POLITECNICO DI TORINO

Master of Science in Biomedical Engineering

Master's Degree Thesis

Comprehensive evaluation of docking algorithms on peptide drugs

Supervisors

J. A. TUSZYNSKI

G. K-S. WONG

M. DEYHOLOS

Candidate

Francesco ARABIA

Academic Year 2023/2024

Abstract

Peptides therapeutics is an upcoming research field and in the last years many papers have been published with promising results. A deep dive in small contigs (500 Da – 5000Da) could reveal a multitude of potential drugs against several diseases. Synthetized or natural molecules are the two main categories of study for drug development. Especially, the second variety includes a remarkable group of peptides which is the one derived from plants and is characterized by a cysteine motif. Thanks to their anti-fungal activity and their role in the plant immune system, these peptides are rising interest in scientific community. Here in this study, a database of 1000 plant peptides is analysed against known cysteine rich peptide family spacings. Then clustering is performed to obtain representative peptides for each family, the clustering algorithm used for this task is CD-HIT. Hundreds of potential representative sequences are collected, and they are ranked by the criteria of higher number of potential species within a cluster.

Acknowledgements

I appreciated and loved to work with brilliant minds like Professor Tuszynski, Professor Wong and Professor Deyholos.

It was an arduous path but with many fruits.

A special thanks to my family.

Table of Contents

List of Tables

List of Figures

Acronyms

PPI
Protein Protein Interaction
ABP
Anti-Bacterial Protein
AFP
Anti-Fungal Protein
Cys
Cysteine
1KP
1000 Plants
SALMdM
S-Adenosyl-L-Methionine-dyMethyl-transferase
NSLtP
Non-Specific Lipid transfer Protein
BBI
Bowman Birk Inhibitors
AMP
Anti-Microbial Peptide
ORF
Open Reading Frame
AEP
Asparagynil EndoPeptidase
KDD
Knowledge Database Discovery

Chapter 1

Introduction

1.1 Peptide therapeutics

The two main actors of drug discovery are small molecules (500 Da) and biologics (5000 Da), molecular weight of peptides places between these two (>500 Da). Certainly, peptides are more interesting than small molecules for different reasons. They are formed by 40 amino acids (simple design), they have ability to interact with underexplored targets, cheaper synthesis, decreased immunogenicity, and enhanced tissue penetration. [1]

Therefore, peptides are good candidates for disrupting protein-protein interactions (PPI) or for inhibiting intracellular receptor tyrosine kinases. PPIs are involved in biochemical process that regulates a series of enzymatic activities, so drugs will bind to these receptors, consequently affecting cellular behaviour. Proteins in PPIs are characterized by large, flat and hydrophobic surfaces so small molecules are not the ideal choice for disrupting these interactions, instead peptides could be an optimal solution. [2]

Despite that, peptides orally assumed are affected by several degradation barriers, such as proteolytic degradation, so they have a short half-life. Additionally, peptides are filtered by cell membranes thus it is hard to disrupt PPIs. [2]

A small molecule is “druglike” if Lipinski’s Rule of 5 is satisfied. Peptides have a weight larger than 500Da, so the Lipinski’s weight requirement is not met. Peptides could be degraded by saliva enzymes, stomach enzymes or by pH changes, besides a molecular weight less than 25kDa does not allow reabsorption through the renal tubule. Many orally assumed peptide drug candidates have entered clinical trials but with limited success. [1] All these limitations do not stop researchers to find a possible therapeutic solution.

1.2 Market trends

In the U.S., Europe and Japan markets, there are more than 100 peptide drugs used to treat a range of diseases. Financially, the peptide market is lucrative as it is estimated to be worth d11–16 billion annually by 2019. [13] Worldwide sales of peptides drugs account for more than $70 billion in 2019, a more than two-fold increase compared with 2013. [5] Globally, peptide therapeutics is expected to have an annual growth rate of nearly 9% from 2016 to 2024, with sales exceeding several billion dollars in 2019. [6] Chronic wounds could extend the necessity of hospitalization, an economic burden ranging from $28.1 billion to $96.8 billion dollars in the USA. Anti-Microbial Peptides with wound healing properties have been largely explored hence, they could be used in healing wound field to cut off costs. A projection revealed that synthetic peptides market will reach $ 426.4 million dollars by 2023, so there is a lot of room to grow. [7]

1.3 Plant peptides

One of the most interesting therapeutic approaches is based on plant derived peptides which is making its way on the market. Pathogens and pests are the main threats against plants, so plants have developed sophisticated defence measures to protect themselves. Anti-Microbial Peptides (AMPs) are the most common chemical barrier against pathogens, and they are effective even in humans. Also, due to the increasing resistance of pathogenic bacteria against antibiotics, AMPs are involved into treating infections. One of the main characteristics of the AMPs are the presence of cysteine motifs.[3] Anti-Bacterial Proteins (ABPs) are positively charged, and they interact with many plant bacteria. They overcome multi-drug resistance pathogens. Consequently, they could be considered as a potential alternative for a new class of antibiotics. The amino acid sequence, location and number of cysteine residues are the key classification criteria for ABPs. Anti-Fungal Peptides (AFPs) target some fungal components like chitin or by pore formation on fungal cell wall causing cell lysis. The anti-viral activity of plant peptides is another feature which could be crucial for several disease. In the 2019 modern society has affected by SARS-CoV-2, also called COVID-19, pandemic. Biopharmaceutical companies and public researchers did an enormous effort to find a vaccine. Among all these studies has been discovered that Lectin extracted from red marine alga *Griffithsia* sp. (GRFT) have been shown to inhibit the cytopathic effect of SARS-CoV.[8]

1.4 Cysteine rich peptides

Cysteine rich peptides (CRPs) are peptides with a molecular weight of 2–6 kDa containing a high number of cysteine residues. CRPs have a length of 40–100 amino acids [9] and they are presumed to be an effective source of potential therapeutic agents. Disulfide bonds and knotted motifs confer to peptides high stability in acidic conditions and resistance to proteolytic degradation. It is assumed this could solve the gastrointestinal tract degradation of the peptides when assumed orally.[4]
Here are listed CRPs families:

1.4.1 Defensins

These peptides are part of AMPs subset, they are structured by a characteristic β-sheet-rich fold and six-cysteine residues that form three disulphide bonds. In addition to this, all the plant peptides have a common three-dimensional structure. The tertiary structure of defensins consists of a triple stranded antiparallel β-sheet and a prominent α-helix, the latter stabilizes the complete peptide by the s–s bridges forming a cysteine stabilized αβ-motif.
Defensins are composed of two motifs, α-core (from first β-strand to the α-helix) and γ-core (β-strands 2 and 3), furthermore the three-dimensional structures of defensins in plants exhibit significant similarity (Fig.1). [9]

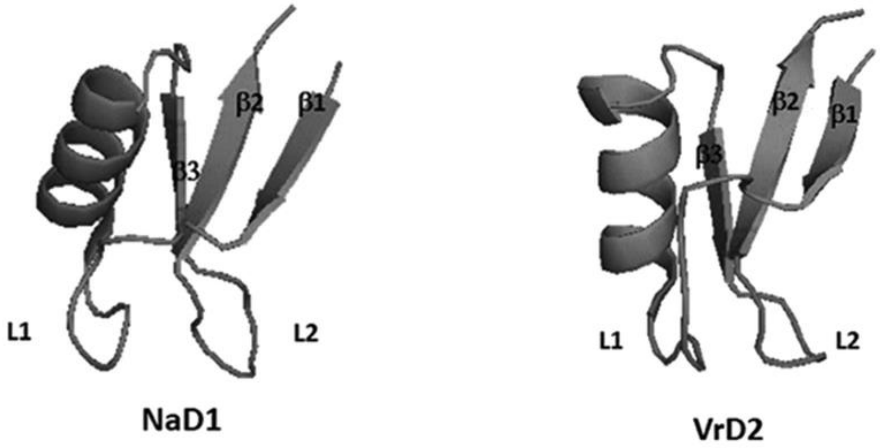

Figure 1 *Three-dimensional structure of two defensins. NaD1: Nicotiana alata defensin 1. VrD2: Vigna radiata defensin 2. [9]*

1.4.2 Cyclotides

Cyclotides are involved into host defence, they have some interesting structural features like a head-to-tail cyclized backbone and a cysteine knot core. These peptides are 28–37 amino acid long and contain six cysteine residues forming S–S bridges. Specifically, a cysteine knot is formed by the intersection of 3 disulphides CysI–CysIV, CysII–CysV and CysIII–CysVI. The interlocking cysteine knot provides stability and rigidity to the cyclotide backbone. Cyclotides are classified in three subfamilies Möbius, Bracelet and Trypsin inhibitor (Fig. 2).[9]

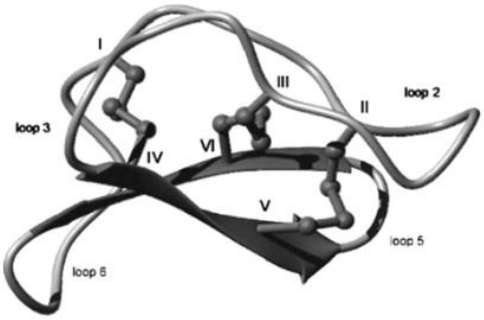

***Figure 2**: Three-dimensional structure of prototypical cyclotide, kalata B1.[9]*

1.4.3 Snakins

Snakins are AMPs which are similar to the GAST (gibberellic acid stimulated transcript) and GASA (gibberellic acid stimulated in Arabidopsis). Snakin-1 (StSN1) (Fig.3) and snakin-2 (StSN2) with 63 amino acid residues are the two main categories of Snakins family. They have a characteristic cysteine motif composed by 12 cysteines of the C-terminus in highly conserved positions. This arrangement is responsible of their structure and is essential for their biochemical activity as antioxidant. Nonetheless, many hypotheses on their biological functions were suggested, their mode of action is still unknown. [9]

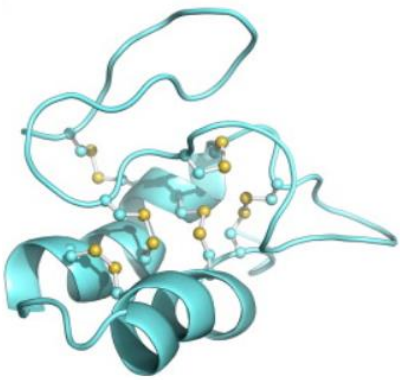

***Figure 3**: The cartoon representation of the final three-dimensional snakin-1 structure after 50 ns of simulation.[15]*

1.4.4 Non-Specific Lipid Transfer Proteins

Lipid Transfer proteins are involved in subcellular membranes activities to enable various metabolic patterns. They could be coupled with phospholipids and the term used to identify these peptides is “nonspecific lipid transfer proteins” (NsLTPs) (Fig. 4). NsLTPs are stabilized by eight conserved cysteine residues forming four disulfide bonds and they usually contain signal peptides in the N-terminus. NsLTPs can be divided into two main groups according to their molecular weight: nsLTP1 (9 kDa) and nsLTP2 (7 kDa). The disulfide bond linkages of nsLTP1 at Cys1-Cys6 and Cys5- Cys8 differ from those of nsLTP2 at Cys1-Cys5 and Cys6-Cys8. For the CXC motif in nsLTP1, X is a hydrophilic residue; however, in nsLTP2, a hydrophobic residue, such as leucine or phenylalanine, was found at the X position [16]

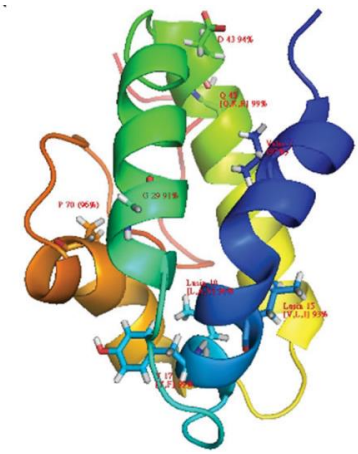

***Figure 4**: Representation of one the non-specific lipid transfer type 1 proteins.[16]*

1.4.5 S-adenosyl-L-methionine-dependent methyltransferases

S-Adenosyl-methionine (SAM)-dependent methyltransferases (MTases) catalyse the transfer of methyl groups from SAM to a specific substrate (metabolites or bio-macromolecules) generating S-adenosyl-homo-cysteine (SAH). SAM-dependent methyltransferases can be grouped into different types based on the substrates. The methylation process has important implications in various disease processes and applications in industrial chemical processing. The structure of these peptides have been solved by NMR and X-ray crystallography. (Fig. 5). [10]

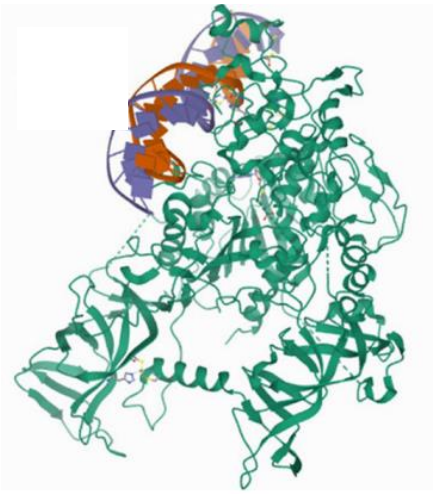

***Figure 5**: Crystal structures of human DNMT1(646–1600) in complex with DNA.[10]*

1.4.6 Protease inhibitor II

Generally, Protease Inhibitor-II protein consists of a double head-like structure having one reaction centre at each head. Each head contains two cysteine residues of a five amino acid residue. In one domain protease inhibitor have two cysteines that form two disulfide bonds with other two cysteines of another domain. The three amino acid residues between these cysteine residues often show variations between the homologues. The three-dimensional structure of a Protease inhibitor II extracted from *Cenchritis muricatusis* is showed (Fig. 6) [11]

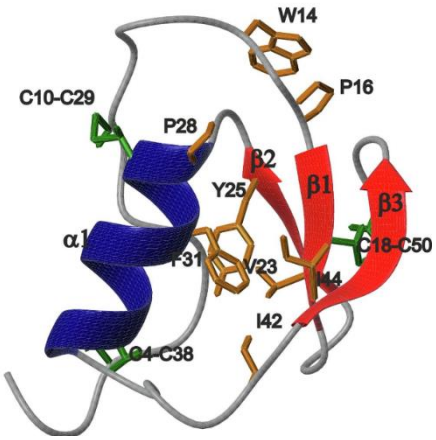

***Figure 6**: Ribbon representation of the CmPI-II lowest-energy structure.[17]*

1.4.7 Heveins

Hevein-like peptides (Fig. 7) are cysteine-rich peptides of 29–45 amino acids with 3–5 disulfide bonds. These contigs have a chitin-binding site, a secondary structure with coil-β1-β2-coil-β3 motif and variations. The central-sheet of the Hevein motif is formed by two antiparallelβ-strands, whereas the disulfide bridges are in the core. The C-terminal is a short helical segment. [9] Hevein-like peptides display antifungal activity, and they are promising against Gram-positive and Gram-negative bacteria. Studies have shown that Heveins have a big presence in the plant kingdom, so it supposed that they could have a relevant role in therapeutics. [18]

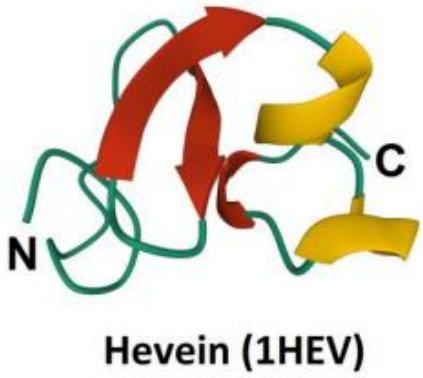

Figure 7: *3D representation of Hevein-like peptides. Helices in yellow and β-sheets in red. [18]*

1.4.8 Kazal

Kazal amino acid sequences are characterized by 40–60 amino acid residues including some spacer. By and large, the vertebrate Kazal domains and architecture are quite similar to invertebrate ones. The Kazal motif has a general amino acid sequence of C-X_a-C-X_b-PVCG-X_c-Y-X_d-C-X_e-C-X_f-C where the subscripts a, b, c, d, e and f are integral numbers of amino acid residues. Between cysteine numbers 1–5, 2–4,3–6 there are three disulfide bridges resulting in a characteristic three-dimensional structure (Fig. 8) [12]

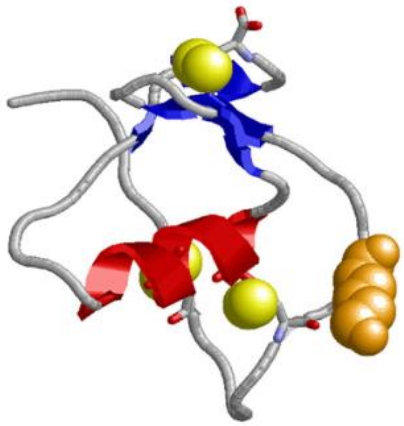

***Figure 8**: The structure of porcine pancreatic secretory inhibitor derived from the PDB file 1TGS. [12]*

1.4.9 Bowman-Birk protease inhibitors

Typically, Bowman Birk protease inhibitor (BBI) sequences have been determined from both monocotyledonous and dicotyledonous seeds. These single polypeptides with a conserved array of seven disulfide bridges, which stabilize their reactive site configuration. BBI have two homologous regions within three β-strands (Fig. 9). When two cysteine residues are conserved then disulfide bridges occur. Dicot BBIs have 14 cysteines, whereas the monocot BBIs with a similar size have 10 cysteines. [13]

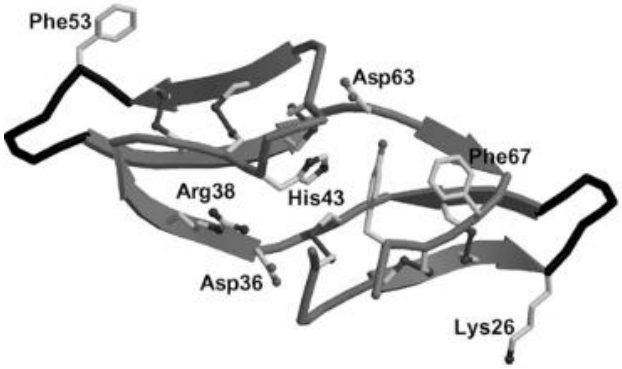

***Figure 9:** Ribbon diagram of a Bowman Birk protease inhibitor. [13]*

1.5 Circularization

A bunch of peptides have an exceptional structure and proteolytic stability, thanks to their cyclized backbone. The N-terminus links to a C-terminus with a peptide bond shaping a head to tail structure. Circular peptides are categorized in two different groups. The first one includes small peptides with a length less of 15 amino acids, they are synthesized by microorganisms, and their three-dimensional arrangement is not well known. The second group of peptides with more than 15 amino acids are identified in mammals and plants. Moreover, the three-dimensional structure of these larger peptides is defined even if some contigs have unusual amino acids produced by posttranslational modifications. [30]

1.5.1 Cyclotides

In general, precursors of these proteins are composed by an ER targeting signal, an N-terminal pro-peptide (NTPP), an N-terminal repeat (NTR) and cyclotide domain. These portions could be repeated up to 3 times and the final peptide usually is an hydrophobic C-terminal pro-peptide (CTPP). [30]

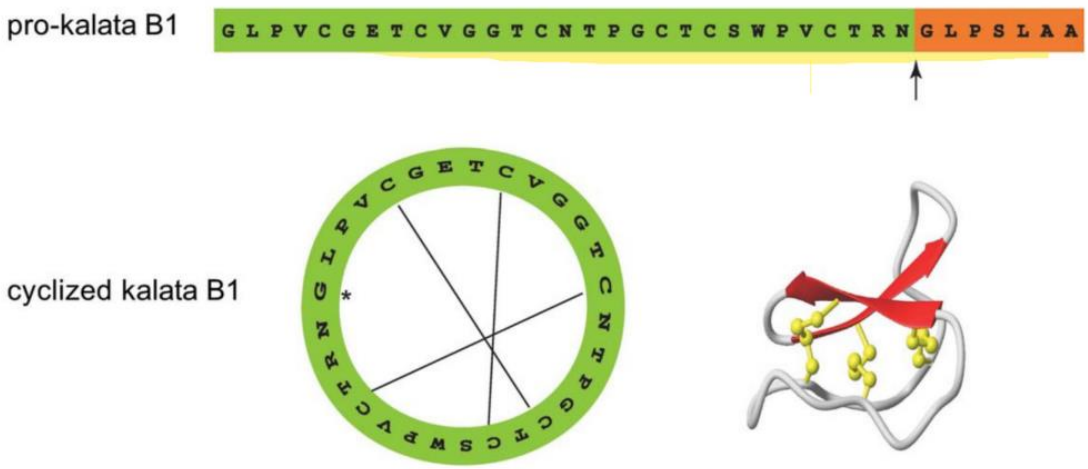

Figure 10: *The figure shows the linear sequence of the kalata B1 peptide and its cyclized form. In orange is signed the hydrophobic C-terminal pro-peptide (CTPP).[30]*

In figure 10 linkage point is pointed by the black arrow in the linear peptide, the C-terminus of the cyclized peptide is an Asparagine residue (N) or Aspartic acid (D) while the N-terminus is a Glycine (G). The point of cyclization is showed by "*" in the green cyclic peptide. Asparaginyl endopeptidase (AEP) is the enzyme which attacks the bond between the Asparagine and the tripeptide GLP. Notably, there are four residues before the first cysteine and three residues after the last cysteine.

Chapter 2

Bioinformatics

2.1 Six frame translation

Biological information is sorted in three levels:

- Chromosomal genome information.
- Expressed genome or transcriptome information.
- Proteome information.

DNA sequence portions that encode proteins or RNA molecules are called coding regions, while other segments are known as non-coding regions. Using genetic code tables, it is possible to translate DNA sequences. The start of a coding sequence is unknown, so it is essential to execute six-frame translation. First the translation begins with the reading of three forward frames starting from the first, the second and the third triplets. Next step is to continue reading three reverse frames, reversing DNA frame (complementary strand) and repeating the same logic as described for forward frames. The six frames are potential proteins, normally only one of these frames is biologically functional. (Fig. 11) [31]

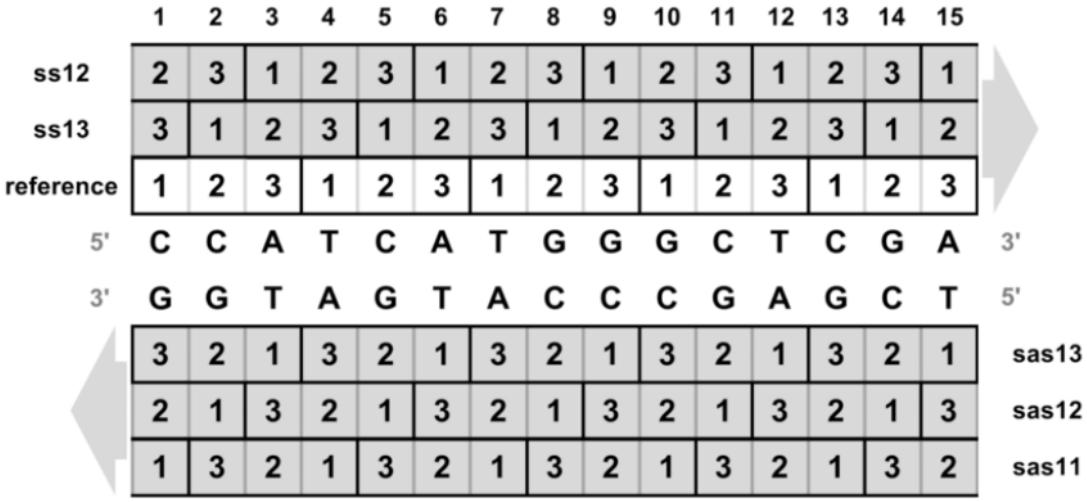

Figure 11: *the first triplet is the starting point of the reference, ss13 is associated to second triplet and ss12 is associated to the third triplet. Sas13 is associated to the first reverse triplet, sas12 is associated to the second reverse triplet and sas11is associated to the third reverse triplet.[32]*

2.2 FASTA format

In bioinformatics the FASTA format is a text-based format used for representing amino acid, nucleotide or nucleic acid sequences. A sequence begins with a greater-than character (“>”), followed by a description of the sequence (name, unique identifier and additional information). NCBI also defined a standard for unique identifiers. Therefore, multiple DNA/RNA or peptides could be classified in one single file. The extension of FASTA files is not defined but there are several extensions to indicate the type of sequences. A lot of software and scripts could be used for manipulating FASTA files such as FaBox [33] or FASTX-Toolkit within Galaxy servers [34] or own made python scripts.

```
>SEQUENCE_1
MTEITAAMVKELRESTGAGMMDCKNALSETNGDFDKAVQLLREKGLGKAAKKADRLAAEG
LVSVKVSDDFTIAAMRPSYLSYEDLDMTFVENEYKALVAELEKENEERRRLKDPNKPEHK
IPQFASRKQLSDAILKEAEEKIKEELKAQGKPEKIWDNIIPGKMNSFIADNSQLDSKLTL
MGQFYVMDDKKTVEQVIAEKEKEFGGKIKIVEFICFEVGEGLEKKTEDFAAEVAAQL
```

***Figure 12**: Representation of amino acid FASTA file sequence. >SEQUENCE_1 is the header of the sequence.*

2.3 Open Reading Frames

Biological pipeline of translation begins from transcription of the double stranded DNA to mRNA, afterwards mRNA is processed to form the mature mRNA. Ribosomes are molecular structure employed for allowing translation of mRNA into amino acids. The protofilament generated by ribosomes is known as protein. The start codon, a codon is a triplet of bases, usually is a methionine (AUG), the translation stops only when a stop codon is found (UAA,UAG,UGA).

Other molecular structures involved in the translation process are the t-RNAs, which read the codons and then they carry the corresponding amino acids. Subsequently, the protein is built by peptide bonds among these amino acids. (Figure 13)

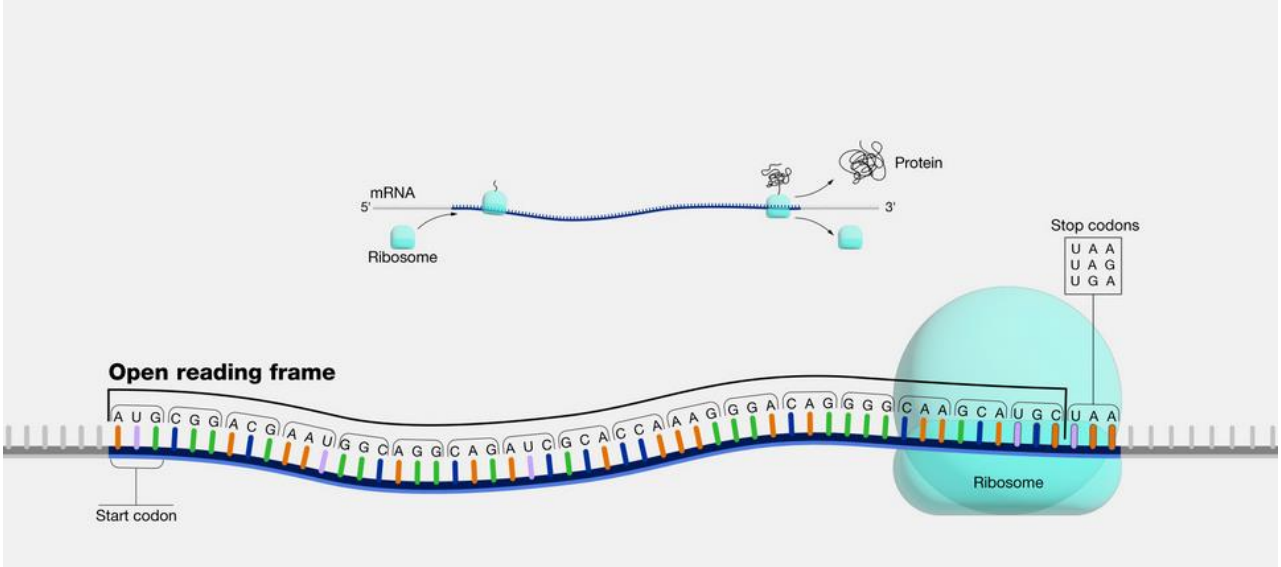

***Figure 13**: Translation process. Open reading frame is highlighted.*

Open reading frame is a portion of DNA, or RNA, transcribed into RNA which doesn't include a stop codon. [35] The biological functional frame usually is the longest open reading frame. These frames are functional to execute computational analyses or collecting them in datasets for clustering.

2.4 Database mining

Data mining means to “mine” a large amount of data from a database and try to find patterns, association, common features helpful for several applications. One of the main applications is in the biomedical field to enable researchers to find new treatments to improve medical care and life knowledge. Data mining is called Knowledge Discovery in Database (KDD), this methodology in bioinformatics is fundamental for fields of study such as drug development, protein modelling, gene expression and biomarker identification. The pipeline of data mining is showed in figure 14.[36][37][38]

***Figure 14**: Data mining pipeline.*

Chapter 3

Analysis of 1kp database

3.1 The 1000 plant transcriptomes project

The 1000 plants initiative (1KP) is a multidisciplinary consortium aiming to generate large-scale gene sequencing data for over 1000 species of plants. These species potentially are involved into agriculture and medicines, thus future studies are likely to be decisive in the discovery of new therapeutics. The 1kp project is rising interest in the scientific community and several papers are already published. This initiative sequenced and analysed transcribed RNA from 1,342 samples representing 1,173 green plant and chloroplast bearing species. Furthermore, in the 1kp project the Viridiplantae: streptophyte and chlorophyte green algae, bryophytes, ferns, angiosperms, and gymnosperms taxa examples, are included. [14]

3.2 Experimental procedure

3.2.1 Download FASTA files from 1kp database.

First step is to download all the entries from a database, from the GigaScience paper published in 2019 (https://doi.org/10.1093/gigascience/giz126) in the supplementary material there is a google drive link of the 1kp database (https://drive.google.com/drive/folders/175nB8kf1UQushuEzv7UaJLPNNwdOrxh5?usp=sharing).
All the trans-assemblies were downloaded as “fasta.gz”.

3.2.2 Removing mislabelled/contaminated files

Initially, the number of trans-assemblies is an amount of 1455 samples, afterwards we removed the known mislabelled/contaminated FASTA files. The list of 29 specimens is published on Giga DB website (http://gigadb.org/dataset/100910)

3.2.3 Six frame translation and Open Reading Frames

Using several own made python scripts, a FASTA file with ".gz" extension is uncompressed and for each sequence of one single FASTA file six frame translation is executed. (Fig. 15)

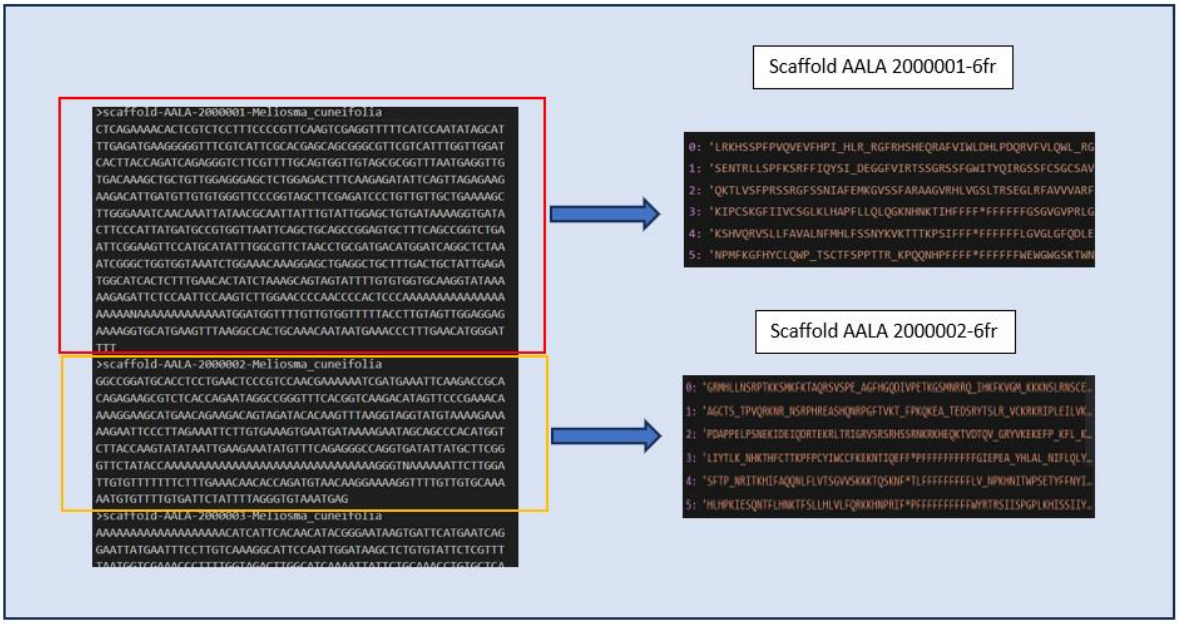

Figure 15: *On the left multiple sequences from one FASTA file. On the right the six-frame translation of the respective sequence. The software used for this analysis was Python.*

The script scans every DNA sequences. Some sequences have the letter “N” which stands for undefined base so when the script encounter a non-translatable triplet it translates the triplet in “*”. The translation will stop when a STOP CODON (“_”) is found. After, the script finds the Open Reading Frames (ORF) for all six frames. One frame could have several ORFs, so they are potential ORFs. (Fig. 11) The finding has the following logic: find a start codon “M”, then find a stop codon “_”, thus save open reading frame (with several meta data such as 4 letter id, scaffold id, number of frame, number of ORF). Repeat if there are more starting codons.

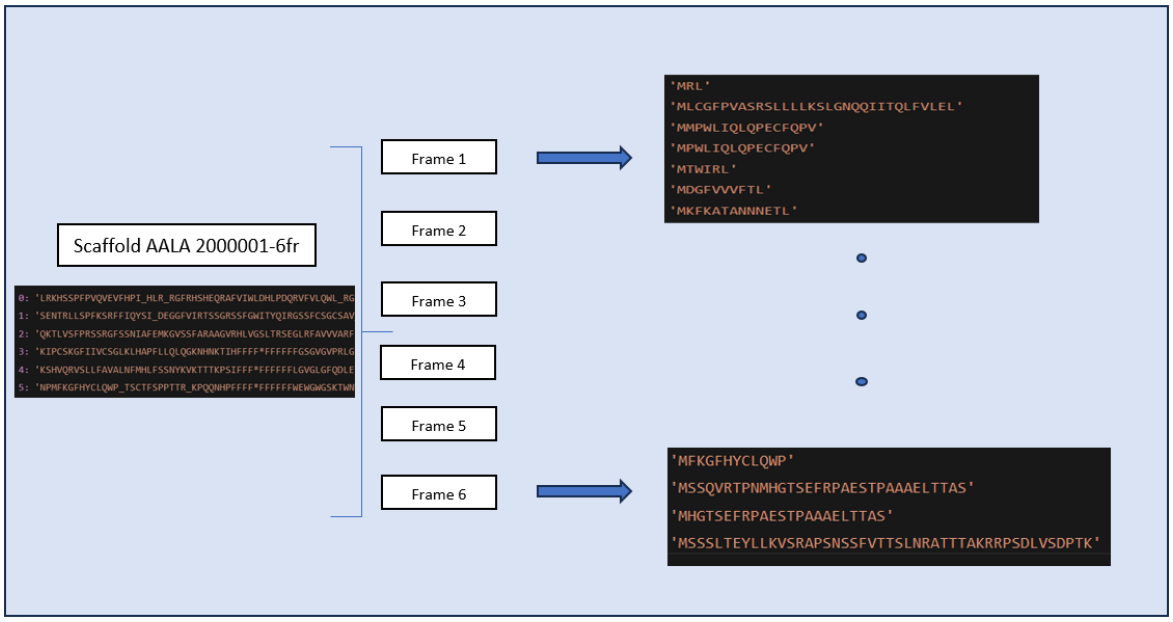

***Figure 16**: Example of potential Open reading Frame finding. For each frame there are several potential Open reading Frames.*

3.2.4 Cysteine spacings

The script inspects all the open reading frames and it finds cysteine spacings with a minimum hold of two cysteines (a filter is implemented to exclude only one cysteine sequences). The logic of the script to find the cysteine spacing is showed in the chart below:

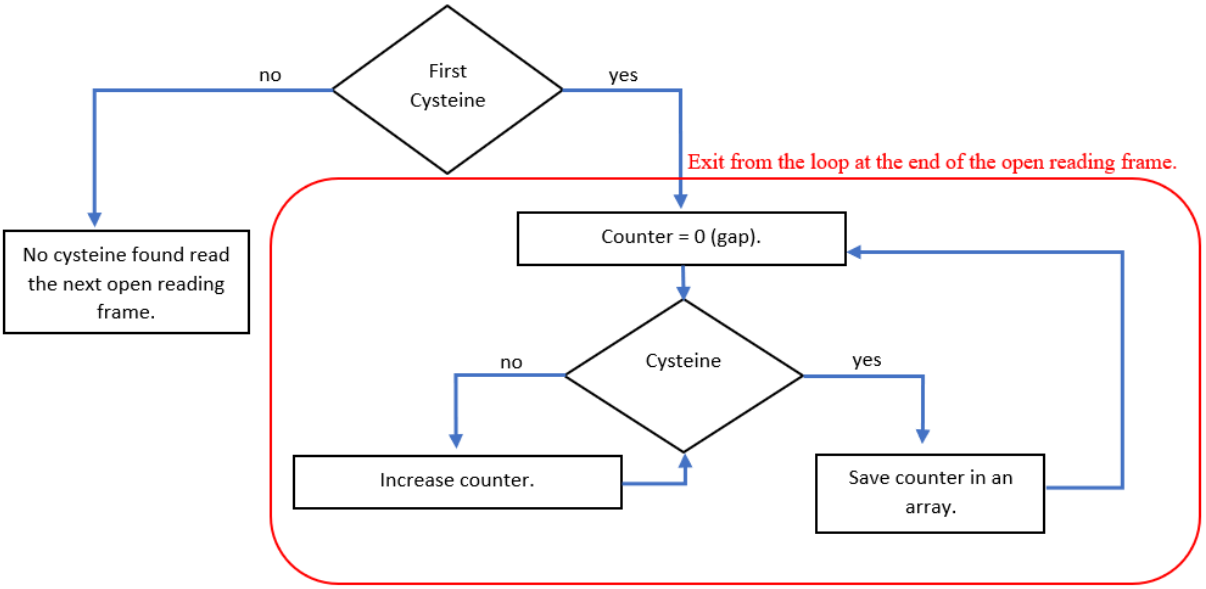

***Figure 17**: Flow chart to find cysteine spacing.*

Every peptide sequence is saved with several metadata (ID sequence, spacing, frame, amino acid sequence, potential ORF) in an excel file named with the 4 letter ID of the reference sample. (Fig. 18)

	A	B	C	D	E
1	ID Sequence	Spacing	Frame	aa seq	pot.ORF
2	scaffold-AALA-2008693-	[0, 2, 1, 7, 1, 6, 3, 2, 1, 7, 1, 6, 3]	4	MALKLNFLVIVALFATSVLTVNARLDLSLLLGSFGIESYDSMARGNQNVKLPNEGKNCCDACLCTRSIPPQCRCADIKNYCPSSCTSCVCTRSIPPQCRCTDIKAYCDPVCT	1
3	scaffold-AALA-2008693-	[0, 2, 1, 7, 1, 6, 3, 2, 1, 7, 1, 6, 3]	4	MARGNQNVKLPNEGKNCCDACLCTRSIPPQCRCADIKNYCPSSCTSCVCTRSIPPQCRCTDIKAYCDPVCT	2
4	scaffold-ACYX-2004465-	[0, 2, 1, 7, 1, 6, 3, 2, 1, 7, 1, 6, 3]	2	MAQKMVVSKSTVVLAVLMMLAIVAATSSLTMAHHEACCNGCVCTKSIPPQCRCADIKDHCHSECKSCLCTKSLPPQCRCADVKDFCYPKC	1
5	scaffold-ACYX-2004465-	[0, 2, 1, 7, 1, 6, 3, 2, 1, 7, 1, 6, 3]	2	MVVSKSTVVLAVLMMLAIVAATSSLTMAHHEACCNGCVCTKSIPPQCRCADIKDHCHSECKSCLCTKSLPPQCRCADVKDFCYPKC	2
6	scaffold-ACYX-2004465-	[0, 2, 1, 7, 1, 6, 3, 2, 1, 7, 1, 6, 3]	2	MMLAIVAATSSLTMAHHEACCNGCVCTKSIPPQCRCADIKDHCHSECKSCLCTKSLPPQCRCADVKDFCYPKC	3
7	scaffold-ACYX-2004465-	[0, 2, 1, 7, 1, 6, 3, 2, 1, 7, 1, 6, 3]	2	MLAIVAATSSLTMAHHEACCNGCVCTKSIPPQCRCADIKDHCHSECKSCLCTKSLPPQCRCADVKDFCYPKC	4
8	scaffold-ACYX-2004465-	[0, 2, 1, 7, 1, 6, 3, 2, 1, 7, 1, 6, 3]	2	MAHHEACCNGCVCTKSIPPQCRCADIKDHCHSECKSCLCTKSLPPQCRCADVKDFCYPKC	5
9	scaffold-AKRH-2138412-	[0, 2, 1, 7, 1, 6, 3, 2, 1, 7, 1, 6, 3]	6	MELSMKVLVKVASLLFLLGFTATVVDARFDPSSFITQFLPNAEANNYYVKSTTKACCNSCPCTKSIPPQCRCSDIGETCHSACKTCICTRSIPPQCHCSDITNFCYEPCNSSETEAH	1
10	scaffold-AKRH-2138412-	[0, 2, 1, 7, 1, 6, 3, 2, 1, 7, 1, 6, 3]	6	MKVLVKVASLLFLLGFTATVVDARFDPSSFITQFLPNAEANNYYVKSTTKACCNSCPCTKSIPPQCRCSDIGETCHSACKTCICTRSIPPQCHCSDITNFCYEPCNSSETEAH	2
11	scaffold-AUGV-2049316-	[0, 2, 1, 7, 1, 6, 3, 2, 1, 7, 1, 6, 3]	3	MAGAVKVQVTVLMAVLVMMVVATSVVTVDARLDFSQILGSFGIQTDSYATPEYSGGSACCDMCVCTKSIPPQCRCTDVKTSCYKGCKGCLCTKSWPPQCRCTDIQNFCYKPCT	1
12	scaffold-AUGV-2049316-	[0, 2, 1, 7, 1, 6, 3, 2, 1, 7, 1, 6, 3]	3	MAVLVMMVVATSVVTVDARLDFSQILGSFGIQTDSYATPEYSGGSACCDMCVCTKSIPPQCRCTDVKTSCYKGCKGCLCTKSWPPQCRCTDIQNFCYKPCTTTTEN	2
13	scaffold-AUGV-2049316-	[0, 2, 1, 7, 1, 6, 3, 2, 1, 7, 1, 6, 3]	3	MMVVATSVVTVDARLDFSQILGSFGIQTDSYATPEYSGGSACCDMCVCTKSIPPQCRCTDVKTSCYKGCKGCLCTKSWPPQCRCTDIQNFCYKPCTTTTEN	3
14	scaffold-AUGV-2049316-	[0, 2, 1, 7, 1, 6, 3, 2, 1, 7, 1, 6, 3]	3	MVVATSVVTVDARLDFSQILGSFGIQTDSYATPEYSGGSACCDMCVCTKSIPPQCRCTDVKTSCYKGCKGCLCTKSWPPQCRCTDIQNFCYKPCTTTTEN	4
15	scaffold-BCGB-2001848-	[0, 2, 1, 7, 1, 6, 3, 2, 1, 7, 1, 6, 3]	4	MAMKVFNVLLVAFLMAFLATSSLADDPFVIYSGDGMMAKGPCCDMCVCTRSIPPQCRCTDIKSYCHSSCKSCRCTKSIPPQCQCNDIKDYCDDSCSPHK	1
16	scaffold-BCGB-2001848-	[0, 2, 1, 7, 1, 6, 3, 2, 1, 7, 1, 6, 3]	4	MKVFNVLLVAFLMAFLATSSLADDPFVIYSGDGMMAKGPCCDMCVCTRSIPPQCRCTDIKSYCHSSCKSCRCTKSIPPQCQCNDIKDYCDDSCSPHK	2
17	scaffold-BCGB-2001848-	[0, 2, 1, 7, 1, 6, 3, 2, 1, 7, 1, 6, 3]	4	MAFLATSSLADDPFVIYSGDGMMAKGPCCDMCVCTRSIPPQCRCTDIKSYCHSSCKSCRCTKSIPPQCQCNDIKDYCDDSCSPHK	3
18	scaffold-BCGB-2001848-	[0, 2, 1, 7, 1, 6, 3, 2, 1, 7, 1, 6, 3]	4	MMAKGPCCDMCVCTRSIPPQCRCTDIKSYCHSSCKSCRCTKSIPPQCQCNDIKDYCDDSCSPHK	4
19	scaffold-BCGB-2001848-	[0, 2, 1, 7, 1, 6, 3, 2, 1, 7, 1, 6, 3]	4	MAKGPCCDMCVCTRSIPPQCRCTDIKSYCHSSCKSCRCTKSIPPQCQCNDIKDYCDDSCSPHK	5
20	scaffold-BEKN-2021264-	[0, 2, 1, 7, 1, 6, 3, 2, 1, 7, 1, 6, 3]	4	MIPQTSSKLAVVILTLFLFVVLVAATTSSWSTTRITDKTEETLKYSFSSIHNFIGVNKIDDRNSNLLQDDQSSSKDLQMGTTACCDECICTRSNPPKCQCDDIKPQCHANCKLCRCL	1

***Figure 18**: Example of potential Open reading Frame finding. For each frame there are several potential Open reading Frame*

A database of 13,4 Gigabyte is built from python scripts. The time required to run all the scripts is around 4 days, never stopping the machine. The technical settings of the workstation are CPU of 12th Gen Intel(R) Core (TM) i5-12400 2.50 GHz, 16Gb of RAM and OS Windows 64 bit.

3.2.5 Filtering by known cysteine spacings and by longest ORF

Thanks to Bernhard Retzl's thesis ("Analysis of cysteine-rich peptides from plants using mass spectrometry") we have a list of potential cysteine spacings for each peptide family. This list was created by a BLAST search, some known cysteine spacings have been included from experimental findings. (Table A)

	Spacing from Retzl's thesis	Spacing from papers
BOWMAN BIRK	[0, 2, 1, 9, 1, 6, 3, 2, 1, 7, 1, 7, 4]'	[0, 2, 1, 9, 1, 6, 3, 2, 1, 7, 1, 7, 4]'
	[0, 2, 1, 7, 1, 7, 3, 2, 1, 7, 1, 6, 3]'	[0, 2, 1, 7, 1, 7, 3, 2, 1, 7, 1, 6, 3]'
		[0, 4, 7, 1, 6 ,3, 2, 11]'
		[0, 4 ,7, 1, 6, 3, 2, 10, 11]'
		[0, 4, 7, 1, 6, 3 ,2, 12, 11]'
		[0, 4, 7, 1, 6, 3, 2, 9, 11]'
		[0, 4, 7, 1, 6, 3, 2, 11, 11]'
CYCLOTIDES		
	'[3, 4, 7, 1, 4]',	'[3, 4, 7, 1, 4]'
	'[3, 4, 6, 1, 4]',	'[3, 4, 6, 1, 4]'
	'[3, 4, 4, 1, 4]',	'[2, 4, 6, 1, 4]'
	'[3, 4, 4, 1, 5]',	'[3, 4, 4, 1, 4]'
	'[3, 4, 5, 1, 4]',	'[3, 4, 4, 1, 5]'
	'[3, 4, 3, 1, 5]',	
	'[3, 4, 4, 1, 6]',	
	'[3, 4, 5, 1, 5]',	
	'[3, 5, 4, 1, 4]',	
	'[3, 4, 6, 1, 5]',	
	'[3, 5, 7, 1, 7]',	
	'[3, 5, 6, 1, 4]',	
	'[3, 5, 4, 1, 5]',	
	'[3, 4, 6, 1, 6]',	
	'[3, 8, 4, 1, 6]',	
	'[3, 3, 6, 1, 4]'	

1

	Spacing from Retzl's thesis	Spacing from papers
DEFENSINS	[8, 4, 3, 9, 4, 1, 3]',	'[10, 5, 3, 9, 6, 1, 3]',
	[8, 5, 3, 9, 6, 1, 3]'	'[10, 5, 3, 9, 7, 1, 3]',
	[8, 4, 3, 9, 9, 1, 3]',	'[10, 5, 3, 9, 8, 1, 3]',
	[8, 4, 3, 9, 10, 1, 3]',	'[10, 5, 3, 10, 6, 1, 3]',
	[9, 4, 3, 9, 4, 1, 3]',	'[10, 5, 3, 10, 7, 1, 3]',
	[9, 4, 3, 10, 5, 0, 0]',	'[10, 5, 3, 10, 8, 1, 3]',
	[9, 4, 3, 11, 5, 0, 0]',	
	[10, 5, 3, 9, 4, 1, 3]',	
	[10, 5, 3, 10, 4, 1, 3]',	
	[10, 5, 3, 10, 6, 1, 3]',	
	[10, 5, 3, 9, 6, 1, 3]',	
	[10, 5, 3, 9, 7, 1, 3]',	
	[10, 5, 3, 9, 8, 1, 3]',	
	[10, 5, 3, 10, 8, 1, 3]',	
	[10, 5, 3, 10, 9, 1, 3]',	
	[10, 5, 3, 9, 9, 1, 3]',	
	[10, 5, 3, 10, 8, 1, 4]',	
	[10, 5, 3, 10, 7, 1, 3]',	
	[10, 4, 3, 10, 8, 1, 3]',	
	[10, 5, 3, 9, 11, 1, 3]',	
	[10, 5, 3, 9, 5, 1, 3]',	
	[10, 5, 3, 9, 6, 1, 2]',	
	[10, 8, 3, 10, 4, 1, 3]',	
	[11, 4, 3, 12, 5, 1, 3]',	
	[11, 6, 3, 10, 4, 1, 4]',	
	[11, 5, 3, 9, 8, 1, 3]',	
NONSPECIFICLIPIDTRANSFER	[9, 19, 0, 13, 1, 23, 9]	[9, 19, 0, 13, 1, 23, 9]
SALMDM	[10, 22, 6, 88, 60, 2, 53, 26, 11, 16, 123, 25, 54]'	'[28, 10, 12, 70, 49, 19, 22, 20, 62, 25, 7, 0, 5]'
		'[28, 6, 3, 6, 5, 70, 16, 32, 19, 43, 62, 29, 0, 5, 0]'
		[10, 22, 6, 88, 60, 2, 53, 26, 11, 16, 123, 25, 54]'

2

	Spacing from Retzl's thesis	Spacing from papers
HEVEINS	[6, 5, 0, 4, 4, 3, 6]',	'[8, 4, 0, 5, 6, 5, 3]',
	[5, 5, 0, 5, 6, 3, 4]',	'[8, 2, 1, 0, 5, 6, 3, 3, 2]',
	[4, 4, 0, 5, 6, 3, 4]',	'[8, 2, 1, 0, 5, 6, 4, 3, 2]',
	[1, 4, 0, 5, 6, 3, 4]',	'[8, 4, 0, 5, 3, 2, 3, 3, 3]',
	[8, 4, 0, 5, 6, 3, 3]',	'[8, 4, 0, 5, 6, 0, 4, 3, 2]',
	[8, 4, 0, 5, 6, 3, 4]',	'[8, 4, 0, 5, 6, 3, 3]',
	[7, 4, 0, 5, 6, 4, 3]'	'[8, 4, 0, 5, 6, 4, 3]',
		'[7, 4, 0, 5, 6, 4, 3]',
		'[4, 4, 0, 5, 6]',
		'[5, 4, 0, 5, 6]',
		'[6, 4, 0, 5, 6]',
KAZAL	[12, 3, 6, 8, 1, 4, 10]',	'[3, 6, 10, 6, 12]',
	[11, 3, 6, 8, 1, 4, 10]',	'[3, 7, 10, 7, 12]',
	[10, 3, 6, 8, 1, 4, 10]',	'[3, 8, 10, 6, 12]',
	[9, 3, 6, 8, 1, 4, 10]',	'[3, 7, 10, 6, 14]',
	[8, 3, 6, 8, 1, 4, 10]',	'[3, 7, 10, 7, 12]',
	[7, 3, 6, 8, 1, 4, 10]'	'[3, 7, 10, 6, 13]',
		'[3, 9, 14, 3, 11]',
		'[5, 7, 10, 6, 12]',
		'[5, 7, 10, 8, 11]',
		'[5, 7, 10, 3, 12]',
		'[5, 7, 10, 3, 10]',
		'[6, 7, 10, 3, 10]',
		'[1, 7, 10, 3, 16]',
		'[1, 7, 10, 3, 15]',
		'[1, 7, 10, 3, 10]'

3

	Spacing from Retzl's thesis	Spacing from papers
PROTEASE INHIBITOR II	'[3, 7, 11, 1, 0, 5, 11]',	[3, 7, 11, 2, 0, 5, 10]',
	'[3, 8, 11, 1, 0, 5, 11]',	[3, 8, 11, 2, 0, 5, 10]',
	'[3, 8, 11, 1, 0, 5, 10]',	
	'[3, 8, 11, 1, 0, 6, 10]',	
	'[3, 7, 11, 1, 0, 5, 10]',	
	'[3, 7, 11, 1, 0, 6, 11]',	
	'[3, 7, 11, 1, 0, 7, 11]',	
	'[3, 7, 11, 1, 0, 11, 11]'	
SNAKINS	'[3, 3, 8, 3, 2, 0, 2, 1, 11, 1, 15]',	'[3, 3, 8, 3, 2, 0, 2, 1, 11, 1, 12]',
	[3, 3, 8, 3, 2, 0, 2, 1, 11, 1, 14]',	'[3, 3, 8, 3, 2, 0, 2, 1, 11, 2, 12]',
	[3, 3, 8, 3, 2, 0, 2, 1, 11, 1, 13]',	'[3, 3, 8, 3, 2, 0, 2, 1, 12, 1, 12]'
	[3, 3, 8, 3, 2, 0, 2, 1, 11, 1, 12]',	
	[3, 3, 8, 3, 2, 0, 2, 1, 11, 2, 12]',	
	[3, 3, 9, 3, 2, 0, 2, 1, 11, 2, 12]',	
	[3, 3, 8, 3, 2, 0, 2, 1, 11, 2, 11]',	
	[3, 3, 8, 3, 2, 0, 2, 1, 11, 1, 11]',	
	[3, 3, 8, 3, 2, 0, 2, 2, 11, 1, 12]',	
	[3, 3, 7, 3, 2, 0, 2, 2, 11, 1, 12]',	
	[3, 3, 7, 3, 2, 0, 2, 1, 11, 1, 12]',	
	[3, 3, 8, 3, 2, 0, 2, 1, 12, 1, 12]',	
	[3, 3, 8, 3, 2, 0, 2, 1, 12, 1, 13]',	
	[3, 3, 7, 3, 2, 0, 2, 2, 11, 1, 13]'	

4

Table A: ***Table A-1***: *Bowman Birk and Cyclotides spacings [21][22];* ***Table A-2***: *Defensins, Non-specific lipid transfer and S-adenosyl-L-methyltransferase spacings [23][24][25];* ***Table A-3***: *Heveins and Kazal spacings [26][27];* ***Table A-4***: *Protease Inhibitor II and Snakins spacings [28][29].*

Therefore, an own made python script was coded to find several peptides by Retzl and experimental spacings. Next step is to collect them in separated excel file by family name. Furthermore, it is a good bioinformatics practice to choose the longest ORF, thus many potential open reading frames are ruled out.

3.2.6 Truncation of amino acid sequences

The peptides have several portions of interest, my focus is the cysteine motif. I observed that the truncation reduced the number of clusters and increased the average number of sequences in each cluster after clustering step. Consequently, I supposed to consider truncated peptides to improve the clustering analysis because the algorithm exclusively scans variations between extreme cysteine ends. (Fig. 19)

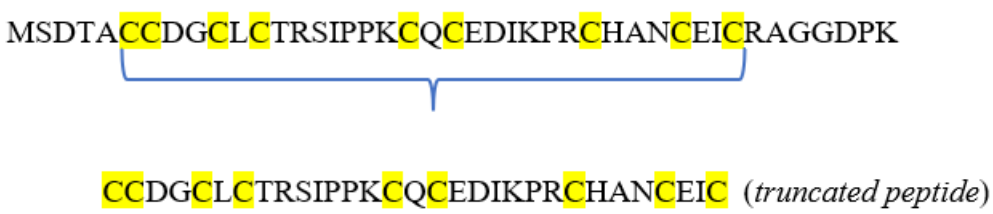

***Figure 19**: Example of a starting complete peptide and his truncated version.*

This procedure is executed for all CRP families.

3.2.7 Clustering

The excel files are converted in FASTA files to be processed by CD-HIT a well-known method to cluster biological sequences. Furthermore, Ubuntu has been used to run CD-HIT. CD-HIT is a clustering algorithm based on the minimum number of identical short substrings, called 'words', such as dipeptides, tripeptides and so on, shared by two proteins. It establishes a similarity threshold by simple word counting. [19] Basically, CD-HIT is a greedy incremental algorithm that starts with the longest input sequence as the first representative cluster, and then process the remaining sequences from long to short to classify each sequence as a redundant or representative sequence based on its similarities to the existing representatives. The clustering algorithm exploits a parallelization process. Given T threads or cores, CD-HIT uses two-word tables and use $T-1$ threads to run multiple checking procedures using one word table, and the remaining thread to run a single clustering procedure using the other table in parallel. [20] CD-HIT parameters are regulated by the user in the command line, I tested several threshold values (c = threshold of similarity) excluding c = 1.0 - 0.9 and c < 0.7. Values below 7 are not suggested by the user manual but at the same time a high similarity between sequences is what I would like. On the other hand, many of the clusters are composed by one single peptide for c = 1 and c = 0.9. The idea behind the choice of the threshold is to reduce number of clusters and at the same time increase the average number of sequences in each cluster. Best results are reached by c values of 0.7 for all families, while c = 0.75 for S-adenosyl-L-methionine-dependent methyltransferase (Table B).

CRP family	similarity threshold = c	number of clusters	average number of sequence in each cluster
Truncated peptides			
BOWMAN BIRK	0.8/0.7	47/22	3.0/6
CYCLOTIDES	0.8/0.7	54/46	1.6/2
DEFENSINS	0.8/0.7	616/327	4.0/7
HEVEINS	0.8/0.7	91/52	4/7.6
KAZAL	0.8/0.7	118/48	3.0/8
NON SPECIFIC LIPID TRANSFER	0.8/0.7	39/14	6/17.6
PROTEASE INHIBITOR II	0.8/0.7	115/75	2.0/3
SALMdM	0.8/0.75	36/5	18/129
SNAKINS	0.8/0.7	671/179	10.0/37

Table B: *Testing of several threshold values against number of clusters and average number of contigs.*

From every CD-HIT run I got two types of files:

1. CRP FAMILY NAME.clstr
2. CRP FAMILY NAME (no extension but opened as txt file)

The first file is the list of peptides grouped in several clusters. The second file is the list of the representative sequences associated to a cluster.

3.2.8 Removing/Checking unknown amino acids

In several FASTA files downloaded from the 1kp-database I encountered some peptides that have unknown bases (N). During six-frame translation non translatable triplets have been translated with " * " symbol to avoid misunderstanding with asparagine (N). Specifically, the removal of unknown amino acids before clustering do not affect the number of elements per cluster in a relevant way. Therefore, before clustering the dataset is filtered by a python script to remove peptides with undetermined triplets. This step is fundamental for future computational analyses, because a complete known amino acid sequence is required as input.

3.2.9 Ranking the clusters

Every cluster has one representative peptide so the criteria to rank all these clusters/peptides is to arrange them from the more representative (highest number of 4 letter IDs) to the less representative.

K	L	M	N	O	P
	>Cluster 5		>Cluster 6		>Cluster 7
AKRH-	1	AUGV-	1	BEKN-	5
BKQU-	1	ERXG-	1	BMRX-	1
CMFF-	1	JTQQ-	1	EPRK-	1
FPLR-	1	KNMB-	1	FNXH-	1
HJMP-	1	MAQO-	1	FPYZ-	1
JTQQ-	1	NJKC-	1	GMAM-	1
KEGA-	1	OBPL-	1	IORZ-	4
LQUX-	1	TUHA-	1	JSVC-	1
LXGM-	1	UDHA-	2	KKCW-	1
MYMP-	1	WAIL-	1	QCOU-	2
OAQS-	1	YZRI-	1	RQNK-	2
PEZP-	1	ZGQD-	1	SSDU-	1
PFSA-	1			STDO-	1
QKMG-	1			TMWO-	1
QTJY-	1			ZSNV-	1
RIDD-	1				24
RRID-	1				
SSDU-	1				
TJMB-	1				
TMWO-	1				
ZUQW-	1				
	21				

***Table C**: (Bowman Birk c=0.7) 4letter ID found in a cluster (K, M, O columns) and number of peptides belonging to the 4 letter ID class (L, N, P columns). i.e. In cluster 7 have been found 5 peptides belonging to BEKN.*

In table C the most representative cluster/sequence is Cluster 5 because it has the highest number of 4 letter IDs and Cluster 6 is the less representative. Another criterion of ranking, which is not examined in this thesis, is to choose the cluster/sequence with more contigs such as Cluster 7 (24 contigs signed in RED).
However, the ranking score for Table C is cluster 5 with 21 4letter IDs > cluster 7 with 15 4letter IDs > cluster 6 with 12 4letter IDs > …

According to this ranking criterion all the representative sequences/clusters are arranged in a descending way.

Chapter 4

Results

Here are reported all the representative sequences for each CRP family found by CD-HIT and ranked with the criteria explained in the 3.2.9 paragraph. The sorted number of 4 letter IDs is approximatively the number of species found and the sorted cluster is the respective cluster. In the third column are showed all the representative sequence identifiers. The number of species could change since some transcriptomes of the same species share the same 4 letter ID.

4.1 Bowman Birk Inhibitors

sorted number of 4 letter IDs	sorted cluster	
21	5	054aa, >scaffold-AKRH-2138412-6... *
15	7	054aa, >scaffold-BEKN-2021264-4... *
14	4	054aa, >scaffold-ACYX-2004465-2... *
13	1	655aa, >scaffold-JETM-2087744-4... *
12	6	054aa, >scaffold-AUGV-2049316-3... *
7	12	054aa, >scaffold-KAYP-2041321-1... *
5	18	054aa, >scaffold-QKMG-2089575-6... *
4	0	055aa, >scaffold-BFMT-2068635-1... *
4	15	054aa, >scaffold-NAUM-2049264-4... *
3	8	054aa, >scaffold-COCP-2082934-1... *
2	2	055aa, >scaffold-JTQQ-2010870-6... *
2	17	054aa, >scaffold-PAWA-2006115-5... *
1	3	055aa, >scaffold-QZXQ-2087260-6... *
1	9	054aa, >scaffold-DWZT-2007499-6... *
1	10	054aa, >scaffold-JETM-2088802-5... *
1	11	054aa, >scaffold-JTQQ-2009529-1... *
1	13	054aa, >scaffold-KNMB-2048205-3... *
1	14	054aa, >scaffold-MFIN-2001626-1... *
1	16	054aa, >scaffold-NPND-2011619-2... *
1	19	054aa, >scaffold-RKLL-2061163-3... *
1	20	054aa, >scaffold-VYLQ-2000244-1... *
1	21	053aa, >scaffold-KNMB-2052248-1... *

***Table D**: 22 bowman birk inhibitors representative sequences/clusters*

4.2 Cyclotides

sorted number of 4 letter IDs	sorted cluster	
11	0	125aa, >scaffold-ACYX-2102426-5... *
6	35	022aa, >scaffold-BJSW-2010321-5... *
3	14	024aa, >scaffold-LPGY-2021620-3... *
3	25	023aa, >scaffold-OMYK-2019638-6... *
3	36	022aa, >scaffold-BQEQ-2009543-3... *
1	1	025aa, >scaffold-AJFN-2012592-5... *
1	2	025aa, >scaffold-GGWH-2012726-3... *
1	3	025aa, >scaffold-JKAA-2182112-6... *
1	4	025aa, >scaffold-KYAD-2012510-3... *
1	5	025aa, >scaffold-NJLF-2054764-1... *
1	6	024aa, >scaffold-BXAY-2000185-5... *
1	7	024aa, >scaffold-FFFY-2049412-5... *
1	8	024aa, >scaffold-FPLR-2016484-2... *
1	9	024aa, >scaffold-GJIY-2007815-5... *
1	10	024aa, >scaffold-LPGY-2007253-1... *
1	11	024aa, >scaffold-LPGY-2019978-6... *
1	12	024aa, >scaffold-LPGY-2021147-3... *
1	13	024aa, >scaffold-LPGY-2021245-3... *
1	15	024aa, >scaffold-LPGY-2031019-6... *
1	16	024aa, >scaffold-LPGY-2150203-5... *
1	17	024aa, >scaffold-LPGY-2151832-1... *
1	18	024aa, >scaffold-LPGY-2153004-5... *
1	19	024aa, >scaffold-NJLF-2011312-4... *
1	20	024aa, >scaffold-NJLF-2048529-2... *
1	21	024aa, >scaffold-VFIV-2008184-3... *
1	22	024aa, >scaffold-WHNV-2044183-4... *
1	23	023aa, >scaffold-EHNF-2017859-6... *
1	24	023aa, >scaffold-ETCJ-2050173-1... *
1	26	023aa, >scaffold-LPGY-2005594-2... *
1	27	023aa, >scaffold-LPGY-2032436-5... *
1	28	023aa, >scaffold-NJLF-2005207-5... *
1	29	023aa, >scaffold-PZIF-2012589-3... *
1	30	023aa, >scaffold-QFAE-2053804-1... *
1	31	023aa, >scaffold-SDPC-2012617-6... *
1	32	023aa, >scaffold-SWOH-2008016-5... *
1	33	023aa, >scaffold-VHZV-2023286-4... *
1	34	023aa, >scaffold-WWSS-2006665-2... *
1	37	022aa, >scaffold-DNQA-2035222-3... *
1	38	022aa, >scaffold-GGEA-2068232-2... *
1	39	022aa, >scaffold-LPGY-2020935-3... *
1	40	022aa, >scaffold-MUMD-2011088-3... *
1	41	022aa, >scaffold-NJLF-2052032-6... *
1	42	022aa, >scaffold-NJLF-2052034-5... *
1	43	022aa, >scaffold-OQWW-2188878-4... *
1	44	022aa, >scaffold-QRTH-2043026-3... *
1	45	022aa, >scaffold-XPAF-2051608-3... *

***Table E**: 46 cyclotides representative sequences/clusters.*

4.3 Defensins

sorted number of 4 letter IDs	sorted cluster		
369	112		4446aa, >scaffold-CFRN-2093739-3... *
114	137		045aa, >scaffold-ACYX-2012985-6... *
86	114		1546aa, >scaffold-CRNC-2037619-2... *
65	56		047aa, >scaffold-ACWS-2061877-6... *
65	143		045aa, >scaffold-ATQZ-2004205-5... *
55	82		1447aa, >scaffold-GHLP-2001283-2... *
45	142		045aa, >scaffold-AKTA-2039636-4... *
43	83		1647aa, >scaffold-GNQG-2013002-5... *
41	15		048aa, >scaffold-BMIF-2078740-3... *
38	153		045aa, >scaffold-BPKH-2074100-2... *
36	59		047aa, >scaffold-AXBO-2099710-6... *
35	148		045aa, >scaffold-BERS-2005621-2... *
35	150		045aa, >scaffold-BJKT-2059237-1... *
28	171		045aa, >scaffold-DKFZ-2058174-5... *
27	131		2746aa, >scaffold-WGUG-2024812-1... *
23	8		048aa, >scaffold-AFQQ-2046446-2... *
22	136		045aa, >scaffold-AAXJ-2049721-4... *
20	141		045aa, >scaffold-AJBK-2061271-6... *
19	23		448aa, >scaffold-DGXS-2074475-1... *
19	189		045aa, >scaffold-FZQN-2008107-5... *
17	68		647aa, >scaffold-BYNZ-2076556-2... *
17	135		045aa, >scaffold-AAXJ-2007413-5... *
15	11		048aa, >scaffold-AXNH-2003537-6... *
15	17		048aa, >scaffold-BXBF-2009936-5... *
15	58		047aa, >scaffold-AWQB-2004681-2... *
15	174		045aa, >scaffold-DPFW-2013549-5... *
13	2		249aa, >scaffold-CWYJ-2062011-5... *
13	5		049aa, >scaffold-GSXD-2005106-1... *
13	32		048aa, >scaffold-GSXD-2007422-3... *
13	226		045aa, >scaffold-PCGJ-2059765-6... *
12	0		050aa, >scaffold-BLWH-2100768-1... *
12	45		848aa, >scaffold-TXMP-2036510-4... *
12	57		047aa, >scaffold-AUDE-2046467-1... *
11	12		048aa, >scaffold-BBDD-2065715-4... *
11	181		045aa, >scaffold-EDIT-2068313-1... *
11	324		039aa, >scaffold-AZBL-2019902-4... *
10	1		049aa, >scaffold-CJNT-2067964-2... *
10	14		048aa, >scaffold-BLWH-2019943-3... *
10	67		047aa, >scaffold-BXAY-2081721-1... *
9	172		045aa, >scaffold-DLJZ-2041208-6... *
8	4		149aa, >scaffold-EWXK-2118042-5... *
8	28		048aa, >scaffold-ERIA-2003386-4... *
8	62		047aa, >scaffold-BRUD-2059107-2... *
8	72		047aa, >scaffold-CWZU-2035775-5... *
8	129		846aa, >scaffold-TJLC-2006914-5... *
8	138		045aa, >scaffold-AEPI-2043385-2... *
8	149		045aa, >scaffold-BJKT-2000021-6... *
8	151		045aa, >scaffold-BOLZ-2166973-1... *
8	155		045aa, >scaffold-BSEY-2011372-4... *
8	158		045aa, >scaffold-BWRK-2053892-2... *

3	263		044aa, >scaffold-EVOD-2007504-5... *
3	267		044aa, >scaffold-GIPR-2004300-3... *
3	305		041aa, >scaffold-FEDW-2069281-5... *
3	320		040aa, >scaffold-SLYR-2091888-5... *
2	9		048aa, >scaffold-ALVQ-2014874-3... *
2	13		048aa, >scaffold-BEGM-2091524-3... *
2	24		048aa, >scaffold-DJSE-2016420-3... *
2	25		048aa, >scaffold-DJSE-2131040-6... *
2	37		048aa, >scaffold-MHYG-2070953-2... *
2	61		047aa, >scaffold-BKQU-2013741-2... *
2	73		047aa, >scaffold-DESP-2010575-5... *
2	74		047aa, >scaffold-DESP-2100202-6... *
2	77		047aa, >scaffold-EFCZ-2217690-6... *
2	79		047aa, >scaffold-ERXG-2006493-6... *
2	80		047aa, >scaffold-EVOD-2113470-2... *
2	85		047aa, >scaffold-HRUR-2029124-6... *
2	89		047aa, >scaffold-KPTE-2009602-3... *
2	106		047aa, >scaffold-WBIB-2057675-1... *
2	118		046aa, >scaffold-ERXG-2002466-5... *
2	122		046aa, >scaffold-JHUL-2016579-3... *
2	133		246aa, >scaffold-YGAT-2040522-6... *
2	175		045aa, >scaffold-DTOA-2080977-3... *
2	177		045aa, >scaffold-DYFF-2025170-2... *
2	183		045aa, >scaffold-ERXG-2050685-5... *
2	184		045aa, >scaffold-FNEN-2070005-1... *
2	186		045aa, >scaffold-FROP-2000963-5... *
2	201		045aa, >scaffold-JNVS-2041374-5... *
2	203		045aa, >scaffold-KEGA-2330946-2... *
2	210		045aa, >scaffold-LTZF-2075501-5... *
2	215		045aa, >scaffold-MUNP-2007151-2... *
2	216		045aa, >scaffold-MYVH-2000678-3... *
2	221		045aa, >scaffold-OAGK-2045705-5... *
2	225		045aa, >scaffold-PAWA-2043518-6... *
2	232		045aa, >scaffold-QOXT-2029408-6... *
2	233		045aa, >scaffold-REOF-2015342-5... *
2	240		045aa, >scaffold-UFQC-2042721-6... *
2	241		045aa, >scaffold-UFQC-2047256-1... *
2	247		045aa, >scaffold-VXKB-2003200-1... *
2	249		045aa, >scaffold-VXKB-2049047-4... *
2	268		044aa, >scaffold-KKDQ-2070894-2... *
2	269		044aa, >scaffold-KNMB-2052556-6... *
2	286		043aa, >scaffold-CNTZ-2038740-6... *
2	288		043aa, >scaffold-DMIN-2109286-2... *
2	289		043aa, >scaffold-DYFF-2263154-2... *
2	291		043aa, >scaffold-LELS-2080870-2... *
2	309		040aa, >scaffold-EBOL-2050922-6... *
2	325		039aa, >scaffold-FYSJ-2017338-6... *
1	7		049aa, >scaffold-ZSAB-2185500-1... *
1	10		048aa, >scaffold-AQFM-2075232-3... *
1	18		048aa, >scaffold-CQPW-2020392-6... *
1	19		048aa, >scaffold-CUTE-2000967-3... *

8	287	043aa, >scaffold-DESP-2096669-2... *
7	179	045aa, >scaffold-DYFF-2266669-4... *
7	284	043aa, >scaffold-BERS-2010442-6... *
6	20	048aa, >scaffold-DCDT-2067667-1... *
6	39	548aa, >scaffold-NNOK-2045122-1... *
6	139	045aa, >scaffold-AFLV-2013147-4... *
6	144	045aa, >scaffold-ATQZ-2092700-5... *
6	165	045aa, >scaffold-CPOC-2002162-5... *
6	208	045aa, >scaffold-LQJY-2070091-6... *
5	3	049aa, >scaffold-ERXG-2052896-1... *
5	27	048aa, >scaffold-EMIG-2002997-1... *
5	51	048aa, >scaffold-XDDT-2051749-5... *
5	90	347aa, >scaffold-LSKK-2016682-6... *
5	109	046aa, >scaffold-AIOU-2050026-4... *
5	120	146aa, >scaffold-HBHB-2204474-3... *
5	140	045aa, >scaffold-AIOU-2006298-6... *
5	147	045aa, >scaffold-BEKN-2277059-3... *
5	170	045aa, >scaffold-DCCI-2006170-4... *
5	192	045aa, >scaffold-GNPX-2084523-6... *
4	6	049aa, >scaffold-LHSH-2063569-6... *
4	21	048aa, >scaffold-DESP-2103610-3... *
4	22	048aa, >scaffold-DFHO-2048686-5... *
4	41	048aa, >scaffold-QIAD-2055001-3... *
4	46	048aa, >scaffold-UPOG-2007733-6... *
4	70	047aa, >scaffold-CUVY-2002597-3... *
4	111	046aa, >scaffold-BHYC-2042508-4... *
4	146	045aa, >scaffold-BEKN-2224644-6... *
4	160	045aa, >scaffold-CCID-2009286-4... *
4	169	045aa, >scaffold-CWYJ-2050602-1... *
4	202	045aa, >scaffold-KDCH-2012461-2... *
4	214	045aa, >scaffold-MTII-2037558-6... *
4	217	045aa, >scaffold-NJJO-2002511-5... *
4	313	040aa, >scaffold-KAOV-2058588-1... *
3	16	048aa, >scaffold-BQEQ-2053891-5... *
3	34	048aa, >scaffold-KTQI-2118320-3... *
3	71	047aa, >scaffold-CWLL-2139108-3... *
3	76	047aa, >scaffold-DZQM-2043424-3... *
3	86	047aa, >scaffold-IDGE-2101801-5... *
3	92	047aa, >scaffold-OMYK-2141620-5... *
3	110	046aa, >scaffold-AYMT-2000876-3... *
3	168	045aa, >scaffold-CWLL-2014218-1... *
3	173	045aa, >scaffold-DMIN-2099765-4... *
3	180	045aa, >scaffold-EBOL-2048208-4... *
3	182	045aa, >scaffold-EFCZ-2213718-6... *
3	187	045aa, >scaffold-FUPX-2053665-1... *
3	188	045aa, >scaffold-FXGI-2044281-2... *
3	197	045aa, >scaffold-IQYY-2105820-5... *
3	206	045aa, >scaffold-LDEL-2060889-4... *
3	207	045aa, >scaffold-LELS-2086313-3... *
3	229	045aa, >scaffold-PZRT-2053737-2... *
3	260	044aa, >scaffold-DBYD-2047926-5... *

1	117		046aa, >scaffold-ERIA-2053749-5... *
1	119		046aa, >scaffold-FXGI-2042124-4... *
1	121		046aa, >scaffold-HJMP-2003490-1... *
1	123		046aa, >scaffold-JHUL-2017272-2... *
1	124		046aa, >scaffold-KNMB-2005845-1... *
1	125		046aa, >scaffold-LVUS-2047876-1... *
1	126		046aa, >scaffold-OLXF-2082744-3... *
1	127		046aa, >scaffold-PTBJ-2096308-3... *
1	128		046aa, >scaffold-RSPO-2092501-3... *
1	130		046aa, >scaffold-TPEM-2134099-3... *
1	132		046aa, >scaffold-YADI-2057853-3... *
1	134		046aa, >scaffold-YXNR-2032147-2... *
1	145		045aa, >scaffold-AXBO-2088230-2... *
1	152		045aa, >scaffold-BOLZ-2167061-1... *
1	154		045aa, >scaffold-BQEQ-2001492-6... *
1	156		045aa, >scaffold-BSVG-2060379-3... *
1	157		045aa, >scaffold-BSVG-2076650-1... *
1	159		045aa, >scaffold-BXAY-2006555-2... *
1	161		045aa, >scaffold-COCP-2014634-4... *
1	162		045aa, >scaffold-COCP-2075998-4... *
1	163		045aa, >scaffold-COCP-2081516-6... *
1	164		045aa, >scaffold-CPKP-2011507-2... *
1	166		045aa, >scaffold-CPOC-2018402-6... *
1	167		045aa, >scaffold-CRNC-2031445-3... *
1	176		045aa, >scaffold-DWZT-2004984-3... *
1	178		045aa, >scaffold-DYFF-2207750-1... *
1	185		045aa, >scaffold-FNEN-2070094-6... *
1	190		045aa, >scaffold-GCFE-2045248-2... *
1	191		045aa, >scaffold-GJNX-2004133-3... *
1	193		045aa, >scaffold-HANM-2060594-4... *
1	194		045aa, >scaffold-HRUR-2022450-2... *
1	195		045aa, >scaffold-HRUR-2129796-3... *
1	196		045aa, >scaffold-HRUR-2135131-6... *
1	198		045aa, >scaffold-IUSR-2011861-6... *
1	199		045aa, >scaffold-IXVJ-2115757-4... *
1	200		045aa, >scaffold-JDTY-2001739-5... *
1	204		045aa, >scaffold-KVAY-2054581-5... *
1	205		045aa, >scaffold-KWGC-2000190-2... *
1	209		045aa, >scaffold-LRTN-2089587-2... *
1	211		045aa, >scaffold-MFEA-2102262-6... *
1	212		045aa, >scaffold-MFIN-2006430-1... *
1	213		045aa, >scaffold-MGVU-2032263-4... *
1	218		045aa, >scaffold-NMGG-2056412-4... *
1	219		045aa, >scaffold-NPND-2000888-1... *
1	220		045aa, >scaffold-NXTS-2153743-3... *
1	222		045aa, >scaffold-OHAE-2024211-6... *
1	223		045aa, >scaffold-OLES-2132373-3... *
1	224		045aa, >scaffold-OUER-2095886-1... *
1	227		045aa, >scaffold-PKMO-2011230-6... *
1	228		045aa, >scaffold-PZRT-2011087-3... *
1	230		045aa, >scaffold-PZRT-2059851-1... *

1	231		045aa, >scaffold-QOXT-2008537-3... *
1	234		045aa, >scaffold-RMVB-2010831-4... *
1	235		045aa, >scaffold-SKNL-2061403-5... *
1	236		045aa, >scaffold-TOKV-2045919-3... *
1	237		045aa, >scaffold-TVSH-2060787-4... *
1	238		045aa, >scaffold-UDUT-2002172-4... *
1	239		045aa, >scaffold-UDUT-2044075-4... *
1	242		045aa, >scaffold-UOYN-2004285-2... *
1	243		045aa, >scaffold-URDJ-2032640-4... *
1	244		045aa, >scaffold-VGVI-2058465-1... *
1	245		045aa, >scaffold-VKGP-2110411-4... *
1	246		045aa, >scaffold-VMNH-2009577-4... *
1	248		045aa, >scaffold-VXKB-2047610-5... *
1	250		045aa, >scaffold-WFBF-2039236-4... *
1	251		045aa, >scaffold-WPHN-2046552-4... *
1	252		045aa, >scaffold-WTKZ-2056516-3... *
1	253		045aa, >scaffold-XRLM-2066973-6... *
1	254		045aa, >scaffold-XSZI-2008092-2... *
1	255		045aa, >scaffold-YADI-2005813-4... *
1	256		045aa, >scaffold-YGAT-2037987-6... *
1	257		045aa, >scaffold-YMES-2000670-6... *
1	258		045aa, >scaffold-YQIJ-2028805-5... *
1	259		045aa, >scaffold-YRHD-2046149-3... *
1	261		044aa, >scaffold-DUQG-2001009-6... *
1	262		044aa, >scaffold-ETCJ-2052409-5... *
1	264		044aa, >scaffold-FVXD-2039902-5... *
1	265		044aa, >scaffold-FVXD-2054957-2... *
1	266		044aa, >scaffold-FYSJ-2064690-4... *
1	270		044aa, >scaffold-MYMP-2053650-2... *
1	271		044aa, >scaffold-OQHZ-2011651-2... *
1	272		044aa, >scaffold-OQHZ-2066975-3... *
1	273		044aa, >scaffold-QKMG-2094007-2... *
1	274		044aa, >scaffold-SYHW-2000930-5... *
1	275		044aa, >scaffold-SYHW-2001018-2... *
1	276		044aa, >scaffold-SYHW-2001019-3... *
1	277		044aa, >scaffold-VGHH-2038571-5... *
1	278		044aa, >scaffold-VGVI-2002821-5... *
1	279		044aa, >scaffold-WGET-2037594-1... *
1	280		044aa, >scaffold-XGFU-2113237-6... *
1	281		044aa, >scaffold-XKPS-2012135-4... *
1	282		044aa, >scaffold-XKPS-2123484-1... *
1	283		044aa, >scaffold-YHFG-2070328-6... *
1	285		043aa, >scaffold-CCID-2012741-4... *
1	290		043aa, >scaffold-JETM-2080069-6... *
1	292		043aa, >scaffold-MTII-2006467-5... *
1	293		043aa, >scaffold-OAGK-2047640-6... *
1	294		043aa, >scaffold-PZRT-2058760-6... *
1	295		043aa, >scaffold-SMMC-2141468-2... *
1	296		043aa, >scaffold-VGHH-2036874-6... *
1	297		043aa, >scaffold-VMNH-2077876-3... *
1	298		043aa, >scaffold-WOGB-2084406-5... *

***Table F**: 327 defensins representative sequences/clusters.*

4.4 Heveins

sorted number of 4 letter IDs	sorted cluster	
42	16	037aa, >scaffold-AUDE-2002285-6... *
24	17	037aa, >scaffold-BBDD-2004959-3... *
18	42	026aa, >scaffold-AAXJ-2048754-2... *
12	48	025aa, >scaffold-BWRK-2079160-5... *
11	20	037aa, >scaffold-HBGV-2043743-4... *
9	6	138aa, >scaffold-HPXA-2010559-3... *
8	45	626aa, >scaffold-SMMC-2134691-5... *
6	9	238aa, >scaffold-WCOR-2083208-6... *
5	3	440aa, >scaffold-DFHO-2000675-3... *
5	8	438aa, >scaffold-RCBT-2010498-5... *
5	28	036aa, >scaffold-BXAY-2011451-6... *
5	33	031aa, >scaffold-FALI-2015979-3... *
4	2	041aa, >scaffold-BCAA-2064707-1... *
4	23	037aa, >scaffold-LQJY-2068729-1... *
3	4	239aa, >scaffold-IPUI-2019393-1... *
3	7	138aa, >scaffold-PYHZ-2061773-5... *
3	18	037aa, >scaffold-CRNC-2006854-2... *
3	32	031aa, >scaffold-DMIN-2024170-6... *
2	15	038aa, >scaffold-VKGP-2033460-5... *
2	19	037aa, >scaffold-DGNP-2008747-2... *
2	21	037aa, >scaffold-JVSZ-2003894-2... *
2	22	037aa, >scaffold-KLGF-2006233-6... *
2	24	037aa, >scaffold-PKOX-2006241-4... *
2	29	136aa, >scaffold-TAVP-2001196-1... *
2	31	136aa, >scaffold-ZQRI-2011043-3... *
2	35	031aa, >scaffold-GMHZ-2006710-1... *
2	41	026aa, >scaffold-AAXJ-2002216-4... *
2	43	026aa, >scaffold-CBJR-2058893-4... *
2	49	025aa, >scaffold-CBJR-2056119-1... *
1	0	041aa, >scaffold-AHRN-2012568-5... *
1	1	041aa, >scaffold-AHRN-2085500-5... *
1	5	039aa, >scaffold-UNPT-2015295-3... *
1	10	038aa, >scaffold-BLWH-2001234-1... *
1	11	038aa, >scaffold-CVDF-2094468-6... *
1	12	038aa, >scaffold-DLAI-2012635-6... *
1	13	038aa, >scaffold-OFTV-2091707-3... *
1	14	038aa, >scaffold-SNBI-2000174-6... *
1	25	037aa, >scaffold-TFYI-2074209-3... *
1	26	037aa, >scaffold-XSZI-2001702-6... *
1	27	037aa, >scaffold-ZFGK-2009069-1... *
1	30	036aa, >scaffold-XWHK-2009231-6... *
1	34	031aa, >scaffold-FPLR-2026847-4... *
1	36	031aa, >scaffold-KTAR-2024584-3... *
1	37	027aa, >scaffold-IUSR-2022000-2... *
1	38	027aa, >scaffold-MQIV-2012805-4... *
1	39	027aa, >scaffold-OQBM-2001884-3... *
1	40	027aa, >scaffold-QASA-2009698-4... *
1	44	026aa, >scaffold-SJEV-2036206-2... *
1	46	026aa, >scaffold-TXMP-2033532-4... *
1	47	026aa, >scaffold-UFJN-2067125-1... *
1	50	025aa, >scaffold-FGRF-2056300-2... *
1	51	025aa, >scaffold-NFXV-2012474-6... *

Table G: *53 heveins representative sequences/clusters*

4.5 Kazal

sorted number of 4 letter IDs	sorted cluster			
80	2			652aa, >scaffold-BOLZ-2167705-4... *
70	0			052aa, >scaffold-ARYD-2017100-6... *
45	5			1252aa, >scaffold-GIOY-2002983-4... *
35	16			249aa, >scaffold-BTTS-2073571-6... *
27	6			752aa, >scaffold-HWUP-2097887-1... *
26	19			048aa, >scaffold-GCFE-2010871-3... *
12	18			248aa, >scaffold-BNDE-2001017-2... *
8	12			450aa, >scaffold-LWCK-2073273-5... *
6	8			051aa, >scaffold-INQX-2115273-6... *
5	15			350aa, >scaffold-UBLN-2065042-6... *
5	21			148aa, >scaffold-RWKR-2118081-1... *
4	9			251aa, >scaffold-RBYC-2012078-5... *
4	20			048aa, >scaffold-MYVH-2056972-5... *
3	4			152aa, >scaffold-CWYJ-2055320-5... *
3	7			052aa, >scaffold-KKDQ-2071163-2... *
3	10			250aa, >scaffold-HBUQ-2031972-5... *
3	22			248aa, >scaffold-VMNH-2023280-5... *
3	32			042aa, >scaffold-NSTT-2038808-2... *
2	1			052aa, >scaffold-BMSE-2050139-3... *
2	13			050aa, >scaffold-MYVH-2013134-5... *
2	14			050aa, >scaffold-MYVH-2053679-4... *
2	31			045aa, >scaffold-XWDM-2111714-1... *
1	3			052aa, >scaffold-CQPW-2023478-5... *
1	11			050aa, >scaffold-JVSZ-2006638-4... *
1	17			049aa, >scaffold-OSHQ-2041709-1... *
1	23			048aa, >scaffold-XPBC-2000109-3... *
1	24			048aa, >scaffold-YGAT-2041561-6... *
1	25			047aa, >scaffold-ACRY-2038115-4... *
1	26			047aa, >scaffold-XPBC-2038083-5... *
1	27			046aa, >scaffold-MUNP-2011183-4... *
1	28			045aa, >scaffold-ISPU-2043706-4... *
1	29			045aa, >scaffold-JETM-2015422-1... *
1	30			045aa, >scaffold-JRDV-3016420-4... *
1	33			042aa, >scaffold-SVVG-2002646-6... *
1	34			037aa, >scaffold-ANON-2005384-3... *
1	35			037aa, >scaffold-FOYQ-2039651-4... *
1	36			037aa, >scaffold-JKHA-2015964-6... *
1	37			037aa, >scaffold-POOW-2057670-3... *
1	38			037aa, >scaffold-PQED-2047186-5... *
1	39			037aa, >scaffold-QGLJ-2012835-3... *
1	40			037aa, >scaffold-RAWF-2042258-1... *
1	41			037aa, >scaffold-TZJQ-2005496-5... *
1	42			037aa, >scaffold-WAFT-2012566-1... *
1	43			037aa, >scaffold-XDLL-2041951-1... *
1	44			037aa, >scaffold-YADI-2064434-4... *
1	45			037aa, >scaffold-YRMA-2004073-4... *
1	46			037aa, >scaffold-ZGQD-2038529-5... *
1	47			037aa, >scaffold-ZYCD-2033615-6... *

Table H: *48 kazal representative sequences/clusters.*

4.6 Protease Inhibitor II

sorted number of 4 letter IDs	sorted cluster			
32	18			647aa, >scaffold-CUVY-2025464-4... *
18	33			947aa, >scaffold-LTZF-2077095-6... *
15	43			046aa, >scaffold-BFJL-2181487-2... *
15	45			046aa, >scaffold-BZDF-2076110-1... *
9	0			052aa, >scaffold-AQFM-2072131-5... *
8	57			046aa, >scaffold-LVUS-2039717-2... *
6	8			152aa, >scaffold-PLYX-2080049-1... *
4	17			048aa, >scaffold-WWSS-2038077-5... *
4	21			047aa, >scaffold-DHAW-2070715-6... *
4	25			047aa, >scaffold-GSXD-2056708-3... *
4	55			046aa, >scaffold-IHPC-2064610-3... *
4	63			046aa, >scaffold-OLES-2032919-4... *
4	64			046aa, >scaffold-OOVX-2040645-4... *
3	2			052aa, >scaffold-DCDT-2058981-2... *
3	23			047aa, >scaffold-FXGI-2046366-3... *
3	31			147aa, >scaffold-JBLI-2093176-4... *
3	34			047aa, >scaffold-MAQO-2008273-3... *
3	41			047aa, >scaffold-GHLP-2043262-2... *
3	46			046aa, >scaffold-CJNT-2062963-2... *
3	50			046aa, >scaffold-EDIT-2056410-5... *
2	4			052aa, >scaffold-JRNA-2058974-5... *
2	11			052aa, >scaffold-UUJS-2109942-1... *
2	13			048aa, >scaffold-BLAJ-2100765-4... *
2	14			048aa, >scaffold-CIEA-2047175-3... *
2	16			148aa, >scaffold-OOSO-2002945-3... *
2	32			147aa, >scaffold-LELS-2083925-2... *
2	36			147aa, >scaffold-SFKQ-2081018-6... *
2	47			046aa, >scaffold-COBX-2058673-2... *
2	53			146aa, >scaffold-GDZS-2092301-1... *
2	60			046aa, >scaffold-NXTS-2041157-5... *
2	61			046aa, >scaffold-NXTS-2144708-2... *
2	68			046aa, >scaffold-SCAO-2161377-5... *
2	74			146aa, >scaffold-ZCDJ-2083162-3... *
1	1			052aa, >scaffold-CDFR-2053546-2... *
1	3			052aa, >scaffold-GSXD-2056597-2... *
1	5			052aa, >scaffold-KJZG-2052748-2... *
1	6			052aa, >scaffold-KJZG-2056759-1... *
1	7			052aa, >scaffold-PIVW-2079688-3... *
1	9			052aa, >scaffold-PSKY-2049278-5... *
1	10			052aa, >scaffold-QIAD-2054155-2... *
1	12			052aa, >scaffold-YQEC-2061897-3... *
1	15			048aa, >scaffold-IZLO-2000947-3... *

1	19			047aa, >scaffold-DBCE-2003992-2... *
1	20			047aa, >scaffold-DHAW-2055900-2... *
1	22			047aa, >scaffold-FXGI-2039500-4... *
1	24			047aa, >scaffold-GETL-2002710-2... *
1	26			047aa, >scaffold-HRUR-2027088-2... *
1	27			047aa, >scaffold-HRUR-2124614-4... *
1	28			047aa, >scaffold-HRUR-2132678-4... *
1	29			047aa, >scaffold-HUSX-2046732-6... *
1	30			047aa, >scaffold-INSP-2004803-5... *
1	35			047aa, >scaffold-RXEN-2008688-3... *
1	37			047aa, >scaffold-TCYS-2089623-4... *
1	38			047aa, >scaffold-VVVV-2037255-1... *
1	39			047aa, >scaffold-BOLZ-2160726-1... *
1	40			047aa, >scaffold-DLJZ-2003519-2... *
1	42			047aa, >scaffold-LQJY-2075444-6... *
1	44			046aa, >scaffold-BJKT-2001806-5... *
1	48			046aa, >scaffold-DCCI-2001222-1... *
1	49			046aa, >scaffold-DHPO-2073062-3... *
1	51			046aa, >scaffold-EILE-2030631-2... *
1	52			046aa, >scaffold-EWXK-2100001-2... *
1	54			046aa, >scaffold-GNPX-2010863-1... *
1	56			046aa, >scaffold-LKKX-2059782-5... *
1	58			046aa, >scaffold-NEBM-2072982-6... *
1	59			046aa, >scaffold-NUZN-2045695-5... *
1	62			046aa, >scaffold-OHAE-2024178-6... *
1	65			046aa, >scaffold-QOXT-2156719-3... *
1	66			046aa, >scaffold-RDYY-2065695-3... *
1	67			046aa, >scaffold-RXEN-2074375-6... *
1	69			046aa, >scaffold-THEW-2015565-3... *
1	70			046aa, >scaffold-TPEM-2142236-6... *
1	71			046aa, >scaffold-VVVV-2038782-3... *
1	72			046aa, >scaffold-WKSU-2015479-2... *
1	73			046aa, >scaffold-XMQO-2018511-4... *

Table I: *75 protease inhibitors ii representative sequences/clusters*

4.7 Non-Specific Lipid Transfer

sorted number of 4 letter IDs	sorted cluster		
61	3		082aa, >scaffold-ARYD-2095691-2... *
54	0		082aa, >scaffold-AFLV-2002357-1... *
42	1		082aa, >scaffold-AIGO-2000916-3... *
26	2		082aa, >scaffold-AIGO-2008115-5... *
15	5		082aa, >scaffold-CIAC-2102653-6... *
7	4		082aa, >scaffold-BYQM-2002213-5... *
5	6		082aa, >scaffold-FUMQ-2080014-6... *
2	11		082aa, >scaffold-NRXL-2058821-5... *
1	7		082aa, >scaffold-GIWN-2078988-4... *
1	8		082aa, >scaffold-IXVJ-2118845-1... *
1	9		082aa, >scaffold-JRNA-2011250-2... *
1	10		082aa, >scaffold-KAWQ-2051236-4... *
1	12		082aa, >scaffold-YBML-2119972-3... *
1	13		082aa, >scaffold-YJUG-2136899-4... *

Table J: *14 non-specific lipid transfer representative sequences/cluster*

4.8 S-Adenosyl-L-Methionine-Dependent Methyltransferase

sorted number of 4 letter IDs	sorted cluster		
426	0		0510aa, >scaffold-AALA-2001649-5... *
42	4		0510aa, >scaffold-LELS-2001751-4... *
6	1		0510aa, >scaffold-BYQM-2014815-2... *
4	2		0510aa, >scaffold-DEMH-2002418-5... *
4	3		0510aa, >scaffold-GAMH-2011437-6... *

Table K: *5 S-adenosyl-L-methionine dependant transferase representative sequences/clusters.*

4.9 Snakins

sorted number of 4 letter IDs	sorted cluster	
622	86	058aa, >scaffold-AFQQ-2013121-6... *
400	10	059aa, >scaffold-ABSS-2004730-4... *
285	94	058aa, >scaffold-ATQZ-2010082-5... *
274	87	058aa, >scaffold-AIGO-2062163-3... *
264	76	058aa, >scaffold-ABSS-2069472-6... *
249	32	5459aa, >scaffold-CGGO-2014777-2... *
236	58	17559aa, >scaffold-QIKZ-2140037-4... *
206	3	11061aa, >scaffold-MFIN-2051769-4... *
172	103	058aa, >scaffold-BJKT-2059822-2... *
130	92	058aa, >scaffold-AQXA-2009646-2... *
108	33	2359aa, >scaffold-CLMX-2038820-5... *
108	134	058aa, >scaffold-FWCQ-2009318-2... *
100	9	059aa, >scaffold-AALA-2072124-4... *
96	96	058aa, >scaffold-AYIY-2084502-1... *
90	78	058aa, >scaffold-ACWS-2058108-3... *
89	11	059aa, >scaffold-AFQQ-2045162-2... *
85	25	1259aa, >scaffold-BSEY-2011266-2... *
85	80	058aa, >scaffold-AEPI-2042988-2... *
84	15	059aa, >scaffold-AQGE-2008871-1... *
72	130	058aa, >scaffold-ERIA-2001212-2... *
70	77	058aa, >scaffold-ACFP-2028398-6... *
64	43	1959aa, >scaffold-FROP-2077766-5... *
48	4	060aa, >scaffold-AJBK-2064399-2... *
46	12	059aa, >scaffold-AHRN-2010013-3... *
45	117	058aa, >scaffold-CKDK-2088751-6... *
44	0	061aa, >scaffold-BLAJ-2111887-3... *
43	38	359aa, >scaffold-DUQG-2048670-6... *
43	39	059aa, >scaffold-EGOS-2079159-3... *
40	111	058aa, >scaffold-BSVG-2072153-4... *
39	101	058aa, >scaffold-BFMT-2066648-6... *
37	141	058aa, >scaffold-HABV-2056376-2... *
36	13	059aa, >scaffold-AIGO-2062649-5... *
34	74	058aa, >scaffold-ABIJ-2035835-1... *
33	71	058aa, >scaffold-AAXJ-2047296-4... *
30	1	161aa, >scaffold-BXAY-2020106-4... *
28	104	058aa, >scaffold-BLVL-2054182-3... *
27	82	058aa, >scaffold-AFLV-2002318-6... *
26	36	1159aa, >scaffold-COCP-2010601-5... *
25	138	058aa, >scaffold-GBCQ-2081973-6... *
24	81	058aa, >scaffold-AEPI-2045244-3... *
24	83	058aa, >scaffold-AFPO-2006191-4... *
24	84	058aa, >scaffold-AFPO-2065464-2... *
24	112	058aa, >scaffold-BZMI-2106485-6... *
24	120	058aa, >scaffold-CXSJ-2003729-4... *
23	127	058aa, >scaffold-DZTK-2037361-6... *
22	30	059aa, >scaffold-CDFR-2005711-3... *
22	75	058aa, >scaffold-ABIJ-2036554-4... *
20	79	058aa, >scaffold-ACWS-2064136-3... *
19	85	058aa, >scaffold-AFQQ-2011673-3... *
19	102	058aa, >scaffold-BJKT-2007985-4... *
19	131	058aa, >scaffold-EVOD-2012664-1... *
19	156	058aa, >scaffold-OLXF-2082060-5... *
18	125	058aa, >scaffold-DGXS-2099725-5... *
17	105	058aa, >scaffold-BLWH-2019760-3... *
16	7	660aa, >scaffold-GDKK-2065853-4... *
14	8	1360aa, >scaffold-ZZEI-2021281-1... *

14	44		359aa, >scaffold-FZQN-2012064-5... *
13	100		058aa, >scaffold-BERS-2008962-4... *
13	126		058aa, >scaffold-DVXD-2064901-3... *
12	34		059aa, >scaffold-CMCY-2000596-5... *
12	48		059aa, >scaffold-HNDZ-2013282-4... *
11	5		060aa, >scaffold-BPKH-2014912-4... *
11	49		059aa, >scaffold-HTIP-2050457-6... *
11	95		058aa, >scaffold-ATQZ-2012711-2... *
11	106		058aa, >scaffold-BLWH-2101625-3... *
11	135		058aa, >scaffold-FYUH-2107791-4... *
11	155		658aa, >scaffold-OLES-2123003-3... *
10	16		059aa, >scaffold-ATFX-2040760-5... *
10	17		159aa, >scaffold-AVJK-2129498-6... *
10	21		159aa, >scaffold-BEKN-2005481-5... *
10	122		058aa, >scaffold-DBCE-2004653-5... *
10	128		058aa, >scaffold-ECTD-2016932-6... *
9	18		059aa, >scaffold-AWQB-2050267-4... *
9	42		059aa, >scaffold-ERXG-2060829-6... *
9	56		559aa, >scaffold-PTFA-2003186-3... *
9	107		058aa, >scaffold-BMIF-2004453-1... *
9	113		058aa, >scaffold-CAPN-2007049-3... *
8	41		059aa, >scaffold-EMJJ-2106854-6... *
8	110		058aa, >scaffold-BNTL-2024623-3... *
7	14		059aa, >scaffold-ALVQ-2006555-3... *
7	19		159aa, >scaffold-AXPJ-2037348-2... *
7	97		058aa, >scaffold-BBDD-2071275-4... *
7	98		058aa, >scaffold-BEGM-2016947-4... *
6	23		059aa, >scaffold-BPKH-2073214-1... *
6	45		059aa, >scaffold-GANB-2010937-2... *
6	52		059aa, >scaffold-JQCX-2080196-1... *
6	54		459aa, >scaffold-PIVW-2014024-3... *
6	66		159aa, >scaffold-WKCY-2044804-1... *
6	115		058aa, >scaffold-CBAE-2002201-1... *
6	119		058aa, >scaffold-CQPW-2003296-6... *
6	144		058aa, >scaffold-JKAA-2009412-1... *
5	73		058aa, >scaffold-ABIJ-2006730-3... *
5	91		058aa, >scaffold-AQFM-2075896-6... *
5	93		058aa, >scaffold-ATQZ-2009611-4... *
5	158		058aa, >scaffold-PCGJ-2059568-2... *
5	177		057aa, >scaffold-FZQN-2080691-5... *
4	24		059aa, >scaffold-BRUD-2006466-4... *
4	26		059aa, >scaffold-BSTR-2004096-5... *
4	40		059aa, >scaffold-EMJJ-2002479-6... *
4	57		059aa, >scaffold-QIAD-2054174-3... *
4	72		058aa, >scaffold-ABIJ-2000137-1... *
4	118		058aa, >scaffold-CMEQ-2080284-4... *
4	137		058aa, >scaffold-GAON-2009142-2... *
4	139		058aa, >scaffold-GGJD-2119379-3... *
4	147		058aa, >scaffold-KUXM-2002150-2... *
3	35		059aa, >scaffold-CMCY-2036897-3... *
3	47		059aa, >scaffold-GNQG-2005849-6... *
3	51		059aa, >scaffold-JKAA-2011094-5... *
3	62		059aa, >scaffold-TXMP-2038431-6... *
3	65		259aa, >scaffold-VBHQ-2032401-3... *
3	90		058aa, >scaffold-ANON-2003954-4... *
3	116		058aa, >scaffold-CBAE-2058413-3... *
3	121		058aa, >scaffold-CYVA-2066838-4... *

3	124		058aa, >scaffold-DFHO-2005577-4... *
3	129		058aa, >scaffold-ECTD-2108098-4... *
3	136		058aa, >scaffold-GANB-2067413-2... *
3	142		058aa, >scaffold-HJMP-2064459-6... *
3	154		058aa, >scaffold-NYBX-2125528-4... *
3	159		058aa, >scaffold-PKOX-2090172-1... *
3	174		058aa, >scaffold-WQUF-2019789-1... *
2	2		161aa, >scaffold-EMJJ-2108628-2... *
2	6		060aa, >scaffold-CGGO-2015569-6... *
2	20		059aa, >scaffold-AZBL-2005889-4... *
2	22		059aa, >scaffold-BEKN-2058746-3... *
2	29		059aa, >scaffold-CAPN-2038182-2... *
2	63		059aa, >scaffold-UDHA-2025424-5... *
2	89		058aa, >scaffold-ALVQ-2013030-3... *
2	99		058aa, >scaffold-BEGM-2099909-5... *
2	123		058aa, >scaffold-DEMH-2017884-4... *
2	140		058aa, >scaffold-GNQG-2079070-3... *
2	143		058aa, >scaffold-JKAA-2006715-1... *
2	145		058aa, >scaffold-JKAA-2011512-2... *
2	149		058aa, >scaffold-LPGY-2151546-6... *
2	151		058aa, >scaffold-NDUV-2098785-4... *
2	152		058aa, >scaffold-NDUV-2099945-3... *
2	153		058aa, >scaffold-NOKI-2010607-3... *
2	160		058aa, >scaffold-PKOX-2092923-2... *
2	164		058aa, >scaffold-THDM-2003993-4... *
2	166		058aa, >scaffold-TOXE-2055826-4... *
2	167		058aa, >scaffold-TRPJ-2049954-2... *
2	176		058aa, >scaffold-YNFJ-2006929-3... *
1	27		059aa, >scaffold-BSTR-2088409-1... *
1	28		059aa, >scaffold-BYQM-2007548-3... *
1	31		059aa, >scaffold-CGGO-2014481-2... *
1	37		059aa, >scaffold-DFHO-2050383-4... *
1	46		059aa, >scaffold-GJPF-2009557-3... *
1	50		059aa, >scaffold-JHCN-2084618-1... *
1	53		059aa, >scaffold-MFIN-2000562-1... *
1	55		059aa, >scaffold-PSKY-2007038-6... *
1	59		059aa, >scaffold-RUUB-2078513-5... *
1	60		059aa, >scaffold-SGTW-2036357-1... *
1	61		059aa, >scaffold-TOKV-2052584-6... *
1	64		059aa, >scaffold-UOMY-2008346-4... *
1	67		059aa, >scaffold-WTDE-2006057-4... *
1	68		059aa, >scaffold-WWSS-2003529-6... *
1	69		059aa, >scaffold-XZME-2072123-2... *
1	70		059aa, >scaffold-ZZEI-2019169-5... *
1	88		058aa, >scaffold-ALVQ-2003328-5... *
1	108		058aa, >scaffold-BMIF-2006012-6... *
1	109		058aa, >scaffold-BNDE-2090049-1... *
1	114		058aa, >scaffold-CAPN-2036054-1... *
1	132		058aa, >scaffold-EYKJ-2053966-2... *
1	133		058aa, >scaffold-EYKJ-2058106-4... *
1	146		058aa, >scaffold-JPDJ-2053844-4... *
1	148		058aa, >scaffold-KUXM-2036951-6... *
1	150		058aa, >scaffold-MTII-2037596-3... *

1	157		058aa, >scaffold-OPDF-2032353-3... *
1	161		058aa, >scaffold-PQTO-2076625-3... *
1	162		058aa, >scaffold-PYHZ-2066053-6... *
1	163		058aa, >scaffold-SGTW-2033538-4... *
1	165		058aa, >scaffold-TJES-2003716-1... *
1	168		058aa, >scaffold-UBLN-2076135-6... *
1	169		058aa, >scaffold-UDHA-2087264-5... *
1	170		058aa, >scaffold-ULKT-2017074-6... *
1	171		058aa, >scaffold-UPMJ-2079340-5... *
1	172		058aa, >scaffold-VDAO-2003131-6... *
1	173		058aa, >scaffold-VTLJ-2066359-1... *
1	175		058aa, >scaffold-XSZI-2010521-1... *
1	178		057aa, >scaffold-RNBN-2037485-1... *

Table L: *179 snakins representative sequences/clusters.*

Chapter 5

Conclusions

It is clear that the development of new scripts is indispensable to extract data from databases. There are a lot of online sources but for specific tasks it is convenient to use own made scripts, using an open-source programming language like Python. The choice of Python is crucial because it has a big community, so it is more accessible, and it is easier to find support. Another important aspect is the computing power to reduce the time spent to analyse the enormous number of peptides/proteins. Six frame translation and the finding of open reading frames are demanding tasks, indeed over than 150.000.000 sequences have been extracted, it took 4-5 days using a low budget PC. After the extraction of peptides from the 1000 plants database, python scripts filtered the dataset by known cysteine motifs. The knowledge of a particular aspect of the peptide (i.e. cysteine motif) is crucial to rule out not relevant elements. Filtering data is essential because we need a decrease in number of peptides/proteins. Thus, nine cysteine rich families have been found in the 1000 plants database. To follow up clustering sequences is decisive to find similarity between the peptides filtered. First, several runs of the clustering algorithm are required to find optimal parameters, also user manual is informative to avoid mistakes. For this reason, the optimal threshold of similarity is c = 0.7 (0.75 only for SALMdM) and n = 5 (size of word). One particularity of this study is the "virtual" truncation of the contigs based on data obtained after a certain number of clustering runs. The truncation is performed before the clustering, it showed that number of clusters decreased and average number of elements for each cluster increased respect to the use of complete peptides. This method reduced single peptide clusters. Last step before clustering is the removal of peptides with unknown amino acids, this will be beneficial for future computational analyses. CD-HIT revealed 22 bowman birk clusters, 46 cyclotides clusters, 327 defensins clusters, 53 heveins clusters, 48 kazal clusters, 75 protease inhibitors II clusters, 14 non-specific lipid transfer clusters, 5 SALMdM clusters and 179 snakins clusters. Arrangement of data with a ranking criterion is advantageous to visualize peptides of interest. Finally, we found similar sequences that could be potential drugs or valuable in other fields such as food or pesticides, but further computational and experimental validations are indispensable. This methodology is convenient to decrease the amount of time for experimental finding, optimistically minimize the number of experiments and optimize all the drug development pipeline.

Bibliography

[1] Bethany M. Cooper, Jessica Iegre, Daniel H. O' Donovan, Maria Ölwegård Halvarsson and David R. Spring. *Peptides as a platform for targeted therapeutics for cancer: peptide-drug conjugates (PDCs). 30th September 2020.*

[2] Andy Chi-Lung Lee, Janelle Louise Harris, Kum Kum Khanna and Ji-Hong Hong. *Comprehensive Review on Current Advances in Peptide Drug Development and Design 2019*

[3] Junpeng Li, Shuping Hu, Wei Jian, Chengjian Xie and Xingyong Yang. *Plant antimicrobial peptides: structures,functions, and applications.* 2021

[4] Bernhard Retzl, Roland Hellinger, Edin Muratspahić, Meri E. F. Pinto, Vanderlan S. Bolzani, and Christian W. Gruber. *Discovery of a Beetroot Protease Inhibitor to Identify and Classify Plant-Derived Cystine Knot Peptides. 2020*

[5] Wang, L., Wang, N., Zhang, W. *et al.* *Therapeutic peptides: current applications and future directions. 2022*

[6] Uttpal Anand, Anustup Bandyopadhyay, Niraj Kumar Jha, José M. Pérez de la Lastra, Abhijit Dey. *Translational aspect in peptide drug discovery and development: An emerging therapeutic candidate*

[7] Guilherme Sastre de Souza, Leandra de Jesus Sonego, Ana Clara Santos Mundim, Júlia de Miranda Moraes, Helioswilton Sales-Campos, Esteban Nicolás Lorenzón. *Antimicrobial-wound healing peptides: Dual-function molecules for the treatment of skin injuries ,2022*

[8] Mani S, Bhatt SB, Vasudevan V, Prabhu D, Rajamanikandan S, Velusamy P, Ramasamy P, Raman P. *The Updated Review on Plant Peptides and Their Applications in Human Health 2022*

[9] Shilpi Srivastava, Kavya Dashora, Keshav Lalit Ameta, Nagendra Pratap Singh, Hesham Ali El-Enshasy, Marcela Claudia Pagano, Abd El-Latif Hesham, Gauri Dutt Sharma, Minaxi Sharma, Atul Bhargava *Cysteine-rich antimicrobial peptides from plants: The future of antimicrobial therapy*

[10] Li J, Sun C, Cai W, Li J, Rosen BP, Chen J. *Insights into S-adenosyl-l-methionine (SAM)-dependent methyltransferase related diseases and genetic polymorphisms. 2021*

[11] Shazia Rehman, Ejaz Aziz, Wasim Akhtar, Muhammad Ilyas, Tariq Mahmood. *Structural and functional characteristics of plant proteinase inhibitor-II (PI-II) family*

[12] Rimphanitchayakit V, Tassanakajon A. *Structure and function of invertebrate Kazal-type serine proteinase inhibitors*. *2010*

[13] Rui-Feng QI, Zhan-Wu SONG, Cheng-Wu CHI. *Structural Features and Molecular Evolution of Bowman-Birk Protease Inhibitors and Their Potential Application 2005*

[14] Eric J. Carpenter, Naim Matasci, Saravanaraj Ayyampalayam, Shuangxiu Wu, Jing Sun , Jun Yu , Fabio Rocha Jimenez Vieira,Chris Bowler, Richard G. Dorrell , Matthew A. Gitzendanner, Ling Li, Wensi Du, Kristian K. Ullrich , Norman J. Wickett , Todd J. Barkmann, Michael S. Barker, James H. Leebens-Mack and Gane Ka-Shu Wong *Access to RNA-sequencing data from 1,173 plant species: The 1000 Plant transcriptomes initiative (1KP)*

[15] William F. Porto, Octavio L. Franco, *Theoretical structural insights into the snakin/GASA family, 2013,*

[16] Wang, NJ., Lee, CC., Cheng, CS. *et al.* *Construction and analysis of a plant non-specific lipid transfer protein database (nsLTPDB). BMC Genomics **13** (Suppl 1), S9 (2012).*

[17] Aymara Cabrera-Muñoz, Pedro A. Valiente, Laritza Rojas, Maday Alonso-del-Rivero Antigua, José R. Pires. *NMR structure of CmPI-II, a non-classical Kazal protease inhibitor: Understanding its conformational dynamics and subtilisin A inhibition, 2019,*

[18] Slezina MP, Odintsova TI. *Plant Antimicrobial Peptides: Insights into Structure-Function Relationships for Practical Applications.2023*

[19] Weizhong Li, Adam Godzik. *Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences July 2006.*

[20] Fu L, Niu B, Zhu Z, Wu S, Li W. *CD-HIT: accelerated for clustering the next-generation sequencing data. 2012*

[21] Balaji Prakash, S. Selvaraj, M.R.N. Murthy, Y.N. Sreerama, D. Rajagopal Rao, Lalitha R. Gowda *Analysis of the Amino Acid Sequences of Plant Bowman-Birk Inhibitors 1995*

[22] David J. Craik, Norelle L. Daly, Trudy Bond and Clement Waine. *Plant Cyclotides: A Unique Family of Cyclic and Knotted Proteins that Defines the Cyclic Cystine Knot Structural Motif*

[23] Thomas M. A. Shafee, Fung T. Lay, Mark D. Hulett, and Marilyn A. Anderson *The Defensins Consist of Two Independent, Convergent Protein Superfamilies*

[24] Monika M. Edstama , Lenita Viitanenb , Tiina A. Salminenb and Johan Edqvista. *Evolutionary History of the Non-Specific Lipid Transfer Proteins*

[25] Jiaojiao Lia, Chunxiao Suna, Wenwen Caia, Jing Lia, Barry P. Rosenb, Jian Chen. *Insights into S-adenosyl-L-methionine (SAM)-dependent methyltransferase related diseases and genetic polymorphisms.*

[26] Marina P. Slezina and Tatyana I. Odintsova. *Plant Antimicrobial Peptides: Insights into Structure-Function Relationships for Practical Applications*

[27] Sebastián Pariani, Marisol Contreras, Franco R. Rossi, Valeria Sander, Mariana G. Corigliano, Francisco Simón, María V. Busi, Diego F. Gomez-Casati, Fernando L. Pieckenstain , Vilma G. Duschak , Marina Clemente. *Characterization of a novel Kazal-type serine proteinase inhibitor of Arabidopsis thaliana*

[28] Shazia Rehman, Ejaz Aziz. Wasim Akhtar, Muhammad Ilyas, Tariq Mahmood. *Structural and functional characteristics of plant proteinase inhibitor-II (PI-II) family*

[29] Marta Berrocal-Lobo, Ana Segura, Manuel Moreno, Gemma Lo´pez, Francisco García-Olmedo, and Antonio Molina. *Snakin, an Antimicrobial Peptide from Potato Whose Gene Is Locally Induced by Wounding and Responds to Pathogen Infection.*

[30] Brendon F. Conlan, Amanda D. Gillon, David J. Craik, Marilyn A. Anderson. *Circular proteins and mechanisms of cyclization. 2010*

[31] Tyagi, R. *Computational Genetics. India: Discovery Publishing House*. 2009

[32] Nelson CW, Ardern Z, Wei X. *OLGenie: Estimating Natural Selection to Predict Functional Overlapping Genes. 2020*

[33] P. Villesen *FaBox: an online toolbox for fasta sequences*

[34] Blankenberg D, Von Kuster G, Bouvier E, Baker D, Afgan E, Stoler N, Galaxy Team, Taylor J, Nekrutenko A *(2014). Dissemination of scientific software with Galaxy ToolShed.*

[35] Oleg A. Shchelochkov, M.D. *https://www.genome.gov/genetics-glossary/Open-Reading-Frame*

[36] Momeni Z, *et al.* *A Survey on Single and Multi-Omics Data Mining Methods in Cancer Data Classification*, 2020.

[37] Khalid R. *Application of Data Mining In Bioinformatics.* 2010.

[38] Zaki M J, et *al*. *Data Mining in Bioinformatics (BIOKDD). Algorithms for Molecular Biology.*