UNIVERSITÀ DEL PIEMONTE ORIENTALE

Dipartimento di scienze e innovazione tecnologica - DISIT

Corso di Laurea in Biologia

TESI

Caratterizzazione dei motivi specifici delle proteine effetrici di fitoparassiti, utilizzando la pipeline MOnSTER

Relatore: Prof. Francesco Dondero
Correlatore: Prof.ssa Silvia Bottini

Candidata: Paola Porracciolo
MATRICOLA: 20023920

Anno Accademico: 2024/2025

A mia Madre, Maria Luisa, questa laurea è anche tua.

Acknowledgments

During the academic year 2021/2022, I completed my internship in the context of the Erasmus+ training exchange program in Université Côte d'Azur in Nice, France. I worked between two institutes to develop the bioinformatics tool MOnSTER (MOtifs cluSTER) to identify CLUMPs (CLUsters of Motifs of Proteins):

1. MDLab Medical Data Laboratory at the MSI Maison de la Modélisation, Simulation et Interactions, in Nice, France; under the supervision of Professor Bottini and the co-supervision of Dr. Djampa Kozlowski.
2. Equipe GAME at INRAE Institut national de recherche pour l'agriculture, l'alimentation et l'environnement, in Valbonne, France; under the co-supervision of Dr. Etienne Danchin.

During that year as an Erasmus+ student - the second year of the present master's degree - I completed the first year of the master's degree of Life Science, "Biology, Informatics and Mathematics" curriculum.

The year after (a.a. 2022/2023), I suspended my career in the present master's degree - according to the regulation of Università del Piemonte Orientale - to attend the second year of the same French master's degree, that I completed by defending a different work from this thesis.
In particular, I worked on the "Integration of hierarchical cell types for marker genes visualization in Human Lung Cell Atlas". The importance of knowing cell types heterogeneity in the human lungs is fundamental to understand the molecular mechanisms of life-threatening illnesses, like the Chronic obstructive pulmonary disease (COPD).

In the second half of 2023, I regained access to the present master's degree to complete my career. Between 2023 and today, in addition to being a student in Università del Piemonte Orientale:

- I worked as a design engineer in bioinformatics in Institut de Pharmacologie Moléculaire et Cellulaire (IPMC) in Valbonne, France (from September 2023 to September 2024).
- I have started an interdisciplinary PhD in biology and mathematics, on the 1st October 2024, that I am currently a candidate of, in Université Côte d'Azur.

In addition to completing the exams from my career in Università del Piemonte Orientale, I worked on a publication, from the present work, with my co-rapporteur Professor Bottini, who was my internship director in France in the a.a. 2021/2022.
The work was completed by:

- A proof of concept of the developed pipeline on *Oomycetes*.
- Testing the pipeline on other plant parasitic nematodes.
- Experimental validation of MiEFF72, a novel putative effector of *M. incognita*.

This work, of which I am co-first author, was published in the second half of 2024. I cite it as follows:

Calia, G., Porracciolo, P., Chen, Y. et al. Identification and characterization of specific motifs in effector proteins of plant parasites using MOnSTER. *Commun Biol* 7, 850 (2024). https://doi.org/10.1038/s42003-024-06515-9

For these reasons, I want to thank my supervisor, Professor Francesco Dondero, who I have known for years and who has always encouraged me to dream big to achieve my dreams. It is also thanks to him if I was able to do all this.
I want to thank Professor Silvia Bottini, my supervisor while being in France, Dr. Djampa Kozlowski and Dr. Etienne Danchin, my co-supervisors in France. This internship was a very enriching experience, not only I acquired a lot of competences, but it truly confirmed my deep passion for research.
I want to thank Dr. Giulia Calia, for extending the work of MOnSTER to a very interesting level, that means a lot to me and my scientific career that has just started.
Last but not least, my dad, Nino, my mum, Maria Luisa, and my sister, Marisa, without you none of this and nothing in my life would be possible.

Abstract

I nematodi parassiti di piante (PPN) del genere *Meloidogyne* causano ogni anno perdite di raccolto di miliardi di dollari, in tutto il mondo. Tra i PPN, in particolare, la specie *Meloidogyne incognita* è verosimilmente la più devastante, poiché in grado di infettare quasi tutte le piante coltivate. Il nematode infetta la pianta tramite delle proteine chiamate "effettori", che modulano la risposta della pianta per la sopravvivenza del nematode, tramite una strategia biotrofica. L'identificazione e la caratterizzazione degli effettori all'interno del proteoma del parassita è particolarmente dispendiosa e complessa e i metodi attuali generano troppi candidati per gli studi sperimentali.

Per affrontare questo problema, mi sono concentrata sulla ricerca di corte sequenze, dette "motivi", al fine d'identificare e caratterizzare suddette proteine effettrici. L'ho fatto utilizzando un dataset positivo e uno negativo di sequenze di *M. incognita*, rispettivamente di effettori e non effettori, e un dataset positivo e uno negativo di sequenze di *M. arenaria*.

Sebbene esistano metodi efficienti per individuare i motivi, esistono delle problematiche di natura tecnica, ovvero: una degenerazione troppo elevata e nessun calcolo che tenga conto delle proprietà fisicochimiche degli amminoacidi (AA). Per colmare questa lacuna, ho sviluppato la pipeline bioinformatica MOnSTER (MOtifs cluSTER). Lo scopo principale è di raggruppare i motivi in cluster di motivi proteici (CLUMP), calcolare i valori delle caratteristiche fisicochimiche dei dataset di proteine e sui CLUMP; e, infine, l'obiettivo del software è di calcolare il punteggio MOnSTER sulla base delle sopraindicate proprietà fisicochimiche e dei tassi di occorrenza dei CLUMP nei dataset, con una prevalenza nel dataset contente le proteine effettrici, per discriminare le proteine effettrici dalle proteine non effettrici.

Il lavoro è stato svolto in due fasi: la prima nel contesto di uno scambio internazionale, tramite il programma "Erasmus+", durante il quale ho studiato all'Université Côte d'Azur di Nizza, in Francia, in questo periodo, ho sviluppato MOnSTER utilizzando le sequenze di *M.incognita* e l'ho testato sulle sequenze di *M. arenaria*; la seconda parte invece finalizzata a testare la robustezza della pipeline MOnSTER in oomiceti e in altri fitoparassiti PPN.

Table of contents

ACKNOWLEDGMENTS **2**

ABSTRACT **4**

TABLE OF CONTENTS **5**

1. CHAPTER 1: INTRODUCTION **7**

1.1. *MELOIDOGYNE INCOGNITA* AND THE ROOT-KNOT NEMATODES 7
1.1.1. Classification of the RKNs based on their reproduction strategy *7*
1.1.2. Life cycle of M. incognita *8*
1.1.3. The triploid genome of M. incognita *9*
1.2. THE PLANT INFECTION 10
1.2.1. Physiopathology of the plant infection *11*
1.2.2. Different plant phyla react differently to the plant infection *11*
1.2.3. Statistics on annual crop losses due to root-knot nematodes *12*
1.3. ECOTOXICOLOGY OF NEMATICIDES 13
1.3.1. Existing biocontrol strategies as alternatives to nematicides *13*
1.4. EFFECTOR PROTEINS 15
1.4.1. Experimental study of effector proteins *16*
1.5. MOTIFS 18
1.5.1. Well studied motifs in effectors of plant parasites *18*
1.5.2. Motif mining *19*

2. CHAPTER 2: OBJECTIVE OF THE STUDY **20**

3. CHAPTER 3: MATERIALS AND METHODS **21**

3.1. DATA TREATED TO DEVELOP MONSTER 21
3.1.1. Datasets *21*
3.2. DATA TREATED TO TEST THE ROBUSTNESS OF MONSTER 21
3.2.1. Oomycetes *22*
3.2.2. Other PPNs *22*
3.3. MONSTER PIPELINE 22
3.3.1. Motif Discovery *23*
3.3.2. Feature calculation *24*
3.3.3. Clustering *26*
3.3.4. Scoring *26*
3.3.5. CLUMPs positions in sequences *29*
3.3.6. CLUMPs co-occurrences *29*

4. CHAPTER 4: RESULTS **30**

4.1. RESULTS OBTAINED WHEN DEVELOPING MONSTER 30
4.1.1. Sequence composition preferences of the effectors of M. incognita *30*
4.1.2. MOnSTER allowed to identify 6 CLUMPS of motifs discriminant for the effectors *33*
4.1.3. Best scored CLUMPs (1,0,4) occur at central positions in sequences of effectors *38*
4.1.4. CLUMPs 1, 0 and 4 co-occur in effector sequences at small distances within each other *39*
4.1.5. The 6 CLUMPs identified by MOnSTER characterize sub-populations of effector proteins in M. incognita *42*
4.2. MONSTER SUCCESSFULLY IDENTIFIES WELL-KNOWN PARASITISM MOTIFS TO HELP EFFECTORS' EXPERIMENTAL VALIDATION 43
4.2.1. Proof of concept of the MOnSTER: application on Oomycetes *43*
4.2.2. Application of MOnSTER on other plant parasitic nematodes *48*
4.2.3. Positional preference of the motifs *49*
4.2.4. Co-occurrences of CLUMPs and functional domains *49*
4.2.5. Experimental validation of MiEFF72: a novel putative effector of M. incognita *50*

5. CHAPTER 5: DISCUSSION **53**

6. CHAPTER 6: CONCLUSIONS 57
REFERENCES 59

1. Chapter 1: Introduction

1.1. *Meloidogyne incognita* and the root-knot nematodes

Plant-Parasitic Nematodes (PPNs) are biotrophic roundworms that cause billions of dollars of crop losses worldwide every year (Vieira & Gleason, 2019).

Among the PPNs, nematodes from the *Meloidogyne* genus are described to be the most devastating ones (Jones et al., 2013). *Meloidogyne* species, also known as root-knot nematodes (RKNs), are microscopic endoparasites that specifically target the plant's roots. *Meloidogyne* agronomic impact is due to their wide host range (up to more than 3,000 species of plants) and their worldwide distribution (Malaysia et al., 2016).

One species, in particular, *M. incognita*, due to its extreme versatility, is one of the most important known crop pathogens among all (Abad et al., 2008).

1.1.1. Classification of the RKNs based on their reproduction strategy

The *Meloidogyne* genus is very vast, researchers have currently identified around 90 species (*Https://Ephytia.Inra.Fr/En/C/20910/Potato-Meloidogyne-Spp-Root-Knot-Nematodes*, n.d.).

These species are characterized by a variety of chromosome counts, hosts and reproductive models. In particular, we can distinguish those with: obligate sexual reproduction, facultative sexual reproduction, obligatory mitotic parthenogenesis (asexual reproduction) (Phan et al., 2020), as indicated in **Table 1**.

Group (by reproduction strategy)	**Relative number of chromosomes (#numeric example)**	**Host range**	**Geographic distribution**	**Example of species**
Obligate sexual reproduction	Few chromosomes (around 7)	Narrow	Limited	*M. spartinae*
Facultative sexual reproduction	Intermediate (13-19)	Broad	Vaste	*M. graminicola, M. hapla, M. chitwoodii*
Obligatory asexual reproduction (mitotic parthenogenesis)	High (polyploidy, triploidy: 42-48)	Very broad	Most extensively distributed	*M. incognita, M. javanica, M. arenaria, M. enterolobii, M. floridensis*

Table 1. Classification of the RKNs based on their reproduction strategy

1.1.2. Life cycle of *M. incognita*

The life cycle of *M. incognita* has been extensively studied for decades, identifying an egg stage, four developmental stages (indicated with an initial "J"), and the adult male/female stage (Subedi et al., 2020). These stages are shown in **Figure 1**.

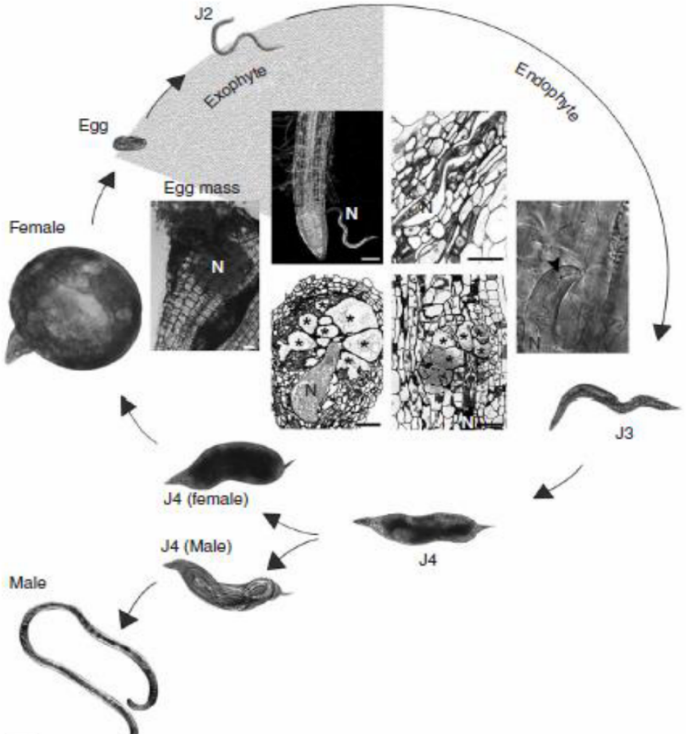


Figure 1. Life cycle of M. incognita (Figure retrieved from (Abad et al., 2008)).

1. **First stage of development: the egg stage**. When females reach maturity, they lay up to 1,000 eggs either on the root or in gall tissues.
2. **Juvenile stages of development: J1** (not displayed in **Figure 1**) **and J2**. Embryogenesis occurs, and the first-stage juvenile transitions to the second-stage juvenile. Then, hatching of the eggs happens.
3. **Infection stage: J2**. This is the stage where the juvenile nematode penetrates the host roots. To do this, it uses a stylet to pierce cell walls, as described in **Figure 2B**.
4. **Induction of the feeding site: the giant cells (GCs)**. The J2 nematode induces a changing in 5-7 surrounding cells in the differentiation zone (**Figure 2C**) to transform them into GCs: a nutrient source essential for the nematode survival.
5. **Developing molts: J3, J4 and adult stages**. When the nutrition is guaranteed, the nematode is in the optimal environment for three molts: from J2 to J3, then to J4, and finally to get to the adult stage.

6. **Final adult stage: male and female nematodes**. At this stage:
    a. Females become pear-shaped, become sedentary and lay eggs.
    b. Males become motile (again) and migrate out of the root tissues.

1.1.3. The triploid genome of *M. incognita*

The most recent and comprehensive work focusing specifically on the triploid genome of *M. incognita:* (Mota et al., 2024) unraveled essential information about its complex structure.
As mentioned before, *M. incognita* has a triploid (AAB) genome, composed of: two subgenomes (A) which are closely related, one subgenome (B) that is more divergent to the two (A) subgenomes.

The main characteristics of the genome structure and the subgenomes are summarized in **Table 2**.

Size	**Subgenome divergence**	**Contig assignment**	**Karyotype**
Approximate genome size: 199.4 Mb (291 contigs of 1.86 Mb, with a high N50)	Average nucleotide divergence between types A and B subgenomes: 6.6%	- 113 contigs assigned to type A subgenomes - 43 contigs assigned to type B subgenome. (156/291 contigs from the complete genome assembly)	Estimated number of chromosomes: 46-47.

Table 2.Characteristics of the genome structure and the subgenomes of M. incognita.

Thanks to the massive improvement in the technical quality of the assembly, this version of the genome displays very low fragmentation, with big contigs of 1.86 million base pairs long. In **Table 2**, N50 refers to the contiguity of the genome assembly. In order words, half of the total assembled genome, 199.4 Mb, is described by 1.86 Mb contigs, which is two orders of magnitude higher than the N50 recovered in earlier assemblies. This is indeed essential to the quality and precision of analysis deriving from this assembly.

The three subgenomes can be categorized into two types, A and B, to explore the subgenomes divergence. To explore this datum, k-mer were used: a k-mer is a sequence of *k* nucleotides that is extracted from reads deriving from sequencing. K-mers have several roles in bioinformatics, but in this context, they were used for heterozygosity assessment, hence the genetic divergence of A and B types.

Although the complete genome assembly counts for 291 contigs, only 156 (which account for 16,892 genes) met the stringent quality criteria to be certainly assigned to the subgenomes.

Through genomic and cytogenetic analysis (including laser scanning microscopy and Fluorescent In Situ Hybridization), an estimated number of 46-47 chromosomes were estimated including the complete genome assembly.

1.2. The plant infection

RKNs are biotrophic parasites, hence they do not directly kill the plant since it represents their mean to survive. Nevertheless, they cause drastic reductions in the plant growth (Hussain et al., n.d.).

In order to infect the plant, RKNs penetrate the plant roots (**Figure 2A**) to reach to the vascular tissue (**Figure 2B**) and then migrate to the differentiation zone (**Figure 2C**). Once there, they inject proteins, known as effectors, into the cells through a stylet to induce all the necessary metabolic changes in the plant to survive at its expenses (Shi et al., 2018).

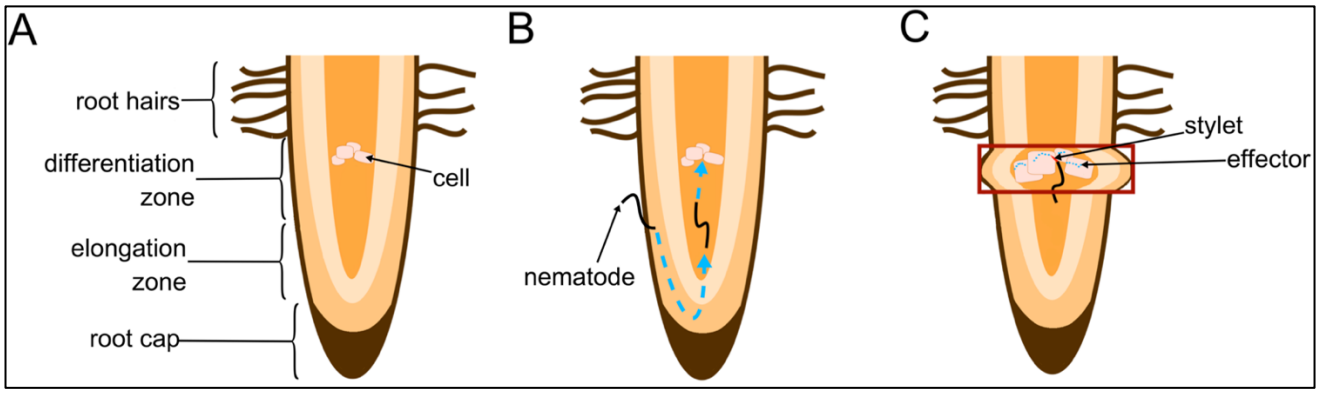


Figure 2. Plant infection by the RKNs scheme.

(A) General scheme of longitudinal section of the plant's root from the root hairs to the root cap. (B) RKNs penetrate the root and migrate through the cortex (where the first arrow is located), to get to the vascular tissue (where the second arrow is located). Once in the vascular

tissue, the RKNs migrate to the differentiation zone. (C) When RKNs get to the differentiation zone, they start their establishment, by injecting effector proteins into the cells, through a stylet.

1.2.1. Physiopathology of the plant infection

As explained in general terms in the previous paragraphs, RKNs establish a biotrophic interaction with the host plant, inducing it into creating a favorable environment for development and nourishment (Rutter et al., 2022).

To achieve this, the nematode injects effector proteins in the vascular cylinder of the plant root. The effectors have two main roles:

- Suppressing pathogen-associated molecular triggered immunity (PAMP-triggered immunity, or PTI).
- Manipulating cellular processes to generate multiciliated GCs from the protoxylem in the differentiation zone (**Figure 2C**).

To generate GCs, the cytoskeleton is rearranged through actin filaments and microtubules disruption. In co-occurrence to nematodes feeding from GCs, the surrounding tissues undergo hypertrophic cell division, which leads to root galls.

Effector secretion, GCs formation and the formation of root galls are all essential steps in the infection. This infection triggers systemic hormonal changes that involve jasmonic acid. This phytohormone is not synthesized directly in the roots, but in the shoots and then transported to the roots; its role is:

- Triggering defenses in the plant.
- Depending on its concentration and the balance with auxin and other hormones, enabling the maintenance of the feeding site in correspondence to the GCs.

1.2.2. Different plant phyla react differently to the plant infection

RKNs infect the majority of *Angiosperms*, causing major crop losses every year. However, monocotyledonous and dicotyledonous plant species react differently, structurally and molecularly.

1.2.2.1. Monocotyledonous and dicotyledonous plant species have different structural barriers

Monocotyledonous plants, like rice, possess a physical barrier to J2 nematodes penetration, thanks to a constitutive a higher suberin deposition than dicotyledonous plants. These lasts, like Arabidopsis, on the other hand, display a higher Casparian strip susceptibility induced by mutations.

1.2.2.2. Jasmonic acid- mediated resistance and ROS levels across plant species and phyla

Hormonal defense signaling can show differences across species and phyla.
In particular, the jasmonic acid-mediated resistance can vary across species: some synthesize it in the shoots, others in the site of infection, the roots.
In addition to that, nematodes use reactive oxygen species (ROS) as a chemical weapon. In fact, plants usually generate a burst of ROS to trigger apoptosis in the infected cells or to kill parasites. Contrarily, RKNs regulate the levels of ROS production to limit apoptosis and promote infection, with adjusted levels in the different phyla, adapting to their metabolism.

1.2.2.3. Timing of the effector-triggered immunity

NOD-like receptors usually play a central role in most plants resistance because they trigger effector-triggered immunity (ETI). Some plants, like pepper, induce a fast hypersensitive response (HR) to induce apoptosis in the cells of the infection site. Instead, others, such as the cowpea, only degrade cells through a classic HR in a second moment, allowing the feeding site to initially form.

1.2.3. Statistics on annual crop losses due to root-knot nematodes

RKNs represent one of the most important groups of plant-parasitic nematodes (PPNs) worldwide in terms of economic burden. The damage they cause is extended to several plant groups in crops, orchards, greenhouses and gardens (Kantor et al., 2024).

Their economic and agricultural impact makes them the primary contributors to an annual global loss caused by PPNs, for an estimated global value of USD 173 billion.
In particular, for what concerns staple crops, the estimation is the following:

- 35 billion USD for *Oryza sativa L.* (rice).

- 21 billion USD for *Zea mays L.* (maize).
- 6 billion USD for *Solanum tuberosum L.* (potatoes).
- 6 billion USD for *Triticum aestivum L.* (wheat).

Worldwide, RKNs significantly contribute to the roughly 40% of all food crops that are lost due to agricultural pests.

1.3. Ecotoxicology of nematicides

Nematicides (pesticides for nematodes) constitute a class of generally non-selective pesticides (Tiwari, 2025). Nematicides are of particular interest for the environment because they have an impact on humans, non-target organisms and the environment in general.
We can distinguish between two main groups of nematicides:

- **Traditional fumigants**: for instance, methyl bromide and 1,3-dichloropropene. They are linked to high environmental risks, such as ozone depletion and severe toxicity due to inhalation.
- **Contact nematicides**: such as carbamates and organophosphates. They are particularly dangerous for fish, birds, earthworms and other soil organisms.

Although absorption from the soil is not the same for the two classes of nematicides, the risk deriving from both is groundwater contamination via liking, especially in sandy soils with low organic matter.

Considering the importance of fighting nematodes in soils and the high toxicity of nematicides, researchers have been exploring alternatives extracted from plants (Anastasiadou et al., 2025). In particular, extracts from: *Melia azedarach* (chinaberry), rocket and parsley have shown negligible toxicity on model organisms like: *Eisenia fetida* (earthworm), *Daphnia magna* and zebrafish embryos.

1.3.1. Existing biocontrol strategies as alternatives to nematicides

To reduce the employ of synthetic pesticides, researchers have been recently exploring alternative biocontrol strategies for two main reasons: improving soil health and minimizing the impact of non-target organisms (Anastasiadou et al., 2025).
In particular, researchers have been working on the following biocontrol strategies (Tiwari, 2025):

- **Non-chemical methods**. These include: crop rotation (alternation of crops in a certain soil), green manuring (an agricultural method that consists of incorporating a rapidly growing, high biomass crop into the soil itself, in order to increase the organic matter content) and organic supplements (for example chitin or compost, that promote beneficial bacteria and free-nematodes growth).
- **Microbial-based control**:
    - **Nematophagous fungi**. These are fungi that feed on nematodes that parasite the nematode hosts. Among them we cite three fungi of the phylum *Ascomycota*: *Pochonia chlamydosporia*, *Hirsutella rhossiliensis*, and *Dactylella oviparasitica*.
    - **Other biological agents**: These are bacteria (like *Pasteuria penetrans*) and rhizosphere organisms (like *Pseudomonas spp.* and *Bacillus subtilis*) that produce toxins and can induce host resistance.

In addition to these methods, there exist nematode-trapping fungi (like *Orbiliales*, from *Ascomycota*) that have developed an architecture that allows them to trap and digest soil-dwelling nematodes from the outside (Jiang et al., 2017), as shown in **Figure 3**, retrieved from the same paper.

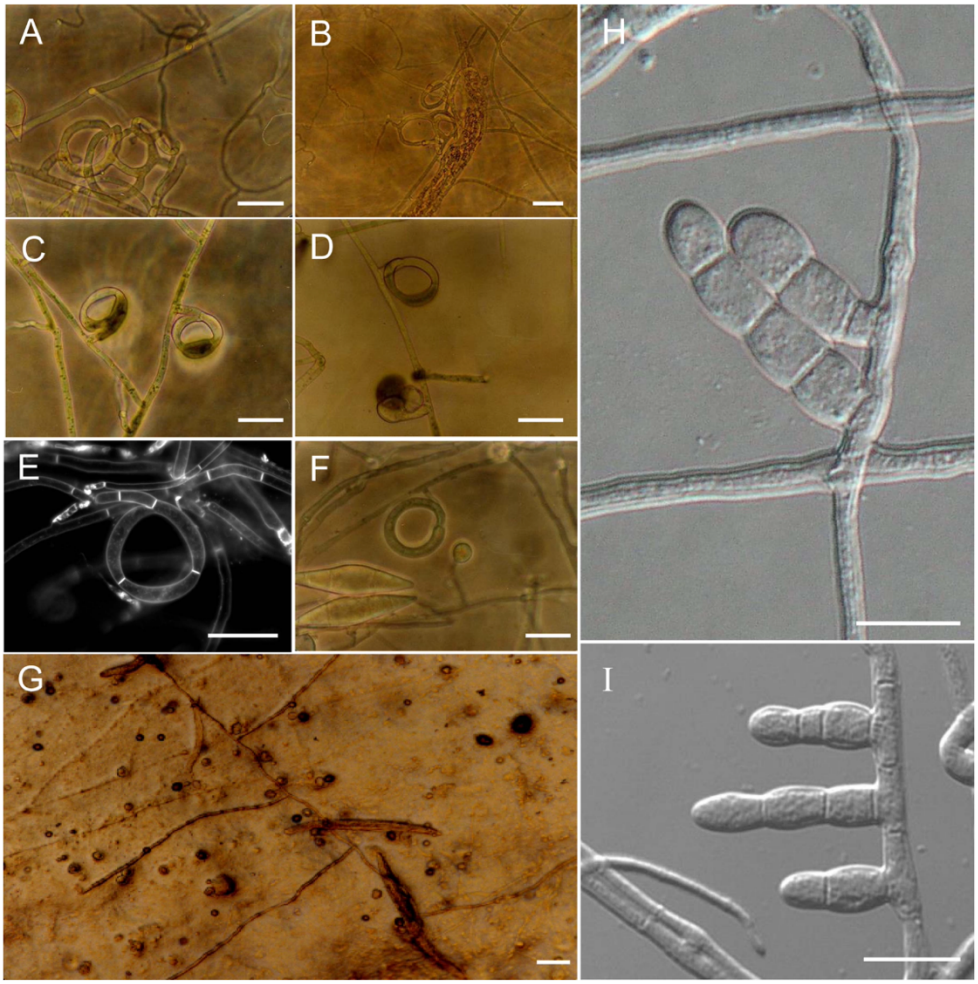


Figure 3. Nematode-trapping fungi: the structures (figure retrieved from (Jiang et al., 2017)). *(A, B, C, D, E, F, H, I: bar, 20 µm)(G: bar,40 µm).*

(A, B) The first two images show adhesive networks formed by the fungus Arthrobotrys oligospora. (C–E) From C to E, we can see examples of Drechslerella stenobrocha forming constricting rings. (F) Figure shows Dactylellina haptotyla producing adhesive knobs and non-constricting rings. (G) Figure D. haptotyla trapping a nematode. (H, I) Gamsylella cionopaga forming adhesive columns.

Nematodes in nutrient-poor, nitrogen-limiting habitats have developed an evolutionary adaptation: the secretion of pheromones called ascarosides. Trap formation is often induced by the detection of these pheromones. After pheromones detection, the fungus penetrates the nematode's cuticula by: mechanical pressure and extracellular enzymes (e.g. proteases). Subsequently, the fungus digests the nematode's internal tissues.

1.4. Effector proteins

The effector proteins represent the core of the intimate relationship these nematodes establish with the plant. To the current knowledge, it has been discovered that *M. incognita* uses three

glands (2 sub-ventrals and 1 dorsal) to produce the effectors that are then secreted in the plant (Bellafiore & Briggs, 2010). Effectors play several roles in the infection process, from the degradation of the plant cell wall to the reduction of the plant's defenses against the parasite. Bioinformatically, effectors proteins are identified based on the presence of a signal peptide for secretion and a lack of transmembrane region. However, considering only these criteria the lists of potential effectors is very long. Thus, additional features should be taken into account to shorten the list of candidates.

The study by (Vens et al., 2011) identified 4 sequence motifs enriched in the effectors of *M. incognita*. Scanning these motifs against the proteome identified 2,579 proteins, 80% of them contained also a signal peptide for secretion. Recently, using an improved and more complete version of the *M. incognita* genome assembly, these motifs were identified in more than 12,600 proteins (out of 43,700 in total), which represents an unrealistically high number (unpublished data from the hosting team). Their identification in a high number of other proteins suggests these motifs are not specific to effectors.

1.4.1. Experimental study of effector proteins

Effectors characterization is a multi-step experimental process that involves different wet lab techniques with aim to identify where the effector is produced and delivered, and its effect in the host (Mejias et al., 2019).

The main steps can be summarized as follows:

1. **Initial validation**: In Situ Hybridization (ISH). Standard procedure requires a first step of ISH to localize putative effector gene expression within the nematode's secretory esophageal glands (either sub-ventral or dorsal). Thanks to this step, the researcher can confirm if the protein is likely secreted through the stylet into the root.
2. **Localization in the host**: immunolocalization or transient expression. Researchers use immunolocalization to detect the protein inside the plant apoplast. They use transient expression to verify if the effector targets specific cellular compartments such as the nucleus; essentially, they induce a nematode's gene into a plant cell (either *Nicotiana benthamiana* leaves or *Arabidopsis* protoplasts), to see if the plant cell starts producing the effector protein.
3. **Identification of the host targets**: Yeast Two-Hybrid (Y2H) and Co-Immunoprecipitation (Co-IP) or Bimolecular Fluorescence Complementation (BiFC).

Effector proteins interfere with the normal physiology of the plant by interacting with certain host proteins. To explore this, researchers essentially use Y2H and often complete validation by Co-IP or BiFC. These two techniques allow to confirm the physical and molecular interactions between the nematode's effectors and the plant host proteins that were previously identified by Y2H.

1.4.1.1. Why wet laboratory validation of an effector protein is money and time consuming

As stated before, computationally, researchers generally identify putative effector proteins, based on the presence of a signal peptide for secretion and a lack of transmembrane region. If this is a relatively simple task, not only the list of putative effectors is extremely long ("putative", because they need to be validated by wet-lab techniques), but also the majority of these proteins are "pioneer proteins" with no functional domains (Mejias et al., 2019). These proteins usually do not look like any other previously studied, and to assess their function every single molecule needs to be tested from scratch in the laboratory, requiring time and money for each protein.

In addition to that, the great effort that is put in exploring the behavior of the effector in the host is not always sufficient. A lot of effectors are usually found by ISH, but only few are successfully deciphered. Only for few of the candidates we can actually explore the localization in the host and the host targets, which is crucial information to essentially understand how the effector works.

Another challenge in wet laboratory validation is proving the effector's role in parasitism. To do that, researchers must either: produce transgenic plant lines that either overexpresses the effector, or knockout the host target by RNAi/CRISPR (RNAi stands for RNA interference). These assays, monitoring the nematode development and exploring this over multiple generations is very time consuming.

Finally, to study into details the interactions between the nematode and the plant it is necessary to employ: high-end technology (e.g. confocal microscopy) and specialized cleaning techniques (e.g. Benzyl Alcohol and Benzyl Benzoate (BABB) clearing) to visualize the internal root structure where the effector acts.

1.5. Motifs

Motifs are short protein or nucleotide sequences with crucial functional or structural roles in the macromolecule (Chepsergon et al., 2022). They are often involved in many dynamic networks, like the regulation pathways in cells (Davey et al., 2012).

Sequence motifs are usually tiny with constant size and are often repeated and conserved. Typically, motifs conform to a particular sequence pattern, where a certain position can be constrained to a specific amino-acid, whereas others are not (Davey et al., 2015). In this case the motif is called degenerate (Roberson, 2018).

1.5.1. Well studied motifs in effectors of plant parasites

Conserved motifs in effector proteins are a great resource to explore plant parasites, since they contribute in understanding how the pathogen manipulates the host plant. Across different kingdoms, there are several well-characterized motifs that have been identified as molecular determinants of virulence of the parasites (Calia et al., 2024).

1.5.1.1. Gram-negative bacterial pathogens

Gram-negative bacterial pathogens secrete effectors via the Type III Secretion System (T3SS) which are often characterized by specific motifs or domains that play a major role in controlling virulence and targeting proteins in the host plant. These is not be a subject of interest of the present work.

1.5.1.2. Fungal pathogens

Fungal pathogens secrete effectors that are often described in the literature as small proteins, that are characterized by cysteine-rich sequences. These is not be a subject of interest of the present work either.

1.5.1.3. Oomycete pathogens

Oomycete pathogens are a eukaryotic filamentous heterotrophic phylum (*Oomycota* in general) of the kingdom *Stramenopila*, and are among the most deeply studied due to their high impact as plant parasites.

In particular, they display two main families of effector proteins that are defined by specific conserved motifs in the N-terminal region, downstream of the signal peptide:

- **RxLR**: marked by the RxLR and the -dEER motifs.
- **Crinkler (CRN)**: marked by the LxLFLAK and the HVLVxxP motifs (start and end of the DWL domain).

1.5.1.4. PPNs

To identify universal consensus motifs across multiple species is a harder challenge than it is for the abovementioned pathogens. However, researchers have found some specific motifs enriched in effectors of *M. incognita*.

- LIIS, EGAG, ASKY and AEGD motifs. These were identified by the bioinformatics tool MERCI (Vens et al., 2011). These were initially found as characteristic, but by the study of more complete genome assemblies, they would later identify a too high number of putative effectors, suggesting a lack of specificity of these motifs.
- Mel-DOG box, a cis-regulatory promoter motif. It was identified as a feature of the effectors that are expressed in the dorsal gland.

1.5.2. Motif mining

There are two categories of bioinformatic algorithms used to identify motifs: generative and discriminative (Huggins et al., 2011). The first searches for motifs enriched in sub-sequences of the dataset. Whereas, the discriminative methods use two datasets (a positive and a negative) where motifs are identified as those enriched in the positive dataset and very few, or ideally absent, in the negative one. The introduction of a second dataset has enabled bioinformaticians to find elements that would be typical of a group of sequences and not of another, giving higher chances to be biologically validated.

Although several discriminative approaches are available for nucleotides sequences, very few are available for protein sequences. MERCI uses a graph-based approach based on physicochemical features of the amino-acids composing proteins sequences and then selects the motifs with high occurrences in the positive dataset and absent in the negative one (Vens et al., 2011). STREME, from the MEME-suite (Bailey et al., 2015), employs a tree-based method to identify discriminant motifs (Bailey, 2021). Finally, DiMotif first employs a peptide-pair encoding, a technique commonly used in Natural Language Processing to segment the sequences, and then a chi-square test to assess the significant motifs (Asgari et al., 2019).

2. Chapter 2: Objective of the study

Effector proteins are crucial in understanding plant-pathogen interactions. Despite their importance, their accurate detection in nematode genomes is challenging. Hence, the main aim of my internship was to find sequence motifs that would help to characterize effector proteins in *M. incognita* using two manually curated datasets generated by the hosting laboratory.

To achieve this goal, I performed the following tasks:

- Exploring the characteristics of the sequences of known effector proteins of *M. incognita.*
- Development of a bioinformatic pipeline to identify the discriminant motif(s).
- Characterization of the properties of the identified motif(s).
- Identification of putative new effector protein candidates in *M. incognita*.

3. Chapter 3: Materials and methods

3.1. Data treated to develop MOnSTER

I developed several scripts in Python language (version 3.9) in the form of Jupiter notebook. All notebooks and datasets are available at: https://github.com/paolaporracciolo/MOnSTER_PROMOCA.git

3.1.1. Datasets

I used two datasets built from two close related species from the *Meloidogyne* genus: *M. incognita* and *Meloidogyne arenaria*. The first was used to set up the pipeline and identify the motifs, the second to study the discriminant power of the identified motifs.

3.1.1.1. M. incognita

The hosting team prepared two manually curated datasets by literature mining: one composed of 161 protein sequences that represent known effectors for this species and one composed of 495 protein sequences that are not effectors. From now on, I will refer to these two datasets as "positive" and "negative", respectively.

3.1.1.2. M. arenaria

Similarly, I disposed of a positive dataset composed of 127 protein sequences and a negative one of 673 protein sequences, provided by the hosting team after manual revision.

3.1.1.3. Proteome of M. incognita

The proteome of *M. incognita* is constituted of 43,718 proteins predicted from its genome (UniProt ID: UP000887563, ENA ID: ERP009887) (Blanc-Mathieu et al., 2017).

3.2. Data treated to test the robustness of MOnSTER

After my Erasmus+ internship, the scope of the research was significantly expanded to validate the robustness of the pipeline MOnSTER, using much larger datasets across more species of nematodes and oomycetes.

3.2.1. Oomycetes

As a proof of concept of the pipeline, we applied it to 5 species of oomycetes: *Phytophthora infestans*, *Phytophthora sojae*, *Phytophthora ramorum*, *Hyaloperonospora arabidopsidis* and *Bremia lactucae*. This dataset included 1,743 effector proteins for the positive dataset, and 3,009 non-effectors for the negative dataset.

3.2.2. Other PPNs

We extended the research in PPNs by increasing the number of species and the size of the datasets: we included 546 well-known putative proteins involved in parasitism for the positive dataset, and 3,849 non-effector sequences for the negative dataset, covering 13 different species from different genera: *Meloidogyne*, *Globodera*, *Heterodera*, *Radopholus*, and *Bursaphelenchus*.

3.3. MOnSTER pipeline

The scope of the MOnSTER pipeline is to identify discriminant clusters of motifs, CLUMPs, of proteins. It is composed of four main steps as described in **Figure 4A** and in the following paragraphs.

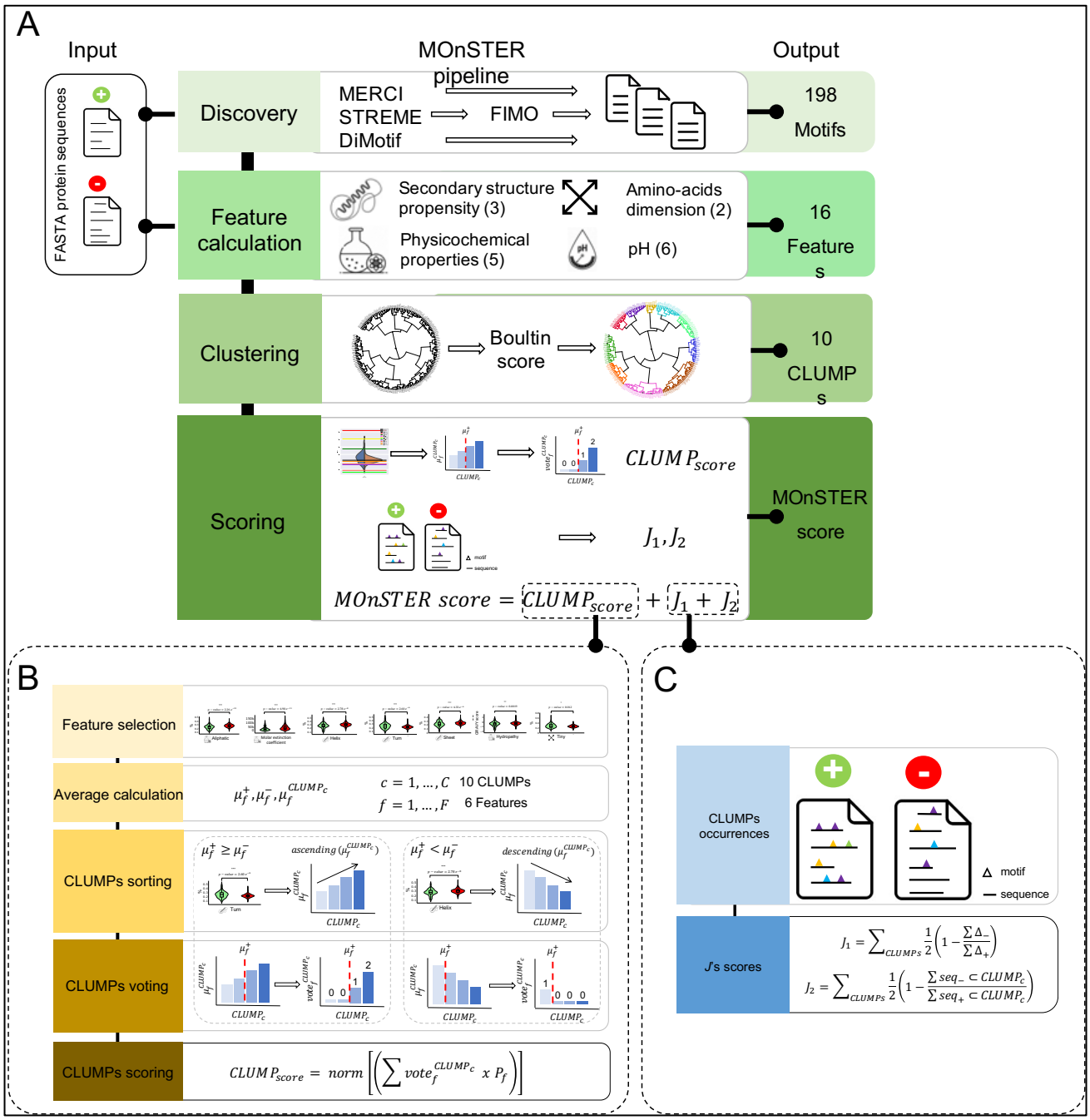


Figure 4. MOnSTER pipeline scheme.

(A) The MOnSTER pipeline consists of four sequential steps: it takes two FASTA protein sequences datasets as input. It generated, as output, a list of CLUMPs, each associated with a MOnSTER score. The MOnSTER score integrates two components: (B) CLUMP score calculation. (C) 2 modified Jaccard indexes.

3.3.1. Motif Discovery

The first step of the pipeline regards the identification of discriminant motifs enriched in the sequences of the positive dataset compared to the sequences of the negative dataset. Thus, I

used three tools by command line version with default parameters: MERCI, DiMotif and STREME (where I set up the motifs' length between 3 and 5 amino acids (AA)).

Unlike the other two, STREME's output is a list of degenerated motifs. Hence, I used the tool FIMO with default parameters to extract 91 motifs from the 7 degenerated.
Finally, I obtained the following numbers of motifs: 10 with MERCI, 100 with DiMotif and 91 with STREME.
Then, I developed a script to extract the motifs, where I convert all of them to capital letters, remove the identical motifs and create a single list containing all the motifs in the same format, hence obtaining 198 motifs.

3.3.2. Feature calculation

The second step of the pipeline concerns the calculation of parameters that describe protein sequences. Thus, I selected 16 features belonging to 4 categories:

- secondary structure propensity ('helix', 'turn', and 'sheet').
- amino-acids dimensions ('tiny' and 'small').
- pH ('basic', 'acid', 'charge at pH', 'isoelectric point', and 'charged').
- physicochemical properties ('hydropathy', 'molar extinction coefficient', 'polar', 'non-polar', 'aromatic', and 'aliphatic').

Each property is extensively described in **Figure 5**.

A

Category	Symbol	Feature	Definition
secondary structure propensity		helix	cumulative percentage of amino acids which tend to be in Helix (e.g. 'V', 'I', 'Y', 'F', 'W', 'L')
		turn	cumulative percentage of amino acids which tend to be in turn (e.g 'N', 'P', 'G', 'S')
		sheet	cumulative percentage of amino acids which tend to be in sheet (e.g 'E', 'M', 'A', 'L')
amino acids dimensions		tiny	cumulative percentage of the smallest amino acids (e.g 'A', 'C', 'G', 'S', 'T')
		small	cumulative percentage of the small amino acids (e.g A', 'C', 'F', 'G', 'I', 'L', 'M', 'P', 'V', 'W', 'Y')
pH		basic	cumulative percentage of the basic amino acids (e.g 'H', 'K', 'R')
		acid	cumulative percentage of the acidic amino acids (e.g 'B', 'D', 'E', 'Z')
		charge at pH	calculate the charge of a protein at given pH (default : 7.5)
		isoelectric point	calculate the isoelectric point
		charged	cumulative percentage of the charged amino acids (e.g 'B', 'D', 'E', 'H', 'K', 'R', 'Z')
physicochemical properties		hydropathy	(grand average of hydropathy) calculated by adding the hydropathy value for each residue and dividing by the length of the sequence
		molecular extinction coefficient	calculates the molar extinction coefficient assuming cysteines (reduced) and cystines residues (Cys-Cys-bond)
		polar	cumulative percentage of the polar amino acids (e.g 'D','E', 'H', 'K', 'N', 'Q', 'R', 'S', 'T', 'Z')
		non-polar	cumulative percentage of the non-polar amino acids (e.g 'A', 'C', 'F', 'G', 'I', 'L', 'M', 'P', 'V', 'W', 'Y')
		aromatic	cumulative percentage of the aromatic amino acids (e.g 'F', 'H', 'W', 'Y')
		aliphatic	cumulative percentage of the aliphatic amino acids (e.g 'A', 'I', 'L', 'V')

B

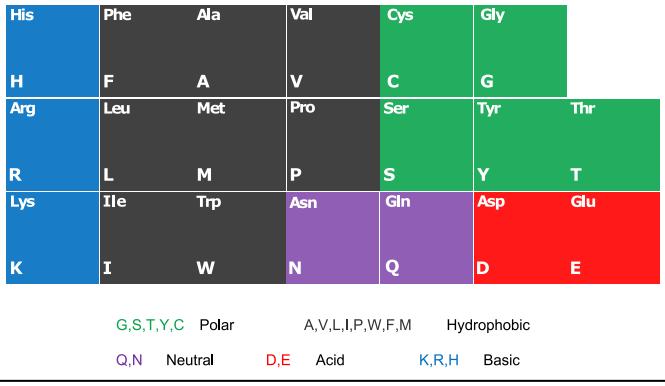


Figure 5. AA properties and categories of AA features.

(A) The table shows the 16 features used for feature calculation. (B) The scheme shows the categories of AA from a chemical point of view.

Error! Reference source not found.I performed feature calculation on two inputs: the two datasets (positive and negative) and the list of motifs. Specifically, I conducted the feature calculation using the python module *ProteinAnalysis* imported from the *Bio.SeqUtils.ProtParam* module of the Bio package.

At the end of this step, I obtained three tables of features, one for each of the input datasets (positive, negative and the list of motifs).

3.3.3. Clustering

This step allowed to cluster motifs based on their properties described by the 16 features. To make the features comparable to each other, I performed data normalization by using the module *fit_transform* imported from the *sklearn.preprocessing.StandardScaler* module (Pedregosa et al., n.d.). This normalization consists in the removal of the mean and the scaling to unit variance.

Then, I performed a hierarchical clustering, using the Euclidian distance. Afterwards, to find a threshold to identify clusters, I employed the Davies-Bouldin score (Davies & Bouldin, 1979). For each CLUMP, I removed the redundant motifs. Briefly, I identified motifs that shared a core sequence (for example: 'HWT in HWTQ' and in 'GHWTQ'), and I only retained the cores (for instance: "HWT") in the CLUMPs. From an original set of 198, I obtained 177 motifs.

3.3.4. Scoring

The final objective is to identify the CLUMP(s) with the highest discriminative power with respect to positive dataset. Thus, I conceived a new score called MOnSTER score, to sort CLUMPs by their discriminate power.

MOnSTER score is composed of three parts: the CLUMP score and two modified version of the Jaccard index.

3.3.4.1. CLUMP score

This score considers the AA composition of the motifs belonging to each CLUMP with respect to the preferences of the sequences of the positive dataset. The procedure that I implemented to calculate this score is showed in **Figure 4B**.

a) Feature selection

I did not consider the following features: 'charge at pH', 'isoelectric point', and 'molar extinction coefficient' as they are not adapted for short sequences like motifs.

Then, I used the Mann-Whitney test to find those features whose values were significantly different between the positive and negative datasets. I only retained the statistically significant features, with a p-value < 0.05, namely: aliphatic, helix, turn, sheet, hydropathy and tiny. Then,

I assigned them a score, by calculating -Log(p-value) of each feature. I will refer to it as the 'feature weight'.

b) Average calculation

For each of the 6 selected features, I calculated the average value for: the positive dataset, the negative dataset and each CLUMP, that I will refer to with this notation: μ_f^+, μ_f^- and $\mu_f^{CLUMP_c}$, respectively.

c) CLUMPs sorting

I compared the averages of the positive and negative dataset for each feature and sort CLUMPs accordingly.

Thus, if the $\mu_f^+ \geq \mu_f^-$, I would sort CLUMPs averages in ascending order.

Otherwise ($\mu_f^+ < \mu_f^-$), I would sort CLUMPs averages in descending order.

d) CLUMPs voting

In this step, for each feature, I divided the CLUMPs into two groups accordingly to the following statements:

If $\mu_f^+ \geq \mu_f^-$: CLUMPs with $\mu_f^{CLUMP_c} \geq \mu_f^+$ have a vote from 1 to the number of CLUMPs with increment of 1, otherwise the score is set to 0.

If $\mu_f^+ < \mu_f^-$: CLUMPs with $\mu_f^{CLUMP_c} < \mu_f^+$ the vote attributed goes from 1 to the number of CLUMPs, otherwise it is 0.

e) CLUMPs scoring

For each CLUMP, for each feature, I multiplied the 'feature weight' by the CLUMPs vote then I summed all the results using the following formula:

$$CLUMP_{score} = norm\left[\left(\sum vote_f^{CLUMP_c} \, x \, P_f\right)\right]$$

where I applied a normalization to have a range between 0 and 1.

3.3.4.2. Modified Jaccard indexes

The two modified Jaccard scores take into account: i) the occurrences of the motifs, for each CLUMP, in the positive dataset compared to the negative, and ii) the number of positive sequences containing the motifs in each CLUMP with respect to the negatives (**Figure 4C**).

a) CLUMPs occurrences

I calculated the occurrences of the motifs in each CLUMPs in the two datasets (positive and negative).

b) J's scores

The Jaccard index consists in calculating the similarity between two sets. Here I propose two ways to calculate the J index that will be called J1 and J2 hereafter.

To obtain J_1, I calculated the number of occurrences of the motifs for each CLUMP in the negative dataset over the number of occurrences of the motifs of the CLUMP in the positive dataset, using the following equation:

$$J_1 = \sum_{CLUMPs} \frac{1}{2}\left(1 - \frac{\sum \Delta_-}{\sum \Delta_+}\right)$$

Where:

Δ_- and Δ_+ the number of occurrences of the motifs of the CLUMP in the negative or in the positive dataset, respetively.

To obtain J_2, for each CLUMP, I calculated the number of sequences of the positive dataset that contain at least a motif of the CLUMP, over the number of sequences of the negative dataset that contain at least a motif of the CLUMP, accordingly to the following formula:

$$J_2 = \sum_{CLUMPs} \frac{1}{2}\left(1 - \frac{\sum seq_- \subset CLUMP_c}{\sum seq_+ \subset CLUMP_c}\right)$$

Where:

seq_- is the number of sequences of the negative dataset containing at least a motif of the CLUMP.

seq_+ is the number of sequences of the positive dataset containing at least a motif of the CLUMP.

The ½ is applied to have values between 0 and 0.5 for each J in order to have equal weight in the final score, and (1 – Jaccard Index) is to consider the degree of dissimilarity rather than similarity.

3.3.4.3. MOnSTER score

The final MOnSTER score, for each CLUMP, is the sum of: the CLUMP score, and the two J's indexes:

$$MOnSTER\ score = CLUMP_{score} + J_1 + J_2$$

3.3.5. CLUMPs positions in sequences

To study at which positions CLUMPs occur within the sequences of the positive and negative datasets I coded a script in which first I calculated the start positions of each motif in each sequence. Afterwards, I normalized the start positions by dividing their value by the length of the sequence. For each sequence, I split its length into 10 bins of equal range. Then I counted how many motifs in each CLUMP fell in each bin and reported the results as a bar plot.
The output of this analysis shows in which portion of the sequences the motifs of the CLUMPs are found.

3.3.6. CLUMPs co-occurrences

To study if CLUMPs tend to occur together in the positive and negative sequences, I developed a script where for each CLUMP I calculated how many sequences contained only one occurrence of its motifs, co-occurrences of two or more motifs within the same CLUMP and co-occurrences of two or more motifs within different CLUMPS. The results are represented with an upset plot, using the module *upsetplot* (Nothman, n.d.).

4. Chapter 4: Results

4.1. Results obtained when developing MOnSTER

4.1.1. Sequence composition preferences of the effectors of *M. incognita*

The main task of my internship was to identify effectors-specific protein sequence motif(s) of *M. incognita,* a plant-parasite nematode. I focused on this one because it is well-known among the agri-environmental community for its huge damage on crops, as described in the introduction. Briefly, effectors are proteins involved in the inflammatory and parasitic process. Hence, being specific to parasites and essential for parasitism, effector proteins constitute targets of choice for the development of cleaner and more specific control methods.

To perform the analysis, I disposed of two datasets whose sequences were manually selected from literature mining: the positive, composed of 161 protein sequences (known effectors), and the negative, composed of 495 protein sequences (unlikely to be effectors).

First, I explored the characteristics of the sequences composing the two datasets. Specifically, I calculated 16 different characteristics or features belonging to 4 classes: physicochemical properties, pH, amino-acids dimension, secondary structure propensity (see methods, and **Figure 5**).

In **Figure 6**, I show the distribution of the features for the two datasets as violin plots, reporting only the ones with significant difference (p-value < 0.05). The distribution of the other features is reported in **Figure 7**.

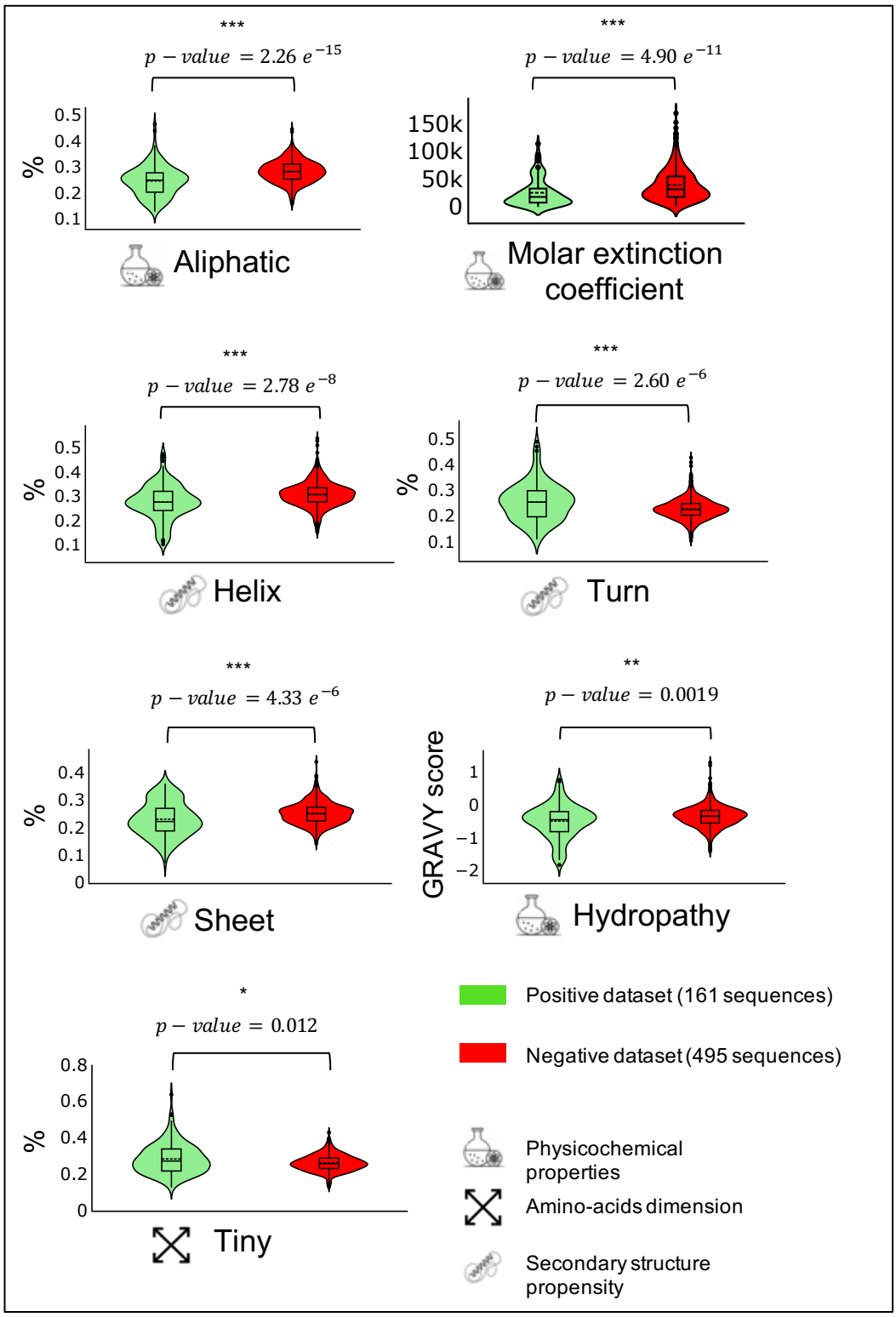


Figure 6. Distribution of significant feature values in the positive and negative dataset.

*The violin plots represent the distribution of features' values on the whole positive and negative datasets. The green violins represent the positive dataset and the red, the negative one. Inside the violins, the boxplots report the averages (dashed lines) and medians. Plots are ordered in terms of significance, based on the p-values reported above each pair (Mann-Whitney test). The category of the feature is reported with the correspondent symbol. * (0.05, 0.01], ** (0.01, 0.001], *** < 0.001.*

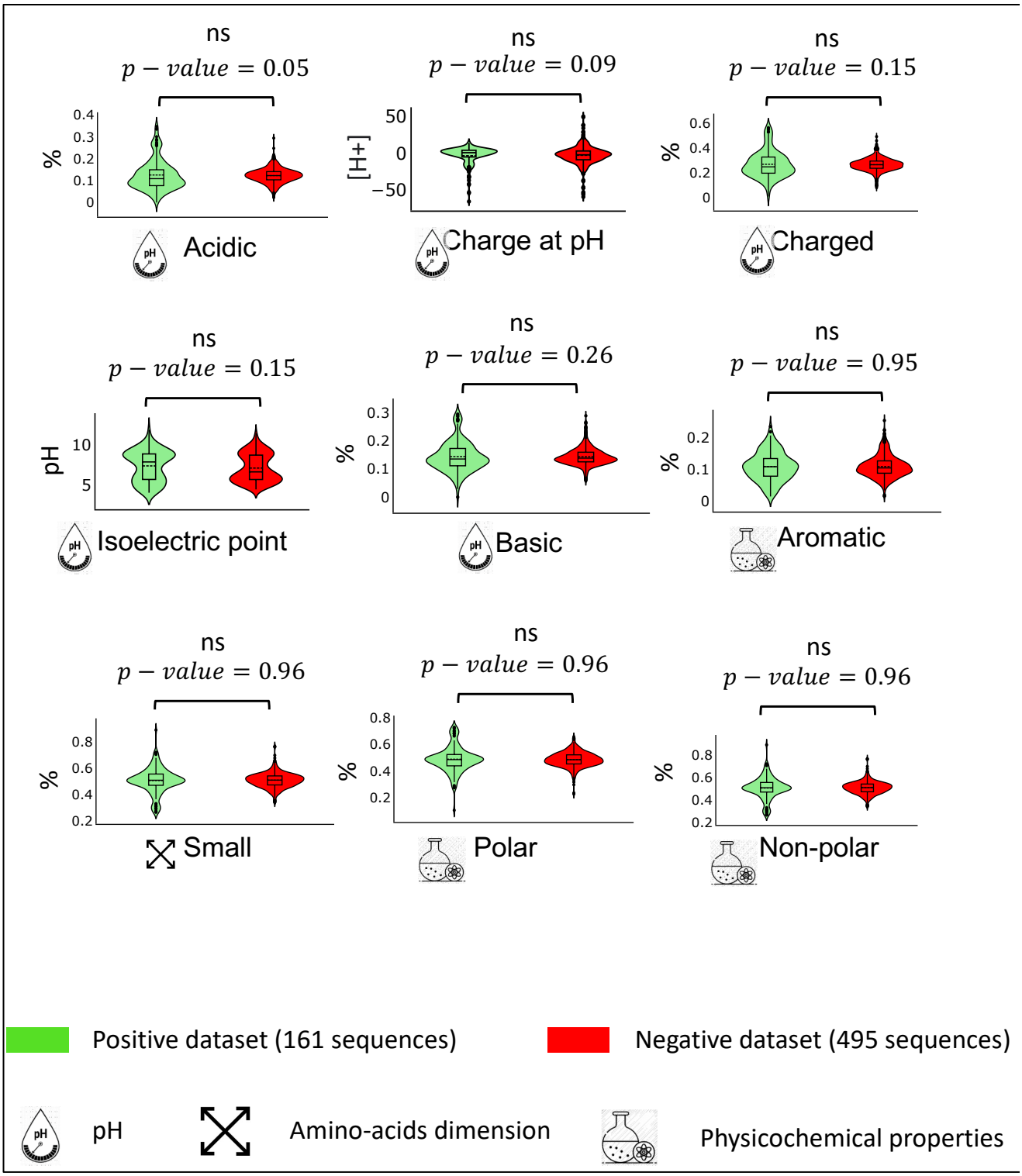


Figure 7. Distribution of non-significant feature values in the positive and negative dataset.

The violin plots represent the distribution of features' values on the whole positive and negative datasets. The green violins represent the positive dataset and the red, the negative one. Inside the violins, the boxplots report the averages (dashed lines) and medians. Plots are ordered in terms of significance, based on the p-values reported above each pair (Mann-Whitney test). The category of the feature is reported with the correspondent symbol. ns > 0.05

First, I noticed that none of the features belonging to the category "pH" is significant, suggesting that effectors proteins do not have any specificity regarding this category. Interestingly, all 3 features belonging to the category "secondary structure propensity" have reported significant p-values. I observe a positive enrichment of AAs with a preference for turn structure in the positive dataset with respect to the negative one. On the other hand, AAs with preferences for the other two kinds of secondary structures (alpha helix and beta sheets) are underrepresented in the positive sequences compared to the negative ones. I also observe a positive enrichment of tiny AAs. Regarding the physiochemical properties, three of them are significant. The calculation of the hydropathy using the GRAVY score (grand average of hydropathy) (Kyte & Doolittle, 1982) suggests that effectors protein sequences have the tendency to prefer hydrophilic AAs. Similarly, they do not prefer aliphatic AAs. Finally, the effector proteins seem to have lower molecular extinction coefficient (estimation of how strongly the protein absorbs light at a particular wavelength) compared to the proteins in the negative dataset. However further analyses are needed to understand more the role of this finding regarding effectors proteins.

These results suggest that effector protein sequences have a peculiar composition with specific AA propensities. This prompted us to search for sequence motifs discriminant of effector proteins.

4.1.2. MOnSTER allowed to identify 6 CLUMPS of motifs discriminant for the effectors

I ran my pipeline on the two aforementioned datasets (see methods). First, the pipeline led to discovering 198 different (non-identical) motifs. The length of these motifs ranges from 2 to 5 AAs.

To investigate whether motifs share similar properties and thus can be grouped together, I used a hierarchical clustering approach. Briefly, a hierarchical clustering is applied on the motifs represented by the 16 features used to describe the properties of the sequences of the two datasets. Since the output of the hierarchical clustering is a dendrogram, I had to find a criterium to cut the tree in clusters. I employed the Davies-Bouldin score that is a metric to calculate if the clustering grouped elements in sufficiently different groups. As reported in **Figure 8,** the best Davies-Bouldin score led to the identification of 10 CLUMPs.

height of cut	davies_bouldin_score
2	325.33
5	119.95
9	34.10
10	22.90
11	24.81

Figure 8. Table of Davies-Bouldin score results.

The table shows the results of the Davies-Bouldin score calculation with different heights of cuts. The green box highlights the chosen threshold.

In **Figure 9**, I can observe two main clades. The upper clade (CLUMPs 9,6,0,1,4) with mainly motifs composed by hydrophobic or polar AAs and the lower clade (CLUMPs 8,3,5,2,7) with motifs composed mainly by acid or basic AAs.

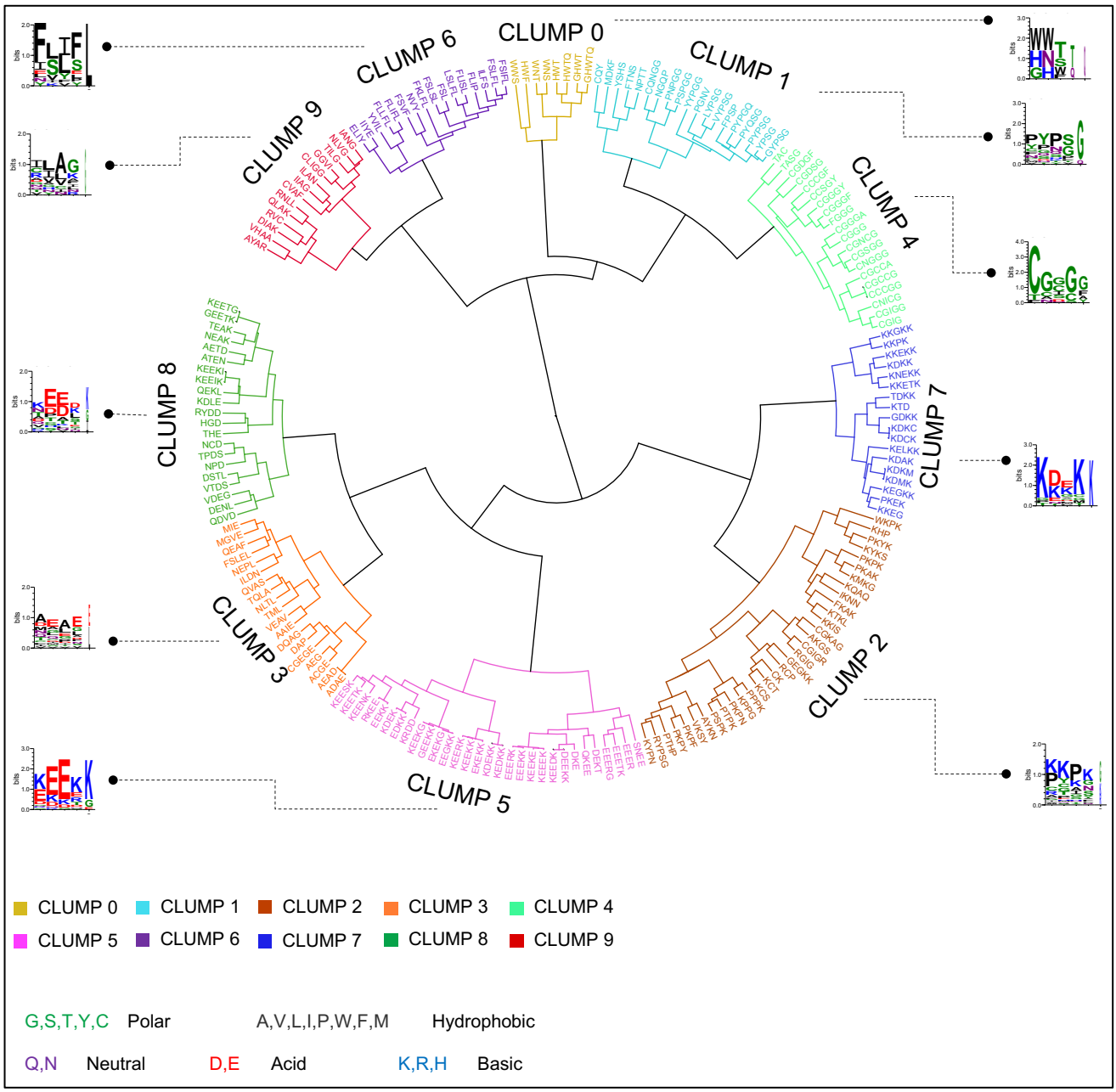


Figure 9. Motifs grouped in CLUMPs.

Motif clustering is represented as a dendrogram tree, where each leaf is a motif. Motifs are grouped in CLUMPs, based on their physicochemical properties. CLUMPs are represented by a color code reported in the legend. CLUMPs' motifs degeneration is represented as a logo, where at each position, the height of a letter corresponds to the probability to find a certain AA, colored by their chemical composition, as reported in the legend.

Then I computed the MOnSTER score to associate a score to each CLUMP (fourth step of the pipeline). This score takes into account three components: 1) the AA composition of the motifs belonging to the CLUMP, with respect to the preferences of the sequences of the positive dataset, 2) the occurrences of the motifs for each CLUMP in the positive dataset compared to

the negative, 3) the number of positive sequences containing the motifs in the CLUMP with respect to the negatives. The results of the MOnSTER scores are presented in **Figure 10**.

The first three CLUMPS accordingly to the MOnSTER score are 1, 0 and 4. Importantly, they belong to the same clade in the tree (**Figure 9**) and the motifs of these CLUMPs are mainly composed of polar AAs, which is in line with the positive dataset sequences composition preferences. Although these CLUMPs have similar or lower numbers of occurrences in the positive sequences than the other CLUMPs, their occurrences in the negative dataset are very limited. The three following CLUMPs 5, 7, and 8, mainly group motifs composed globally of basic or acid AAs.

Finally, CLUMPs 2, 9, 3 and 6, which tend to be composed mainly by hydrophobic AAs, are at the bottom of the MOnSTER score list. Considering that our previous analysis showed that effectors proteins have preferences for hydrophilic AAs and that the occurrences of these CLUMPs in the negative datasets are similar to the ones in the positive, I decided to not consider these four CLUMPs to pursue the analysis. I thus set a threshold of 1 to the MOnSTER score to select the CLUMPs with the highest discriminative power.

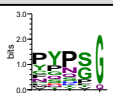
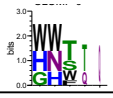
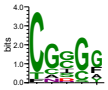
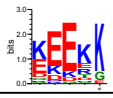
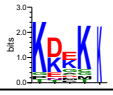
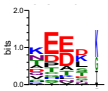
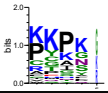
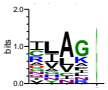
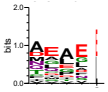
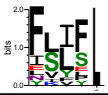
CLUMPs	Logo	MOnSTER score	# Motifs (D-M-S)*	# occurrences in positive dataset (%)	# occurrences in negative dataset (%)
1		1,667	20 (8-2-10)	48 (30)	9 (2)
0		1,584	20 (5-0-15)	37 (23)	16 (3)
4		1,579	9 (5-4-0)	20 (12)	10 (2)
5		1,563	30 (8-0-22)	38 (24)	77 (15)
7		1,531	18 (2-0-16)	52 (32)	52 (10)
8		1,004	21 (17-0-4)	59 (36)	72 (14)
2		0,991	34 (19-2-13)	101 (62)	191 (38)
9		0,757	14 (13-0-1)	52 (32)	54 (10)
3		0,717	19 (15-2-2)	57 (35)	88 (18)
6		0,690	16 (8-0-8)	57 (35)	65 (13)
* D→ DiMotif; M→ MERCI; S→ STREME					

Figure 10. results of MOnSTER score calculations.

The table resumes the results of the MOnSTER score calculations by 6 columns. Column 1: CLUMP number. Column 2: Logo of the motifs composing the CLUMP, (see ***Figure 9*** *for further information). Column 3: CLUMP's MOnSTER score. Column 4: Total number of motifs in the CLUMP and, in parenthesis, the number of motifs from DiMotif, MERCI and STREME. Column 5: Number of occurrences of the CLUMP in the positive dataset and in parenthesis the respective percentage. Column 6: Number of occurrences of the CLUMP in the negative dataset and in parenthesis the respective percentage.*

The MOnSTER pipeline allowed the identification of 6 CLUMPs characteristic of effector proteins.

4.1.3. Best scored CLUMPs (1,0,4) occur at central positions in sequences of effectors

To explore the characteristics of the identified CLUMPs, I started by studying their position in the protein sequences of the two datasets. My hypothesis is that: if CLUMPs occur specifically at preferred positions within the protein sequence, this could reflect a biological meaning in terms of functions or structure.

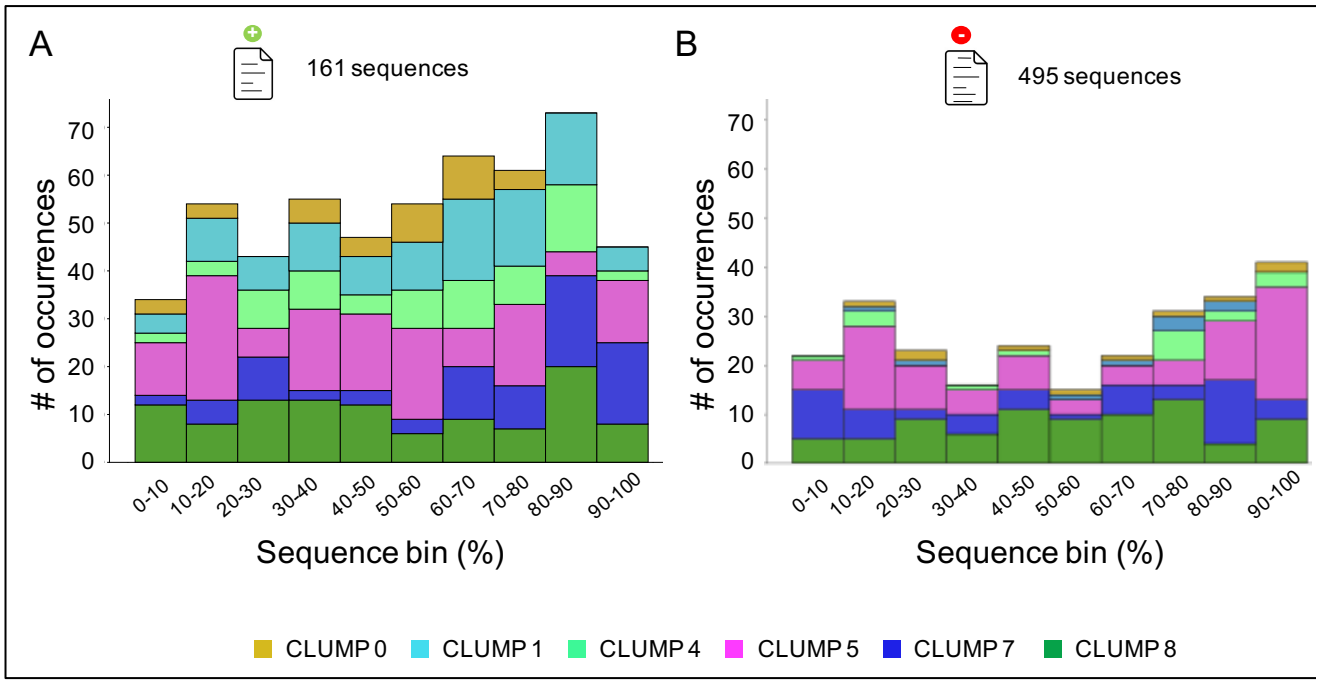


Figure 11. Histograms of positions of the CLUMPs in the sequences.

The two histograms show the positions of the CLUMPs in the sequences of the positive (A) and the negative (B) datasets. On the x-axis, the sequences length is divided into 10 bins of equal range. On the y-axis, the number of occurrences is reported for each CLUMP. CLUMPs are represented by a color code reported in the legend.

Figure 11 represents CLUMPs occurrences along the positive (**Figure 11A**) and negative (**Figure 11B**) dataset sequences. Overall, we observe a difference in the position preferences of the best scored CLUMPs by MOnSTER in the positive dataset, with a tendency to be in central positions of the sequences, compared to the negative one. While no occurrences are observed for CLUMP 0 at the last bins of the sequences of the positive dataset, the opposite scenario is remarked in the negative ones. CLUMP 1 and 4 mainly occur in very central positions in the positive dataset, unlike the negative dataset.

Surprisingly, no differences in the position preferences are observed for CLUMP 5, 7, 8 between the positive and the negative datasets. This observation suggests that these CLUMPs are less effectors-specific with respect to the CLUMPs with the highest score.

4.1.4. CLUMPs 1, 0 and 4 co-occur in effector sequences at small distances within each other

To continue the characterization of the properties of the CLUMPs, I studied how many occurrences of CLUMPs are present in the sequences of the two datasets. Co-occurrences of motifs may be related to a biological or structural role they may play together in the sequences. To perform this analysis, I counted how many sequences show co-occurrences of motifs within the same CLUMP or of two or more different CLUMPS. To test all possible combinations of co-occurrences, I used the upset plots (**Figure 12**).

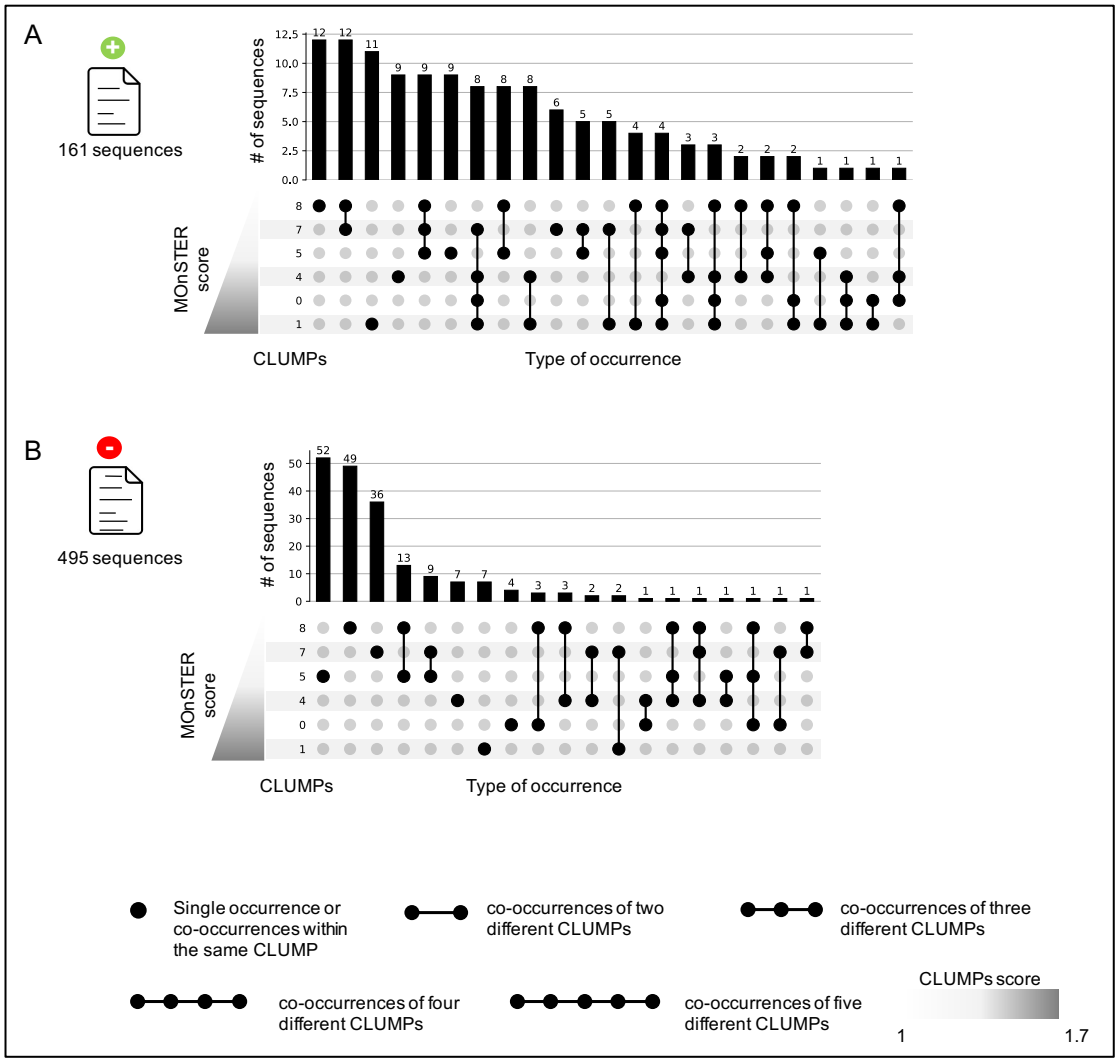


Figure 12. Upset plots of co-occurrences of CLUMPs in the datasets' sequences.

The two upset plots show the co-occurrences of the CLUMPs in the sequences of the positive (A) and the negative (B) datasets. The type of occurrence is reported in the bottom part of the plot, as showed in the legend one point represent single occurrences or co-occurrence within the same CLUMP, two points linked by a line represent co-occurrence of two CLUMPs on the same sequence and so on. On the left I show the CLUMPs sorted ascendingly by MOnSTER score. In the upper part of the plot, the bar plot shows the number of sequences that contain a specific co-occurrence of CLUMPs accordingly to the type of combinations of co-occurrences reported in the bottom part.

Overall, I notice that CLUMPs' tend to co-occur more in the sequences of the positive dataset than the negative one. In support of this observation, CLUMP 0 shows only co-occurrences with other CLUMPs in the positive dataset, while very few co-occurrences are reported in the negative dataset. Importantly, the top scored CLUMPs 1, 0 and 4 show only one co-occurrence in the negative dataset. At the same time, several are reported in the positive, with different combinations (all pairs of co-occurrences of two and co-occurrences of the three). For the other CLUMPs, single co-occurrences are privileged in the negative dataset, while very few are reported in the positive.

I then focused on the top three CLUMPs. First, I wanted to see whether there is a relationship between the number of co-occurrences and the length of the sequences. The absence of this relationship as showed in **Figure 13** suggests that the number of co-occurrence of CLUMPs do not systematically increase with the sequence length but specific properties rule these phenomena.

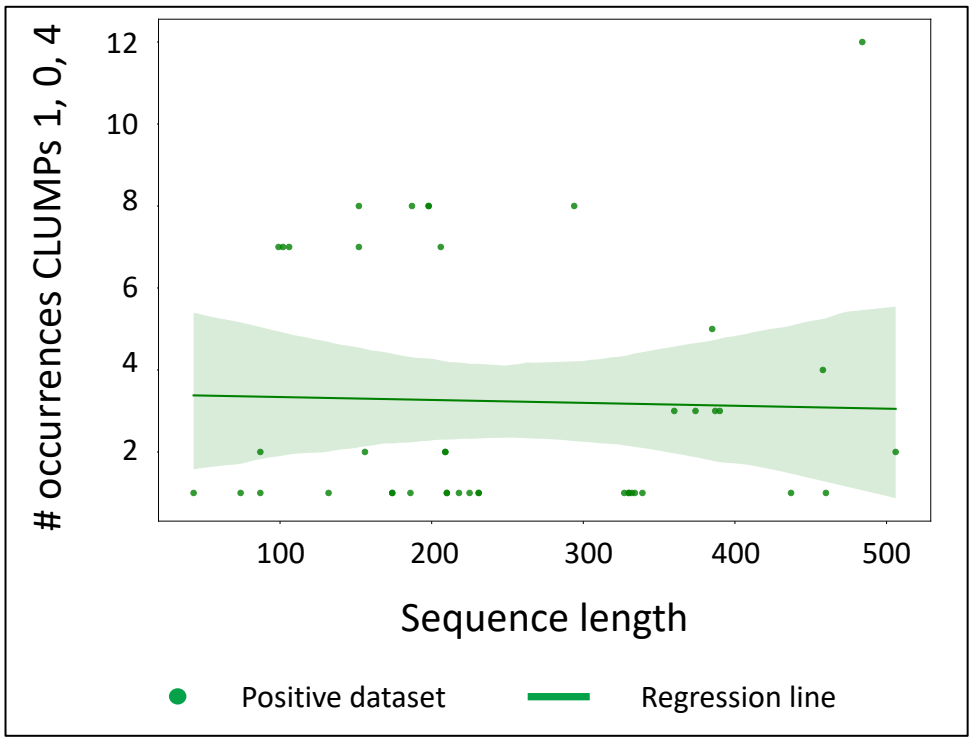


Figure 13. Regression plot of the number occurrences of CLUMPs 1, 0, 4 as a function of the sequence lengths of the positive dataset sequences.

This plot reports both the points corresponding to elements of the sample and the corresponding regression line. The points show at a certain sequence length, how many occurrences of CLUMPs 1, 0, 4 there are. The regression line explores whether there is: an augmentation, a diminution or no effect on the variable y (number of occurrences of CLUMPs 1, 0, 4) at the augmentation of the values of variable x (sequence length).

Then I studied the distances between co-occurrences. As reported in **Figure 14**, co-occurrences tend to happen at very small distances suggesting i) either structural purposes to allow a specific folding of effector proteins, ii) or biochemical reasons as for instance an active site of interaction. Further analysis is needed to investigate these findings. For instance, we could inspect tridimensional structures of effectors proteins containing these co-occurrences of CLUMPs to see where these motifs are located.

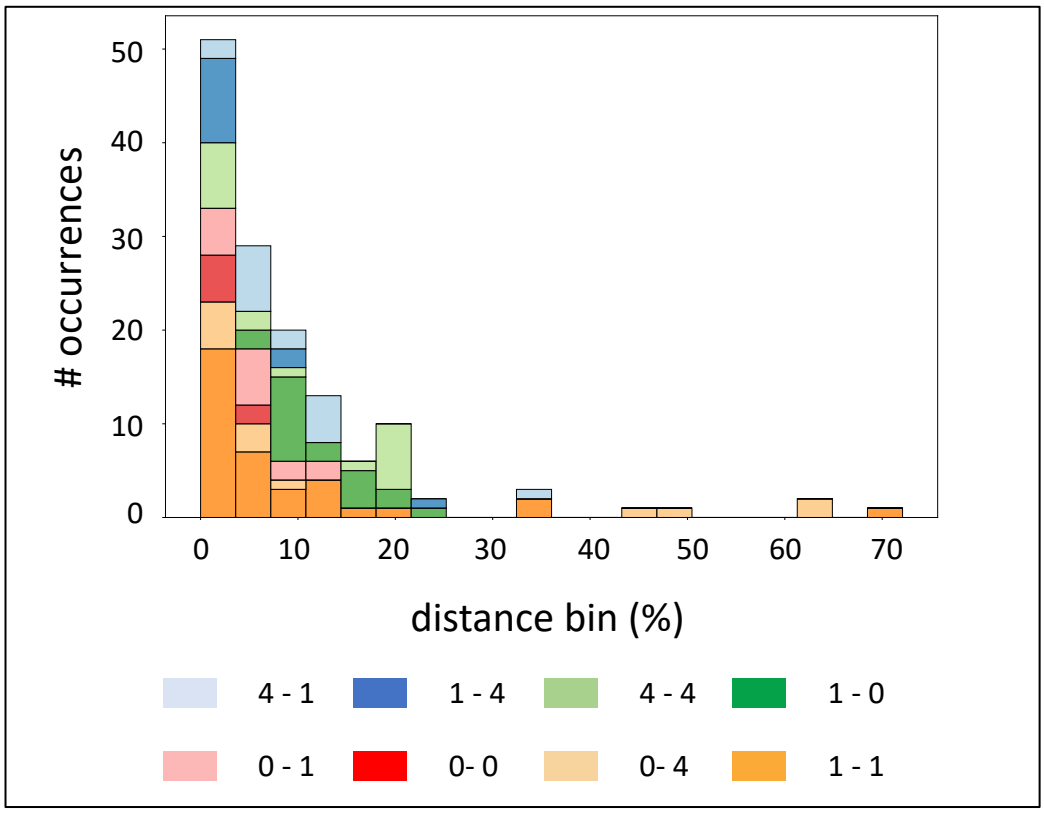


Figure 14. Histogram of distances between co-occurring CLUMPs 1, 0, 4.

This plot shows the trend of the number of co-occurrences as a function of the distance between the co-occurring CLUMPs 1, 0, 4. Since the sequences are of a different length, distances were normalized, by division of the distance itself by the length of the sequence, obtaining 10 bins of 10% each. Since, they are stacked histograms, they show, for each bin, the number of occurrences of each couple of CLUMPs at that bin. If the couple of CLUMPs is not present in a bin, there are no co-occurrences of that couple at that bin.

4.1.5. The 6 CLUMPs identified by MOnSTER characterize sub-populations of effector proteins in *M. incognita*

To finally test the discriminative power of the identified CLUMPs I quantified the number of effectors that I can identify in *M. incognita* and *M. arenaria* datasets and the number of non-effectors that contains the CLUMPs.

In *M. incognita*, the six CLUMPs identified 78.26% of the sequences in the positive dataset as effectors, however 39.19% of sequences in the negative dataset contain the selected CLUMPs. Importantly, running the same analysis but taking into account only co-occurrences, thus excluding single occurrences, reduced the number of sequences identified in the negative dataset to only 7.88%, while identifying 49.07% of true effectors. These results suggest that co-occurrences of motifs in multiple CLUMPs or within the same CLUMPs are discriminant of effector proteins in *M. incognita*.

Noteworthy, similar results were obtained for *M. arenaria* where 71.65% of the sequences of the positive datasets and 35.51% of the sequences of the negative one, contain at least one occurrence of motifs of the 6 CLUMPS. Regarding the co-occurrences, as for *M. incognita*, 40.16% of the true effectors and 6.84% of the negative sequences were retrieved.

This analysis shows that, overall, the presence of any motif from the 6 CLUMPs, and especially their co-occurrence, in protein sequences can be used to identify candidate effector proteins in nematodes.

Taking these results into account, I decided to scan the entire proteome of *M. incognita*, to see how many protein sequences contain the motifs belonging to the selected CLUMPs.

16,240 out of 43,718 (37.14%) contain at least one occurrence of any motif belonging to one of the 6 selected CLUMPs. Considering only the co-occurrences of any among the selected 6 CLUMPs, 4,591 (10.5%) proteins sequences are retained. Finally, 4,674 (10.69%) protein sequences contain any occurrence of the three best scored CLUMPs (1, 0, 4), among them only 358 (0.82%) show co-occurrences of the three.

These results suggest that it exists a hierarchy of motifs that identify different sub-populations of effectors with different characteristics.

4.2. MOnSTER successfully identifies well-known parasitism motifs to help effectors' experimental validation

After the Erasmus+ internship, the work I had completed was integrated under the supervision of Professor Bottini.

The work was completed by following three main axes, as stated before:

- A proof of concept of the developed pipeline on oomycetes.
- Testing the pipeline on other PPNs.
- Experimental validation of MiEFF72, a novel putative effector of *M. incognita*.

In the following paragraphs, I will focus on every single one of these points.

4.2.1. Proof of concept of the MOnSTER: application on Oomycetes

The first purpose of this second part of the research was to validate the reliability of MOnSTER. To do that, we applied it to a dataset of 5 oomycete species (including *Phytophtora infestans* and *Bremia lactucae*, species of great agronomical importance), to see if we would find known motifs essential for the oomycete infection.

Indeed, in the dataset, composed of 1,743 effectors and 3,009 non-effectors, we found 4 CLUMPs that recovered all the essential motifs deputed to infection from the two main classes of effectors: RxLR and the -dEER motifs from the RxLR class, and the LxLFLAK and the HVLVxxP motifs from the CRN class, as shown in **Figure 15**.

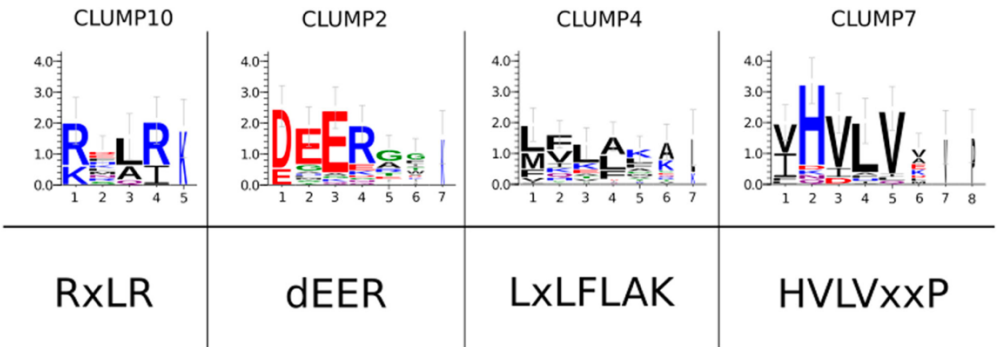


Figure 15. Motif logos of CLUMPs compared to the target.

*Top section: motif alignments within each CLUMP, generated by PROMOCA (see **Figure 16**, for more details about this method I developed); from which sequence logos were produced (using WebLogo3). In these logos, the x-axis indicates the AA position within the motif, while the y-axis represents logtransformed frequencies, expressed in bits of information.*

Bottom section: effectors-families-characteristic motifs of oomycete from the literature (McGowan & Fitzpatrick, 2017).

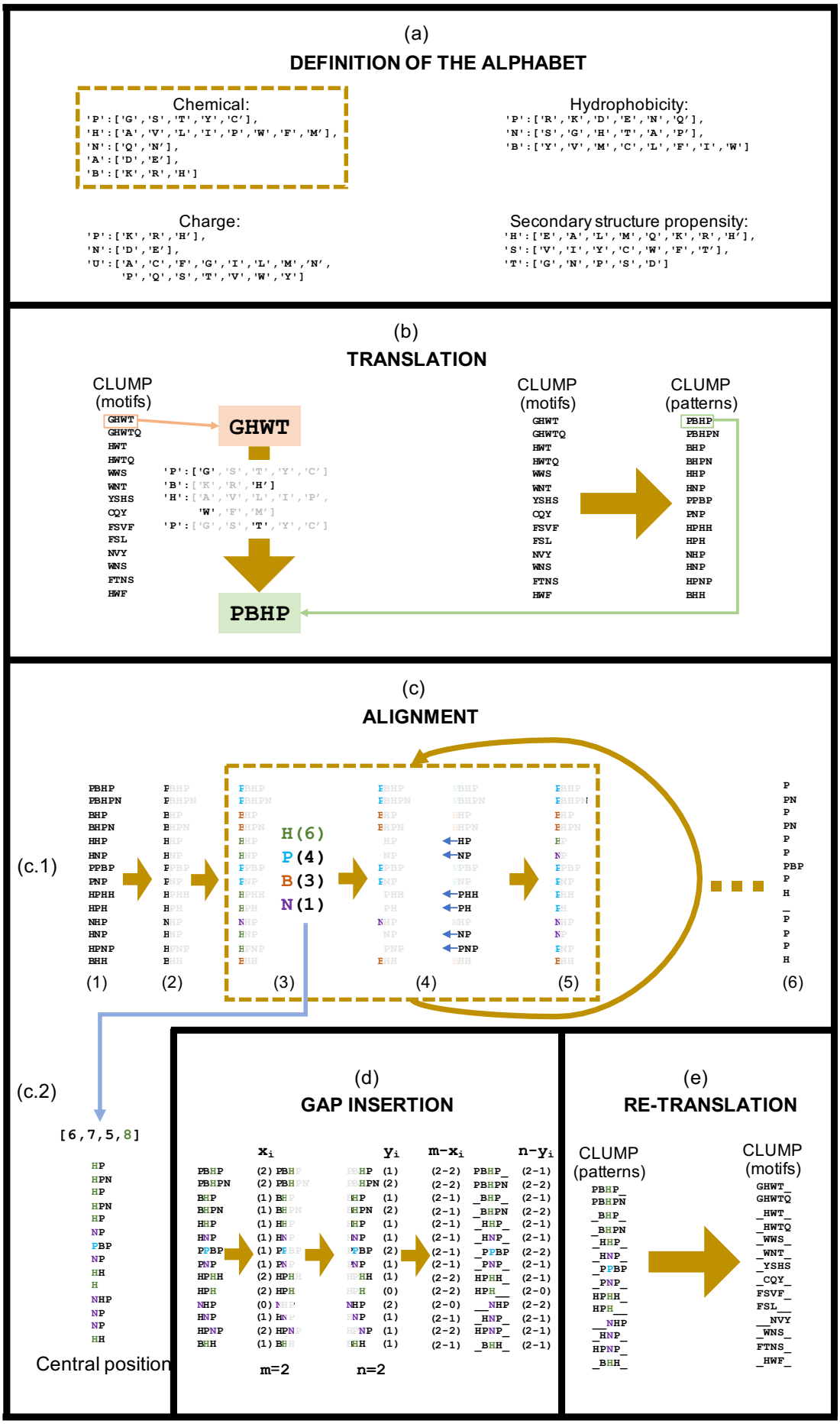

(a)
DEFINITION OF THE ALPHABET
Chemical:
'P':['G','S','T','Y','C'],
'H':['A','V','L','I','P','W','F','M'],
'N':['Q','N'],
'A':['D','E'],
'B':['K','R','H']
Hydrophobicity:
'P':['R','K','D','E','N','Q'],
'N':['S','G','H','T','A','P'],
'B':['Y','V','M','C','L','F','I','W']
Charge:
'P':['K','R','H'],
'N':['D','E'],
'U':['A','C','F','G','I','L','M','N',
'P','Q','S','T','V','W','Y']
Secondary structure propensity:
'H':['E','A','L','M','Q','K','R','H'],
'S':['V','I','Y','C','W','F','T'],
'T':['G','N','P','S','D']
(b)
TRANSLATION
CLUMP (motifs)
GHWT
GHWTQ
HWT
HWTQ
WWS
WNT
YSHS
CQY
FSVF
FSL
NVY
WNS
FTNS
HWF
GHWT
'P':['G','S','T','Y','C']
'B':['K','R','H']
'H':['A','V','L','I','P',
'W','F','M']
'P':['G','S','T','Y','C']
PBHP
CLUMP (motifs)
CLUMP (patterns)
PBHP
PBHPN
BHP
BHPN
HHP
HNP
PPBP
PNP
HPHH
HPH
NHP
HNP
HPNP
BHH
(c)
ALIGNMENT
(c.1)
H(6)
P(4)
B(3)
N(1)
(1)
(2)
(3)
(4)
(5)
(6)
(c.2)
[6,7,5,8]
Central position
(d)
GAP INSERTION
xi
yi
m-xi
n-yi
m=2
n=2
(e)
RE-TRANSLATION
CLUMP (patterns)
CLUMP (motifs)

Figure 16. PROMOCA: a tool for motifs alignment.

(a) DEFINITION OF THE ALPHABET. In this step, the user chooses the alphabet to use for the translation of the protein sequences into "patterns". In each alphabet in the figure, there is a letter followed by a ':' and a python list of letters (amino acids) after that. Such first letter is the translation of all the letters inside the python list. Meaning that if the sequence shows any of the amino acids in a python list, they will all be translated into that letter. This letter is an abbreviation of a physicochemical characteristic. All these characteristics are reported as follows. Chemical alphabet: 'P' = 'polar', 'H' = 'hydrophobic', 'N'= 'neutral', 'A' = 'acid', 'B' = 'basic'. Hydrophobicity alphabet: 'P' = 'hydrophilic', 'N' = 'neutral', 'B' = 'hydrophobic'. Charge alphabet: 'P' = 'positive', 'N' = 'negative', 'U' = 'neutral'. Secondary structure propensity alphabet: 'H' = 'helix (α)', 'S' = 'sheet (β)', 'T' = 'turn'. In this example we use the chemical alphabet.

(b) TRANSLATION. Once the alphabet is chosen (step (a)), this step translates the motifs in the CLUMP from the protein alphabet (AA) into patterns (of the given alphabet). In this example I use the chemical alphabet. In figure there is also an example of the translation of a motif 'GHWT' into the pattern 'PBHP'.

(c) ALIGNMENT. (c1) For the alignment, promoca has to: (1) take the given list of patterns, (2) extract the first position of each one of the patterns, (3) calculate how many times each letter is repeated (at that position) and find which letter is the most repeated: the result of this step is stored at each iteration of the algorithm (4) patterns that contain the most repeated letter at the first position loose letter in question, then the remaining letters of the pattern are shifted of -1 position. (5) This all results in a list of patterns, in this list: a subset of the patterns has a new first position. From here on, the process (3)-(5) is repeated (6) until promoca gets a list where there is at least 1 pattern without any more letters. (c2) When we get to (6), promoca aims to find the central position of the alignment. To do that, promoca stored in (3) the number of repetitions of the most repeated letter. In (c2) promoca calculates which is the highest value: from here it retrieves the corresponding list of patterns that have those first positions.

(d) GAP INSERTION. Gap insertion is only permitted at the extremities of the patterns; the scope of gap insertion is to get an alignment with patterns all of the same length. For each pattern promoca, for each pattern, calculates how many letters there are: before (xi) and after (yi) the central position; then it calculates the max (m and n) of these values. Promoca then subtracts m – xi and n – yi. The result of the subtraction indicates how many gaps each pattern has before and after the given central position.

(e) RE-TRANSLATION. In the end, it is necessary to re-translate such aligned patterns into aligned motifs to generate a logo.

In addition to finding all the expected motifs, when clustering the CLUMPs based on their similarities in composition, as shown in **Figure 17**. CLUMPs of the two families were clustered together, indicating similar physicochemical characteristics.

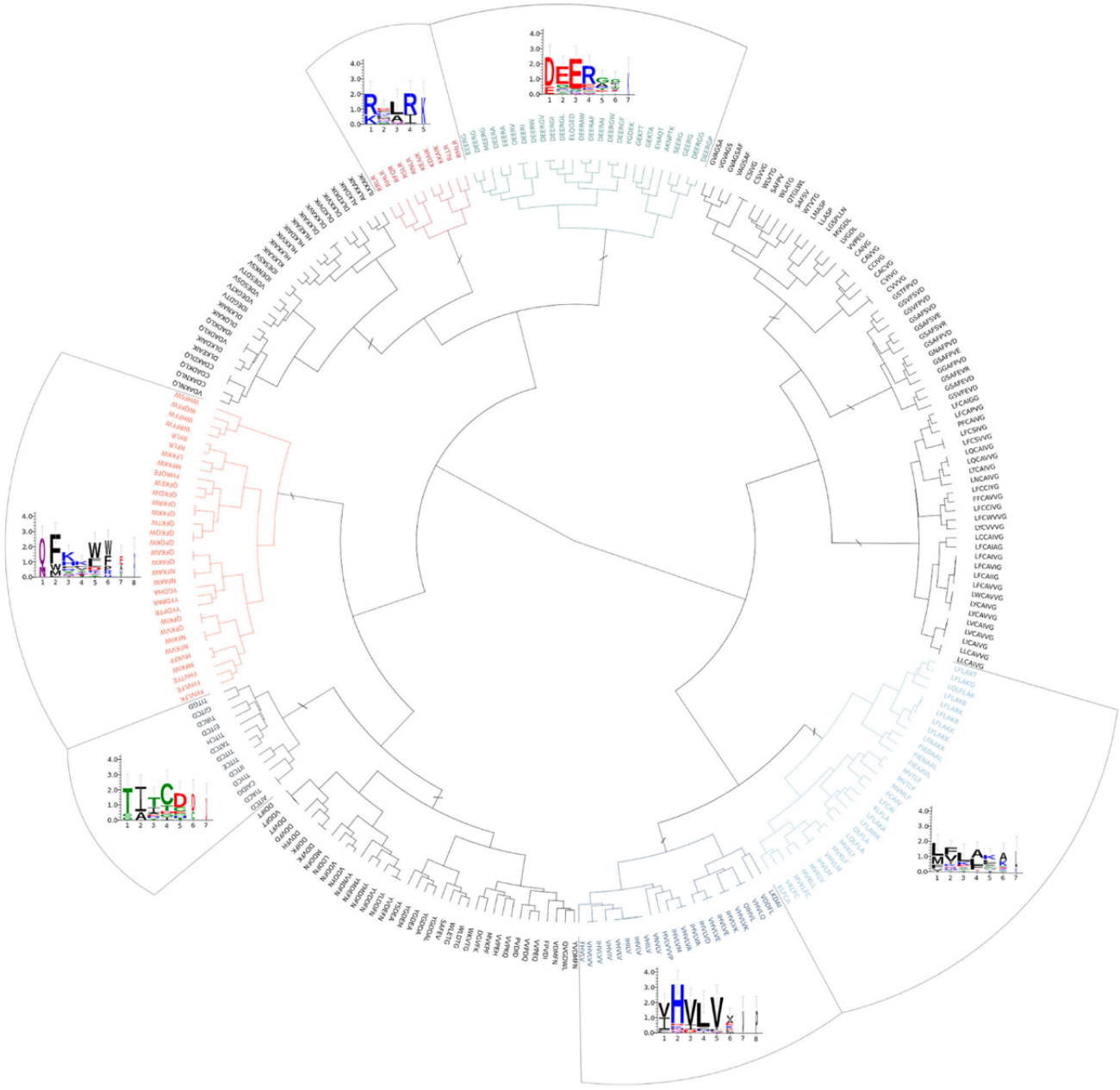


Figure 17. Dendrogram of CLUMPs in Oomycetes.

MOnSTER generated 11 CLUMPs (each one indicated with the "/" sign). Colored CLUMPs are those selected as best-scoring CLUMPs (employing the MOnSTER-score). For each best-scoring CLUMP the corresponding motif logo is displayed; to generate the alignment of the motifs in the CLUMP we used PROMOCA and WebLogo 3 (x-axis: the AA position of the motif, y-axis: the logtransformed frequency of each AA (bits of information)).

These findings suggest a robust proof of concept, proving that the MOnSTER score and the clustering of CLUMPs using features of the AA, instead of the AA themselves is effective for sequences classification.

4.2.2. Application of MOnSTER on other plant parasitic nematodes

After proving the robustness of MOnSTER in another relevant plant parasite, we went back to analyzing PPNs and extended the research to 13 PPNs, not only from *Meloidogyne*, including *Globodera*, *Heterodera*, and *Bursaphelenchus*.

In the dataset, composed of 546 well-known putative effectors and 3,849 non-effectors, among the CLUMPs we found, the 6 best scoring, showed in **Figure 18Error! Reference source not found.**, characterize approximately 60% of the known nematode effectors, while appearing in only 5% of the negative sequences. If precision alone is moderate (around 63%), we need to consider the 1:7 ratio of the positive and negative dataset. This translates to: 63% precision vs a 12.4% baseline (546 / (546 + 3849) ≈ 12.4%), which is a roughly 5x enrichment. This result, given the present biological motif discovery on relatively unbalanced data, is a very high achievement.

In addition to that, we found a consistency of the frequency in the CLUMPs across phylogenetically distant species: this means that phylogenetically close PPNs share common characteristics in their motifs to achieve similar parasitic functions.

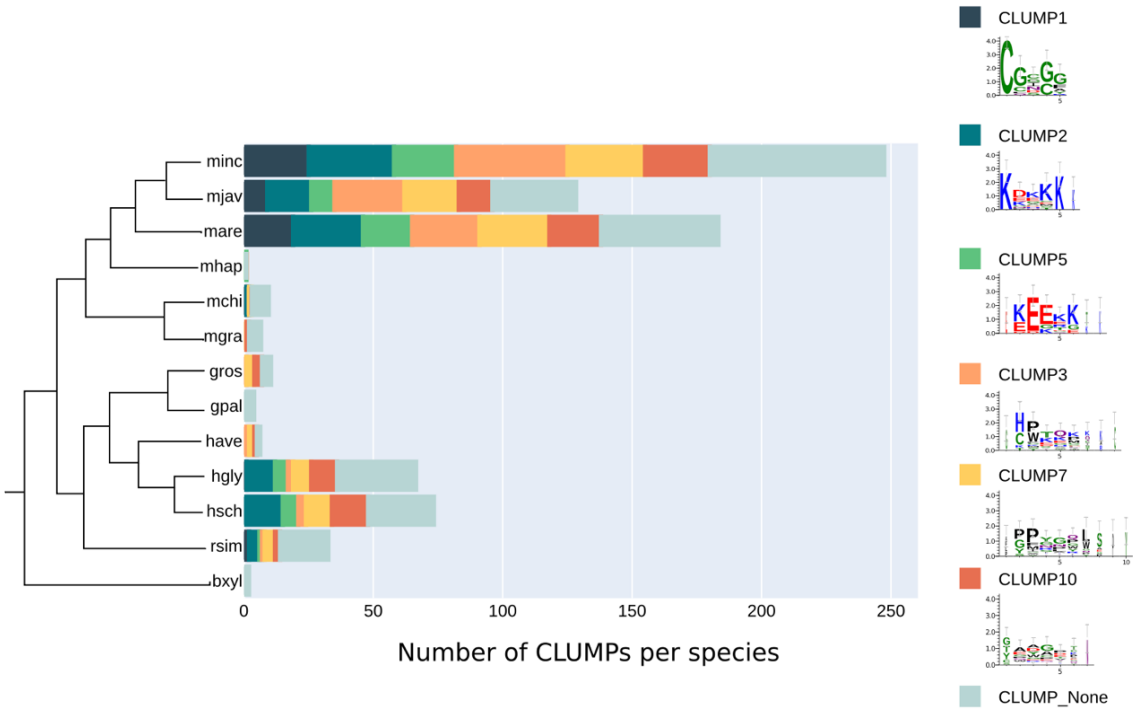


Figure 18. Cardinality of CLUMPs-motifs in each PPN species considered.

Number of motifs in each selected CLUMP, for each PPN species, based on their phylogeny (the names of the species are indicated as follows: minc: Meloidogyne incognita, mjav: Meloidogyne javanica, mare: Meloidogyne arenaria, mhap: Meloidogyne hapla, mchi: Meloidogyne chitwoodi, mgra: Meloidogyne graminicola, gros: Globodera rostochiensis,

gpal: Globodera pallida, have: Heterodera havenae, hgly: Heterodera glycines, hsch: Heterodera schachtii, rsim: Radopholus similis, bxyl: Bursaphelenchus xylophilus).

4.2.3. Positional preference of the motifs

Another finding from the post-internship research was the substantial difference in the positional preference of the motifs between the two different types of parasites (PPNs and oomycetes).

- CLUMPs in oomycetes were found significantly more in the N-terminal region, within the first 40% of the sequence.
- CLUMPs in PPNs showed a preference for the middle and C-terminal positions (around the 50% and 70% portions of the sequences).

Unlike results from other tools, that can be biased towards specific regions, these results show that MOnSTER is capable to identify motifs regardless of their position in the protein sequence.

4.2.4. Co-occurrences of CLUMPs and functional domains

CLUMPs by themselves already had a great discriminatory power between the positive and the negative dataset. On the other hand, similarly to what I found during my internship, we observed that, a more powerful tool to find CLUMPs specific to the positive dataset and practically absent in the negative dataset is the co-occurrences of the CLUMPs in the sequences.

These co-occurrences, that were found in 30% of the positive dataset, and in less than the 1% of the negative sequences, also had a link to pathogenicity, as shown in**Error! Reference source not found. Figure 19**, giving the finding a deeper biological importance. In particular:

- The co-occurrence of CLUMPs 7 and 10 was exclusively found in proteins with a glycosyl hydrolase family 5 domain.
- The co-occurrence CLUMPs 3,7 and 10 was shown in concomitance with cysteine-rich domains.

These associations suggest that: for the classes of parasitism proteins to properly work, specific associations of motifs (displayed here as CLUMPs) are necessary.

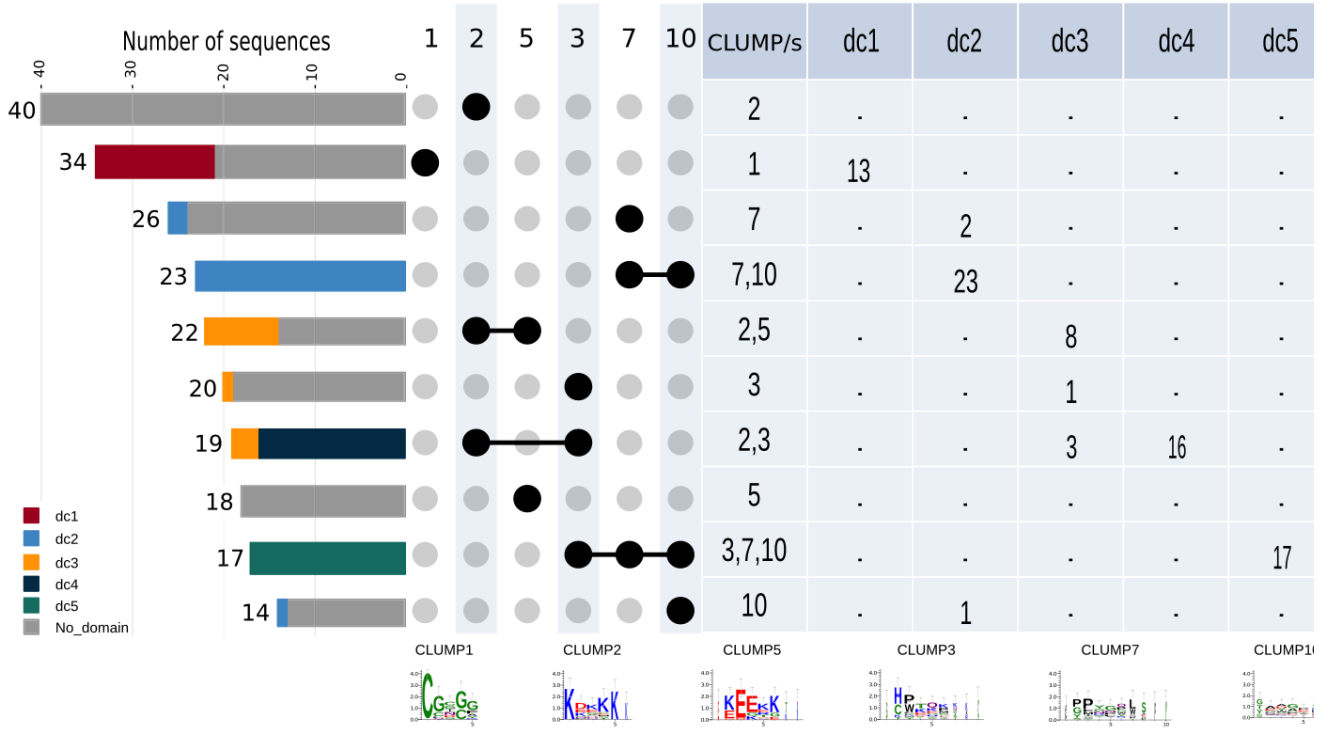


Figure 19. Candidate parasitism proteins show CLUMP(s) that are associated with pathogenicity-related protein domain(s).

The table on the right illustrates the co-occurrence of individual CLUMPs with specific domain classes (dc): dc1, pectate lyase; dc2, glycosyl hydrolase family 5; dc3, Stichodactyla toxin (ShK); dc4, 14-3-3 family; and dc5, cysteine-rich domain. The upset plot on the left displays the occurrences and co-occurrences of CLUMPs across the positive dataset, with sequences carrying a noteworthy protein domain highlighted according to the counts reported in the table.

4.2.5. Experimental validation of MiEFF72: a novel putative effector of *M. incognita*

The main objective of developing the MOnSTER pipeline was to improve the prioritization of candidate effector proteins through a computational strategy, thereby reducing the large number of putative effectors to a manageable subset suitable for experimental testing. Within this framework, an important component of the present work was the bioinformatic identification and prioritization of candidate proteins from the *M. incognita* proteome.

To evaluate the predictive power of the CLUMPs identified by MOnSTER, the *M. incognita* proteome was screened for potential effector candidates. The screening combined commonly used biological criteria for effector prediction with the information provided by the MOnSTER analysis. In particular, attention was focused on proteins that contained a signal peptide, lacked transmembrane domains, and included motifs belonging to CLUMP 5, one of the CLUMPs identified by the MOnSTER pipeline and the most abundant CLUMP in *M. incognita*.

CLUMP 5 emerged from the motif clustering and scoring procedure implemented in MOnSTER, in which motifs were grouped according to their physicochemical properties and ranked based on their discriminatory power between effectors and non-effectors. Using this computational prioritization strategy, the protein MiEFF72 (*Minc3s00056g02931*) (UniProt ID: A0A914KNF1) was identified as a strong candidate effector for further investigation, as illustrated in Figure 20.

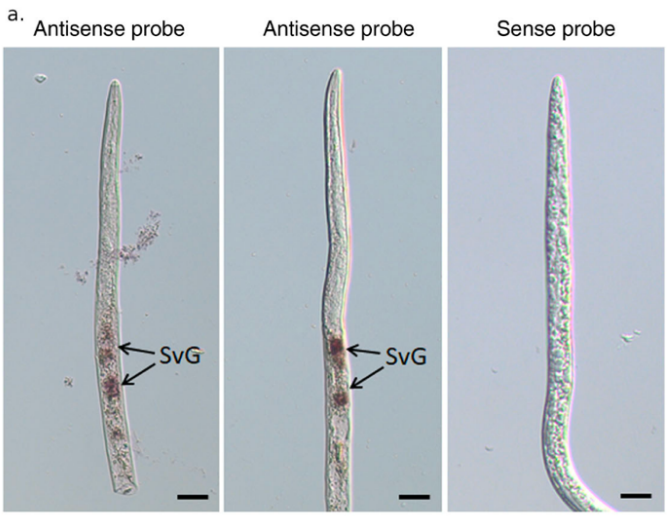
The identification of MiEFF72 represents a key outcome of the bioinformatic analysis presented in this thesis, as it translated the general output of the MOnSTER pipeline into a concrete and biologically testable target. In this sense, the computational filtering and interpretation performed here played a crucial role in narrowing down the candidate space and selecting a specific protein suitable for downstream experimental assessment.

The subsequent experimental validation was carried out by the collaborating laboratory, for this I have to mention: Dr. Michaël Quentin and Dr. Bruno Favery from INRAE (Université Côte d'Azur, CNRS, Institut Sophia Agrobiotech, Sophia-Antipolis, France); and Dr. Yongpan Chen also from INRAE in Sophia-Antipolis, currently in the Department of Plant Pathology of the China Agricultural University, Beijing, China. The validation is reported here to document the biological relevance of the computationally selected target. In particular, in situ hybridization (ISH) localized MiEFF72 transcripts in the sub-ventral gland cells of J2 pre-parasitic nematodes, which are known to be involved in effector secretion. In addition, transient expression assays in *Nicotiana benthamiana* showed that the protein localizes to the cytoplasm and cytoplasmic vesicles, a localization consistent with the expected behavior of secreted effector proteins acting within host plant cells.
In particular, *Minc3s00056g02931* is a gene encoding the protein MiEFF72, a 352 AA protein in *M. incognita*. The protein has 7 orthologues within the genus *Meloidogyne*, but shows no sequence homology with proteins from other organisms. This genus-specific distribution (which is a characteristic feature of many nematode effector proteins) and the experimental validation support its classification as a candidate effector in *M. incognita*.

Taken together, these experimental observations provide biological support for the candidate protein identified through the bioinformatic analysis, thereby reinforcing the relevance of the computational prioritization strategy. This section therefore highlights two complementary aspects: the identification of MiEFF72 as a target protein through the MOnSTER-based

screening performed in this thesis, and the experimental validation conducted by the partner laboratory, which confirms the biological plausibility of the selected candidate.

Overall, the successful validation of MiEFF72 illustrates how the MOnSTER pipeline can effectively guide the discovery of biologically relevant effector candidates, bridging large-scale computational screening with targeted experimental investigation.



c.

Sequence_identity	Known as	Evidence for secretion (ISH)	Reference	SP-noTM	Nematode specificity	CLUMP (number of motifs)	Motif/s
Minc3s00081g03887	MiTSPc	Subventral glands	Pu et al. 2022	yes	conserved	CLUMP5 (1)	KEEDK
Minc3s00519g13668	Minc3s00519g13668	Dorsal gland	Rocha et al., 2023	yes	plant-parasitic nematodes	CLUMP2 (5)	KDAK
Minc3s00147g06033	MiCE108	Subventral glands	Yu et al., 2024	yes	conserved	CLUMP3 (1), CLUMP7 (1), CLUMP10 (1)	GHWTQL, GIGHW, CGIGR
Minc3s00056g02931	Minc04572_MiEFF72	Subventral glands	This study	yes	Meloidogyne spp.	CLUMP5 (1)	GEEKK

Figure 20.* MiEFF72 *is specifically expressed in the sub-ventral glands. (a-b, bar =20 µm

a) In situ hybridization (ISH) showing MiEFF72 transcripts localized in the sub-ventral glands (SvG) of M. incognita second-stage juveniles (J2s) (two left images). A sense probe was used as a negative control (right image).

b) MiEFF72 localizes to the cytoplasm of plant cells and to cytoplasmic vesicles (red arrows and inset). The MiEFF72 coding sequence was fused to the C-terminus of GFP and transiently expressed in N. benthamiana leaves via agroinfiltration. Water-infiltrated leaves served as a negative control. Representative confocal GFP fluorescence images and merged differential interference contrast/fluorescence overlays are shown.

c) Characteristics of recently published, experimentally confirmed M. incognita effectors that yielded no significant BLAST hit or orthogroup match within the positive dataset. Columns indicate: sequence ID, alternative identifiers, secretion site based on ISH, original publication, presence of a signal peptide and absence of a transmembrane domain (SP no TM), nematode specificity, presence of CLUMPs (number of motifs per CLUMP indicated as "xn", where n is the number of repeats), and motif sequence (if multiple motifs from the same CLUMP are present in the sequence, they are indicated in "xn" format; if motifs belong to distinct CLUMPs, they are listed in the order corresponding to the preceding column).

5. Chapter 5: Discussion

The work aimed to identify, characterize and cluster specific sequence motifs, based on physicochemical properties, to discriminate effector proteins in plant parasites. Effector proteins are of particular agronomical interest because they are the main drivers of infection in phytoparasites.

The parasites I studied belong to PPNs of different genera, with a focus on *M. incognita*, and oomycetes as a proof of concept of MOnSTER.

MOnSTER was applied to four curated datasets, all the four of them split into two categories: a positive dataset, containing putative effector proteins, and a negative one with known non-effector proteins. The objective of MOnSTER was and is, to today, to decrease the number of putative effector proteins to validate in the laboratory, since it is a very expensive and time-consuming process.

The first and the second datasets were generated from two closely related species *M. incognita* and *M.arenaria*, respectively, and used during the Erasmus+ internship.

I used the first one to set up the pipeline and identify the motifs, and the second one to explore the discriminant power of such motifs.

The third and the fourth were much bigger in size and included, respectively: 5 species of oomycetes of agronomical interest, and 13 PPNs. The third dataset was used as a proof of concept to test the robustness of MOnSTER, and the fourth to extend the promising results I got on *M. incognita* and *M. arenaria*, since PPNs were of particular interest in the host laboratory.

For what concerns the application of MOnSTER to *M. incognita*, my first finding was that effector proteins display a particular composition in AAs compared to non-effectors. In particular, effector proteins showed a positive enrichment of AAs with a prevalence of turn structure and tiny AAs, with an underrepresentation of alpha-helix and beta-sheet propensities. Furthermore, effectors showed a high preference for hydrophilic AAs and lower molecular extinction coefficient (which is an estimation of how strongly the protein absorbs light at a particular wavelength). These compositional characteristics suggest that effector proteins may possess structural features that facilitate their secretion and interaction with host targets, which is consistent with the biological role of effectors as molecular tools used by parasites to manipulate host cellular processes.

MOnSTER identified 198 non-identical motifs that were clustered in 10 CLUMPs based on 16 physicochemical features. Then, by implementing the MOnSTER score, I identified 6 CLUMPs (1, 0, 4, 5, 7 and 8) with high discriminative power between effectors and non-effectors. Of these, CLUMPs 1, 0 and 4: were the top three, were composed of polar AAs and tended to occupy central positions within the protein sequences.

A comparison can be made with previous motif discovery approaches such as MERCI (Vens et al., 2011), which identified four motifs characteristic of effector proteins in *M. incognita* (LLIIS, EGAG, ASKY and AEGD) using a discriminative approach based on positive and negative datasets. However, when applied to the entire proteome, such motif-based approaches often generate a very large number of candidate proteins, making experimental validation difficult. In contrast, the strategy implemented in MOnSTER groups motifs into CLUMPs based on shared physicochemical properties and prioritizes combinations of motifs rather than isolated motifs. This approach substantially refines the candidate space and allows the identification of more biologically meaningful motif architectures associated with effector proteins.

In addition to the exploring the CLUMPs, their compositions and their discriminant power by themselves, results highlighted the importance of co-occurrences in discriminating between the two datasets. In fact, while the abovementioned 6 CLUMPs initially identified 78.26% of known effectors, they also appeared in 39.19% of the negative dataset.

However, when discrimination was performed by restricting the criteria to co-occurrences of CLUMPs, the false positive rate in the negative dataset dropped almost by six times, to 7.88%, while still capturing almost half (49.07%) of true effectors. MOnSTER showed robustness when applied to *M. arenaria*, where co-occurrences identified 40.16% of true effectors with only a 6.84% false positive rate.

The second part of the present work contributed to the abovementioned publication and aimed to expand the scope to ensure MOnSTER reliability across different taxa of plant parasites.

In particular, the proof of concept, on oomycetes, was highly successful, as MOnSTER recovered well-known essential motifs: RxLR, -dEER, LxLFLAK, and HVLVxxP, belonging to the two main families of effectors. This confirmed that clustering motifs by physicochemical properties, rather than “letters” sequence alone is a good strategy for effector protein sequences discrimination.

Afterwards, we extended the analysis to 13 PPN species, where the 6 best-scoring CLUMPs characterized approximately 60% of known effectors, while appearing in only 5% of negative

sequences. This result is particularly important when considering the 1:7 ratio of the positive and the negative dataset. In order words, this represents a roughly 5x enrichment over the baseline precision (63% precision vs. a 12.4% baseline). This observation suggests that the motif patterns captured by MOnSTER represent conserved molecular features associated with parasitic functions across phylogenetically related nematodes.

Another observation we made is a distinct positional preference between motifs in the two different parasites groups (oomycetes and PPNs): oomycete motifs were concentrated in the N-terminal region (first 40% of the sequence), whereas PPN motifs were mostly found in the middle and C-terminal regions (50% to 70%). Biologically, these differences may reflect distinct mechanisms of host interaction between the two parasite groups, suggesting that motif position within effector proteins may be linked to different secretion or interaction strategies.

Biologically speaking, perhaps most importantly, we observed a link between CLUMPs co-occurrences and functional domains in the protein sequences. In particular, the co-occurrence of CLUMPs 7 and 10 was exclusively found in proteins containing a glycosyl hydrolase family 5 domain, while the combination of CLUMPs 3, 7, and 10 was associated with cysteine-rich domains. This finding suggests that specific motif associations could be necessary for the structural or biochemical functionality of various parasitism protein classes. Further investigation of these associations, for example through structural modelling of effector proteins containing these motif combinations, could help to clarify the functional role of such motif architectures.

Finally, the ultimate objective of the MOnSTER pipeline is to identify candidate effector proteins with a higher probability of experimental validation. In this context, the identification of MiEFF72 represents an important outcome of the present work.

MiEFF72 (*Minc3s00056g02931*) was selected through the computational screening of the *M. incognita* proteome by combining classical criteria used in effector prediction (presence of a signal peptide and absence of transmembrane domains) with the occurrence of motifs belonging to CLUMP 5. Importantly, the identification of MiEFF72 illustrates the role of MOnSTER as a prioritization tool capable of reducing the large candidate space generated by conventional secretion-based filters.

The subsequent experimental validation was performed by the collaborating laboratory and is reported here to support the biological relevance of the computationally selected candidate.

Indeed, the expression of MiEFF72 was localized in the sub-ventral glands of J2 pre-parasitic nematodes, which are known to produce effector proteins, and the protein was observed in the cytoplasm and cytoplasmic vesicles of plant cells, confirming its behavior as a secreted protein involved in host interaction.

Therefore, MiEFF72 can be considered a proof-of-principle example demonstrating how motif-based computational prioritization can lead to the identification of biologically relevant effector candidates. In this sense, the main contribution of the present work lies not only in the identification of motif clusters but also in the development of a strategy capable of guiding experimental target discovery from large proteome-scale datasets.

6. Chapter 6: Conclusions

The aim of this work was to identify sequence features that characterize effector proteins in plant parasitic nematodes, using *M. incognita* as a model species. To address this objective, I developed the bioinformatic pipeline MOnSTER, designed to identify and cluster sequence motifs based on their physicochemical properties and their ability to discriminate between effector and non-effector proteins.

Using curated datasets of *M. incognita*, MOnSTER identified 198 motifs that were grouped into 10 CLUMPs, of which six showed the highest discriminative power for effector proteins. The results indicate that effector proteins are not defined by single conserved motifs but rather by combinations of motifs sharing common physicochemical properties.

A key finding of this work is that the co-occurrence of motifs across different CLUMPs provides a stronger signal for effector discrimination than the presence of individual motifs alone. This observation suggests that effector proteins may rely on specific motif architectures, reflecting structural or functional constraints related to their interaction with host targets.

The robustness of the MOnSTER pipeline was further supported by its successful application to other plant parasites. In oomycetes, the method recovered well-known motifs associated with parasitism, while in a broader dataset of plant parasitic nematodes it retained significant discriminative power across phylogenetically distant species. These results indicate that motif clustering based on physicochemical properties represents a promising strategy for identifying functional signatures of effector proteins across taxa.

Applying this approach to the *M. incognita* proteome revealed thousands of proteins containing motifs or motif combinations characteristic of the identified CLUMPs, highlighting the presence of multiple potential effector-like proteins within the proteome. Rather than defining individual effectors, these results suggest the existence of several sub-populations of proteins sharing effector-like sequence architectures.

Importantly, the computational prioritization strategy enabled the identification of MiEFF72 (*Minc3s00056g02931*) as a novel candidate effector of *M. incognita*. Its subsequent experimental validation by the collaborating laboratory, showing expression in the sub-ventral glands of the nematode and localization in plant cell cytoplasm and cytoplasmic vesicles, supports the biological relevance of the computational predictions.

Overall, this work demonstrates that motif clustering approaches such as MOnSTER can contribute to bridging large-scale computational screening and experimental validation. By enabling the prioritization of candidate effector proteins from complex proteomes, this strategy

provides a useful framework for exploring the diversity and functional organization of parasitism-related proteins in plant pathogens.

References

Abad, P., Gouzy, J., Aury, J.-M., Castagnone-Sereno, P., Danchin, E. G. J., Deleury, E., Perfus-Barbeoch, L., Anthouard, V., Artiguenave, F., Blok, V. C., Caillaud, M.-C., Coutinho, P. M., Dasilva, C., De Luca, F., Deau, F., Esquibet, M., Flutre, T., Goldstone, J. V., Hamamouch, N., … Wincker, P. (2008). Genome sequence of the metazoan plant-parasitic nematode Meloidogyne incognita. *Nature Biotechnology*, *26*(8), 909–915. https://doi.org/10.1038/nbt.1482

Anastasiadou, P., Ntalli, N., Kyriakopoulou, K., & Kasiotis, K. M. (2025). Nematicidal Extracts of Chinaberry, Parsley and Rocket Are Safe to Eisenia fetida, Enchytraeus albidus, Daphnia magna and Danio rerio. *Agriculture*, *15*(4), 436. https://doi.org/10.3390/agriculture15040436

Asgari, E., McHardy, A. C., & Mofrad, M. R. K. (2019). Probabilistic variable-length segmentation of protein sequences for discriminative motif discovery (DiMotif) and sequence embedding (ProtVecX). *Scientific Reports*, *9*(1), 3577. https://doi.org/10.1038/s41598-019-38746-w

Bailey, T. L. (2021). STREME: accurate and versatile sequence motif discovery. *Bioinformatics*, *37*(18), 2834–2840. https://doi.org/10.1093/bioinformatics/btab203

Bailey, T. L., Johnson, J., Grant, C. E., & Noble, W. S. (2015). The MEME Suite. *Nucleic Acids Research*, *43*(W1), W39–W49. https://doi.org/10.1093/nar/gkv416

Bellafiore, S., & Briggs, S. P. (2010). Nematode effectors and plant responses to infection. *Current Opinion in Plant Biology*, *13*(4), 442–448. https://doi.org/10.1016/j.pbi.2010.05.006

Blanc-Mathieu, R., Perfus-Barbeoch, L., Aury, J.-M., Da Rocha, M., Gouzy, J., Sallet, E., Martin-Jimenez, C., Bailly-Bechet, M., Castagnone-Sereno, P., Flot, J.-F., Kozlowski, D. K., Cazareth, J., Couloux, A., Da Silva, C., Guy, J., Kim-Jo, Y.-J., Rancurel, C.,

Schiex, T., Abad, P., … Danchin, E. G. J. (2017). Hybridization and polyploidy enable genomic plasticity without sex in the most devastating plant-parasitic nematodes. *PLOS Genetics*, *13*(6), e1006777. https://doi.org/10.1371/journal.pgen.1006777

Calia, G., Porracciolo, P., Chen, Y., Kozlowski, D., Schuler, H., Cestaro, A., Quentin, M., Favery, B., Danchin, E. G. J., & Bottini, S. (2024). Identification and characterization of specific motifs in effector proteins of plant parasites using MOnSTER. *Communications Biology*, *7*(1), 850. https://doi.org/10.1038/s42003-024-06515-9

Chepsergon, J., Nxumalo, C. I., Salasini, B. S. C., Kanzi, A. M., & Moleleki, L. N. (2022). Short Linear Motifs (SLiMs) in "Core" RxLR Effectors of *Phytophthora parasitica* var. *nicotianae*: A Case of PpRxLR1 Effector. *Microbiology Spectrum*, *10*(2), e01774-21. https://doi.org/10.1128/spectrum.01774-21

Davey, N. E., Cyert, M. S., & Moses, A. M. (2015). Short linear motifs – ex nihilo evolution of protein regulation. *Cell Communication and Signaling*, *13*(1), 43. https://doi.org/10.1186/s12964-015-0120-z

Davey, N. E., Van Roey, K., Weatheritt, R. J., Toedt, G., Uyar, B., Altenberg, B., Budd, A., Diella, F., Dinkel, H., & Gibson, T. J. (2012). Attributes of short linear motifs. *Mol. BioSyst.*, *8*(1), 268–281. https://doi.org/10.1039/C1MB05231D

Davies, D. L., & Bouldin, D. W. (1979). A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-1*(2), 224–227. https://doi.org/10.1109/TPAMI.1979.4766909

Https://ephytia.inra.fr/en/C/20910/Potato-Meloidogyne-spp-Root-knot-nematodes. (n.d.).

Huggins, P., Zhong, S., Shiff, I., Beckerman, R., Laptenko, O., Prives, C., Schulz, M. H., Simon, I., & Bar-Joseph, Z. (2011). DECOD: Fast and accurate discriminative DNA motif finding. *Bioinformatics*, *27*(17), 2361–2367. https://doi.org/10.1093/bioinformatics/btr412

Hussain, M. A., Mukhtar, T., & Kayani, M. Z. (n.d.). ASSESSMENT OF THE DAMAGE CAUSED BY MELOIDOGYNE INCOGNITA ON OKRA (ABELMOSCHUS ESCULENTUS). *J. Anim. Plant Sci.*, 6.

Jiang, X., Xiang, M., & Liu, X. (2017). Nematode-Trapping Fungi. *Microbiology Spectrum*, *5*(1), 5.1.10. https://doi.org/10.1128/microbiolspec.FUNK-0022-2016

Jones, J. T., Haegeman, A., Danchin, E. G. J., Gaur, H. S., Helder, J., Jones, M. G. K., Kikuchi, T., Manzanilla-López, R., Palomares-Rius, J. E., Wesemael, W. M. L., & Perry, R. N. (2013). Top 10 plant-parasitic nematodes in molecular plant pathology: Top 10 plant-parasitic nematodes. *Molecular Plant Pathology*, *14*(9), 946–961. https://doi.org/10.1111/mpp.12057

Kantor, C., Eisenback, J. D., & Kantor, M. (2024). Biosecurity risks to human food supply associated with plant-parasitic nematodes. *Frontiers in Plant Science*, *15*, 1404335. https://doi.org/10.3389/fpls.2024.1404335

Kyte, J., & Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, *157*(1), 105–132. https://doi.org/10.1016/0022-2836(82)90515-0

McGowan, J., & Fitzpatrick, D. A. (2017). Genomic, Network, and Phylogenetic Analysis of the Oomycete Effector Arsenal. *mSphere*, *2*(6), e00408-17. https://doi.org/10.1128/mSphere.00408-17

Mejias, J., Truong, N. M., Abad, P., Favery, B., & Quentin, M. (2019). Plant Proteins and Processes Targeted by Parasitic Nematode Effectors. *Frontiers in Plant Science*, *10*, 970. https://doi.org/10.3389/fpls.2019.00970

Mota, A. P. Z., Koutsovoulos, G. D., Perfus-Barbeoch, L., Despot-Slade, E., Labadie, K., Aury, J.-M., Robbe-Sermesant, K., Bailly-Bechet, M., Belser, C., Péré, A., Rancurel, C., Kozlowski, D. K., Hassanaly-Goulamhoussen, R., Da Rocha, M., Noel, B., Meštrović, N., Wincker, P., & Danchin, E. G. J. (2024). Unzipped genome assemblies of polyploid

root-knot nematodes reveal unusual and clade-specific telomeric repeats. *Nature Communications*, *15*(1), 773. https://doi.org/10.1038/s41467-024-44914-y

Nothman, J. (n.d.). *Upsetplot Documentation*. 79.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., & Cournapeau, D. (n.d.). Scikit-learn: Machine Learning in Python. *MACHINE LEARNING IN PYTHON*, 6.

Phan, N. T., Orjuela, J., Danchin, E. G. J., Klopp, C., Perfus-Barbeoch, L., Kozlowski, D. K., Koutsovoulos, G. D., Lopez-Roques, C., Bouchez, O., Zahm, M., Besnard, G., & Bellafiore, S. (2020). Genome structure and content of the rice root-knot nematode (*Meloidogyne graminicola*). *Ecology and Evolution*, *10*(20), 11006–11021. https://doi.org/10.1002/ece3.6680

Roberson, E. D. O. (2018). Motif scraper: A cross-platform, open-source tool for identifying degenerate nucleotide motif matches in FASTA files. *Bioinformatics*, *34*(22), 3926–3928. https://doi.org/10.1093/bioinformatics/bty437

Rutter, W. B., Franco, J., & Gleason, C. (2022). Rooting Out the Mechanisms of Root-Knot Nematode–Plant Interactions. *Annual Review of Phytopathology*, *60*(1), 43–76. https://doi.org/10.1146/annurev-phyto-021621-120943

School of Agriculture Science and Biotechnology, Faculty of Bioresources and Food Industry, University of Sultan Zainal Abidin, Besut Campus, 22000 Besut, Terengganu, Malaysia, Ralmi, N. H. A. A., Khandaker, M. M., School of Agriculture Science and Biotechnology, Faculty of Bioresources and Food Industry, University of Sultan Zainal Abidin, Besut Campus, 22000 Besut, Terengganu, Malaysia, Mat, N., & School of Agriculture Science and Biotechnology, Faculty of Bioresources and Food Industry, University of Sultan Zainal Abidin, Besut Campus, 22000 Besut, Terengganu, Malaysia. (2016). Occurrence and control of root knot nematode in crops: A review.

Australian Journal of Crop Science, *10*(12), 1649–1654. https://doi.org/10.21475/ajcs.2016.10.12.p7444

Shi, Q., Mao, Z., Zhang, X., Zhang, X., Wang, Y., Ling, J., Lin, R., Li, D., Kang, X., Sun, W., & Xie, B. (2018). A Meloidogyne incognita effector MiISE5 suppresses programmed cell death to promote parasitism in host plant. *Scientific Reports*, *8*(1), 7256. https://doi.org/10.1038/s41598-018-24999-4

Subedi, S., Thapa, B., & Shrestha, J. (2020). Root-knot nematode (Meloidogyne incognita) and its management: A review. *Journal of Agriculture and Natural Resources*, *3*(2), 21–31. https://doi.org/10.3126/janr.v3i2.32298

Tiwari, S. (2025). Impact of nematicides on plant-parasitic nematodes: Challenges and environmental safety. *Tunisian Journal of Plant Protection*, *19*(2). https://doi.org/10.4314/tjpp.v19i2.4

Vens, C., Rosso, M.-N., & Danchin, E. G. J. (2011). Identifying discriminative classification-based motifs in biological sequences. *Bioinformatics*, *27*(9), 1231–1238. https://doi.org/10.1093/bioinformatics/btr110

Vieira, P., & Gleason, C. (2019). Plant-parasitic nematode effectors—Insights into their diversity and new tools for their identification. *Biotic Interactions*, *50*, 37–43. https://doi.org/10.1016/j.pbi.2019.02.007