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

Yoann DUFRESNE[1], Juraj MICHALIK[1], Areski FLISSI[1], Valerie LECLÈRE[1,2] and Maude PUPIN[1]
[1] Équipe Bonsai, Univ. Lille, CNRS, Centrale Lille, UMR 9189 - CRIStAL - Centre de Recherche en Informatique Signal et Automatique de Lille, F-59000 Lille, France
[2] Équipe ProBioGEM, Univ. Lille, INRA, ISA, Univ. Artois, Univ. Littoral Côte d'Opale, EA 7394 - ICV - Institut Charles Viollette, F-59000 Lille, France

Corresponding author: yoann.dufresne0@gmail.com

1 The Norine database

Norine, first released in 2006 [1], remains the unique platform dedicated to computational analysis of non-ribosomal peptides (NRPs). Among others, NRPs can act as antibiotics, siderophores, surfactants or protease inhibitors. The Norine database is the reference NRP knowledge base, containing more than 1200 peptides composed of almost 530 different monomers (various building blocks including amino acids). Each referenced NRP have a dedicated web page with various informations, including the most importants, their composition and their biological activities. The monomeric representation, correspond to the nearest representation of the NRP assembly process. The other representation is the atomic representation, obtained by reconstruction after a mass spectrum analysis. The knowledge of the monomeric representation allow to understand the synthesis pathway. It has also been proved [2] that, the activity of the molecule can be predicted from this representation.

2 Improving the data quality and quantity

We developed a tool called smiles2monomers (s2m) [3] that automatically creates NRP annotations. From a SMILES [4], s2m infers the monomeric structure of the NRP. In Norine, a significant amount of NRP entries (around 30%) are annotated with both structures. We used s2m on the atomic structures to verify the integrity of the data and we found a few errors (50 NRPs with a wrong atomic or monomeric structure). To avoid the insertion of new errors, we included the s2m software in the crowdsourcing tool MyNorine.

We identified 3 main databases that could be sources of new NRPs for Norine: MIBiG [5] (store gene clusters of secondary metabolites), BIRD [6] (Centralisation of external resources about "interesting" molecules), StreptomeDB [7] (molecules produced by bacteria in the *Streptomyces* genus). The Norine database is well known for the quality of its manual annotations. So, we did not want to add wrong informations from an automatic filling of the database. For this reason, we created a strict validation pipeline for the potential new entries. After the filtering process, we found 472 NRPs unreferenced in Norine: 235 from MIBiG, 162 from BIRD and 75 from StreptomeDB. Those data represent an increase of 30% of the entries in the Norine database.

3 Conclusion

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

References

[1] Ségolène Caboche, Maude Pupin, Valérie Leclère, Arnaud Fontaine, Philippe Jacques, and Gregory Kucherov. NORINE: a database of nonribosomal peptides. 36:D326–D331.
[2] Ammar Abdo, Ségolène Caboche, Valérie Leclère, Philippe Jacques, and Maude Pupin. A new fingerprint to predict nonribosomal peptides activity. 26(10):1187–1194.
[3] Yoann Dufresne, Laurent Noé, Valérie Leclère, and Maude Pupin. Smiles2monomers: a link between chemical and biological structures for polymers. 7:62.
[4] David Weininger. SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. 28(1):31–36.
[5] MIBiG: Minimum information about a biosynthetic gene cluster.
[6] Helen Berman, Kim Henrick, and Haruki Nakamura. Announcing the worldwide protein data bank. 10(12):980–980.
[7] Xavier Lucas, Christian Senger, Anika Erxleben, Björn A. Grüning, Kersten Döring, Johannes Mosch, Stephan Flemming, and Stefan Günther. StreptomeDB: a resource for natural compounds isolated from streptomyces species. 41:D1130–D1136.