Course 04: Alignment and 3D Protein Structure

I. Alignment

Introduction

During natural evolution, mutations cause errors during DNA replication because evolution occurs through successive mutations. These errors can be:

- Substitutions (a single change of one nucleotide for another). This is called a transition or a transversion.
- Insertions (addition of one or more nucleotides),
- Deletions (removal of a base or a segment of DNA).

This then results in differences, more or less significant, in the structures (primary, secondary, …) of these sequences, hence the divergence and biodiversity of species.

In bioinformatics, sequence comparison (DNA, RNA, and/or proteins…) relies primarily on the concept of alignment[1] and allows us to determine the degree of similarity between them (similarity or identity by revealing nearby regions in their primary sequences). This can then indicate that:

- The structure (primary, secondary or tertiary) of the two sequences is similar,

- The biological function is similar or different (in the case of dissimilarity),

- The origin of the aligned sequences is common or distant (notion of homology), …

However, comparison to obtain optimal alignment between two biological sequences requires the implementation of calculation procedures (algorithms) and biological models to quantify the notion of similarity between these sequences.

1. Definitions

Alignment: the process by which two (or n) sequences are compared in order to obtain as many correspondences (identities or conservative substitutions) as possible between the letters that compose them.

Local alignment: aligning sequences along a portion of their length

Global alignment: alignment of sequences along their entire length

Optimal alignment: the sequence alignment that produces the highest possible score

Multiple alignment: global alignment of three or more sequences. Gap: an artificial space introduced into one sequence to counterbalance and materialize an insertion into another sequence. It allows for optimization of the alignment between sequences.

indel: "in" = insertion **"del"** = deletion

Similarity: This is the percentage of identities and/or conservative substitutions between sequences. The degree of similarity is quantified by a score. The result of a similarity search can be used to infer sequence homology.

Homology : Two sequences are homologous if they share a common ancestor. Mismatch: A mismatch between two letters. A mismatch can be either the substitution of one character for another, i.e., a mutation, or the introduction of a "gap".

Score: An overall score quantifies homology. It results from the sum of the elementary scores calculated for each position in relation to the two sequences in their optimal matching. It is the total number of "good matches" penalized by the number of mismatches.

2. PROCESSING OF NUCLEIC SEQUENCES (DNA or RNA)

Concept of scoring : The elementary score (denoted "s") is a numerical value assigned to each pair of nucleotides in the two sequences being compared. It takes the value of 1 when the two nucleotides in the two sequences are identical, and the value of zero otherwise. Example:

Séquence1	A	G	C	T	A	C	C	T	G	T	Score global : Total des scores
Séquence1	A	A	G	T	A	G	C	T	T	T	
Point de comparaison	1	2	3	4	5	6	7	8	9	10	
Score élémentaire (s)	1	0	0	1	1	0	1	1	0	1	1+0+0+1+1+0+1+1+0+1=6

In this example, note that at the first comparison point (or comparison site), both sequences contain the same nucleotide A, so the elemental score(s) at this point takes the value of 1 (s = 1).

At the second point of comparison, sequence 1 contains a G and sequence 2 contains an A. They are therefore different at this point, hence an elementary score of zero (s = 0)...

At the 10th comparison point, the two sequences contain the same T nucleotide, therefore an elementary score of 1.

We observe that the sum of the individual scores is equal to six (s = 6). Therefore, there are six identical points between the two sequences; that is, 60% identity between the two sequences ([6/10] x 100). We then say that the overall score between the two sequences is equal to six. The score has thus allowed us to quantify the similarity between the two sequences.

The relationship between the overall score (S) and the elementary scores (s) for two sequences is of the form:

$$S = \sum si\ n\ i=1$$

3. Even alignment

If a new sequence is obtained from genomic sequencing, the first step is to search for similarities with known sequences in other organisms. If the function/structure of similar sequences/proteins is known, it is highly likely that the new sequence corresponds to a protein with the same function/structure. Indeed, it has been found that only about 1% of human genes have no counterpart in the mouse genome, and that the average similarity between mouse and human genes is 85%.

Similarities exist because all cells share a common ancestor cell (a mother cell). Therefore, in different organisms, there could be amino acid mutations in certain proteins because not all amino acids are essential for function and they can be replaced by amino acids with similar chemical properties without changing the structure. Sometimes the mutations are so numerous that it is difficult to find similarities.

The method of calculating gene functions by similarity is called comparative genomics or homology search. Two sequences are homologous when they share a common ancestor.

- **Finding identical segments: The point matrix**

It allows a view (visual method) encompassing the similarities between the regions of the sequences to be compared.

Example implementation: Given two sequences x and y:

x=ACTCGGATT and y=AGCTCGGT

This method involves creating a matrix containing the two sequences (sequence x horizontally and sequence y vertically) and checking the boxes in this matrix only when the nucleotides are identical (Match). When there is no identity, it is called a Mismatch.

		Séquence s								
		A	C	T	C	G	G	A	T	T
Séquence t	A	X						X		
	G					X	X			
	C		X		X					
	T			X					X	X
	C		X		X					
	G					X	X			
	G					X	X			
	T			X					X	X

On this matrix, we observe a diagonal formed by five cells. Therefore, the longest identical segment between the two sequences x and y contains five identical and consecutive nucleotides, which are: **CTCGG**

		Séquence s								
		A	C	T	C	G	G	A	T	T
Séquence t	A									
	G									
	C		X							
	T			X						
	C				X					
	G					X				
	G						X			
	T									

Note: If the two sequences are completely identical, the result is a diagonal:

		Séquence s								
		A	C	T	C	G	G	A	T	T
Séquence t	A	X						X		
	C		X		X					
	T			X					X	X
	C		X		X					
	G					X	X			
	G					X	X	X		
	A	X						X		
	T			X					X	X
	T			X					X	X

- **Nucleic acid pairing methods and aminos**

Even alignment is performed using the following methods:

1. Dot Analysis Matrix

The Dot matrix analysis, first described by Gibbs and McIntyre (1970), is primarily a method for comparing two sequences to find possible alignments between them. The method is also used to search for direct or inverted repeats in DNA and protein sequences, to predict self-complementary regions in RNA sequences, and to predict secondary structure by base pairing. The major advantage of this method is that it examines the best alignment that appears diagonally.

2. Dynamic programming (or DP algorithm)

Dynamic programming is a computational method used to align two DNA or protein sequences. This method is crucial because it yields the best or "optimal" score. The algorithms provide a reliable method for aligning sequences. It compares each nucleotide pair and generates paired and unpaired residues, as well as gaps, in both sequences. The number of paired residues determines the highest possible score. The method has been mathematically proven to optimize the score between two sequences.

3. Word and K-tuple Method

The method allows for the rapid alignment of sequences by searching for fragments in the two sequences that resemble each other (called words or K-tuples) and then reassembling them through a dynamic program.

4. Multiple Alignments

A multiple alignment involves comparing several sequences. This type of alignment is achieved by successively considering all possible pairwise alignments. In practice, one sequence is compared to all others to try to determine the most probable evolutionary path, given the probabilities of different possible substitutions. The more sequences added to the multiple alignment, the more accurate the model becomes regarding the evolutionary history of the families of the compared sequences. This alignment can show that certain residues are identical

in all sequences. Any residue or short sequence that is identical in all sequences of a given group is said to be conserved. Multiple alignments generally provide a better assessment of similarity than pairwise alignments and allow the identification of distantly related members in a gene family that would not have been revealed by a pairwise alignment. This alignment involves superimposing each residue of a sequence with those of one or more other sequences. Furthermore, to construct an optimal alignment, it will often be necessary to add indels (gaps).

5. Molecular sequence scoring system

The choice of scoring system assigns scores to identical residues (matches), different residues (mismatches), substitutions, insertions, and deletions in DNA and protein sequences. This system is applied in global alignments (Needleman-Wunsch algorithm) and local alignments (Smith-Waterman algorithm). For DNA alignment, a simple positive score is given to identical residues and a negative score to different residues, and gaps are taken into account. On the other hand, to score identical and different residues in a protein sequence, it is important to know how one amino acid is substituted by another.

6. Bank Comparison Programs

Searching for similarities in sequence databases, why?

• To know if my sequence resembles others already known

• Find all the sequences of the same family

• Search for all sequences that contain a given pattern

The size of the sequence banks necessitated the development of specific algorithms to perform the comparison of a sequence with a database because the standard algorithms for comparing two sequences are generally too long on conventional machines.

Most of these programs are heuristic methods. Their goal is to filter the database data in successive steps because few sequences will have similarities with the compared sequence.

These methods use certain approximations to quickly eliminate irrelevant situations and thus identify sequences in the database likely to be related to the searched sequence. These programs calculate a score to highlight the best local similarities they have observed.

The two most commonly used types of programs by biologists are software:

1. FASTA

The software actually combines several search programs with databases:

• The FASTA program which compares respectively a nucleic acid sequence with a nucleic acid base or a protein sequence with a protein base.

• The TFASTA or TFASTX programs which compare a protein sequence with translated nucleic bases.

• FASTX or FASTY programs that compare a translated nucleic acid sequence with protein bases.

2. BLAST : Basic Local Alignment Search Tool

BLAST is a basic local alignment search tool. BLAST searches protein and DNA databases for sequences (subjects) that resemble our sequence (query) used as a keyword.

This software actually has several database comparison programs:

- BLASTN (to compare a nucleic acid sequence against a nucleic acid base),

- BLASTP (To compare a protein sequence against a protein base),

- BLASTX (comparison of nucleic acid sequence (translated into 6 phases) against protein base),

- TBLASTN (protein sequence versus nucleic base comparison (translated in 6 phases)),

- TBLASTX (comparison of nucleic acid sequence (translated in the 6 phases) against nucleic acid base (translated in the 6 phases))

II. Protein structure

While DNA is the physical carrier of biological information, protein is its reflection, and at the molecular level it is already a veritable functional machine, ensuring both vital structural and dynamic functions. Thus, a protein is a linear polymer made up of different basic units, amino acids (aa) or residues.

An amino acid (aa) consists of a central carbon atom (alpha carbon or Cα) bonded to a carboxyl group (COOH), an amine group (NH2), a hydrogen atom (H), and a radical (R). The twenty amino acids differ in the nature of this radical, which confers various properties such as charge,

flexibility, steric hindrance, and hydrophobicity. The structure of a protein can be described at several levels, each providing specific information.

1. **The primary structure**

Amino acids are linked together by peptide bonds between the carboxyl group (COOH) of one residue and the amine group (NH2) of the next. The chain thus formed is called the "main chain" or "backbone," while the side chains are referred to as "side chains." A molecule is called a peptide when the number of residues is less than 50, and a protein when it exceeds 50. The main information provided by the primary structure is the order or sequence of amino acids that form the protein molecule.

Secondly, the primary structure allows the calculation of specific intrinsic values such as pI, PM, hydrophobicity, etc.

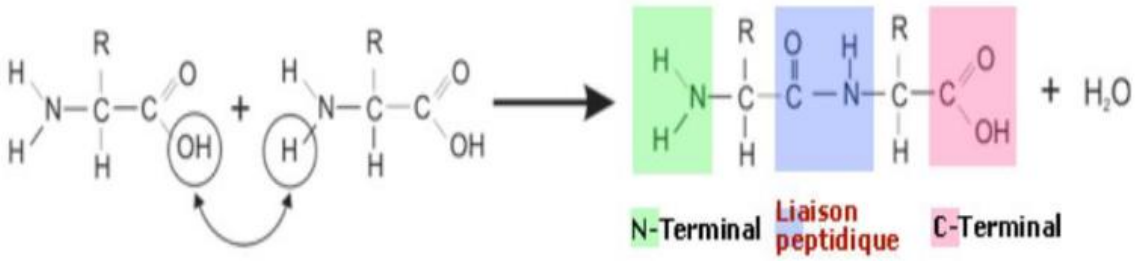

Figure: Formation of the peptide bond. The linking of two amino acids is accompanied by the loss of a water molecule.

2. **Secondary structure**

Secondary structure refers to the regular spatial organization of the polypeptide chain. A protein can thus be described by a sequence of secondary structure elements that adopt conformations that are clearly favored because they are stabilized by hydrogen bonds between the amine (-NH_2) and carbonyl (-CO) groups of the peptide backbone. It is generated by the rotation of the atoms of the peptide chain relative to one another during in vivo chain synthesis. The possible angles and the structures they most frequently produce are shown in the RAMACHANDRAN table.

Two main types of secondary structure are recognized:

2.1 The alpha helix: When the carbon skeleton of a protein adopts a periodic helical folding, it is called an alpha helix. The α helix rotates the carbon chain around itself by one turn approximately every 4 amino acids. It is stabilized by hydrogen bonds between the carbonyl

group of the peptide bond following amino acid #1 and the amine group of the peptide bond preceding amino acid #5, and then similarly between amino acids 2 and 6, etc.

- **Tours and loops**

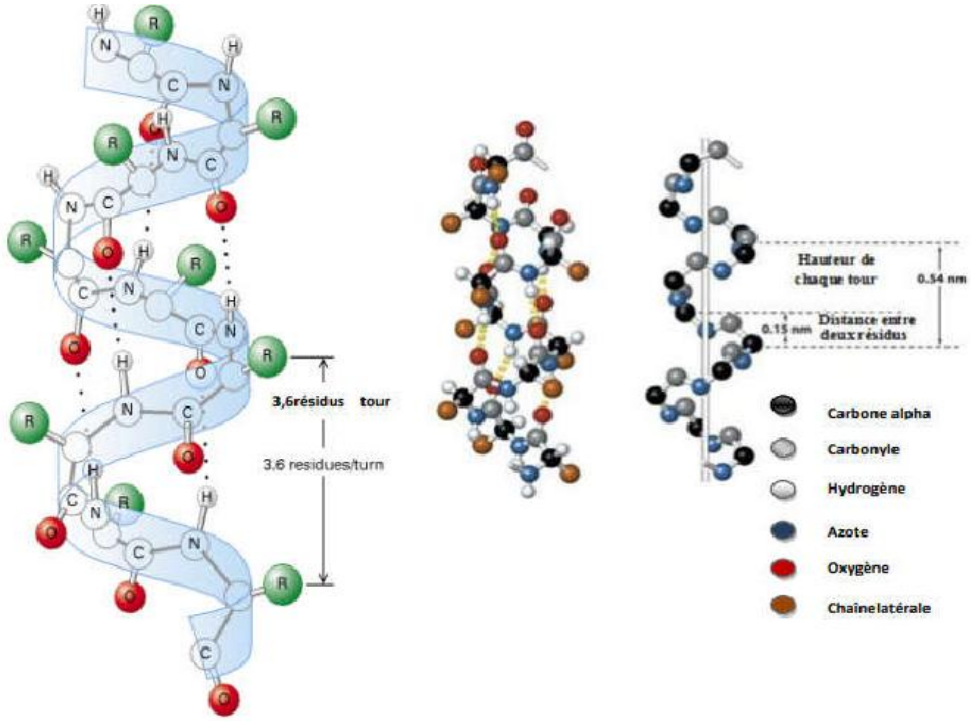

Figure: Representation of the alpha helix structure. (a): Ribbon structure; the R radicals are outside the chain. (b): Details of the ball structure

2.2 β-pleated sheets: Unlike the alpha helix, these are not continuous segments of a single polypeptide chain, but combinations of different segments not necessarily following one another and originating from one or more polypeptide chains.

These β strands are arranged side by side in such a way that hydrogen bonds can form between the CO and NH groups of neighboring strands. The two strands can be parallel or antiparallel.

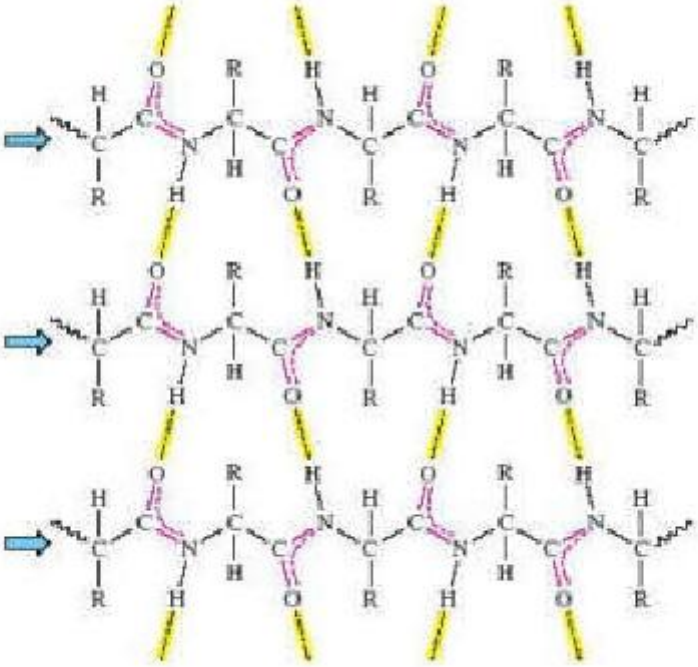

Figure: Representation of the structural model of beta sheets

- **The turns:** This is the structure that connects two antiparallel β strands. The turns are generally short: 2 to 4 amino acids outside the strands.

- **Loops:** Protein structures are often combinations of helices and sheets linked by loops of highly variable lengths: from 1 to 12 residues (or even up to 22), most frequently 1, 3, 4, or 7 residues. Comparison of three-dimensional structures shows that the loops adopt a limited number of conformations.

3. The tertiary structure

It corresponds to the folding and assembly of the different elements of the secondary structure. This structure is, in fact, the three-dimensional structure (3D structure) of the protein. Protein folding and stability are governed by several forces: hydrogen bonds, Van der Waals forces, electrostatic forces, and hydrophobic interactions.

III. Bioinformatics approaches for predicting 2D and 3D protein structures

The primary sequence of proteins contains all the information that will determine the secondary and tertiary structures of proteins. The function of a protein depends largely on its structure, that is, on how the amino acid chain folds.

Identifying a protein's function begins with searching for sequence similarity by comparing its primary sequence to other sequences (alignments). If the similarity is significant (>70%), then two postulates can be made:

- Firstly, the two sequences are homologous, in other words, they are phylogenetically linked.

- Secondly, homology between two sequences may suggest that the proteins have the same structure or even the same function.

Searching for similarity within a sample of sequences involves multiple alignment. This allows us to:

- Determine if the protein organizes itself into

- Identify preserved patterns

- To trace its evolution and phylogenetic links

- Identify its function

There are many multi-alignment programs. The most used are Clustal Omega, Multalign, Clustal W, Dialign, T-coffee, MAFFT and MUSCLE.

- **Methods for predicting secondary structures**

A protein's function depends largely on its structure, that is, how it folds around itself. Predicting protein function begins with 2D (and 3D) structural predictions. Predicting the 2D structure involves predicting local conformational features: α helices, β sheets, and bends.

Early methods for predicting secondary structures relied on statistical analysis of the propensity of amino acids to be found in one of the secondary structure components. The methods of Chou and Fasman and GOR are the most widely used.

- **Methods for predicting tertiary structures**

The 3D structure is important because it determines the biochemical properties and biological function of proteins. Generally, the 3D structures of proteins are determined either by X-ray crystallography or by NMR. However, these methods are expensive and remain a time-consuming process.

Two types of methods are used for predicting the 3D structure of proteins: comparative modeling and so-called ab initio methods.

1. **Comparative modeling**

- **Homology modeling**

Its principle consists of aligning the sequence of a given protein whose structure is unknown (target) with the sequence of one or more proteins having a known experimental structure (NMR or X-rays) (patterns or references).

- **Fold recognition methods (Threading)**

This method is recommended when the sequence similarity between the target and the patterns is between 20 and 30%. It consists of threading the target sequence onto a folding library in order to determine the structures that best match the sequence based on an energy criterion or score.

2. **Ab initio methods**

These methods are used to predict the tertiary structure of the target for very low percentages of sequence identity from its primary amino acid sequence, based on their physicochemical interactions between the atoms of the residues.

Within this method, we distinguish two types of methods called: pure ab initio, based solely on physical principles, and de novo, which use a battery of information from databases.