Identification of Putative Parasitism Genes in Plant-Parasitic Nematodes

In silico Screening of Whole Genomes and Transcriptomes

Amandine CAMPAN-FOURNIER[1], Laetitia PERFUS-BARBEOCH[1], Marie-Noëlle ROSSO[1], Marie-Jeanne ARGUEL[1], Corinne DA SILVA[2], Celine VENS[3], Nathalie MARTEU[1], Karine LABADIE[2], François ARTIGUENAVE[2], Pierre ABAD[1] and Etienne G.J. DANCHIN[1]

[1] Plant-Nematode Interactions, UMR 1301 INRA - UNS - CNRS, 400 route des Chappes, 06903, Sophia-Antipolis Cedex, France
amandine.fournier@sophia.inra.fr

[2] Institut de Génomique, Genoscope, CEA, 2 rue Gaston Crémieux, 91000, Evry, France

[3] Department of Computer Science, Katholieke Universiteit Leuven, Celestijnenlaan, 3001, Leuven, Belgium

Abstract *Plant-parasitic nematodes (PPN) are microscopic roundworms that cause disease on nearly all economically important crop plants. They are responsible for estimated losses of several billion Euros/year. Currently, nematicides are the most important means of controlling nematodes. However, current nematicides are non-specific, notoriously toxic and pose a threat to soil ecosystem, ground water and human health. Therefore, novel and specific targets are needed to develop new strategies directed against plant-parasitic nematode species.*

In this context, our project aims at identifying putative parasitism genes in plant-parasitic nematodes. A multi-disciplinary approach combining bioinformatics and functional genomics based on large-scale screening of genomic and proteomic data from nematodes showing different modes of plant parasitism is proposed. Candidate targets are identified by bioinformatics methods and the most promising candidates will be selected for further functional analyses.

Here we report the semi-automated bioinformatics pipeline developed for that purpose. We have undertaken a comparative analysis of the sets of predicted proteins in Meloidogyne incognita and Meloidogyne hapla (two fully sequenced plant-parasitic nematodes) with a large dataset of whole genomes and transcriptomes. As our objective is to identify druggable parasitism genes we have searched for proteins conserved in other parasitic or plant-associated species, but absent from species that could be negatively affected by newly developed drugs or control means. We also have undertaken bioinformatics annotations of these proteins, including but not limited to: detection of signal peptide and Pfam domains, assignment of gene ontology terms and identification of specific motifs.

Keywords Plant-parasitic nematodes (PPN), parasitism genes, *in silico* screening, automatic functional annotation, comparative genomics.

1 Introduction

Plant-parasitic nematodes (PPN) are microscopic roundworms. Their strategy to infest plants and their host range depend on the species. Most of them feed on root tissue and damage their host mainly by altering root growth (resulting in reduced water uptake), by promoting microbial infections through wound sites or by serving as vectors for pathogenic viruses. They cause disease on nearly all economically important crop plants, including corn, soybean, cotton, rice, tomato, carrots and tobacco. They are thus responsible for estimated losses of several billion Euros/year. The most economically impacting plant-parasitic nematodes are root-knot nematodes and cyst nematodes that both pertain to the phylum Tylenchida (or clade 12) [1]. During their life cycle, these plant-parasitic nematodes penetrate the root and migrate within plant tissue. They induce the development of a specialized feeding structure from root plant cells and settle sedentary at this feeding site. It has been shown that they secrete proteins called “effectors” in plant tissue. Several of these effectors have been shown to be involved in degradation of the plant cell wall or potentially implicated in modulation of plant defenses. Many effectors remain uncharacterized and are suspected to be involved in the development of feeding structures or other processes related to successful parasitism.

Measures such as growing resistant crop varieties and the use of nematicides are extensively employed to control plant-parasitic nematodes. However, it happens that some nematodes overcome resistance genes and become “virulent” (able to infect varieties that were previously resistant to these nematodes). Moreover, current nematicides are costly, non-specific, notoriously toxic and pose a threat to the soil ecosystem, ground water and human health. This has lead to the banning of the most efficient chemicals that were previously commonly used. Therefore, novel and specific targets are needed to develop new strategies directed against plant-parasitic nematode species.

In 2008, the careful analysis of the first sequenced genome of a plant-parasitic animal, the root-knot nematode *Meloidogyne incognita* [2], highlighted new potential targets for anti-parasitic strategies. To confirm the relevance of these genes as good candidate targets, efforts are needed to produce high-throughput data on additional plant-parasitic nematode species. Indeed, very few genomic and transcriptomic resources were available so far: only two genomes of plant-parasitic nematodes (*M. incognita* and *Meloidogyne hapla* [3]) are fully sequenced and annotated, and most EST come from species from the Meloidogyne genus (preventing comparative studies).

That is why we propose: (i) an in-depth search for potential new targets by comparaison of the *M. incognita* and *M. hapla* sets of predicted proteins with proteins from other organisms (parasitic or not) ; (ii) the generation and analysis of large-scale transcriptomic data (RNA-seq) from four other plant-parasitic nematodes representing diverse parasitic strategies (*Pratylenchus coffeae*, *Ditylenchus dipsaci*, *Bursaphelencus xylophilus* and *Xiphinema index*). The rationale of our analysis is that the more a protein is broadly conserved across parasitic or plant-associated species yet restricted to them, the more it is likely to be involved in the parasitism process. Thus, conservation of a protein in a parasitic or a plant-associated species is considered like a “bonus”. In contrast, we call “forbidden species” hereinafter the species that are neither plant-parasitic, nor plant-associated, and that could be negatively affected by the development of novel drugs or control means. In concrete terms, by “forbidden species”, we mean species like plants, mammals, fishes, pollinating insects… Indeed, the long-term application of our project is to manage parasitic nematode infestations, without affecting crop plants or being toxic to ecosystem and human health. For example, a novel chemical developed against nematodes should not kill honeybees. Some species are neither “forbidden” nor “bonus” and are therefore considered as “neutral” (bacteria and viruses that are neither plant-parasitic nor plant-associated for example). After identification and annotation of candidate targets by bioinformatics methods, the most promising candidates will be selected for further functional analyses.

Here we report the semi-automated bioinformatics pipeline and the data management system developed for the identification of genes involved in plant-parasitism. To date, it has not been possible to develop such a comparative pipeline in the context of plant-parasitic nematodes because of the scarcity of genomics and transcriptomics data available for these species.

2 Material and Methods

The bioinformatics pipeline begins with two steps of screening, based on sequence similarity (Fig. 1). After the screening steps, the remaining proteins from *M. incognita* and *M. hapla* are analysed in terms of transcription evidence and automatic functional annotation. Lastly, all the data produced are stored into a relational database dedicated to this project.

Throughout the process, all the scripts necessary to parse the results or format the data files have been written with the Perl language with use of some BioPerl modules.

2.1 *In silico* Screenings of Potential Targets

The first screening step consists in a comparative analysis of the sets of predicted proteins from *M. incognita* and *M. hapla* with sets of predicted proteins from twenty-three other fully sequenced species. The dataset includes putative proteomes from 1 human-parasitic nematode, 2 plant-pathogenic fungi and 2 plant-eating insects. They are considered as “bonus species”. In addition, the dataset includes putative proteomes from 3 nematodes, 5 mammals, 1 bird, 1 amphibian, 2 fishes, 3 insects and 1 plant. They are neither parasitic, nor plant-associated, and are considered as “forbidden species”. To perform the comparative analysis, the OrthoMCL tool [4] was run with default parameters. The OrthoMCL procedure starts with all-against-all BLASTp comparisons of protein sequences from the submitted genomes. Putative orthologous relationships are identified between pairs of genomes by reciprocal best similarity pairs.

« Recent » paralogs (or in-paralogs) are identified as sequences within the same genome that are (reciprocally) more similar to each other than to any sequence from other species. Then, putative orthologous relationships are converted into a graph, to which the MCL (Markov Clustering) algorithm [5] is applied. The final output consists in clusters of putative orthologs and « recent » paralogs.

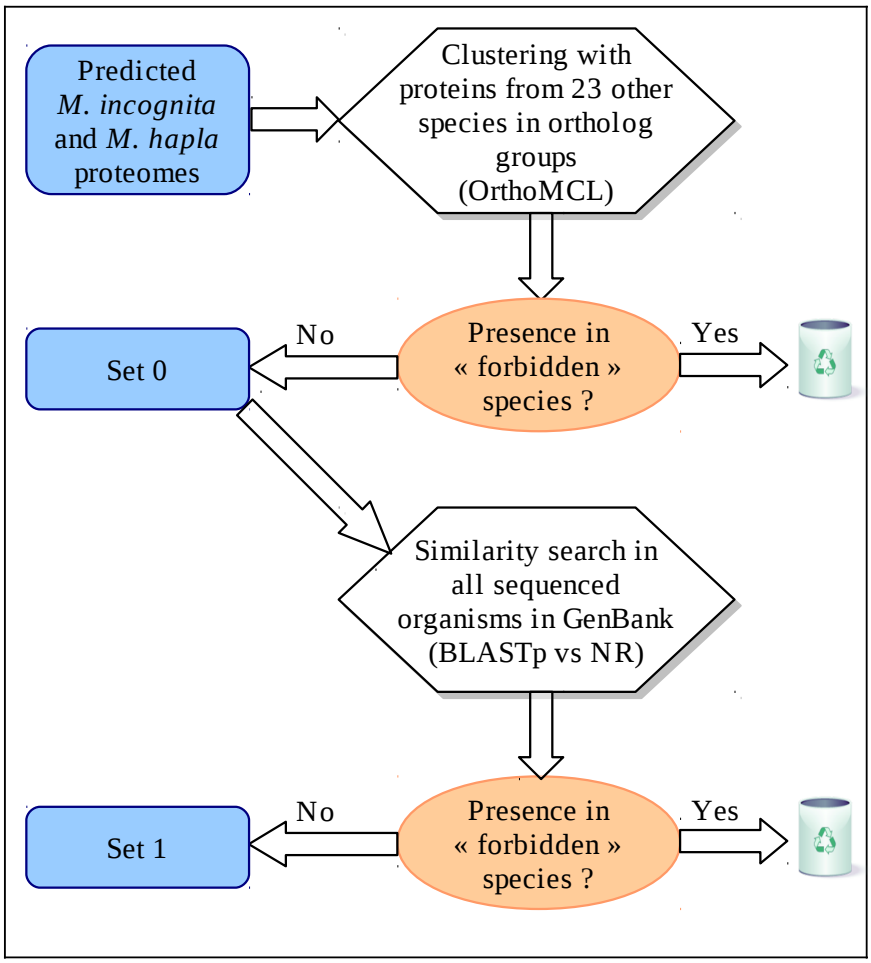

Figure 1. Schematic overview of the screening pipeline to identify potential targets in *Meloidogyne incognita* and *Meloidogyne hapla* whole sets of predicted proteins.

The results were parsed to exclude proteins conserved in forbidden species. The “remaining” proteins constitute the ***set 0***. We also assigned a “bonus tag” to proteins passing this filter that presented a potential ortholog in a known parasitic or plant-pathogenic species.

The second screening step consists in a BLAST search [6] of the ***set 0*** against GenBank (NR database, blastp, evalue max = 0.01, no filter for low complexity regions). This second step is perfectly complementary to the first one. Indeed, several proteomes included in the OrthoMCL run are absent from NR. In addition, most species in NR can not be included in the OrthoMCL run, because we do not have their complete putative proteomes (species not fully sequenced).

For each protein, all BLAST hits were analysed sequentially. We excluded proteins showing significant similarity (at least 40 % identity and 70 % of query length covered by the alignment) with one or more forbidden species. The remaining proteins presenting a significant similarity (at least 30 % identity and 50 % of query length covered by the alignment) with proteins from known plant-parasitic or plant-pathogenic species were assigned a “bonus tag”. (Fig. 2) The criteria are more stringent for exclusion than tag assignment to avoid considering hits in forbidden species that are not true orthologs, since only one hit in a forbidden species lead to the exclusion of the protein (whereas bonus tags are rather informative).

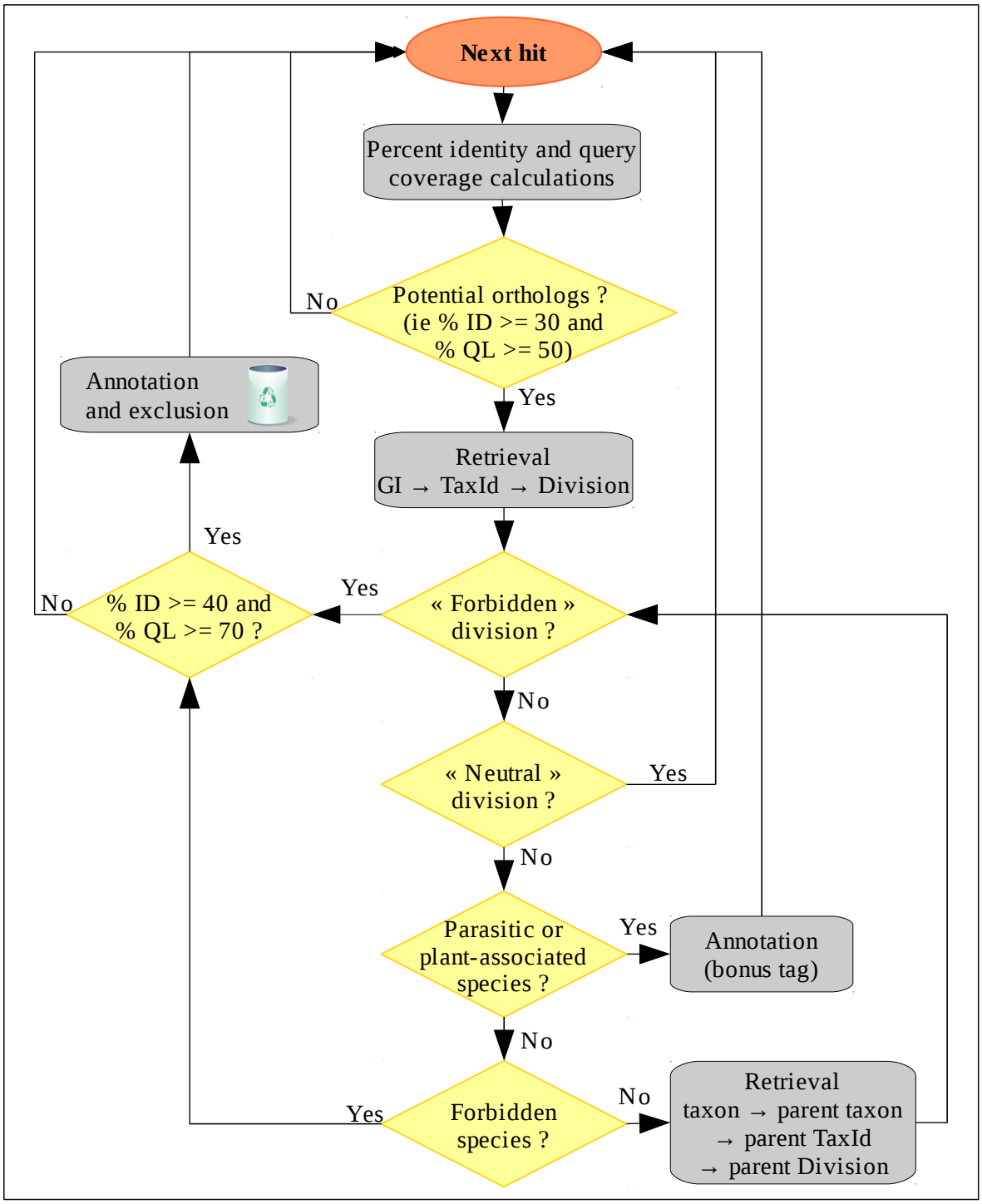

Figure 2. Schematic overview of the algorithm implemented in the BLAST parser. Remaining proteins after the first screening step based on an OrthoMCL run underwent a second screening step based on BLAST searches. All BLAST hits were sequentially analysed. When there is an hit in a species that is neither a "bonus" one, nor a "forbidden" one, we get the parent node in the taxonomy and test this parent node in the same way (until reaching the root if necessary), because the taxonomy identifiers (TaxId) we have listed sometimes correspond to clades (a higher level than a species). Moreover, to reduce computations, the parser first sorts the hits on the "division" criteria, as a division is assigned to each taxon node by the NCBI taxonomy. Forbidden divisions are Mammals, Primates, Rodents and Vertebrates. Neutral divisions are Phages, Synthetic, Unassigned and Environmental samples. In the other cases, the parser has to check if the TaxId is in the "bonus" list or in the "forbidden" list. On the figure, "% ID" means percent identity and "% QL" means percentage of query length covered by the alignment.

We are aware that this methodology would overlook genes involved in parasitism that are duplicates or mutated versions of existing genes shared with forbidden species. As our aim is to identify druggable targets, we can not take the risk to select genes that would be too similar to genes conserved in such species.

As there is no large-scale database that propose an inventory of plant-associated species, we collected information from the bibliography, from plant-pathologists and from two partial databases. The first one is the Comprehensive Phytopathogen Genomics Resource (http://cpgr.plantbiology.msu.edu/). It consists in a data warehouse of finished, draft and in progress genome and EST sequencing projects for viral, bacterial, oomycete, fungal, and nematode plant pathogens. The second one is the Pathogen Hosts Interactions base (http://www.phi-base.org/, [7]). It contains expertly curated information on experimentally verified pathogenicity, virulence and effector genes from fungal, oomycete and bacterial pathogens. Moreover, according to bibliography, we considered that four clades of nematodes are plant-associated: Tylenchida, Nordiidae, Longidoroidea and Trichodoroidea. In the end, we derived a list of 834 NCBI's taxonomy identifiers (TaxId) corresponding to species and clades known to be involved in parasitic or pathogenic interactions with plants. This is not an exhaustive list, but it represents more than 28000 species in total (as numerous species pertain to a clade) of nematodes, oomycetes, fungi, bacteria, trypanosomes, insects, virus and viroids.

To parse the BLAST results, we also needed to download the NCBI taxonomy and to list clades that we consider as “forbidden” (four clades: Chordata, Annelida, Mollusca, Viridiplantae).

The proteins kept at the end of the pipeline constitute ***set 1***. As computation requires huge memory, BLAST search and parsing have been computed on a cloud: the ProActive PACA Grid (http://proactive.inria.fr/pacagrid/).

2.2 Evidence of Expression at the Transcriptional Level

Evidence of the transcription of a gene coding for a putative protein supports the existence of this putatively expressed gene.

We already had accumulated data about EST evidence for the *M. incognita* set of proteins. They come from the NCBI dbEST database and from “in-house” *M. incognita*-specific EST clusters. The latter provides information about the stage(s) of the life cycle during which the gene is expressed. In our case, we are particularly interested in genes expressed during the free-living stage (as nematodes are more reachable by control means), but expression during plant-nematode interaction can provide insights into the mechanisms of parasitism.

We also downloaded datasets of protein predictions derived from clustered EST of seventeen plant-parasitic nematodes (including *M. incognita* and *M. hapla*) from the NEMBASE4 resource [8]. These collections are publicly available from http://nematodes.org/downloads/databases/NEMBASE4/index.shtml. We performed BLAST searches of our ***set 1***, using the polypeptides from NEMBASE4 as subject sequences (blastp, evalue max = 0.01, no filter for low complexity regions). Data were then parsed and criteria (identity and alignment percentages) were fit according to the phylogenetic distance between the subject species and the Meloidogyne phylum. Here, data provide not only expression evidence, but also information about conservation of the gene in plant-parasitic nematodes (as seen before, the more a gene is conserved in plant-parasitic species, the more it is likely to be involved in parasitism).

However, the amount of available transcriptomic data for plant-parasitic nematodes is relatively limited and most information is restricted to root-knot nematodes and to a lesser extent to cyst nematodes. This limits the possibility of comparing various different plant-parasitic nematodes that have adopted different strategies to feed on plant material. Hence, we have performed the RNA-seq transcriptome sequencing of four plant-parasitic nematode species presenting diverse parasitic strategies (*Pratylenchus coffeae*, *Ditylenchus dipsaci*, *Bursaphelencus xylophilus* and *Xiphinema index*). We have also generated the RNA-seq of different developmental stages of *M. incognita* in order to bring additional transcription support to the identified genes. Bioinformatics analyses are currently *in progress*.

2.3 Automatic Functional Annotation

As we are working on proteins predicted from the genomes, most of them have currently unknown functions. We have therefore undertaken bioinformatics annotations of these proteins.

- Functional regions (commonly termed *domains*) have been identified into our proteins, by using the PfamScan tool with the Pfam-A database (the part of Pfam containing high quality, manually curated families) and default parameters [9].
- "Standard" gene ontology (GO) terms have then been assigned to the proteins, based on the correspondence between the Pfam domains and the GO terms. "Slim" terms associated to the "standard" terms have also been subsequently assigned to the proteins. GO slims are cut-down versions of the GO ontologies containing a subset of the terms in the whole GO. They give a broad overview of the ontology content, without the detail of the specific fine grained terms. [10]
- The presence and location of signal peptide cleavage sites have been detected with the SignalP v3.0 tool [11], using both available methods (artificial neural networks and hidden Markov models).
- Prediction of transmembrane protein topology has been performed by searching for transmembrane helices in protein with the TMHMM v2.0 tool [12].
- Motifs specific to effectors of root-knot nematodes have been identified in the whole proteome of *M. incognita* using the MERCI software [13].

2.4 Database Development

A relational database has been developed using MySQL and phpMyAdmin in order to store all the data generated during the project. It allows to make data integration and analyses easier. Complex queries that combine results obtained at different steps of the pipeline (screening and/or annotation) can be launched. Outputs can be saved as simple tables or spreadsheets that can readily be used by the biologists.

3 Results and Discussion

The relational database contains all the data generated so far: results from the screening steps (which proteins have been "excluded" or "kept to go further" at each step, description of the hits) and the different types of annotations (as previously described: domains, gene ontology terms, signal peptides, transmembrane topology, specific motifs). Some more general information are included too, as the description of each step (who carried out this step, when, which tool and parameters have been used, what was the previous step...). This database allows us to compute a wide range of queries.

First of all, our ***set 1*** contains 16320 proteins. These root-knot nematode proteins are not present in species that could be negatively affected by the development of novel drugs and part of these are conserved in other plant-parasites. Among them:

– 5497 proteins (~ 34%) are shared with at least one other parasitic or plant-associated species,

– 3462 proteins (~ 21%) are supported by a transcription evidence from the same species and 4203 proteins (~ 26 %) are supported by a transcription evidence whatever the species (itself or another plant-parasitic nematode),

– less than a quarter of proteins have been associated with a functional annotation. Indeed, 3835 proteins (~ 24 % of ***set 1***) are annotated with one Pfam domain or more, and 2255 proteins (~ 14 %) have a GO term assigned.

A more detailed analysis of GO slim terms associated with the proteins shows that some terms seem to be under- or over-represented in the ***set 1*** compared to the whole putative proteomes of *M. incognita* and

M. hapla. In particular, we notice that the terms *nucleus* (GO:0005634 from the 'cellular component' ontology), *transcription factor activity* (GO:0003700 from the 'molecular function' ontology) and *transcription* (GO:0006350 from the 'biological process' ontology) seem to be over-represented ; suggesting that we may have identified specific transcription factors. By combining criteria on GO slim terms, transcription evidence and conservation in other species, we observed that twelve proteins are shared between several parasitic or plant-associated species, are annotated with the three GO slim terms related to transcription mentioned above and are supported by a transcriptional evidence. These proteins are of particular interest and are currently under experimental investigation.

Moreover, it is possible to identify putative effectors by combining the following criteria: conservation in other parasitic or plant-associated species, transcriptional evidence, no transmembrane helix, presence of a signal peptide and presence of one of the effector-specific motifs previously identified [13]. We obtain a list of 158 proteins. Among them, 18 are known to be carbohydrate-active enzymes (CAZymes: http://www.cazy.org/) involved in plant cell wall degradation, which is one of the known function of effectors [2]. They are able to degrade carbohydrates like cellulose, hemicellulose or pectin. Presence of known effectors within the reduced set of predicted effectors validates the screening method. Proteins of as yet unknown function constitute a set of interest to identify and characterize new effectors.

4 Conclusion and Perspectives

To date, the bioinformatics pipeline has generated a number of data which are all stored in a relational database. By combining several criteria, the database allows identification of sets of target genes restricted to parasitic or plant-associated species (such as putative transcription factors or effectors) for the design of durable new strategies to manage parasitic nematode infestations. As a control, we could align the identified genes back to the genomes of some neither parasitic nor plant-associated species to check that their successful outcome through filters is not due to annotation problems. But it would be quite surprising that these genes would have been missed in all the proteomes of the 18 forbidden species included in the OrthoMCL run.

In near future, it is planned to implement a graphical user interface, probably by using BioMart [14], as it is described as a simple and robust data integration system for large scale data querying. This interface would allow users to query the database more easily (without the need of writing SQL instructions). It is also planned to include more data, according to the users needs (such as bibliography or comments).

Furthermore, to partially overcome the scarcity of omics data available for plant-parasitic nematodes, we have performed the RNA-seq transcriptome sequencing of four plant-parasitic nematode species presenting diverse parasitic strategies as well as the RNA-seq of different developmental stages of *M. incognita*. Bioinformatics analyses are currently *in progress* and will soon provide additional information about the transcriptional support of the identified genes and their conservation across the four plant-parasitic nematodes sequenced.

In the end, the most promising candidates will be selected for further functional analyses in the plant-parasitic nematode model *Meloidogyne incognita*. This will include expression analysis, tissue localization of gene expression and gene inactivation by RNA interference assays.

Acknowledgements

This work is supported by the French National Research Agency (“Nematargets” project).

Celine Vens is a Postdoctoral Fellow of the Research Foundation - Flanders (FWO).

The authors are grateful to Franck Panabières and René Feyereisen for their explanation about parasitic and plant-associated oomycetes and insects (respectively).

References

[1] H. van Megen, S. van den Elsen, M. Holterman, G. Karssen, P. Mooyman, T. Bongers, O. Holovachov, J. Bakker and J. Helder, A phylogenetic tree of nematodes based on about 1200 full-length small subunit ribosomal DNA sequences. *Nematology*, 11: 927-950, 2009.

[2] P. Abad, J. Gouzy, J. Aury, P. Castagnone-Sereno, E.G.J. Danchin, E. Deleury, L. Perfus-Barbeoch, V. Anthouard, F. Artiguenave, V.C. Blok, M. Caillaud, P.M. Coutinho, C. Dasilva, F. De Luca, F. Deau, M. Esquibet, T. Flutre, J.V. Goldstone, N. Hamamouch, T. Hewezi, O. Jaillon, C. Jubin, P. Leonetti, M. Magliano, T.R. Maier, G.V. Markov, P. McVeigh, G. Pesole, J. Poulain, M. Robinson-Rechavi, E. Sallet, B. Ségurens, D. Steinbach, T. Tytgat, E. Ugarte, C. van Ghelder, P. Veronico, T.J. Baum, M. Blaxter, T. Bleve-Zacheo, E.L. Davis, J.J. Ewbank, B. Favery, E. Grenier, B. Henrissat, J.T. Jones, V. Laudet, A.G. Maule, H. Quesneville, M. Rosso, T. Schiex, G. Smant, J. Weissenbach and P. Wincker, Genome sequence of the metazoan plant-parasitic nematode Meloidogyne incognita. *Nat. Biotechnol.*, 26: 909-915, 2008.

[3] C.H. Opperman, D.M. Bird, V.M. Williamson, D.S. Rokhsar, M. Burke, J. Cohn, J. Cromer, S. Diener, J. Gajan, S. Graham, T.D. Houfek, Q. Liu, T. Mitros, J. Schaff, R. Schaffer, E. Scholl, B.R. Sosinski, V.P. Thomas and E. Windham, Sequence and genetic map of Meloidogyne hapla: A compact nematode genome for plant parasitism. *Proc. Natl. Acad. Sci. U.S.A.*, 105: 14802-14807, 2008.

[4] L. Li, C.J.J. Stoeckert and D.S. Roos, OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.*, 13: 2178-2189, 2003.

[5] A.J. Enright, S. Van Dongen and C.A. Ouzounis, An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, 30: 1575-1584, 2002.

[6] S.F. Altschul, T.L. Madden, A.A. Schäffer, J. Zhang, Z. Zhang, W. Miller and D.J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, 25: 3389-3402, 1997.

[7] R. Winnenburg, M. Urban, A. Beacham, T.K. Baldwin, S. Holland, M. Lindeberg, H. Hansen, C. Rawlings, K.E. Hammond-Kosack and J. Köhler, PHI-base update: additions to the pathogen host interaction database. *Nucleic Acids Res.*, 36: D572-6, 2008.

[8] J. Parkinson, C. Whitton, R. Schmid, M. Thomson and M. Blaxter, NEMBASE: a resource for parasitic nematode ESTs. *Nucleic Acids Res.*, 32: D427-30, 2004.

[9] R.D. Finn, J. Mistry, J. Tate, P. Coggill, A. Heger, J.E. Pollington, O.L. Gavin, P. Gunasekaran, G. Ceric, K. Forslund, L. Holm, E.L.L. Sonnhammer, S.R. Eddy and A. Bateman, The Pfam protein families database. *Nucleic Acids Res.*, 38: D211-22, 2010.

[10] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin and G. Sherlock, Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, 25: 25-29, 2000.

[11] O. Emanuelsson, S. Brunak, G. von Heijne and H. Nielsen, Locating proteins in the cell using TargetP, SignalP and related tools. *Nat Protoc*, 2: 953-971, 2007.

[12] 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.*, 305: 567-580, 2001.

[13] C. Vens, M. Rosso and E.G.J. Danchin, Identifying discriminative classification-based motifs in biological sequences. *Bioinformatics*, 27: 1231-1238, 2011.

[14] S. Haider, B. Ballester, D. Smedley, J. Zhang, P. Rice and A. Kasprzyk, BioMart Central Portal--unified access to biological data. *Nucleic Acids Res.*, 37: W23-7, 2009.