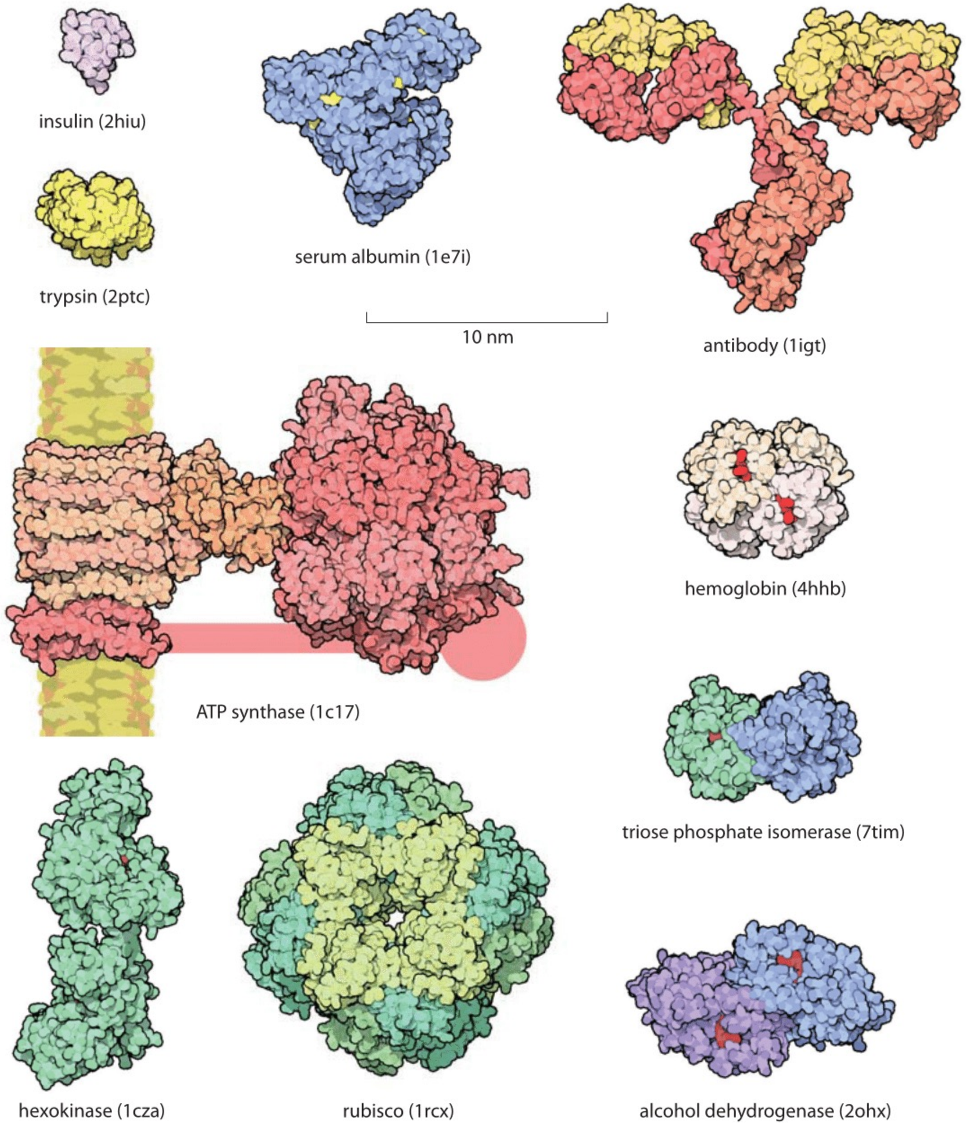
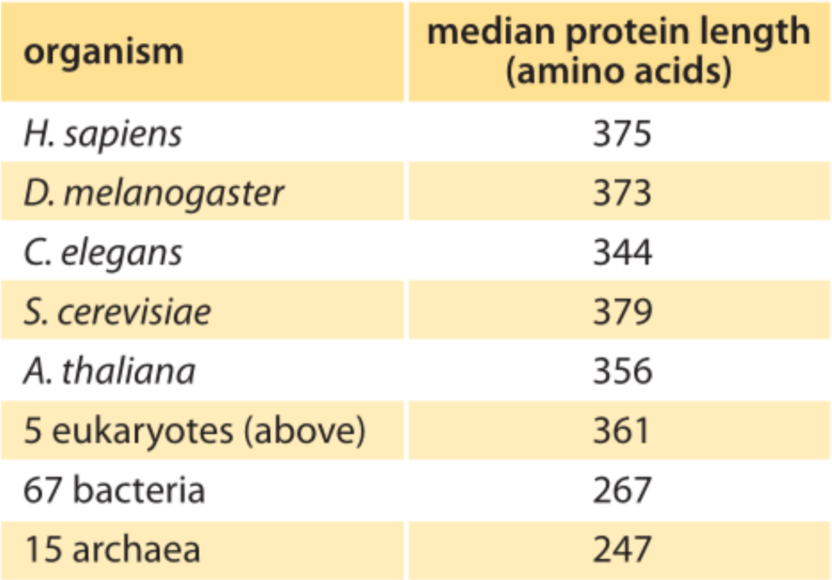
Amino acids and proteins

Proteins

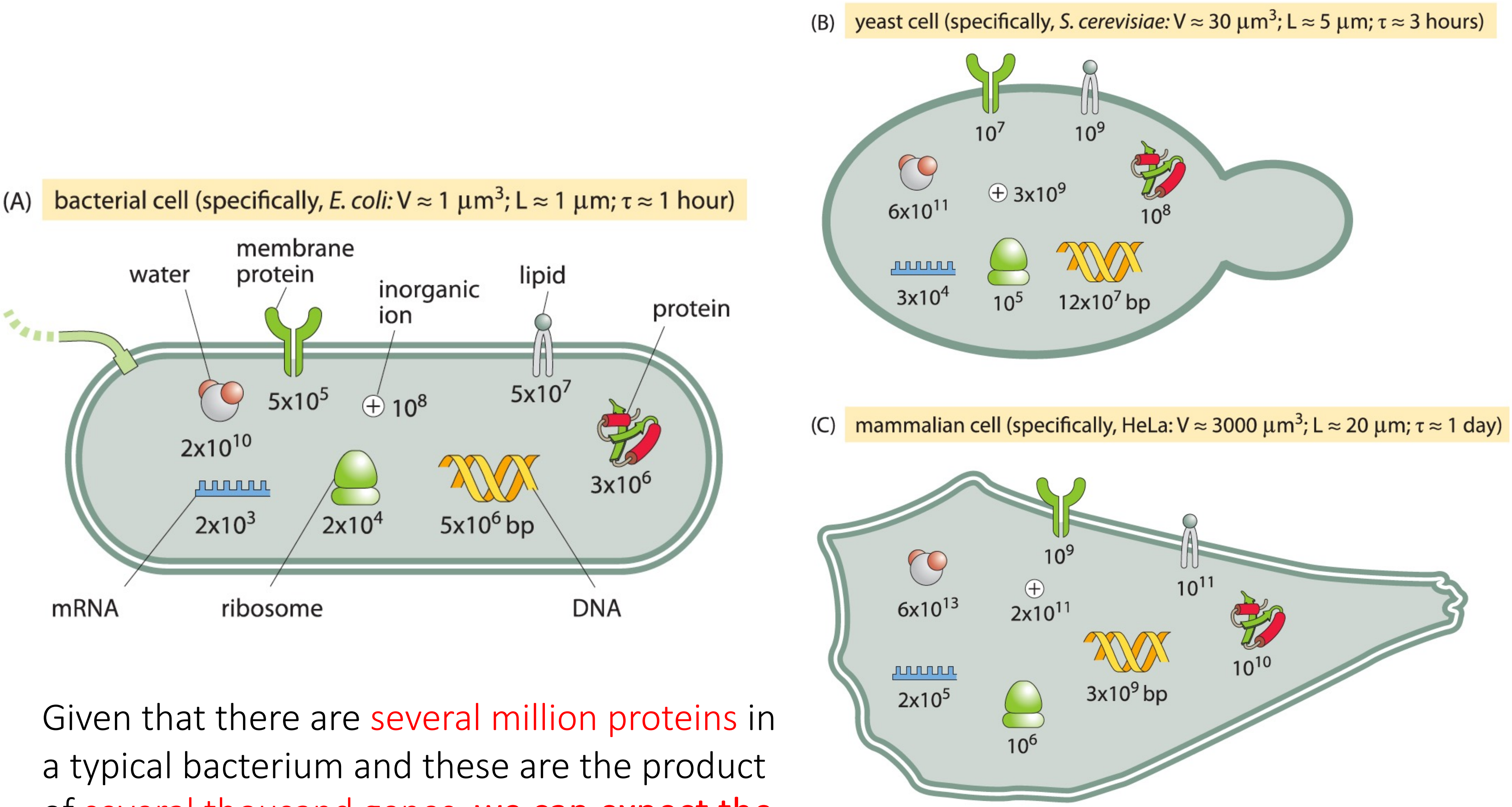




organism	median protein length (amino acids)
H. sapiens	375
D. melanogaster	373
C. elegans	344
S. cerevisiae	379
A. thaliana	356
5 eukaryotes (above)	361
67 bacteria	267
15 archaea	247

http://book.bionumbers.org/how-big-is-the-average-protein/

The nucleus houses the billions of base pairs of the genome and is the site of the critical transcription processes taking place as genes are turned on and off in response to environmental stimuli and over the course of both the cell cycle and development.



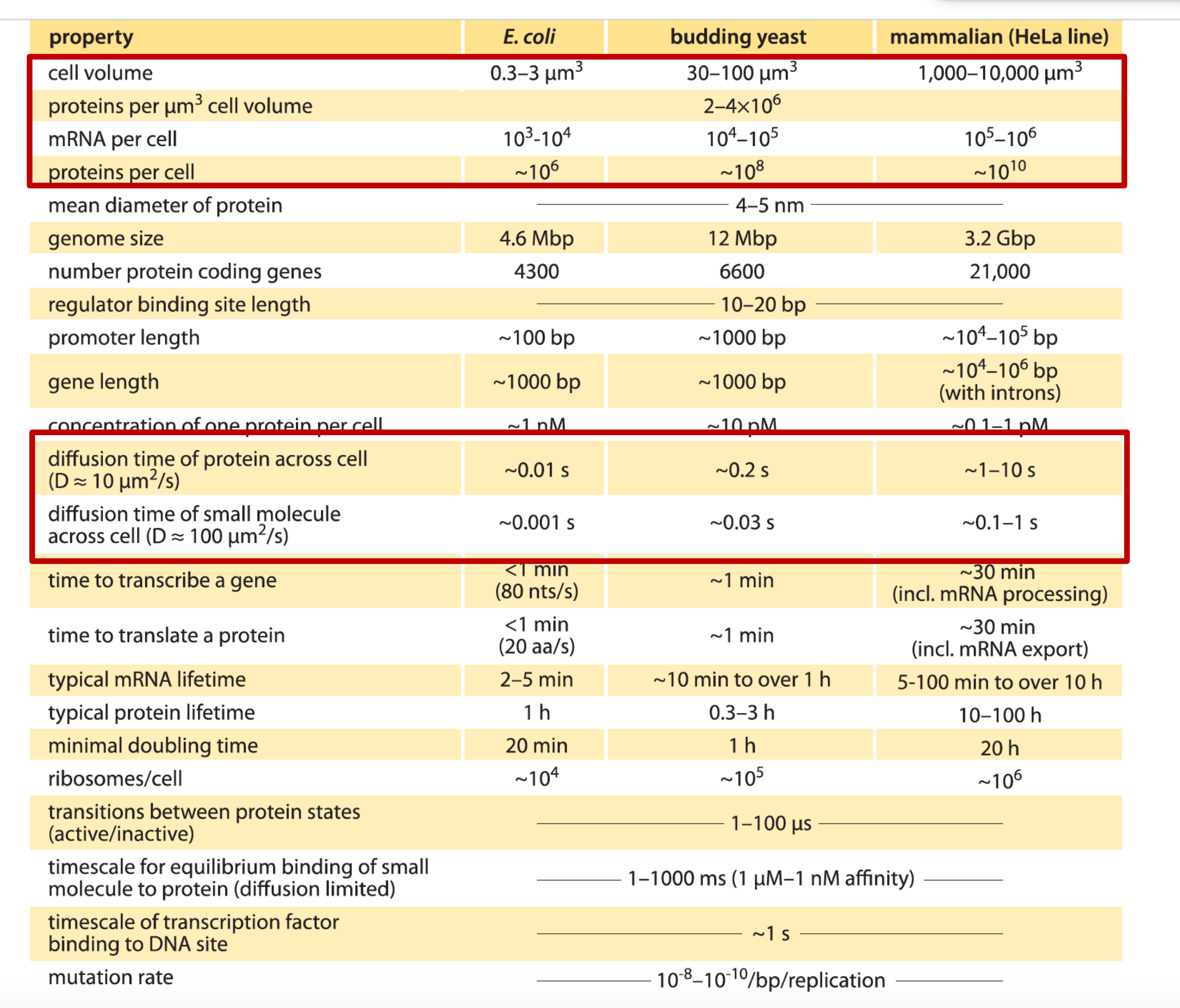
(A) bacterial cell (specifically, *E. coli:* $V \approx 1\ \mu m^3$; $L \approx 1\ \mu m$; $\tau \approx 1$ hour)

(B) yeast cell (specifically, *S. cerevisiae:* $V \approx 30\ \mu m^3$; $L \approx 5\ \mu m$; $\tau \approx 3$ hours)

(C) mammalian cell (specifically, HeLa: $V \approx 3000\ \mu m^3$; $L \approx 20\ \mu m$; $\tau \approx 1$ day)

Given that there are several million proteins in a typical bacterium and these are the product of several thousand genes, **we can expect the "average" protein to have about 10^3 copies.**

In 1 um3 (1 nL) stanno tra 0,1 a 0,3 microgrammi (µg) di proteine (tra 100 e 300 mg/mL)

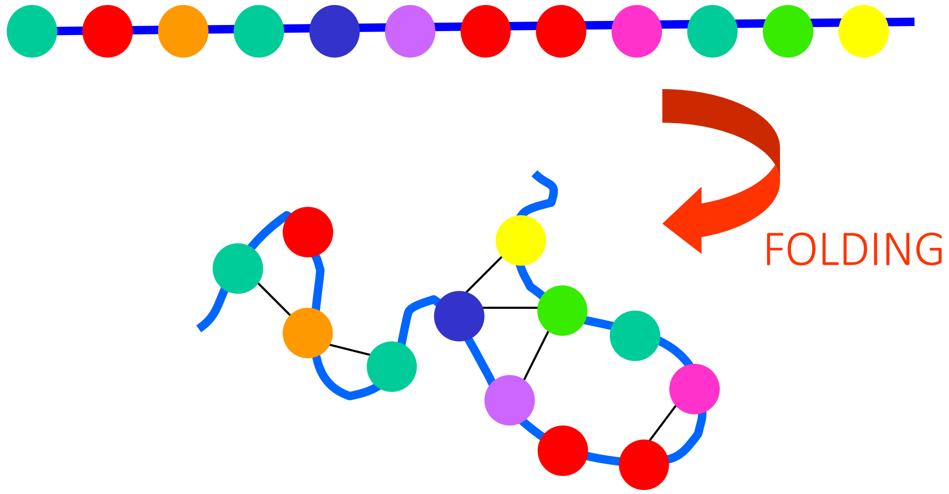


property	*E. coli*	budding yeast	mammalian (HeLa line)
cell volume	0.3–3 μm^3	30–100 μm^3	1,000–10,000 μm^3
proteins per μm^3 cell volume		2–4×10^6	
mRNA per cell	10^3-10^4	10^4–10^5	10^5–10^6
proteins per cell	~10^6	~10^8	~10^{10}
mean diameter of protein		4–5 nm	
genome size	4.6 Mbp	12 Mbp	3.2 Gbp
number protein coding genes	4300	6600	21,000
regulator binding site length		10–20 bp	
promoter length	~100 bp	~1000 bp	~10^4–10^5 bp
gene length	~1000 bp	~1000 bp	~10^4–10^6 bp (with introns)
concentration of one protein per cell	~1 nM	~10 pM	~0.1–1 pM
diffusion time of protein across cell (D ≈ 10 μm^2/s)	~0.01 s	~0.2 s	~1–10 s
diffusion time of small molecule across cell (D ≈ 100 μm^2/s)	~0.001 s	~0.03 s	~0.1–1 s
time to transcribe a gene	<1 min (80 nts/s)	~1 min	~30 min (incl. mRNA processing)
time to translate a protein	<1 min (20 aa/s)	~1 min	~30 min (incl. mRNA export)
typical mRNA lifetime	2–5 min	~10 min to over 1 h	5-100 min to over 10 h
typical protein lifetime	1 h	0.3–3 h	10–100 h
minimal doubling time	20 min	1 h	20 h
ribosomes/cell	~10^4	~10^5	~10^6
transitions between protein states (active/inactive)		1–100 μs	
timescale for equilibrium binding of small molecule to protein (diffusion limited)		1–1000 ms (1 μM–1 nM affinity)	
timescale of transcription factor binding to DNA site		~1 s	
mutation rate		10^{-8}–10^{-10}/bp/replication	

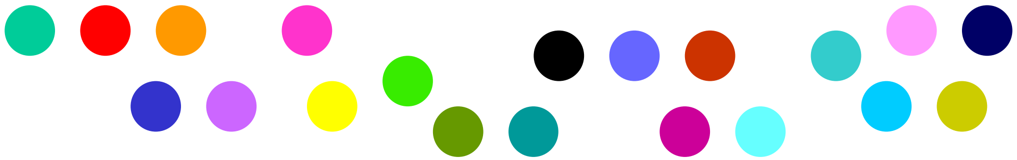
Proteins

Proteins are linear chains of amino acids.

These chains fold in 3D due to the non-covalent interactions between regions of the linear sequence



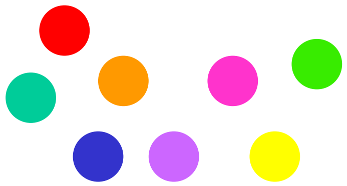
There are 20 different types of amino acid, each with different physico-chemical properties.



- FUNCTION DEPENDS ON 3D STRUCTURE
- 3D STRUCTURE DEPENDS ON SEQUENCE
- SEQUENCE IS DETERMINED GENETICALLY

Overview of protein architecture

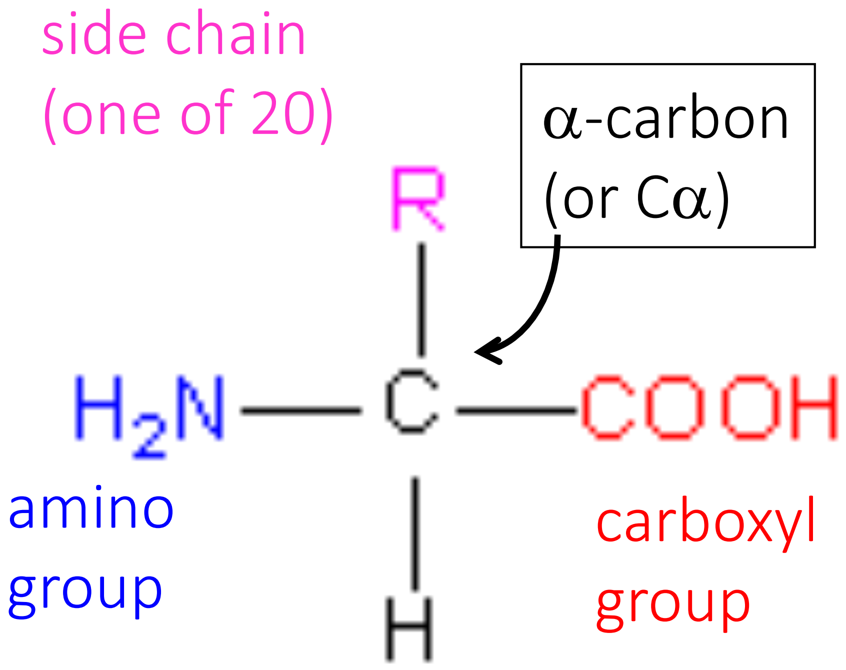
1) structure and chemistry of amino acids



2) how amino acids are linked together through peptide bonds to form a polypeptide chain

3) how the polypeptide chain folds in 3D
   - secondary structure elements (α-helix and β-sheet)
   - how secondary structure elements pack together

Structure of amino acids



side chain (one of 20)

α-carbon (or Cα)

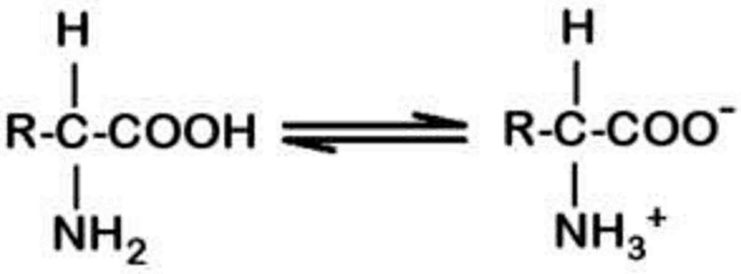
amino group

carboxyl group

At neutral pH:

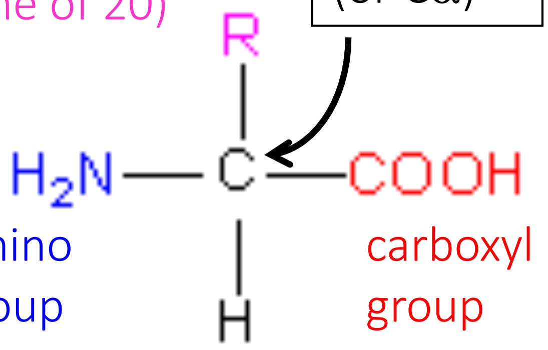
NH_2 is a base → NH_3^+

COOH is an acid → COO^-



zwitterion (dipolar form)

Structure of amino acids



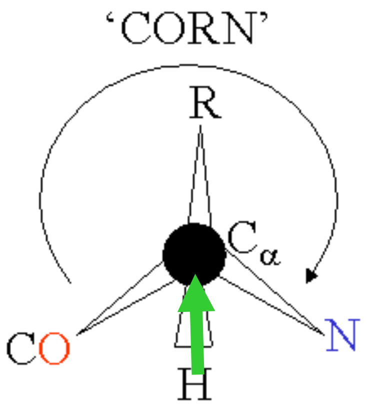
side chain
(one of 20)

α-carbon
(or Cα)

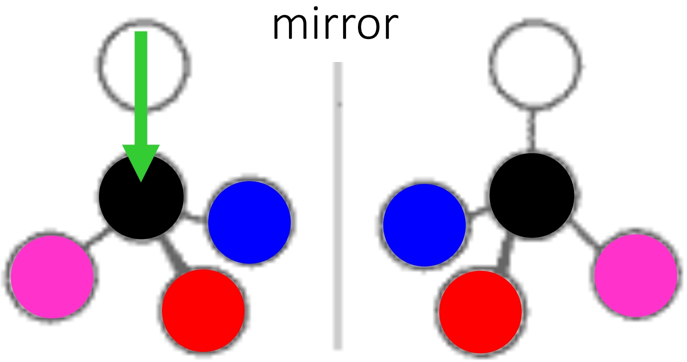
amino
group

carboxyl
group

CORN rule:
looking down
the H-Cα bond
for an L amino
acid we read
the groups
CO-R-N
clockwise



The Cα is an asymmetric carbon (bound to 4 different groups) and therefore is a chiral centre.
Two configurations (stereoisomers) are possible, which are one the mirror image of the other:



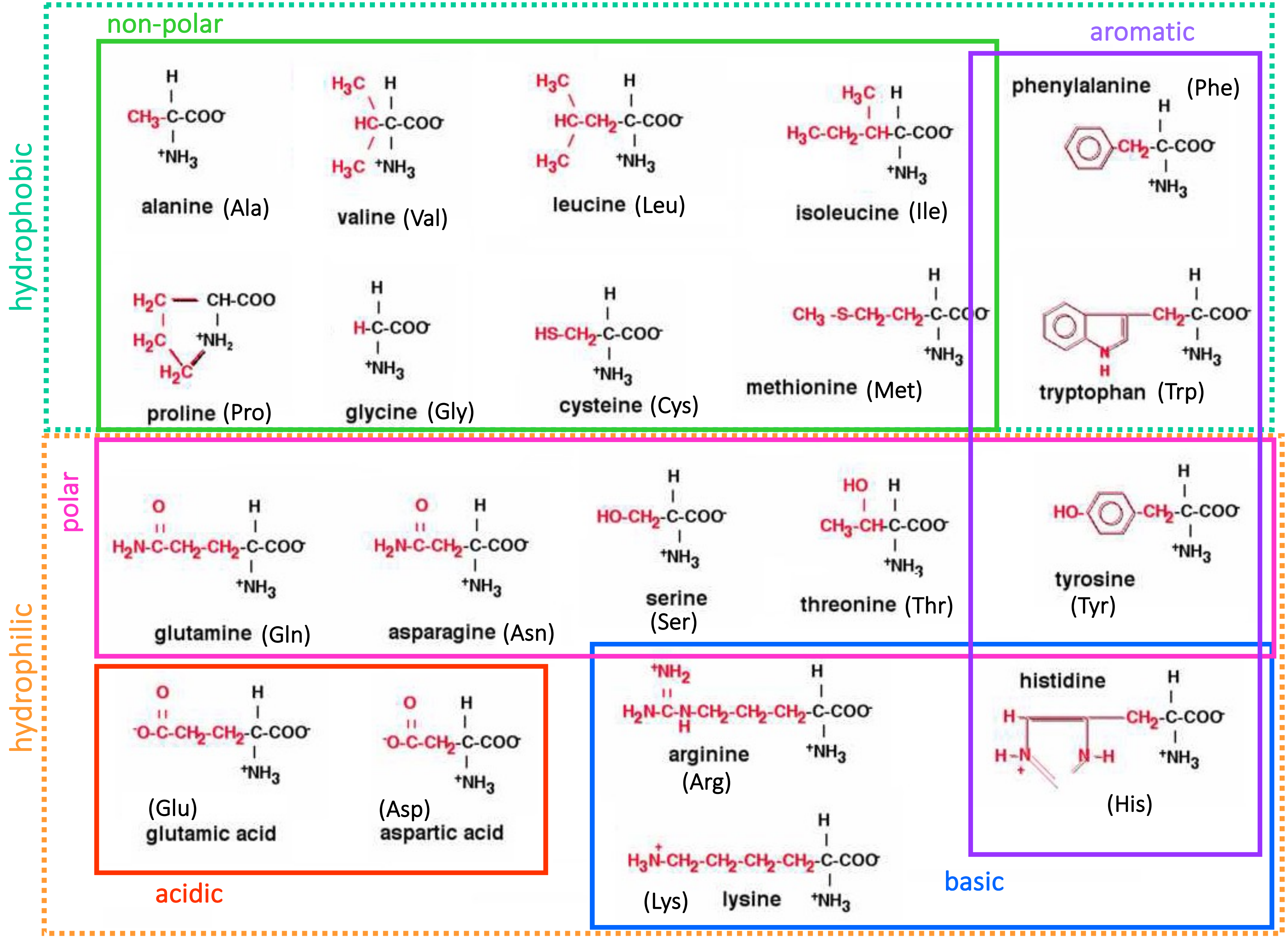
L-amino acid

D-amino acid

all amino acids in proteins are L!!

glycine, is the only exception: R = H, so no chirality

The 20 amino acids:



hydrophobic

non-polar

alanine (Ala)

valine (Val)

leucine (Leu)

isoleucine (Ile)

proline (Pro)

glycine (Gly)

cysteine (Cys)

methionine (Met)

aromatic

phenylalanine (Phe)

tryptophan (Trp)

hydrophilic

polar

glutamine (Gln)

asparagine (Asn)

serine (Ser)

threonine (Thr)

tyrosine (Tyr)

acidic

(Glu) glutamic acid

(Asp) aspartic acid

basic

arginine (Arg)

(Lys) lysine

histidine (His)

Properties of amino-acid side chains

R varies in

- shape
- size
- charge
- hydrophobicity
- reactivity

Hydrophobic amino acids: insoluble or slightly soluble in water

(side chains made of C, H, S - atoms with similar electronegativity)

avoid water by coalescing into oily droplets - the same forces causes hydrophobic aa to pack together in the interior of proteins, away from acqueous solutions.

Hydrophylic amino acids: soluble in water

(side chains contains atoms such as N and O, which can make HB)

- polar
- basic
- acidic

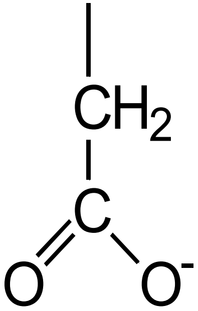
Charged side chains

at neutral pH

negative charge

Aspartic acid

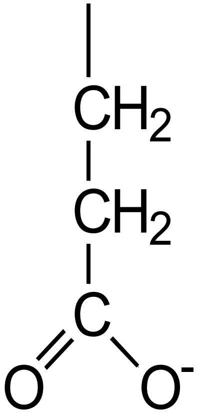
(Asp or D)



$pK_a=3.9$

Glutamic acid

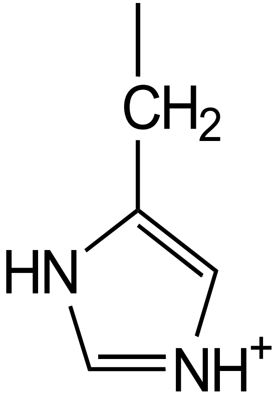
(Glu or E)



$pK_a=4.2$

Histidine

(His or H)

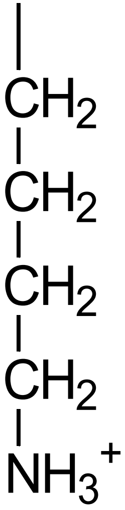


$pK_a=6.1$

positive charge

Lysine

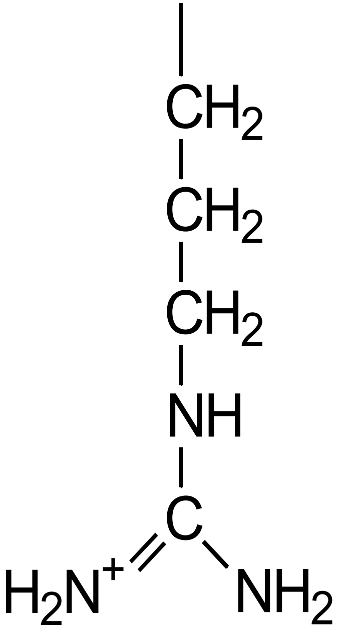
(Lys or K)



$pK_a=10.8$

Arginine

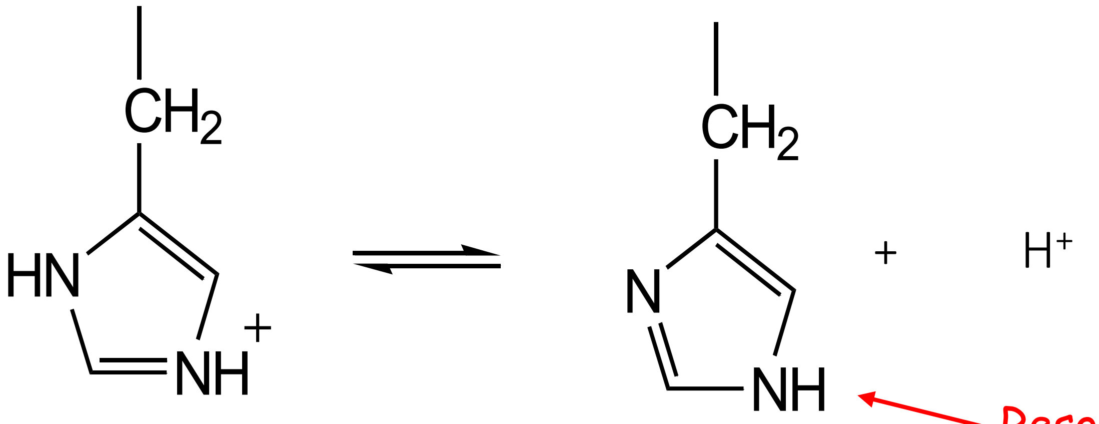
(Arg or R)



$pK_a=12.5$

Histidine

The side chain of a histidine residue is uncharged at high pH



Resonance - proton can be on either nitrogen

$pK_a = 6.1$

Charge on His residue in a protein at neutral pH will depend on the local environment - small shifts of pH or local environment can change the charge of a His (important in enzyme mechanism)

pH

The acidity of a solution is measured on a "pH" scale where:

$$pH = -\log_{10}[H^{+}]$$

For pure water $[H^{+}] = 10^{-7}$ M and thus the pH = $-\log(10^{-7}) = 7$

pH < 7.0 then [H+] > [OH-] → solution is acidic

pH > 7.0 then [H+] < [OH-] → solution is basic (or alkaline)

The interior of a cell is kept close to neutrality by the presence of buffers: weak acid and bases that can release or take up protons near pH 7, keeping the environment of the cell relatively constant under a variety of conditions.

pK_a

For an acid:

$$HA \overset{K_a}{\rightleftharpoons} H^+ + A^-$$

$K_a = [H^+]\,[A^-]/[HA]$

K_a is an equilibrium constant (in this case an acid dissociation constant)

We define: $$pK_a = -\log_{10}(K_a)$$

where K_a is the acid dissociation constant, i.e. how much the acid tends to give up a proton (H^+) in water. pK_a expresses the strength of the acid

More dissociated → more equilibrium to right → larger K_a → smaller pK_a

Smaller pK_a → strong acid / weak conjugate base

Larger pK_a → weak acid / strong conjugate base

pK is a log measure of how tightly a molecule holds onto a proton. It tells you the pH at which the molecule will switch between protonated and deprotonated forms.

pK_a

For an acid:

$$HA \overset{K_a}{\rightleftharpoons} H^+ + A^-$$

K_a is an equilibrium constant (in this case an acid dissociation constant)

We define:

$$pK_a = -\log_{10}(K_a)$$

$K_a = [H^+]\ [A^-]/[HA]$

$$\mathrm{pH} = \mathrm{p}K_a + \log_{10}\left(\frac{[A^-]}{[HA]}\right)$$

[HA] = concentration of the protonated form.
[A⁻]= concentration of the deprotonated form.
The equation links pH of the solution with the protonation state of the group.

The key case: when pH = pKa

•The log term = log(1) = 0.

•So pH = pK_a exactly when [A⁻] = [HA].

➡️ That means the group is 50% protonated and 50% deprotonated.

•pKa defines the "tipping point" pH where a group is half-protonated

pK_a

For an acid:

$$HA \xrightleftharpoons{K_a} H^+ + A^-$$

$$K_a = [H^+]\,[A^-]/[HA]$$

K_a is an equilibrium constant (in this case an acid dissociation constant)

We define: $$pK_a = -\log_{10}(K_a)$$

$$\mathrm{pH} = \mathrm{p}K_a + \log_{10}\left(\frac{[A^-]}{[HA]}\right)$$

Rules of thumb

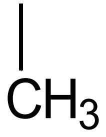
- **If pH < pKa** → mostly protonated (acid holds onto H^+).
- **If pH > pKa** → mostly deprotonated (acid loses H^+).

Why it matters in proteins

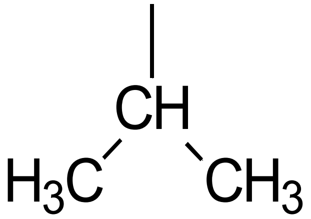
- Side chains with pKa values near physiological pH (~7) can switch protonation state easily.
- Example: Histidine (pKa ≈ 6) is often used in enzyme active sites, because it can donate or accept a proton depending on small pH changes.

Non-polar side chains

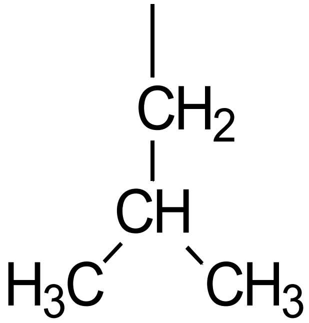
Alanine
(Ala or A)



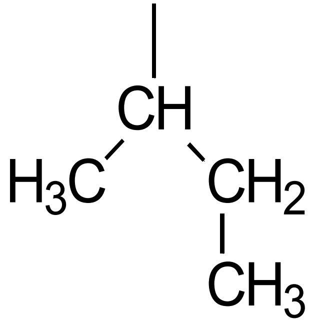
Valine
(Val or V)



Leucine
(Leu or L)

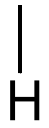


Isoleucine
(Ile or I)



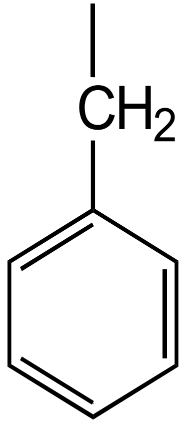
SMALL

Glycine
(Gly or G)

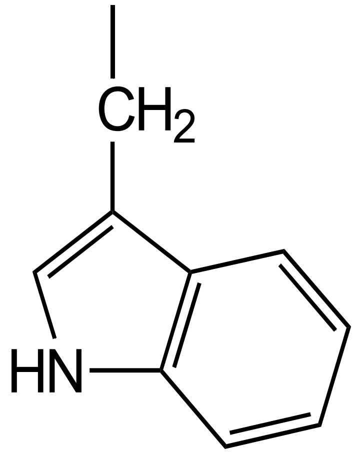


Glycine is the smallest amino acid: has no side chain and is not chiral

Phenylalanine
(Phe or F)



Tryptophan
(Trp or W)



LARGE

Non-polar side chains

Hydrophobic core formation

- Water “dislikes” nonpolar groups → they cluster together inside the protein (hydrophobic effect).
- This drives **protein folding**, creating a stable interior shielded from water.

Stability of 3D structure

- The buried hydrophobic side chains pack tightly, like puzzle pieces.
- This packing gives proteins much of their **structural stability**.

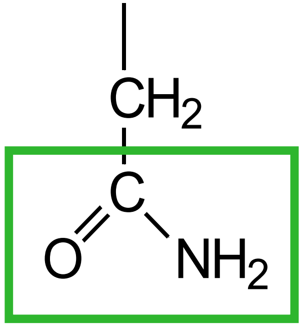
🔷 **Membrane interactions**

- In membrane proteins, nonpolar amino acids often sit in the **transmembrane regions**, interacting with the lipid tails.
- They allow proteins to be embedded in lipid bilayers.

Uncharged polar side chains

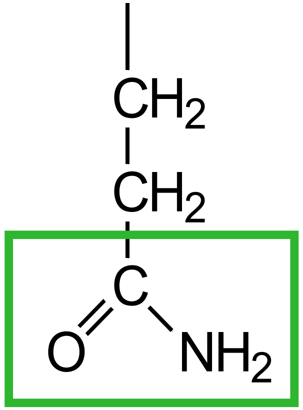
Asparagine

(Asn or N)



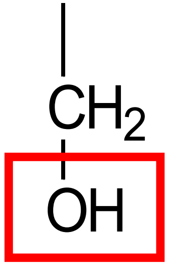
Glutamine

(Gln or Q)



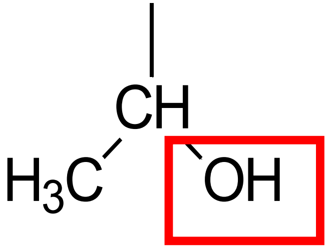
Serine

(Ser or S)



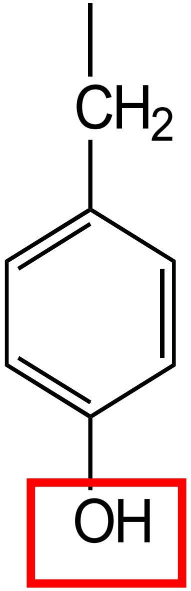
Threonine

(Thr or T)



Tyrosine

(Tyr or Y)



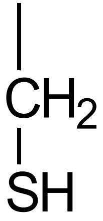
They are **polar** because they have electronegative atoms.
They are **uncharged** at physiological pH (≈7)

Main roles in proteins

- Hydrogen bonding; stabilize secondary/tertiary structures,

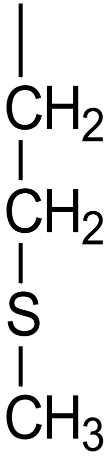
Remaining non-polar side chains

Cysteine
(Cys or C)



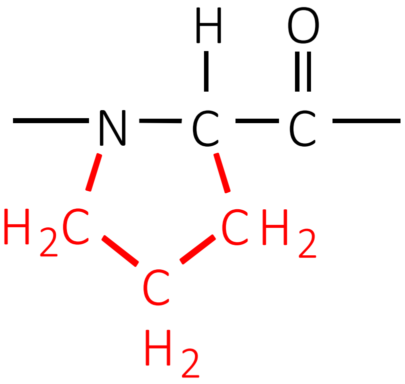
Cysteine residues can form disulphide bonds.

Methionine
(Met or M)



Proteins usually start with a Met

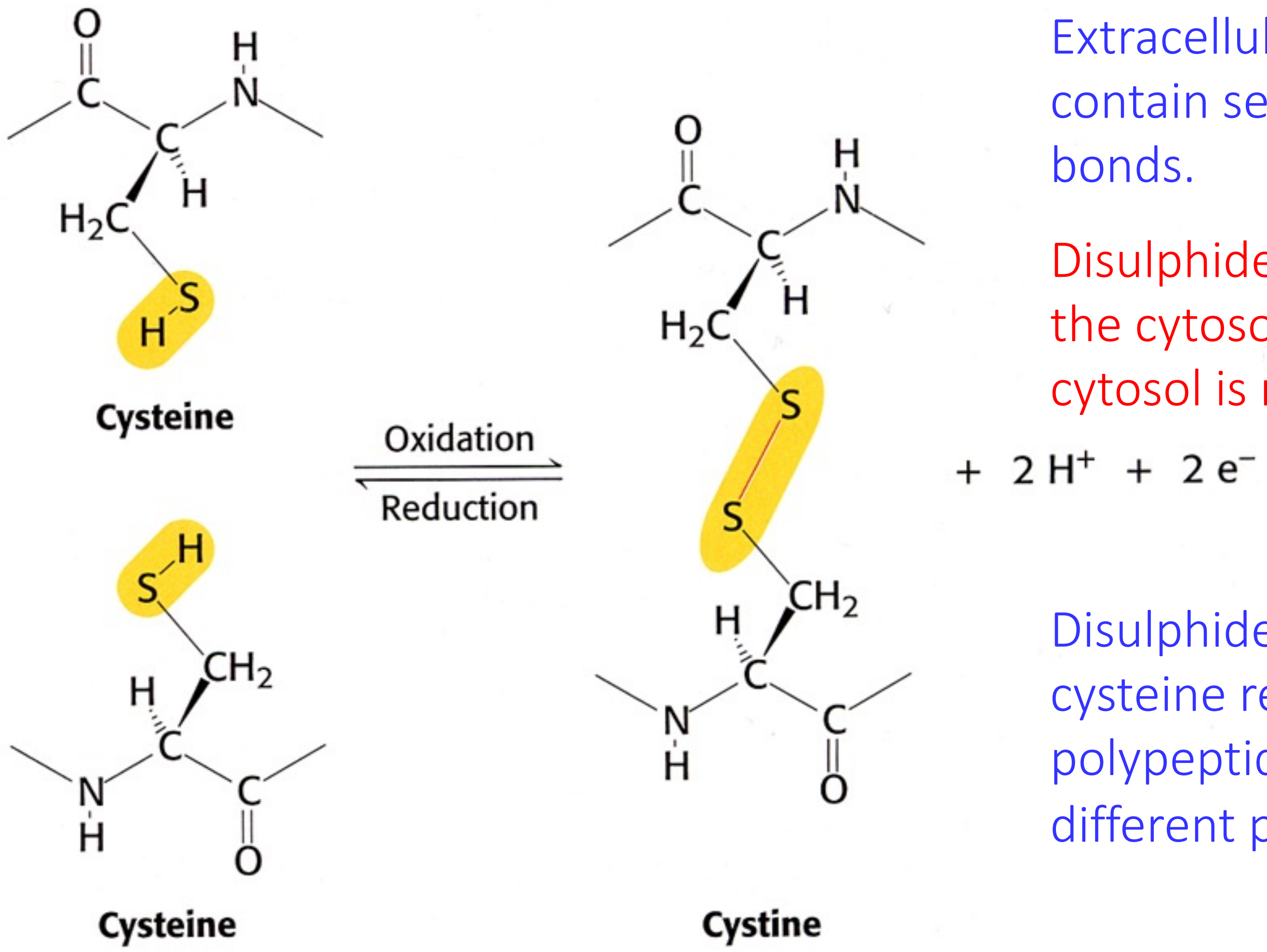
Proline
(Pro or P)



The side chain is covalently bonded to the main chain nitrogen. This locks the conformation around the N-Cα bond – reducing flexibility of the polypeptide chain.

Disulphide bonds

A disulphide bond can form between two cysteine residues in proteins.



Extracellular proteins often contain several disulphide bonds.

Disulphide bonds do not form in the cytosol (need an oxidizing env, cytosol is reducing).

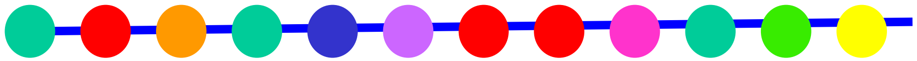
Disulphide bonds can link cysteine residues within a single polypeptide chain or on different polypeptide chains.

60 kcal/mol (251 kJ mol^{-1})

Overview of protein architecture

1) structure and chemistry of amino acids

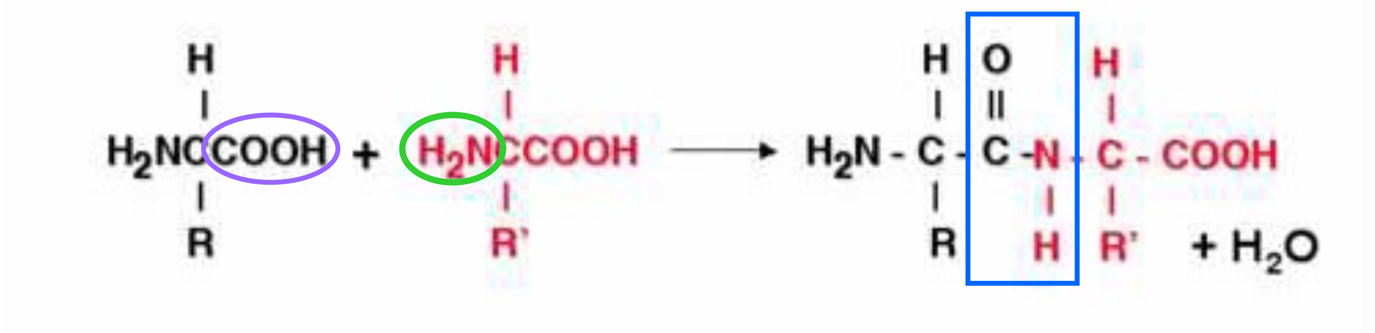
2) how amino acids are linked together through peptide bonds to form a polypeptide chain



3) how the polypeptide chain folds in 3D
   - secondary structure elements (α-helix and β-sheet)
   - how secondary structure elements pack together

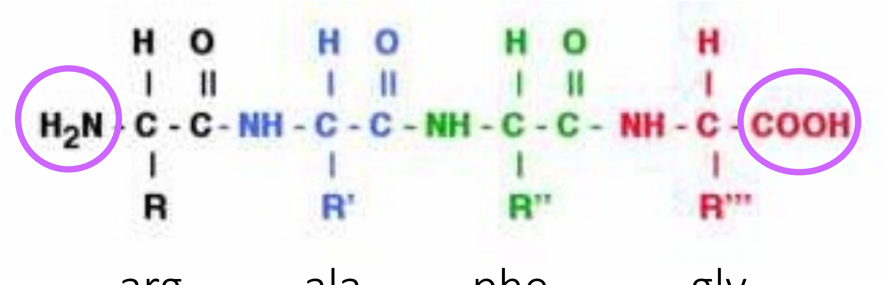
The peptide bond

The amino acids of a protein are joined together through a covalent bond between the carboxyl group of one aa and the amino group of the next aa (peptide bond).



This produce a chain of amino acids which is asymmetric: on one end there is a free NH_2 group (N terminus) and at the other end a free COOH (C terminus).

N



C

arg ala phe gly

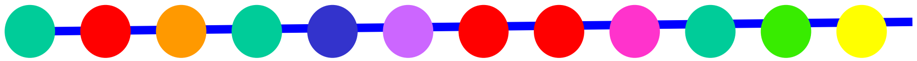
R A F G

A peptide/protein sequence is always given from the N to the C terminus (here RAFG).

Primary structure

the linear sequence of amino acids

- the sequence is always written N↦C
- each protein has a unique and defined sequence, which is genetically determined
- a typical protein contains 100-1000 aa
- Average mass of 1 aminoacid: 100 Da
- 1 Dalton = 1 g/mol ≈ 1.7x10-24 g



- sequencing=determining the number and order of the aa in the chain

In 1953 Saenger sequenced insulin (Nobel price); now it is more common to sequence the corresponding gene.

We can guess the function of an unknown protein if it shows sequence similarity to a protein of known function.

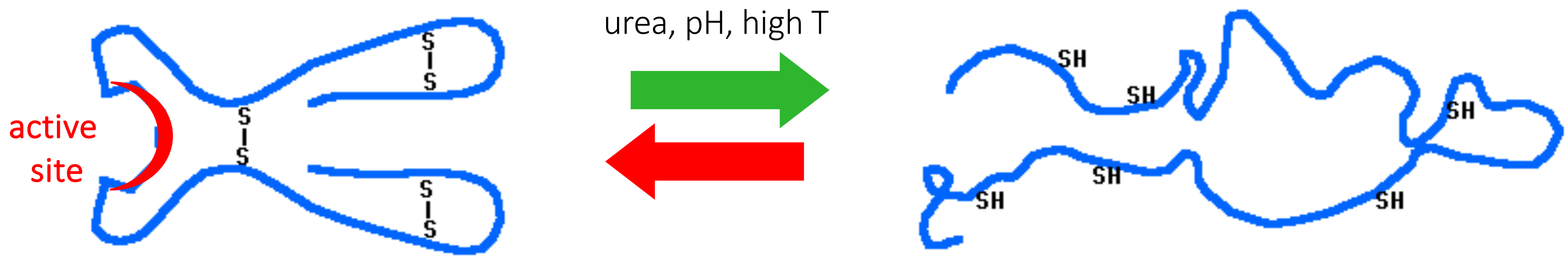
Often we know the sequence of the same protein from different organisms: these are more and more different the more the organisms have diverged in evolution.
Proteins evolve by changing (little by little) their aminoacid sequence

Denaturation

Many proteins can be unfolded and refolded:

FOLDED, active protein

UNFOLDED, denatured protein



active site

urea, pH, high T

refold = regain biological activity

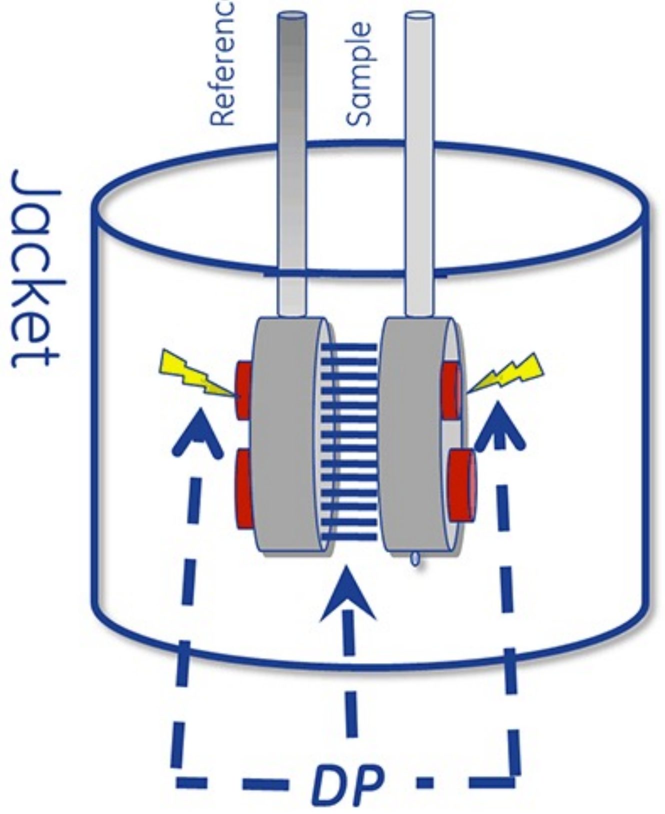
unfold = loose biological activity

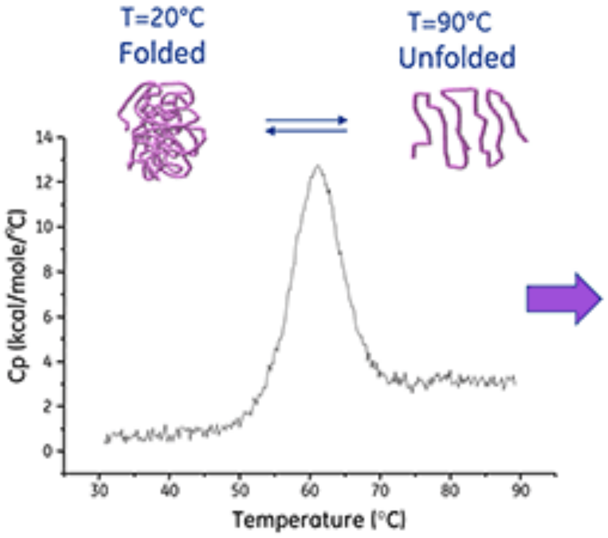
When a protein contains disulphide bridges (S-S)
we also need to reduce them to SH to unfold.

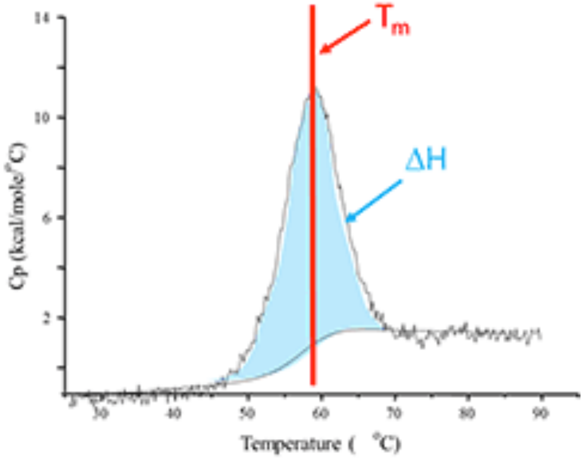
It does not work for all proteins - some proteins, once unfolded cannot be easily refolded again.

Denaturation

Many proteins can be unfolded and refolded: **differential scanning calorimetry**





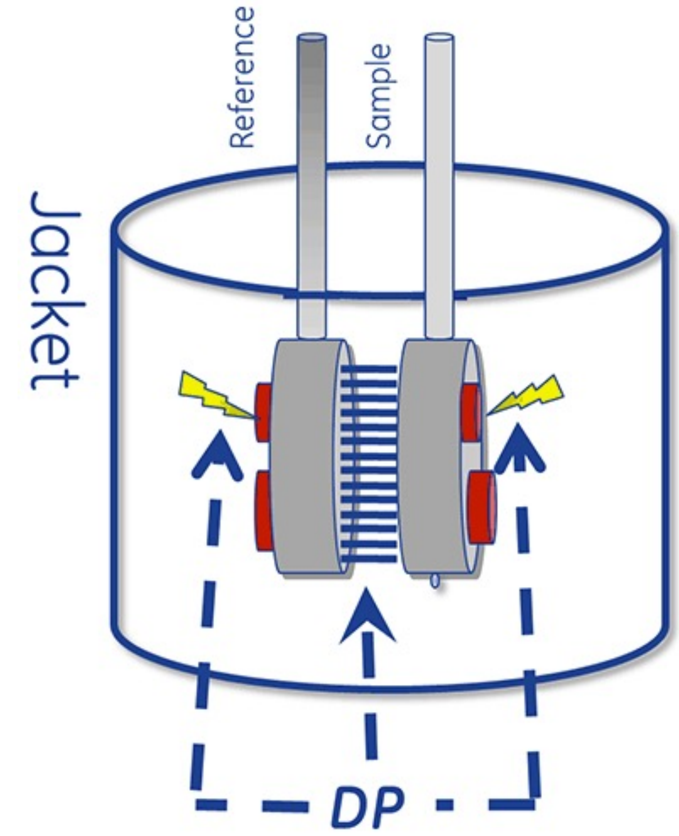


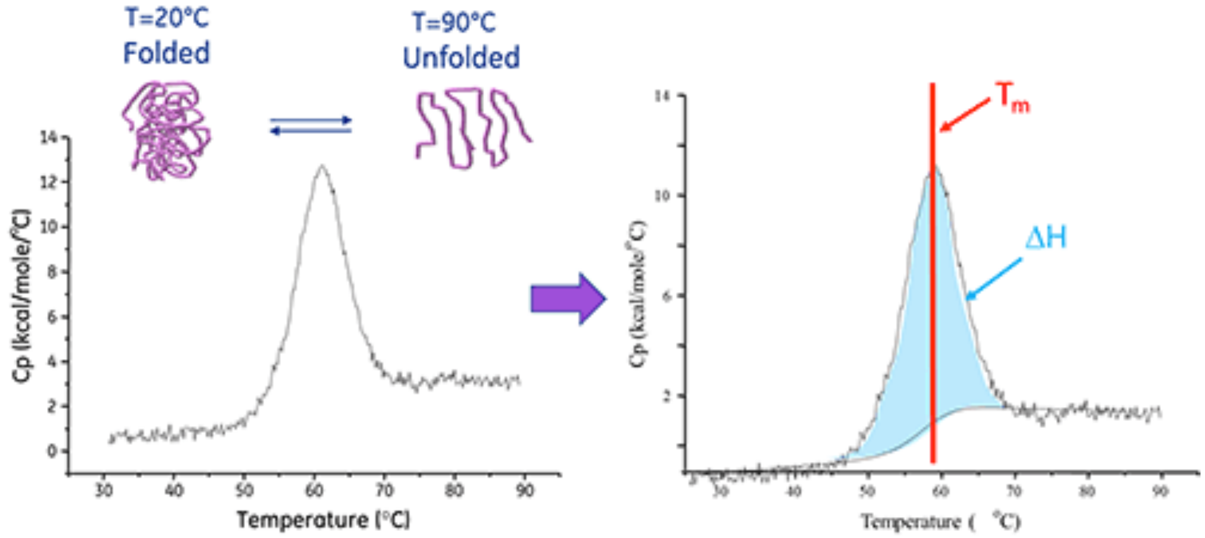
DSC measures the **heat required to raise the temperature** of a protein solution compared to a reference (usually just buffer).

As the protein unfolds with increasing temperature, it absorbs heat (endothermic process). This extra heat shows up as a peak in the heat capacity (Cp) curve.

Denaturation

Many proteins can be unfolded and refolded: differential calorimetry





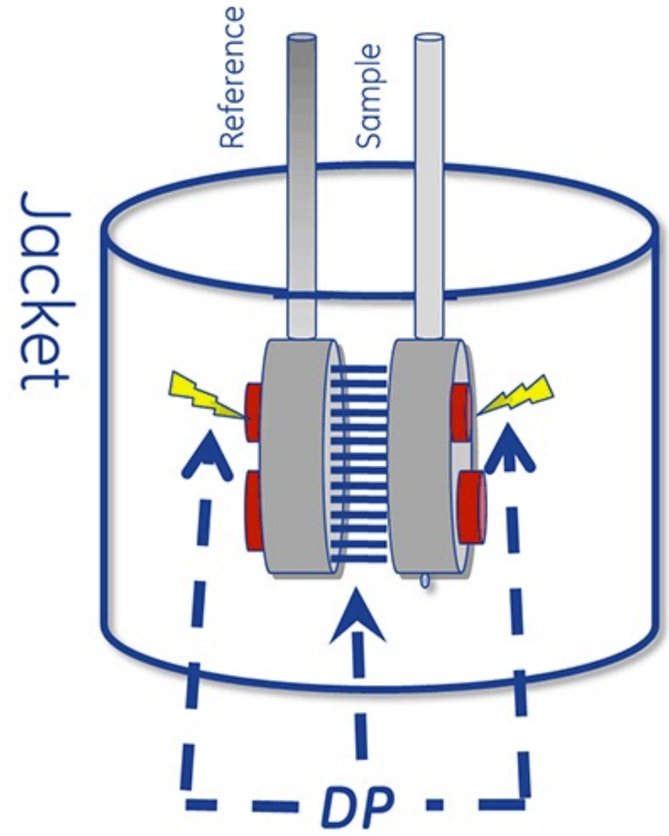
Heat capacity (Cp)

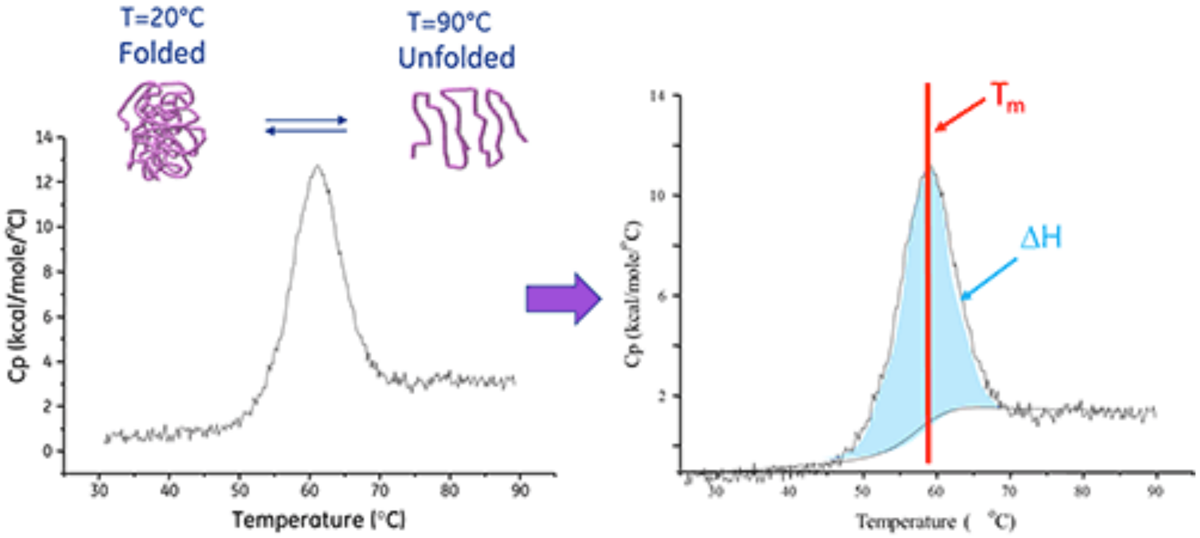
•Definition: Cp=∂Q/∂T = amount of heat needed to increase temperature by 1 degree.

•Proteins in folded vs unfolded states have different Cp values, because unfolded proteins expose hydrophobic residues to solvent, increasing solvent reorganization.

Denaturation

Many proteins can be unfolded and refolded: differential calorimetry



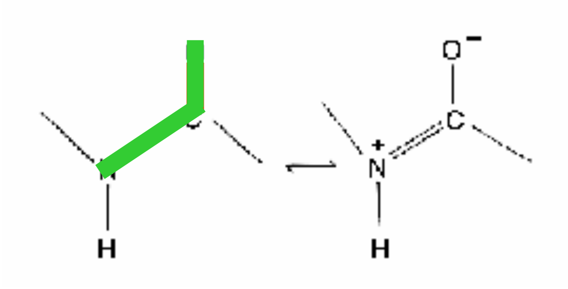


Unfolding as a cooperative transition

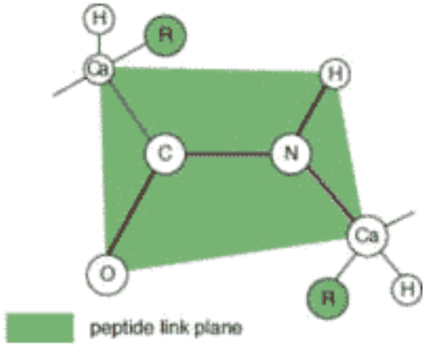
•Protein folding/unfolding is not gradual for each residue, but cooperative: many residues unfold together once the transition starts.

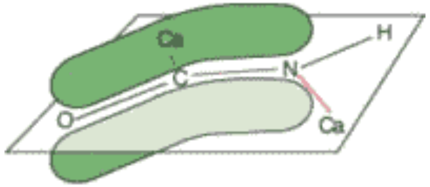
•This shows up as a sharp heat absorption peak at the melting temperature T_m .

Planarity of the peptide bond



Partial double bond character of the N-C bond
leads to restricted rotation
the region NH-CO is planar:

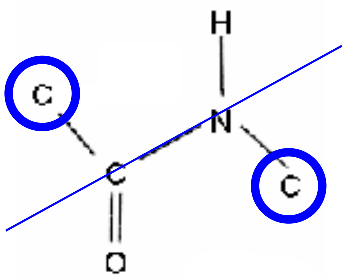




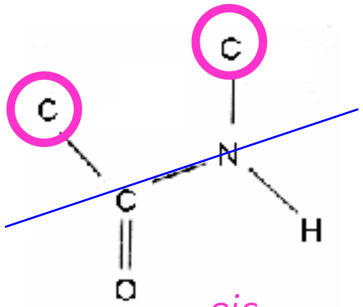
delocalisation of the π electrons over the entire peptide bond, rather than simply over the C=O bond

The peptide bond can assume a *trans* or a *cis* conformation: the *trans* form is favoured 1000:1.

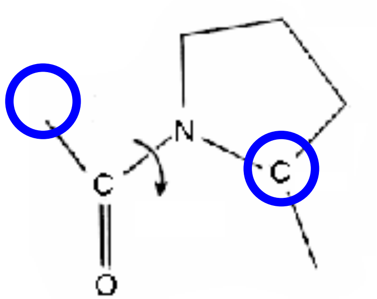
In the case of prolines, the *trans* form is only favoured 15:1



trans



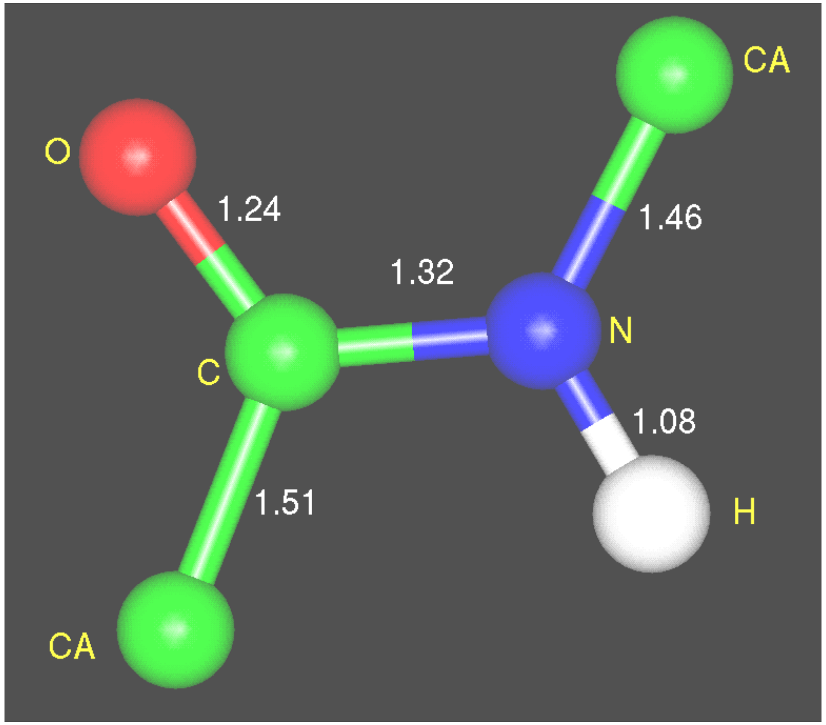
cis



trans
Proline

proline a structural regulator, important in folding, turns, and regulation.

The ideal peptide



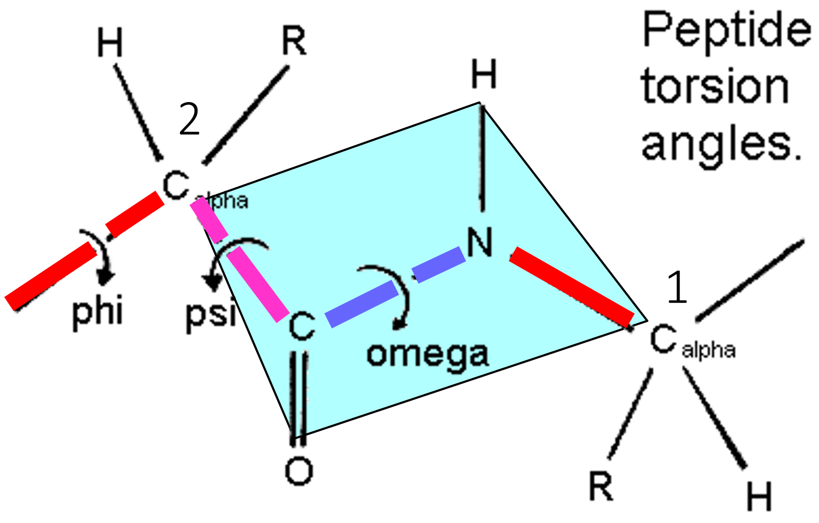
C-N single bond ~ 1.48 Å

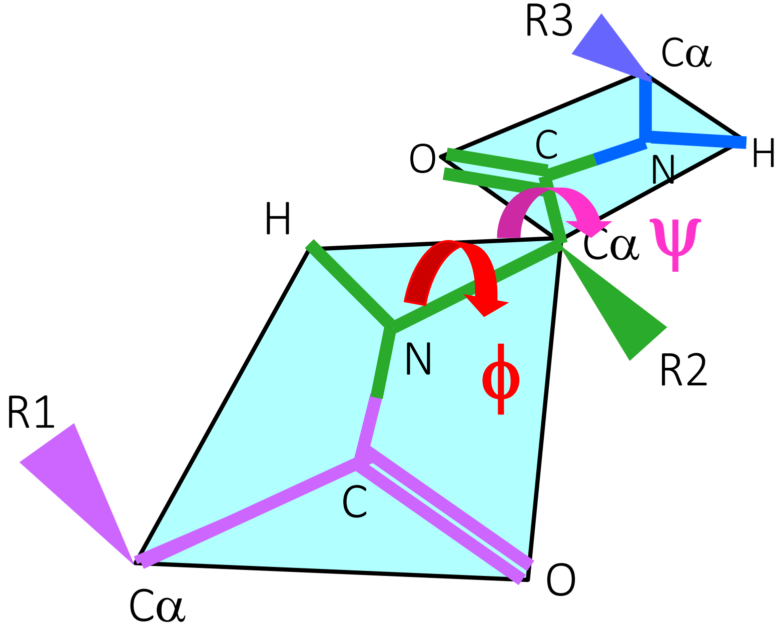
C=O double bond ~ 1.20 Å

peptide bond C-N = 1.32 Å (i.e. shorter than a single bond due to partial double bond character) while C=O bond is slightly longer

Peptide bond	Average length	Single Bond	Average length	Hydrogen Bond	Average (±0.3)
Cα – C	1.51 (Å)	C - C	1.54 (Å)	O-H --- O-H	2.8 (Å)
C - N	1.32 (Å)	C - N	1.48 (Å)	N-H --- O=C	2.9 (Å)
N - Cα	1.46 (Å)	C - O	1.43 (Å)	O-H --- O=C	2.8 (Å)

The torsion angles ψ and ϕ





omega ($\boldsymbol{\omega}$) = rotation around C-N bond

not allowed because of resonance,
therefore ω=180° (for trans)

🟦 planar region

phi ($\boldsymbol{\phi}$)= free rotation around Cα-N bond

psi ($\boldsymbol{\psi}$)= free rotation around Cα-C bond

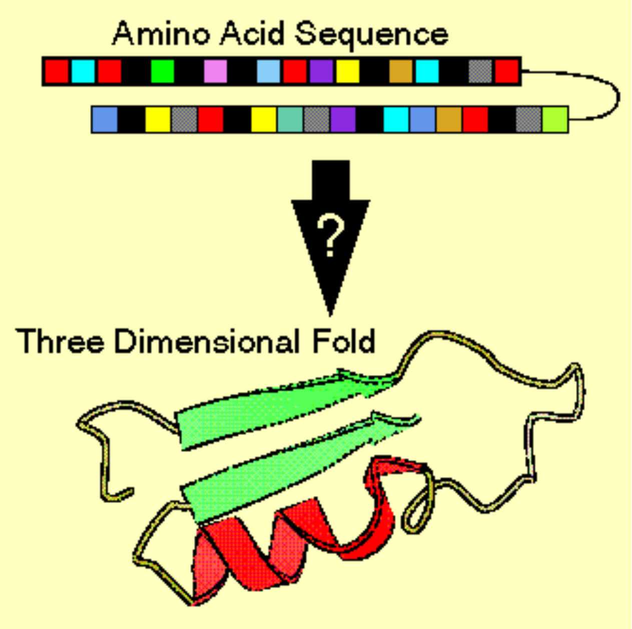
The main chain conformation is defined
by the sequence of the (ψ,ϕ) angles:
the list of the (ψ,ϕ) for each amino acid
dictate the fold of the polypeptide chain, i.e.
the 3D structure of the protein

How do proteins fold in the cell?

The amino-acid sequence specify the 3D structure, which is (probably?) the energy minimum for that particular sequence...

BUT how does a protein reach the correct three-dimensional fold?

by trying out all the possible conformations?



- consider the number of possible conformations of a chain of 100 amino acids
- assume each amino acid can have only 3 different conformations
- 3^{100} = $5x10^{47}$ possible different conformations
- if it took only 0.1 psec (10^{-13} sec) to try each possibility, it still would take $1.6x10^{27}$ years to find the minimum of energy!

➡ There must be a ‘folding pathway’!!!

first forming local structures quickly, then packing them together

The "folding problem"

Experimental approach

Studying experimentally how folding of a particular protein occur in vitro by using techniques like NMR which can detect the presence of secondary structure elements in a partially unfolded protein (trying to determine the 'folding pathway')

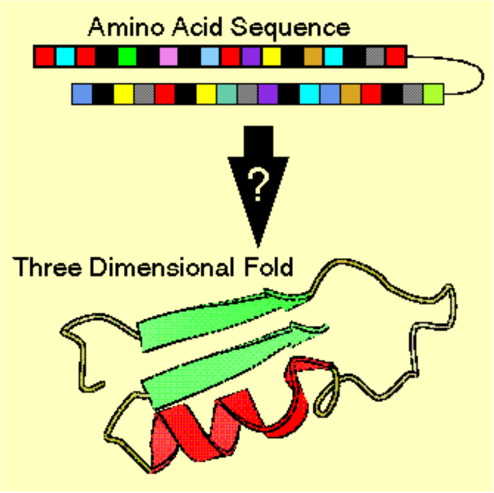
Studying experimentally how folding occur in the cell: some proteins fold by themselves, others require the help of other proteins called chaperones.

Theoretical approach

Using bio-informatics to predict the 3D structure from the amino-acid sequence. The sequence dictate the fold, but we are not very good at going from the sequence to the structure!

Problems?

- poor energy functions and parameters
- complexity
- treatment of solvent



The “folding problem”

Role of AI

1. Protein folding problem

•The challenge: given just the amino acid sequence, predict the 3D folded structure and possibly the folding pathway.

•Folding involves navigating an enormous “energy landscape” with countless conformations → too complex for brute-force physics alone.

2. AI’s role in *structure prediction*

•Tools like AlphaFold (DeepMind) and RoseTTAFold use deep learning trained on massive databases of known protein structures.

•AI learns patterns of evolutionary constraints (co-variation between residues) and the geometry of folded proteins.

•Outcome: near-experimental accuracy in many cases for final folded structures.

The “folding problem”

Role of AI

3. Protein folding pathway (beyond final structure)

This is harder than predicting the final fold, but AI helps in several ways:

🔷 Mapping the energy landscape

•AI models can predict not just the final structure, but intermediate states by generating ensembles of possible conformations.

•These can reveal folding intermediates and transition states.

🔷 Combining AI with physics

•Machine learning can speed up molecular dynamics (MD) simulations by providing better force fields or guiding simulations toward likely conformations.

•Hybrid methods allow exploration of folding pathways on realistic timescales.

🔷 Predicting folding kinetics

•Some AI approaches analyze sequence features to predict which parts fold first, rate-limiting steps, or propensity for misfolding/aggregation.

The “folding problem”

Role of AI

4. Why this matters

- Understanding pathways is key for:
  - **Diseases** (Alzheimer’s, Parkinson’s → protein misfolding/aggregation).
  - **Biotechnology** (designing stable proteins and enzymes).
  - Drug discovery (targeting folding intermediates or misfolded states).

5. Limits today

- AI is excellent at predicting final structures.
- Pathway prediction is still emerging:
  - Models are less reliable for dynamics and rare states.
  - Needs integration of AI with experimental data (e.g., NMR, cryo-EM, calorimetry, single-molecule FRET).

AI solved much of the protein structure prediction problem. It is starting to illuminate folding pathways by mapping conformational landscapes, guiding simulations, and predicting kinetics.
But: it works best when combined with physics-based models and experiments: folding is a dynamic, multi-step process that’s harder to learn from static structures alone.

Molecular evolution

Proteins evolve by changing little by little their amino-acid sequence

Changes are due to random mutations in the gene that code for that protein

- some mutations disrupt the structure and/or function of the protein and are eliminated by the selective pressure
- some mutations are ‘neutral’ and therefore allowed
- some (rare) mutations improve the functionality of the protein or change the function in a way that is advantageous for the cell

➡ evolution will select the favourable mutations

A lot of small changes occurring in all protein sequences accumulate with time and are responsible for the variety of living forms we see.

By comparing amino-acid sequences of proteins we can build evolutionary trees:

- key residues (structurally or functionally) are usually conserved
- other residues are usually very similar in organisms that have diverged recently but more and more diverse in distantly related organisms

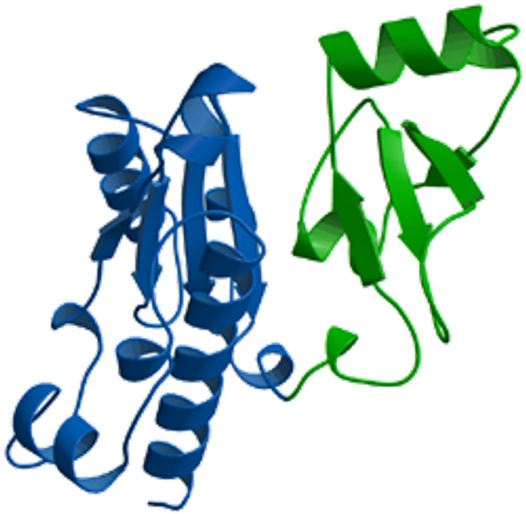
Overview of protein architecture

1) structure and chemistry of amino acids

2) how amino acids are linked together through peptide bonds to form a polypeptide chain

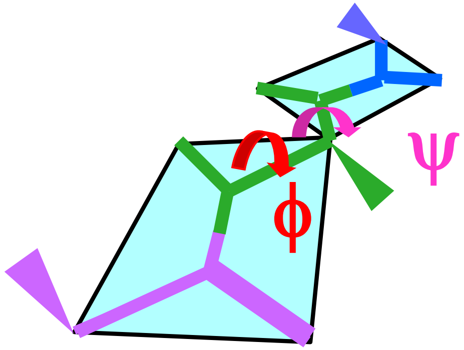
3) how the polypeptide chain folds in 3D:

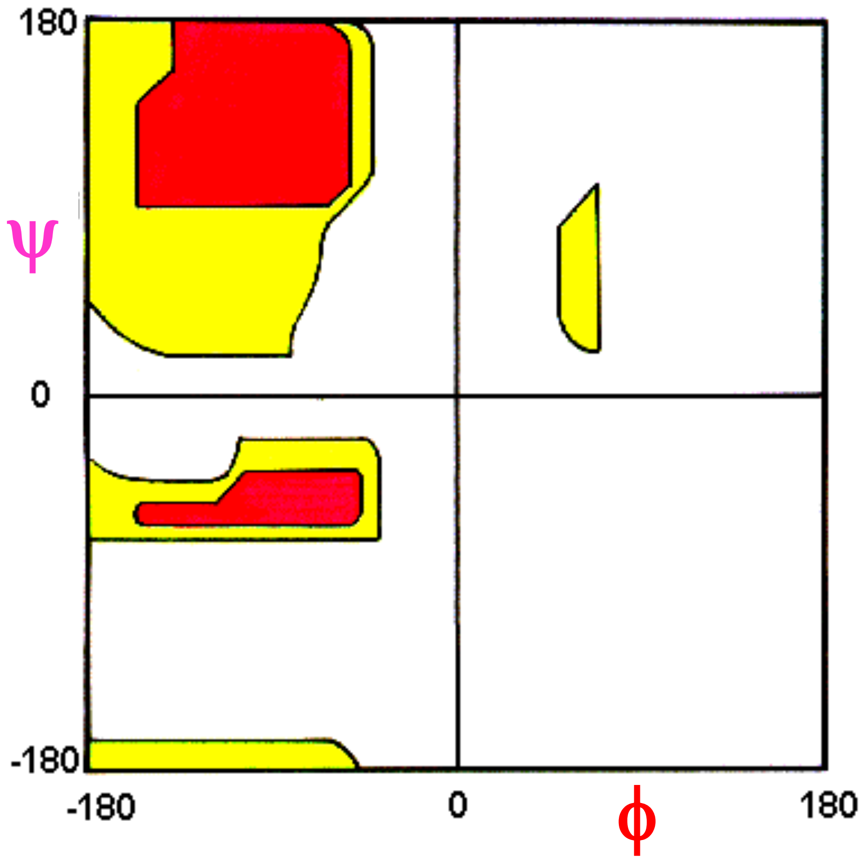
   - the Ramachandran plot
   - secondary structure elements (α-helix and β-sheet)
   - how secondary structure elements pack together



The Ramachandran plot

Because of steric clashes (atoms aren’t points; they have size. If you rotate bonds too freely, some atoms would crash into each other., only certain combinations of torsion angles are allowed): we can plot these allowed combinations in the (ψ,ϕ) plane - this is called the Ramachandran plot.

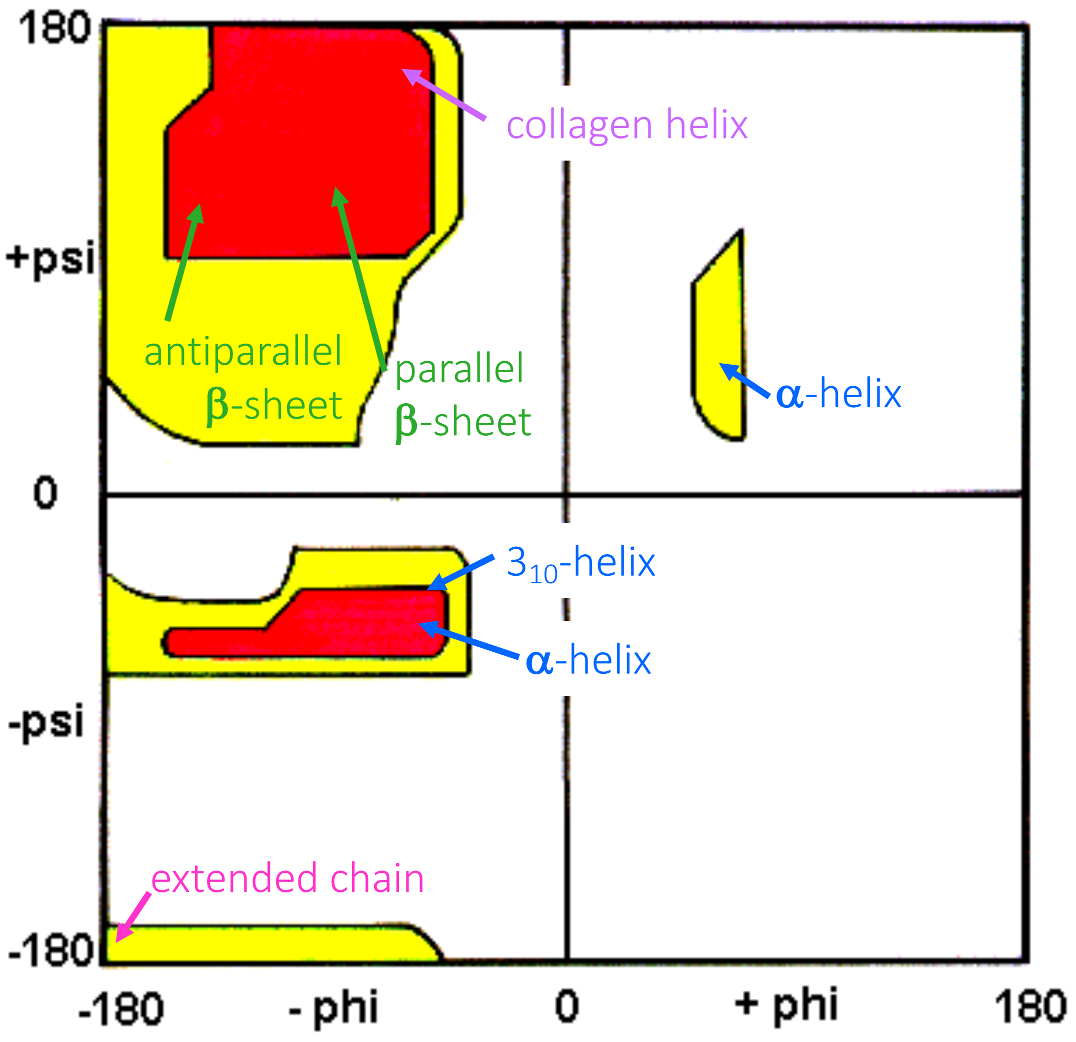




🟥 favorable regions for all aa

🟨 allowed regions for all aa

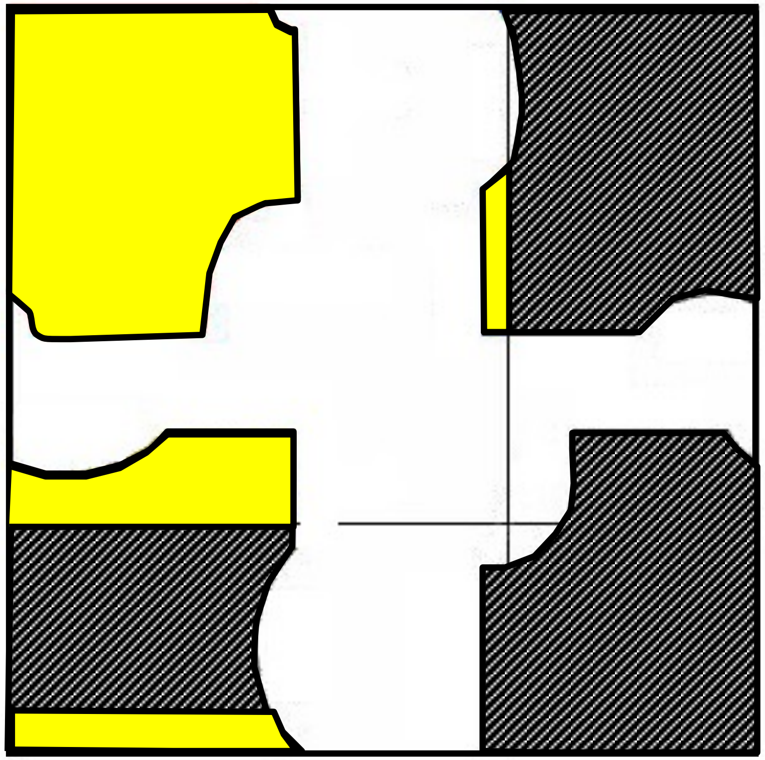
The Ramachandran plot: secondary structure elements



The main chain conformation is defined by the sequence of the (ψ,ϕ) angles: the list of the (ψ,ϕ) for each amino acid dictate the folding of the polypeptide chain, i.e. the 3D structure of the protein

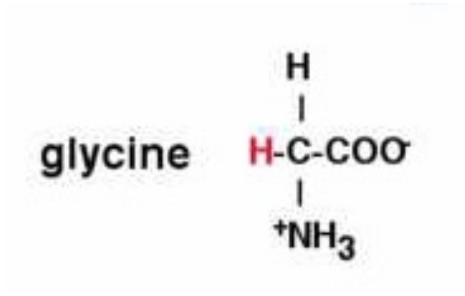
Therefore secondary structure elements will be associated with specific average values of ψ, ϕ and therefore with specific regions of the Ramachandran plot.
For instance, one region corresponds to **α-helices**: the chain coils up because the angles allow a tight spiral without clashes.
Another one corresponds to **β-sheets**: the chain stretches out and aligns with neighbors.
Everything else is rare or unstable because of steric hindrance.

The Ramachandran plot: glycine residues

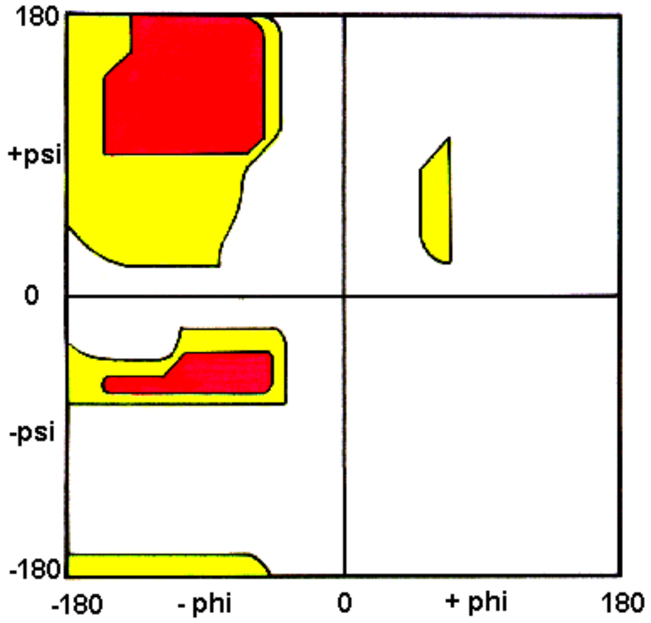


▨ regions allowed only for glyine

🟨 regions allowed only for all aa



side chain=H (very small, hardly any steric hindrance)



Protein architecture

Secondary structure

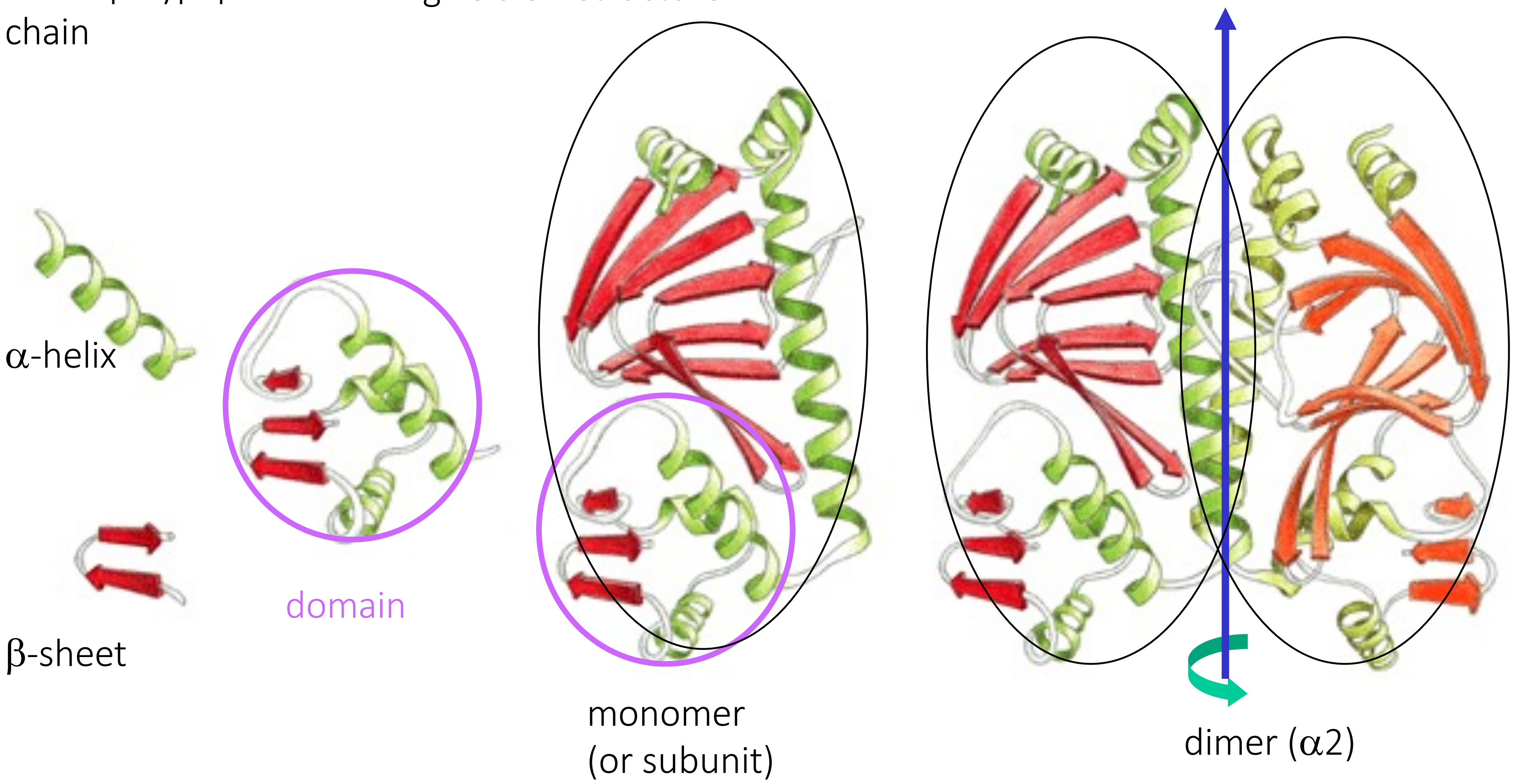
local organisation of the polypeptide chain

Tertiary structure

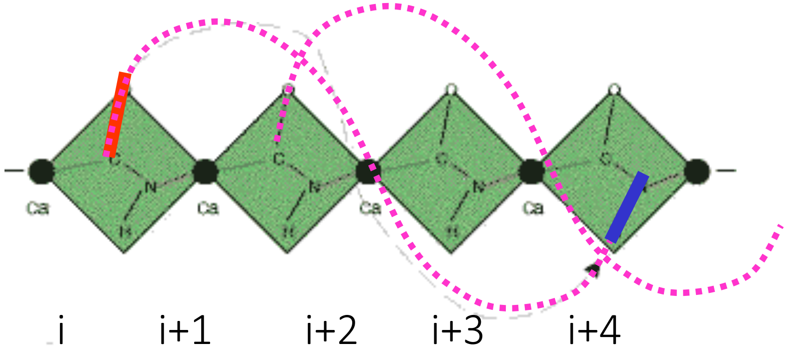
how the secondary structure elements pack together to give a 3D structure

Quaternary structure

the number and relative position of the subunits in a multimeric protein



The α-helix

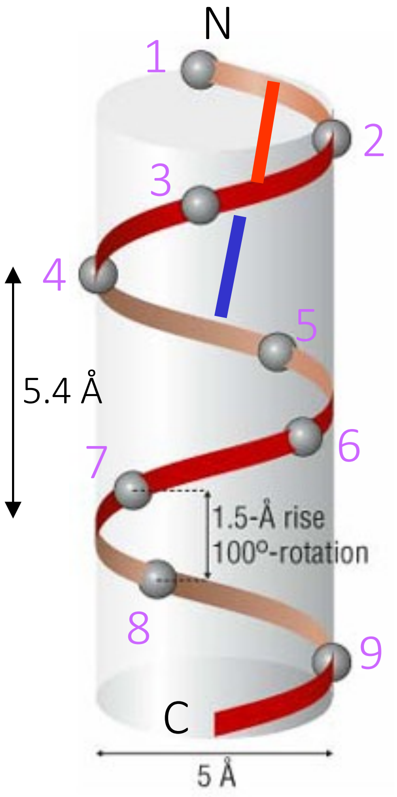


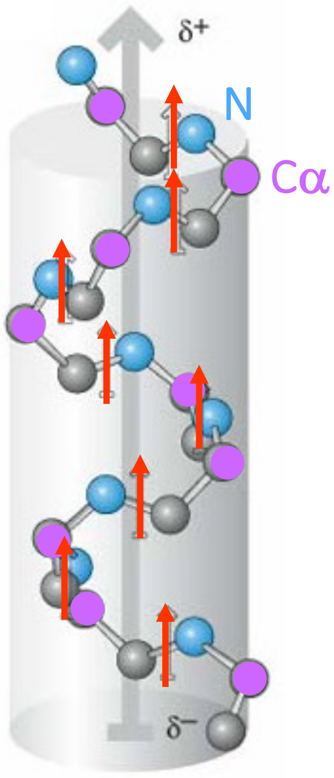
H-bonding pattern $CO_i \mapsto NH_{i+4}$
(local interactions)

all main-chain CO and NH are bonded

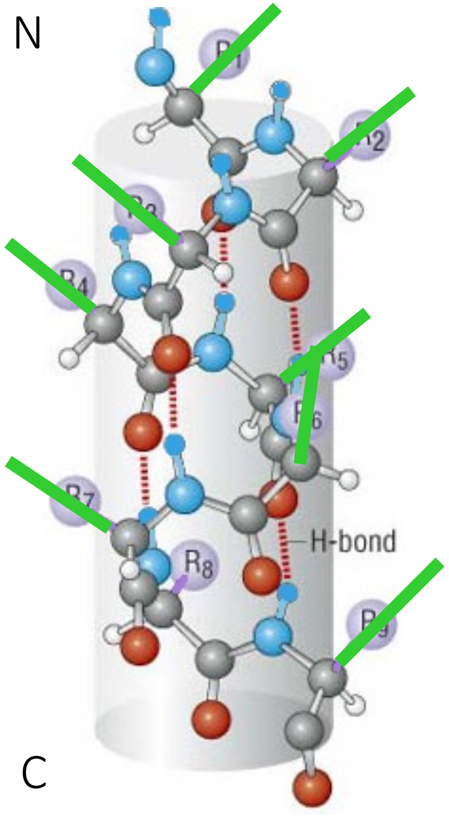
3.6 amino acids per turn;
1.5 Å rise per amino acid
↦ 5.4 Å pitch

each peptide bond has a small dipole moment; in a helix all peptide bonds point in the same direction and generate a dipole pointing towards N





The α-helix



● nitrogen

● oxygen

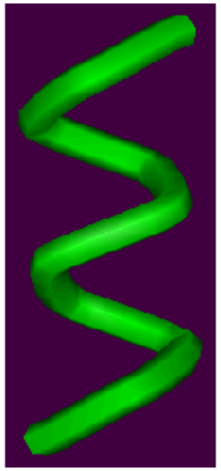
● carbon

● R=side chain

rod-like structure with side chains extending outside

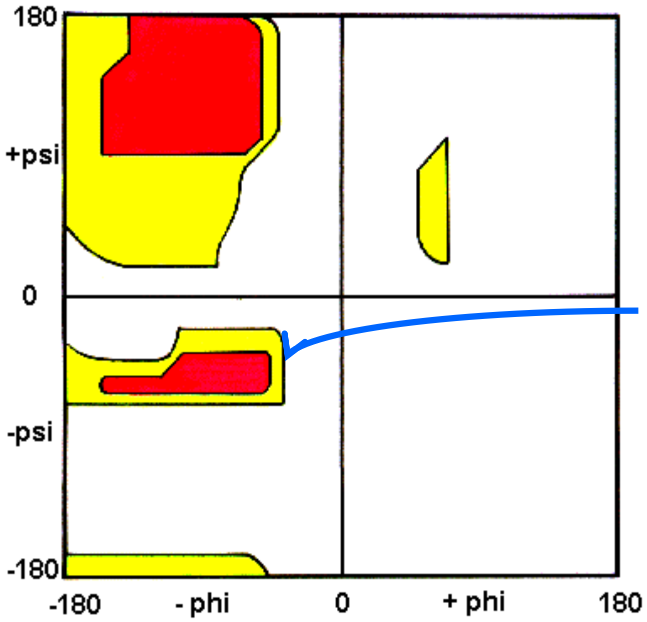
if the helix is oriented so that it goes from N (top) to C (bottom), the side chains point upwards

always right-handed



right

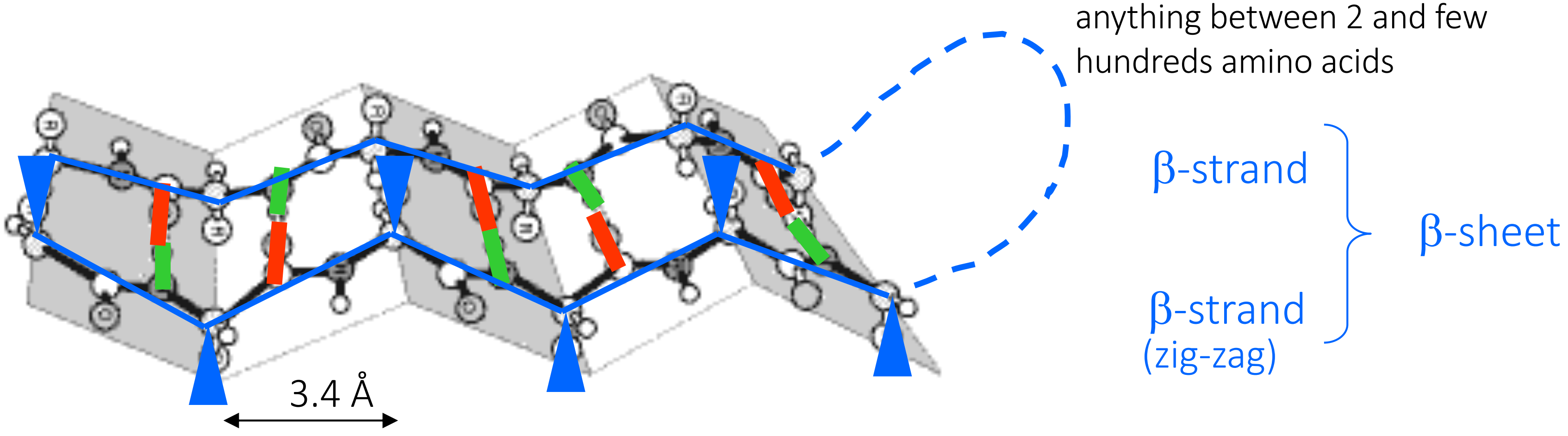
left



right-handed helix

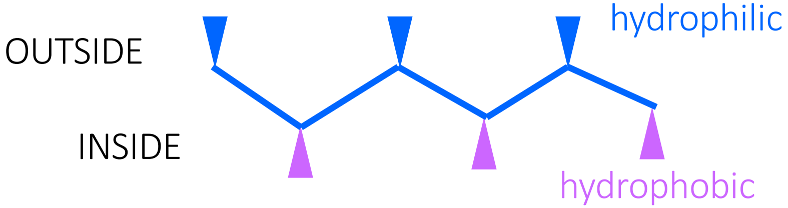
can accommodate all residues except proline

The β-sheet



the polypeptide is almost fully extended (3.4 A per residue)

side chains points alternatively up and down

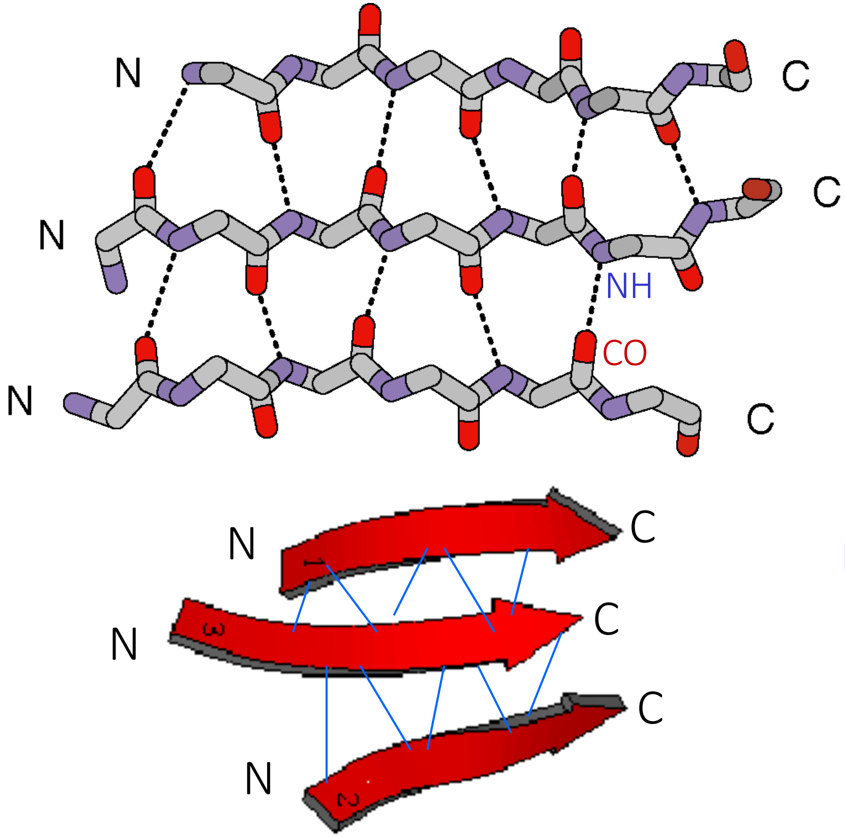


stabilised by main-chain:main-chain NH/CO hydrogen bonds between adjacent strands; contrary to the α-helix these are H bonds between NH/CO groups far apart in the amino-acid sequence

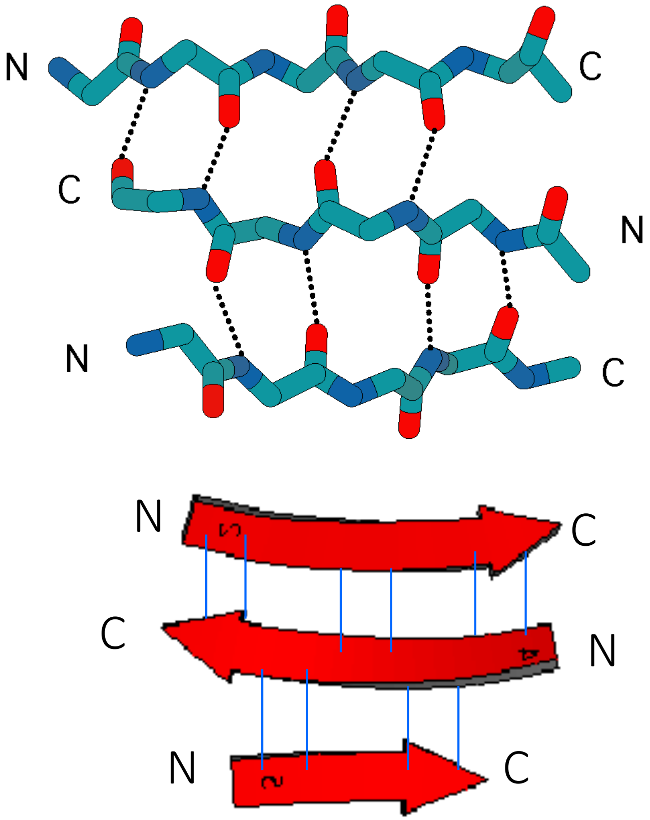
successive α-carbons point alternately above and below the plane of the sheet.

The β-sheet

can be parallel (strands
run in the same direction)



or antiparallel (strands
run in opposite direction)



hydrogen
bonding
pattern

We often have mixed β-sheet, with some strands parallel and some antiparallel.

Tertiary structure:

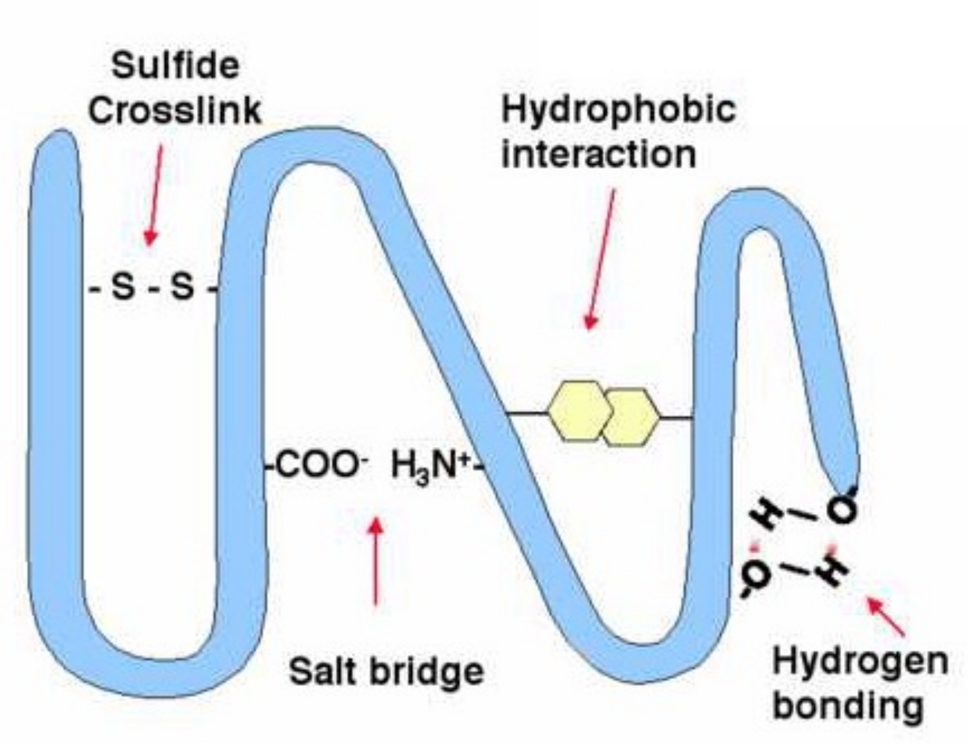
how the secondary structure elements pack together to give a 3D structure

3D structures are held together by “hydrophobic forces” and hydrogen bonds

hydrophobic side chains tend to cluster together in the interior of the protein

polar and charged amino acids interact with each other through hydrogen bonds and ionic interactions or gather on the outside of the protein where they can interact with water molecules

in some proteins S-S bonds and metal ions help to stabilise the 3D structure

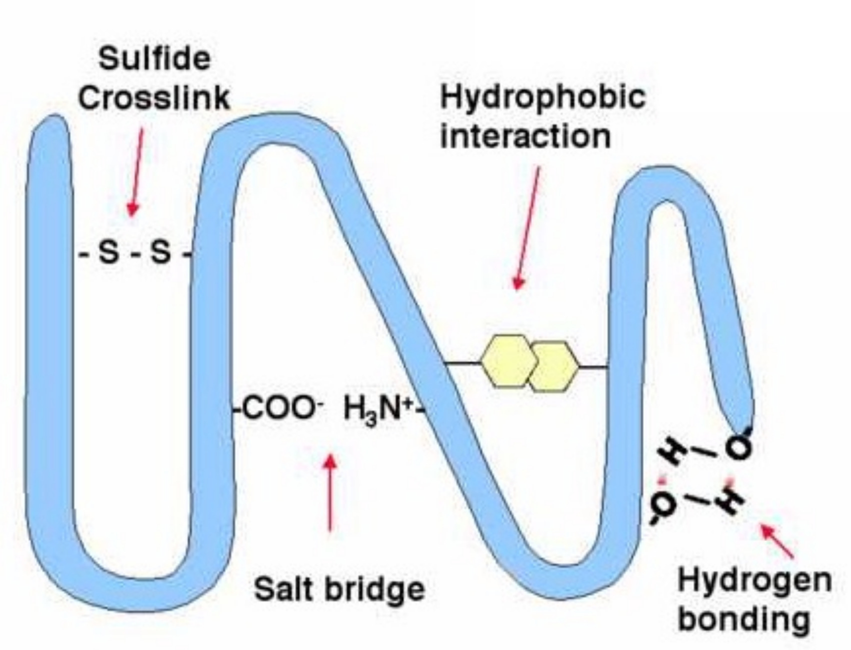


Tertiary structure:

All proteins have a well defined structure. A randomly arranged polypeptide has no biological activity

The function of a protein depends on the structure.

Proteins with similar sequences have similar structures (and similar functions), but not always the opposite is true: proteins with very different sequences can adopt similar conformations!

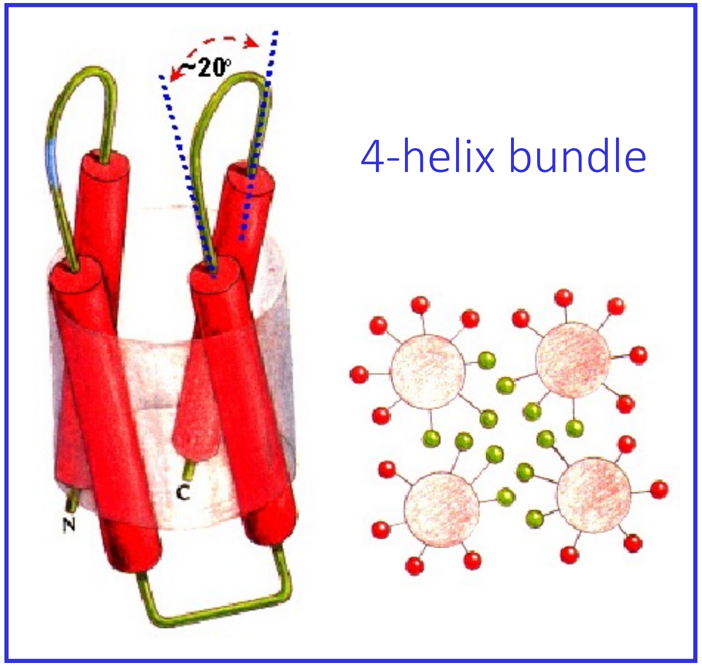


The structure is more conserved than the sequence.

Tertiary structure:
motifs in protein structures

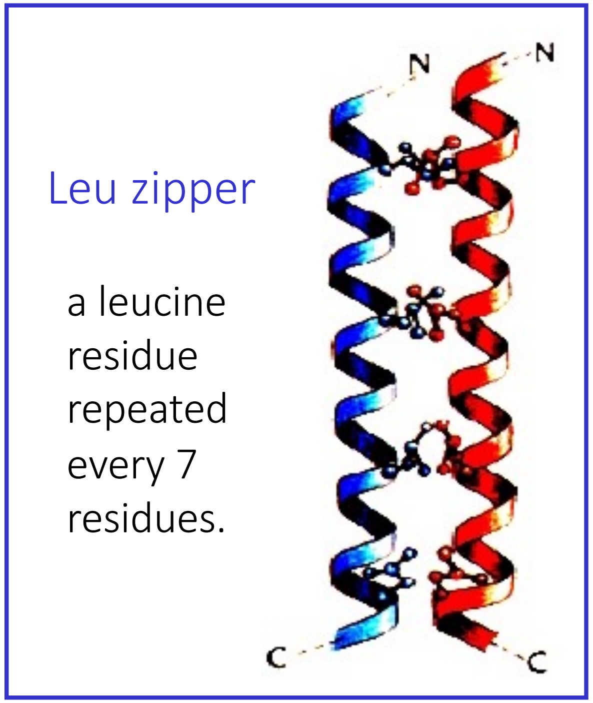
Secondary structure elements are often connected to form structural motifs, i.e some specific geometric arrangements that occur often in protein structures; some of these motifs may be associated with certain functions, others have no specific biological function.

It is difficult to systematically list and classify all the motifs - here are examples of some of the common ones:



4-helix bundle

stabilized by hydrophobic interactions in the core (ferritin)



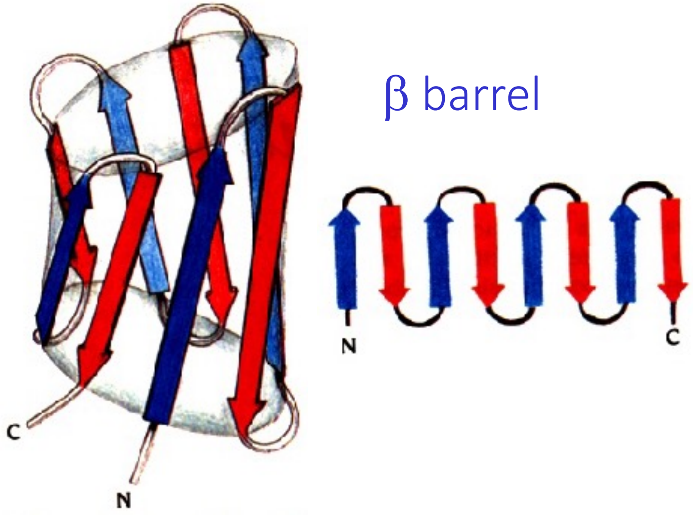
Leu zipper

a leucine residue repeated every 7 residues.

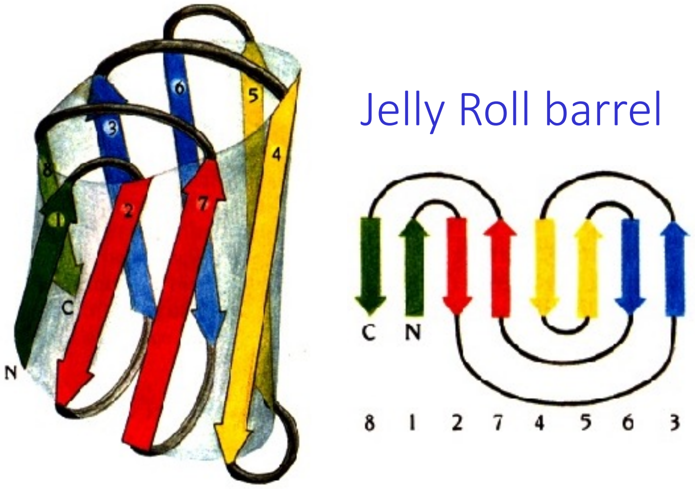
two α-helices from different protein subunits that coil around each other, forming a coiled coil.

Tertiary structure:
motifs in protein structures

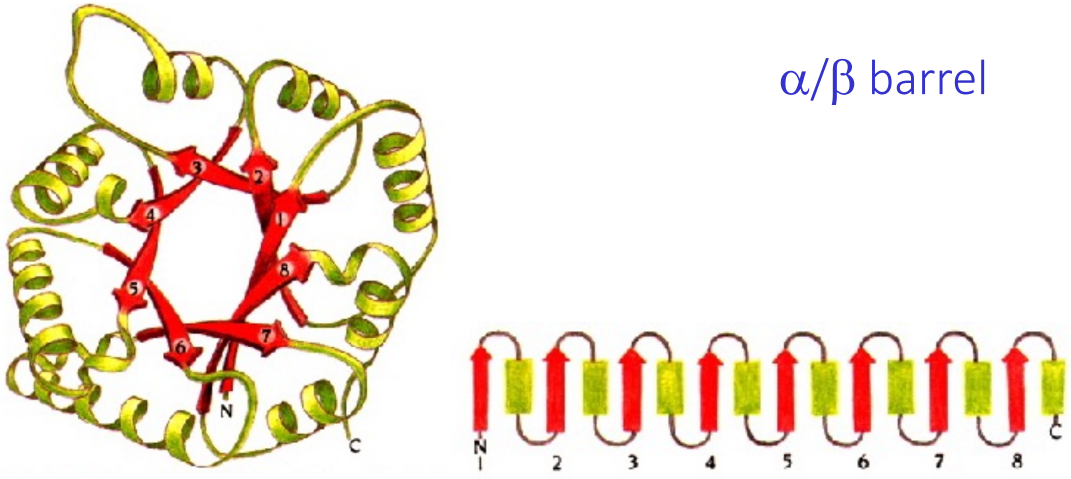
β barrel



Jelly Roll barrel



α/β barrel

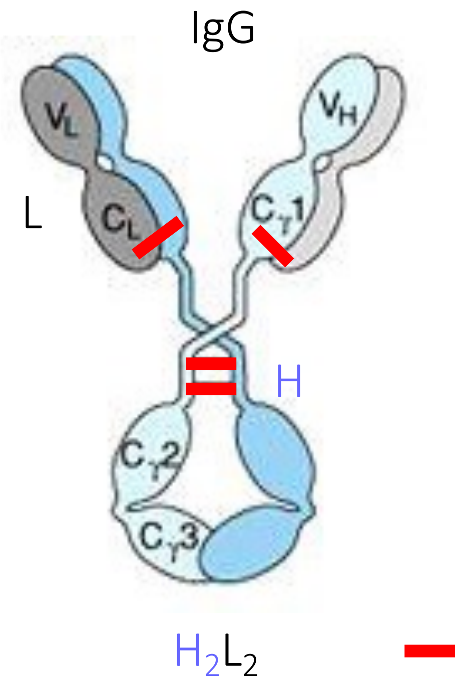


.... and many, many more!!!!

Quaternary structure:

how subunits aggregate to form multimeric proteins

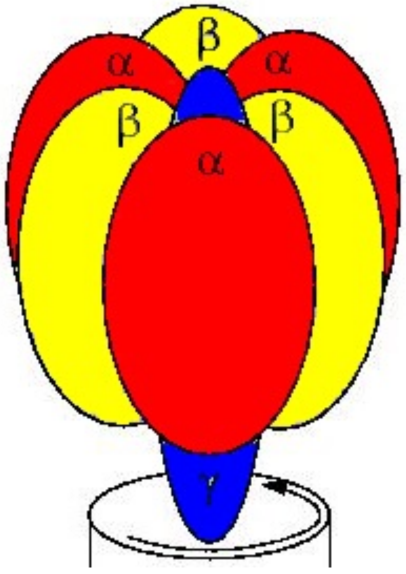
Covalently-linked polypeptide chains



For example an antibody is formed by two copies of a heavy chain H (in blue) and two copies of a light chain (in grey) connected by disulphide bridges

▬ S-S bridges

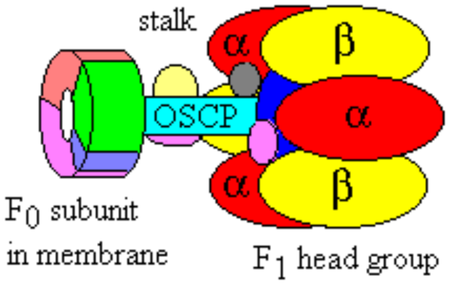
Hetero-multimers: **different** polypeptides aggregating together to form a unit.



$\alpha_3\beta_3\gamma\delta\varepsilon$

An example is the F1 head of the ATP synthase which is formed by 3 α subunits, 3 β subunit and one each of γ, ε, δ subunits.

The entire molecule is even more complex, with a transmembrane portion as well:

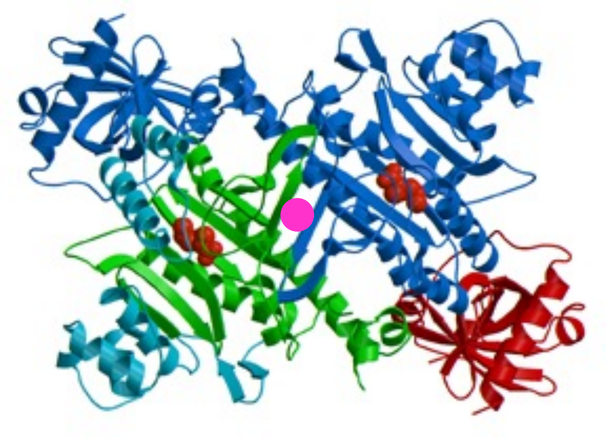


Quaternary structure:

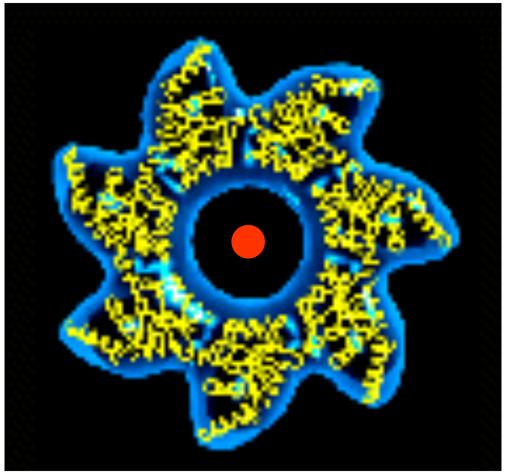
how subunits aggregate to form multimeric proteins

Homo-multimers: multiple copies of the **same** polypeptide associating non-covalently.

Such complexes usually exhibit rotational symmetry about one or more axes, forming dimers, trimers, tetramers, pentamers, hexamers, octamers, decamers, dodecamers, (or even tetradecamers in the case of the chaperonin GroEL).



Lysyl-tRNA synthetase:
● 2-fold axis



GROEL:
7-fold axis ●
72 symmetry

Quaternary structure:

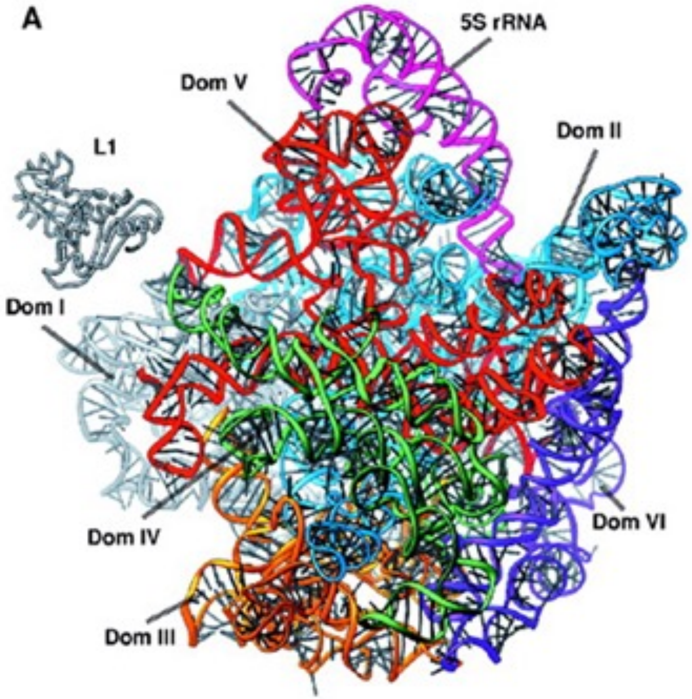
how subunits aggregate to form multimeric proteins

Larger Structures

The molecular machinery of the cell and indeed of assemblies of cells, rely on components made from multimeric assemblies of proteins, nucleic acids, and sugars. A few examples include :

- Viruses
- Microtubules
- Flagellae
- Ribosomes
- Histones

Here is the 3D structure of the large subunit of the ribosome



Fibrous proteins

Triple helix in collagen - next

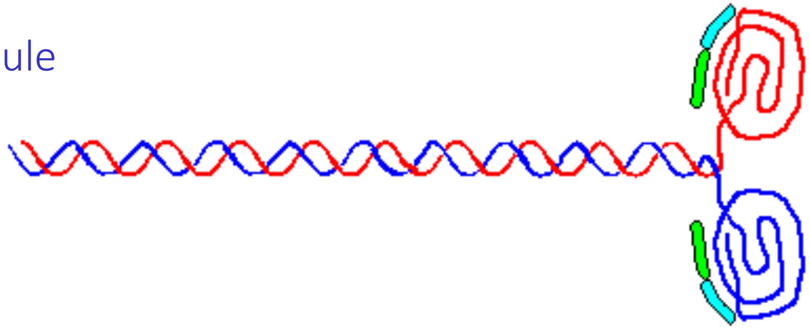
Coiled-coil α-helices present in keratin and myosin:

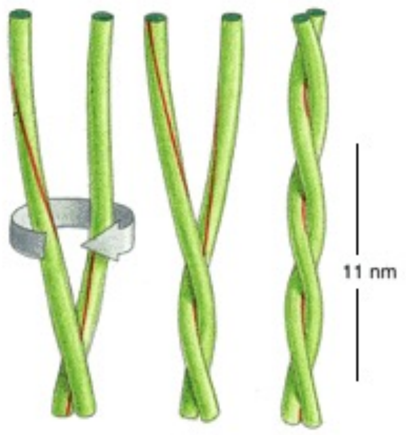
two a-helices twisted around each other to form a left handed coiled coil (7 residues repeat)

Example:a myosin molecule

▬ ▬ heavy chains

▬ ▬ light chains





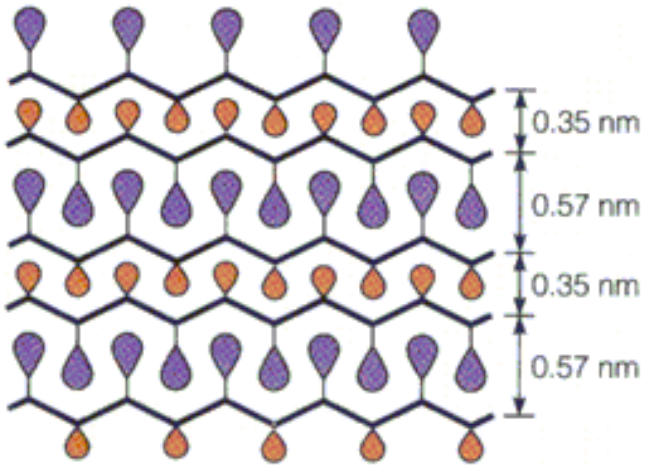
coiled-coil helices

β-sheets in amyloid fibres, spider webs and silk

antiparallel β-sheet whose chains extend parallel to the fibre axis

● Ala or Ser

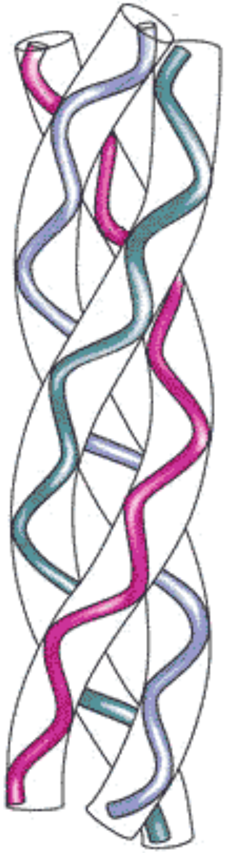
● Gly



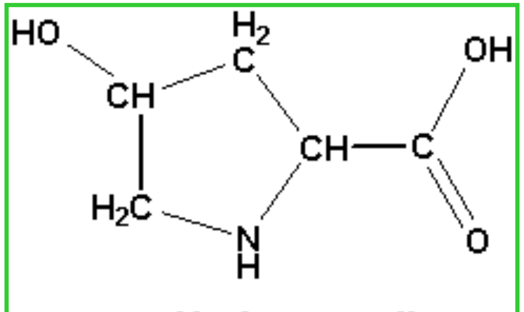
Fibrous proteins: the collagen helix

Collagens are family structural proteins forming the tendons and the extracellular matrix. Bones and teeth are made by adding mineral crystals to collagen.

Collagen is composed of three chains wound together in a triple helix.



Each chain is very long and consists of a repeating sequence of three amino acids: every 3rd amino acid is a glycine that fits in the interior of the triple helix; many of the remaining positions contain prolines and hydroxyprolines:



Hydroxyproline

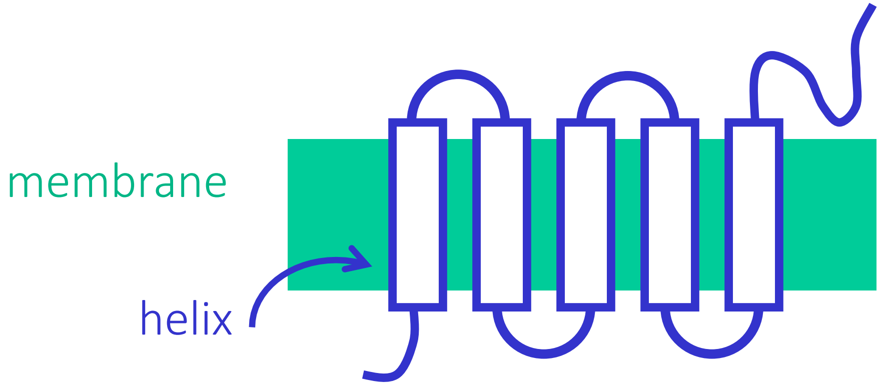
The enzyme that modifies a proline into hydroxyproline requires vitamin C; lack of vitamin C causes scurvy.

There are other non-standard aa (such as hydroxylysines) which are used to crosslink the chains.

Structures of membrane proteins

Less is known about the 3D structure of membrane proteins since in general they are much more difficult to crystallise than soluble proteins.

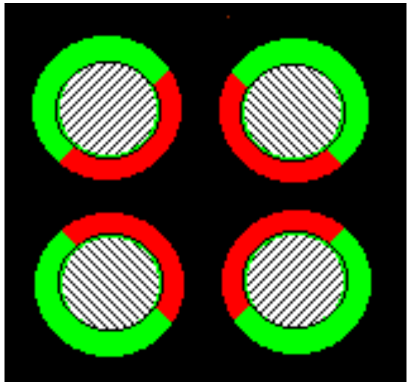
They are often built of α-helices spanning the membrane; but some are built of extended β-barrels (such as porins)



Contrary to soluble proteins, the hydrophobic residues will be on the outside, where they will interact with the chains of the lipids, while hydrophilic side chains will cluster inside

🟥 hydrophilic

🟩 hydrophobic

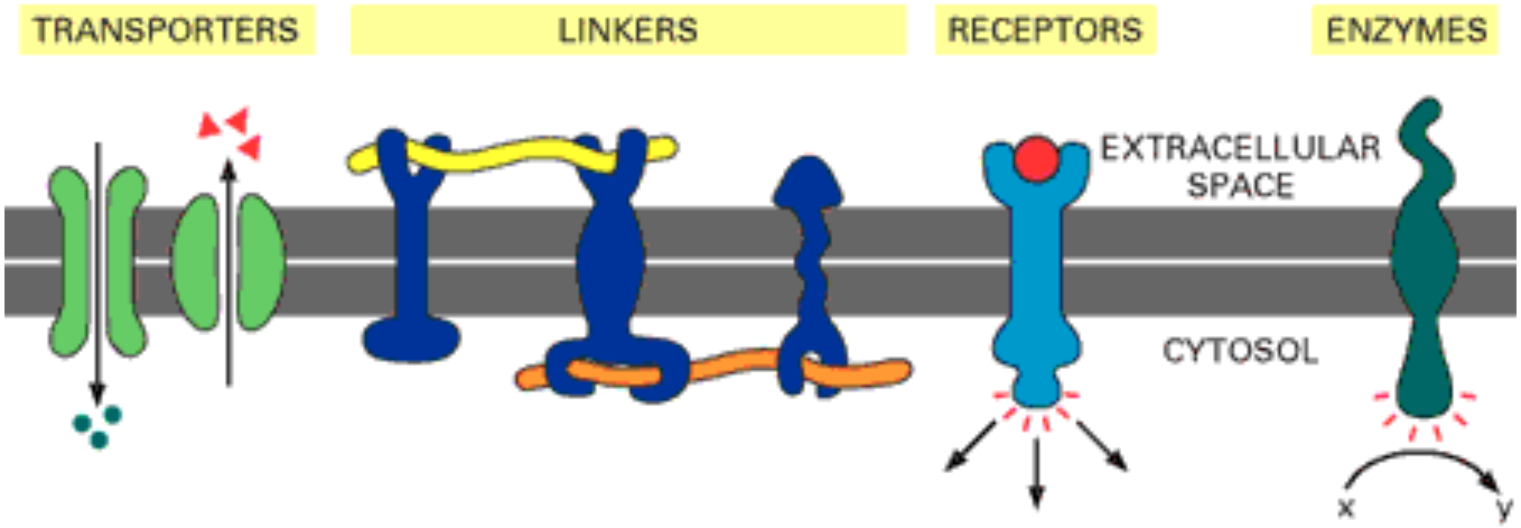


Membrane proteins: biological roles

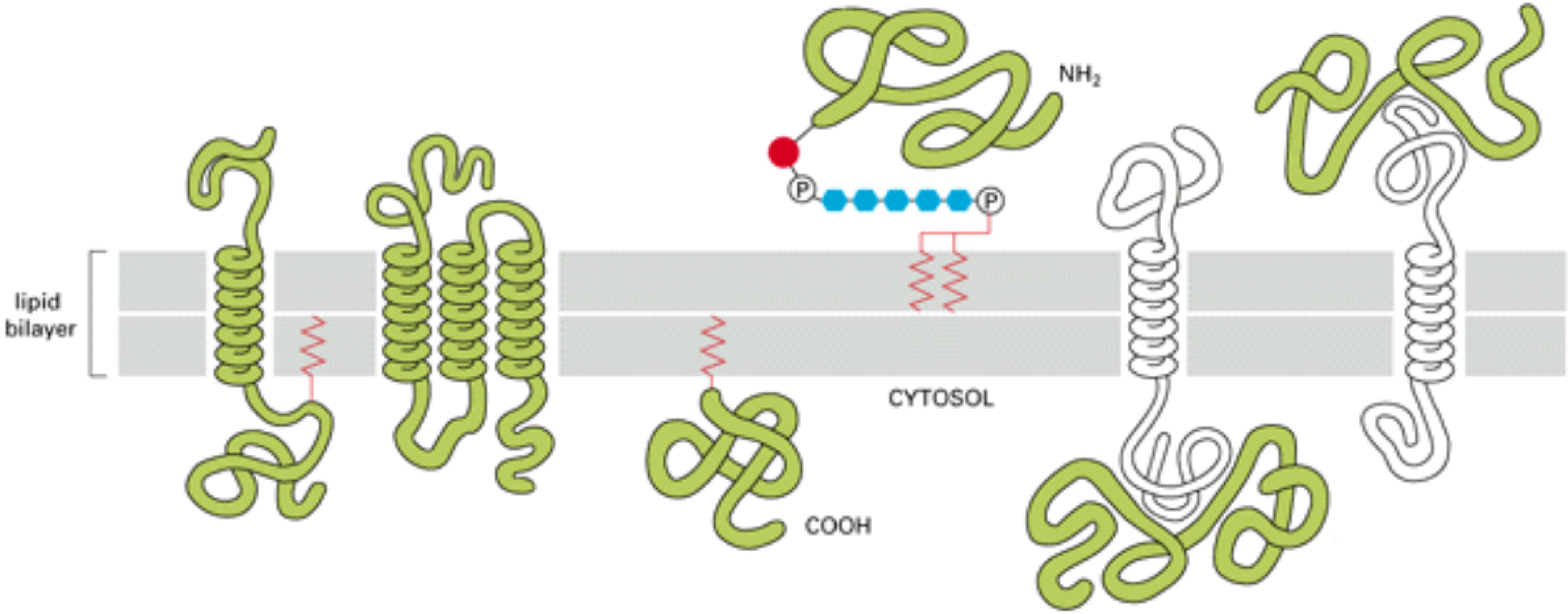
Membrane proteins are defined as proteins that sit in the lipid bilayer: they perform very different biological roles:

- pumps
- channels
- receptors
- cell-to-cell adhesion

control the flow of chemicals and information between the inside and the outside of the cell and mediate communication between different cells.



Membrane proteins associate with the lipid bilayer in various ways:



Noncovalent bonds and folding

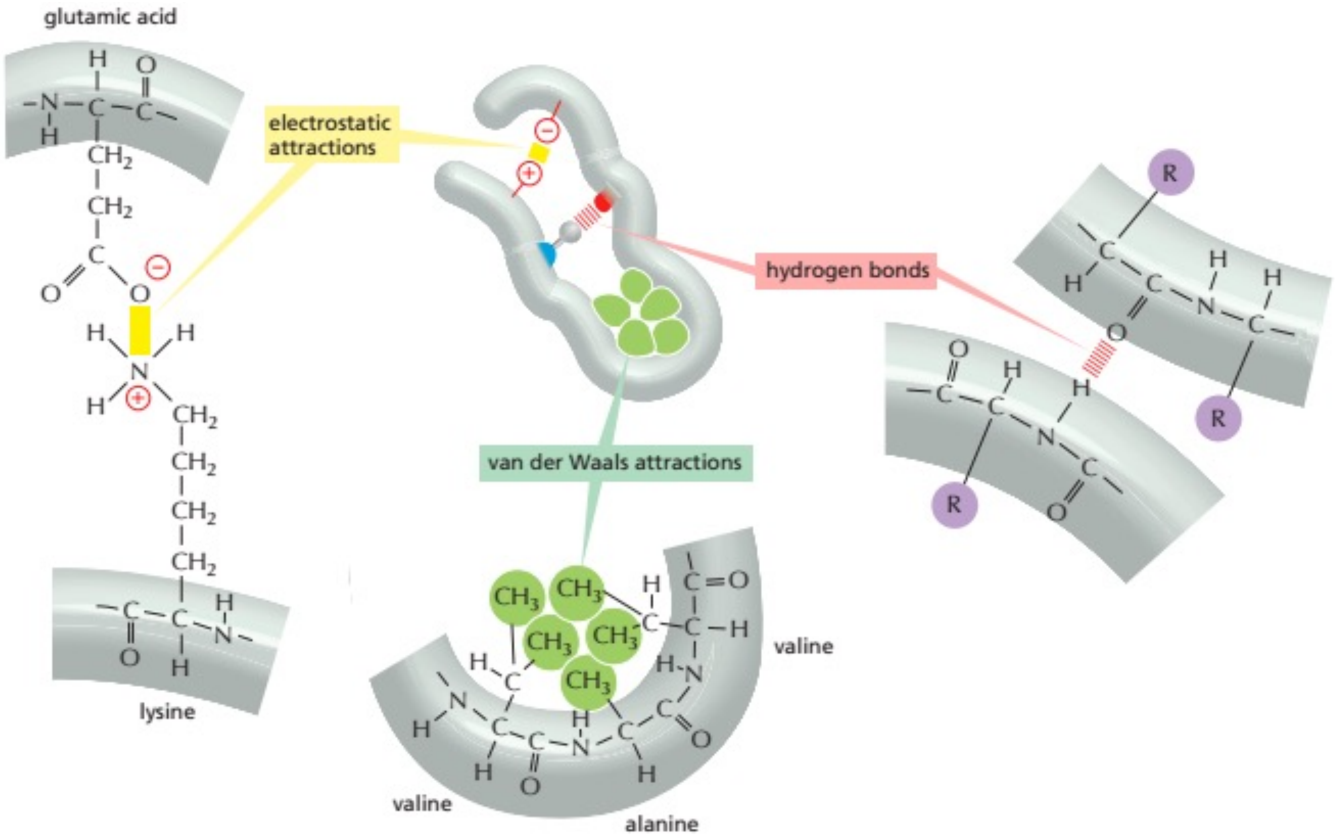


Figure 3–4 Three types of noncovalent bonds help proteins fold. Although a single one of these bonds is quite weak, many of them act together to create a strong bonding arrangement, as in the example shown. As in the previous figure, R is used as a general designation for an amino acid side chain.

Intermolecular forces

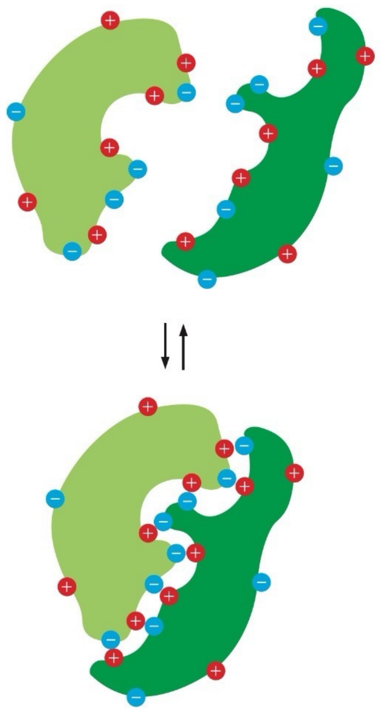
Intermolecular interactions are governed by electromagnetic interactions.
Determine how proteins fold (DNA/RNA, lipid bilayer etc.) and which of its different conformations will predominate; drive ligand-macromolecules association

Quantum mechanical forces:

- Covalent bonds: strength and direction
- Steric, repulsive interactions (i.e. Pauli)

Purely electrostatic (non-covalent) interactions:

- multipole interactions
  - ion-ion
  - ion-dipole
  - dipole-dipole ⇨ hydrogen bond



Polarization interactions:

- induction interaction
- dispersion forces

The final structure will be the result of the interplay of the different forces, and of solute-solute/solute-solvent interactions: complexity!

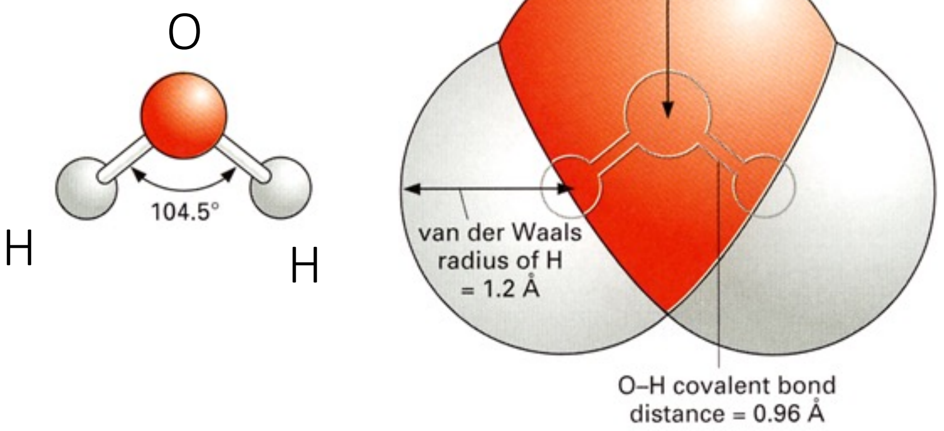
Covalent bonds

Covalent bonds are what hold “molecules” together

- strong (200-800 kJ/mol)
- have well defined lengths
- have well defined directions

⇨ compare with RT =~2.6 kJ/mol at 37º
With R = 1.987 cal/mol ºK
RT=average thermal energy per mole at temperature T

A water molecule (H_2O)

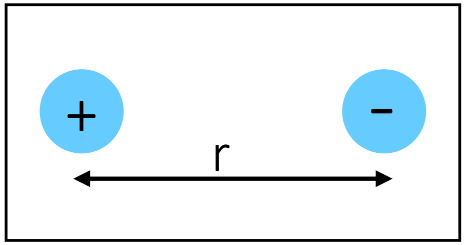


O
H
H
104.5°
van der Waals radius of O = 1.4 Å
van der Waals radius of H = 1.2 Å
O–H covalent bond distance = 0.96 Å

1 kcal/mol = 4.2 kJ/mol = 0.043 eV

The Coulomb potential

ion-ion interactions



$$U = \frac{Q_1 Q_2}{4\pi\varepsilon_0\varepsilon_r \mathbf{r}}$$

50-350 kJ/mol

ϵ_0 = vacuum permittivity
ϵ_r = medium dielectric constant

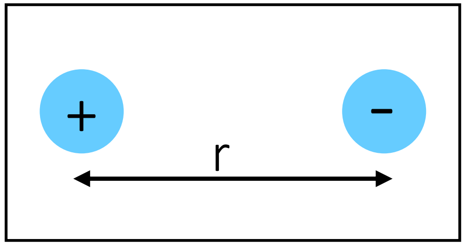
Characterizes the response of the surrounding medium to an electric field: depends on how easily the molecules are polarized

Water has a large value of ϵ_r (about 80). It counteracts the electric field (water mol. are highly polarizable, easily rotate)

In water ϵ_r is strongly T dependent decreasing by 0.46% per degree K near RT. At T= 300K the entropic term -TS = -1.38 G, greater than the free energy G. Therefore, the Coulomb potential is a balance between ion-ion and ion-water molecule interaction. Ions make work on surrounding water forcing them to rotate and orient their dipoles

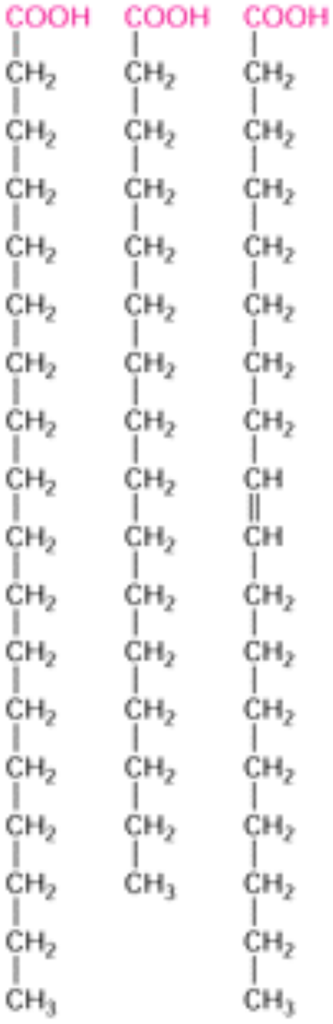
The Coulomb potential

ion-ion interactions



$$U = \frac{Q_1 Q_2}{4\pi\varepsilon_0\varepsilon_r\mathbf{r}}$$

50-350 kJ/mol



characterizes the response of the surrounding medium to an electric field: depends on how easily the molecules are polarized

Hydrocarbons have ε_r of 2: the hydrophobic core of proteins and membranes experiences strong electrostatic interactions

1 kcal/mol = 4.2 kJ/mol = 0.043 eV

Electrostatic self-energy

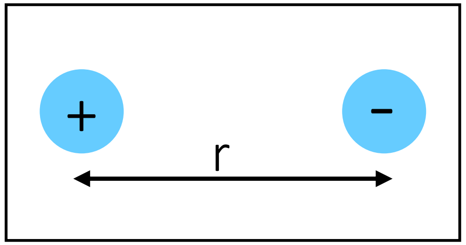
$$G = \frac{1}{\varepsilon_r r}\int_0^q q' dq' = q2/2\varepsilon_r r$$

Is the self-energy of a charge, or the energy of placing an ion in a dielectric medium (calculated from the work done to bring an increment dq’ on the surface of a sphere with radius r and charge q’)

For water, it is the hydration energy.
To transfer a Na+ ion with r = 0.95 Å from water to an hydrocarbon medium (ε goes from 80 to 2), the work necessary is of 85 kcal/mol.
In fact inorganic ions are generally insoluble in organic solvents.
It is difficult to move an ion inside a protein of a lipid bilayer!
Ions are always attracted towards the region with higher ε

Multipole interactions

ion-ion interactions

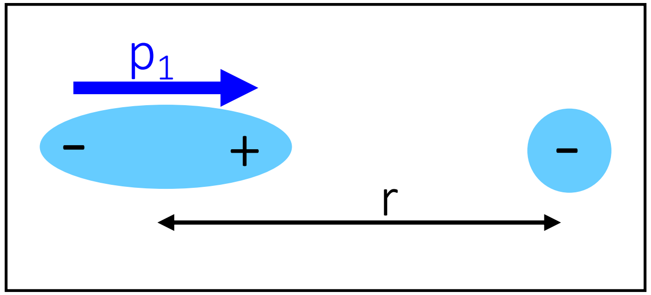


$$\frac{Q_1Q_2}{4\pi\varepsilon_0\varepsilon_r\mathbf{r}}$$

50-350 kJ/mol

Even in neutral molecules, dipoles result from the unequal distribution of e$^-$ due to differences in electronegativity between atoms.

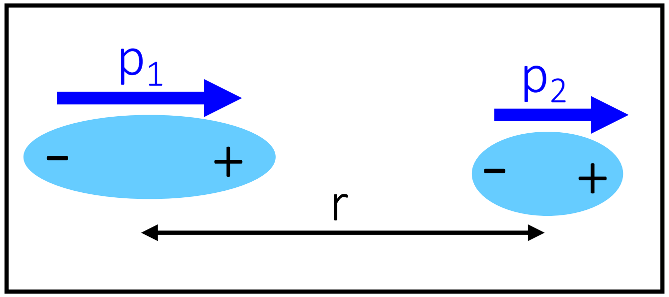
ion-dipole interactions



$$\frac{Q_1p_1}{4\pi\varepsilon_0\varepsilon_r\mathbf{r}^2}$$

1-50 kJ/mol

dipole-dipole interactions



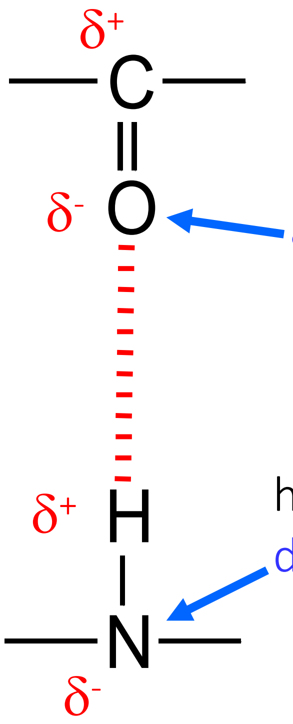
$$\frac{p_1p_2}{4\pi\varepsilon_0\varepsilon_r\mathbf{r}^3}$$

0.1-10 kJ/mol

Hydrogen bond

Hydrogen bonds are a particular case of a dipole-dipole interaction, unusually strong because the small size of the H atom allows the dipoles to come close to each other (~15-30 kJ/mol)

17 kJ/mol—**0.30 nm** bond length
Becomes 4.2 kJ/mol in water!!

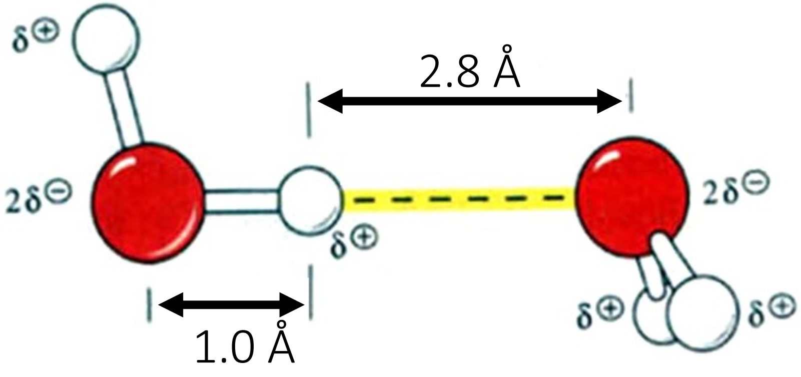


hydrogen-bond acceptor

hydrogen-bond donor

Donors and acceptors must be electronegative atoms (O, N)

Hydrogen bonding in water:



Hydrogen bonds have a defined lenght and orientation

Induction forces

Ions and dipoles can polarise the electron cloud of an adjacent molecule. This causes an attractive force between the ion/permanent dipole and the induced dipole.

Interaction proportional to
- r^{-4} **for ion-induced dipole**
- r^{-6} **for permanent dipole/induced dipole interactions**

Dispersion forces

Random fluctuations of the electron clouds cause temporary dipoles even in uncharged molecules; these temporary dipoles will induce dipoles in the adjacent molecules causing a weak attractive force (He liquefies at 4K).

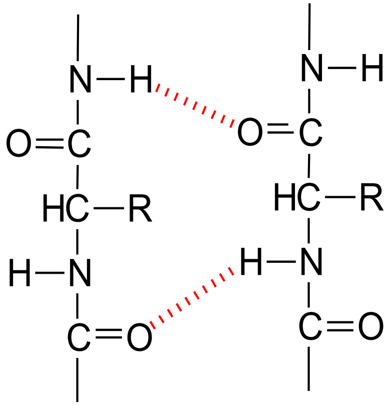
Van der Waals attractive forces!

0.4 kJ/mol—0.35 nm bond length
Does not change in water!!

Hydrogen bonds in biology

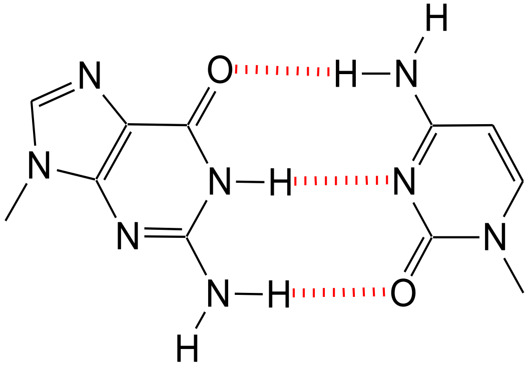
Hydrogen bonding interactions play a fundamental role in determining both the conformation of biological macromolecules and their interactions with other molecules.

The 3D structures of proteins are stabilized by hydrogen bonds between main-chain amide groups:



protein secondary structure: a β-sheet

The pairing of the bases in DNA is mediated by H-bonds:



Guanine-Cytosine base pair

Dispersion forces

Fluctuactions of transient dipole moments can be attractive or repulsive. The attractive configurations have a lower potential E than the repulsive ones, meaning have larger weights in Boltzmann average and therefore a net attraction.

The fluctuactions in the electronic structure responsible for the transient dipole moments are much faster than molecular rotation in liquids. Therefore such forces are not dependent on the specific medium.

Hydrophobic forces

Hydrophobic forces are very relevant in biology. They are primarily driven by an energy cost of creating hydrocarbon-water contact.
There is a reduction of entropy of water close of a hydrophobic surface: water becomes structured, even ice-like. It restricts the possible orientations close to the surface and decrease entropy.

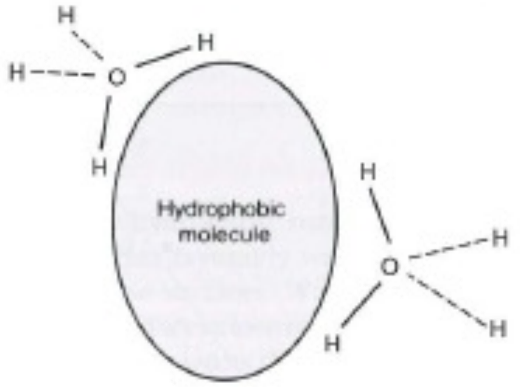


Fig. 2.7 Water molecules adjacent to a hydrophobic molecule suffer restrictions in orientation as they form hydrogen bonds with other water molecules.

Bulk water is a dynamic H-bond network with many rotational degrees of freedom.
Near a hydrophobic surface, water molecules try to **maintain their hydrogen-bond network** by orienting in a more ordered way around the "excluded" region

Hydrophobic forces

Thermodynamic consequences

When a nonpolar solute is dissolved in water:

$\Delta G = \Delta H - T\Delta S$

- $\Delta S < 0$ (unfavorable) → entropy decreases due to structuring.
- ΔH can be close to zero (since there's no strong enthalpic interaction with the solute).
- Thus, solubility of hydrophobic species is poor, because ΔG is often positive.

4. Hydrophobic aggregation

When hydrophobic molecules cluster together:

- The water molecules that were "frozen" around each solute are released back to bulk water.
- This restores entropy ($\Delta S > 0$ for the system).
- That's why hydrophobic groups spontaneously aggregate (micelle formation, protein folding).

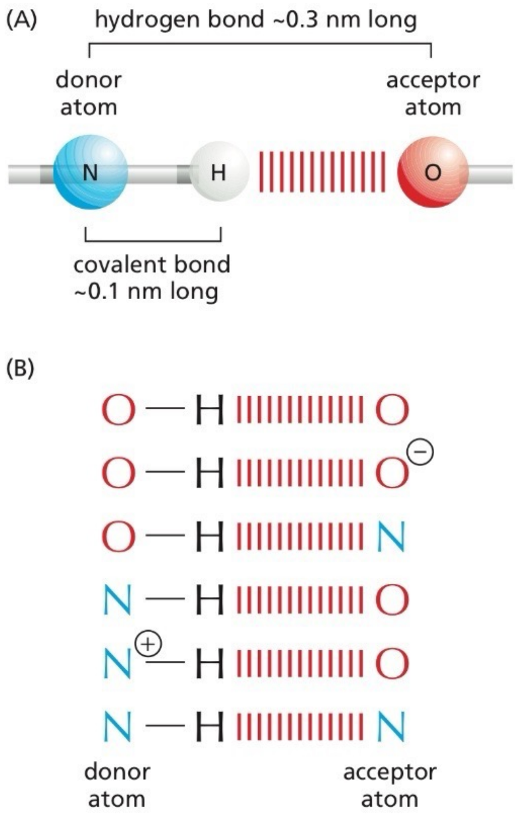


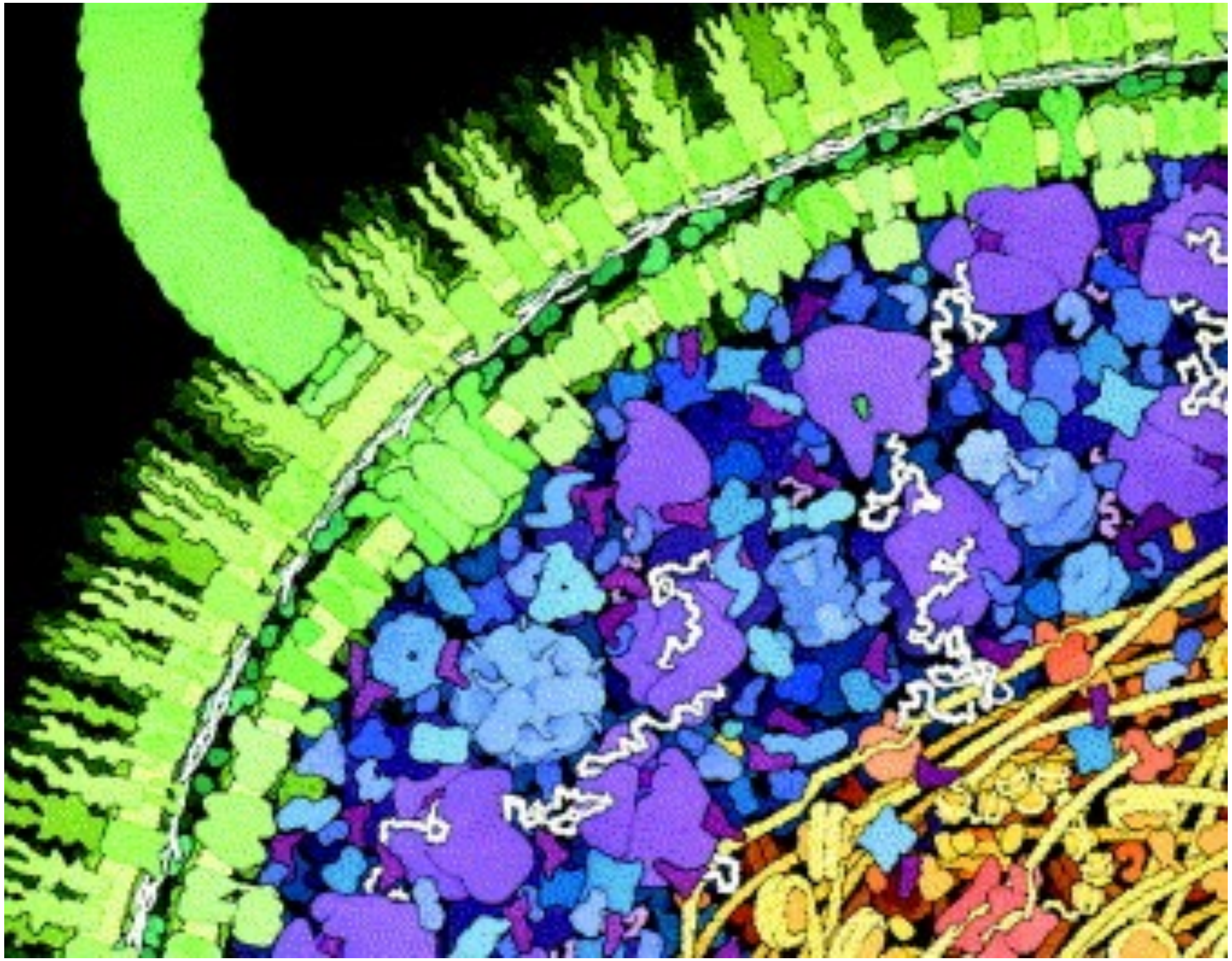
Figure 2–4 Hydrogen bonds. (A) Ball-and-stick model of a typical hydrogen bond. The distance between the hydrogen and the oxygen atom here is less than the sum of their van der Waals radii, indicating a partial sharing of electrons. (B) The most common hydrogen bonds in cells.

TABLE 2–1 Covalent and Noncovalent Chemical Bonds

			Strength kJ/mole**	
Bond type		Length (nm)	in vacuum	in water
Covalent		0.15	377 (90)	377 (90)
Noncovalent	ionic*	0.25	335 (80)	12.6 (3)
	hydrogen	0.30	16.7 (4)	4.2 (1)
	van der Waals attraction (per atom)	0.35	0.4 (0.1)	0.4 (0.1)

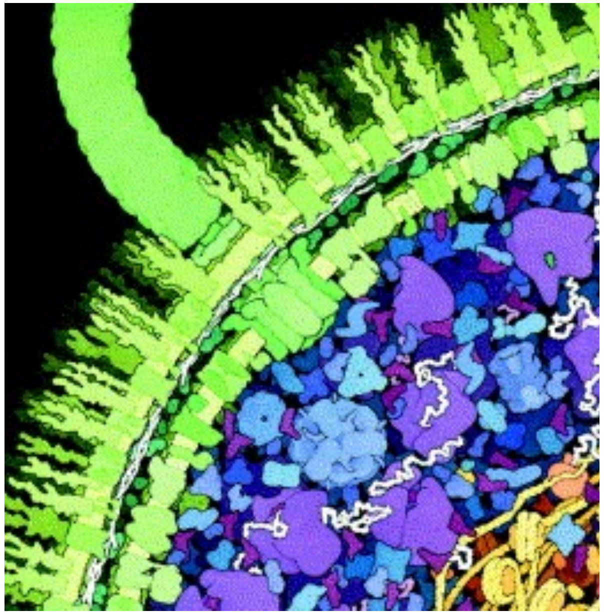
*An ionic bond is an electrostatic attraction between two fully charged atoms. **Values in parentheses are kcal/mole. 1 kJ = 0.239 kcal and 1 kcal = 4.18 kJ.

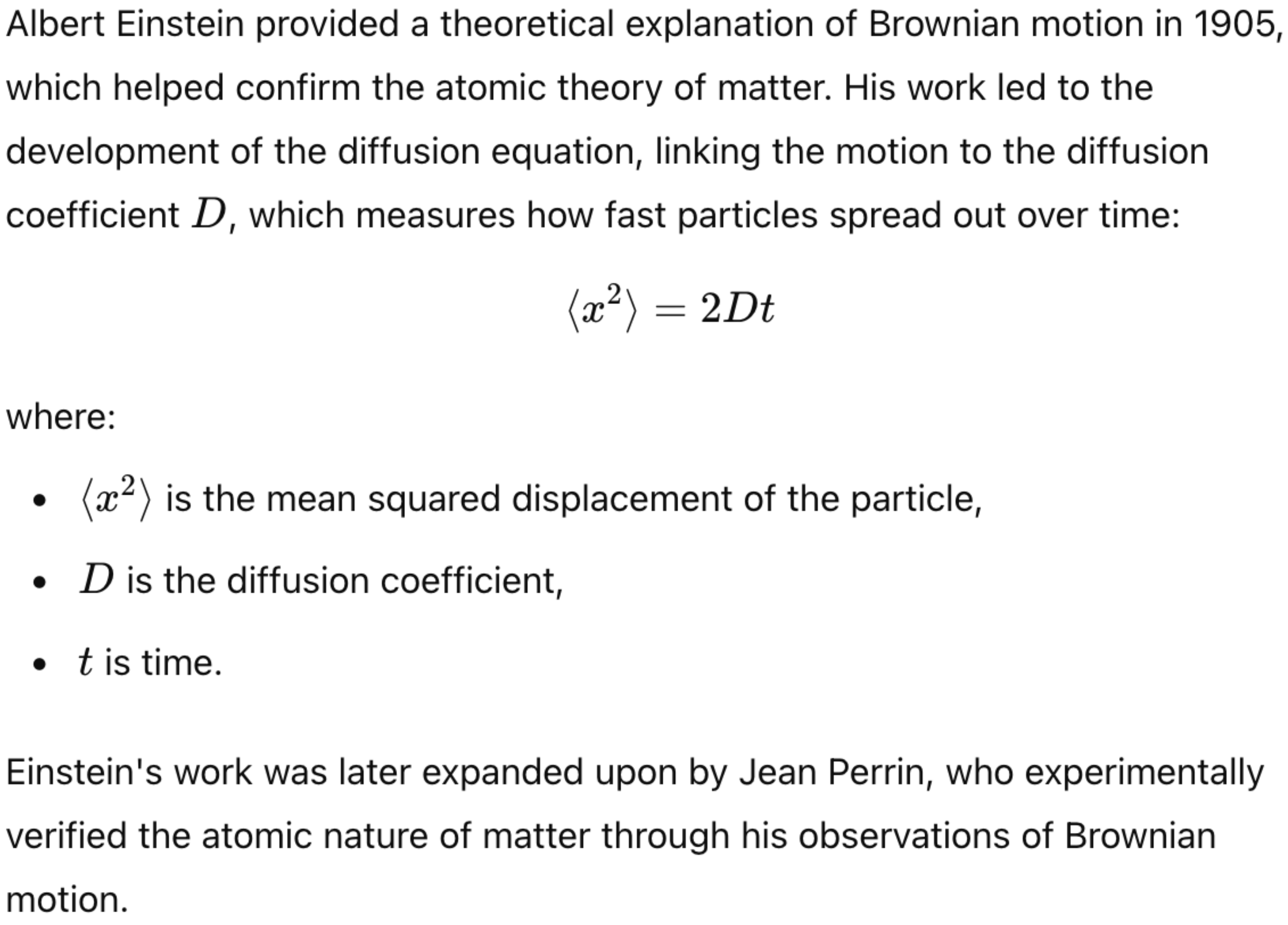
Cell crowding and diffusion constant



Cell crowding and diffusion constant

Brownian motion





Albert Einstein provided a theoretical explanation of Brownian motion in 1905, which helped confirm the atomic theory of matter. His work led to the development of the diffusion equation, linking the motion to the diffusion coefficient D, which measures how fast particles spread out over time:

$$\langle x^2 \rangle = 2Dt$$

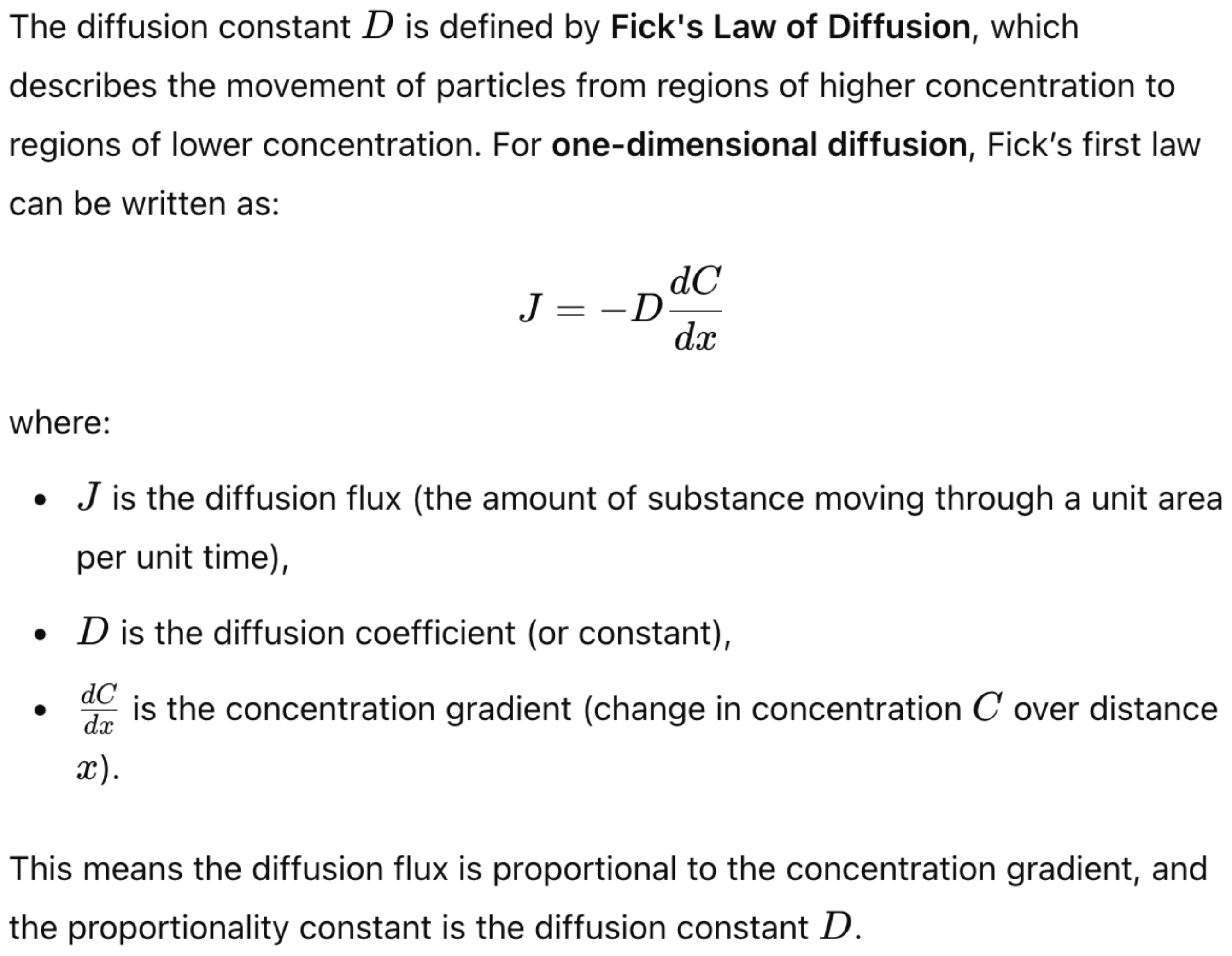
where:

- $\langle x^2 \rangle$ is the mean squared displacement of the particle,
- D is the diffusion coefficient,
- t is time.

Einstein's work was later expanded upon by Jean Perrin, who experimentally verified the atomic nature of matter through his observations of Brownian motion.

D measures the rate of diffusion
is expressed as unit of area per unit of time!

Diffusion constant



The diffusion constant D is defined by **Fick's Law of Diffusion**, which describes the movement of particles from regions of higher concentration to regions of lower concentration. For **one-dimensional diffusion**, Fick's first law can be written as:

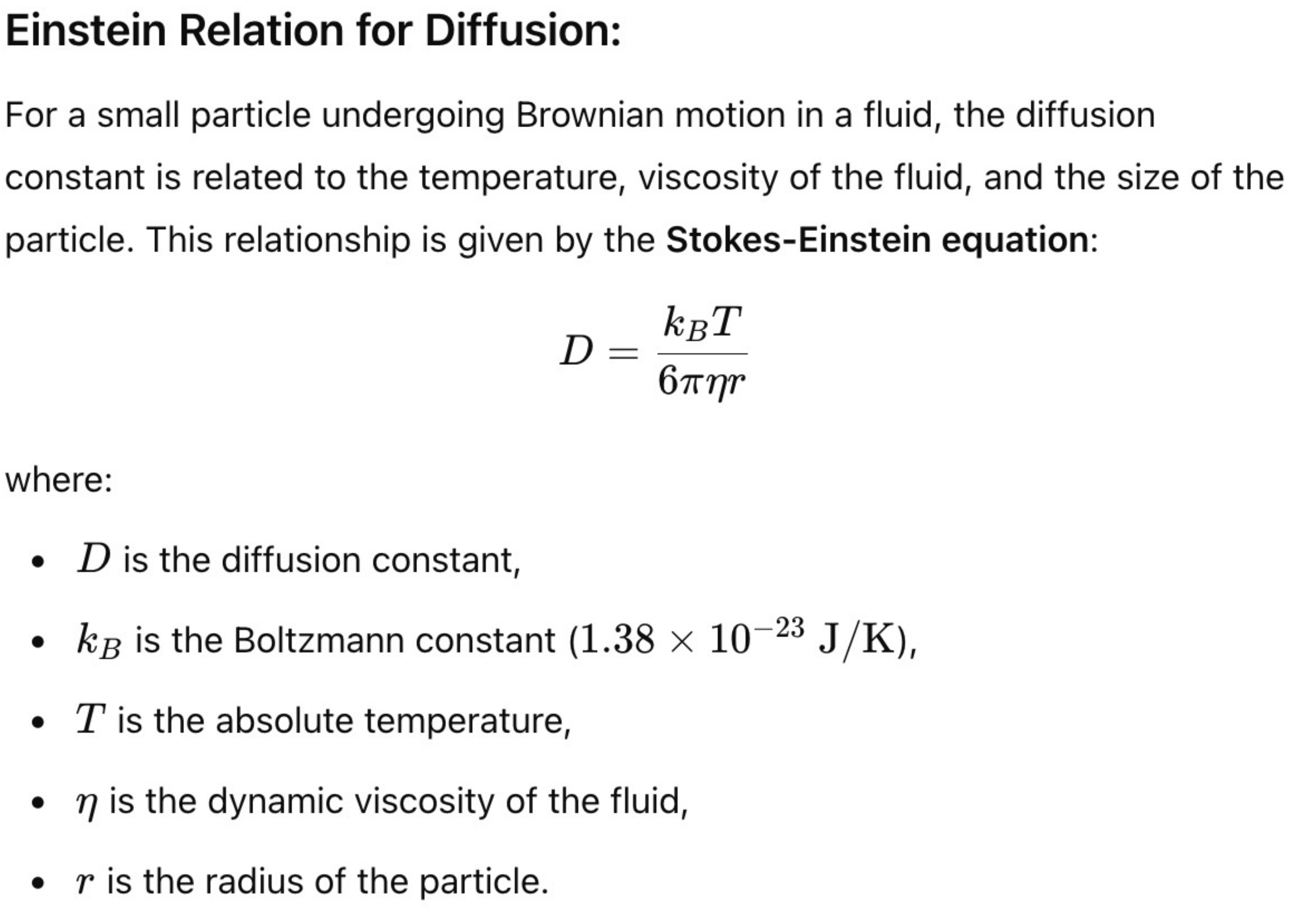
$$J = -D\frac{dC}{dx}$$

where:

- J is the diffusion flux (the amount of substance moving through a unit area per unit time),
- D is the diffusion coefficient (or constant),
- $\frac{dC}{dx}$ is the concentration gradient (change in concentration C over distance x).

This means the diffusion flux is proportional to the concentration gradient, and the proportionality constant is the diffusion constant D.

Diffusion constant



Einstein Relation for Diffusion:

For a small particle undergoing Brownian motion in a fluid, the diffusion constant is related to the temperature, viscosity of the fluid, and the size of the particle. This relationship is given by the **Stokes-Einstein equation**:

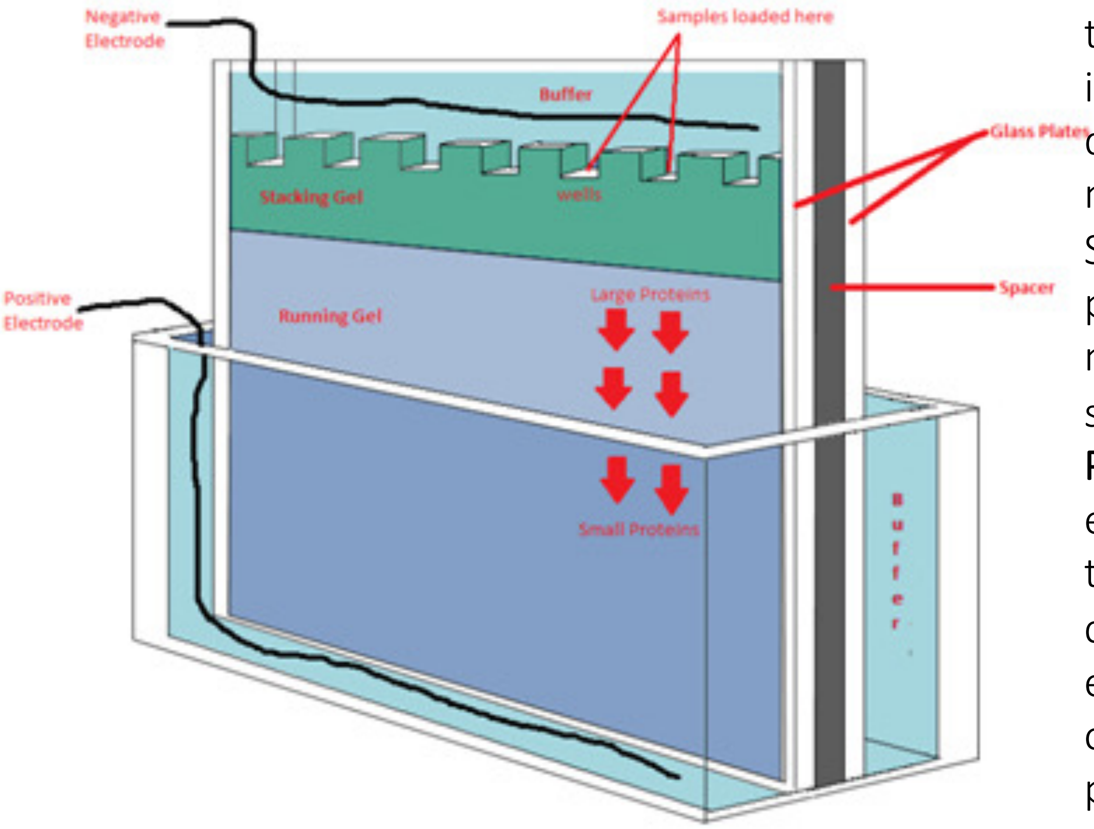
$$D = \frac{k_B T}{6\pi\eta r}$$

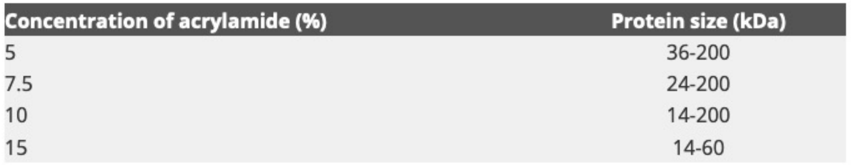
where:

- D is the diffusion constant,
- k_B is the Boltzmann constant ($1.38 \times 10^{-23}\ \mathrm{J/K}$),
- T is the absolute temperature,
- η is the dynamic viscosity of the fluid,
- r is the radius of the particle.

Mass of a protein

1D –SDS-PAGE
Sodium Dodecyl Sulphate - PolyAcrylamide Gel
Electrophoresis





Concentration of acrylamide (%)	Protein size (kDa)
5	36-200
7.5	24-200
10	14-200
15	14-60

a method that separates protein by molecular weight over a range of about 10 to 300 kilodaltons (kDa). Samples are weighed and dissolved in sodium dodecyl sulfate (SDS). SDS is a negatively charged detergent that has both hydrophilic and hydrophobic regions. SDS likes to bind to proteins (1.4 g SDS/1 g protein) and to be in water. This SDS- protein-water interaction allows water insoluble proteins to dissolve in water, and to dissolve protein mixtures.

SDS confers uniform negative charge to proteins, masking their native charges and making the separation dependent on molecular size, rather than charge or shape.

Proteins are completely denatured. When an electric field is applied, the negative charge of the SDS causes the proteins to move through a clear acrylamide matrix toward the positive electrode. This matrix has holes in it that sieve out the proteins by molecular weight. Large proteins move more slowly through the matrix than the smaller proteins thereby separating proteins by molecular weight.

1 Da = 1 g/mol Average mol. weigth of 1 aminoacid: 110 Da