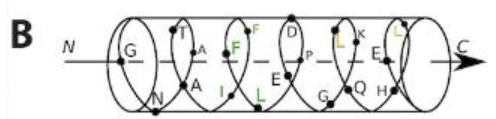
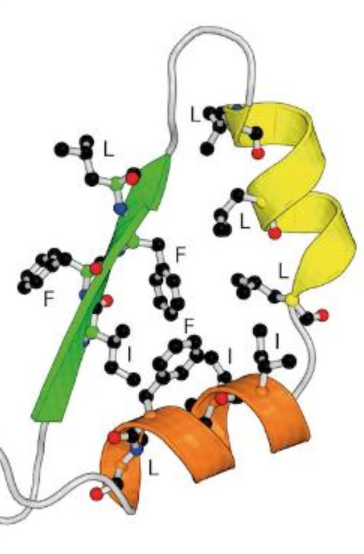
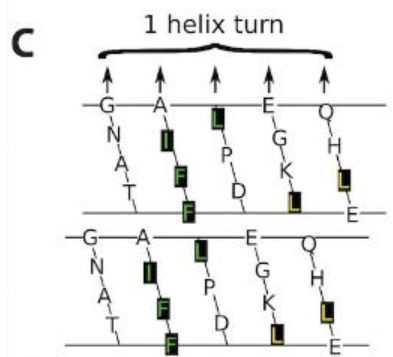
**GENOM Project:   
Evolution of structural signatures in families of soluble protein domains**

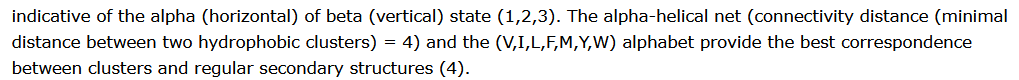
# Project description

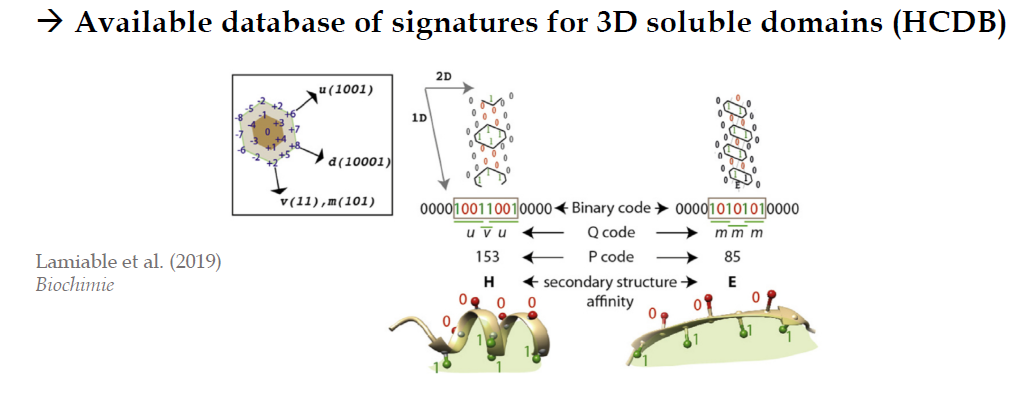
The aim of the projet is to build a **HC substitution matrix** from sequence alignments. Now, what exactly is a HC substitution matrix?  
 A HC (Hydrophobic Cluster) is defined from pure amino acid sequences and match regular secondary structures. As explained [here](http://impmc.sorbonne-universite.fr/fr/equipes/biophysique_et_bioinformatique/liste_des_membres/permanent-e-s/isabelle-callebaut-responsable/hydrophobic-cluster-analysis-hca-1.html), HCA (Hydrophobic Cluster Analysis) was developed in the late 80s and gives information about **protein regular secondary structure** from only the information of a **single amino acid sequence** (thus, bypassing the need for homologous sequences).

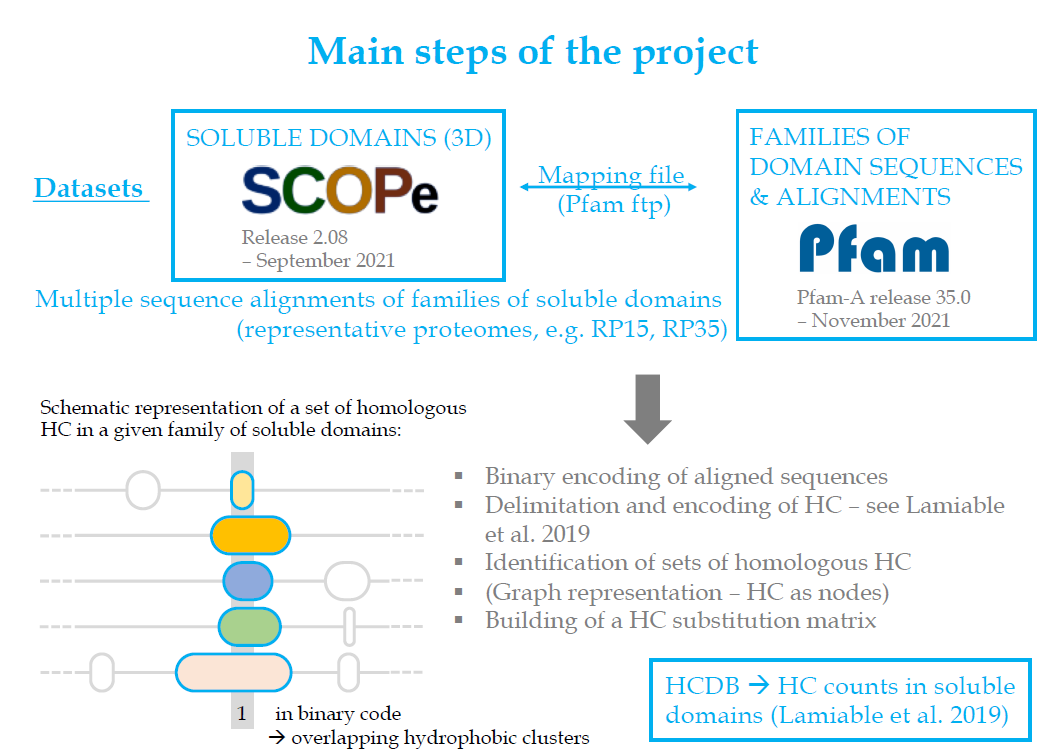
This is necessary because a lot of sequences in existing protein databases like Swissprot exhibit biological unknowns: **dark regions** and gray regions that far outmatch known regions. Thus, these are issues that prevent genome annotation based on sequence similarity, because we simply might not have homologous sequences to compare to; thereby preventing the deciphering of the functional potential of genomes.

Enter: HC. This method of representation is a bit complex, so I’m explaining here for Alexis and Liam. Basically, we start from a 1D sequence of amino acids:  
  
(which will be subsequently transformed into another form with special symbols denoting amino acids with particular structural behaviours such as P = star, G = diamond, T = square, S = dotted square, like so:)  
  
And then further into binary, where the **coloured letters are considered “joined contiguous amino acids”**:  


There is of course a biological equivalence to be found here. This is once again well explained on the website linked above, but here’s a summary: from the 1D sequence shown above, a 2D plot is created by writing the sequence on an alpha helix and cutting it along the horizontal axis like so:  
  
 Then, a **repeated** set of this helix creates planes when contiguous strong hydrophobic amino acids (**V, I, L, F, M, Y, W**) are linked to each other, in order to form **clusters**:  
  
 Finally, consider that **horizontal clusters (in yellow/orange)** are associated to **alpha helixes** and **vertical clusters (in green)** are associated to **beta sheets**.  
 All in all, this is useful for the detection of **remote** sequence homology, foldable regions as well as fold cores.

****From <http://impmc.sorbonne-universite.fr/fr/equipes/biophysique_et_bioinformatique/liste_des_membres/permanent-e-s/isabelle-callebaut-responsable/hydrophobic-cluster-analysis-hca-1.html>

There is a database known as HCDB for 3D soluble domains which gives such information in very simple values: **binary code, Q code, P code, and secondary structure affinity** as shown below:  


 Thus, the project is to build a **HC substitution matrix** (like the weight matrices we worked on last year), to determine factors such as:

* HC groups?
* Relations with secondary structure types?
* Impact of sequence conservation?

Essentially, we need to play around with creating relevant matrices where more significant links between amino acids need to be scored higher.

The recommended steps as according to the teacher are:

1. Binary encoding of aligned sequences
2. Delimitation and encoding of HC (cf. Lamiable et al. 2019 paper on **HCDB, important!**)
3. Identification of sets of homologous HC
4. (Graph representation of HC as nodes)
5. Building of the HC substitution matrix: basically we would have two HCs to compare to each other.

Another useful link: <http://www-ext.impmc.upmc.fr/~callebau/HCA.html>

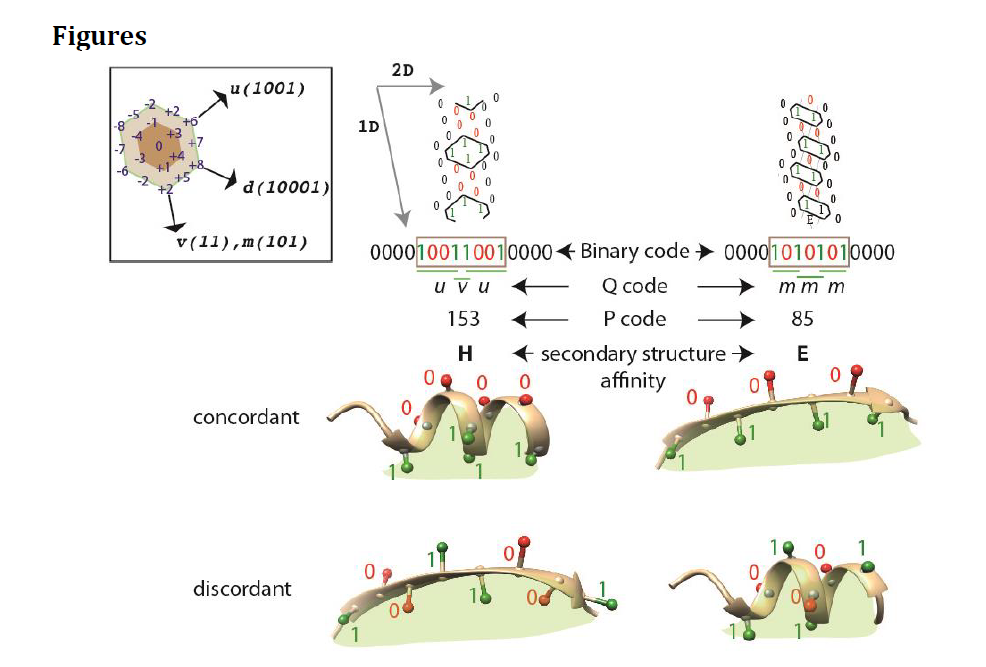
# Lamiable 2019 paper – notes; Encoding of Binary sequences

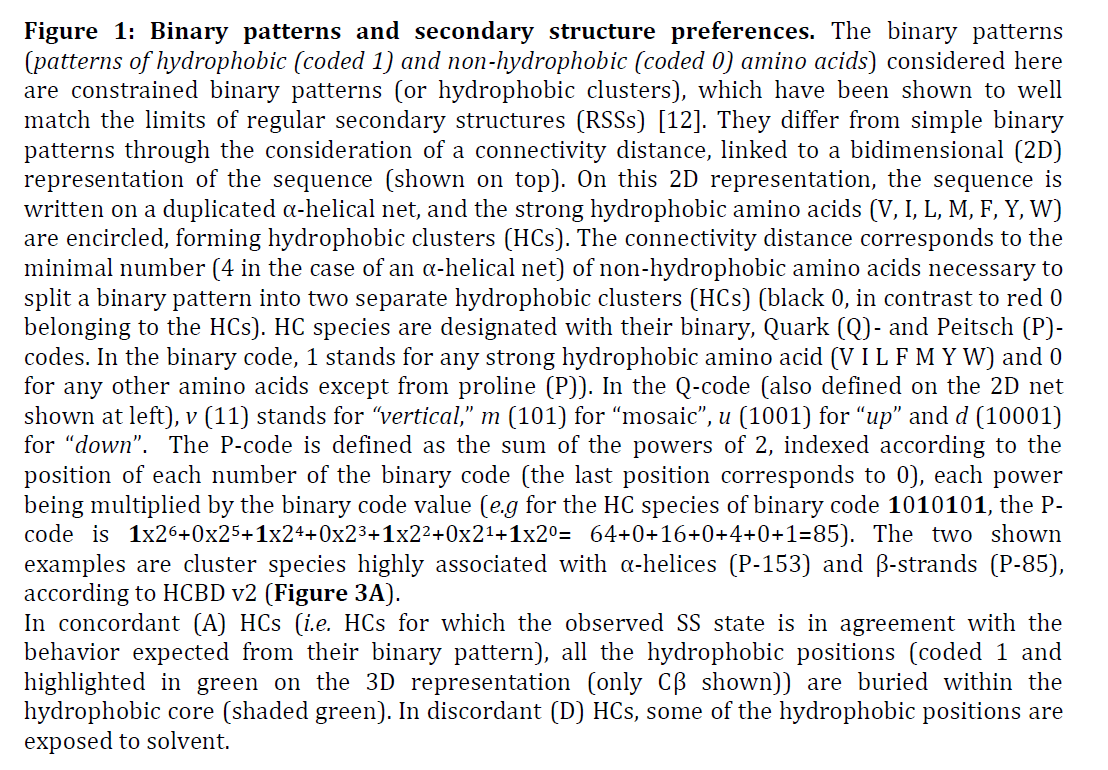
## Abstract

HC (as defined by HCA) are conditioned binary patterns made of both hydrophobic and non-hydrophobic position; useful for predicting secondary structures in proteins from only a single aa sequence information.  
 RSS 🡺 Regular Secondary Structure.  
 They almost **double** the number of hydrophobic cluster species, with each species being defined by a unique binary pattern, by considering available experimental 3D structures of **protein globular domains** representing the most frequent structural bricks.  
 Then, they “used this updated HCDB to show that the hydrophobic amino acids of discordant clusters, i.e. those less abundant clusters for which the observed secondary structure is in disagreement with the binary pattern preference of the species to which they belong, are more exposed to solvent and are more involved in protein interfaces than the hydrophobic amino acids of concordant clusters.” 🡸 Don’t really get this for now.

“HC clusters are defined as previously, as a succession of strong hydrophobic amino acids (V I L F M Y W) separated from each other by breakers. Breakers are composed of **at least four consecutive non-hydrophobic amino acids (connectivity distance) or a proline** (Figure 1).” (cf. page 4 of paper).

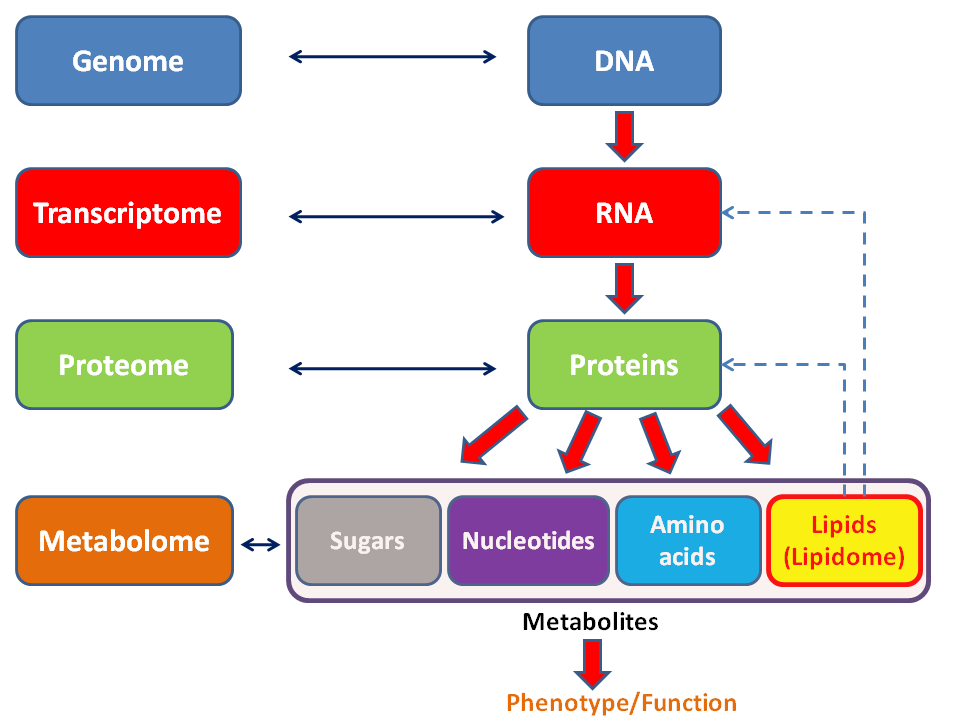
Updated with instructions from Elodie: we must encode the hydrophobic amino acid group (V, I, L, M, F, Y, W) in **binary**. The rules for defining a hydrophobic cluster and P-code encoding are detailed in the following figure from the paper (figure 1):





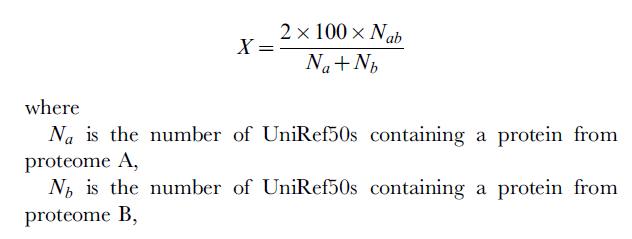
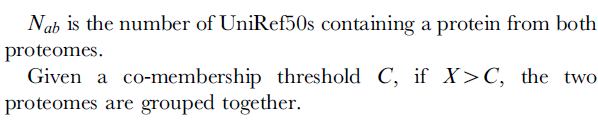
# RP15 proteome sequences

According to Elodie, we need to work on the RP15 proteome sequences first from the large list of files available at the link [here](https://ftp.ebi.ac.uk/pub/databases/Pfam/current_release/). The distinction between the RP15, RP35, RP55 etc. can be explained by the paper titled “Representative Proteomes: A Stable, Scalable and Unbiased Proteome Set for Sequence Analysis and Functional Annotation”.



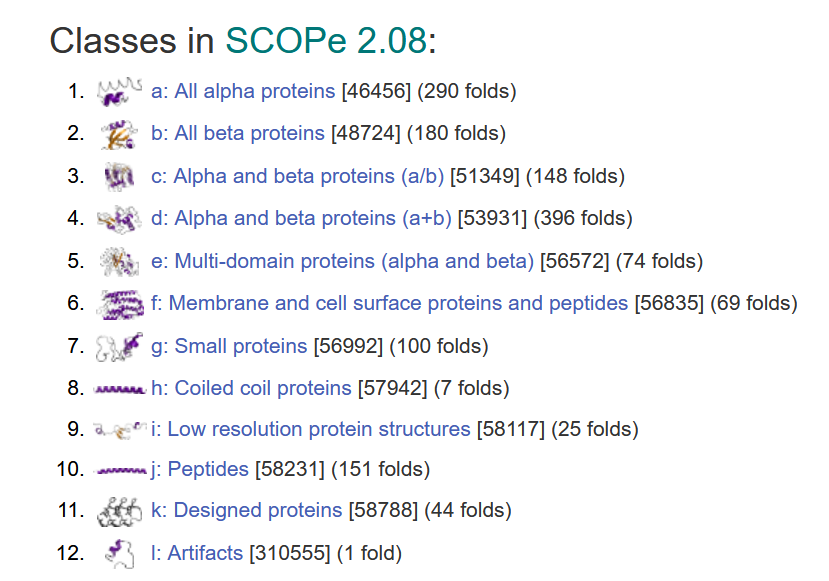
RP15 essentially stands for 15% co-membership threshold (CMT) Representative Proteome. As a reminder, a **proteome is the entire set of proteins that is expressed by a genome.** (This is important because multicellular organisms may have very different proteomes in different cells, despite being part of the same genome = organism). Essentially RP15, RP35 etc. were proteomes that were selected from a Representative Proteome Group (RGP) containing similar proteomes calculated based on co-membership in UniRef50 clusters. This concept is also explained at the end of the link here: <https://ftp.ebi.ac.uk/pub/databases/Pfam/current_release/userman.txt>   
 The goal was always to achieve **reduced sequence space** (reducing redundancy in protein sequence space).

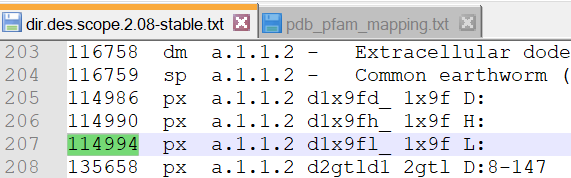
CMT between two proteomes A and B is measured as such:

# SCOPe classification and Pfam extraction

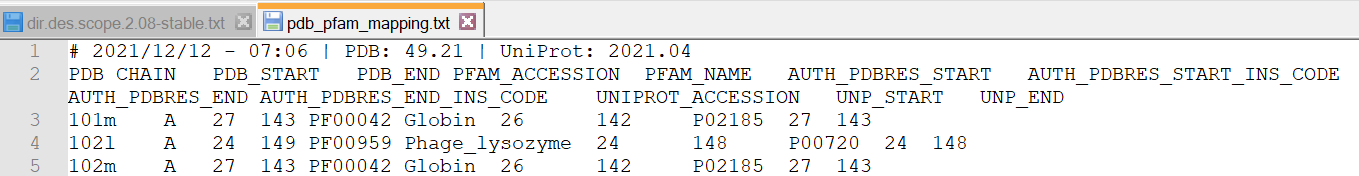
We must only work on **soluble protein domains**, which, according to Elodie, are a, b, c and d SCOPe classifications: (found on <https://scop.berkeley.edu/>)



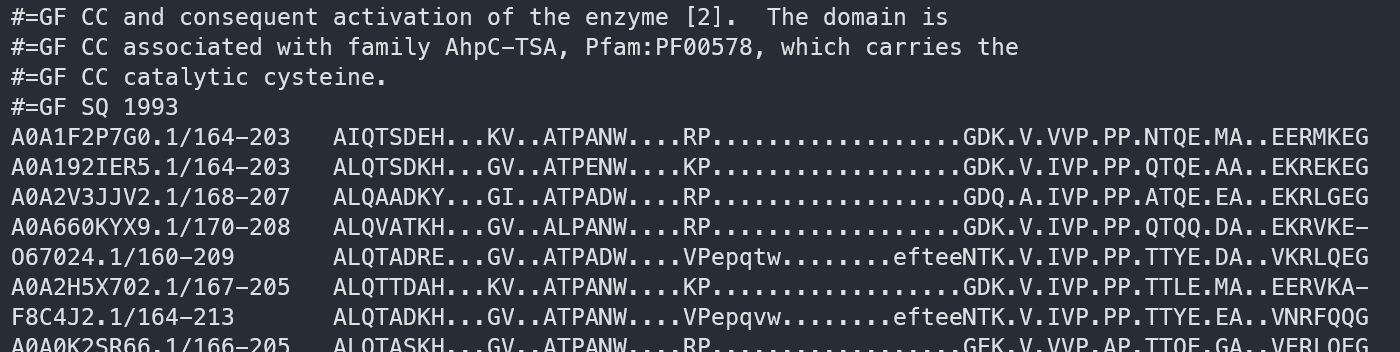
An example of how to read a line in a SCOPe classification from the first text file downloaded:   


In this example, for 114994, we know that the polypeptidic chain L of the PDB file whose access number is 1x9f has been classed with the label: a.1.1.2 in the SCOPe database. It is of class A, and is as such **soluble**.  
 On the line right below, for 135658, the polypeptidic chain D of the PDB file whose access number is 2gtl, has its regions from the 8th to the 147th amino acid classed with the label: a.1.1.2 in the SCOPe database. It is of class A, and is also **soluble.**

Furthermore, from the second downloaded text file (pdb\_pfam\_mapping.txt), we will be able to find the **Pfam families indicated in sequences whose 3D structure is known.** The PDB code is in the first column. From this document, we must extract a list of **Pfam identifiers corresponding to soluble domains**, then work on sequence alignment for those domains only on the RP15 file.



# Reading the RP15 file



The RP15 file starts like so (and is 7 gigabytes large, so no way to visualise it in a text editor). As mentioned before, the RP15 file is the Representative Proteome file that regroups amino acid sequences with at least 15% co-membership based on UniRef50 clusters. (The RP15 is the most redundant set, logically)

According to the link here: <https://compbio.soe.ucsc.edu/a2m-desc.html> the “-” (dashes) and uppercase characters represent alignment columns. The “.” (periods) and lowercase characters represent gaps (insertion positions).