INTRODUCTION TO PROTEIN STRUCTURES

Andrea Giansanti
Dipartimento di Fisica, Sapienza Università di Roma

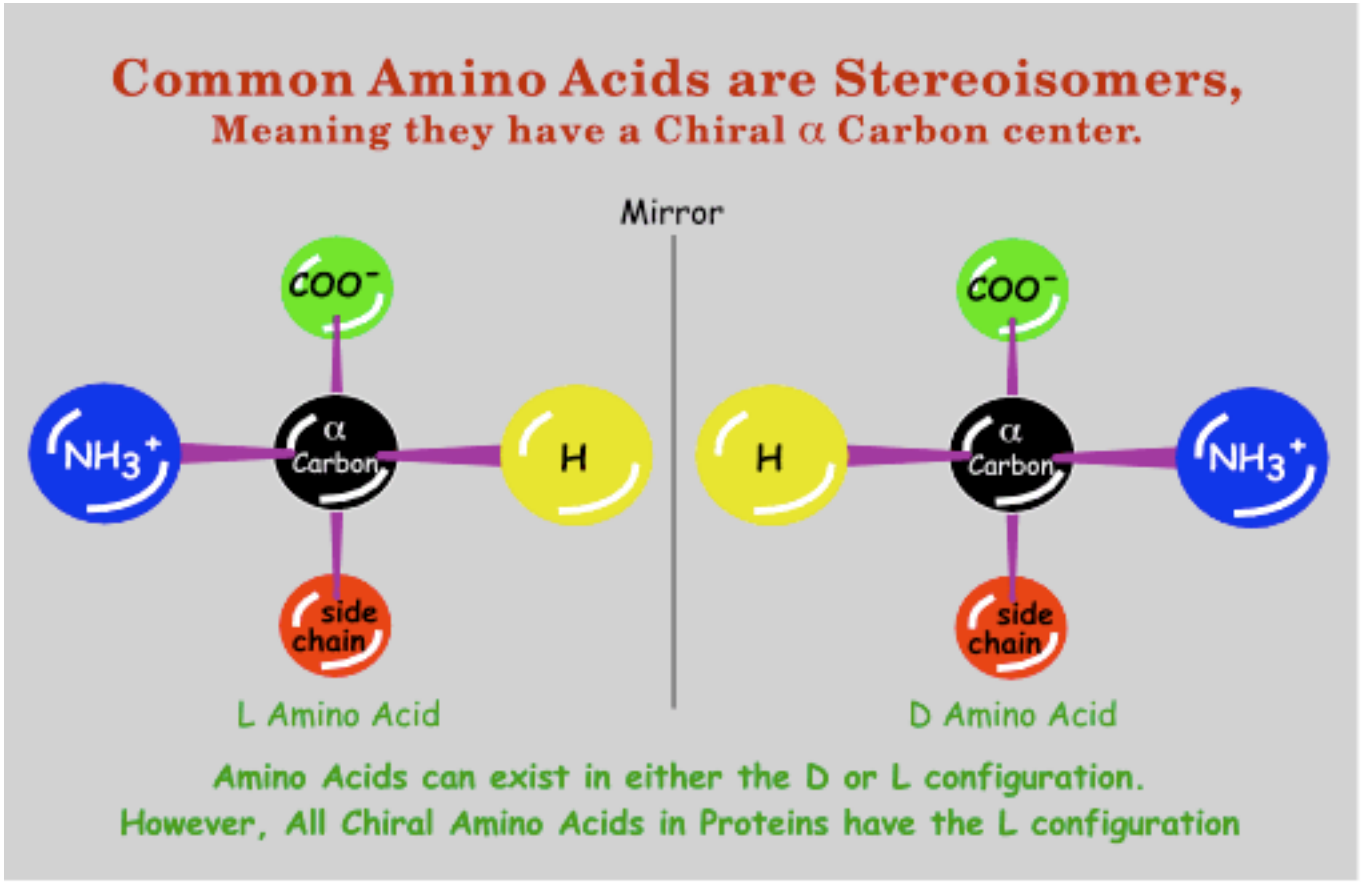
Andrea.Giansanti@roma1.infn.it

COMPUTATIONAL BIOPHYSICS CB_23_24_L11
thu Oct 19 2023

DIPARTIMENTO DI FISICA

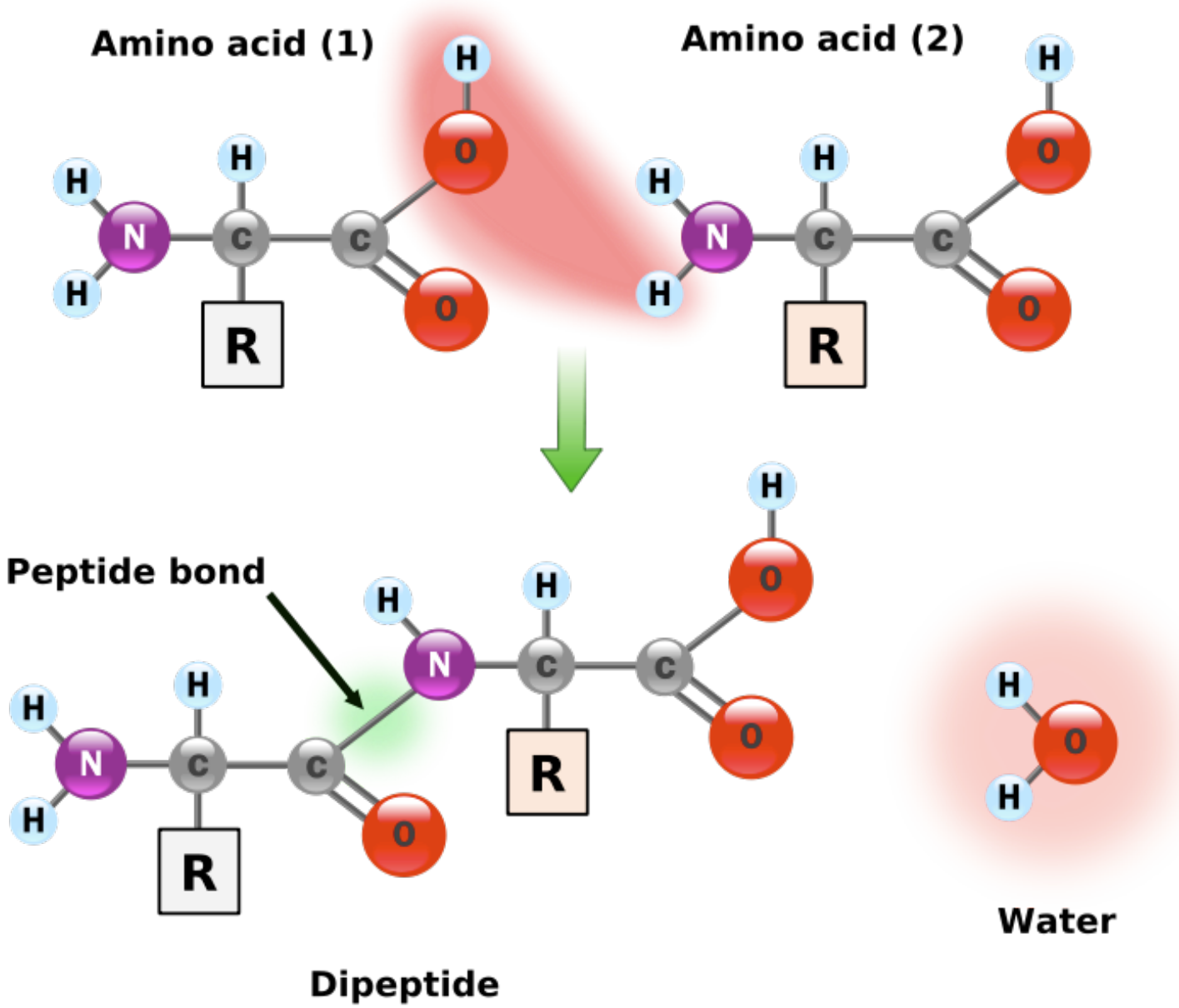
SAPIENZA
UNIVERSITÀ DI ROMA

- SEQUENCE->STRUCTURE/UNSTRUCTURE->DYNAMICS-FUNCTIONS

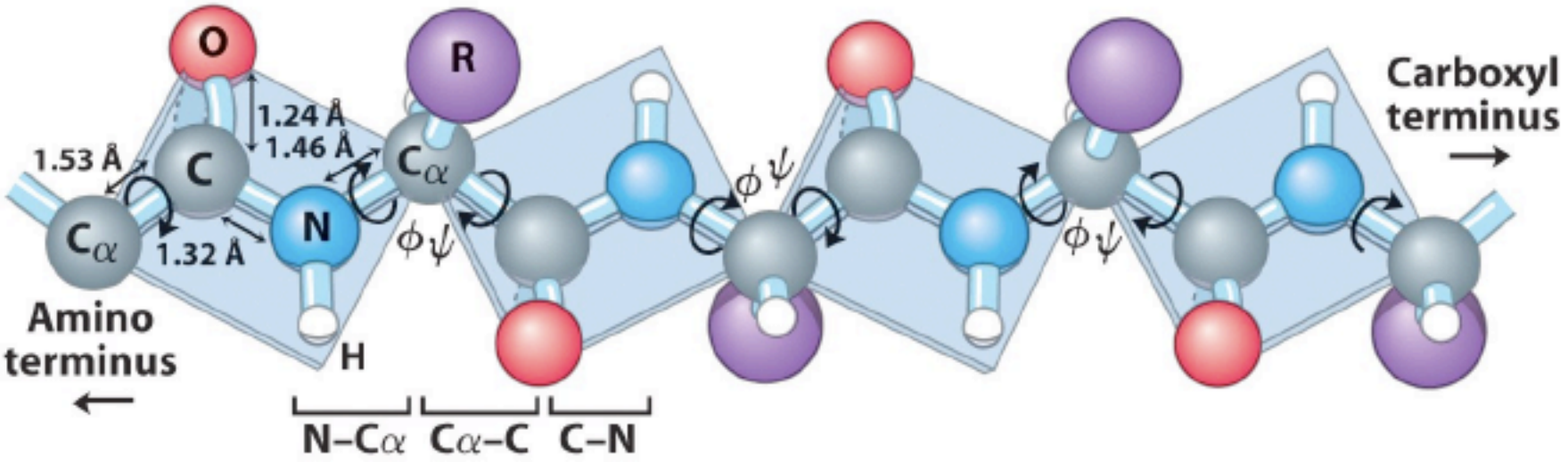


19 out of 20 are chiral: natural state is L! Still an unsolved puzzle.
Nevertheless D-chiral enzymes are active on D-chiral substrates

PEPTIDE BOND

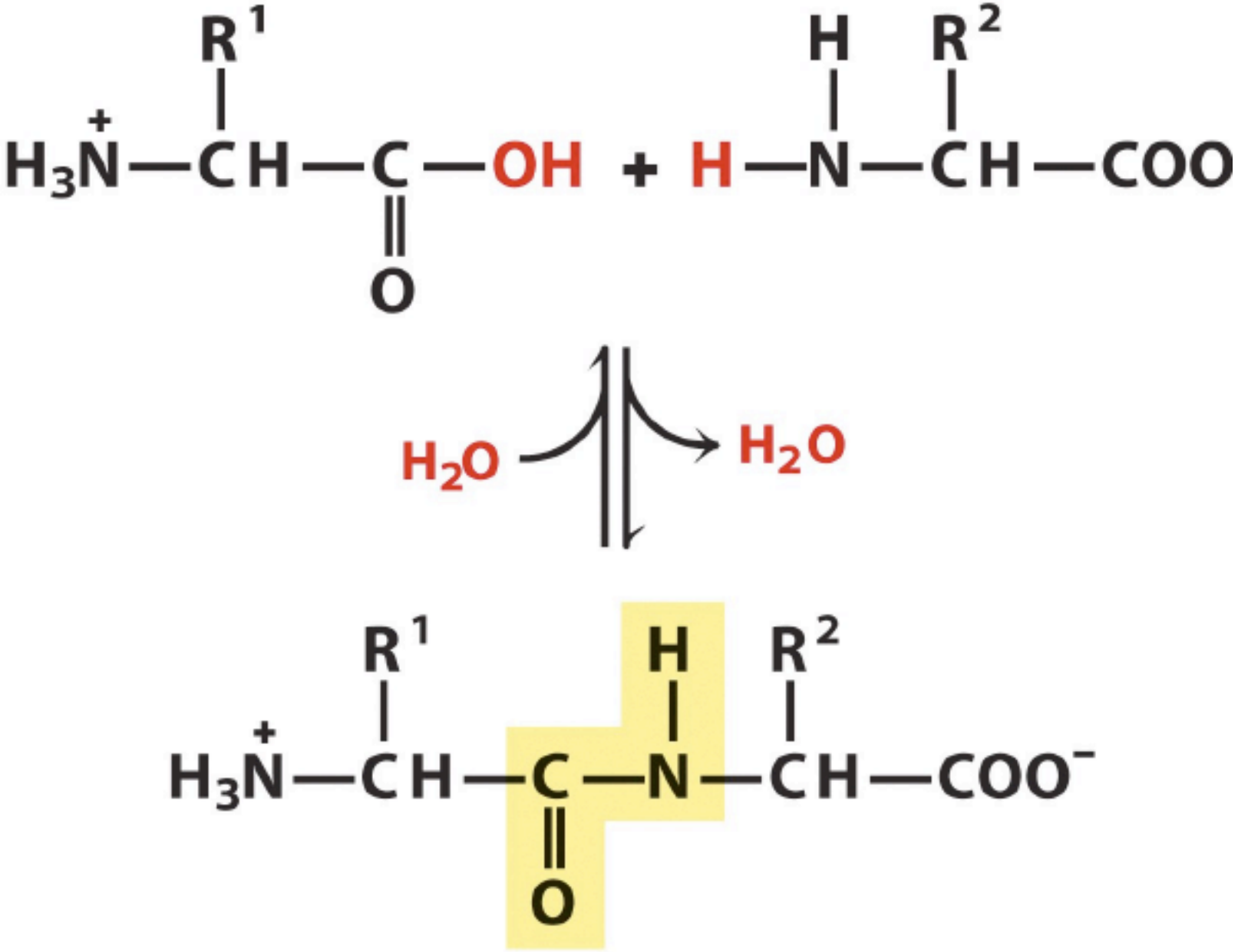


Peptide bonds are planar

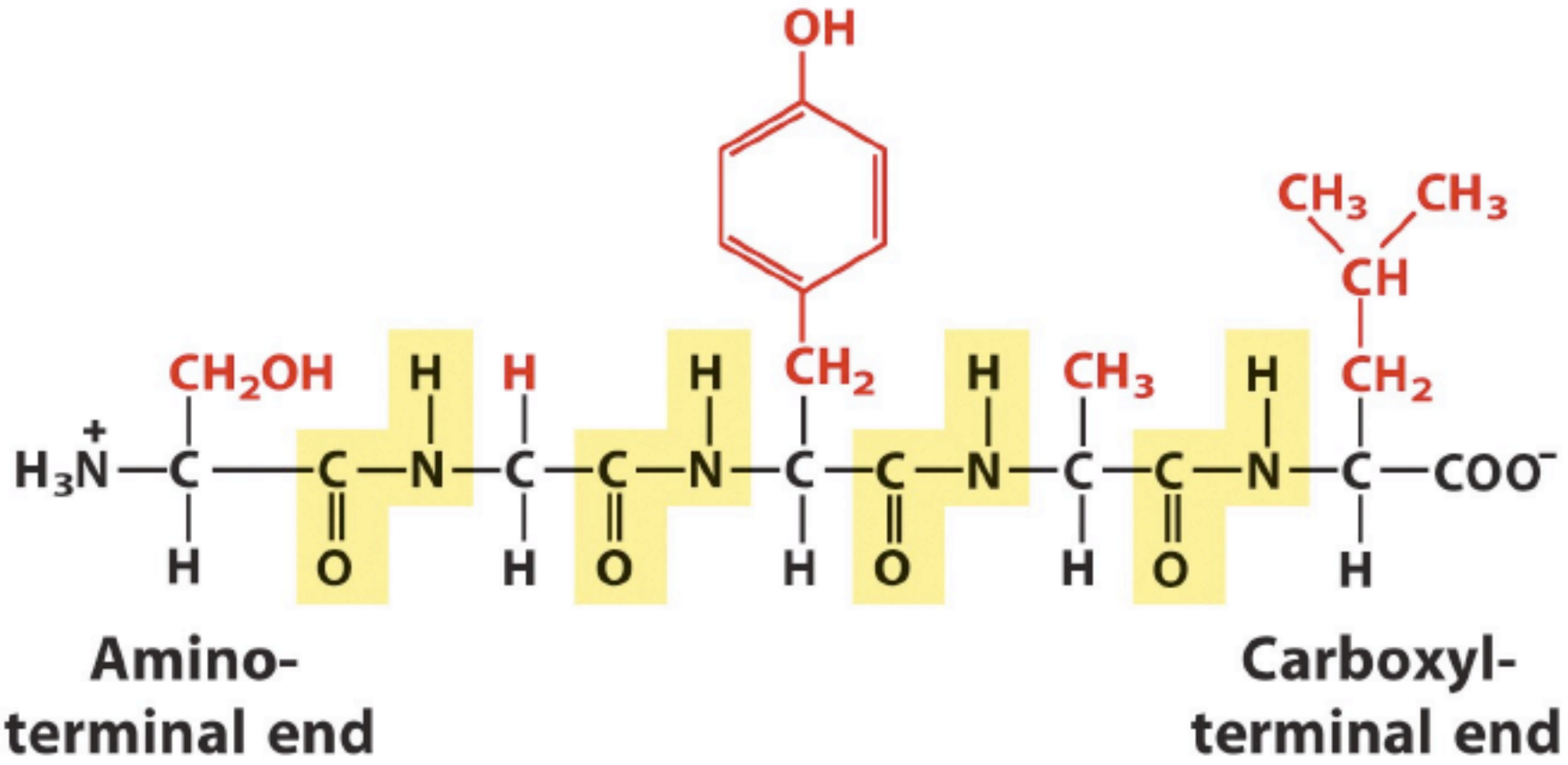


The polypeptide backbone (main chain) refers to atoms that participate in peptide bonds, and can be drawn as a linked set of rigid planar groups.

Two torsion angles, phi and psi,describe backbone conformation.

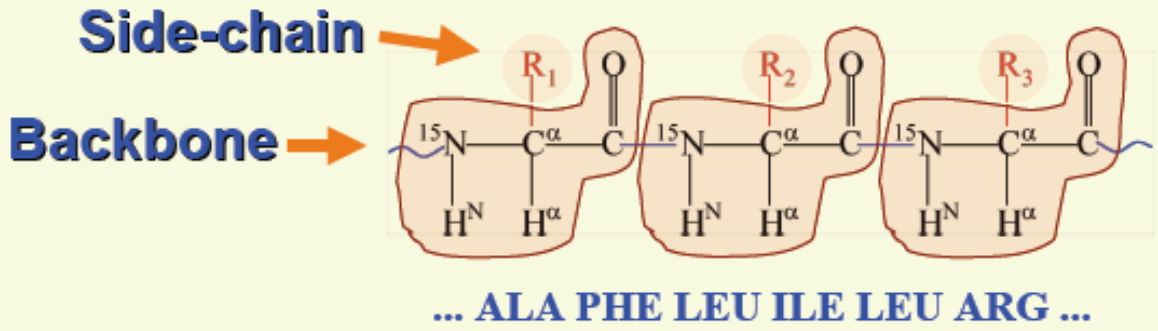

R1
H3N+—CH—C—OH
O
+
H
R2
H—N—CH—COO
H2O
H2O
R1
H3N+—CH—C—N—CH—COO−
O
H
R2

Generally a protein initiates with a Methionine



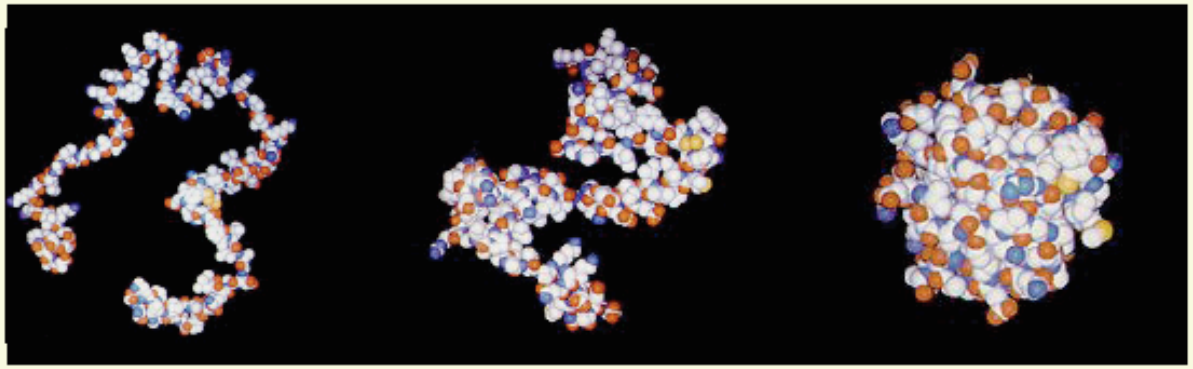
Protein Structure

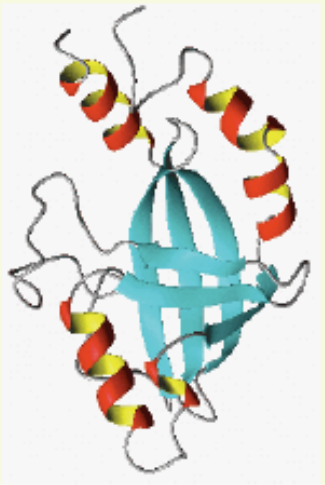
Primary Sequence: Linear String of Amino Acids

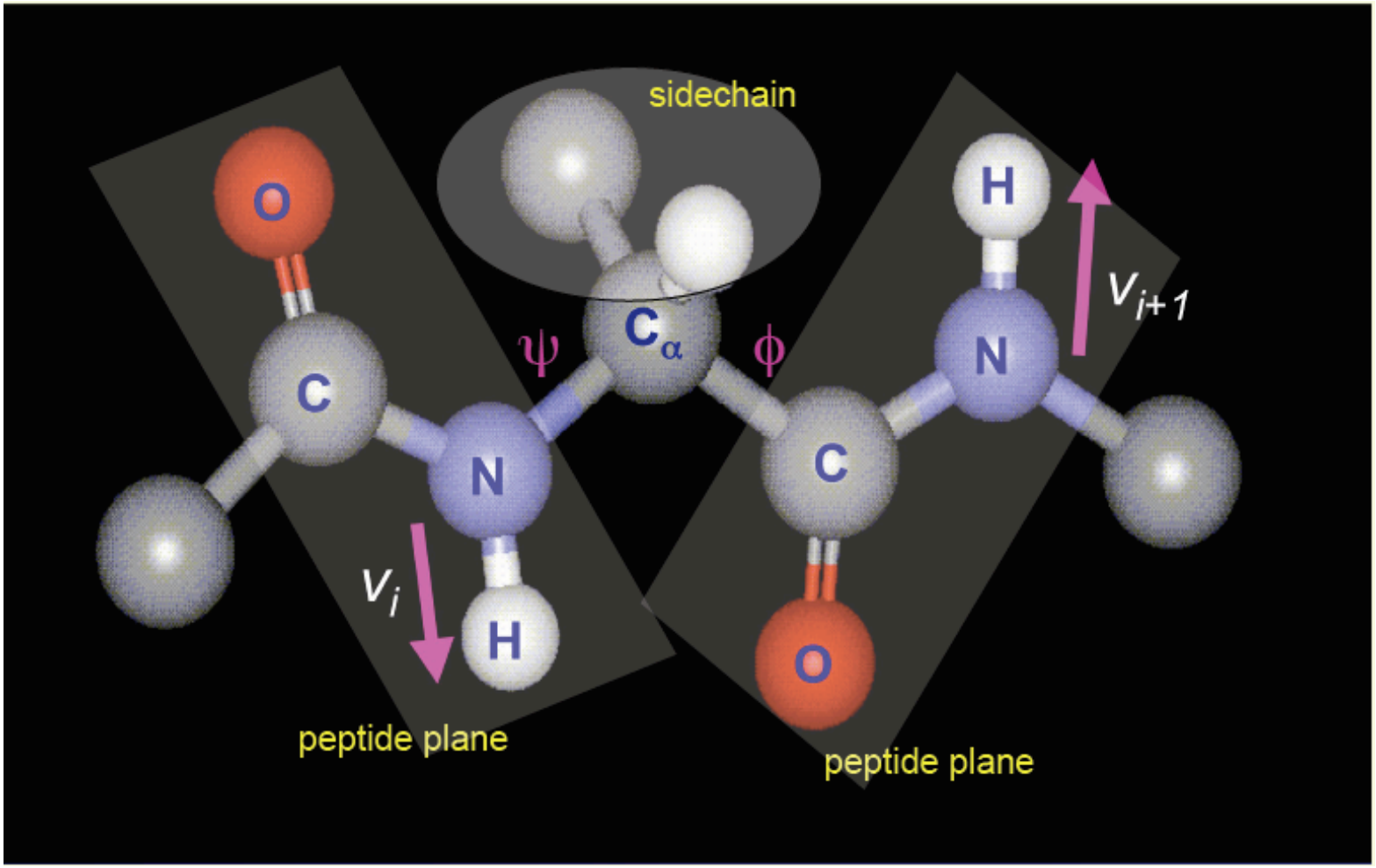


Secondary structure: regular α-helices and β-strands

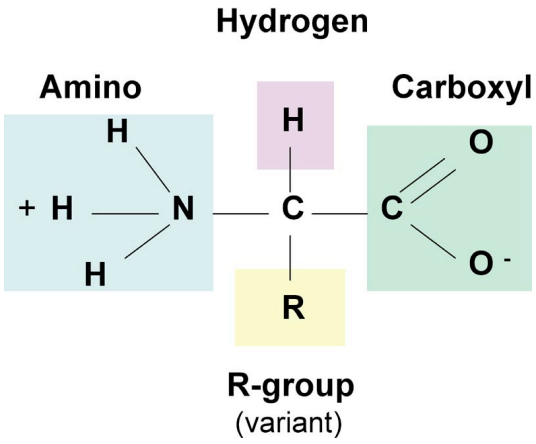
Global Fold

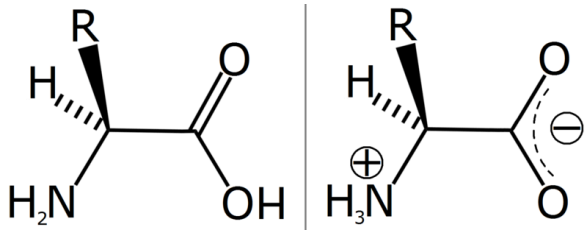


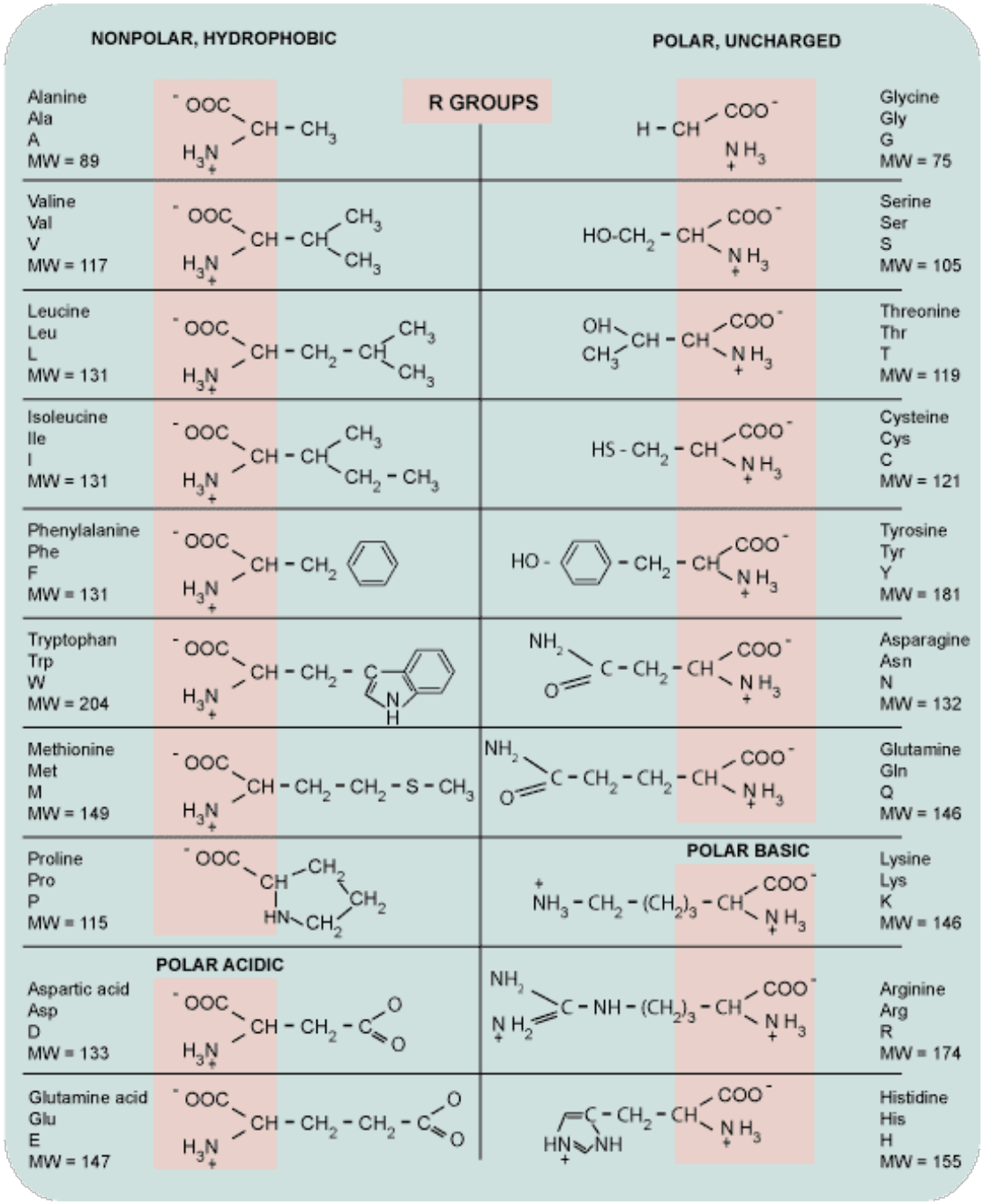



sidechain
O
C
N
H
C_α
ψ
ϕ
C
N
H
O
V_i
V_{i+1}
peptide plane
peptide plane

Amino Acid Structure







Amino acids can be classified by R group

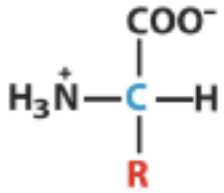


TABLE 3–1 Properties and Conventions Associated with the Common Amino Acids Found in Proteins

Amino acid	Abbreviation/ symbol	M_r	pK_a values: pK_1 (—COOH)	pK_2 ($—NH_3^+$)	pK_R (R group)	pI	Hydropathy index*	Occurrence in proteins (%)†
Nonpolar, aliphatic R groups								
Glycine	Gly G	75	2.34	9.60		5.97	−0.4	7.2
Alanine	Ala A	89	2.34	9.69		6.01	1.8	7.8
Proline	Pro P	115	1.99	10.96		6.48	1.6	5.2
Valine	Val V	117	2.32	9.62		5.97	4.2	6.6
Leucine	Leu L	131	2.36	9.60		5.98	3.8	9.1
Isoleucine	Ile I	131	2.36	9.68		6.02	4.5	5.3
Methionine	Met M	149	2.28	9.21		5.74	1.9	2.3
Aromatic R groups								
Phenylalanine	Phe F	165	1.83	9.13		5.48	2.8	3.9
Tyrosine	Tyr Y	181	2.20	9.11	10.07	5.66	−1.3	3.2
Tryptophan	Trp W	204	2.38	9.39		5.89	−0.9	1.4

*A scale combining hydrophobicity and hydrophilicity of R groups; it can be used to measure the tendency of an amino acid to seek an aqueous environment (− values) or a hydrophobic environment (+ values). See Chapter 11. From Kyte, J. & Doolittle, R.F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157,** 105–132.

†Average occurrence in more than 1,150 proteins. From Doolittle, R.F. (1989) Redundancies in protein sequences. In *Prediction of Protein Structure and the Principles of Protein Conformation* (Fasman, G.D., ed.), pp. 599–623, Plenum Press, New York.

Amino acids can be classified by R group

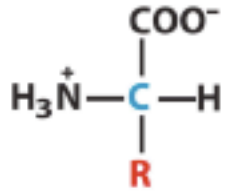


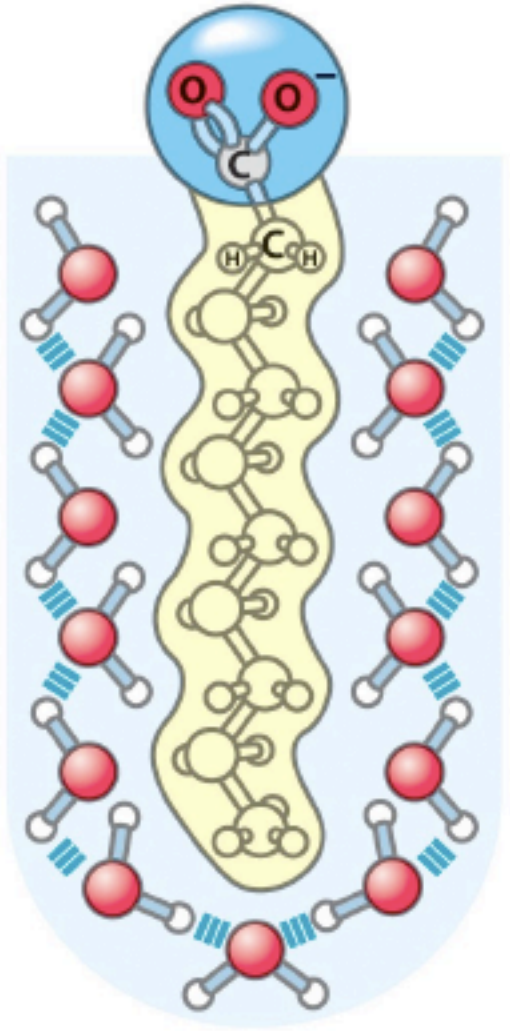
TABLE 3–1 Properties and Conventions Associated with the Common Amino Acids Found in Proteins

Amino acid	Abbreviation/ symbol	M_r	pK_a values: pK_1 (—COOH)	pK_2 ($—NH_3^+$)	pK_R (R group)	pI	Hydropathy index*	Occurrence in proteins (%)†
Polar, uncharged R groups								
Serine	Ser S	105	2.21	9.15		5.68	−0.8	6.8
Threonine	Thr T	119	2.11	9.62		5.87	−0.7	5.9
Cysteine	Cys C	121	1.96	10.28	8.18	5.07	2.5	1.9
Asparagine	Asn N	132	2.02	8.80		5.41	−3.5	4.3
Glutamine	Gln Q	146	2.17	9.13		5.65	−3.5	4.2
Positively charged R groups								
Lysine	Lys K	146	2.18	8.95	10.53	9.74	−3.9	5.9
Histidine	His H	155	1.82	9.17	6.00	7.59	−3.2	2.3
Arginine	Arg R	174	2.17	9.04	12.48	10.76	−4.5	5.1
Negatively charged R groups								
Aspartate	Asp D	133	1.88	9.60	3.65	2.77	−3.5	5.3
Glutamate	Glu E	147	2.19	9.67	4.25	3.22	−3.5	6.3

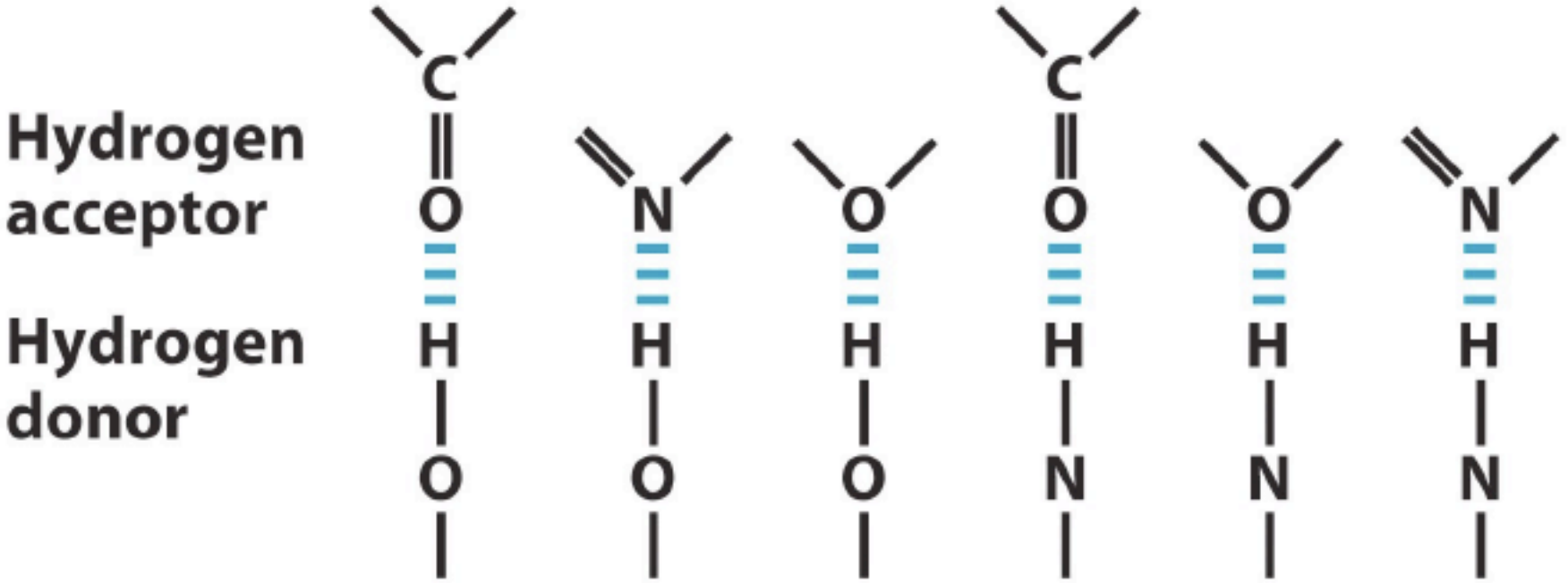
*A scale combining hydrophobicity and hydrophilicity of R groups; it can be used to measure the tendency of an amino acid to seek an aqueous environment (− values) or a hydrophobic environment (+ values). See Chapter 11. From Kyte, J. & Doolittle, R.F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157,** 105–132.

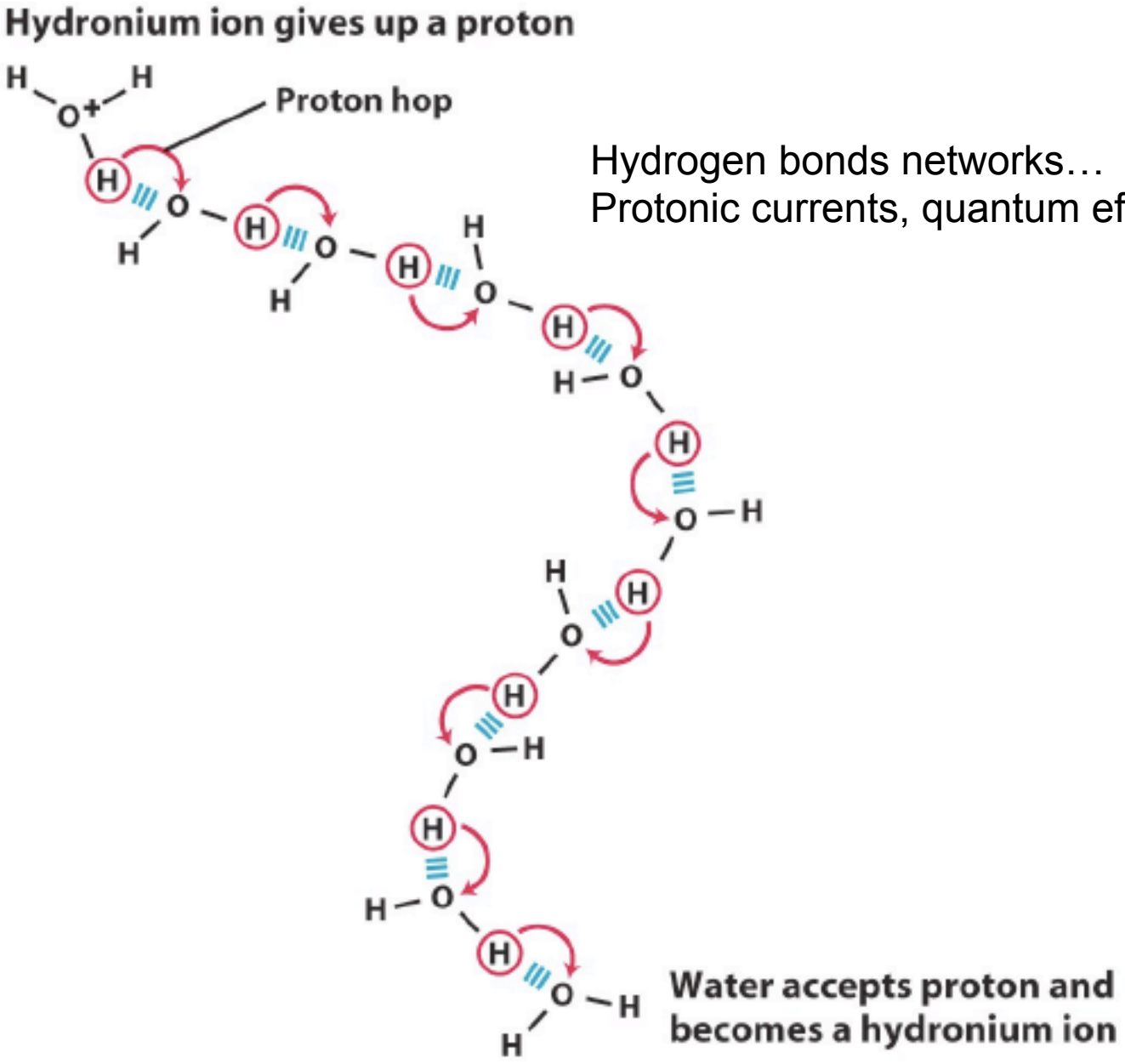
†Average occurrence in more than 1,150 proteins. From Doolittle, R.F. (1989) Redundancies in protein sequences. In *Prediction of Protein Structure and the Principles of Protein Conformation* (Fasman, G.D., ed.), pp. 599–623, Plenum Press, New York.

Water plays an important role in the structure and behavior of biomolecules



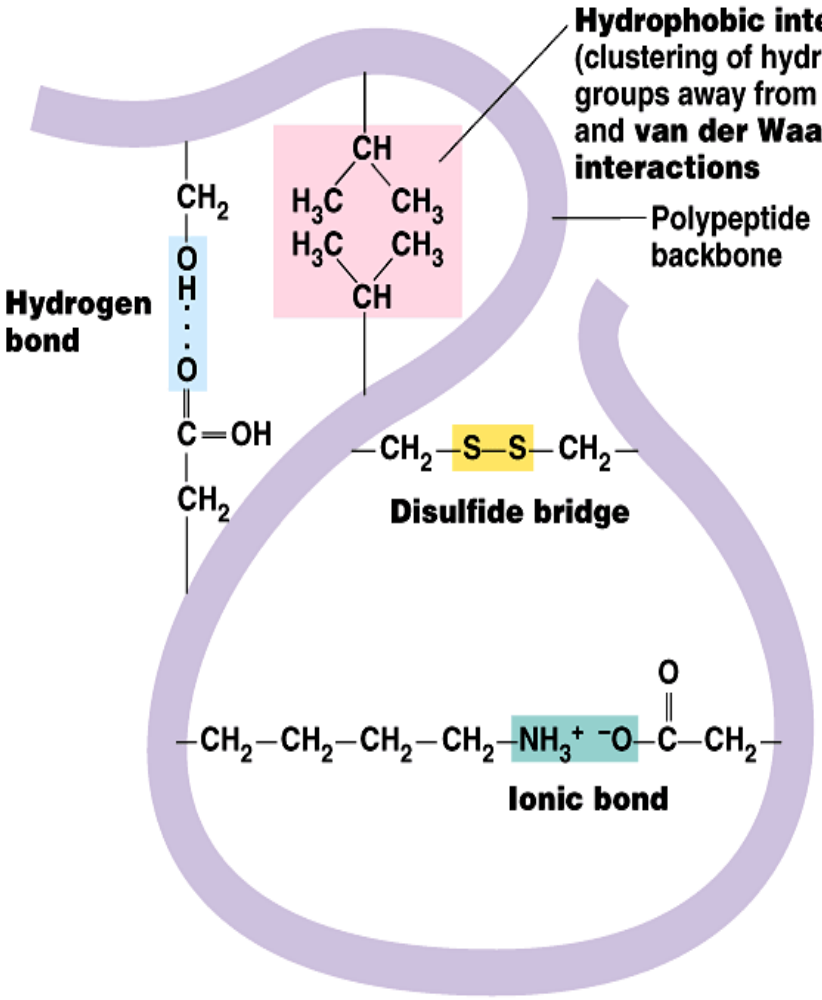
Hydrogen bonds occur in water and between biomolecules





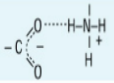
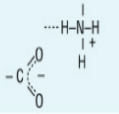
Hydrogen bonds networks…
Protonic currents, quantum effects

Internal Protein Interactions



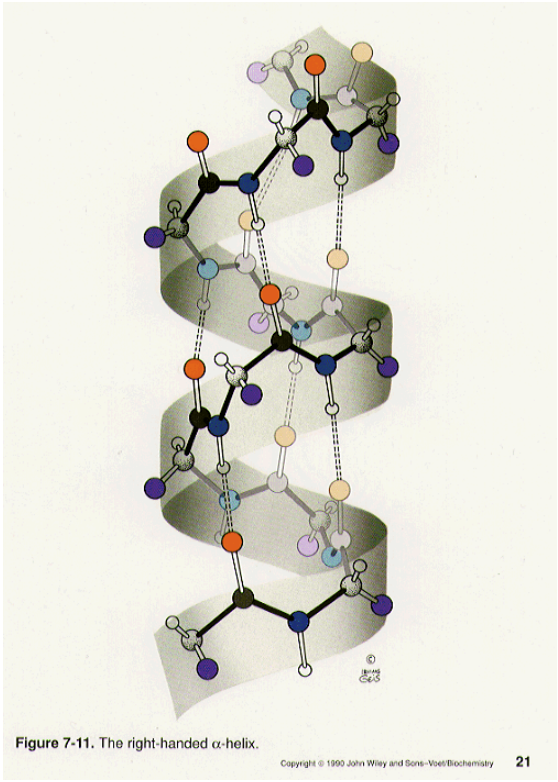
Copyright © Pearson Education, Inc., publishing as Benjamin Cummings.

- The thermal stability of a protein is about 60 kJ/mol

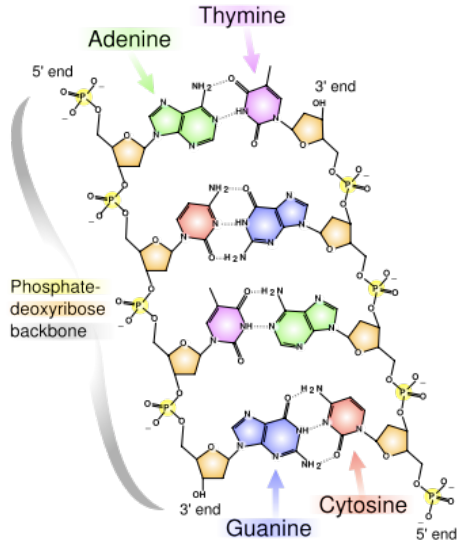
Interaction	Example	Distance dependence	Typical distance	Free energy (bond dissociation enthalpies for the covalent bonds)
Covalent bond	$-C_\alpha-C-$	-	1.5 Å	356 kJ/mole (610 kJ/mole for a C=C bond)
Disulfide bond	–Cys–S–S–Cys–	-	2.2 Å	167 kJ/mole
Salt bridge		Donor (here N), and acceptor (here O) atoms <3.5 Å	2.8 Å	12.5–17 kJ/mole; may be as high as 30 kJ/mole for fully or partially buried salt bridges (see text), less if the salt bridge is external
Hydrogen bond		Donor (here N), and acceptor (here O) atoms <3.5 Å	3.0 Å	2–6 kJ/mole in water; 12.5–21 kJ/mole if either donor or acceptor is charged
Long-range electrostatic interaction		Depends on dielectric constant of medium. Screened by water. 1/r dependence	Variable	Depends on distance and environment. Can be very strong in nonpolar region but very weak in water
Van der Waals interaction		Short range. Falls off rapidly beyond 4 Å separation. $1/r^6$ dependence	3.5 Å	4 kJ/mole (4–17 in protein interior) depending on the size of the group (for comparison, the average thermal energy of molecules at room temperature is 2.5 kJ/mole)

- Cooperative interactions: one bond favors the formation of another bond etc..

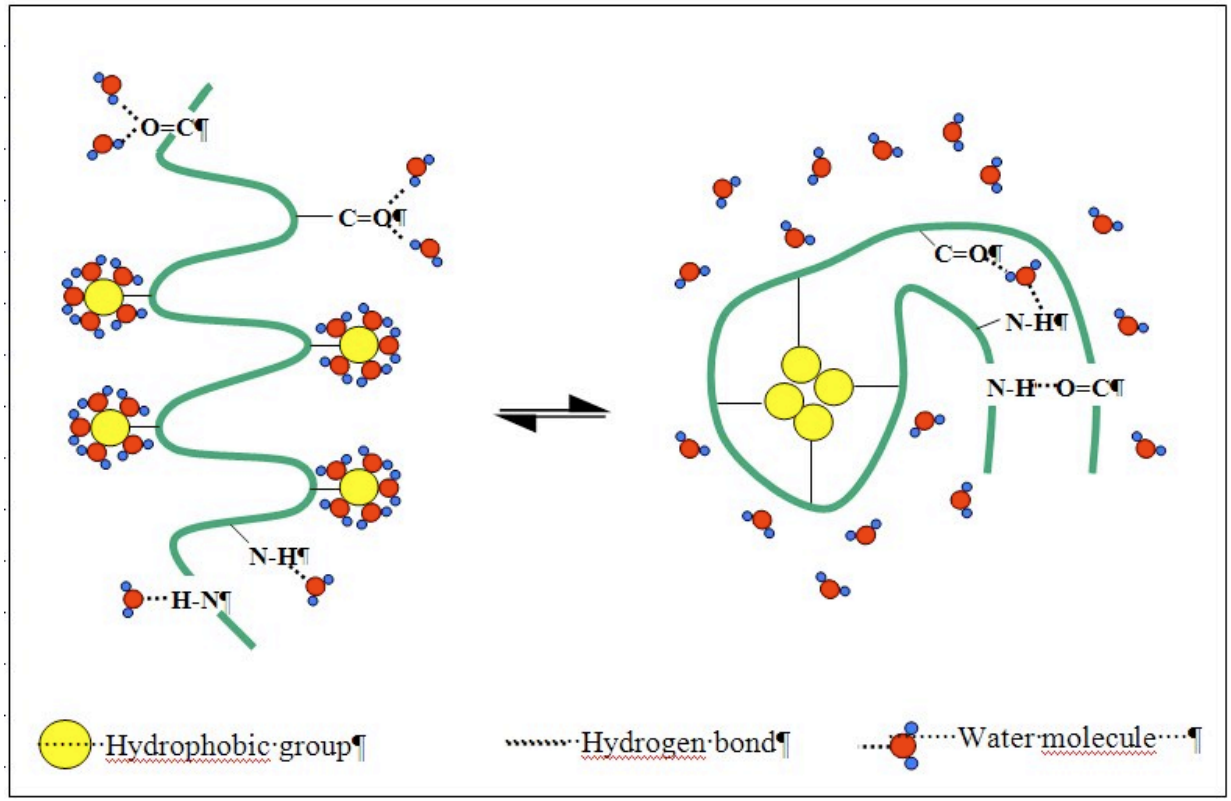
Hydrogen-BONDS

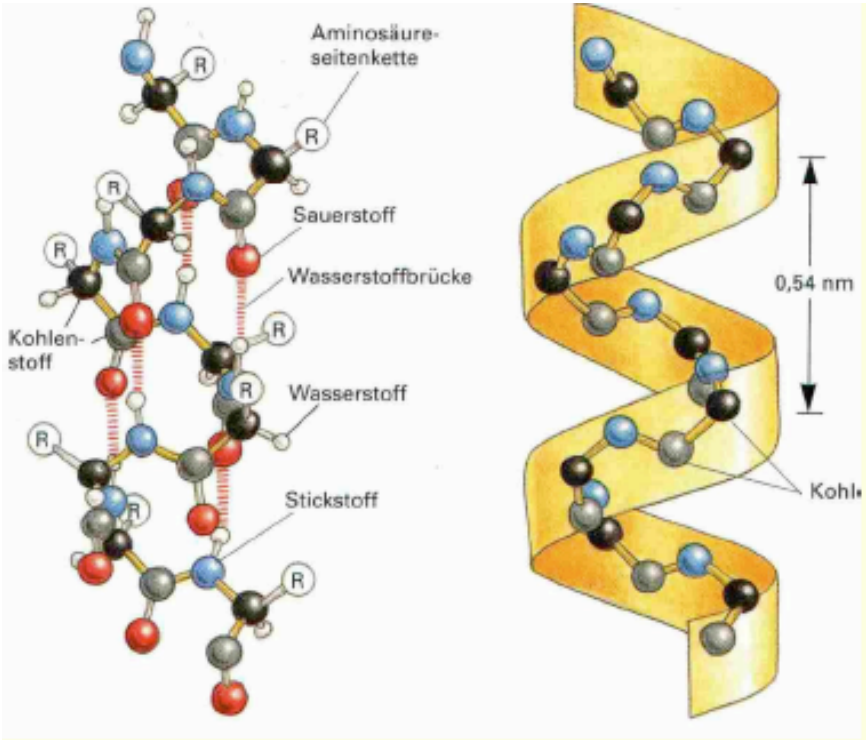


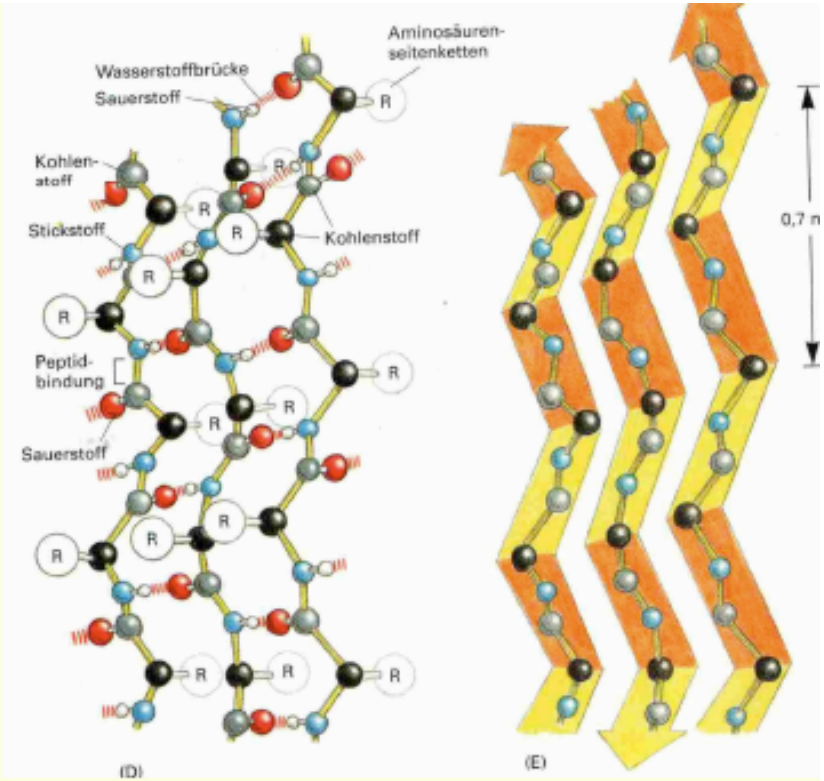
Proteins



DNA





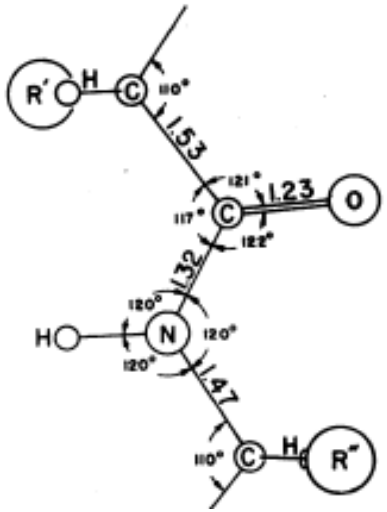


α-helix

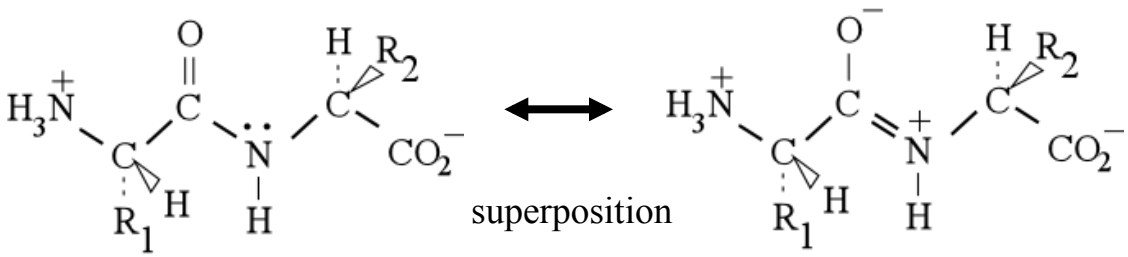
β-sheet

Pauling and Corey (1951) predicted that amino acid sequences could form regular geometric shapes via hydrogen bonding.

Determinants of helical structure



Distances and angles
Between atoms

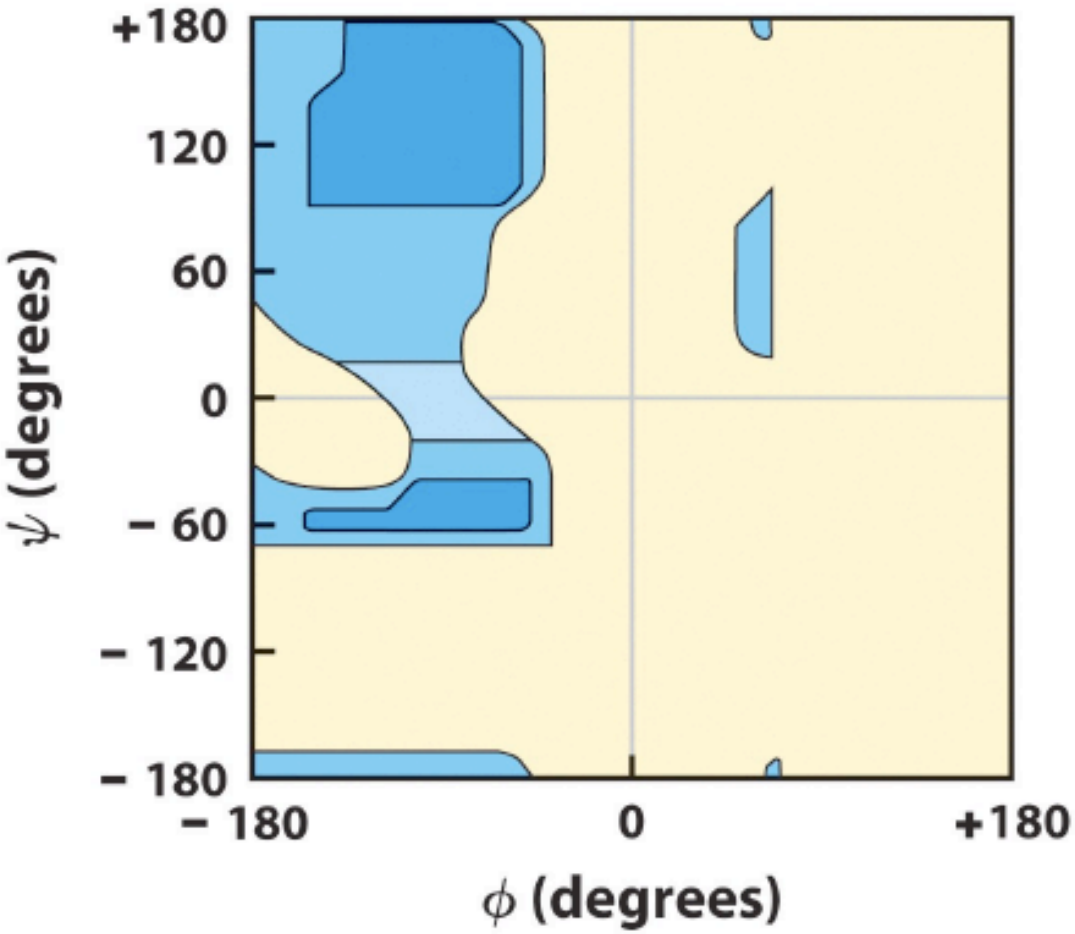


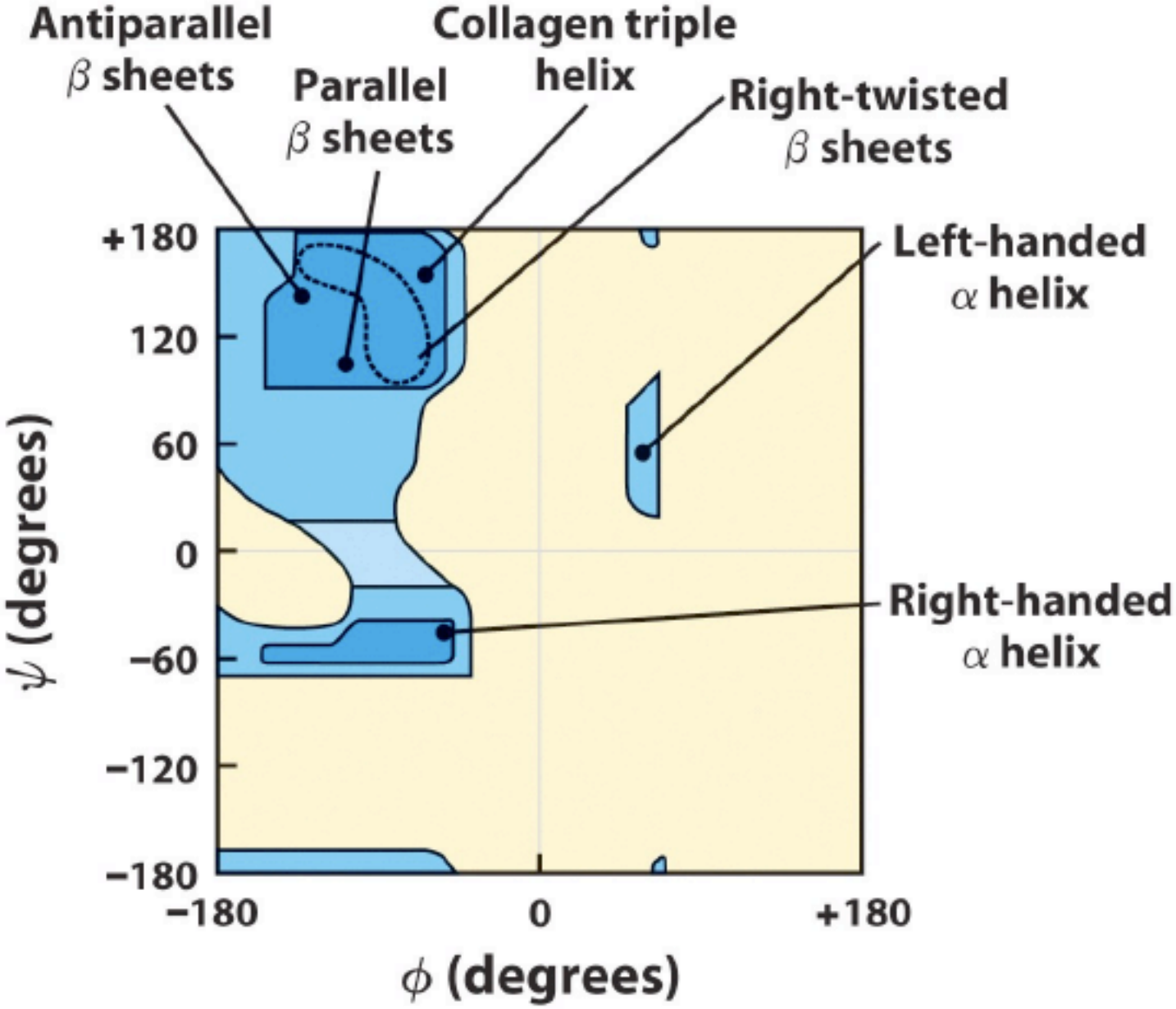
Resonant partial double bond character of peptide bond induces planar arrangement of atoms

N—H · · · O

All hydrogen bonds should be satisfied, i.e. distance N-O of about 2.7Å and angle between C = O and H – N less then ~30°

Ramachandran plot

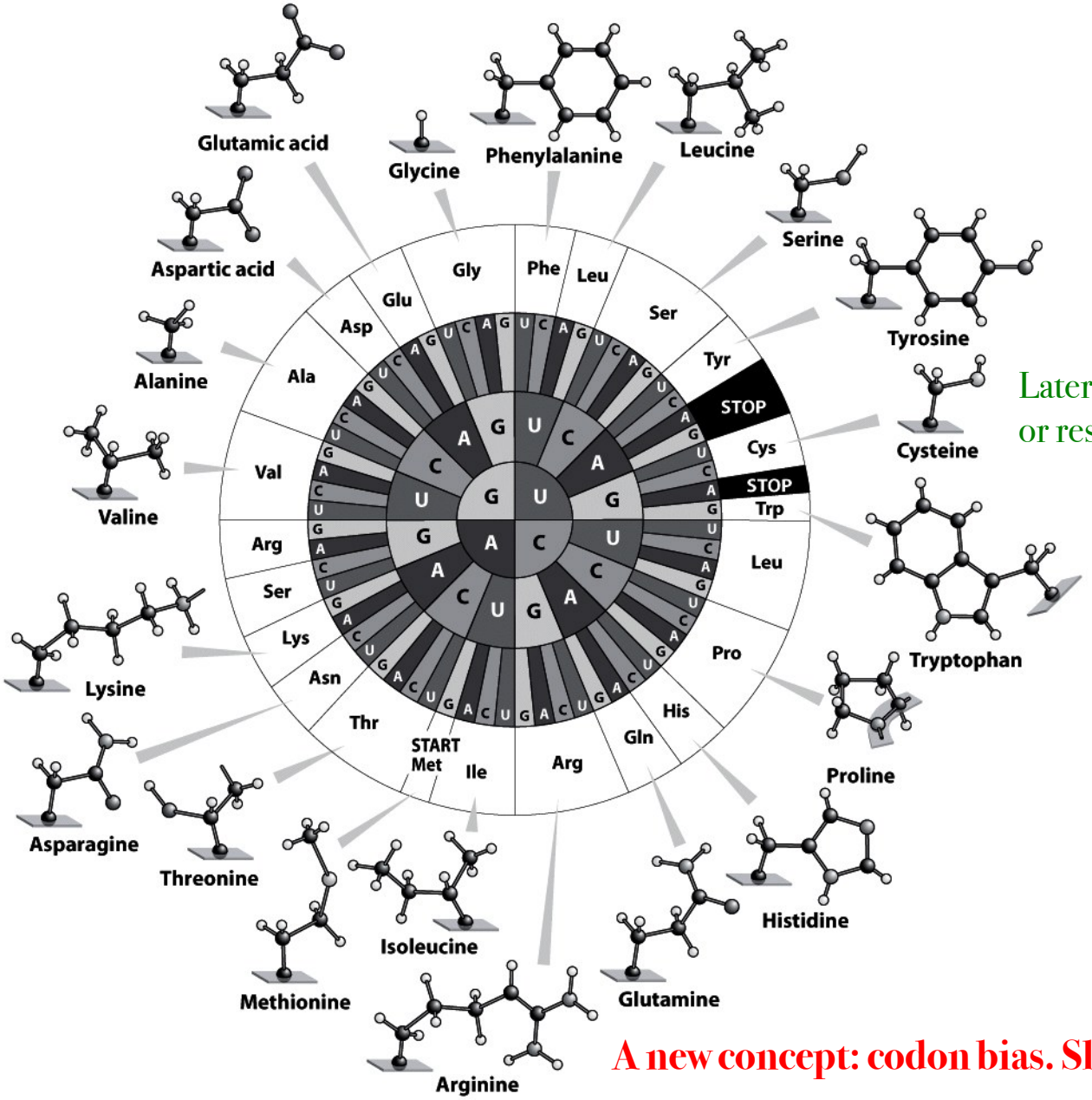



Antiparallel β sheets
Parallel β sheets
Collagen triple helix
Right-twisted β sheets
Left-handed α helix
Right-handed α helix
+180
120
60
0
−60
−120
−180
ψ (degrees)
−180
0
+180
ϕ (degrees)

Structural Domains

- independent functional subproteins: thermal stability, correlated dynamics
- Example: Pawson's domains

The genetic code: codons code for amino acids



Lateral chains: or residues

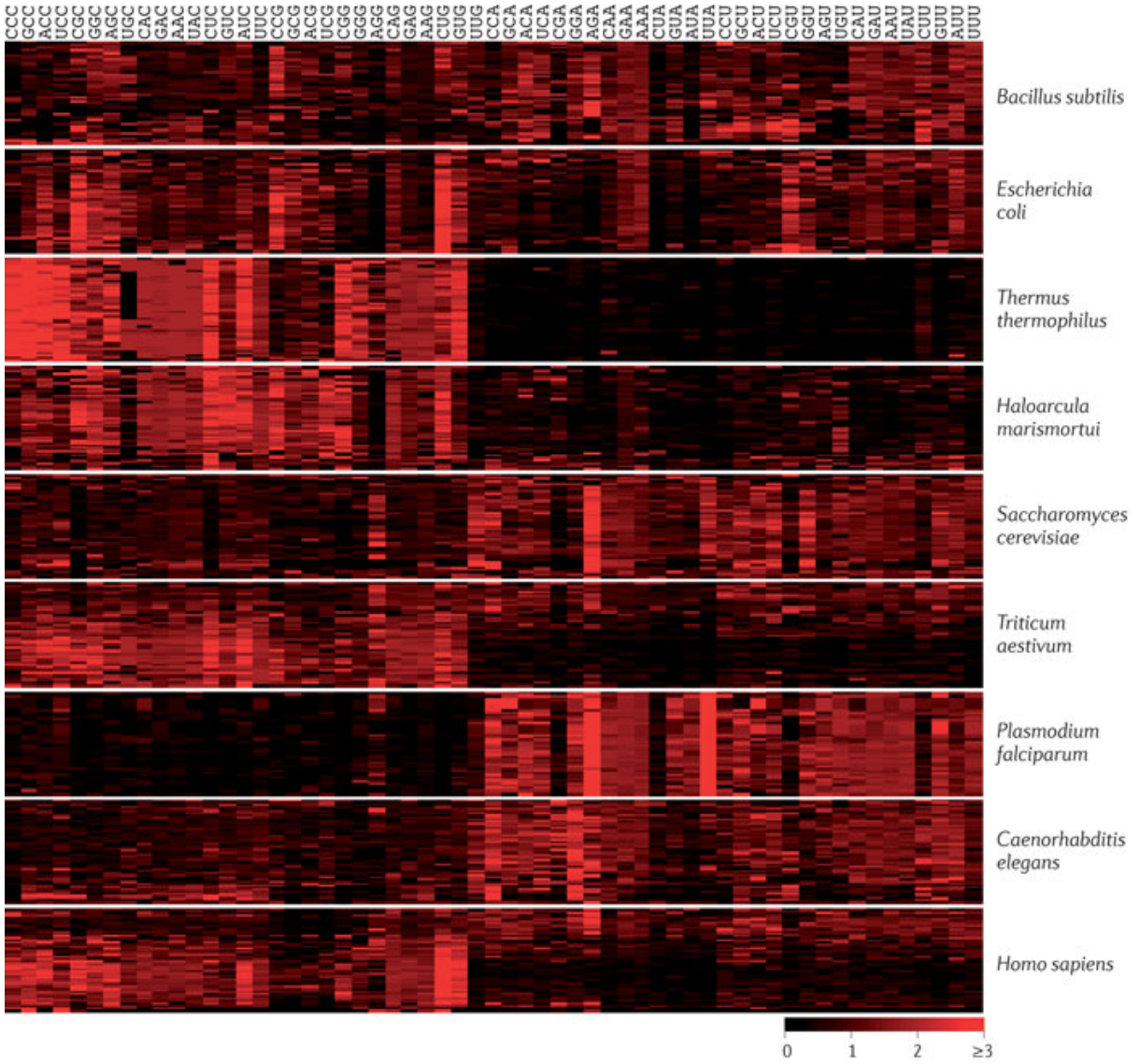
A new concept: codon bias. Slow/fast codon

Figure 1.4 Physical Biology of the Cell (© Garland Science 2009)

Possible mechanisms of translational regulation

- optimality of ribosomal attachment site
- mRNA secondary structure
- codon usage
- t-RNA BIOLOGY

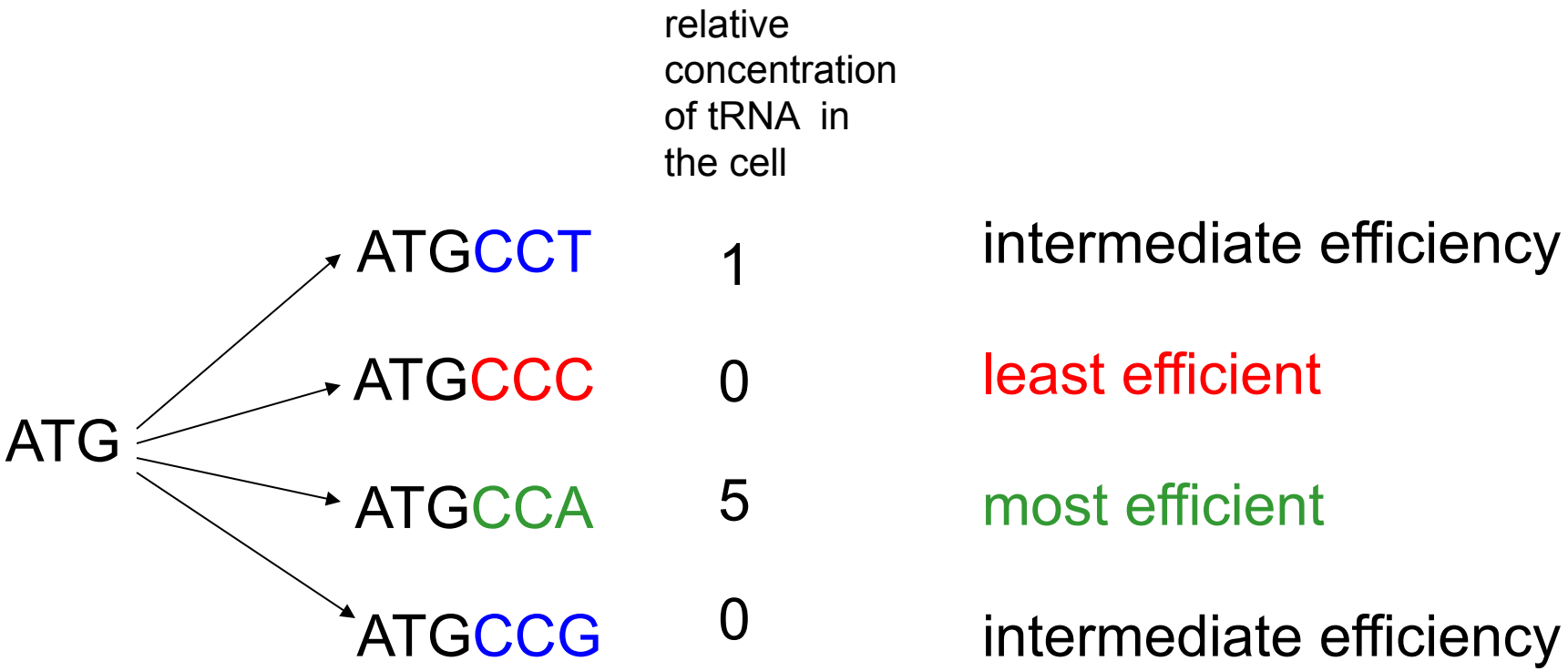
The codon bias in genomes



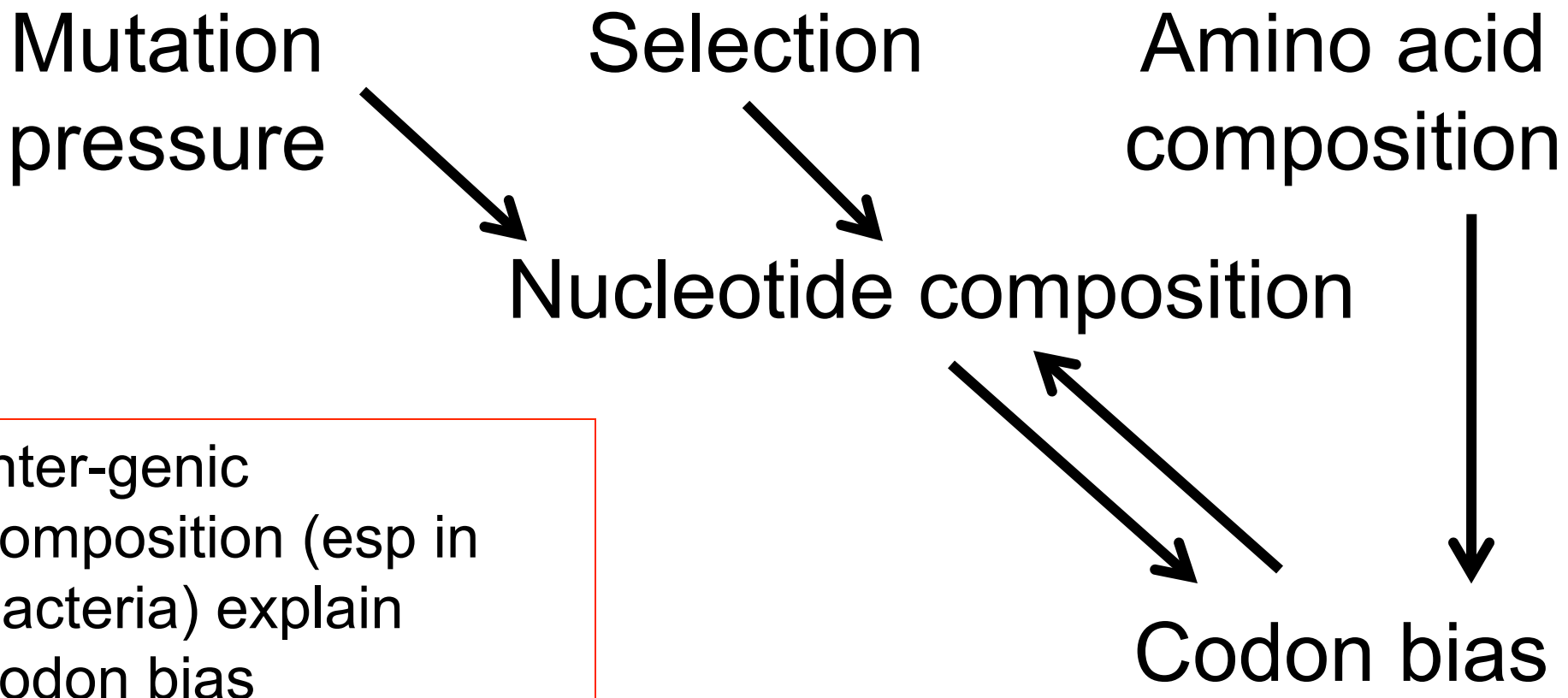
Nature Reviews | Genetics

The same protein can be encoded in many ways…

amino acid sequence: MPKSNFRFGE



The effect of (or on?) GC content

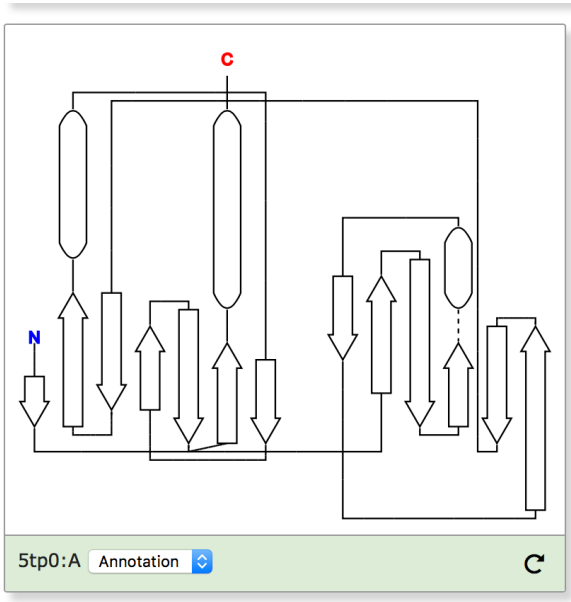


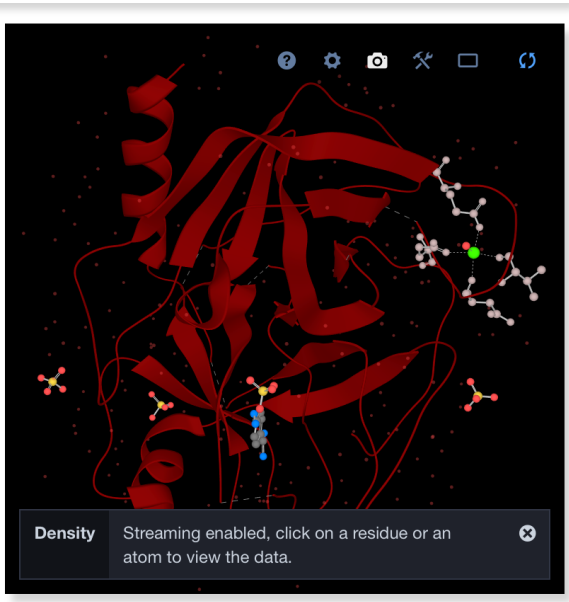
Inter-genic composition (esp in bacteria) explain codon bias

Tutorial

FASTA Sequence

```
>pdb|5tp0|A
IVGGYTCEENSLPYQVSLNSGSHFCGGSLISEQWVVSAAHCYKTRIQVRLGEHNIKVLEGNEQFINAA
KIIRHPKYNRDTLDNDIMLIKLSSPAVINARVSTISLPTAPPAAGTECLISGWGNTLSFGADYPDELK
CLDAPVLTQAECKASYPGKITNSMFCVGFLEGGKDSCQRDSGGPVVCNGQLQGVVSWGHGCAWKNRPG
VYTKVYNYVDWIKDTIAANS
```





Coordinate systems

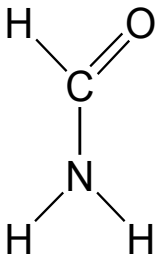
At least two ways of specifying atomic coordinates:

Cartesian coordinates from the Protein Data Bank .pdb files.

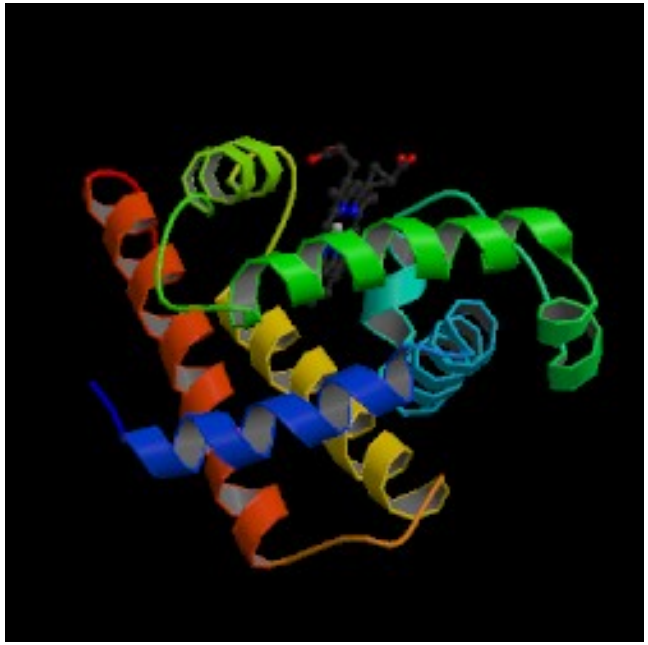
X,Y,Z with respect to a laboratory system, modulo a ROTO-TRANSLATION

N atoms ⇒ 3N cartesian coordinates

```
REMARK   4
REMARK   4 FORM COMPLIES WITH FORMAT V. 2.0, 26-MAR-2003
ATOM      1  C1  ALHD    1       3.450   0.774  -9.239  1.00  0.00           C
ATOM      2  O2  ALHD    1       4.639   0.791  -8.925  1.00  0.00           O
ATOM      3  H1  ALHD    1       2.972  -0.166  -9.513  1.00  0.00           H
ATOM      4  N1  AMDE    1B      2.756   1.892  -9.246  1.00  0.00           N
ATOM      5 1H1  AMDE    1B      1.760   1.878  -9.510  1.00  0.00           H
ATOM      6 2H1  AMDE    1B      3.207   2.781  -8.987  1.00  0.00           H
END
```



Proof came 7 years later…



John Cowdery Kendrew
The Nobel Prize in Chemistry 1962

Kendrew, J. C., Bodo, G., Dintzis, H. M. Parrish, R. G., Wyckoff, H., and Phillips, D. C. **A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-ray Analysis**. Nature, 181, 662 (1958).

Pauling and Corey papers series – PNAS April 1951

1. Pauling, L., Corey, R.B. and Branson H. R. **The Structure of Proteins: Two Hydrogen-Bonded Helical Configurations of the Polypeptide Chain.** PNAS, **37**, 205-211, (1951).
2. Pauling, L. & Corey, R. B. **Atomic Coordinates and Structure Factors for Two Helical Configurations of Polypeptide Chains**. PNAS, **37**, 235-240, (1951).
3. Pauling, L. & Corey, R. B. **The Structure of Synthetic Polypeptides**. PNAS, **37**, 241-250, (1951).
4. Pauling, L. & Corey, R. B. **The Pleated Sheet, A New Layer Configuration of Polypeptide Chains**. PNAS, **37**, 251-256, (1951).
5. Pauling, L. & Corey, R. B. **The Structure of Feather Rachis Keratin.** PNAS, **37**, 256-261, (1951).
6. Pauling, L. & Corey, R. B. **The Structure of Hair, Muscle, and Related Proteins**. PNAS, **37**, 261-271, (1951).
7. Pauling, L. & Corey, R. B. **The Structure of Fibrous Proteins of the Collagen-Gelatin Group**. PNAS, **37**, 272-281, (1951).
8. Pauling, L. & Corey, R. B. **The Polypeptide-Chain Configuration in Hemoglobin and Other Globular Proteins**. PNAS, **37**, 282-285, (1951).

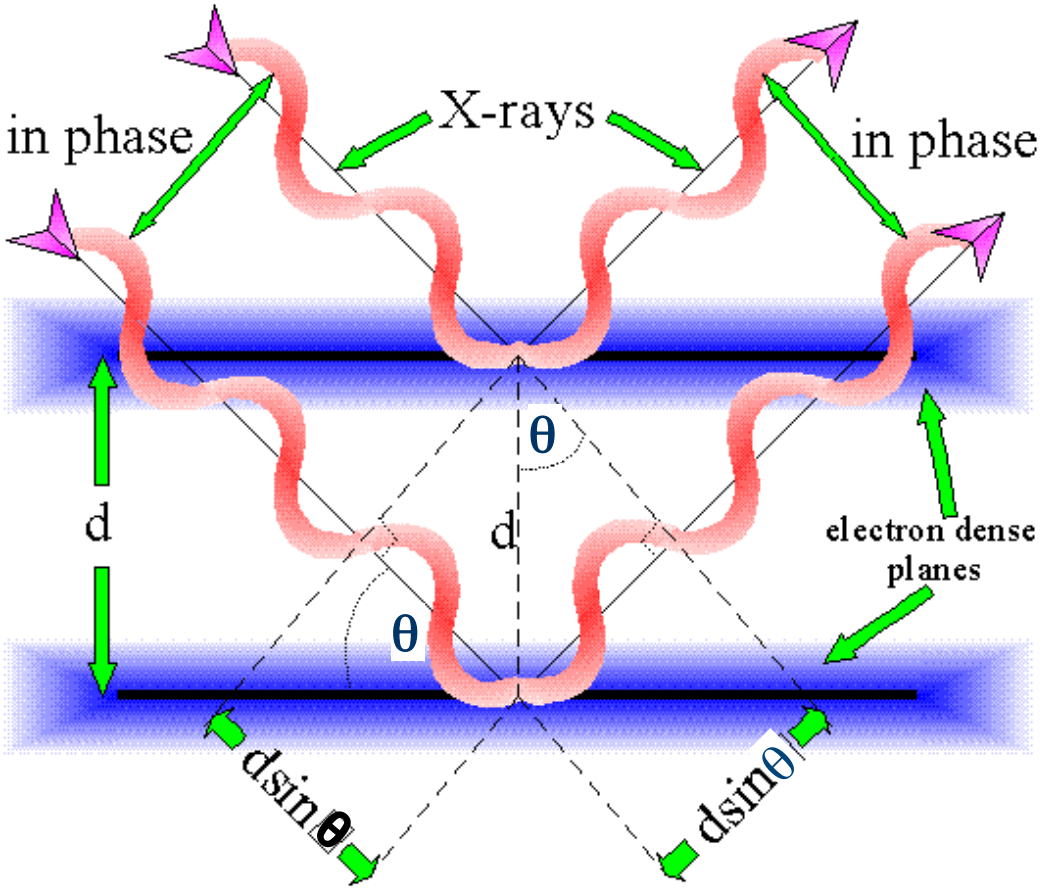
X-Rays crystallography – the tool of structural biology

- X-ray are (weakly) scattered by electrons
- Diffraction from a single molecule is weak
  so use a crystal
- Why X-ray?
  Wavelength of visible light: ~500 nm
  Bond lengths in proteins: ~0.15 nm
  Typical X-ray wavelength: ~0.15 nm
- + Multiple copies of the molecule increases diffraction
- - Crystalline structure imposes constraints on diffraction pattern

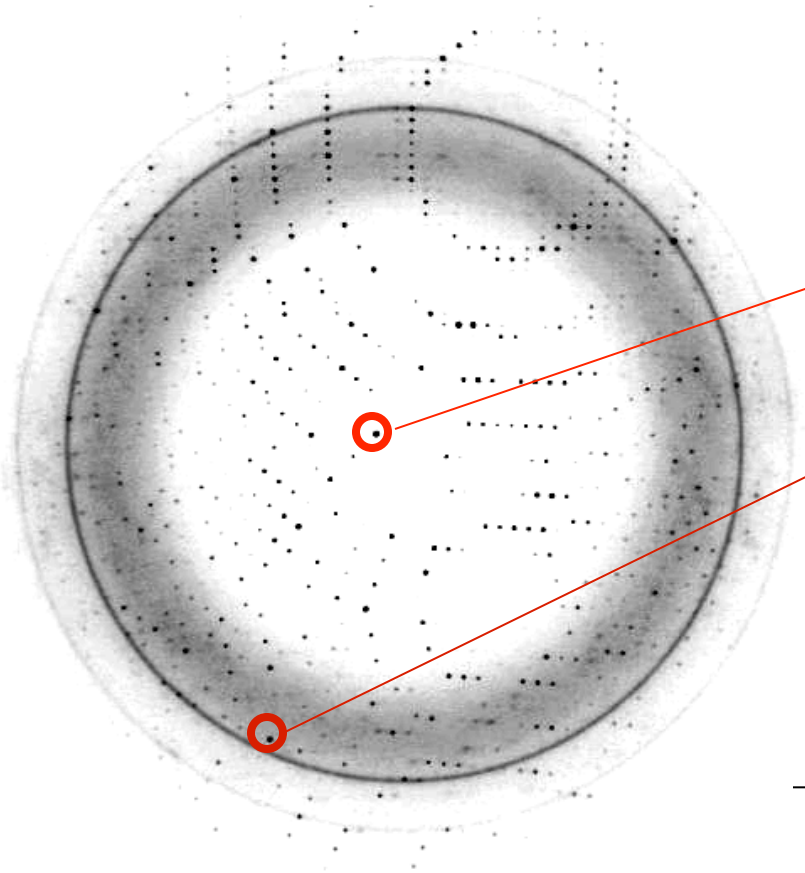
- OTHER TECHNIQUES: **NMR, CRYO ELECTRON MICROSCOPY**

Diffraction occurs at particular angles

- Diffraction spots are the result of <u>constructive</u> interference from multiple scatterers satisfying Bragg's Law: $\lambda = 2\,d\,\sin\theta$



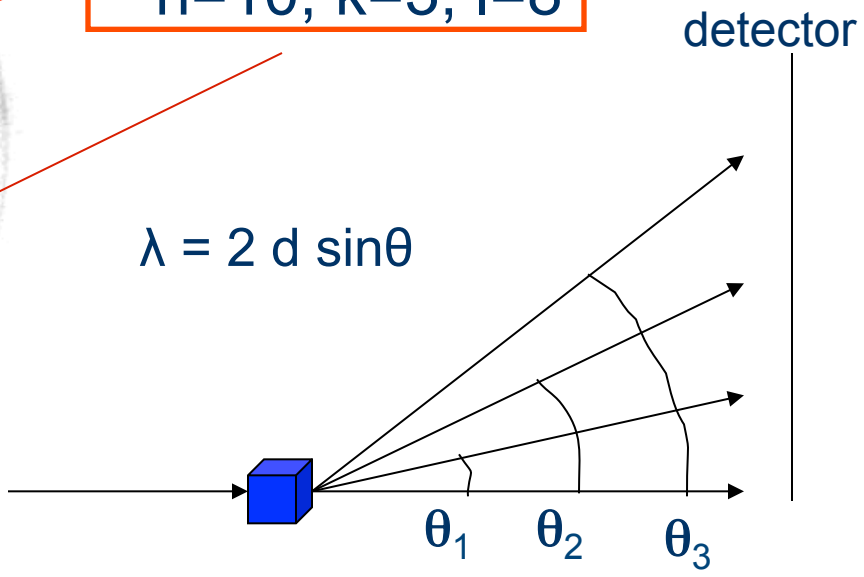
Each Spot Represents a Unique Set of Bragg Planes



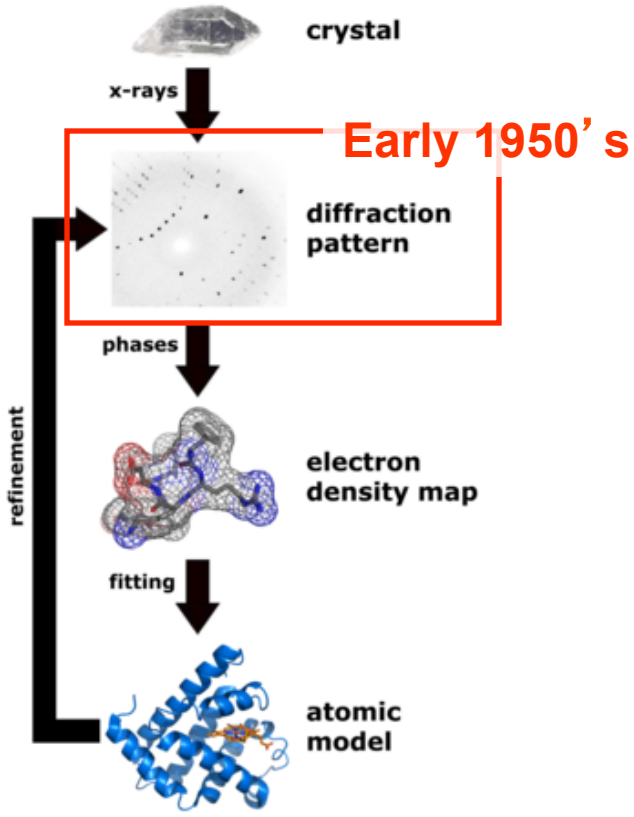
Points in k-space
(Fourier Space)

h=2, k=1, l=3

h=10, k=3, l=8



Modern X-Ray Crystallography



Need good crystals for better resolution, which is difficult in proteins (need right conditions) and sometimes nearly impossible (e.g. membranal proteins)

High resolution details are faint – requires good experimental apparatus

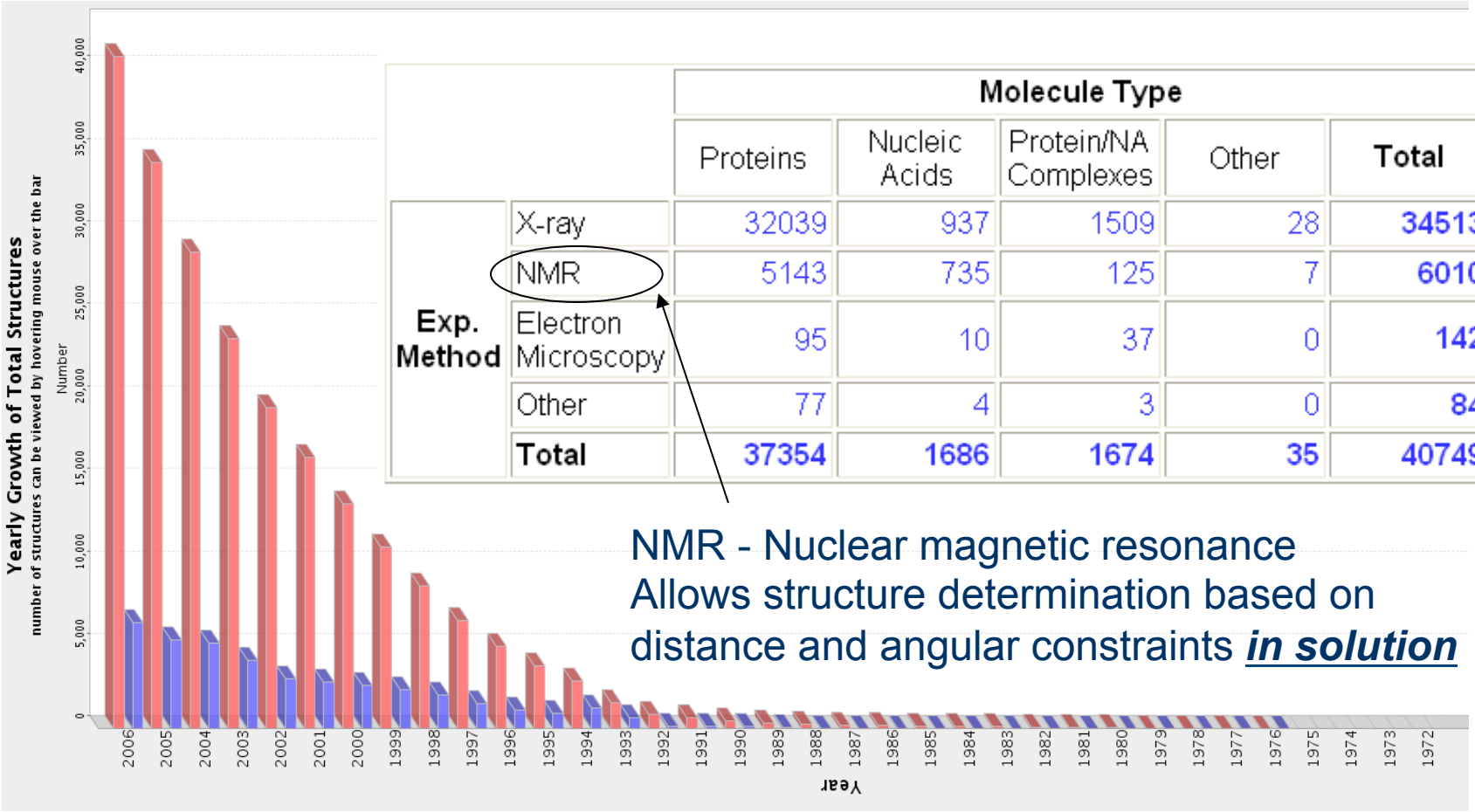
Recorded intensity give only the magnitude but not the phase of the complex "form factor"

Error in density map lead to un-realistic atom assignment, requiring iterative refinement process

The Protein Data Bank (www.pdb.org)

The PDB contains over 40,000 structures (as of December 2006)

RECENTLY CRYO_EM JOINED!



		Molecule Type				
		Proteins	Nucleic Acids	Protein/NA Complexes	Other	**Total**
Exp. Method	X-ray	32039	937	1509	28	**34513**
	NMR	5143	735	125	7	**6010**
	Electron Microscopy	95	10	37	0	**142**
	Other	77	4	3	0	**84**
	Total	**37354**	**1686**	**1674**	**35**	**40749**

NMR - Nuclear magnetic resonance
Allows structure determination based on distance and angular constraints ***in solution***

RCSB PDB PROTEIN DATA BANK

An Information Portal to
129367 Biological
Macromolecular Structures

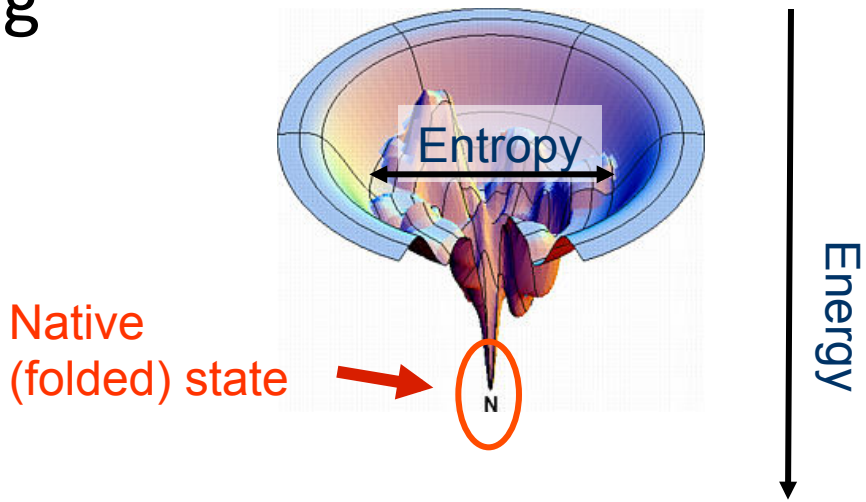
PDB Current Holdings Breakdown

Exp.Method	Proteins	Nucleic Acids	Protein/NA Complexes	Other	Total
X-RAY	108448	1826	5512	4	115
NMR	10341	1200	241	8	11
ELECTRON MICROSCOPY	1061	30	380	0	1
HYBRID	99	3	2	1	
other	188	4	6	13	
Total	120137	3063	6141	26	129

(Click on any number to retrieve the results from that category.)
105563 structures in the PDB have a structure factor file.
9124 structures in the PDB have an NMR restraint file.
2882 structures in the PDB have a chemical shifts file.
1466 structures in the PDB have a 3DEM map file.

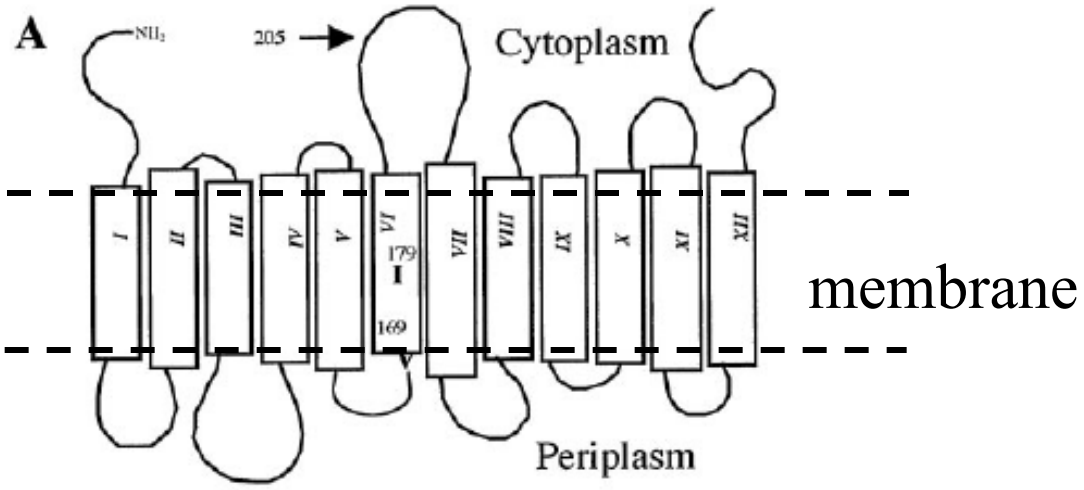
The Protein folding problem…

- Late **1980's** - Wolynes et al. present the "Energy Landscape" or "Folding Funnel" model for protein folding



- **2006** – There is still no precise understanding how proteins fold fast (up to µsec!), reliably and accurately to their native structure

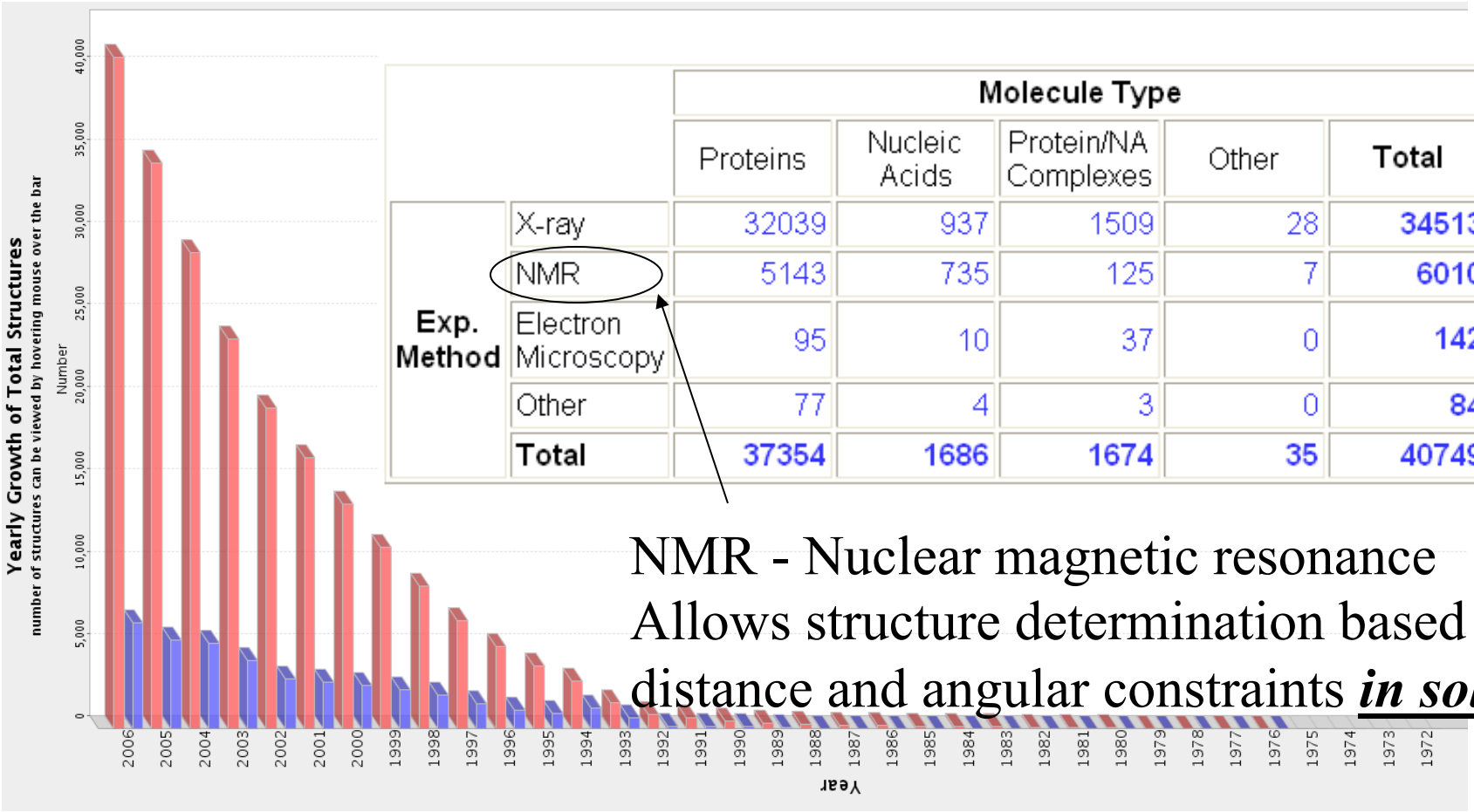
Alpha-helices appear a lot in trans-membranal proteins



E.g. Lactose permease (LacY)

The Protein Data Bank (www.pdb.org)

The PDB contains over 40,000 structures (as of December 2006)



Exp. Method	Molecule Type: Proteins	Nucleic Acids	Protein/NA Complexes	Other	Total
X-ray	32039	937	1509	28	**34513**
NMR	5143	735	125	7	**6010**
Electron Microscopy	95	10	37	0	**142**
Other	77	4	3	0	**84**
Total	**37354**	**1686**	**1674**	**35**	**40749**

NMR - Nuclear magnetic resonance
Allows structure determination based on
distance and angular constraints ***in solu***