CHAPTER III.
AIGNEMENT OF BIOLOGICAL SEQUENCES

Introduction

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

- Substitutions (a point change of one nucleotide for another). These are known as transitions or transversions,
- Insertions (the addition of one or more nucleotides),
- Deletions (removal of a base or a segment of DNA).

As a result, there are varying degrees of differences in the structures (primary, secondary, etc.) of these sequences, leading to species divergence and biodiversity. In bioinformatics, sequence comparison (DNA, RNA, and/or proteins) primarily relies on the concept of alignment and allows for determining the degree of similarity between them (similarity or identity by revealing closely related regions in their primary sequences). This can indicate that:

- The structure (primary, secondary, or tertiary) of the two sequences is similar,
- The biological function is either close or different (in the case of dissimilarity),
- The origin of the aligned sequences is either common or distant (concept of homology), etc.

However, comparing sequences to obtain an optimal alignment between two biological sequences requires the implementation of computational procedures (algorithms) and biological models that allow quantification of the degree of similarity between these sequences.

1. Definitions

- **Alignment**: The process by which two (or more) sequences are compared to obtain the highest number of matches (identities or conservative substitutions) between the characters that compose them.
- **Local alignment**: Alignment of sequences over a portion of their length.
- **Global alignment**: Alignment of sequences over their entire length.
- **Optimal alignment**: The alignment that produces the highest possible score.
- **Multiple alignment**: Global alignment of three or more sequences.

- **Gaps**: An artificial space introduced in a sequence to counterbalance and materialize an insertion in another sequence, optimizing the alignment between sequences.
- **Indel**: "in" = insertion, "del" = deletion.
- **Similarity**: 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 have a common ancestor.
- **Mismatch**: A lack of correspondence between two characters. A mismatch can be either a substitution of one character for another (a mutation) or the introduction of a gap.
- **Score**: A global score quantifies homology. It results from the sum of elementary scores calculated for each corresponding position in the optimal alignment of the two sequences. It is the total number of "good matches" penalized by the number of mismatches.

2. PROCESSING NUCLEOTIDE SEQUENCES (DNA or RNA)

Score concept: The elementary score (denoted "s") is a numerical value assigned to each pair of nucleotides from the two sequences being compared. It takes the value of 1 when the nucleotides are identical and 0 otherwise.

Sequence1	A	G	C	T	A	C	C	T	G	T	Global scores: Total of scores
Sequence2	A	A	G	T	A	G	C	T	T	T	
Comparison points	1	2	3	4	5	6	7	8	9	10	
Elementary score (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, at the first comparison point, both sequences have the same nucleotide A, so the elementary score (s) is 1. At the second comparison point, Sequence 1 has a G and Sequence 2 has an A, so they are different, resulting in an elementary score of 0. At the tenth comparison point, both sequences contain the same nucleotide T, so the elementary score is 1.

The sum of the elementary scores is six, meaning there are six identical points between the two sequences, resulting in 60% identity between them. Therefore, the global score between the two sequences is six. The score helps quantify the similarity between the sequences. The relationship between the global score (S) and the elementary scores (s) for two sequences is represented as:

$S = \sum si$ (from $i = 1$ to n)

3. Pairwise Alignment

If a new sequence is obtained through 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. For instance, it has been found that only about 1% of human genes have no counterpart in the mouse genome, and the average similarity between mouse and human genes is 85%.

Similarities exist because all cells share a common ancestor (a "mother cell"). Therefore, in different organisms, mutations in certain proteins may occur, as not all amino acids are crucial for function and can be replaced by others with similar chemical properties without changing the structure. Sometimes, mutations are so numerous that it becomes difficult to find similarities. The method for calculating gene functions through similarities is called comparative genomics or homology search. Two sequences are homologous when they share a common ancestor.

3.1. Sequence Similarities and Score

After sequencing, biologists often have no idea of the functions of the genes found. In hopes of discovering clues about their functions, they try to find similarities between newly sequenced genes and others whose functions are already known.

The following game illustrates this: transform one English word into another by going through a series of intermediate words, with each word differing from the next by only one letter. For example, transforming "head" into "tail" requires four intermediaries:

head → heal → teal → tell → tall → tail.

For biological sequences, it is known how one sequence can mutate into another. First, there are point mutations, where one nucleotide or amino acid is changed into another. Second, deletions occur when a nucleotide, amino acid, or an entire subsequence is removed. Third, insertions occur when an element or subsequence is inserted into the sequence.

An alignment can be interpreted as an editing task: finding the minimum number of elementary editing operations needed to transform one sequence into another. The three operations are:

(a) Insertion: inserting one or more letters.
(b) Deletion: removing one or more letters.
(c) Substitution: replacing one letter with another.

From an evolutionary perspective, these three operations can be interpreted as mutations, and the editing task can be seen as an attempt to reconstruct evolutionary history by considering these three elementary mutations. For example, the following alignment:

```
BIOINFORMATICS          BIOI-N-FORMATICS
                   ⟶
BOILING FOR MANICS      B-OILINGFORMANICS
```

(1) Deletion of I	BOINFORMATICS
(2) Insertion of LI	BOILINFORMATICS
(3) Insertion of G	BOILINGFORMATICS
(4) Substitution of T by N	BOILINGFORMANICS

The two sequences seem very similar. Note that insertions or deletions cannot be distinguished if both sequences are presented (was the I removed from the first sequence or inserted into the second?). Thus, both cases are denoted by "-".

The task of bioinformatics algorithms is to find, from two series (the left side in the above example), the optimal alignment (the right side in the above example). The optimal alignment is the arrangement of the two series in such a way that the number of mutations is minimized.

Alignment can be global (over the entire length of the sequence) or local (over the most conserved regions), depending on the presumed relationship between the sequences. An alignment score is defined to determine the best alignment of two sequences and quantify their similarity.

3.2. Identity Matrix

The identity matrix or dot matrix is a tool for representing alignments, where one sequence is written horizontally at the top and the other vertically on the left. This forms a matrix where each letter of the first sequence is paired with each letter of the second sequence. For every letter match, a dot is placed in the corresponding position in the matrix. Which pairs appear in the optimal alignment? As we will see, every path through the matrix corresponds to an alignment (Figures 1a and 1b).

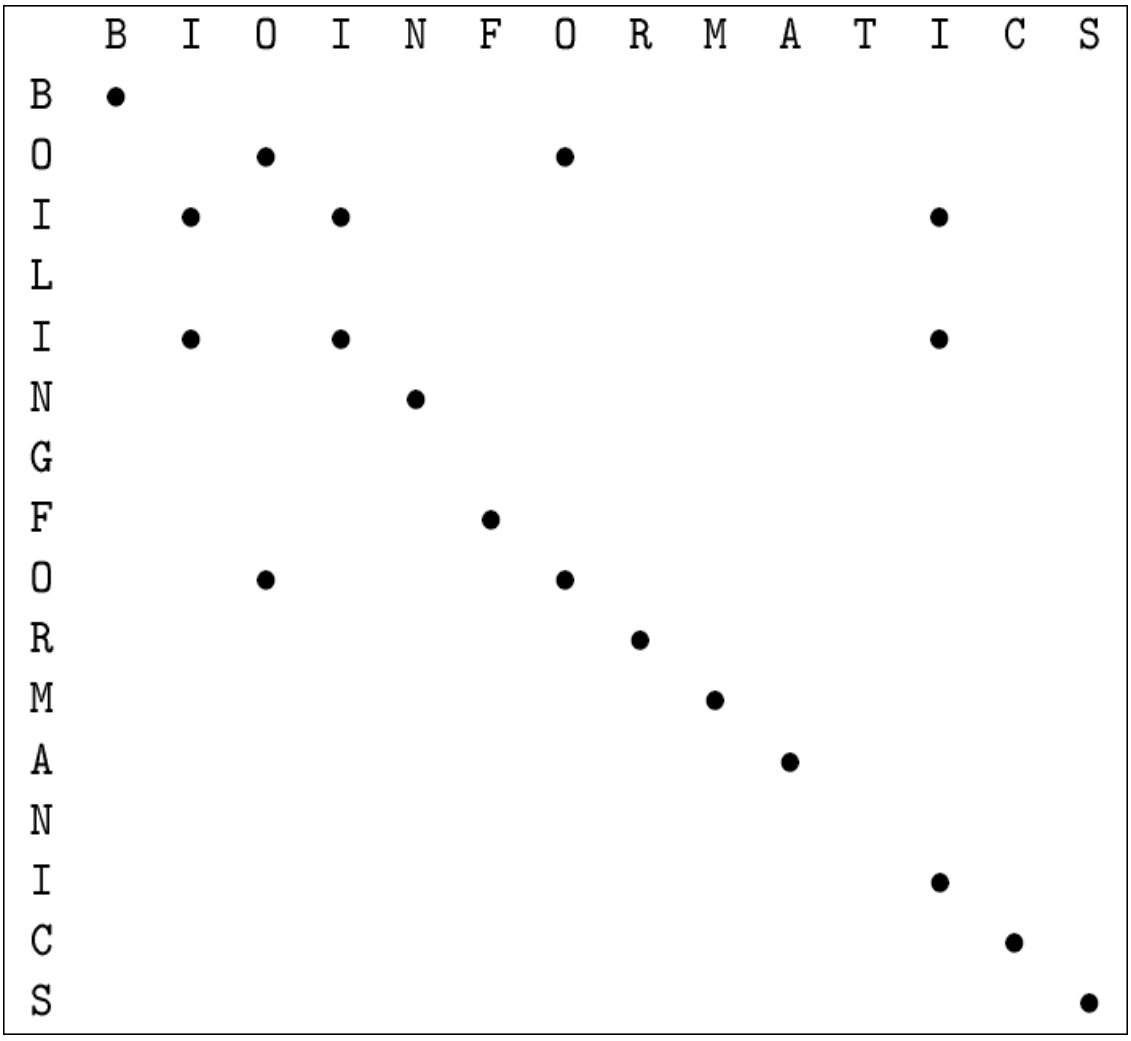

Figure 1a. Operational principle of the identity matrix.

Rules: You can move horizontally "→", vertically "↓", and you can only move diagonally " " if you are in the dot position.

Task: Make as many diagonal movements as possible when moving from the top left corner to the bottom right corner.

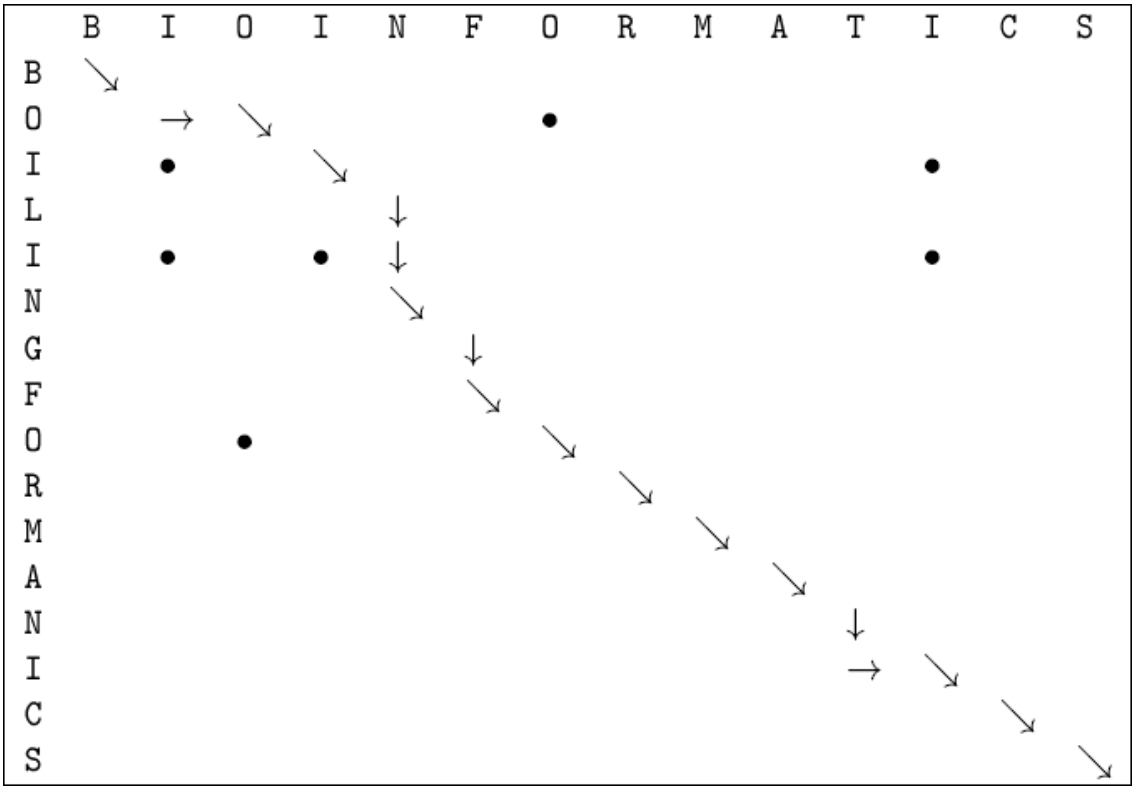

Figure 1b. Operational principle of the identity matrix.

The number of diagonal movements " " represents the matches, and the number of scores: "→" corresponds to "-" in the vertical sequence, "↓" to "-" in the horizontal sequence, and the combination "→↓" or "↓→" corresponds to a mismatch. Therefore, each path through the matrix corresponds to an alignment, and each alignment can be expressed by a path in the matrix.

In Figure 2, the dots on the diagonals correspond to matching (similarity) regions. It represents Dot Matrices for the comparison of human triosephosphate isomerase (TIM) protein with that of yeast, *E. coli*, and *Archaeon*. For yeast, the diagonal is complete, and for E. coli, small gaps are visible, but Archaeon does not show an extended diagonal. Thus, human TIM corresponds most closely with yeast TIM, followed by *E. coli* TIM, and has the lowest similarity with *Archaeon* TIM.

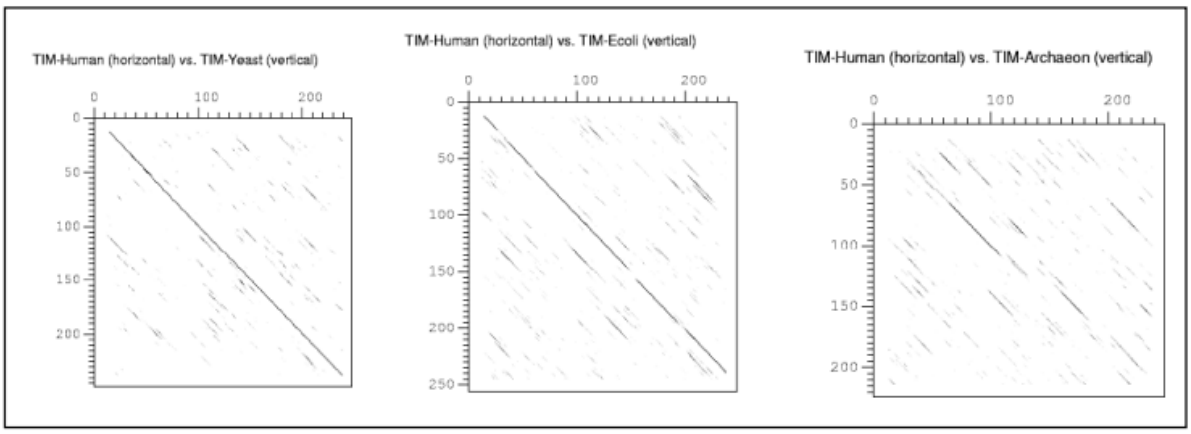

Figure 2. Dot matrix of human triosephosphate isomerase with the same protein in yeast, *E. coli*, and *Archaeon*.

Yeast provides the best match as the diagonal is almost complete. *E. coli* has some gaps in the diagonal. *Archaeon* shows the lowest similarity. However, the 3D structure and function are the same for all proteins.

4. Multiple Alignment

The goal of comparing protein sequences is to discover "biological" similarities (i.e., structural or functional) among proteins. Biologically similar proteins may not exhibit strong sequence similarity, and we would like to recognize structural/functional resemblance even when the sequences are very different.

Simultaneous comparison of many sequences often reveals similarities that are invisible in pairwise sequence comparisons ("pairwise alignment whispers... multiple alignment shouts"). Multiple alignment is the basis for studying protein families and functional domains. Its purpose is to reveal sequence or structural similarities in a family of sequences that are evolutionarily or functionally related.

It is important to carefully analyze the result of the multiple alignment before proceeding to construct the phylogenetic tree and to adjust the software parameters appropriately. We will perform the multiple alignment of the sequence set using the ClustalW tool. These sequences belong to the family of transcription factors of the "Basic Leucine Zipper" type. These are genes that code for proteins that regulate mRNA transcription.

The result of a part of the multiple alignment of this set of sequences is as follows:

```
Solanum.tuberosum1466pb          -GGCTGCAC-----ACCAAT-CAGCT--------CAGGGTC------TCC 1172
Triticum.monococcum1062pb        TGACCACAG-----GC-AGT-CTGCC--------CGTGCAC------TTC 931
Rattus.norvegicus1785pb          GGGCAGCCC-----ACCAG--CAGCTG-------CAGGAAGCTGATATCC 1427
Zea.mays1236pb                   TGGTAGCGG-----TC--AT-CAGCCC-------CGAGCGCACGGTGTAC 1047
Oryza.sativa1272pb               TGGTAG-AA-----GCTAG---AGCTT-------AGCTAGC--------- 1099
Xenopus.laevis1188pb             CGACAGCAACGACTGCTAA---AGTTGC------CGAAAGC--------- 1049
Arabidopsis.thaliana1489pb       TAACCAGAA--------AAA-GAGTCAT------TGGTTTT--------- 1281
Triticum.aestivum1585pb          TTGTAGAAGAAGGATCCATCTCTGCCTTTCTTCTCAGACATAGTCATGCA 1324
                                                       *

Solanum.tuberosum1466pb          TT-----------GCCTTAGG------AGAGT----ACTTTAAACGTC- 1199
Triticum.monococcum1062pb        TT-----------GTGATAAG------TGATT----ACTCATCCCGGC- 958
Rattus.norvegicus1785pb          TTAAACTGAGTCAGGCATCAAGA----CTAAGC----ACTCAGCAAGTG- 1468
Zea.mays1236pb                   ATA----------GCTTTCAG-----TAGATCG--AATTCCAGGCATG- 1078
Oryza.sativa1272pb               --------------TAGCGAG------AGAGTG--AGCTCAGCTAAGC- 1125
Xenopus.laevis1188pb             -------------GCAGCAGA-------GATCCCTAATACTATAAAAG- 1077
Arabidopsis.thaliana1489pb       ----------------GTGATT----TTGATTG---AGGTAACTATTG- 1306
Triticum.aestivum1585pb          TCATGCT--------CCTCGAGAGTCTCTGAATGAGCACATGATCCATGG 1366
                                                        *

Solanum.tuberosum1466pb          TTCG-----TGCTCTTA-----GCTCACTTTGGGC--------TGGTCGT 1231
Triticum.monococcum1062pb        TTCG-----TGCCCTAA-----GTTCTCTTTGG-C--------T--TTGC 987
Rattus.norvegicus1785pb          CTGGA---CTGGTTTGACTCTCGATTGCCCAAGCCAGCAGAAGTGGTAGT 1515
Zea.mays1236pb                   TCCA---------TCAACAAGCAGTTTCTTC------------TCGTCAT 1107
Oryza.sativa1272pb               TTAATTAGCTGGCTTGAT---TGCTTGCTTTG-----------TGGCTGG 1161
Xenopus.laevis1188pb             TAGG-----------GAT-----GTCCTTTTGATA---------CGTCAC 1102
Arabidopsis.thaliana1489pb       TCTG-----TATTTTTAT------TTACTGTATGACTCAGCGACGGTAAA 1345
Triticum.aestivum1585pb          TTAATTAACAGGATCTAC-----ATCCTCCTG-----------TGCTCAT 1400
                                               *       *
```

This alignment shows many gaps that distort the interpretation. This is because our sequences belong to individuals with completely different taxonomy. We aligned sequences from frog, wheat, etc. We will redo this alignment, but this time with sequences from the plant kingdom only. The order of individuals appearing in the result of the multiple alignments is as follows:

1. *Triticum aestivum*
2. *Oryza sativa*
3. *Zea mays*
4. *Arabidopsis thaliana*
5. *Solanum tuberosum*
6. *Triticum monococcum*

The result of part of the multiple alignments:

```
gi|62736387|gb|AY914051.1|          GAGAAGATCGGCTACTGGAGGTACATCACCATCTTCAGGCACCTAAAGG---CCAACCCG
gi|33943625|gb|AY346329.1|          GA------------------CAAGGACGCCCTCGCCGCCGAGATCGCCG---ACCTCCGG
gi|308044466|ref|NM_001196644.1|    GA------------------GAAGCACACGCTCCTCAAGCAGCTGGAGA---AGCTAGCC
gi|334185982|ref|NM_001203162.1|    --CAAGGCTCCATTGTGGCACAAACCTCACCTGGTGCTTCATCTGTTAGATTTTCTCCCA
gi|575417|emb|X82544.1|             TAGAATTGCGCATTCTTGTCGAGAGTT--GCTTGAATCAC-TATTTTGATCTCTTTCGCT
gi|461682445|gb|JX424318.1|         CTGAGCTGCGTAGTGTTGTTGAGAAGA--TCATGTCACAC-TATGATGAGATTTTTAAGC
                                                :.       .:          .*    .      :.

gi|62736387|gb|AY914051.1|          GAGTACCAGGTGTACCCCATCTTCAAGTACTTCGAGAACTGGTGTCAGGACGAGAACCGG
gi|33943625|gb|AY346329.1|          GACAGGGTGGACGGCCAGATGTCC---------GTCAAGCTGGAGGCCGTGGCCG---CG
gi|308044466|ref|NM_001196644.1|    GAGATGCTGCACGAGCCGCGGGGCAAGTACAGCGGCAATGCGGACGCCGCCGGCG---CC
gi|334185982|ref|NM_001203162.1|    CAACAAGCACGCAAAAGAAACCTGATGTTC---CAGCCAGACAAACTAGTATTTC---AT
gi|575417|emb|X82544.1|             TGAAAGCTACAGCCGCAAATGCTGATGTTC---TCTACCTTATGTCTGGCACATG-----
gi|461682445|gb|JX424318.1|         AAAAAGGAAATGCAGCCAAAGCAGATGTCT---TTCATGTGTTATCAGGCATGTG-----
                                     .     .     . .                .     .    *

gi|62736387|gb|AY914051.1|          CATGGCGATTTCTTCTCCGCGCTGCTCAAGGCGCAGCCGCAGTTCCTCAATGACTGGAAG
gi|33943625|gb|AY346329.1|          GACGAACACCAGCCGCCTCCGCCGCCGCCGCCGCCGCCACTGGCGTATAACAGCAAGGTG
gi|308044466|ref|NM_001196644.1|    GGGGACGACGT-------------GCGCTCGGGCGTCGGCGGCATGAA-GGACGAGTTT
gi|334185982|ref|NM_001203162.1|    CACGAGATGATTCTGATGACGATGATCTTGATGGAGACGCAGATAAT-------------
gi|575417|emb|X82544.1|             ---GAAGACATCAGCTGAGCGTTTCTTCTTGTGGATTGGGGGATTT--------------
gi|461682445|gb|JX424318.1|         ---GAAGACACCAGCTGAGAGGTGTTTCCTATGGCTTGGAGGTTTC--------------
                                      *.  :                        * .   .  *

gi|62736387|gb|AY914051.1|          GCCAAGCTCTGGTCACGCTTCTTCTGCCTCTCGGTGTATATAAC---------CATGTAC
gi|33943625|gb|AY346329.1|          GTGGACGGCTCGACGGACAGCGACTCGAGCGCGGTGTTCAACGAGGAGGCGTCGCCGTAC
gi|308044466|ref|NM_001196644.1|    GCAGACGCCGGGGCCGCGCCCTACTCGTCCGAGGGCGGTGGCGGTGGCAAGTTCGCGCAC
gi|334185982|ref|NM_001203162.1|    -------------GGAGATCCTACTGATGTGAAGCGTGCTAGGA-----------GGATG
gi|575417|emb|X82544.1|             -------------CGCCCCTCCGAACTTCTAAAGGTTCTCACGC-----------CACAT
gi|461682445|gb|JX424318.1|         -------------CGACCTTCTGAGCTTTTAAAGCTTCTTTCGA-----------CCCAA
                                                 *  .       ..*        .              :
```

Let's note that there are fewer gaps and many more identities. We can also use protein sequences to perform a multiple alignment for phylogenetic construction. For this, we need a set of sequences belonging to the same protein family. The presence of motifs generally suggests a conserved function during evolution. They are highlighted by a multiple alignment and are represented by consensus sequences. In the case of proteins, their search helps identify sites involved in specific biological functions: catalysis, ligand binding, regulation, etc.

Conserved regions could harbor active sites, allowing the preservation of vital functions in living beings, such as respiration, photosynthesis, membrane transport, etc.

Example of sequence alignment by BLAST/NCBI

Figure 3 represents the result of an alignment of the partial 16S rRNA gene sequence of *Aeromonas veronii* obtained from GenBank, via the BlastN program.

Aeromonas veronii

TACTTTTGCCGGCGAGCGGCGGACGGGTGAGTAATGCCTGGGGATCTGCCCAGTCGAGGGGGATAACTACTGG
AAACGGTAGCTAATACCGCATACGCCCTACGGGGGAAAGCAGGGGACCTTCGGGCCTTGCGCGATTGGATGAA
CCCAGGTGGGATTARCTAGTTGGTGAGGTAATGGCTCACCAAGGCGACGATCCCTARCTGGTCTGAGAGGATG
ATCAGCCACACTGGAACTGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGTGGGGAATATTGCACAATGGG
GGAAACCCTGATGCMGCCATGCCGCGTGTGTGAAGAAGGCCTTCGGGTTGTAAAGCACTTTCAGCGAGGAGGA

AAGGTTGGTAGCTAATAACTGCCAGCTGTGACGTTACTCGCAGAAGAAGCACCGGCTAACTCCGTGCCAGCAG
CCGCGGTAATACGGAGGGTGCAAGCGTTAATCGGAATTACTGGGCGTAAAGCGCACGCAGGCGGTTGGATAAG
TTAGATGTGAAAGCCCCGGGCTCAACCTGGGAATTGCATTTAAAACTGTCCAGCTAGAGTCTTGTAGAGGGGG
GTAGAATTCCAGGTGTAGCGGTGAAATGCGTAGAGATCTGGAGGAATACCGGTGGCGAAGGCGGCCCCC

Aeromonas veronii bv. *sobria* strain ER.1.24 16S ribosomal RNA gene, partial sequence, Length=1029, Score = 1195 bits (647), Expect = 0.0, Identities = 650/653 (99%), Gaps = 0/653 (0%), Strand = Plus/Plus**.**

This figure showcases the alignment result of the partial 16S rRNA gene sequence of *Aeromonas veronii*, obtained from GenBank, using the BlastN program. It highlights the high degree of similarity between the query sequence and the subject sequence with almost no gaps, showcasing a 99% identity score over 653 base pairs. This bioinformatics analysis is a key step in understanding genetic relationships and evolutionary links between species.

```
Query  1    TACTTTTGCCGGCGAGCGGCGGACGGGTGAGTAATGCCTGGGGATCTGCCCAGTCGAGGG  60
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  61   TACTTTTGCCGGCGAGCGGCGGACGGGTGAGTAATGCCTGGGGATCTGCCCAGTCGAGGG  120

Query  61   GGATAACTACTGGAAACGGTAGCTAATACCGCATACGCCCTACGGGGGAAAGCAGGGGAC  120
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  121  GGATAACTACTGGAAACGGTAGCTAATACCGCATACGCCCTACGGGGGAAAGCAGGGGAC  180

Query  121  CTTCGGGCCTTGCGCGATTGGATGAACCCAGGTGGGATTARCTAGTTGGTGAGGTAATGG  180
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  181  CTTCGGGCCTTGCGCGATTGGATGAACCCAGGTGGGATTAGCTAGTTGGTGAGGTAATGG  240

Query  181  CTCACCAAGGCGACGATCCCTARCTGGTCTGAGAGGATGATCAGCCACACTGGAACTGAG  240
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  241  CTCACCAAGGCGACGATCCCTAGCTGGTCTGAGAGGATGATCAGCCACACTGGAACTGAG  300

Query  241  ACACGGTCCAGACTCCTACGGGAGGCAGCAGTGGGGAATATTGCACAATGGGGGAAACCC  300
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  301  ACACGGTCCAGACTCCTACGGGAGGCAGCAGTGGGGAATATTGCACAATGGGGGAAACCC  360

Query  301  TGATGCMGCCATGCCGCGTGTGTGAAGAAGGCCTTCGGGTTGTAAAGCACTTTCAGCGAG  360
           |||||| ||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  361  TGATGCAGCCATGCCGCGTGTGTGAAGAAGGCCTTCGGGTTGTAAAGCACTTTCAGCGAG  420

Query  361  GAGGAAAGGTTGGTAGCTAATAACTGCCAGCTGTGACGTTACTCGCAGAAGAAGCACCGG  420
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  421  GAGGAAAGGTTGGTAGCTAATAACTGCCAGCTGTGACGTTACTCGCAGAAGAAGCACCGG  480

Query  421  CTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGGTGCAAGCGTTAATCGGAATTACTG  480
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  481  CTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGGTGCAAGCGTTAATCGGAATTACTG  540

Query  481  GGCGTAAAGCGCACGCAGGCGGTTGGATAAGTTAGATGTGAAAGCCCCGGGCTCAACCTG  540
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  541  GGCGTAAAGCGCACGCAGGCGGTTGGATAAGTTAGATGTGAAAGCCCCGGGCTCAACCTG  600

Query  541  GGAATTGCATTTAAAACTGTCCAGCTAGAGTCTTGTAGAGGGGGGTAGAATTCCAGGTGT  600
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  601  GGAATTGCATTTAAAACTGTCCAGCTAGAGTCTTGTAGAGGGGGGTAGAATTCCAGGTGT  660

Query  601  AGCGGTGAAATGCGTAGAGATCTGGAGGAATACCGGTGGCGAAGGCGGCCCCC  653
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  661  AGCGGTGAAATGCGTAGAGATCTGGAGGAATACCGGTGGCGAAGGCGGCCCCC  713
```

Figure 3. Bioinformatic analysis of 16S rRNA sequences on GenBank via the BlastN program.