Data updates on Norine, the reference Non-Ribosomal Peptide knowledge base

Yoann DUFRESNE[1], FirstName LASTNAME2[2] and FirstName LASTNAME3[2]
[1] Équipe Bonsai, CRIStAL, université de Lille, INRIA Lille Nord Europe, Batiment M3, avenue Carl Gauss, 59655, Villeneuve d'Ascq, France
[2] Laboratory, Address, zip code, Town, Country

Corresponding author: yoann.dufresne0@gmail.com

Abstract *The abstract of the paper (optional for short contributions) must be typeset in italic, with* Times New Roman *11-point font. The left and right margins must be set to 3cm. 350 words maximum.*

Keywords Norine database, Non-Ribosomal Peptides, Update, Data curation

1 Introduction

Norine, first released in 2006[TODO], remains the unique platform dedicated to computational biology analysis of non-ribosomal peptides (NRPs). The NRPs have increased in popularity in recent years because they harbour diverse interesting biological activities. Indeed, they are produced by micro-organisms, bacteria and fungi, to colonise and survive in various environments. Among others, NRPs can act as antibiotics (penicillin -NOR00006-, daptomycin -NOR00001- or vancomycin -NOR00681-), siderophores (pyoverdins -NOR00160 to 206, NOR00903 to 912- or vibriobactin -NOR00250-), surfactants or protease inhibitors. In addition to their primary activity, some NRPs are also successfully prescribed for treating cancers (actinomycin D -NOR00228-) or reducing transplant rejection (cyclosporin A -NOR00033-). Beyond the pharmacology, NRPs promise other advantageous applications such as biocontrol of plant diseases, bioremediation of areas contaminated with toxic metals and/or non-biodegradable organic compounds. These metabolites are produced by a specific biosynthetic pathway. In few words, huge enzymes called NRP synthetases select specific amino acids, variant amino acids, lipids (and many other) and assemble them.

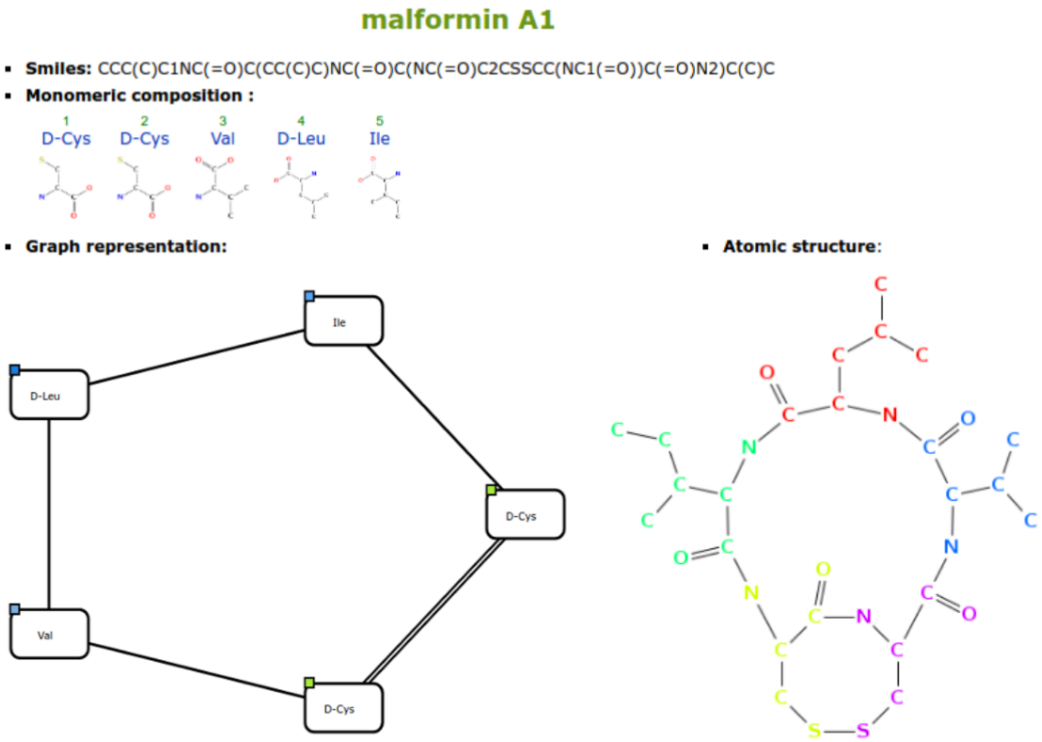

Fig. 1. Structure representations of the malformin A1 in the Norine database.

The Norine database is the reference NRP knowledge base, containing more than 1600 peptides composed of almost 600 different monomers (different building blocks including amino acids). In the database, each NRP

referenced have a dedicated web page with a lot of informations about their provenance, their composition and their pharmaceutical properties. The most important information is their monomeric structure/composition (see figure 1, on the left). The monomeric representation, that we also called the biological structure, correspond to the nearest representation of the NRP assembly process. In this representation, each node correspond to a molecule that had been included during the synthesis. The other representation (figure 1, on the right), is the atomic representation, obtained by reconstruction after a mass spectrum analysis. The knowledge of the monomeric representation is the most important information about a peptide because it is needed to fully understand the synthesis pathway. It also as been proved[TODO] that, in the majority of the cases, the activity of the molecule can be deduced from this only one information.

Since 2016, the Norine database is open to the crowdsourcing[TODO]. External users can submit new peptides to improve the data quantity of the database. A complete procedure of submission and reviewing had been set up to guaranty the quality of these data. Nevertheless, we know that many NRP discovered are not present in the Norine database and this process is not allowing a massive addition of data. This is also not allowing the correction of wrong data that add been entered before the set up.

In the next parts of this article we will have a quick overview of the current work on data curation and database filling.

2 Improving the data quality

In the Bonsai group, we developed a tool called smiles2monomers (s2m) that automatically create annotations of NRP. From a SMILES[TODO] (a textual atomic representation of a molecule), s2m infer the monomeric structure of the NRP. On one side, as we said during the introduction, the most useful information is the monomeric structure of NRP. On the other side, almost every NRP are characterised by mass spectrum experiments, so we often only know their atomic structure. So, s2m is a very powerful tool for the NRP community.

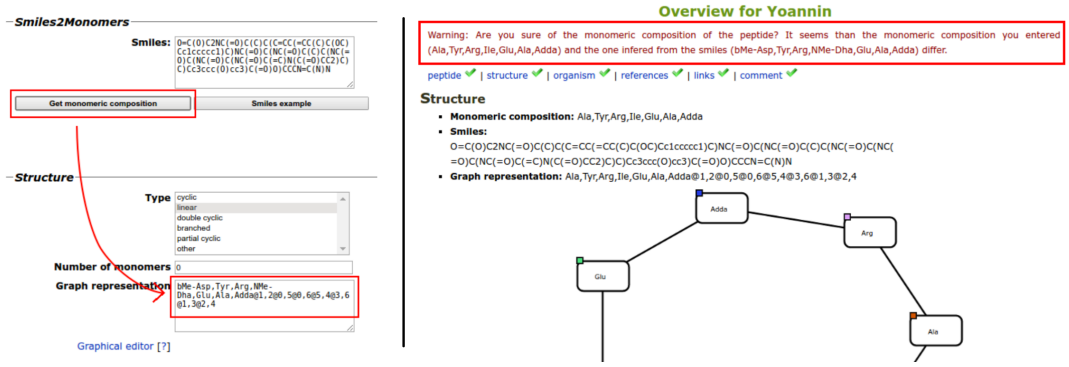

Fig. 2. Quality controls in the MyNorine software.

In the Norine database, a significative amount of NRP entries (around 30%) are filled with both of the atomic and monomeric structures. We used s2m on the atomic structures to verify the integrity of the data and we found a few errors (50 with a wrong atomic or monomeric structure). To avoid the insertion of new errors, we included the s2m software in the crowdsourcing tool MyNorine. When a user want to add a new compound s2m is used during two validation steps. Firstly, when the user fill the SMILES area, myNorine can automatically create the monomeric structure (see figure 2, on the left). Secondly, if the user did not explicitly generate the monomeric structure, s2m run in background to compare the result with the manually entered structure. If the automatic and manual annotations are not equivalent, the MyNorine tool will raise a warning to the user (see figure 2, on the right)

3 Improving the data quantity

Norine was created in 2009 and updated until 2016 by a small group of people. Many NRP published were added to the database but we know that a lot of other molecules, for many diferents reasons, had never been published, even if they are fully characterised. Aiming the goal of adding all theses unknown NRP to Norine,

we opened the database to crowdsourcing. Since this opening, a multitude of NRP have been added but still know that many NRP are present in other partially specialised and unspecialised databases of molecules.

We identified 3 main databases that can be sources of new NRP for Norine. The first one is the Minimum Information about a Biosynthetic Gene cluster (MIBiG) database[TODO]. This database store informations about gene clusters of secondary metabolites (whose NRP are). The second one is the Biologically Interesting Molecule Reference Dictionary (BIRD)[TODO]. This database references external resources about "interesting" molecules and we know that some of them are NRP. The last database used as ressource is StreptomeDB[TODO], the database that reference the molecules produced by bacterias in the genus of Streptomyces. These bacterias are known as producers of NRP and that's why we explored this database. Until now, the Norine database is well known for the quality of the manual annotations included. So, we did not want to add wrong informations from an automatic filling of the database and that's why we created a strict validation pipline for the potential new entries. After the filtering processus that I will describe in the next paragraph, we found 472 unreferenced NRP in Norine: 235 from MIBiG, 162 from BIRD and 75 from StreptomeDB.

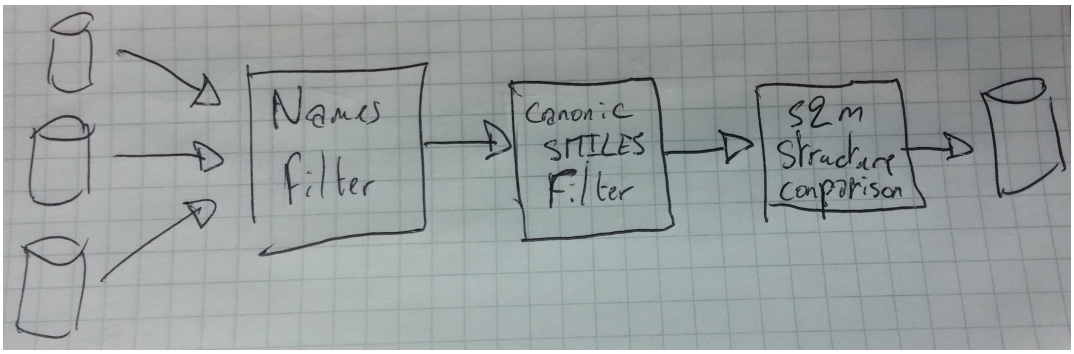

Fig. 3. Automatic filing quality control.

To automatically include new entries, we created a pipline of filters based on name, SMILES and structure comparisons (see figure 3). Before all the filtering tasks, we only extract molecules that are already annotated as NRP in the databases. First, we filter by names and known synonymes to avoid duplicates entries (fast and easy task). Then, we try to avoid duplicates using a string comparison of SMILES. For this task, we use the CDK library[TODO] to get a canonical SMILES from each candidate entry (multiple SMILES can represent a single molecule; a canonical SMILES is always the same for one molecule). Those canonical SMILES are used to compare the candidate entry to all the SMILES already present in the database (the canonization is time consuming but the string comparison is fast). Finally, for the molecules not filtered, we compute and compare the monomeric graph with monomeric graphs present in Norine. For this task, we use s2m to generate the monomeric graphs and a simple algorithm of graph isomorphism to compare the graph against the database (the most time consuming task).

After the execution of these scripts, the database was filled with 472 new peptide annotations. Those data represent an increase of 30% of the entries in the Norine database for a new total of 1658 NRP annotated. For the data that where already present in Norine, we are currently looking at the similarities and differences between our annotation and the other databases ones to select the best combination of the two.

4 Conclusion and perspectives

On this article we presented an update of the data from the knowledge base Norine. Using tools like smiles2monomers, we detected a few errors in the annotations. We correct them and created safeguards to avoid errors in futur user submissions. In a second time, we used multiple tools to retrieve and filters many possible new NRP entries in the database. The work on automatic filing scripts led us to a data increase of 30% for a new amount of 1658 annotations in Norine. So, in the coming release of Norine we strongly improve the data quantity and quality available for all.

We also noticed that many data are still not present in the database. For example, we manually noticed that some entries in the PubChem database are NRP undercover (without the NRP annotation). We that soon

we will be able to screen the entire PubChem database and automatically relate them with proof of their NRP origins.