NRPS toolbox for the discovery of new nonribosomal peptides and synthetases

Maude Pupin, Malika Smaïl-Tabbone, Philippe Jacques, Marie-Dominique Devignes, Valérie Leclère

NRPS toolbox for the discovery of new nonribosomal peptides and synthetases

Maude Pupin[1], Malika Smaïl-Tabbone[3], Philippe Jacques[2], Marie-Dominique Devignes[3] and Valérie Leclère[2]

[1] LIFL, UMR8020 CNRS, INRIA, Bat M3, Univ Lille Nord de France, Sciences et Technologies, 59655 Villeneuve d'Ascq cedex, France

maude.pupin@lifl.fr

[2] ProBioGEM, UPRES EA 1026, Polytech'Lille/IUT A, Av P Langevin, Univ Lille Nord de France, Sciences et Technologies, 59655 Villeneuve d'Ascq cedex,

valerie.leclere@univ-lille1.fr, philippe.jacques@polytech-lille.fr

3 LORIA (CNRS UMR7503, INRIA Nancy Grand-Est, Nancy Université), Campus scientifique, 54506 Vandoeuvre-lès-Nancy, France

{Marie-Dominique.Devignes, malika.smail}@loria.fr

Abstract *Nonribosomal peptide synthetases are huge multi-enzymatic complexes synthesizing peptides, but not through the classical process of transcription and then translation. The synthetases are organised in modules, each one integrating an amino acid in the final peptide. The modules are divided in domains providing specialized activities. So, those enzymes are as diverse as their products. We present our toolbox designed to annotate them accurately and promising results obtained on some* Burkholderia*,* Bacillus *and* Pseudomonas *genomes.*

Keywords database, protein annotation, nonribosomal peptides, nonribosomal synthetase.

1 Introduction

Micro-organisms are able to synthesize peptides by a pathway alternative to the central dogma, the nonribosomal peptide synthesis. Multifunctional enzymes, called NonRibosomal Peptide Synthetases (NRPSs), assemble directly monomers to produce atypical peptides harboring original physico-chemical properties that give them various properties such as antibiotic, anti-tumor, immunosuppressive or surfactant (surface-active substances such as detergents). Several nonribosomal peptides are already exploited in pharmacology or other biotechnological area, but their great potential of new drugs or bio-active compounds is underexploited.

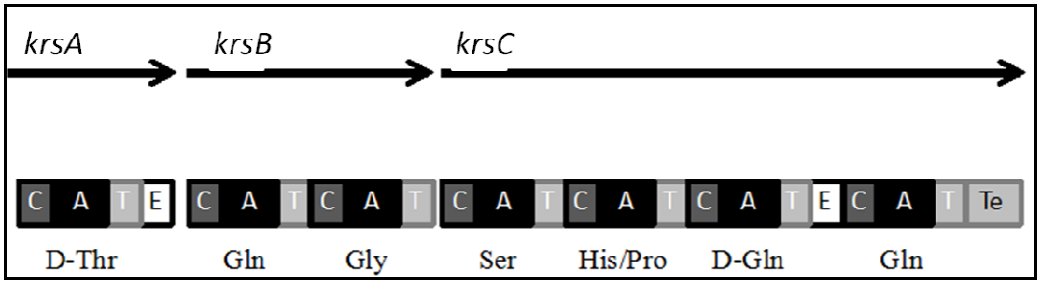

Figure 1. Scheme of a synthetase composed of 3 proteins (KrsA, KrsB and KrsC) and 7 modules. Each colored box represents a domain (C for condensation domain, A for adenylation domain, T for thiolation, E for epimerization and Te for thioesterase). The amino acid incorporated by each module is mentioned under it.

NonRibosomal Peptide Synthetases are organized in sets of catalytic domains which constitute modules containing the information needed to complete an elongation step in an original peptide biosynthesis (see Figure 1). The main catalytic functions are responsible for the activation of an amino acid residue (adenylation -A- domain), the transfer of the corresponding adenylate to the enzyme-bound 4'-phosphopantetheinyl cofactor (thiolation -T- domain) and the peptide bond formation (condensation -C- domain). The active site of the adenylation domain is specific of the incorporated amino acid. As non proteogenic amino acids or other compounds can be incorporated, we also use the term monomer. Additional

domains can lead to modification of substrates if required for peptide final structure. For example, epimerization -E- domains, tranform L-amino acids in D-amino acids. A thioesterase -Te- domain is usually present in final position to ensure the cleavage of the thioester bond between the nascent peptide and the last T domain and, in several cases, to cyclize the peptide. To summarize, a given synthetase produces a specific peptide, with as many modules as amino acids incorporated in the final peptide. The synthetase illustrated in Figure 1 is composed of 7 modules, each incorporating the mentioned amino acid. The modules with epimerization domains transform the L-amino acid in D-amino acid.

Several bioinformatics tools [1, 2, 3, 4] were developed during the past decade to predict, from the protein sequence, the modular organization of the NRPS and the potential monomer composition of the synthesized peptidic product. The two first tools predict the modular organization of the synthetases and all the four predict the amino acids incorporated by the A-domains. As bioinformatics tools today available help predicting the genes, the produced proteins and their functions from genomes data, we can now expect to be able to predict the produced peptides and their potential activity from the NRPS protein sequence. However, one step remains difficult, which consists in the detection of the putative synthetases among the proteome of a given micro-organism. The difficulty comes from the fact that each synthetase is specific of one peptide so we cannot use a classical BLAST search to find all of them.

Our aim is to provide bio-informatics tools that help discovering new nonribosomal peptides by predicting and analyzing their synthetases from data obtained by genome sequencing or metagenomics.

2 Methods : our NRPS toolbox

Our work began with the creation of Norine [5] (http://bioinfo.lifl.fr/norine/), the unique resource dedicated to nonribosomal peptides. It provides a database with detailed annotations, including but not limited to biological activity, producing organism and the monomer structure of the peptides that deals with their non-linear 2D-structure. It provides also bio-informatics environment to analyze NRPs such as visualization or edition tools for monomer structures, statistics representations of the results, peptide search by monomer composition, structural pattern (with a list of or undetermined monomers at several positions) [6] or structure comparison.

We are now developing a complete toolbox dedicated to NRPSs and their products by integrating existing and ongoing tools. Doris, Database of nOnRIbosomal Synthetases (not yet public) contains not only synthetases automatically extracted from generalist protein databases such as UniProt, but also manually curated ones, annotated with the tools dedicated to NRPSs. The synthetases are connected to their product in the Norine database, when the structure of the peptide is experimentally verified. To do so, we search the monomer composition or structural pattern predicted from the synthetase, in Norine database. The results can be a perfect match with a given NRP, suggesting that we have found the synthetase of this peptide, a nearly perfect match, suggesting that the produce peptide is a variant of the one stored in Norine, or a partial match suggesting either the synthetase we have found is incomplete or we have found a new peptide.

To complete the DORIS database, candidate NRPS are also extracted from newly sequenced genomes on the basis of sequence similarity with already described NRPS combined with manual analysis of coding sequence description as provided by automatic genome annotation. For example, we search expressions such as “NRPS”, “adenylation”, “nonribosomal” or “siderophore” among the descriptions of the studied proteome. To this aim, the MODIM (MOdel Driven Data Integration for Mining) methodology is used as it facilitates data collection and integration [7, 8]. It requires a relational database model and allows the specification of workflows for collecting data from various resources. The collected data subsequently populate the target database. Specific views can then be defined on the database for extracting datasets to be mined.

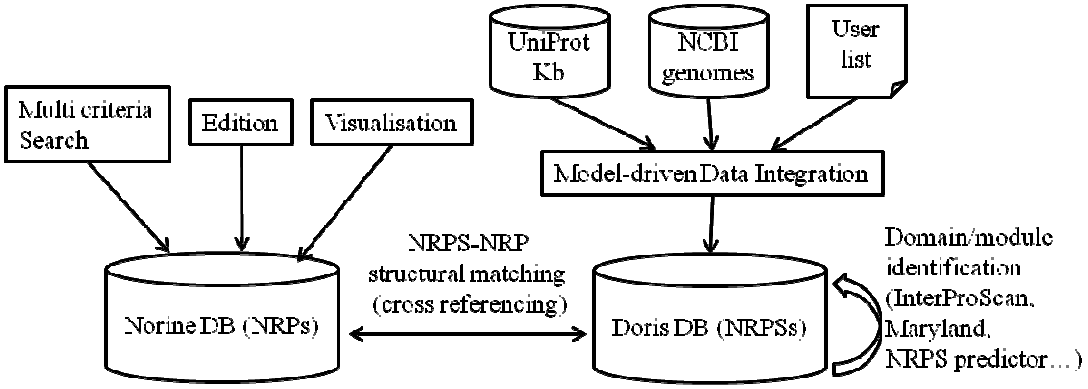

Figure 2. Scheme of the NRPS toolbox, summarizing the interaction between the tools and databases.

The association of all these tools will constitute a unique complete toolbox dedicated to NRPSs (Figure 2) and their products and will be very useful for discovering new natural antimicrobial or anti-tumoral peptides. The tools are validated with novel bacilli genomes which will be used for illustration purposes.

3 Results

The NRPS toolbox was used to implement a strategy for discovering NRPS encoded in newly sequenced genomes. We performed systematic BLAST analyses of all protein coding sequences (CDS) of a genome of interest against the database constituted by reference NRPSs stored in Doris. The best significant hits were selected. Filtering conditions were tuned manually thanks to the KoriBlast software facility. Best hits were found with a length greater than 500 amino acids, displaying more than 5 HSPs, with the best HSP at a percentage of identity and similarity greater than 28% and 45% respectively, an E-value close to zero and a gap percentage below 10 %. The MODIM system then retrieved and integrated the annotations and positions of these CDS on the genome as well as their domain composition.

3.1 Discovery of new nonribosomal peptide synthetases

To validate our strategy, a first experiment was performed on four bacterial genomes (three *Burkholderia* chromosomes and one *Bacillus cereus* chromosome) and produced 86 BLAST hits. Domain analysis by the specialized NRPS tools provided us with NRPS domains. At this stage of the work we therefore decided to match the primary NRPS A, T, C, and Te domains with InterPro domains (see Table 1). The InterProScan localization tool was then used to retrieve start and end positions of each domain on the protein sequence. Occurrences of A, T and C domains in the same or in contiguous CDS were found for 9 protein hits delineating 15 complete NRPS elongation modules and one termination module. An example of data view obtained for *Burkholderia ambifaria* chromosome 1 is presented in Table 2.

NRPS primary domain	InterPro id	source database id
adenylation domain	IPR01007	TIGR01733
thiolation domain	IPR006163	PF00550
condensation domain	IPR001242	PF00668
thioesterase domain	IPR001031	PF00975

Table 1. Relationship between A, T, C and Te domains and InterPro domains.

Domain	Module nb	InterPro id	Hit : UniProt_ID	Domain start (aa)	Domain end (aa)	Genome Id	Locus Id
A	1	IPR010071	B1YQA7	39	449	NC_010551	BamMC406_1558
T	1	IPR006163	B1YQA7	536	600	NC_010551	BamMC406_1558
C	2	IPR001242	B1YQA7	632	929	NC_010551	BamMC406_1558
A	2	IPR010071	B1YQA7	1111	1513	NC_010551	BamMC406_1558
T	2	IPR006163	B1YQA7	1603	1663	NC_010551	BamMC406_1558
C	3	IPR001242	B1YQA7	1687	1983	NC_010551	BamMC406_1558

iprE	3	IPR010060	B1YQA7	1989	2138	NC_010551	BamMC406_1558
C	3	IPR001242	B1YQA7	2158	2451	NC_010551	BamMC406_1558
A	3	IPR010071	B1YQA7	2643	3058	NC_010551	BamMC406_1558
T	3	IPR006163	B1YQA7	3139	3201	NC_010551	BamMC406_1558
C	4	IPR001242	B1YQA8	50	350	NC_010551	BamMC406_1559
A	4	IPR010071	B1YQA8	540	946	NC_010551	BamMC406_1559
T	4	IPR006163	B1YQA8	1036	1097	NC_010551	BamMC406_1559
C	5	IPR001242	B1YQA8	1124	1425	NC_010551	BamMC406_1559
T	5	IPR006163	B1YQA8	1593	1655	NC_010551	BamMC406_1559

Table 2. View on Doris data summarizing the module signatures obtained for *Burkholderia ambifaria* chromosome 1. iprE : epimerization domain (see below)

We introduce here the concept of "module signature" which is a set of ordered protein domains always encountered in modules sharing similar function. For example the NRPS elongation module signature is composed of the three C, A, T domains (< C, A, T > signature), whereas the NRPS termination module signature is composed of C, A, T and Te domains (< C, A, T, Te > signature). When secondary domains are detected in modules associated to some specific function, enriched module signatures can be proposed (see below for modules containing an epimerisation domain).

The prediction of monomers was carried out with specialized tools [2, 3, 4] for each A domain of our 16 NRPS modules. The prediction remains impossible for 4 A domains pointing out the limits of the available tools.

3.2 Characterization of new signatures for optional domains

A unique advantage of the NRPS toolbox is the possibility, when this information is available, of matching the structure of a NRPS with the structure of the peptide it produces. This led us to investigate the domain structure of some NRPS responsible for the synthesis of peptides containing D-monomers. We thus identified two groups of such NRPS. In the first group, containing for example bacitracine and gramicidine synthetases from firmicutes, classical NRPS tools are able to detect a so-called E (epimerization) domain in modules responsible for the condensation of D-monomers. In fact, E domains are always followed by C domains in these modules. Moreover, InterProScan analysis of E domains reveals that these epimerisation domains are composed of an IPR001242 (C) domain followed by an extension of about 130 amino acids (iprE for InterPro Epimerisation domain) recognized as the IPR010060 Interpro domain. The enriched signature < C, iprE, C, A, T > can thus be defined for such modules.

The second group of NRPS (for example massetolide and arthrofactin synthetases from *Pseudomonas*) includes NRPS modules corresponding to the condensation of D-monomers but lacking iprE domains. In such modules only regular A, T and C domains are observed. No other InterPro domain is detected by InterProScan. We therefore carefully analyzed inter-domain regions searching for a yet undescribed epimerisation domain. We observed that a region of constant length of 186 amino acids is always present downstream the C domain in all modules of these NRPS. Multiple sequence alignment of 137 instances of this region (downC-186) lead to distinguishing two clusters of highly conserved sequences (downC-186, and downC-186-E). Interestingly sequences of downC-186-E cluster are always found in modules responsible for the condensation of D-monomers but not in any other module of this group of NRPS. We thus propose < C, downC-186-E, A, T > as an enriched module signature for a new type of NRPS modules responsible for the condensation of D-monomers.

4 Conclusion and perspectives

In conclusion we have shown here that our NPRS toolbox is a unique and useful resource for characterizing NRPS and further exploring the relationships between structure of NRPS modules and the type of incorporated monomers. In future work we will apply machine learning methods to refine signature description for modules associated with a given monomer. This will lead to improve peptide prediction and to better understand the function of NRPS for which no peptide is yet described. Ultimately, the NRPS

toolbox will become a precious resource for designing recombinant NRPS to produce synthetic active compounds such as novel antibiotics.

Acknowledgements

This work was supported by PPF Bioinformatique of Lille 1 University and FEDER (INTERREG IV PHYTOBIO project). We wish to thank INRIA and the CPER-Region Lorraine for their financial support , Birama Ndiaye for his help with the MODIM system. We acknowledge the contribution of Constant Denis, Nicola Gref, Jean-Philippe Monnerville and Hélène Polvèche during their student internships.

References

[1] M. Z. Ansari, G. Yadav, R.S. Gokhale and D. Mohanty, NRPS-PKS: a knowledge-based resource for analysis of NRPS/PKS megasynthetase. *Nucl. Acids Res.*, 32:405-413, 2004. (http://www.nii.res.in)

[2] B. O. Bachmann and J. Ravel, In Silico Prediction of Microbial Secondary Metabolic Pathways from DNA Sequence Data. *Meth. Enzymol.,* 458:181-217, 2009. (http://nrps.igs.umaryland.edu)

[3] C. Rausch, T. Weber, O. Kohlbacher, W. Wohlleben and D. H. Huson, Specificity prediction of adenylation domains in nonribosomal peptide synthetases (NRPS) using Transductive Support Vector Machines (TSVM). *Nucl. Acids Res*., 33:5799-5808, 2005.

[4] M. Röttig, MH. Medema, K. Blin, T. Weber, C. Rausch, and O. Kohlbacher. NRPSpredictor2 - a web server for predicting NRPS adenylation domain specificity. Nucl. Acids Res., 39:W362–W367, 2011. (http://nrps.informatik.uni-tuebingen.de/)

[5] S. Caboche, M. Pupin, V. Leclère, A. Fontaine, P. Jacques and G. Kucherov, NORINE: a database of nonribosomal peptides. *Nucl. Acids Res.,* 36:D326-D331, 2008. (http://bioinfo.lifl.fr/norine/)

[6] S. Caboche, M. Pupin, V. Leclère, P. Jacques and G. Kucherov, Structural pattern matching of nonribosomal peptides. *BMC Structural Biology,* 9:15, 2009.

[7] B. Ndiaye, E. Bresso, M. Smail-Tabbone, M. Souchet and MD. Devignes, MODIM : Model Driven Data Integration for Mining. JOBIM, 2011 (poster).

[8] S. Yilmaz, P. Jonveaux, C. Bicep, L. Pierron, M. Smail-Tabbone and MD. Devignes Gene-disease relationship discovery based on model-driven data integration and database view definition, *Bioinformatics*, 25:230-236, 2002.