

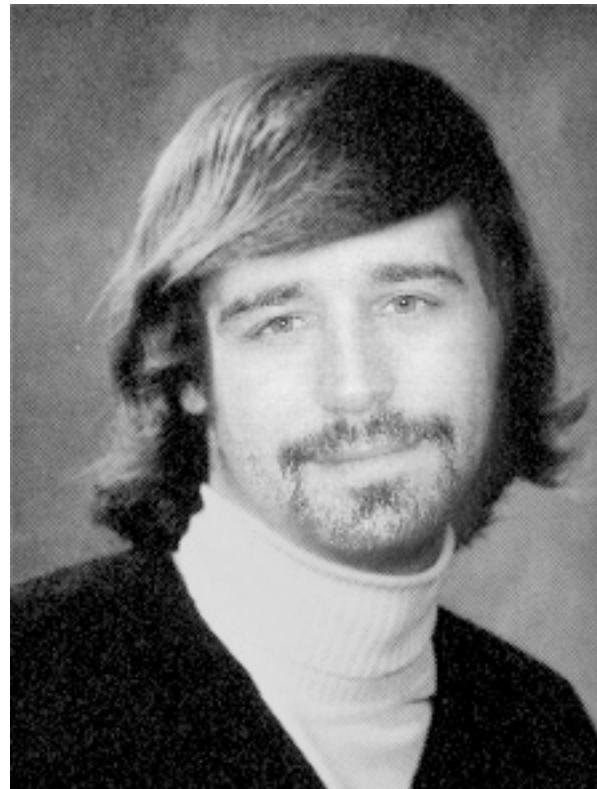
# Structural Bioinformatics

Francisco Martínez

David Dufour

- Estructura y biofísica de ácidos nucleicos y proteínas 25 febrero (DD)
- Bases de datos de estructura de proteínas, ácidos nucleicos y pequeñas moléculas 11 marzo (DD)
- Alineamiento y clasificación de estructura 25 marzo (DD)
- Predicción de estructura tridimensional de ácidos nucleicos y proteínas 15 abril (DD)
- Docking de pequeñas moléculas en la superficie de estructura de proteínas 29 abril (FM)
- Aplicaciones para el desarrollo de nuevos fármacos 13 mayo (FM)

# Structural Bioinformatics



- Intro Genómica Estructural
- Laboratorios 15 abril y 13 mayo

# Estructura y biofísica de ácidos nucleicos y proteínas



# Summary

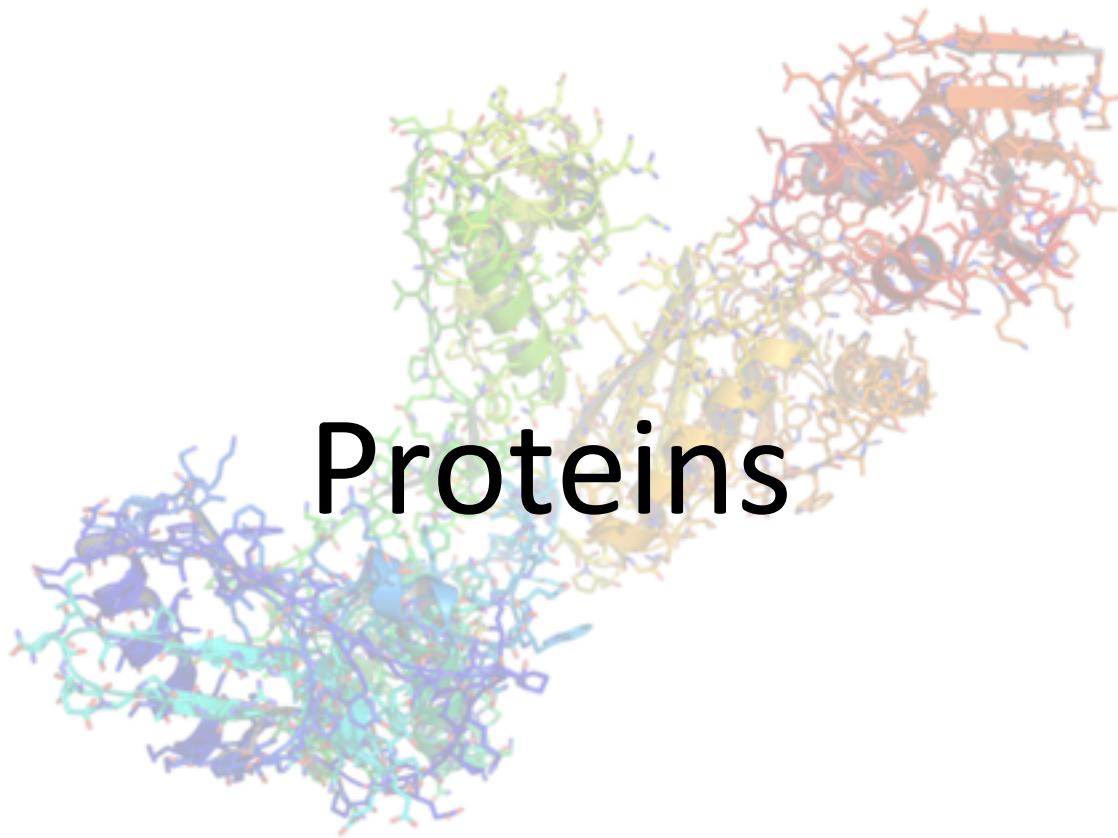
## Proteins

- Primary to quaternary structure
- Peptide Bond
- Amino acids
- Torsion angles
- Ramachandran plot
- Secondary structures:  $\alpha$ -helix,  $\beta$ -sheets
- Fold space
- SCOP, CATH, PDB

# Summary

## Nucleic Acids (DNA, RNA)

- Bases, sugar, phosphate
- Phosphodiester bond
- Numbering system
- Sugar Puckering
- Orientation around glycosidic angle
- C4'-C5' torsion angle
- RNA rotamers
- Base pairs
- Triples
- Helical parameters
- Helical grooves
- Secondary structures in RNA
- Forces that drive RNA folding
- NDB



Proteins

# Nomenclature

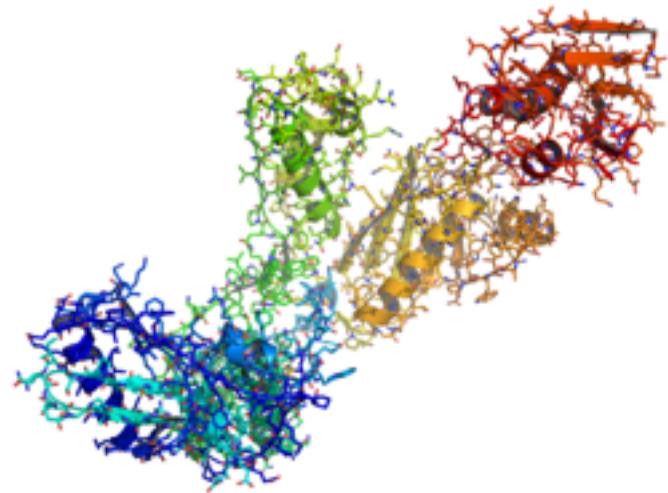
**Fold:** Three dimensional conformation of a protein sequence (usually at domain level).

**Domain:** Structurally globular part of a protein, which may independently fold.

**Secondary Structure:** Regular sub-domain structures composed by alpha-helices, beta-sheets and coils (or loops).

**Backbone:** Protein structure skeleton composed by the carbon, nitrogen and oxygen atoms.

**Side-Chain:** Specific atoms identifying each of the 20 residues types.

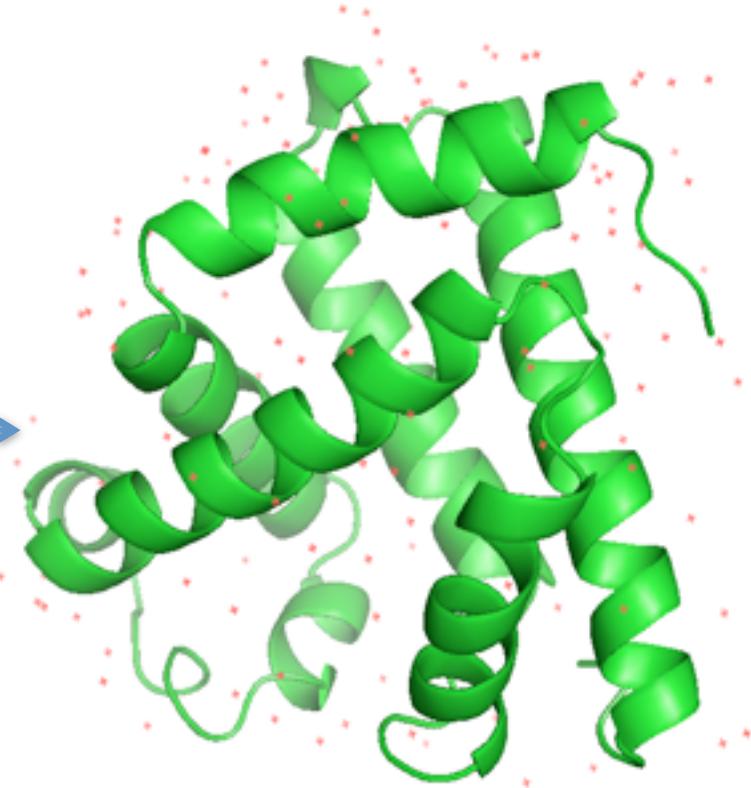


Primary structure

Secondary structure

Tertiary structure

MVLSEGEWQLVLHVWAKVEAD  
VAGHGQDILIRLFKSHPETLEKFD  
RFKHLKTEAEMKASEDLKKHGVT  
VLTALGAILKKKGHHEAEKPLAQ  
SHATKHKIPIKYLEFISEAIIHVLHS  
RHPGNFGADAQGAMNKALELFR  
KDIAAKYKELGYQG

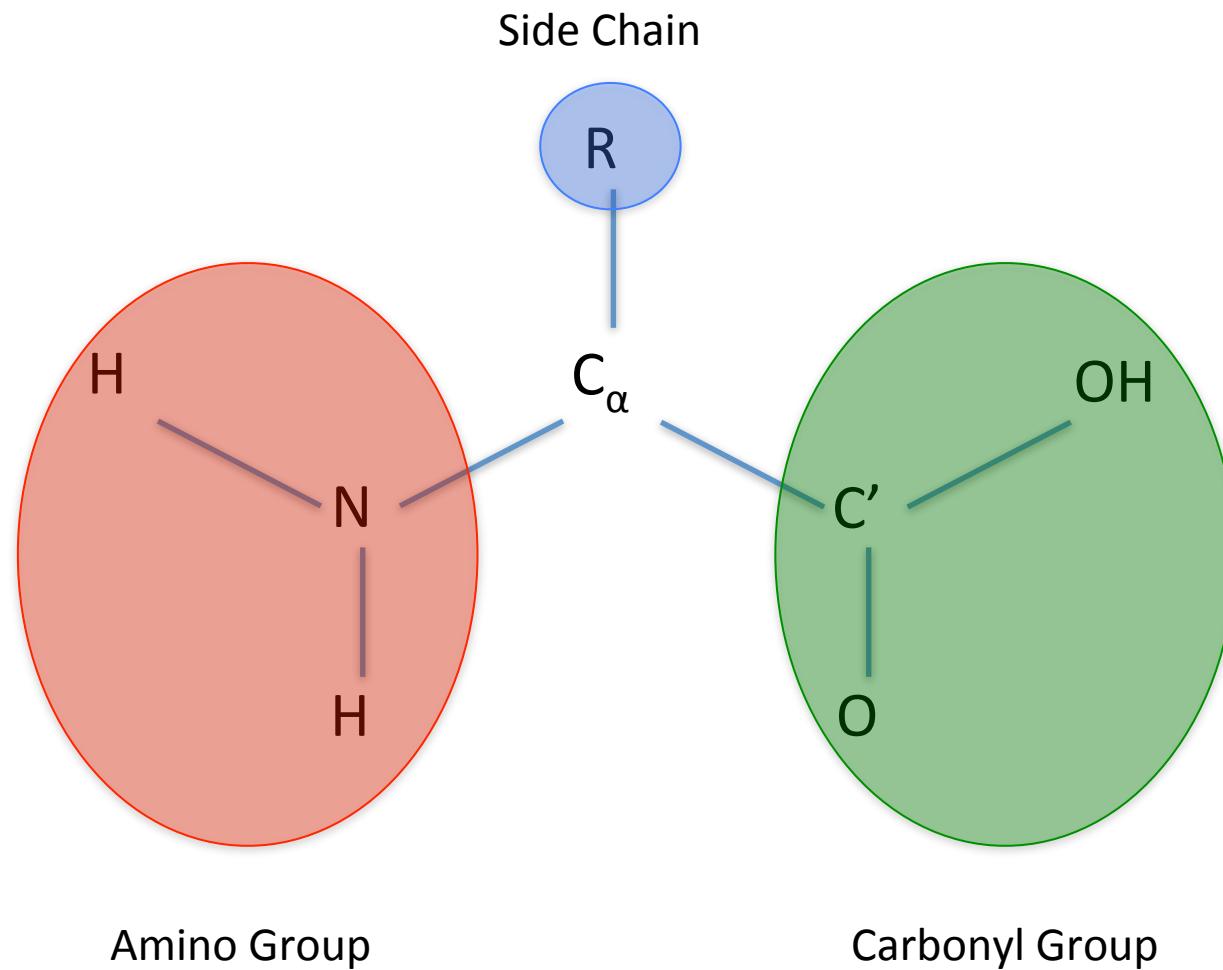


(Myoglobin)

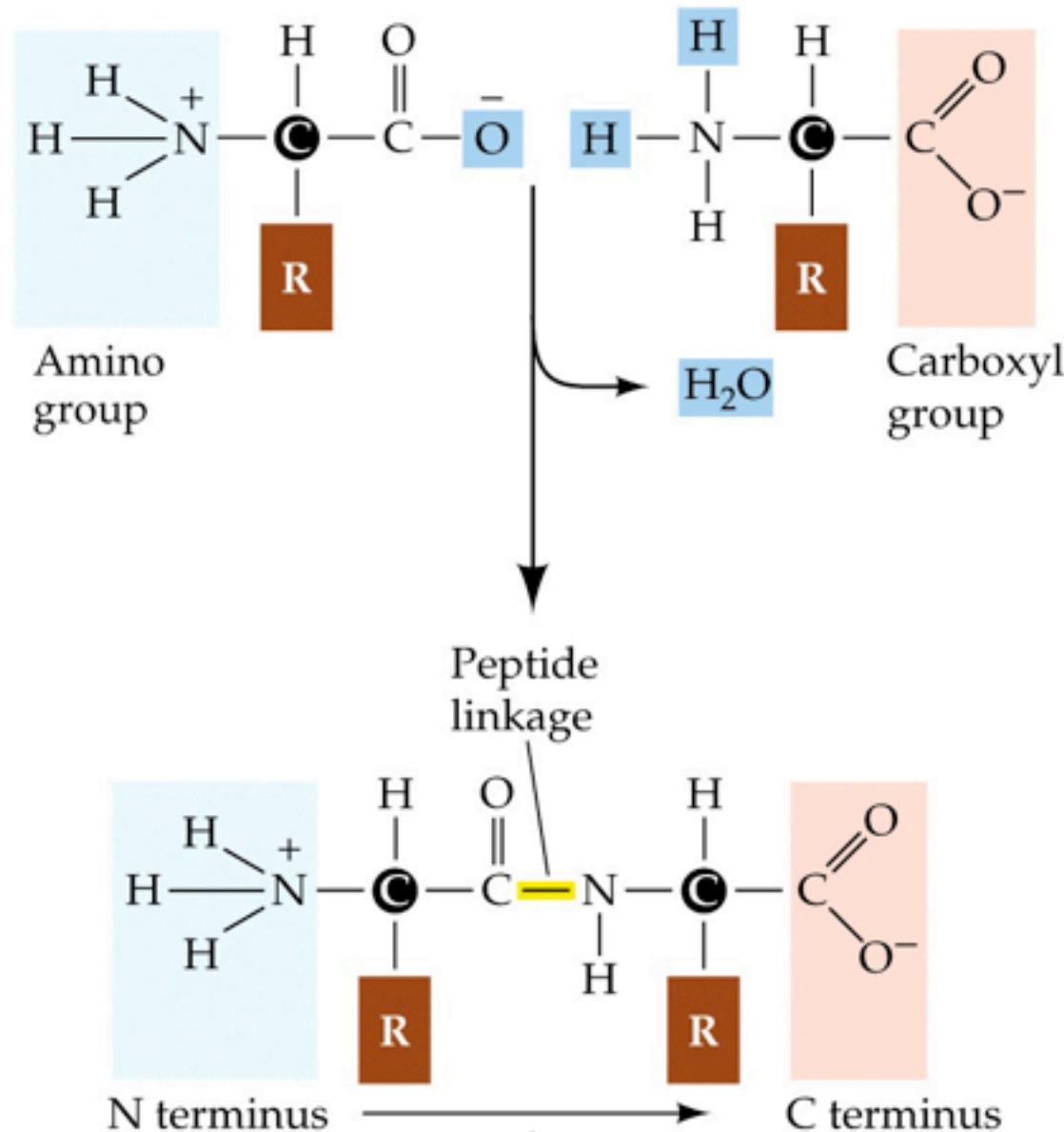
## Quaternary structure



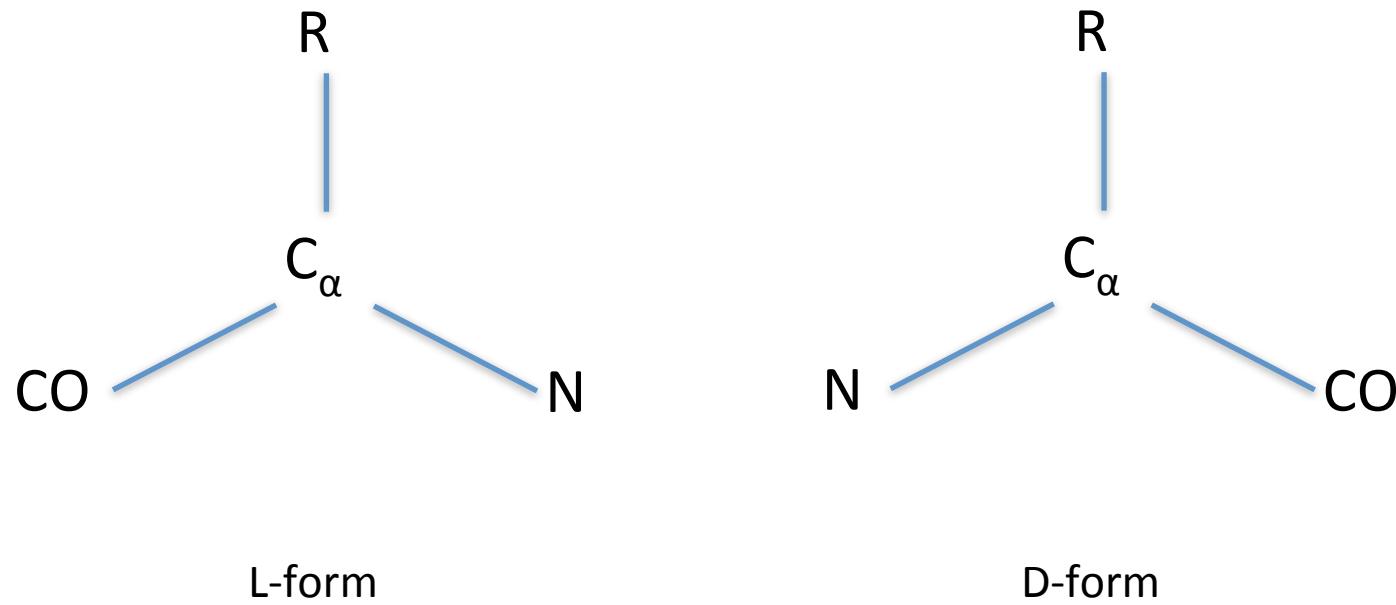
# Proteins are polypeptide chains



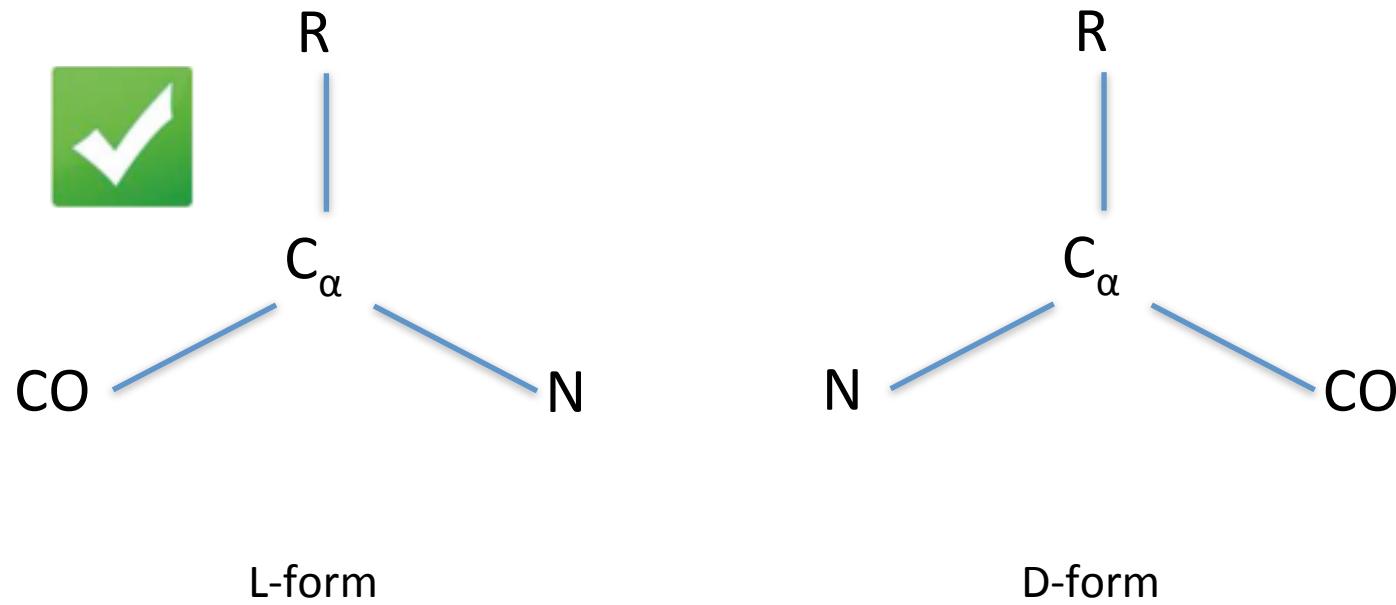
# Peptide Bond



# L- and D- forms

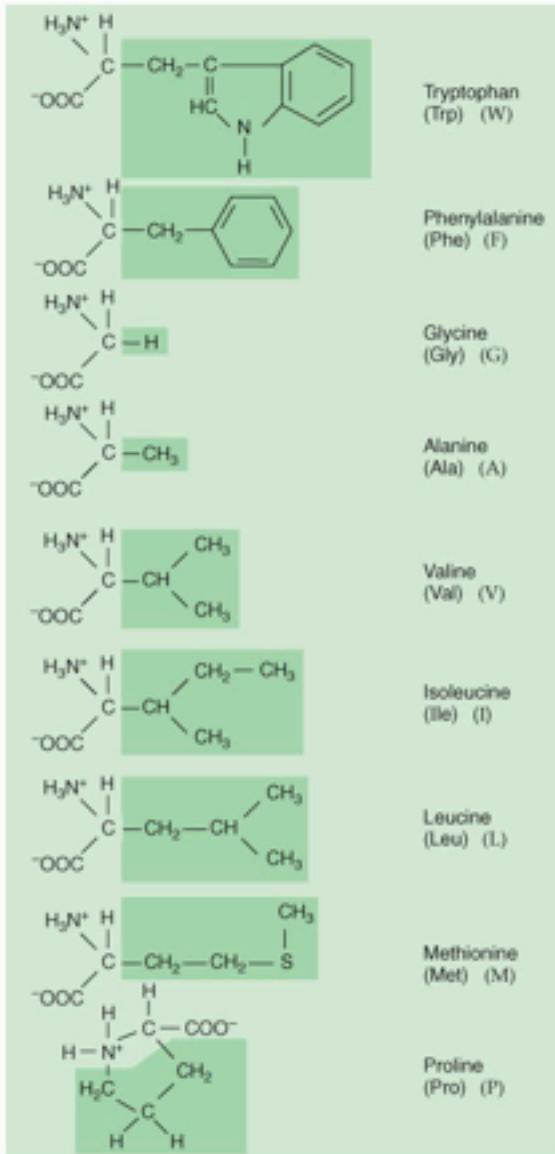


# L- and D- forms

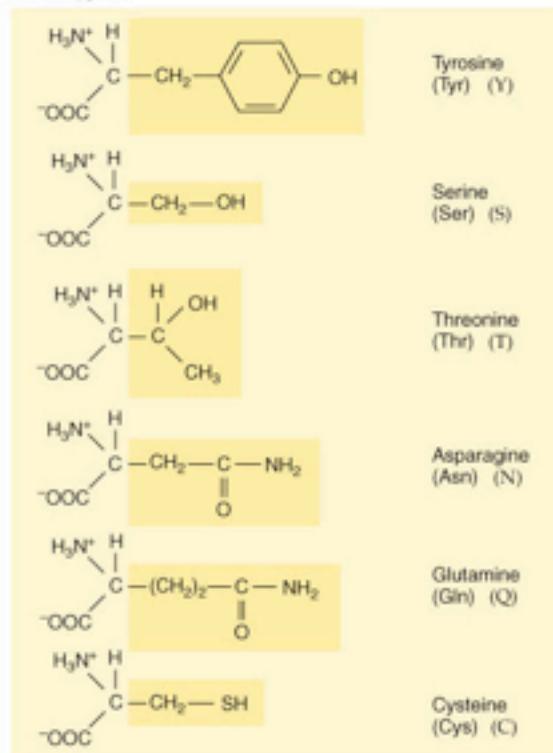


# The 20 amino acids

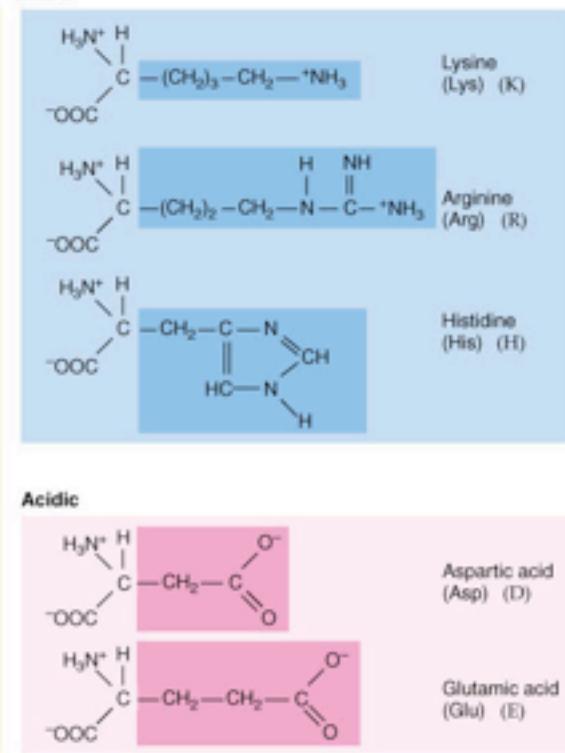
Neutral, nonpolar



Neutral, polar

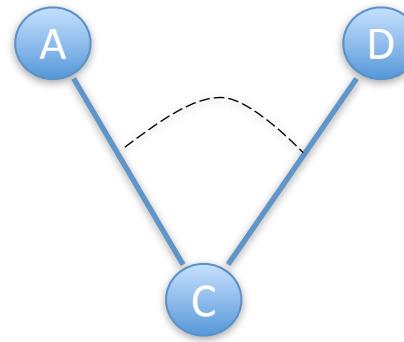
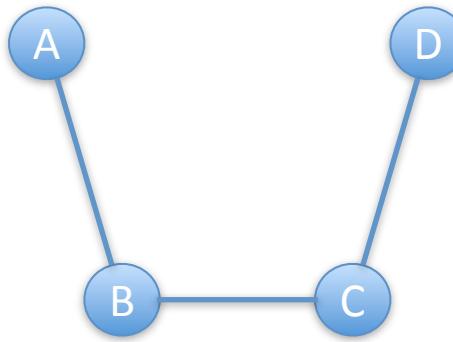


Basic

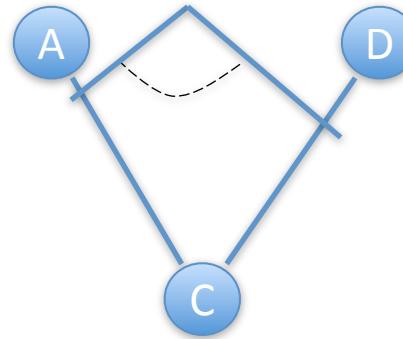
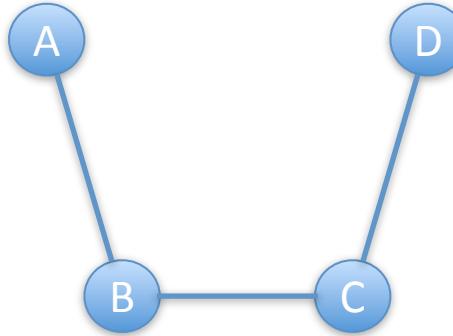


# Torsion and dihedral angles

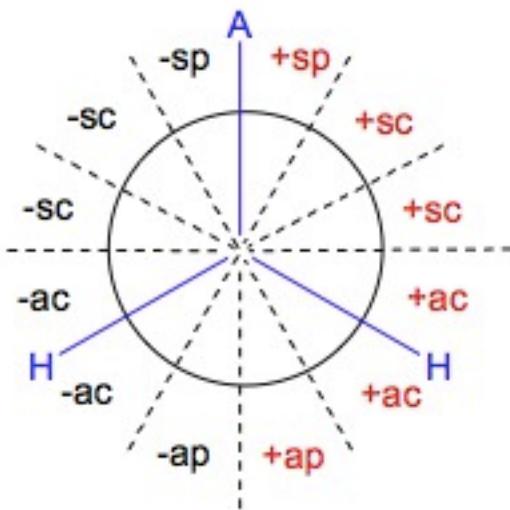
Torsion  
angle



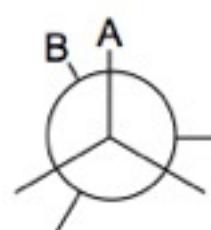
Dihedral  
angle



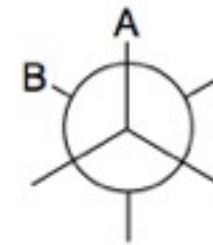
# Torsion angle classification



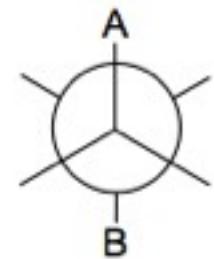
*s: syn*  
*p: periplanar*  
*a: anti*  
*c: clinal*



Cis



Gauche



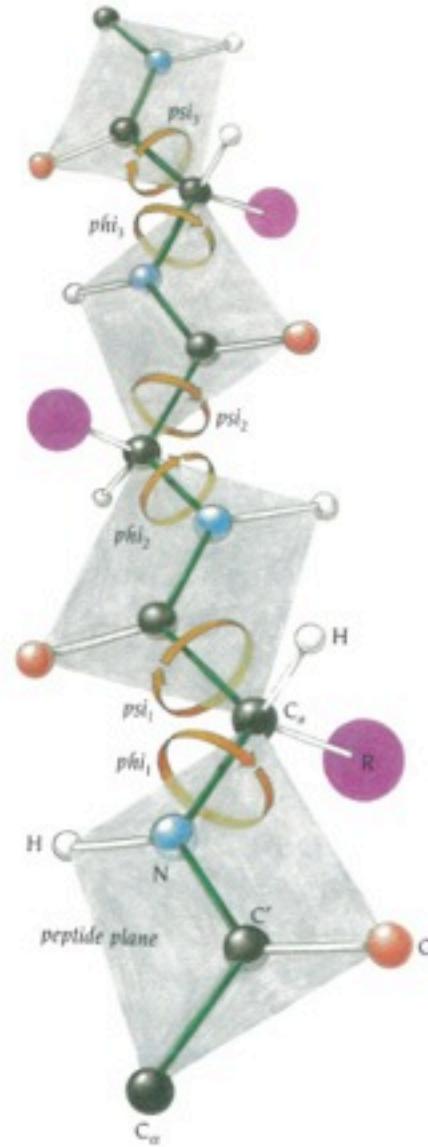
Trans

## Klyne-Prelog system

# Peptide units = suits

Phi ( $\phi$ ) =  $C'-C_\alpha-N-C'^{+1}$ ,  
rotation around  $C_\alpha-N$

Psi ( $\psi$ ) =  $N-C'-C_\alpha-N^{+1}$ ,  
rotation around  $C'-C_\alpha$



# Ramachandran plot

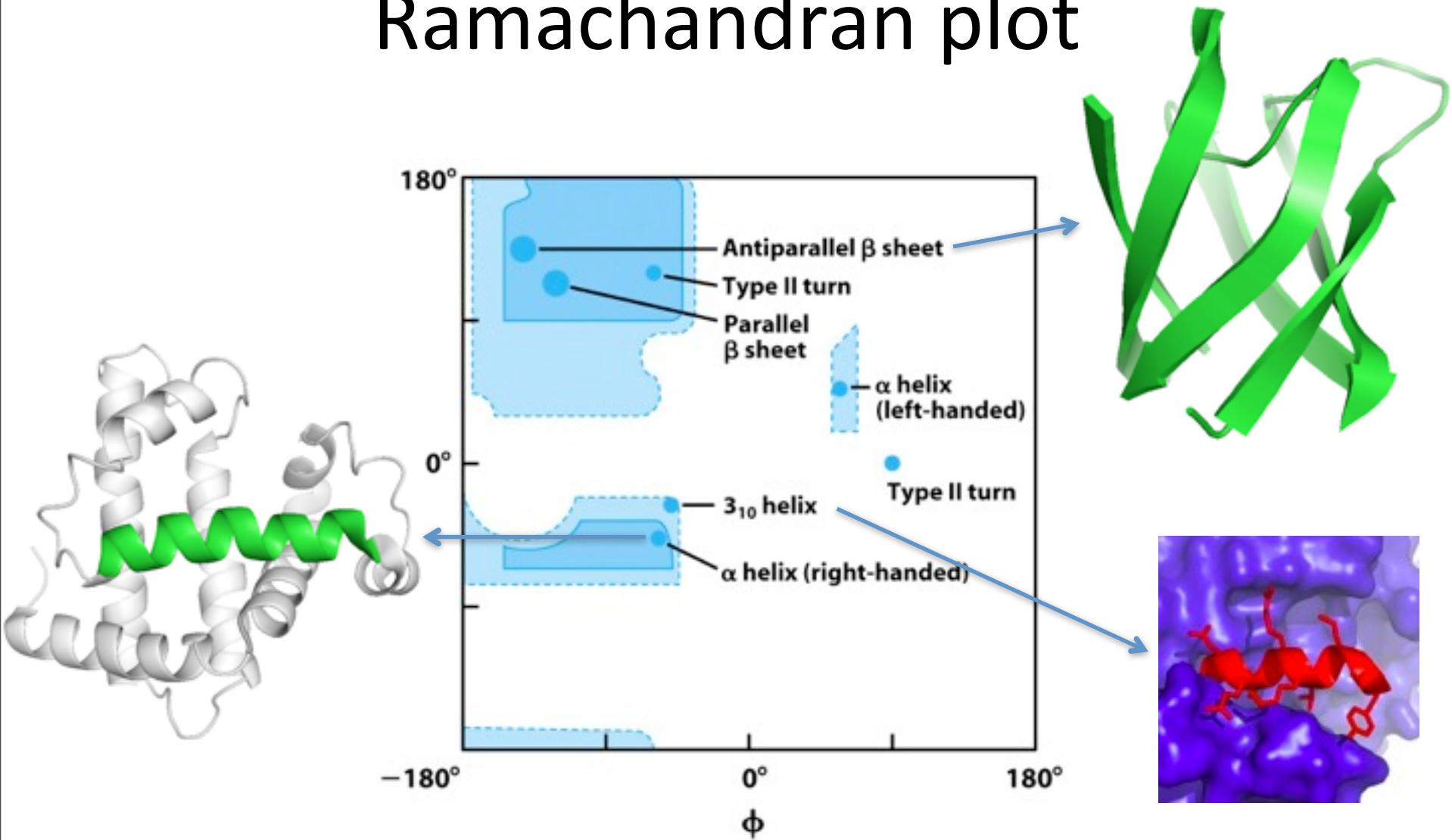
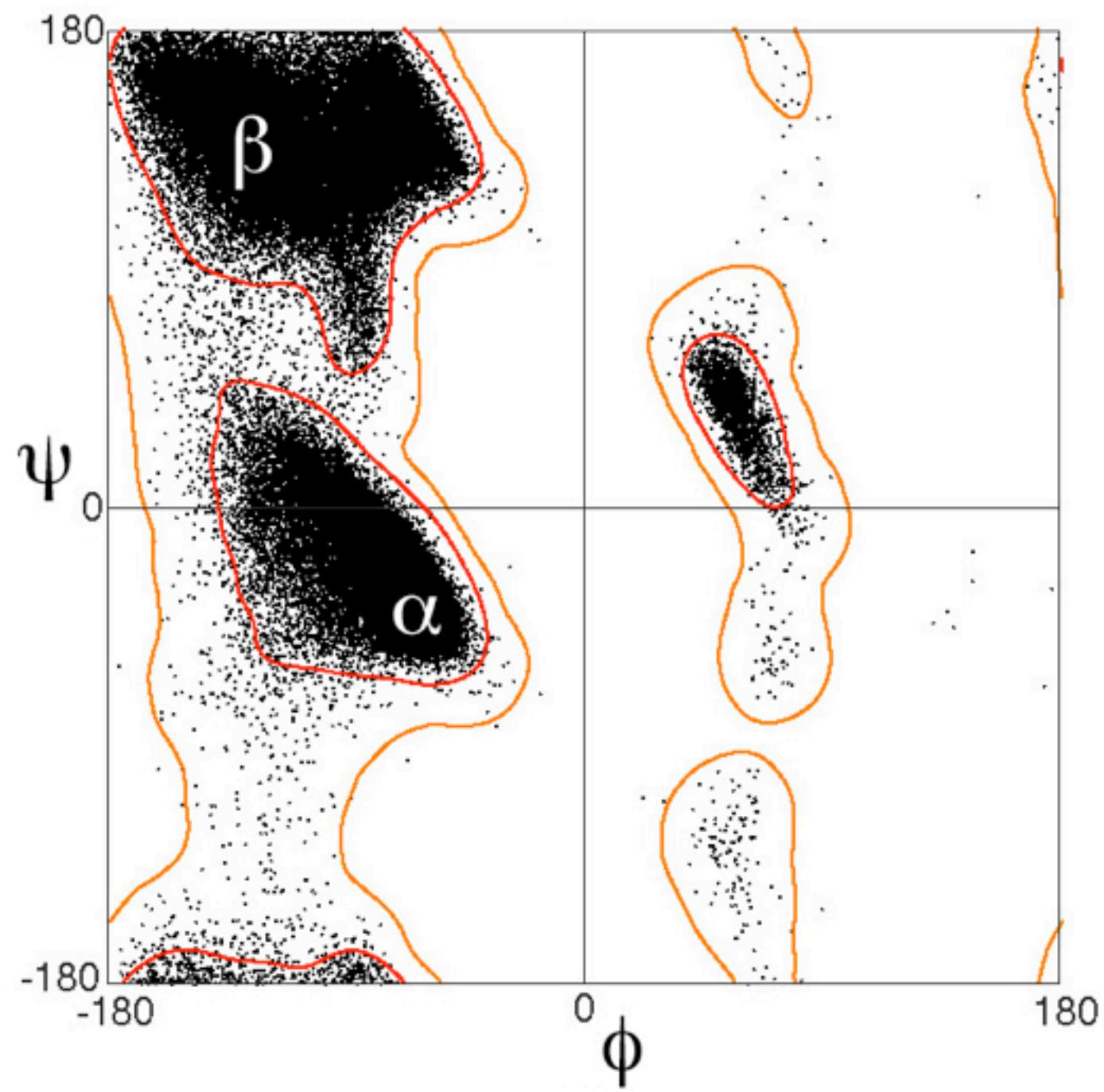
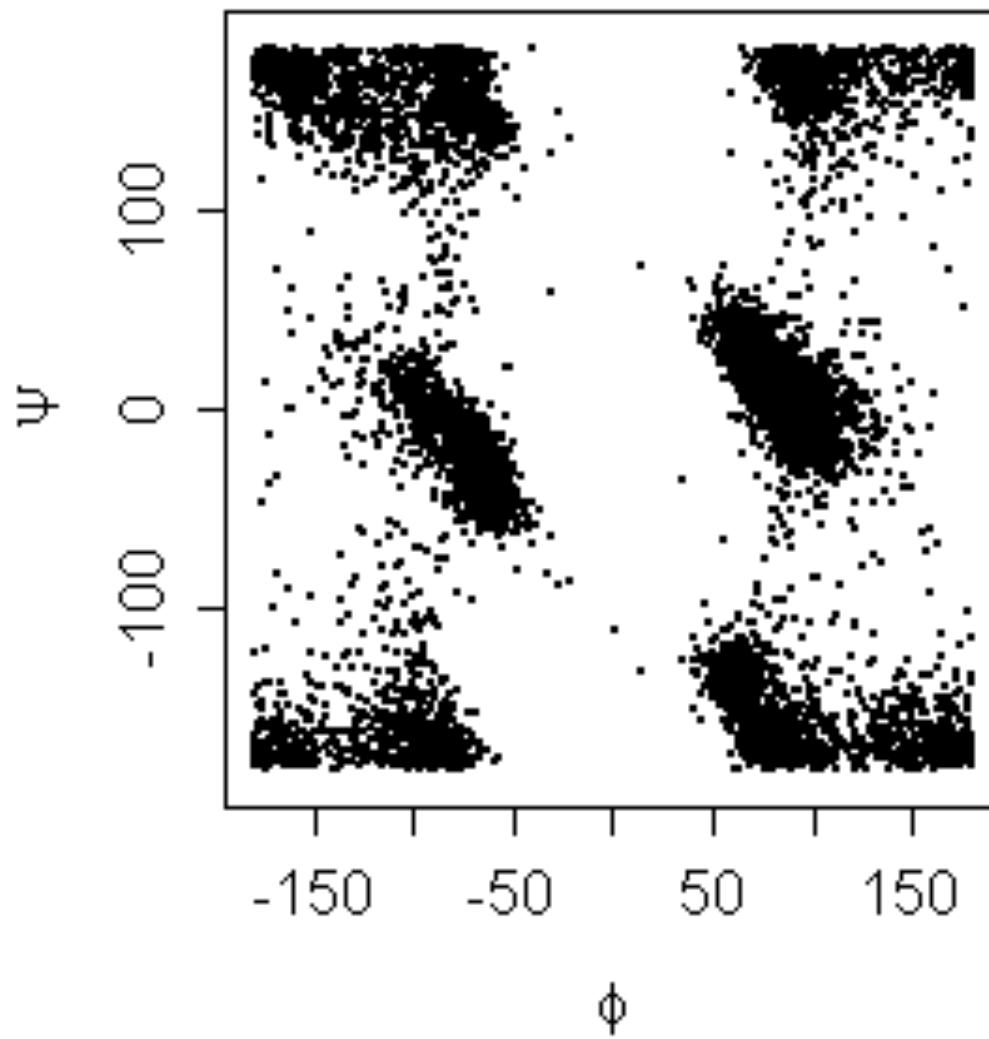


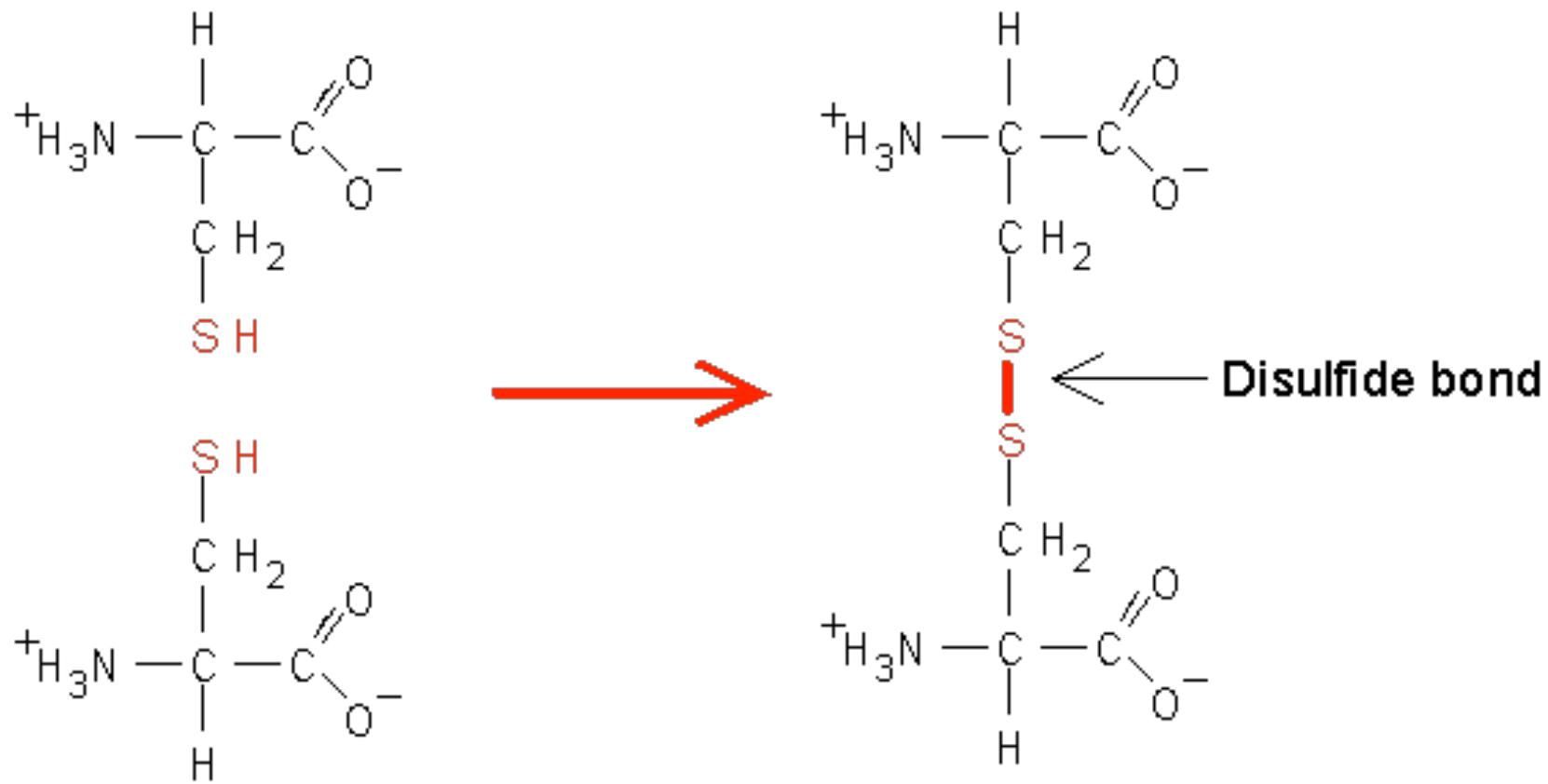
Figure 4-9a Principles of Biochemistry, 4/e  
© 2006 Pearson Prentice Hall, Inc.



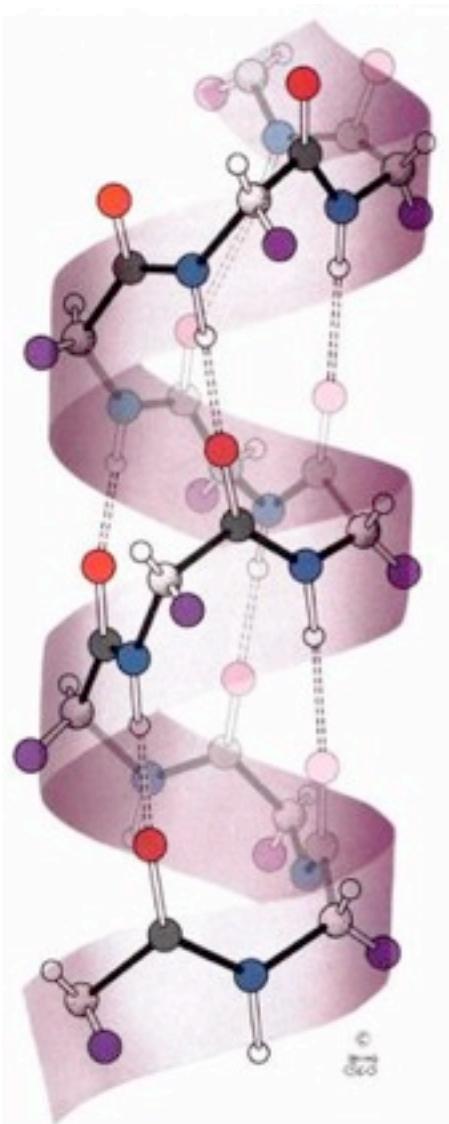
# Glycine



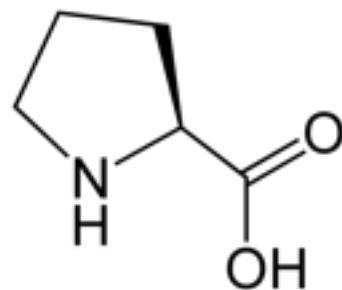
# Cysteine disulfide bond



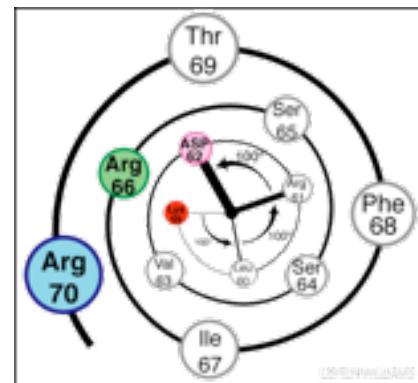
# $\alpha$ -helix



- Backbone angles:
  - $\psi = -50^\circ$
  - $\phi = -60^\circ$
- Dipole moment
- Proline bending

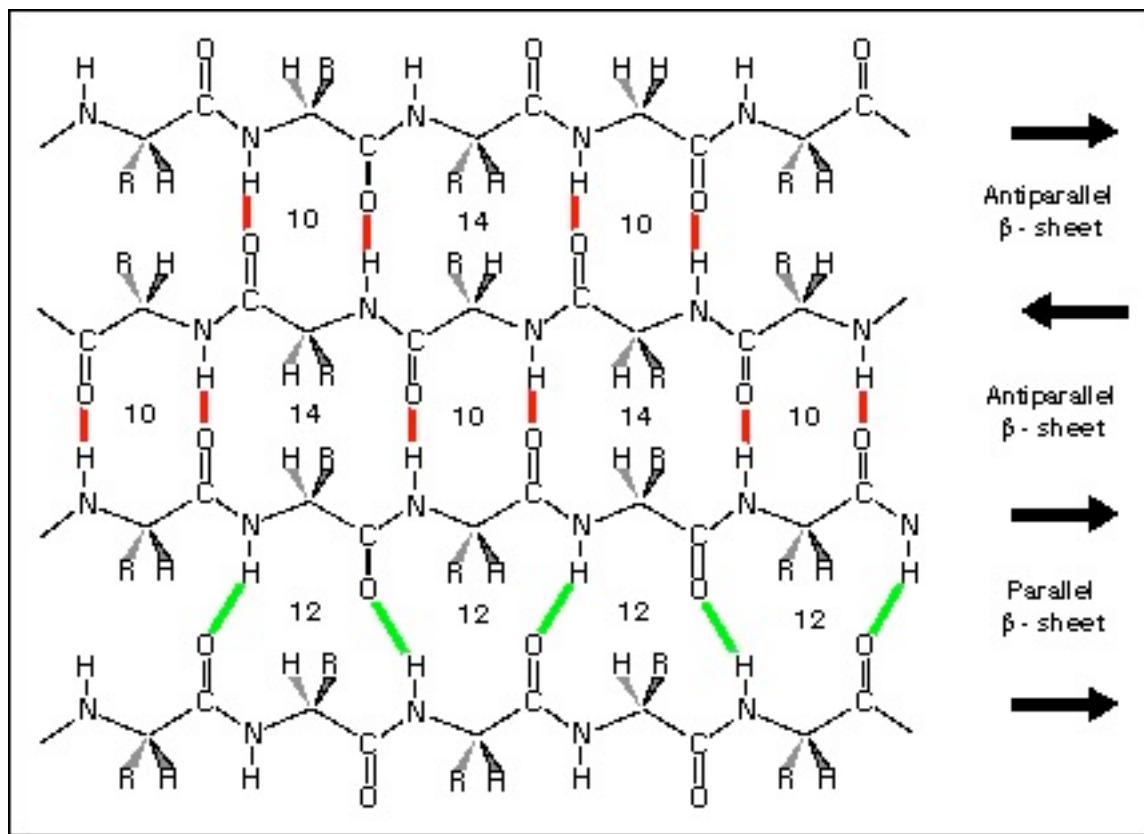


Proline

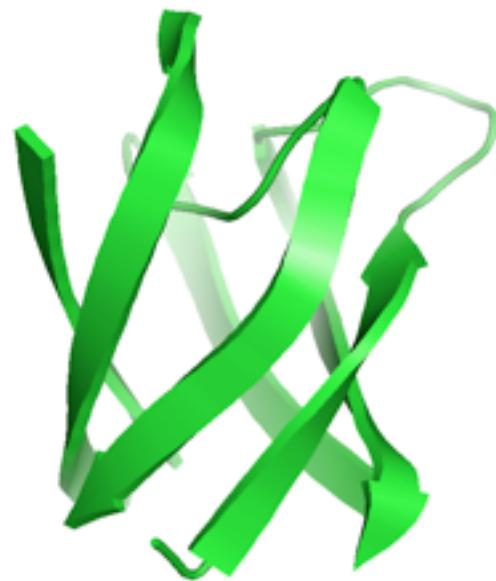


Helical wheel  
projection

# $\beta$ -sheet

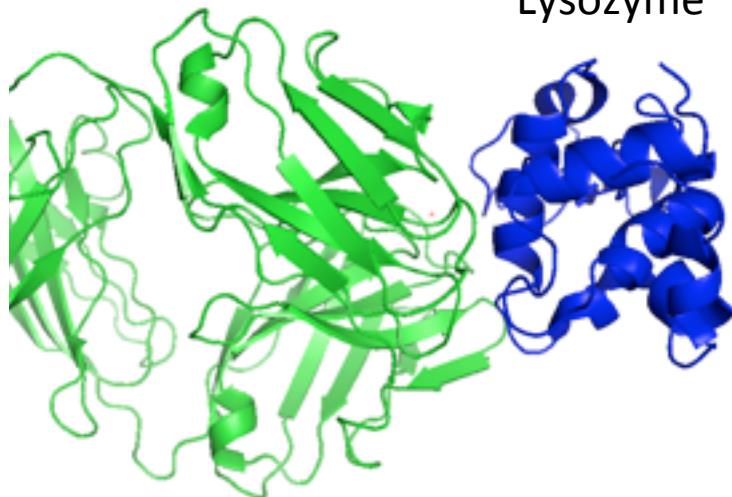


- First quadrant in Ramachandran's plot
- Parallel or antiparallel

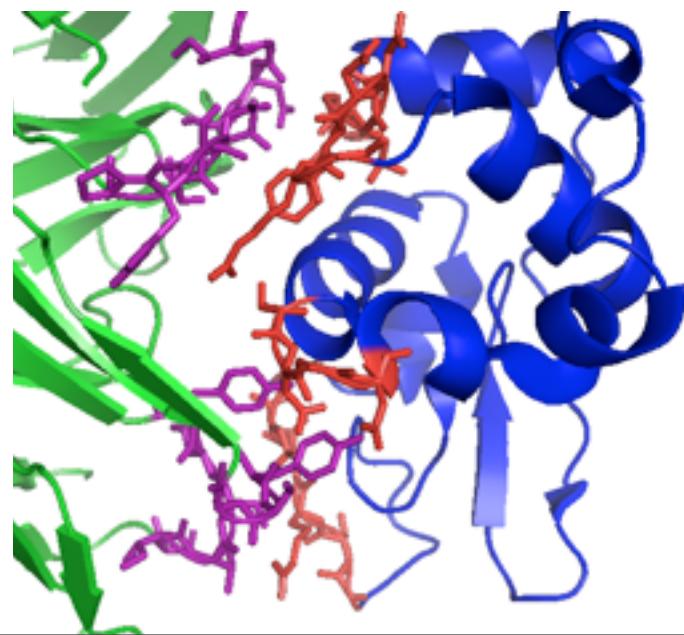
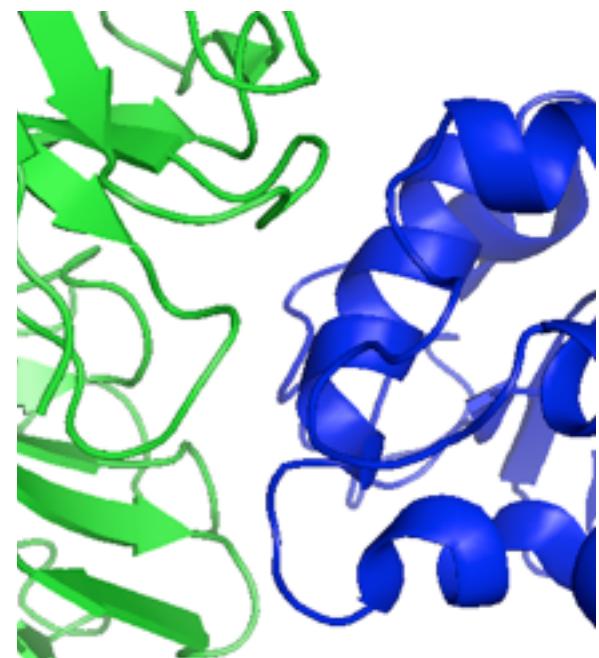
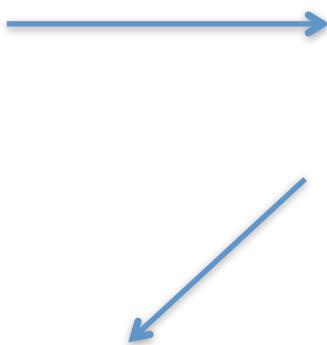


# Loops and function

Anti-lysozyme HyHEL-10 Fab

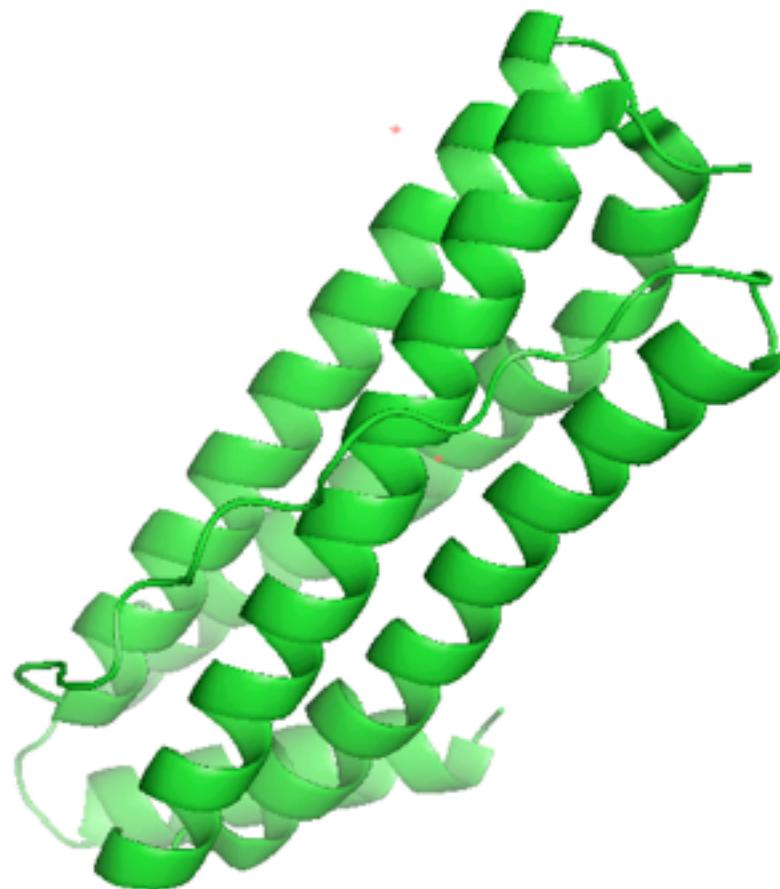


Lysozyme



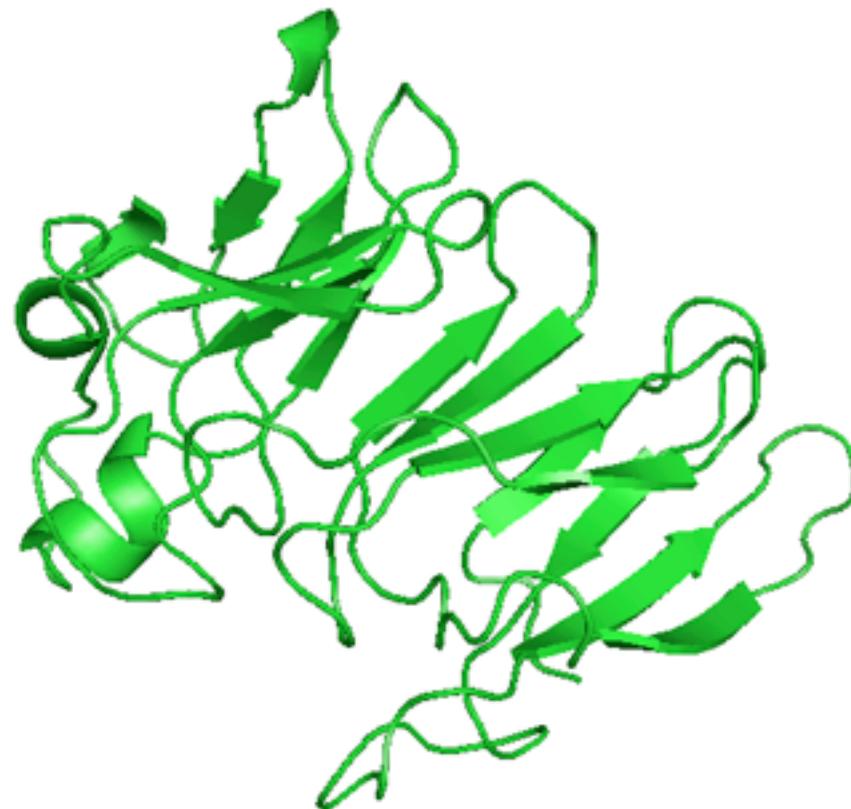
# Tertiary structures - Fold space

Mainly alpha:



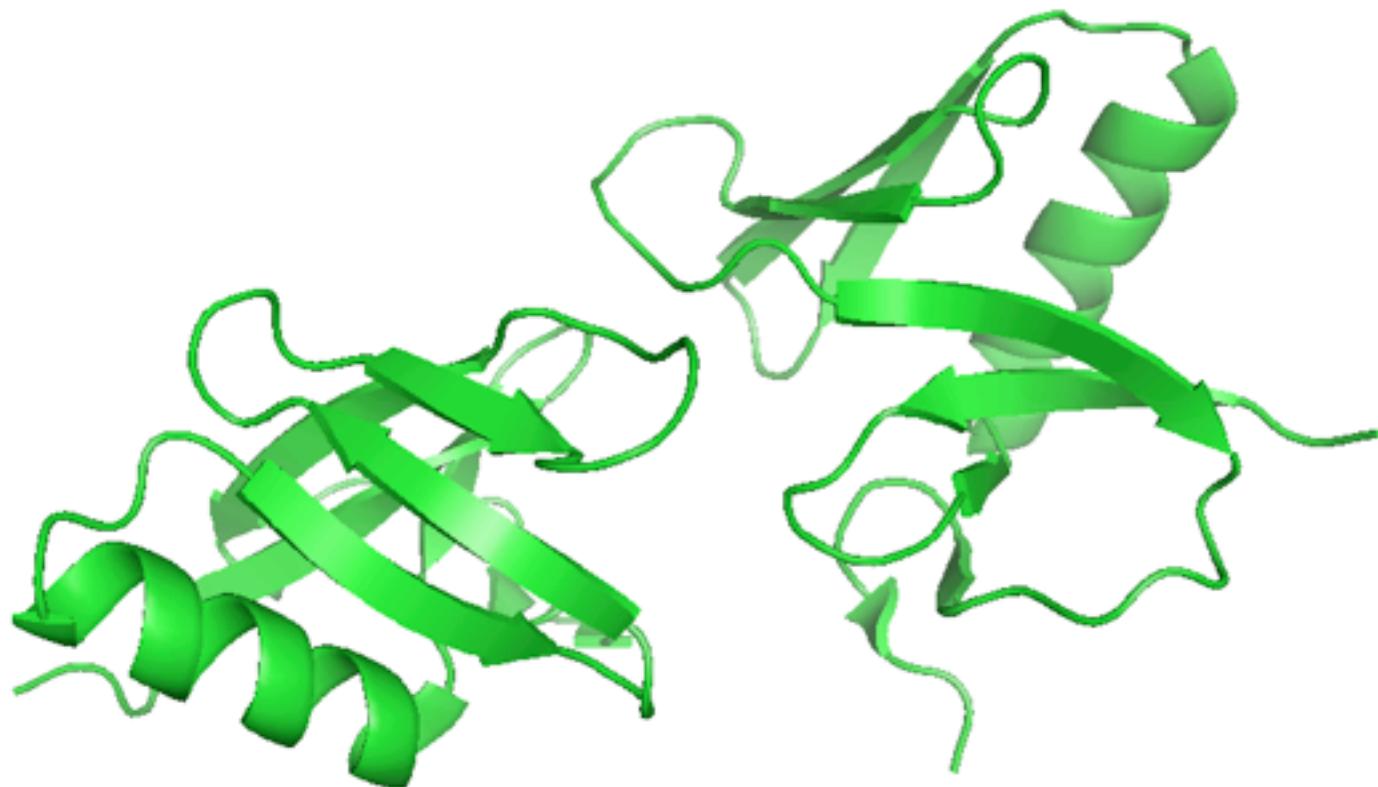
# Tertiary structures - Fold space

Mainly beta:



# Tertiary structures - Fold space

Alpha beta:



# SCOP<sub>1.75</sub> database

<http://scop.mrc-lmb.cam.ac.uk/scop/>

Welcome to SCOP: Structural Classification of Proteins.  
1.75 release (June 2009)

38221 PDB Entries, 1 Literature Reference, 110800 Domains. (excluding nucleic acids and theoretical models).  
Folds, superfamilies, and families [statistics here](#).  
[New folds superfamilies families](#).  
[List of obsolete entries and their replacements](#).

Authors: Alexey G. Murzin, John-Marc Chandonia, Antonina Andreeva, Dave Howorth, Lorendana Lo Conte, Bartley G. Ailey, Steven E. Brenner, Tim J. P. Hubbard, and Cyrus Chothia. [scop@mrc-lmb.cam.ac.uk](mailto:scop@mrc-lmb.cam.ac.uk)  
Reference: Murzin A. G., Brenner S. E., Hubbard T., Chothia C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247, 536-540. [PDF]  
Recent changes are described in: Lo Conte L., Brenner S. E., Hubbard T.J.P., Chothia C., Murzin A. (2002). SCOP database in 2002: refinements accommodate structural genomics. *Nucl. Acid Res.* 30(1), 264-267. [PDF].  
Andreeva A., Howorth D., Brenner S.E., Hubbard T.J.P., Chothia C., Murzin A.G. (2004). SCOP database in 2004: refinements integrate structure and sequence family data. *Nucl. Acid Res.* 32:D226-D229. [PDF], and  
Andreeva A., Howorth D., Chandonia J.-M., Brenner S.E., Hubbard T.J.P., Chothia C., Murzin A.G. (2007). Data growth and its impact on the SCOP database: new developments. *Nucl. Acids Res.* 2008; 36: D419-D425; doi:10.1093/nar/gkm993 [PDF].

**Postdoc Wanted**

- Want to help us design and build the next generation of SCOP and ASTRAL?  
[Get more details and apply here](#).

**Access methods**

- Enter SCOP at the [top of the hierarchy](#)
- [Keyword search of SCOP entries](#)
- [SCOP parseable files](#)
- [All SCOP releases and reclassified entry history](#)
- [pre-SCOP - preview of the next release](#)
- [SCOP domain sequences and pdb-style coordinate files \(ASTRAL\)](#)
- [Hidden Markov Model library for SCOP superfamilies \(SUPERFAMILY\)](#)
- [Structural alignments for proteins with non-trivial relationships \(SISYPHUS\)](#)
- [Online resources](#) of potential interest to SCOP users

SCOP [mirrors](#) around the world may speed your access.

Murzin A. G., et al. (1995). *J. Mol. Biol.* 247, 536-540.

- ✓ Largely recognized as “standard of gold”
- ✓ Manually classification
- ✓ Clear classification of structures in:  
**CLASS**  
**FOLD**  
**SUPER-FAMILY**  
**FAMILY**
- ✓ Some large number of tools already available

**Manually classification**  
**Not 100% up-to-date**  
**Domain boundaries definition**

Class	Number of folds	Number of superfamilies	Number of families
All alpha proteins	284	507	871
All beta proteins	174	354	742
Alpha and beta proteins (a/b)	147	244	803
Alpha and beta proteins (a+b)	376	552	1055
Multi-domain proteins	66	66	89
Membrane and cell surface proteins	58	110	123
Small proteins	90	129	219
Total	1195	1962	3902

# CATH<sub>3.5</sub> database

<http://www.cathdb.info>

The screenshot shows the homepage of the CATH / Gene3D website. At the top, there's a browser header with tabs like 'Most Visited', 'Getting Started', 'Latest Headlines', 'Home - PubMed...', and 'Bookmarks'. Below the header, the CATH logo is on the left, and a search bar says 'Search CATH by keyw...'. On the right, there's a menu icon. The main content area has a large 'CATH / Gene3D' title, a subtext '16 million protein domains classified into 2,626 superfamilies', and four buttons: 'Get Started', 'Search', 'Download', and 'Take the Tour'. Below this, there's a section titled 'What's New?' with a paragraph about recent changes and a 'get in touch' link.

## What's New?

The CATH website has recently undergone a big overhaul. We really hope you find the new pages more useful, easier to use and quicker to load. Please [get in touch](#) and let us know what you think.

## Searching CATH

- Search by ID / keyword
- Search by FASTA sequence
- Search by PDB structure

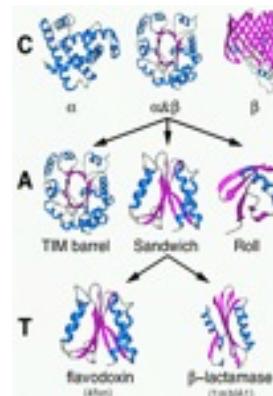
## Example pages

- PDB "1dan"
- Domain "1cukA01"
- Relatives of "1cukA01"
- Superfamily "HUPs"
- Functional Family

Uses FSSP for superimposition

- ✓ Recognized as “standard of gold”
- ✓ Semi-automatic classification
- ✓ Clear classification of structures in:  
CLASS  
ARCHITECTURE  
TOPOLOGY  
HOMOLOGOUS SUPERFAMILIES
- ✓ Some large number of tools already available
- ✓ Easy to navigate

## Semi-automatic classification Domain boundaries definition



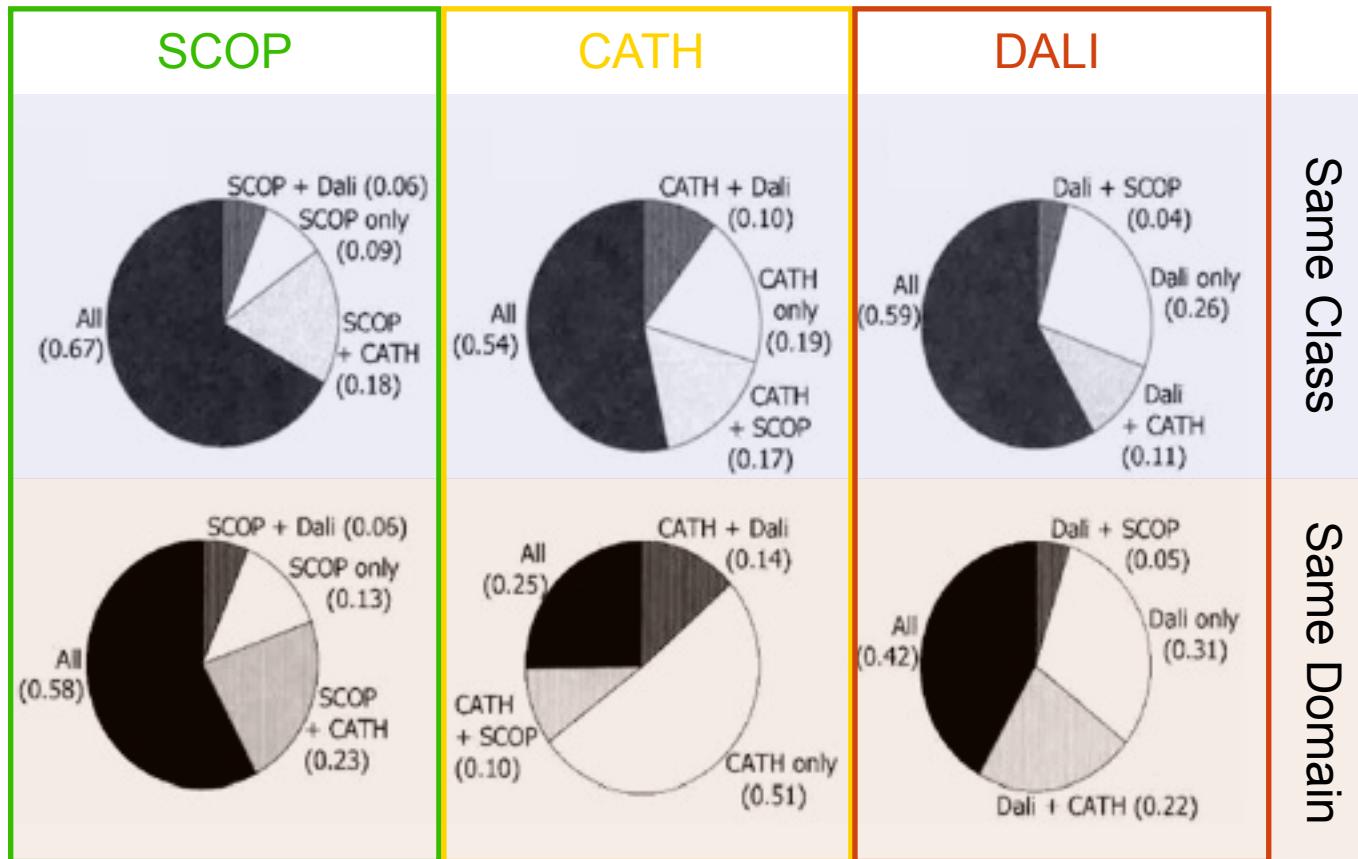
173,536 CATH Domains  
2,626 CATH Superfamilies  
51,334 PDBs

Orengo, C.A., et al. (1997) *Structure*. 5. 1093-1108.

# Classification of the structural space

## *Not an easy task!*

Domain definition AND domain classification



Day, et al. (2003) Protein Sciences, 12 pp2150

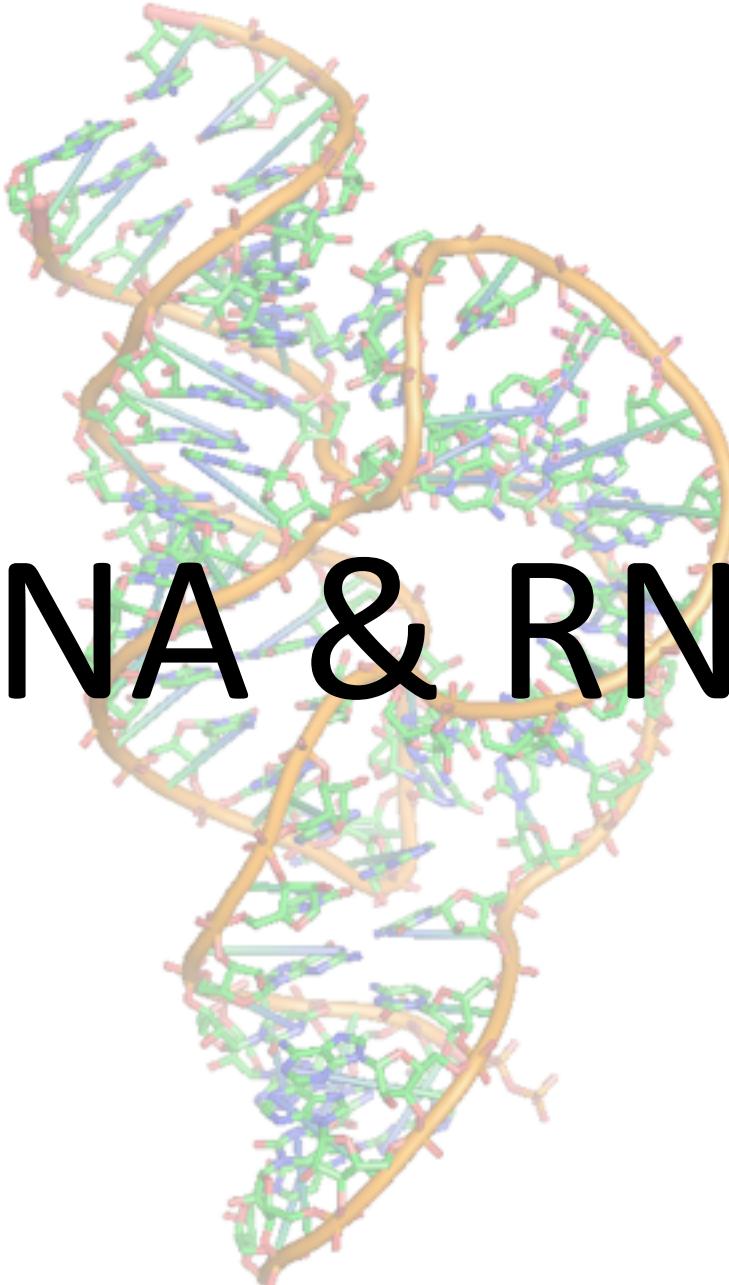
# PDB

The screenshot shows the homepage of the RCSB Protein Data Bank (PDB). The top navigation bar includes links for Science, Journals, News, Programming, Help, Grants, Booking, DOI, Shopping, and RCSB. The main content area features a "Biological Macromolecular Resource" section with a "Molecule of the Month" (Proton-Gated Urine Channel), a "Designer Proteins" section, and a "Protein Structure Initiative Featured System". Below this are sections for "Explore Archives" (Proteins, Nucleic Acids, Organism, PDB Method, PDB Release, PDB Classification, Organism Classification), "Labelled Structures" (e.g., 4S4K), and "Advanced Search". The left sidebar contains links for PDB-101, RCSB, News & Information, Reference Policies, Deposit Policies, Deposit FAQs, Contact Us, General, External Links, and New Website Features. The bottom sidebar includes links for Tools, Help, and a Launch Help System.

Yearly and total PDB structures per year

Yearly

Total



# DNA & RNA

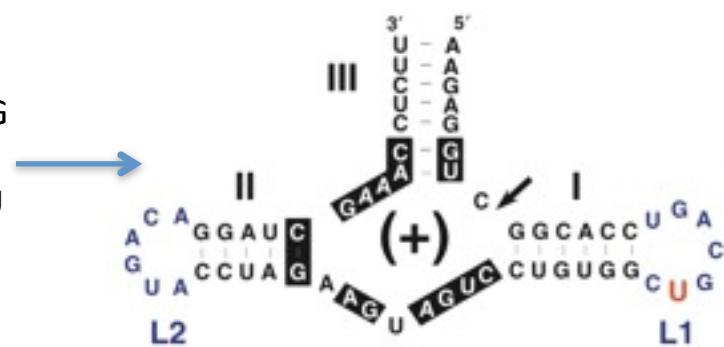
# RNA function

- 1) Informative
- 2) Structural
- 3) Catalytic
- 4) Regulatory

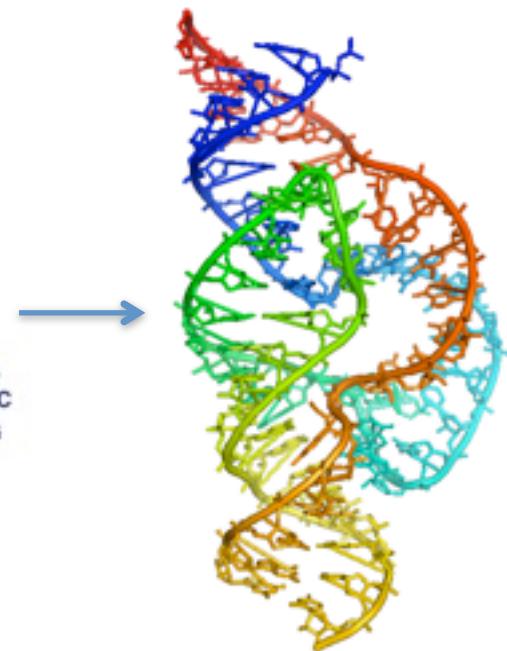
## Primary structure

CAGACCGAAGUGAUGAAGCG  
AUUGGUUAUCUGGGCAAA  
GCGUCUGAAGGUUGUGGUU  
UCGAC

## Secondary structure



## Tertiary structure (3D)

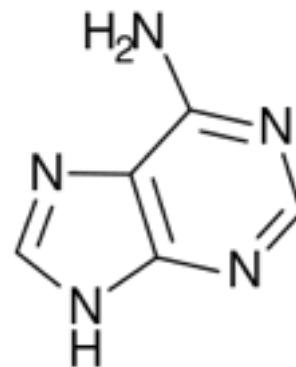


# Bases

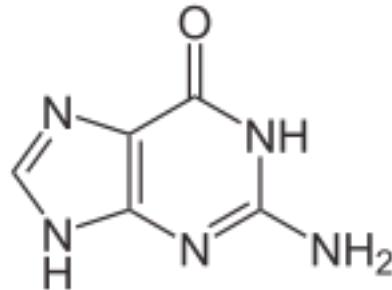
(aka nucleotidic bases, aromatic bases)

## Purines

Adenine  
(A)

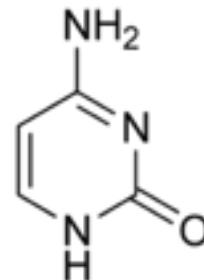


Guanine  
(G)

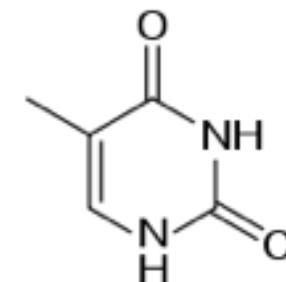
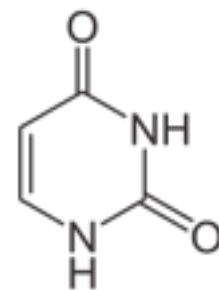


## Pyrimidines

Cytosine  
(C)



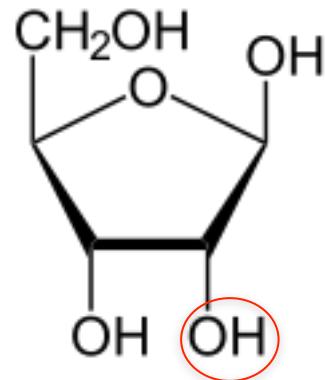
Uracil  
(U)



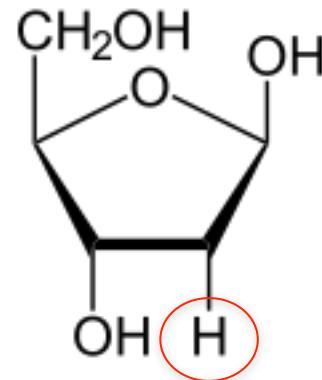
Thymidine  
(T)

# Sugars

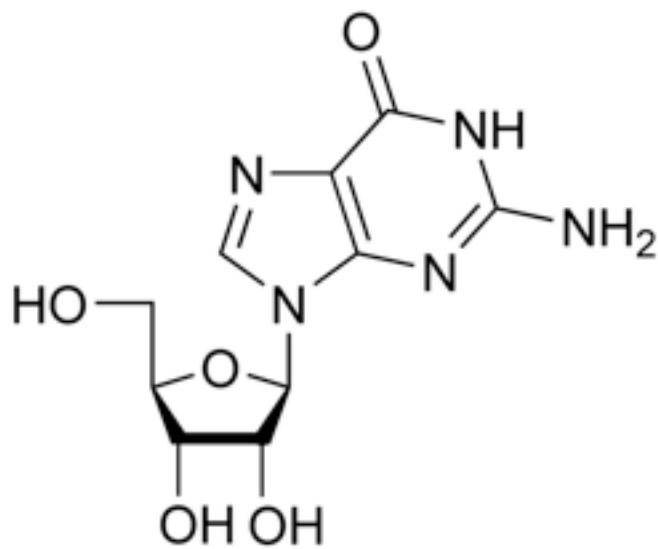
$\beta$ -D-ribose



$\beta$ -D-2-Deoxyribose

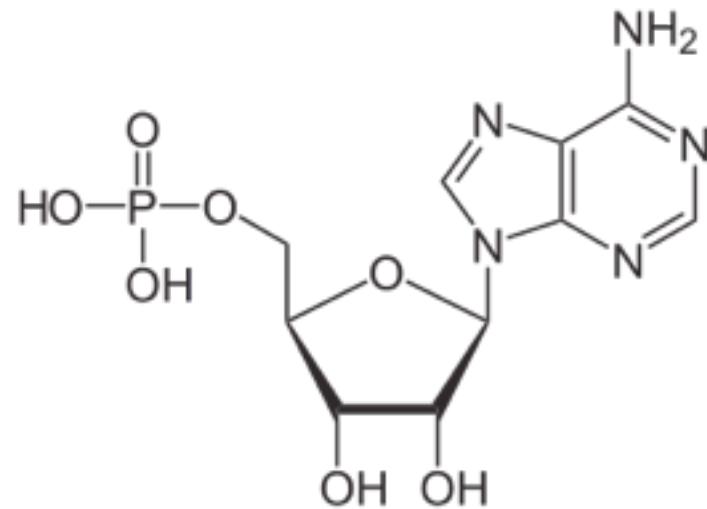


Nucleoside



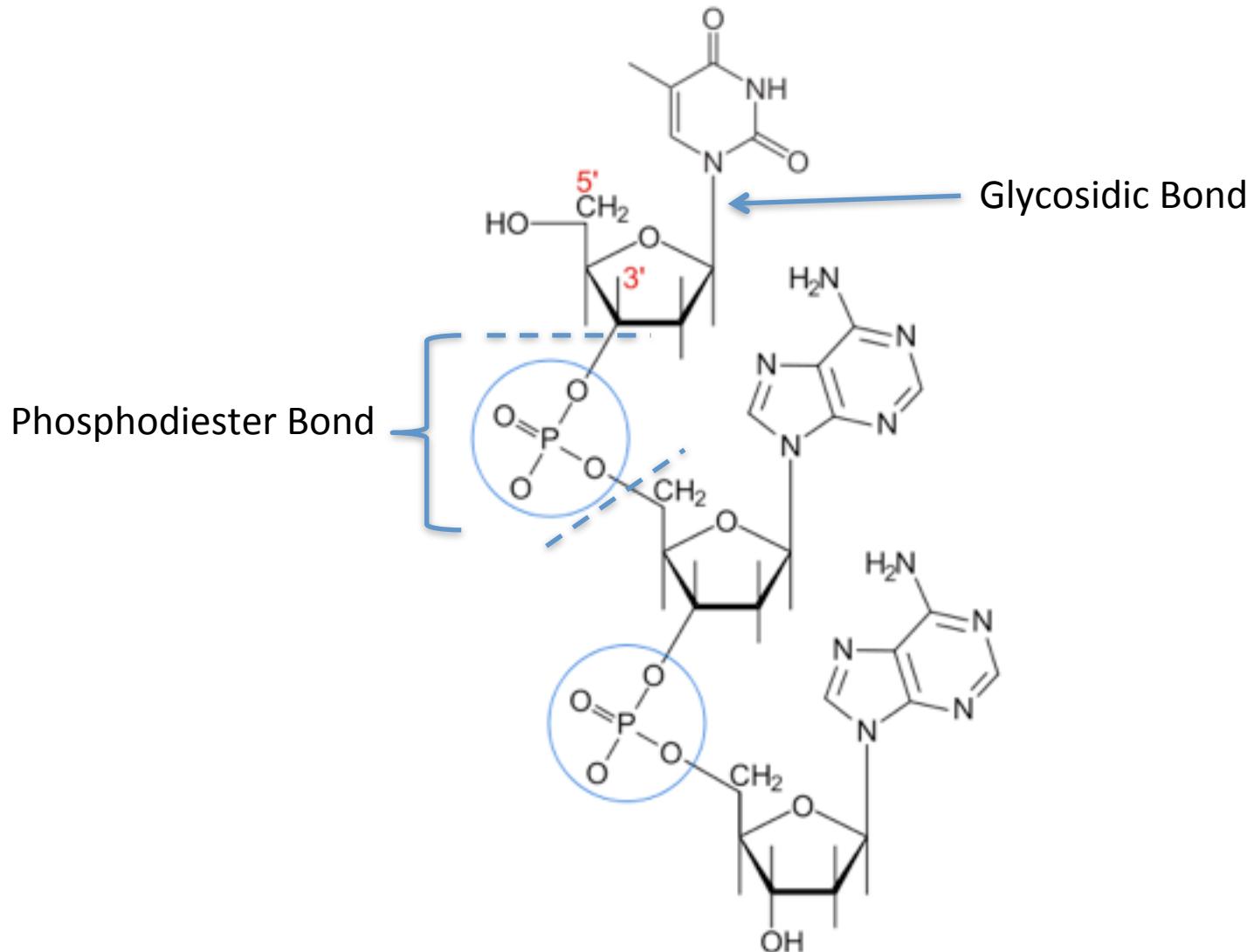
Sugar + base

Nucleotide monophosphate



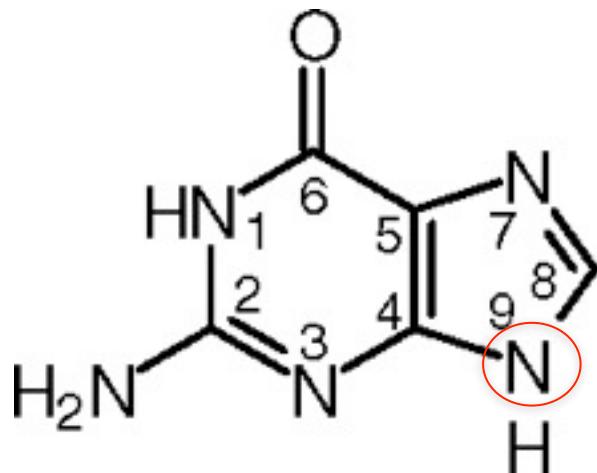
Phosphate + sugar + base

# Phosphodiester Bond

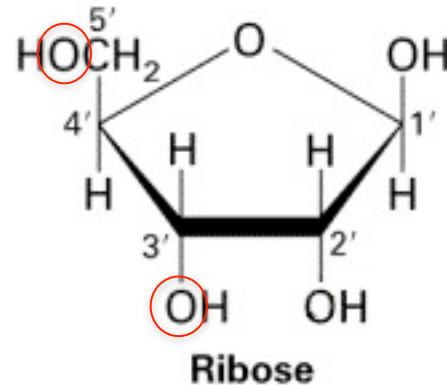
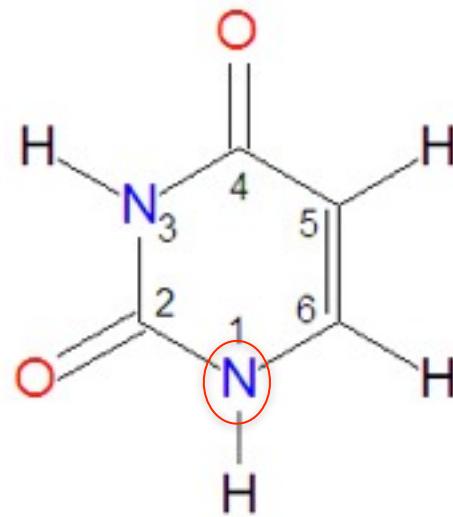


# Numbering system

Guanine

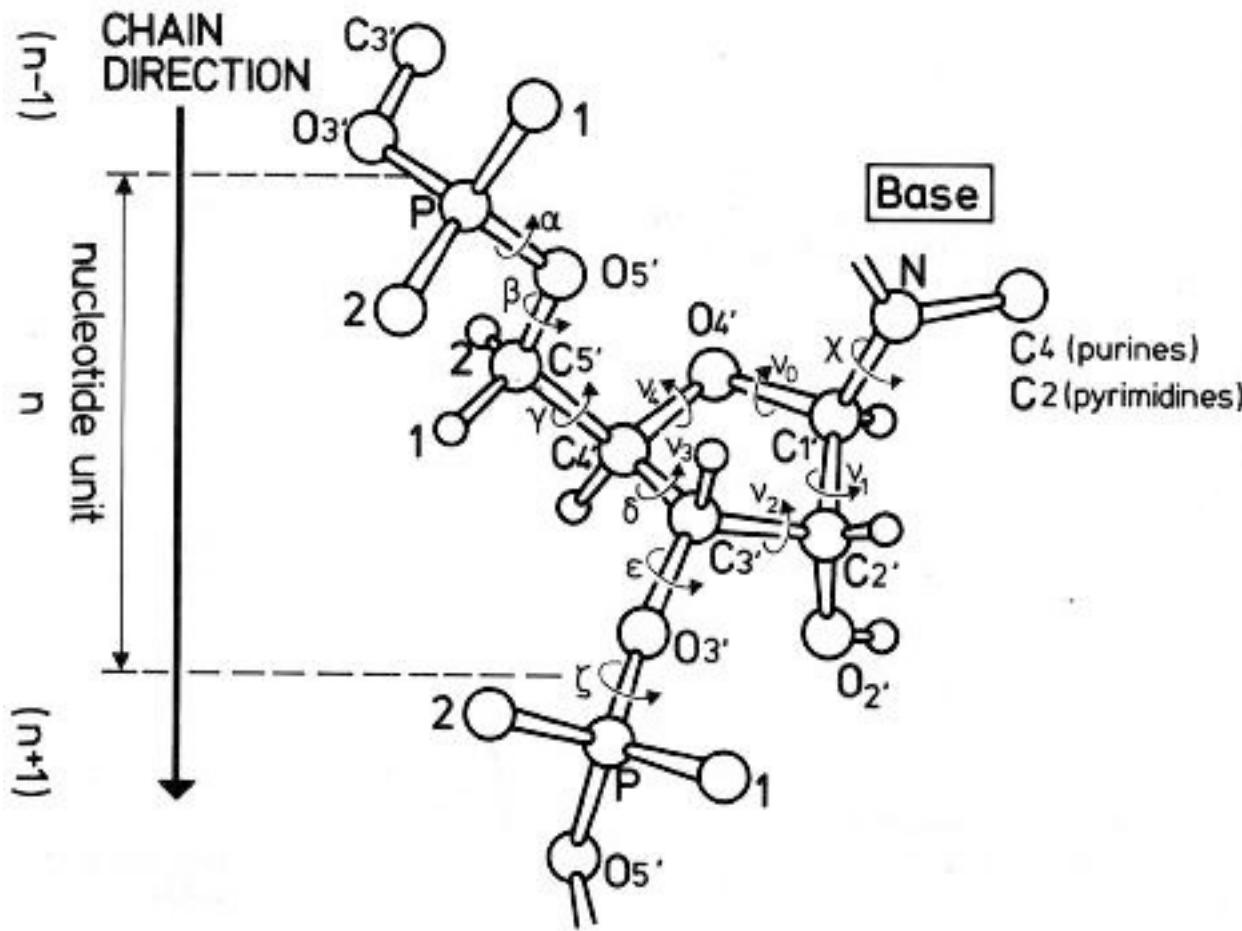


Uracil



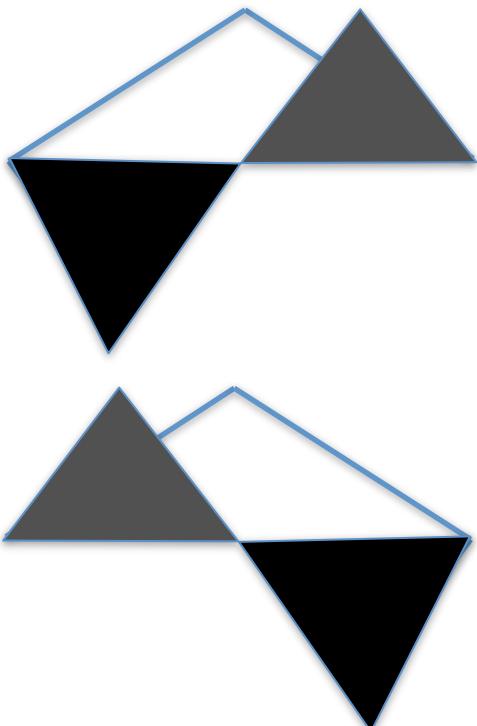
Ribose

# Nucleic acid torsion angles

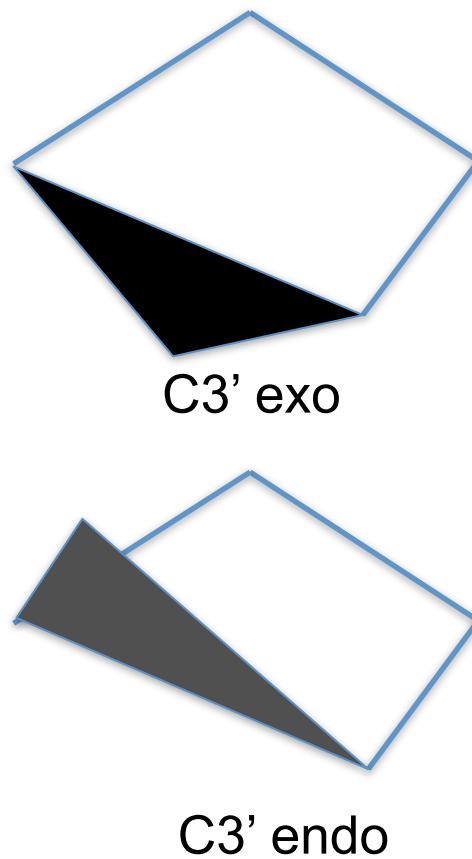


# Furanose ring

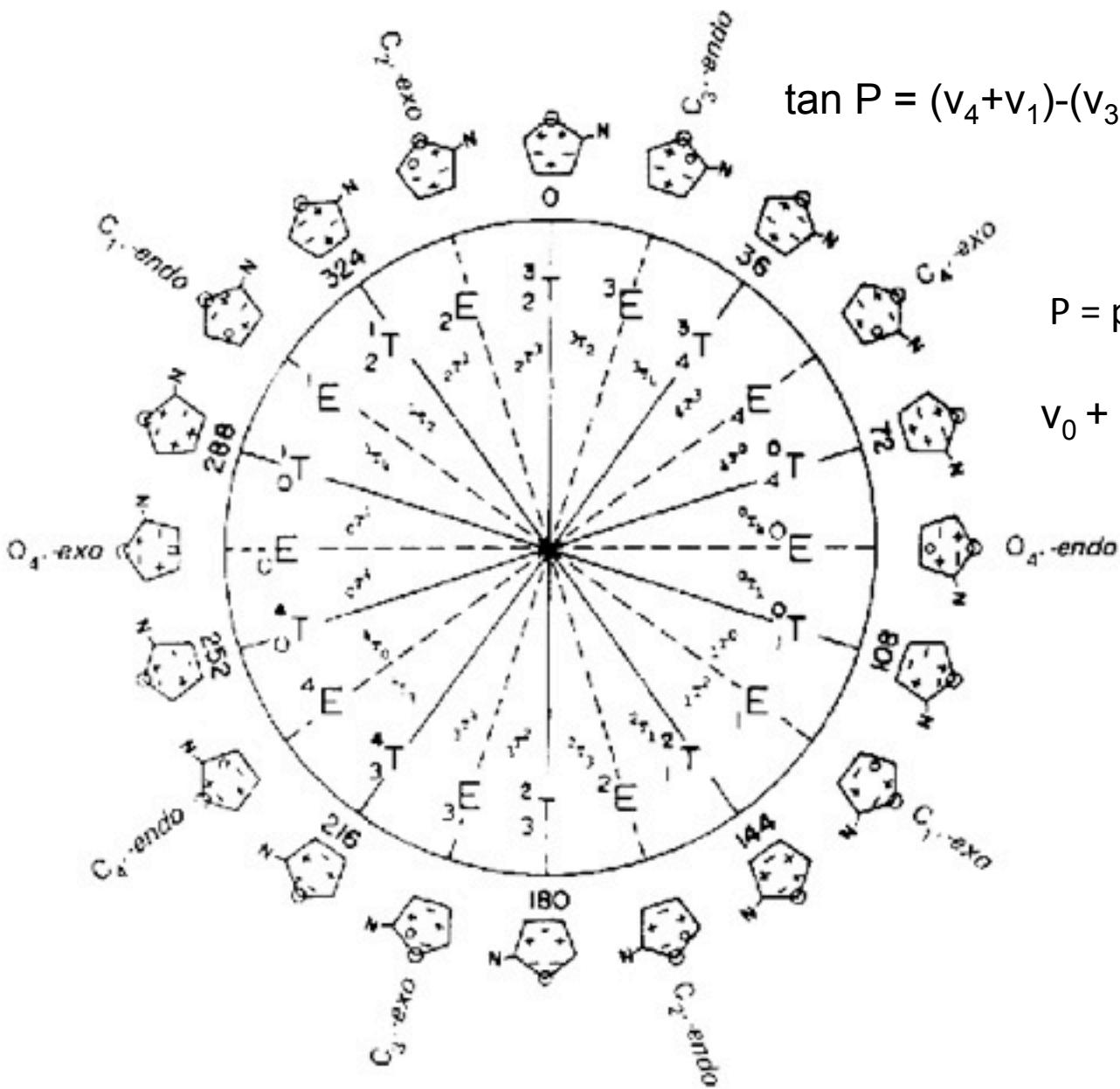
Twisted  
(T)



Envelope  
(E)



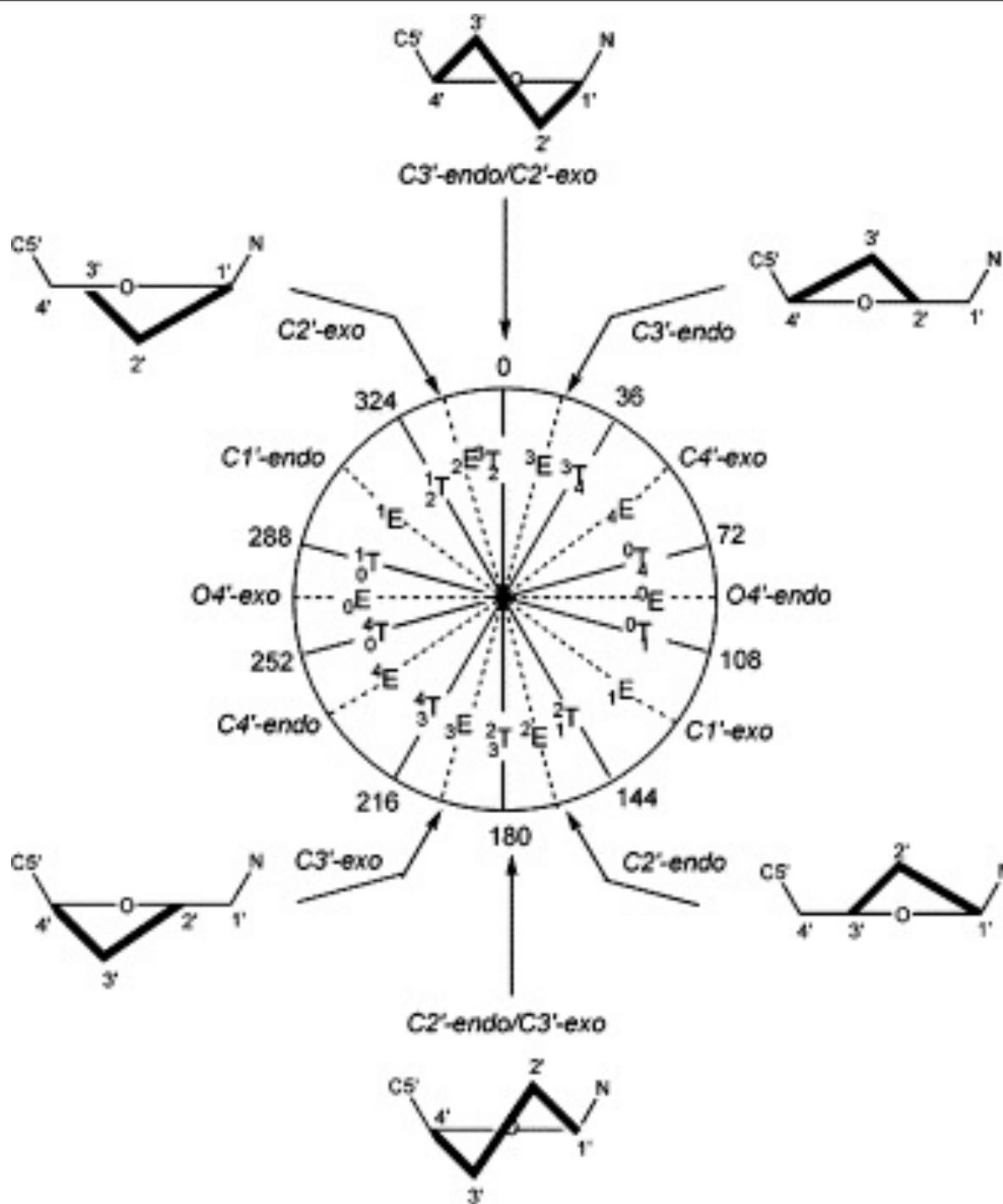
# Sugar Puckering



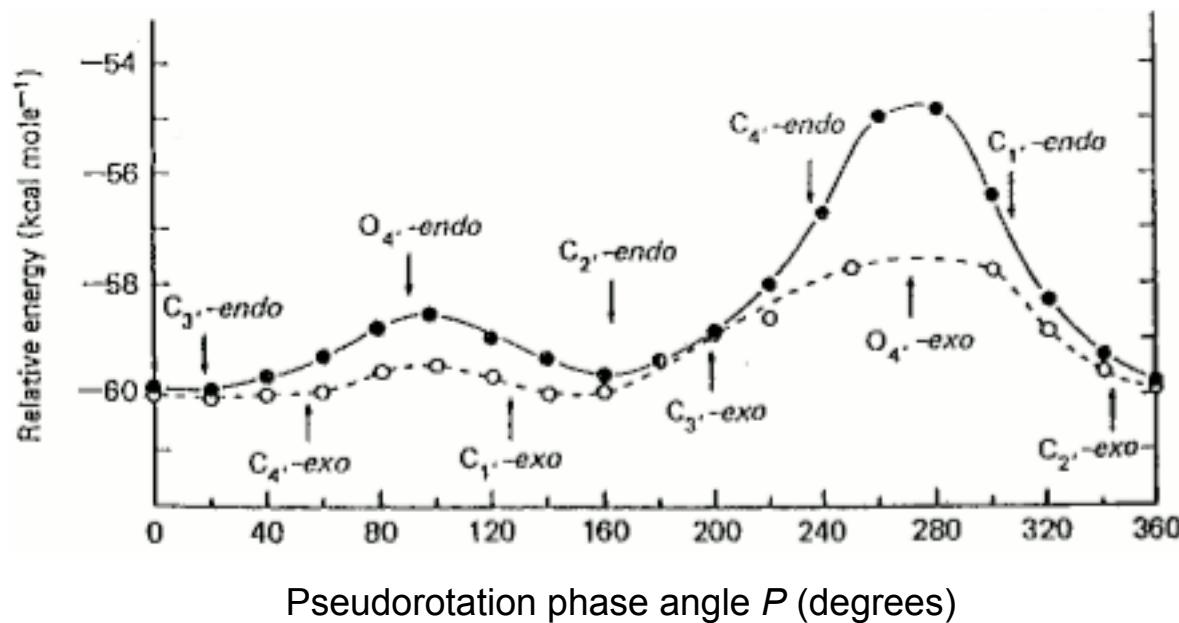
$$\tan P = (v_4 + v_1) - (v_3 + v_0) / [2v_2(\sin 36^\circ + \sin 72^\circ)]$$

$P$  = pseudorotation angle

$$v_0 + v_1 + v_2 + v_3 + v_4 = 0$$

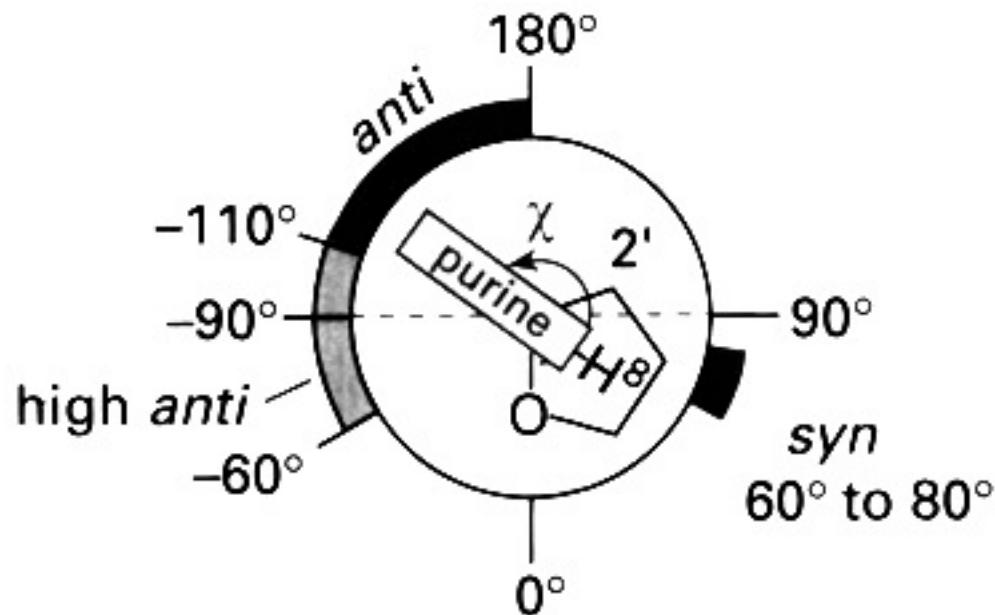


# Energy variation related to $P$

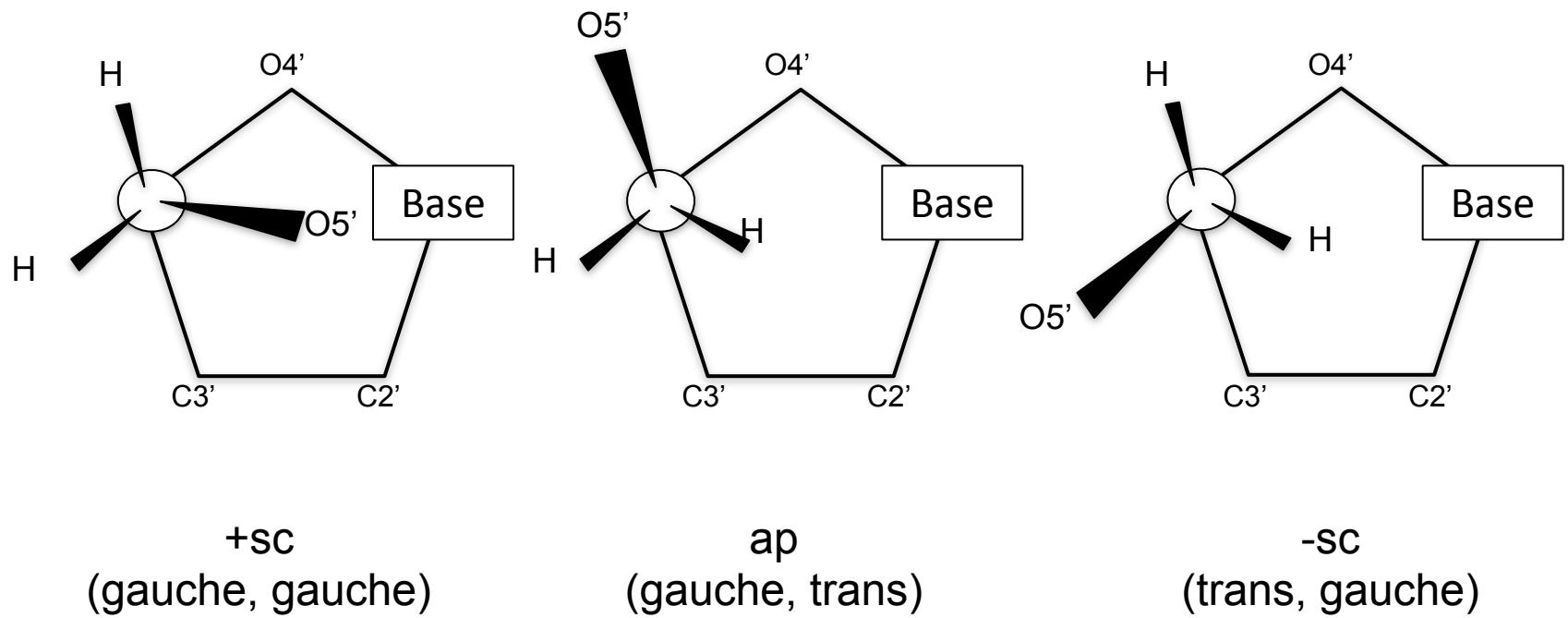


Pseudorotation phase angle  $P$  (degrees)

# Orientation around glycosidic angle



# C4'-C5' torsion angle



# Ramachandran plot

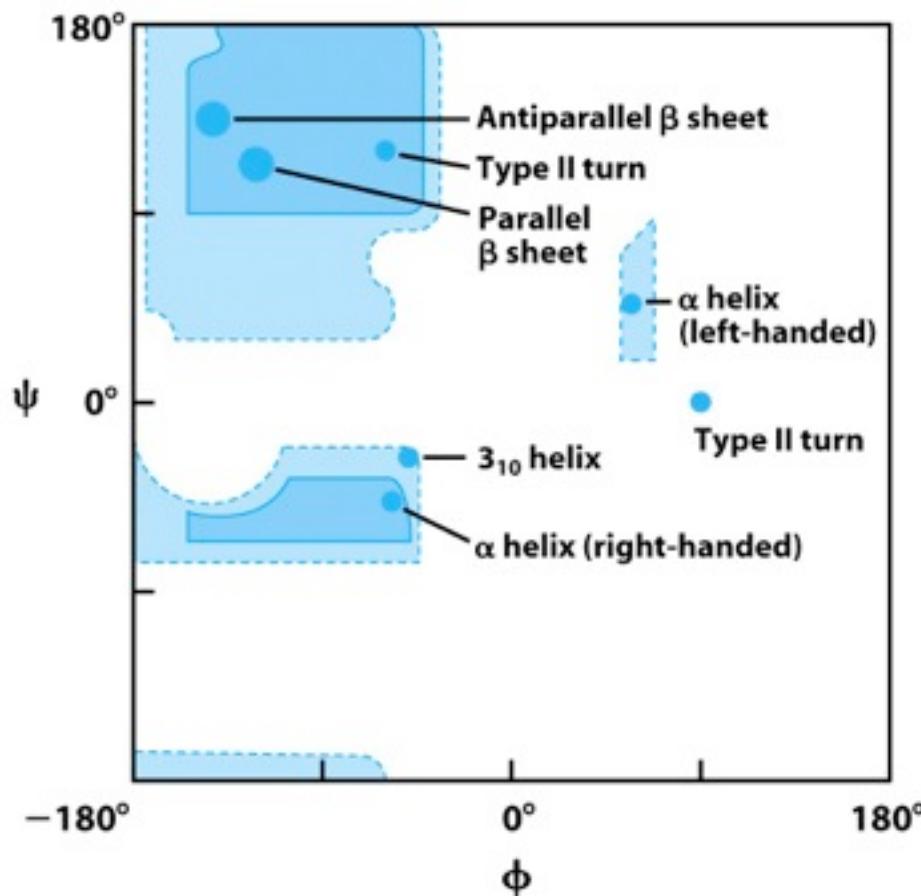
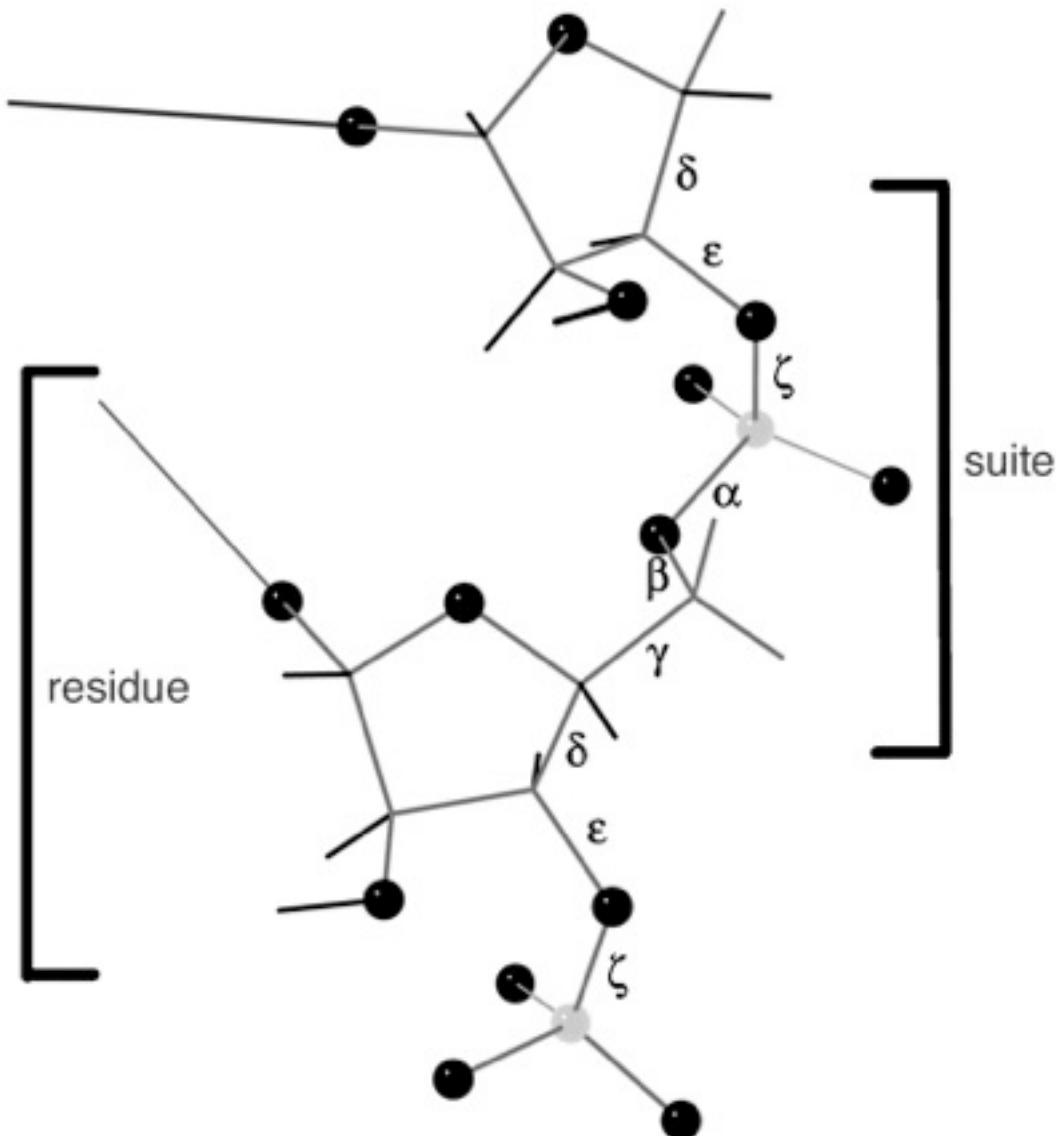
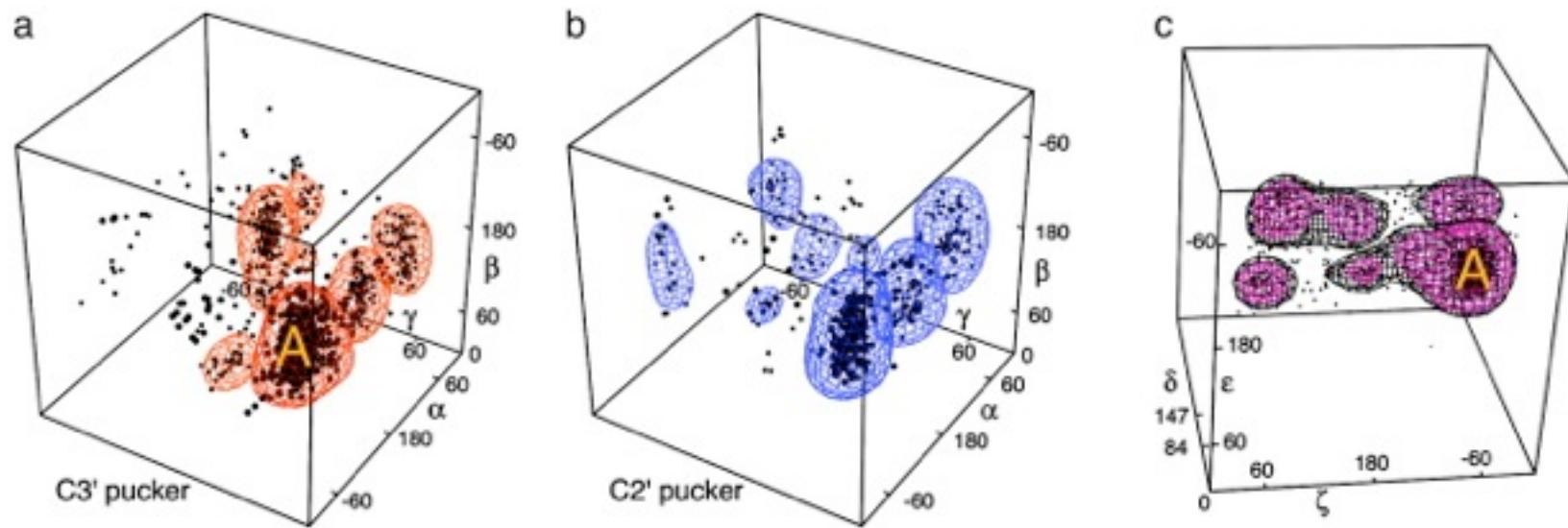


Figure 4-9a Principles of Biochemistry, 4/e  
© 2006 Pearson Prentice Hall, Inc.

# RNA backbone is rotameric



Plots of the heminucleotide angle triplets for data filtered by clashes and  $B > 60$ , with smoothed contours enclosing the top 7–10 peak clusters: (a)  $\alpha$ – $\beta$ – $\gamma$  plot for adjacent sugar pucker C3' endo.



Murray L J W et al. PNAS 2003;100:13904-13909

PNAS

# RNA backbone rotamers

$\alpha$	$\delta$	C3'	C3'	C3'	C3'	C2'	C2'	C2'	
$\beta$	$\epsilon$	e p	e t	e -140	e m	e p	e t	e m	
		p t p	*		*	*	**	*	*
		p 110 t	*						
		p t t					*	a	
		t t p			*** <sup>b</sup>		*	**	*
		t 135 t			*				
		t t t			*** <sup>c</sup>				
		m t p	**	**	A <sup>d</sup>		*** <sup>e</sup>	***	
		m -135 p			*				
		-110 80 t			** <sup>f</sup>				
		m t t					*		
		p t p	*	*	*		**	*	
		p 110 t	*						
		p t m			*				
		t t p	** <sup>g</sup>			*			
		t t t			*				
		m t p	*	*	***		*	*	*
		m -135 p			*** <sup>j</sup>			*	
		m t m			*			*	

Conformer frequencies:

\* ~1% of non-A

\*\* 2-3% of non-A

\*\*\* 4-20% of non-A

Angle codes:

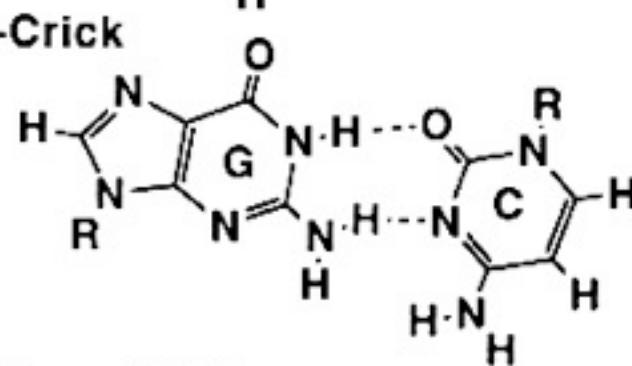
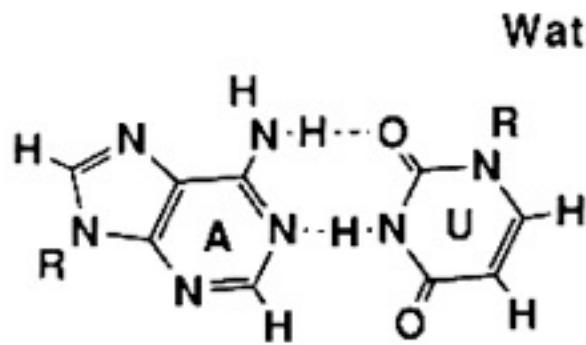
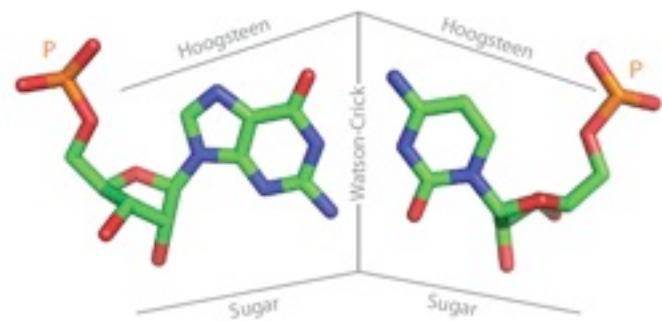
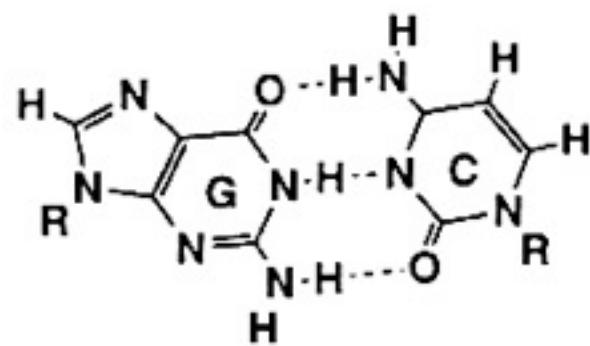
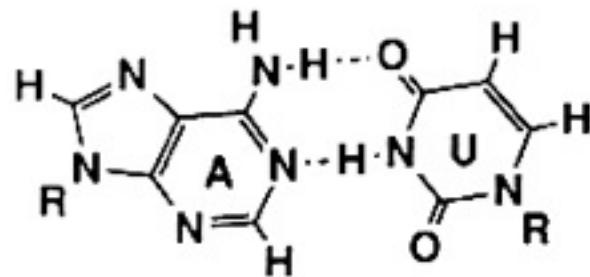
t trans

m gauche minus

p gauche plus

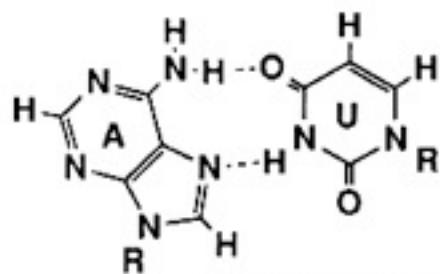
e eclipsed

# Base pairs

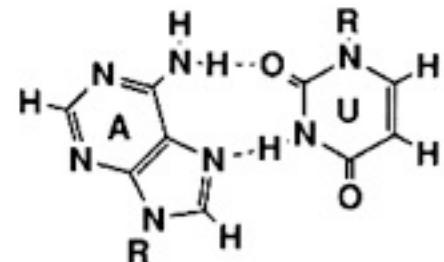


Watson-Crick  
Reverse Watson-Crick

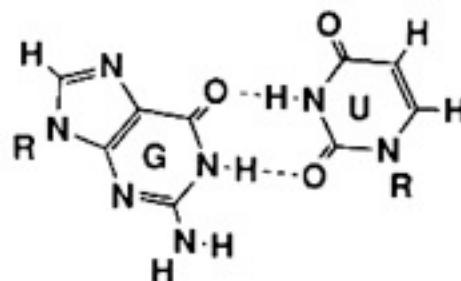
# Hoogsteen, Wobble



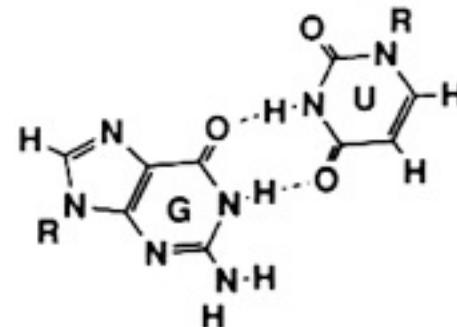
A•U Hoogsteen



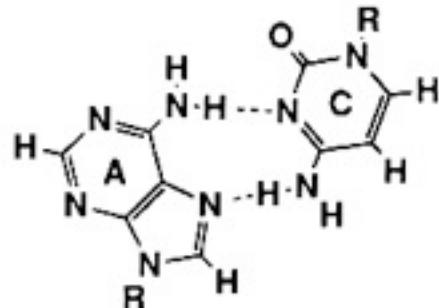
A•U Reverse Hoogsteen



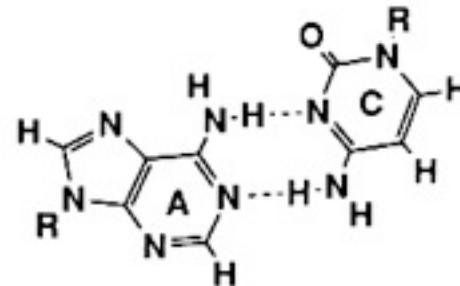
G•U Wobble



G•U Reverse Wobble

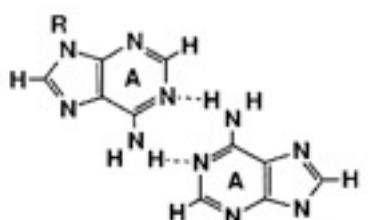


A•C Reverse Hoogsteen

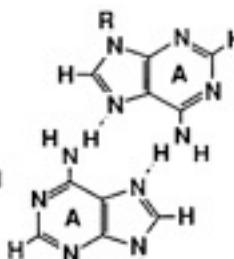


A•C Reverse Wobble

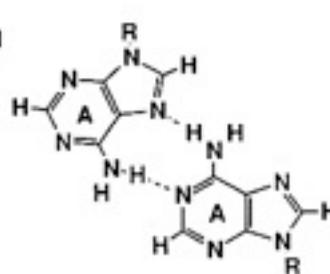
# Homopurines



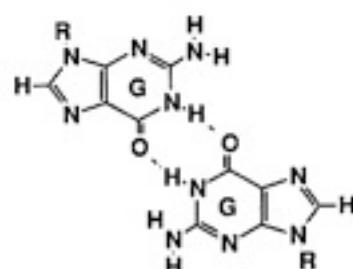
A-A N1-amino,  
symmetric



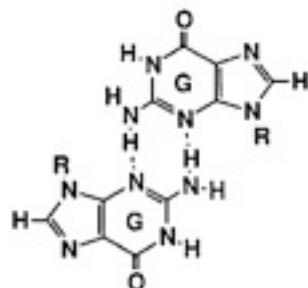
A-A N7-amino,  
symmetric



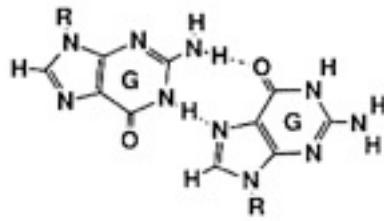
A-A N1-amino,  
N7-amino



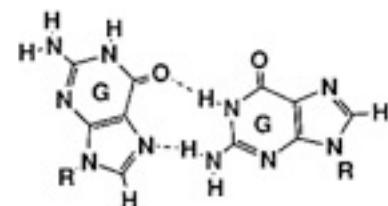
G-G N1-carbonyl,  
symmetric



G-G N3-amino,  
symmetric



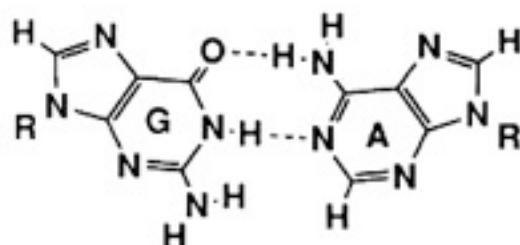
G-G N7-N1,  
carbonyl-amino



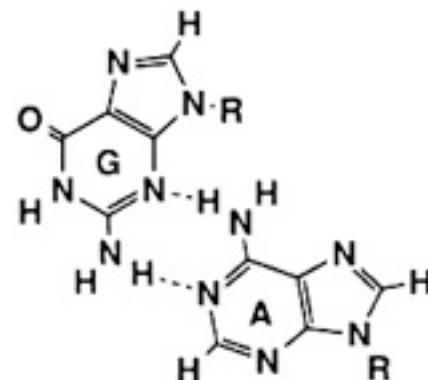
G-G N1-carbonyl,  
N7-amino

-

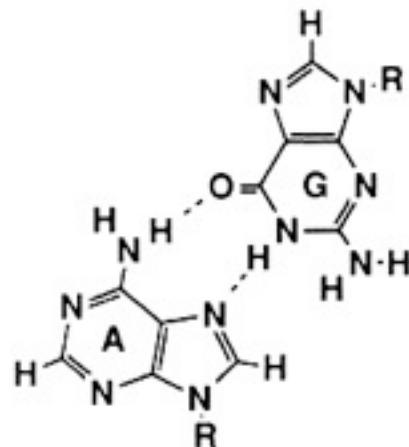
# Heteropurines



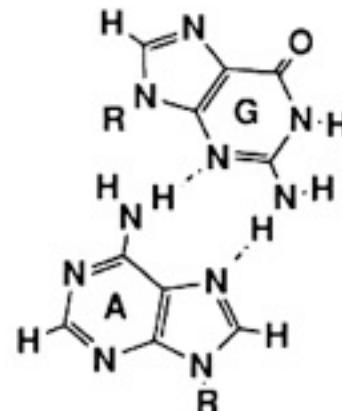
G-A N1-N1,  
carbonyl-amino



G-A N3-amino,  
amino-N1

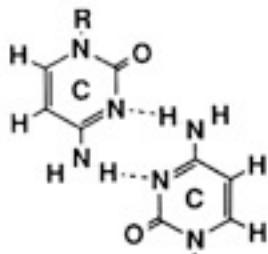


A-G N7-N1,  
amino-carbonyl

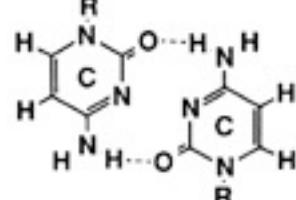


A-G N7-amino,  
amino-N3

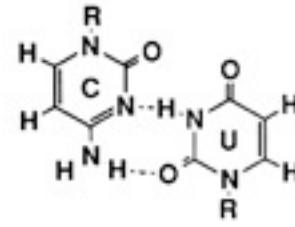
# Pyrimidines



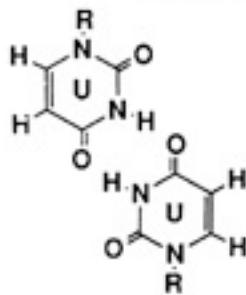
C-C N3-amino,  
symmetric



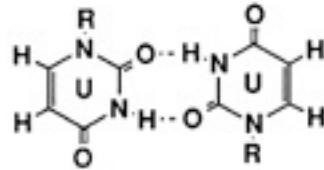
C-C carbonyl-amino,  
symmetric



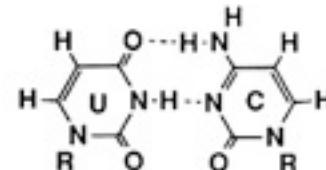
C-U N3-N3,  
2-carbonyl-amino



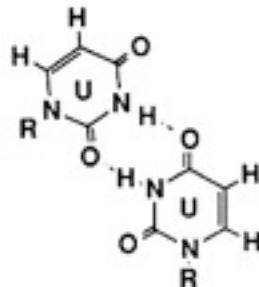
U-U 4-carbonyl-N3,  
symmetric



U-U 2-carbonyl-N3,  
symmetric

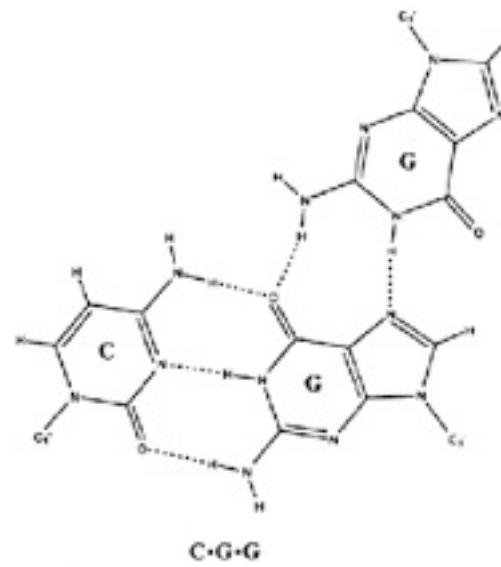
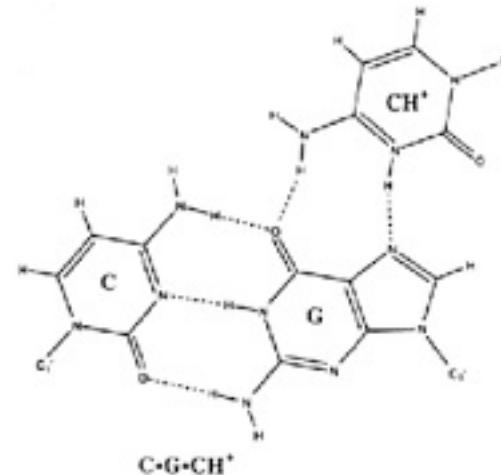
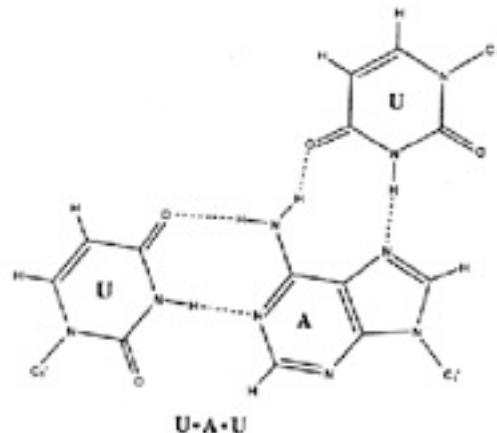


U-C N3-N3,  
4-carbonyl-amino

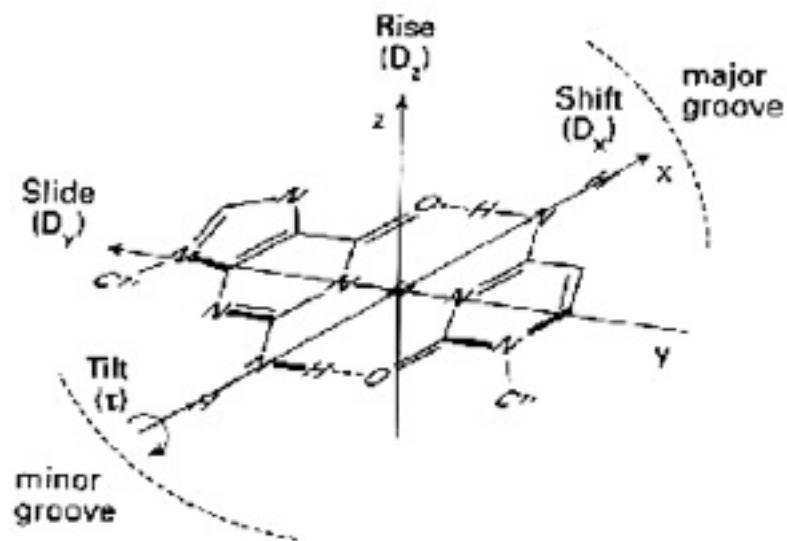
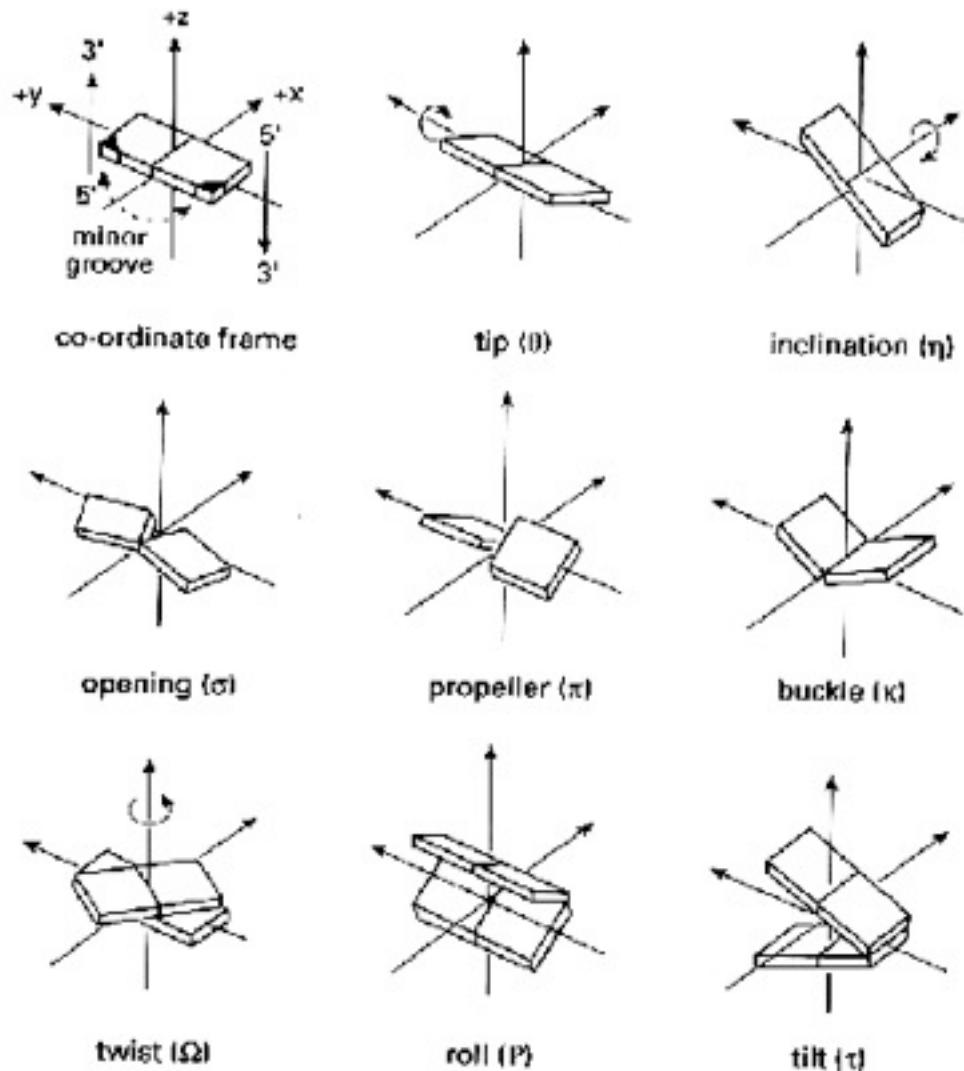


U-U 2-carbonyl-N3,  
4-carbonyl-N3

# Base triples



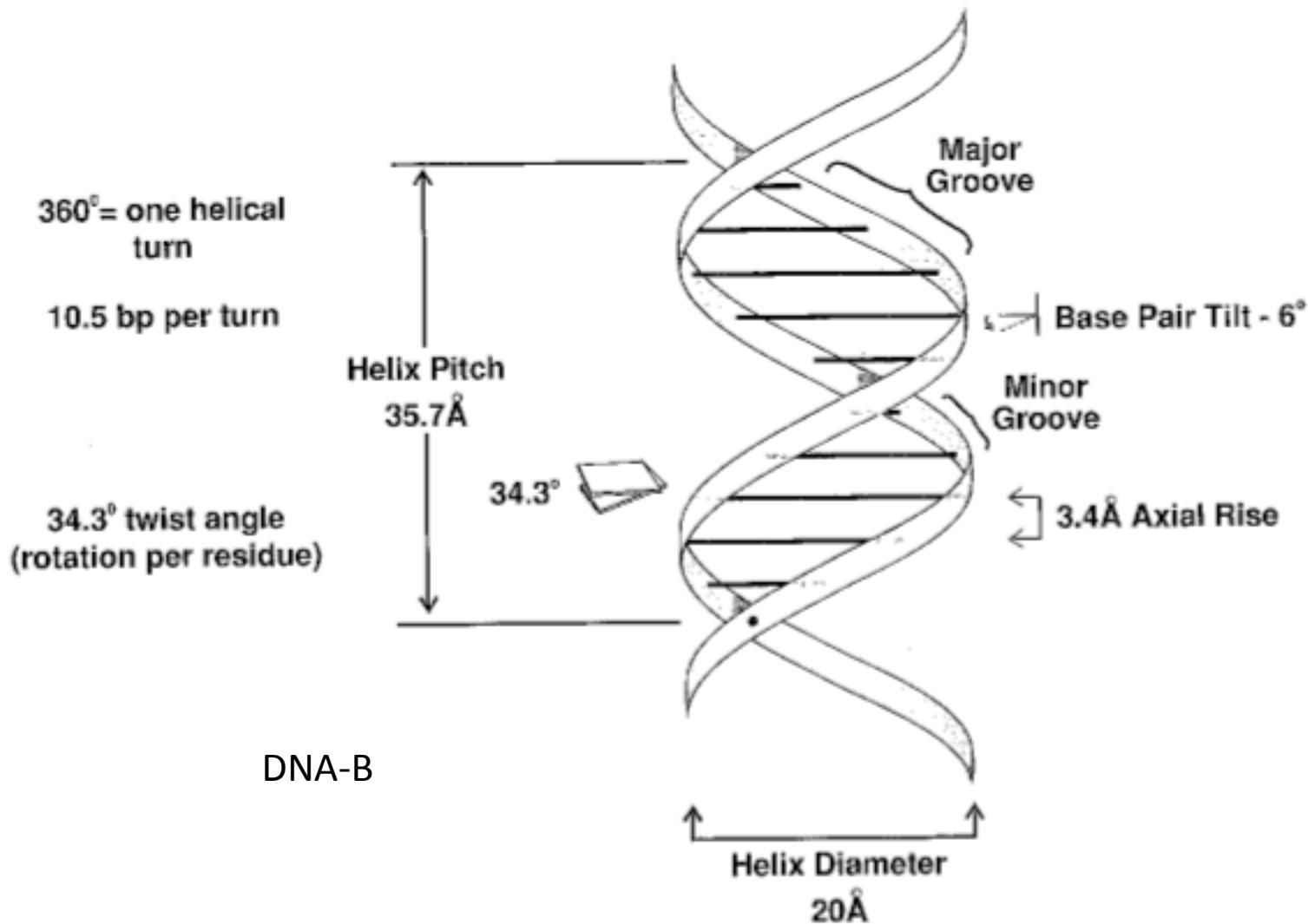
# Movements of base pairs



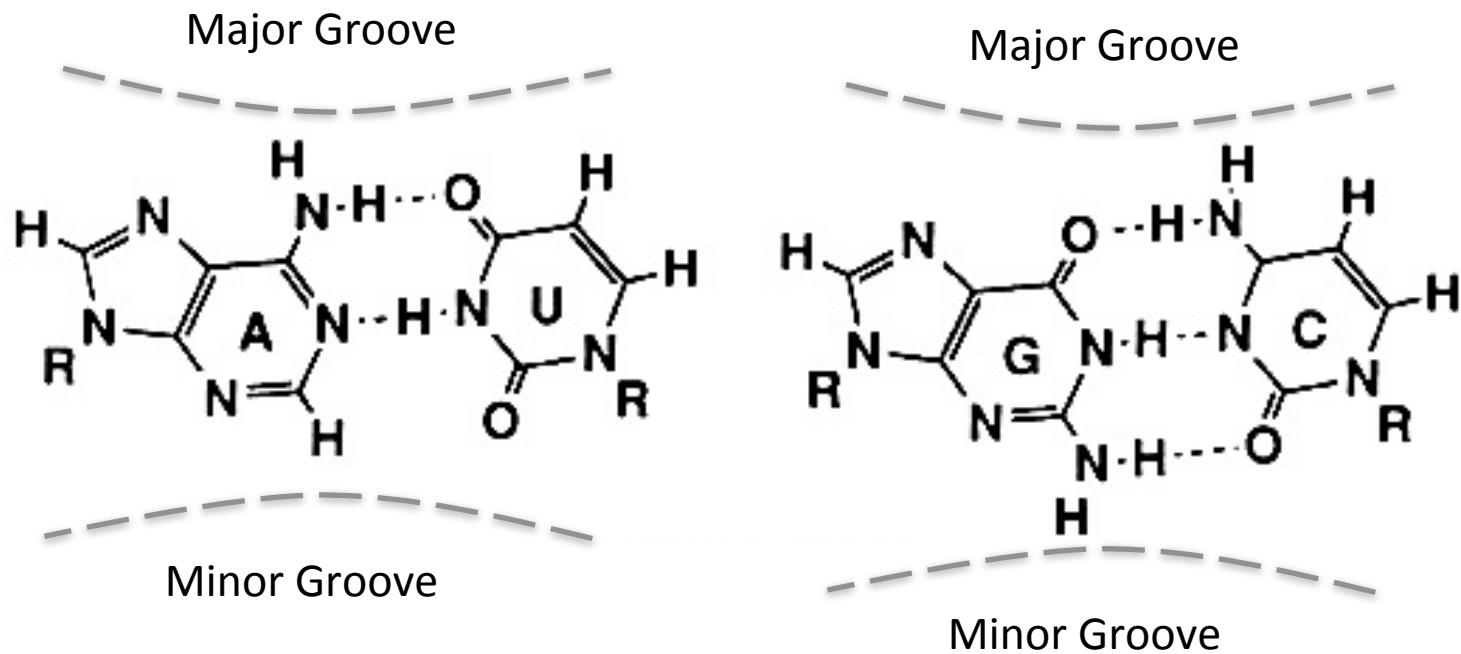
# Helical parameters

- Sense. Helical rotation of the double helix. Clock- or anticklockwise
- Residues per turn. Number of residues in one helical turn.
- Axial rise. Distance between adjacent planar bases.
- Helix pitch. Length of one helical turn.
- Diameter of the helix. Width across the helix.
- Rotation per residue or Twist angle. Angle between two adjacent base pairs.

# Helical parameters

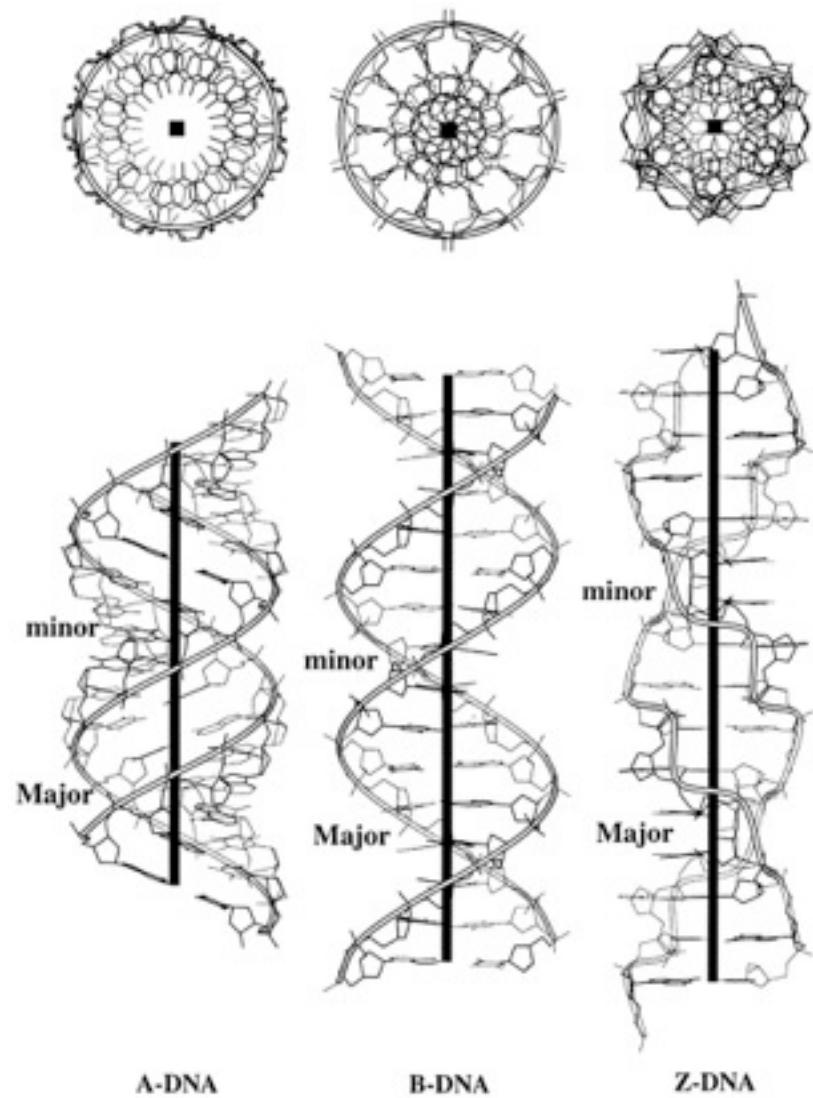


# Major VS Minor Groove



# Nucleic Acid Helical Structures

Geometry attribute:	A-form	B-form	Z-form
Helix sense	right-handed	right-handed	left-handed
Repeating unit	1 bp	1 bp	2 bp
Rotation/bp	33.6°	35.9°	60°/2
Mean bp/turn	11	10.5	12
Inclination of bp to axis	+19°	-1.2°	-9°
Rise/bp along axis	2.4 Å (0.24 nm)	3.4 Å (0.34 nm)	3.7 Å (0.37 nm)
Rise/turn of helix	24.6 Å (2.46 nm)	33.2 Å (3.32 nm)	45.6 Å (4.56 nm)
Mean propeller twist	+18°	+16°	0°
Glycosyl angle	anti	anti	pyrimidine: anti, purine: syn
Sugar pucker	C3'-endo	C2'-endo	C: C2'-endo, G: C2'-exo
Diameter	23 Å (2.3 nm)	20 Å (2.0 nm)	18 Å (1.8 nm)

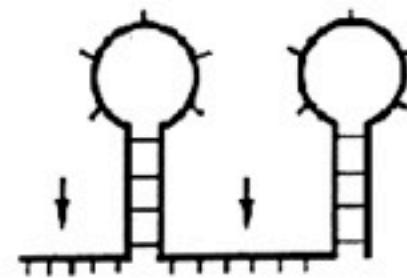


# Secondary structures in RNA

a. DUPLEXES



b. SINGLE STRANDED REGIONS



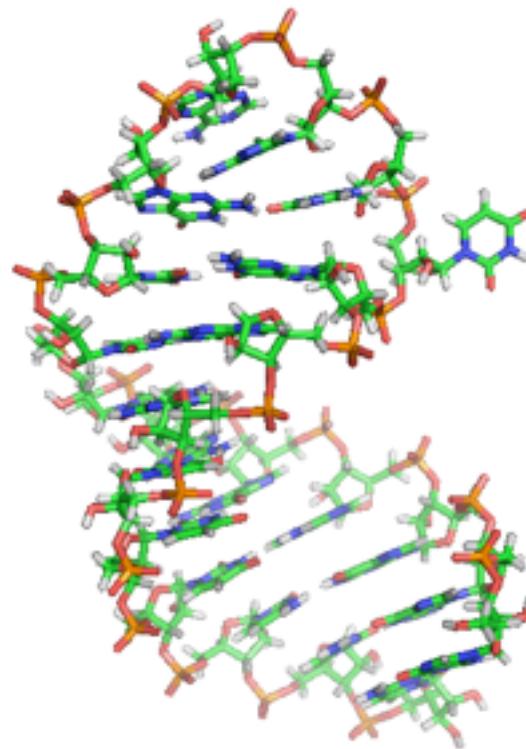
# Apical loops (hairpins)

c. HAIRPINS



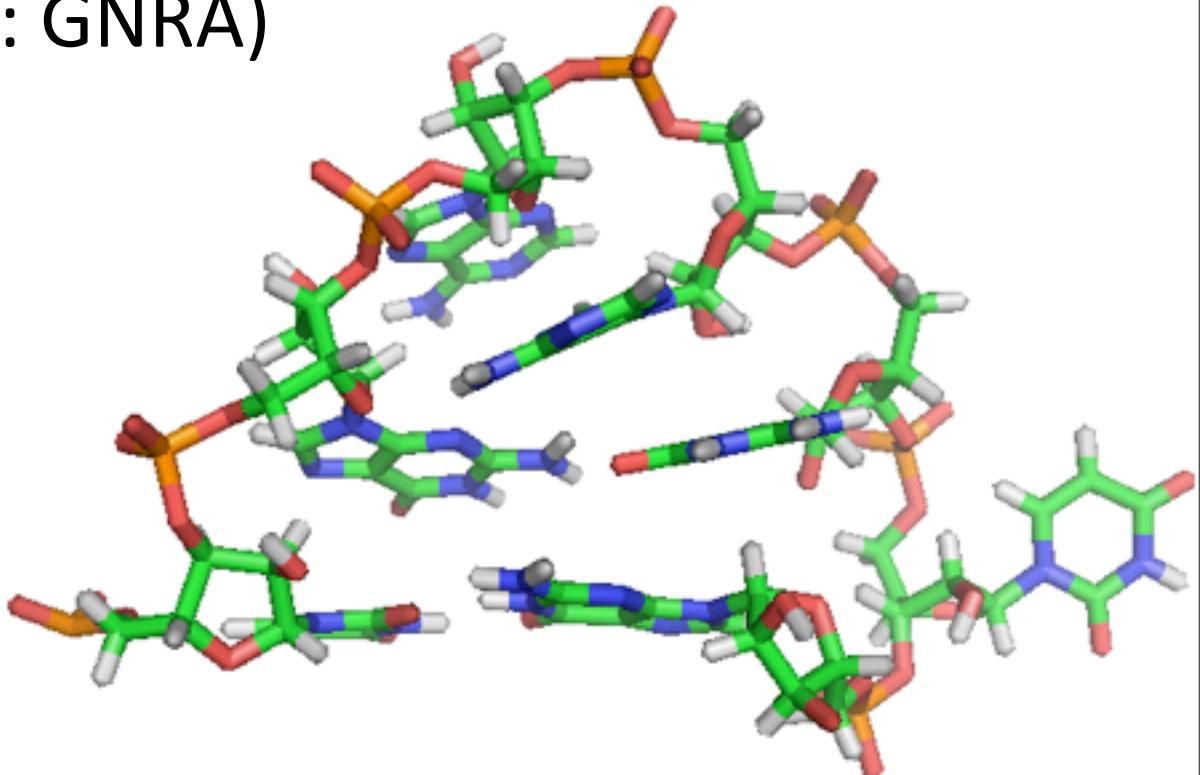
HAIRPIN LOOP

HAIRPIN STEM



# Several kinds of apical loops

- Minimum: triloops
- Tetraloops (e.g.: GNRA)
- Pentaloops
- Hexaloops
- .
- .
- .
- whatever loops



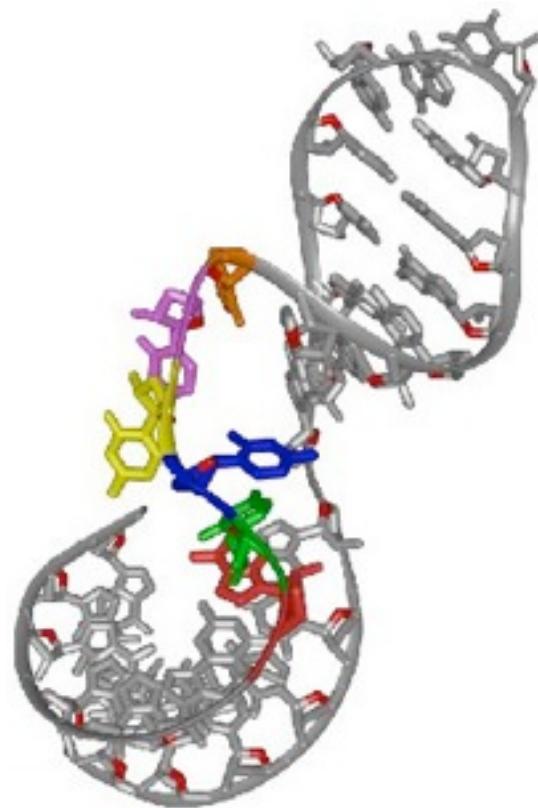
**d. BULGES**



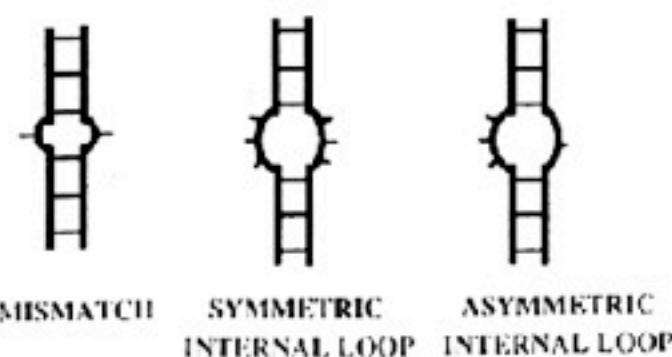
BULGE



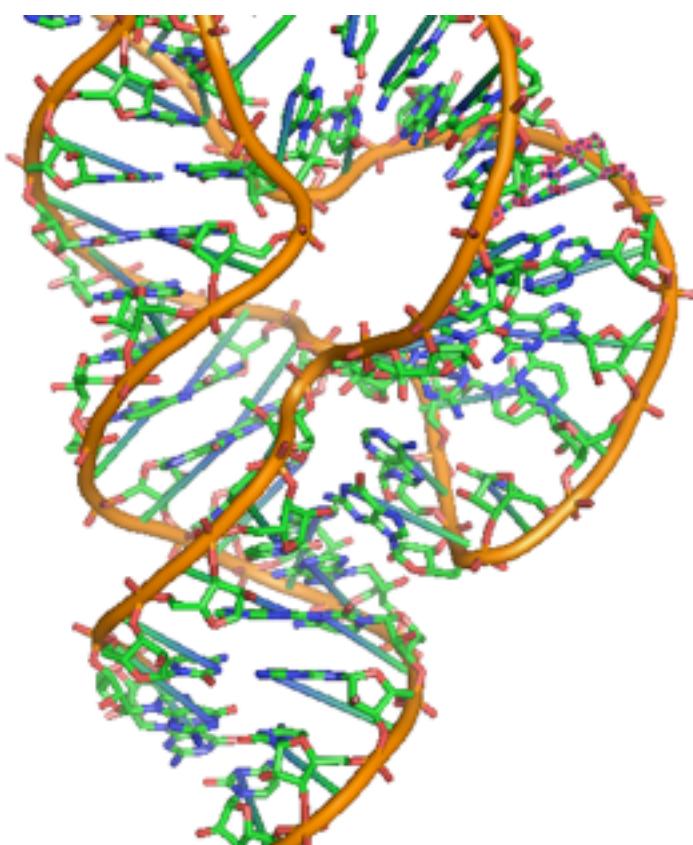
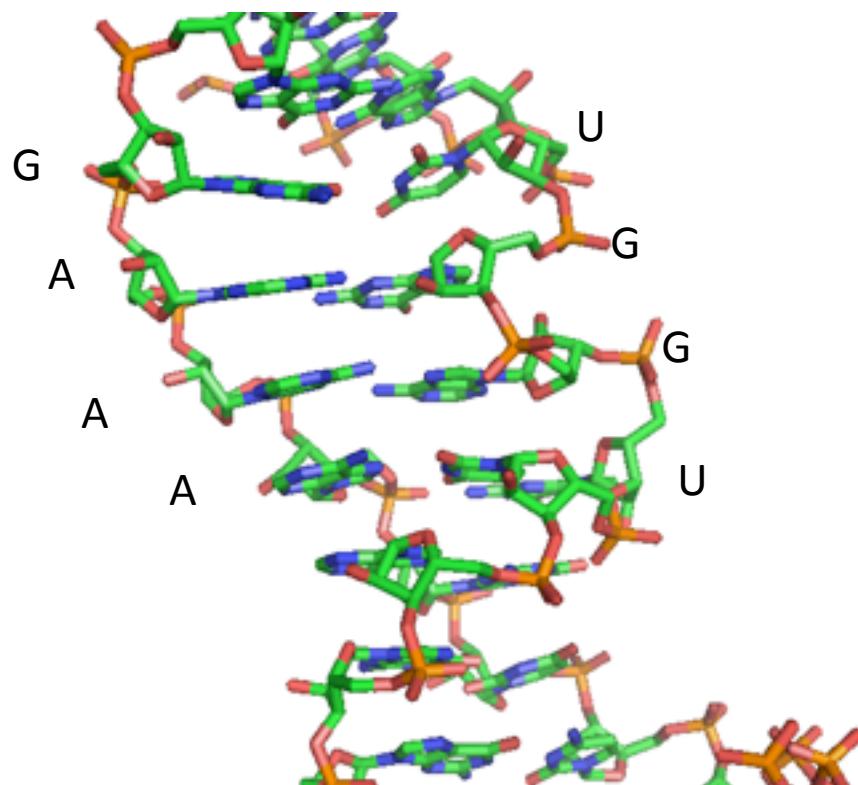
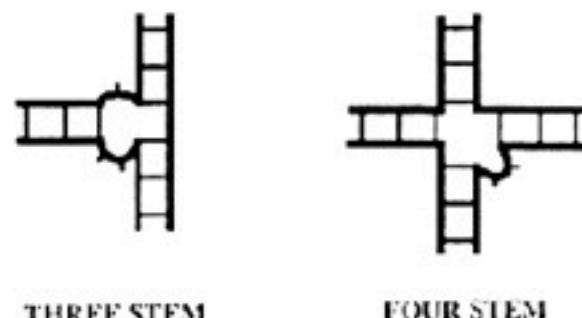
SINGLE-BASE BULGE



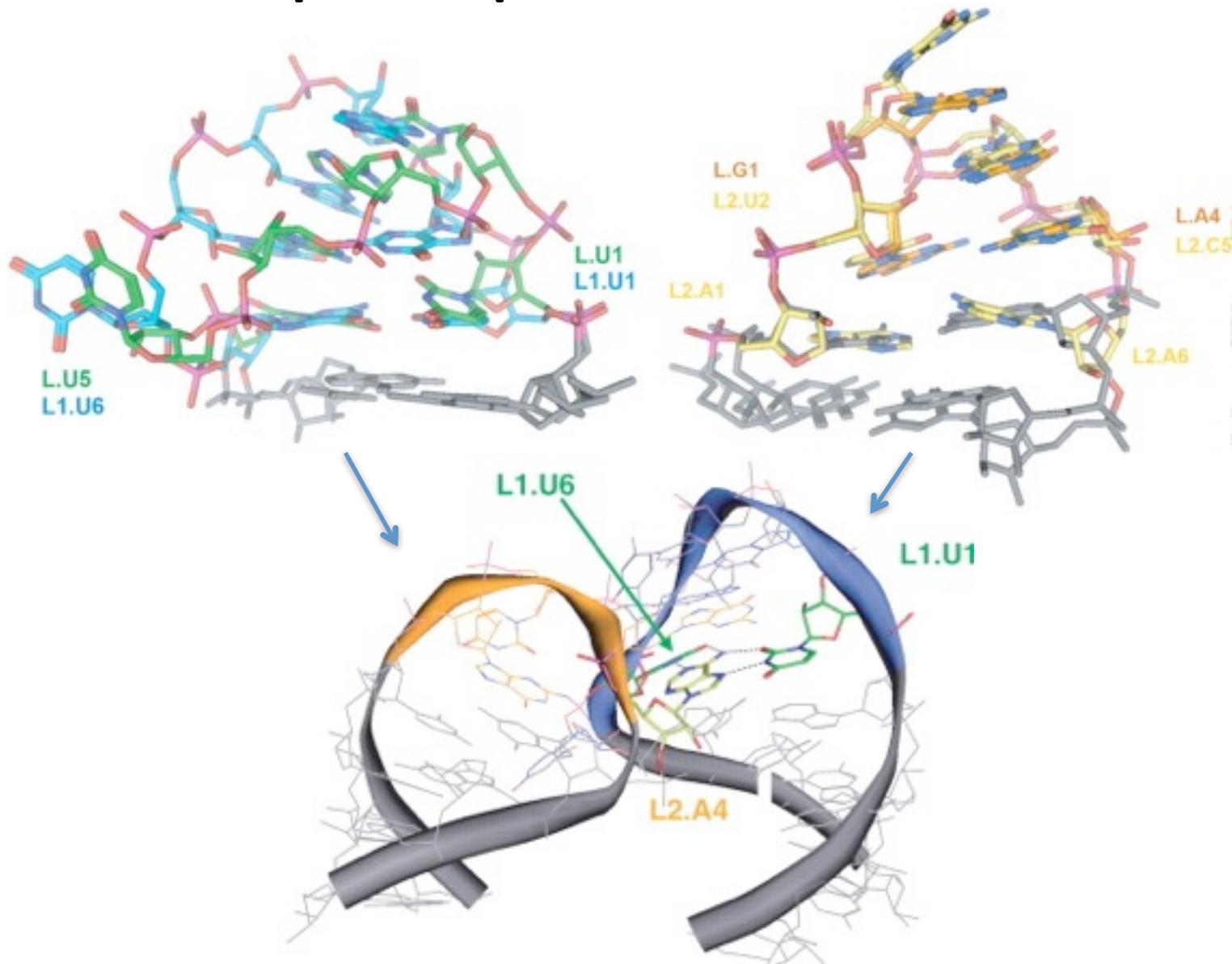
e. INTERNAL LOOPS



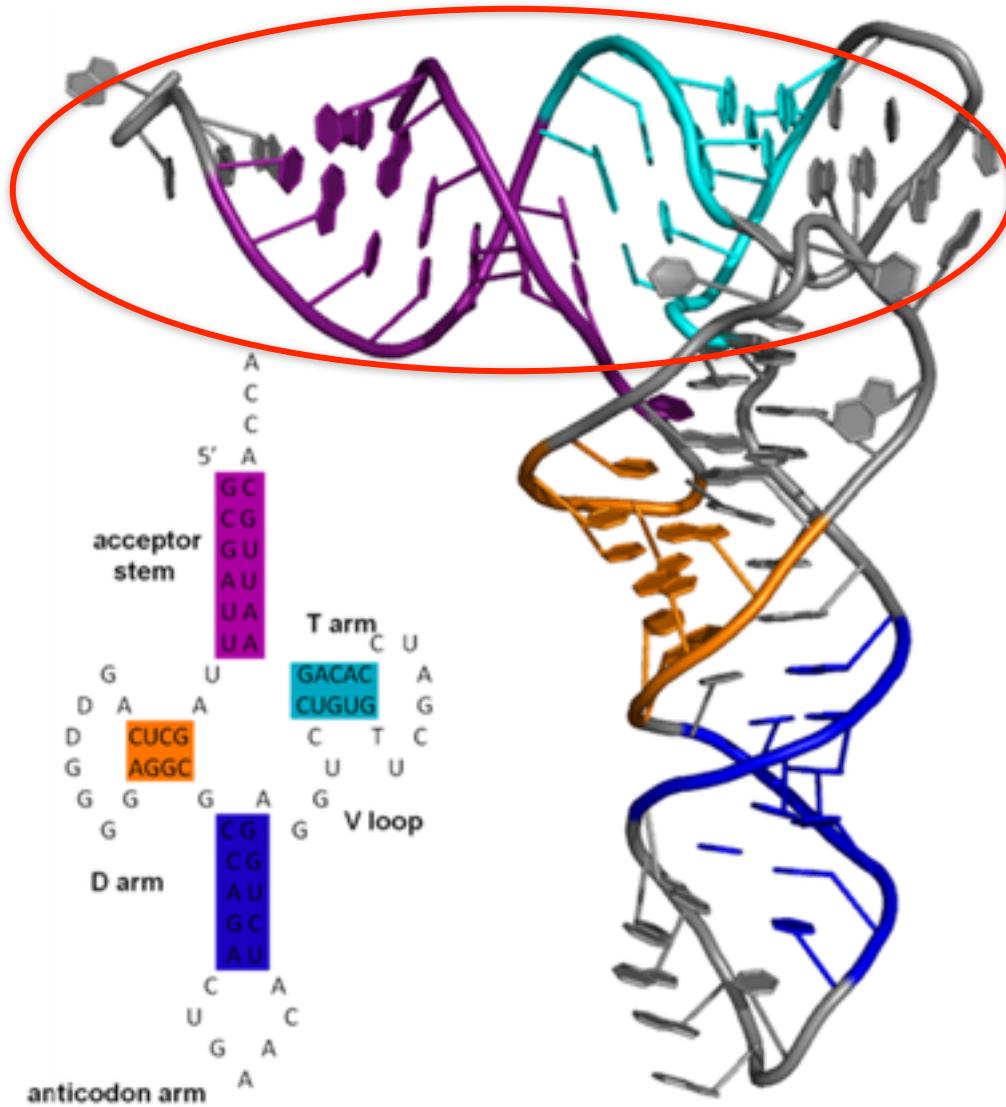
f. JUNCTIONS



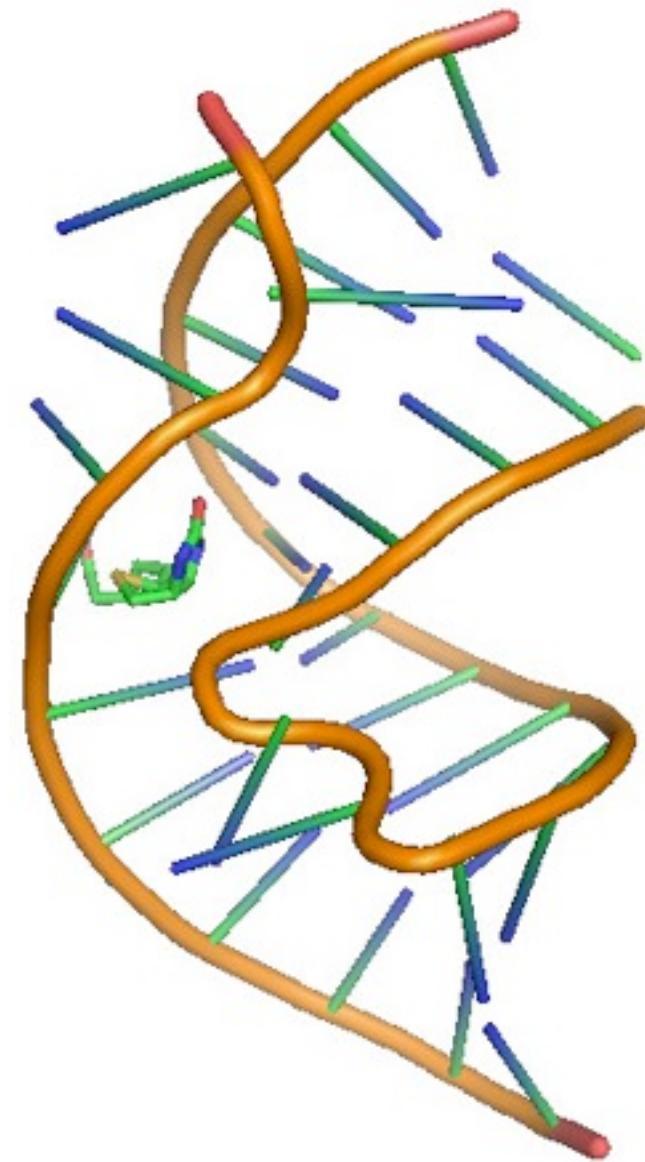
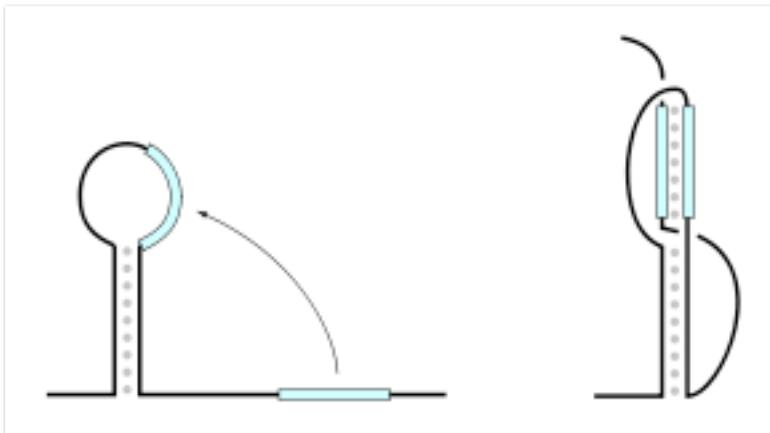
# Loop-loop interactions



# Coaxial stacking

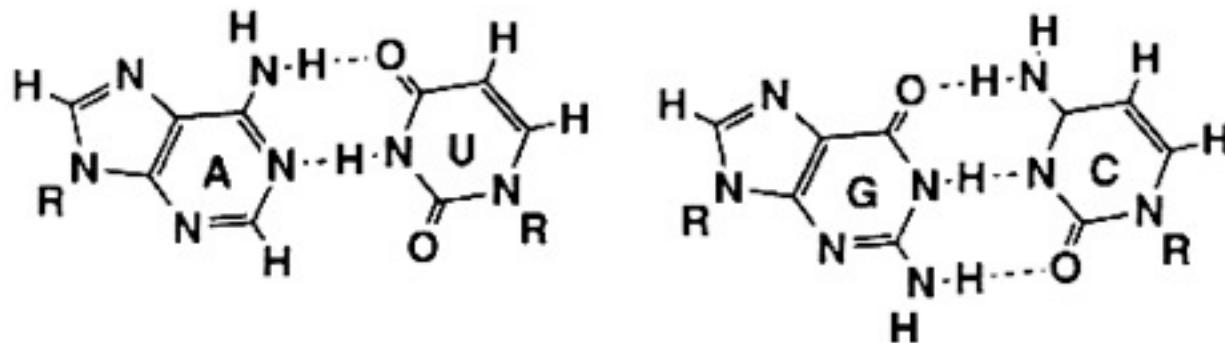


# Pseudoknots



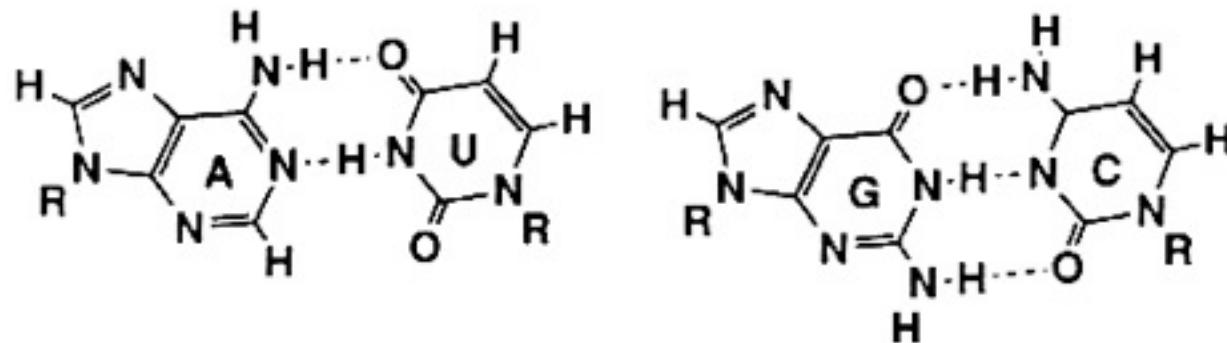
# Forces that drive RNA folding

- 1) Hydrogen bonds

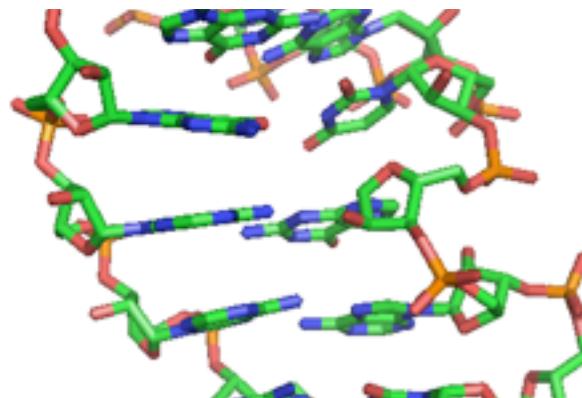


# Forces that drive RNA folding

- 1) Hydrogen bonds

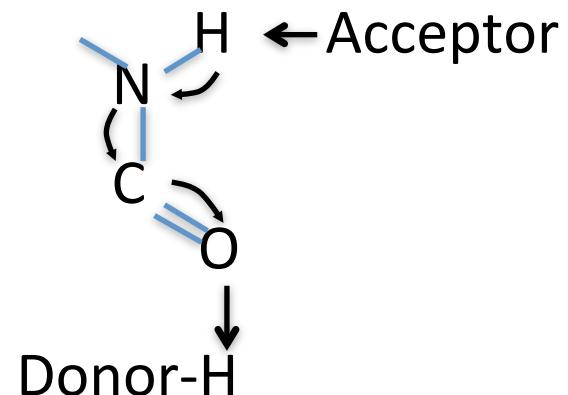


- 2) Stacking interactions



# Hydrogen Bonding

- Short, noncovalent, highly directional between an electronegative atom ( $-C=O$ ,  $-N:$ ) and a H atom bound to an electropositive atom ( $-N-H$ ).
- 20-30 times weaker compared to covalent bonds (3-7 kcal/mol VS 80-100 kcal/mol)
- Additive and cooperative



# Base Stacking

- Base stacking = Van der Waals forces + Hydrophobic interactions
- 4-15 kcal/dinucleotide
- Van der Waals = dipole-dipole interactions + London dispersion forces

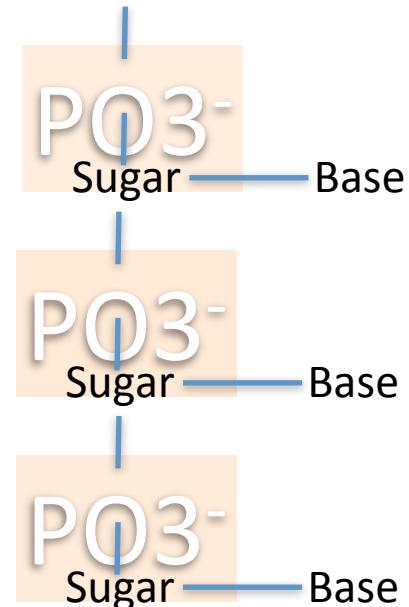
Dinucleotide base pairs	Stacking energies (kcal/mol/stacked pair)
GC · GC	-14.59
AC · GT	-10.51
TC · GA	-9.81
CG · CG	-9.69
GG · CC	-8.26
AT · AT	-6.57
TG · CA	-6.57
AG · CT	-6.78
AA · TT	-5.37
TA · TA	-3.82

# Base Stacking

- Isodesmic reaction
- Additive
- Diffusion controlled reaction
- Weak force controlled process
- $E_{\text{tot}} = E_d + E_p + E_{\text{el}}$
- $E_{\text{tot}}$  = Total energy
- $E_d$  = Dispersion energy
- $E_p$  = Polarization energy

# Other interactions

- RNA is a polianion
- Metal ion coordination ( $Mg^{2+}$ )
  - Diffusion controlled
  - Through water or direct contact
  - $PO_3^-$  or bases



# NDB

<http://ndbserver.rutgers.edu/>



## WELCOME TO THE NUCLEIC ACID DATABASE

a repository of three-dimensional structural information about nucleic acids

[Site Index](#)

- [Atlas](#)
- [Deposit Data](#)
- [Download Data](#)
- [Search](#)
- [Education](#)
- [Standards](#)
- [Tools](#)
- [Links](#)

Number of Released Structures:  
**6425 Structures**

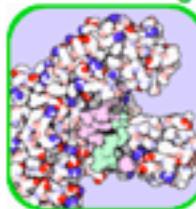
Last Update: 20-Feb-2013

[Search the NDB by ID](#)  
Enter an NDB ID or PDB ID

Search for Released Structures

### Nucleic Acids Highlight



### About NDB

The NDB follows the dictionaries and formats used by the Worldwide Protein Data Bank. Please see [www.wwpdb.org](http://www.wwpdb.org) for format announcements and documentation.

### Archive of NDB newsletters

The NDB is supported by funds from the [National Science Foundation](#) and the [Department of Energy](#).

In citing the NDB please refer to: H. M. Berman, W. K. Olson, D. L. Beveridge, J. Westbrook, A. Gelbin, T. Demeny, S.-H. Hsieh, A. R. Srinivasan, and B. Schneider. (1992) The Nucleic Acid Database: A Comprehensive Relational Database of Three-Dimensional Structures of Nucleic Acids. *Biophys. J.*, 63, 751-759.

[ndbadmin@ndbserver.rutgers.edu](mailto:ndbadmin@ndbserver.rutgers.edu)

©1995-2012 The Nucleic Acid Database Project Rutgers, The State University of New Jersey