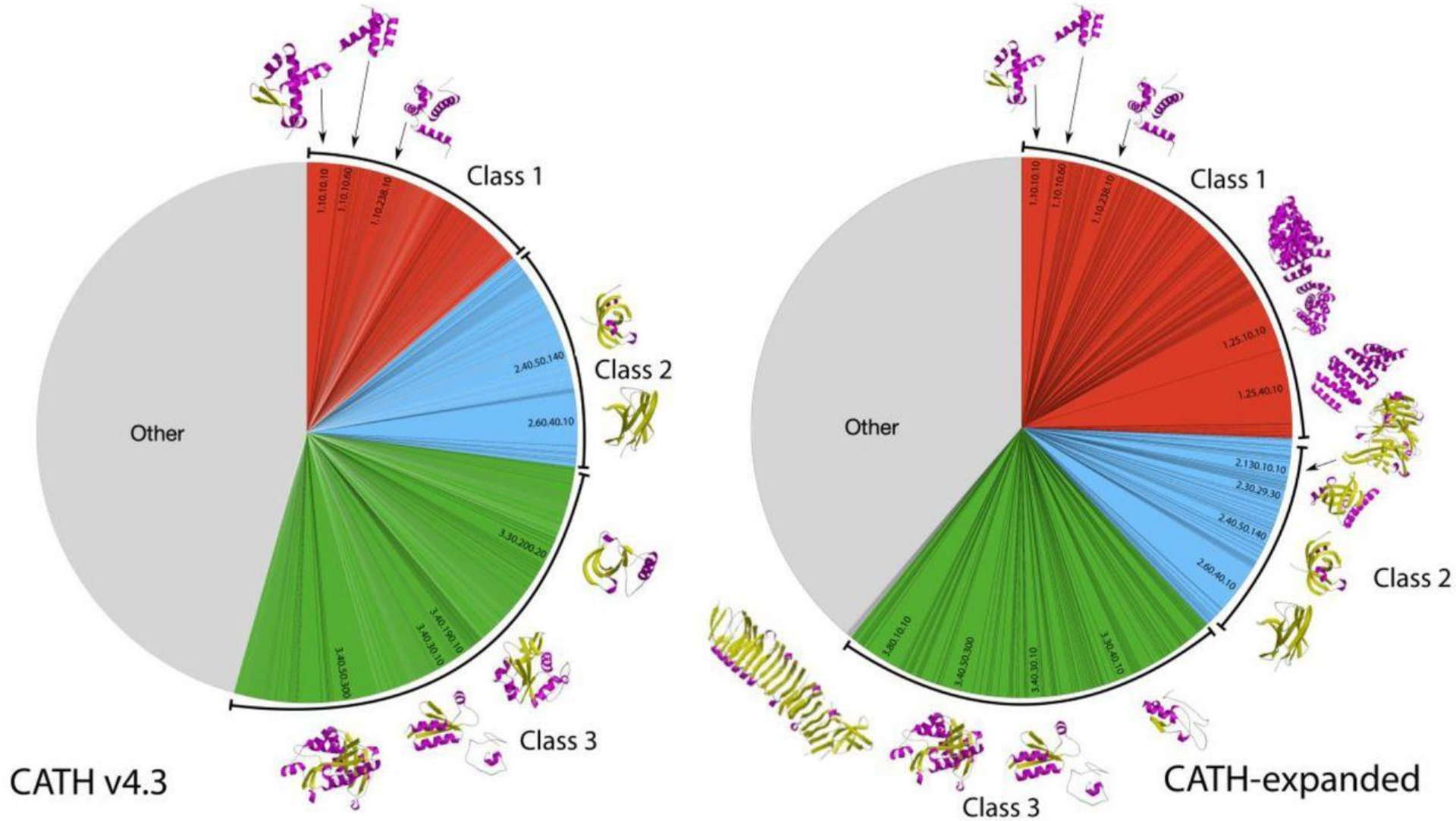
The distribution of all non-homologous structures (2386) within CATH v3.3

Classes:
pink (mainly α),
yellow (mainly β),
green (αβ)
brown (little secondary structure).

Proportion of structures within any given architecture (inner circle)
Fold group (outer circle).

Cuff A L et al. Nucl. Acids Res. 2011;39:D420-D426

CATH update 2022

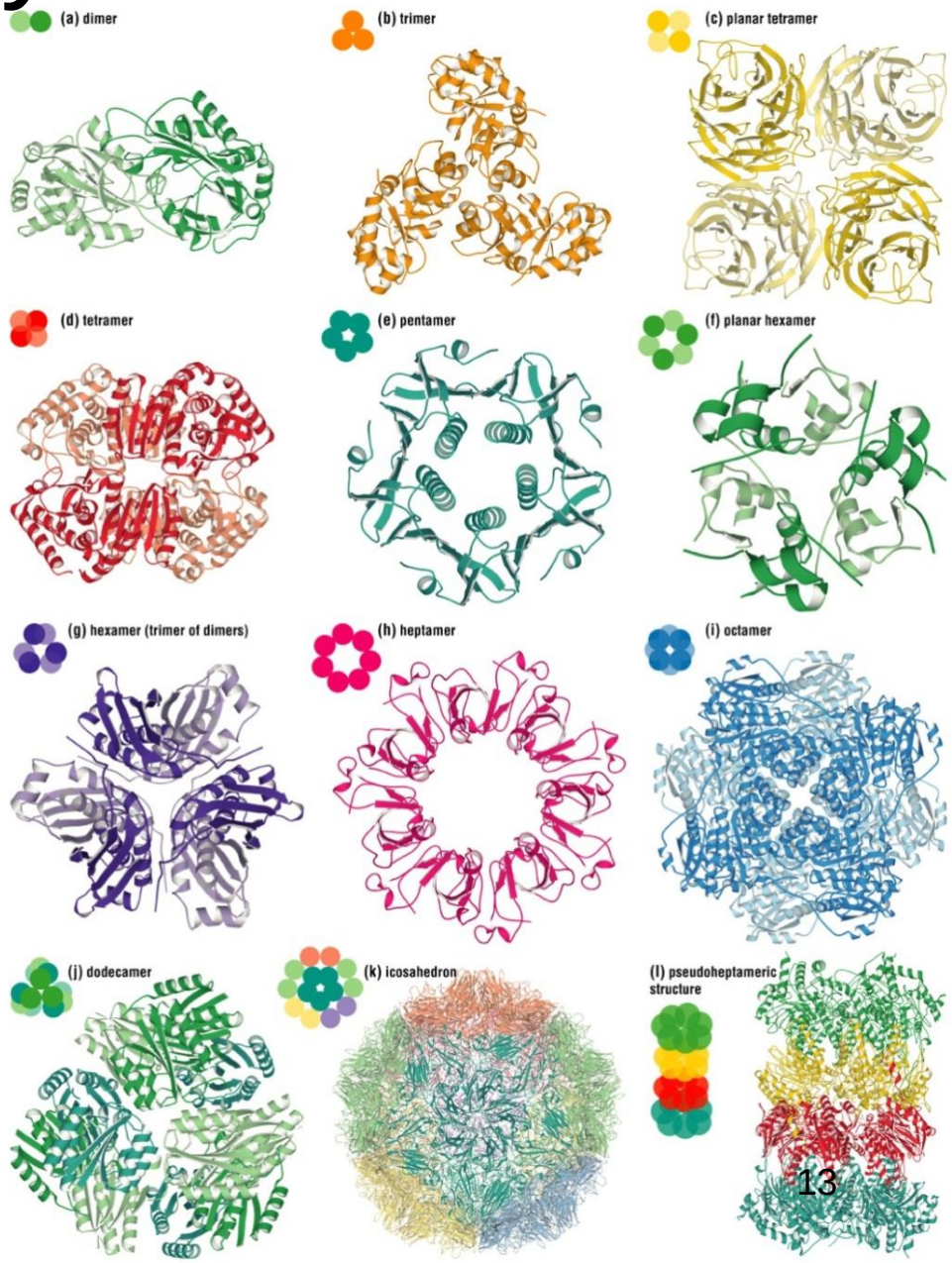


Structural diversity in CATH Superfamilies (left) and expanded by AF2 models (right).

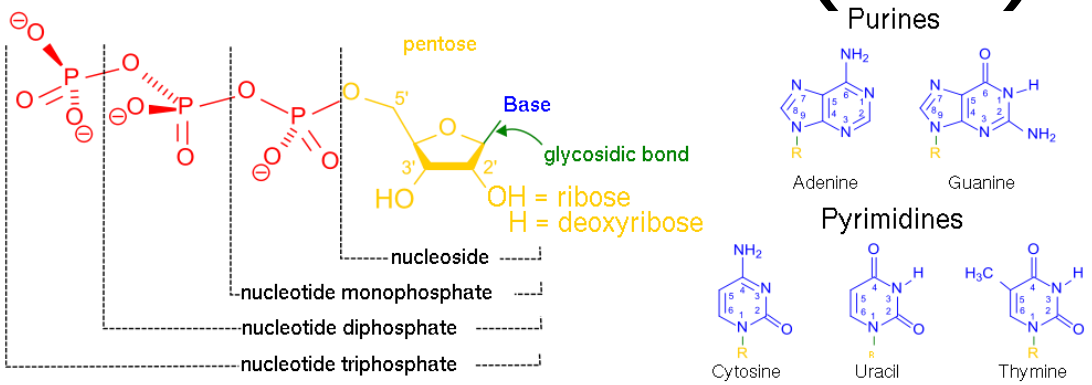
Bordin N et al.: AlphaFold2 reveals commonalities and novelties in protein structure space for 21 model organisms. *bioRxiv* 2022, doi:10.1101/2022.06.02.494367v1.full

Quarternary Structure

- asociace více řetězců:
  - Kooperativita
    (asociace zesílí vazebné vlastnosti)
    hemoglobin
  - Kolokalizace funkce
    (každá podjednotka dělá něco jiného)
    tryptophansyntáza
  - Kombinace podjednotek
    (přizpůsobování)
    imunoglobuliny
  - Skládání větších struktur
    (podjednotky uspořádávají procesem self-assembly)
    aktin,
    virové kapsidy



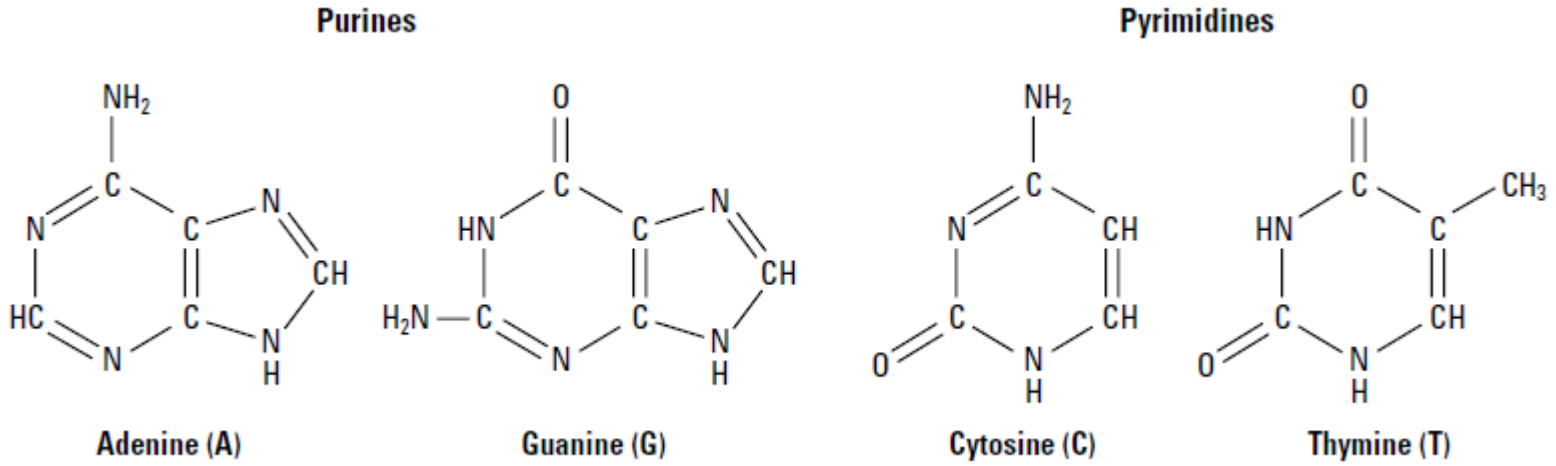
Nucleic Acids (NA)



- Primary structure
  - sequence of NA basis in chains
- Secondary structure
  - set of interactions between nucleic basis
- Tertiary structure
  - 3D localization of atoms
- Quarternary structure
  - Higher organization levels
    - DNA in chromatin
    - Interaction of RNA units in ribosome or spliceosome.

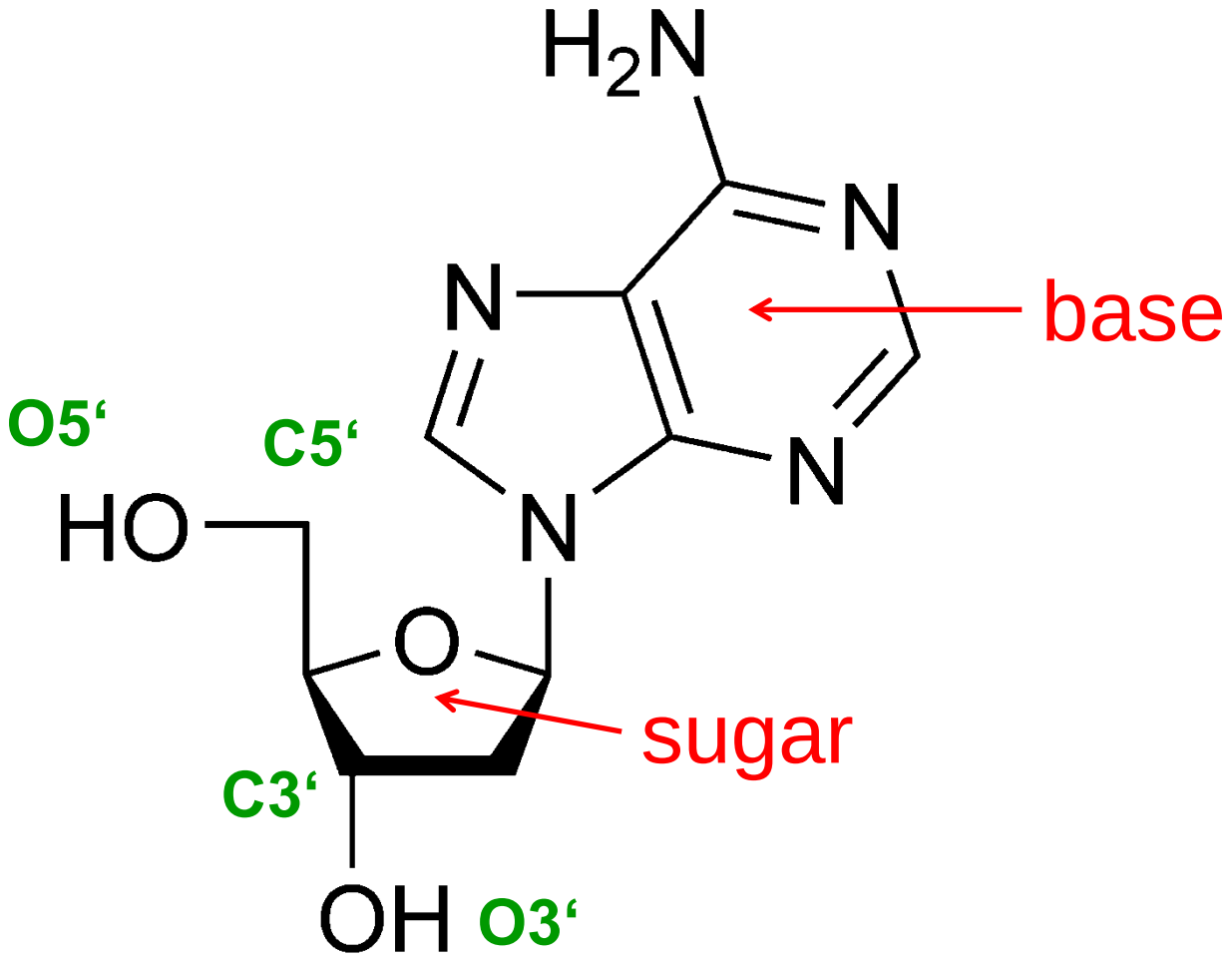
DNA – DeoxyriboNucleic Acid

- bases, deoxyribose sugar, phosphate – nucleotide
- Bases are flat → stacking
- p**Y**rimidines – C, T
- pu**R**ines – A, G



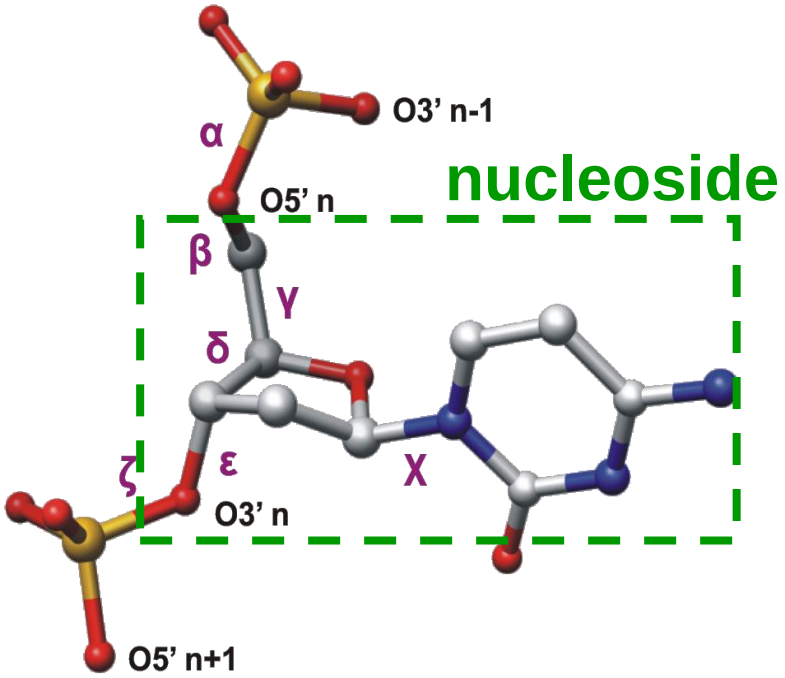
•http://www.umass.edu/molvis/tutorials/dna/, http://ich.vscht.cz/~svozil/teaching.html

Nucleoside



Nucleotide

- nucleosides are interconnected by phospohodiester bond
- nucleotide monophosphate

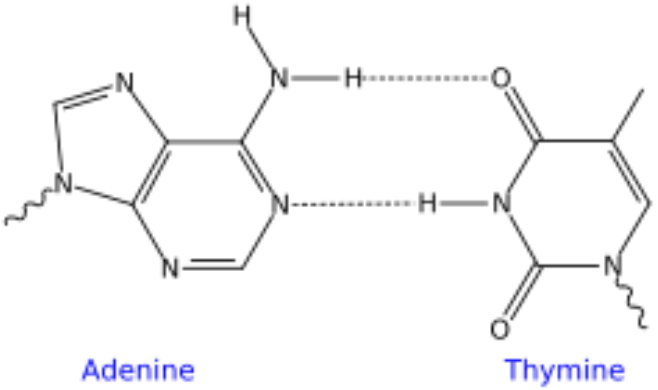


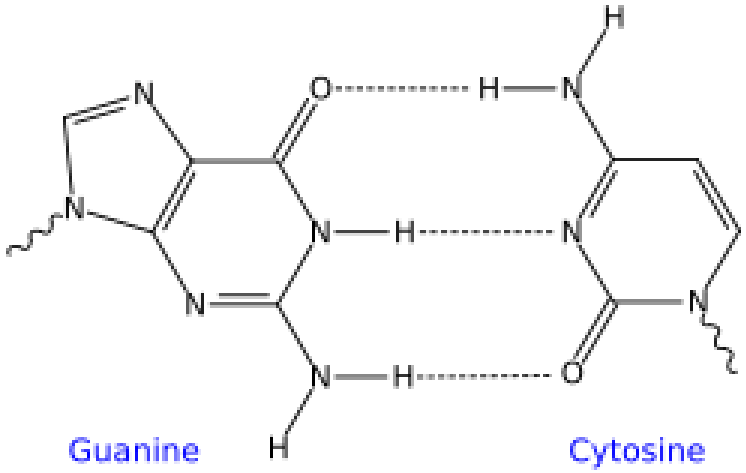
Watson-Crick pairing

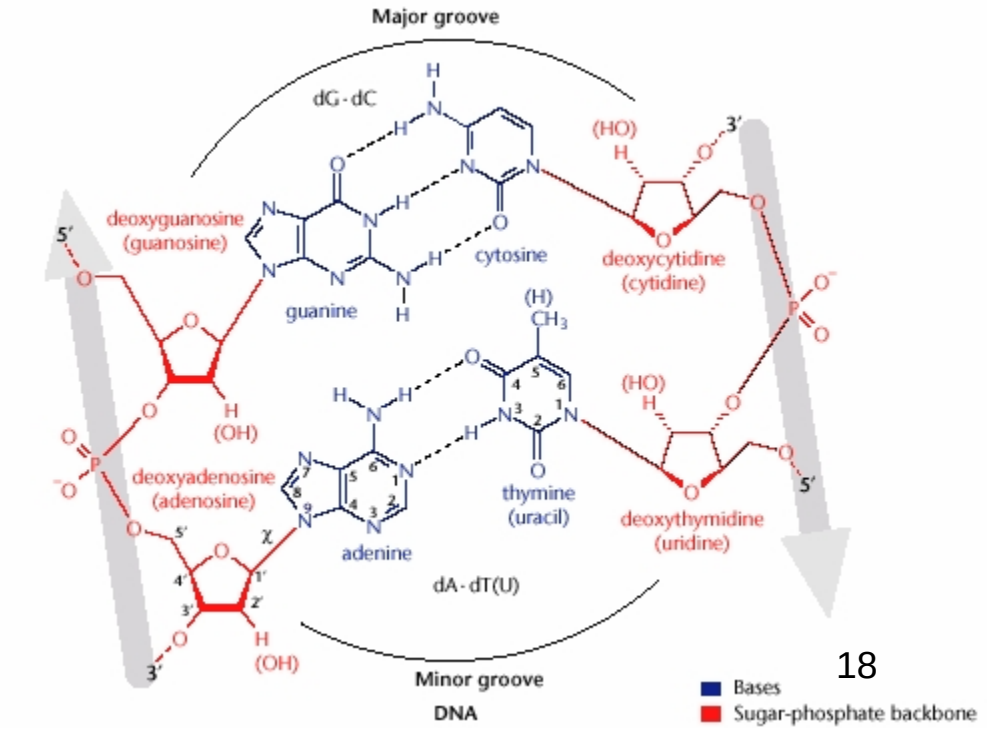
Bases complement each other.

Chargaffs' rules

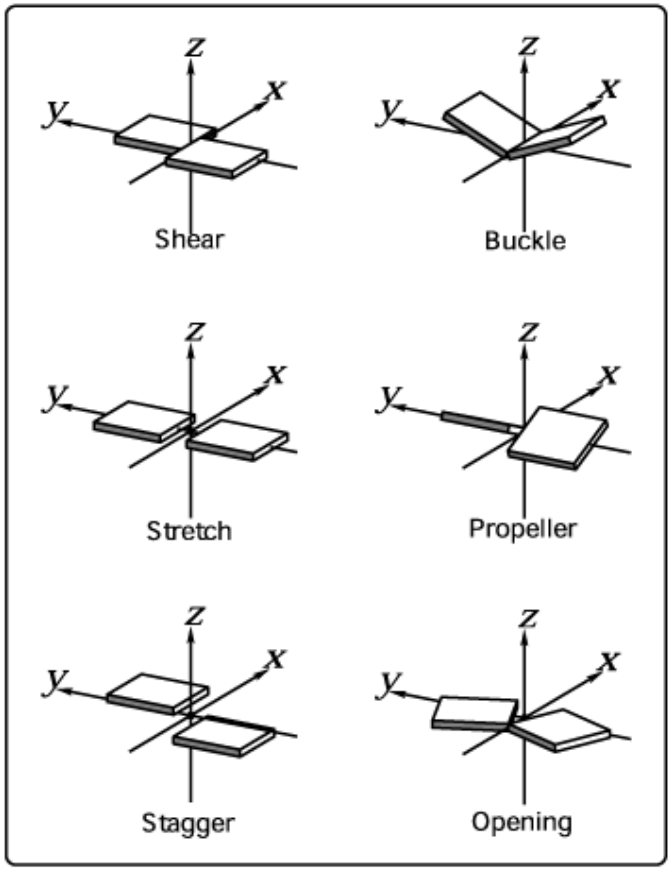
- amount of G = C
- amount of A = T

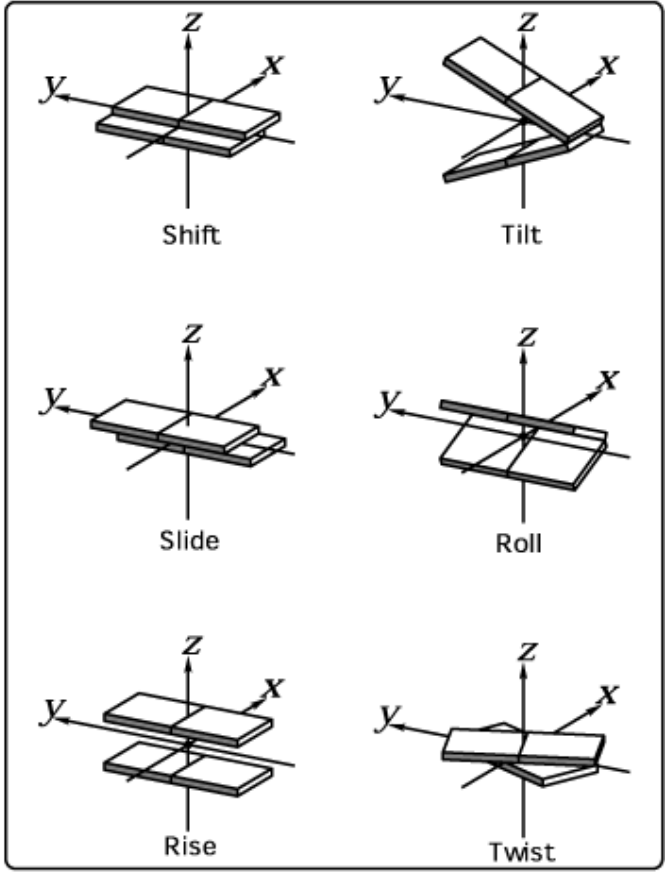


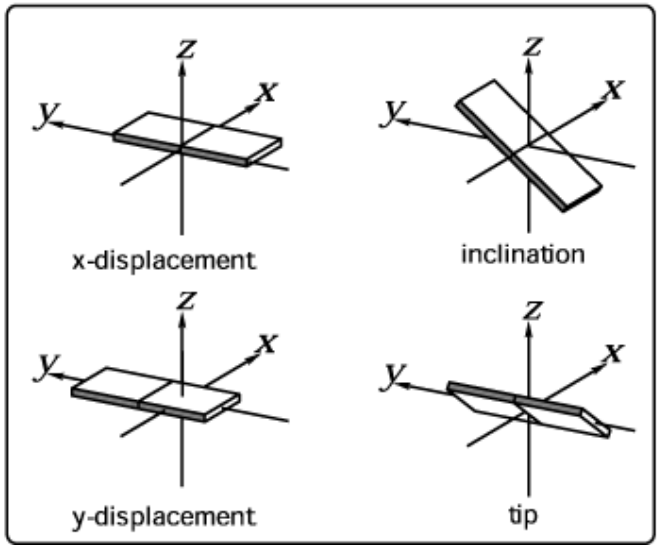


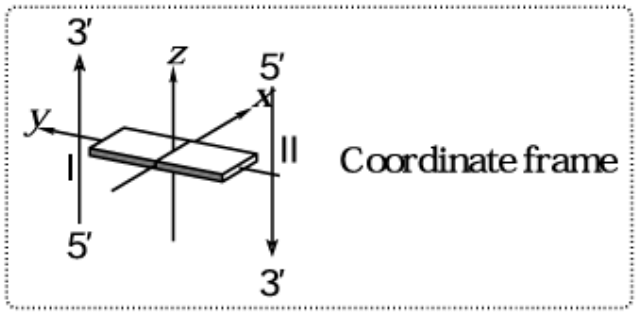


Párování



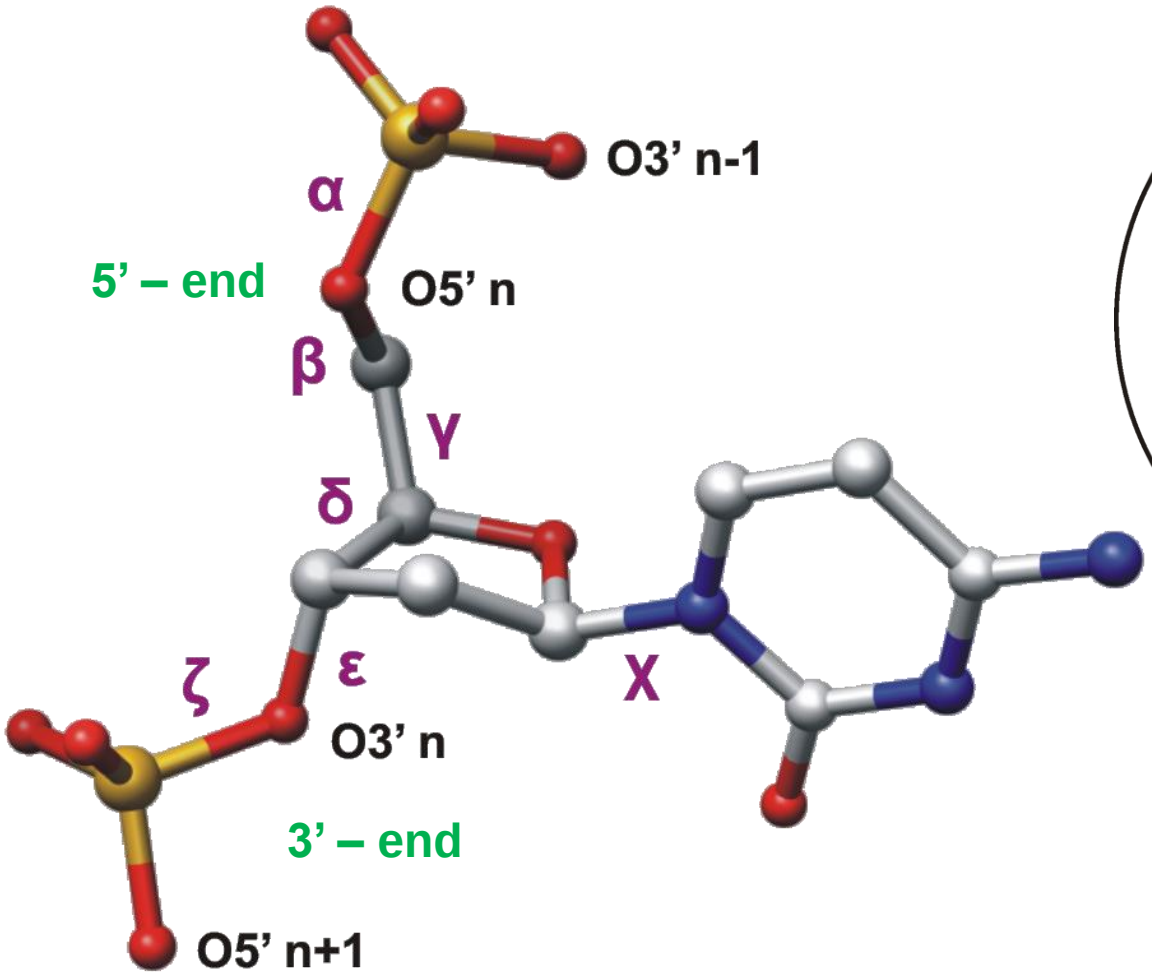


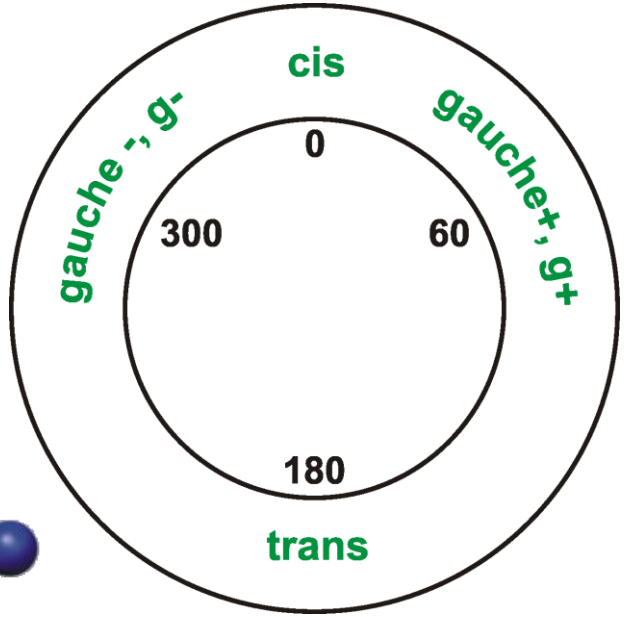




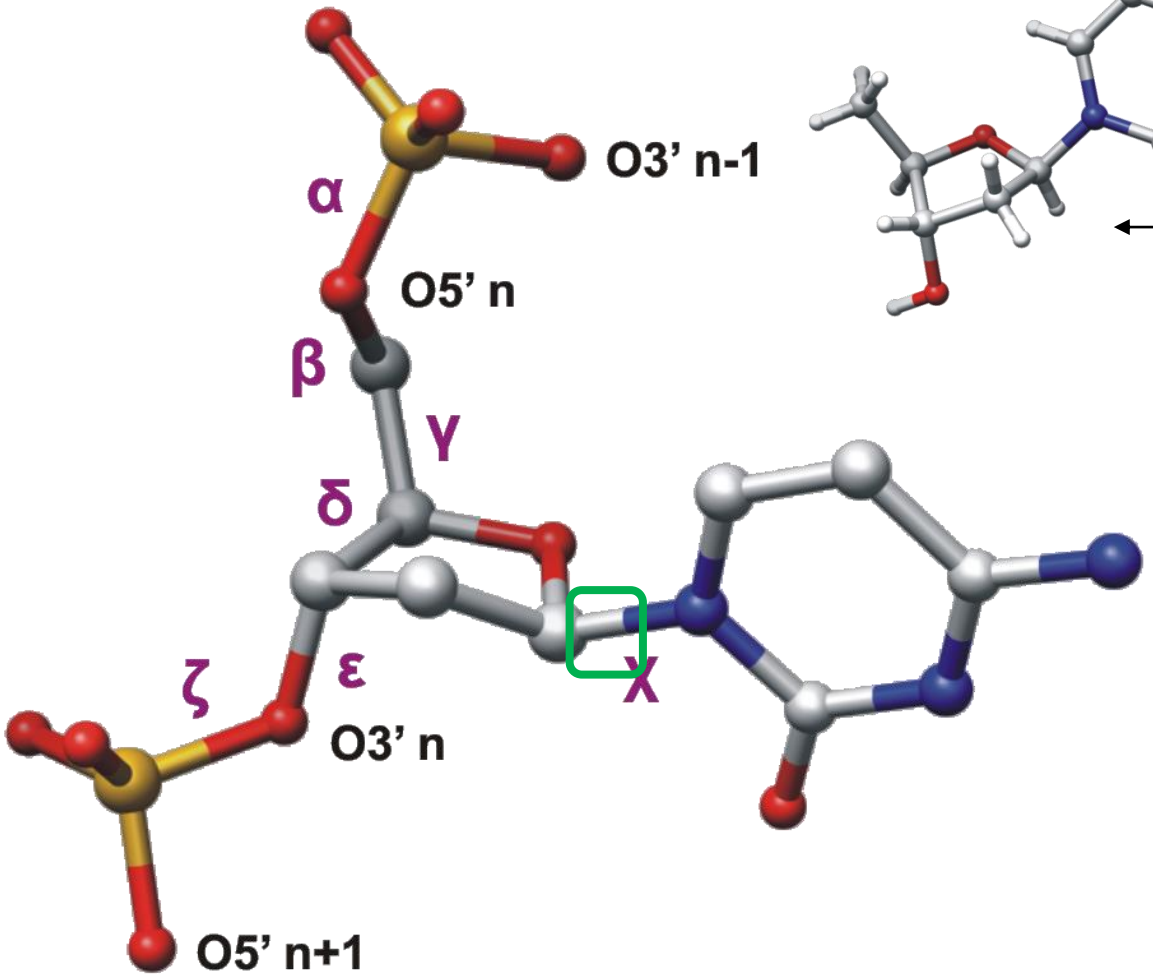
Images created with 3DNA illustrating positive values of designated parameters

DNA backbone

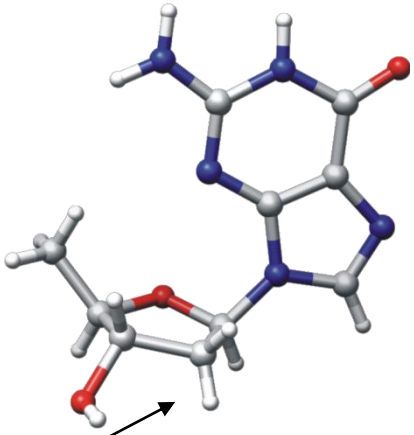




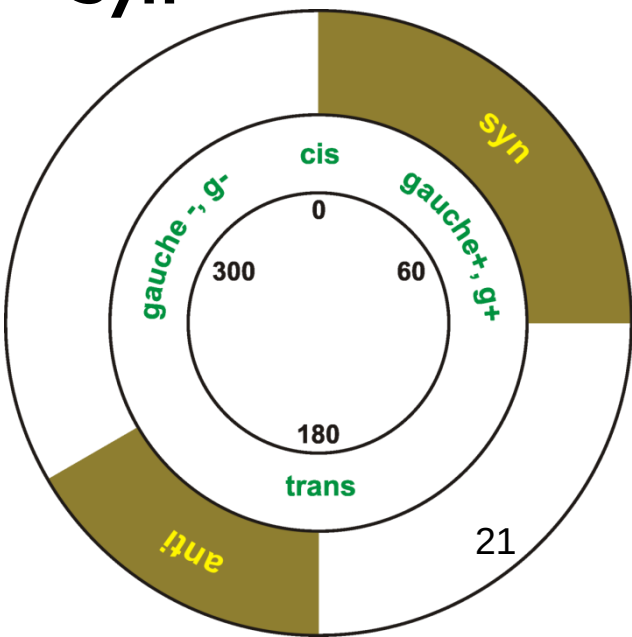
Base at sugar dihedrals



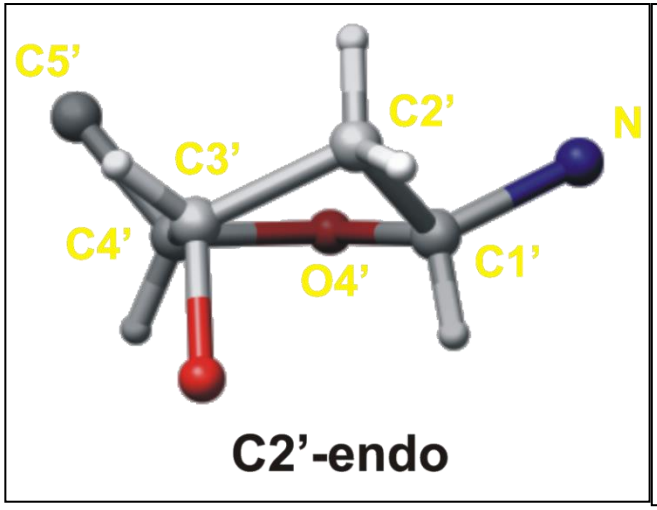
Anti



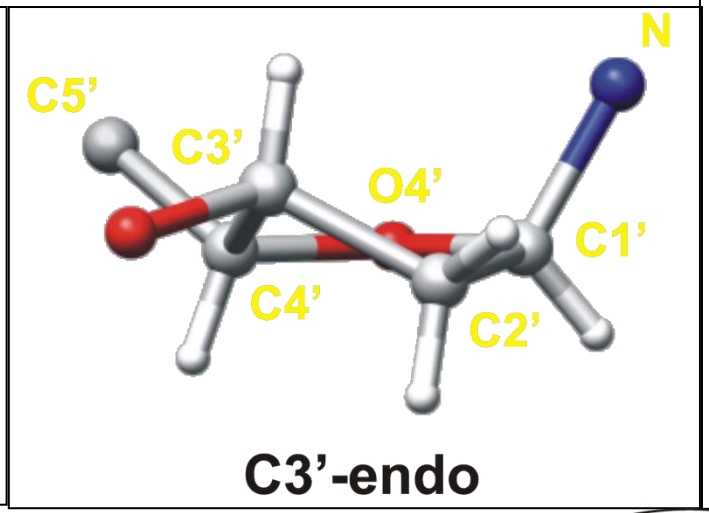
Syn



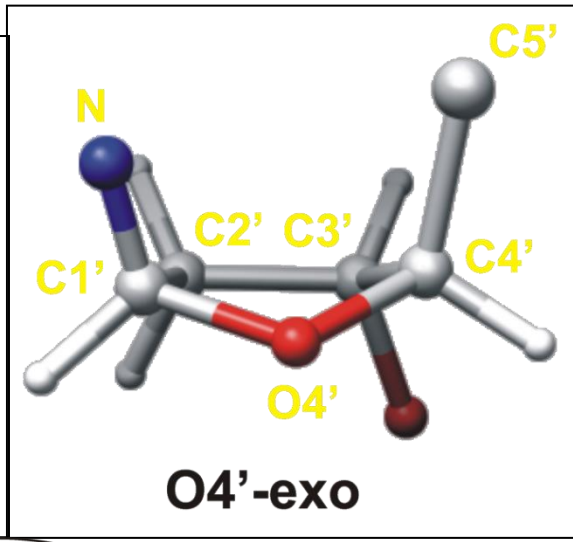
Sugar conformation



C2'-endo



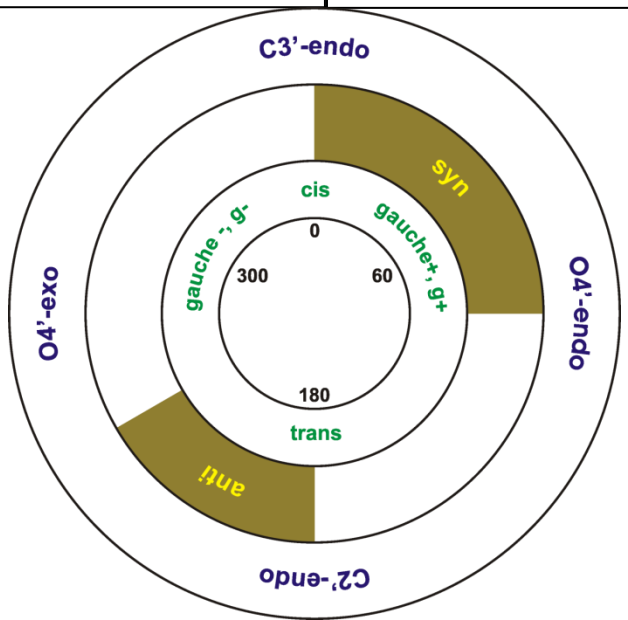
C3'-endo



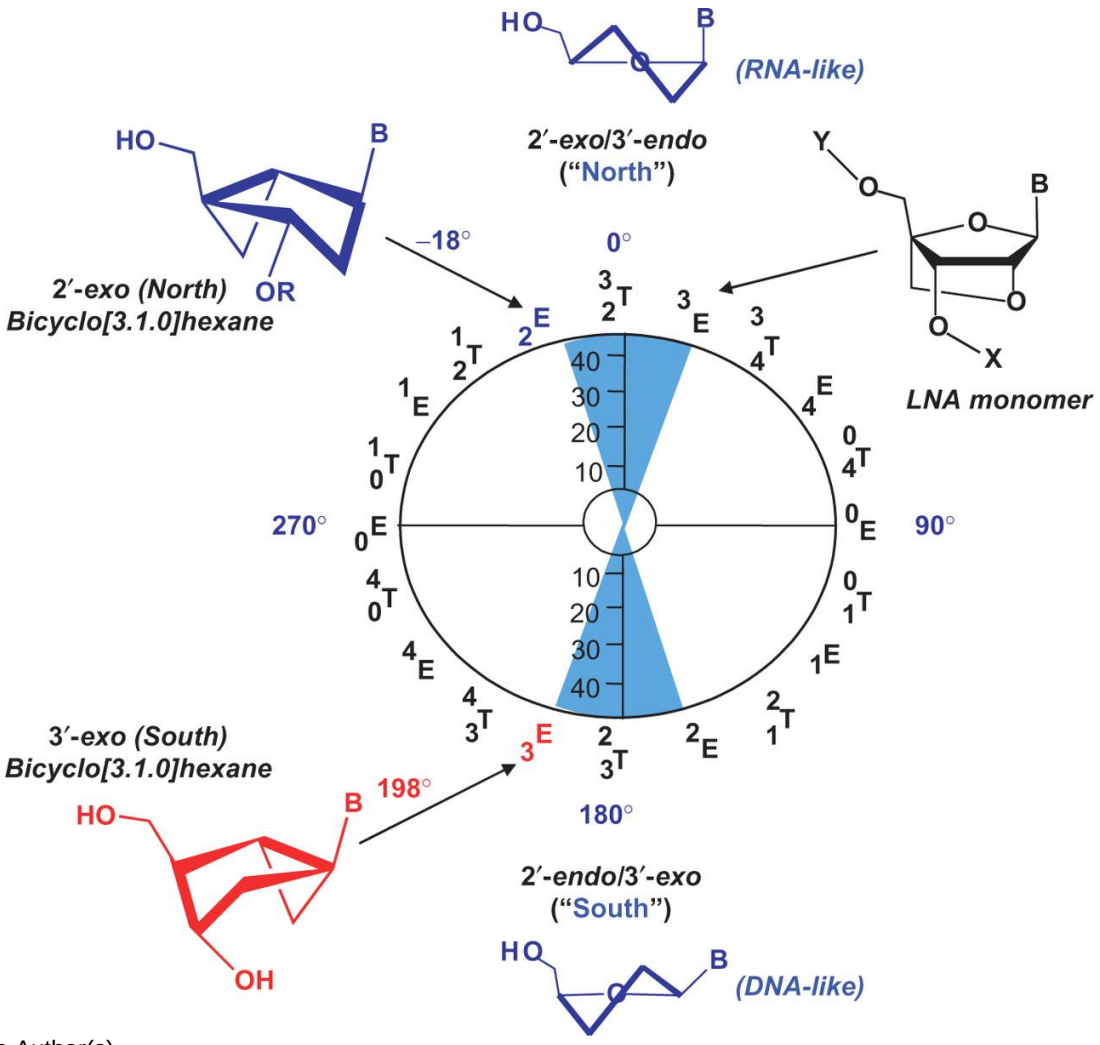
O4'-exo

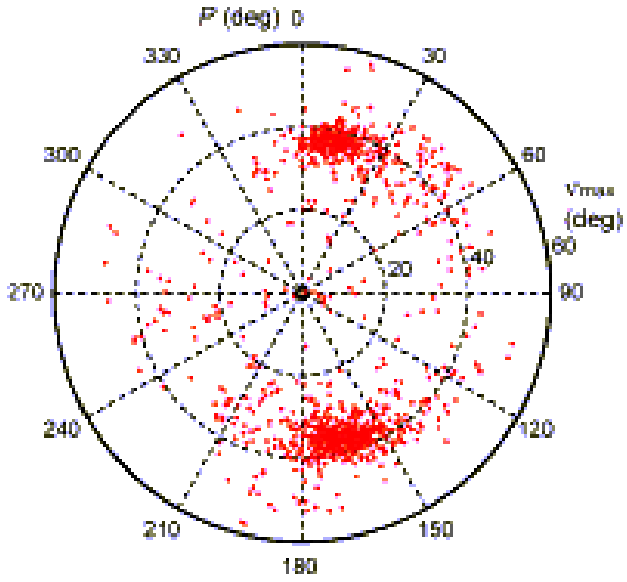
orientation with respect to C5'

- same side – endo
- opposite side – exo



Pseudorotational cycle for furanose ring puckers.





Pucker conformation of sugars in CSD database from PROSIT server

Maderia M et al. Nucl. Acids Res. 2007;35:1978-1991

Nucleic Acids Research

DNA Double helix

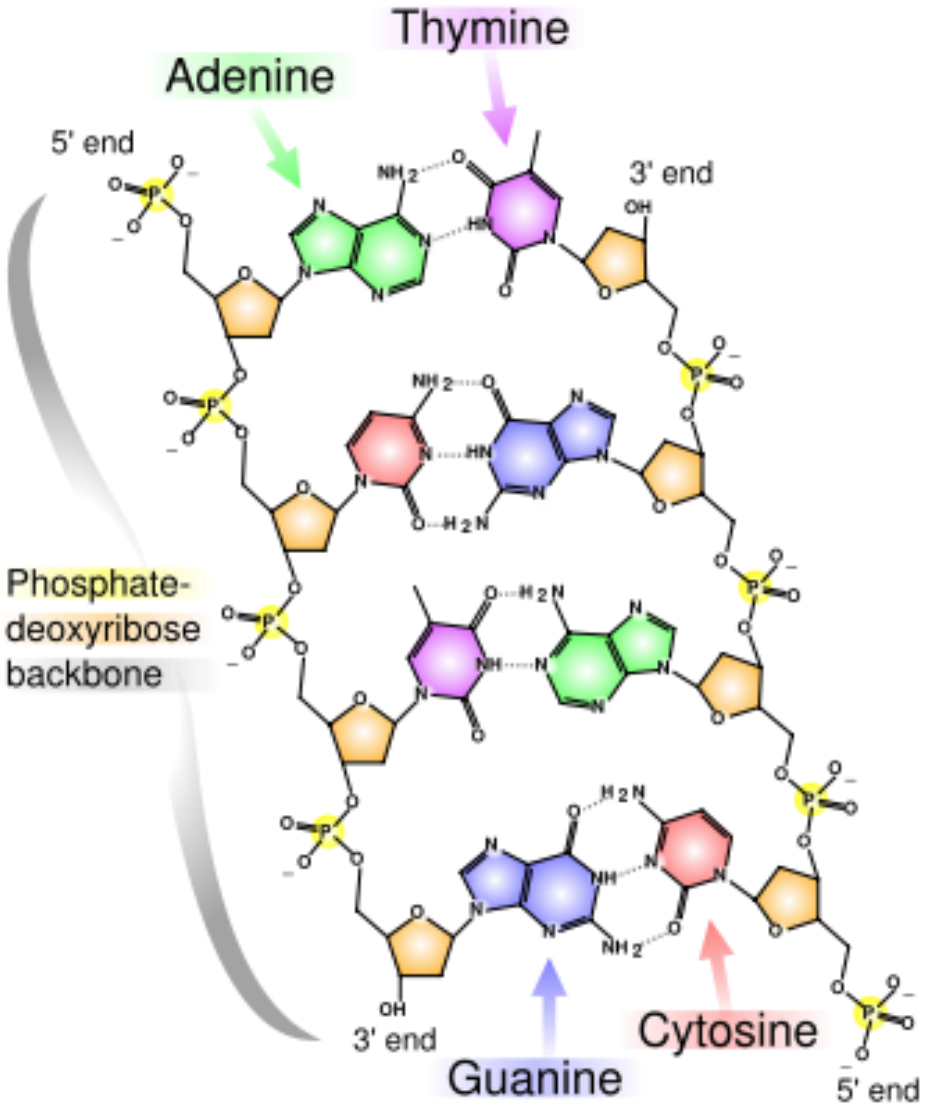
5’ ⟶ 3’

AATCGCTA

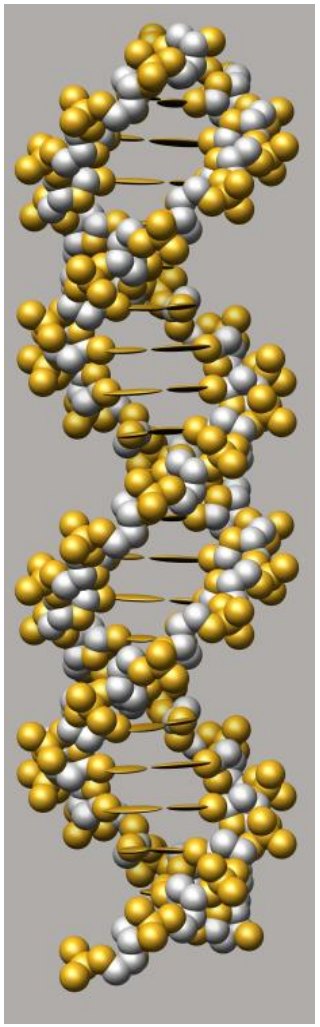
TTAGCGAT

3’ ⟵ 5’

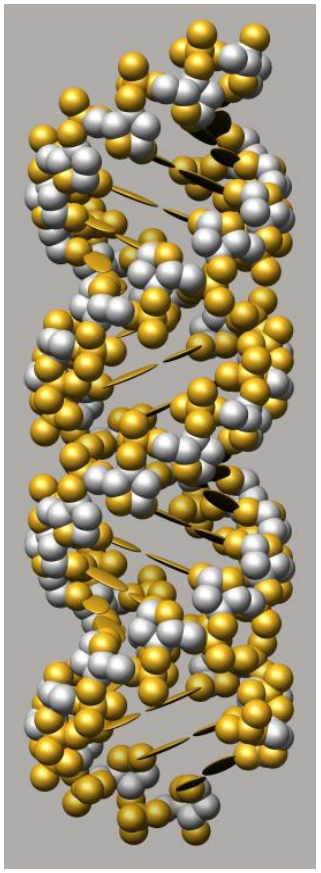
antiparallel



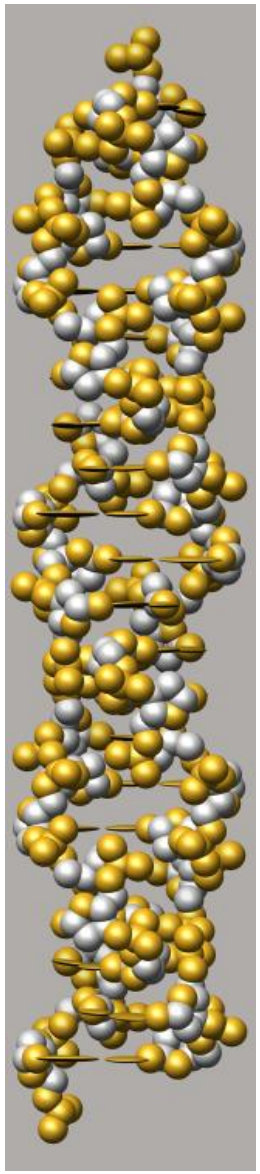
Types of DNA



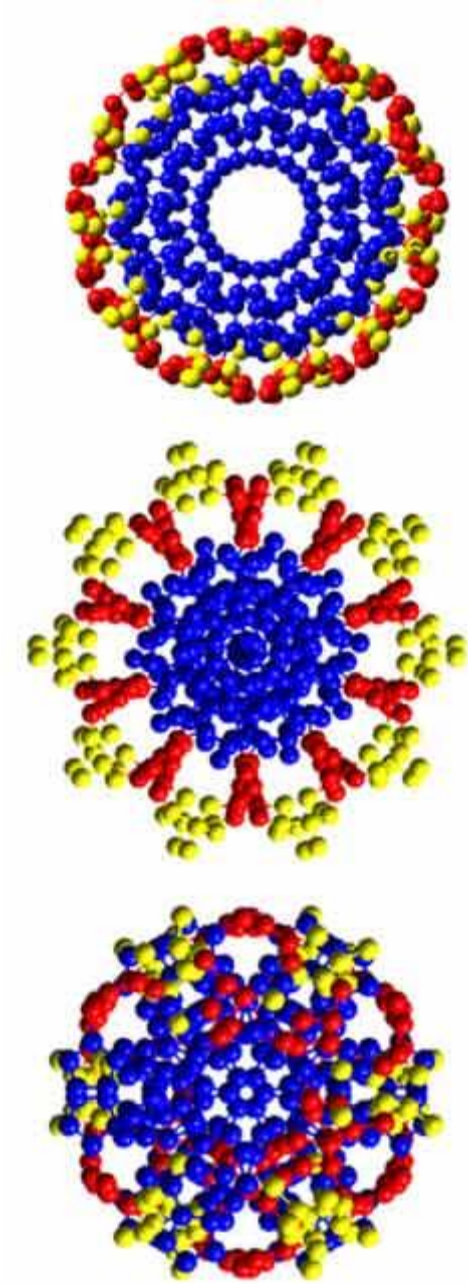
B-DNA



A-DNA



Z-DNA



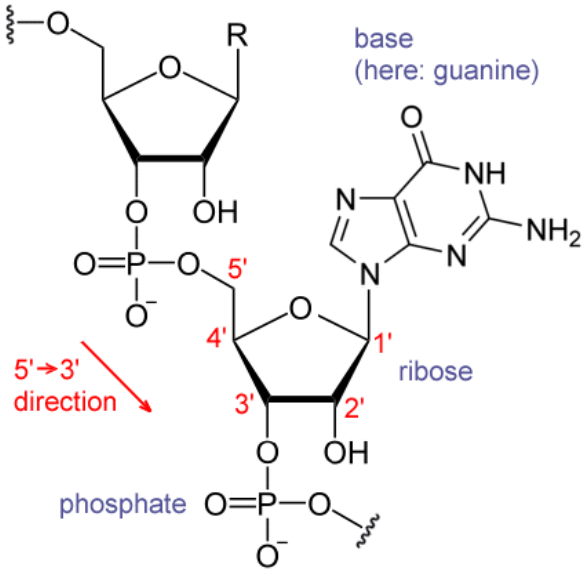
Biological role of different DNAs

- B-DNA
  - canonical DNA
  - predominant
- A-DNA
  - Conditions of lower humidity, common in crystallographic experiments. However, they’re artificial.
  - In vivo – local conformations induced e.g. by interaction with proteins.
- Z-DNA
  - No definite biological significance found up to now.
  - It is commonly believed to provide torsional strain relief (supercoiling) while DNA transcription occurs.
  - The potential to form a Z-DNA structure also correlates with regions of active transcription.

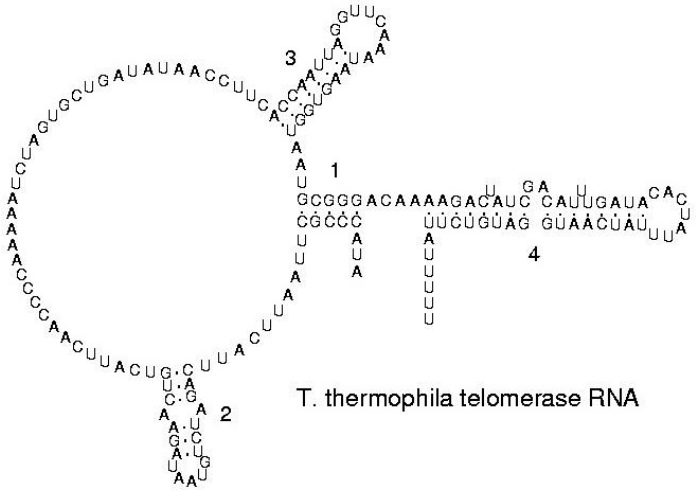
Different sets of DNA

- nuclear DNA
  - cell's nucleus
  - majority of functions cell carries out
  - sequencing the genome – scientists mean nuclear DNA
- mitochondrial DNA
  - *mt*DNA
  - circular, in human very short (17 kbp) with 37 genes (controling cellular metabolism)
  - all *mt*DNA comes from mom
- chloroplast DNA
  - *cp*DNA
  - circular and fairly large (120 – 160 kbp), with only 120 genes
  - inheritance is either maternal, or paternal

RNA - ribonucleic acid

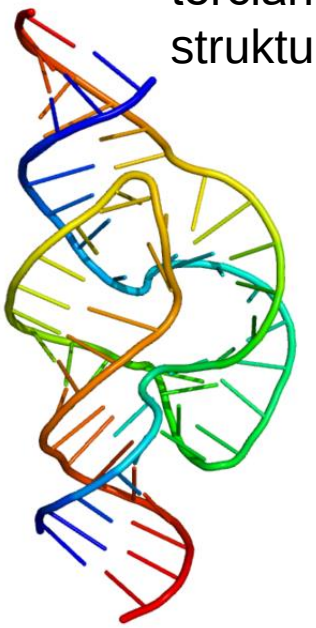


primární struktura



sekundární struktura

terciární
struktura



hammerhead
ribozyme 2GOZ

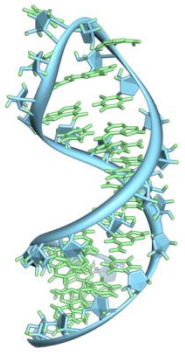
RNA

RNAs involved in protein synthesis

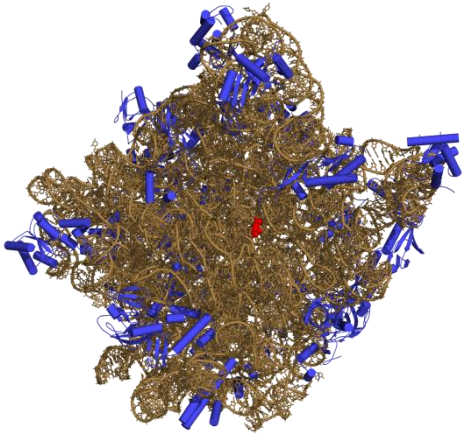
Type	Abbr.	Function	Distribution	Ref.
Messenger RNA	mRNA	Codes for protein	All organisms	
Ribosomal RNA	rRNA	Translation	All organisms	
Signal recognition particle RNA	7SL RNA or SRP RNA	Membrane integration	All organisms	[1]
Transfer RNA	tRNA	Translation	All organisms	
Transfer-messenger RNA	tmRNA	Rescuing stalled ribosomes	Bacteria	[2]

RNAs involved in post-transcriptional modification or DNA replication

Type	Abbr.	Function	Distribution	Ref.
Small nuclear RNA	snRNA	Splicing and other functions	Eukaryotes and archaea	[3]
Small nucleolar RNA	snoRNA	Nucleotide modification of RNAs	Eukaryotes and archaea	[4]
SmY RNA	SmY	mRNA trans-splicing	Nematodes	[5]
Small Cajal body-specific RNA	scaRNA	Type of snoRNA; Nucleotide modification of RNAs		
Guide RNA	gRNA	mRNA nucleotide modification	Kinetoplastid mitochondria	[6]
Ribonuclease P	RNase P	tRNA maturation	All organisms	[7]
Ribonuclease MRP	RNase MRP	rRNA maturation, DNA replication	Eukaryotes	[8]
Y RNA		RNA processing, DNA replication	Animals	[9]
Telomerase RNA		Telomere synthesis	Most eukaryotes	[10]



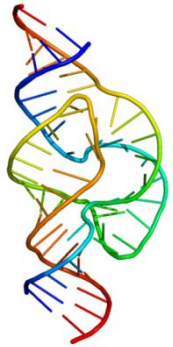
pre-mRNA hairpin



50S-ribozome

Regulatory RNAs

Type	Abbr.	Function	Distribution	Ref.
Antisense RNA	aRNA	Transcriptional attenuation / mRNA degradation / mRNA stabilisation / Translation block	All organisms	[11][12]
Cis-natural antisense transcript		Gene regulation		
CRISPR RNA	crRNA	Resistance to parasites, probably by targeting their DNA	Bacteria and archaea	[13]
Long noncoding RNA	Long ncRNA	Various	Eukaryotes	
MicroRNA	miRNA	Gene regulation	Most eukaryotes	[14]
Piwi-interacting RNA	piRNA	Transposon defense, maybe other functions	Most animals	[15][16]
Small interfering RNA	siRNA	Gene regulation	Most eukaryotes	[17]
Trans-acting siRNA	tasiRNA	Gene regulation	Land plants	[18]
Repeat associated siRNA	rasiRNA	Type of piRNA; transposon defense	Drosophila	[19]
7SK RNA	7SK	negatively regulating CDK9/cyclin T complex		



hammerhead ribozyme

2GOZ

Parasitic RNAs

Type	Function	Distribution	Ref.
Retrotransposon	Self-propagating	Eukaryotes and some bacteria	[20]
Viral genome	Information carrier	Double-stranded RNA viruses, positive-sense RNA viruses, negative-sense RNA viruses, many satellite viruses and reverse transcribing viruses	
Viroid	Self-propagating	Infected plants	[21]
Satellite RNA	Self-propagating	Infected cells	

Other RNAs

Type	Abbr.	Function	Distribution	Ref.
Vault RNA	vRNA	Expulsion of xenobiotics, maybe		[22]

http://en.wikipedia.org/wiki/List_of_RNAs

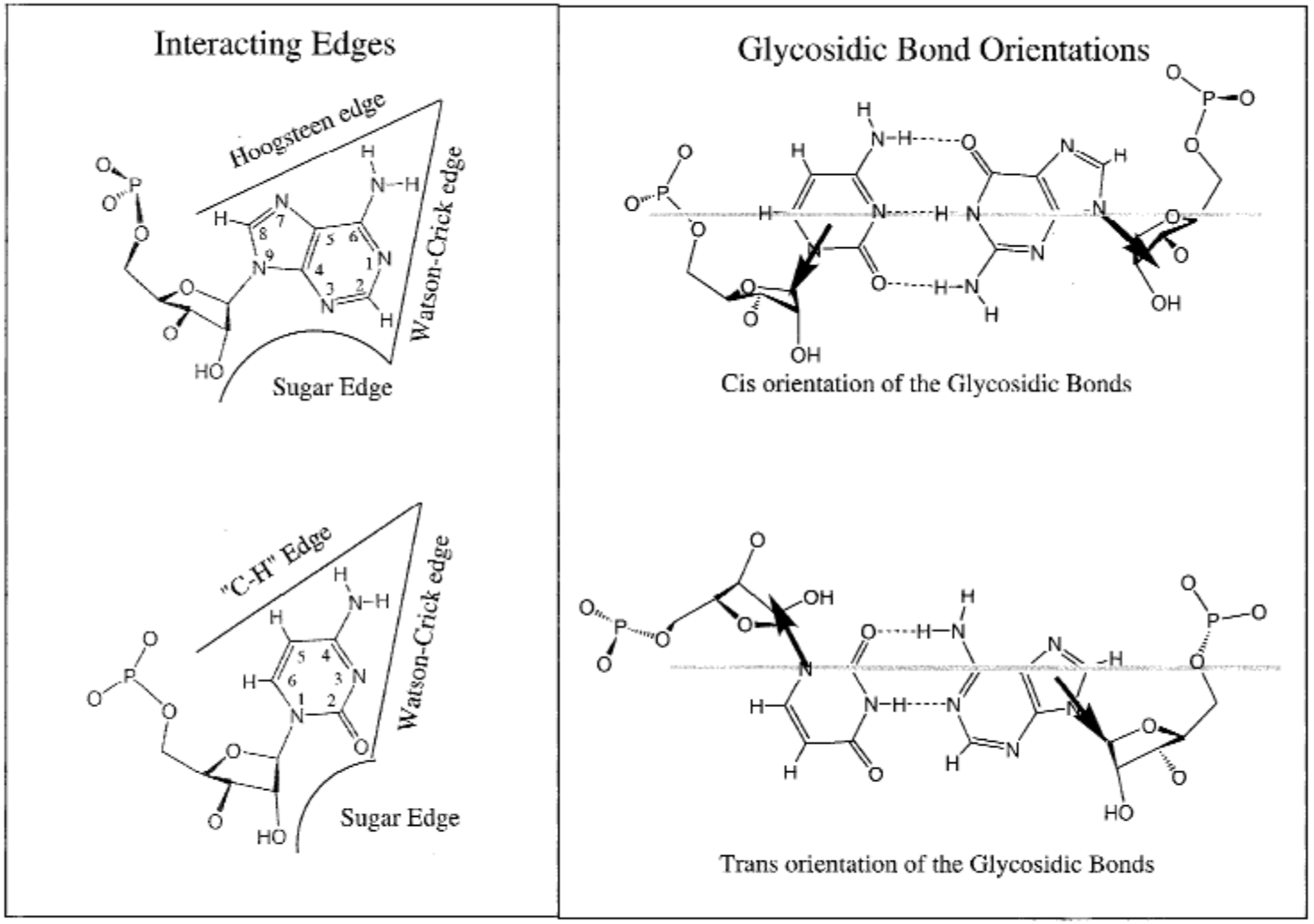


FIGURE 1. Left panel: Purine (A or G, indicated by "R") and pyrimidine (C or U, indicated by "Y") bases provide three edges for interaction, as shown for adenosine and cytosine. The Watson–Crick edge comprises A(N6)/G(O6), R(N1), A(C2)/G(N2), U(O4)/C(N4), Y(N3), and Y(O2). The Hoogsteen edge comprises A(N6)/G(O6), R(N7), U(O4)/C(N4), and Y(C5). The Sugar-edge comprises A(C2)/G(N2), R(N3), Y(O2), and the ribose hydroxyl group, O2′. Right panel: The *cis* and *trans* orientations are defined relative to a line drawn parallel to and between the *base-to-base* hydrogen bonds in the case of two hydrogen bonds or, in the case of three hydrogen bonds, along the middle hydrogen bond.

N. B. Leontis, E. Westhof, RNA (2001), 7:499-512

Annotation of 2D RNA Structures

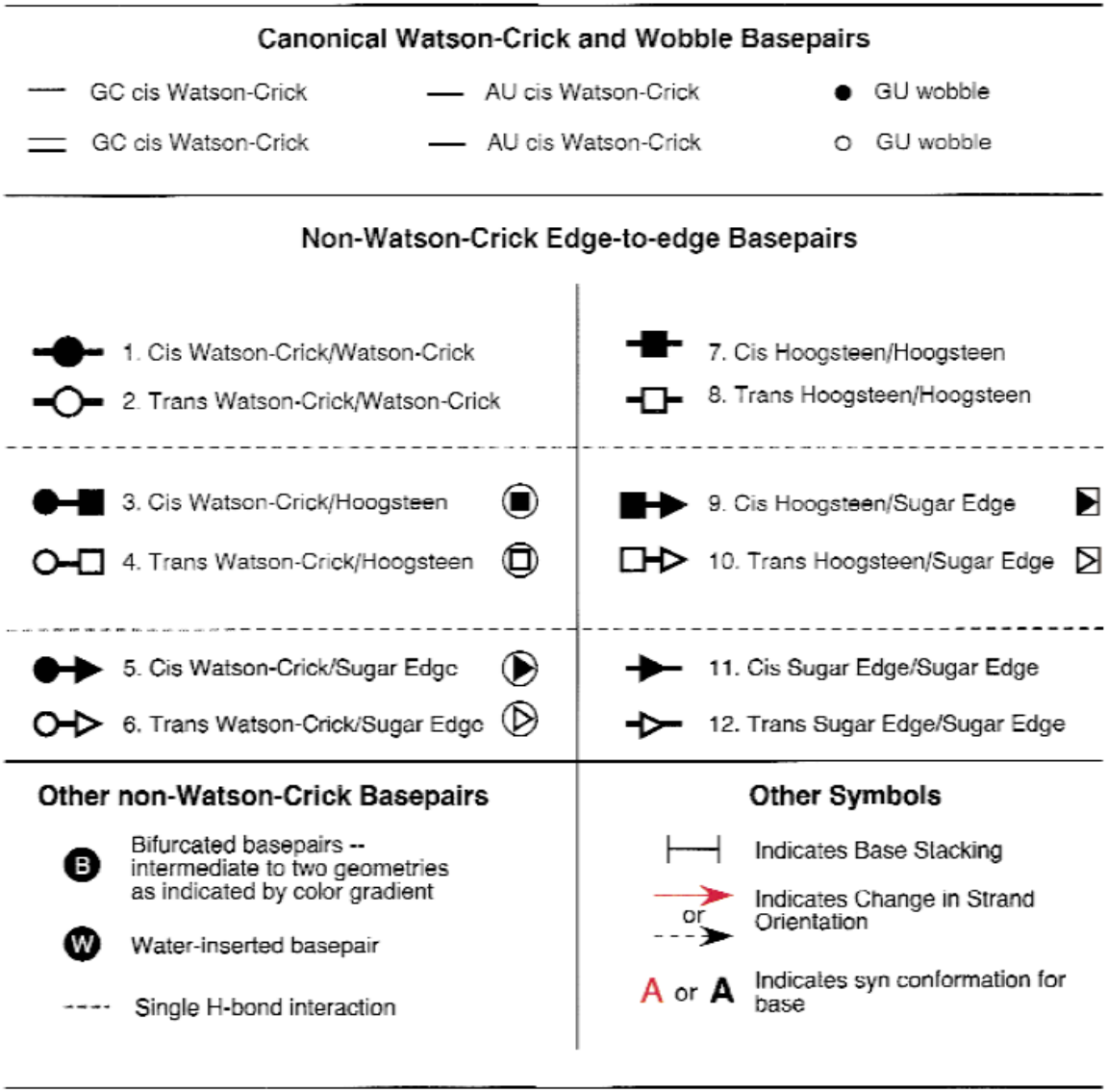
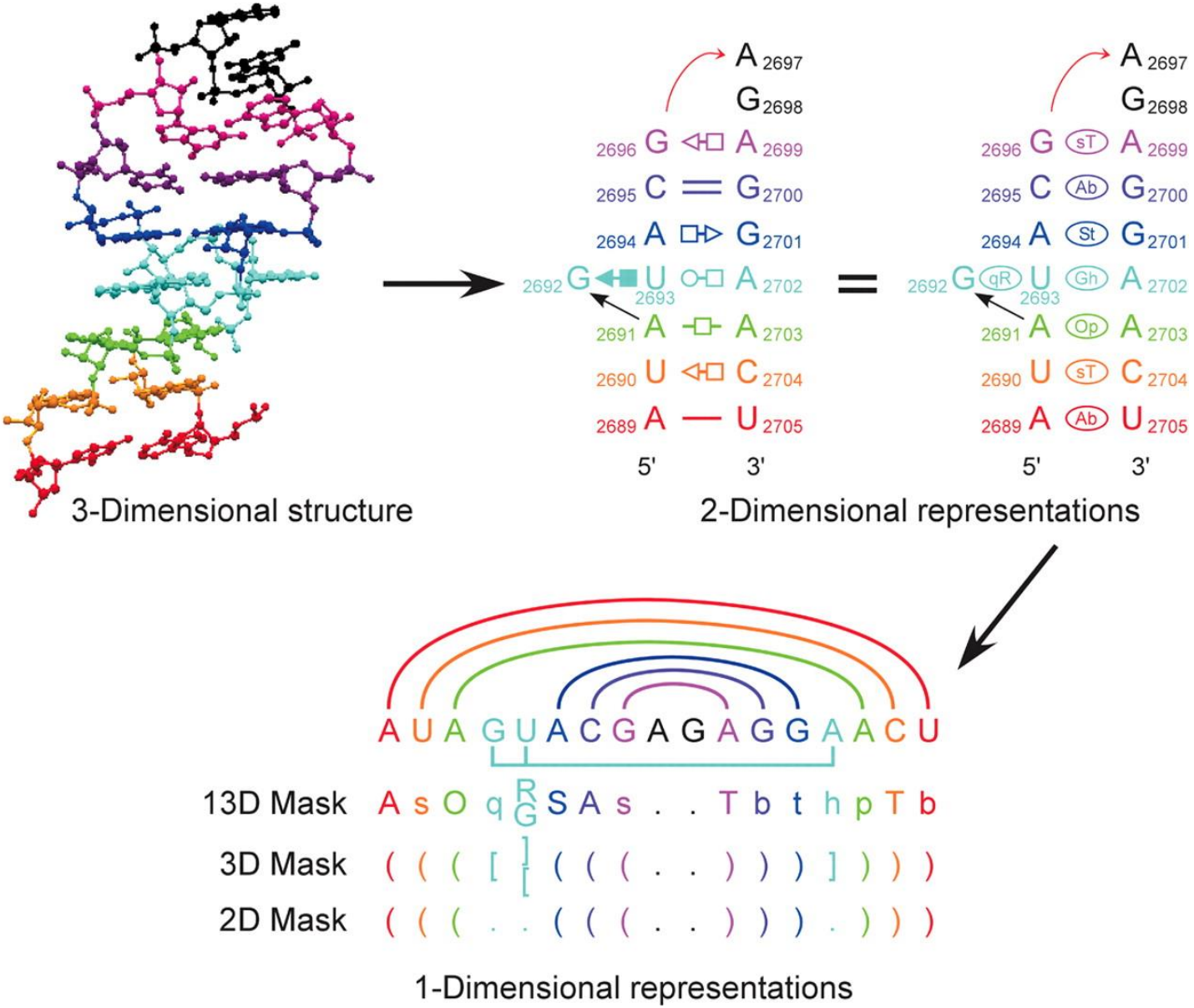


FIGURE 6. Suggested symbols for indicating tertiary interactions and other three-dimensional structural features in two-dimensional representations of RNA structures.

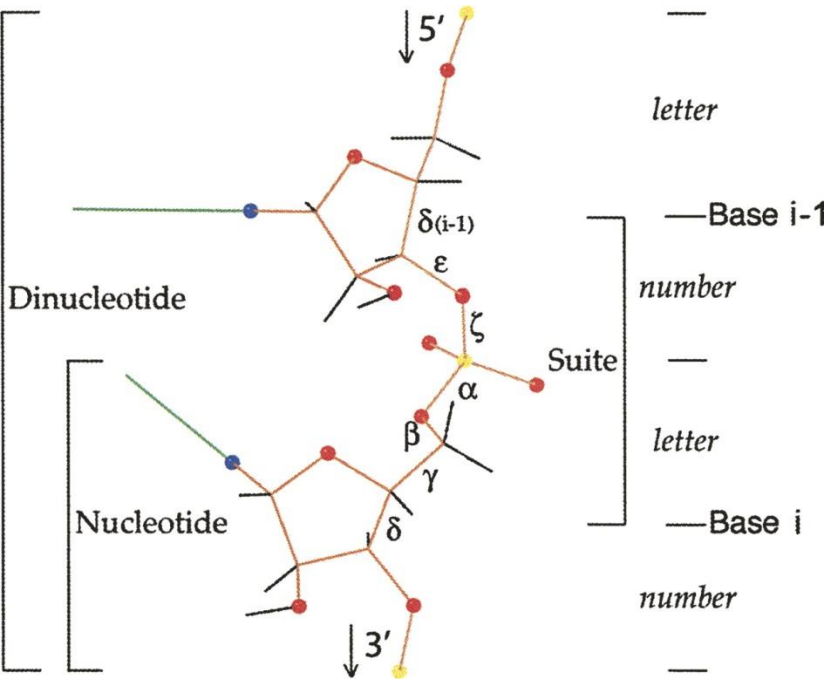
N. B. Leontis, E. Westhof, RNA (2001), 7:499-512

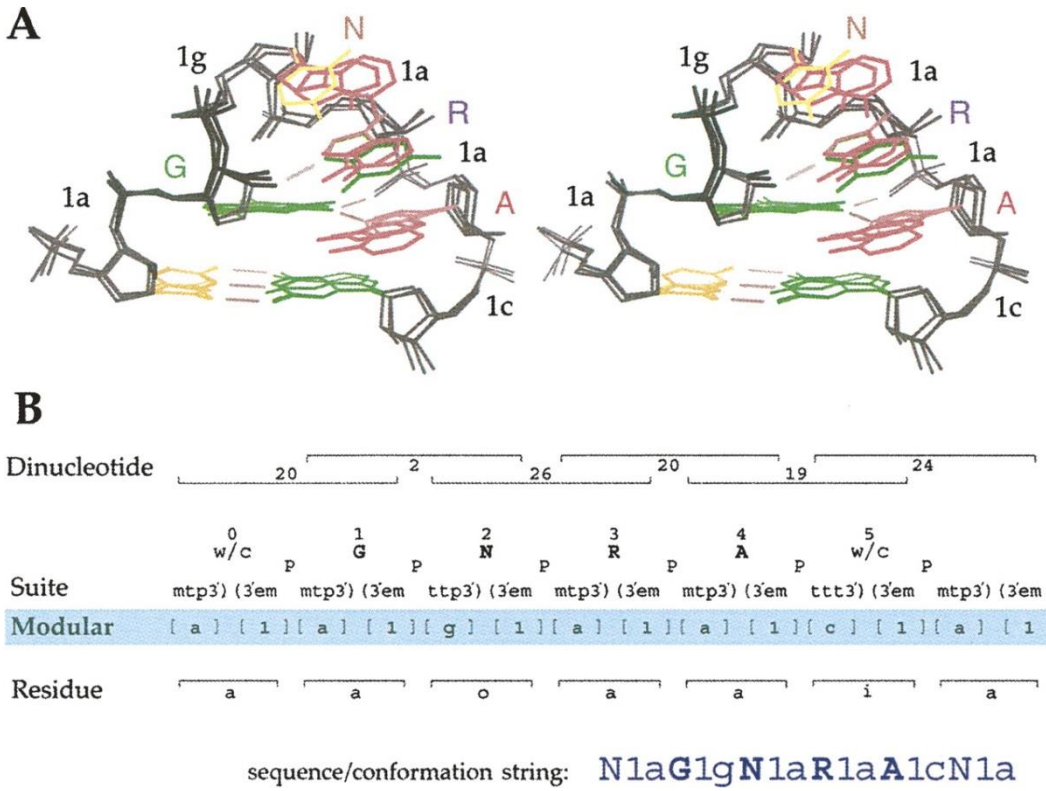
RNA Representations



Mokdad A , Leontis N B Bioinformatics 2006;22:2168-2170

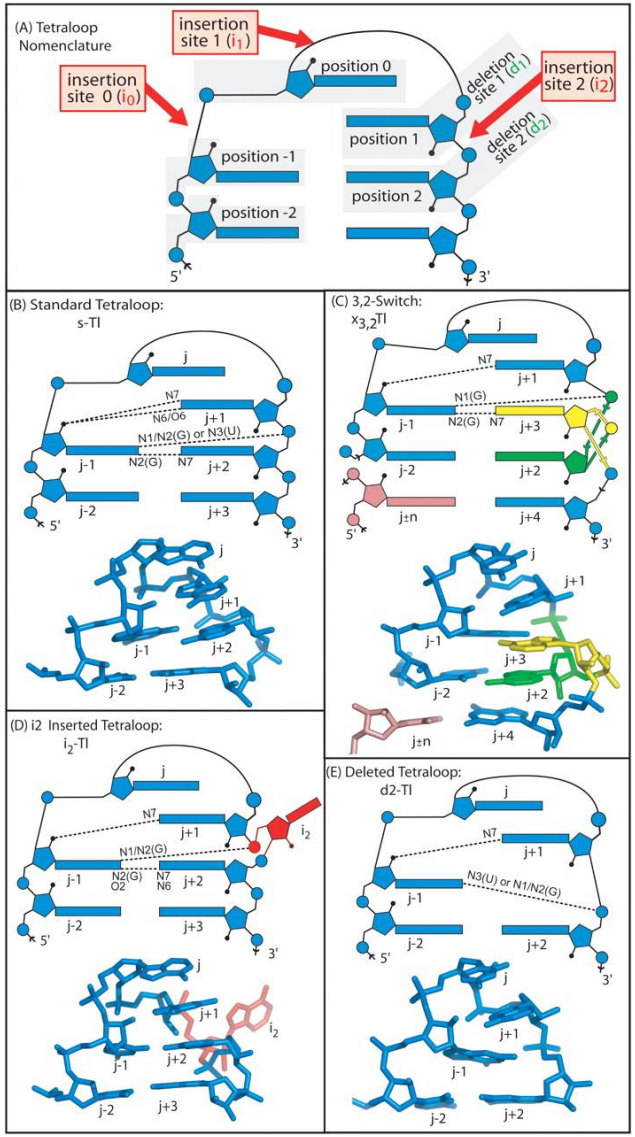
RNA Backbone

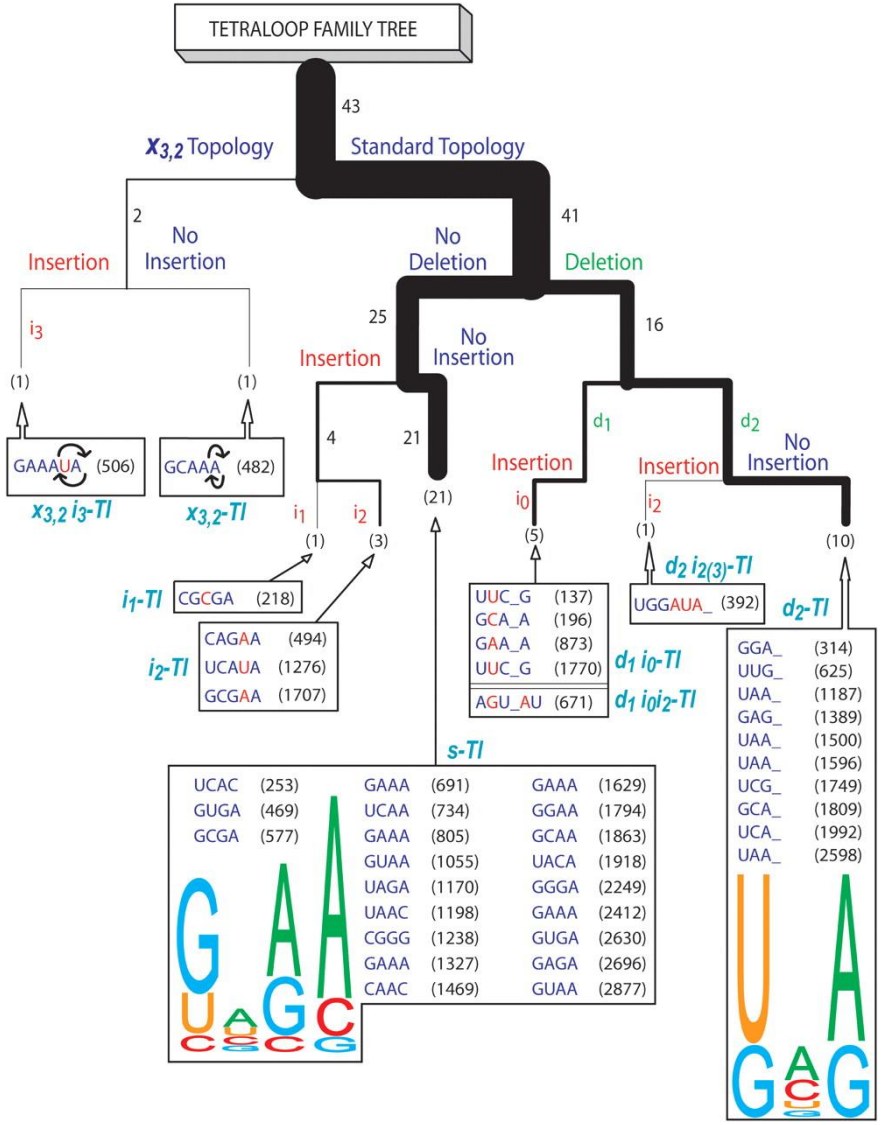




Richardson J S et al. RNA 2008;14:465-481

RNA Tetraloop Family Tree.

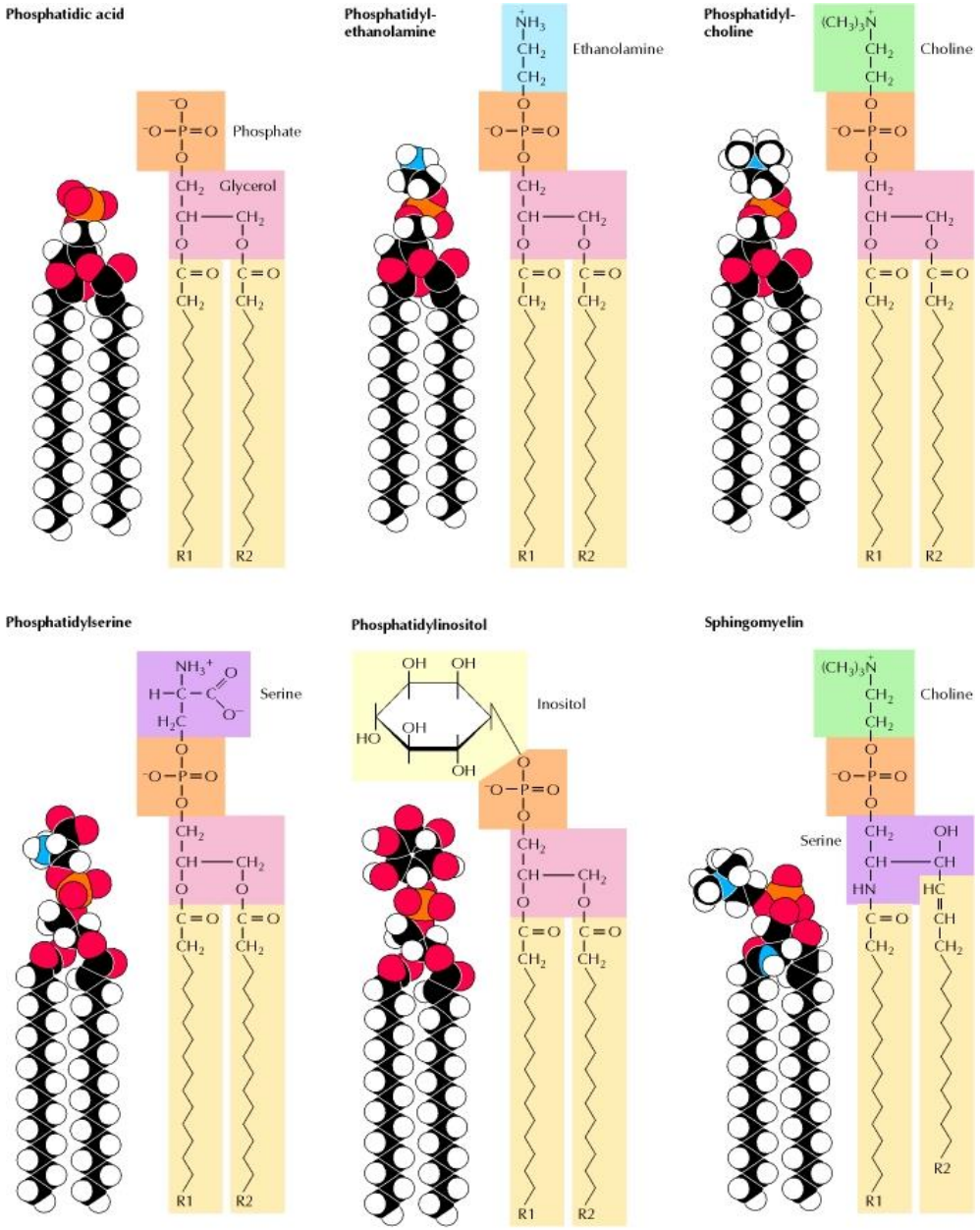


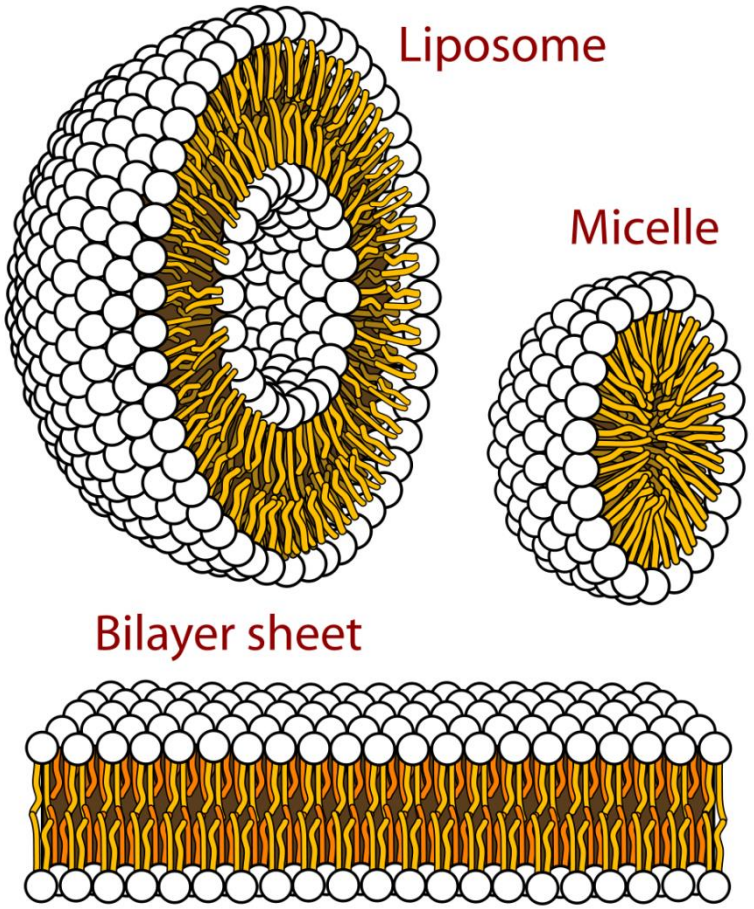


Nucleic Acids Research

Hsiao C et al. Nucl. Acids Res. 2006;34:1481-1491

Lipids





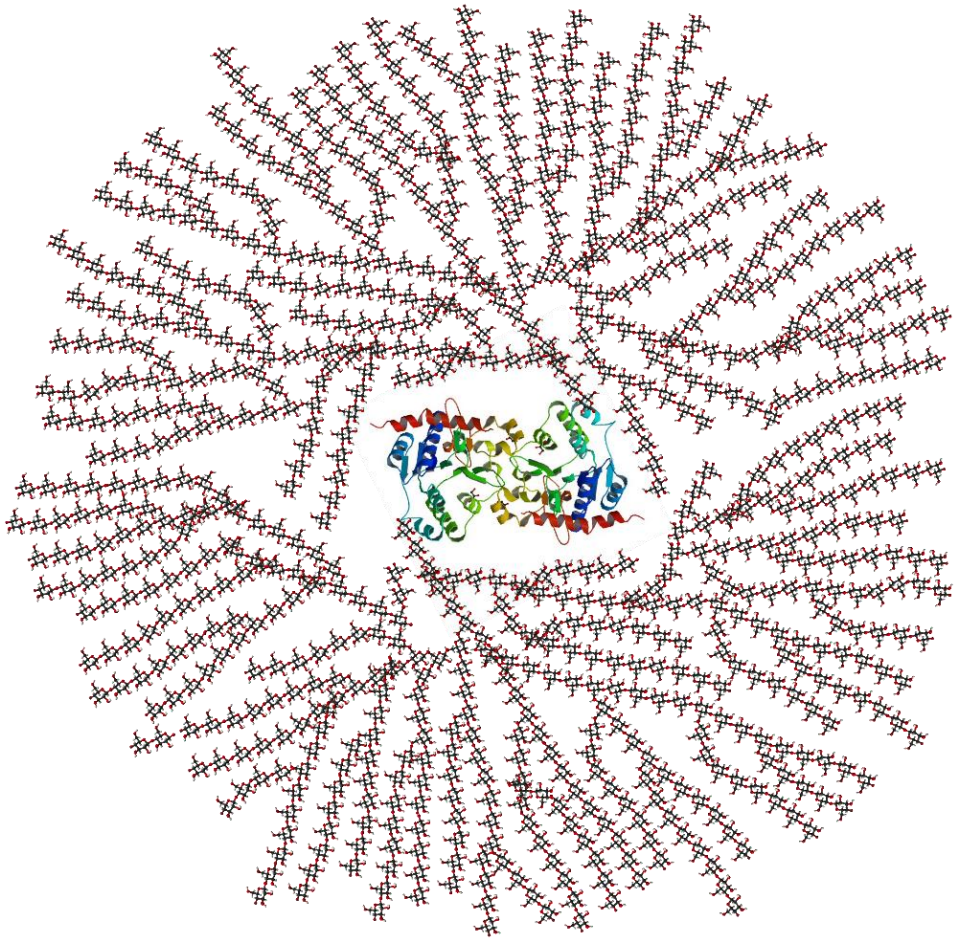
main phospholipids

M. Paloncyová, Lipid membranes report, 2010

Membrane proteins

Polysaccharides

- role:
  - Energy storage
  - Molecular recognition
- Harder to read in sequences than NA or proteins
- Quite often on extracellular proteins

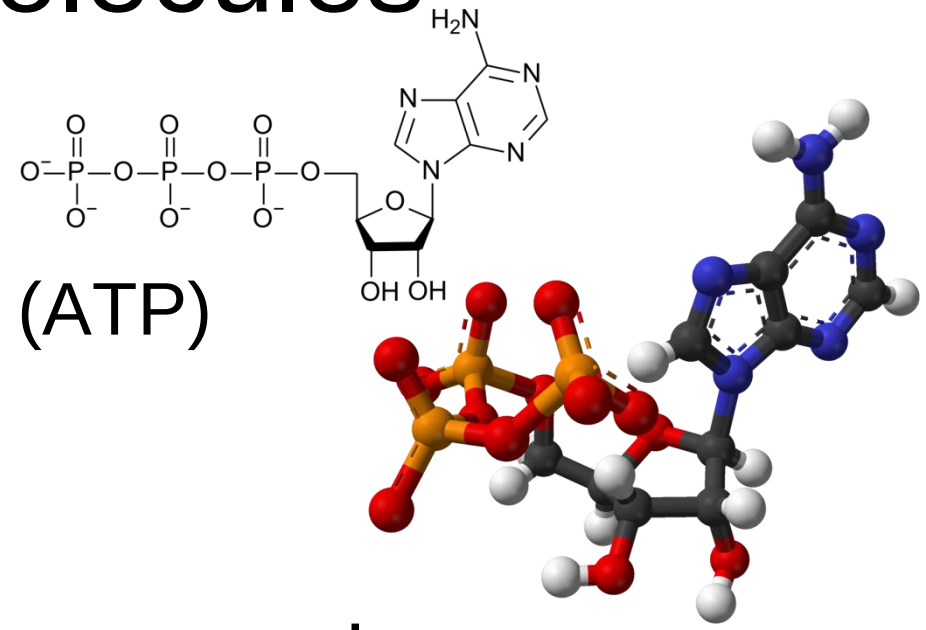


glycogen

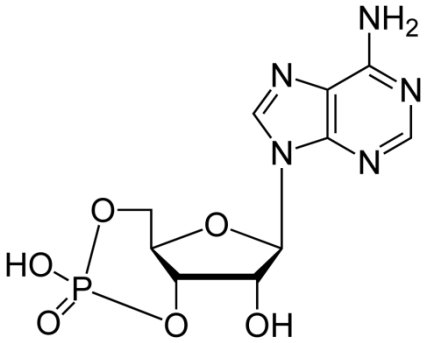
Glycoproteins

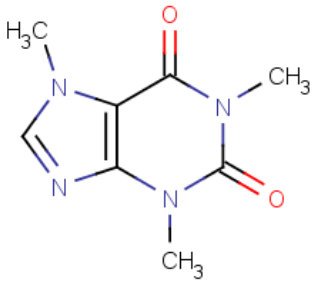
Small molecules

- NTP
  - Cell energy transporter (ATP)
  - Basic stones for NA

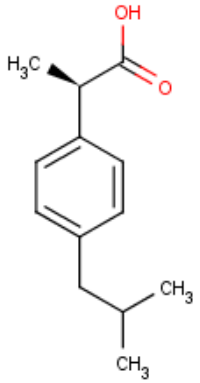


- Messengers, Agonists, antagonists
  - (cAMP, xenobiotics)





caffeine



ibuprofen

Drugs and where to find them

- Drugbank
- ChEBI
- ChEMBL
- PubChem

- IDSM