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

## Introduction

The aim of this project was to work on **Hydrophobic Clusters (HCs)** present in soluble domains of proteins, with the goal of building a suitable **substitution matrix** from information that is derived from HC analysis.  
 **Hydrophobic Cluster Analysis (HCA)** is a technique developed in the late 80s and is able to provide us with protein Regular Secondary Structure (RSS) information from a single amino acid sequence, thus bypassing the need for homologous sequences. This is important because a large number of proteins present in databases such as **Swissprot** exhibit dark or gray regions – biological unknowns which can impede genome annotation and deciphering based on reliable homology-based methods.  
 The creation of a substitution matrix for different HCs (there being hundreds upon hundreds of them, depending on how they are defined) is thus a novel approach that would permit a BLOSUM62 type matrix for HCs rather than simply amino acids, and holds much promise.  
 In this short report, we will seek to outline our methods and approach to this problem under the guidance of Elodie Duprat, our supervisor.

## Methods

### Data acquisition and pre-processing

The data for the project was acquired from three separate links and required a pre-processing function in order to isolate relevant sequences.  
 The three files that were respectively required for this part were obtained at: <https://scop.berkeley.edu/downloads/parse/dir.des.scope.2.08-stable.txt> for SCOPe classification, <https://ftp.ebi.ac.uk/pub/databases/Pfam/mappings/pdb_pfam_mapping.txt> for Pfam mapping, as well as <https://ftp.ebi.ac.uk/pub/databases/Pfam/current_release/Pfam-A.rp??.gz> for Representative Proteomes.

First, we needed to make sure that we were only working on **soluble domains** within the SCOPe classification. Add justification and/or explanation of SCOpe? In order to do so, we only considered domains that belonged to classes a (all alpha proteins), b (all beta proteins), c (alpha/beta), and d (alpha+beta).   
 To do so, we created a function that filtered all the domains on the first file based on the identifier on the third column, keeping only those whose columns started with the letters a, b, c or d.

Then, the second tabulated file contained relevant information and notably **Pfam accession number** for families whose three-dimensional sequences were known. From our filtered output (for soluble domains) of the first file, we extracted a second filtered list of Pfam accession numbers based on the SCOPe access numbers (1st column) of the soluble domains. That way, we had the Pfam accession numbers of soluble domains that also had known three dimensional sequences. Alexis, read over what I said to make sure I didn’t get any of this wrong.

The third file that was worked on was *Pfam-A.rp15*. This file is a **Representative Proteome** file containing a large number of sequence alignments per Pfam accession file that fit the threshold value of 15% for CMT (Co-Membership Threshold), which is a metric for (clarify CMT here?).   
 Thus, we used the extracted Pfam accession numbers from the second step in order to only keep the alignments in RP15 that both fit the criteria of: 1) Soluble domain, and 2) Known 3D structure. Check if this is correct.

### HC – rules of definition, binarization

HCs are strictly defined as a grouping of several strong amino acids, that is to say **V, I, L, F, M, Y or W,** that start with a strong aa and end with a strong aa.In the context of HCA and matrix building, any amino acid that wasn’t part of the former group became a **binary 0**, while any that was became a **binary 1.**   
 A HC also wasn’t necessarily only strong amino acids: amino acids represented in binary as 0 may also be present within a HC, with the condition that they do not exceed 4 in length if following each other, as defined by Lamiable et al.[1] For example: 11, 1101, 10001, 101010101 are all considered acceptable HCs, while 10, 01, 100000001 do not fit the definition of an acceptable HC, so were not identified as such.

Thus, we ran an HC identification code on the filtered RP15 file in order to extract a list of unique HCs by both their raw sequence as well as their alternative representations in P-code and Q-code insert reference here (though the latter is not currently used). We identified a large number of unique HCs, some even going as long as having a P-code of 435641765927107293001, but decided to filter the relevant HCs to the **500 most common ones**, mirroring closely the reference csv file we were provided, which also included a list of 400+ most common HCs built from the HC Database[2].

### Matrix creation

With a filtered list of the most common HCs in hand, we then proceeded to creating the substitution matrix. In order to do so, we had to consider every single Pfam alignment that remained after having applied all three of our pre-processing steps. Thus, every alignment was able to provide a matrix of counts for each pairing of HCs in our 500 most common list, after which a **“supermatrix”** of counts could be built to create our substitution matrix as the BLOSUM62 matrix insert reference here.

To go into a bit more detail, we first needed a function to binarize our alignments slightly differently than before. Previously, we merely converted strong amino acids to 1 and kept all the other amino acids at 0, while checking for HCs using the method described in the previous section.   
 The second binarization required was this time to binarize in terms of belonging to an HC: essentially, for a PFAM alignment (of for example 1000 sequences), we needed to have **any amino acid belonging to an HC become a binary 1** and any that did not become a binary 0.

For illustrative purposes, see the drawing below:  
 Insert schema if there is space

In order to build our substitution matrix, then, we needed to consider (with a pseudo count) the sum of every column (so, every position in a Pfam alignment). Then, if said sum was greater than a minimal acceptance threshold (e.g., 75%), we would consider this particular position as strongly conserved in HCs.  
 If a particular position is strongly conserved in HCs, we then built a matrix according to the overlap of HCs in those positions. For example, if at position 25, there was an overlap of HC of P codes 3, 11, and 13, then we would increment 1 on the rows and columns corresponding to those HCs.  
 This was done for every Pfam alignment that was present in the now filtered RP15 file, which led to the creation of the previously mentioned “supermatrix”, which is simply the sum of the 500 by 500 matrices for every Pfam alignment.

Finally, from this “supermatrix”, we applied a log odds ratio as for BLOSUM62 type calculations, following the formula:  
 Where corresponds to the frequency of having HC1 and HC2 overlap at least over one amino acid over a strongly conserved HC region, and to the product of their individual frequencies.

## Results

Add in when we have final matrix.

## References

[1] A. Lamiable *et al.*, ‘A topology-based investigation of protein interaction sites using Hydrophobic Cluster Analysis’, *Biochimie*, vol. 167, pp. 68–80, Dec. 2019, doi: 10.1016/j.biochi.2019.09.009.

[2] P. J. Silva, ‘Assessing the reliability of sequence similarities detected through hydrophobic cluster analysis’, *Proteins*, vol. 70, no. 4, pp. 1588–1594, Oct. 2007, doi: 10.1002/prot.21803.