H3ABioNet

H3ABioNet

Pan African Bioinformatics Network for H3Africa

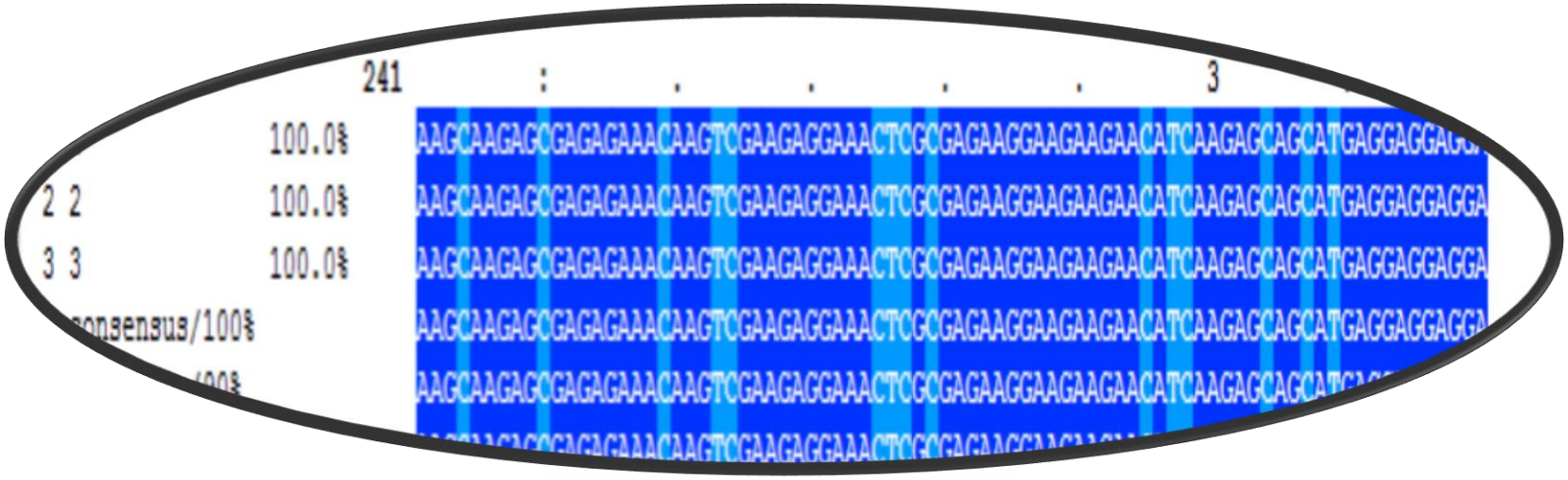
Introduction to Bioinformatics Online Course: IBT

Multiple Sequence Alignment

Building Multiple Sequence Alignment

Lec6:Interpreting Your Multiple Sequence Alignment

Interpreting Your Multiple Sequence Alignment



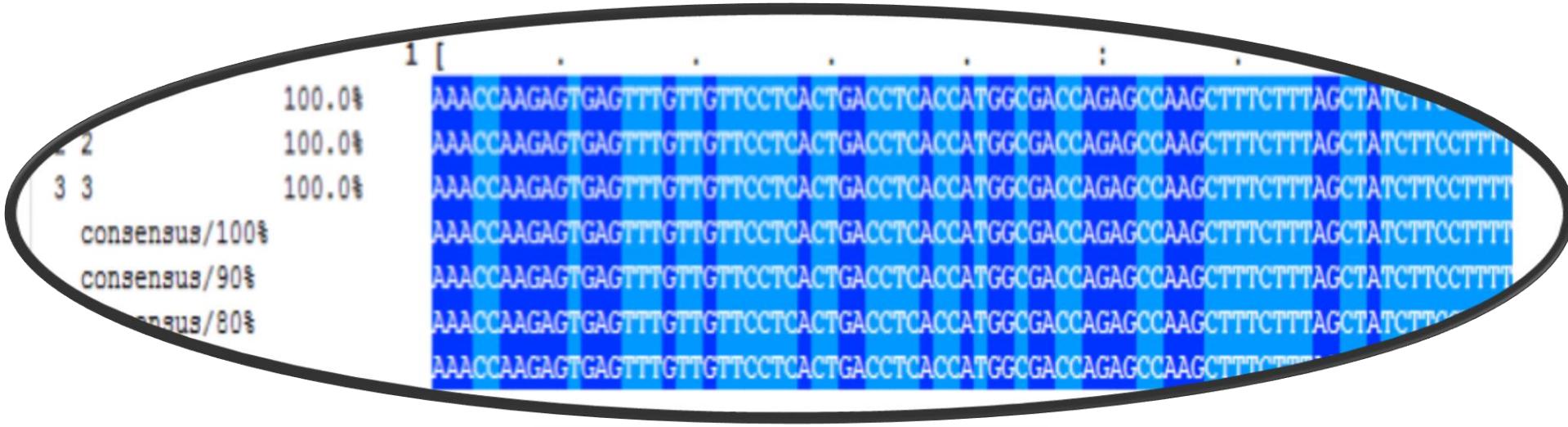
Claverie J, Notredame C (2007). Bioinformatics for Dummies (2nd Edn). *Wiley publishing, Inc*. 436 pp.

Interpret your multiple sequence alignment

The **interpretation** of a multiple alignment depends very much on its appearance. Some tools on the Net can help you **make sense** of your multiple alignments by **extracting blocks** or singling out **special positions**.

Claverie J, Notredame C (2007). Bioinformatics for Dummies (2nd Edn). *Wiley publishing, Inc*. 436 pp.

- **Interpreting** an alignment is a bit of an **art**.
- **E-values** (the scores that tell you how reliable your database search is)



Claverie J, Notredame C (2007). Bioinformatics for Dummies (2nd Edn). *Wiley publishing, Inc*. 436 pp.

That means deciding whether your alignment is **correct** still involves some educated **guesswork**.

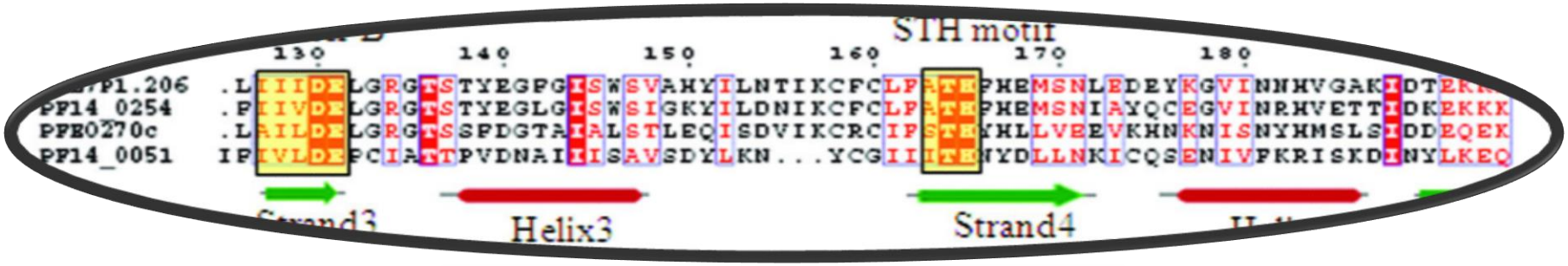
Claverie J, Notredame C (2007). Bioinformatics for Dummies (2nd Edn). *Wiley publishing, Inc*. 436 pp.

- **DNA** alignments are by far the most **difficult to interpret**.

- If you're analyzing this type of sequence, you want a very **high level** of **conservation**, knowing that **single conserved columns** are likely to be **meaningless**.

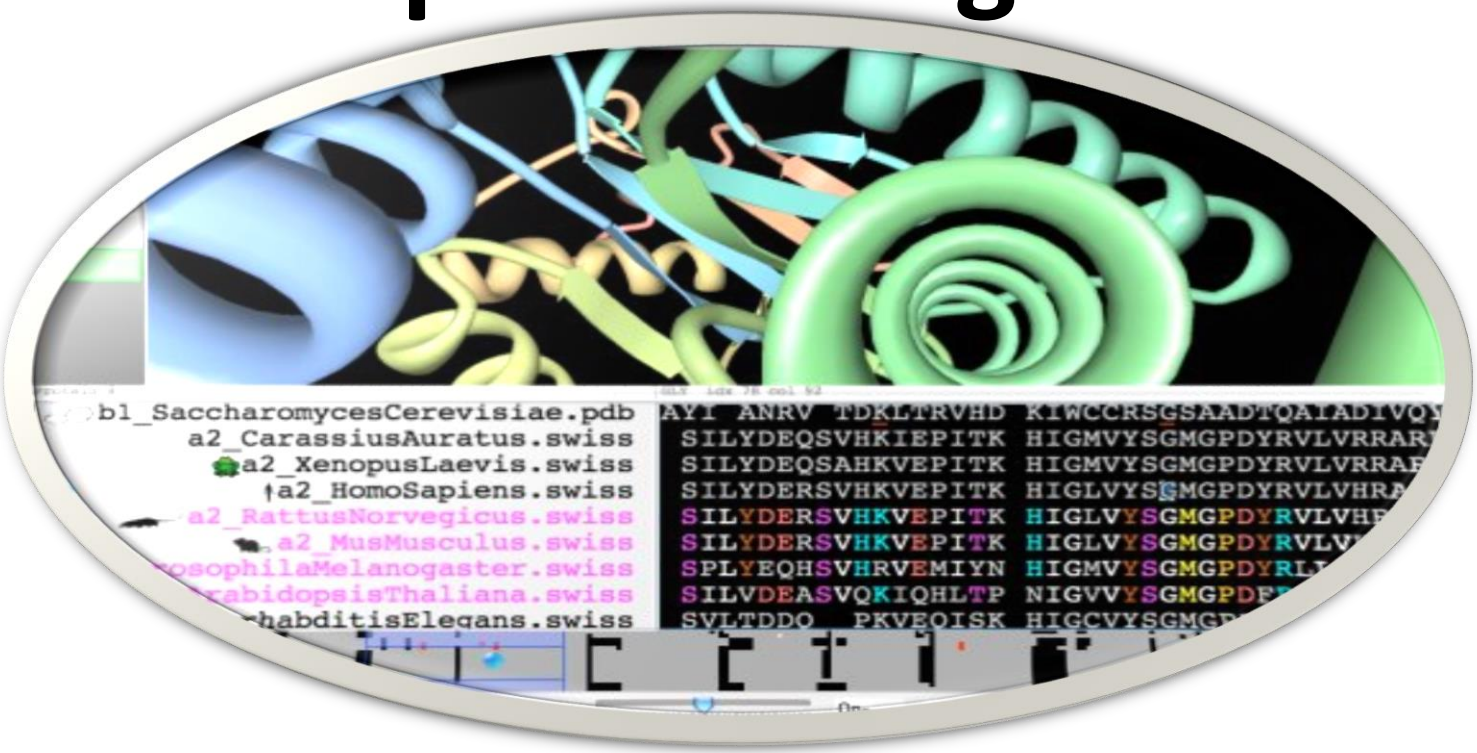
Claverie J, Notredame C (2007). Bioinformatics for Dummies (2nd Edn). *Wiley publishing, Inc*. 436 pp.

- A DNA block is **only informative** when it contains **several** identical columns in a cluster.
- Even with the DNA of closely related sequences, obtaining such an alignment is still difficult.
- This is why most biologists **prefer protein alignments.**

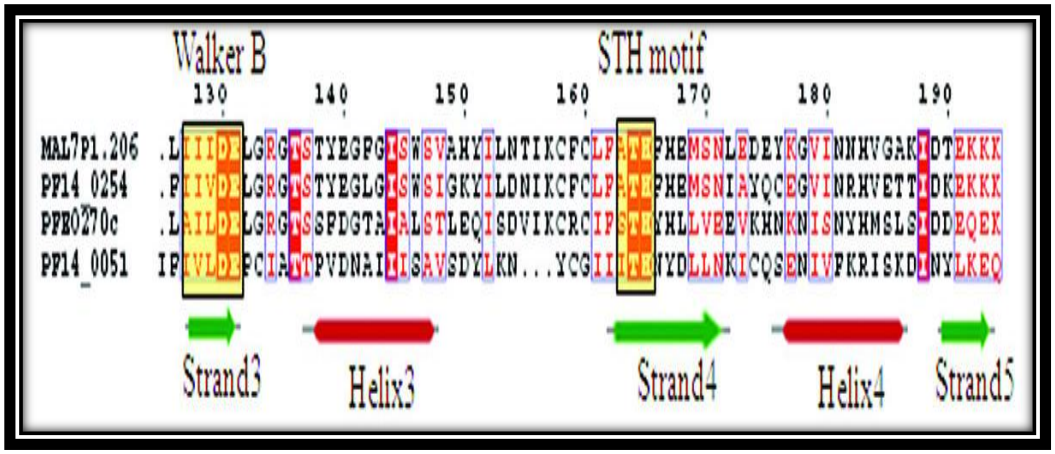


Claverie J, Notredame C (2007). Bioinformatics for Dummies (2nd Edn). *Wiley publishing, Inc*. 436 pp.

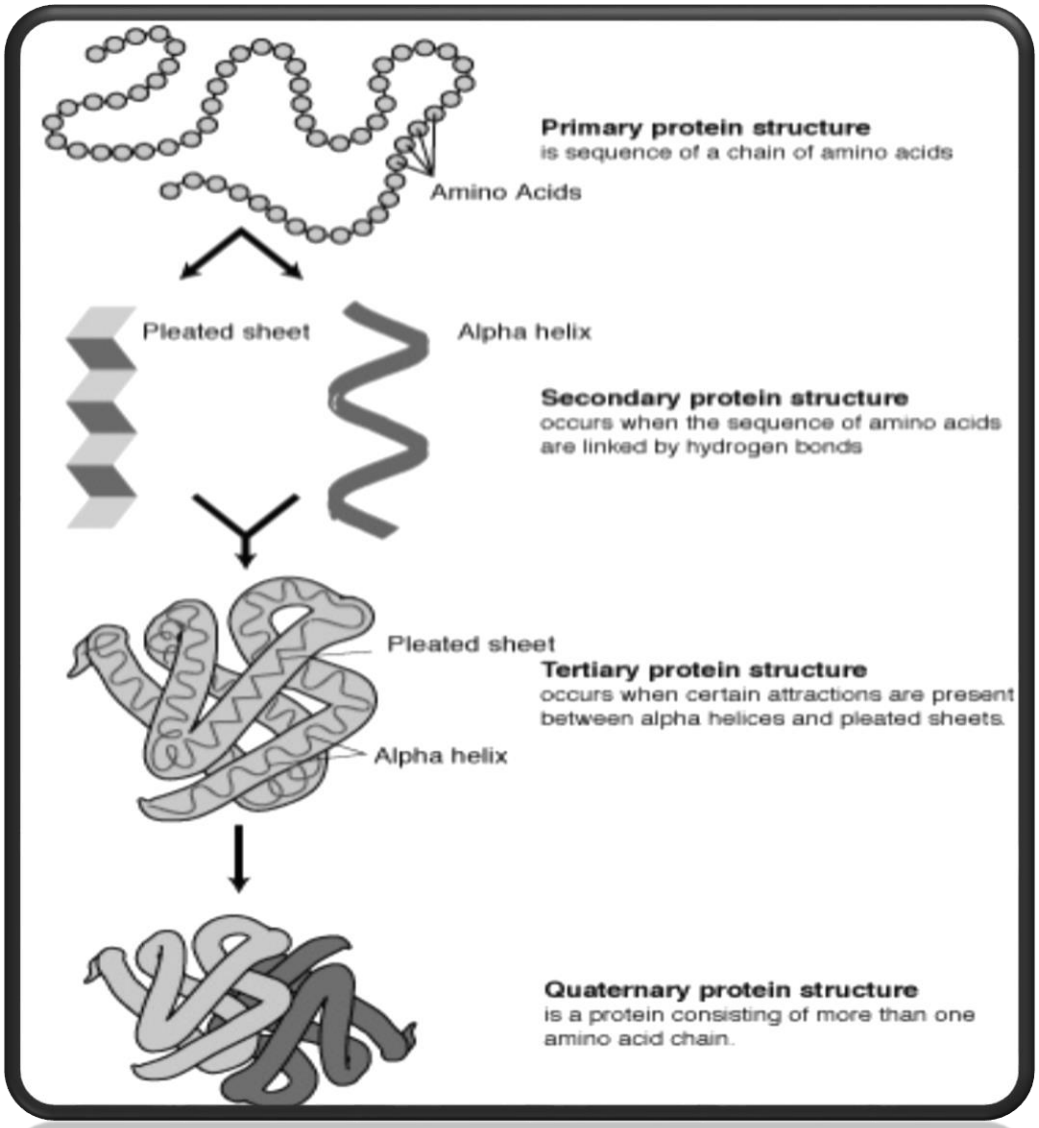
Recognizing the good parts in a protein alignment



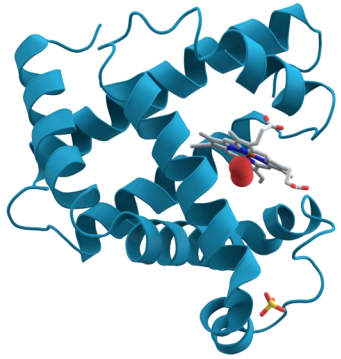
- The most convincing evaluative grid we have for a protein multiple alignment stems from our knowledge of protein structures.

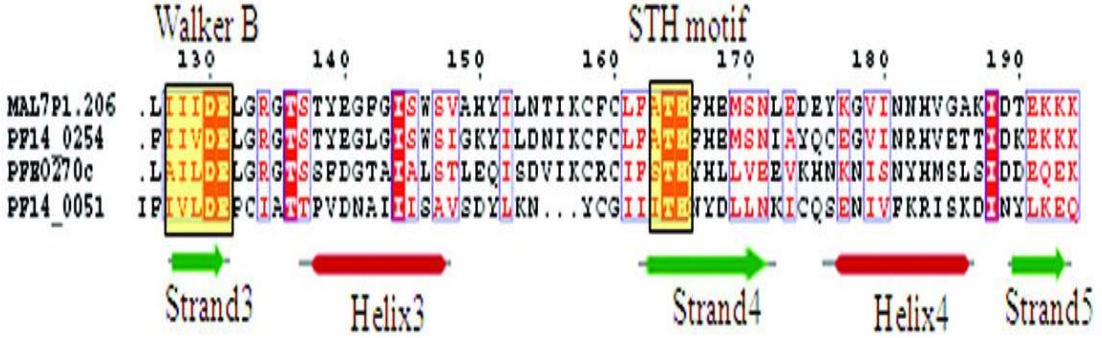


Claverie J, Notredame C (2007). Bioinformatics for Dummies (2nd Edn). *Wiley publishing, Inc*. 436 pp.


Primary protein structure
is sequence of a chain of amino acids
Amino Acids
Pleated sheet
Alpha helix
Secondary protein structure
occurs when the sequence of amino acids are linked by hydrogen bonds
Pleated sheet
Tertiary protein structure
occurs when certain attractions are present between alpha helices and pleated sheets.
Alpha helix
Quaternary protein structure
is a protein consisting of more than one amino acid chain.

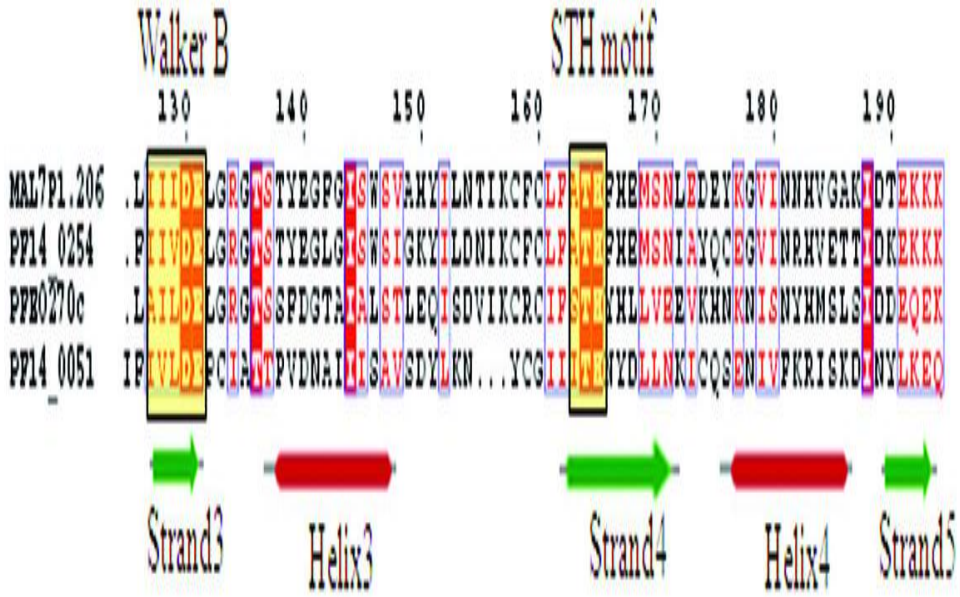
- We know that structures contain surface loops that evolve rapidly. (Loops are softer portions of the protein that connect its more rigid portions).

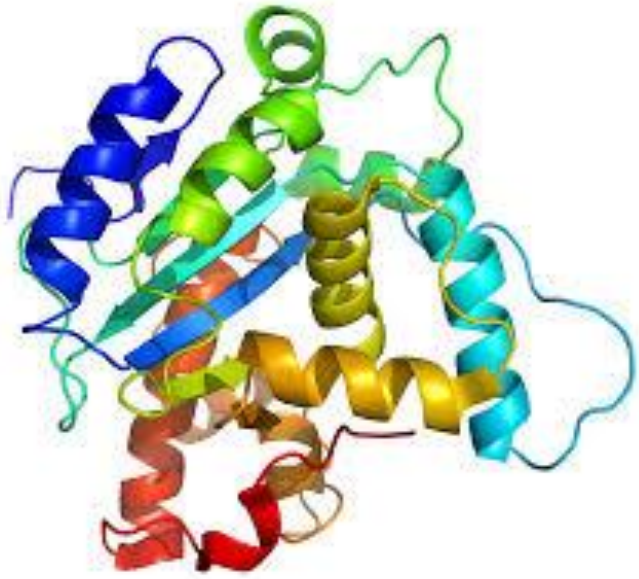




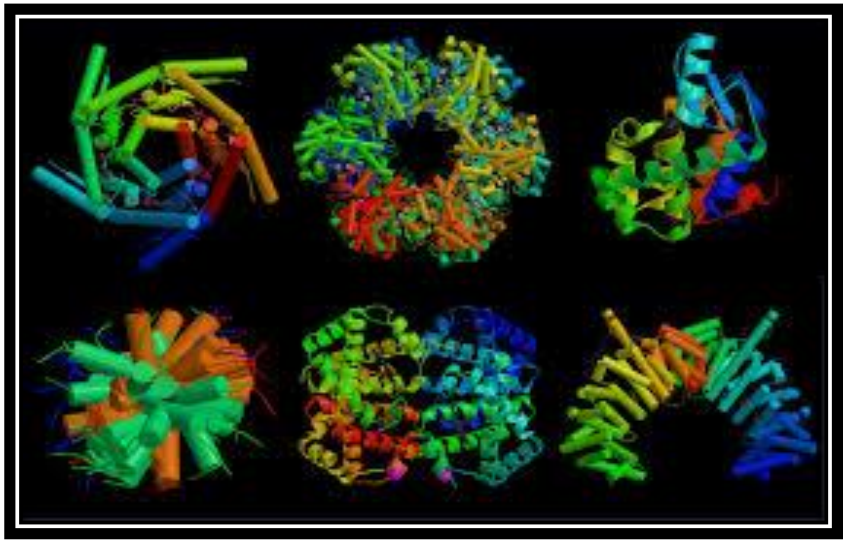
Claverie J, Notredame C (2007). Bioinformatics for Dummies (2nd Edn). *Wiley publishing, Inc*. 436 pp.

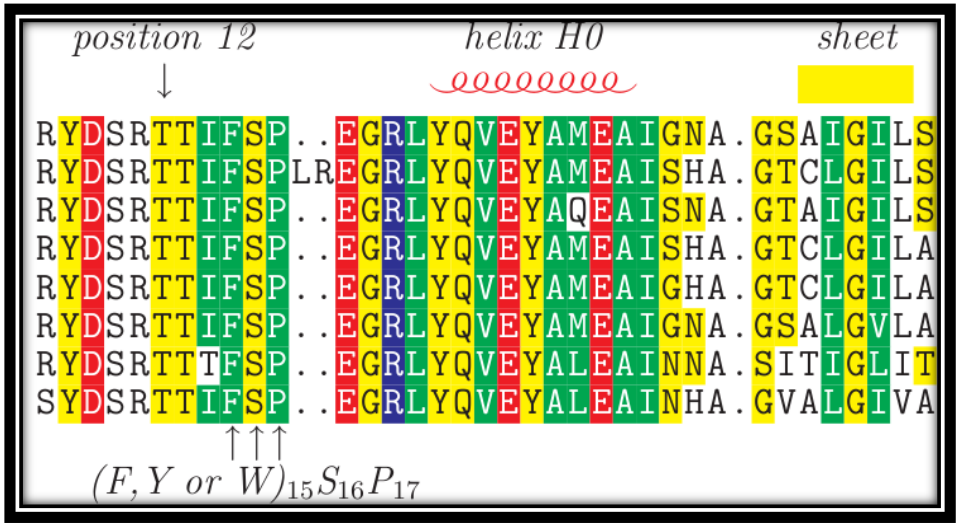
Protein structures also contain **core regions** that act **as support walls** for the protein. These support walls **evolve less rapidly** than the **loops on the surface**.





In your multiple alignment, you can expect to find nice, gap-free blocks that correspond to the core regions — and gap-rich regions that correspond to the loops.





Claverie J, Notredame C (2007). Bioinformatics for Dummies (2nd Edn). *Wiley publishing, Inc*. 436 pp.

Cabalistic signs

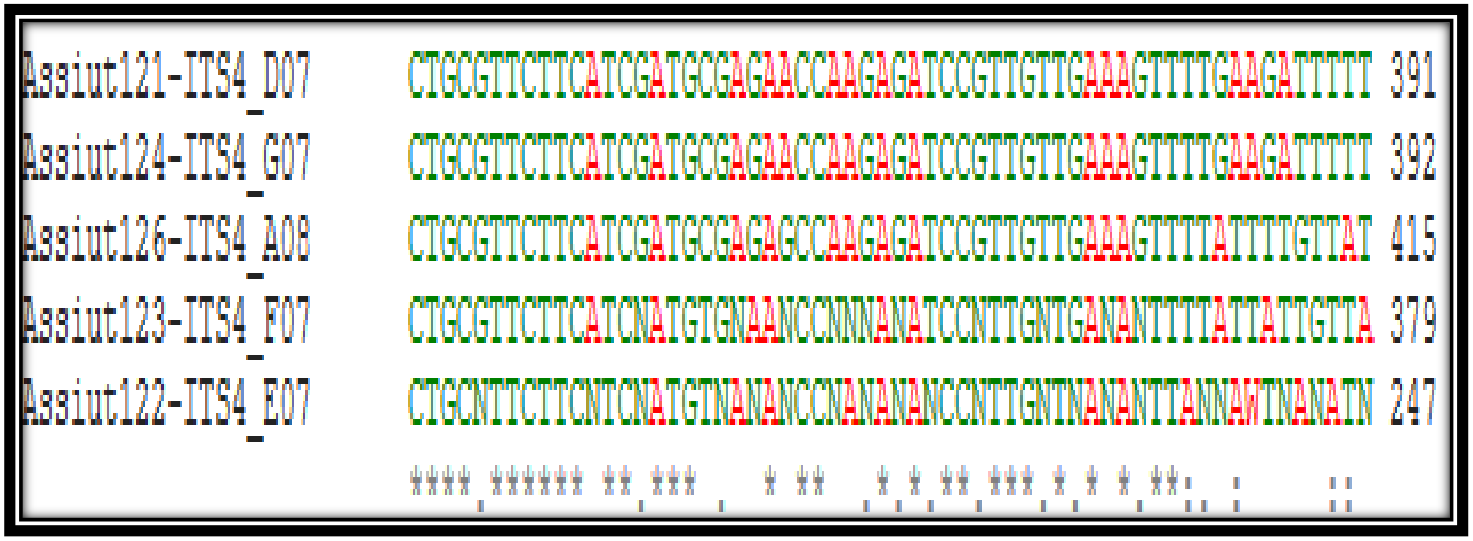
The last line contains seemingly ClustalW, MUSCLE, or Tcoffee alignment, cabalistic signs such as (*), (:), or (.).

(*) A **star** indicates an **entirely conserved** column.
(:) A **colon** indicates columns where all the residues have roughly the **same size** and the same **hydropathy**.
(.) A **period** indicates columns where the **size *OR*** the **hydropathy** has been **preserved** in the course of **evolution**.

Claverie J, Notredame C (2007). Bioinformatics for Dummies (2nd Edn). *Wiley publishing, Inc*. 436 pp.

The **average *good block*** is:

- A unit at least **10–30 amino** acids long, exhibiting **at least one to three stars (*)**, a **few more colons (:)** close to the stars, and **a several periods (.)** scattered along the MSA result.

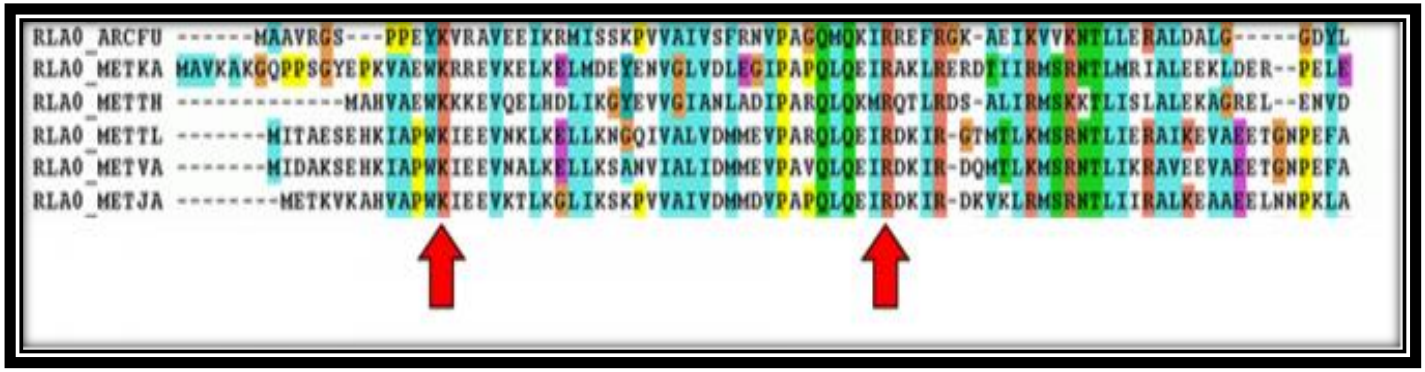


Claverie J, Notredame C (2007). Bioinformatics for Dummies (2nd Edn). *Wiley publishing, Inc*. 436 pp.

- **The magic thing about multiple sequence alignments is that 4 or 5 conserved positions over 50 amino acids can be enough to convince us that we're looking at a genuine signal. This is less than 10 percent identity!**
- **You have to remember that we require at least 25 percent identity to consider a pairwise alignment**

Claverie J, Notredame C (2007). Bioinformatics for Dummies (2nd Edn). *Wiley publishing, Inc*. 436 pp.

Conserved columns in a multiple sequence alignment are meaningful only when the surrounding columns are not conserved



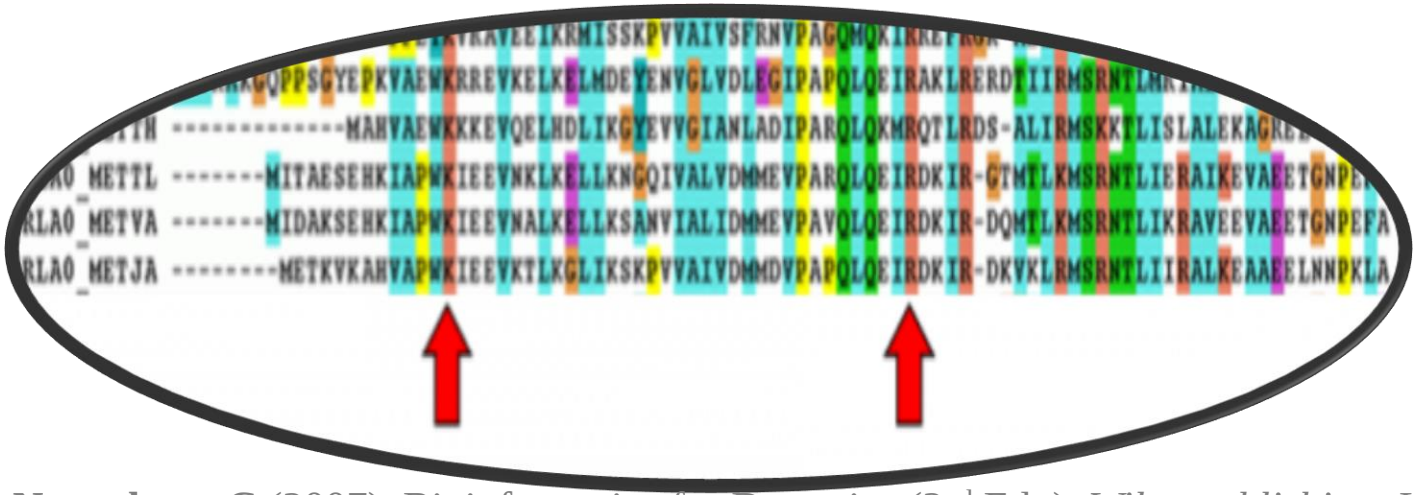
Claverie J, Notredame C (2007). Bioinformatics for Dummies (2nd Edn). *Wiley publishing, Inc*. 436 pp.

Another criterion

for a useful multiple alignment is knowing the type of amino acids you can expect to see conserved.

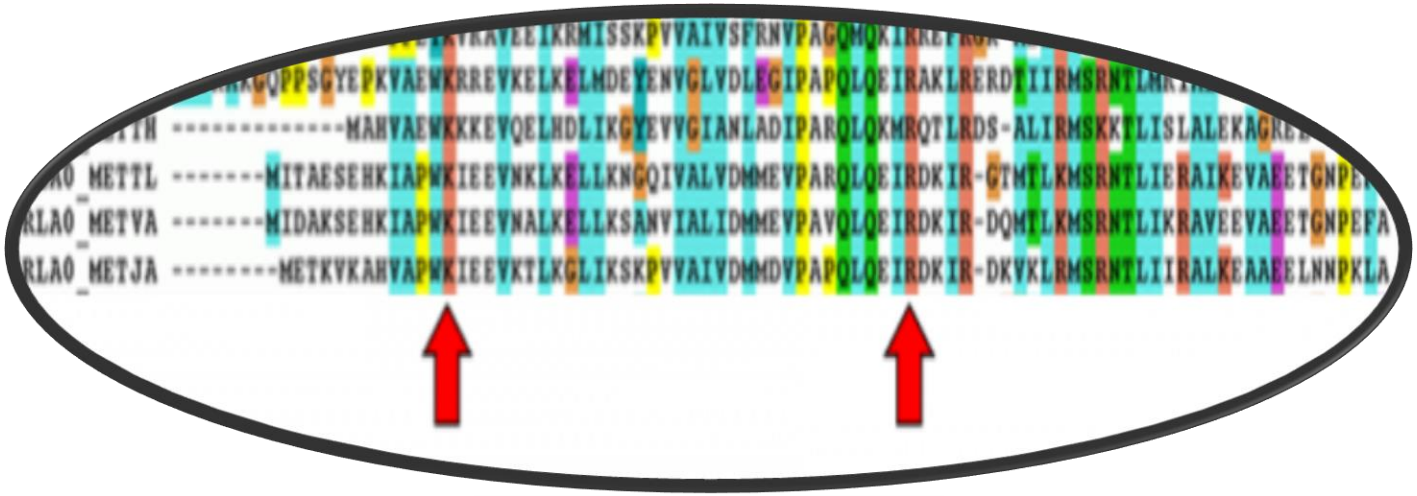
Claverie J, Notredame C (2007). Bioinformatics for Dummies (2nd Edn). *Wiley publishing, Inc*. 436 pp.

Amino acids aren't equal and they all have very characteristic patterns of mutation/conservation in a multiple sequence alignment.



Claverie J, Notredame C (2007). Bioinformatics for Dummies (2nd Edn). *Wiley publishing, Inc*. 436 pp.

Patterns of Conservation in Multiple Sequence Alignments



W(tryptophans),F(phenylalanine), Y(tyrosine)

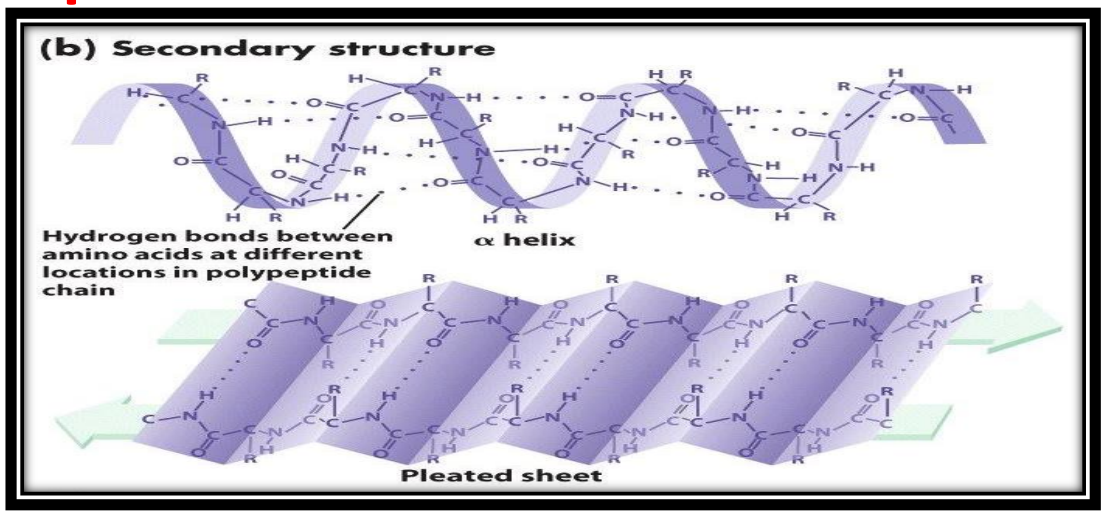
It is common to find conserved ***tryptophans***. **Tryptophan** is a large ***hydrophobic*** residue that ***sits deep in the core of proteins***. It plays an important role in their **stability** and is therefore ***difficult to mutate***. When ***tryptophan mutates***, it is usually replaced by another aromatic amino acid, such as ***phenylalanine or tyrosine***.

Patterns of conserved ***aromatic amino acids*** constitute the most common signatures for recognizing ***protein domains.***

Claverie J, Notredame C (2007). Bioinformatics for Dummies (2nd Edn). *Wiley publishing, Inc*. 436 pp.

G (glycine), P (proline)

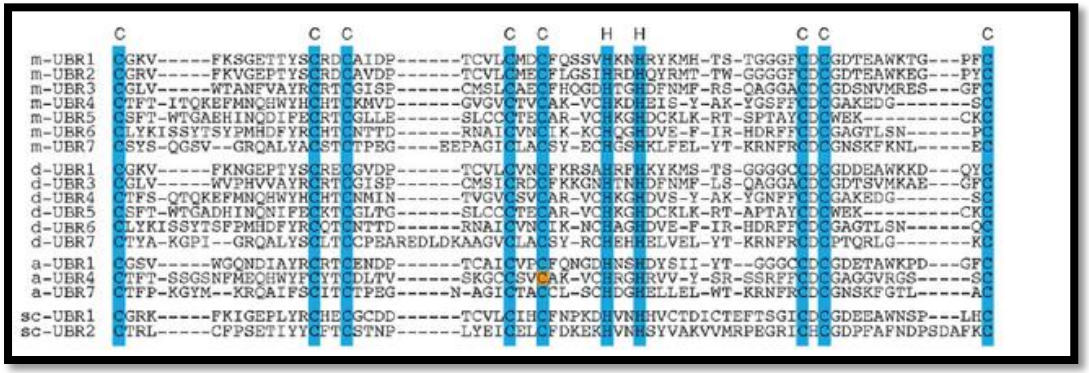
It is common to find conserved columns with a **glycine** or a **proline** in a multiple alignment. These two amino acids often coincide with the **extremities** of well-structured **beta strands or alpha helices**.



Claverie J, Notredame C (2007). Bioinformatics for Dummies (2nd Edn). *Wiley publishing, Inc*. 436 pp.

C (cysteines)

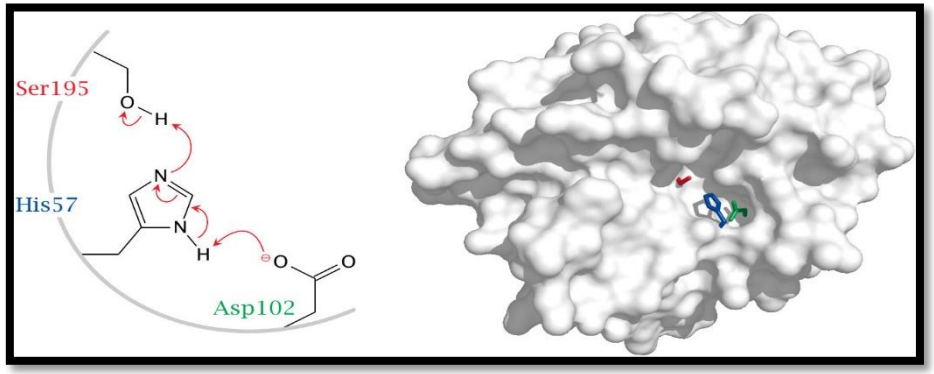
Cysteines are famous for making **C-C (disulphide) bridges**. Conserved columns of cysteines are rather common and usually indicate such bridges. Columns of conserved cysteines with a specific distance provide a useful signature for recognizing protein **domains and folds**.



Claverie J, Notredame C (2007). Bioinformatics for Dummies (2nd Edn). *Wiley publishing, Inc*. 436 pp.

H(Histidine), S(serine)

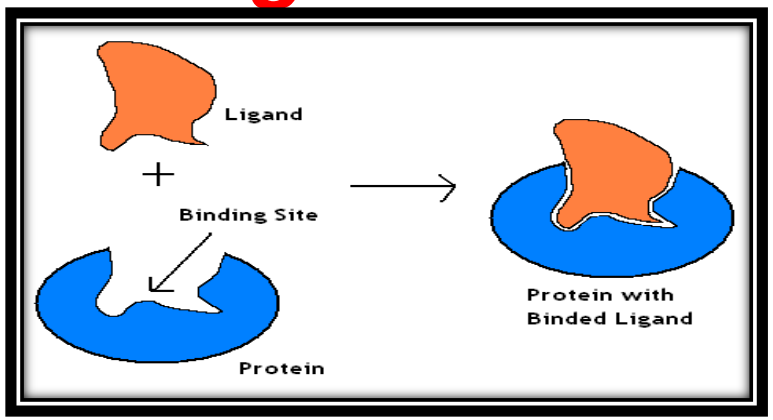
Histidine and serine are often involved in **catalytic sites**, especially those of **proteases**. Conserved histidine or a conserved serine are good candidates for being part of an **active site**.

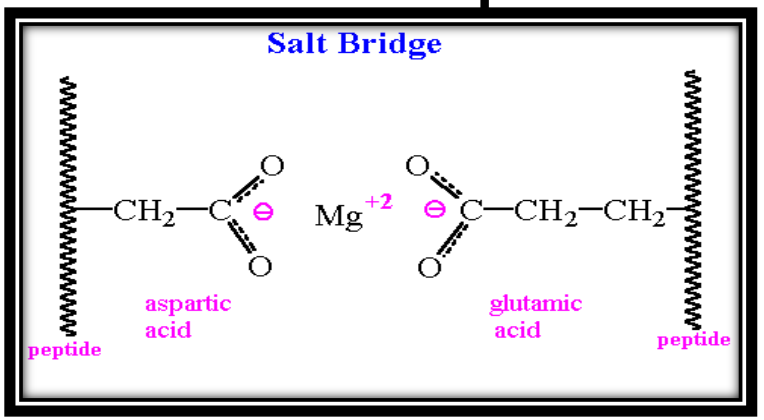


Claverie J, Notredame C (2007). Bioinformatics for Dummies (2nd Edn). *Wiley publishing, Inc*. 436 pp.

K (Lysine), R (Arginine), D (Aspartic Acid), E (Glutamic Acid)

These **charged amino** acids are often involved in **ligand binding**. Highly conserved columns can also indicate a ***salt bridge*** inside the core of the protein.

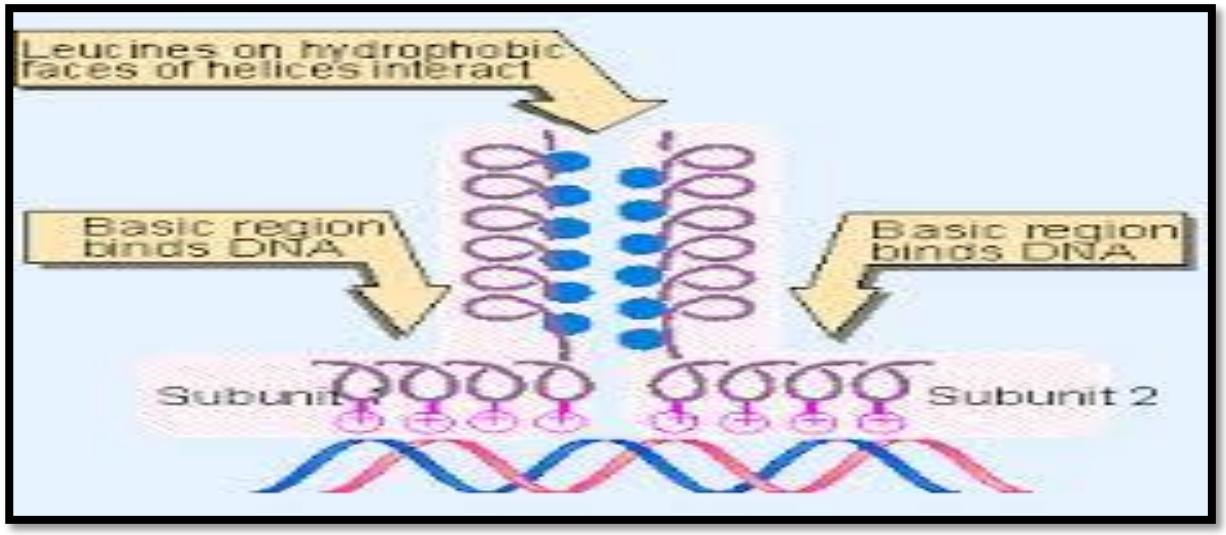




Claverie J, Notredame C (2007). Bioinformatics for Dummies (2nd Edn). *Wiley publishing, Inc*. 436 pp.

L (Leucines)

Leucines are rarely very conserved unless they're involved in ***protein-protein interactions*** such as a ***leucine zipper***.



Claverie J, Notredame C (2007). Bioinformatics for Dummies (2nd Edn). *Wiley publishing, Inc*. 436 pp.


BTU
BIOINFORMATICS
TRAINNING UNIT
GENOMICS
1st Module What Bioinformatics Can Do for You
2nd Module Manipulation of Biological Sequences
3rd Module Working with single DNA sequence
4th Module How to Build a Multiple Sequence Alignment?
5th Module Inferring Phylogenetics analysis using Jalview
6th Module Advanced Molecular Concepts
7th Module Inferring Protein Sequence (structure & Function)
8th Module RNAanalysis and Function
9th Module Editing and Publishing Alignments in your Manuscript
10th Module Building and Publishing Phylogenetic Trees
11th Module Working with Protein 3-D Structures
12th Module Advanced Bioinformatics Using R
Bioinformatics Workshops Series
EST by, Dr. Ahmed Mansour Alzohairy
Call us: 01026060324 / 01000727270
Facebook: https://www.facebook.com/Bioinformaticsunit

By

Ahmed Mansour Alzohairy

Department of Genetics, Zagazig University, Zagazig, Egypt