AMERICAN SOCIETY FOR MICROBIOLOGY | mSystems®

PERSPECTIVE
Novel Systems Biology Techniques

Computational Genomics of Specialized Metabolism: from Natural Product Discovery to Microbiome Ecology

Marnix H. Medema[a]

[a]Bioinformatics Group, Wageningen University, Wageningen, The Netherlands

ABSTRACT Microbial and plant specialized metabolites, also known as natural products, are key mediators of microbe-microbe and host-microbe interactions and constitute a rich resource for drug development. In the past decade, genome mining has emerged as a prominent strategy for natural product discovery. Initially, such mining was performed on the basis of individual microbial genome sequences. Now, these efforts are being scaled up to fully genome-sequenced strain collections, pan-genomes of bacterial genera, and large sets of metagenome-assembled genomes from microbial communities. The Medema research group aims to play a leading role in these developments by developing and applying computational approaches to identify, classify, and prioritize specialized metabolite biosynthetic gene clusters and pathways and to connect them to specific molecules and microbiome-associated phenotypes. Moreover, we are extending the scope of genome mining from microbes to plants, which will allow more comprehensive interpretation of the chemical language between hosts and microbes in a microbiome setting.

KEYWORDS bioinformatics, biosynthetic gene cluster, microbiome, natural products, specialized metabolism

Bacteria, fungi, and plants produce a wide range of specialized metabolites (also known as natural products) that allow them to thrive in their environments. In microbiomes, these molecules play key roles in competition and collaboration by serving as signals, weapons, nutrient-scavenging agents, and stress protectants. Many different chemical classes of natural products exist, including terpenes, polyketides, peptides, saccharides, and alkaloids. Thousands of these molecules are applied in human society as crop protection agents, antibiotics, chemotherapeutics, immunosuppressants, surfactants, and ingredients for food manufacturing.

The genes encoding natural product biosynthetic pathways are frequently physically clustered on the chromosome of the producing organism. Over 1,500 of these biosynthetic gene clusters BGCs) and their products have now been characterized experimentally (1). Intriguingly, this physical clustering makes it straightforward to identify biosynthetic pathways for novel molecules through computational genomic analysis, regardless of the fact that many BGCs are transcriptionally silent under typical laboratory conditions.

The continuous technological developments in DNA sequencing and assembly now make it affordable for individual research groups to acquire hundreds of complete bacterial genomes. Culture collections worldwide hold more than 1.5 million bacterial and fungal strains, large numbers of which are planned to be genome sequenced soon in several initiatives. Moreover, genomes can now be reconstructed from metagenomes thousands at a time (2) and massive metagenomic efforts such as the Earth Microbiome Project (3) plan to reconstruct around 500,000 genomes from diverse communities around the globe. It is not at all unrealistic to expect that within 5 to 10 years, the nucleotide sequence databases will contain millions of genome sequences of tens of

Received 9 November 2017 **Accepted** 23 November 2017 **Published** 6 March 2018

Citation Medema MH. 2018. Computational genomics of specialized metabolism: from natural product discovery to microbiome ecology. mSystems 3:e00182-17. https://doi.org/10.1128/mSystems.00182-17.

Address correspondence to marnix.medema@wur.nl.

Conflict of Interest Disclosures: M.H.M. reports grants from NWO, ZonMW, the NIH, and the Graduate School for Experimental Plant Sciences during the conduct of the study, in addition to personal fees from Hexagon Bio and Third Rock Ventures outside the submitted work.

mSystems® *vol. 3, no. 2, is a special issue sponsored by Janssen Human Microbiome Institute (JHMI).*

Computational genomics of specialized metabolism: from natural product discovery to microbiome ecology

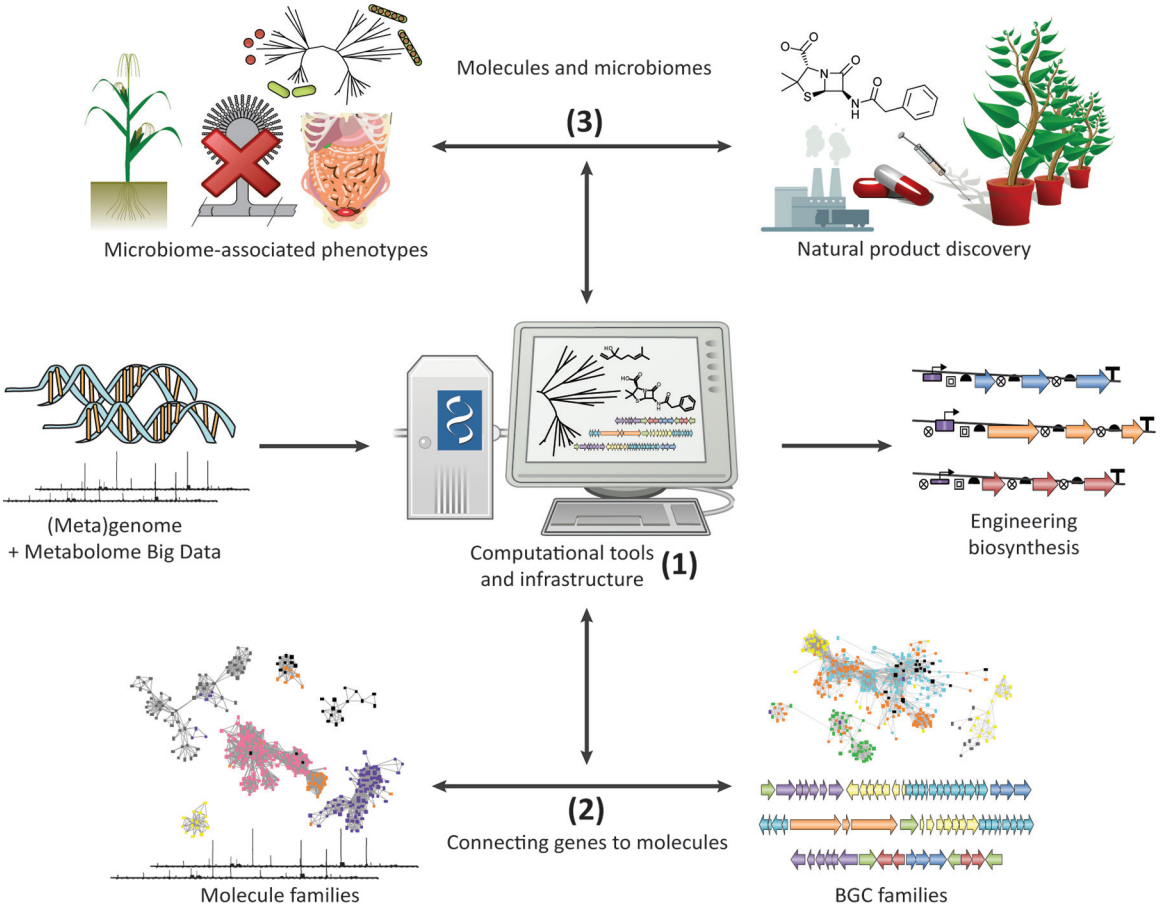

FIG 1 Overview of the research line of the Medema research group. We develop computational tools and infrastructure (part 1) to connect genes to molecules (part 2). With these technologies, we aim to accelerate natural product discovery and acquire an ecological understanding of the molecular mechanisms behind microbiome-associated phenotypes (MAPs) driven by specialized metabolism (part 3).

thousands of biological species. Similarly, plant and fungal genome sequencing is also being scaled up, with the sequencing of thousands of eukaryotic genomes planned for the next few years. At the same time, complementary data are being gathered by using metabolomics, transcriptomics, and large-scale phenotyping studies. This presents tremendous opportunities for genome-based natural product discovery, as millions of BGCs can be scoured to identify high-value molecules and to predict and assess their functions in ecology.

For the field studying specialized metabolite biosynthesis, this will require radical changes in the methods employed. Traditional approaches alone no longer suffice. Indeed, computation will play a more and more central role in the integration of large and diverse data sets and the generation of meaningful hypotheses for experimentation (Fig. 1).

FROM INDIVIDUAL GENOMES TO PANGENOMES AND METAGENOMES

The starting point of natural product genome mining is the identification of BGCs. This procedure is fully automated by antiSMASH (4), a computational pipeline and web server that is currently jointly coordinated by the Medema group and the research group of Tilmann Weber at the Technical University of Denmark. AntiSMASH not only identifies BGCs, it also compares identified BGCs to experimentally characterized reference gene clusters from the MIBiG repository (1) and provides chemical structure prediction for several classes of natural products. Precomputed results are available online in the antiSMASH database (5). As an open-source project, antiSMASH is con-

tinuously extended with new functionalities by researchers worldwide through a model of open collaboration.

While antiSMASH effectively automates the analysis of individual genomes, it was already conceived in 2010 and therefore was never designed for the simultaneous exploration of hundreds or thousands of genomes or metagenomes. It is highly likely that the millions of BGCs that are becoming available will offer novel solutions for combating multidrug-resistant pathogens, treating cancer, and protecting crops against dangerous pathogens. The key challenge, however, is to find these much-desired needles in such a giant haystack.

To address this, we are currently developing novel solutions. To first acquire high-quality sets of BGCs from complex and large (meta)genomic data sets, new algorithms are being developed by us and several collaborating research groups to reconstruct full BGCs from metagenomic assemblies or from large sets of medium-quality draft genomes. Subsequently, the construction of BGC sequence similarity networks and the clustering of BGCs into gene cluster families (GCFs) are key methods to reduce the complexity of sets of thousands of BGCs and provide a bird's-eye perspective on the underlying biosynthetic diversity (6–8). Our new software BiG-SCAPE (J. Navarro-Muñoz et al., unpublished data; https://git.wageningenur.nl/medema-group/BiG-SCAPE) streamlines and optimizes these methods to allow detailed analysis of the relationships between large numbers of BGCs without the need for supercomputing. Through annotation propagation with reference data from MIBiG (1), it allows rapid identification of GCFs with known and unknown functions and allows tracing of the taxonomic distribution of their pangenomic absence/presence patterns. Moreover, it provides a rich network visualization that allows interactive exploration of the data by dynamically browsing the network and searching it on the basis of taxonomic or Pfam identifiers. Finally, ongoing integration with CORASON (F. Barona-Gómez, personal communication; https://github.com/nselem/EvoDivMet) will facilitate phylogenetic reconstruction of GCFs to identify the relationships of the underlying BGCs at high resolution. In this way, scientists will be able to perform interactive exploration of biosynthetic diversity across, e.g., all genomes of the genus *Burkholderia*, multiple metagenomic data sets from plant rhizospheres, or all genomes associated with the human oral microbiome. Also, in the future, we plan to use precomputed BGC predictions for all publicly available genomes to populate an online database of standardized GCFs with curated annotations.

CONNECTING GCFs TO MOLECULES

Exploration of biosynthetic diversity should never be a goal in itself, or it will remain nothing more than a "stamp-collecting exercise." It should generate new hypotheses and illuminate real mechanisms and chemistry. Importantly, genomic data have the potential to greatly illuminate metabolomes. It has been estimated that in metabolomic data, >95% of the metabolite-derived masses cannot be linked to structures or functions (9). Matching of these masses to pathways and the strains producing these metabolites will play crucial roles in the identification of their structures and functions. Effective connection of genomic and metabolomic data will entail a bidirectional process, in which chemical features are predicted from BGCs, as well as from tandem mass spectrometric peak patterns.

The MIBiG initiative documented the connections between a large number of BGCs and the chemical structures of their products and allows for the standardized storage of data on enzymatic classes involved in these pathways, as well as their substrate specificities. This presents a rich data set to train algorithms to make powerful predictions about chemical (sub)structures of BGC products based on the DNA sequence of a BGC alone. For example, the SANDPUMA algorithm (10) for substrate specificity prediction of nonribosomal peptide synthetases made key improvements upon previous methods by extending earlier training sets with hundreds of new data points from MIBiG. Our recent work with the Dorrestein lab (11) showed that connecting such chemistry predictions to large-scale metabolomics of pseudomonads allowed the

identification of new families of cyclic lipopeptides, which are known to mediate key interorganismal interactions in plant microbiomes.

The next challenge will be to take the prediction of substructures beyond peptides and include a wider variety of genetic features that can be correlated with mass spectrometric features. The former would include subclusters that encode pathways toward key precursors, enzyme-coding genes for group transfer of chemical monomers, and more "exotic" scaffold biosynthesis enzymes such as *trans*-acyltransferase polyketide synthases and terpene cyclases. This will make it possible to systematically extend metabologenomic matching (12) from correlation-based mapping alone to feature-based mapping. The emergence of more and more paired genomic-metabolomic data sets (with both types of data from the same samples) will undoubtedly accelerate these efforts.

CONNECTING MICROBES TO PLANT AND HUMAN HOSTS

Even molecules are not the endpoint of natural product discovery. In the end, it is their function that matters most. Such function goes beyond pharmacological activities such as antibiotic or cytotoxic activity, and from a fundamental perspective, is really about the fitness effects of molecules on both the producing organism and members of the community surrounding it. We feel that microbiomes and their chemical language should be approached from a systems perspective. In the end, this allows acquiring a real understanding of how specialized metabolites and other molecular mechanisms shape key microbiome-associated phenotypes (MAPs) (15), such as disease suppression or growth promotion, in plant and human microbiomes. These interactions comprise both host and microbial components.

From the microbial side, metagenomics and metatranscriptomics allow the identification of differential abundance and differential expression of BGCs in microbial communities, which can be correlated with MAPs. Again, the grouping of BGCs into GCFs that represent functional traits across multiple organisms plays a key role here, as they allow metagenome-wide association studies of these genetic traits with (un)desirable MAPs. This then generates hypotheses that can be tested in the laboratory through the isolation and characterization of specific strains (e.g., in synthetic microbial communities) or through heterologous expression of BGCs refactored through synthetic biology.

From the host side, it is increasingly appreciated that small molecules also play key roles in shaping the microbiota. In plants, this includes root exudation (positive regulation), as well as secretion of defense compounds (negative regulation). The technological ability to sequence high-quality plant genomes along with (time series) transcriptomes across a range of conditions paves the way for the development of genome mining strategies for the identification of plant biosynthetic pathways by studying patterns of genomic colocalization, coexpression, and coevolution of enzyme-coding genes, in combination with metabolomic and phenotypic data (13). Our recently launched platform plantiSMASH, the plant equivalent of antiSMASH, facilitates many of these analyses. In the human microbiome, chemical interactions between hosts and microbes are also of key importance. For example, gut microbes transform bile acids produced by the liver into a wide range of secondary bile acids, many of which have a major impact on human health (14). Building an "antiSMASH"-like platform for the identification of microbial pathways involved in such chemical transformations has great potential to foster our understanding of the role of the human and animal microbiota in health and disease.

Concluding, we are excited about the prospects for the reinvigorated study of specialized metabolism and are convinced that the integration of cutting-edge omics technologies, computation, and foundational chemical and ecological concepts will provide many new insights into the chemical language of life and its many biotechnological applications that can improve human well-being.

ACKNOWLEDGMENTS

I thank members of the Medema group and the larger Wageningen Bioinformatics Group for inspiring and stimulating discussions and for their feedback on this article.

Research in the Medema group is currently supported by VENI grant 863.15.002 from The Netherlands Organization for Scientific Research (NWO), an NWO Groen grant (ALWGR.2015.1), an NWO Eranet CoBiotech grant (053.80.739), an NWO-TTW NACTAR grant (16440), an NWO ASDI grant (ASDI.2017.030), the Graduate School for Experimental Plant Sciences (EPS), National Institutes of Health Genome to Natural Products Network supplementary awards (U01GM110706 and U101GM110699), and ZonMW DTL Hotel projects 435003008 and 435003011.

REFERENCES

1. Medema MH, Kottmann R, Yilmaz P, Cummings M, Biggins JB, Blin K, De Bruijn I, Chooi YH, Claesen J, Coates RC, Cruz-Morales P, Duddela S, Düsterhus S, Edwards DJ, Fewer DP, Garg N, Geiger C, Gomez-Escribano JP, Greule A, Hadjithomas M, Haines AS, Helfrich EJN, Hillwig ML, Ishida K, Jones AC, Jones CS, Jungmann K, Kegler C, Kim HU, Kötter P, Krug D, Masschelein J, Melnik AV, Mantovani SM, Monroe EA, Moore M, Moss N, Nützmann HW, Pan G, Pati A, Petras D, Reen FJ, Rosconi F, Rui Z, Tian Z, Tobias NJ, Tsunematsu Y, Wiemann P, Wyckoff E, et al. 2015. Minimum information about a biosynthetic gene cluster. Nat Chem Biol 11:625–631. https://doi.org/10.1038/nchembio.1890.
2. Parks DH, Rinke C, Chuvochina M, Chaumeil PA, Woodcroft BJ, Evans PN, Hugenholtz P, Tyson GW. 2017. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat Microbiol 2:1533–1542. https://doi.org/10.1038/s41564-017-0012-7.
3. Gilbert JA, Jansson JK, Knight R. 2014. The Earth Microbiome Project: successes and aspirations. BMC Biol 12:69. https://doi.org/10.1186/s12915-014-0069-1.
4. Blin K, Wolf T, Chevrette MG, Lu X, Schwalen CJ, Kautsar SA, Suarez Duran HG, De Los Santos ELC, Kim HU, Nave M, Dickschat JS, Mitchell DA, Shelest E, Breitling R, Takano E, Lee SY, Weber T, Medema MH. 28 April 2017. antiSMASH 4.0-improvements in chemistry prediction and gene cluster boundary identification. Nucleic Acids Res. https://doi.org/10.1093/nar/gkx319.
5. Blin K, Medema MH, Kottmann R, Lee SY, Weber T. 2017. The antiSMASH database, a comprehensive database of microbial secondary metabolite biosynthetic gene clusters. Nucleic Acids Res 45:D555–D559. https://doi.org/10.1093/nar/gkw960.
6. Cimermancic P, Medema MH, Claesen J, Kurita K, Wieland Brown LC, Mavrommatis K, Pati A, Godfrey PA, Koehrsen M, Clardy J, Birren BW, Takano E, Sali A, Linington RG, Fischbach MA. 2014. Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. Cell 158:412–421. https://doi.org/10.1016/j.cell.2014.06.034.
7. Doroghazi JR, Albright JC, Goering AW, Ju KS, Haines RR, Tchalukov KA, Labeda DP, Kelleher NL, Metcalf WW. 2014. A roadmap for natural product discovery based on large-scale genomics and metabolomics. Nat Chem Biol 10:963–968. https://doi.org/10.1038/nchembio.1659.
8. Medema MH, Fischbach MA. 2015. Computational approaches to natural product discovery. Nat Chem Biol 11:639–648. https://doi.org/10.1038/nchembio.1884.
9. van der Hooft JJJ, Wandy J, Barrett MP, Burgess KEV, Rogers S. 2016. Topic modeling for untargeted substructure exploration in metabolomics. Proc Natl Acad Sci U S A 113:13738–13743. https://doi.org/10.1073/pnas.1608041113.
10. Chevrette MG, Aicheler F, Kohlbacher O, Currie CR, Medema MH. 2017. SANDPUMA: ensemble predictions of nonribosomal peptide chemistry reveals biosynthetic diversity across actinobacteria. Bioinformatics 33:3202–3210. https://doi.org/10.1093/bioinformatics/btx400.
11. Nguyen DD, Melnik AV, Koyama N, Lu X, Schorn M, Fang J, Aguinaldo K, Lincecum TL Jr, Ghequire MGK, Carrion VJ, Cheng TL, Duggan BM, Malone JG, Mauchline TH, Sanchez LM, Kilpatrick AM, Raaijmakers JM, De Mot R, Moore BS, Medema MH, Dorrestein PC. 2016. Indexing the Pseudomonas specialized metabolome enabled the discovery of poaeamide B and the bananamides. Nat Microbiol 2:16197. https://doi.org/10.1038/nmicrobiol.2016.197.
12. Goering AW, McClure RA, Doroghazi JR, Albright JC, Haverland NA, Zhang Y, Ju KS, Thomson RJ, Metcalf WW, Kelleher NL. 2016. Metabologenomics: correlation of microbial gene clusters with metabolites drives discovery of a nonribosomal peptide with an unusual amino acid monomer. ACS Cent Sci 2:99–108. https://doi.org/10.1021/acscentsci.5b00331.
13. Medema MH, Osbourn A. 2016. Computational genomic identification and functional reconstitution of plant natural product biosynthetic pathways. Nat Prod Rep 33:951–962. https://doi.org/10.1039/c6np00035e.
14. Wahlström A, Sayin SI, Marschall HU, Bäckhed F. 2016. Intestinal crosstalk between bile acids and microbiota and its impact on host metabolism. Cell Metab 24:41–50. https://doi.org/10.1016/j.cmet.2016.05.005.
15. Oyserman B, Medema MH, Raaijmakers J. 2018. Road MAPs to engineer host microbiomes. Curr Opin Microbiol 43:46–54. https://doi.org/10.1016/j.mib.2017.11.023.