Scientific Symposium FAIR Data Sciences for Green Life Sciences
12 December 2018 | Wageningen | the Netherlands

Conference Proceedings | DOI: https://doi.org/10.18174/FAIRdata2018.16286

# A community-driven paired data platform to accelerate natural product mining by combining structural information from genomes and metabolomes

Van der Hooft, J.J.J.[1], M. Schorn[1], P.C. Dorrestein[1], M.H. Medema[1]

*[1] Wageningen University & Research Bioinformatics Group, Droevendaalsesteeg 1, 6708PB Wageningen, The Netherlands*
*Corresponding author's e-mail: justin.vanderhooft@wur.nl*

Natural products are small molecules produced by bacteria, fungi, and plants that have a large variety of functions, including chemical defence and communication. Moreover, many of those natural products are exploited for therapeutic or medicinal use. For example, many antibiotics have natural products as origin and finding novel types of antibiotics is currently recognised as top priority in our combat against so-called superbugs that are multi-resistant against most commonly used antibiotics. Most of the microbes and plants around us are still unexplored for their biosynthetic potential – thus representing a wealth of unexplored chemistries that could be used to combat the antibiotics resistance problem. However, the main challenges lie in the quick dereplication of known natural products and the structural elucidation process of novel natural products from natural extracts. The production of natural products is encoded by groups of biosynthesis genes in an organism's genome. The last decade has witnessed progress in both genome mining tools that predict the biosynthetic potential of organisms based on their whole genome sequence, as well as metabolome mining tools that provide a comprehensive map of produced natural products based on mass spectrometry data. Both these mining strategies provide an arsenal of structural information; however, with each method individually, the full structural annotation of natural products remains very hard and often impossible.

With the accumulation of genome sequences from both microbiota and plants and improvements in metabolomics technologies, an exciting new route is opening up to improve mining for novel chemistries by computational algorithms that link information from the genome and metabolome. However, there is currently no resource available that stores computer-readable links between public genome and metabolome data. With such a platform in place, new algorithms can be developed that leverage information from both omics worlds.

Thus, here we propose a paired data platform that documents links between genomic data and metabolomics data stored in public repositories. Furthermore, we will also add links between biosynthesis gene clusters (encoding natural products) and their (tandem) mass spectra and structures. In this talk, I will highlight the key challenges that we are currently tackling to develop a community-driven platform that will store the increasing amount of links between available paired data sets and natural product structures. Since such links rely on publicly available data, the FAIR principles are key to ensure the success of such a platform. Moreover, defining the minimal needed information to establish reliable links is essential to make adding new links a worthwhile effort for the community. Finally, examples of how combined structural information from the genome and metabolome could help to accelerate i) mining for novel chemistry, and ii) defining entire natural product structures and how they structurally relate to other molecules. It will conclude that only when we start to exploit the complementary information from omics data we can effectively spend our resources to fully capture structures of the potentially next antibiotic blockbusters.