Identification and characterization of specific motifs in effector proteins of plant parasites using MOnSTER.

Giulia Calia[1,2 ¶], Paola Porracciolo[3,4¶], Djampa Kozlowski[3,4], Hannes Schuler[1,5], Alessandro Cestaro[2,6], Etienne G.J. Danchin[4&] and Silvia Bottini[3&*]

[1] Free University of Bolzano, Faculty of Agricultural Environmental and Food Science, Bolzano, Italy

[2] Fondazione Edmund Mach, Research and Innovation Centre, San Michele all'Adige, Italy

[3] Université Côte d'Azur, Center of Modeling, Simulation and Interactions, Nice, France

4 INRAE, Université Côte d'Azur, CNRS, Institut Sophia Agrobiotech, Sophia-Antipolis, France

[5] Free University of Bolzano, Competence Centre for Plant Health, Bolzano, Italy

[6] Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies (IBIOM), National Research Council (CNR), Bari, Italy

[¶]contributed as co-first author

#contributed as co-last author

[*]Corresponding Author: silvia.bottini@univ-cotedazur.fr

Short title: MOnSTER allows identification of motifs in effectors of plant parasites

To whom correspondence should be addressed:

Silvia Bottini, PhD
Hopital l'Archet 3
151 route de Saint Antoine de Ginestiere,
06200, Nice, France
+33(6)30566999
silvia.bottini@univ-cotedazur.fr

Abstract

Plant pathogens cause billions of dollars of crop loss every year and are a major threat to global food security. Identifying and characterizing pathogens effectors is crucial towards their improved control. Because of their poor sequence conservation, effector identification in protein sequences predicted from genomes is challenging and current methods generate too many candidates without indication for prioritizing further experimental studies. In most phyla, effectors contain specific sequence motifs which influence their localization and targets in the plant. Although bacterial, fungal and oomycetes effectors have been studied extensively and conserved characteristic motifs have been identified, research on plant-parasitic nematode effectors (PPN) identified some enriched degenerate motifs in only one species so far. The different lifestyles of PPNs might reflect effectors with different functions according to the nematode's specific needs, thus presenting a high variety of characteristic motifs.

To circumvent these limitations, we have developed MOnSTER a novel tool that identifies clusters of motifs of protein sequences (CLUMP) and associates a score to each CLUMP. This score encompasses the physicochemical properties of AAs and the motif occurrences. We built up our method to identify discriminant CLUMPs in effector proteins of plant-pathogenic oomycetes. We showed the reliability of MOnSTER by identifying five CLUMPs that correspond to the known motifs: RxLR, -dEER and LxLFLAK-HVLVxxP. Consequently, we applied MOnSTER on PPN effector proteins and identified peculiar motifs in their sequences. We identified six CLUMPs in about 60% of the known nematode effectors. Furthermore, we found that specific co-occurrences of at least two CLUMPs are present in PPN effector sequences bearing protein domains important for invasion and pathogenicity.

The potentiality of this tool goes behind the effector proteins and can be used to easily cluster motifs and calculate the CLUMPs score on any set of protein sequences.

Keywords

Motif clustering – motif scoring – effectors – plant-parasite interaction – *oomycetes - nematodes*

Authors summary

Population growth, environmental degradation and climate change are already bringing harm to human communities and the natural world that needs to be addressed rapidly. Ensuring food security for a population that will exceed 9 billion people by 2050 while preserving the environment and biodiversity is a major challenge. Agricultural pathogens, to cause the infection, secrete effector proteins that promote colonization of the host plant. Identifying and characterizing pathogens' effectors is crucial towards understanding how they manipulate the plant and better combat them. Because of their poor sequence conservation, effector identification in protein sequences predicted from genomes is challenging and current methods generate too many candidates without indication for prioritizing further experimental studies. To address these challenges, we have developed a novel tool called MOnSTER, that identifies and score clusters of motifs of protein sequences (CLUMPs). MOnSTER is an easy to use tool that can be included in any pipeline needing motif calling and will be of great use to accelerate both computational and experimental studies relating to protein motif discovery. Altogether our findings provide improvements in the understanding of the mechanisms set up by the pathogens to infect the plant and can elucidate important signatures to block the development of plant-pathogen interactions and allow to engineer of durable disease resistance.

Introduction

Plant pathogens are a major threat to global food security. To cause the infection, pathogenic organisms secrete effector proteins that promote colonization of the host plant by overcoming the physical barriers of plant cell-walls, suppressing or evading immune perception, and deriving nutrients from host tissues [1]. Therefore, identifying and characterizing pathogens effectors is crucial towards understanding how they manipulate the plant and better combat them. Effector proteins are often specific to pathogens and essential for causing plant pathology, constituting targets of choice for the development of cleaner and more specific control methods [2]–[4]. Because of their poor sequence conservation, effector identification among the set of predicted proteins from the genome (proteome) is challenging and current methods generate too many candidates without further indication for prioritizing experimental studies. Classically, effector proteins are indirectly identified among the predicted secretome based on the presence of a signal peptide for secretion and a lack of transmembrane region [5], [6] . However, these criteria alone suffer from two main limitations. On one side, the secretome comprises many proteins that are not effectors, on the other side some known effectors do not possess signal peptides for secretion. In most phyla, effectors contain specific sequence motifs which target host proteins with distinct roles in the infection process and control virulence [7]. The best-studied example is effectors secreted via the type III secretion system (T3SS) class of Gram-negative bacterial pathogens which are characterized by a specific motif/domain conferring a repertoire of molecular determinants with important roles during infection [8], [9]. However, these features are not conserved in other bacteria. Indeed, gram-positive pathogens and certain phloem- and xylem-colonizers, such as *Candidatus liberibacter* and *Xylella spp*., do not encode the T3SS. In these bacteria, effector delivery is dependent on the presence of the N-terminal signal peptide, which is required for protein secretion [10]. In fungi, two motifs have been identified in effectors, namely the cysteine-rich motifs and the MAX motif [11]. Another well-characterized example are the effectors of the oomycetes pathogens. Oomycetes are eukaryotic filamentous and heterotrophic microorganisms among which, more than 60% of them parasitize plants [12]. Well-known plant pathogens in oomycetes include late blight of potato, sudden oak death, root rot agents (*Phytophthora* species), and downy mildew *Peronospora* and *Bremia* species [13], [14]. These pathogens code for two notable classes of effector proteins RxLR and Crinkler (CRN), that can be predicted by the occurrence of the related motifs, RxLR, -dEER and LxLFLAK-HVLVxxP in the N-terminal region downstream the signal peptide [15]–[17].

Although bacterial, fungal and oomycetes effectors have been studied extensively and characteristics motifs have been identified [18], [19], research on Plant-Parasitic Nematode effectors (PPN) did not identify any consensus motif, conserved across multiple species. The most economically important PPNs are the sedentary Root-Knot Nematodes (RKNs) and cyst nematodes [20]. These sedentary parasites induce the formation of a feeding structure that serves as a constant food source for the nematode. Other PPN are migratory and a whole spectrum of variations exists between endo and ecto parasites, with semi-endoparasites an intermediate between the two extremes [21]. The different lifestyles of PPNs are expected to be reflected in their secretions, which presumably contain effectors with different functions according to the nematode's specific needs, thus presenting a high variety of characteristic motifs complicating their identification.

A first step toward the identification of motifs characteristics of RKN effectors was performed by Vens et al. [22]. The authors developed a bioinformatic tool, called MERCI, to identify motifs with high occurrences in a positive dataset (known effector sequences) and absent in the negative one (non-effector sequences). MERCI uses a graph-based approach incorporating physicochemical features of the amino acids composing protein sequences. By analyzing the known effector sequences of the RKN species *Meloidogyne incognita*, one of the most important known crop pathogens among all [23], they identified 4 motifs. However, at the time of their publication, very few genomes for RKN species were available, and the study was therefore conducted on one single RKN species. Furthermore, the genome used at that time was later shown to be partially incomplete [24]. These limitations prevent the generalization of the previous findings. Therefore, there is an urgent need for a novel study of the properties of PPN effector sequences and motif research.

By taking advantages of the multitude of proteomes available nowadays for several PPN, we developed a comprehensive motif mining analysis to identify characteristic motifs of effector sequences of these species. Sequence motifs are usually of constant short size and are often repeated and conserved. Typically, motifs conform to a particular sequence pattern, where certain positions can be constrained to a specific amino acid, whereas others are not [25]. This confers a high degeneration of the motifs yielding to a huge list of non-redundant motif sequences and consequently some motifs that are not characteristics of effector sequences only [26]. Furthermore, different amino acids (AAs) can have similar physicochemical properties, thus different motif sequences can share similar properties. However, most available motif discovery tools do not take these properties into consideration. To circumvent these limitations, we have developed MOnSTER a novel tool that identifies clusters of motifs of protein sequences (CLUMP) and associates a score to each CLUMP. This score encompasses the physicochemical properties of AAs and the motif occurrences.

We built-up our method to identify discriminant CLUMPs in 1743 effector proteins of plant-pathogenic oomycetes. We showed the reliability of MOnSTER by identifying 5 CLUMPs that correspond to the known motifs: RxLR, -dEER and LxLFLAK-HVLVxxP. After this proof of concept, we applied MOnSTER on PPN effector proteins and identified peculiar motifs in their sequences at an unprecedented level. We selected a set of 4395 protein sequences from 13 PPN species belonging to the genera *Meloidogyne, Globodera, Heterodera, Radopholus and Bursaphelenchus*. We identified 6 CLUMPs present in 60% of the known effectors (positive dataset). Of note these CLUMPs were found in only 5% of the sequences of the negative datasets, thus highlighting the enrichment of the identified motifs in effector sequences. Furthermore, we found a specific co-occurrences of at least two CLUMPs in PPN effector sequences bearing protein domains important for invasion and pathogenicity.

The potentiality of this tool goes behind the effector proteins and can be used to easily cluster motifs and calculate the CLUMPs score on any set of protein sequences. Furthermore, we also provide a new scoring system capable of measuring the physicochemical properties of motifs grouped in CLUMPs and a motif alignment algorithm to better explore chemical-physical properties within the CLUMPs. MOnSTER is freely available at https://github.com/paolaporracciolo/MOnSTER_PROMOCA.git.

Materials and methods

Datasets

Oomycetes

We used proteins from five oomycetes species to create the input datasets for MOnSTER, namely *Phytophthora infestans*, *Phytophthora sojae*, *Phytophthora ramorum*, *Hyaloperonospora arabidopsidis* and *Bremia lactucae*.

Positive dataset

The positive dataset consists of 1743 effector proteins belonging to the aforementioned oomycetes obtained from a concatenation of proteins selected from PHI-base database (v4.14) [26], Uniprot (release 2023_02)[28], and the work of Haas et al., (2009) [28], in which they have manually curated the annotations of the proteins. Since the proteins come from different sources, we used CD-HIT (v4.8.1) [29] with the parameters in **Supplementary information,** to filter out identical protein sequences. A total of 1283 proteins are annotated as RxLR effectors, 377 as Crinkler effectors and the last 83 sequences are proteins with no previously identified motif and known to be involved in the host-pathogen interaction.

Negative dataset

Proteins in the negative dataset derive all from Uniprot (release 2023_02) and from the oomycetes species cited before filtered from proteins included in the positive dataset and for evident effector-related annotations. Due to the large amount of non-effector proteins remaining from the filtering we firstly used ‘cd-hit’ to reduce protein sequence redundancy and then, to also reduce the unbalance of the final dataset we refined the selection taking only the representative sequences of the orthogroups found with Orthofinder (v2.5.4) [30]. In total 3009 non effector proteins are included in the negative dataset.

Motif Discovery

The last input file consists in a list of motifs identified as enriched in the sequences of the positive dataset compared to the sequences of the negative one. We used MERCI and STREME (v5.5.1) [31], with parameters detailed in **Supplementary information** [32]. We imposed different lengths for motifs prediction to be inclusive but more stringent on the motifs in which we are interested. STREME’s output is a list of motifs. Hence, we used the tool FIMO (v5.5.1) [33], with default parameters to extract 246 degenerated motifs from the 4524 different motifs.

We obtained the following numbers of non-redundant motifs: 19 with MERCI and 246 with STREME. Then, we removed the identical motifs and created a single non-redundant list containing all the motifs in the same format, which resulted in 265 different motifs.

Plant Parasitic Nematodes (PPNs)

Positive dataset

The positive dataset contains proteins selected to be likely secreted by PPNs in their plant host and belonging to 13 species (*Meloidogyne incognita, Meloidogyne javanica, Meloidogyne arenaria,* *Meloidogyne hapla, Meloidogyne chitwoodi, Meloidogyne graminicola, Globodera rostochiensis, Globodera pallida, Heterodera havenae, Heterodera glycines, Heterodera schachtii, Radopholus similis, Bursaphelenchus xylophilus)*. We could identify a part of these proteins as Ground-Truth (GT) effectors based on their literature description. More precisely we considered as GT effectors those proteins for which *in-situ* hybridization experiments showed that the corresponding transcript is present in nematode secretory glands (dorsal or sub-ventral), implying that these proteins are likely secreted by the nematodes into the host plant. The literature mining led to the extraction of 163 proteins from NCBI GeneBank thanks to the NCBI 'entrez' API. We also manually extracted 41 sequences from the publications' core text and Supplementary information. In addition, we downloaded 41 sequences from WormBase ParaSite (www.parasite.wormbase.org, vWBPS17-WS282 [34], [35]), and eight sequences from nematode.net [36]. In total we obtained 229 GT effectors. We extended the positive dataset with proteins that are non-redundant homologs of the previous GT effectors in PPN proteomes. We first used cd-hit-2D with parameters in **Supplementary information**, to cluster sequences from PPNs proteomes and GT effectors [37]. We then pooled all the GT effectors from closely related *Meloidogyne* species (e.g., *M. incognita*, *M. javanica* and *M. arenaria*) and scanned each corresponding proteome with this multi-species set of sequences using cd-hit. Since the remaining species are genetically distinct, we then scanned each proteome with the relative set of GT effectors, except for *H. havenae* and *M. chitwoodi* for which no proteomes were currently available. We merged the two set of selected effectors and we performed CD-HIT intra- and inter-species to reduce dataset redundancy (parameters in **Supplementary information**), retaining only sequence having more than 1% divergence and aligning on more than 80% of their length (the longest sequence from each cluster was kept). The final positive dataset includes 546 proteins from 13 species.

Negative dataset

The negative dataset is composed of 3849 protein sequences that we obtained by selecting genes widely conserved across the nematode tree of life and close outgroup species, including many species that are non-parasites. Specifically, we filtered the results from a previous analysis [38] and only retained genes from orthogroups i) conserved in more than 90% (62/64) of the analyzed species including two tardigrade species (outgroups), and ii) presenting less than 10 genes/species/orthogroups to avoid multigenic families, which would lead to overrepresentation of some proteins. To remove the redundancy, we used the same strategy as for the positive dataset (cdhit2D first and then CD-HIT).

Motif Discovery

Using the aforementioned software in the same configuration we obtained the following numbers of non-redundant motifs: 40 with MERCI and 229 with STREME applying FIMO. In total, we obtained 269 different motifs.

All datasets are available at https://github.com/paolaporracciolo/MOnSTER_PROMOCA.git and in **Supplementary tables 1.1-1.2 and 2.1-2.2**.

MOnSTER pipeline

The MOnSTER (MOtifs of cluSTERs) pipeline is composed of three main steps as described in **Figure 1** and in the following paragraphs.

Feature calculation

The first step of the pipeline concerns the calculation of parameters that describe protein sequences (**Figure 1**a). To allow an easy calculation of the features on any dataset, we used *ProteinAnalysis* class from the *Bio.SeqUtils.ProtParam,* a python sub-package. We selected 13 features based on individual AA properties, belonging to 4 categories:

- secondary structure propensity ‘helix’ (V, I, Y, F, W, L), ‘turn’ (N, P, G, S), and ‘sheet’ (E, M, A, L)).
- amino-acids dimensions (‘tiny’ (A, C, G, S, T) and ‘small’ (A, C, F, G, I, L, M, P, V, W, Y)).
- pH (‘basic’ (H, K, R), ‘acid’ (B, D, E), and ‘charged’ (H, K, R, B, D, E)).
- physicochemical properties (‘hydropathy-score’, ‘polar’ (D, E, H, K, N, Q, R, S, T), ‘non-polar’ (A, C, F, G, I, L, M, P, V, W, Y), ‘aromatic’ (F, H, W, Y), and ‘aliphatic’ (A, I, L, V)).

We performed feature calculations on the positive and negative datasets and the list of motifs. At the end of this step, we obtained three tables of features, one for each of the input datasets (positive, negative datasets and the list of motifs).

Clustering

This step allowed to cluster motifs based on their properties described by the 13 features. To make the features comparable to each other, we performed data normalization by using the *StandardScaler* method from *sklearn.preprocessing* [39]. This normalization consists of the removal of the mean and the scaling to unit variance.

Then, we performed a hierarchical clustering of the motifs using the Euclidian distance. We then divided the resulting tree into clusters of motifs of proteins (CLUMPs) selecting the threshold distance that minimized the Davies-Bouldin score [40].

For each CLUMP, we removed the redundant motifs. Briefly, we identified motifs that shared a core sequence (for example: ‘HWT in HWTQ’ and ‘GHWTQ’), and we only retained the cores (for instance: “HWT”) in the CLUMPs.

Scoring

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

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

CLUMP score

This score considers the AA composition of the motifs belonging to each CLUMP concerning the preferences of the sequences of the positive dataset. The procedure that we implemented to calculate this score is shown in **Figure 1**b.

a) Feature selection

We used the Mann-Whitney test to identify the features whose values were significantly different between the positive and negative datasets. We only retained the statistically significant features, with a p-value < 0.05. Then, we assigned them a score, by calculating -Log(p-value) of each feature. We will refer to it as the 'feature weight' hereafter.

b) Average calculation

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

c) CLUMPs sorting

We compared the averages of the positive and negative datasets for each feature and sorted CLUMPs accordingly.

Thus, if the $\mu_f^{+\geq\mu_f^-}$, the CLUMPs averages would be sorted in ascending order.

Otherwise $(\mu_f^{+\mu_f^-})$, CLUMPs averages would be sorted in descending order.

d) CLUMPs voting

For each feature, we 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 an 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, we multiplied the 'feature weight' by the CLUMPs vote then we summed all the results using the following formula:

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

where we normalized the value to have a range between 0 and 1.

Modified Jaccard indexes

The two modified Jaccard scores respectively consider: 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 concerning the negatives (**Figure 1**c).

a) CLUMPs occurrences

We 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 we propose two ways to calculate the J index that will be called J1 and J2 hereafter.

To obtain J_1, we 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:

$$J_1 = \sum_{CLUMPs} \frac{1}{2}$$

Where:

$\Delta_{-}$and $\Delta_{+}$the number of occurrences of the motifs of the CLUMP in the negative or

in the positive dataset, respectively.

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

$$J_2 = \sum_{CLUMPs} \frac{1}{2}$$

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

MOnSTER score

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

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

PRO-MOCA: a novel method to create motif logo of CLUMPs

To create motif logos for each CLUMP, we developed a novel method. PRO-MOCA (PROtein-MOtifs Characteristics Aligner) aligns protein motifs based on the characteristics of the amino acids as shown in **Supplementary figure 1**. The first step is to define the alphabets associated with each characteristic that can be used to represent the motifs (**Supplementary figure 1**a). We have defined four alphabets, namely: “chemical”, “hydrophobicity”, “charge”, “secondary structure propensity” (details for each alphabets are included in the **Supplementary information**).

These alphabets are easily modifiable and other alphabets can be included. Different CLUMP logos can be obtained depending on the alphabet chosen. Secondly, PRO-MOCA uses the selected alphabet to translate the AA sequences of each motif in a CLUMP in the new alphabet (**Supplementary figure 1**b). The third step is the alignment (**Supplementary figure 1**c). Briefly, PRO-MOCA screens the translated motif sequences of a CLUMP looking for a “summit position” with the highest frequency of the same “letter” of the novel alphabet (further details in supplementary materials). Once this position is identified, all motifs are aligned accordingly (**Supplementary figure 1**d). Since the motifs of a CLUMP can have different lengths, after the alignment, PRO-MOCA calculates the number of gaps to add at the extremities to make all motifs having the same length. Importantly, gaps are not allowed inside the motif sequences. The last step of the method is to re-translate the aligned motifs in the original AA sequences (**Supplementary figure 1**e). The alignment is ready to feed a program to create logos. Here we used the tool Weblogo3 [41].

PPNs effector protein domains mining analysis

To investigate the relationship between the selected CLUMPs and functional domain in effector proteins we firstly selected proteins from the positive datasets containing at least one occurrence of a selected CLUMP (311 proteins in total). Then we predicted the protein domains with InterProScan (v5.54-87.0) [42] with default parameters. From the results, we extracted the proteins containing the most frequent predicted domains and considered only entries coming from: MobiDB-lite, Coils, CDD, PANTHER, Pfam and ProSitePatterns. Afterwards we also predicted the presence of Signal Peptide (SP) (SignalP4.1 [43]) and TransMembrane (TM) domain regions (TMHMM2.0 [44]). We obtained 258 proteins having at least a CLUMP and one of the aforementioned predicted domains, SP or TM.

Results & Discussion

MOnSTER identified five CLUMPs containing known motifs characteristics of oomycetes effector protein sequences

Characteristic motifs of oomycetes effector proteins are well-known in the literature, such as RxLR, -dEER and LxLFLAK-HVLVxxP [15]. Thus, we reasoned to apply our novel tool, MOnSTER, on oomycetes effectors to test its ability to recover well-characterized motifs. We compiled a set of 4752 oomycetes proteins, comprising 1743 effectors and 3009 non effectors, from five oomycetes species. We performed motif discovery on this set of proteins using MERCI and STREME and we identified 265 significantly enriched motifs (see methods for further details). Then we fed MOnSTER with these motifs and we obtained 11 CLUMPs (**Supplementary table 3**), employing the Davis-Bouldin score, as a criterion to cut the tree. By selecting CLUMPs having a MOnSTER score greater than the median of the overall scores we identified six CLUMPs (CLUMP7, 4, 10, 6, 2 and 9), the first five best-scoring CLUMPs, accordingly to the MOnSTER score, correspond to the known motifs (**Figure 2**). In **Supplementary figure 2** we can also observe that the motifs are respectively grouped in two clades, the two characteristics motifs of CRN-effectors (LxLFLAK and HVLVxxP), form a separate subclade on the right, while the RxLR and -dEER motifs fall into the left clade, resembling the family distinction of effectors to which they belong. More precisely RxLR motifs are divided into 2 different CLUMPs; CLUMP6 containing only RYLR and RFLR motifs, and CLUMP10, containing other RxLR motifs and included in the same sub-clade of the dEER motif (CLUMP2). The last best-scoring CLUMP contains no known motifs, perhaps suggesting a novel putative motif for oomycetes effectors to investigate. Since oomycetes effectors characterization is not in the scope of this article, we did not consider this last CLUMP for further analysis. In support of that, CLUMPs 7, 4, 10, 6 and 2 are present in 1205/1743 effectors (~70% of the sequences in the positive dataset) while in combination with the last significant CLUMP (CLUMP9) only 2 more sequences can be detected.

Thus, we investigated the occurrences and co-occurrences of the five selected CLUMPs in oomycetes effectors and non-effectors (**Supplementary figure 3**). For the effectors we deeply analyzed the two distinct families; in total we found that 68% of the RxLR-effectors in the positive dataset contain the motifs in CLUMPs associated with the RxLR motif (CLUMP10, 6 and 2). In particular, CLUMP10 and 6 are present alone in 41% of the RxLR-effectors (1238/1743 RxLR-effectors), while 19% of the RxLR-effectors contained the co-occurrence of these CLUMPs with the CLUMPs representing the dEER motif (CLUMP2). This reflects the importance of the RxLR motifs in the effector sequences and the role of the attached dEER [45]. On the other hand, the co-occurrence of CLUMPs specific for LxLFLAK and HVLVxxP (CLUMP7 and 4), in CRN-effector sequences accounts for 67% of the relative sequences in the positive dataset (377/1743). The high co-occurrences rate of CLUMP7 and 4 is strongly in agreement with the presence of LxLFLAK and HVLVxxP motif marking the beginning and the end of the DWL-domain in the Crinkler-effector family [28]. For the negative dataset, instead, only 15% of the sequences show the presence of CLUMP-motifs with a huge decrease in CLUMPs co-occurrences. Overall co-occurrences, indeed, are present in around 30% of positive sequences and in 1% of negative ones.

Previous research showed that the motifs characteristics of oomycetes effectors have strong sequence position preferences [46]–[48]. Thus, we plotted the CLUMPs occurrences in the positive versus negative dataset (**Supplementary figure 4**). Indeed, we can observe that the CLUMPs are concentrated at the beginning of the sequence in positive sequences and conversely spread around the sequence of negative dataset proteins. More precisely the 5 most interesting CLUMPs are condensed in the first 40% of the sequence with a higher preference at the very beginning and around 30% of the sequence probably corresponding to the N-terminal of the protein in which the target motifs lie.

Altogether these results highlight the ability of MOnSTER to identify CLUMPs containing biologically relevant motifs.

MOnSTER allowed to identify six CLUMPs characteristics of nematodes effector proteins

The application of MOnSTER of the oomycetes effectors served as a proof of concept of our methodology. Thus, we moved to the characterization of nematodes effector sequences for which no characteristic motifs have been identified yet. We collected a set of 4395 proteins, including 546 well-known effectors and 3849 non-effectors, coming from 13 nematode species. By running motif discovery analysis as for the previous dataset, we found 269 motifs enriched in the effectors sequences. By applying MOnSTER with the previous configuration, the 269 input motifs were grouped into 11 CLUMPs. Six best-scoring CLUMPs were selected using the median as the significative threshold (**Supplementary table 4**). Similar to the oomycetes results, we observe two main clades (**Figure 3**): the second and the third best scoring ones (CLUMP2 and 5 respectively) form a single clade while the other significant CLUMPs (CLUMP1, 3, 7 and 10) are distributed in the bigger clade with the non-significant ones. Overall, we found CLUMPs in almost 60% of sequences from the positive dataset compared to 5% of sequences from the negative.

Then we investigated the presence of the six CLUMPs in each of the 13 PPN species present in the dataset. **Figure 4** shows the abundance of the six best-scoring CLUMPs in the species accordingly to their phylogeny tree. The first three species are the most represented in the positive dataset. Interestingly very distant species show similar CLUMPs frequencies thus suggesting that they might share common characteristics at the sequence level for accomplishing similar functions. Furthermore, we could identify characteristic CLUMPs also for species represented in the dataset with very few sequences reinforcing the previous observation. Overall, this analysis suggests that CLUMPs might be associated with functional properties of PPN nematodes.

Finally, we focused on the positional sequence preferences of CLUMPs in effector sequences (**Supplementary figure 5**). In general, we observe a difference in the position preferences of the best-scoring CLUMPs between positive and negative dataset sequences. The 6 CLUMPs tend to occur more frequently in the middle of the sequences in effectors (positive dataset), with more abundance in central (around 50% of the sequence) and terminal (around 70%), positions. The same CLUMPs are rare in the central position of the non-effector protein sequences (negative dataset). Contrary to the properties of oomycetes effectors, which characteristics CLUMPs occur mainly at the beginning of the sequence, PPN effectors showed a different pattern of occurrences, privileging a central – C terminal occurrence.

Co-occurrences of different CLUMPs are associated with functional protein domains.

We investigated the co-occurrence patterns of CLUMPs in the PPNs effector sequences (all possible combinations of co-occurrences are reported in **Supplementary figure 6)**. Overall, we notice that

CLUMPs tend to co-occur more frequently in the sequences of the positive dataset than in the negative one, despite the positive set being smaller than the negative one. 30% of effector sequences show co-occurrences of the 6 selected CLUMPs, while in the non-effectors, co-occurrences, are present in less than 1% of the sequences. As observed for oomycetes, some CLUMPs tend to be present alone, while others tend to co-occur with specific CLUMPs. This suggests that different classes of nematode effectors might exist, similar to the oomycetes effectors. Importantly, there is no relationship between the sequence length and the number of co-occurrences that might suggest a functional role for CLUMPs co-occurrences (**Supplementary figure 7**).

To inspect further a putative functional role of CLUMPs in effector sequences, we queried the effectors having at least one CLUMP or a co-occurrence of multiple CLUMPs against several protein domain databases (see Methods, results in **Figure 5** and **Supplementary table 5**). The most recurrent hits are the coil domain, intrinsically disordered domain and the presence of the signal peptide (SP) followed by the pectate lyase domain, glycosyl hydrolase family 5, Stichodactyla toxin (ShK) domain, 14-3-3 family and cysteine-rich domain. Interestingly, we observe the almost exclusive association between CLUMPs and functional domains, mainly when multiple CLUMPs co-occur in effector sequences.

The strongest association that we observe is between the co-occurrences of CLUMPs 7 and 10 and the glycosyl hydrolase family 5 domain on one hand and the co-occurrences of CLUMPs 3, 7, 10 and the cysteine-rich domain, on the other hand. Specifically, all 23 effector sequences containing the co-occurrences of CLUMP 7 and 10 bear also the glycosyl hydrolase family 5 domain. By inspecting the position of CLUMPs occurrences within the sequences, we observed that the two CLUMPs are flanking the domain: CLUMP7 is consistently present at the beginning of the sequence and consequently of the domain, while CLUMP10 mostly concentrates at the end of the domain, around 60-80% of the sequences (**Supplementary figure 8**). Examples of these genes in nematodes is poorly characterized and likely resulting from horizontal transfer [49], [50]. Similarly, all 17 sequences presenting the co-occurrence of CLUMPs 3, 7,10 also contain the cysteine-rich domain. Cysteine-rich domain and CAP protein are known to be involved in the virulence of nematodes [51]. They are expressed in both plants and pathogens; in the latter, they are important for their virulence by suppressing the host's immune responses and promoting colonization. Interestingly, these sequences do not contain disordered regions or coil domains, consistently with unique conserved sandwich fold with a large central cavity of these kinds of proteins [52]. 16 out of 19 sequences presenting co-occurrences of CLUMPs 2, 3 have also the 14-3-3 family domain, a eukaryotic-specific protein family with a general role in the signal transduction [53]. We also observe only one motif from CLUMP 2 in these sequences (KDKM) and 4 from CLUMP 3 (NKDKAC, KMKG, PTHPIR, PTHP). 13 out of 34 sequences bearing only CLUMP 1 also contain the pectate lyase domain. Of note, these sequences do not contain coiled or disordinate regions, and only 7 show the presence of the SP. Pectate lyase enzymes in nematodes facilitate penetration in plant-cell walls made of pectin [54]. numerous recent reports showed that these enzymes are produced in specialized nematode gland cells and secreted during the parasitism process. In the case of sedentary endo-parasitic nematodes, this occurs mainly during juvenile migration through the root tissue, when these enzymes play a crucial role in the maceration of the plant tissue facilitating the infection [55]. Finally, 8 out of 22 sequences bear the co-occurrences of CLUMPs 2, 5 and the ShK domain. Although the exact biological function of the ShKT domain remains unclear, previous reports have shown that this domain might be associated with immunosuppression [56], [57].

Overall, these findings highlight that specific CLUMPs co-occurrences are associated with specific functional domains with roles in invasion and/or infection and might suggest different classes of effectors cross-species.

Conclusions

This work is structured around three main aims: (1) the development of a novel method to cluster and score discriminant motifs of protein sequences called MOnSTER, (2) the validation of the MOnSTER results by applying it to identify CLUMPs specific to effector protein sequences of oomycetes (3) the application of MOnSTER to protein sequences from plant-parasitic nematodes with unprecedented discriminant motifs detection.

The application of MOnSTER on oomycetes yielded the identification of five CLUMPs corresponding to the well-known effector-related motifs like RxLR-dEER and LxLFLAK-HVLVxxP motifs in oomycetes. This demonstrated that the novel scoring method introduced by MOnSTER is a good parameter with which calculate CLUMP specificity for effector protein sequences. When applied on the nematodes effectors, MOnSTER found six novel CLUMPs, not previously characterized. The main advantage of MOnSTER is that the definition of CLUMPs allowed us to reduce the degeneration of 265 and 269 motifs (oomycetes and nematodes respectively), to 11 CLUMPs. Effectors sequences of both pathogens show some common characteristics. Indeed, selected CLUMPs-motifs are present in about 70% of the input effector proteins for oomycetes and 60% in PPN compared to 15% and 5% in of the non-effector proteins, respectively. Furthermore, around 30% of effector sequences have co-occurring CLUMPs, in contrast with less than 1% of the non-effector sequences, in both applications. The main difference between effector-specific motifs of the two pathogens is the positional preference: the beginning of the sequence for oomycetes and central C-terminal for PPNs. This highlights MOnSTER ability to cluster motifs specifically relevant for effector sequences without privilege any portion of the sequence, like other motif discovery tools.

Concerning the novel identified motifs for PPNs effectors, we observed that the pattern of occurrences and co-occurrences of CLUMPs in effector sequences is associated with specific functional domains and might suggest the existence of different classes of effectors. Importantly we did not observe any species-related preferences thus implying the generality of these results.

In conclusion, MOnSTER quantifies the motifs and sequence properties in each dataset provided, thus allowing a wide application to other protein classes. Since the MOnSTER score considers the physicochemical properties and occurrences of motifs in CLUMPs concerning the protein sequences provided, it works without the need for a reference dataset. Furthermore, the MOnSTER scores are normalized values, therefore, allowing direct comparison between different studies.

Our results highlighted that MOnSTER is a powerful new method to cluster and score discriminant motifs in protein sequences according to their physicochemical properties and pattern of occurrences. It is also a tool that can be easily used on any set of protein sequences and a list of motifs. As such, MOnSTER can be included in any pipeline needing motif calling and will be of great use to accelerate both computational and experimental studies relating to protein motif discovery.

Data availability

The source code and related data are available at: https://github.com/paolaporracciolo/MOnSTER_PROMOCA.git

Funding

This work was supported by the French government, through the UCA JEDI Investments in the Future project managed by the National Research Agency (ANR) under reference number ANR-15-IDEX-01.

Competing Interests

The authors declare that they have no competing interests.

References

[1] T. Y. Toruño, I. Stergiopoulos, and G. Coaker, “Plant-Pathogen Effectors: Cellular Probes Interfering with Plant Defenses in Spatial and Temporal Manners,” *Annu. Rev. Phytopathol.*, vol. 54, pp. 419–441, Aug. 2016, doi: 10.1146/annurev-phyto-080615-100204.

[2] A. Haegeman, S. Mantelin, J. T. Jones, and G. Gheysen, “Functional roles of effectors of plant-parasitic nematodes,” *Gene*, vol. 492, no. 1, pp. 19–31, Jan. 2012, doi: 10.1016/j.gene.2011.10.040.

[3] C. Selin, T. R. de Kievit, M. F. Belmonte, and W. G. D. Fernando, “Elucidating the Role of Effectors in Plant-Fungal Interactions: Progress and Challenges,” *Front. Microbiol.*, vol. 7, 2016, [Online]. Available: https://www.frontiersin.org/articles/10.3389/fmicb.2016.00600

[4] D. M. Bird, J. T. Jones, C. H. Opperman, T. Kikuchi, and E. G. J. Danchin, “Signatures of adaptation to plant parasitism in nematode genomes,” *Parasitology*, vol. 142 Suppl 1, no. Suppl 1, pp. S71-84, Feb. 2015, doi: 10.1017/S0031182013002163.

[5] J. Sperschneider *et al., "* Evaluation of Secretion Prediction Highlights Differing Approaches Needed for Oomycete and Fungal Effectors", Frontiers in Plant Science, vol. 6, 2015, doi:10.3389/fpls.2015.01168.

[6] H. Sonah and R.K. Deshmukh, "Computational Prediction of Effector Proteins in Fungi: Opportunities and Challenges", Front Plant Sci., vol. 12, pp. 7:126, Feb. 2016, doi:10.3389/fpls.2016.00126.

[7] L. Liu *et al.*, “Arms race: diverse effector proteins with conserved motifs,” *Plant Signal. Behav.*, vol. 14, no. 2, p. 1557008, Jan. 2019, doi: 10.1080/15592324.2018.1557008.

[8] P. Dean, “Functional domains and motifs of bacterial type III effector proteins and their roles in infection,” *FEMS Microbiol. Rev.*, vol. 35, no. 6, pp. 1100–1125, Nov. 2011, doi: 10.1111/j.1574-6976.2011.00271.x.

[9] E. R. Green and J. Mecsas, “Bacterial Secretion Systems: An Overview,” *Microbiol. Spectr.*, vol. 4, no. 1, Feb. 2016, doi: 10.1128/microbiolspec.VMBF-0012-2015.

[10] P. Natale, T. Brüser, and A. J. M. Driessen, “Sec- and Tat-mediated protein secretion across the bacterial cytoplasmic membrane—Distinct translocases and mechanisms,” *Biochim. Biophys. Acta BBA - Biomembr.*, vol. 1778, no. 9, pp. 1735–1756, Sep. 2008, doi: 10.1016/j.bbamem.2007.07.015.

[11] J. Sperschneider, P. N. Dodds, D. M. Gardiner, J. M. Manners, K. B. Singh, and J. M. Taylor, “Advances and Challenges in Computational Prediction of Effectors from Plant Pathogenic Fungi,” *PLOS Pathog.*, vol. 11, no. 5, p. e1004806, May 2015, doi: 10.1371/journal.ppat.1004806.

[12] G. W. Beakes, S. L. Glockling, and S. Sekimoto, “The evolutionary phylogeny of the oomycete ‘fungi,’” *Protoplasma*, vol. 249, no. 1, pp. 3–19, Jan. 2012, doi: 10.1007/s00709-011-0269-2.

[13] M. Thines and S. Kamoun, “Oomycete-plant coevolution: recent advances and future prospects,” *Curr. Opin. Plant Biol.*, vol. 13, no. 4, pp. 427–433, Aug. 2010, doi: 10.1016/j.pbi.2010.04.001.

[14] K. J. Wood *et al.*, “Effector prediction and characterization in the oomycete pathogen Bremia lactucae reveal host-recognized WY domain proteins that lack the canonical RXLR motif,” *PLOS Pathog.*, vol. 16, no. 10, p. e1009012, Oct. 2020, doi: 10.1371/journal.ppat.1009012.

[15] M. Franceschetti, A. Maqbool, M. J. Jiménez-Dalmaroni, H. G. Pennington, S. Kamoun, and M. J. Banfield, “Effectors of Filamentous Plant Pathogens: Commonalities amid Diversity,” *Microbiol. Mol. Biol. Rev.*, vol. 81, no. 2, pp. e00066-16, Mar. 2017, doi: 10.1128/MMBR.00066-16.

[16] R. H. Y. Jiang, S. Tripathy, F. Govers, and B. M. Tyler, “RXLR effector reservoir in two Phytophthora species is dominated by a single rapidly evolving superfamily with more than 700 members,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 105, no. 12, pp. 4874–4879, Mar. 2008, doi: 10.1073/pnas.0709303105.

[17] T. A. Torto *et al.*, “EST mining and functional expression assays identify extracellular effector proteins from the plant pathogen Phytophthora,” *Genome Res.*, vol. 13, no. 7, pp. 1675–1685, Jul. 2003, doi: 10.1101/gr.910003.

[18] L. Pritchard and P. Birch, “A systems biology perspective on plant–microbe interactions: Biochemical and structural targets of pathogen effectors,” *Plant Sci.*, vol. 180, no. 4, pp. 584–603, Apr. 2011, doi: 10.1016/j.plantsci.2010.12.008.

[19] A. H. Lovelace, S. Dorhmi, M. T. Hulin, Y. Li, J. W. Mansfield, and W. Ma, “Effector Identification in Plant Pathogens,” *Phytopathology*, vol. 113, no. 4, pp. 637–650, Apr. 2023, doi: 10.1094/PHYTO-09-22-0337-KD.

[20] J. T. Jones *et al.*, “Top 10 plant-parasitic nematodes in molecular plant pathology,” *Mol. Plant Pathol.*, vol. 14, no. 9, pp. 946–961, Dec. 2013, doi: 10.1111/mpp.12057.

[21] M. Holterman *et al.*, “Disparate gain and loss of parasitic abilities among nematode lineages,” *PloS One*, vol. 12, no. 9, p. e0185445, 2017, doi: 10.1371/journal.pone.0185445.

[22] C. Vens, M.-N. Rosso, and E. G. J. Danchin, “Identifying discriminative classification-based motifs in biological sequences,” *Bioinformatics*, vol. 27, no. 9, pp. 1231–1238, May 2011, doi: 10.1093/bioinformatics/btr110.

[23] R. Blanc-Mathieu *et al.*, “Hybridization and polyploidy enable genomic plasticity without sex in the most devastating plant-parasitic nematodes,” *PLOS Genet.*, vol. 13, no. 6, p. e1006777, Jun. 2017, doi: 10.1371/journal.pgen.1006777.

[24] P. Abad *et al.*, “Genome sequence of the metazoan plant-parasitic nematode Meloidogyne incognita,” *Nat. Biotechnol.*, vol. 26, no. 8, pp. 909–915, Aug. 2008, doi: 10.1038/nbt.1482.

[25] N. E. Davey, M. S. Cyert, and A. M. Moses, “Short linear motifs – ex nihilo evolution of protein regulation,” *Cell Commun. Signal.*, vol. 13, no. 1, p. 43, Dec. 2015, doi: 10.1186/s12964-015-0120-z.

[25] E. D. O. Roberson, “Motif scraper: a cross-platform, open-source tool for identifying degenerate nucleotide motif matches in FASTA files,” *Bioinformatics*, vol. 34, no. 22, pp. 3926–3928, Nov. 2018, doi: 10.1093/bioinformatics/bty437.

[26] M. Urban, R. Pant, A. Raghunath, A. G. Irvine, H. Pedro, and K. E. Hammond-Kosack, “The Pathogen-Host Interactions database (PHI-base): additions and future developments,” *Nucleic Acids Res.*, vol. 43, no. Database issue, pp. D645-655, Jan. 2015, doi: 10.1093/nar/gku1165.

[27] The UniProt Consortium, “UniProt: the Universal Protein Knowledgebase in 2023,” *Nucleic Acids Res.*, vol. 51, no. D1, pp. D523–D531, Jan. 2023, doi: 10.1093/nar/gkac1052.

[28] B. J. Haas *et al.*, “Genome sequence and analysis of the Irish potato famine pathogen Phytophthora infestans,” *Nature*, vol. 461, no. 7262, Art. no. 7262, Sep. 2009, doi: 10.1038/nature08358.

[29] L. Fu, B. Niu, Z. Zhu, S. Wu, and W. Li, “CD-HIT: accelerated for clustering the next-generation sequencing data,” *Bioinformatics*, vol. 28, no. 23, pp. 3150–3152, Dec. 2012, doi: 10.1093/bioinformatics/bts565.

[30] D. M. Emms and S. Kelly, “OrthoFinder: phylogenetic orthology inference for comparative genomics,” *Genome Biol.*, vol. 20, no. 1, p. 238, Nov. 2019, doi: 10.1186/s13059-019-1832-y.

[31] T. L. Bailey, “STREME: accurate and versatile sequence motif discovery,” *Bioinformatics*, vol. 37, no. 18, pp. 2834–2840, Sep. 2021, doi: 10.1093/bioinformatics/btab203.

[32] C. E. Grant, T. L. Bailey, and W. S. Noble, “FIMO: scanning for occurrences of a given motif,” *Bioinformatics*, vol. 27, no. 7, pp. 1017–1018, Apr. 2011, doi: 10.1093/bioinformatics/btr064.

[33] K. L. Howe *et al.*, “WormBase 2016: expanding to enable helminth genomic research,” *Nucleic Acids Res.*, vol. 44, no. D1, pp. D774–D780, Jan. 2016, doi: 10.1093/nar/gkv1217.

[34] K. L. Howe, B. J. Bolt, M. Shafie, P. Kersey, and M. Berriman, “WormBase ParaSite − acomprehensive resource for helminth genomics,” *Mol. Biochem. Parasitol.*, vol. 215, pp. 2–10, Jul. 2017, doi: 10.1016/j.molbiopara.2016.11.005.

[35] J. Martin *et al.*, “Helminth.net: expansions to Nematode.net and an introduction to Trematode.net,” *Nucleic Acids Res.*, vol. 43, no. D1, pp. D698–D706, Jan. 2015, doi: 10.1093/nar/gku1128.

[36] W. Li and A. Godzik, “Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences,” *Bioinformatics*, vol. 22, no. 13, pp. 1658–1659, Jul. 2006, doi: 10.1093/bioinformatics/btl158.

[37] P. Grynberg *et al.*, “Comparative Genomics Reveals Novel Target Genes towards Specific Control of Plant-Parasitic Nematodes,” *Genes*, vol. 11, no. 11, p. 1347, Nov. 2020, doi: 10.3390/genes11111347.

[38] R. Blanc-Mathieu *et al.*, “Hybridization and polyploidy enable genomic plasticity without sex in the most devastating plant-parasitic nematodes,” *PLOS Genet.*, vol. 13, no. 6, p. e1006777, Jun. 2017, doi: 10.1371/journal.pgen.1006777.

[39] F. Pedregosa *et al.*, “Scikit-learn: Machine Learning in Python,” *Mach. Learn. PYTHON*, p. 6.

[40] D. L. Davies and D. W. Bouldin, “A Cluster Separation Measure,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-1, no. 2, pp. 224–227, Apr. 1979, doi: 10.1109/TPAMI.1979.4766909.

[41] G. E. Crooks, G. Hon, J.-M. Chandonia, and S. E. Brenner, “WebLogo: A Sequence Logo Generator,” p. 3.

[42] P. Jones *et al.*, “InterProScan 5: genome-scale protein function classification,” *Bioinformatics*, vol. 30, no. 9, pp. 1236–1240, May 2014, doi: 10.1093/bioinformatics/btu031.

[43] H. Nielsen, “Predicting Secretory Proteins with SignalP,” in *Protein Function Prediction: Methods and Protocols*, D. Kihara, Ed., in Methods in Molecular Biology. New York, NY: Springer, 2017, pp. 59–73. doi: 10.1007/978-1-4939-7015-5_6.

[44] A. Krogh, B. Larsson, G. von Heijne, and E. L. Sonnhammer, “Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes,” *J. Mol. Biol.*, vol. 305, no. 3, pp. 567–580, Jan. 2001, doi: 10.1006/jmbi.2000.4315.

[45] L. S. Boutemy *et al.*, “Structures of Phytophthora RXLR effector proteins: a conserved but adaptable fold underpins functional diversity,” *J. Biol. Chem.*, vol. 286, no. 41, pp. 35834–35842, Oct. 2011, doi: 10.1074/jbc.M111.262303.

[46] S. Schornack *et al.*, “Ancient class of translocated oomycete effectors targets the host nucleus,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 107, no. 40, pp. 17421–17426, Oct. 2010, doi: 10.1073/pnas.1008491107.

[47] A. P. Rehmany *et al.*, “Differential recognition of highly divergent downy mildew avirulence gene alleles by RPP1 resistance genes from two Arabidopsis lines,” *Plant Cell*, vol. 17, no. 6, pp. 1839–1850, Jun. 2005, doi: 10.1105/tpc.105.031807.

[48] T. M. M. M. Amaro, G. J. A. Thilliez, G. B. Motion, and E. Huitema, “A Perspective on CRN Proteins in the Genomics Age: Evolution, Classification, Delivery and Function Revisited,” *Front. Plant Sci.*, vol. 8, 2017, Accessed: Jun. 01, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpls.2017.00099

[49] E. G. J. Danchin *et al.*, “Multiple lateral gene transfers and duplications have promoted plant parasitism ability in nematodes,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 107, no. 41, pp. 17651–17656, Oct. 2010, doi: 10.1073/pnas.1008486107.

[50] H. Aspeborg, P. M. Coutinho, Y. Wang, H. Brumer, and B. Henrissat, “Evolution, substrate specificity and subfamily classification of glycoside hydrolase family 5 (GH5),” *BMC Evol. Biol.*, vol. 12, no. 1, p. 186, Sep. 2012, doi: 10.1186/1471-2148-12-186.

[51] Z. Han, D. Xiong, R. Schneiter, and C. Tian, “The function of plant PR1 and other members of the CAP protein superfamily in plant–pathogen interactions,” *Mol. Plant Pathol.*, vol. 24, no. 6, pp. 651–668, 2023, doi: 10.1111/mpp.13320.

[52] G. M. Gibbs, K. Roelants, and M. K. O’Bryan, “The CAP superfamily: cysteine-rich secretory proteins, antigen 5, and pathogenesis-related 1 proteins--roles in reproduction, cancer, and immune defense,” *Endocr. Rev.*, vol. 29, no. 7, pp. 865–897, Dec. 2008, doi: 10.1210/er.2008-0032.

[53] R. Lozano-Durán and S. Robatzek, “14-3-3 Proteins in Plant-Pathogen Interactions,” *Mol. Plant-Microbe Interactions®*, vol. 28, no. 5, pp. 511–518, May 2015, doi: 10.1094/MPMI-10-14-0322-CR.

[54] K. Wieczorek *et al.*, “A Distinct Role of Pectate Lyases in the Formation of Feeding Structures Induced by Cyst and Root-Knot Nematodes,” *Mol. Plant-Microbe Interactions®*, vol. 27, no. 9, pp. 901–912, Sep. 2014, doi: 10.1094/MPMI-01-14-0005-R.

[55] T. Hewezi and T. J. Baum, “Manipulation of plant cells by cyst and root-knot nematode effectors,” *Mol. Plant-Microbe Interact. MPMI*, vol. 26, no. 1, pp. 9–16, Jan. 2013, doi: 10.1094/MPMI-05-12-0106-FI.

[56] H. Song *et al.*, “The Meloidogyne javanica effector Mj2G02 interferes with jasmonic acid signalling to suppress cell death and promote parasitism in Arabidopsis,” *Mol. Plant Pathol.*, vol. 22, no. 10, pp. 1288–1301, Oct. 2021, doi: 10.1111/mpp.13111.

[57] J. Niu *et al.*, “Msp40 effector of root-knot nematode manipulates plant immunity to facilitate parasitism,” *Sci. Rep.*, vol. 6, no. 1, Art. no. 1, Jan. 2016, doi: 10.1038/srep19443.

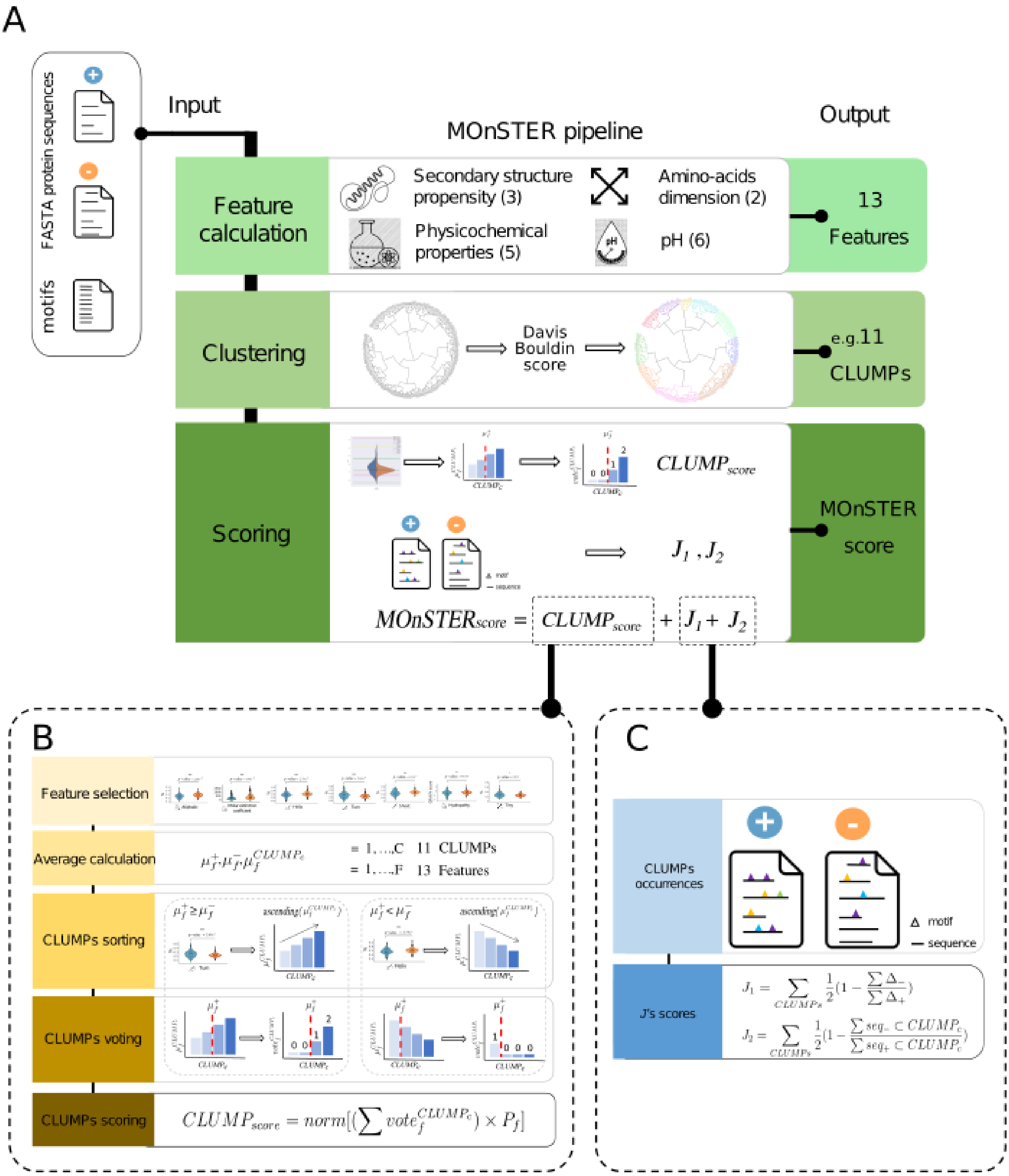

A
FASTA protein sequences
motifs
Input
MOnSTER pipeline
Output
Feature calculation
Secondary structure propensity (3)
Amino-acids dimension (2)
Physicochemical properties (5)
pH (6)
13 Features
Clustering
Davis Bouldin score
e.g.11 CLUMPs
Scoring
$CLUMP_{score}$
J_1, J_2
motif
sequence
$MOnSTER_{score} = CLUMP_{score} + J_1 + J_2$
MOnSTER score
B
Feature selection
Average calculation
$\mu_f^+, \mu_f^-, \mu_f^{CLUMP_c}$
= 1,…,C 11 CLUMPs
= 1,…,F 13 Features
CLUMPs sorting
$\mu_f^+ \geq \mu_f^-$
$ascending(\mu_f^{CLUMP_c})$
$\mu_f^+ < \mu_f^-$
CLUMPs voting
μ_f^+
0 0 1 2
1 0 0 0
CLUMPs scoring
$CLUMP_{score} = norm[(\sum vote_f^{CLUMP_c}) \times P_f]$
C
CLUMPs occurrences
motif
sequence
J's scores
$J_1 = \sum_{CLUMPs} \frac{1}{2}(1 - \frac{\sum \Delta_-}{\sum \Delta_+})$
$J_2 = \sum_{CLUMPs} \frac{1}{2}(1 - \frac{\sum seq_- \subset CLUMP_c}{\sum seq_+ \subset CLUMP_c})$

Figure 1: MOnSTER pipeline scheme.

(A) MOnSTER pipeline is composed of three steps. It takes two FASTA protein sequences datasets (positive and negative) and a list of predicted motifs (enriched in the positive dataset) as input. The output is a list of CLUMPs and an associated MOnSTER score. The MOnSTER score is constituted by: (B) $CLUMP_{score}$ calculation. (C) Two modified Jaccard Indexes.

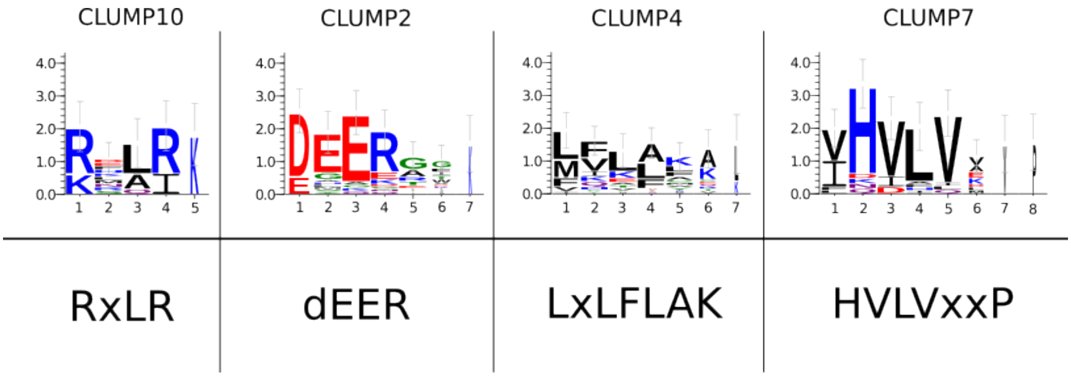


Figure 2: Motif logos of CLUMPs compared to the target motifs.

Upper-panel: alignments of motifs in the respective CLUMP are produced by PROMOCA, and then the aligned motif sequences are used to produce the logos with WebLogo3. The x-axis represents the AA position in the motif, while the y-axis represents log-transformed frequencies translated into bits of information. Lower-panel: characteristic motifs of oomycetes effectors families from literature.

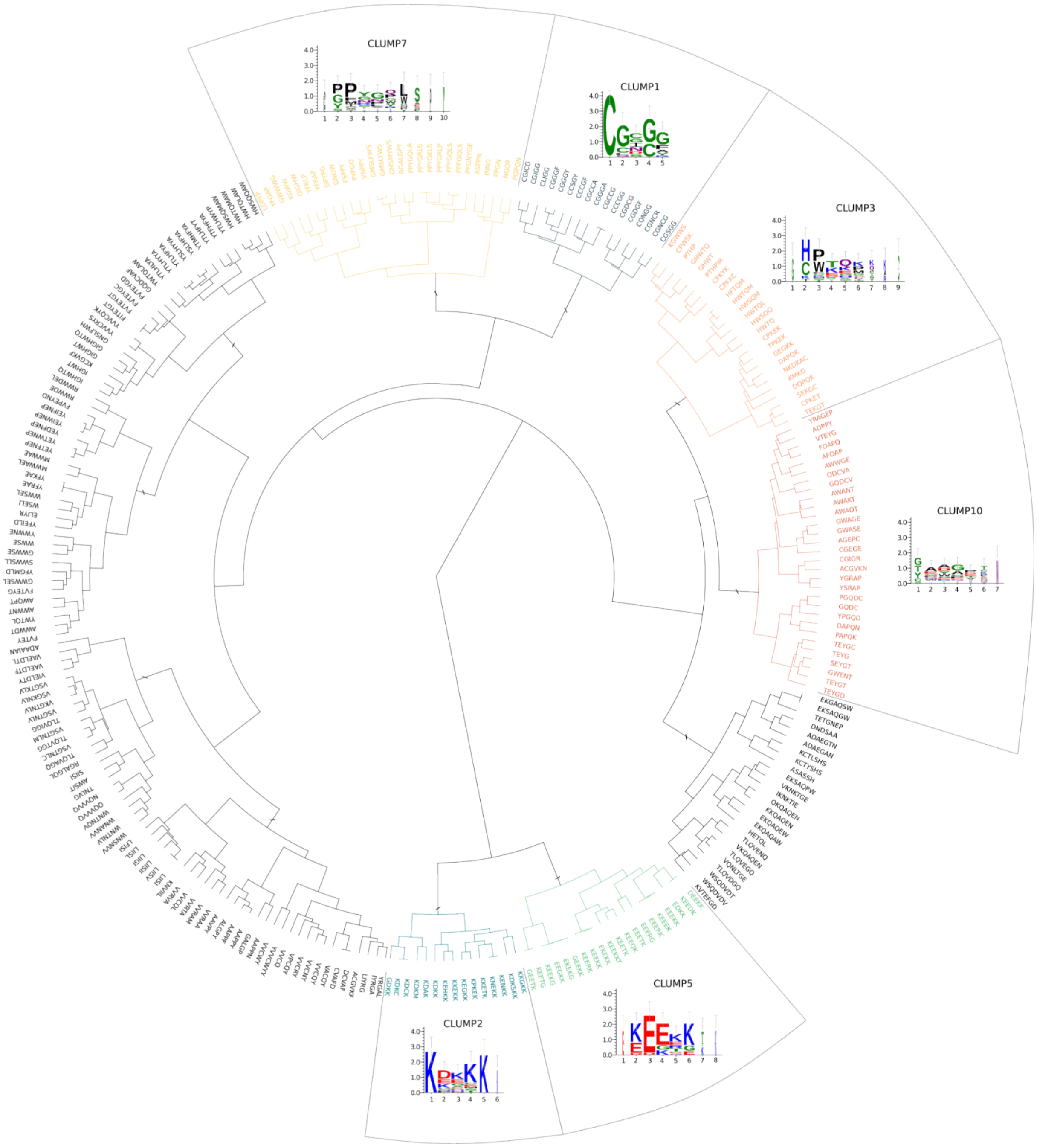


Figure 3: Dendrogram of CLUMPs in Plant Parasitic Nematodes (PPNs)
11 CLUMPs produced by MOnSTER (indicated with "/" sign). The coloured ones are those selected as best-scoring CLUMPs after MOnSTER-score calculation. Each best-scoring CLUMP is associated with the corresponding motif logo; alignment of motifs in each CLUMP is produced by PROMOCA and then WebLogo 3 is used to produce the image (the x-axis shows the AA position of the motif and the y-axis represents the log-transformed frequency of each AA in terms of bits of information).

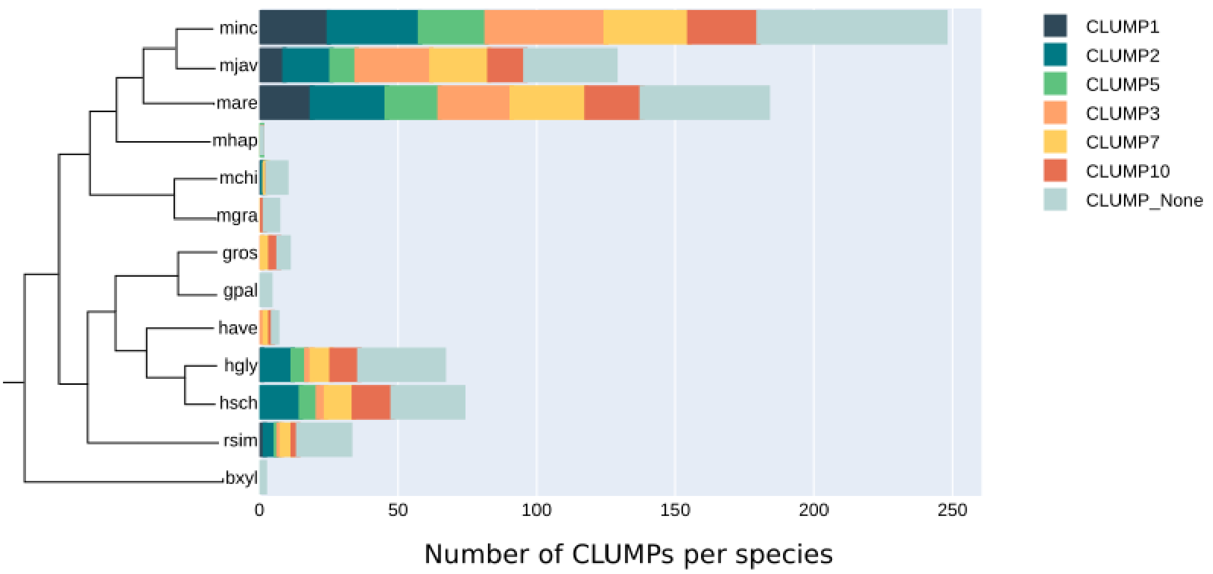
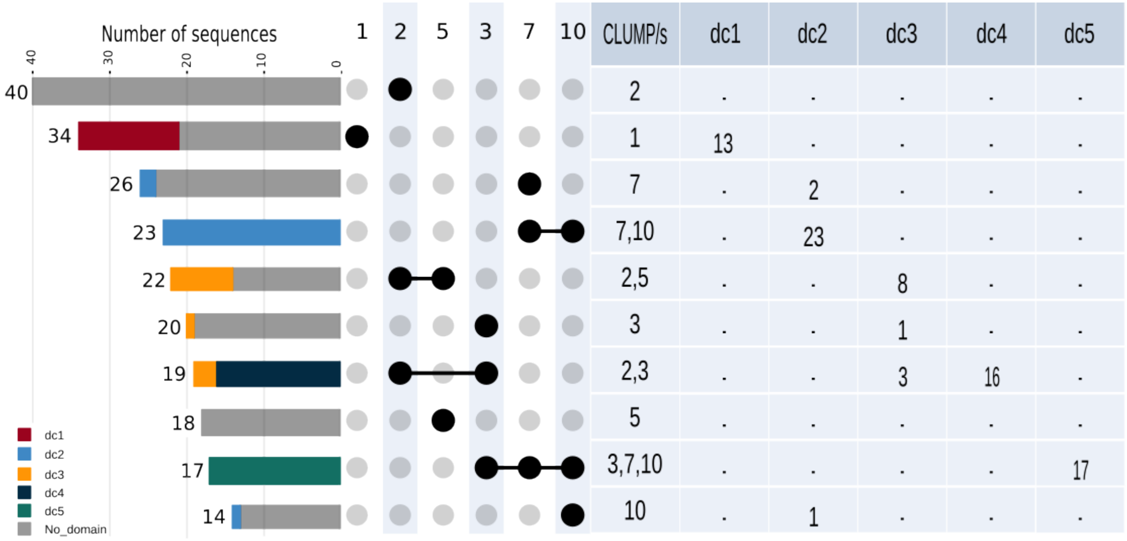


Figure 4: Cardinality of CLUMPs-motifs in each PPN species considered. The total number of motifs belonging to each significant CLUMP per PPN species accordingly to their phylogeny. (minc: *Meloidogyne incognita,* mjav: *Meloidogyne javanica,* mare: *Meloidogyne arenaria,* mhap: *Meloidogyne hapla,* 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*)



CLUMP/s	dc1	dc2	dc3	dc4	dc5
2	.	.	.	.	.
1	13	.	.	.	.
7	.	2	.	.	.
7,10	.	23	.	.	.
2,5	.	.	8	.	.
3	.	.	1	.	.
2,3	.	.	3	16	.
5	.	.	.	.	.
3,7,10	.	.	.	.	17
10	.	1	.	.	.

Figure 5: Effector proteins showing the presence of CLUMP/s associated with pathogenicity-related protein domain/s.

The table on the right shows the co-occurrence of CLUMP or CLUMPs with specific domain classes (dc); dc1, pectate lyase domain class, dc2, glycosyl hydrolase family 5 domain class, dc3 Stichodactyla toxin (ShK) domain class, dc4 14-3-3 family domain class and dc5, cysteine-rich domain class. The upset plot on the left represents the occurrences and co-occurrences of respective CLUMPs in the positive dataset, highlighting the sequences that also have an interesting protein domain following the table counts.