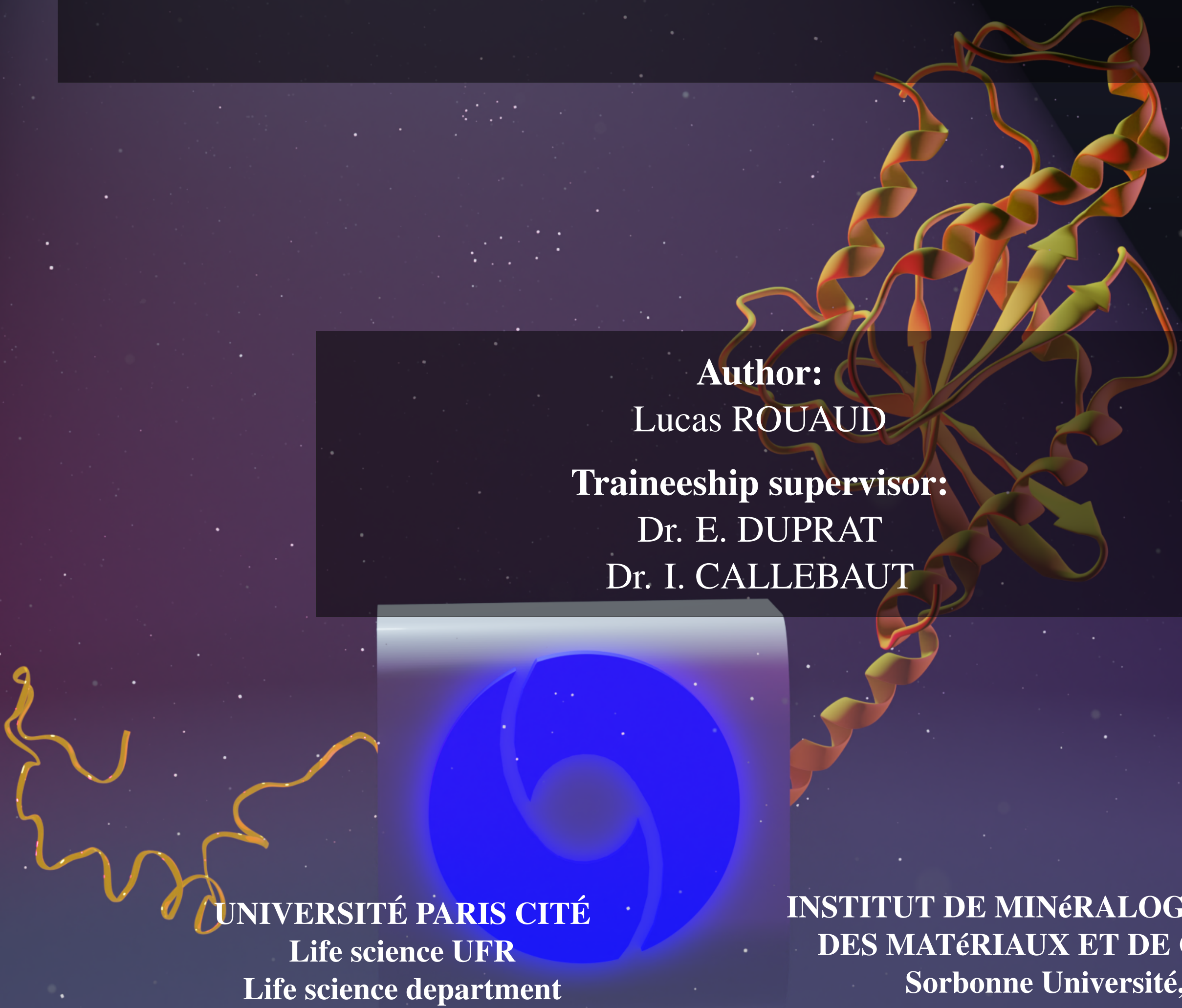
UNIVERSITé PARIS CITé

2^D YEAR IN MASTER
DEGREE IN
BIOINFORMATIC

SCHOOL YEAR 2022-2023

INTERNSHIP REPORT

Protein structural signatures as a new language for remote homology detection in sequence databases



Author:
Lucas ROUAUD

Traineeship supervisor:
Dr. E. DUPRAT
Dr. I. CALLEBAUT

UNIVERSITÉ PARIS CITÉ
Life science UFR
Life science department

85 boulevard Saint-Germain – 75 006 PARIS
+33 (0) 157 279 000
https://u-paris.fr/sdv/

INSTITUT DE MINéRALOGIE, DE PHYSIQUE DES MATéRIAUX ET DE COSMOCHIMIE
Sorbonne Université, UMR7590
CNRS, Muséum national d'Histoire naturelle
équipe bioinformatique et biophysique

4 place de Jussieu –75 005 PARIS
+33 (0) 144 275 897
http://impmc.sorbonne-universite.fr/en

19 june 2023

Document done through LaTeX

UNIVERSITé PARIS CITé

2^D YEAR IN MASTER DEGREE IN BIOINFORMATIC

SCHOOL YEAR 2022-2023

INTERNSHIP REPORT

Protein structural signatures as a new language for remote homology detection in sequence databases

Author:
Lucas ROUAUD

Traineeship supervisor:
Dr. E. DUPRAT
Dr. I. CALLEBAUT

UNIVERSITÉ PARIS CITÉ
Life science UFR
Life science department

85 boulevard Saint-Germain – 75 006 PARIS
+33 (0) 157 279 000
https://u-paris.fr/sdv/

INSTITUT DE MINéRALOGIE, DE PHYSIQUE DES MATéRIAUX ET DE COSMOCHIMIE
Sorbonne Université, UMR7590
CNRS, Muséum national d'Histoire naturelle
équipe bioinformatique et biophysique

4 place de Jussieu –75 005 PARIS
+33 (0) 144 275 897
http://impmc.sorbonne-universite.fr/en

19 june 2023
Document done through LaTeX

Acknowledgments

Foremost, I'd like to thank Isabelle CALLEBAUT and Élodie DUPRAT for their guidance, advice, and help throughout the course. I would also like to thank all my work colleagues, especially Kamel, Raphaël, Geoffroy, and Alex with whom I had some very interesting discussions with them, but also for their help during this internship.

I would also like to thank my great-uncle and great-aunt, Sergio and Renée, for their immeasurable support. I would especially like to congratulate them for putting up with me for these two (very long) years! Furthermore, I would also like to thank the other members of my family, including my parents, my sister, my little brother and my grandparents, for putting up with me all these years. Even though, deep down, I know I'm still their favourite...Right?

Finally, I'd like to thank a wonderful person who has always supported me and taught me patience when it comes to work. Someone I loved and will always love. A person who was interested in my work, curious, concerned about those around her, from her animals to her family member. A person for whom all the words on earth would barely suffice to say how much I owned her. Thank you, Grandma, for everything you gave me. May you rest in peace.

“The Dark Side of the Force is a pathway to many abilities some consider to be unnatural”

– Supreme Councillor Palpatine.

Table of contents

Acronyms

3D three-dimensional

AA amino acid

HC hydrophobic clusters

HCA hydrophobic cluster analysis

MSA multiple sequence alignment

NLP natural language processing

OCD ordered context diversity

RL redundancy level

RSS regular secondary structure

SCOPe structural classification of proteins extended

UCD unordered context diversity

W2V Word2Vec

1. Introduction

1.1. General context

1.1.1. Protein universe

Proteins are biological macromolecules that perform many functions in living organisms [1]. They can, for instance, play the role of catalyst as enzyme, or a structural role as cytoskeleton. Because they are so major in cells' life, it is important to understand their functions, which are linked to their three-dimensional (3D) structures [2]. Hence, it is possible to have a better understanding of pathological mechanisms at the molecular level and, eventually, to counter them. The experimental techniques used to get insights into the protein structures and functions are not always applicable, especially due to the difficulty and duration of these approaches [3,4] and the huge number of proteins in the universe [5] ($> 10 \times 10^{10}$). In this context, bioinformatics approaches are useful in so far they allow prediction of structural and functional features, based on homologs detection. Indeed, proteins can be grouped into approximately 20×10^3 functional families [6,7], which share similar 3D structures and sequence similarities, reflecting a common evolutionary origin.

1.1.2. Remote homology detection

So, as said just before, detecting sequence homology permits to predict function to proteins from the information of their amino acid (AA) sequences. However, if sequences are very different, remote homology can be difficult to detect, even when using the most sensitive tools like PSI-BLAST [8,9] or HH-BLITS [10]. This is due to the fact that sequences are less conserved than 3D structures. For instance, as reported in B. ROST's paper [11], "*most pairs of similar structures have sequence identity as low as expected from randomly related sequences (from 8 to 9 %)*". More sensitive tools, based on the 3D structure information, are desirable, in order to increase the detection of remote relationships. This issue is also relevant in the context of accurate prediction of 3D structures using artificial intelligence approaches such as AlphaFold2 [12], available for whole proteomes [13], insofar as this type of approach also relies on multiple sequence alignment (MSA) to predict contacts between residues. So, even if a substantial gain in predictions has been obtained, there are still some sequences in the "dark", corresponding to folded structures, as illustrated by recent works [14–16]. Adding 3D structure information can be made by considering the secondary structures, which can be predicted from the AA sequence information. However, accurate secondary structures prediction also generally relates to MSA [17]. Yet, an approach is developed in the internship lab and detailed below that allows this information on secondary structure to be considered from the information in a single AA sequence, without the need of homologues.

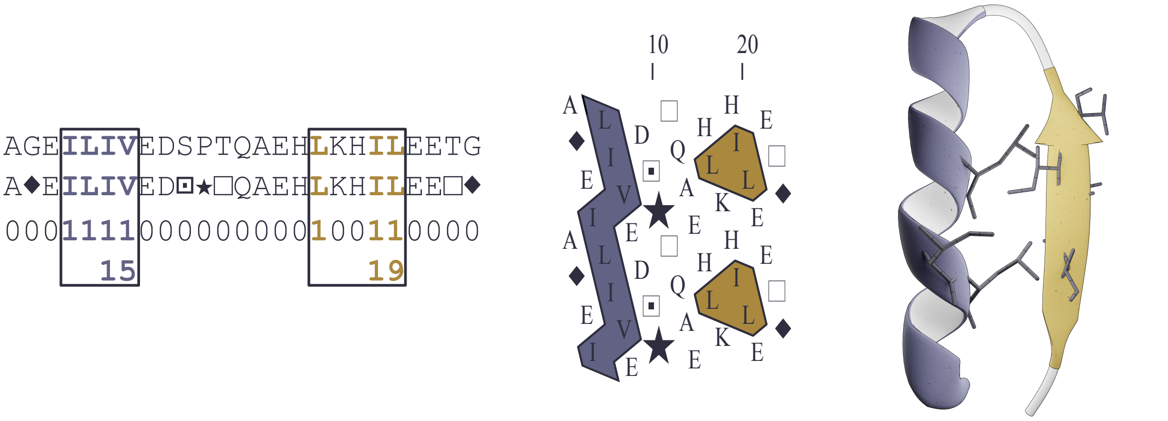


Fig. 1: **HCA principle.** *Star* = proline; *black diamond* = glycine; *square* = threonine; *square with a dot* = serine. An AA sequence (*top text line*) is translated into a binary code (*third text line*; with 1 being one of this strong hydrophobics AA: V, I, L, F, M, Y, W). Those can also be translated into a integer called Peitsch code (*last text line*). Then, an HCA diagram can be drawn, on which HC are directly linked to the RSS observed on the 3D structures, on the right (PDB accession code 3GT7, partial structure).

1.2. HCA

1.2.1. Principle

The hydrophobic cluster analysis (HCA) is a two-dimensional approach, based on the representation of the protein sequence on an alpha-helical net [18]. Its main principle is shown on [fig. 1]. On this net, one can highlight clusters of strong hydrophobic AA (V, I, L, F, M, Y, W), which mainly corresponds to regular secondary structure (RSS) [18]. A dictionary of hydrophobic clusters (HC) present in protein globular domains has been established [19,20], with some HC species (defined by a unique binary code) displaying marked propensities for the alpha-helices or beta-strands. As exemplified in A. LAMIABLE *et al.* [20] and illustrated on [fig. 2], these HC are more conserved than the AA sequence and permit to detect structural similarities, even at very low sequence identity [21]. This means that HC could be used as a discriminant feature (structural signature) to detect remote homology.

1.2.2. Previous works and limitations

In two previous internships, V. LEDUC [22] and N. FRANCILLONNE [23], have both tempted to build HC substitution matrices, with the aim of using such matrices instead of classical AA matrices like BLOSUM62 in similarity searches, and thereby increase the sensitivity. Also, they successfully highlighted substitution patterns between HC. This means that it could be possible to develop algorithms, based on these substitution matrices, to find similar proteins in terms of HC content, and so, in terms of 3D structure. Even if their work led to encouraging results, the matrices have the major limitation to be based on MSA of related sequences. As a matter of fact, using MSA have the problems to limit the search for substitution patterns at a sequence scale. And yet, the objectives are to be able to detect remote homology also at the superfamily, fold or even at an architecture level. So, a main objective of this internship is to test the possibility of constructing such a matrix without considering MSA. For this, it has been decided to use the fast-growing methods of natural language processing (NLP). More precisely, Word2Vec (W2V) is going to be used in order to extract subtleties on HC sequences.

```
1djm_A     1 MADKELKFLVVDDFSTMRRIVRNLLKELGFNNVEEAEDGVDALNKLQAGG      50
                 ..:.|:|:|..|....::::|:|.|: ..|...:|.:|:..|....
3gt7_A     1 ----AGEILIVEDSPTQAEHLKHILEETGY-QTEHVRNGREAVRFLSLTR      45

1djm_A    51 YGFVISDWNMPNMDGLELLKTIRADGAMSALPVLMVTAEAKKENIIAAAQ     100
             ...:|||..||.|||..|.:.::....:..:||:::|..:...:::.:.:
3gt7_A    46 PDLIISDVLMPEMDGYALCRWLKGQPDLRTIPVILLTILSDPRDVVRSLE      95

1djm_A   101 AGASGYVVKPFTAATLEEKLNKIFEKLGM----------     129
             .||..::.||.....|...:.::..  |:
3gt7_A    96 CGADDFITKPCKDVVLASHVKRLLS--GVKRTESITLAF     132
```

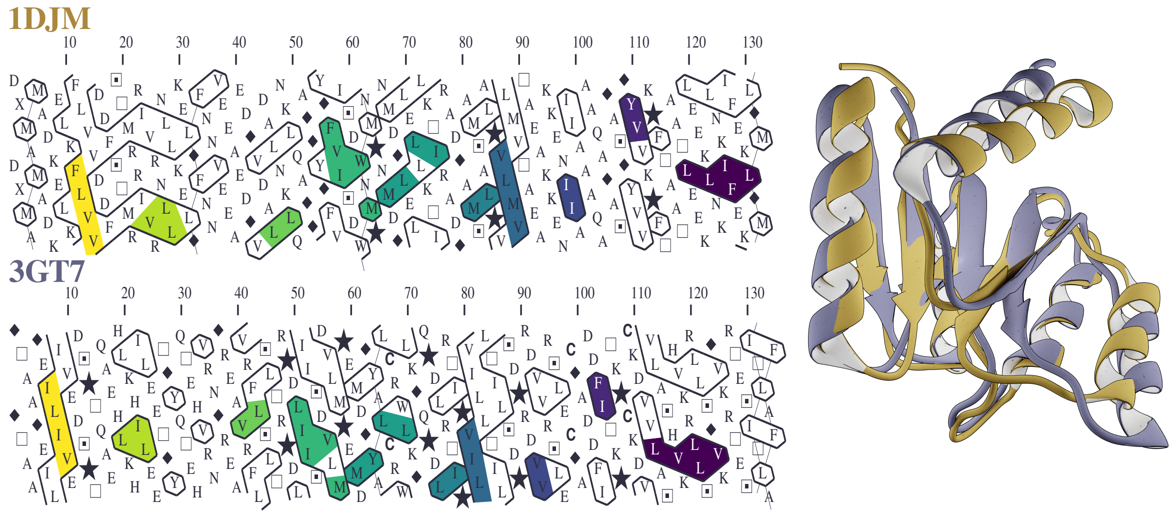


Fig. 2: Sequence versus HC conservation. *Top* = Sequence alignment of 1DJM and 3GT7, which are 3D structures experimentally solved. There is a low identity percentage between the two sequence (identity < 21 %); *bottom-left* = HCA diagrams of the two corresponding sequences. In this figure are highlighted the HC, with the conserved hydrophobic AA colored between; *bottom-right* = tridimensional structure (*yellow* = 1DJM; blue = 3GT7) alignment, done with iPBA [24–26].

1.3. NLP and W2V

1.3.1. Principle

The tool W2V [27] is a single-layer neural network that takes as input a corpus of words and gives as output numerical vectors. In other words, it permits to go from a text information to a vectorial space. These W2V vectors have multiples properties. The first major one is that when two words are close, thus are synonyms, the corresponding vectors are close to each other. This means that, by computing cosine similarities, values close to 1 are obtained for synonyms. In contrast, antonyms words give values close to -1. The second major property is that simple arithmetic operations are doable between vectors to obtain other ones. The most classic and known example is $\overrightarrow{\text{king}} + \overrightarrow{\text{woman}} - \overrightarrow{\text{man}} = \overrightarrow{\text{queen}}$ [28]. Note that W2V can be used in deep learning method, like to operate classification [29], and that there is other tool to achieve this goal, like GloVe [30], that treat the context in another way.

1.3.2. Vectors norm meaning and W2C application on biological context

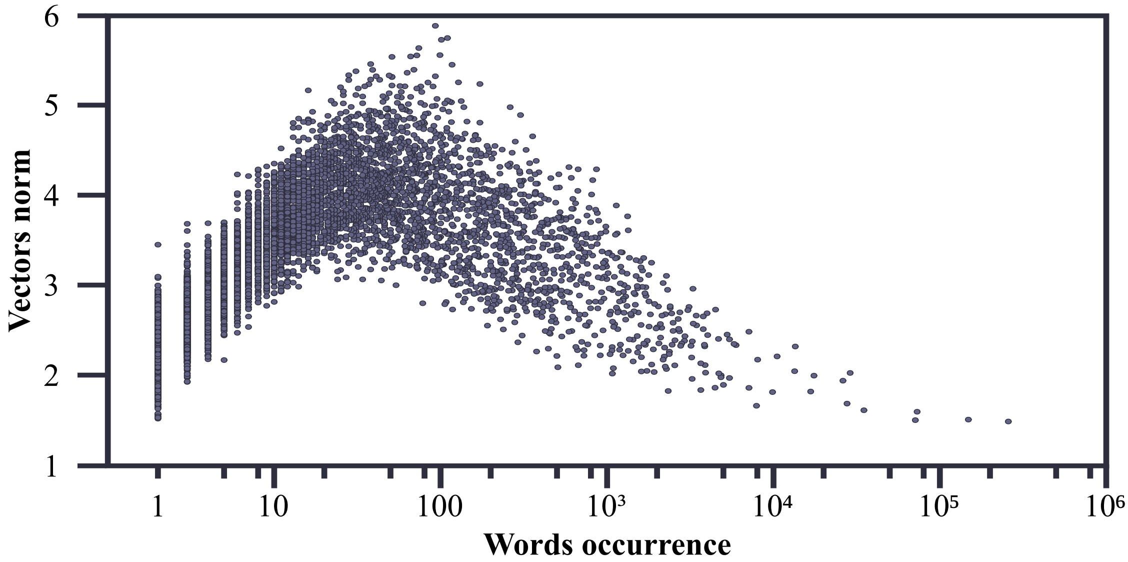


Fig. 3: Vectors norm as a function of the occurrence of the words. *Y-axis* = logarithm to base 10. This figure is extracted from the paper by A. M. J. SCHAKEL and B. J. WILSON [31]. **Some modifications were done to this plot (axis names, colors, font and general layout).**

First, A. M. J. SCHAKEL and B. J. WILSON [31] have shown that the vectors' norms are important to understand which words are significative or not during the learning. It is due to a problem in NLP of treating linking words with poor meaning [27]. So there is a necessity to decrease the weight of those kinds of words during the learning. Thus, it seemed worthwhile to study this norm, specifically in order to evaluated if W2V is applicable to a HC corpus and compare the obtained results. This is done especially in order to understand which HC words are important and are worth to be kept.

Furthermore, other studies have also applied W2V to biological corpus. This is for instance the case of MILLER *et al.* [32] who considered genes functions as a language, gene identifiers as words and annotated contigs as sentences. They used the resulting vectors to make a UMAP [33], a non-linear dimension reduction method for clustering data, to assign genes functions. Moreover, another study of BILESCHI *et al.* [34] used protein sequences as a language, defining AA as words and sequences as sentences. With the resulting vectors, they built a substitution matrix which shared multiple characteristics with the BLOSUM62 [35] one. This indicates that W2V seems to be an appropriate tool for highlighting fundamental features of protein structures.

1.4. Objectives

In this report, the question is to know whether the W2V tool could be used in a biological context, in which the text corpus is made up of HC, to produce *in fine* a cosine similarity matrix that may inform of their substitution patterns. HC indeed are words in a protein sequence text, which have a structural significance since they mainly correspond to RSS of folded domains.

In order to answer that, sequences of 3D structures experimentally resolved are considered, as reported in the structural classification of proteins extended (SCOPe) database [36–38]. These sequences are translated into HC (words) and then used as a corpus for the training dataset. The SCOPe database provides access to a hierarchical classification of 3D structures to select only the soluble domains, and to analyse results according to structural classes. Several redundancy level (RL) of sequences, defined by the maximum identity between pairs of sequences, are considered in the SCOPe database to analyse the impact of redundancy on the parameters studied. Indeed, it is recommended to work with large databases as the quality of the word vectors generated by W2V increases with the amount of training data, although high RL can also introduce a bias in results.

First, it is important to assess if the contexts, in which HC are embedded, are conserved. As a matter of fact, it is on this conservation that the possibility of highlighting HC synonyms through vector representation depends, which could be used to establish HC substitution matrices without resorting to MSA. An assumption is that the developed context metric should allow the measurement of conserved context, because a conserved context should lead to the learning of substitution patterns. Another one is that there will be a link between the context and the norms, as far as W2V learning is based on this context. Then, after learning the vector representation of HC, their overall significances are measured by analysing their norms, as done in the work of A. M. J. SCHAKEL and B. J. WILSON. These investigations should lead to the selection of a set of informative HC, on which a cosine distance matrix could be established in order to appreciate HC substitution patterns. The hypothesis is that HC with the more "structural signification", so with an intermediates size, will be kept. These results will also permit the selection of a database with a specific RL. The assumption here is that the higher RL should permits to have better results, as far as in NLP used corpus are generally very big (for instance, up to 783×10^6 words in T. MIKOLOV *et al.*'s paper [27]).

2. Material and methods

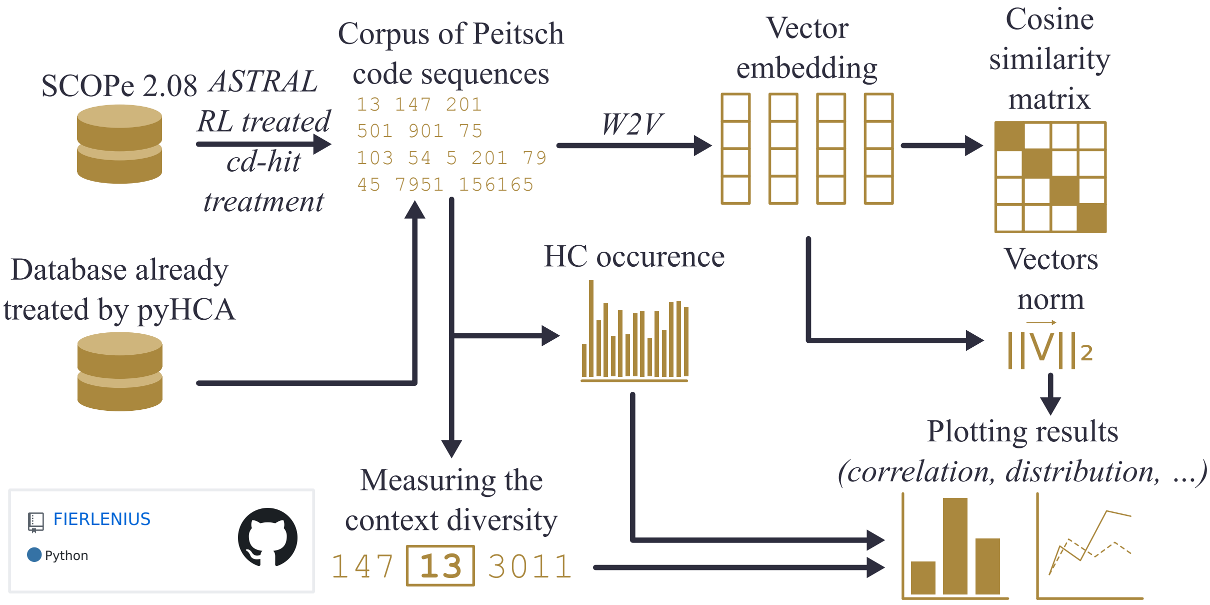


Fig. 4: Workflow used to produced all results. All the script are available on GitHub. After the data have been produced, they are analysed by doing plots presented in the result section. The "database already treated by pyHCA" correspond to the 30# RL database.

The general protocol followed in this study is shown in [fig. 4]. All implementation have been done in python `3.10.8` [39]. Numpy `1.23.5` [40] is used for vectorization and to compute vectors norm. Plotly `5.9.0` [41] and Matplotlib `3.6.2` [42] are used for plotting. igviz `0.5.1` is used to plot network (figure not shown here), available at https://github.com/Ashton-Sidhu/plotly-graph/. Pearson correlation and linear regression are computed using SciPy `1.8.1` [43]. All the scripts used for the realization of this manuscript are available on GitHub at https://github.com/FilouPlains/FIERLENIUS.

2.1. Protein database

2.1.1. SCOPe database

SCOPe [36–38] is a database providing a hierarchical classification of proteins domains of known 3D structures, solved experimentally. Those domains are organized following next characteristics: fold, superfamilies, families, species. The version `2.08` was used. Only the first four classes are taken in consideration (**a:** all alpha proteins; **b:** all beta proteins; **c:** alpha and beta proteins [a/b]; **d:** alpha and beta proteins [a+b]), leading to 4,472 domains in the dataset, and corresponding to soluble globular domains.

2.1.2. Redundancy sequence treatment

Originally, the considered dataset used in A. BRULEY *et al.*'s work [15] has a RL of 30 %. This dataset (noted as $30^{\#}$) was obtained from the Astral SCOPe database, considered at a 30 % of RL, and treated by pyHCA [15]. This software defined foldable domains from the SCOPe domains, and associated HC. As such, this database only contains HC features, without directly providing information of the AA content of the whole sequences (**[tab. 1]**). Moreover, by testing the identity percentage between sequences, it appears that a few sequences shared 100 % identity over a short segment, representing only a part of the total length of the compared sequences. Such 100 % overlap was present in the dataset extracted from Astral SCOPe at https://scop.berkeley.edu/astral/subsets/ver=2.08.

So, in order to avoid a possible bias in the analysis and consider different RL, the SCOPe database was downloaded from https://scop.berkeley.edu/astral/subsets/ver=2.08, with three defined RL (30 %, 70 %, and 90 % identity). In addition, to avoid the sequences which still shared 100 % of identity, cd-hit `4.8.1-2019-0228` [44] was used to treat the databases. Their characteristics after those treatments are given in **[tab. 1]**.

Tab. 1: Database characteristcs after redundancy treatment. # = database untreated by cd-hit; *#AA* = number of AA; *#Seq.* = number of sequences; *X* = data unavailable.

	$30^{\#}$ RL		30 RL		70 RL		90 RL	
Class	**#Seq.**	**#AA**	**#Seq.**	**#AA**	**#Seq.**	**#AA**	**#Seq.**	**#AA**
All	10,847	X	9,893	1,804,475	22,534	4,308,638	28,650	5,385,957
a	2,497	X	2,081	316,000	3,942	630,557	4,706	748,199
b	2,497	X	2,225	342,617	5,157	787,397	8,338	1,282,403
c	2,834	X	2,926	727,783	7,773	1,964,118	8,891	2,251,869
d	3,019	X	2,661	418,075	5,662	926,566	6,715	1,103,486

2.2. HCA

To translate a sequence of AA into a sequence of HC, everything was made according to A. LAMIABLE *et al.*'s paper [20]. To begin with, a sequence is translated into HC by "binarizing" into a series of `1` and `0`; with `1` being one of the next strong hydrophobic AA: V, I, L, F, M, Y or W. After this, HC are delimited by following the next rule:

1. HC are delimited by two `1`, which mean that the smaller possible one is `11`.
2. Two HC are separated with at least four `0` or one proline (P).

To finish, HC are translated into a unique integer code that describe it, called Peitsch code.

To obtain it, by considering that a HC is a vector (like `1011` being `[1, 0, 1, 1]`), the following formula is used:

$$\text{Peitsch code} = \sum_{i=1}^{\text{length(HC)}} \left(\text{HC}_{\text{length(HC)}-i} \times 2^{(i-1)} \right)$$

So, for instance, the HC `1100001` gives, by following the upper formula, the Peitsch code `97`. Note that, in addition to all what have been said, the cluster `11` is not taken in consideration. This is due to the fact that this very frequent HC species is not associated with marked propensities for RSS and most often observed in coils [19,20].

2.3. Word embedding

For the word embeddings generation, W2V `2018` [27] from Gensim `4.2.0` is used. The parameters used for this module/software are given in the [tab. 2].

Tab. 2: Used parameters for launching W2V model.

window	epochs	size	mintf	sample
10	20	300	2	10^{-3}
Words number taken in consideration for the context prediction.	Number of learning iteration to performed.	Size of the embedding vector.	Minimal word occurrence. If a word has a lower occurrence, it is not taken in consideration.	Down sampling of the words that are too frequent.

Note that for the downsampling, the probability to discard a word is:

$$p(\text{word}_i) = \left(\sqrt{\frac{f_{\text{word}_i}}{\text{corpus size} \times \text{down sampling}}} + 1 \right) \times \frac{\text{corpus size} \times \text{down sampling}}{f_{\text{word}_i}}$$

To use this tool, a corpus with defined sentences and words has to be given in input. A corpus is here a set of sequences from the SCOPe database (with different RL). A sentence is here a sequence translated into HC, more precisely, into Peitsch codes. A word is here a Peitsch code. Then, numerical vectors are obtained as outputs. When forming the corpus, it is possible to extract other information like the distinct Peitsch codes and their occurrences. Note that for the learning method, the parameter “unigram” is kept to default (n-gram was not used), which mean that words are taken one by one in consideration to predict the output, instead of n by n in the case of n-gram. Note that the skip-gram option is also kept to default, which is the way of

W2V to learn the context (take a one-hot encoded vector for a central HC, as input, and output predicted numerical vectors for the other surrounding HC –or the context–).

2.4. Implemented metrics

2.4.1. Context analysis

Because w2v is based on the context into a window of fixed size, it is necessarily to have a metric to measure the context diversity. To that aim, two methods are used. In NLP, words order can change while having the same meaning. For instance, “The doc eat the food” and “the food is eaten by the dog” mean the same, while having words in a different order. The metric linked to this type of context is called unordered context diversity (UCD). But in this biological cases, order have an important meaning, as far as AA sequences with a different order could give different proteins with potential different functions, and this should be the same with HC. The metric linked to this type of context is called ordered context diversity (OCD). So, for the UCD, the Bray-Curtis distance [45] is used:

$$\mathrm{UCD} = 1 - \frac{n \times \left|\bigcap_{i=0}^{n} S_i\right|}{\sum_{i=0}^{n} |S_i|}$$

Where S_i is a set and n the number of sets. Here, a set is all w HC surrounding a given one, with w being the window size (if $w = 10$, then 10 HC are taken at the left and 10 more at the right of the given HC). But, using this formula for all possible sets is too strict and almost always give maximum UCD values of 1.0. To compensate that, the UCD are computed between all pair of set, then all obtained paired-values are averaged.

The second one is an implementation from an algorithm from S. WU *et al.*’s paper [46], which is qualified as O(NP) algorithm. This metric is used to analyse the OCD. Note that a relative version is used (in other words, the result is divided by the sum of the length of both compared sequences). As UCD, all OCD are computed by pairs, then all obtained paired-values are averaged.

2.4.2. Cosine similarity matrix

A cosine similarly matrix is a symmetrical matrix of size n^2, with n the number of distinct words. To compute the cosine similarity matrix, for each cell of the matrix, the next formula is applied [47]:

$$M_{i,\,j} = \frac{\vec{v_i} \cdot \vec{v_j}}{\|\vec{v_i}\|_2 \times \|\vec{v_j}\|_2}$$

With $\vec{v}$ a vector output by W2V, and $\mathtt{M}$ the matrix to fill. Obtained values are included from -1 to 1.

3. Results and discussion

For all results, the SCOPe a, b, c, d classes have been considered, at three RL (30 %, 70 % and 90 %). Also, for a first time, a database previously used by A. BRULEY *et al.* [15] (designated next by 30 RL#) at a 30 RL from Astral is also considered. However, it was realized afterwards that it contained sequences which shared 100 % of identity, over a short length of sequences. These introduced bias in vectors norm analysis. Yet, this dataset was kept as a reference.

3.1. Characteristics of HC in the datasets

3.1.1. Relationship between the HC size and their occurrence

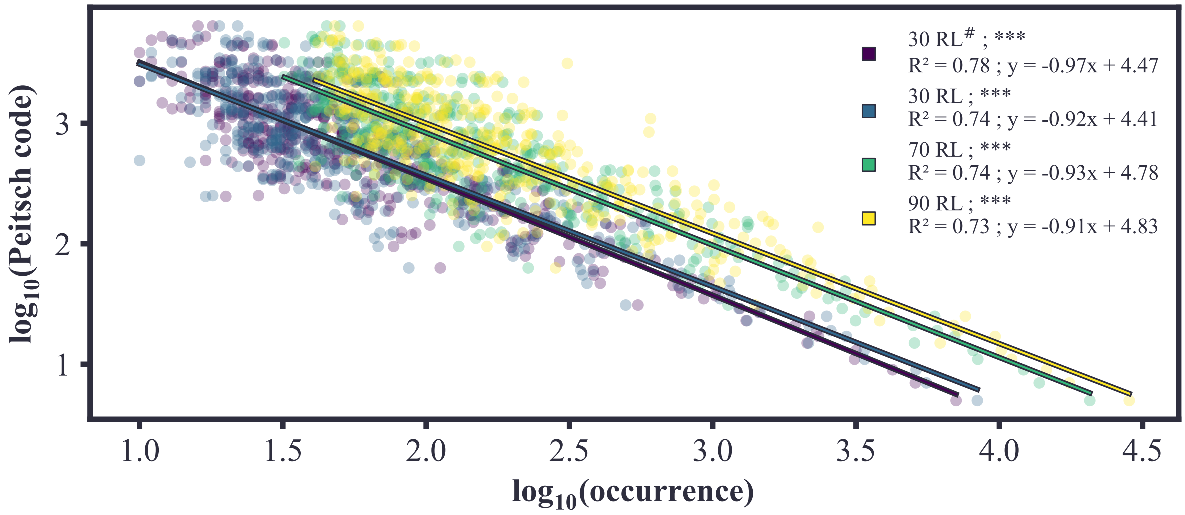


Fig. 5: Log$_{10}$ of Peitsch codes as a function of the log$_{10}$ of their occurrence. # = dataset untreated by cd-hit; *** = highly significant p-value, < 0.01, obtained with a two-sided test of the slope of the regression line. Equation, R^2 and p-values are obtained using a linear least-squares regression method from `SciPy`.

First, the relationship between the size of the HC, appreciated through their Peitsch code, and their occurrence is studied. As shown on [fig. 5], the log$_{10}$ of the Peitsch codes is correlated to the log$_{10}$ of their occurrences, for all RL, as indicates by strong R^2 (> 0,7) and highly significant p-values. This means that the more an HC is long, the less it occurs in the dataset. Also, a shift of the regression lines, obtained for the different RL, is observable. It follows the log$_{10}$(occurrence) axis, but not the other one. This shows that, as a function of the RL, the occurrence of HC is modified, with more HC at higher RL. Yet, the relationship between the HC size and their occurrence is still the same. In particular, no clear increase of occurrence of some particular HC can be observed at the highest RL, which could indicate the existence of very similar sequences. The consideration of high RL for analysing HC properties has already been supported by a previous study done by R. EUDES *et al.* [19], highlighting the stability of their properties (secondary structure preferences, AA composition) relative to RL.

3.1.2. Informative HC species

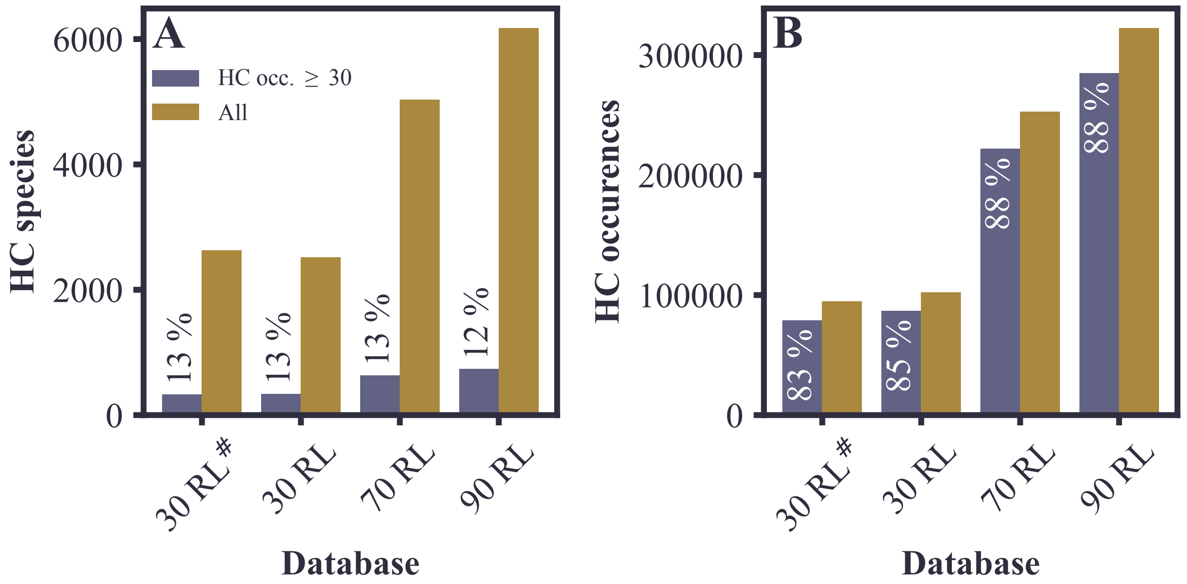


Fig. 6: HC species occurrences and HC occurrences. *A* = Distribution of HC species, $\mathrm{PERCENTAGE} = \frac{\text{HC species count}\geq 30}{\text{all HC species count}} \times 100$; *B* = Distribution of HC occurences, $\mathrm{PERCENTAGE} = \frac{\text{HC occurences count}\geq 30}{\text{all HC occurences count}} \times 100$; # = dataset untreated by cd-hit; *occ.* = occurrence. Legend is for both plot.

Next, two things are evaluated. The first one is how much HC species, each species corresponding to a unique Peitsch code, do contain at least 30 HC occurrences. The second one is what proportion of total HC occurrence present in these species, with at least 30 HC. This threshold of 30 HC was considered in the previous works of R. EUDES *et al.* [19] and A. LAMIABLE *et al.* [20], to allow a statistical analysis of the HC properties. In **[fig. 6: A]**, only the HC species present in all RL databases are counted, while in **[fig. 6: B]**, the total occurrences are counted for each cluster species in all RL databases.

For all RL, the proportion of HC species with at least 30 members is similar (from 12 to 13 % in **[fig. 6: A]**), as it is approximately also the case for the proportion of the HC occurrence (from 83 to 88 % in **[fig. 6: B]**). When the RL increased, not only the total number of HC increased, but also the number of HC species for which there is a significant number of occurrences. This means that a with higher RL, the diversity of HC increased at the same time as the HC occurrence. Interestingly, those informative HC species, although relatively few in terms of total occurrence (from 12 to 13 %) represent the bulk of the HC occurrence (from 83 to 88 %)

in the corpus studied. One can note that the 30 RL# database has less HC than the 30 RL. This is due to the 30 RL# being processed by the pyHCA software [16]. This software cut domains in foldable segment, leading to have less HC than the 30 RL remade database.

By combining all this information, the database that looks like to be the better for the training should be the one with the biggest RL. Simply because there are more data. Yet, in order to validate the database choice, other parameters must be checked.

3.1.3. Correlation of the HC occurrences between databases with different RL

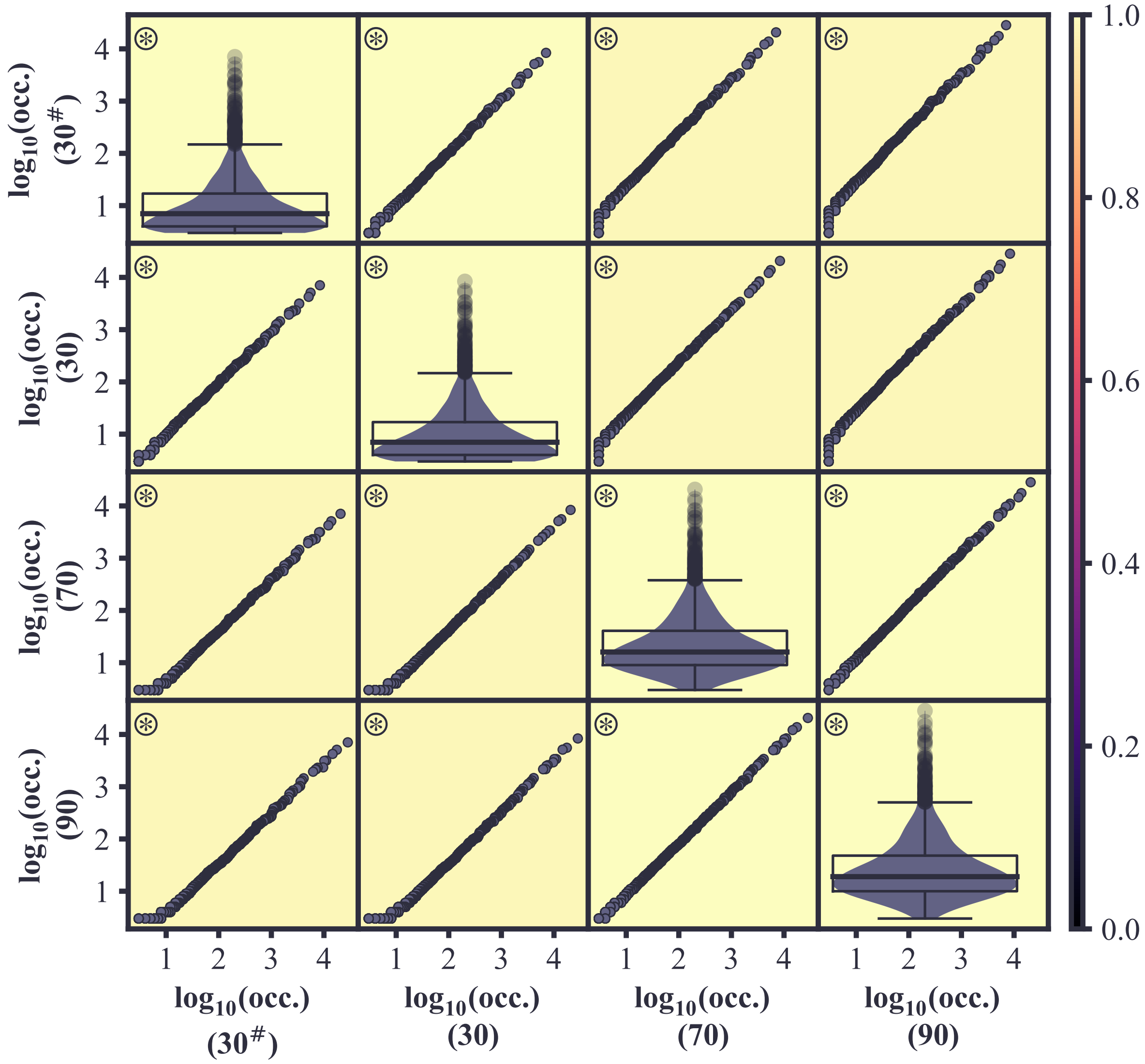


Fig. 7: Comparison of the $\log_{10}$(occurence) for the differents dataset using a heatmap and scatter plots. *Plot background* = R^2 value between two paired dataset; *color bar* = R^2 value scale; ⊛ = highly significant p-value, < 0.01; # = treated dataset; *diagonal* = box and violin plot of the $\log_{10}$ of the occurrence by RL; *one point* = one HC. The number in parenthesis indicates the database RL. A Pearson correlation test have been done using SciPy, between each paired of this matrix. For the diagonal, distribution of the $\log_{10}$(occurence) is shown. Only HC in common between all database are kept.

Next, the correlation of the occurrences of the HC between all considered databases is assessed. Also, potential outliers through scatter plots are checked. Here, there are only high correlations (R^2 around 1) with highly significant p-values. Also, no “significant” outliers are observed. This means that, independently of the RL, HC keep the same behaviour in terms of occurrences. Noteworthy is the break on the line for low $\log_{10}$(occurence), when comparing low RL with high one. This reflects the gain in information for infrequent clusters (large clusters). This is in line with previous observation done by R. EUDES *et al.* [19], when they showed a clear increase of information as cluster length increases.

3.1.4. Conclusion about occurrences and database RL

The first three figures of this results section showed important information about HC and HC species occurrences. First, increasing RL led to increase the number of HC into the dataset, both in terms of the number of informative HC species ($n \geq 30$) and of the global occurrences, without apparently introducing any bias due to an over-representation of some HC at high RL. The most frequent HC species ($n \geq 30$) moreover cover a large part (> 80 %) of the HC dataset, thus contributing to a large part of the corpus studied. Also, the HC occurrence is inversely correlated to their size, with an increase of information for the large cluster with the RL. All of these observations tend to conclude that a database with a high RL could be better, when the occurrence is the only parameter taken in consideration.

3.2. Analyses of the context diversity

3.2.1. OCD versus UCD

Context analysis is a central element of NLP approaches. While the order of words in natural language may be undifferentiated, the same may not be true for HC, the basic elements of 3D structures, as proteins are synthesized from the N- to the C-terminus and folds according to specific topologies. So here the correlation between conservation of an unordered context, generally considered in NLP approaches, with that of an ordered context, reflecting *a priori* the situation under study, is tested. Contexts were calculated according to the methodology presented in the materials and methods section **[2.4.1.]**, considering unigram and window sizes of 10 HC. So, there is 10 HC at left and 10 more at right of a considered HC.

Context conservation is inversely linked to the context diversity, which is presented in **[fig. 8]** for ordered and unordered contexts, for the four databases with different RL. The SCOPe databases were considered as a whole or with the different classes considered separately. Then, a linear Pearson coefficient is computed between OCD and UCD, for a same SCOPe level and a same database.

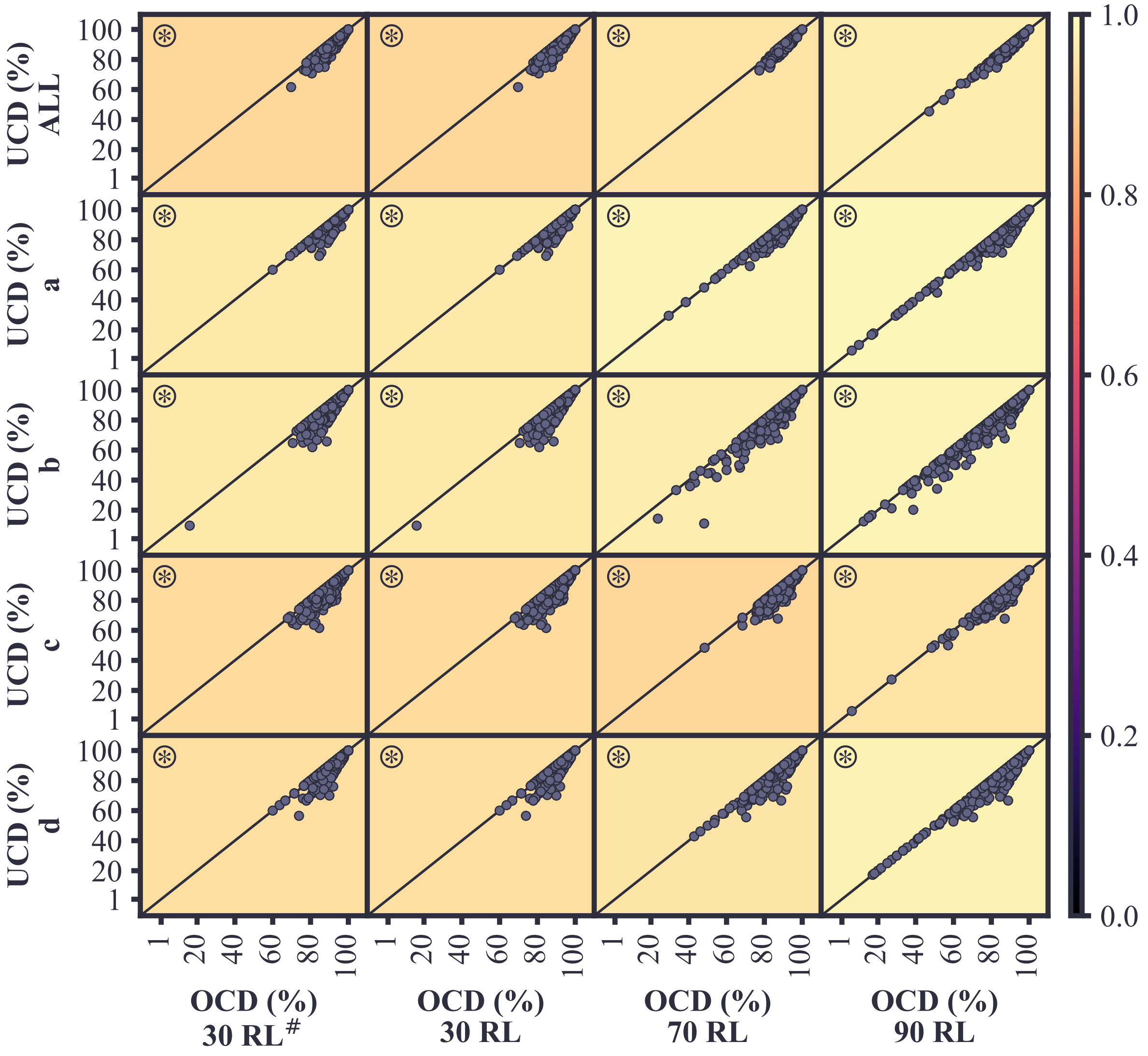


Fig. 8: Comparison of UCD and OCD for the different datasets and SCOPe levels using a heatmap and scatter plots. *Color bar* = R^2 values; ⊛ = highly significant p-value, < 0.01; *plot background* = R^2 value between two paired dataset; *line* = a specific SCOPe level; *column* = a database with a specific RL; *one point* = one HC. These correlations are computed for the different SCOPe level. A Pearson correlation test have been done using `SciPy`, between each paired of this matrix. Only HC in common between all database are kept.

First, all tested correlations are very high (around at least 0.9) and highly significant. Also, there are no very outlying values (only one for SCOPe b, 70 % RL database). This means that, firstly, the UCD metric is, in this biological context, correlated to the ordered one, even though diversity values are generally a little higher for the ordered context than for the unordered one (shift of the HC cloud under the diagonal). This means that if W2V is based on the UCD, this tool can still be used as far as these two are correlated. Second, it is not possible to determine, through those results, classes of HC that have bad correlation, which could be discarded from the dataset for the calculation of cosine matrix. The [fig. 8] also indicates that the context conservation logically increases (thus diversity decreases), for some HC, with the increase of the RL, but also when taking in consideration one specific SCOPe level. This suggests that higher RL are more appropriate for NLP-based context analysis.

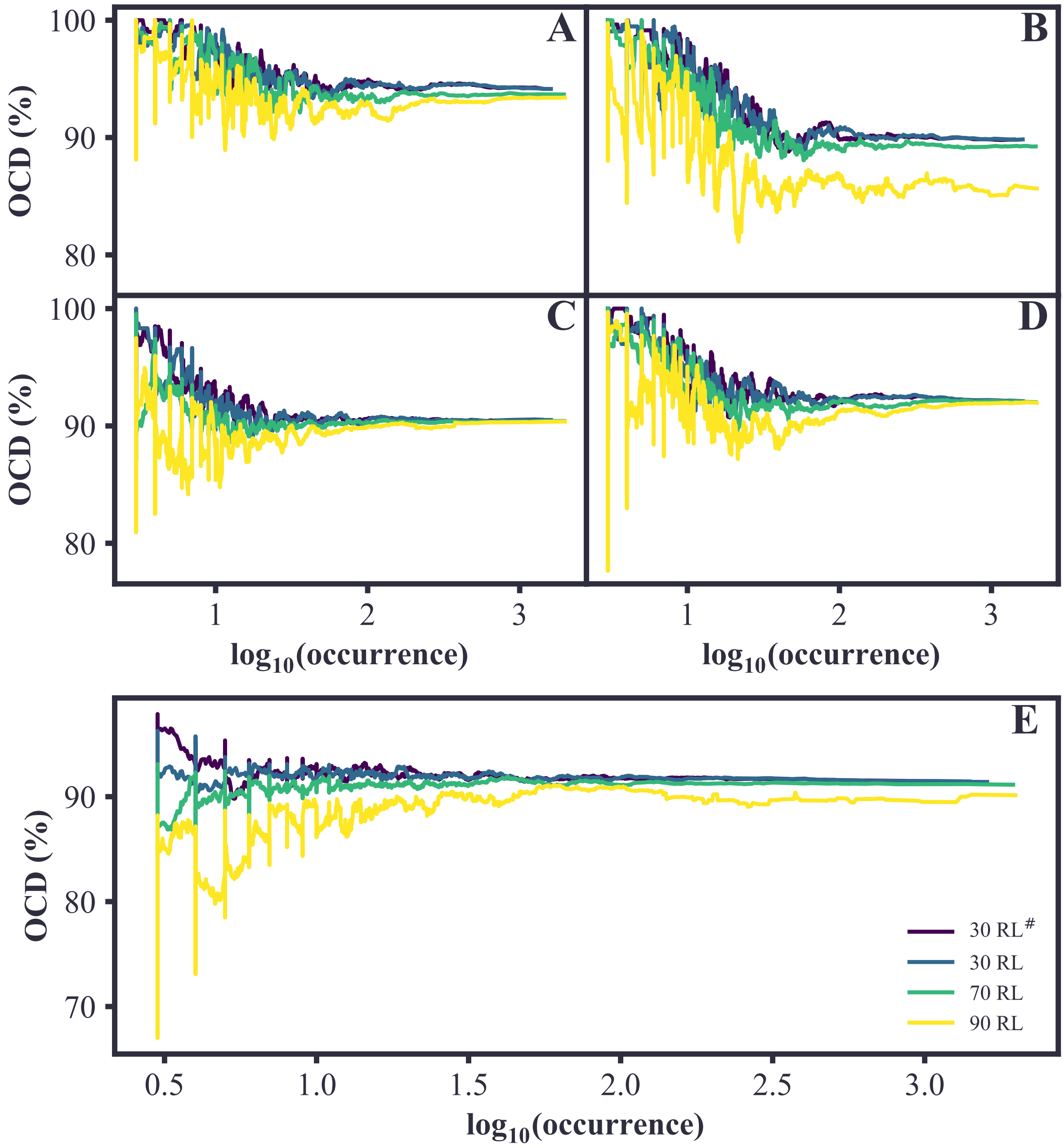


Fig. 9: OCD as a function of the $\log_{10}$ of the occurrence. *A* = SCOPe level a; *B* = SCOPe level b; *C* = SCOPe level c; *D* = SCOPe level d; *E* = SCOPe level a to d; # = dataset untreated by cd-hit. Only **floating average** is shown here due to the high numbers of points to render, making the chart unreadable.

Next, the context diversity is analysed as a function of HC occurrence, which is directly correlated to HC size. In **[fig. 9]**, context is extracted from the sequences, as in **[fig. 8]**. In order to take into consideration the SCOPe class levels, only sequences from a given class are taken in consideration (for instance, sequence from "SCOPe class a" only with a given HC are used to compute the OCD for the plot **[fig. 9: A]**).

When taking in consideration all SCOPe level (**[fig. 9: E]**), all HC having a $\log_{10}$(occurrence) from 1.0 to 3.5 have a context diversity around 90 %. Below 1.0, context diversity is much variable due to the fact that less frequent HC have a higher probability to share a conserved context (so low diversity). Now distinguishing between the different SCOPe classes (**[fig. 9: A-D]**), one can note that for the SCOPe class b (all beta), HC has a context diversity much lower than other level. In contrast, the SCOPe a (all alpha) has, on average, higher OCD values, while SCOPe class c and d ([a/b] and [a+b]) behave in-between. This can be interpreted that there are higher constraints linked to beta-strand assembly than to alpha-helix, consistent with literature data [48,49]. Finally, high RL has especially an impact on less frequent HC and on SCOPe level b which has a lower OCD than other database (**[fig. 9: E]**). It is seeable that on average, for all the HC occurrences, the HC diversity is lower for the 90 RL database than for the other databases, indicating that increasing the RL from 70 to 90 leads to a decrease of the diversity by introducing more similar sequences.

In conclusion, the context diversity is logically influenced by taking specific SCOPe class in consideration. Again, the 90 RL database should be appropriate for NLP, by presenting a global OCD lower (thus a context conservation greater) than the other RL databases.

3.2.3. Conclusion about the HC context diversity metrics

Foremost, it has been seen that even if W2V is based on UCD, its use can nevertheless be envisaged in an ordered context such as the sequences of globular domains of proteins, given that both ordered and unordered metrics are correlated. Then, it has been shown that the diversity in an ordered context, as present in protein sequences, can be influenced by refining the analysis and considering the SCOPe classes separately, indicating that this metrics can extract a biological meaning (beta-strand have a better conserved context than other secondary structure). Yet, by taking in consideration the issue of redundancy, it may be advisable to use the 90 RL bank as a training set, as this contains more information about context conservation. However, these conclusions are only valid for the conditions tested (window of length 10, unigram), and these parameters should be varied in future studies.

3.3. Analyses of the norm of the vectors produced by W2V

3.3.1. Vectors norm meaning and its link to the HC occurrence

While the direction of the vectors produced by W2V has been shown to contain semantic information [27,50], A. M. J. SCHAKEL and B. J. WILSON's work in 2015 has shown that vectors norm can be used as a measure of the word significance in a corpus. Therefore, the norms, computed by giving to W2V a corpus containing sequences as sentences and HC as words, are studied. This tool gave numerical vectors, which can be manipulated to obtain norms.

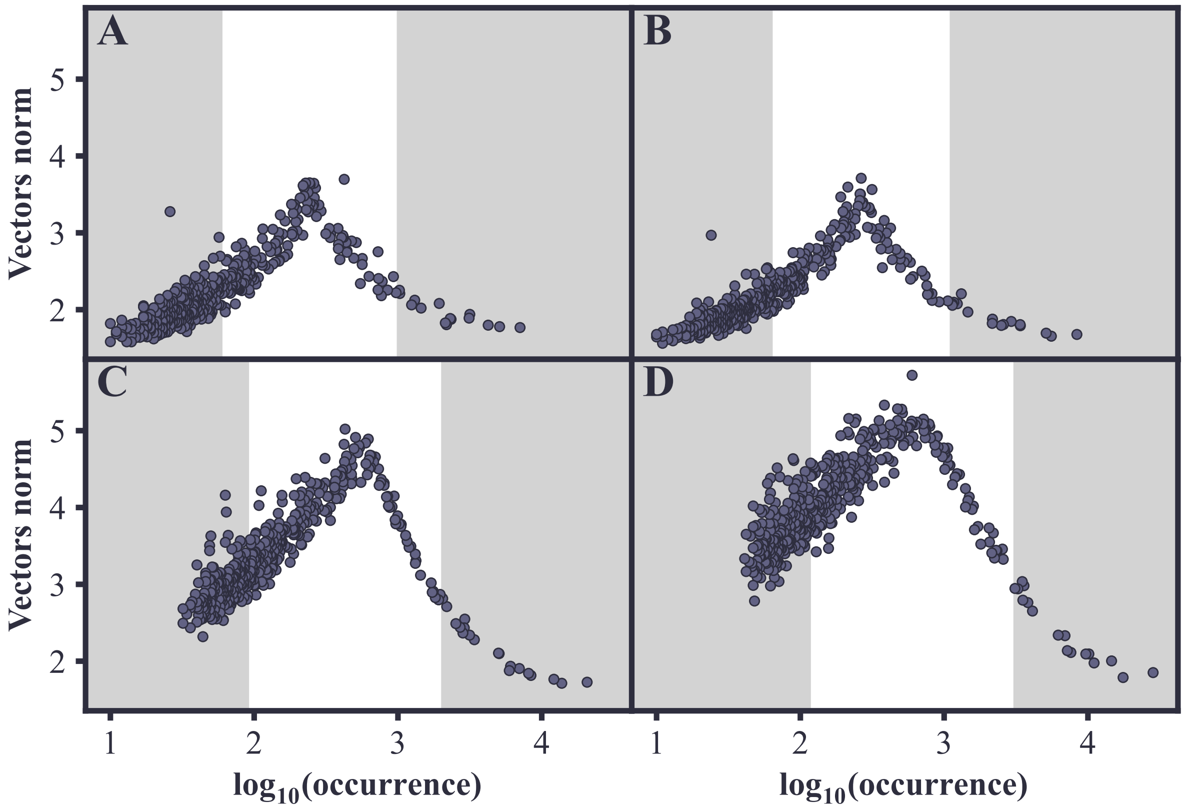


Fig. 10: Vectors norm as a function of the log$_{10}$ of the occurence, for different dataset. *Grey area* = Peitsch code with low meaning; *white area* = Peitsch code with high meaning; *A* = 30$^{\#}$ RL dataset untreated by cd-hit; *B* = treated 30 RL database; *C* = treated 70 RL database; *D* = treated 90 RL database; *one point* = one HC. **Note that the grey/white area are given for information purposes only.**

Foremost, according to A. M. J. SCHAKEL and B. J. WILSON's observations, low occurrences led to low norms due to W2V `--downsampling` parameters. Words that are overrepresented are randomly deleted to emphasize the learning on more significant words. These overrepresented words correspond here to short HC, as seen in the previous section, and correspond to small norms on **[fig. 10]**. Words that are not frequently encountered are badly learnt by W2V. They correspond here to long HC, as also seen in the previous section, and correspond to small norms on **[fig. 10]**. Therefore, it can be concluded that the vectors norm pic corresponds to HC of an "intermediate size". This is very interesting, due to the fact that too long and too short HC are not good structural descriptors. Indeed, long HC are generally made of severalRSS, while short ones are found in a wide variety of contexts and do not show marked tendencies towards a particular type of secondary structure [19,20]. Furthermore, another interesting observation is that the more the RL increases, the more the vectors norms pic increases. This means that with a database with a higher RL, the obtained vectors are more descriptive, thus leading us to select the 90 RL database as a training set.

3.3.2. Correlation between HC norms at different RL

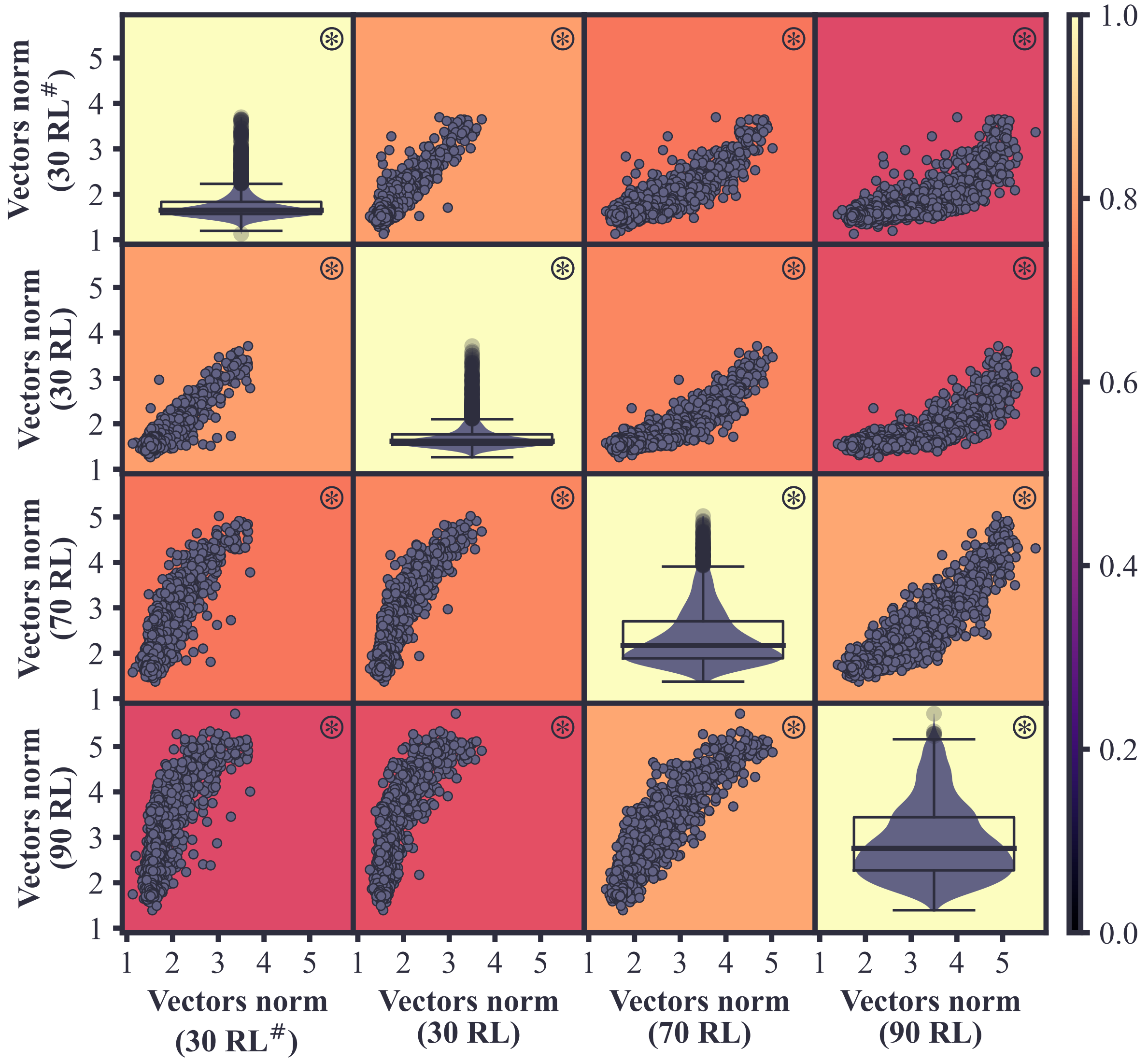


Fig. 11: Comparison of obtained vectors norm for the different dataset using a heatmap and scatter plots. *Plot background* = R^2 value between two paired dataset; *color bar* = R^2 value scale; ⊛ = highly significant p-value, < 0.01; # = treated dataset; *one point* = one HC. A Pearson correlation test have been done using `SciPy`, between each paired of this matrix. For the diagonal, distribution of the norm is shown. Only HC in common between all database are kept.

The [fig. 11] shows the comparison between all HC vectors norms computed for the different RL databases. The scatter plots are shown to identify potential outliers. There are only medium to high correlations (R² from 0.6 to 0.9) with highly significant p-values. Also, no significant outliers are observed. This means that, independently of the RL, HC keep the same behaviour in terms of vectors norms. The major difference is how the scatters plots are spread, as far as the maximal vectors norms values are dependent on the RL of the chosen database. In conclusion, because the vector norms appear well correlated, it would again be tempting to select the 90 RL database as a training set, at least in the context of the single set of parameters tested.

3.3.3. Conclusion about the vectors norms

The higher the RL of the database is, the higher are the HC vector norms, indicating a global higher significance of HC words. Moreover, what is observed in classical NLP, as documented in A. M. J. SCHAKEL and B. J. WILSON's article in 2015, appear to be transferable to this biological context, and makes sense with the HC structural behaviour in relation with their sizes.

3.4. Links between OCD and the vectors norm

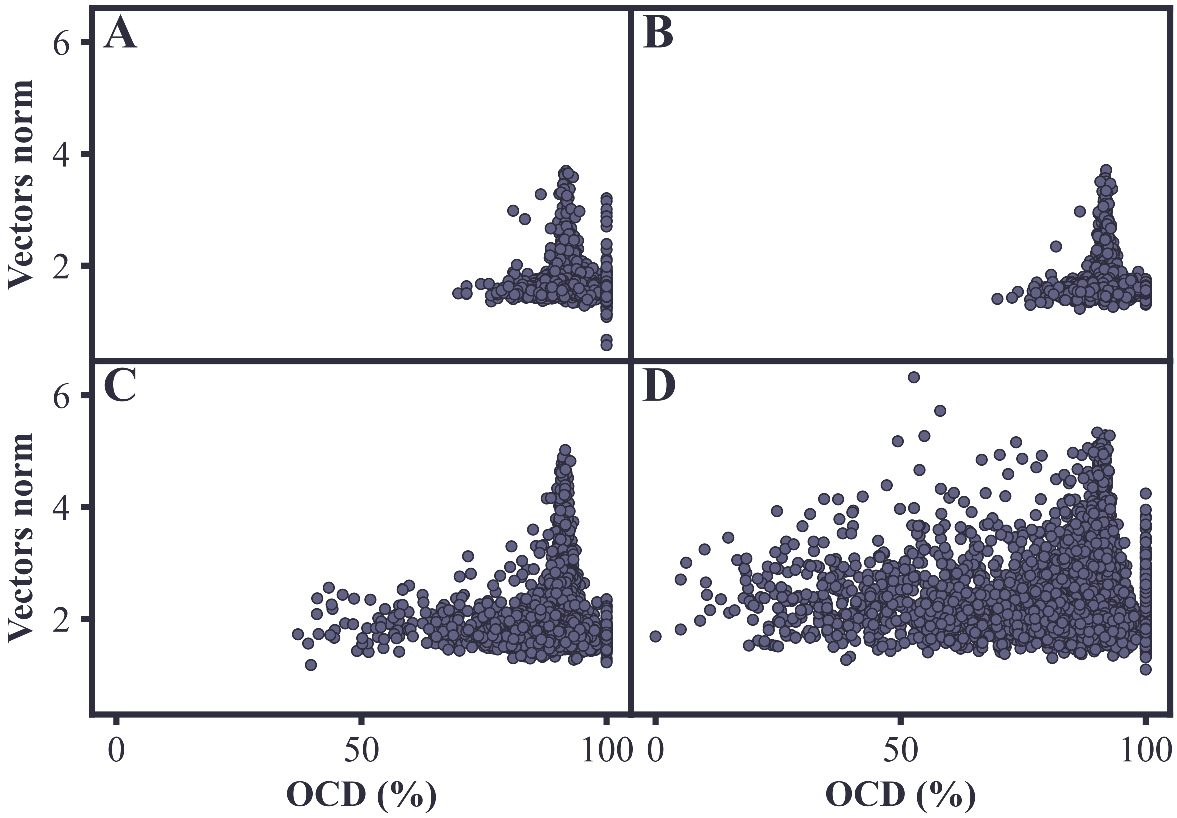


Fig. 12: Vectors norm as a function of the ordered context diversity. *A* = 30# RL database dataset untreated by cd-hit; *B* = treated 30 RL database; *C* = treated 70 RL database; *D* = treated 90 RL database; *one point* = one HC.

To evaluate whether the HC vectors norms are correlated to the OCD (thus inversely correlated to OCD), [fig. 12] is built. One major problem with this representation is linked to what has been explained before: Not all norms values are "significant" to be analysed, because some HC are down sampled and some are "not learnt" due to their low occurrence. Moreover, there is no existing threshold to choose optimal sets of HC with good norms values, to then compute a correlation between those values and the OCD. So, to counter this, a progressive threshold is used (no plots shown here):

1. A progressive vectors norm threshold is defined, going from 1 to 7.

2. For each threshold tested, a set of HC is selected. Those are selected only if their vector norm value is greater or equal than the current tested threshold.
3. For each set of HC that have been selected, a Pearson correlation coefficient is computed between their vectors norm and their OCD values, using `SciPy`.
4. When the p-value are at least inferior to 0.1, the R^2 greater or equal than 0.5, and there is a sufficient number of HC that have been kept, the threshold is kept.

With this approach, no correlation has been detected (only with at the most 10 points, which is not significant enough). In other words, the HC vectors norms are not, whatever threshold tested, correlated to the OCD. This could be explained in two ways: First, W2V is not able to learn HC context appropriately; second, the context diversity metrics used here are too strict.

3.5. Towards a cosine similarity matrix

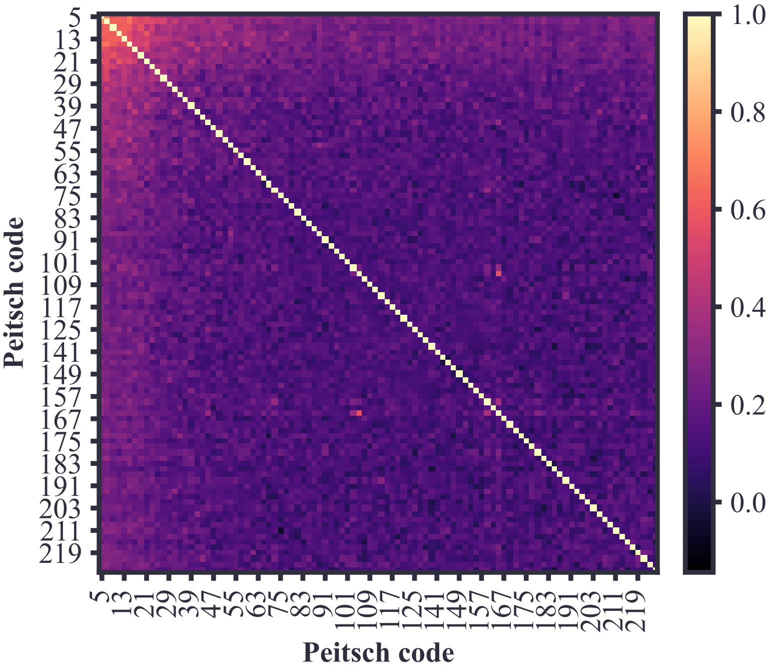


Fig. 13: Cosine similarity matrix using the 90 RL database. *Color bar* = cosine similarity. Note that not all Peitsch code are noted on axis, due to space limitations. Note also that the shown matrix is truncated too, due to size limitation.

A first cosine similarity matrix on [fig. 13] was built, according to the formula described in the material and methods and applied for each cell of the matrix, using as a reference the 90 RL SCOPe database. As far as, all previous observations tend to show that this one is the more appropriate (more data, higher vectors norm. . .).

First of all, the observed diagonal of the matrix is 1.0, meaning that the computed cosine similarity matrix is correct (as far as computing a cosine similarity between the same vector always give 1.0). Nevertheless, the rest of the matrix is quite particular: all cosine similarity values for Peitsch code > 31 are inferior to 0.3. However, even if substitution patterns do not appear as clearly as those previously observed in matrices established from multiple alignments of related sequences [22,23], a few observations can be made:

1. In agreement with previous observations [22,23], higher positive values (orange/pink, > 0.5) appear in the upper left corner of the matrix, testifying to substitution patterns adopted by short clusters (length $\leq$ 5, 31 of Peitch max code), mostly associated with beta strands.
2. Higher values are also observed at vertical (far left) and horizontal (far top) bands, testifying to substitutions where one of the clusters is extended to the left or right of the reference cluster.
3. Outside these regions, a few pairs with higher values can be observed. The most marked cosine value corresponds to the pair 105 – 165, as seen on [fig. 13]. These 165 (length 8) and 105 (length 7) HC can be aligned with 100 % conservation of hydrophobic AA (no substitution of 1 by 0), their substitution resulting from N-ter and C-ter extension, as shown on [fig. 14]. This type of substitution is clearly observed in matrices deduced from MSA [22,23] ([appendix I]).

```
Peitsch code 105 (beta strand):  1101001
Peitsch code 165 (beta strand):   10100101
```

Fig. 14: Alignment of the Peitsch code 105 and 165. *Yellow bold* = conserved hydrophobic AA between both HC.

In conclusion, these preliminary results support the relevance of the developed approach, using W2V applied to HC words, in highlighting substitution patterns, even if substantial improvement has to be made, especially by considering W2V parameters and learning datasets.

4. Conclusions and perspectives

4.1. Conclusions

During this internship, the W2V tool was evaluated whether it could be used in a particular biological context, related to the 3D structural behaviour of folded proteins. W2V is a NLP-related tool, enabling words to be converted into a vector space. In this space, words that are similar have vectors that are close to each other, and simple arithmetic operations can be performed. Here, the aim was to use a corpus of protein domains from the SCOPe database, corresponding to soluble globular domains, and grouped according to their structure and sequence similarities. These domains are characterized by the presence of RSS (alpha-helices and beta strands) folded around a hydrophobic core according to different topologies. The sequences in this database can be translated into HC, which include strong hydrophobic AA and the vast majority of which match the RSS. As these "structural word" are much conserved than the AA sequences, the aim was here to see if it is possible to extract patterns of substitutions between these HC trough a cosine similarity matrix. This matrix corresponds to the scalar products of the vectors, normalized by their norms, and defined by W2V.

Different RL of the SCOPe-derived database were tested, leading to select a high RL, the 90 % one. This database is enriched in terms of both HC species (each species corresponding to a unique binary pattern) and HC occurrences. In addition, this database allowed the obtaining of the lowest OCD values (in other words, the highest context conservation, and the highest vector norms. Higher norms are related to more descriptive vectors and therefore to a better learning potentiality of W2V. In addition, analysis of the vector norms allowed to highlight a subset of more informative HC (those with higher norms), consistent with the general marked propensities of those HC for RSS [20]. However, two first limitations of this tudy can be pointed out. First, the OCD metric was not effective. In fact, it was not possible to highlight a link between this value dans the vectors norm. However, it was still able to highlight some specific features linked to the SCOPe classification and the RL of the databases tested, by observing smaller OCD values. Finally, analysis of a first cosine similarity matrix highlighted some favoured substitutions between HC, consistent with previous matrices built using sequence alignments of related sequences. These preliminary results thus support the relevance of the developed approach, using W2V applied to HC words, in highlighting substitution patterns, even if substantial improvement has to be made. It is necessary to optimize or change various parameters, especially as only a few of them have been tested

So, to sump up, the most suitable database is the one with a RL of 90 %. Then, the OCD metric was not perfectly suited to the hypotheses posed, even if it has already enabled to grasp some subtleties linked to the SCOPe classification and the RL. Finally, the matrix obtained still needs to be optimized. However, the obtained results remain encouraging for applying NLP methods to this biological context related to 3D structures by managing to recover some logic of this domain as better context conservation in the SCOPe b class, corresponding only to beta strands.

4.2. Perspectives

4.2.1. Context diversity metric

The first step is to rework the context conservation metric. It does not take into account the `--downsampling` parameter, considered by W2V. This one randomly removes the most frequent words. However, when the context is calculated, a fixed size window is used without deleting words. As a result, the context is calculated on words that W2V has probably not used for its training. In addition, this metric is too strict, as it does not take into account nearby words. For instance, if the contexts **`147`** `45` and **`147`** `47`, with `45` having a cosine of 1 with `47`; then when the OCD is computed, it will give a value of 100 %, meaning that this context is not conserved. However, since these two HC are very similar, it would be much more interesting to consider an OCD value closer to 0 %, thus a conserved context, due to the similarities of these clusters. These parameters will therefore have to be added to the program already developed.

4.2.2. W2V parameters

The next step is to optimize the W2V parameters used during training. The parameters used are either random (like the `--downsampling`) or based in other articles [31] (like the `--epochs`). For this, several options are available, such as random test or the use of genetic algorithms. However, the major problem that needs to be resoled is the selection of parameters to validate the model produced. These should be intrinsic validations models (direct evaluation of the model produced by W2V) and extrinsic validation models (using another neural network to predict for instance class, and then validate this last model). For the moment, however, only loss minimization (a function that evaluates the difference between predictions and actual values) can be used in this context, which is an intrinsic evaluation method.

4.2.3. Database extension

Finally, a third perspective is to extend the database. For example, it is possible to use the non-redundant protein database [51] (constituted of 2,695,523) to form a much larger corpus, since in NLP, the corpus traditionally used can reach sizes of up to 196×10^6 words [27]. The work might therefore be carried out on AA sequences, independently of the knowledge of their 3D structures. However, it will not be possible, without further processing, to focus on folded domains, as protein sequences also include disordered regions. In order to focus on these foldable domains, it should be possible to use the pyHCA software. This software, described in the paper by A. BRULEY *et al.* [15], allows from the only information of single AA sequences, to sort foldable domains according to their structural characteristics (soluble, transmembrane domains, mainly disordered).

5. References

[1] Morris, R., Black, K. A. & Stollar, E. J. Uncovering protein function: From classification to complexes. *Essays in Biochemistry* **66**, 255–285 (2022).

[2] Stollar, E. J. & Smith, D. P. Uncovering protein structure. *Essays in Biochemistry* **64**, 649–680 (2020).

[3] Perdigão, N. & Rosa, A. Dark Proteome Database: Studies on Dark Proteins. *High-Throughput* **8**, 8 (2019).

[4] Perdigão, N. *et al.* Unexpected features of the dark proteome. *Proceedings of the National Academy of Sciences* **112**, 15898–15903 (2015).

[5] Dryden, D. T., Thomson, A. R. & White, J. H. How much

of protein sequence space has been explored by life on Earth? *Journal of The Royal Society Interface* **5**, 953–956 (2008).

[6] Mistry, J. *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Research* **49**, D412–D419 (2021).

[7] Paysan-Lafosse, T. *et al.* InterPro in 2022. *Nucleic Acids Research* **51**, D418–D427 (2023).

[8] Altschul, S. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research* **25**, 3389–3402 (1997).

[9] Schaffer, A. A. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Research* **29**, 2994–3005 (2001).

[10] Remmert, M., Biegert, A., Hauser, A. & Söding, J. HHblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods* **9**, 173–175 (2012).

[11] Rost, B. Protein structures sustain evolutionary drift. *Folding and Design* **2**, S19–S24 (1997).

[12] Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

[13] Varadi, M. *et al.* AlphaFold Protein Structure Database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research* **50**, D439–D444 (2022).

[14] Porta-Pardo, E., Ruiz-Serra, V., Valentini, S. & Valencia, A. The structural coverage of the human proteome before and after AlphaFold. *PLOS Computational Biology* **18**, e1009818 (2022).

[15] Bruley, A., Bitard-Feildel, T., Callebaut, I. & Duprat, E. A sequence-based foldability score combined with ALPHAFOLD2 predictions to disentangle the protein order/disorder continuum. *Proteins: Structure, Function, and Bioinformatics* prot.26441 (2022).

[16] Bruley, A., Mornon, J.-P., Duprat, E. & Callebaut, I. Digging into the 3D Structure Predictions of AlphaFold2 with Low Confidence: Disorder and Beyond. *Biomolecules* **12**, 1467 (2022).

[17] Zhang, B., Li, J. & Lü, Q. Prediction of 8-state protein secondary structures by a novel deep learning architecture. *BMC Bioinformatics* **19**, 293 (2018).

[18] Callebaut, I. *et al.* Deciphering protein sequence information through hydrophobic cluster analysis (HCA): Current status and perspectives. *Cellular and Molecular Life Sciences (CMLS)* **53**, 621–645 (1997).

[19] Eudes, R., Le Tuan, K., Delettré, J., Mornon, J.-P. & Callebaut, I. A generalized analysis of hydrophobic and loop clusters within globular protein sequences. *BMC Structural Biology* **7**, 2 (2007).

[20] Lamiable, A. *et al.* A topology-based investigation of protein interaction sites using Hydrophobic Cluster Analysis. *Biochimie* **167**, 68–80 (2019).

[21] Bitard-Feildel, T., Lamiable, A., Mornon, J.-P. & Callebaut, I. Order in Disorder as Observed by the “Hydrophobic Cluster Analysis” of Protein Sequences. *PROTEOMICS* **18**, 1800054 (2018).

[22] LEDUC, V. Étude de l’interchangeabilité des amas hydrophobes locaux au sein des domaines globulaires protéiques. Rapport de stage, CNRS UMR7590, Département de Biologie Structurale (2006).

[23] FRANCILLONNE, N. Développement d’un protocole d’analyse de substitution d’amas hydrophobes et son utilisation pour la reconnaissance d’apparentements lointains entre séquences. Rapport de stage, CNRS UMR7590, IMPMC (2012).

[24] Joseph, A. P. *et al.* A short survey on protein blocks. *Biophysical Reviews* **2**, 137–145 (2010).

[25] Huang, X. & Miller, W. A time-efficient, linear-space local similarity algorithm. *Advances in Applied Mathematics* **12**, 337–357 (1991).

[26] De Brevern, A., Etchebest, C. & Hazout, S. Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. *Proteins: Structure, Function, and Genetics* **41**, 271–287 (2000).

[27] Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient Estimation of Word Representations in Vector Space (2013). 1301.3781.

[28] Mikolov, T., Yih, W.-t. & Zweig, G. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 746–751 (Association for Computational Linguistics, Atlanta, Georgia, 2013).

[29] Yilmaz, S. & Toklu, S. A deep learning analysis on question classification task using Word2vec representations. *Neural Computing and Applications* **32**, 2909–2928 (2020).

[30] Pennington, J., Socher, R. & Manning, C. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543 (Association for Computational Linguistics, Doha, Qatar, 2014).

[31] Schakel, A. M. J. & Wilson, B. J. Measuring Word Significance using Distributed Representations of Words (2015). 1508.02297.

[32] Miller, D., Stern, A. & Burstein, D. Deciphering microbial gene function using natural language processing. *Nature Communications* **13**, 5731 (2022).

[33] McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction (2020). 1802.03426.

[34] Bileschi, M. L. *et al.* Using deep learning to annotate the protein universe. *Nature Biotechnology* **40**, 932–937 (2022).

[35] Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences* **89**, 10915–10919 (1992).

[36] Chandonia, J.-M., Fox, N. K. & Brenner, S. E. SCOPe: Classification of large macromolecular structures in the structural classification of proteins—extended database. *Nucleic Acids Research* **47**, D475–D481 (2019).

[37] Chandonia, J.-M. *et al.* SCOPe: Improvements to the structural classification of proteins – extended database to facilitate variant interpretation and machine learning. *Nucleic Acids Research* **50**, D553–D559 (2022).

[38] Fox, N. K., Brenner, S. E. & Chandonia, J.-M. SCOPe: Structural Classification of Proteins—extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Research* **42**, D304–D309 (2014).

[39] van Rossum, G. & Drake, F. L. *The Python Language Reference*. No. Pt. 2 in Python Documentation Manual / Guido van Rossum; Fred L. Drake [Ed.] (Python Software Foundation, Hampton, NH, 2010), release 3.0.1 [repr.] edn.

[40] Harris, C. R. *et al.* Array programming with NumPy. *Nature* **585**, 357–362 (2020).

[41] Inc., P. T. Collaborative data science. https://plot.ly (2015).

[42] Hunter, J. D. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering* **9**, 90–95 (2007).

[43] Virtanen, P. *et al.* SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods* **17**, 261–272 (2020).

[44] Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).

[45] Bray, J. R. & Curtis, J. T. An Ordination of the Upland Forest Communities of Southern Wisconsin. *Ecological Monographs* **27**, 325–349 (1957).

[46] Wu, S., Manber, U., Myers, G. & Miller, W. An O(NP) sequence comparison algorithm. *Information Processing Letters* **35**, 317–323 (1990).

[47] Metcalf, L. & Casey, W. Metrics, similarity, and sets. In *Cybersecurity and Applied Mathematics*, 3–22 (Elsevier, 2016).

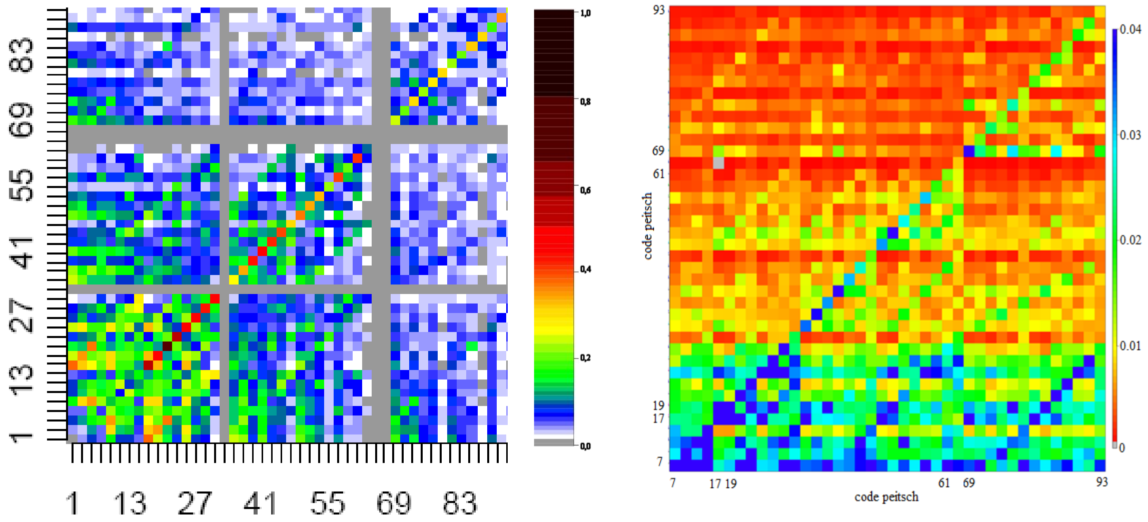
[48] Woolfson, D. N., Evans, P. A., Hutchinson, E. & Thornton, J. M. Topological and stereochemical restrictions in β-sandwich protein structures. *"Protein Engineering, Design and Selection"* **6**, 461–470 (1993).

[49] Kister, A. E., Finkelstein, A. V. & Gelfand, I. M. Common features in structures and sequences of sandwich-like proteins. *Proceedings of the National Academy of Sciences* **99**, 14137–14141 (2002).

[50] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. & Dean, J. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, 3111–3119 (Curran Associates Inc., Red Hook, NY, USA, 2013).

[51] Sayers, E. W. *et al.* Database resources of the national center for biotechnology information. *Nucleic Acids Research* **50**, D20–D26 (2022).

Appendix



Appendix I: Substitution matrix done in previous internship. *Left* = matrix done by V. LEDUC [22]; *right* = matrix done by N. FRANCILONNE [23].

INTERNSHIP REPORT

Protein structural signatures as a new language for remote homology detection in sequence databases

Proteins are biological macromolecules that perform various functions within cells. It is important to know their three-dimensional structure, which supports these functions. Comparative modelling or prediction tools such as AlphaFold 2 provide access to this structural information in the absence of experimental data, but these tools are based on the search for sequence homologies. Sequences are much less conserved than structures. This means that there are still proteins without detectable homologs, whose structure cannot be predicted. In this internship, a new approach to detect distant homology, based on structural descriptors, defined by hydrophobic cluster analysis, was studied. To this end, Word2Vec was used to transform a corpus of words (the clusters) into vectors. This corpus is a subset of globular domain sequences extracted from the SCOPe hierarchical classification database, considered at several levels of redundancy. In the case of close words (synonyms), characterized by a preserved context, the scalar product of the vectors generated by Word2Vec gives maximum values. Therefore, two hydrophobic clusters represented by these vectors could be substitutable. The characteristics of these clusters in terms of the occurrence and norm of their vectors were studied, highlighting, among other things, the value of banks with a high level of redundancy. In addition, subsets of clusters for which relevant information can be obtained were also studied. The conservation of the context of hydrophobic clusters varies as a function of redundancy levels and the SCOPe hierarchy, but is not, under the conditions tested, correlated with the norm of their vectors. Finally, the first observations of a computed cosine matrix support the relevance of the approach developed to highlight substitution patterns between hydrophobic clusters, even if improvements are needed. To this end, several points are being considered, relating to the improvement of context conservation metrics, the optimization of Word2Vec parameters or the consideration of larger databases for learning.

Les protéines sont des macromolécules biologiques opérant des fonctions diverses au sein des cellules. Il est important de connaître leur structure tridimensionnelle qui est le support de ces fonctions. La modélisation comparative ou des outils de prédictions comme AlphaFold 2 permettent d'accéder à ces informations de structure en absence de données expérimentales, mais ces outils reposent sur la recherche d'homologies de séquences. Or, celles-ci sont beaucoup moins conservées que ne le sont les structures. Donc, il reste encore des protéines sans homologues détectables, dont leurs structures ne peuvent être prédites. Dans ce stage, une nouvelle approche de détection d'homologie lointaine, reposant sur des descripteurs structuraux, les amas hydrophobes définis dans l'approche d'analyses des amas hydrophobes, a été étudiée. Dans ce but, Word2Vec a été utilisé pour transformer un corpus de mots (les amas) en vecteur. Ce corpus est un sous-ensemble de séquences de domaines globulaires extrait de la banque de classification hiérarchique SCOPe, considérée à plusieurs niveaux de redondance. Dans le cadre de mots proches (synonymes), caractérisés par un contexte conservé, le produit scalaire des vecteurs générés par Word2Vec donne des valeurs maximales. Donc, deux amas hydrophobes représentés par ces vecteurs pourraient alors être substituables. Les caractéristiques de ces amas en termes d'occurrence et de norme de leurs vecteurs ont été étudiées, mettant entre autres en évidence l'intérêt des banques à haut niveau de redondance. De plus, des sous-ensembles d'amas pour lesquels une information pertinente peut être obtenue ont aussi été étudiés. La conservation du contexte des amas hydrophobes varie en fonction des niveaux de redondance et de la hiérarchie SCOPe, mais n'est pas, dans les conditions testées, corrélée à la norme de leurs vecteurs. Enfin, des premières observations d'une matrice cosine calculée appuient la pertinence de l'approche développée pour mettre en évidence des schémas de substitution entre amas hydrophobes, même si des améliorations sont nécessaires. Ainsi, plusieurs points sont envisagés dans ce but, portant sur l'amélioration des métriques de conservation du contexte, l'optimisation des paramètres de Word2Vec, ou des banques d'apprentissage plus larges.

Author: Lucas ROUAUD
Traineeship supervisors: Drs E. DUPRAT and I. CALLEBAUT