

The background of the slide features several overlapping, semi-transparent DNA double helix structures. These structures are rendered in a light blue-grey color and are positioned at various angles and depths, creating a sense of three-dimensional space. The helices are composed of two strands connected by rungs representing base pairs.

Prediction of protein and RNA structure

“Two quite opposite qualities equally bias our minds – habits and novelty.”

(Jean de La Bruyère)

Table of contents

- ✦ Amino acids
- ✦ The composition of polypeptides
- ✦ Protein secondary structure
- ✦ Protein tertiary and quaternary structure
- ✦ Protein folding algorithms
- ✦ RNA secondary structure prediction

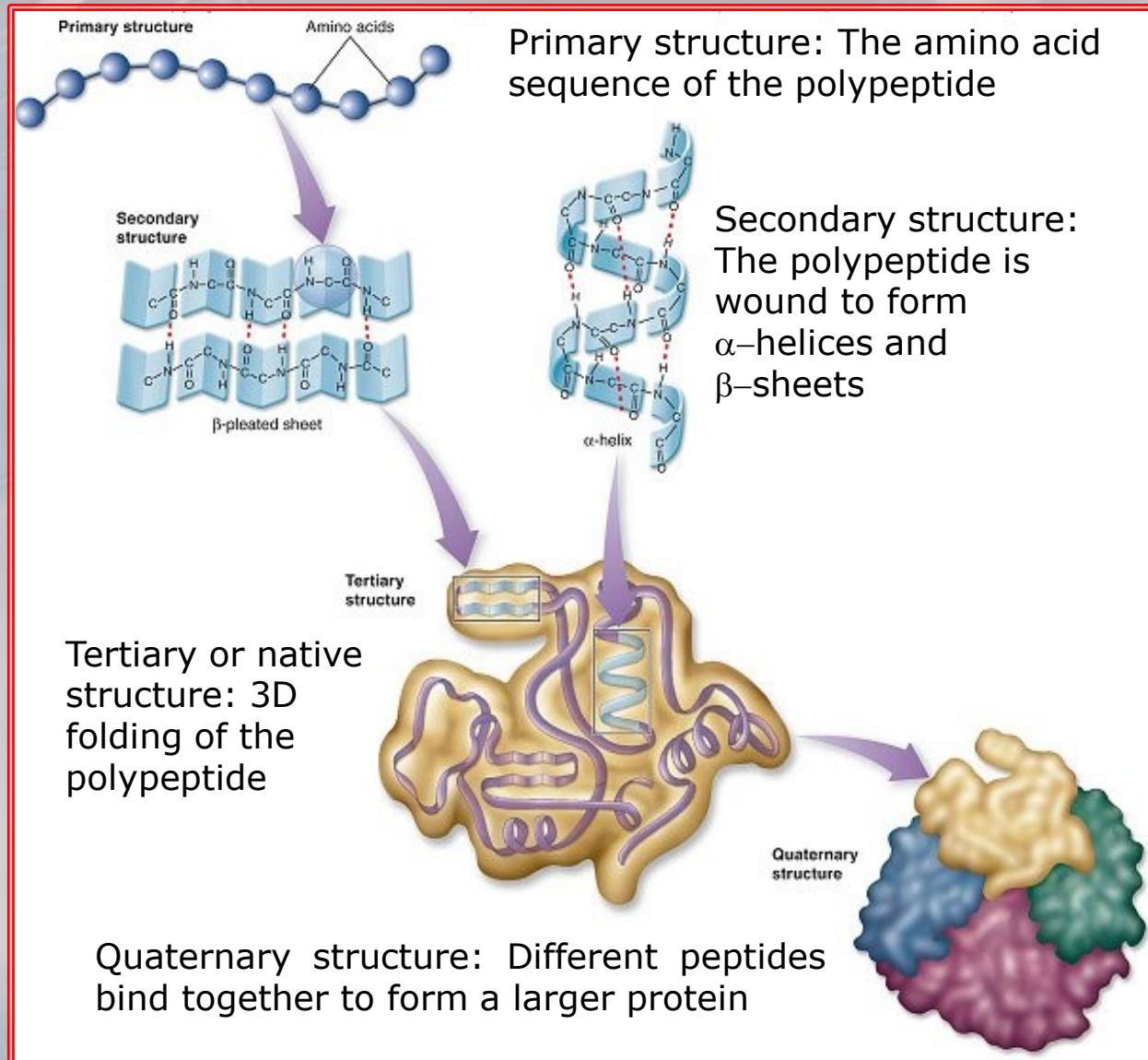
Introduction – 1

- ✦ Proteins are the molecular mechanism that oversees, regulates and executes all biological functions in living organisms
 - **Structural proteins**, such as **collagen**, maintain and strengthen our connective tissues
 - **Mechanoenzymes**, such as skeletal muscle **myosin**, are responsible for movements, both on a microscopic and a macroscopic scale
 - Several different types of enzymes catalyze different chemical reactions, activating and governing digestion and metabolism, immune system, reproduction, and an astonishing assortment of other functions
- ✦ Protein interaction with DNA and RNA molecules allows the production of new proteins and regulates their expression levels, responding appropriately to changes in both the internal and the external environment

Introduction – 2

- ✦ Proteins are synthesized as linear amino acid chains, but, *in vivo*, they fold up quickly in a compact and globular form
- ✦ In the late '60s, C. B. Anfinsen was the first to prove that, when unfolded, or **denatured**, proteins repeatedly take the same conformation when they are left free to fold
 - This **native structure** is essential for their biological function: Only when they are folded into their native, globular structure, proteins are biologically active
 - **Problem:** Only AlphaFold reliably predicts the three-dimensional shape of a protein, starting from its amino acid sequence, but it is computationally unaffordable, apart from Google

Introduction – 3

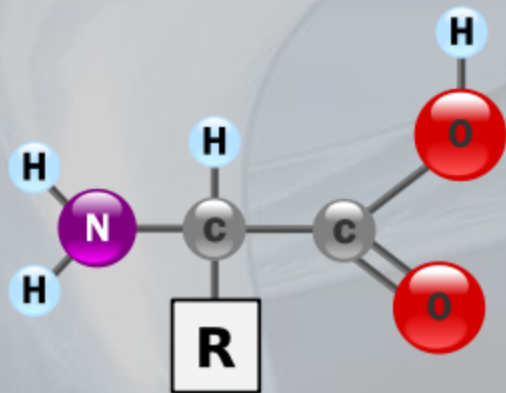


Amino acids – 1

- ✦ **Amino acids** are the basic building blocks of proteins
- ✦ Such as DNA and RNA, proteins are synthesized as linear polymers (chains), composed by smaller molecules
- ✦ Unlike DNA and RNA, whose alphabet is constituted by four nucleotides, proteins are made up by twenty amino acids, with various size, shape and chemical properties

Amino acids – 2

- ✦ Each amino acid has a **main chain** or a **backbone**, consisting of an **amino group** ($-\text{NH}_2$), an **alpha-carbon**, C_α , and a **carboxyl group** ($-\text{COOH}$)
- ✦ To the alpha-carbon, a **side chain** is attached (often denoted by $-\text{R}$)
- ✦ The side chain significantly changes from an amino acid to another, a feature that confers unique stereochemical properties to each of them

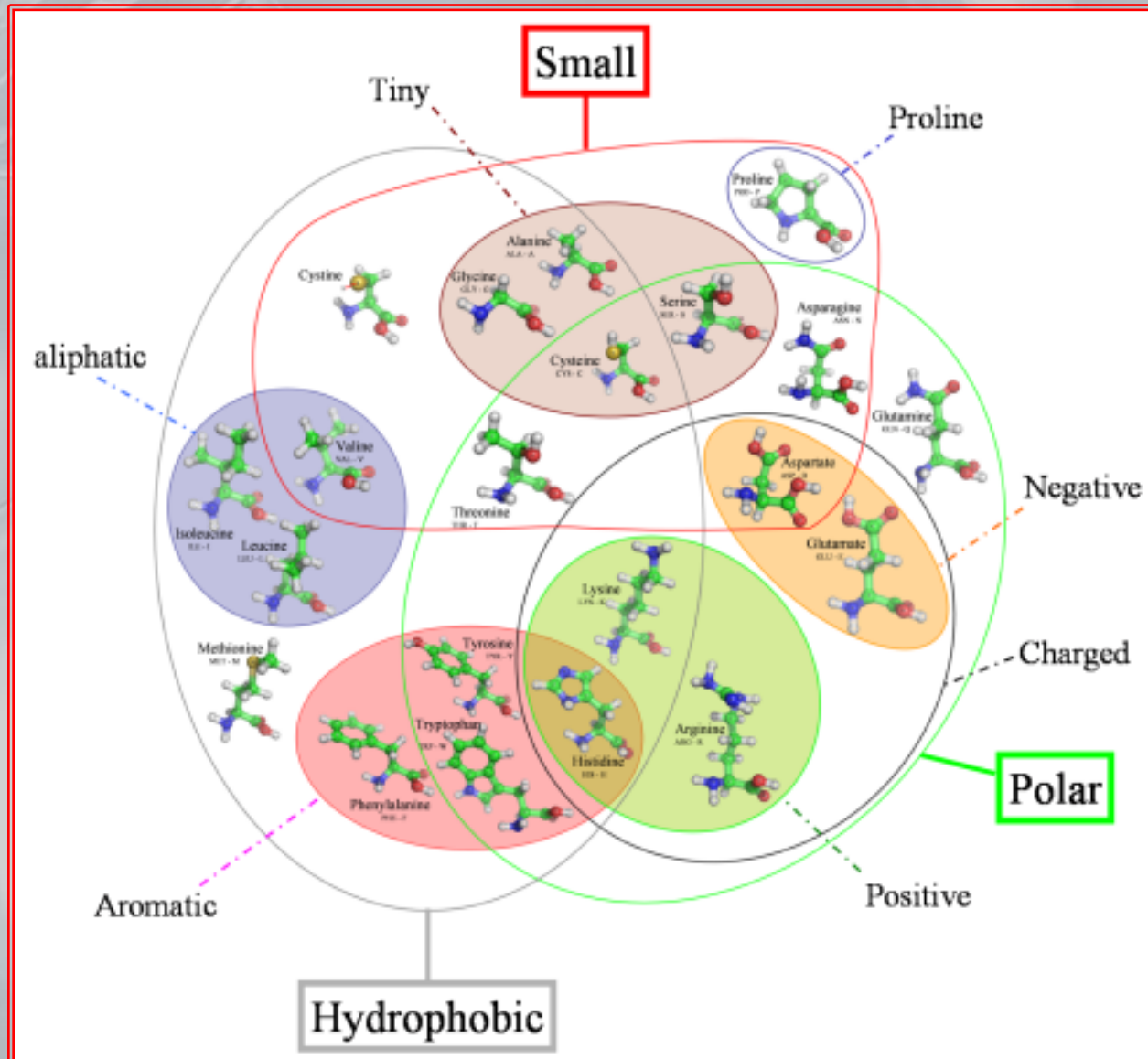


Stereochemistry studies the spatial properties of molecules (absence or presence of centers, planes and axes of reflective or rotational symmetry) and how these impact on the behavior of chemical compounds

Amino acids – 3

- ✦ **Amino acids** are grouped into three main categories
 - **Hydrophobic amino acids** have side chains composed for the majority (or even entirely) by carbon and hydrogen; rarely, they form hydrogen bonds with water molecules
 - ✗ Glycine, Alanine, Valine, Leucine, Isoleucine, Phenylalanine, Methionine
 - **Polar amino acids** contain oxygen and/or nitrogen in their side chains; they easily form hydrogen bonds with water
 - ✗ Serine, Threonine, Tyrosine, Tryptophan, Cysteine, Asparagine, Glutamine, Proline
 - **Charged amino acids** have a positive or a negative charge at biological pH (pH=7)
 - ✗ Aspartic acid, Glutamic acid, Lysine, Histidine, Arginine

Amino acids – 4



Amino acids – 5

- ✦ The amino acid order, in the primary sequence of a protein, plays a key role in the determination of its secondary and tertiary structure
 - The amino acid sequence properties and organization are responsible for both the structure and the biological function of proteins
- ✦ **Example:** Excerpt from the GenBank entry of *bacteriocuprein superoxide dismutase gene* for the *Photobacterium leiognathi*

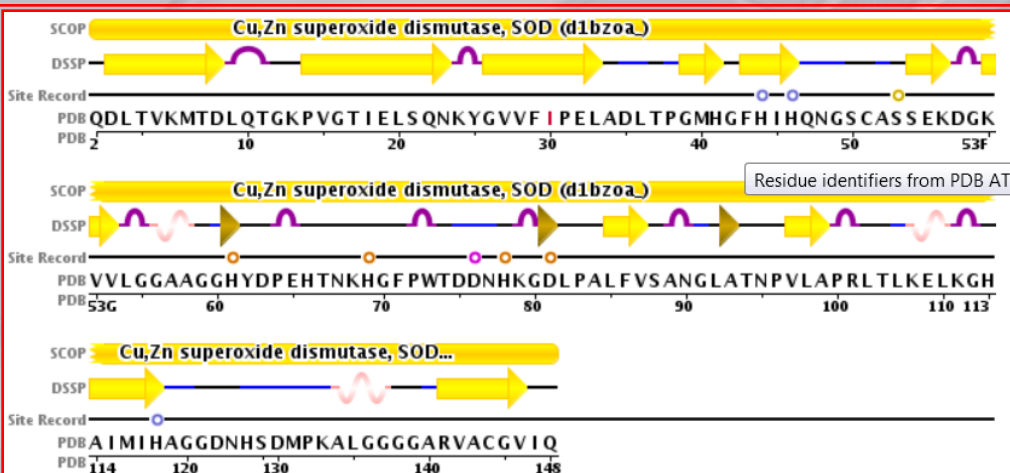
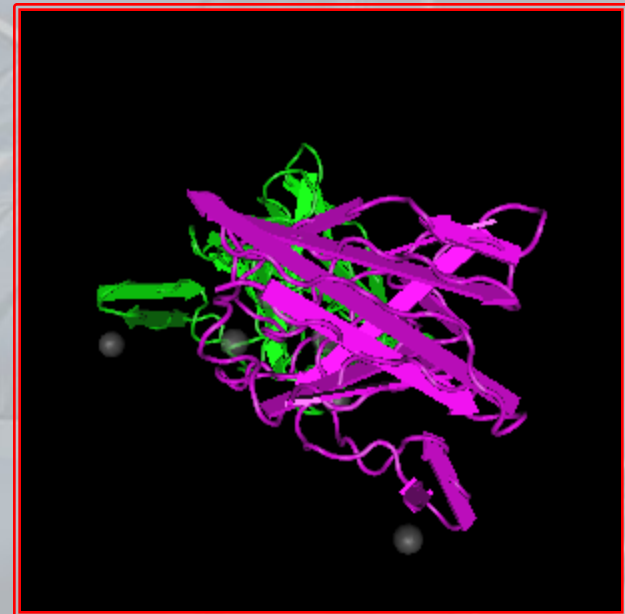


Amino acids – 6

Nucleotide gene sequence

ACCESSION J02658
 KEYWORDS bacteriocuprein superoxide dismutase; superoxide dismutase.
 SOURCE Photobacterium leiognathi
 ORGANISM Photobacterium leiognathi
 Bacteria; Proteobacteria; Gammaproteobacteria; Vibrionales;
 Vibrionaceae; Photobacterium.
 ORIGIN 256 bp upstream of BglII site.
 1 agtaaaaatt tagcaattaa gtagtggtga tgaatggta agagtaaaaa gtacacacgc
 61 tatgggatta atcttcttag cgaatgtttg agatattatc gataactata atcgtaaata
 121 tcagctatac ctttttggtta aaagcatggt taatgcctgt ggaaataaaa aacaataagg
 181 ataaaatatg aacaaggcaa aaacgttact cttcacgcgt ctacgttttg gtttatctca
 241 ccaagcggtta gcacaagatc tcacgggttaa aatgaccgat ctacaaacag gtaagcctgt
 301 tggtagcatt gaactaagcc aaaataaata cggagtagta ttacacacgt aactggcaga
 361 tttaacacgc gggatgcatg gttccatat tcatcaaaat gtagctgtg cttcatcaga
 421 aaaagacggc aaagtgtgtt taggtggcgc tgctgggtga cattacgacg ctgagcacac
 481 aaataaacac ggtttcccat ggactgatga taatcataaa ggtgatctgc cagcactggt
 541 tgtgagtgca aatggttttag caactaaccg tgttttagcg ccacgtttaa cgttgaaaga
 601 actaaaaggc cagcaatga tgatccatgc tgggtggtgat aatcactctg atatgccaaa
 661 agcattaggt ggcggcggcg cactgtgtgc gtgtggtgtg atccaataat ttagtgagaa
 721 ccagcagcga atttgcgct gttggtttta ttttaatcag attaagtttt ttagaaacag
 781 ccagttaatt gtaaaatatg taaaatgtg aaattcaggt gaatttgaaa tcttctctta
 841 a

Protein 3D structure



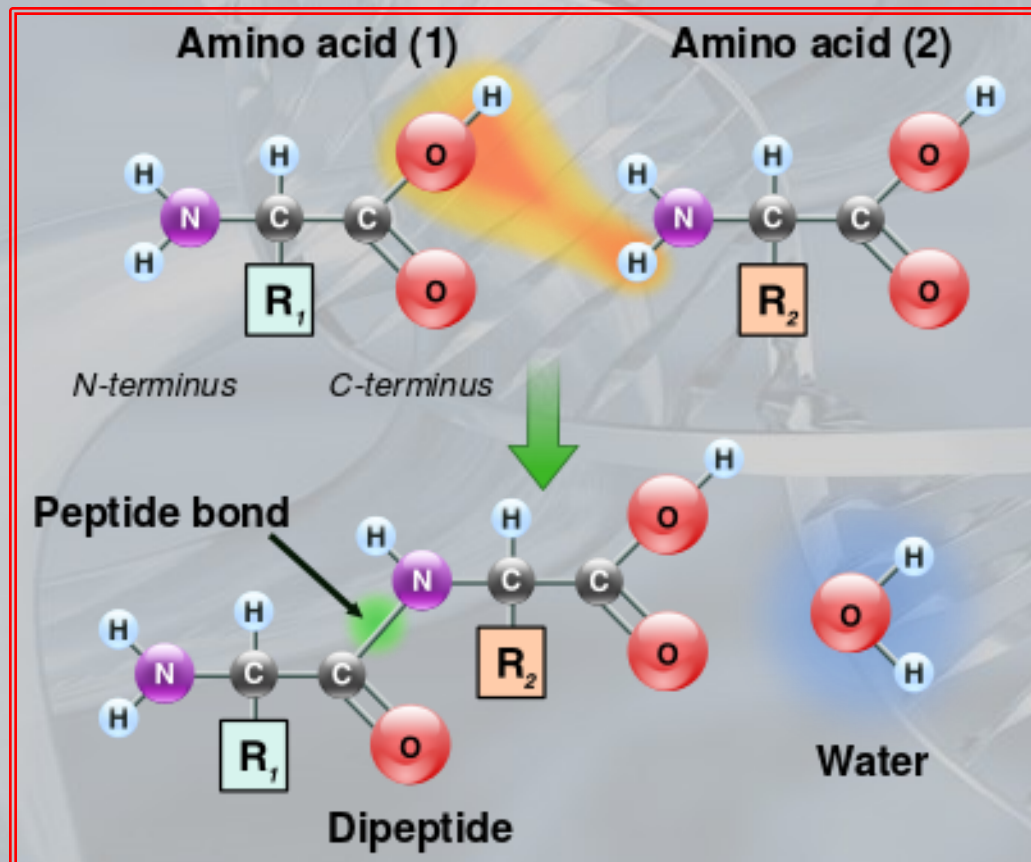
Protein primary and secondary structure

The composition of polypeptides – 1

- ✦ **Peptides** are molecules of molecular weight less than 5000 **daltons** (dalton: 1/12 of the ^{12}C atom mass), constituted by a chain of few amino acids, linked together by peptide (or *carboamide*) bonds
- ✦ Longer chains are called **polypeptides** or **proteins**
- ✦ When two amino acids are covalently linked, one of the two loses a hydrogen (H^+) from its amino group, while the other loses an oxygen and a hydrogen (OH^-) from its carboxylic group
 - ➡ A carbonyl group ($\text{C}=\text{O}$) and a water molecule (H_2O) are produced

The composition of polypeptides – 2

- ✦ The result is a dipeptide – two amino acids joined by a peptide bond – plus a water molecule



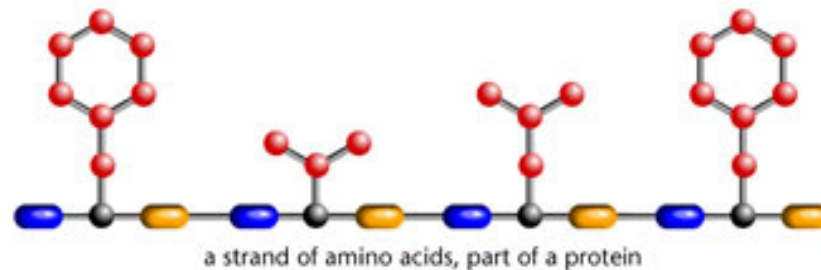
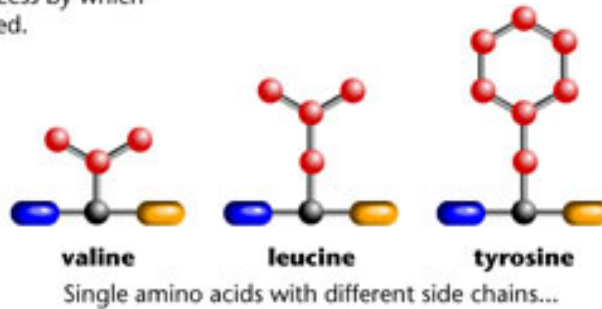
The composition of polypeptides – 3

- ✦ In a polypeptide, amino acids are sometimes called **residues**, because some atoms of the original chemical compounds are lost (in the form of water molecules) during the formation of peptide bonds
- ✦ As DNA and RNA molecules, polypeptides have a specific directionality
 - The **amino-terminal** (or **N-terminal**) of the polypeptide has an unbound amino group, while the **carboxy-terminal** (or **C-terminal**) ends with a carboxyl instead of with a carbonyl group
 - Protein sequences are usually considered from the N-terminal to the C-terminal
- ✦ The amino acid sequence that constitutes a protein, its **primary structure**, completely determines its three-dimensional structure, its physical and chemical properties, and finally, its biological function

The composition of polypeptides – 4

DIFFERENT AMINO ACIDS JOIN TOGETHER

This is the basic process by which protein are assembled.



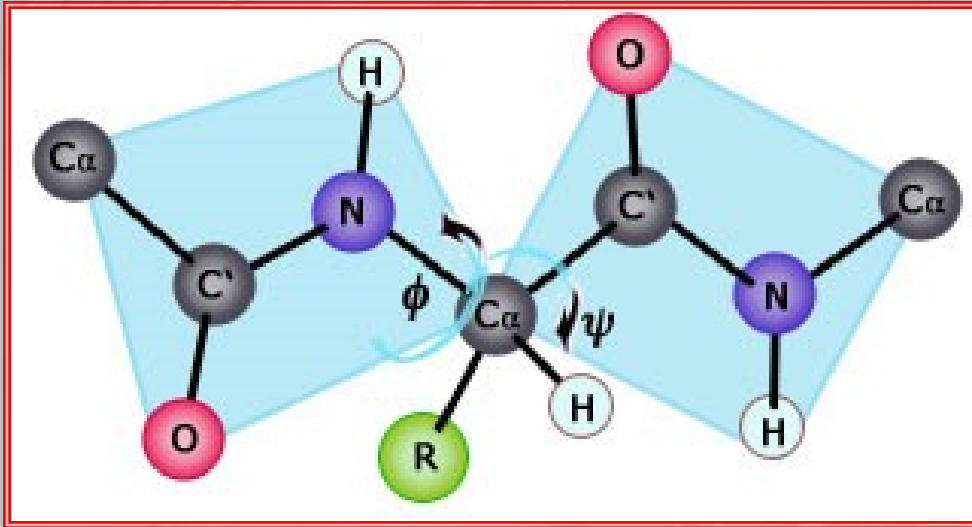
Protein secondary structure

Backbone flexibility – 1

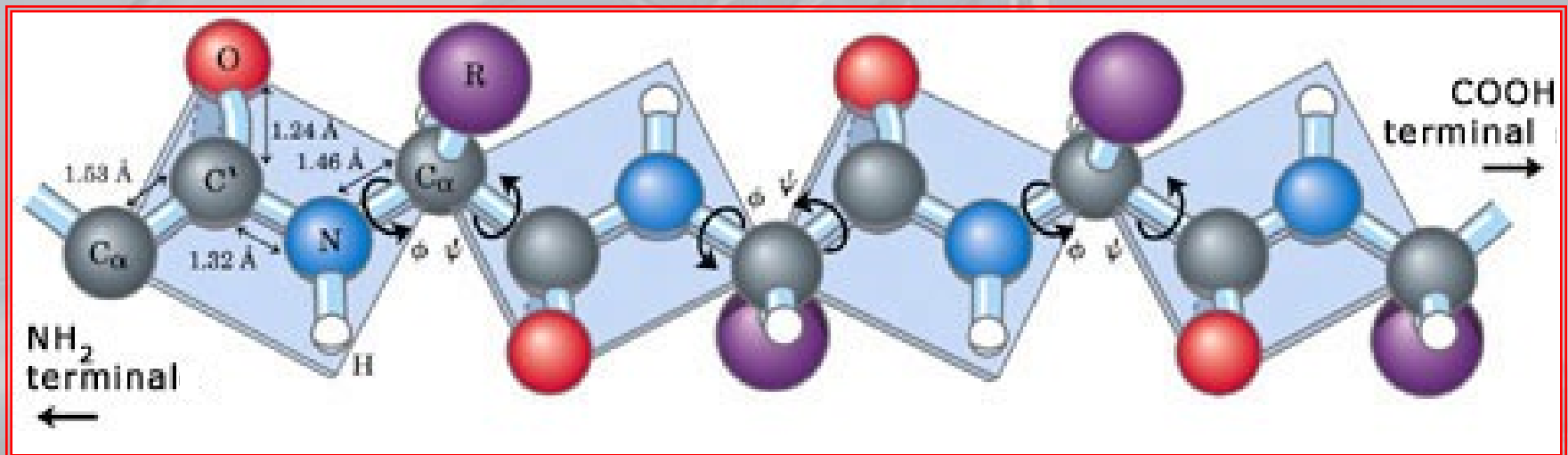
- ✦ In a polypeptide chain, the amino acid atoms that do not belong to the side chains form the **protein backbone** (or the main chain)
- ✦ The peptide group is rigid and planar, and the two bonds of C_α with N and C belonging to the same residue are the only points, on the backbone, on which rotations can occur
 - The rotation angle around the bond between the nitrogen of the amide and the alpha-carbon is called ϕ
 - The rotation angle around the bond between the alpha-carbon and the carbonyl carbon is called ψ
- ✦ The backbone conformation of an entire protein can be specified in terms of the angles ϕ , ψ of each amino acid

Protein secondary structure

Backbone flexibility – 2



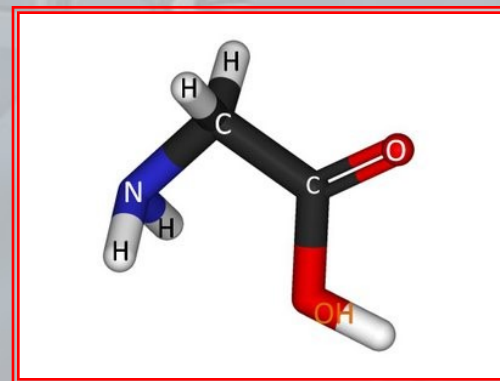
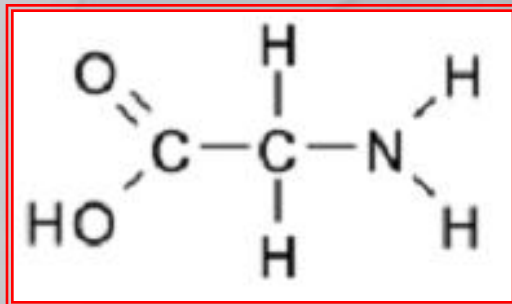
Rotations around the bonds of C_α are the only degrees of freedom allowed to the backbone; the structure of the peptide bonds forces the other backbone atoms to assume a rigid and planar configuration



Protein secondary structure

Backbone flexibility – 3

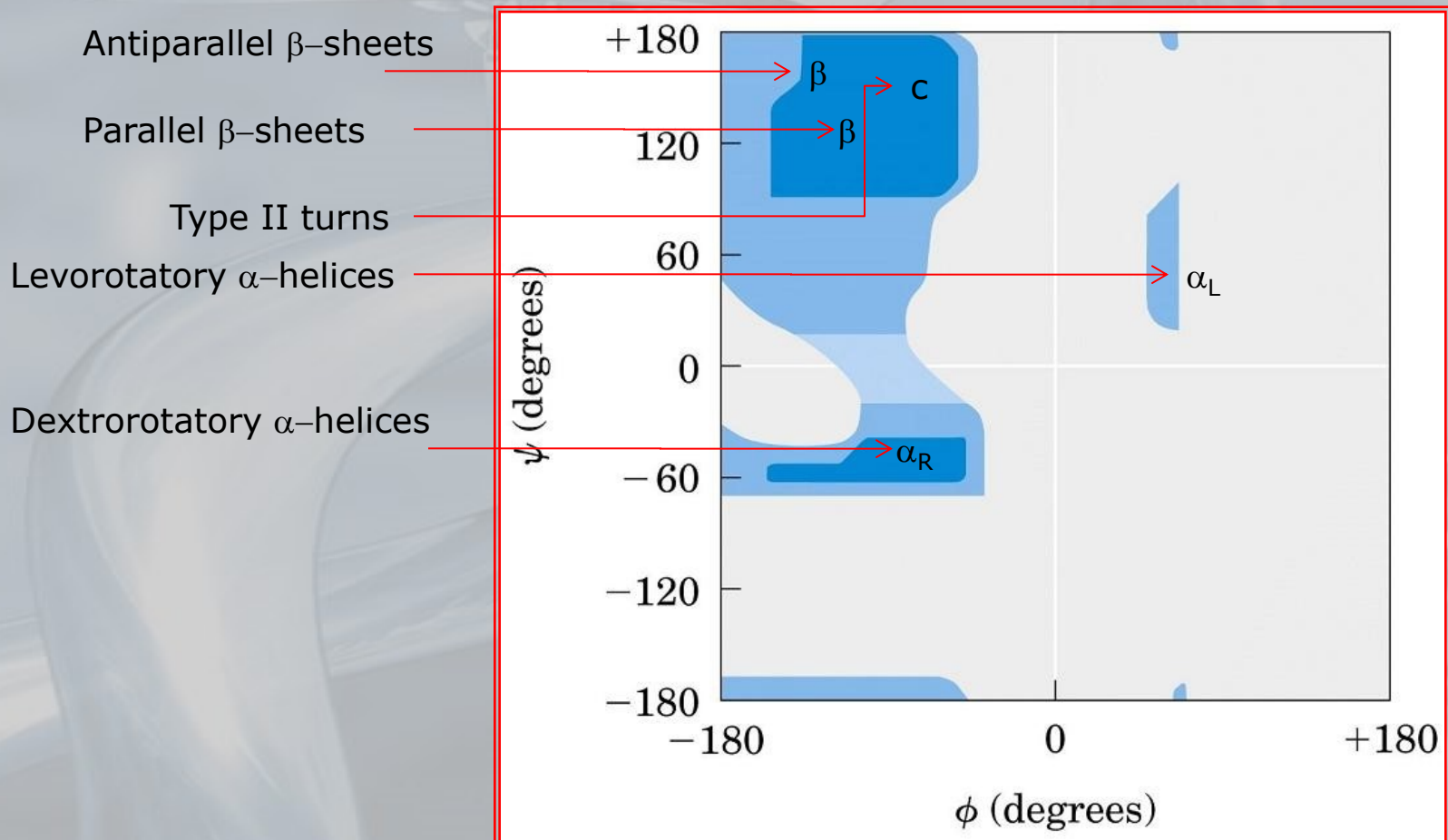
- ✦ Not all the values of ϕ and ψ are physically admissible
 - Some combinations of ϕ and ψ give rise to **steric collisions**, i.e. to a physical overlap of the space occupied by the atoms of the side chain of a residue and by the main chain of the next
 - Because of the lack of a side chain (different from hydrogen), the glycine residues have a much broader range of possible angles ϕ and ψ compared to the other residues



Protein secondary structure

Backbone flexibility – 4

The Ramachandran plot shows the values of ϕ and ψ which are physically allowed without causing steric collisions (dark regions); the glycine can assume additional conformations because of its small side chain (light regions)



Protein secondary structure

Backbone flexibility – 5

- ✦ Most of the protein backbones contain secondary structure elements, including α -helices ($\phi \approx \psi \approx -60^\circ$) and β -strands ($\phi \approx -135^\circ$, $\psi \approx 135^\circ$), which are associated with other β -strands to form parallel or antiparallel β -sheets
- ✦ Given a protein sequence (or its primary structure), the first step toward the prediction of the three-dimensional structure consists in determining its **secondary structure**, that is in defining which backbone regions are most likely to form helices, strands, and **β -turns**, U-bent structures obtained when a β -strand reverses its direction in an antiparallel β -sheet

Protein secondary structure

Prediction accuracy – 1

- ✦ Algorithms for the secondary structure prediction employ a variety of computational techniques, that includes neural networks, finite state automata, HMMs, clustering techniques, and genetic algorithms
- ✦ Most of the existing prediction algorithms are based on a preliminary alignment of the amino acid sequences, obtained by classical algorithms, such as BLAST, FASTA and CLUSTALW
- ✦ Based on the alignment, the degree of conservation of each amino acid in the target sequence is estimated, from which the secondary structure prediction can start

Protein secondary structure

Prediction accuracy – 2

- ✦ Having a protein sequence and the corresponding conservation levels as inputs, the most used prediction methods (based on information gathered from proteins whose secondary structure is already resolved) are:
 - **Chou–Fasman method**, a statistical approach that is based on the observation that the twenty amino acids show significant preferences for particular secondary structures (A,R,Q,E,M,L,K – helices, C,I,F,T,W,Y,V – sheets); average accuracy 56%
 - ✗ **GOR method**, a different statistical approach based on a window of 17 amino acids; average accuracy 65%
 - **Stereochemical Lim method**, that takes into account hydrophobic, hydrophilic and electrostatic properties of amino acids, considering their role in the protein folding procedure
 - **Neural networks**, which are able to process both statistical and chemico–physical information, in addition to the evolutionary information coming from multiple alignments (f.i., software PSIPRED); average accuracy 70–75%

Protein secondary structure

Prediction accuracy – 3

- The output of a secondary structure prediction algorithm is usually similar to the following:

```
APAFSVSPASGDGQSVSVSVAAAGETYYIAQCAPVGGQDACNPAT
-----HHHHHHH-HHHhhh---EEEEEEEe---EEEEee----
```

- In this case, **H** and **h** represent predictions of a helical conformation (respectively with high and low confidence), while **E** and **e** represent predictions of sheets ("extended", plain, surfaces)

Protein secondary structure

The Chou–Fasman method

- ✦ The **Chou–Fasman method** is based on the PDB databank
- ✦ It uses a simple statistical approach for predicting the secondary structure of proteins
- ✦ Many conformational parameters are assigned to each amino acid: $P(a)$, $P(b)$ and $P(t)$ (where t means “turn”)
- ✦ These parameters, which represent the propensity of each amino acid to be part of, respectively, α -helices, β -sheets or β -turns, are determined on the basis of observed frequencies in a set of known protein samples (PDB)
- ✦ Moreover, to each amino acid, four “turn” parameters are assigned, $f(i)$, $f(i+1)$, $f(i+2)$, $f(i+3)$, that correspond to the frequency with which it is observed in the first, second, third or fourth position of a hairpin turn

Chou–Fasman parameters for the 20 common amino acids

Amino acid	$P(a)$	$P(b)$	$P(t)$	$f(i)$	$f(i+1)$	$f(i+2)$	$f(i+3)$
Alanine	142	83	66	0.06	0.076	0.035	0.058
Arginine	98	93	95	0.070	0.106	0.099	0.085
Asparagine	67	89	156	0.161	0.083	0.191	0.091
Aspartic acid	101	54	146	0.147	0.110	0.179	0.081
Glutamine acid	151	37	74	0.056	0.060	0.077	0.064
Cysteine	70	119	119	0.149	0.050	0.117	0.128
Phenylalanine	113	238	60	0.059	0.041	0.065	0.065
Glutamine	111	110	98	0.074	0.098	0.037	0.098
Glycine	57	75	156	0.102	0.085	0.190	0.152
Histidine	100	87	95	0.140	0.047	0.093	0.054
Isoleucine	108	160	47	0.043	0.034	0.013	0.056
Leucine	121	130	59	0.061	0.025	0.036	0.070
Lysine	114	74	101	0.055	0.115	0.072	0.095
Methionine	145	105	60	0.068	0.082	0.014	0.055
Proline	57	55	152	0.102	0.301	0.034	0.068
Serine	77	75	143	0.120	0.139	0.125	0.106
Tyrosine	69	147	114	0.082	0.065	0.114	0.125
Threonine	83	119	96	0.086	0.108	0.065	0.079
Tryptophan	108	137	96	0.077	0.013	0.064	0.167
Valine	106	170	50	0.062	0.048	0.028	0.053

Protein secondary structure

The Chou–Fasman algorithm – 1

- ✦ Using the Chou–Fasman parameters, the algorithm for the estimation of the secondary structure proceeds according to the following steps
 1. Identification of α -helices:
 - a. Scan through the peptide and find the regions where 4 out of 6 contiguous residues have $P(a) > 100$
 - b. For each identified region, the region is extended in both directions until a sequence of 4 contiguous residues is found for which $P(a) < 100$
 - c. For each extended region, $\sum P(a)$ and $\sum P(b)$ – i.e. the sums of the $P(a)$ and the $P(b)$ values for all the residues in such a region – are calculated; if the region is longer than 5 residues and $\sum P(a) > \sum P(b)$, then it is predicted to be an α -helix

Protein secondary structure

The Chou–Fasman algorithm – 2

2. Identification of β -sheets using the same algorithm as in point 1., but looking for regions where 3 out of 5 residues show $P(b) > 100$; after having extended the regions (as in 1.b), a region is declared a β -sheet if the average $P(b)$, over all the residues, is greater than 100 and $\Sigma P(b) > \Sigma P(a)$
3. If a certain α -helix, assigned in 1., overlaps to a β -sheet, assigned in 2., then the overlapped region is defined to be a helix if $\Sigma P(a) > \Sigma P(b)$, or to be a β -sheet if $\Sigma P(b) > \Sigma P(a)$

Protein secondary structure

The Chou–Fasman algorithm – 3

4. Identification of β -turns:

- a. For each residue i , the propensity to bend, $p(t)$, is calculated by $p(t) = \prod_{k=0, \dots, 3} f(i+k)$
- b. A hairpin turn is predicted for each position i satisfying the following criteria
 - i. The propensity to bend should be $p(t) > 0.000075$
 - ii. The average value of $P(t)$ for the four residues in positions $i+k$, $k = 0, \dots, 3$, is greater than 100
 - iii. $\sum P(a) < \sum P(t) > \sum P(b)$ holds on the four residues in positions $i+k$, $k = 0, \dots, 3$

Exercise 1

- Using the Chou–Fasman algorithm and its related parameters (calculated from PBD), predict the α -helix and the β -sheet regions, for the following sequence

CAENKLDHVADCCILFMTWDYNGPCIFIDYNGP

- Solution**

P(a)

70-142-151-67-114-121-101-100-106-142-101-70-70-108-121-113-145-83-108-101-69-67-57-57-70-108-113-108-101-69-67-57-57

P(b)

119-83-37-89-74-130-54-87-170-83-54-119-119-160-130-138-105-119-137-54-147-89-75-55-119-160-138-160-54-147-89-75-55

$$\Sigma P(a) = 2134$$

$$\Sigma P(b) = 2061$$

\Rightarrow **helix**

$$\Sigma P(a) = 557$$

$$\Sigma P(b) = 686$$

\Rightarrow this is probably too short to be a helix

Exercise 1 (cont.)

✦ Solution P(b)

119-83-37-89-74-130-54-87-170-83-54-119-119-160-130-138-105-119-137-54-147-
89-75-55-119-160-138-160-54-147-89-75-55

$$\Sigma P(b)/16 = 112.875$$

$$\Sigma P(a) = 1659 < 1806 = \Sigma P(b)$$

⇒ sheet

$$\Sigma P(b)/8 = 113.5$$

$$\Sigma P(a) = 683 < 908 = \Sigma P(b)$$

⇒ sheet

- ✦ Finally, we need to control what happens in positions from 6 to 20, where the two predictions overlap:

$$\Sigma P(a) = 1590$$

$$\Sigma P(b) = 1659$$

⇒ sheet

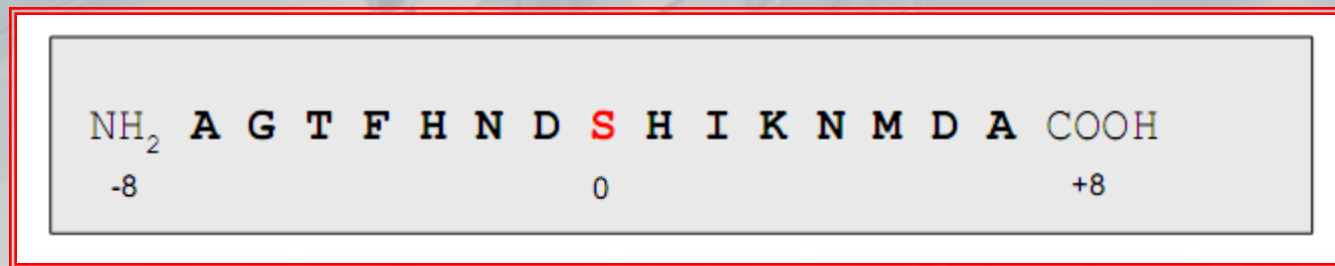
- ✦ Therefore...

CAENKLDHVADCCILFMTWDYNGPCIFIDYNGP
HHHHHeeeeeeeeeeeeeeeeeee-EEEEEEEE- - -

Protein secondary structure

The GOR method – 1

- A different statistical approach is the **GOR method** (from the names of those who developed it: Garnier, Osguthorpe and Robson), that predicts the secondary structure based on a window of 17 residues



- For each residue of the sequence, 8 N-terminal and 8 C-terminal positions are considered, together with such central residue

Protein secondary structure

The GOR method – 2

- ✦ As in the Chou–Fasman method, a collection of proteins of known secondary structure is analyzed
- ✦ The frequencies with which each amino acid occupies each of the 17 positions inside a window, being in a helical, sheet, turn or random coil conformation, are calculated
 - ▶ Four score matrices of dimension 17×20 are evaluated
- ✦ The values of these matrices are used to calculate the probability that the central residue belongs to a helix, a sheet, a turn, or a coil
- ✦ The method is effective in predicting α -helices (>65%), whereas it is less precise in the case of β -sheets (36.5%)

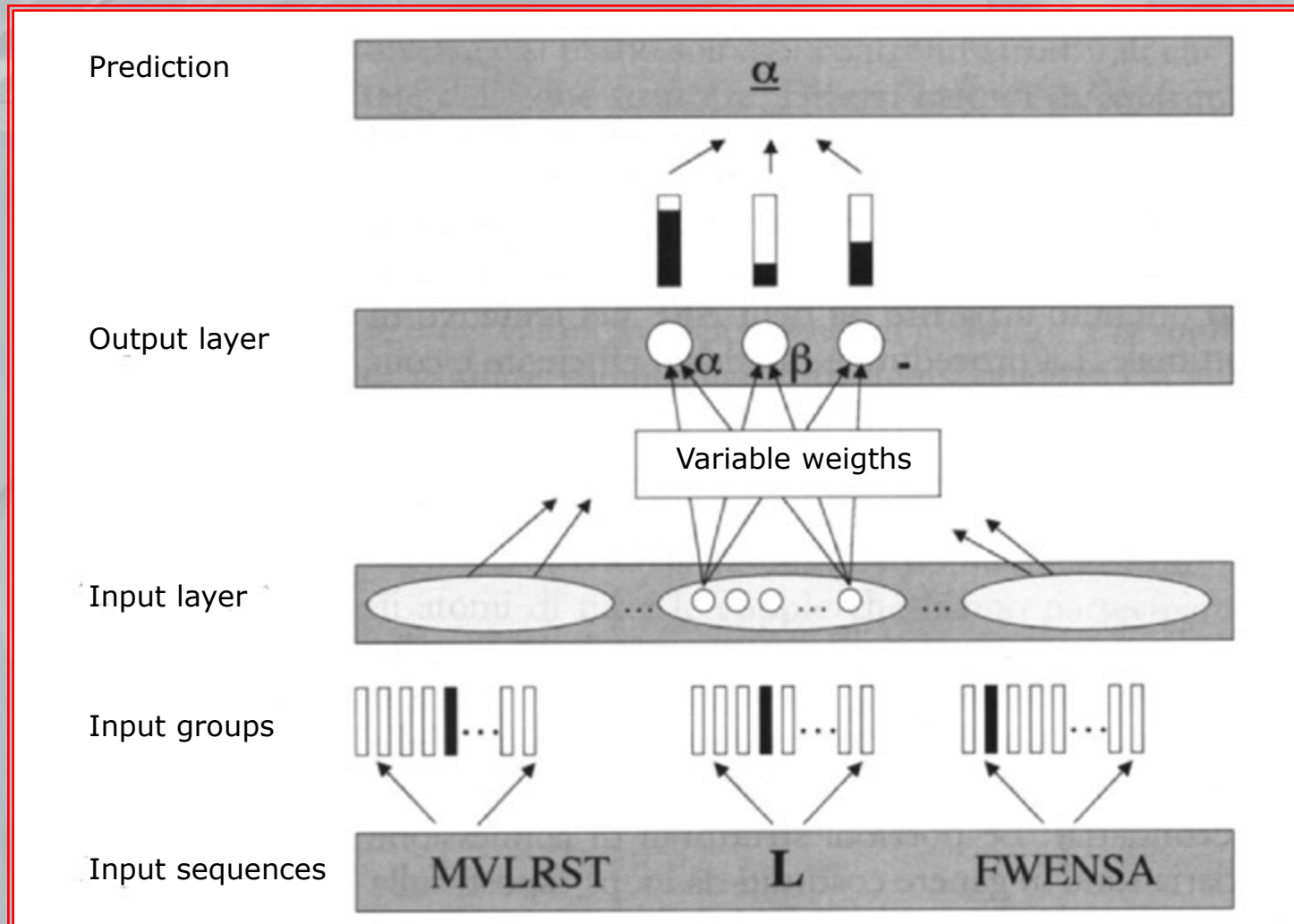
Protein secondary structure

Neural networks – 1

- ✦ The best state-of-the-art methods for secondary structure prediction employ neural networks
- ✦ For evaluating neural network weights many samples of non-homologous proteins are required
- ✦ In order to make a prediction on the residue i of the protein, local information is used (f.i.: $i-6, \dots, i-1, i, i+1, \dots, i+6$)
- ✦ Each residue is represented by a one-hot encoding, using an array of 21 elements, 20 for each type of residue, one for its absence (gap)

Protein secondary structure

Neural networks – 2



Protein secondary structure

Neural networks – 3

- ✦ Obviously, instead of using a sliding window, moving along the amino acid sequence, we can process the sequence as it is based, for instance, on LSTMs
- ✦ Also, Bi-directional recurrent neural networks can be used
 - **Bi-directional RNNs** use a finite sequence to predict or label each element of the sequence based on the element's past and future contexts
 - This is done by concatenating the outputs of two RNNs, one processing the sequence from left to right and the other from right to left
 - This technique has been proven to be especially useful when realized based on LSTMs

Tertiary and quaternary structure – 1

- ✦ The secondary structure prediction process is only the first step in predicting the complete three-dimensional structure of a functional protein
- ✦ The secondary structure elements of a protein are packaged, along with less structured loop, so as to form a compact and globular native conformation
- ✦ The overall three-dimensional structure of a folded polypeptide chain represents its **tertiary structure**
- ✦ The **quaternary structure** describes the intermolecular interactions that occur when multiple polypeptides are associated, to form a functional protein (example: protein–protein contacts, that occur in multienzymatic complexes)

Tertiary and quaternary structure – 2

(a)



(b)



(c)



(d)



Tertiary structure motifs: (a) bundle of helices; (b) α - β barrel; (c) and (d) open β -sheets

Tertiary and quaternary structure – 3

- ✦ The **protein folding** problem includes the prediction of secondary, tertiary and quaternary structure of polypeptides, based on their primary structure
- ✦ Different types of forces guide the protein folding:
 - Electrostatic forces
 - Hydrogen bonds
 - Van der Waals forces (a kind of weak intermolecular attraction caused by induced molecular dipoles)
 - Covalent bonds between cysteines
- ✦ Moreover, the prediction of the tertiary structure is made even more complex by the action of a special class of proteins, called **chaperones** (escorts), which act by altering the structure of the proteins (to favor their folding) in an important but unpredictable way

Tertiary and quaternary structure

Hydrophobicity – 1

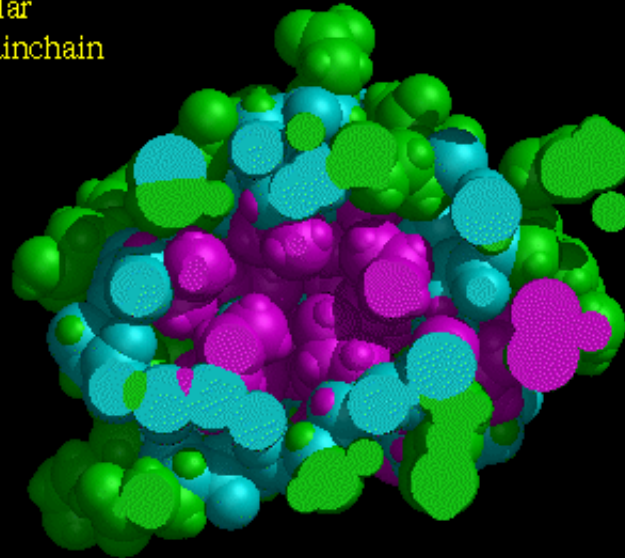
- ✦ The hydrophobic effect is generally considered fundamental in the folding process, during which the protein assumes a compact, globular structure
- ✦ The native structure of most of the proteins comprises a hydrophobic core, where the hydrophobic residues, are “buried”, i.e. subtracted from the contact with the solvent, while the protein surface, exposed to the solvent, is composed primarily, or entirely, by polar and charged residues
- ✦ The process of folding into a compact conformation that isolates the hydrophobic residues from the solvent is called **hydrophobic collapse**

Tertiary and quaternary structure

Hydrophobicity – 2

Distribution of Hydrophobic Residues in Ubiquitin

Magenta: Hydrophobic
Green: Polar
Cyan: Mainchain



Crystal structure of ubiquitin from Vijay-Kumar et al. 1987
Figure made using MidasPlus, UCSF

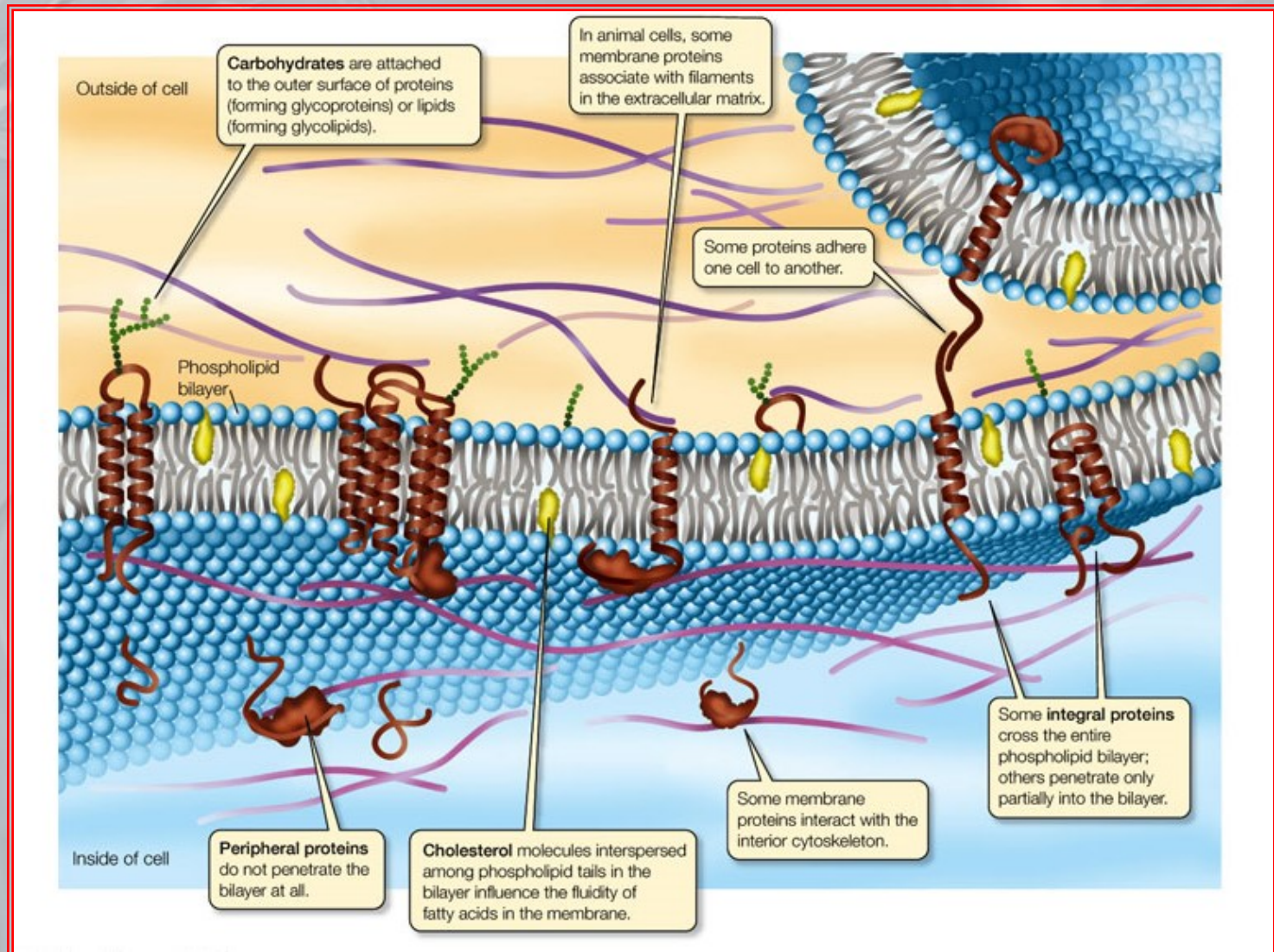
Tertiary and quaternary structure

Hydrophobicity – 3

- **Integral membrane proteins** are an exception to this rule
 - They contain one or more regions, often with a helical structure, which are inserted in the cell membrane
 - They cross the lipid bilayer, “looking out” the two membrane surfaces
 - They possess hydrophobic superficial regions located inside the phospholipid bilayer

Tertiary and quaternary structure

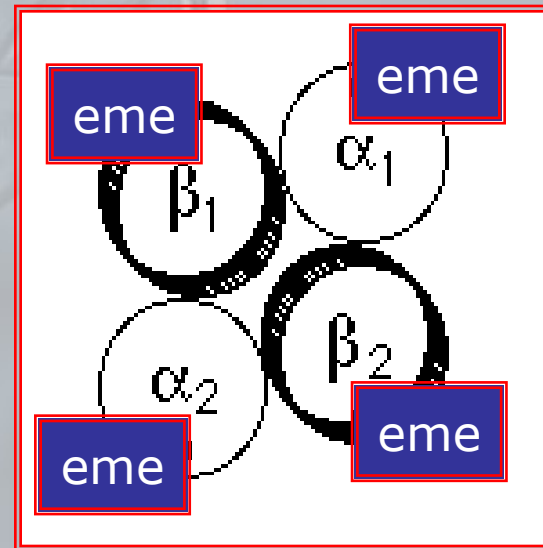
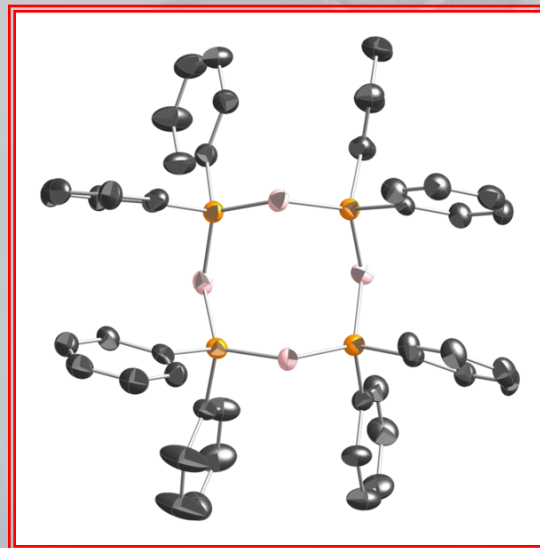
Hydrophobicity – 4



Tertiary and quaternary structure

Hydrophobicity – 5

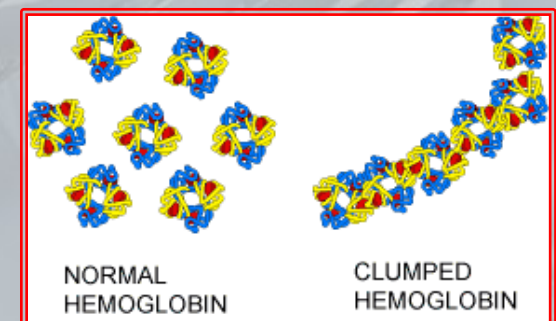
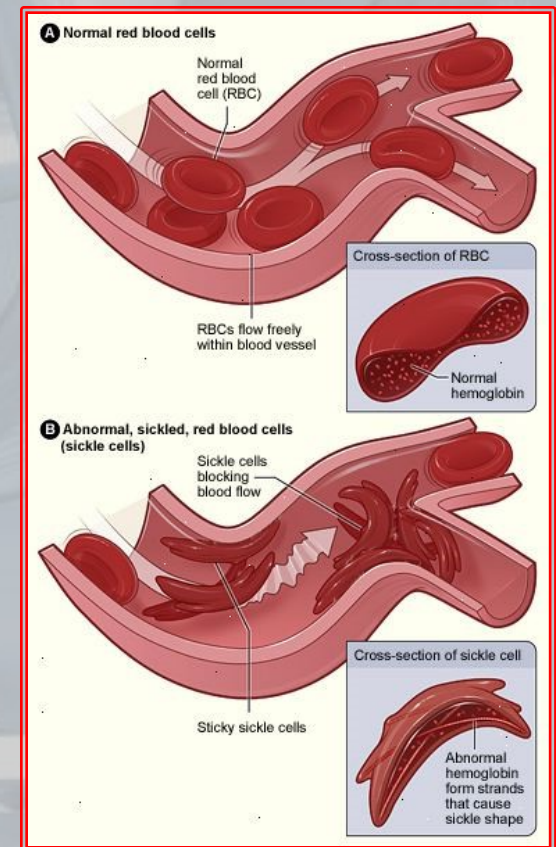
- ✦ The importance of isolating the hydrophobic residues from the solvent is clearly illustrated by the molecular pathology called sickle cell anemia
- ✦ The human hemoglobin, the protein responsible for the oxygen transport in the blood, is biologically active as a tetramer (an oligomer consisting of four subunits), formed by two chains of α -globin and two chains of β -globin



Tertiary and quaternary structure

Hydrophobicity – 6

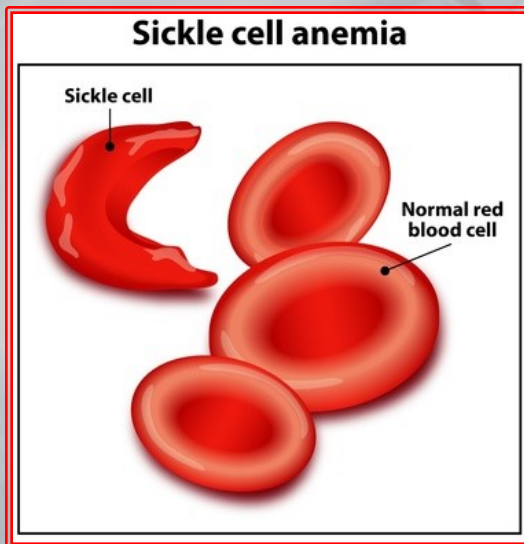
- ✦ The mutation of a single residue on the surface of a chain of the β -globin, from a charged residue of glutamic acid to a hydrophobic valine residue (**GAG**→**GUG**), forces the presence of a hydrophobic region on the protein surface exposed to the solvent
- ✦ The hydrophobic effect brings the valine residues to avoid contacts with the solvent, and causes the involved β -globin molecules to aggregate to the others



Tertiary and quaternary structure

Hydrophobicity – 7

- ✦ This results in long hemoglobin chains, which distort red blood cells, converting them from their normal disc shape to the characteristic sickle shape
 - The effect is particularly evident when the oxygen levels are low (at the ends and under stress) and the sickle cells become entangled with one another in the tiny blood vessels and in the capillaries
- ➡ Pain, anemia, gangrene



Tertiary and quaternary structure

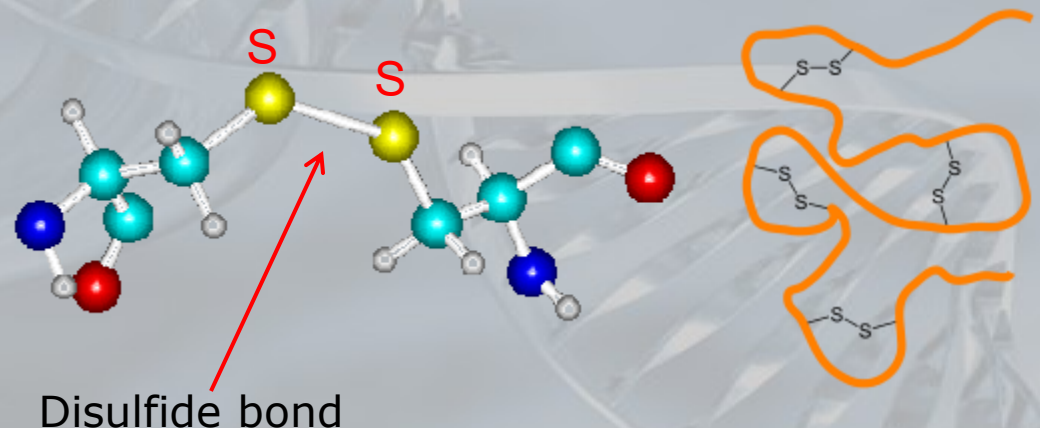
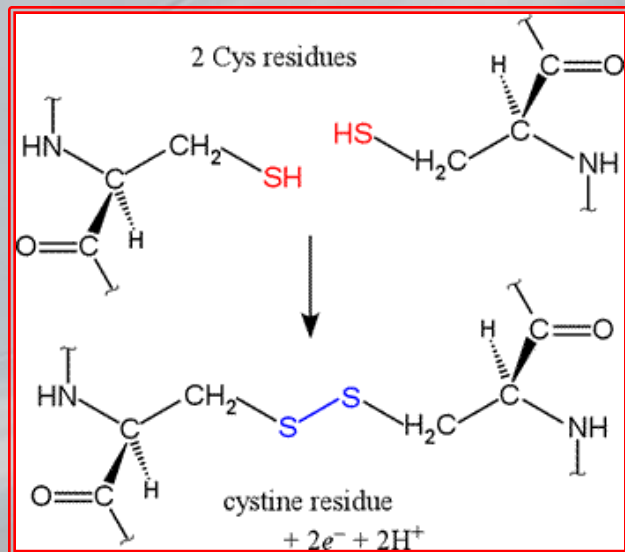
Hydrophobicity – 8

- ✦ The exact energy of the hydrophobic effect and its contribution to protein folding are difficult to calculate
- ✦ However, most of the folding algorithms, which base their calculations on molecular forces, include the hydrophobic collapse as one of the central mechanisms for protein folding

Tertiary and quaternary structure

Disulfide bonds – 1

- ✦ When the sulfhydryl groups $-SH$ of (the lateral chain of) cysteine residues are close, they can be oxidized to form covalent **disulfide bonds** (or **disulfide bridges**)
 - Cross-links are formed between residues that may be distant in the primary structure of the protein
 - The **cystines**, obtained by this chemical reaction, produce a significant stabilizing effect on the folded structure of a protein



Tertiary and quaternary structure

Disulfide bonds – 2

- ✦ When experimental methods require that a protein must be denatured, reducing agents, such as the β -mercaptoethanol ($\text{HOCH}_2\text{CH}_2\text{SH}$), are often used to break the disulfide bridges
- ✦ Actually, the original work by Anfinsen showed that the structure of a protein is specified by its sequence
 - In fact, the ribonuclease was first denatured and then it was let free to form disulfide bonds in presence of a high concentration of urea ($\text{CO}(\text{NH}_2)_2$)
 - The urea was able to reduce the effect of hydrophobicity on the conformation of the protein, allowing the formation of disulfide bonds different from those of the native structure
 - The “messy” ribonuclease with its cystine residues linked incorrectly, had only 1% of the enzymatic activity of the native conformation ribonuclease

Tertiary and quaternary structure

Active and stable structures – 1

- ✦ Due to the large number of degrees of freedom in the protein folding procedure, it is still impossible to evaluate, in general, if the native state is actually the most stable (or the energetically most favorable) conformation for a protein
 - Natural selection favors those proteins that are active and robust
 - It is likely that the protein primary structure mutations that reduce the structural stability of a protein are disadvantageous and thus that the natural selection acts against them

Tertiary and quaternary structure

Active and stable structures – 2

- ✦ In his famous work of 1968, C. Levinthal noted that the number of possible folds for a polypeptide chain (even for a small one) is so vast that, performing an exhaustive search of all possible conformations, would take many years
- ✦ Therefore, if a protein were to attain its correctly folded configuration by sequentially sampling all the possible conformations ($\approx 10^{300}$), it would require a time longer than the age of the universe to arrive at its correct native structure
- ✦ This observation, known as the [Levinthal paradox](#), suggests that proteins fold following a path whose intermediate steps are progressively more stable, until they have reached their native state
- ✦ However, that this path finishes (or not) in an absolute minimum configuration for the energy is still a matter of debate ↵

Algorithms for protein folding modeling – 1

- ✦ Numerous algorithms have been developed to understand how the amino acid sequence of a protein determines its unique native conformation, none of which shows an accuracy equal to purely experimental methods, such as X-ray crystallography — at least before the very recent and powerful **AlphaFold** (now release 3), developed by **Google DeepMind**

[nature](#) > [articles](#) > [article](#)

Article | [Open Access](#) | [Published: 15 July 2021](#)

Highly accurate protein structure prediction with AlphaFold

[John Jumper](#) ✉, [Richard Evans](#), [...] [Demis Hassabis](#) ✉

[Nature](#) **596**, 583–589 (2021) | [Cite this article](#)

454k Accesses | **316** Citations | **2793** Altmetric | [Metrics](#)

Abstract

Proteins are essential to life, and understanding their structure can facilitate

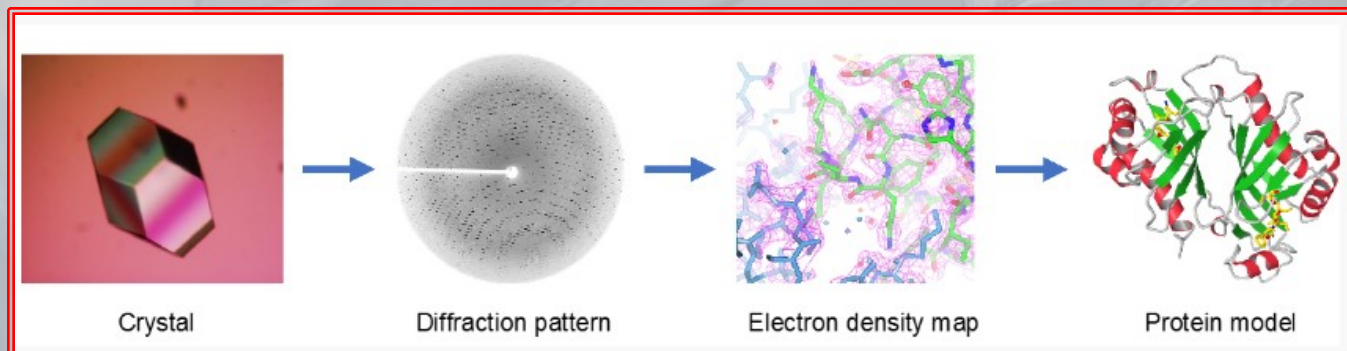
“...the first computational method that can regularly predict protein structures with atomic accuracy even in cases in which no similar structure is known.”

“AlphaFold has revealed millions of intricate 3D protein structures, and is helping scientists understand how life’s molecules interact.”

(see also <https://www.youtube.com/watch?v=gg7WjuFs8F4>
<https://www.youtube.com/watch?v=Mz7Qp73lj9o>)

Algorithms for protein folding modeling – 2

- ✦ **X-ray crystallography** — The image, produced by the X-ray diffraction through the atomic lattice of a crystal, is recorded/analyzed to reveal the nature of the lattice
 - ➡ It allows the determination of the molecular structure of a chemical compound at the atomic level
- ✦ However, even the simplest protein folding algorithms have provided new insights into the forces that determine the protein structure and on the folding process itself
- ✦ ...while the accuracy and the power of these algorithms progressively improve, thanks to new optimization and machine learning techniques, and to the deeper knowledge in the field of experimental biochemistry, especially on those forces which contribute to the stability of a folded protein

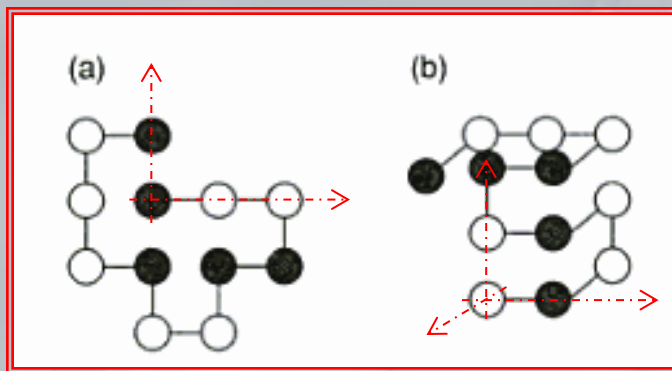


Lattice models – 1

- ✦ Even the most modern high-performance architectures are hardly able to explicitly model all the interactions involved in the folding of a polypeptide chain for more than few femtoseconds ($=10^{-15}$ seconds)
- ✦ Having a limited computational power, very simplified models of the folding process must be defined in order to dominate its complexity
 - ✦ Restricting degrees of freedom for the protein conformation: Instead of allowing all the physically possible configurations, the C_{α} positions are constrained to lie on a bi- or a three-dimensional grid (or lattice)
 - ✗ Number of adoptable protein conformations significantly reduced
 - ✗ For polypeptides of modest size, an exhaustive search can be realized, that identifies the conformation corresponding to the absolute minimum of the energy

Lattice models – 2

- ✦ The simplest and the most studied lattice model is the **H-P model**, where H-P stands for *Hydrophobic-Polar*
 - It simplifies the protein model by representing each amino acid residue as a single atom of fixed radius, of hydrophobic or polar type

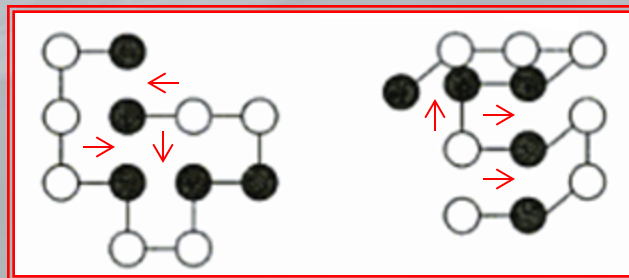


2D (a) and 3D (b) representations of the H-P model for a polypeptide of 12 residues; the hydrophobic residues are shown in black, the polar residues in white

- In the two-dimensional space, by convention, the N-terminal amino acid is positioned at the origin of the coordinate system and the subsequent residue in (1,0)
- The configuration score is based on the number of hydrophobic contacts in the grid

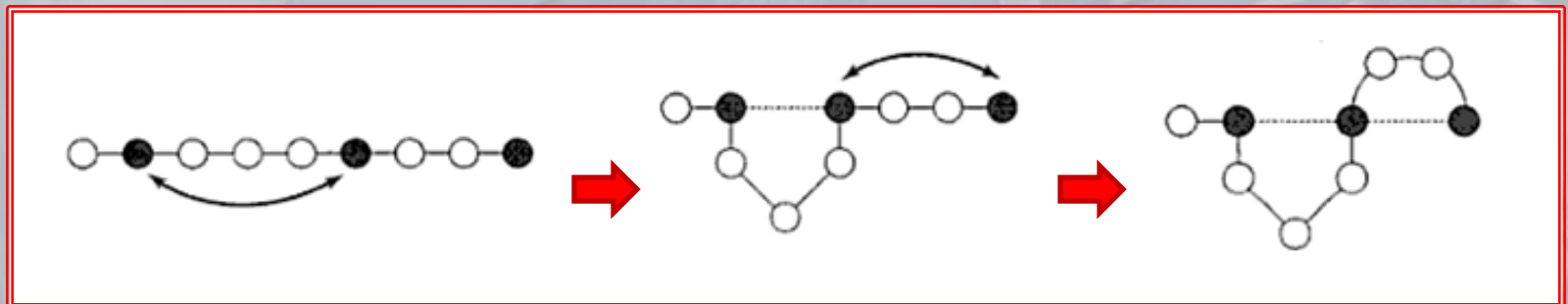
Lattice models – 3

- ✦ It is assumed that every contact H–H makes a contribution to the energy equal to -1 , except when the two hydrophobic residues are contiguous in the primary structure of the polypeptide (since they will contribute to all possible configurations exactly in the same way)
 - The optimal conformation will be the one with the greatest number of contacts H–H, which is normally obtained forming a hydrophobic core containing the maximum number of H residues, and relegating the P residues on the surface of the protein
 - **Example:** The obtained score for both the 2D and 3D conformations is equal to -3



Lattice models – 4

- ✦ Assuming that the hydrophobic collapse represents the only significant factor in protein folding is a very strong approximation, so as the conformational constraints imposed by a 2D or a 3D lattice
- ✦ However, H-P models have provided interesting information on the mechanism of protein folding
 - K. Dill suggested the **hydrophobic hinge** as a possible mechanism for the secondary structure formation (U-turns in antiparallel β -sheets)

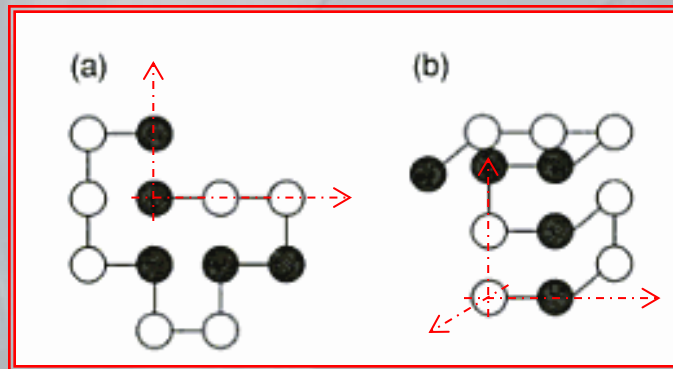


Lattice models – 5

- ✦ As the size of the polypeptide chain increases, the exhaustive search of all possible conformations becomes an intractable problem (the problem has been shown to be NP-complete in the '90s, from Patterson and Prytycka)
 - ➡ Use of soft-computing methods: Monte Carlo methods, **machine learning**, genetic algorithms, branch and bound, etc. to find a possible sub-optimal solution

Lattice models – 6

- ✦ An important consideration in the implementation of a protein folding prediction algorithm on the lattice concerns how to represent a particular configuration
 - **Absolute representation of the direction:** Put the first residue at the origin and then represent the direction of the motion for each subsequent residue
 - ✗ For the 2D model, the possible choices at each position are: Left (L), Right (R), Up (U) and Down (D)
 - ✗ For the 3D model: Left (L), Right (R), Up (U), Down (D), Backward (B) and Forward (F)



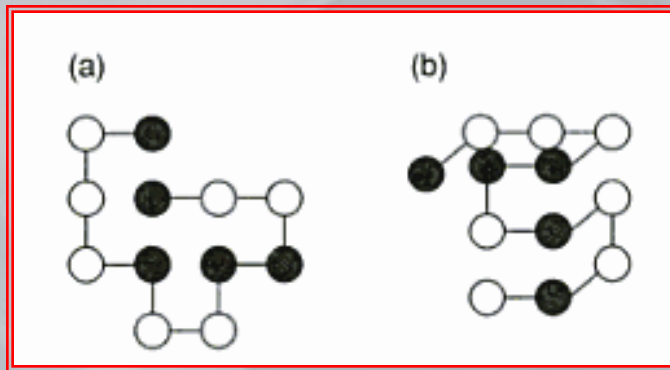
✗ Example

(a) R, R, D, L, D, L, U, L, U, U, R

(b) R, B, U, F, L, U, R, B, L, L, F

Lattice models – 7

- ✦ In fact, the number of choices at each step can be reduced by using a **relative representation of the direction**
 - Each residue, after the second one, has three/five options
 - ✗ In 2D: Left (L), Right (R), Forward (F)
 - ✗ In 3D: Left (L), Right (R), Forward (F), Up (U), Down (D)
 - In this approach, we have not only to keep track of the current position, but also of the direction to which the current residue points, defining what it considers as Up



✗ Example

(a) F, F, R, R, L, R, R, L, R, F, R

(b) F, L, U, U, R, U, U, L, L, F, L

Lattice models – 8

- ✦ A key difficulty, which occurs using both representations, lies in the fact that some of the generated configurations will have two residues in the same position
- ✦ **Example:** With the relative representation, any 2D configuration that begins with (L, L, L, L) will take two residues at the origin (0,0), resulting in a **bump** or a **steric collision**
- ✦ In order to avoid collisions:
 - Assign a very high energy to any configuration showing a collision
 - ✗ Since the optimization algorithm prefers low-energy configurations, those containing collisions will be excluded from the subsequent search steps
 - ✗ The search process may be hindered, eliminating favorable conformations that could lead to a low energy state
 - Local optimization strategies to resolve collisions before assigning a score to the obtained configuration
 - Alternative representations that do not (often) lead to collisions

Lattice models – 9

✦ Representation with a preference order

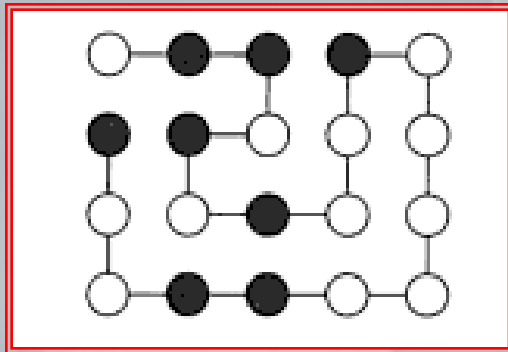
- A permutation of all possible directions, instead of a single direction, is assigned to each residue
 - **Example:** In a 2D model, the permutation {L, F, R} could be assigned to a single residue \Rightarrow the preferred direction is left but, in the event of a collision, also moving forward is acceptable
 - Although this type of representation can again produce bumps, when a collision occurs for each direction, they actually happen with a significantly lower frequency
- ✦ The use of representations with a preference order, in conjunction with local optimizations, can be used to obtain models that do not produce bumps

Lattice models – 10

- ✦ The correspondence established between amino acids and their occupied positions is called **protein embedding**, and when the embedding is injective, the conformation that the protein assumes is called **self-avoiding**
- ✦ In summary:
 - The hydrophobic interaction is crucial for protein folding and the hydrophobicity of amino acids is the driving force for the development of the native conformation for small globular proteins
 - The native structure of many proteins is compact and shows a well-defined nucleus, formed by the hydrophobic residues that are minimally exposed to the solvent, protected from the surface composed by polar residues

Exercise 2

- Assuming an energy contribution of -1 for each hydrophobic contact that does not concern two contiguous backbone residues, determine the energy for the following configuration of the 2D H-P model



- Assuming a relative representation (L, R, F), how can the configuration be described?

- Solution**

- The value of the energy is -4
- F,F,R,R,L,L,F,L,F,R,R,F,F,R,F,F,F,R,F

Off-lattice models – 1

- ✦ Free moving in the 3D space allows a protein model to assume more realistic configurations
- ✦ Using a complete model of the main chain and allowing ϕ and ψ angles to assume any value in the admissible regions of the Ramachandran plot, the *off-lattice* models have produced, for small polypeptides, configurations very similar to those experimentally observed
- ✦ The error is measured by the *Root Mean Square Deviation* (RMSD) between the predicted and the observed configurations of C_{α}
- ➡ Improvements in the realism of the protein models are obtained at the cost of a greater complexity

Off-lattice models – 2

- Off-lattice models may only include the C_α , all the backbone atoms or even the side chain
 - The backbone conformations are represented by the angles ϕ and ψ for each C_α
 - Side chains, if included, may be rigid, semi-flexible or completely flexible
 - In the case of rigid side chains, we consider the resolved X-ray crystallographic structures and the most common conformation, for each type of amino acid, is taken as the only possible one for that particular residue

Off-lattice models – 3

- ✦ In the case of semi-flexible side chains, for each amino acid, the side-conformation of all the resolved X-ray structures is considered; after that, a clustering phase is carried out in order to calculate the centroids, called **rotamers**
 - In a semi-flexible model, each side chain is allowed to adopt any of the rotamers most commonly observed
 - Different conformations allowed, with a modest computational load

Energy functions and optimization – 1

- ✦ The off-lattice protein models require more sophisticated energy functions
- ✦ **Theoretical approach**

- In addition to hydrophobic contacts, energy functions may consider the formation of hydrogen and disulfide bonds, the presence of electrostatic interactions and van der Waals forces, and the interactions with the solvent

$$E = a_1 E_{Hyd} + a_2 E_{Hbonds} + a_3 E_{Sbonds} + a_4 E_{Coulomb} + a_5 E_{vdWalls} + a_6 E_{Solv}$$

- They should be defined so that the protein conformations resolved by the X-ray crystallography represent minimum energy states
- Relative contributions are difficult to be experimentally calculated

Energy functions and optimization – 2

- ✦ Even if modeling proteins using all the forces that drive their folding is actually sensible, *ab initio* approaches have had a limited success, due to several reasons:
 - The exact forces and their interactions have not been fully understood yet
 - These approaches are computationally too expensive to be used for polypeptides of a realistic size
- ➡ Devise a **semi-empirical energy** function based on the conformations observed in known proteins
 - The 3D neighborhood of each amino acid is examined in order to create some scores based on the relative positions of the different residues
 - Local conformations which are common in the reference database are scored as low-energy, uncommon patterns obtain high-energy values

Energy functions and optimization – 3

✦ Examples

- If a particular serine residue has three residues within a neighborhood of 6\AA , an aspartate, a histidine, and a glutamate, and these amino acids are commonly found in the vicinity of the serine in the reference database \Rightarrow serine receives a low-energy score
- Conversely, if **SER** and **GLU** are rarely found to be close within the database, the serine residue will receive a higher value for its energy
- ✦ Local values are then summed over the entire protein to calculate the total energy
- ➡ Semi-empirical energy functions promote conformations similar to those observed in known proteins and penalize new or unusual ones

Folding algorithms – 1

- ✦ In summary, formulating an algorithm for the protein folding prediction problem consists of the following steps
 - Select a protein folding model
 - ✗ Lattice models allow fast calculations and exhaustive searches, but are not able to reliably reproduce the true protein conformations
 - ✗ Off-lattice models require lengthy calculations to evaluate the energy
 - Define how to describe every possible conformations
 - ✗ For lattice models, simple representations are sufficient, that encode the direction in which the “next move” should occur
 - ✗ For off-lattice models, ϕ and ψ angles for each C_α are used
 - Choose an energy function to evaluate how much a given conformation is favorable
 - Select an optimization method to search, through all the possible configurations, the structure that represents an absolute minimum for the energy function chosen

Folding algorithms – 2

- ✦ An interesting approach, to handle the computational complexity of *ab initio* methods, has been devised by V. Pande
- ✦ [Folding@Home](#), which acts as a screen saver, use the idle CPU cycles to develop protein folding models
- ✦ All the calculations, required to model the 3D structure of a particular protein, are “parallelized” and distributed via Internet to computers that execute the Folding@Home code
- ✦ The obtained results are sent back to the remote server, where they are recombined and analyzed

Folding algorithms – 3

- ✦ Using the power of distributed computing, the Folding@Home software can perform *ab initio* modeling of long polypeptides, a task that would otherwise be computationally intractable
- ✦ To join the Folding@Home project, only a computer (with Linux, Mac OS or Windows) and an Internet connection are needed
- ✦ For more information, the reference site is <https://foldingathome.org/>
- ★ During the pandemic period, the site has been focused on COVID-19 and reported the following message: “Together we have created the most powerful supercomputer on the planet and are using it to help understand SARS-CoV-2/COVID-19 and develop new therapies. We need your help pushing toward a potent, patent-free drug. Use your PC to help fight COVID-19.”
- ★ Now, there is still a dedicated section to COVID-19

Folding and misfolding

- When proteins do not fold correctly, a **misfolding** occurs, a situation which can lead to Alzheimer's disease, Parkinson's disease, Bovine Spongiform Encephalopathy (BSE), Huntington's disease, Amyotrophic Lateral Sclerosis (ALS), etc.
- A deep understanding of the misfolding causes is therefore essential to synthesize drugs and to define treatment programs to combat these neurodegenerative diseases

Tertiary structure prediction – 1

- ✦ Though protein folding models actually support a partial understanding of this phenomenon (and of the involved forces), for *ab initio* algorithm is anyway difficult to obtain a high degree of accuracy (i.e., an RMSD on the main chain $< 3\text{\AA}$) for large protein structures
- ✦ On the other hand, for applications such as **drug discovery** and **ligand design** (ligands are molecules capable of reacting with a receptor to produce a given physiological response) a very accurate model of the active site of a protein is required

Tertiary structure prediction – 2

- ✦ Proteins often interact with ligands with a so high specificity that a deviation $< 1\text{\AA}$ of the position of a key backbone atom can cause a substantial reduction in their binding affinity
- ✦ Alternatively, when the tertiary structure of one or more proteins, similar to the target protein with respect to the primary structure, is known, then the target protein can be modeled, often with a high degree of accuracy, based on comparative techniques

Comparative modeling – 1

- ✦ **Comparative modeling**, sometimes called **homology modeling**, is able to predict the structure of a target protein by comparison with the structures of some related proteins
 - The fundamental hypothesis on which these techniques are based is the “robustness of the folding code”
 - Small changes in the amino acid sequence of a protein (usually) will cause changes equally small of its tertiary structure
 - That is: In general, many changes in the primary structure must be accumulated before a significant distortion in the native conformation occurs

Comparative modeling – 2

- ✦ A generalized protocol for the comparative modeling of a target protein is composed by the following steps
 1. Identification of a set of protein structures related with the target protein – Since the 3D structure of the target protein is unknown, the similarity is based on the primary structure, which can be detected with database search algorithms, such as BLAST and FASTA; since the selected structures will serve as a template for the target protein modeling, they are just called template structures (www.ncbi.nlm.nih.gov/BLAST/)

Comparative modeling – 3

2. Alignment of the target sequence with the template protein sequences

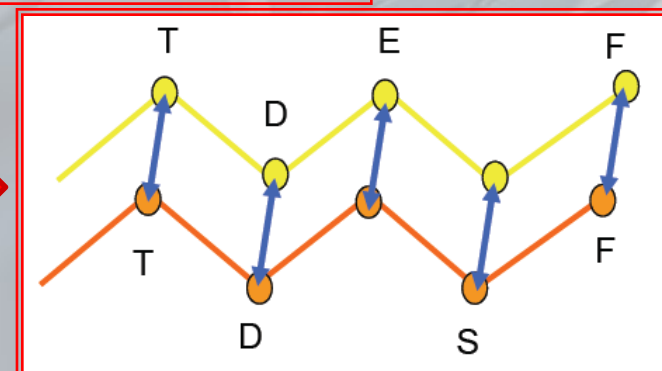
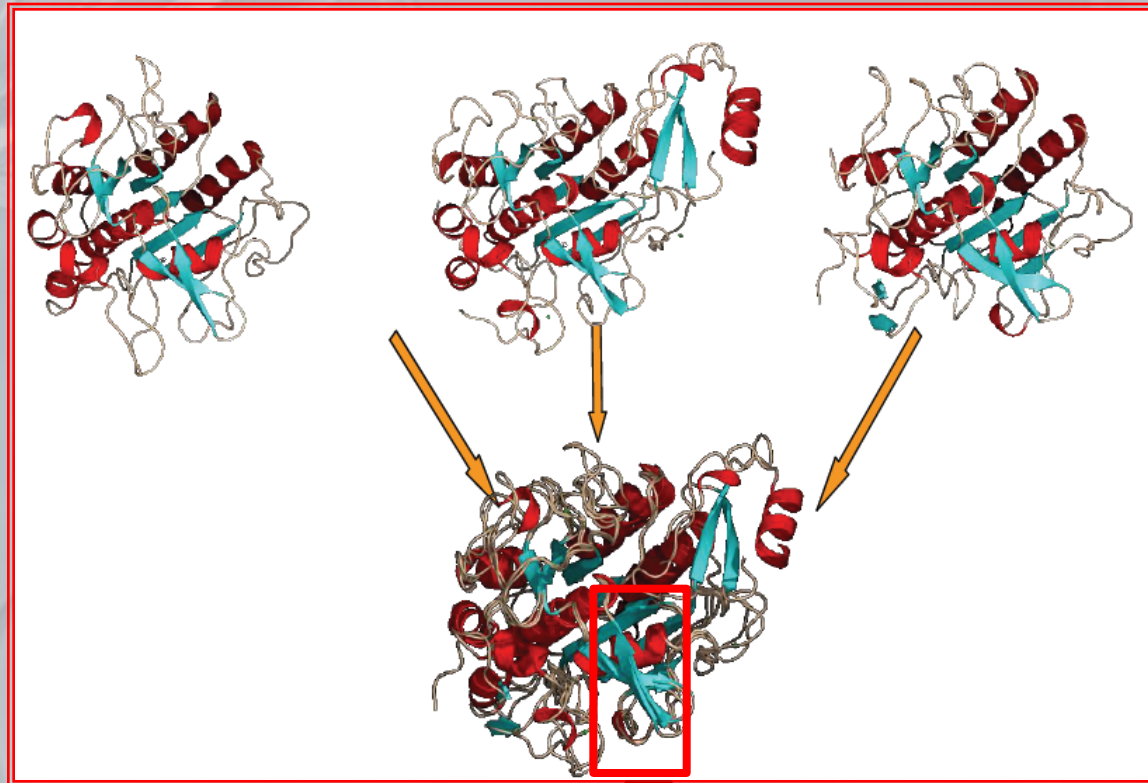
- ✗ Classical tools are used for multiple alignments, such as CLUSTALW
- ✗ Multiple alignments allow to identify those regions within the target protein that are highly conserved w.r.t. the template structures
- ✗ When the identity between the target sequence and the template sequences is less than 30%, automatic multiple alignment methods might not provide results of sufficient quality ⇒ manual adjustments are needed
 - Move the gaps from secondary structure elements and place them inside the superficial loops, where they assume a less important influence, and where it is evolutionarily more likely that indel events occur
 - However, for a sequence identity lower than 30%, the model will be inaccurate

Comparative modeling – 4

3. Building the model

- ✗ Overlap the template structures and find the structurally conserved regions
- ✗ The backbone of the target structure is then aligned with the conserved fragments, forming the core of the model
- ✗ When the template structures diverge, in order to select the correct structure for the target protein, methods for secondary structure prediction must be used
- ✗ Since it is much more likely that the template structures differ in loop regions, w.r.t. regions with a well-defined secondary structure, loops are, in general, modeled separately, after building the core of the model

Comparative modeling – 5



Comparative modeling – 6

4. Loop modeling

- ✗ Select the better loop from a database of known conformations
- ✗ Perform a conformational evaluation
- ✗ Although there exists a wide variety of methods for modeling loops, it is difficult to get an accurate shape for loops longer than six residues

5. Side chain modeling – Once built the backbone model, the positions of the side chain atoms must be determined

- ✗ Searches in rotamer libraries and applications of simple molecular dynamics approaches

Comparative modeling – 7

Loop database

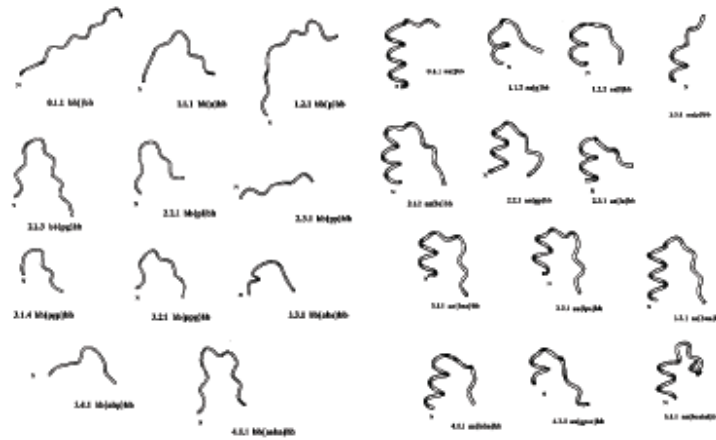


Figure 2. Library of β - β links. One member of each loop class from Table 2 is shown as a smoothed trace of the C $^{\alpha}$ positions. N denotes the N terminus of the loop.

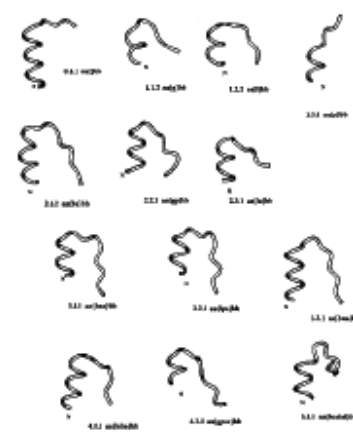


Figure 4. Library of α - β loops. One member of each loop class from Table 4 is shown as a smoothed trace of the C $^{\alpha}$ positions. N denotes the N terminus of the loop.

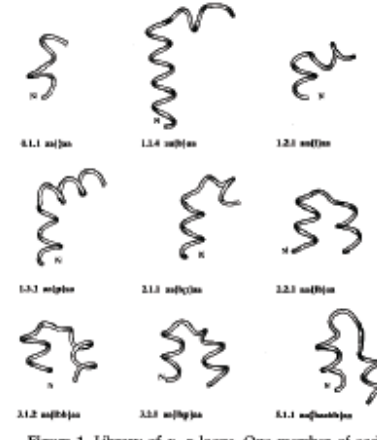


Figure 1. Library of α - α loops. One member of each loop class from Table 1 is shown as a smoothed trace of the C $^{\alpha}$ positions. N denotes the N terminus of the loop. Figures 1 to 8 were generated by PREPI (S. A. Islam & M. J. E. S., available upon request).

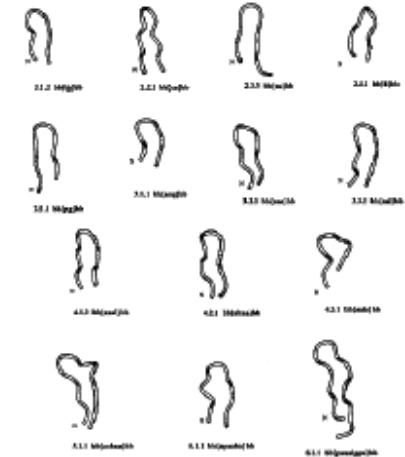
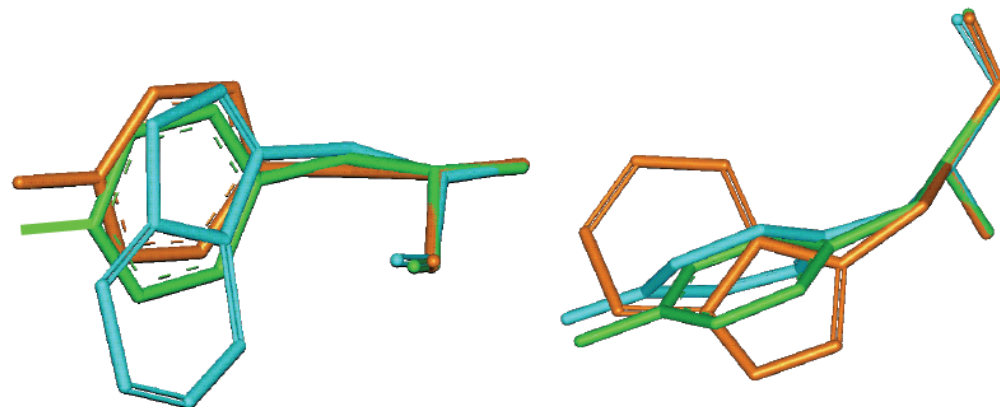


Figure 3. Library of β -hairpins. One member of each loop class from Table 3 is shown as a smoothed trace of the C $^{\alpha}$ positions. N denotes the N terminus of the loop.

Side chain models



Comparative modeling – 8

6. **Model evaluation** – There are several software packages for the quality assessment of 3D structures (PROCHECK, WHAT_CHECK, Verify3D, etc.)
- ✗ Model validation algorithms control abnormalities such as ϕ/ψ combinations that are not in the eligible regions of the Ramachandran plot, steric collisions, and unfavorable bond lengths or angles
 - ✗ After having identified any structural abnormalities, the model is usually adjusted by hand to correct the detected problems

Threading:

the inverse problem of protein folding – 1

- ✦ The evolutionary studies have taught us that proteins with a small sequence similarity can anyway fold in a similar way
- ✦ The number of protein topologies existing in nature is finite and presumably small ($<10^4$)
- ✦ Formulation of the inverse problem w.r.t. that of 3D structure prediction
 - Given a particular sequence, to which structural class can it belong?
 - ➡ Compatibility between the sequence and the 3D structure representative of a particular class
- ✦ The process that consists in considering a particular 3D conformation and in estimating how much it is beneficial for the target protein is called **protein threading**

Threading:

the inverse problem of protein folding – 2

✦ Elements of a threading system

- A **library of 3D structures** representative of the known universe of proteins (obviously not exhaustive of the entire universe)
 - An **energy function** to measure the compatibility between the sequence and the structure
 - A **threading algorithm** that computes the minimum energy alignment between the sequence and the tested structures
 - A **statistical evaluation criterion** for the results
- ✦ Various criteria have been developed, together with many hierarchical databases that identify groups of proteins folded in a similar manner (CATH, SCOP, LPFC, pClass, FSSP, etc.)

Threading:

the inverse problem of protein folding – 3

- ✦ Such databases group proteins with similar structure into categories, such as families, superfamilies and classes of folding
 - In general, libraries contain elements at the “superfamily” or “family” level, in order to limit the structural variations within each folding
 - To identify to which folding family or superfamily a given protein belongs, a medium structure that represents all the peculiarities of a family can be considered, in order to evaluate the quality of the structure obtained if the target protein assume this conformation

Threading:

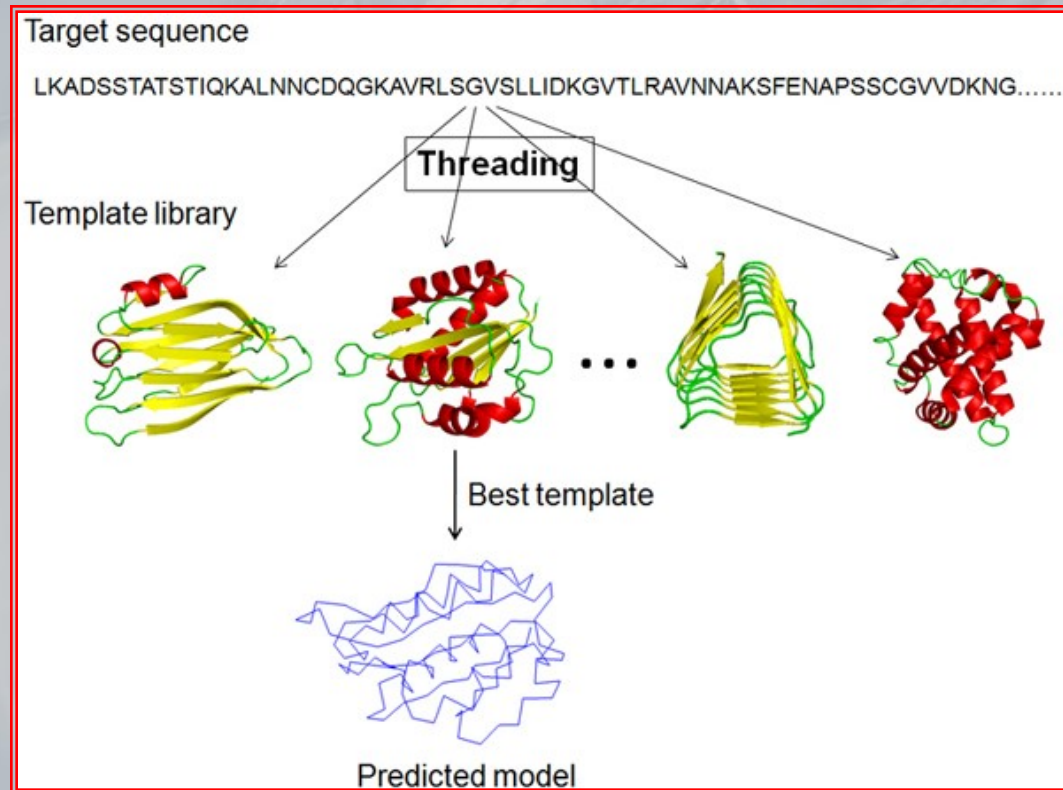
the inverse problem of protein folding – 4

- ✦ The evaluation is repeated for each folding family, in order to select the conformation with the most favorable score, which attests the membership of the target protein to that particular family
- ✦ To measure the threading compatibility between a sequence and a template belonging to the library, an energy function is used
 - Normally, empirical–statistical (*knowledge-based*) energies are employed, rather than physico–chemical ones

Threading:

the inverse problem of protein folding – 5

- ✦ The assignment of a protein sequence to a particular family not only provides a rough approximation of its native structure, but also gives information about its possible functions and on the relationships with other proteins and other biological pathways



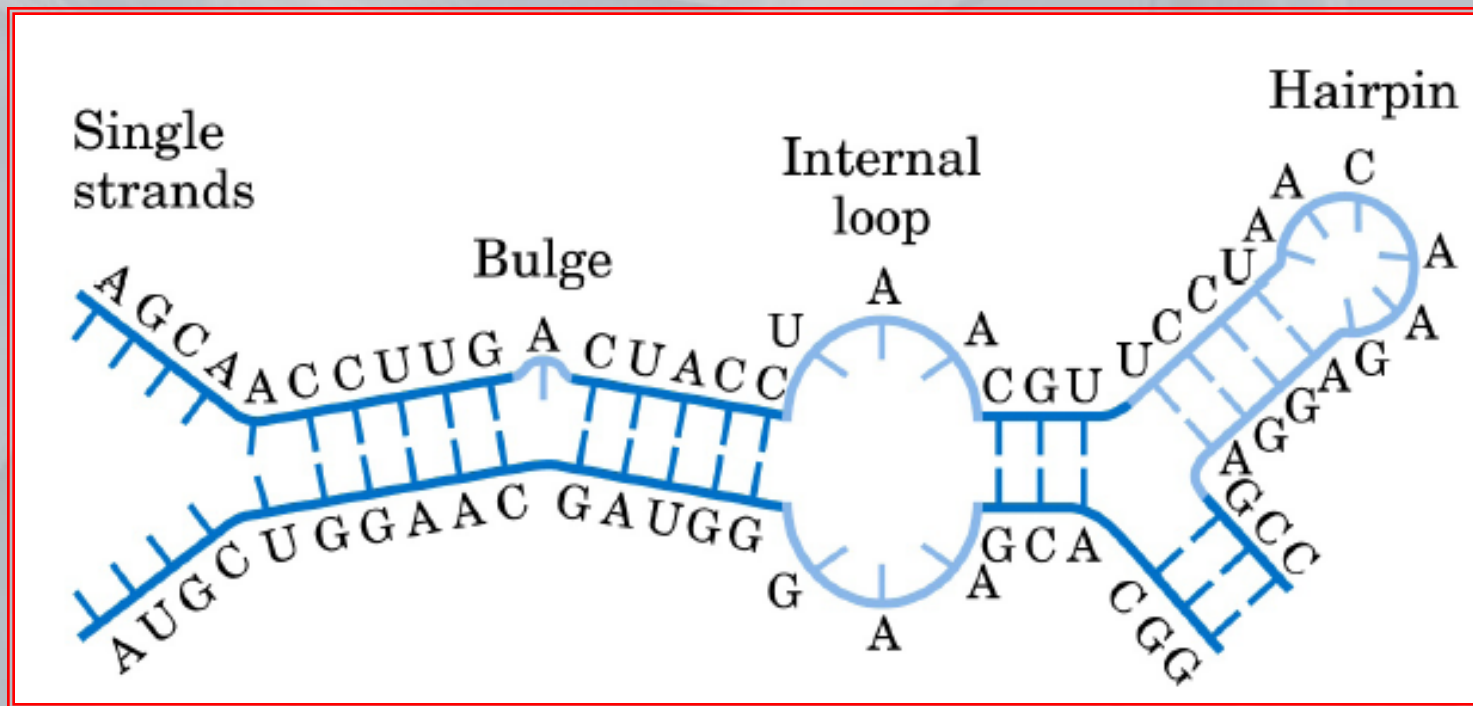
RNA secondary structure prediction – 1

- ✦ Unlike DNA, which usually assumes the well-known double helix conformation, the 2D structure of a single-strand RNA is determined by the sequence of its nucleotides, as well as the structure of a protein is determined by its amino acid sequence
- ✦ The RNA structure, however, is less complex than a protein structure and can be well-characterized by identifying the positions of the secondary structure elements that commonly occur
- ✦ For the RNA, the secondary structure elements are different from those of proteins

RNA secondary structure prediction – 2

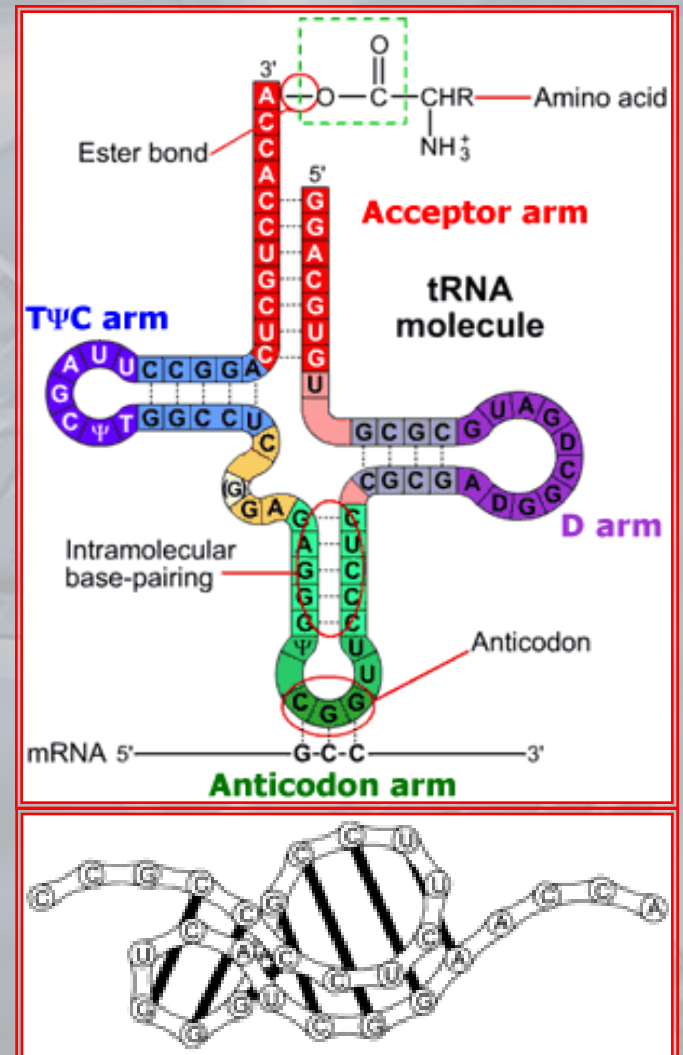
- ✦ As for the DNA, the RNA complementary base pairs form hydrogen bonds, inducing a helical structure
- ✦ However, in a single-strand RNA, base pairing occurs within a single RNA molecule, forming regions obtained by a sequence of bases coupled in a **stem**
- ✦ At the point in which the RNA chain reverses its direction, to allow the base pairing, it forms a **hairpin turn**
 - When a small number of bases along the RNA chain is not complementary, a small **bulge** or a (larger) **loop** can be formed

RNA secondary structure prediction – 3



RNA secondary structure prediction – 4

- ✦ The RNA structure most difficult to be predicted is the **pseudo-knot**, where the bases involved in a loop are coupled with some bases that are outside the loop
- ✦ Given the difficulty of predicting pseudo-knots, most of the RNA secondary structure prediction algorithms ignore them totally, searching only for simpler structure elements



RNA secondary structure prediction – 5

- ✦ Since RNA represents an intermediate language between DNA and proteins, an accurate prediction of the RNA secondary structure is important to understand gene regulation and expression of protein products
- ✦ In fact, it is a recent discovery that many RNAs also have catalytic properties
 - They are called **ribozymes**, and they are involved in the splicing of tRNA molecules, in the activity of ribosomes, in the eukaryotic hnRNA processing, etc.
 - While, usually, ribozymes are found in the context of a protein–RNA interaction, it has been shown that, in some circumstances, they may present a catalytic activity even in the absence of their protein partners

RNA secondary structure prediction – 6

- ✦ Moreover, the RNA acts as a structural scaffold for the DNA, RNA and polypeptide reactions
- ✦ In addition, because some viruses, such as HIV, are encoded in the form of RNA, the RNA secondary structure understanding can support the process of discovery and testing pharmacological agents against these pathogens
- ✦ Therefore, various approaches (most of all based on machine learning) have been devised for the prediction of the RNA secondary structure; the function to be minimized is just the free energy of the folded macromolecule, which implies obtaining the most stable configurations

Concluding...

- ✦ Proteins are complex macromolecules that fold in different types of three-dimensional structures; however, a particular amino acid sequence encodes a unique 3D native structure
- ✦ The secondary structure of a protein can be predicted with a significant accuracy ($>80\%$), using, for instance, recurrent neural networks, hidden Markov models, etc.
- ✦ The tertiary and quaternary structures are much more difficult to predict
 - Folding algorithms
 - Comparative modeling
 - Threading
- ✦ Also important is the prediction of the RNA secondary structure that, so as for proteins, can be obtained with good accuracy by means of machine learning techniques