UNIVERSITY OF CALIFORNIA SAN DIEGO

Scalable Computational Methods for Discovering Novel Natural Products

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Bioinformatics and Systems Biology

by

Bahar Behsaz

Committee in charge:

Professor Pavel A. Pevzner, Chair
Professor Rob Knight, Co-Chair
Professor Vineet Bafna
Professor Pieter S. Dorrestein
Professor Siavash Mir Arabbaygi

2020

The Dissertation of Bahar Behsaz is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Co-Chair

Chair

University of California San Diego

2020

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my advisor, Professor Pavel A. Pevzner, for his unwavering kindness, support, and guidance, throughout my doctoral studies. I am honored and immensely fortunate to have finished my dissertation under supervision of the person whose remarkable body of work sparked my interest in Bioinformatics. Pavel is not only an exceptional researcher and a visionary in the field bioinformatics but is also a devoted mentor who fosters his students in all aspects of their graduate career.

I would also like to thank Professor Rob Knight who has been a great source of insight and inspiration during my graduate career. I am thankful for his valuable comments and generous support.

I would also like to extend my gratitude to Professor Pieter C. Dorrestein for his support and mentorship in the past five years. His enthusiasm and encouragement have inspired me to push further in my research endeavors.

My gratitude also extends to Professors Vineet Bafna, Nuno Bandeira, and Siavash Mir Arabbaygi for their valuable suggestions and support during my graduate studies.

I am also indebted to Professor Hosein Mohimani for his guidance, help, and cooperation across the years in my doctoral research. I could not have finished this dissertation without his help and encouragements.

I would also like to thank my previous professional and academic supervisors, Professors Ebad Mahmoodian, Ladislav Stacho, Cedric Chauve, and Inanc Birol, for encouraging me to go further and seek excellence in my career.

I would also extend my appreciation to all of my friends whom I met in the past five years at UC San Diego, either at Bioinformatics & Systems Biology Program or Computer Science and Engineering Department or elsewhere. I am grateful for the memorable moments we have shared.

I send my genuine gratitude to my mother, father, and brother for their love and support. I am also thankful to my cat, Tiny, who accompanied me during many hours of work, day or night, purring along the way. Finally, I would also want to thank my husband, David, for his support, his belief in me, his kind love, and his consistent encouragement.

Chapter 1, in full, is a reformatted reprint of "De novo peptide sequencing reveals many cyclopeptides in the human gut and other environments" as it appears in *Cell Systems, 2020*[48] by Bahar Beshsaz, Hosein Mohimani, Alexey Gurevich, Andrey Prjibelski, Mark Fisher, Fernando Vargas, Larry Smarr, Pieter C. Dorrestein, Joshua S. Mylne, and Pavel A. Pevzner. The dissertation author was the primary author of this material.

Chapter 2, in full, has been submitted for journal publication by: Bahar Behsaz[*], Edna Bode[*], Alexey Gurevich, Yan-ni Shi, Florian Grundmann, Deepa Archarya, Andrés Mauricio Caraballo-Rodríguez, Amina Bouslimani, Morgan Panitchpakdi, Annabell Linck, Changhui Guan, Julia Oh, Pieter C. Dorrestein, Helge B. Bode, Pavel A. Pevzner, and Hosein Mohimani. The dissertation author was the primary author of this work.

Chapter 3, in full, is a reformatted reprint of a part of "metaFlye: scalable long-read metagenome assembly using repeat graphs" as it appears in *Nature Methods* (2020) by Mikhail, Kolmogorov, Derek M. Bickhart, Bahar Behsaz, Alexey Gurevich, Mikhail Rayko, Sung Bong Shin, Kristen Kuhn, Jeffrey Yuan, Evgeny Polevikov, Timothy P. L. Smith & Pavel A. Pevzner. The dissertation author was a primary author of this work.

VITA

2006-2010:	Bachelor of Science in Computer Science, *Sharif University of Technology*, Tehran, Iran.
2010-2013	Master of Science in Mathematics, *Simon Fraser University*, Vancouver, Canada.
2013-2015	Computational Biologist, *Genome Sciences Centre*, Vancouver, Canada.
2015-2020	Doctor of Philosophy in Bioinformatics & Systems Biology, *University of California San Diego*, San Diego, USA.

PUBLICATIONS

Behsaz, B., Edna, B., Gurevich, A., Shi, Y., Grundmann, F., Caraballo-Rodríguez, A.M., Bouslimani, A., Panitchpakdi, M., Linck, A., Guan, C., Oh, J., Dorrestein, P.C. Bode, H.B., Pevzner, P.A., Mohimani, H., 2020. Integrating Metagenomics and Metabolomics for Scalable Non-Ribosomal Peptide Discovery. *Under Review*.

Kolmogorov, M., Bickhart, D.M., **Behsaz, B.**, Gurevich, A., Rayko, M., Shin, S.B., Kuhn, K., Yuan, J., Polevikov, E., Smith, T.P. and Pevzner, P.A., 2020. metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nature Methods*, pp.1-8.

Behsaz, B., Mohimani, H., Gurevich, A., Prjibelski, A., Fisher, M., Vargas, F., Smarr, L., Dorrestein, P.C., Mylne, J.S. and Pevzner, P.A., 2020. De novo peptide sequencing reveals many cyclopeptides in the human gut and other environments. *Cell Systems*, *10*(1), pp.99-108.

McDonald, D., Hyde, E., Debelius, J.W., Morton, J.T., Gonzalez, A., Ackermann, G., Aksenov, A.A., **Behsaz, B.**, Brennan, C., Chen, Y. and Goldasich, L.D., 2018. American Gut: an open platform for citizen science microbiome research. *Msystems*, *3*(3), pp.e00031-18.

Fisher, M.F., Zhang, J., Taylor, N.L., Howard, M.J., Berkowitz, O., Debowski, A.W., **Behsaz, B.**, Whelan, J., Pevzner, P.A. and Mylne, J.S., 2018. A family of small, cyclic peptides buried in preproalbumin since the Eocene epoch. *Plant direct*, *2*(2), p.e00042.

Brown, T.M., Hammond, S.A., **Behsaz, B.**, Veldhoen, N., Birol, I. and Helbing, C.C., 2017. De novo assembly of the ringed seal (Pusa hispida) blubber transcriptome: A tool that enables identification of molecular health indicators associated with PCB exposure. *Aquatic Toxicology*, *185*, pp.48-57.

Warren, R.L., Yang, C., Vandervalk, B.P., **Behsaz, B.**, Lagman, A., Jones, S.J. and Birol, I., 2015. LINKS: Scalable, alignment-free scaffolding of draft genomes with long reads. *GigaScience*, *4*(1), pp.s13742-015.

Birol, I., **Behsaz, B.**, Hammond, S.A., Kucuk, E., Veldhoen, N. and Helbing, C.C., 2015. De novo transcriptome assemblies of Rana (Lithobates) catesbeiana and Xenopus laevis tadpole livers for comparative genomics without reference genomes. *PloS one*, *10*(6), p.e0130720.

Behsaz, B., Maňuch, J. and Stacho, L., 2012, August. Turing universality of step-wise and stage assembly at temperature 1. In *International Workshop on DNA-Based Computers* (pp. 1-11). Springer, Berlin, Heidelberg.

ABSTRACT OF THE DISSERTATION

Scalable Computational Methods for Discovering Novel Natural Products

by

Bahar Behsaz

Doctor of Philosophy in Bioinformatics and Systems Biology

University of California San Diego, 2020

Professor Pavel A. Pevzner, Chair
Professor Rob Knight, Co-Chair

Today, as the world is stricken by the proliferation of novel infectious pathogens, we are faced with the urgent need for new anti-infective therapeutic agents. *Natural products*, also known as specialized metabolites, are chemical compounds produced by living organisms and have served as an excellent source for drug discovery. Many clinically used small molecules including various antimicrobial, anticancer, antiviral, and immunosuppressant drugs, are either natural products or are inspired by them. Traditionally, natural products were discovered mostly through slow and laborious experiments that often lead to rediscovering previously known compounds.

Over the past decade, advancements in *short/long-read (meta)genomics* and *tandem mass spectrometry* (MS/MS) technologies provided an unprecedented resource for large-scale natural product discovery. In accordance with these advancements, scalable bioinformatics algorithms are required to leverage this massive data and enable analyses of natural products across thousands of samples. In this dissertation, I present several scalable computational methods for discovering novel natural products using the (MS/MS-based) metabolomics and/or (meta)genomics data.

In the first chapter, I present CycloNovo, the first algorithm for scalable *de novo* sequencing of MS/MS data to discover cyclic and branch cyclic peptides (referred to as cyclopeptides). Cyclopeptides constitute a diverse and biomedically important class of natural products. CycloNovo employs *de Bruijn* graphs, the workhorse of DNA sequencing algorithms, for efficient cyclopeptide sequencing and revealed a wealth of novel cyclopeptides, including a large hidden cyclopeptidome in the human gut.

In the following chapters, I discuss bioinformatics methods for discovering Non-Ribosomal Peptides (NRPs) that include a multitude of antibiotics and other clinically used drugs. NRPs are produced by metabolic pathways partially encoded by Biosynthetic Gene Clusters (BGCs). In the second chapter, I present NRPminer, a modification-tolerant and scalable algorithm for NRP discovery by integrating (meta)genomic and MS/MS data. NRPminer identified many novel NRPs from different origins, including novel NRPs produced by soil-associated microbes and human microbiota. Finally, I discuss the problem of identifying NRP-producing BGCs in the human gut microbiome and I show long-read metagenomic assemblies can be used to reveal many BGCs that synthesize previously unknown NRPs in the human gut microbiome.

INTRODUCTION

Natural products, also known as *specialized metabolites*, have been used as a rich resource for discovering a wide range of therapeutic agents, including many *anti-microbial*, *anticancer*, *antiviral*, *antifungal* and *immunosuppressants* drugs[1,2]. For example, over the past four decades, nearly 52% and 64% of small molecules approved in the area of infectious disease and cancer, respectively, are either inspired by or directly derived from natural products. Natural products have also found applications in the crop protection[3] and food preservation[4] industries. Traditionally, natural products are discovered using bioassay-guided fractionation followed by structure elucidation with nucleic magnetic resonance spectrometry[5,6]. These methods are laborious and time-consuming and often lead to rediscovery of previously known compounds[7]. Exciting, over the past decade, progress in *short/long-read (meta)genomics* and the advancements in *tandem mass spectrometry* for metabolomic screening, has created a new opportunity for large-scale search for natural products, across thousands of samples from various origins and environments[6].

Tandem mass spectrometry for natural product discovery.

Advancements in tandem mass spectrometry (MS/MS) established this technology as a fast, sensitive, and reliable approach for large-scale natural product discovery over the past decade. The MS/MS spectra can be regarded as bar codes or fingerprints of the metabolites present in a sample[6]. A single MS/MS project can yield millions of spectra representing a wide range of metabolites and natural products[6,7]. Natural products can be classified to a range of chemical classes based on their underlying monomers and biosynthetic origin. *Peptidic natural products* (PNPs) constitute an important class of medicinal natural product that include some of the most potent antibiotics. Among the classes of natural products, PNPs are the most amenable MS/MS technologies, due to the nature of the bonds between their amino acids[6,7]. Several scalable computational methods have been developed over the past few years that specifically targets this class of natural products.

Identifying known PNPs and their variants in MS/MS data.

To avoid rediscovering previously characterized compounds, an initial step in natural product discovery is to identify the known compounds represented by a given spectral datasets. Currently, three main approaches are available for MS/MS-based identification of known compounds. This section provides a brief description for each of these approaches. In s*pectral library matching,* a given spectrum *S*, is compared against a library of already characterized spectra (*spectral library*) to find spectra *similar* to *S*[8]. While this approach has been successful for finding known compounds, but often a known compound might be absent from the sample while its *related* PNP (which has some modifications compared to the known compound) is present. To identify such metabolites in a scalable fashion, Bandeira et al.[9] introduced the concept of *spectral networks* that reveal the spectra of related peptides without knowing their amino acid sequences. Nodes in a spectral network correspond to spectra, while edges connect *spectral pairs*, i.e. spectra of peptides differing by a single modification or a mutation. Ideally, each connected component of a spectral network corresponds to a *molecular family*[10] representing a set of similar PNPs. In contrast to the former method, spectral networks can identify a PNP using the spectra of its related PNPs.

Spectral matching and spectral networking have proven effective in identifying spectra originating from known PNP (or their related peptides) in metabolomics datasets[6]; however, only a small fraction of known PNPs are represented in the current reference spectral libraries[6,7].To overcome this barrier, several bioinformatics methods have been developed over the past few years that use the chemical structure of known PNPs to identify the spectra originating from them. We refer to a database of chemical structures representing known PNPs as *PNP database*. Given a spectrum *S* and a PNP database *DB*, *database search* refers to the process of finding a peptide (or its variant) in *DB* that generated *S*. Several computational tools have been developed for scalable database search. Dührkop et al., 2015[11], proposed CSI:fingerID that utilizes support vector machines trained on previously annotated spectra to predict the MS/MS fragmentation patterns and find the spectra representing the known compounds[11]. While this method worked well for a number of small metabolites, it failed to generalize due to lack of sufficient annotated spectra in natural product studies. As a result, the database search problem remained open for most PNPs[6,7].

To address this bottleneck, several dedicated scalable computational methods were developed that specifically targets PNPs[7,12]. Similar to the methods in traditional proteomics[13], these methods use statistical measures to form *matches* between the chemical structure of a PNP *P* and a given spectrum *S*. Dereplicator[14] was the first scalable algorithm that systematically links structures from a large PNP database to MS/MS spectra by using specific *in silico* MS/MS fragmentation rules in PNPs[6,14]. Furthermore, this method was complimented by the VarQuest[15] which is able to find variants of known PNPs represented in a spectral dataset. Later, Mohimani et al. 2019[16], introduced Dereplicator+ and expanded this idea to other classes of natural products.

Discovering novel PNPs using MS/MS data.

As high-throughput experimental and computational technologies became a staple in natural product discovery research[5], number of natural products platforms such as the Global Natural Products Social (GNPS) molecular networking[17] were developed. GNPS provides an online repository to share massive spectral datasets. Furthermore, GNPS delivers a scalable platform to apply a variety of computational tools for finding known and novel natural products (including most approaches described above)[18]. The GNPS project has already gathered nearly half a billion of information-rich tandem mass spectra and is an untapped gold mine for discovering new molecules. However, the utility of the GNPS network is mostly limited to the identification of previously discovered molecules and their variants using the methods described above[6]. Currently, only about 5% percent of the GNPS spectra are annotated[15,17]. This highlights the need for novel algorithms for annotating large spectral datasets that are not bound by the current database of known PNPs. In this dissertation, I present several such methods using metabolomics and (meta)genomic data.

De novo PNP sequencing of MS/MS data.

De novo PNP *sequencing* is the process of determining the amino acid sequence of a PNP from a spectrum alone, *i.e.* without using any database or genomic information. Although recent studies made progress towards PNP database search[14,15,19,20], existing *de novo* PNP sequencing algorithms[21–24] are not compatible with the large-scale nature of current metabolomics and mass spectrometry studies. These tools are rarely used[7,25] because they are inaccurate, too slow for analyzing large spectral datasets, are limited to the proteinogenic amino acids, and cannot distinguish cyclopeptides from other compounds.

To address this issue, I developed CycloNovo[26] algorithm for finding *cyclic* and *branch cyclic peptides* (*cyclopeptides*). Cyclopeptides are an important class of bioactive PNPs that include many antibiotics and anti-tumor compounds[2,22]. The discovery of the cyclopeptide gramicidin S in 1942 (the first antibiotic used for treating soldiers during the World War II) led to two Nobel prizes and has been followed by the discovery of ≈400 families of cyclopeptides (*cyclofamilies*) in the last 75 years[15]. A relatively small number of known cyclofamilies reflects the experimental and computational challenges in cyclopeptide discovery. The question of how many cyclofamilies stayed below the radar of previous studies, even though their spectra were already deposited in public databases, remains open. To answer this question, CycloNovo first recognizes *cyclospectra* (tandem mass spectra that originated from cyclopeptides) in large spectral datasets. Afterwards, CycloNovo *de novo* sequences the recognized cyclospectra (determining the cyclopeptide sequence from a spectrum alone).

In **Chapter 1**, I describe how CycloNovo uses *de Bruijn graphs*, that are the workhorse of DNA sequencing algorithms, to de novo sequence cyclospectra. CycloNovo is the first scalable PNP sequencing method and reconstructed many new cyclopeptides, which were validated with transcriptome and metagenome analyses[26]. Our benchmarking revealed a large hidden *cyclopeptidome* in the *human gut* and other environments, including a wealth of anti-microbial cyclopeptides from food that survive the complete human gastrointestinal tract.

Integrating Metabolomics and (Meta)genomics for Discovering *Non-Ribosomal Peptides*.

Non-Ribosomal Peptides (NRPs) represent a diverse class of PNPs that include many antibiotics, immunosuppressants, anticancer agents, toxins, siderophores, pigments, and cytostatics[1,27–29]. NRPs have been reported in various habitats, from marine environments[30] to soil[29] and even human microbiome[31–34]. However, the discovery of novel NRPs remains a time-consuming and onerous process because NRPs are not directly encoded in the genome and are instead assembled by *Non-Ribosomal Peptide Synthetases* (NRPSs). NRPSs are multi-modular proteins that are encoded by a set of chromosomally adjacent genes called *biosynthetic gene clusters* (*BGCs*)[35,36]. As the microbial projects expanded in the past decade, multiple *genome mining* methods have been developed for predicting the molecular products synthesized by a given BGC[37–41].

Despite a great progress in genome mining methods in recent years, only a small fraction of the identified BGCs have been successfully connected to their metabolites so far[42,43]as genome mining tools predict too many putative NRPs synthesized by a given BGC. Due to this large false positive rate, it remains unclear which of these putative NRPs are correct or how to identify *post-assembly modifications* (PAM) of amino acids in the final NRPs in a *blind mode*, without knowing which modifications exist in the sample. Therefore, genome mining and identification of BGCs should without revealing the true chemical diversity encoded by these BCGs[44] does not capture their full potential for discovering novel NRPs. To do so, it has been shown integrating (meta)genomics and metabolomics is necessary for realizing the promise of large-scale natural products discovery[5,25,45]. In **Chapter 2**, I present *NRPminer* algorithm, a scalable and modification-tolerant tool for discovering novel NRPs by combining the power of MS/MS and (meta)genome mining. Using NRPminer, I identified many known and novel NRPs from different environments, including four novel NRP families from *soil-associated* microbes as well as a novel NRP from *human microbiota*, thus demonstrating the power of NRPminer for discovering novel bioactive NRPs.

Search for novel biosynthetic gene clusters in human gut assemblies.

Finally, I discuss the problem of *identifying NRP-producing BGCs* from complex long-read metagenomics datasets. Genome mining approaches fail unless a BGC is fully assembled within a single contig. However, NRP-producing BGCs are difficult to assemble as they are long (average length ~60 kb) and repetitive (made up of series of highly similar domains). Consequently, short-read metagenome assemblers hardly ever capture these BGCs within a single contig and hence are not adequate for BGC identification in complex samples[46].

In **Chapter 3**, I address the problem of assembling BGC sequences using *long-read metagenomics*. I show that assembly algorithms specialized for this technology revealed several novel BGCs in human gut microbiome encoding for previously unknown NRPs[47] as well as multiple NRP-producing BGCs associated with colorectal cancer[33,34]

In summary, this dissertation presents novel computational methods utilizing (meta)genomic and metabolomics data to identify novel natural product and/or recognize the enzymes involved in their biosynthesis. Furthermore, it also presents a variety of novel natural products discovered by applying the described methods to multiple large-scale (multi-omics) datasets.

REFERENCES

1. Harvey, A. L., Edrada-Ebel, R. & Quinn, R. J. The re-emergence of natural products for drug discovery in the genomics era. *Nature Reviews Drug Discovery* **14,** 111–129 (2015).

2. Newman, D. J. & Cragg, G. M. Natural Products as Sources of New Drugs over the Nearly Four Decades from 01/1981 to 09/2019. *Journal of Natural Products* **83,** 770–803 (2020).

3. Cantrell, C. L., Dayan, F. E. & Duke, S. O. Natural products as sources for new pesticides. *Journal of Natural Products* **75,** (2012).

4. Tiwari, B. K., Valdramidis, V. P., O'Donnell, C. P., Muthukumarappan, K., Bourke, P. & Cullen, P. J. Application of natural antimicrobials for food preservation. *Journal of Agricultural and Food Chemistry* **57,** 5987–6000 (2009).

5. Medema, M. H. & Fischbach, M. A. Computational approaches to natural product discovery. *Nature chemical biology* **11,** 639–648 (2015).

6. van der Hooft, J. J. J., Mohimani, H., Bauermeister, A., Dorrestein, P. C., Duncan, K. R. & Medema, M. H. Linking genomics and metabolomics to chart specialized metabolic diversity. *Chemical Society reviews* **49,** 3297–3314 (2020).

7. Mohimani, H. & Pevzner, P. A. Dereplication, sequencing and identification of peptidic natural products: from genome mining to peptidogenomics to spectral networks. *Natural product reports* **33,** 73–86 (2016).

8. Frank, A. M., Monroe, M. E., Shah, A. R., Carver, J. J., Bandeira, N., Moore, R. J., Anderson, G. A., Smith, R. D. & Pevzner, P. A. Spectral archives: extending spectral libraries to analyze both identified and unidentified spectra. *Nature Methods* **8,** 587–591 (2011).

9. Watrous, J., Roach, P., Alexandrov, T., Heath, B. S., Yang, J. Y., Kersten, R. D., van der Voort, M., Pogliano, K., Gross, H., Raaijmakers, J. M., Moore, B. S., Laskin, J., Bandeira, N. & Dorrestein, P. C. Mass spectral molecular networking of living microbial colonies. *Proceedings of the National Academy of Sciences U.S.A.* **109,** 1743–1752 (2012).

10. Mohimani, H., Wei-Ting Liu, Y.-L. Y., Susana P. Gaudêncio, W. F., Dorrestein, P. C. & Pevzner, P. A. Multiplex de novo sequencing of peptide antibiotics. *Journal of Computational Biology* **18,** 1371–1381 (2011).

11. Dührkop, K., Shen, H., Meusel, M., Rousu, J. & Böcker, S. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proceedings of the National Academy of Sciences of the United States of America* **112,** 12580–12585 (2015).

12. Mohimani, H., Kim, S. & Pevzner, P. A. A new approach to evaluating statistical significance of spectral identifications. *Journal of Proteome Research* **12,** 1560–1568 (2013).

13. Kim, S. & Pevzner, P. A. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nature Communications* **5,** 5277 (2014).

14. Mohimani, H., Gurevich, A., Mikheenko, A., Garg, N., Nothias, L.-F., Ninomiya, A., Takada, K., Dorrestein, P. C. & Pevzner, P. A. Dereplication of peptidic natural products through database search of mass spectra. *Nature Chemical Biology* **13,** 30–37 (2017).

15. Gurevich, A., Mikheenko, A., Shlemov, A., Korobeynikov, A., Mohimani, H. & Pevzner, P. A. Increased diversity of peptidic natural products revealed by modification-tolerant database search of mass spectra. *Nature Microbiology* **3,** 319–327 (2018).

16. Mohimani, H., Gurevich, A., Shlemov, A., Mikheenko, A., Korobeynikov, A., Cao, L., Shcherbin, E., Nothias, L. F., Dorrestein, P. C. & Pevzner, P. A. Dereplication of microbial metabolites through database search of mass spectra. *Nature Communications* **9,** (2018).

17. Wang, M., Carver, J. J., Phelan, V. V, Sanchez, L. M., Garg, N., Peng, Y., Nguyen, D. D., Watrous, J., Kapono, C. A., Luzzatto-Knaan, T., Porto, C., Bouslimani, A., Melnik, A. V, Meehan, M. J., Liu, W.-T., Crüsemann, M., Boudreau, P. D., Esquenazi, E., Sandoval-Calderón, M., Kersten, R. D., Pace, L. A., Quinn, R. A., Duncan, K. R., Hsu, C.-C., Floros, D. J., Gavilan, R. G., Kleigrewe, K., Northen, T., Dutton, R. J., Parrot, D., Carlson, E. E., Aigle, B., Michelsen, C. F., Jelsbak, L., Sohlenkamp, C., Pevzner, P., Edlund, A., McLean, J., Piel, J., Murphy, B. T., Gerwick, L., Liaw, C.-C., Yang, Y.-L., Humpf, H.-U., Maansson, M., Keyzers, R. A., Sims, A. C., Johnson, A. R., Sidebottom, A. M., Sedio, B. E., Klitgaard, A., Larson, C. B., Boya P, C. A., Torres-Mendoza, D., Gonzalez, D. J., Silva, D. B., Marques, L. M., Demarque, D. P., Pociute, E., O'Neill, E. C., Briand, E., Helfrich, E. J. N., Granatosky, E. A., Glukhov, E., Ryffel, F., Houson, H., Mohimani, H., Kharbush, J. J., Zeng, Y., Vorholt, J. A., Kurita, K. L., Charusanti, P., McPhail, K. L., Nielsen, K. F., Vuong, L., Elfeki, M., Traxler, M. F., Engene, N., Koyama, N., Vining, O. B., Baric, R., Silva, R. R., Mascuch, S. J., Tomasi, S., Jenkins, S., Macherla, V., Hoffman, T., Agarwal, V., Williams, P. G., Dai, J., Neupane, R., Gurr, J., Rodríguez, A. M. C., Lamsa, A., Zhang, C., Dorrestein, K., Duggan, B. M., Almaliti, J., Allard, P.-M., Phapale, P., Nothias, L.-F., Alexandrov, T., Litaudon, M., Wolfender, J.-L., Kyle, J. E., Metz, T. O., Peryea, T., Nguyen, D.-T., VanLeer, D., Shinn, P., Jadhav, A., Müller, R., Waters, K. M., Shi, W., Liu, X., Zhang, L., Knight, R., Jensen, P. R., Palsson, B. Ø., Pogliano, K., Linington, R. G., Gutiérrez, M., Lopes, N. P., Gerwick, W. H., Moore, B. S., Dorrestein, P. C. & Bandeira, N. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nature Biotechnology* **34,** 828–837 (2016).

18. Aron, A. T., Gentry, E. C., McPhail, K. L., Nothias, L. F., Nothias-Esposito, M., Bouslimani, A., Petras, D., Gauglitz, J. M., Sikora, N., Vargas, F., van der Hooft, J. J. J., Ernst, M., Kang, K. Bin, Aceves, C. M., Caraballo-Rodríguez, A. M., Koester, I., Weldon, K. C., Bertrand, S., Roullier, C., Sun, K., Tehan, R. M., Boya P, C. A., Christian, M. H., Gutiérrez, M., Ulloa, A. M., Tejeda Mora, J. A., Mojica-Flores, R., Lakey-Beitia, J., Vásquez-Chaves, V., Zhang, Y., Calderón, A. I., Tayler, N., Keyzers, R. A., Tugizimana, F., Ndlovu, N., Aksenov, A. A., Jarmusch, A. K., Schmid, R., Truman, A. W., Bandeira,

N., Wang, M. & Dorrestein, P. C. Reproducible molecular networking of untargeted mass spectrometry data using GNPS. *Nature Protocols* **15,** 1954–1991 (2020).

19. Mohimani, H., Liu, W. T., Mylne, J. S., Poth, A. G., Colgrave, M. L., Tran, D., Selsted, M. E., Dorrestein, P. C. & Pevzner, P. A. Cycloquest: Identification of cyclopeptides via database search of their mass spectra against genome databases. *Journal of Proteome Research* **10,** 4505–4512 (2011).

20. Ibrahim, A., Yang, L., Johnston, C., Liu, X., Ma, B. & Magarvey, N. A. Dereplicating nonribosomal peptides using an informatic search algorithm for natural products (iSNAP) discovery. *Proceedings of the National Academy of Sciences U.S.A.* **109,** 19196–19201 (2012).

21. Mohimani, H., Yang, Y. L., Liu, W. T., Hsieh, P. W., Dorrestein, P. C. & Pevzner, P. A. Sequencing cyclic peptides by multistage mass spectrometry. *Proteomics* **11,** 3642–3650 (2011).

22. Ng, J., Bandeira, N., Liu, W. T., Ghassemian, M., Simmons, T. L., Gerwick, W. H., Linington, R., Dorrestein, P. C. & Pevzner, P. a. Dereplication and de novo sequencing of nonribosomal peptides. *Nat Methods* **6,** 596–599 (2009).

23. Kavan, D., Kuzma, M., Lemr, K., Schug, K. A. & Havlicek, V. CYCLONE - A utility for de novo sequencing of microbial cyclic peptides. *Journal of the American Society for Mass Spectrometry* **24,** 1177–1184 (2013).

24. Townsend, C., Furukawa, A., Schwochert, J., Pye, C., Edmondson, Q. & Lokey, R. S. CycLS: Accurate, whole-librarysequencing of cyclic peptides using tandem mass spectrometry. *Bioorganic & Medicinal Chemistry* **26,** 1232–1238 (2018).

25. Kersten, R. D., Yang, Y.-L., Xu, Y., Cimermancic, P., Nam, S.-J., Fenical, W., Fischbach, M. A., Moore, B. S. & Dorrestein, P. C. A mass spectrometry--guided genome mining approach for natural product peptidogenomics. *Nature chemical biology* **7,** 794–802 (2011).

26. Behsaz, B., Mohimani, H., Gurevich, A., Prjibelski, A., Fisher, M., Vargas, F., Smarr, L., Dorrestein, P. C., Mylne, J. S. & Pevzner, P. A. De Novo Peptide Sequencing Reveals Many Cyclopeptides in the Human Gut and Other Environments. *Cell Systems* **10,** 99–108 (2020).

27. Newman, D. J. & Cragg, G. M. Natural Products as Sources of New Drugs from 1981 to 2014. *Journal of Natural Products* **79,** 629–661 (2016).

28. Li, J. W. H. & Vederas, J. C. Drug discovery and natural products: End of an era or an endless frontier? *Science* **325,** 161–165 (2009).

29. Ling, L. L., Schneider, T., Peoples, A. J., Spoering, A. L., Engels, I., Conlon, B. P., Mueller, A., Hughes, D. E., Epstein, S., Jones, M., Lazarides, L., Steadman, V. a, Cohen, D. R., Felix, C. R., Fetterman, K. A., Millett, W. P., Nitti, A. G., Zullo, A. M., Chen, C. &

Lewis, K. A new antibiotic kills pathogens without detectable resistance. *Nature* **517,** 455–459 (2015).

30. Wang, H., Fewer, D. P., Holm, L., Rouhiainen, L. & Sivonen, K. Atlas of nonribosomal peptide and polyketide biosynthetic pathways reveals common occurrence of nonmodular enzymes. *Proceedings of the National Academy of Sciences of the United States of America* **111,** 9259–9264 (2014).

31. Donia, M. S., Cimermancic, P., Schulze, C. J., Wieland Brown, L. C., Martin, J., Mitreva, M., Clardy, J., Linington, R. G. & Fischbach, M. A. A Systematic Analysis of Biosynthetic Gene Clusters in the Human Microbiome Reveals a Common Family of Antibiotics. *Cell* **158,** 1402–1414 (2014).

32. Zipperer, A., Konnerth, M. C., Laux, C., Berscheid, A., Janek, D., Weidenmaier, C., Burian, M., Schilling, N. A., Slavetinsky, C., Marschal, M., Willmann, M., Kalbacher, H., Schittek, B., Brötz-Oesterhelt, H., Grond, S., Peschel, A. & Krismer, B. Human commensals producing a novel antibiotic impair pathogen colonization. *Nature* **535,** 511–516 (2016).

33. Wilson, M. R., Jiang, Y., Villalta, P. W., Stornetta, A., Boudreau, P. D., Carrá, A., Brennan, C. A., Chun, E., Ngo, L., Samson, L. D., Engelward, B. P., Garrett, W. S., Balbo, S. & Balskus, E. P. The human gut bacterial genotoxin colibactin alkylates DNA. *Science* **363,** eaar7785 (2019).

34. Vizcaino, M. I. & Crawford, J. M. The colibactin warhead crosslinks DNA. *Nature Chemistry* **7,** 411–417 (2015).

35. Marahiel, M. A., Stachelhaus, T. & Mootz, H. D. Modular Peptide Synthetases Involved in Nonribosomal Peptide Synthesis. *Chemical Reviews* **97,** 2651–2674 (1997).

36. Süssmuth, R. D. & Mainz, A. Nonribosomal Peptide Synthesis—Principles and Prospects. *Angewandte Chemie - International Edition* **56,** 3770–3821 (2017).

37. Röttig, M., Medema, M. H., Blin, K., Weber, T., Rausch, C. & Kohlbacher, O. NRPSpredictor2 - A web server for predicting NRPS adenylation domain specificity. *Nucleic Acids Research* **39,** 362–367 (2011).

38. Medema, M. H., Blin, K., Cimermancic, P., de Jager, V., Zakrzewski, P., Fischbach, M. A., Weber, T., Takano, E. & Breitling, R. antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic acids research* **39,** 339–346 (2011).

39. Tietz, J. I., Schwalen, C. J., Patel, P. S., Maxson, T., Blair, P. M., Tai, H. C., Zakai, U. I. & Mitchell, D. A. A new genome-mining tool redefines the lasso peptide biosynthetic landscape. *Nature Chemical Biology* **13,** 470 (2017).

40. Skinnider, M. A., Merwin, N. J., Johnston, C. W. & Magarvey, N. A. PRISM 3: Expanded prediction of natural product chemical structures from microbial genomes. *Nucleic Acids*

Research **45,** W49–W54 (2017).

41. Johnston, C. W., Skinnider, M. A., Wyatt, M. A., Li, X., Ranieri, M. R. M., Yang, L., Zechel, D. L., Ma, B. & Magarvey, N. A. An automated Genomes-to-Natural Products platform (GNP) for the discovery of modular natural products. *Nature Communications* **6,** 1–11 (2015).

42. Helfrich, E. J. N., Vogel, C. M., Ueoka, R., Schäfer, M., Ryffel, F., Müller, D. B., Probst, S., Kreuzer, M., Piel, J. & Vorholt, J. A. Bipartite interactions, antibiotic production and biosynthetic potential of the Arabidopsis leaf microbiome. *Nature Microbiology* **3,** 909–919 (2018).

43. Mukherjee, S., Stamatis, D., Bertsch, J., Ovchinnikova, G., Katta, H. Y., Mojica, A., Chen, I. M. A., Kyrpides, N. C. & Reddy, T. B. K. Genomes OnLine database (GOLD) v.7: Updates and new features. *Nucleic Acids Research* **47,** D649–D659 (2019).

44. Medema, M. H. Computational Genomics of Specialized Metabolism: from Natural Product Discovery to Microbiome Ecology. *mSystems* **3,** e000182 (2018).

45. Mohimani, H., Liu, W.-T., Kersten, R. D., Moore, B. S., Dorrestein, P. C. & Pevzner, P. A. NRPquest: coupling mass spectrometry and genome mining for nonribosomal peptide discovery. *Journal of natural products* **77,** 1902–1909 (2014).

46. Meleshko, D., Mohimani, H., Tracanna, V., Hajirasouliha, I., Medema, M. H., Korobeynikov, A. & Pevzner, P. A. BiosyntheticSPAdes: reconstructing biosynthetic gene clusters from assembly graphs. *Genome Research* **29,** 1352–1362 (2019).

47. Kolmogorov, M., Bickhart, D. M., Behsaz, B., Gurevich, A., Rayko, M., Shin, S. B., Kuhn, K., Yuan, J., Polevikov, E., Smith, T. P. L. & others. metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nature Methods* 1–8 (2020).

CHAPTER 1.

De Novo Peptide Sequencing Reveals Many Cyclopeptides in the Human Gut and Other Environments

1.1. ABSTRACT

Cyclic and branch cyclic peptides (*cyclopeptides*) represent an important class of bioactive natural products that include many antibiotics and anti-tumor compounds. However, despite the recent advances in metabolomics analysis, still little is known about cyclopeptides in the human gut and their diversity while we are constantly exposed to them. To address this bottleneck, we developed the CycloNovo algorithm for automated *de novo* cyclopeptide analysis and sequencing that employs de Bruijn graphs, the workhorse of DNA sequencing algorithms. CycloNovo reconstructed many new cyclopeptides that were validated with transcriptome, metagenome, and genome mining analyses. Our benchmarking revealed a large hidden cyclopeptidome in the human gut and other environments and suggested that CycloNovo offers a much-needed step-change for cyclopeptide discovery. Furthermore, CycloNovo revealed a wealth of anti-microbial cyclopeptides from food that survive the complete human gastrointestinal tract, raising the question of how these cyclopeptides might affect the human microbiome.

1.2. INTRODUCTION

The golden age of antibiotics was followed by a decline in the pace of antibiotics discovery in the 1990s. However, antibiotics and other natural products are again at the center of attention as exemplified by the recent discovery of teixobactin[1]. A key prerequisite for the resurgence of natural product research is the development of computational discovery pipelines[2] such as the Global Natural Products Social (GNPS) molecular networking[3], Dereplicator[4], and VarQuest[5]. The GNPS project alone has already accumulated over one billion mass spectra, an untapped resource for discovery of new antibiotics. Currently, however, the GNPS network is mainly used for identifying previously discovered natural products and their analogs. Therefore, developing algorithms that are not bound by the current database of natural products is necessary to truly realize the promise of computational natural product discovery.

This study focuses on *de novo* analysis of *cyclopeptides*, which includes cyclic and branch-cyclic peptides (might include several branches), an important and large class of bioactive natural products with an unparalleled track record in pharmacology. Many antibiotics as well as anti-tumor agents, immunosuppressors, and toxins are cyclopeptides. The favorable properties of bioactive cyclopeptides such as high affinity and selectivity, has made them particularly interesting as drug candidates[6] among peptidic natural products. Despite this great interest, cyclopeptide sequencing and analysis from tandem mass spectra are extra challenging as the propensity of these molecules to break at all pairs of points in their cyclic backbone gives a far more complex series of ions than in linear peptides. Cyclopeptides are divided into cyclic *Ribosomally synthesized and Post-translationally modified Peptides* (*RiPPs*) and cyclic *Non-Ribosomal Peptides* (*NRPs*). RiPPs are encoded using the genetic code and are built from the 20 proteinogenic amino acids, which however are subjected to numerous post-translational modifications. NRPs are not directly inscribed in the genomes and cannot be inferred with traditional DNA sequencing. Instead, they are encoded using "nonribosomal code"[7] and are built from over 300 different naturally occurring amino acids.

The discovery of the cyclopeptide gramicidin S in 1942 (first antibiotic used for treating soldiers during the World War II) led to two Nobel prizes and has been followed by the discovery of ≈400 families of cyclopeptides (*cyclofamilies*) in the last 75 years[5]. A relatively small number of known cyclofamilies reflects the experimental and computational challenges in cyclopeptide discovery. The question of how many cyclofamilies stayed below the radar of previous studies (even though their spectra have already been deposited to public databases!) remains open. To answer this question, first we considered the problem of recognizing *cyclospectra* (tandem mass spectra that originated from cyclopeptides) in large spectral datasets (targeted and/or untargeted).

Bandeira et al.[8] introduced the concept of *spectral networks* that reveal the spectra of related peptides without knowing their amino acid sequences. Nodes in a spectral network correspond to spectra, while edges connect *spectral pairs*, i.e. spectra of peptides differing by a single modification or a mutation. Ideally, each connected component of a spectral network corresponds to a cyclofamily[9] representing a set of similar cyclopeptides. Although spectral networks of various GNPS datasets have become the workhorse of the cyclopeptide studies, they typically contain false-positive edges that render the analysis of cyclofamilies challenging[3]. Moreover, constructing the spectral network of all GNPS spectra remains an open algorithmic problem.

By recognizing cyclospectra first, one can construct a small *cyclospectral sub-network* of the entire input spectral dataset (for example the GNPS network) and to evaluate the number of cyclofamilies in the dataset as the number of connected components in this sub-network. Our analysis revealed that many cyclopeptides evaded detection in previous studies and that the known cyclopeptides represent the tip of the iceberg of cyclopeptides that are waiting to be decoded.

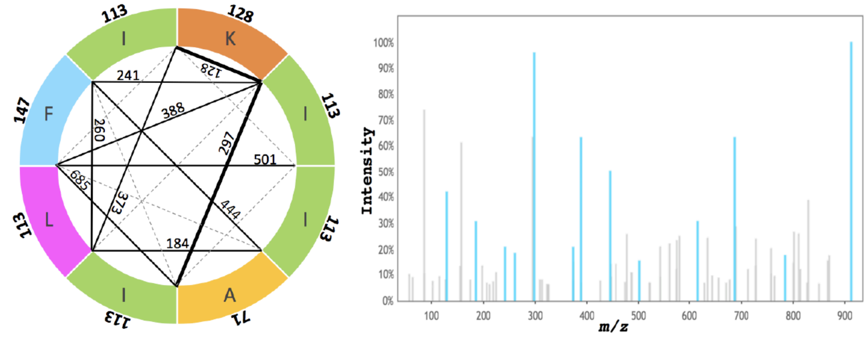
Recognition of cyclospectra, not only creates the possibility to get an estimate on the number of cyclofamilies for large-scale projects but also makes otherwise time-consuming downstream analyses feasible for such projects. For example, recognized cyclospectra can be matched against NRP and RiPP biosynthetic genes using various genome mining and peptidogenomics tools[10,11]. These tools typically generate a huge database of putative cyclopeptides mounting to possibly millions of putative peptides[10], making it prohibitively time-consuming to search large spectral datasets against such databases. This problem is especially aggravated in the case of NRPs where the traditional DNA sequencing approaches do not apply. Genome mining tools for NRP prediction such as NRPSpredictor2[12] combined with NRPquest[13], predict a database of putative NRPs based on the nonribosomal code. Because the nonribosomal code is not yet fully understood and is not as specific as the genetic code, these analyses often result in colossal databases of putative error-prone peptides predicted from a single gene cluster and therefore are limited in their power for peptidogenomics analysis[2,10]. Fast algorithms for recognizing cyclospectra are critical as they greatly reduce the set of spectra that need to be matched against databases of putative cyclopeptides.

In addition to recognizing cyclospectra, CycloNovo *de novo* sequences the recognized cyclospectra. We distinguish between *cyclopeptide identification* (identifying cyclopeptides by matching their spectra against databases of known cyclopeptides[14]) and *de novo cyclopeptide sequencing* (determining the cyclopeptide sequence from a spectrum alone). Although recent studies have made progress towards cyclopeptide identification[4,5,15,16], previous cyclopeptide sequencing algorithms[17–20] are not compatible with the large-scale nature of current metabolomics and mass spectrometry studies. These tools are rarely used[14,21] because they are inaccurate, are too slow for analyzing high-throughput spectral datasets, are limited to the proteinogenic amino acids, and also cannot distinguish cyclopeptides from other compounds. To this date, no novel cyclopeptide has been introduced through fully automated *de novo* sequencing of tandem mass spectra.

To close these gaps, here we present CycloNovo, an algorithm that performs fast cyclopeptide sequencing based on the concept of the *de Bruijn graph* of a spectrum, a compact representation of putative *k*-mers (strings formed by *k* consecutive amino acids) in an unknown cyclopeptide. Although de Bruijn graphs represent the workhorse of DNA sequencing[22], they have not previously been applied to cyclopeptide sequencing. We demonstrate that using this technique, CycloNovo enables high-throughput analysis of cyclopeptides in large spectral datasets, sequencing many cyclopeptides in diverse samples that include marine, soil, and human gut bacterial communities.

1.3. RESULTS

To illustrate how CycloNovo works, we used a spectrum of the cyclopeptide surugamide A[23] (referred to as surugamide hereon) with the amino acid sequence AIIKIFLI (Figure 1.1).



TheoreticalSpectrum (*Surugamide*)	71 113 **128** 147 **184** 226 **241** **260** **297** 354 **373** **388** 410 425 **444** 467 486 **501** 523 538 557 **614** 651 670 **685** 727 764 **783** 798 840 **911**
Spectrum$_{Surugamide}$	55 61 85 85 85 101 114 126 **128** 128 154 156 173 **184** 196 205 211 218 224 **241** **260** 294 294 **297** 297 308 312 321 323 325 373 **373** 373 **388** 388 425 **444** 445 445 455 473 477 486 486 **501** 520 521 541 541 559 572 573 578 578 **614** 627 629 633 642 654 669 675 **685** 687 712 726 727 756 756 763 **783** 799 800 809 812 825 828 849 866 869 **911**

Figure 1.1. Theoretical and experimental spectra of surugamide. (Top left) Diagram of the surugamide from a marine *Streptomyces* CNQ329 (mass 911.62 Da). Each color represents an amino acid (the numbers on the outer edge are the nominal masses of amino acids in Daltons). Each chord corresponds to a fragment of surugamide appearing between its start and end. The solid chords represent the fragments in *TheoreticalSpectrum*(*Surugamide*) whose masses match masses in the experimental spectrum *Spectrum*$_{Surugamide}$. The numbers on solid chords show the nominal masses of the corresponding fragment represented by that chord. For example, the chord labeled 297 corresponds to the fragment Ile-Ile-Ala of mass 297 Da. Each chord corresponds either to a single mass *x* or two masses *x* and *mass(Spectrum)-x* (in the latter case the chord is shown in bold). Given a set of fragments with the same mass, we show one of them (arbitrarily chosen) by a solid chord and the others by dashed chords. For example, one of two fragments with the same integer mass 241 (Ile-Lys and Lys-Ile in clockwise order), is shown by a solid chord and another by a dashed chord. (Top right) The experimental spectrum of surugamide (*Spectrum*$_{Surugamide}$) with 82 peaks (GNPS ID MSV000078839). The *y*-axis in the *Spectrum*$_{Surugamide}$ shows the ion intensities as the percentage of the intensity of the highest intensity peak. Blue peaks represent masses shared with *TheoreticalSpectrum*(*Surugamide*) for the error threshold $\varepsilon = 0.015$ Da. (Bottom) The theoretical and (pre-processed) experimental spectra for surugamide rounded to the nearest integer (this rounding results in the repetitive integers in the list). Masses in the pre-processed experimental spectrum were reduced by the mass of hydrogen $m_H \approx 1.0078$ Da. A mass in the theoretical spectrum is shared with a mass in the experimental spectrum if they were within the error threshold. The numbers in bold represent 13 shared masses.

Theoretical and experimental spectra. Given an amino acid string, its *mass* is defined as the sum of masses of its amino acids. Given a cyclopeptide *Peptide*, its *theoretical spectrum TheoreticalSpectrum*(*Peptide*) is the set of masses of all substrings of *Peptide* (Figure 1.1). For example, *TheoreticalSpectrum*(AGCD) contains masses of A, G, C, D, AG, GC, CD, DA, AGC, GCD, CDA, DAG, and AGCD. Note that if multiple fragments have the same mass, they contribute a single mass to the theoretical spectrum.

An *experimental spectrum* is a list of *peak*s, where each peak is characterized by its *intensity* and *m/z* (*m* and *z* represent the mass and the charge of the ion corresponding to the peak). For simplicity, we represent a pre-processed spectrum as an increasing sequence of numbers *Spectrum*=$\{s_1, \ldots, s_n\}$, assuming that all peaks in the spectrum have charge 1 and ignoring intensities. Similar to pre-processing practices in proteomics[24], CycloNovo filters out low-intensity peaks in each spectrum by retaining at most 5 peaks with the highest intensities in each 50 Da window. CycloNovo further filters out all peaks that are less than 0.05 Da apart from another peak with higher intensity. It further removes spectra with a small number of peaks (less than 20) and spectra with a small precursor mass (less than 500 Da). We subtract the mass of a hydrogen atom from all masses in the spectrum (for simplicity, we assume that each ion is protonated with a single proton).

We estimate the *PeptideMass* of the cyclopeptide that generated *Spectrum* based on the precursor mass and the charge of *Spectrum.* We define the *symmetric* version of *Spectrum* (denoted *Spectrum**) as a spectrum that, in addition to all masses in *Spectrum*, contains *PeptideMass-s* for each mass *s* in *Spectrum.*

Scoring Peptide-Spectrum Matches. A mass *s* in a (pre-processed) experimental spectrum *Spectrum* matches a mass *s'* in *TheoreticalSpectrum*(*Peptide*) if *s* is "equal" to *s'*. By "equal" we mean "approximately equal" with error below the *error threshold* ε (with default value ε=0.02 Da). The score between *Peptide* and *Spectrum* (denoted *score*(*Peptide, Spectrum*) is defined as the number of matches between masses in *Spectrum* and masses in *TheoreticalSpectrum*(*Peptide*). Although CycloNovo uses accurate masses, examples below use nominal masses for simplicity.

Figure 1.1 illustrates that *score*(*Surugamide*, $Spectrum_{Surugamide}$)=13. For a linear peptide *Peptide, score*(*Peptide, Spectrum*) is the number of matches between masses of all linear substrings of *Peptide* and all masses in *Spectrum.* For example, *score*(ILFIK, $Spectrum_{Surugamide}$)=7 because the theoretical spectrum of the linear peptide ILFIK has 7 shared masses with $Spectrum_{Surugamide}$ corresponding to 7 chords within the ILFIK segment in Figure 1.1. These chords correspond to the following substrings: K (nominal mass 128), IL (226), IK (241), LF (260), LFI (373), LFIK (501), and ILFIK (614).

Spectral convolution. The *convolution* of a spectrum is the set of all pairwise differences between its masses[18]. Given a mass *a*, the convolution of *Spectrum with offset a* (denoted *convolution*(*Spectrum, a*)) is defined as the number of masses in the convolution equal to *a* (with error up to ε). As shown by Ng et al.[18], the value *convolution*(*Spectrum, a*) is expected to be high if *a* is the mass of an amino acid in a cyclopeptide that gave rise to *Spectrum*. Thus, offsets with high convolutions reveal the masses of amino acids in an unknown cyclopeptide that gave rise to an experimental spectrum.

To account for measurement errors, we cluster the masses in the convolution using *single linkage clustering* by combining pairs of masses in a cluster if they are less than ε apart. We define the *cluster mass* as the median mass of its members, and *cluster multiplicity* as the number of elements in the cluster. We call a cluster *cyclopeptidic* if one of its elements is within ε of the mass of a selected amino acid. Since high-multiplicity clusters reveal amino acids in the unknown cyclopeptide that gave rise to an experimental spectrum, we use them to generate the set of putative amino acids in an unknown cyclopeptide[18].

CycloNovo outline. Given an experimental spectrum *Spectrum*, the Cyclopeptide Sequencing Problem refers to finding a cyclopeptide *Peptide* that maximizes *score*(*Peptide, Spectrum*). Figure 1.2 illustrates the CycloNovo pipeline for solving this problem:

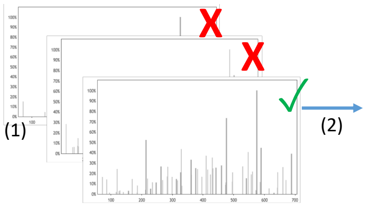
- ***Recognizing cyclospectra.*** Natural product researchers use *Marfey's analysis* for inferring the amino acid composition and configuration of an unknown peptide. However, since Marfey's analysis requires a purified peptide and has a number of limitations[25], we describe its *in silico* alternative for deriving an *approximate* amino acid composition of a cyclopeptide that gave rise to a given spectrum (see Methods section). If applying this approach reveals that a spectrum originated from a cyclopepeptide, we classify it as a cyclospectrum.
- ***Predicting amino acids in a cyclopeptide.*** For each cyclospectrum, CycloNovo predicts the set of putative amino acids in a cyclopeptide that gave rise to this spectrum. CycloNovo considers each cyclopeptidic cluster with multiplicity exceeding the *cyclopeptidic aa threshold* and classifies the selected amino acids corresponding to this cluster as a *putative amino acid* of the cyclopeptide that generated the cyclospectrum. Figure 1.2 illustrates that CycloNovo classifies amino acids **A, I/L, F, K,** T, W, R, and G as putative amino acids for $\mathit{Spectrum}_{\mathit{Surugamide}}$ (amino acids occurring in suragamide are shown in bold).
- ***Predicting amino acid composition of a cyclopeptide.*** For each cyclospectrum, CycloNovo uses dynamic programming to find all combinations of putative amino acids with total mass matching the precursor mass of the spectrum. We refer to each such

combination as a *putative composition,* which may include the same amino acid multiple times. We represent a composition by a sequence of its amino acids where each amino acid is superscripted by its multiplicity in the composition. For example, a composition **$A^1I/L^5K^1F^1$** includes eight amino acids: one A, five I/L, one K, and one F, and its total mass matches the precursor mass $mass(Spectrum_{Surugamide})$. Note that composition reveals the set of amino acids but provides no information about the order of amino acids in a cyclopeptide. Figure 1.2 illustrates that CycloNovo predicts the following putative compositions for $Spectrum_{Surugamide}$: **$A^1I/L^5K^1F^1$ ($71^1113^5128^1147^1$)**, $I/L^4F^1R^2$ ($113^4147^1156^2$), $A^2T^1K^4R^1$ ($71^2101^1128^4156^1$), and G^1T^1I/L^1K^5 ($57^1101^1113^1128^5$). The putative composition of surugamide is shown in bold.

- ***Predicting k-mers in a cyclopeptide.*** For each *Composition(Spectrum*), CycloNovo analyzes all linear *k*-mers formed by amino acids in this composition (the default value *k*=5) and scores them against *Spectrum** using linear scoring. It assumes that if a Peptide-Spectrum Match has a high score *score*(*Peptide,Spectrum*) (a condition that usually holds for well-fragmented spectra), then each linear *k*-mer in *Peptide* also has high score (for an appropriately chosen *k*). High-scoring *k*-mers (defined as *k*-mers with scores exceeding the *k-mer score threshold*) represent putative *k*-mers in an unknown cyclopeptide. For example, for *Composition*=$71^1113^5128^1147^1$, there exist 4^5=1024 5-mers and CycloNovo identifies 524 of them as high-scoring 5-mers. We refer to the set of high-scoring *k*-mers as $Kmers_{Composition,k}(Spectrum)$. Figure 1.2 illustrates that three out of six highest scoring 5-mers for $Spectrum_{Surugamide}$ are correct, i.e., represent 5-mers from surugamide. CycloNovo computes the *k-merScore*, the score of the highest-scoring *k*-mer.

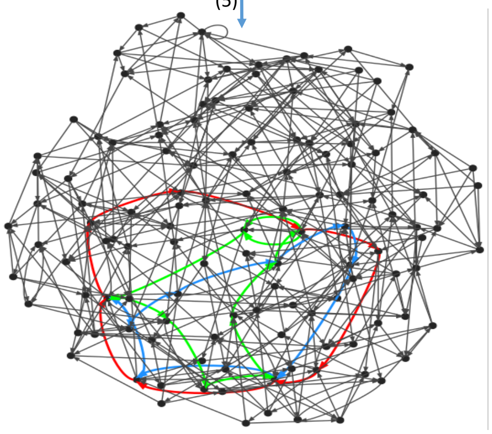
- ***Constructing the de Bruijn graph of a spectrum.*** Given a set *Kmers*= $Kmers_{Composition,k}(Spectrum)$, CycloNovo constructs the *de Bruijn graph* $DB_{Kmers}(Spectrum)$[22]. Nodes in $DB_{Kmers}(Spectrum)$ correspond to all (*k-1*)-mers from *Kmers* and each directed edge corresponds to a *k*-mer from *Kmers* and connects its first (*k-1*)-mer with its last (*k-1*)-mer. Each cycle in $DB_{Kmers}(Spectrum)$ spells out a cyclic amino acid sequence. Figure 1.2 presents the *pruned de Bruijn graph* for the putative composition $71^1113^5128^1147^1$ that is obtained by iterative removal of tips (nodes without outgoing or incoming edges), and single isolated edges from the de Bruijn graph. The composition $113^5128^171^1147^1$ results in a de Bruijn graph with 202 vertices and 524 edges and the pruned de Bruijn graph with 126 vertices and 392 edges (Figure 1.2).

- ***Generating cyclopeptide reconstructions.*** A cycle in the de Bruijn graph of a spectrum is *feasible* if it spells a cyclopeptide with the mass matching the precursor mass of the spectrum. Using the breadth-first search algorithm, CycloNovo finds all feasible cycles in the de Bruijn graph with length equal to the number of amino acids in *Composition* (a cycle may traverse the same edge multiple times). Each such cycle spells a putative cyclopeptide and CycloNovo scores each of them against *Spectrum.* Finally, it reports the highest scoring cyclopeptides along with the P-values of their Peptide-Spectrum Matches (PSMs) computed using MS-DPR[13]. The default P-value threshold (10^{-15}) is chosen based on the previous studies where the P-value cut-off 10^{-15} was necessary for reaching a False Discovery Rate (FDR) below 1% against *CyclopeptideDatabase*[4,5]. However, the user can change the *P*-value thresholds depending on their study. See See Methods section for a brief summary on P-value estimations in CycloNovo.



putative amino acids	113 (I/L)	128 (K)	186 (W)	71 (A)	147 (F)	156 (R)	57 (G)	101 (T)

putative compositions	top-scoring putative *k*-mers	#putative *k*-mers
$113^4156^2147^1$	(156 156 113 113 147) (156 113 147 113 113) (156 113 147 113 156) (113 113 147 156 113) (113 147 113 113 113) (113 147 113 156 113)	57
$128^471^2156^1101^1$	(101 128 128 128 156) (128 128 128 101 71) (128 128 101 128 71) (128 113 71 113 113) (128 128 128 128 71) (127 101 128 71 128)	76
$113^5128^171^1147^1$	(113 113 128 113 113) (113 147 113 128 113) (113 147 113 113 128) (113 113 71 113 113) (147 113 128 113 71) (128 113 71 113 113)	524
$128^5113^1101^157^1$	(57 128 113 128 128) (128 128 128 113 57) (101 128 113 128 128) (128 128 101 128 113) (128 113 128 101 57) (128 57 128 113 128)	180



sequence of masses of amino acids in a cyclopeptide								P-value
128	113	147	113	113	71	113	113	3.9×10^{-23}
128	113	113	147	113	113	71	113	3.9×10^{-23}
128	113	113	147	113	71	113	113	7.1×10^{-21}

Figure 1.2. CycloNovo outline illustrated using *Spectrum*$_{Surugamide}$. CycloNovo includes six steps: (1) recognizing cyclospectra in the entire spectral dataset, (2) predicting amino acids in a cyclopeptide from a recognized cyclospectrum, (3) predicting amino acid composition of a cyclopeptide by generating all combinations of predicted amino acids with total mass equal to the precursor mass of the spectrum, (4), predicting *k*-mers in a cyclopeptide, (5) constructing the de Bruijn graph of a spectrum, and (6) generating cyclopeptide reconstructions. Only six top-scoring putative *k*-mers for each putative amino acid composition are shown. Masses of amino acids occurring in surugamide are shown in red and *k*-mers occurring in surugamide are underlined. To simplify the de Bruijn graph (corresponding to the composition $71^1113^5128^1147^1$), all tips and isolated edges in the graph were removed. Red, blue and green feasible cycles in the graph spell out three cyclopeptides shown in the bottom table along with their P-values. The red cycle spells out surugamide.

In the case of *Spectrum*$_{Surugamide}$, CycloNovo found three similar cyclopeptides (Figure 1.2) spelled by feasible cycles in the de Bruijn graph with a putative composition $113^5128^171^1147^1$ (the highest-scoring one corresponds to surugamide). The remaining three putative compositions do not yield feasible cycles in their de Bruijn graphs. Figure 1.3 shows the pruned *de Bruijn* graphs of three compositions of *Spectrum*$_{Surugamide}$ that do not contain feasible cycles. CycloNovo sequenced *Spectrum*$_{Surugamide}$ in ≈3 seconds on a laptop with a single 2.5GHz processor.

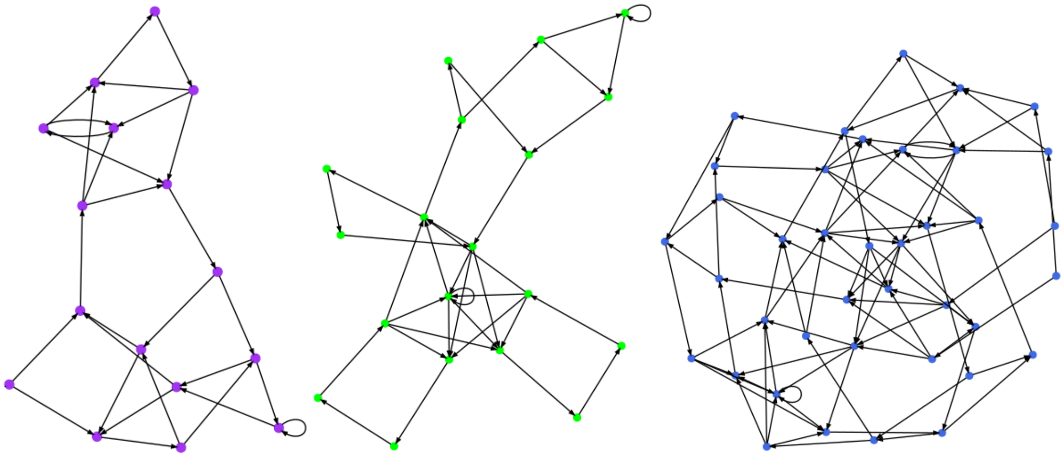


Figure 1.3. The pruned *de Bruijn* graphs of the compositions of *Spectrum*$_{Surugamide}$ that do not contain feasible cycles. (Left) The composition $113^4156^2147^1$ results in a de Bruijn graph with 40 vertices and 57 edges and a pruned de Bruijn graph with 18 vertices and 40 edges. (Middle) The composition $128^471^2156^1101^1$ results in a de Bruijn graph with 52 vertices and 76 edges and a pruned de Bruijn graph with 20 vertices and 42 edges. (Right) The composition $128^5113^1101^157^1$ results in a de Bruijn graph with 94 vertices and 180 edges and a pruned de Bruijn graph with 40 vertices and 92 edges.

Datasets. We analyzed various spectral datasets obtained from diverse bacterial communities. To benchmark CycloNovo, we also analyzed a plant spectral dataset that had a paired RNA-seq dataset, thus enabling us to validate the CycloNovo reconstructions by matching them against the transcriptome.

The S.VULGARIS dataset was generated from a single sample collected from seeds of the plant *Senecio vulgaris* (both medicinal and poisonous)[27] from the *Asteraceae* family. We also analyzed the RNA-Seq reads from the same sample (~74 million 100 bp long Illumina reads)[27], assembled them using rnaSPAdes[28], and used the assembled transcripts (61.9 Mb total length) and prior knowledge of cyclopeptide processing[29–31] to validate the reconstructed cyclopeptides.

The HUMANSTOOL dataset was generated from 65 stool samples of a single person (L.S., co-author of this paper and a contributor to the "Quantified self" initiative) collected over a course of four years. This dataset is accompanied by the detailed medical and food metadata[32] as well as metagenomics reads generated from the same samples (bioproject ID PRJEB24161).

The GNPS dataset[5] was formed by combining forty datasets from GNPS[3]. The $\text{GNPS}_{\text{CYANO}}$, $\text{GNPS}_{\text{PSEUDO}}$, and $\text{GNPS}_{\text{ACTI}}$ datasets represent sub-datasets of the GNPS dataset corresponding to three phyla with extensively analyzed cyclopeptides (*Cyanobacteria*, *Pseudomonas* and *Actinobacteria*).

The CYCLOLIBRARY dataset contains 81 spectra from 81 distinct cyclopeptides (forming 41 cyclofamilies) that were identified by Dereplicator[4] after searching the GNPS network against a chemical structure database of known cyclopeptides, *CyclopeptideDatabase*. To generate such databse we used the *PNPDatabase*[5] which combines all known peptidic natural products from various databases. Many peptides in this database are *lipopeptides* containing a lipid chain, e.g., surfactin is a cyclopeptide containing a fatty acid side chain connected to a fully peptidic part via a peptide bond. We classify a peptide in the *PNPdatabase* as a cyclopeptide if its backbone could be represented as a circular graph (cycle) with nodes corresponding to either a single amino acid or a single lipid tail (i.e. monomers) and edges corresponding to the amide bonds in the peptide structure. 1,257 out of 5,021 peptides in the *PNPDatabase* represent cyclopeptides and form *CyclopeptideDatabase* (note that the *CyclopeptideDatabase* database contains lipopeptides). We searched ~130 million GNPS spectra against the *CyclopeptideDatabase* using Dereplicator[4] and identified 81 distinct cyclopeptides (41 cyclofamilies) corresponding to PSMs with FDR=0% and P-value below 10^{-15}. For each identified cyclopeptide, we selected the PSM with the minimum P-value (among all PSMs identified for this cyclopeptide), resulting in a set of 81 PSMs and hence created a spectral dataset CYCLOLIBRARY with 81 spectra (Table 1.1). CYCLOLIBRARY includes only 13 cyclopeptides (6 cyclofamilies) that are made up entirely of selected amino acids (Table 1.1). 34 peptides (25 cyclofamilies) in the CYCLOLIBRARY dataset contain lipid tails and 34 peptides (14 cyclofamilies) contain non-selected amino acids.

Table 1.1. Cyclopeptides in the CYCLOLIBRARY dataset. The peptides that gave rise to 81 spectra in the CYCLOLIBRARY dataset with their corresponding peptide mass, P-value, the GNPS ID of the dataset a spectrum belongs to, *k-merScore,* and *cycloIntensity*. The column 'compound type' specifies whether the compound is fully peptidic or represents a lipopeptide. Column '#correct k-mers predicted with top 25 aa's/length of peptide' shows the total number of correct k-mers predicted using top 25 most frequent amino acids in *CyclopeptideDatabase* versus the total number of correct *k*-mers appearing in the cyclopeptide, i.e. length of cyclopeptide. Column '#unique correct monomers predicted using top 25 aa's / #unique monomers in correct sequence. The blue rows show the 13 cyclopeptides that are made up entirely of selected amino acids.

peptide ID	peptide mass	compound type	P-value	GNPS ID	*k-merScore*	*cycloIntensity*	#correct k-mers predicted with top 25 aa's / length of peptide
Antibiotic_FR_901459	609.9	peptide	1.8×10^{-24}	MSV000079098	10	0.59	0/11
Arthrofactin	1354.8	lipopeptide	1.2×10^{-16}	MSV000079772	8	0.96	7/12
Bacillomycin_D2	1031.5	lipopeptide	6.0×10^{-24}	MSV000078635	4	0.95	0/8
Bacillomycin_D3	1045.6	peptide	3.9×10^{-31}	MSV000079450	4	0.94	0/8
Bacillomycin_D5	1059.6	peptide	2.5×10^{-21}	MSV000078635	5	0.85	0/8
Bacillopeptin_B	1035.5	peptide	1.2×10^{-19}	MSV000079054	4	0.88	3/8
Bacillus_amyloliquefaciens_Surfactin_1	1036.7	lipopeptide	1.3×10^{-18}	MSV000080116	6	0.98	3/8
Bacillus_amyloliquefaciens_Surfactin_22	1022.7	lipopeptide	3.0×10^{-26}	MSV000078936	6	0.87	3/8
BK_10 _101A-form	1021.7	lipopeptide	2.5×10^{-22}	MSV000078688	6	0.78	2/8
BK_10 _101C	1035.7	lipopeptide	2.6×10^{-21}	MSV000078937	6	0.96	3/8
Champacyclin	898.6	peptide	5.8×10^{-26}	MSV000078936	6	0.98	8/8
Cyclolinopeptide_A	1040.7	peptide	4.0×10^{-31}	MSV000080050	6	0.93	9/9
Cyclolinopeptide_B	1058.6	peptide	1.8×10^{-29}	MSV000080050	8	0.87	9/9
Cyclolinopeptide_B_S-Oxide	1074.6	peptide	4.6×10^{-26}	MSV000080050	6	0.81	9/9
Cyclolinopeptide_D	1064.6	peptide	9.2×10^{-20}	MSV000080050	6	0.86	8/8
Cyclolinopeptide_E	977.6	peptide	2.6×10^{-26}	MSV000079777	4	0.73	8/8
Cyclolinopeptide_H	1082.5	peptide	5.1×10^{-20}	MSV000080050	6	0.80	8/8
Cyclosporin_B	1188.8	peptide	3.8×10^{-29}	MSV000079098	8	0.65	6/11
Cyclosporin_C	1218.8	peptide	1.4×10^{-32}	MSV000079581	6	0.81	0/11
Cyclosporin_E	1188.8	peptide	4.7×10^{-27}	MSV000079098	8	0.93	5/11
Cyclosporin_L	1188.7	peptide	1.7×10^{-30}	MSV000079098	8	0.81	5/11
Cyclosporin_P	1204.8	peptide	4.6×10^{-16}	MSV000079777	6	0.74	5/11
Cyclosporin_U	594.9	peptide	3.5×10^{-26}	MSV000079098	9	0.66	5/11
Cyclosporin_Y	601.9	peptide	2.0×10^{-24}	MSV000079098	10	0.45	0/11
Cyclosporin,_9CI_4	1188.8	peptide	1.5×10^{-27}	MSV000079098	8	0.61	0/11
Cyclosporin,_9CI_9	1202.8	peptide	1.8×10^{-43}	MSV000079098	8	0.95	5/11
Cyclosporin,_9CI_Deoxy	1186.9	peptide	1.6×10^{-35}	MSV000079098	8	0.85	5/11
Cyclosporin,_9CI_N9-De-Me	1188.8	peptide	3.8×10^{-32}	MSV000079098	9	0.74	5/11
[8'-Hydroxy-MeBmf]1-cyclosporin	1218.8	peptide	3.0×10^{-37}	MSV000079581	8	0.97	6/11
Daitocidin_B2	1064.7	lipopeptide	1.4×10^{-22}	MSV000078937	6	0.90	0/11

Table 1.1. Cyclopeptides in the CYCLOLIBRARY dataset. The peptides that gave rise to 81 spectra in the CYCLOLIBRARY dataset with their corresponding peptide mass, P-value, the GNPS ID of the dataset a spectrum belongs to, *k-merScore,* and *cycloIntensity*. The column 'compound type' specifies whether the compound is fully peptidic or represents a lipopeptide. Column '#correct k-mers predicted with top 25 aa's/length of peptide' shows the total number of correct k-mers predicted using top 25 most frequent amino acids in *CyclopeptideDatabase* versus the total number of correct *k*-mers appearing in the cyclopeptide, i.e. length of cyclopeptide. Column '#unique correct monomers predicted using top 25 aa's / #unique monomers in correct sequence. The blue rows show the 13 cyclopeptides that are made up entirely of selected amino acids, Continued.

peptide ID	peptide mass	compound type	P-value	GNPS ID	*k-merScore*	*cycloIntensity*	#correct k-mers predicted with top 25 aa's / length of peptide
Daitocidin_Pumilacidin_F	1050.7	lipopeptide	7.5×10^{-27}	MSV000078936	6	0.76	3/11
Dolastatin_1_11-N-Me	999.6	lipopeptide	9.6×10^{-18}	MSV000078568	3	0.58	3/8
Dolastatin_1_15-Epimer	1013.6	lipopeptide	3.5×10^{-16}	MSV000079050	0	0.18	3/8
Dolastatin_1_31	492.3	lipopeptide	1.4×10^{-19}	MSV000078568	4	0.42	0/9
Dolastatin_12	969.6	lipopeptide	4.2×10^{-16}	MSV000078568	5	0.65	0/9
Dolastatin_14_Dolastatin_14	1089.7	lipopeptide	1.5×10^{-20}	MSV000078568	7	0.47	0/9
g-Hydroxy-Meleu4-cyclosporin	609.9	peptide	3.6×10^{-26}	MSV000079581	9	0.56	3/9
Ilamycin_B1	1012.6	peptide	7.7×10^{-25}	MSV000078937	5	0.73	0/8
Ilamycin_B2	1028.6	peptide	1.6×10^{-19}	MSV000078936	3	0.67	0/7
Isocyclosporin_D	1216.9	peptide	5.0×10^{-27}	MSV000079098	7	0.40	0/7
Laxaphycin_A	1196.7	peptide	2.7×10^{-48}	MSV000079050	6	0.95	0/11
Laxaphycin_B	1395.9	peptide	3.9×10^{-33}	MSV000079050	5	0.97	2/11
Laxaphycin_B_32-Epimer,_53-deoxy	690.4	peptide	1.1×10^{-27}	MSV000079050	7	0.18	0/12
Laxaphycin_D	1367.8	peptide	7.1×10^{-28}	MSV000079050	3	0.86	0/12
Laxaphycin_E	1224.8	peptide	7.9×10^{-39}	MSV000079050	6	0.89	0/12
Lichenysin_A	1007.7	lipopeptide	1.2×10^{-20}	MSV000079481	3	0.95	1/11
Lichenysin-G1a	993.7	lipopeptide	1.8×10^{-19}	MSV000078936	5	0.70	0/8
Lichenysin-G3	1007.7	lipopeptide	8.1×10^{-22}	MSV000078936	6	0.93	0/8
Lichenysin-G5b	1021.7	lipopeptide	9.7×10^{-20}	MSV000078936	6	0.99	0/8
Lipodepsipeptides_KMM_A	1036.7	lipopeptide	9.2×10^{-25}	MSV000078635	5	0.98	2/8
Lipodepsipeptides_KMM_E	1064.7	lipopeptide	6.2×10^{-16}	MSV000078937	6	0.96	3/8

Table 1.1. Cyclopeptides in the CYCLOLIBRARY dataset. The peptides that gave rise to 81 spectra in the CYCLOLIBRARY dataset with their corresponding peptide mass, P-value, the GNPS ID of the dataset a spectrum belongs to, *k-merScore,* and *cycloIntensity*. The column ‘compound type’ specifies whether the compound is fully peptidic or represents a lipopeptide. Column ‘#correct k-mers predicted with top 25 aa’s/length of peptide’ shows the total number of correct k-mers predicted using top 25 most frequent amino acids in *CyclopeptideDatabase* versus the total number of correct *k*-mers appearing in the cyclopeptide, i.e. length of cyclopeptide. Column ‘#unique correct monomers predicted using top 25 aa’s / #unique monomers in correct sequence. The blue rows show the 13 cyclopeptides that are made up entirely of selected amino acids, Continued.

peptide ID	**peptide mass**	**compound type**	**P-value**	**GNPS ID**	***k-merScore***	***cycloIntensity***	**#correct k-mers predicted with top 25 aa’s / length of peptide**
Lipodepsipeptides_KMM_F	1078.8	lipopeptide	2.4×10^{-19}	MSV000078936	6	0.97	2/8
Lipopeptide_NO	994.6	lipopeptide	4.1×10^{-17}	MSV000078688	6	0.98	2/8
Majusculamide_C	985.6	lipopeptide	1.1×10^{-23}	MSV000078892	5	0.91	3/8
Majusculamide_C_Demethoxy	955.6	lipopeptide	7.3×10^{-17}	MSV000078568	6	0.95	3/9
Nocardiamide_A	687.5	peptide	8.0×10^{-24}	MSV000078936	6	0.97	4/9
NVA2-g-hydroxy-Meleu4-cyclosporin	1232.9	peptide	5.5×10^{-17}	MSV000079777	9	0.90	6/6
Peptidolipin_NA	964.7	lipopeptide	1.6×10^{-17}	MSV000078937	4	0.64	0/8
Pitipeptolide_E	794.5	peptide	9.3×10^{-19}	MSV000078568	5	0.90	0/7
Pitiprolamide	905.5	peptide	7.3×10^{-17}	MSV000078568	4	0.95	0/8
Precarriebowmide	865.5	lipopeptide	2.1×10^{-24}	MSV000079050	7	0.75	0/7
Precarriebowmide_S-Oxide	881.5	lipopeptide	9.3×10^{-21}	MSV000079050	6	0.70	0/7
Puwainaphycin_A	1235.7	peptide	7.1×10^{-26}	MSV000078982	4	0.97	0/10
Puwainaphycin_B	1233.7	peptide	6.4×10^{-34}	MSV000078982	5	0.91	0/10
Puwainaphycin_C	1227.7	peptide	7.9×10^{-28}	MSV000078982	4	0.91	0/10
Sch_378167_5'-Amide	569.3	peptide	3.5×10^{-31}	MSV000079098	7	0.61	0/10
SCH-378161	1123.6	peptide	2.5×10^{-27}	MSV000079098	6	0.99	0/10
Streptocidin_C	649.9	peptide	8.6×10^{-24}	MSV000079598	7	0.60	0/10
Surfactin_ A1	1008.7	lipopeptide	1.1×10^{-26}	MSV000078936	6	0.83	0/8
Surfactin_7-L-Valine_analogue	1022.7	lipopeptide	1.9×10^{-26}	MSV000078936	5	0.93	3/8
Surfactin_B1	1022.7	lipopeptide	5.2×10^{-23}	MSV000078937	3	0.68	3/8
Surfactin_C1	1036.7	lipopeptide	1.2×10^{-25}	MSV000078688	6	0.84	3/8
[Ile2,Val7]-Surfactin_C14i	1008.7	lipopeptide	2.3×10^{-17}	MSV000079450	4	0.82	0/8
[Val7]-Surfactin_C13ai	994.7	lipopeptide	1.9×10^{-23}	MSV000078936	6	0.80	0/8
Surfactin_D	1050.7	lipopeptide	6.8×10^{-24}	MSV000078937	6	0.96	3/8

Table 1.1. Cyclopeptides in the CYCLOLIBRARY dataset. The peptides that gave rise to 81 spectra in the CYCLOLIBRARY dataset with their corresponding peptide mass, P-value, the GNPS ID of the dataset a spectrum belongs to, *k-merScore,* and *cycloIntensity*. The column ‘compound type’ specifies whether the compound is fully peptidic or represents a lipopeptide. Column ‘#correct k-mers predicted with top 25 aa’s/length of peptide’ shows the total number of correct k-mers predicted using top 25 most frequent amino acids in *CyclopeptideDatabase* versus the total number of correct *k*-mers appearing in the cyclopeptide, i.e. length of cyclopeptide. Column ‘#unique correct monomers predicted using top 25 aa’s / #unique monomers in correct sequence. The blue rows show the 13 cyclopeptides that are made up entirely of selected amino acids, Continued.

peptide ID	**peptide mass**	**compound type**	**P-value**	**GNPS ID**	***k-merScore***	***cycloIntensity***	**#correct k-mers predicted with top 25 aa’s / length of peptide**
Surugamide_A	912.6	peptide	1.6×10^{-25}	MSV000078936	6	0.91	8/8
Surugamide_B	898.6	peptide	7.5×10^{-33}	MSV000079519	7	0.90	8/8
Surugamide_C	898.6	peptide	6.8×10^{-32}	MSV000079519	6	1.00	8/8
Surugamide_D	898.6	peptide	5.9×10^{-30}	MSV000078937	6	0.98	8/8
Viequeamide_B	808.5	lipopeptide	2.7×10^{-18}	MSV000078568	5	0.96	0/7
[Dihydro-MeBmt]1-[g-hydroxy-Meleu]4	1220.9	peptide	8.3×10^{-18}	MSV000079777	8	0.87	3/11

Information about the GNPS dataset. The GNPS dataset is formed by 40 MassIVE datasets that were selected from 120 datasets analyzed in Gurevich et al.[5] to exclude potentially miscalibrated (with respect to mass accuracy) spectral datasets. Since miscalibrated datasets typically do not result in any cyclopeptide identifications, we searched each of these 120 datasets with Dereplicator and excluded datasets that did not result in any identifications (with 0% FDR and P-value below 10^{-15}) from further analysis, leaving us with 40 datasets. The GNPS IDs are of these 40 MassIVE datasets are listed further below. Table 1.2 provides information about the various datasets analyzed by CycloNovo and provides a summary of CycloNovo results. Below we further describe the results for each dataset.

Table 1.2. Information about various high-resolution spectral datasets analyzed by CycloNovo. The number of distinct cyclopeptides and cyclofamilies was estimated using MS-Cluster[24] and SpecNets[3], respectively. The last column shows the number of known cyclopeptides/cyclofamilies (identified by Dereplicator) in each dataset. For each identified cyclopeptide in the CYCLOLIBRARY dataset, we selected the PSM with the minimum P-value (among all PSMs for that cyclopeptide), resulting in a spectral dataset CYCLOLIBRARY with 81 spectra.

dataset	**#spectra**	**#spectra after preprocessing**	**#cyclospectra**	**#distinct cyclopeptides/ cyclofamilies**	**#known cyclopeptides/ cyclofamilies**
CYCLOLIBRARY	81	81	45	45/27	45/27
S.VULGARIS	667	212	23	12/9	4/4
HUMANSTOOL	1,242,178	451,962	703	79/69	7/5
GNPS	51,220,679	27,883,895	12,004	512/213	67/37
$GNPS_{ACTI}$	5,903,921	4,435,893	1,478	116/56	38/24
$GNPS_{CYANO}$	23,582,408	12,118,482	317	74/35	5/4
$GNPS_{PSEUDO}$	697,812	581,012	2,076	120/39	5/2

Analyzing the CYCLOLIBRARY dataset. As the cyclopeptides that gave rise to the spectra in the CYCLOLIBRARY dataset are known, we used this dataset to benchmark CycloNovo. We considered a cyclopeptide/spectrum as correctly sequenced if the sequence of the cyclopeptide appeared among reconstructions with three highest-scores. CycloNovo recognized 45 spectra in the CYCLOLIBRARY dataset as cyclospectra and correctly sequenced 38 of these cyclospectra. For 22 cyclopeptides, the correct sequence of the cyclopeptide was among the highest-scoring reconstructions (Table 1.3). For each of the remaining 16 correctly sequenced cyclopeptides, Table 1.3 lists a high-scoring reconstruction and shows that for all those cases, at least one of the high-scoring reconstructions represented a rearranged sequence of amino acids compared to the correct sequence, except for a single cyclopeptide, where CycloNovo failed to predict correct amino acids for the top-scoring reconstruction.

Table 1.3. 38 cyclopeptides reconstructed by CycloNovo from 45 cyclospectra in the CYCLOLIBRARY dataset. The PSM score represents the score of the PSM in the CYCLOLIBRARY dataset. The 'max score' represents the score of the top-scoring reconstruction. For 22 cyclopeptides, the correct sequence of the cyclopeptide has the highest-scoring reconstruction. For the remaining 16 cyclopeptides, a highest-scoring reconstruction is listed below the correct sequence of the cyclopeptide in blue (differently arranged amino acid masses in the reconstructed cyclopeptide are shown in bold blue). Only in one case (cyclosporin C), CycloNovo predicted the wrong amino acids (shown in red) for the top-scoring reconstruction. Column 'rank of correct peptide' shows the rank of the score of the correct cyclopeptide, that the spectrum is generated, from among the scores of all reconstructions for that spectrum. CycloNovo failed to sequences 45-38=7 cyclospectra in the CYCLOLIBRARY dataset since it was not able to predict all their amino acids.

peptide ID	**sequence of aa masses in the peptide vs. sequence of aa masses in the highest-scoring reconstruction (if PSM score ≠ max score)**	**PSM score**	**max score**	**# reconstructions with score ≥ PSM score**	**rank of correct peptide**
BK 101C	113 113 115 99 113 113 128 240	21	21	6	1
nocardiamide A	113 113 99 99 99 163	19	19	1	1
cyclolinopeptide A	113 113 113 147 147 97 97 99 113	30	30	1	1
cyclolinopeptide B	113 99 147 147 97 97 113 113 131	24	24	1	1
bacillopeptin B	239 101 87 129 87 114 163 114	17	17	1	1
daitocidin_Pumilacidin F	254 99 113 115 113 113 113 129	24	24	4	1
BK 101A	113 113 115 99 113 113 128 226	19	19	6	1
cyclolinopeptide H	113 186 147 147 97 131 113 147	16	16	1	1
cyclosporin 9CI_Deoxy	167 113 127 127 71 71 127 99 127 71 85	29	29	6	1
cyclosporin B	183 113 127 127 71 71 127 99 127 71 71	30	30	4	1
laxaphycin A	57 113 113 113 113 147 101 113 83 101 141	44	44	1	1
surfactin 2	240 113 113 115 99 113 99 129	21	21	8	1
cyclolinopeptide D	113 186 147 147 97 113 113 147	20	20	2	1
cyclolinopeptide E	113 147 113 97 147 99 113 147	23	23	1	1
lipodepsipeptide KMM 1364A	240 99 113 115 113 113 113 129	20	20	8	1
Lipodepsipeptide KMM 1364E	268 99 113 115 113 113 113 129	20	20	2	1
cyclolinopeptide C	113 99 147 147 97 97 113 113 147	24	24	1	1
bacillomycin D2	97 114 163 114 225 101 87 129	22	22	1	1
bacillomycin D3	97 114 163 114 239 101 87 129	21	21	2	1
SCH-378161	113 57 97 147 114 143 99 97 113 142	29	29	2	1
[Val7]-Surfactin C13ai	99 113 113 129 212 99 113 115	21	21	3	1
lipopeptide_NO	99 113 113 129 198 113 113 115	17	17	15	1
cyclosporin,9CI 9	183 113 127 127 71 71 127 99 127 71 85 183 113 127 127 71 71 127 **113 99 85** 85	32	33	10	2
surfactin C1	240 113 113 115 99 113 113 129 240 113 113 **99 113 115** 113 129	20	21	15	2
cyclosporin C	183 113 127 127 71 71 127 99 127 71 101 183 113 127 **85** 71 113 **128** 71 99 99 **128**	33	34	49	2
surfactin 1	254 113 113 115 99 113 99 129 254 113 113 **99** 99 **129** 99 129	23	24	16	2
puwainaphycin_B	325 97 128 115 57 128 99 101 83 99 325 97 **99** 128 **57 115** 128 101 83 99	26	27	6	2

Table 1.3. 38 cyclopeptides reconstructed by CycloNovo from 45 cyclospectra in the CYCLOLIBRARY dataset. The PSM score represents the score of the PSM in the CYCLOLIBRARY dataset. The 'max score' represents the score of the top-scoring reconstruction. For 22 cyclopeptides, the correct sequence of the cyclopeptide has the highest-scoring reconstruction. For the remaining 16 cyclopeptides, a highest-scoring reconstruction is listed below the correct sequence of the cyclopeptide in blue (differently arranged amino acid masses in the reconstructed cyclopeptide are shown in bold blue). Only in one case (cyclosporin C), CycloNovo predicted the wrong amino acids (shown in red) for the top-scoring reconstruction. Column 'rank of correct peptide' shows the rank of the score of the correct cyclopeptide, that the spectrum is generated, from among the scores of all reconstructions for that spectrum. CycloNovo failed to sequences 45-38=7 cyclospectra in the CYCLOLIBRARY dataset since it was not able to predict all their amino acids, Continued.

surugamide A	128 113 113 71 113 113 147 113 128 113 113 71 113 113 **113 147**	23	24	5	2
surugamide B	128 99 113 71 113 113 147 113 128 99 113 71 113 **147 113** 113	26	27	5	2
surugamide D	128 113 99 71 113 113 147 113 128 113 **113** 71 **99** 113 147 113	28	30	7	2
lichenysin G5b	99 113 115 99 113 113 128 240 99 113 **99 115** 113 113 128 240	20	21	6	2
pitiprolamide	100 97 99 142 97 175 97 97 100 97 **142 99** 97 175 97 97	17	18	6	2
surfactin 7-L-Valine	240 99 113 115 99 113 113 129 240 99 **99 129** 99 113 113 129	22	24	15	3
surfactin D	254 113 113 115 99 113 113 129 254 113 113 113 **115 113 99** 129	22	25	14	3
majusculamide C Demethoxy	57 114 113 71 141 161 113 57 127 57 **113 114** 71 **161 141** 113 57 127	20	22	69	3
cyclosporin E	183 99 127 127 71 71 127 99 127 71 85 183 99 127 127 **99** 71 71 127 127 71 85	24	26	53	3
champacyclin	128 99 113 147 113 113 71 113 128 99 **71** 113 147 113 113 113	21	23	21	3
surugamide C	128 113 113 71 113 113 147 99 128 113 113 71 113 **99 147 113**	29	31	10	3

CycloNovo recognized 45 out of 81 spectra in the CYCLOLIBRARY dataset as cyclospectra. It classified 12 out of 13 cyclopeptides built from selected amino acids as cyclospectra and *de novo* sequenced them with one of the top three highest scores.

CycloNovo is unable to sequence most spectra in the CYCLOLIBRARY dataset since 68 of them originated from lipopeptides or peptides containing non-selected amino acids. To evaluate how CycloNovo performs on 45 cyclospectra in this dataset, we extended the set of selected amino acids to include the mass of the lipid chain and/or the masses of non-selected amino acids for each spectrum. Using this admittedly imperfect benchmarking approach, CycloNovo sequenced 22 of

45 cyclospectra as a highest-scoring *de novo* reconstructions and an additional 16 spectra with one of the three highest scores. Table 1.4 lists the highest-scoring reconstruction for these spectra and illustrates that the highest-scoring reconstruction is similar to the correct amino acid sequence for all these spectra.

Analyzing the S.VULGARIS dataset. The 23 recognized cyclospectra in this dataset correspond to twelve distinct cyclopeptides. CycloNovo sequenced ten of them with P-values below 10^{-15} (Table 1.4). Nine of ten reconstructed cyclopeptides matched the assembled transcriptome. One reconstructed cyclopeptide (with the highest-scoring reconstruction AFLLADV and score 22), did not match the assembled transcriptome but a suboptimal ALFLGLD reconstruction with score 20 did (see Note "Cyclopeptide-encoding transcripts in the S.VULGARIS dataset").

Table 1.4. Cyclopeptides reconstructed in the S.VULGARIS dataset. Ten reconstructed cyclopeptides (highlighted in yellow) along with their flanking sequences in transcripts translated into amino acids. For each of these cyclopeptides (reconstructed with P-values below 10^{-15}), we selected one representative spectrum with the highest score. The conserved flanking amino acids in the transcripts on the left and right sides of the highlighted cyclopeptides (preceding and succeeding motifs) are shown in red and green, respectively. For nine out of ten cyclopeptides, the reconstruction with the highest score matches one of the transcripts. For the cyclopeptide with mass 730.41 (highlighted in pink), the highest scoring reconstruction AFLLADV (score 22), did not match the assembled transcriptome but a suboptimal ALFLGLD reconstruction (score 20) did. The novel cyclopeptides discovered by CycloNovo are shown with bold IDs and named PLP-47 through PLP-52. For this dataset we used the error threshold ε=0.015 Da as recommended in Fisher *et al* [29].

precursor mass	sequence matching transcripts	PSM score	highest score	#reconstructions with score ≥ PSM score	P-value	peptide ID	gene
899.36	DNFVDTTGYDRLSDN	24	24	1	1.4×10^{-47}	PLP-14	*Sv_PawL1c*
811.37	DNFVGGTSFDRLSDN	14	14	2	2.4×10^{-24}	PLP-12	*Sv_PawL1c*
803.42	DNTFGVVIADRLSEN	30	30	1	1.2×10^{-61}	PLP-13	*Sv_PawL1b*
762.32	DNGFHGTFDGLDN	13	13	1	3.2×10^{-23}	**PLP-47**	*Sv_PawL1e*
730.41	DNALFLGLDGLDN	20	22	12	2.2×10^{-39}	**PLP-48**	*Sv_PawL1f*
702.38	DNAIFGVVDGLDN	20	20	1	5.6×10^{-36}	**PLP-49**	*Sv_PawL1j*
688.36	DNFVGGVIDGLDN	21	21	1	1.0×10^{-40}	**PLP-50**	*Sv_PawL1g*
674.35	DNGVVVGFDGLDN	14	14	5	1.1×10^{-25}	**PLP-51**	*Sv_PawL1l*
668.40	DNALVVGLDGLDN	14	14	1	1.9×10^{-27}	PLP-15	*Sv_PawL1d* *Sv_PawL1g*
654.39	DNALLGIADGLDN	18	18	5	6.9×10^{-34}	**PLP-52**	Sv_PawL1i

The ten reconstructed cyclopeptides (nine highest-scoring reconstruction and one suboptimal reconstruction) matched 11 transcripts (some transcripts encode multiple cyclopeptides and some cyclopeptides are encoded by multiple transcripts) that belong to cyclopeptide-encoding *PawS1-Like* genes in various *Asteraceae* species[27,29]. While three out of 11 identified *PawL1* ORFs and the four cyclopeptides encoded by them (PLP-12 through PLP-15) have been extensively analyzed in recent studies[27,29], the remaining eight ORFs represented previously unknown cyclopeptide-encoding genes in *S. vulgaris*. Table 1.5 lists ORFs (translated into amino acid sequences) in the orbitide-encoding transcripts. The PawL1 proteins have dual fates; they encode an albumin as well as a cyclopeptide(s). An enzyme asparaginyl endopeptidase (targets Asp, Asn) matures both the albumin and the cyclopeptide.

Table 1.5. ORFs in the cyclopeptide-encoding transcripts. All identified ORFs originate from various *PawS1-Like* genes. The sequences are color-coded based on the subunits they belong to: endoplasmic reticulum signal sequence (pink), the reconstructed cyclopeptide (blue), 2S albumin small subunit (lime green), and 2S albumin large subunit (orange). While the first three sequences (*Sv_PawL1b*, *Sv_PawL1c*, and *Sv_PawL1d*) are known *PawS1-Like* genes in *S. vulgaris*, the other eight sequences (named *Sv_PawL1e* through *Sv_PawL1l*) are novel *PawS1-Like* genes that were identified by searching for novel cyclopeptides.

gene	ORF sequence
Sv_PawL1b	AKLIVVVFAFAVIVAFAEVSAYKTTITTTTVEDNFVGGTSFDRLSENFMYGTPVDRLSDNRGSQKQCHRQIP
Sv_PawL1c	AKLIVVVFAFAVIVAFAEVSAYKTTITTTTVEDNTFGVVIADRLSDNFVDTTGYDRLSDNRGSQKQCHRQIP
Sv_PawL1d	ITTVEDNALVVGLDGLDNPITTTVEDNYFAGLIDGLDNPITTTVEDNGVFLGLDGLDNPSGSTYQCRRQIQGQQLNHCQMHIIQQGRSLVE
Sv_PawL1e	FVAIVAFSEQVSAYKTTIPTTVEDNALLVALDGLDNGFHGTFDGLDNGFHGTFDGLDNPSGSTYQCRRQIQ*
Sv_PawL1f	TTVEDNALFLGLDGLDNPSGSTYQCRRQIQGQQLNHCQMHITQQGRSLMENPRQQQLLQMCCNQLRQVEEECQCE*
Sv_PawL1g	ITTTVEDNALVVGLDGLDNPITTTVEDNFVGGVIDGLDNFVGGVIDGLDNPSGSTYKCRRQIQGQQLNHCQMHITQQGRSLVE
Sv_PawL1h	MTKVSAIVVLAFVAIVAFSEQVSAYKTTITTPVEDNAIFLGVDGLDNPI*
Sv_PawL1i	LDGLDNALLGIADGLDNPSGSTYQCRMQIQGQQLNHCQMHIIQQGRSLVENPRQQQQLQMCCNQLR*
Sv_PawL1j	SEQVSAYKTTITTTVEDNAIFGVVDGLDNPSGSTYQCRKQIQGQQ*
Sv_PawL1k	AIVAFSEQVSAYKTTITTTVEDNAIFLGVDGLDNPITTTVEDNGVSDFFDDGLDKPSGSTYQCRRQIQGQQLNHCQMHISQQGRSLVENPRQQQQLQM*
Sv_PawL1l	FVAIVAFSEQVSAYKTTITTPVEDNGVVVGFDGLDNPSGSTYQCRKQIQGQQ*

Figure 1.4 shows cyclospectra in the S.VULGARIS dataset, annotated using their CycloNovo reconstructions.

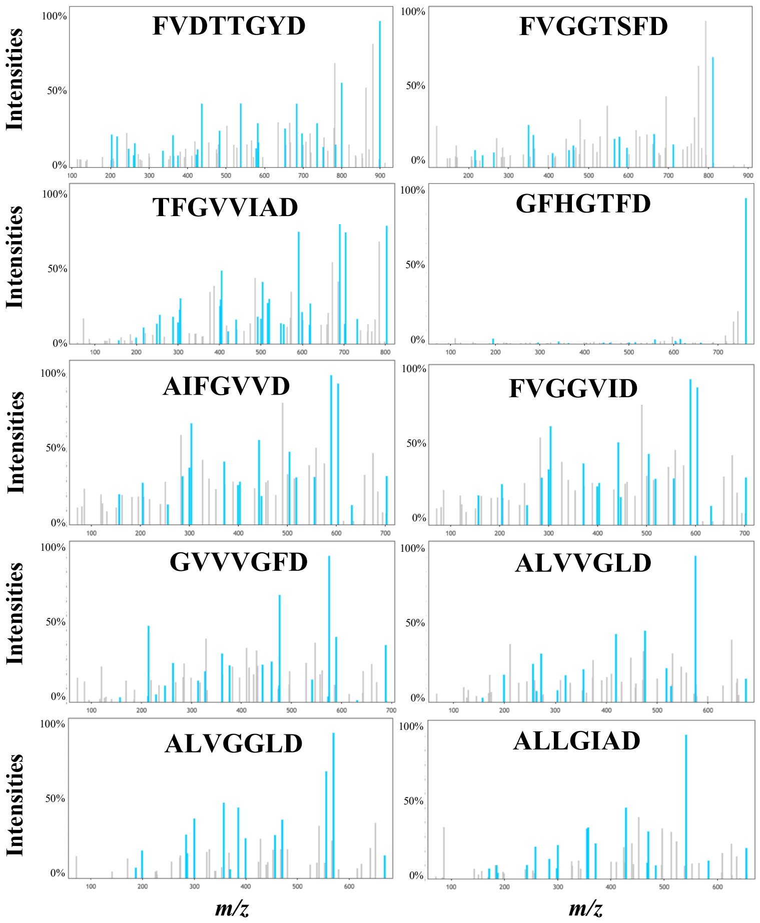


Figure 1.4. Annotated cyclospectra of the ten reconstructed cyclopeptides in the S.VULGARIS dataset. The *x*-axis shows the *m/z* ratios and the *y*-axis shows the percentage of the peak intensity compared to the intensity of the largest peak in that spectrum.

Analysing the HUMANSTOOL dataste. A Dereplicator search of the HUMANSTOOL dataset against *CyclopeptideDatabase* identified seven PSMs at 0% FDR, namely an antimicrobial orbitide citrusin V found in various *Citrus* species[33,34] and cyclolinopeptides A[35], B[36], C[37], D[37], H[37], and E[37]. Cyclolinopeptides are bioactive flaxseed orbitides from *Linum usitatissimum.* The individual who provided the HUMANSTOOL sample frequently ate flaxseeds because they contain α-linolenic acids. CycloNovo sequenced six flaxseed cyclopeptides from the *CyclopeptideDatabase* as well cyclolinopeptide P (a recently discovered cyclopeptide[38] that has not been added to *CyclopeptideDatabase* yet) as the highest-scoring reconstructions. Table 1.6

lists cyclopeptides identified by Dereplicator in the HUMANSTOOL datasets. Table 1.7 lists CycloNovo-reconstructions of 31 cyclopeptides in the HUMANSTOOL dataset.

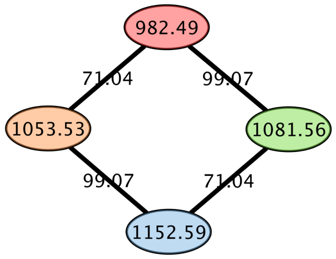
Table 1.6. Cyclopeptides identified by Dereplicator in the HUMANSTOOL dataset. The correct sequence of all reconstructed cyclopeptides has the highest score among all reconstructions. For each cyclopeptide, the score of the correct cyclopeptide (column "PSM score"), the number of reconstructions with scores larger or equal to the PSM score (column "#reconstructions score ≥ PSM score"), and P-values are listed.

precursor mass	peptide	PSM score	#reconstructions with score ≥ PSM score	P-value	peptide ID
1040.66	ILVPPPFFLI	31	1	1.2×10^{-54}	cyclolinopeptide A
1058.61	MLIPPFFVI	24	1	2.3×10^{-42}	cyclolinopeptide B
1074.62	M^{+16}LIPPFFVI	16	1	9.7×10^{-18}	cyclolinopeptide C
1064.57	M^{+16}LLPFFWI	20	2	1.5×10^{-33}	cyclolinopeptide D
1082.52	M^{+16}LMPFFWI	19	1	1.2×10^{-31}	cyclolinopeptide H
977.56	M^{+16}LVFPLFI	25	1	1.6×10^{-43}	cyclolinopeptide E
961.55	MLVFPLFI	25	10	3.1×10^{-42}	cyclolinopeptide P
567.36	GIVIPS	11	1	2.1×10^{-17}	citrusin V

Table 1.7. *De novo* reconstructions of 32 cyclopeptides in the HUMANSTOOL dataset. For each spectrum, its precursor mass, the *de novo* reconstruction (shown as a sequence of nominal masses of amino acids), the score, and the P-value are shown. *De novo* reconstructions are ordered in the decreasing order of their precursor masses. Only reconstructions with score≥ 13 are shown**.** Precursor masses of spectra identified by Dereplicator are highlighted in blue. Cyclopeptides highlighted in green represent a novel cyclofamily described in Figure 1.5.

peptide mass	precursor mass	sequence of amino acid masses	score	#reconstructions with score ≥ PSM score	P-value
1151.52	1152.53	87 99 99 101 147 129 71 97 113 71 137	19	1	2.3×10^{-27}
1098.49	549.75	147 71 87 137 57 99 71 129 163 137	23	1	6.6×10^{-24}
1085.53	543.27	99 99 137 57 113 115 186 71 137 71	20	2	6.2×10^{-30}
1081.51	1082.52	147 113 131 97 147 147 186 113	19	1	1.2×10^{-31}
1080.55	1081.56	87 99 99 101 147 129 71 97 113 137	22	1	3.0×10^{-36}
1073.61	1074.62	147 113 113 97 97 147 147 99 113	16	1	9.7×10^{-18}
1063.56	1064.57	147 113 113 97 147 147 186 113	20	2	1.5×10^{-33}
1060.75	530.74	147 163 57 99 57 137 71 129 71 129	22	1	2.7×10^{-29}
1057.61	1058.62	131 113 113 97 97 147 147 99 113	24	1	2.3×10^{-42}
1052.52	1053.53	87 99 101 147 129 71 97 113 71 137	25	1	8.6×10^{-38}
1039.65	1040.66	113 113 99 97 97 147 147 113 113	31	1	1.2×10^{-54}
1003.54	1004.55	71 97 147 99 147 97 147 99 99	13	1	4.0×10^{-17}
981.493	982.50	101 147 129 71 97 137 113 186	24	1	2.6×10^{-37}
978.543	979.55	128 113 147 87 113 57 99 87 147	18	3	6.1×10^{-25}
976.553	977.56	147 113 99 147 97 113 147 113	25	1	1.6×10^{-43}
960.553	961.56	131 113 99 147 97 113 147 113	25	10	3.1×10^{-42}
948.493	949.50	147 128 99 87 71 147 57 99 113	17	3	2.3×10^{-23}
891.463	892.47	113 147 101 71 57 57 99 99 147	15	1	3.4×10^{-19}
889.453	890.46	128 101 103 71 113 147 129 97	14	4	2.0×10^{-20}
888.493	889.50	57 97 147 113 113 99 99 163	15	9	9.7×10^{-21}
877.453	878.46	113 99 57 71 147 163 113 57 57	20	11	1.1×10^{-27}
873.403	874.41	71 97 115 71 97 101 87 97 137	17	2	2.7×10^{-23}
872.453	873.46	57 71 87 113 71 99 186 101 87	18	1	1.7×10^{-24}
871.453	872.46	147 101 71 57 99 114 97 57 128	19	1	2.5×10^{-26}
856.463	857.47	99 99 147 97 147 71 97 99	21	2	3.2×10^{-37}
841.433	842.44	97 147 101 57 97 113 128 101	20	1	2.2×10^{-29}
829.433	830.44	57 113 57 97 71 147 101 99 87	13	1	3.7×10^{-17}
826.393	827.40	147 99 137 71 71 99 115 87	17	3	5.8×10^{-23}
812.493	813.50	87 99 57 99 99 71 128 71 101	15	1	6.6×10^{-19}
811.413	812.42	147 113 57 57 57 97 99 97 87	16	7	2.0×10^{-20}
801.383	802.39	97 87 57 129 99 57 71 57 147	17	1	1.9×10^{-23}
695.273	696.28	71 57 163 129 57 71 147	18	1	6.1×10^{-27}

In addition to the eight reconstructed orbitides, CycloNovo reconstructed 32 cyclopeptides in the HUMANSTOOL dataset with P-values below 10^{-15} forming 26 cyclofamilies. Figure 1.5 shows a connected component in the spectral network formed by four novel cyclopeptides in the HUMANSTOOL dataset and illustrates that CycloNovo reconstructions are consistent with the spectral network.



precursor mass	peptide	PSM score	#reconstructions with score ≥ PSM score	P-value	dates
982.49	SVTFEAPLH	24	1	2.6×10^{-37}	07.14.2014 07.19.2015
1053.53	SVTFEAPLAH	25	1	8.6×10^{-38}	07.14.2014 07.19.2015
1081.56	SVVTFEAPLH	21	1	3.0×10^{-36}	07.14.2014 07.19.2015
1152.59	SVVTFEAPLAH	19	1	2.3×10^{-27}	07.14.2014

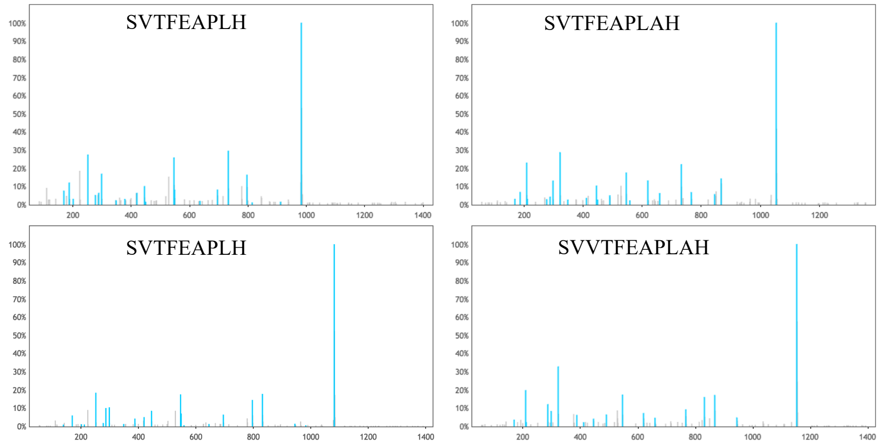


Figure 1.5. A novel cyclofamily reconstructed by CycloNovo in the HUMANSTOOL dataset. (Top) Four cyclopeptides reconstructed by CycloNovo form a cyclofamily represented by a connected component in the spectral network of the HUMANSTOOL dataset (label "L" stands for one of amino acids L and I). Each node represents a spectrum and two nodes are connected by an edge if their spectral similarity[3] exceeds 0.8. The numbers on the edges show the mass shifts between the corresponding spectra. (Middle) The *de novo* reconstructions corresponding to the four spectra forming the spectral network. For each cyclopeptide, the cyclic sequence of the highest-scoring reconstruction along with their scores, the number of reconstructions with scores larger or equal to the PSM score (column "#reconstructions with score ≥ PSM score"), and P-values are listed. The "dates" column shows the dates when the corresponding samples were taken. Note that the cyclopeptides in this cyclofamily appear on the same dates. (Bottom) The annotated spectra of the four cyclopeptides based on the CycloNovo reconstructions.

The Dereplicator search of all 703 cyclospectra in the HUMANSTOOL dataset against *CyclopeptideDatabase* resulted in a single hit and identified a cyclic lipopeptide massetolide F[39] with P-value 7.5×10^{-22}. As this compound includes lipid chains not included in the set of selected amino acids, CycloNovo was not able to generate its full-length reconstructions, but correctly reconstructed its partial amino acid sequence.

We classify a peptide as branch-cyclic if its backbone includes a cycle (with all monomers connected via amide bonds) and a side chain that includes at least one additional amide bond not included in the cycle. Although CycloNovo classify spectra of some branch cyclic peptides as cyclospectra, it is unable to *de novo* sequence them. Nevertheless, CycloNovo provides information about substrings of branch-cyclic peptides made of selected amino acids. For example, CycloNovo classified the spectrum of massetolide F in the HUMANSTOOL dataset as a cyclospectrum. The lipopetide massetolide F consists of the cycle TILSLSLV and a branch EL (along with a fatty acid chain tail with nominal mass 171 Da) connected to the cycle via an amide bond between T and E. We represent this branch cyclic peptides as a concatenate between the sequence of nominal masses of the cyclic and branch region separated by "*" sign, i.e., massetolide F is represented as 100, 113, 87, 113, 87, 113, 99 * 129, 113, 171. CycloNovo found five selected amino acids in massetolide F (S, I, L, V, T, and E) and missed the lipid chain (171 Da). Massetolides are non-ribosomal lipopeptides produced by *Pseudomonas fluorescens*, an indigenous member of human and plant microbiota[40,41].

Analysis of the metagenome assembly of reads paired with the HUMANSTOOL dataset confirmed that *P. fluorescens* is present in the stool samples where massetolide F was detected. Therefore, massetolide F might be originated from *P. fluorescens* in the human gut but further investigation is necessary to test this hypothesis. We used metaSPAdes[42] to assemble the metagenomic dataset, generated from the stool sample (dated by 6/16/2014) where massetolide F was detected. This dataset includes 34.5 million paired reads which are assembled into 81 thousand scaffolds of lengths longer than 500 bp amounting to 407 Mb total assembly length.

Analyzing the GNPS dataset. We analyzed all cyclospectra in the GNPS dataset using MS-Cluster[24] and SpecNets[43] with the goal of estimating the number of still unknown cyclopeptides and cyclofamilies originating from spectra already deposited into GNPS. To provide a conservative estimate for the number of cyclopeptides and cyclofamilies, we limited the analysis to clusters with at least three spectra. Table 1.8 lists the preditected numbers for each GNPS sub-datasets.

Table 1.8. **Information about the GNPS dataset.** The last row shows the total number of spectra and unique cyclopeptides/cyclofamiles across all datasets. The datasets marked in red, blue, and green form GNPS$_{\text{CYANO}}$, GNPS$_{\text{PSEUDO}}$, and GNPS$_{\text{ACTI}}$ subsets of the GNPS dataset, respectively.

GNPS ID	#spectra	#spectra after pre-processing	#cyclo spectra	#putative cyclo-peptides/ cyclo-families found by CycloNovo	#identified cyclo-peptides/ cyclo-families identified by Dereplicator (among cyclo-spectra)	#identified cyclo-peptides/ cyclo-families identified by Dereplicator (among all spectra)	#identified branch-cyclic . peptides/ branch-cyclic families identified by Dereplicator (among cyclo-spectra)
MSV000078567	730582	316993	4	2/1	2/1	4/2	0/0
MSV000078568	23582408	12118472	317	74/35	9/8	15/10	1/1
MSV000078584	680906	263160	0	0/0	0/0	0/0	0/0
MSV000078604	311617	281617	606	56/25	6/3	6/3	3/3
MSV000078606	289170	237988	122	32/12	1/1	1/1	7/6
MSV000078635	680168	569316	2388	124/40	9/4	12/5	10/7
MSV000078656	2844	1023	88	14/1	10/5	11/6	3/3
MSV000078710	1469076	689912	6	1/1	2/2	3/3	0/0
MSV000078787	1767830	1281235	208	58/31	25/13	25/13	8/7
MSV000078839	717600	504350	1	1/1	1/1	1/1	0/0
MSV000078847	167917	115603	19	7/5	1/1	1/1	1/1
MSV000078892	847114	461769	27	8/4	3/2	4/3	0/0
MSV000078936	2059306	1538683	526	58/30	25/13	30/15	5/5
MSV000078937	1694918	1303349	256	52/26	26/14	33/19	11/9
MSV000078982	984	727	32	4/2	3/1	3/1	2/2
MSV000079044	576282	270860	2	1/1	1/1	2/1	0/0
MSV000079050	1241328	683124	207	24/8	3/1	7/3	3/3
MSV000079054	702020	364382	112	16/7	13/6	13/6	3/3
MSV000079069	847145	215229	1066	23/2	0/0	1/1	0/0
MSV000079140	607488	443147	1118	25/7	14/6	15/7	0/0
MSV000079274	5433248	3457806	14	2/2	1/1	1/1	0/0
MSV000079312	54806	25354	1112	15/3	0/0	1/1	0/0
MSV000079450	697812	581012	2245	120/39	6/2	6/2	9/6
MSV000079471	22379	16138	68	14/4	10/4	10/4	0/0
MSV000079481	45742	7692	48	2/1	0/0	2/2	0/0
MSV000079502	47450	3167	14	3/1	3/1	4/2	0/0
MSV000079516	120154	19113	138	22/6	4/2	5/2	0/0
MSV000079517	22516	2911	37	7/3	3/1	4/2	0/0
MSV000079519	76289	13985	112	15/6	3/1	5/2	0/0
MSV000079568	224645	10273	0	0/0	0/0	4/2	0/0
MSV000079581	129012	41779	74	4/3	2/2	5/5	0/0
MSV000079598	919494	286130	9	3/1	4/1	6/3	0/0
MSV000079651	81818	5239	0	0/0	0/0	2/1	0/0
MSV000079679	595244	300682	109	24/14	12/7	14/7	2/2
MSV000079772	75916	13870	40	7/5	2/2	2/2	0/0
MSV000079778	1242178	451962	265	50/44	6/1	7/2	0/0
MSV000079813	578683	170990	23	6/4	2/2	3/2	0/0
MSV000079888	238820	74317	170	17/7	8/4	9/5	1/1
MSV000080115	1567520	709527	400	33/9	12/6	13/7	9/6
MSV000080116	70250	31009	19	9/5	6/3	6/3	0/0
TOTAL	51220679	27883895	12004	512/213	67/37	81/51	41/27

The 12,004 cyclospectra in the GNPS dataset originated from 512 cyclopeptides and 213 cyclofamilies. Dereplicator search of these cyclospectra against *CyclopeptideDatabase* identified only 67 cyclopeptides from 37 cyclofamilies. For each putative cyclopeptide, we selected a representative spectrum with the highest *k-merScore*, resulting in 512 spectra corresponding to the 512 cyclopeptides. CycloNovo *de novo* sequenced 94 cyclopeptides with P-values below 10^{-15} in this set of 512 cyclospectra.

Figure 1.6 shows the number of identified cyclopeptides across all GNPS sub-datasets. Figure 1.7 shows the number of cyclopeptides and cyclofamilies that gave rise to cyclospectra found by CycloNovo across all GNPS sub-datasets.

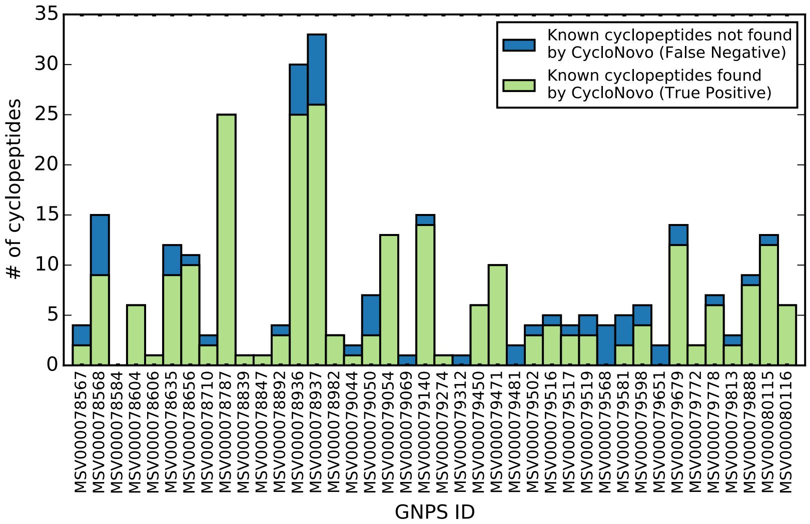


Figure 1.6. Number of cyclopeptides identified by Dereplicator across all GNPS sub-dataset. Dereplicator identified 81 unique cyclopeptides in the GNPS dataset. Since some cyclopeptides are identified in multiple sub-datasets, the total numbers of identified cyclopeptides across all GNPS sub-datasets (180) exceeds 81. The green (blue) part of each bar represent spectra that were (were not) classified by CycloNovo as cyclospectra.

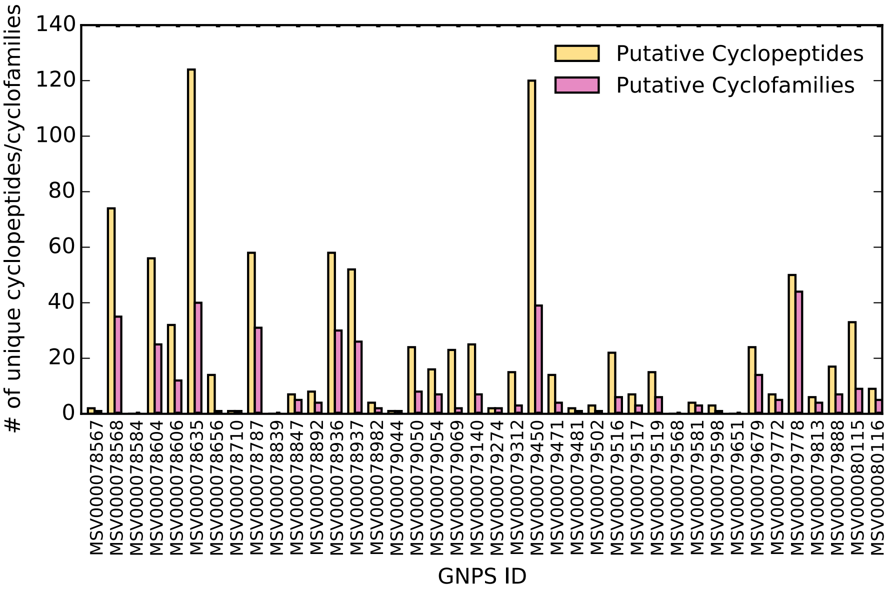


Figure 1.7. Number of cyclopeptides (yellow) and cyclofamilies (pink) found by CycloNovo across all GNPS dataset. The ID under each column shows the GNPS ID for corresponding subdataset.

Comparing CycloNovo and Dereplicator. We compared the number of distinct cyclopeptides, including some branch-cyclic peptides, and cyclofamilies revealed by CycloNovo and identified by Dereplicator in searches against the *CyclopeptideDatabase* (Figure 1.8). As Figure 1.8 illustrates, even for the extensively studied phyla of *Cyanobacteria* and *Pseudomonas*, only a small fraction of cyclopeptides and cyclofamilies revealed by CycloNovo are currently known.

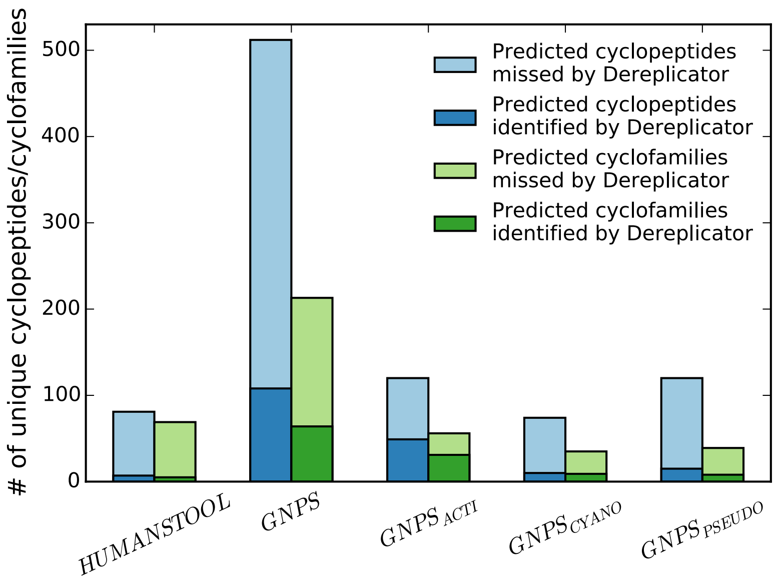


Figure 1.8. Number of cyclopeptides (blue bars) and cyclofamilies (green bars) predicted by CycloNovo and identified/missed by Dereplicator searching against *CyclopeptideDatabase* in various spectral datasets. Missed cyclopeptides/cyclofamilies are not represented in *CyclopeptideDatabase*.

Moreover, CycloNovo revealed many novel cyclopeptides in known cyclofamilies. For example, CycloNovo reconstructed six novel variants of surugamide by analyzing the $\text{GNPS}_{\text{ACTI}}$ dataset and revealed the widespread proliferation of the recently described *A-domain skipping* phenomenon[5,44], suggesting that it is more prevalent than was previously thought. The A-domain skipping phenomenon defies the traditional view that each A-domain encodes a single amino acid in an NRP according to the non-ribosomal code. Genome mining efforts typically rule out such events due to the consecutive arrangements of A-domains in NRP synthetases. CycloNovo found all the known cyclopeptides identified by Dereplicator in the HUMANSTOOL and S.VULGARIS datasets. For each MassIVE dataset included in the GNPS dataset, Figure 1.8 and Table 1.8 presents the number of known cyclopeptides/cyclofamilies identified by Dereplicator and missed by CycloNovo. Figure 1.9 shows a connected component in the spectral network containing known and novel surugamide variants (spectral dataset $\text{GNPS}_{\text{ACTI}}$ generated from samples collected from various *Actinobacteria* species).

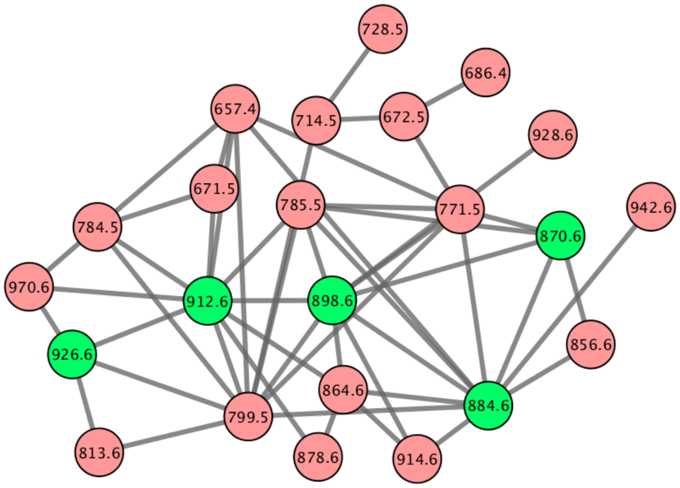


Figure 1.9. A connected component in the spectral network that contains various surugamide variants. Each node in the network is labeled by the precursor mass of a spectrum and each edge connects spectral pairs that reveal related cyclopeptides. The five green nodes are the known surugamide variants[4]. The pink nodes represent unknown cyclopeptides. The spectral network was constructed based on all cyclospectra in the GNPS$_{\mathrm{ACTI}}$ dataset.

Figure 1.10 shows a subgraph of the spectral network shown in Figure1.9 that includes only known surugamides and six novel variants reconstructed by CycloNovo. Figure 1.11 illustrates that five of these novel variants differ from known surugamides by deletions of some amino acids.

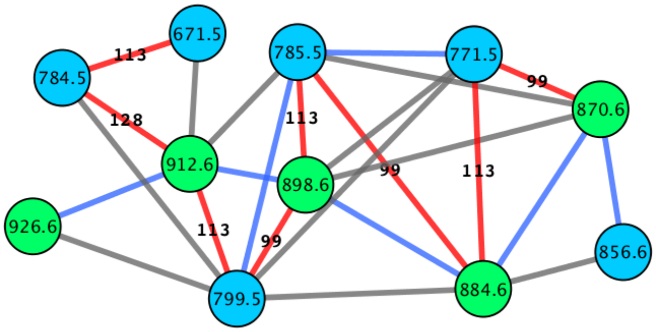
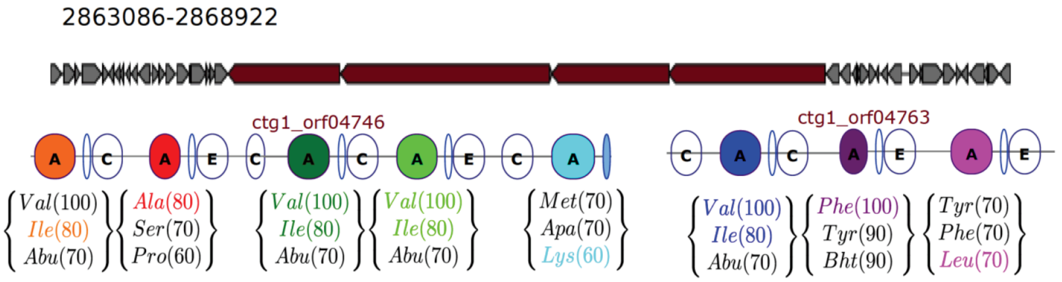


Figure 1.10. Subgraph of the surugamides spectral network (depicted in Figure 1.9) representing only known and novel surugamides. The green nodes correspond to known surugamides and the blue nodes represent the novel surugamide variants reconstructed by CycloNovo. The numbers on edges represent the nominal mass shift between the corresponding spectra. The red edges highlight the mass shifts that suggest loss/addition of an amino acid in the peptide and the blue edges connects peptides that differ from each other by a single Ile □ Val or Val □ Ile substitution (resulting in a nominal offset 14 Da). Although the 14 Da offset can also correspond to methylation, the substitutions represent the more likely explanations in this case. The grey edges show mass shifts that represent combinations of those mass shifts. Figure 1.9 presents the entire connected component.

Figure 1.11. Known and novel surugamide variants. (**a**) Surugamide gene cluster in *Streptomyces albus* along with the three most likely amino acids for each A-domain and their scores predicted by NRPSpredictor2[12]. See Mohimani et al, 2017[4] for more details on this representation. (**b**) Five known (first five rows) and six novel (last six rows) surugamide variants. Each column is color-coded based on the color of the A-domain they represent in the top figure. The dash symbols indicate a violation of the non-ribosomal code (A-domain skipping) when an A-domain in the surugamide gene cluster does not add an amino acid to a cyclopeptide. (**c**) *De novo* reconstructions of the novel surugamide variants. The column 'PSM score/highest score' shows the score of the cyclopeptide and the highest score observed for that spectrum among all CycloNovo reconstructions. The "P-value" column presents the P-value of the PSM (for each cyclopeptide, the spectrum that yielded the lowest P-value is reflected). The column "#reconstructions with score ≥ PSM score" shows the number of reconstructions with score greater or equal to the PSM score. The column "reconstruction with the highest score" shows a highest-scoring reconstruction for the cases when the PSM score is below the highest score. The number of spectra corresponding to each novel surugamide variant in the five GNPS datasets are presented in the columns '78604', '78787', '78936', '78937', and '79516', representing the GNPS sub-datasets MSV000078604, MSV000078787, MSV000078936, MSV000078937, and MSV000079516, respectively. Finding the same surugamide variants in different studies makes it unlikely that they represent artifacts. (**d**) Annotated spectra of six novel surugamide variants.

(a)



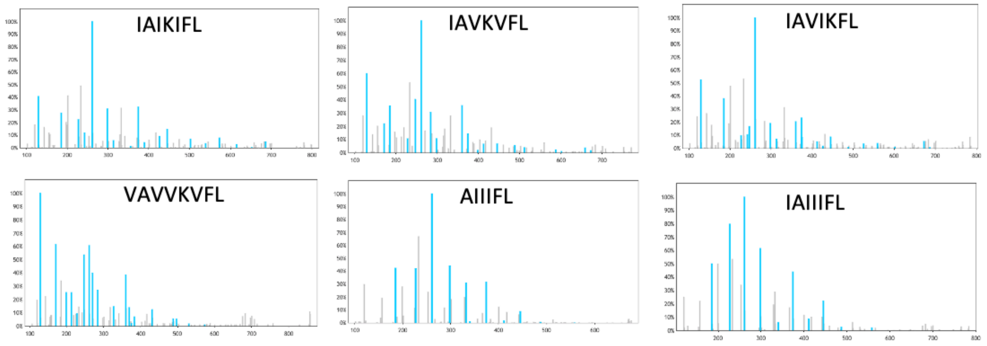
(b)

870.6	Ile	Ala	Val	Val	Lys	Val	Phe	Leu
884.6	Ile	Ala	Val	Ile	Lys	Val	Phe	Leu
898.6	Ile	Ala	Val	Ile	Lys	Ile	Phe	Leu
912.6	Ile	Ala	Ile	Ile	Lys	Ile	Phe	Leu
926.6	Ile	Ala	Ile	Ile	Lys+14	Ile	Phe	Leu
799.5	Ile	Ala	–	Ile	Lys	Ile	Phe	Leu
785.5	Ile	Ala	Val	Ile	Lys	–	Phe	Leu
771.5	Ile	Ala	Val	–	Lys	Val	Phe	Leu
856.5	Val	Ala	Val	Val	Lys	Val	Phe	Leu
784.5	Ile	Ala	Ile	Ile	–	Ile	Phe	Leu
671.5	–	Ala	Ile	Ile	–	Ile	Phe	Leu

(c)

precursor mass	sequence	PSM score/ highest score	P-value	#reconstructions with score ≥ PSM score	reconstruction with the highest score	78604	78787	78936	78937	79516
799.5	IA–IKIFL	19/19	6.2×10^{-37}	1		6				
785.5	IAVIK–FL	22/22	1.6×10^{-38}	4		5	1	4	3	
771.5	IAV–KVFL	20/22	9.4×10^{-39}	3	IAVKVLF	2	2	2	1	
856.6	VAVVKVFL	20/22	4.3×10^{-36}	5	VVVAKVFL	2				1
784.5	IAII–IFL	10/12	3.4×10^{-19}	2	AIIIIFL	5				
671.5	–AII–IFL	12/12	5.0×10^{-21}	1		5				

(d)



1.4. DISCUSSION

Although the advances in mass spectrometry and advent public repositories such as the GNPS molecular network has created a new resource for natural product discovery, there exists a large body of still unknown bioactive compounds represented by various spectra in current studies and the GNPS network[4] (less than one percent of GNPS spectra have been identified so far). As the existing database search approaches are limited to identifying known cyclopeptides and their variants, *de novo* cyclopeptide recognition and sequencing is needed to reveal the "dark matter of cyclopeptidomics."

To address this problem, we developed CycloNovo, an algorithm for *de novo* recognition and sequencing of cyclopeptides. The *de novo* recognition feature relies on the idea of spectral convolutions and can be used as a stand-alone tool for selection and prioritization of cyclopeptides in large metabolomics datasets for further downstream analyses. Using this feature, we analyzed the GNPS molecular network that contains mass spectra generated from various isolated and environmental samples. While only 81 out of 1,257 known cyclopeptides (42 out of 387 known cyclofamilies) have been identified in the GNPS network[5]. CycloNovo revealed over 400 unknown cyclopeptides from 176 novel cyclofamilies by analyzing only ≈51 million GNPS spectra from already published datasets, illustrating that the currently known cyclopeptides represent just a small fraction of cyclopeptides whose spectra have been already deposited into the GNPS network.

After *de novo* recognition of cyclospectra, we showed how CycloNovo *de novo* sequences those cyclospectra by utilizing a de Bruijn graph representation of them. This representation enables CycloNovo to find a small set of potential peptide sequences, by finding cycles in reconstructed graphs, which it then scores against the input spectra. We applied CycloNovo to

CYCLOLIBRARY and S.VULGRAIS datasets and demonstrated that CycloNovo correctly sequenced many known cyclopeptides in a blind mode and reconstructed novel cyclopeptides that were validated using transcriptomics data. This analysis therefore reports the first cyclopeptides found through fully automated *de novo* sequencing of mass spectra.

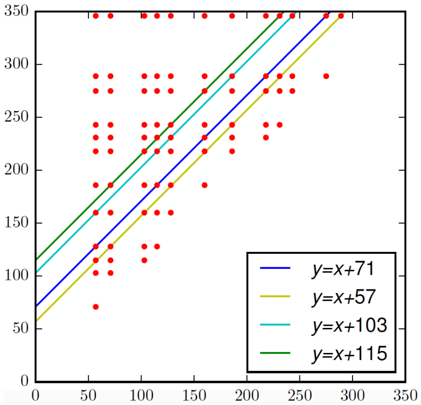
Our CycloNovo analysis of the HUMANSTOOL dataset demonstrated that several bioactive cyclopeptides from consumed food remain stable throughout the proteolytic, absorptive and microbial ecosystem provided by the gastrointestinal system and thus may be interacting with the human microbiome. Our analysis also found cyclospectra originating from the branch cyclic peptide massetolide produced by an indigenous member of the human microbiota and confirmed by metagenomics analysis. In addition, it revealed a large number of still unknown cyclopeptides in the human gut that are either a part of the human diet or are products of the human gut microbiome and provided the composition of some of these peptides including a completely novel family of cyclopeptides present in the stool samples.

In this study, we successfully applied CycloNovo to spectral datasets generated from cultured and environmental samples of different origins and phyla and produced in different laboratories and demonstrated how CycloNovo can be used for *de novo* analysis and sequencing of cyclopeptides in metabolomics datasets. As with all mass spectrometry-based tools that do not use stable isotope labeling[45], CycloNovo is unable to infer stereochemical information and to distinguish leucine from isoleucine. Moreover, CycloNovo is unable to differentiate between similar cyclopeptides (like cyclopeptides with some rearranged amino acids) that yield near identical theoretical spectra and are not discernable using our PSM function. However, in these cases, deriving further information about the final structure (for example via nuclear magnetic resonance) is simplified when some partial sequence information is available through CycloNovo predictions. See Methods section for more information about other possible failure modes of CycloNovo and the limitations of our study.

1.5. METHODS

Spectral convolutions.

We represent each spectrum *Spectrum*={s_1, …, s_n} as its *spectral diagram*, the set of $n\times(n-1)/2$ 2-dimensional points (s_i, s_j) for $1 \leq i < j \leq n$. Given a mass *a*, the convolution of *Spectrum with offset a* (denoted *convolution*(*Spectrum, a*)) is equivalently defined as the number of points in the diagonal (45°) band $y \approx x+a$ in the spectral diagram. Figure 1.12 presents the spectral diagram of *TheoreticalSpectrum*(AGCD) and reveals that bands corresponding to its amino acids (71, 57, 103, and 115 Da) are the most *populous* (contain a large number of points as compared to other bands), meaning *convolution*(*Spectrum,a*) is high when *a* is the mass of amino acids A, G, C, or D. For example, *TheoreticalSpectrum*(AGCD) includes five pairs of fragment masses ((G,AG), (D,AD), (AGC,GC), (CD,CDA), and (GCD,AGCD)) that are located on the "blue" diagonal $y = x+mass(A)$ in Figure 1.12.



cluster mass	cluster multiplicity	cluster distance
113.07	**12**	**0.006 (I/L)**
128.08	**7**	**0.006 (K)**
56.02	6	-
186.10	**6**	**0.029 (W)**
71.04	**5**	**0.004 (A)**
147.07	**4**	**0.001 (F)**
156.10	**4**	**0.002 (R)**
184.11	4	-
98.08	4	-
169.11	4	-
57.05	4	0.027 (G)

Figure 1.12. The spectral diagram of *TheoreticalSpectrum*(AGCD) (left) and the list of clusters in the convolutions of *Spectrum*$_{Surugamide}$ (right). (Left) The highlighted lines with slope 1 correspond to the masses of the amino acids, A, G, C, and D and contain 5, 9, 5, and 6 points, respectively. (Right) Clusters in the convolutions of *Spectrum*$_{Surugamide}$ in the decreasing order of their multiplicities. Only clusters with masses between 55 and 190 Da and multiplicity exceeding 3 are shown. Cyclopeptidic clusters are shown in bold and cyclopeptidic clusters with masses similar to the masses of amino acids in surugamide are shown in red. *Cluster distance* is defined as the distance between the cluster mass and a closest mass of a selected amino acid.

Figure 1.12 lists high-multiplicity clusters for *Spectrum*$_{Surugamide}$ and shows that many of them have masses that are similar to the masses of amino acids in surugamide. Since populous diagonals (high-multiplicity clusters) in the spectral diagram reveal amino acids in the unknown cyclopeptide that gave rise to an experimental spectrum, we use them to generate the set of putative amino acids[22].

The spectral diagrams for *TheoreticalSpectrum*(*Surugamide*) and experimental *Spectrum*$_{Surugamide}$ highlight four populous diagonal bands $y \approx x+a$, where a is the mass of one of four amino acids in surugamide with integer masses 71, 113, 128, and 147. These populous bands in the spectral diagram reveal the masses of amino acids in an unknown cyclopeptide that gave rise to an experimental spectrum. Figure 1.13 illustrates that each amino acid in surugamide results in a populous diagonal in the spectral diagram of *Spectrum*$_{Surugamide}$. For each constructed cluster (diagonal band in the spectral diagram), we consider all pairs of masses in *Spectrum* that contributed to this cluster and form a *band* as the set of these *k* pairs.

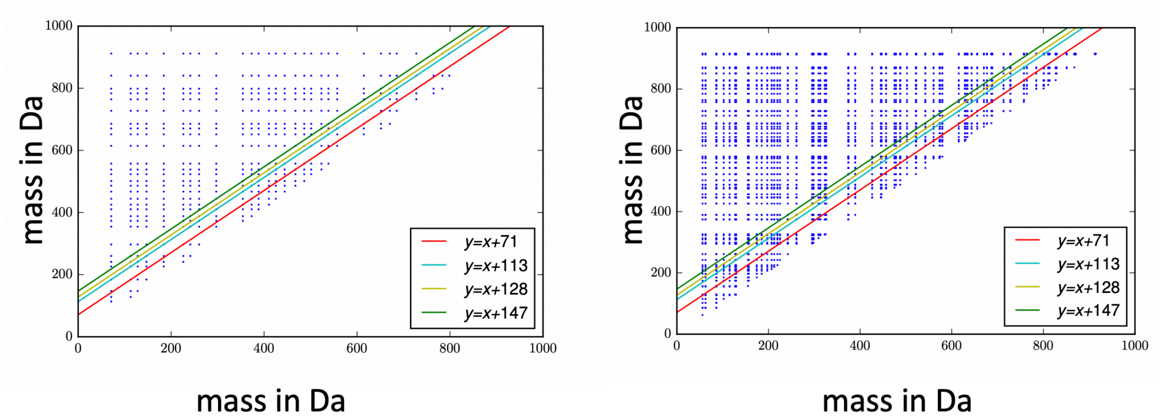


Figure 1.13. The spectral diagrams of the *TheoreticalSpectrum(Surugamide)* (left) and *Spectrum*$_{Surugamide}$ (right). The highlighted lines with slope +1 have *y*-intercepts equal to the masses of the constituent amino acids of surugamide (A, L/I, K, and F). Amino acids A, L/I, K, and F correspond to populous diagonals containing 11, 23, 11, and 11 points (left figure) and 5, 14, 8, and 4 points (right figure), respectively.

We define the *cluster diameter* as the difference between its maximum and minimum elements. Figure 1.14 presents the band for the cluster with multiplicity 8 and mass 128.09 (diameter 0.03) in the spectral convolution of *Spectrum*$_{Surugamide}$ and reveals that the 8 elements of this band can be partitioned into 7 groups of closely located points. We are interested in the number of such groups (rather than the raw cluster multiplicities) since experimental spectra often contain *satellite masses* resulting from *neutral losses* and *isotopic peaks.* For example, in addition to the integer mass 242 Da corresponding to the peptide IK, *Spectrum*$_{Surugamide}$ also contains the integer mass 225 Da corresponding to the loss of NH_3 from this peptide.

$y \approx x + 128$	(185.13,313.22) (261.17,389.26) (298.21,426.28) (374.24,502.32) **(446.30,574.38) (445.28,573.35)** (487.35,615.42) (784.52,912.62)

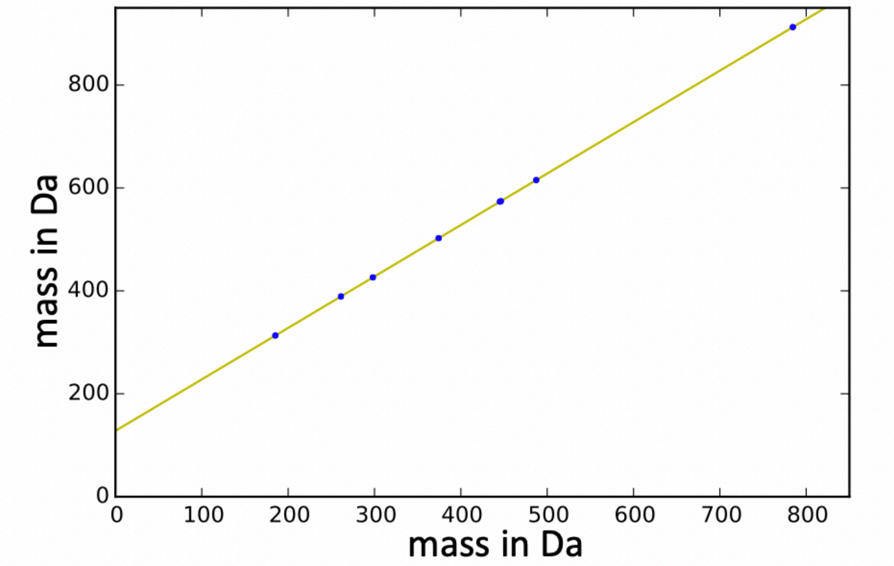


Figure 1.14. A band with multiplicity eight in *Spectrum*$_{Surugamide}$ (cluster with mass 128.09 and diameter 0.03). (**top**) Coordinates of the points in the band. Since the difference between the *x*-coordinates and *y*-coordinates of the two points shown in bold match the mass of hydrogen, these two points are clustered together in this band. (**bottom**) The same band in the spectral diagram for *Spectrum*$_{Surugamide}$. The points of the band can be partitioned into seven groups of closely located points: six singleton groups and one group with two elements.

Since satellite masses artificially inflate cluster multiplicities, there is a need to reduce biases caused by these masses. We thus define the set of common satellite offsets (1 Da (H), 18 Da (H_2O), 17 Da (NH_3), and 28 Da (CO)) and perform additional single linkage clustering in each populous band by combining pairs of masses in a single cluster if both their *x*-coordinates and *y*-coordinates differ by a satellite offset. We redefine the concept of cluster multiplicity as the number of the resulting clusters in the band (Table 1.9)

Table 1.9. List of clusters in the spectral convolution of *Spectrum*$_{Surugamide}$. Clusters are shown in the decreasing order of their multiplicities (only clusters with multiplicity at least 2 are shown). Cyclopeptidic clusters are shown in bold and cyclopeptidic clusters with masses similar to masses of amino acids in surugamide are shown in red.

cluster mass	multiplicity after satellite removal	multiplicity before satellite removal	cluster diameter	cluster distance
113.078	**12**	**14**	**0.106**	**0.006 (I/L)**
128.089	**7**	**8**	**0.032**	**0.006 (K)**
56.025	6	6	0.033	-
186.108	**6**	**6**	**0.077**	**0.029 (W)**
71.041	**5**	**5**	**0.022**	**0.004 (A)**
147.069	**4**	**4**	**0.027**	**0.001 (F)**
156.099	**4**	**5**	**0.036**	**0.002 (R)**
184.106	4	5	0.032	-
98.079	4	4	0.031	-
169.11	4	5	0.01	-
57.049	4	4	0.02	-
70.042	3	3	0.01	-
132.058	3	3	0.006	-
133.584	3	3	0.013	-
183.116	3	3	0.024	-
168.159	3	3	0.023	-
52.049	3	3	0.005	-
96.035	3	3	0.017	-
165.115	3	3	0.004	-
76.041	3	4	0.013	-
73.08	3	3	0.017	-
101.053	**3**	**3**	**0.011**	**0.005 (T)**
167.886	3	4	0.024	-
189.105	2	3	0.009	-
57.018	**2**	**2**	**0.006**	**0.004 (G)**
114.09	2	2	0.01	-
45.041	2	2	0.012	-

Recognizing Cyclospectra.

A cluster in the spectral convolution is called *frequent* if its multiplicity exceeds the *cluster multiplicity threshold* (the default threshold for *Spectrum*$_{Surugamide}$ is 7). CycloNovo classifies a spectrum as a cyclospectrum if the number of frequent cyclopeptidic clusters in its spectral convolution is at least *minNumberFrequentClusters* (the default value *minNumberFrequentClusters*=2). Since there exist two frequent cyclopeptidic clusters for *Spectrum*$_{Surugamide}$ (corresponding to amino acids I/L and K), it is classified as cyclopeptidic (Figure

1.12). In addition to *Spectrum*$_{Surugamide}$, out of 938 spectra passing the preprocessing step in the small spectral dataset for *Streptomyces CNQ329* that contains *Spectrum*$_{Surugamide}$, CycloNovo recognized only one cyclospectrum, also originated from surugamide.

The challenge of distinguishing cyclospectra from spectra of linear peptides and polymers. Fragmentation of linear peptides typically results in *prefix* (e.g., b-ions) and *suffix* (e.g., y-ions) ions and rarely generates *internal* ions. However, spectra of some linear peptides feature a substantial number of internal ions, leading to a possibility to erroneously classify them as cyclospectra. Another source of a potential misclassification of some spectra as cyclospectra are polymers that represent a common source of contamination in mass spectral datasets. Since polymers are made up of repeated units, the spectral convolution of a polymer spectrum typically has high-multiplicity clusters (for clusters corresponding to masses of the repeat units). In some cases, the adducts of these repeat units form high multiplicity clusters with masses equal to the masses of a selected amino acid, triggering a possibility to misclassify a polymer spectrum as a cyclospectrum.

LINEARLIBRARY and POLYMERLIBRARY datasets. To ensure that CycloNovo does not misclassify spectra of linear peptides and polymers as cyclospectra,

we analyzed two spectral datasets described below:

- LINEARLIBRARY is a set of 105,871 Collision-Induced Dissociation (CID) tandem mass spectra of distinct linear peptides from the MassIVE Knowledge-Based (MassIVE-KB) spectral library[46] of linear peptides distilled from all human proteomics data in the MassIVE database. In particular the CID library under MassiveIVE-KB is the annotated spectra from CID datasets generated from a collection of over 330,000 synthetic tryptic peptides representing almost all of the canonical human gene products as a synthetic human proteome collection (Zolg *et al.*, *Nature Methods*, 2017). In this study, pools of synthetic peptides were subjected to LC-MS/MS analysis using CID fragmentation method coupled with ion trap or Orbitrap readouts.

- POLYMERLIBARY is a set of 448 tandem spectra generated from polyethylene glycol (MSV000081544).

These spectral datasets have spectra with the precursor masses varying between 500 Da and 2000 Da and the charges at most 2.

Additional tests for recognizing cyclospectra. To distinguish cyclospectra from spectra of linear peptides and polymers, CycloNovo only classifies a spectrum as cyclopeptidic if it passes additional tests described below.

- **High multiplicity cyclopeptidic clusters test (distinguishing cyclospectra from spectra of linear peptides).** As described in the main text, CycloNovo first selects a spectrum for further analysis if its spectral convolution has at least *minNumberFrequentClusters* frequent cyclopeptidic clusters, i.e., clusters with multiplicities exceeding the cluster multiplicity threshold. Since the cluster multiplicities typically increase with the increase in the length of a peptide, this threshold increases with the increase in the peptide mass. We thus defined the *cluster multiplicity threshold* as $\alpha \times precursorMass + \beta$ (see below for selecting parameters α and β).
- **Polymer test (distinguishing cyclospectra from polymer spectra).** For each cyclospectrum *Spectrum*, CycloNovo analyzes clusters with masses of repeat units observed in background contamination from polyethylene glycol, NaCl, polypropylene glycol, and trimethylsiloxane (44.03, 57.96, 58.04, and 72.04 Da, respectively). We refer to these masses as *polymeric units*[47] and refer to clusters with masses equal to polymeric

units as *polymer-clusters*. CycloNovo classifies a spectrum as polymeric if there exist at least *minNumberFrequentClusters* polymer-clusters with multiplicities at least the *cluster multiplicity threshold.* Polymeric spectra are filtered out from the set of found cyclospectra.

- ***cycloIntensity* test.** For each cyclospectrum, CycloNovo considers all frequent cyclopeptidic clusters. For each such cluster of mass *a*, we consider all pairs of masses *x* and *y* in the spectrum contributing to this cluster, i.e., satisfying the condition $y \approx x+a.$ The *cyclointensity* of the spectrum, referred to as *cycloIntensity*, is defined as the total intensity of all such peaks (across all frequent cyclopeptidic clusters) divided by the total intensity of all peaks in *Spectrum*. Spectra with cyclointensity below the *cycloiIntensity threshold* are filtered out.
- ***k-merScore* test.** CycloNovo computes the *k-merScore*, the score of the highest-scoring *k*-mer that contributes to the de Bruijn graph of the spectrum and filters out cyclospectra with *k-merScore* below the *k-merScore threshold.*

Selecting thresholds for recognizing cyclospectra. To select the default value of *cluster multiplicity threshold*=$\alpha \times precursorMass+\beta$, we varied parameters α (from 0.005 to 0.02) and β (from -5 to +5) and analyzed all found cyclospectra in the CYCLOLIBRARY, LINEARLIBRARY, and POLYMERLIBARY datasets (Figure 1.15). Despite its smaller size, CYCLOLIBRARY is the only dataset where CycloNovo recognizes cyclospectra for all analyzed values of α and β. Since α=0.07 and β=-1 yielded the largest number of recognized cyclospectra in CYCLOLIBRARY (46 out of 81) and no cyclospectra in the LINEARLIBRARY and POLYMERLIBRARY datasets (Figure 1.15) we selected these values as the default parameters.

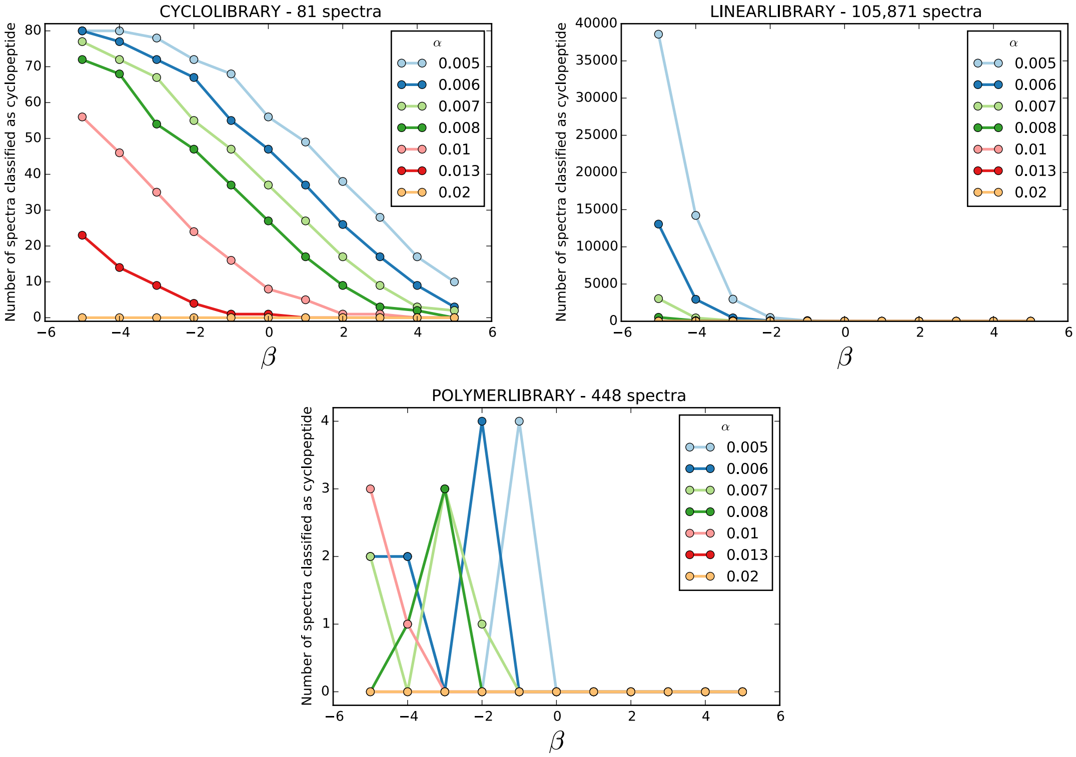


Figure 1.15. Number of spectra passing both the "high multiplicity cyclopeptidic cluster" and the "polymer". Number of spectra in CYCLOLIBRARY, LINERALIBRARY, and POLYMERLIBRARY datasets passing the two tests using various values of parameters α and β, are shown.

Figure 1.16 presents the values of *cycloIntensity* and *k-merScore* for each spectrum in the CYCLOLIBRARY, LINEARLIBRARY, and POLYMERLIBRARY datasets and reveals a separation between the former and the two latter datasets with respect to these two parameters. CycloNovo thus classifies a spectrum as a cyclospectrum if its *cycloIntensity* exceeds the *cycloIntensity* threshold (60%) and its *k-merScore* exceeds the *k-merScore* threshold (5). 45 spectra in the CYCLOLIBRARY datasets, that pass all four tests described above, are classified as cyclospectra.

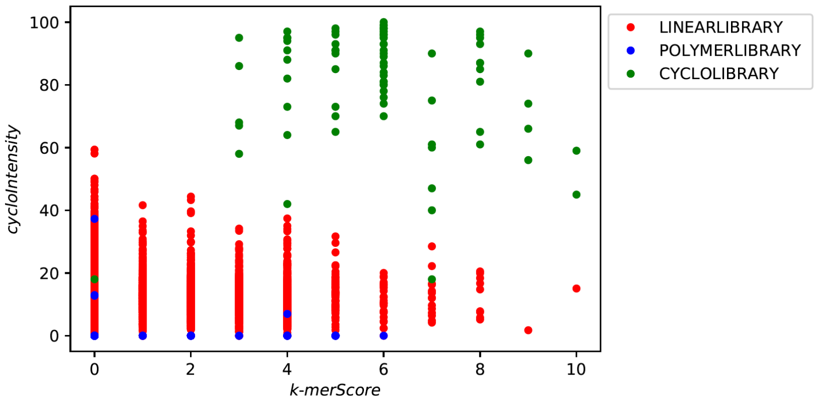


Figure 1.16. ***cycloIntensity*** **and** ***k-merScore*** **values in various annotated datasets.** Each point presents the *cyclointensity* (x-axis) and *k-merScore* (y-axis) values for a spectrum in the entire CYCLOLIBRARY, POLYMERS, and LINEARLIBRARY datasets (k=5). Colors highlight spectra in different datasets.

We also investigated how CycloNovo's ability to recognize a cyclospectrum is affected by the fragmentation quality of the corresponding PSM (measured by the P-value of this PSM). For each spectrum in the CYCLOLIBRARY dataset, we identified the minimum value of the parameter α that leads to classifying this spectrum as cyclopeptidic (for β=-1). Figure 1.17 illustrates that well-fragmented spectra can be recognized even with more restrictive threshold values (larger values of α).

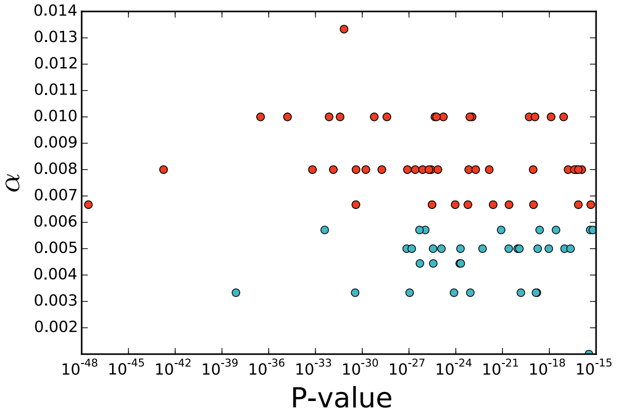


Figure 1.17. Dependence between the P-values and the parameter alpha (for $\beta = -1$) that leads to correctly recognizing each cyclospectrum in CYCLOLIBRARY. Each point represents a spectrum in the CYCLOLIBRARY dataset. The x-axis shows the P-value of the PSM for that spectrum and the y-axis shows the minimum value of the parameter α that leads to classifying this spectrum as a cyclospectrum. The points corresponding to cyclospectra recognized with the default parameter α=0.07 are shown in red.

Estimating the number of distinct cyclopeptides and cyclofamilies.

Spectral datasets often contain multiple spectra originating from the same compound. CycloNovo clusters similar cyclospectra using MS-Cluster[24] and estimates for the number of distinct cyclopeptides as the number of constructed clusters. CycloNovo further constructs the spectral network of cyclospectra using SpecNets[3] and estimates for the number of distinct cyclofamilies as the number of connected components in this network.

Limitations of CycloNovo.

CycloNovo is only applicable to high-resolution CID MS/MS experiments (<0.02 Da window). Also, it has been only on MS/MS data generated by ion trap or Orbitrap instruments. In this paper, we showed that CycloNovo can analyze spectra representing cyclopeptides from RiPP orbitides and cyanobactins RiPP classes. CycloNovo has also been successfully applied to non-ribosomal cyclopeptide (or lipopeptides) produced by bacterial organisms from many genera including Pseudomonas, Bacilli, cyanobacterial genera *Lyngbya* and *Anabaena,* actinobacteria genera *Streptomyces*, *Narcodia*, etc. In what follows we describe a list of challenges and failure modes for CycloNovo in recognizing or sequencing cyclospectra. Any classes of cyclopeptide that are prone to any of the following characteristics are less likely to be successfully analyzed by CycloNovo:

A. Cyclopeptide has monomers not included in the initial amino acid list: either they have modifications that are not considered by CycloNovo and/or they include non-proteinogenic amino acids not considered by CycloNovo.
B. Cyclopeptide includes side chains (or lipid chain). Specifically, CycloNovo often fails to reconstruct branch-cyclic peptides with long side chains (peptidic/non-peptidic) or

multiple side chains, where only a small portion of amino acids are appearing in their ring, especially if the total mass of the ring is smaller compared to the total mass of the cyclopeptide. The spectral convolution analysis step of such peptides typically fails because their cyclopeptidic convolution does not meet the threshold set by their peptide mass as the mass representing the ring is much smaller than the intact peptide mass.

C. Cyclopeptide has many fragments with similar masses (hence similar fragments): since the number of unique peaks in the spectrum matching the theoretical spectrum of such cyclopeptide is low, there is insufficient amount of information in their corresponding spectrum. Therefore, the cyclospectrum might fail the spectral convolution or *k-merScore* tests.

D. Cyclopeptide is a hybrid compound and contains large non-peptidic elements, for example polyketide-NRP hybrid or a lipopeptide with large lipid chains (see the explanation B about side chains above)

E. Cyclopeptide includes bonds other than amide bonds (for example disulfide bonds).

F. Cyclopeptide only is represented in spectra with charge states larger than 2: CycloNovo only consider spectra with charge states 1 and 2.

Table 1.10 provides information about the seven peptides that CycloNovo failed to reconstruct. Note that, although the mentioned features hinder the recognition and sequencing of cyclospectra by CycloNovo, CycloNovo may perform well on cyclopeptides with above conditions. We distinguish between *simple* cyclopeptides (like surugamide) and *complex* peptides such as peptides with monomers not included in the initial set of amino acids or the ones with a complex backbone structure such as branch-cyclic peptides, lipopeptides, depsipeptides and others. CycloNovo has two main applications: recognizing a cyclospectrum and sequencing it.

With respect to sequencing, CycloNovo can output either a complete (for simple cyclopeptides) or partial cyclopeptide reconstruction (for complex cyclopeptides). Table 1.3 shows the number of correct *k*-mers predicted by CycloNovo (using only the default set of amino acids) versus the length of the cyclopeptide. This column highlights that CycloNovo reports at least one correct *k*-mer for many lipopeptides (19 out of 34) although it is not designed for complex compounds. With respect to cyclospectrum recognition, CycloNovo successfully recognized many spectra that originated from complex peptides as cyclospectra (see Table 1.1). Although we were able to predict many putative cyclopeptides with extremely low P-values and confirmed several of them with accompanying datasets, but we did not have the resources to experimentally validate our putative cyclopeptides through NMR or cyclopeptide synthesis.

Table 1.10. Seven cyclopeptides for which CycloNovo failed to correctly sequence the corresponding cyclospectra from CYCLOLIBRARY. CycloNovo failed to predict all amino acids in five of these cyclopeptides and failed to predict some correct *k*-mers in two of these cyclopeptide since their score was below the threshold. For each peptide, we list either the *k*-mer or the mass of the amino acid that CycloNovo failed to predict (*k*-mers are shown as a sequence of nominal masses of amino acids in bold). In cases where an amino acid was missed, the composition of amino acid is presented under column 'elemental compositions of missing amino acids'. The third column presents the elemental composition of missed amino acids. In one case when CycloNovo was able to generate a reconstruction (arthrofactin), the highest-scoring reconstruction with sequence of nominal masses 115 113 101 115 170 113 113 87 87 113 113 113 and score 20 represents a rearrangements of the correct peptide 115 113 **170 115 113** 113 **87 113 87** 113 113 **101** with score 17 (differently arranged amino acid mases in the correct cyclopeptide are shown in bold). For each cyclopeptide, the last column presents the type of experimental or structural challenges that contributed to CycloNovo's failure for that peptide, as listed and labeled above.

peptide ID	missing *k*-mer or amino acid	elemental composition of missing amino acids	Structural Challenges
arthrofactin	**113 113 101 115 113**	-	C, G
bacilomycin D5	97.05	C5H7NO	C
dolastatin 1-31	127.1	C7H13NO	B, E
dolastatin 12	**114 113 85 141 161**	-	B, C, E
puwainaphycin A	325.36	C19H35NO3	A, B, E
puwainaphycin C	317.21	C17ClH32NO2	A, B, E
SCH-378167	143.06 and 147.07	C6H9NO3 and C9H9NO	A

Computing P-values for *de novo* generated Peptide-Spectrum-Matches.

The PSM score, while informative, is biased with respect to length. Therefore, CycloNovo computes a P-value to evaluate the statistical significance of each individual PSM formed. Given a PSM(*P,S*) between a cyclopeptide *P* and a spectrum *S*, MS-DPR[13] computes the probability (p-value) that a random peptide forms a PSM with the spectrum *S* with a score that is greater than or equal to the score of PSM(*P,S*). Therefore, the p-value statistics is not biased by length.

To estimate the p-value of a PSM, one can use *Monte Carlo simulations* by randomly generating a population of billions of peptides and estimating the distribution of PSM scores of all peptides against *S*, using the distribution of PSM scores in this population. But this approach becomes prohibitively time-consuming for estimating very low *p*-values, i.e., when calculating the probabilities of extremely rare events. For example, estimating p-values as low as 10^{-12} requires calculating PSM scores of trillions of randomly generated peptides. Therefore, this method is impractical in mass spectrometry experiments where PSMs with *p*-values as low as 10^{-12} are common (Kim and Pevzner, *Nature Communications*, 2014).

To overcome this challenge, MS-DPR uses a method for evaluating probability of rare events (peptides yielding “high” PSM scores) called *multilevel splitting* for Markov Chain Monte Carlo sampling. This method that was originally developed in nuclear physics, rapidly approximates an extreme tail of the probability distribution of PSM scores against a spectrum *S*. It constructs a Markov Chain over a space of PSM scores of millions of peptides similar to *P* (in molecular weight and length) against *S*. It then uses selection mechanisms under a multilevel splitting implementation that favors the trajectories in this Markov chain deemed likely to lead to rare events. Using this method, it dedicates a greater fraction of the computational effort to a portion of the peptide space that leads to higher PSM scores against *S* and therefore can efficiently estimate the total probability of all peptides with high scores in the constructed Markov chain.

Note that computing p-values of *de novo* reconstructions of linear peptides does not make sense since there exists an efficient dynamic programming algorithm for finding a *linear* peptide with maximum score against a spectrum *S*[26]. Indeed, the peptide with maximum score against a spectrum *S,* over the set of all possible peptides, corresponds to the most extreme point in the probability distribution and therefore its P-value is exactly 0.

In contrast, since efficient algorithms for finding a *cyclopeptide* with the maximum score against a spectrum *S* remain unknown (CycloNovo is a heuristic approach that does not guarantee finding the highest-scoring cyclopeptide), CycloNovo reconstructions have to be evaluated with respect to p-values. Hence, under the assumption that the reconstructed cyclopeptide did not reach the maximum score among all peptides (with ANY amino acids), using MS-DPR provides a statistical measure for *de novo* cyclopeptide sequencing by evaluating how “close” the score of the reconstructed cyclopeptide is to the maximum score in the space of all peptides. Note that this is consistent with calculating P-values in database search in traditional “linear” proteomics, where the assumption is that the exact peptide generating the spectrum (*i.e.* the peptide with the maximum score) may not be present in the database.

Availability. CycloNovo is available as both a stand-alone tool (https://github.com/bbehsaz/cyclonovo) and as a web application (http://gnps.ucsd.edu/ProteoSAFe/static/gnps-theoretical.jsp). All described datasets are available through the corresponding public repositories.

1.6. ACKNOWLEDGEMENTS

We thank Ben Pullman, Sergey Nurk, Alexey Melnik, and Louis-Felix Nothias for fruitful discussions.

Chapter 1, in full, Chapter 1, in full, is a reformatted reprint of "De novo peptide sequencing reveals many cyclopeptides in the human gut and other environments" as it appears in *Cell Systems, 2020*[48] by Bahar Beshsaz, Hosein Mohimani, Alexey Gurevich, Andrey Prjibelski, Mark Fisher, Fernando Vargas, Larry Smarr, Pieter C. Dorrestein, Joshua S. Mylne, and Pavel A. Pevzner. The dissertation of author was the primary author of this material.

Contributions. P.A.P. and B.B. designed the CycloNovo algorithm and B.B. implemented it. B.B. did the benchmarking and spectral network analysis for all datasets included in this study. M.F. and J.M. generated the S.VULGARIS spectral dataset and helped with the biological interpretation and validation of the identified cyclopeptides. A.P. assembled RNA-seq data for S.VULGARIS. L.S. contributed the HUMANSTOOL sample set for analysis and commented on the results. H.M and A.G. performed the Dereplicator search for all datasets. P.C.D., H.M. and P.A.P. directed the work.

Funding. The work of B.B., H.M. and P.A.P. was supported by the US National Institutes of Health (grant 2-P41-GM103484). The work of B.B. was also supported by the Natural Science and Engineering Research Council of Canada (NSERC) Graduate Scholarship, Center for Microbiome Innovation (CMI) Grand Challenges Award grant and CMI Microbial Sciences Graduate Fellowship. The work of A.G. and A.P. was supported by the Russian Science Foundation (grant 14-50-00069). J.S.M. was supported in part by an Australian Research Council Future Fellowship (FT120100013). M.F.F. was supported by the Australian Government's

Research Training Program and a Bruce and Betty Green Postgraduate Research Scholarship. H.B.B. was supported by the LOEWE research cluster MegaSyn funded by the State of Hesse.

Pavel Pevzner is a co-founder, has an equity interest and receives income from Digital Proteomics, LLC. The terms of this arrangement have been reviewed and approved by the University of California, San Diego in accordance with its conflict of interest policies.

1.7. REFERENCES

1. Ling, L. L., Schneider, T., Peoples, A. J., Spoering, A. L., Engels, I., Conlon, B. P., Mueller, A., Hughes, D. E., Epstein, S., Jones, M., Lazarides, L., Steadman, V. a, Cohen, D. R., Felix, C. R., Fetterman, K. A., Millett, W. P., Nitti, A. G., Zullo, A. M., Chen, C. & Lewis, K. A new antibiotic kills pathogens without detectable resistance. *Nature* **517,** 455–459 (2015).

2. Medema, M. H. & Fischbach, M. A. Computational approaches to natural product discovery. *Nature chemical biology* **11,** 639–648 (2015).

3. Wang, M., Carver, J. J., Phelan, V. V, Sanchez, L. M., Garg, N., Peng, Y., Nguyen, D. D., Watrous, J., Kapono, C. A., Luzzatto-Knaan, T., Porto, C., Bouslimani, A., Melnik, A. V, Meehan, M. J., Liu, W.-T., Crüsemann, M., Boudreau, P. D., Esquenazi, E., Sandoval-Calderón, M., Kersten, R. D., Pace, L. A., Quinn, R. A., Duncan, K. R., Hsu, C.-C., Floros, D. J., Gavilan, R. G., Kleigrewe, K., Northen, T., Dutton, R. J., Parrot, D., Carlson, E. E., Aigle, B., Michelsen, C. F., Jelsbak, L., Sohlenkamp, C., Pevzner, P., Edlund, A., McLean, J., Piel, J., Murphy, B. T., Gerwick, L., Liaw, C.-C., Yang, Y.-L., Humpf, H.-U., Maansson, M., Keyzers, R. A., Sims, A. C., Johnson, A. R., Sidebottom, A. M., Sedio, B. E., Klitgaard, A., Larson, C. B., Boya P, C. A., Torres-Mendoza, D., Gonzalez, D. J., Silva, D. B., Marques, L. M., Demarque, D. P., Pociute, E., O'Neill, E. C., Briand, E., Helfrich, E. J. N., Granatosky, E. A., Glukhov, E., Ryffel, F., Houson, H., Mohimani, H., Kharbush, J. J., Zeng, Y., Vorholt, J. A., Kurita, K. L., Charusanti, P., McPhail, K. L., Nielsen, K. F., Vuong, L., Elfeki, M., Traxler, M. F., Engene, N., Koyama, N., Vining, O. B., Baric, R., Silva, R. R., Mascuch, S. J., Tomasi, S., Jenkins, S., Macherla, V., Hoffman, T., Agarwal, V., Williams, P. G., Dai, J., Neupane, R., Gurr, J., Rodríguez, A. M. C., Lamsa, A., Zhang, C., Dorrestein, K., Duggan, B. M., Almaliti, J., Allard, P.-M., Phapale, P., Nothias, L.-F., Alexandrov, T., Litaudon, M., Wolfender, J.-L., Kyle, J. E., Metz, T. O., Peryea, T., Nguyen, D.-T., VanLeer, D., Shinn, P., Jadhav, A., Müller, R., Waters, K. M., Shi, W., Liu, X., Zhang, L., Knight, R., Jensen, P. R., Palsson, B. Ø., Pogliano, K., Linington, R. G., Gutiérrez, M., Lopes, N. P., Gerwick, W. H., Moore, B. S., Dorrestein, P. C. & Bandeira, N. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nature Biotechnology* **34,** 828–837 (2016).

4. Mohimani, H., Gurevich, A., Mikheenko, A., Garg, N., Nothias, L.-F., Ninomiya, A., Takada, K., Dorrestein, P. C. & Pevzner, P. A. Dereplication of peptidic natural products through database search of mass spectra. *Nature Chemical Biology* **13,** 30–37 (2017).

5. Gurevich, A., Mikheenko, A., Shlemov, A., Korobeynikov, A., Mohimani, H. & Pevzner, P. A. Increased diversity of peptidic natural products revealed by modification-tolerant database search of mass spectra. *Nature Microbiology* **3,** 319–327 (2018).

6. Zorzi, A., Deyle, K. & Heinis, C. Cyclic peptide therapeutics: past, present and future. *Current Opinion in Chemical Biology* **38,** 24–29 (2017).

7. Marahiel, M. A., Stachelhaus, T. & Mootz, H. D. Modular Peptide Synthetases Involved in Nonribosomal Peptide Synthesis. *Chemical Reviews* **97,** 2651–2674 (1997).

8. Bandeira, N., Tsur, D., Frank, A. & Pevzner, P. A. Protein identification by spectral networks analysis. *Proceedings of the National Academy of Sciences U.S.A.* **104,** 6140–6145 (2007).

9. Mohimani, H., Wei-Ting Liu, Y.-L. Y., Susana P. Gaudêncio, W. F., Dorrestein, P. C. & Pevzner, P. A. Multiplex de novo sequencing of peptide antibiotics. *Journal of Computational Biology* **18,** 1371–1381 (2011).

10. Mohimani, H., Liu, W.-T., Kersten, R. D., Moore, B. S., Dorrestein, P. C. & Pevzner, P. A. NRPquest: coupling mass spectrometry and genome mining for nonribosomal peptide discovery. *Journal of natural products* **77,** 1902–1909 (2014).

11. Mohimani, H., Kersten, R. D., Liu, W.-T., Wang, M., Purvine, S. O., Wu, S., Brewer, H. M., Pasa-Tolic, L., Bandeira, N., Moore, B. S. & others. Automated genome mining of ribosomal peptide natural products. *ACS chemical biology* **9,** 1545–1551 (2014).

12. Röttig, M., Medema, M. H., Blin, K., Weber, T., Rausch, C. & Kohlbacher, O. NRPSpredictor2 - A web server for predicting NRPS adenylation domain specificity. *Nucleic Acids Research* **39,** 362–367 (2011).

13. Mohimani, H., Kim, S. & Pevzner, P. A. A new approach to evaluating statistical significance of spectral identifications. *Journal of Proteome Research* **12,** 1560–1568 (2013).

14. Mohimani, H. & Pevzner, P. A. Dereplication, sequencing and identification of peptidic natural products: from genome mining to peptidogenomics to spectral networks. *Natural product reports* **33,** 73–86 (2016).

15. Mohimani, H., Liu, W. T., Mylne, J. S., Poth, A. G., Colgrave, M. L., Tran, D., Selsted, M. E., Dorrestein, P. C. & Pevzner, P. A. Cycloquest: Identification of cyclopeptides via database search of their mass spectra against genome databases. *Journal of Proteome Research* **10,** 4505–4512 (2011).

16. Ibrahim, A., Yang, L., Johnston, C., Liu, X., Ma, B. & Magarvey, N. A. Dereplicating nonribosomal peptides using an informatic search algorithm for natural products (iSNAP) discovery. *Proceedings of the National Academy of Sciences U.S.A.* **109,** 19196–19201 (2012).

17. Mohimani, H., Yang, Y. L., Liu, W. T., Hsieh, P. W., Dorrestein, P. C. & Pevzner, P. A. Sequencing cyclic peptides by multistage mass spectrometry. *Proteomics* **11,** 3642–3650 (2011).

18. Ng, J., Bandeira, N., Liu, W. T., Ghassemian, M., Simmons, T. L., Gerwick, W. H.,

Linington, R., Dorrestein, P. C. & Pevzner, P. a. Dereplication and de novo sequencing of nonribosomal peptides. *Nat Methods* **6,** 596–599 (2009).

19. Kavan, D., Kuzma, M., Lemr, K., Schug, K. A. & Havlicek, V. CYCLONE - A utility for de novo sequencing of microbial cyclic peptides. *Journal of the American Society for Mass Spectrometry* **24,** 1177–1184 (2013).

20. Townsend, C., Furukawa, A., Schwochert, J., Pye, C., Edmondson, Q. & Lokey, R. S. CycLS: Accurate, whole-librarysequencing of cyclic peptides using tandem mass spectrometry. *Bioorganic & Medicinal Chemistry* **26,** 1232–1238 (2018).

21. Kersten, R. D., Yang, Y.-L., Xu, Y., Cimermancic, P., Nam, S.-J., Fenical, W., Fischbach, M. A., Moore, B. S. & Dorrestein, P. C. A mass spectrometry--guided genome mining approach for natural product peptidogenomics. *Nature chemical biology* **7,** 794–802 (2011).

22. Compeau, P. E. C., Pevzner, P. A. & Tesler, G. How to apply de Bruijn graphs to genome assembly. *Nature Biotechnology* **29,** 987–991 (2011).

23. Takada, K., Ninomiya, A., Naruse, M., Sun, Y., Miyazaki, M., Nogi, Y., Okada, S. & Matsunaga, S. Surugamides A − E, Cyclic Octapeptides with Four D-Amino Acid Residues, from a Marine Streptomyces sp.: LC-MS-Aided Inspection of Partial Hydrolysates for the Distinction of D- and L-Amino Acid Residues in the Sequence. **50,** 3–7 (2013).

24. Frank, A. M., Monroe, M. E., Shah, A. R., Carver, J. J., Bandeira, N., Moore, R. J., Anderson, G. A., Smith, R. D. & Pevzner, P. A. Spectral archives: extending spectral libraries to analyze both identified and unidentified spectra. *Nature Methods* **8,** 587–591 (2011).

25. Bhushan, R. & Bruckner, H. Use of Marfey's reagent and analogs for chiral amino acid analysis: Assessment and applications to natural products and biological systems. *Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences* **879,** 3148–3161 (2011).

26. Dančík, V., Addona, T. A., Clauser, K. R., Vath, J. E. & Pevzner, P. A. De novo peptide sequencing via tandem mass spectrometry. *Journal of Computational Biology* **6,** 327–342 (1999).

27. Jayasena, A. S., Fisher, M. F., Panero, J. L., Secco, D., Bernath-Levin, K., Berkowitz, O., Taylor, N. L., Schilling, E. E., Whelan, J. & Mylne, J. S. Stepwise Evolution of a Buried Inhibitor Peptide over 45 My. *Molecular Biology and Evolution* **34,** 1505–1516 (2017).

28. Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V.,

Vyahhi, N., Tesler, G., Alekseyev, M. a. & Pevzner, P. a. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology* **19,** 455–477 (2012).

29. Fisher, M. F., Zhang, J., Taylor, N. L., Howard, M. J., Berkowitz, O., Debowski, A. W., Behsaz, B., Whelan, J., Pevzner, P. A. & Mylne, J. S. A family of small, cyclic peptides buried in preproalbumin since the Eocene epoch. *Plant Direct* **2,** e00042 (2018).

30. Mylne, J. S., Colgrave, M. L., Daly, N. L., Chanson, A. H., Elliott, A. G., McCallum, E. J., Jones, A. & Craik, D. J. Albumins and their processing machinery are hijacked for cyclic peptides in sunflower. *Nature Chemical Biology* **7,** 257–259 (2011).

31. Elliott, A. G., Delay, C., Liu, H., Phua, Z., Rosengren, K. J., Benfield, A. H., Panero, J. L., Colgrave, M. L., Jayasena, A. S., Dunse, K. M., Anderson, M. A., Schilling, E. E., Ortiz-Barrientos, D., Craik, D. J. & Mylne, J. S. Evolutionary Origins of a Bioactive Peptide Buried within Preproalbumin. *The Plant Cell* **26,** 981–995 (2014).

32. Yazdani, M., Taylor, B. C., Debelius, J. W., Li, W., Knight, R. & Smarr, L. Using machine learning to identify major shifts in human gut microbiome protein family abundance in disease. in *Proceedings - 2016 IEEE International Conference on Big Data, Big Data 2016* 1272–1280 (2016).

33. Noh, H. J., Hwang, D., Lee, E. S., Hyun, J. W., Yi, P. H., Kim, G. S., Lee, S. E., Pang, C., Park, Y. J., Chung, K. H., Kim, G. Do & Kim, K. H. Anti-inflammatory activity of a new cyclic peptide, citrusin XI, isolated from the fruits of Citrus unshiu. *Journal of Ethnopharmacology* **163,** 106–112 (2015).

34. Belknap, W. R., McCue, K. F., Harden, L. A., Vensel, W. H., Bausher, M. G. & Stover, E. A family of small cyclic amphipathic peptides (SCAmpPs) genes in citrus. *BMC Genomics* **16,** 303 (2015).

35. Kaufmann, H. P. & Tobschirbel, A. Über ein oligopeptid aus leinsamen. *European Journal of Inorganic Chemistry* **92,** 2805–2809 (1959).

36. Morita, H., Shishido, A., Matsumoto, T., Takeya, K., Itokawa, H., Hirano, T. & Oka, K. A new immunosuppressive cyclic nonapeptide, cycloinopeptide B from Linum usitatissimum. *Bioorganic and Medicinal Chemistry Letters* **7,** 1269–1272 (1997).

37. Morita, H., Shishido, A., Matsumoto, T., Itokawa, H. & Takeya, K. Cyclolinopeptides B-E, new cyclic peptides from Linum usitatissimum. *Tetrahedron* **55,** 967–976 (1999).

38. Okinyo-Owiti, D. P., Young, L., Burnett, P. G. G. & Reaney, M. J. T. New flaxseed orbitides: Detection, sequencing, and15N incorporation. *Biopolymers - Peptide Science Section* **102,** 168–175 (2014).

39. Gerard, J., Lloyd, R., Barsby, T., Haden, P., Kelly, M. T. & Andersen, R. J. Massetolides

A-H, antimycobacterial cyclic depsipeptides produced by two pseudomonads isolated from marine habitats. *Journal of Natural Products* **60,** 223–229 (1997).

40. Scales, B. S., Dickson, R. P., Lipuma, J. J. & Huffnagle, G. B. Microbiology, genomics, and clinical significance of the Pseudomonas fluorescens species complex, an unappreciated colonizer of humans. *Clinical Microbiology Reviews* **27,** 927–948 (2014).

41. O'Sullivan, D. J. & O'Gara, F. Traits of fluorescent Pseudomonas spp. involved in suppression of plant root pathogens. *Microbiological reviews* **56,** 662–676 (1992).

42. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. MetaSPAdes: A new versatile metagenomic assembler. *Genome Research* **27,** 824–834 (2017).

43. Watrous, J., Roach, P., Alexandrov, T., Heath, B. S., Yang, J. Y., Kersten, R. D., van der Voort, M., Pogliano, K., Gross, H., Raaijmakers, J. M., Moore, B. S., Laskin, J., Bandeira, N. & Dorrestein, P. C. Mass spectral molecular networking of living microbial colonies. *Proceedings of the National Academy of Sciences U.S.A.* **109,** 1743–1752 (2012).

44. Nguyen, D. D., Melnik, A. V., Koyama, N., Lu, X., Schorn, M., Fang, J., Aguinaldo, K., Lincecum, T. L., Ghequire, M. G. K., Carrion, V. J., Cheng, T. L., Duggan, B. M., Malone, J. G., Mauchline, T. H., Sanchez, L. M., Kilpatrick, A. M., Raaijmakers, J. M., De Mot, R., Moore, B. S., Medema, M. H. & Dorrestein, P. C. Indexing the Pseudomonas specialized metabolome enabled the discovery of poaeamide B and the bananamides. *Nature Microbiology* **2,** 1–10 (2016).

45. Bode, H. B., Reimer, D., Fuchs, S. W., Kirchner, F., Dauth, C., Kegler, C., Lorenzen, W., Brachmann, A. O. & Grün, P. Determination of the absolute configuration of peptide natural products by using stable isotope labeling and mass spectrometry. *Chemistry - A European Journal* **18,** 2342–2348 (2012).

46. Wang, M., Wang, J., Carver, J., Pullman, B. S., Cha, S. W. & Bandeira, N. Assembling the Community-Scale Discoverable Human Proteome. *Cell Systems* **7,** 412–421 (2018).

47. Keller, B. O., Sui, J., Young, A. B. & Whittal, R. M. Interferences and contaminants encountered in modern mass spectrometry. *Analytica Chimica Acta* **627,** 71–81 (2008).

48. Behsaz, B., Mohimani, H., Gurevich, A., Prjibelski, A., Fisher, M., Vargas, F., Smarr, L., Dorrestein, P. C., Mylne, J. S. & Pevzner, P. A. De Novo Peptide Sequencing Reveals Many Cyclopeptides in the Human Gut and Other Environments. *Cell Systems* **10,** 99–108 (2020).

CHAPTER 2.

Integrating Metagenomics and Metabolomics for Scalable Non-Ribosomal Peptide Discovery

2.1. ABSTRACT

Non-Ribosomal peptides (NRPs) represent a biomedically important class of natural products that include a multitude of antibiotics and other clinically used drugs. NRPs are not directly encoded in the genome but are instead produced by metabolic pathways encoded by *biosynthetic gene clusters* (BGCs). Since the existing genome mining tools predict many putative NRPs synthesized by a given BGC, it remains unclear which of these putative NRPs are correct and how to identify post-assembly modifications of amino acids in these NRPs in a blind mode, without knowing which modifications exist in the sample. To address this challenge, we developed NRPminer, a modification-tolerant tool for NRP discovery from large (meta)genomic and mass spectrometry datasets. NRPminer identified 59 known and 121 novel NRPs from different environments, including four completely novel NRP families from soil-associated microbes and a novel NRP from human microbiota. We confirmed the anti-parasitic activities and the structure of two of these novel NRP families using direct bioactivity screening and nuclear magnetic resonance spectrometry, thus demonstrating the power of NRPminer for discovering novel bioactive NRPs.

2.2. INTRODUCTION

Non-ribosomal peptides. Microbial natural products represent a major source of bioactive compounds for drug discovery[1]. Among these molecules, *Non-Ribosomal peptides* (*NRPs*) represent a diverse class of natural products that include antibiotics, immunosuppressants, anticancer agents, toxins, siderophores, pigments, and cytostatics[1–4]. NRPs have been reported in various habitats, from marine environments[5] to soil[3] and even human microbiome[6–9]. However, the discovery of novel NRPs remains a slow and laborious process because NRPs are not directly encoded in the genome and are instead assembled by *Non-Ribosomal Peptide Synthetases* (*NRPSs*).

NRPSs are multi-modular proteins that are encoded by a set of chromosomally adjacent genes called *biosynthetic gene clusters* (*BGCs*)[10,11]. Each NRP-producing BGC encodes for one or more genes composed of *NRPS modules.* Together the NRPS modules synthesize the *core NRP* in an assembly line fashion, with each module responsible for adding one amino acid to the growing NRP. Each NRPS module contains an *Adenylation domain* (*A-domain*) that is responsible for recognition and activation of the specific amino acid[12] that can be incorporated by that module through the *non-ribosomal code*[10] (as opposed to the genetic code). At minimum, each NRPS module also includes a *Thiolation domain* (*T-domain*) and a *Condensation domain* (*C-domain*) that are responsible for loading and elongation of the NRP scaffold, respectively. Additionally, an NRPS module may include additional domains such as *Epimerization domain* (*E-domain*) or dual-function *Condensation/Epimerization domain* (*C/E domain*). An *NRPS assembly line* refers to a sequence of NRPS modules that together assemble a core NRP. The core NRP often undergoes post-assembly modifications (PAMs) that transform it into a *mature NRP*. The order of the modules in an NRPS assembly line can be different from the order of NRPS modules encoded in the BGC through iterative use of NRPS modules.

NRP discovery via genome mining. In the past decade, genome mining methods have been developed for predicting the NRP sequences from their BGC sequences[13,14]. Genome mining tools, such as antiSMASH[15], start by identifying the NRPS BGCs in a microbial genome using Hidden Markov Models (HMMs). Afterwards, they identify NRPS modules and predict the amino acids incorporated by the A-domain in each module using the *substrate prediction* algorithms (such as NRPSpredictor2[13] or SANDPUMA[16]) that are based on machine learning techniques trained on a set of A-domains with known specificities[14,16]. For each observed A-domain, these algorithms predict a set of amino acids potentially recruited by that A-domain, along with the *specificity score* reflecting confidence of each amino acid prediction. The use of genome mining is becoming increasingly popular for discovering novel NRPsover the past decade[17–19], demonstrating the potential of (meta)genomic projects for NRP discovery.

Although genome mining tools like SMURF[20] and antiSMASH[15] greatly facilitate BGC analysis, the core NRPs (let alone mature NRPs) for the vast majority of sequenced NRP-producing BGCs (>99%) remain unknown[21,22]. Identification of NRP-producing BGCs without revealing the molecular products encoded by these BGCs does not capture their full potential for discovering novel NRPs[23]. Thus, integrating (meta)genome mining with metabolomics is necessary for realizing the true promise of large-scale NRP discovery[4]. However, the existing genome mining strategies fail to reveal the chemical diversity of NRPs. For example, these methods fall short in correctly identifying PAMs, which are a unique feature of NRPs that make them the most diverse class of natural products[24] and play a crucial role in their mode of action[25,26]. As a result, the promise of large-scale NRP discovery has not yet been realized[27].

Computational challenges in NRP discovery. Discovery of novel NRPs involves a multitude of challenges such as PAM identification (with exception of methylation and epimerization[15], genome mining tools fail to identify PAMs) and accounting for *substrate promiscuity* of A-domains. The *substrate promiscuity* in NRP biosynthesis refers to the ability of an A-domain in an NRPS to incorporate several different amino acids into the NRP. The existing genome mining tools often predict a set of incorporated amino acids and output a ranked list of multiple amino acids for each A-domain. Allowing for all amino acid possibilities for each A-domain in an NRPS module results in a large number of putative NRPs predicted from each BGC. Without additional complementary data (such as mass spectra of NRPs), the genome mining approaches cannot identify the correct NRP among the multitude of putative NRPs[27,28].

Another challenge in discovering novel NRPs is due to the *non-canonical assembly lines*. While in many NRPSs each A-domain incorporates exactly one designated amino acid and the sequence of amino acids in NRP matches the order of the A-domains in BGC[29–31] (see Figure 2.1.a), several studies revealed that many NRP families violate this pattern[7,29,31–37]. Since an NRPS may have multiple assembly lines[38], one needs to consider different combinations of *NRPS units* encoded by each *open reading frames* (*ORFs*) for finding the core NRPs[25,38]. In some non-canonical assembly lines, A-domains encoded by at least one ORF may be incorporated multiple times (in tandem) in the NRPS (Figure 2.1.b)[7,33–35]. For example, during biosynthesis of rhabdopeptides[33,37] and lugdunin[7], a single ORF encoding only one *Val*-specific A-domain loads multiple *Val* in the final NRPs. Moreover, in some NRPS assembly lines, the A-domains in some ORFs do not contribute to the core NRP[31,36,39] (see Figure 2.1.c). For example, surugamide BGC[31,40,41,46] from *Streptomyces albus* produces two completely distinct NRPs through different non-canonical assembly lines. The non-canonical biosynthesis of surugamide makes its discovery particularly difficult as one need to account for these non-canonical assembly lines by generating different combinations of ORFs in the process of building a database of putative NRPs for each BGC.

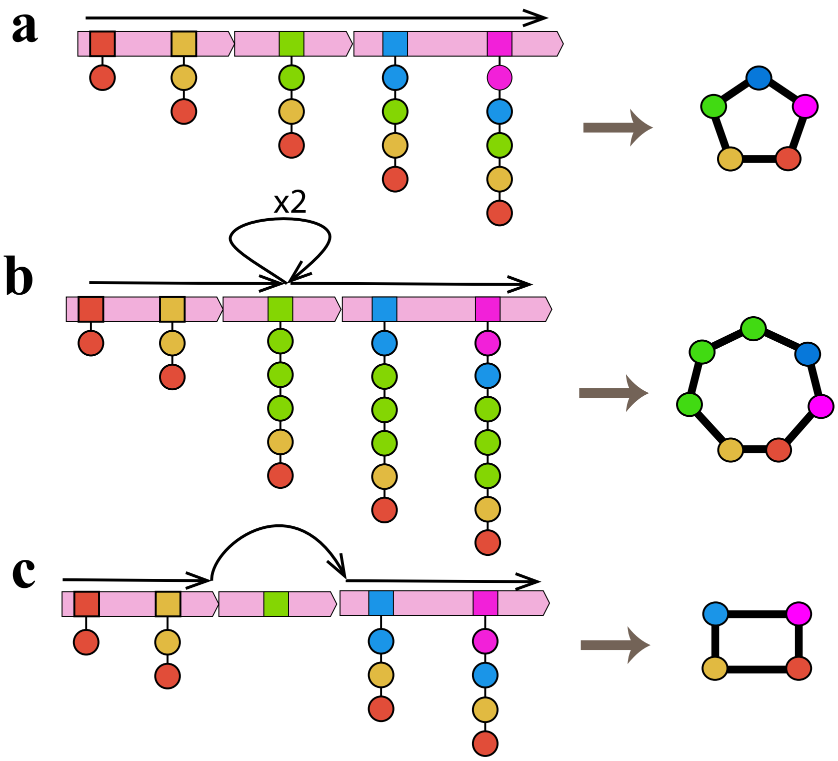


Figure 2.1. Schematic examples of canonical and non-canonical NRPS assembly lines. Squares represent A-domains and circles represent amino acids (different amino acids are shown by different colors). Each amino acid is colored by the same color as the corresponding A-domain. In each panel, the final NRP is represented by its amino acids with amide bonds shown with black lines. (a) A canonical assembly line where each A-domain adds one amino acid to the growing structure. (b) A non-canonical assembly line where a single A-domain (on one ORF) loads a series of three amino acids (the loop shows the repeat of A-domain on the assembly line) to the growing structure also referred to as stuttering in polyketide synthases[43,44]. (c) A non-canonical assembly line where the A-domain appearing on one ORF is skipped in the final NRP.

Other hurdles include lack of sufficient training data for many A-domains, which can lead to specificity mispredictions[16] and complications in the genome mining due to fragmented assemblies (e.g. failure to capture a BGC in a single contig[45]). These challenges, in combination with those mentioned above, make it nearly impossible to accurately predict NRPs based solely on genome mining. The problem gets even more severe for NRP discovery from microbial communities.

Peptidogenomics approaches to NRP discovery. To address these challenges, multiple peptidogenomics approaches have been developed for discovering novel peptidic natural products by combining genome mining and MS information[28,46]. These approaches often use antiSMASH[14] to find all NRPS BGCs in the input genome, use NRPSPredictor2[13] to generate putative core NRPs encoded by each BGC, and attempt to match mass spectra against these putative NRPs. Kersten *et al.*[46] described a peptidogenomics approach based on manually inferring *amino acid sequence tags* (that represent a partial sequence of an NRP) from mass spectra and matching these tags against information about the substrate specificity generated by NRPSpredictor2[13]. Nguyen et al.[47,48] and Tobias *et al.*[30] presented a manual approach for combining genome mining with molecular networking. In this approach, which is limited to the identification of novel variants of known NRPs, molecules present in spectral families with known compounds are compared to BGCs.

Medema *et al.*[38] complemented the manual approach from Kersten *et al.*[46] by the NRP2Path[38] tool for searching the sequence tags against a collection of BGCs. NRP2Path starts with a set of sequence tags manually generated for each spectrum, considers multiple assembly lines for each identified BGC, and forms a database of all possible core NRPs for this BGC. Then, NRP2Path[38] computes a *match score* between each tag and each core NRP (using the specificity scores provided by NRPSpredictor2[13]) and reports high-scoring matches as putative core NRPs. The success of this approach relies on inferring long sequence tags of 4-5 amino acids, which are usually absent in spectra of non-linear peptides. Such long sequence tags are often missing in NRPs with macrocyclic backbones and complex modifications, limiting the applicability of

NRP2Path[46,49]. Moreover, NRP2Path is not able to identify enzymatic modifications (e.g. methylation) and PAMs in the final NRPs and is unable to predict the backbone structure of the mature NRPs (e.g. linear/cyclic/branch-cyclic).

Mohimani *et al.*[28] developed an automated NRPquest approach that takes paired MS and genomic datasets as input and searches each mass spectrum against all structures generated from putative core NRPs to identify high-scoring *Peptide-Spectrum Matches* (*PSMs*). NRPquest leverages the entire mass spectrum (instead of just the sequence tags) to provide further insights into the final structure of the identified NRPs. They proposed using modification-tolerant search of spectral datasets against the core NRPs structures, for identifying PAMs in a blind mode (that is without knowing which PAMs exists in the sample). This is similar to identifying post-translational modifications in traditional proteomics[50]. The presence of covalent modifications in peptides affects the molecular weight of the modified amino acids, therefore, the mass increment or deficit can be detected using MS data[41,50]. However, as NRPquest uses a naïve pairwise scoring of all NRP structures against all mass spectra for PAM identification, it is prohibitively slow when searching for PAMs[28]. Furthermore, NRPquest does not handle non-canonical NRPS assembly lines and it does not provide statistical significance of identified NRPs, a crucial step for large-scale analysis.

On the other hand, development of high-throughput MS-based experimental and computational natural products discovery pipelines[27] such as the Global Natural Products Social (GNPS) molecular networking[51], PRISM[52], GNP[53], RODEO[54] , Dereplicator+[55], CSI:FingerID[56], NAP[57] and CycloNovo[49] have permanently changed the field of peptide natural product discovery. GNPS project already generated nearly half a billion of information-rich tandem mass spectra (MS), an untapped resource for discovering new molecules. However, the utility of the GNPS network is mainly limited to the identification of previously discovered molecules and their analogs. Currently, only about 5% percent of the GNPS spectra are annotated[51], emphasizing the need for novel algorithms, like NRPminer, for annotating such large spectral datasets.

To address these shortcomings, we developed NRPminer, a scalable modification-tolerant tool for analyzing paired MS and (meta)genomic datasets (Figure 2.2). Unlike NRPquest, NRPminer uses the specificity scores of the amino acids appearing in core NRPs to perform an efficient search of all spectra against all core NRPs. In addition to predicting the amino acid sequence of an NRP generated by a BGC, NRPminer also analyzes various non-canonical assembly lines and efficiently predicts potential PAMs and backbone structures. Currently NRPminer is limited to the prediction of total mass and site of modification (rather than complete chemical structure) of PAMs. Such information allows researchers to characterize known PAMs based on their total mass, making it possible for researchers to prioritize NRPs with novel PAMs. After searching only four MS datasets in GNPS against their corresponding reference genomes, NRPminer discovered 180 NRPs representing 18 distinct NRP families, including four novel NRP families from *Amycolatopsis* and *Xenorhabdus* species.

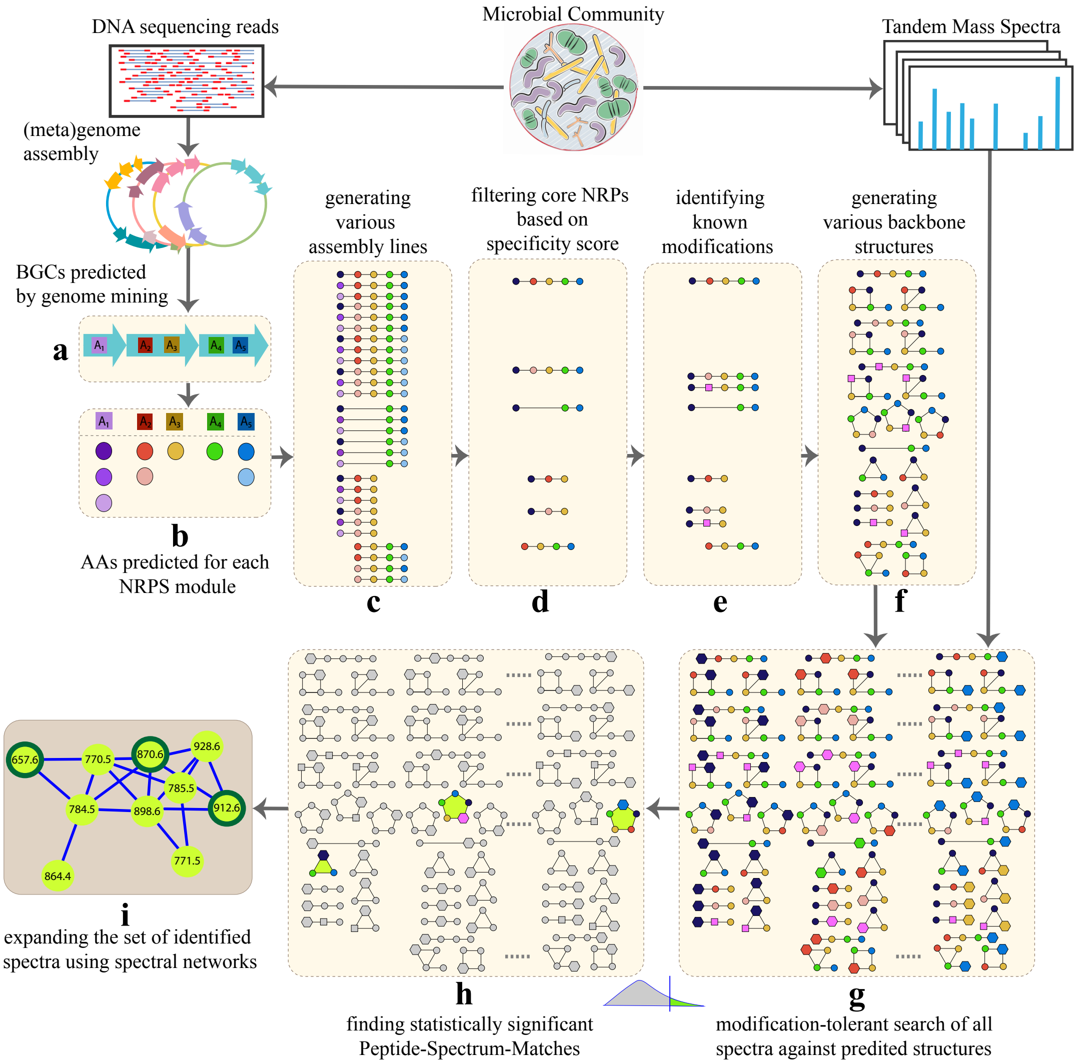

DNA sequencing reads
Microbial Community
Tandem Mass Spectra
(meta)genome assembly
BGCs predicted by genome mining
a
A1
A2
A3
A4
A5
b
AAs predicted for each NRPS module
generating various assembly lines
c
filtering core NRPs based on specificity score
d
identifying known modifications
e
generating various backbone structures
f
g
modification-tolerant search of all spectra against predited structures
h
finding statistically significant Peptide-Spectrum-Matches
i
expanding the set of identified spectra using spectral networks
657.6
770.5
870.6
928.6
785.5
784.5
898.6
912.6
771.5
864.4

Figure 2.2. NRPminer pipeline. (a) Predicting NRPS BGCs using antiSMASH[14]. Each ORF is represented by an arrow, and each A-domain is represented by a square, (b) predicting putative amino acids for each NRP residue using NRPSpredictor2[13], colored circles represents different amino acids, (c) generating multiple assembly lines by considering various combinations of ORFs and generating all putative core NRPs for each assembly line in the identified BGC (for brevity only assembly lines generated by deleting a single NRPS unit are shown; in practice, NRPminer considers loss of up to two NRPS units, as well as single and double duplication of each NRPS unit), (d) filtering the core NRPs based on their specificity scores, (e) identifying domains corresponding to known modifications and incorporating them in the selected core NRPs (modified amino acids are represented by purple squares), (f) generating linear, cyclic and branch-cyclic backbone structures for each core NRP, (g) generating a set of high-scoring PSMs using modification-tolerant VarQuest[41] search of spectra against the database of the constructed putative NRP structures. NRPminer considers all possible mature NRPs with up to one PAM (shown as hexagons) in each NRP structure. For brevity some of the structures are not shown. (h) computing statistical significance of PSMs and reporting the significant PSMs, and (i) expanding the set of identified spectra using spectral networks[58]. Nodes in the spectral network represent spectra and edges connect "similar" spectra (see Methods).

2.3. RESULTS

Outline of the NRPminer algorithm. Figure 2.2 illustrates the NRPminer algorithm. All these steps are described in detail in the Methods section which includes: (*a*) NRPminer starts by identifying the NRPS BGCs in each genome (using antiSMASH[14]), (b) predicting the putative amino acids for each identified A-domain (using NRPSpredictor2[13]), (c) NRPminer accounts for different NRPS assembly-lines by considering various combinations of ORFs in the BGCs, (d) NRPminer filters the set of all core NRPs based on the specificity scores of their amino acids and selects those with high scores, (*e*) NRPminer searches each BGC to find known modification enzymes and incorporates them in the corresponding core NRPs, (*f*) It then constructs a database of putative NRP structures by considering linear, cyclic and branch-cyclic backbone structures for each core NRP, (*g*) NRPminer performs a modification-tolerant search of the input spectra against the constructed database of putative NRPs and computes the statistical significance of Peptide-Spectrum-Matches (PSMs), (*h*) NRPminer reports the statistically significant PSMs, and (*i*) which are then expanded using spectral networks[58] approach.

Datasets. We analyzed four microbial isolate datasets from *Xenorhabdus* and *Photorhabdus* families (***XPF***), *Staphylococcus* (***SkinStaph*)**, soil dwelling and soil dwelling *Actinobacteria* (***SoilActi***), and a collection of soil-associated bacteria within *Bacillus*, *Pseudomonas*, *Buttiauxella*, and *Rahnella* genera generated under the Tiny Earth antibiotic discovery project[59] (***TinyEarth***); all available from GNPS/MassIVE repository. The spectra collected on each of these datasets are referred to as $spectra_{XPF}$, $spectra_{SkinStaph}$, $spectra_{SoilActi}$, $spectra_{TinyEarth}$, and the genomes are referred as $genome_{XPF}$, $genome_{SkinStaph}$, $genome_{SoilActi}$, and $genome_{TinyEarth}$, respectively. Below we describe sample preparation and mass spectra generation for all analyzed datasets.

XPF: A total of 27 strains from soil nematode symbiont *Xenorhabdus* and *Photorhabdus* families were grown in lysogeny broth and agar and were extracted with methanol as described previously[30]. Briefly, the crude extracts were diluted 1:25 (vol/vol) with methanol and analyzed by UPLC-ESI coupled with Impact II qTof mass spectrometer. MS dataset *spectra*$_{XPF}$[30] contains 27 spectral sub-datasets representing each sample for a total of 263,768 spectra across all strains (GNPS-accession #: MSV00081063). The *genome*$_{XPF}$ dataset contains 27 draft genomes generated by DNA sequencing from the same samples as reported by Tobias et al.[30] (available from RefSeq[60]).

SkinStaph*:* A total of 171 *Staphylococcus* strains isolated from skin of healthy individuals were grown in 500 ml Tryptic Soy Broth (TSB) liquid medium in Nunc 2.0 mL DeepWell plates (Thermo Catalog# 278743). An aliquot of each culture was used to measure optical density. Cultures that effectively grew were transferred to a new deep well plate. Cultures were placed in a -80C freezer for 10 min then allowed to thaw at room temperature 3 times, to lyse bacterial cells. 200 ul of the supernatant collected from cell cultures were filtered using a Phree Phospholipid Removal kit (Phenomenex). Sample clean up was performed following the manufacturer protocol described here (https://phenomenex.blob.core.windows.net/documents/c1ac3a84-e363-416e-9f26-f809c67cf020.pdf). Briefly, the Phree kit plate was conditioned using 50% MeOH, bacterial supernatant were then added to the conditioned wells followed by sample clean up using 100% MeOH (a 4:1 v/v ratio of MeOH:Bacterial supernatant). The plate was centrifuged 5 min at 500 g and the clean up extracts were lyophilized using a FreeZone 4.5 L Benchtop Freeze Dryer with Centrivap Concentrator (Labconco). Wells were resuspended in 200 µL of resuspension solvent (80% MeOH spiked with 1.0 µM Amitriptyline), vortexed for 1 min, and centrifuged at 2000 rpm for 15 min at 4 °C. 150 µL of the supernatant was transferred into a 96-well plate and maintained at 20 °C prior to LC-MS/MS analysis. Bacterial extracts were analyzed using a ThermoScientific UltiMate 3000 UPLC system for liquid chromatography and a Maxis Q-TOF (Quadrupole-Time-of-Flight) mass spectrometer (Bruker Daltonics), controlled by the Otof Control and Hystar software packages (Bruker Daltonics) and equipped with ESI source. Untargeted metabolomics data were collected using a previously validated UPLC-MS/MS method[61,62]. The $spectra_{SkinStaph}$ dataset contains 2,657,398 spectra from bacterial extracts of 171 *Staphylococcus* strains (GNPS-accession #: MSV000083956). The $genome_{SkinStaph}$ dataset contains draft genomes of these species (available from RefSeq).

SoilActi*:* A total of 20 strains of soil-dwelling *Actinobacteria* were grown on A1, MS, and R5 agar, extracted sequentially with ethyl acetate, butanol, methanol and analyzed on Agilent 6530 Accurate-Mass QTOF spectrometer coupled with Agilent 1260 LC System. The $spectra_{SoilAct}$ dataset contains 362,421 spectra generated from extracts of these 20 *Actinobacteria* strains (GNPS-accession #: MSV000078604[63]) includes 20 sub-datasets representing each strain. The $genome_{SoilActi}$ dataset contains draft genomes of these strains (available via RefSeq).

TinyEarth*:* A total of 23 bacterial strains extracted from the soil in Wisconsin were extracted with methanol and analyzed by LC-MS/MS on a Thermo Fisher Q-Exactive mass spectrometer. The $spectra_{TinyEarth}$ dataset contains 380,414 spectra generated from extracts of these 23 strains (GNPS-accession #: MSV000084951) includes 23 sub-datasets representing each strain (four *Bacillus*, 16 *Pseudomonas*, one *Buttiauxella*, and one *Citrobacter*) The $genome_{TinyEarth}$ dataset contains draft genomes of these strains (available via Gold OnLine Database[64] under study ID Gs0135839).

Summary of NRPminer results. Table 2.1 summarizes the NRPminer results for each dataset. NRPminer classifies a PSM as statistically significant if its p-value is below the default conservative threshold 10^{-15}. The number of distinct NRPs and NRP families was estimated using MS-Cluster[65] and SpecNets[51] using the threshold *cos* < 0.7 (see Methods section). Two peptides are considered to be variants/modifications of each other if they differ in a single modified residue due to changes by tailoring enzymes, enzyme promiscuity or through changes in the amino acid specificity at the genetic level[48]. Known NRPs (NRP families) are identified either by Dereplicator[40] search against the database of all known peptidic natural products[41] (referred to as *PNPdatabase*) using the p-value threshold 10^{-15}, and/or by SpecNet[58] search against the library of all annotated spectra available on GNPS[51]. NRPminer ignores any BGCs with less than three A-domains and spectra that include less than 20 peaks.

Table 2.1. Summary of NRPminer search results on the XPF, SkinStaph, SoilActi, and datasets. Column "#strains" shows the number of microbial strains. Column "#identified PSMs/ #spectra" shows the number of PSMs identified by NRPminer and the total number of spectra. The column "#distinct NRPs (families)" shows the number of unique NRPs (unique families). The number of unique NRPs is estimated using MS-Cluster[65], and the number of unique families is estimated using SpecNets[51]. The column "#known NRPs (families)" shows the number of known NRPs (families) among all identified NRPs (families). Column "#novel variants of known NRPs" shows the number of NRPs in the known families that were identified as a novel variant. Column "#novel NRPs (families)" shows the number of novel NRPs (families).

dataset	#strains	#identified PSMs/ #spectra	#distinct NRPs (families)	#known NRPs (families)	#novel variants of known NRPs	#novel NRPs (families)
XPF	27	3,023 / 263,768	122 (12)	21 (9)	79	22 (3)
SkinStaph	171	23 / 2,657,398	3 (1)	2 (1)	1	0
SoilActi	20	206 / 362,421	24 (2)	7 (1)	14	3 (1)
TinyEarth	28	498 / 380,414	31 (3)	29 (3)	2	0

Generating putative core NRPs. Table 2.2 presents the number of NRP-producing BGCs and the number of putative core NRPs generated by NRPminer for each analyzed genome in XPF (before and after filtering). For example, NRPminer identified eight NRP-producing BGCs and generated 253,027,076,774 putative core NRPs for *X. szentirmaii* DSM genome. After filtering putative core NRPs based on the sum of the specificity scores reported by NRPSpredictor2[13], only 29,957 putative core NRPs were retained (see Methods section for the details of filtering). Therefore, filtering putative core is an essential step for making the search feasible.

Table 2.2. The number of predicted core NRPs before and after filtering for 27 genomes in the XPF dataset. The column "#NRP producing BGCs" show the number of NRP-producing BGCs. Columns under "#unique core NRPs" show the number of core NRPs generated by NRPminer before and after filtering for each genome. For example, in the case of the *X. szentirmaii* DSM genome with 8 NRP-producing BGCs, NRPminer considers 253,027,076,774 core NRPs before filtering, while after filtering only 57,888 cores are retained. The five species corresponding to the datasets yielding the novel NRP families are shown in blue.

Strain	**#NRP producing BGCs**	**#unique core NRPs**	
		before filtering	**after filtering**
Xenorhabdus bovienii SS-2004	8	8,973,905	7,701
Xenorhabdus nematophila ATCC	6	18,043,657,358	18,062
Xenorhabdus doucetiae FRM16	8	3,726,625,228	8,013
Xenorhabdus poinarii G6	6	14,280	658
Photorhabdus luminescens PB45.5	10	2,994,745,388,283	8,333
Photorhabdus asymbiotica PB68.1	8	157,964	2,602
Xenorhabdus sp. DL20	9	94,818	2,187
Xenorhabdus sp. 30TX1	8	76,044,111	7,287
Xenorhabdus vietnamensis	15	3,373,109,836	21,648
Xenorhabdus beddingii DSM 4764	8	13,721,302	2,998
Photorhabdus temperata	9	42,555,972,979,030	6,924
Photorhabdus asymbiotica PB68.1	8	160,034	5,136
Xenorhabdus budapestensis 16342	7	149,918,342	51,600
Xenorhabdus ehlersii DSM 16337	10	5,026,725	7,542
Xenorhabdus innexi DSM 16336	10	4,957,948,632	9,184
Xenorhabdus szentirmaii US	8	360,039,991,874	57,888
Xenorhabdus mauleonii	10	51,502,147,078	19,400
Xenorhabdus miraniensis	14	11,679,221,261	14,658
Xenorhabdus szentirmaii	8	253,027,076,774	57,888
Xenorhabdus sp. KK7.4	9	5,036,899,357	17,300
Xenorhabdus hominickii DSM	13	60,224,436	6,688
Xenorhabdus stockiae DSM 17904	10	1,159,012,484,964	7,896
Xenorhabdus ishibashii	7	19,911,786	2,547
Xenorhabdus sp. KJ12.1	10	11,916,878,760	10,458
Xenorhabdus kozodoi DSM 17907	11	87,750	2,192
Xenorhabdus cabanillasii JM26	9	80,529,848	47,856
Photorhabdus temperata	11	567,909,518,582	4,823

Analysis of the paired genomic and spectral datasets. NRPminer has a *one-vs-one* mode (each MS dataset is searched against a single genomic dataset) and a *one-vs-all* mode (each MS dataset is searched against a collection of genomic datasets within a taxonomic clade). While the one-vs-all mode is slower than the one-vs-one mode, it is usually more sensitive. For example, a BGC may be fragmented (or misassembled) in the draft assembly of one strain, but a related BGC may be correctly assembled and captured within a single contig in a related well-assembled strain. If these two BGCs synthesize the same (or even similar) NRP, NRPminer may be able to match the spectra from a poorly assembled strain to a BGC from a related well-assembled strain.

For example, NRPminer search of $spectra_{XPF}$ against $genome_{XPF}$ generated 3,023 PSMs that represent 122 NRPs from 12 NRP families. Figure 2.3 shows the spectral network representing 12 NRP families identified by NRPminer in the XPF dataset. SpecNet analysis against the annotated spectra in GNPS[51] showed that 9 out of 12 identified NRP families are known (reported by Tobias *et al.*[30]). NRPminer failed to identify only one additional known family which was reported by Tobias *et al.*[30] (xefoampeptides) that has a side-chain modification with total mass exceeding the default NRPminer threshold (150 Da). Xefoampeptides contain only three amino acids and a large side-chain modification (total mass over 200 Da), resulting in a poorly fragmented spectrum that did not generate statistically significant PSMs against the putative structures generated from their corresponding core NRPs.

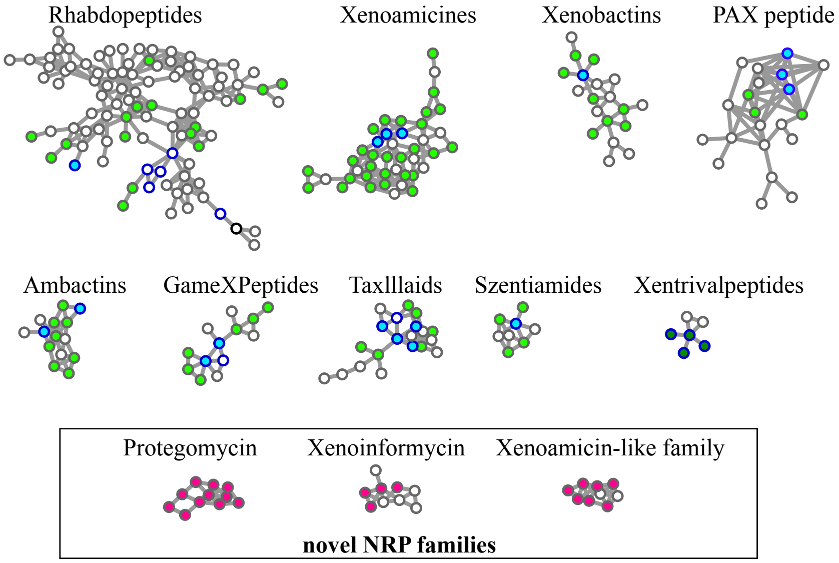


Figure 2.3. Spectral networks for nine known and three novel NRP families identified by NRPminer in the XPF dataset. Each node represents a spectrum. The spectra of known NRPs (as identified by spectral library search against the library of all known compounds in GNPS) are shown with a dark blue border. A node is colored if the corresponding spectrum forms a statistically significant PSM and not colored otherwise. We distinguish between identified spectra of known NRPs with known BGCs[30] (colored by light blue) and identified spectra of known NRPs (from xentrivalpeptide family) with previously unknown BGC (colored by dark green). Identified spectra of novel NRPs from known NRP families (novel NRP variants) are colored in light green. Identified spectra of novel NRPs from novel NRP families are colored in magenta. Proteogomycins and xenoinformycin subnetworks represent previously unreported NRP families with novel putative BGCs. The Xenoamicin-like subnetwork revealed a BGC family distantly related to xenoamicins (6 out 13 amino acids are identical). For simplicity only spectra at charge state +1 are used for the analysis.

Table 2.3 provides information about NRPminer-generated PSMs representing known NRP families. Among the nine known NRP families (in the XPF dataset) listed in Table 2.3, eight families have been connected to their BGCs in the previous studies, and for these families, the corresponding BGCs discovered by NRPminer are consistent with the literature[30] (see Table 2.3 for the list of identified BGCs).

Table 2.3. PSMs identified by NRPminer in the XPF dataset representing the known NRP families. For each NRP family, the information about the PSM with the lowest p-value among all PSMs corresponding to the spectra representing the known NRPS in that family is listed. The column "matched genome" shows the name of the organism whose BGCs generated the putative NRP structure corresponding to that PSM and the column "BGC position" shows the contig and the starting and ending nucleotide position of the BGC in that contig. Columns "precursor mass" and "charge" show the precursor mass and the charge state of matched spectrum.

NRP family name	matched genome	BGC position	p-value	precursor mass	charge
GameXPeptide	*Photorhabdus asymbiotica* PB68.1	ctg1: 3584973 - 3640476	1.5×10^{-25}	586.394	1
PAX peptide	*Xenorhabdus nematophila* ATCC 19061	ctg1: 11609 - 67919	9.9×10^{-18}	826.538	1
Xenobactin	*Xenorhabdus mauleonii* DSM 17908	ctg11: 65321 - 162527	5.0×10^{-21}	756.425	1
Szentiamide	*Xenorhabdus szentirmaii* DSM 16338	ctg1: 762001 - 821352	7.0×10^{-31}	838.404	1
Taxlllaid	*Xenorhabdus bovienii* SS-2004	ctg1: 739318 - 804275	1.2×10^{-30}	808.55	1
Xentrivalpeptide	*Xenorhabdus* sp. KK7.4	ctg14: 6760-112451	6.4×10^{-37}	430.749	2
Ambactin	*Xenorhabdus miraniensis* DSM 17902	ctg6: 132143-191993	5.4×10^{-16}	751.41	1
Xenoamicin	*Xenorhabdus vietnamensis* DSM 22392	ctg9: 1-75156	3.3×10^{-56}	1300.8	1
Rhabdopeptide	*Xenorhabdus stockiae* DSM 17904	ctg14: 1-77935	6.1×10^{-17}	599.427	1

Figure 2.4 presents an example of an identified NRP family, *szentiamide*, and its corresponding BGC in *X. szentirmaii*. For one family (*xentrivalpeptides*) we report the BGC for the first time (Figure 2.5). In addition to these known families, NRPminer also discovered four novel NRP families and 79 novel NRP variants in this dataset.

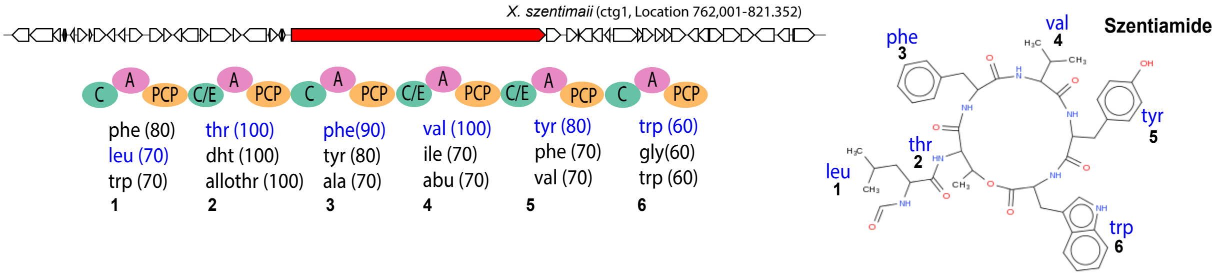


Figure 2.4. Szentiamide biosynthetic gene clusters. (Left) szentiamide BGC in *Xenorhabdus szentirmaii* DSM 16338 with NRPS genes (shown in red) which is consistent with the previous study[66]. Three highest scoring NRPSpredictor2[13] amino acid predictions for each A-domain in these BGC are shown. Amino acids corresponding to the correct structure are shown in blue. NRPminer identified this NRP with p-value $7.0\text{x}10^{-31}$. (Right) The structure of the szentiamide is shown with amino acids highlighted in blue.

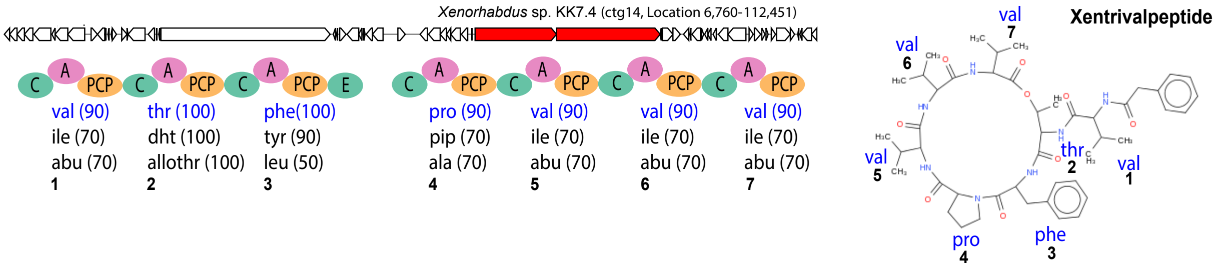
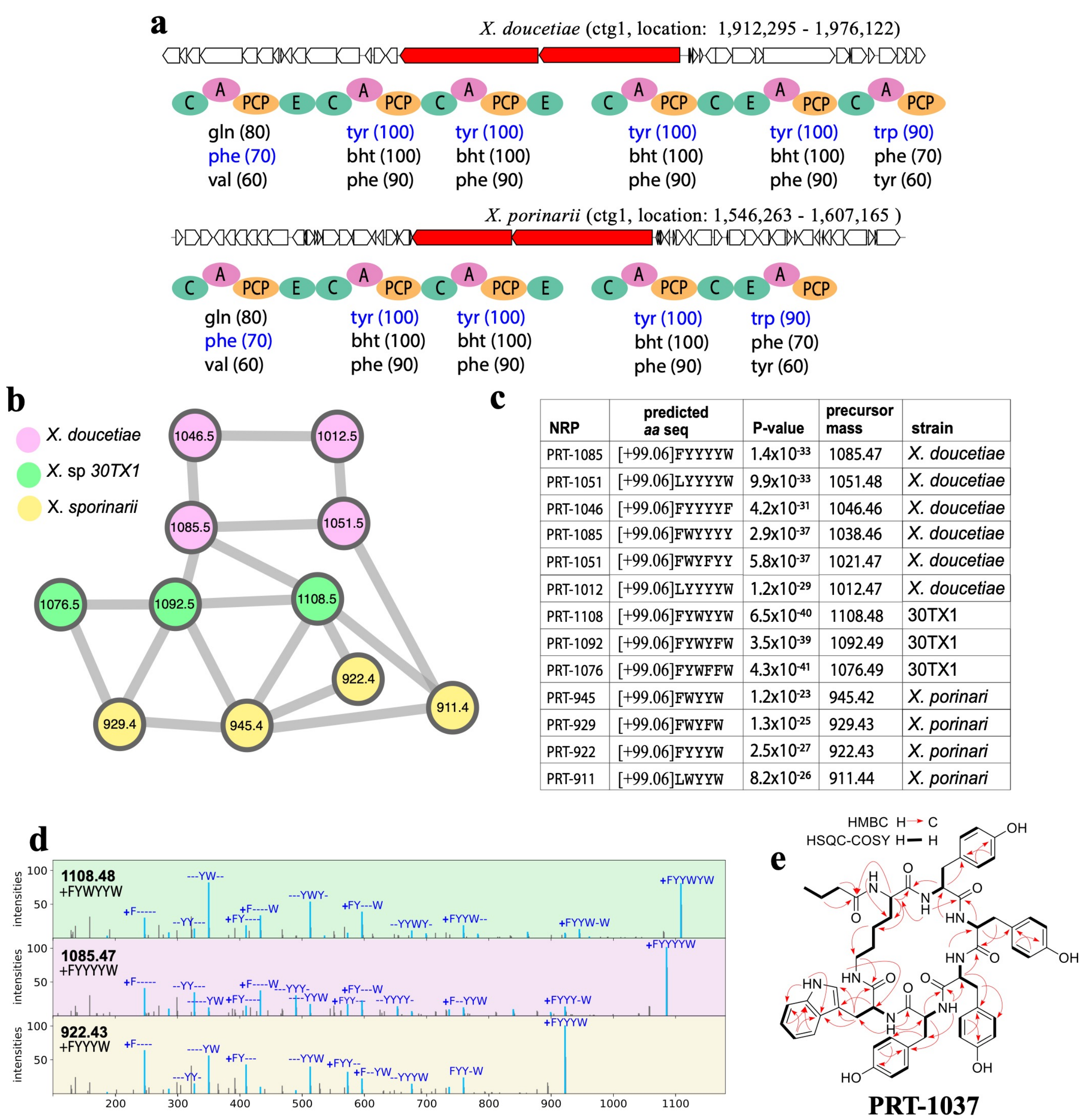


Figure 2.5. Predicted xentrivalpeptides biosynthetic gene clusters. (**Left**) The BGC in *Xenorhabdus* sp. KK7.4 predicted to encode xentrivalpeptide with NRPS genes (shown in red). Three highest scoring NRPSpredictor2[13] amino acid predictions for each A-domain in these BGCs are shown. Amino acids corresponding to the correct structure are shown in blue. NRPminer identified this NRP with p-value $6.4x10^{-37}$. (**Right**) The structure of the xentrivalpeptide is shown with amino acids highlighted in blue.

We named each identified NRP in a reported novel family by combining the name of that family with the nominal precursor mass of the spectrum representing that NRP (with the lowest p-value among all spectra originating from the same NRP). In what follows, we describe the four novel NRP families identified by NRPminer (protegomycin, xenoinformycin, and novel xenoamicin-like family in the XPF dataset and aminformatide in SoilActi), as well as the novel variants in two additional NRP families (lugdunin in SkinStaph and surugamide in SoilActi).

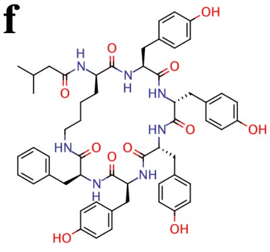
Novel protegomycin (PRT) NRP family in the XPF dataset. NRPminer matched 28 spectra representing 11 novel cyclic NRPs to two previously unreported BGCs. These spectra are from species *X. doucetiae*, *Xenorhabdus* sp. 30TX1, and *X. poinarii*. The BGCs were from in *X. doucetiae* and *X. poinarii* with six and five A-domains, respectively, with one PAM. Figure 2.6 present information about protegeomycin BGC and NRPs.

Figure 2.6. Novel protegomycin NRP family. (*a*) The BGCs generating the NRP in *X. doucetiae* (top) and *X. porinarii* (bottom) along with NRPS genes (shown in red) and A-, C-, PCP-, and E-domains in these NRPSs. The rest of the genes in the corresponding contigs are shown in white. No BGC was found in *Xenorhabdus* sp. 30TX1. Three highest-scoring amino acids for each A-domain in these BGCs (according to NRPSpredictor2[13] predictions) are shown below the corresponding A-domains. Amino acids appearing in the NRPs [+99.06]FYYYYW and [+99.06]FYYYW identified by NRPminer (with the lowest p-value) are shown in blue. (*b*) Spectral network formed by the spectra that originate from NRPs in the protegomycin family. (*c*) Sequences of the identified NRPs in the protegomycin family (with the lowest p-value among all spectra originating from the same NRP). PRT represents protegomycin. (*d*) For each strain, an annotated spectrum representing the lowest p-value is shown. The spectra were annotated based on predicted NRPs [+99.06]FYYWYW, [+99.06]FYYYYW, and [+99.06]FYYYW from top to bottom. The "+" sign represents the addition of [+99.06Da]. Colors in parts *b* and *d* are coordinated. Figures 2.7-2.9 show the annotated spectra for all NRPs shown in part *c*. (*e*) Key HMBC and HSQC-COSY correlations in PRT-1037. (*f*) Structures for selected PRT derivatives produced by *X. doucetiae* including amino acid configuration as concluded from the presence of epimerization domains in the corresponding NRPSs and acyl residues as concluded from feeding experiments).

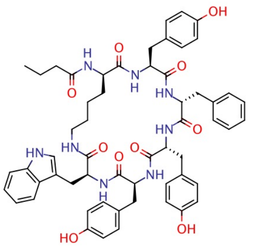

a
X. doucetiae (ctg1, location: 1,912,295 - 1,976,122)
C A PCP E C A PCP C A PCP E C A PCP C E A PCP C A PCP
gln (80) phe (70) val (60)
tyr (100) bht (100) phe (90)
tyr (100) bht (100) phe (90)
tyr (100) bht (100) phe (90)
tyr (100) bht (100) phe (90)
trp (90) phe (70) tyr (60)
X. porinarii (ctg1, location: 1,546,263 - 1,607,165)
C A PCP E C A PCP C A PCP E C A PCP C E A PCP
gln (80) phe (70) val (60)
tyr (100) bht (100) phe (90)
tyr (100) bht (100) phe (90)
tyr (100) bht (100) phe (90)
trp (90) phe (70) tyr (60)
b
X. doucetiae
X. sp 30TX1
X. sporinarii
1046.5
1012.5
1085.5
1051.5
1076.5
1092.5
1108.5
922.4
911.4
929.4
945.4
c
d
intensities
100
50
1108.48
+FYWYYW
1085.47
+FYYYYW
922.43
+FYYYW
+F-----
---YW--
--YY---
+FY----
+F----W
---YWY-
+FY---W
--YYWY-
+FYYW--
+FYYW-W
+FYYWYW
+FYYYYW
--YY---
----YW
+FY---
+F----W
---YYY-
---YYW
+FYY--
+FY---W
--YYYY-
+F--YYW
+FYYY-W
+F----
----YW
---YY-
+FY---
---YYW
+FYY--
+F--YW
--YYYW
FYY-W
200
300
400
500
600
700
800
900
1000
1100
e
HMBC H → C
HSQC-COSY H — H
OH
HO
HN
NH
O
PRT-1037

NRP	predicted *aa* seq	P-value	precursor mass	strain
PRT-1085	[+99.06]FYYYYW	$1.4x10^{-33}$	1085.47	*X. doucetiae*
PRT-1051	[+99.06]LYYYYW	$9.9x10^{-33}$	1051.48	*X. doucetiae*
PRT-1046	[+99.06]FYYYYF	$4.2x10^{-31}$	1046.46	*X. doucetiae*
PRT-1085	[+99.06]FWYYYY	$2.9x10^{-37}$	1038.46	*X. doucetiae*
PRT-1051	[+99.06]FWYFYY	$5.8x10^{-37}$	1021.47	*X. doucetiae*
PRT-1012	[+99.06]LYYYYW	$1.2x10^{-29}$	1012.47	*X. doucetiae*
PRT-1108	[+99.06]FYWYYW	$6.5x10^{-40}$	1108.48	30TX1
PRT-1092	[+99.06]FYWYFW	$3.5x10^{-39}$	1092.49	30TX1
PRT-1076	[+99.06]FYWFFW	$4.3x10^{-41}$	1076.49	30TX1
PRT-945	[+99.06]FWYYW	$1.2x10^{-23}$	945.42	*X. porinari*
PRT-929	[+99.06]FWYFW	$1.3x10^{-25}$	929.43	*X. porinari*
PRT-922	[+99.06]FYYYW	$2.5x10^{-27}$	922.43	*X. porinari*
PRT-911	[+99.06]LWYYW	$8.2x10^{-26}$	911.44	*X. porinari*

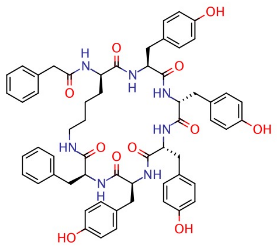
f



PRT-1012



PRT-1021

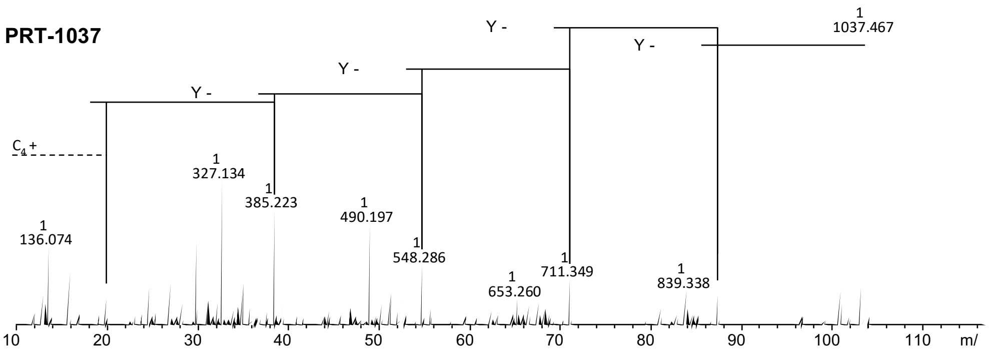


PRT-1046

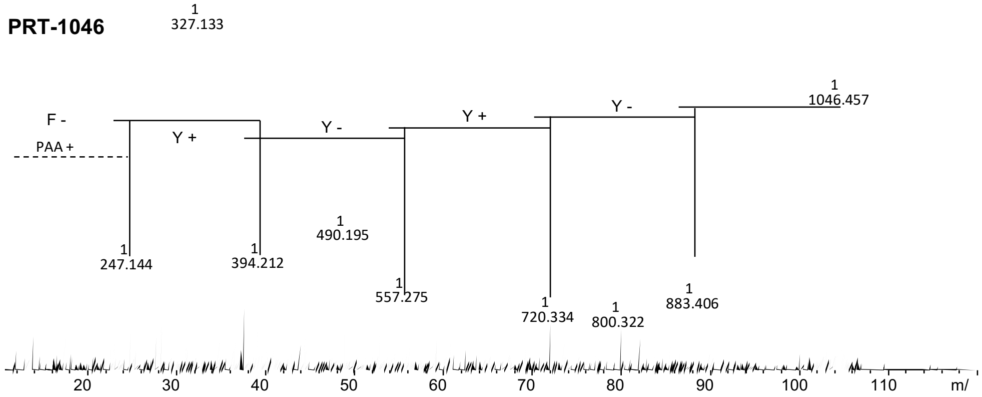
Figures 2.7-2.9 provide further information about MS/MS fragmentation pattern of PRT NRPs listed in Figure 2.6. Figure 2.6.f pictures the selected PRT derivatives produced by *X. doucetiae* including amino acid configuration as concluded from the presence of epimerization domains in the corresponding NRPSs and acyl residues as concluded from feeding experiments.

Figure 2.7. Fragmentation pattern (MS/MS of the molecular ions) of selected PRT derivatives from *X. doucetiae* observed by HPLC-MS analysis. For each spectrum, the corresponding NRP ID is listed in the top left corner.

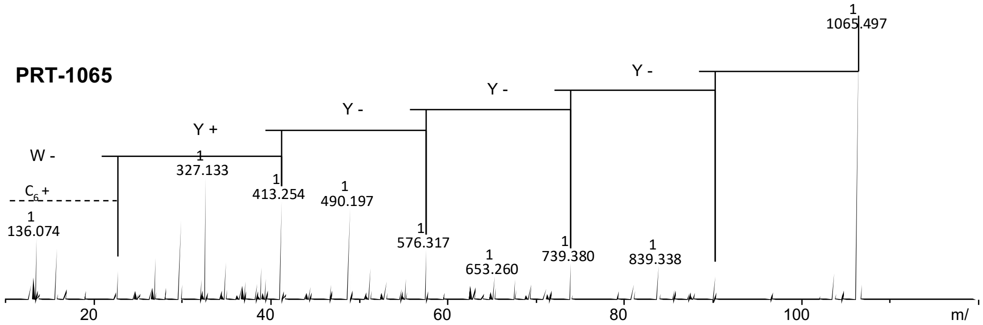
PRT-1037



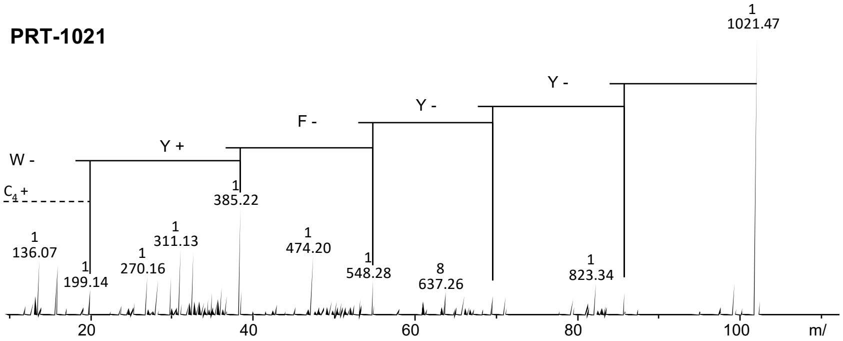
PRT-1046



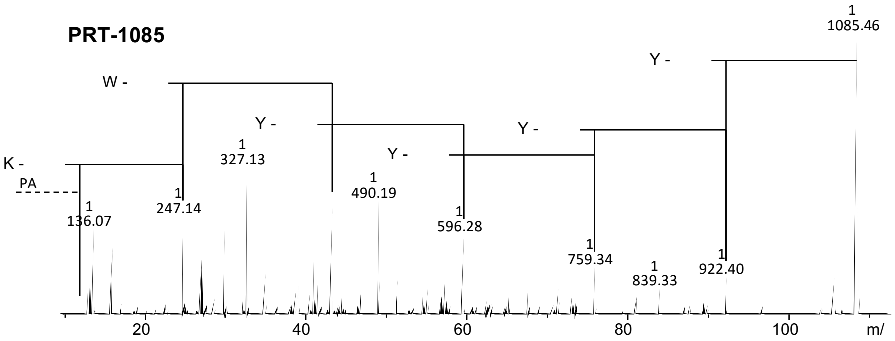
PRT-1065



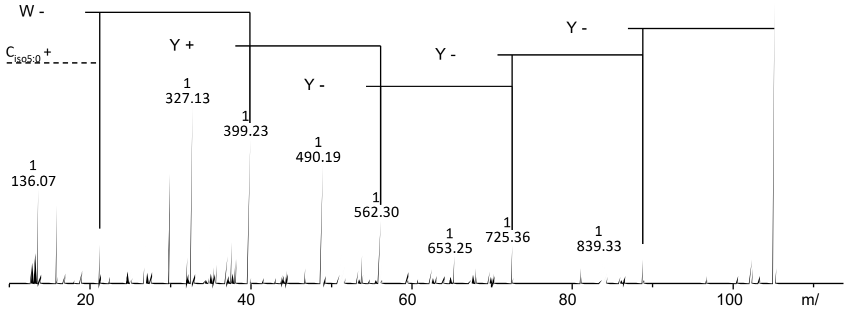
PRT-1021



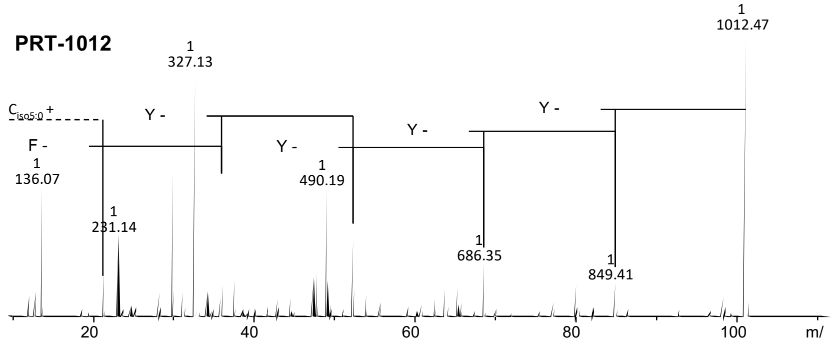
PRT-1085



PRT-1051



PRT-1012



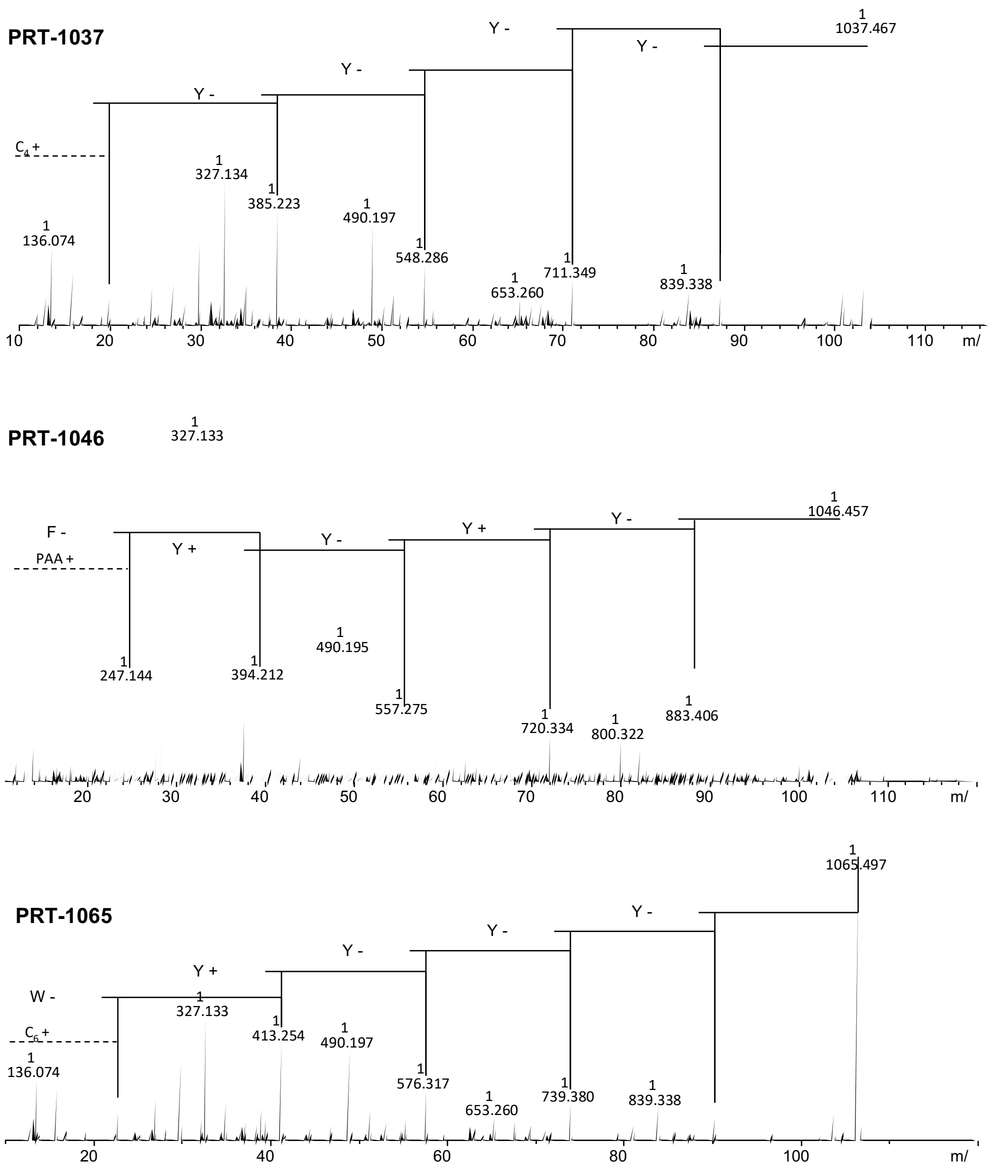


Figure 2.7. Fragmentation pattern (MS/MS of the molecular ions) of selected PRT derivatives from *X. doucetiae* observed by HPLC-MS analysis. For each spectrum, the corresponding NRP ID is listed in the top left corner, Continued.

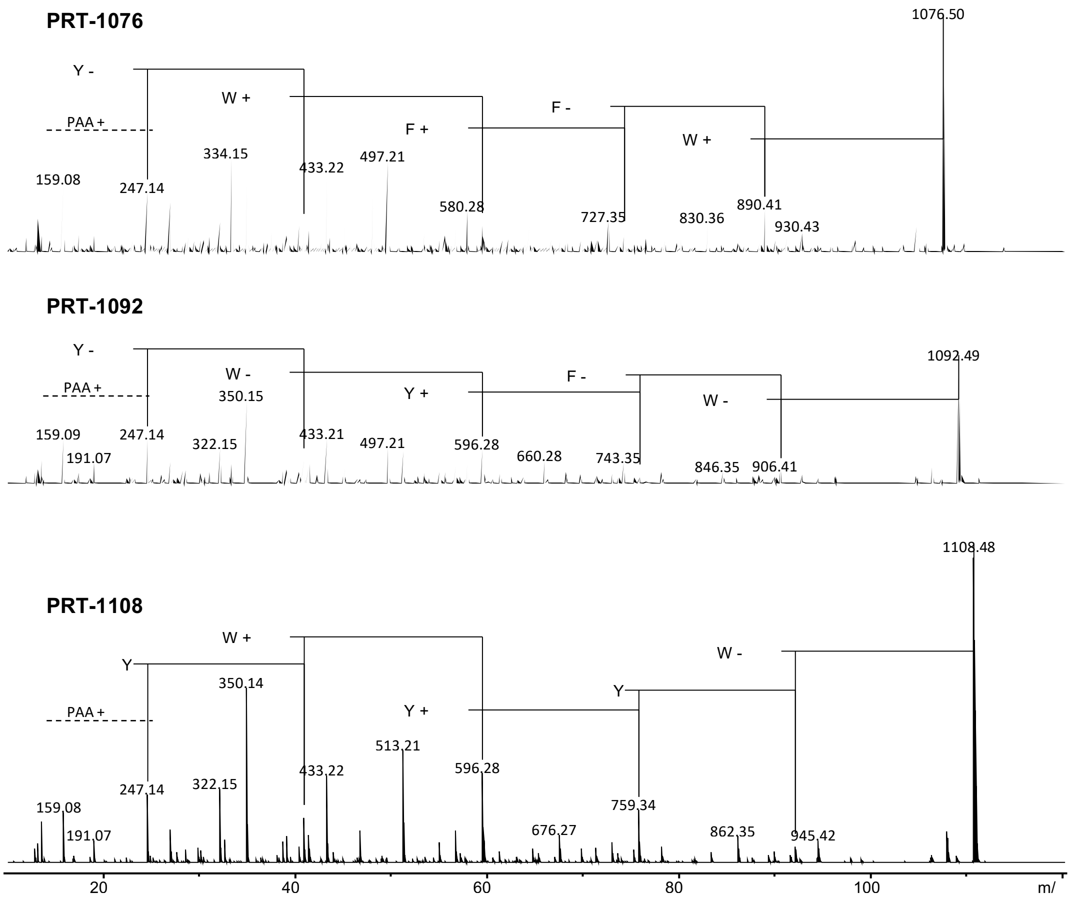


Figure 2.8. Fragmentation pattern (MS/MS of the molecular ions) of selected PRT derivatives from *Xenorhabdus* sp. 30TX1 observed by HPLC-MS analysis. For each spectrum, the corresponding NRP ID is listed in the top left corner.

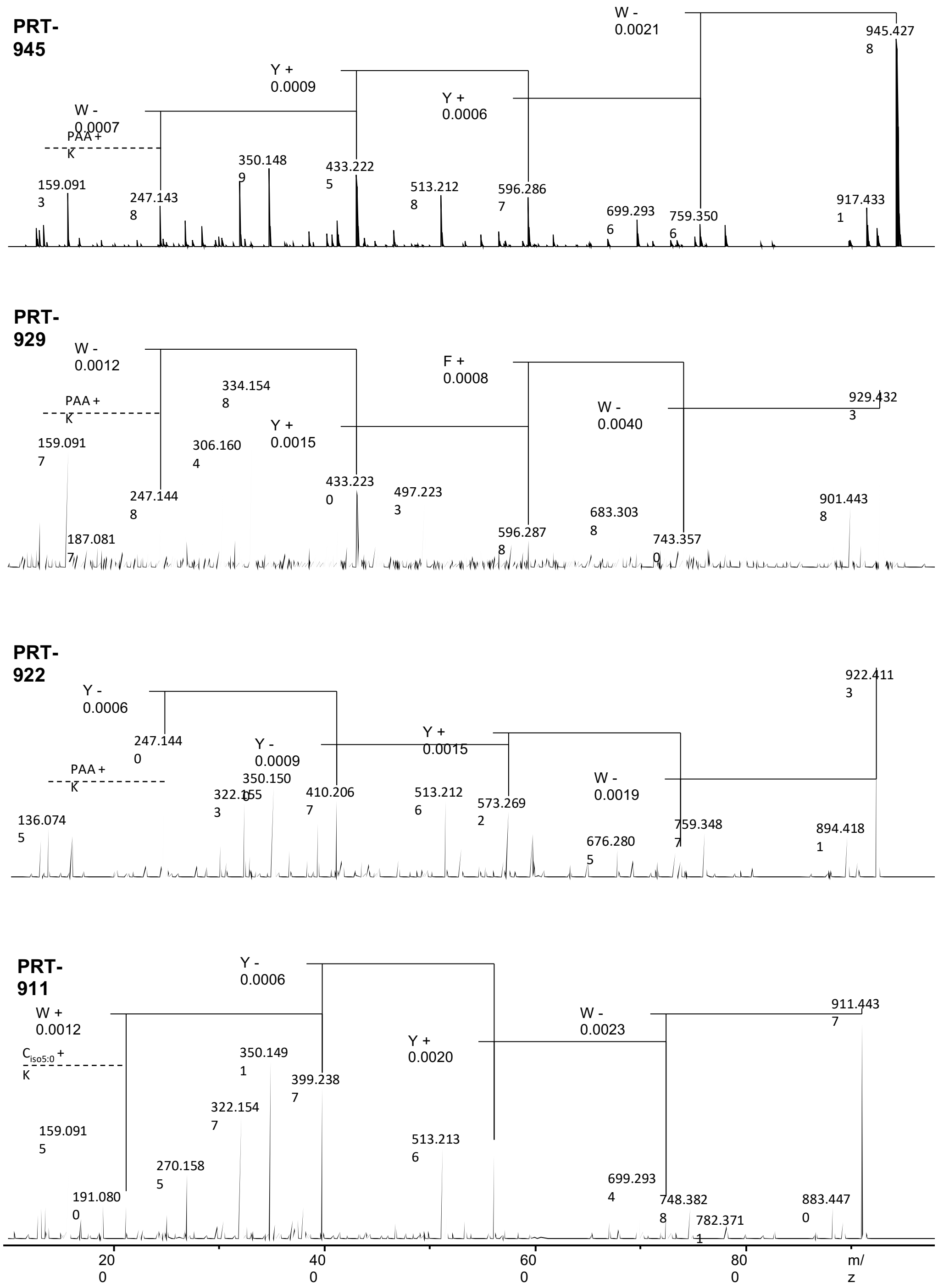
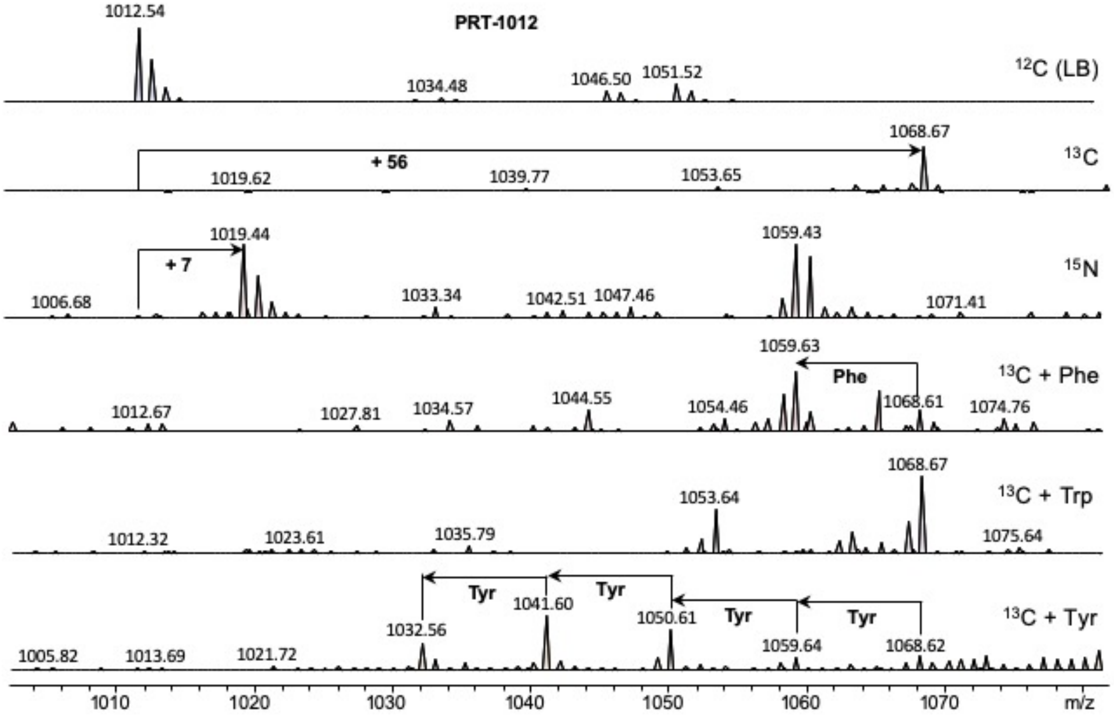
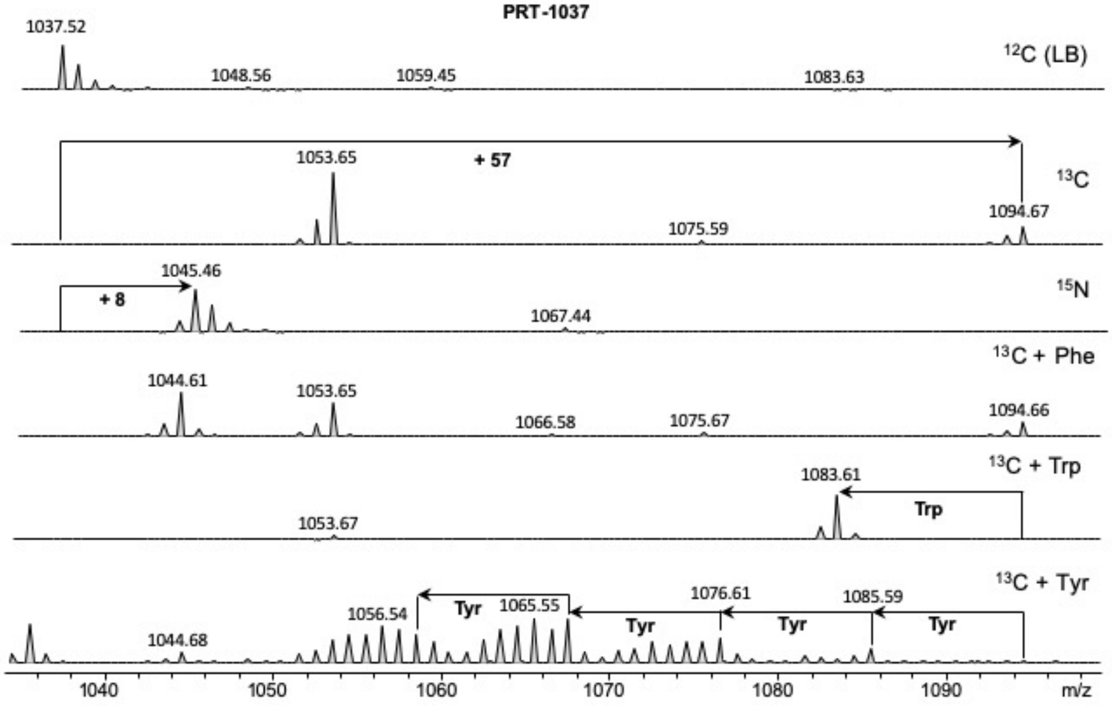


Figure 2.9. Fragmentation pattern (MS/MS of the molecular ions) of selected PRT derivatives from *X. poinarii* observed by HPLC-MS analysis. For each spectrum, the corresponding NRP ID is listed in the top left corner.

Figure 2.10. MS analysis of selected PRT derivatives after cultivation in ^{12}C (LB), ^{13}C- and ^{15}N- medium. Analysis of the incorporation of non-labelled Phe, Trp, Tyr and Leu added to fully labeled ^{13}C medium. For each spectrum, the corresponding NRP ID is listed in listed on the top.


PRT-1012
1012.54
1034.48
1046.50
1051.52
^{12}C (LB)
1068.67
+ 56
1019.62
1039.77
1053.65
^{13}C
1019.44
+ 7
1006.68
1033.34
1042.51
1047.46
1059.43
1071.41
^{15}N
1059.63
Phe
1012.67
1027.81
1034.57
1044.55
1054.46
1068.61
1074.76
^{13}C + Phe
1068.67
1053.64
1012.32
1023.61
1035.79
1075.64
^{13}C + Trp
Tyr
Tyr
Tyr
Tyr
1041.60
1050.61
1032.56
1059.64
1068.62
1005.82
1013.69
1021.72
^{13}C + Tyr
1010
1020
1030
1040
1050
1060
1070
m/z


PRT-1037
1037.52
1048.56
1059.45
1083.63
^{12}C (LB)
+ 57
1053.65
1075.59
1094.67
^{13}C
1045.46
+ 8
1067.44
^{15}N
^{13}C + Phe
1044.61
1053.65
1066.58
1075.67
1094.66
^{13}C + Trp
1083.61
Trp
1053.67
^{13}C + Tyr
1076.61
1085.59
1065.55
1056.54
Tyr
Tyr
Tyr
Tyr
1044.68
1040
1050
1060
1070
1080
1090
m/z

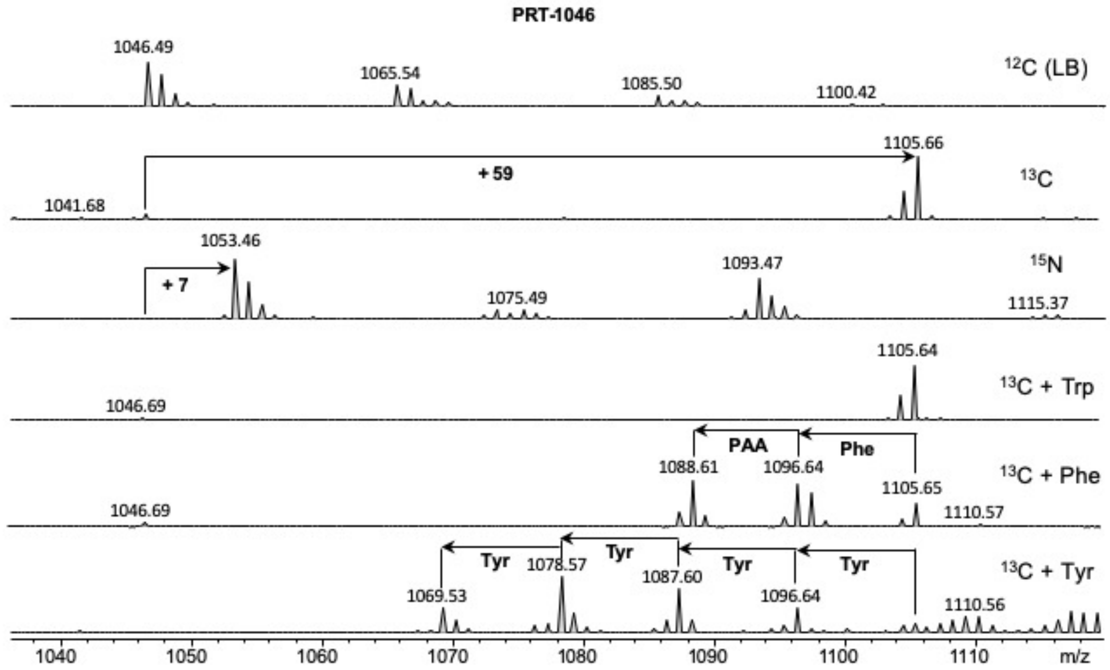


Figure 2.10. MS analysis of selected PRT derivatives after cultivation in ^{12}C (LB), ^{13}C- and ^{15}N- medium. Analysis of the incorporation of non-labelled Phe, Trp, Tyr and Leu added to fully labeled ^{13}C medium. For each spectrum, the corresponding NRP ID is listed in listed on the top, Continued.

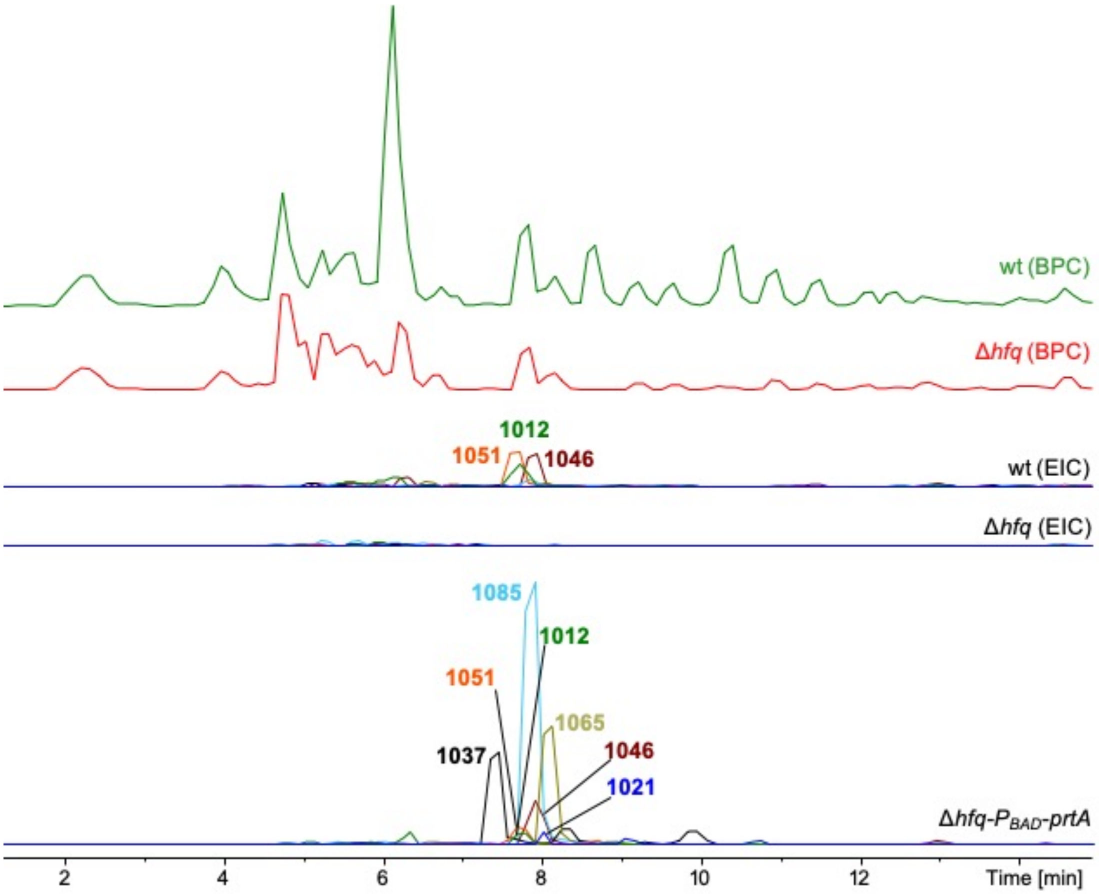


Figure 2.11. (top) Base peak chromatogram (BPC) of *X. doucetiae* wt (green) and *X. doucetiae*-Δ*hfq* (red) crude extracts. **(bottom)** Extracted ion chromatograms (EIC) of PRT derivatives from the extract of induced *X. doucetiae*-Δ*hfq*-PBAD-*prtA*.

Additional derivatives were found in large scale cultivation of wildtype and □*hfq* mutants of *X. doucetiae* (Figure 2.11). Appendix 1 describes this additional analysis. No BGC was found in *Xenorhabdus* sp. 30TX1 due to highly fragmented assembly. We further conducted nuclear magnetic resonance (NMR) spectroscopy on one of the major derivatives (Appendix 1 describe this experiment and Figures 2.12-2.18, and Table 2.4 present the results). Our NMR results confirmed the MS results, with the distinction that NMR revealed a short chain fatty acid like phenylacetic acid (PAA) as a starting unit (incorporated by the C-starter domain), followed by a Lys that is cyclized to the terminal thioester by the C-terminal TE domain. NRPminer predicted

Phe instead of the correct amino acid *Lys*, since NRPSpredictor2[13] made an error in identifying the amino acid for the corresponding A-domain (see Figure 2.6.a for the list of predicted amino acids). It has been shown that NRPSpredictor2[13] often fails to predict *Lys* residues, due to lack of training data for this amino acid[13]. Furthermore, as with any other MS-based method, NRPminer was not able to distinguish between residues with the same molar mass in the structure of final NRP, such as the pair *Ala* and b-*Ala*. All other NRPminer predictions of individual amino acids were consistent with NMR.

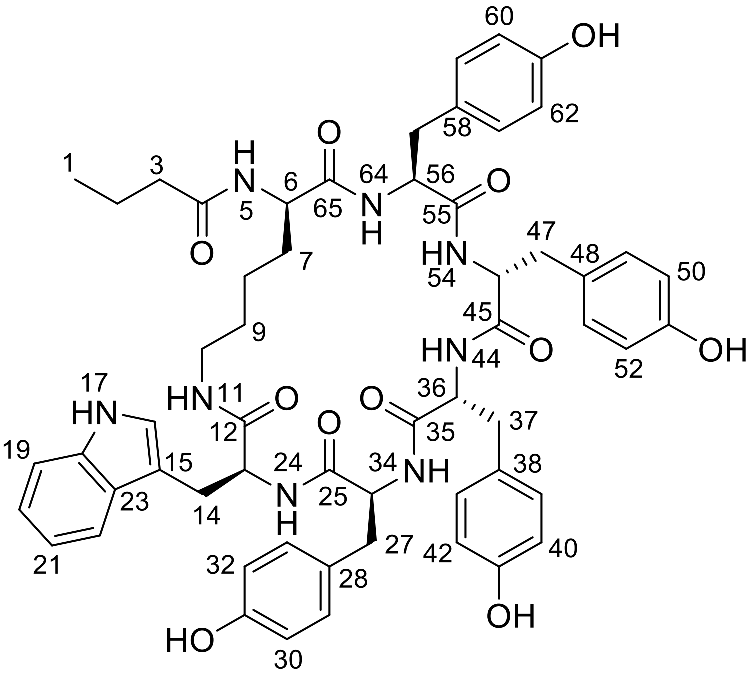


Figure 2.12. Numbering of protegomycin PRT-1037 (NMR data are provided in Table 2.4).

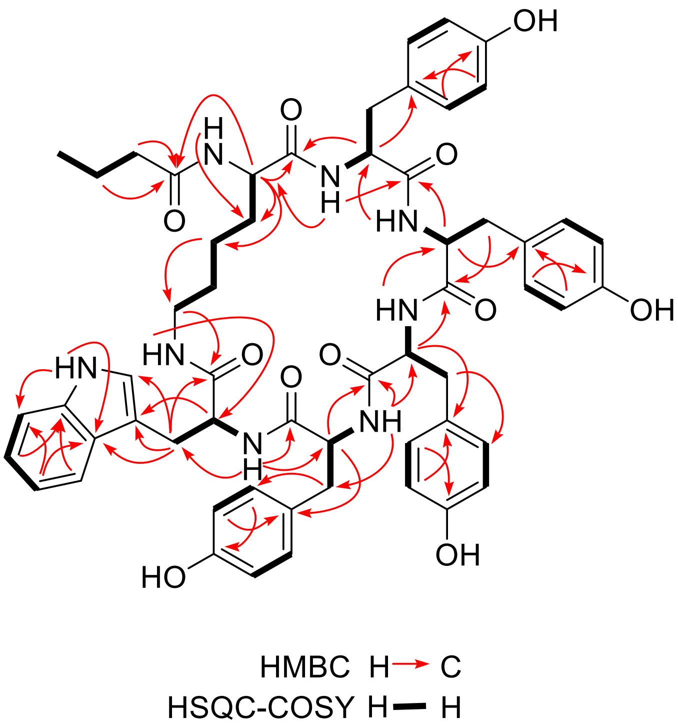


Figure 2.13. Key HMBC and HSQC-COSY correlations PRT-1037.

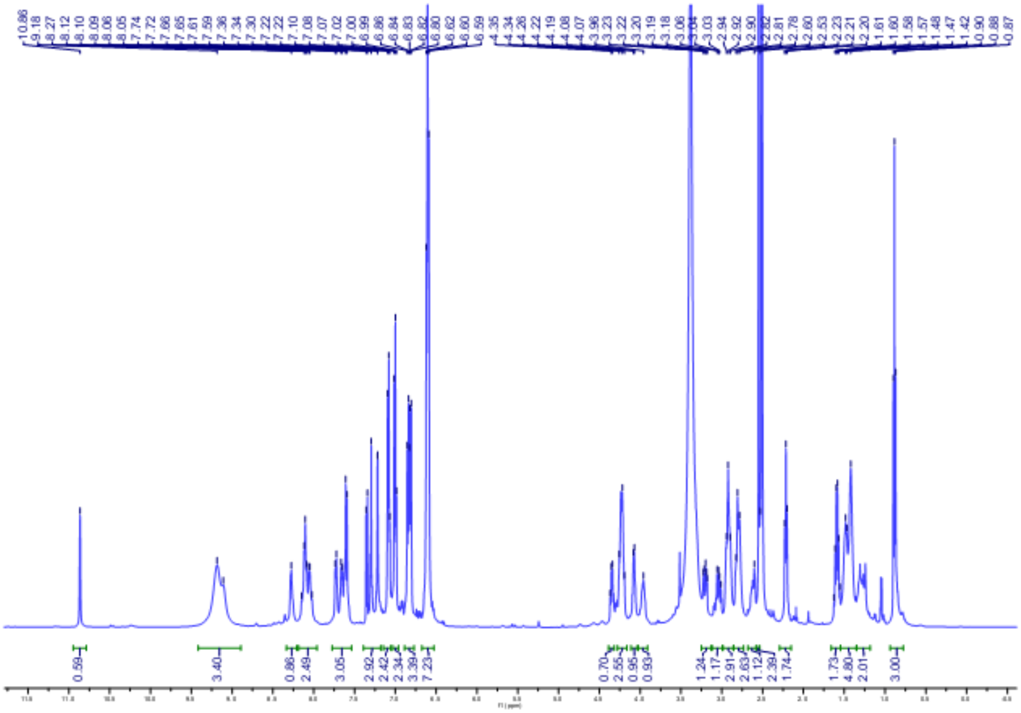


Figure 2.14. ^{1}H NMR spectrum of compound PRT-1037.

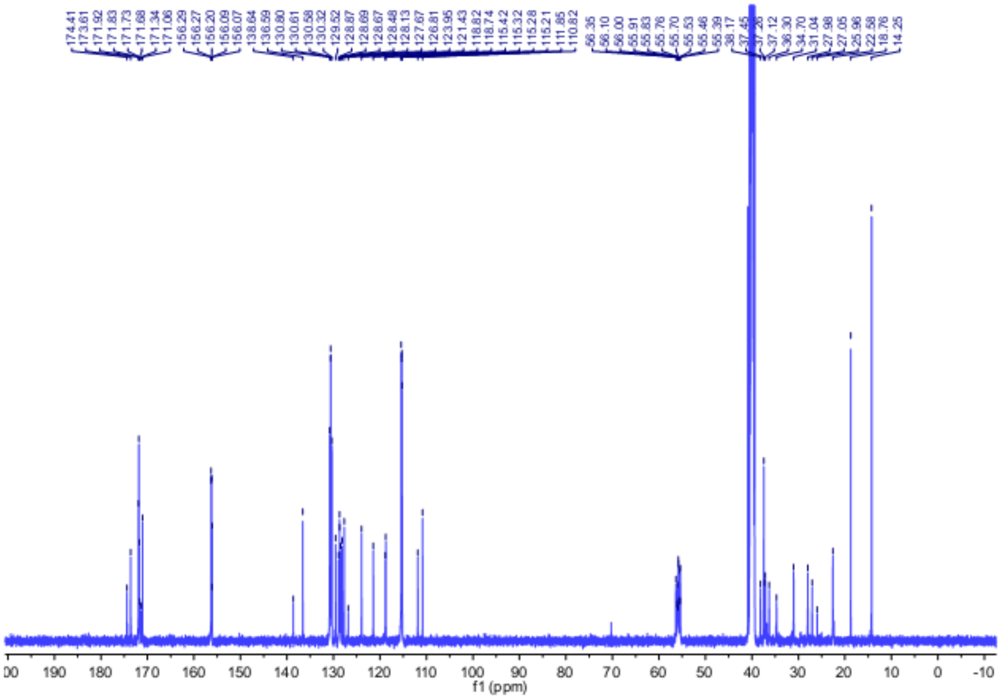


Figure 2.15. ^{13}**C NMR spectrum of compound PRT-1037.**

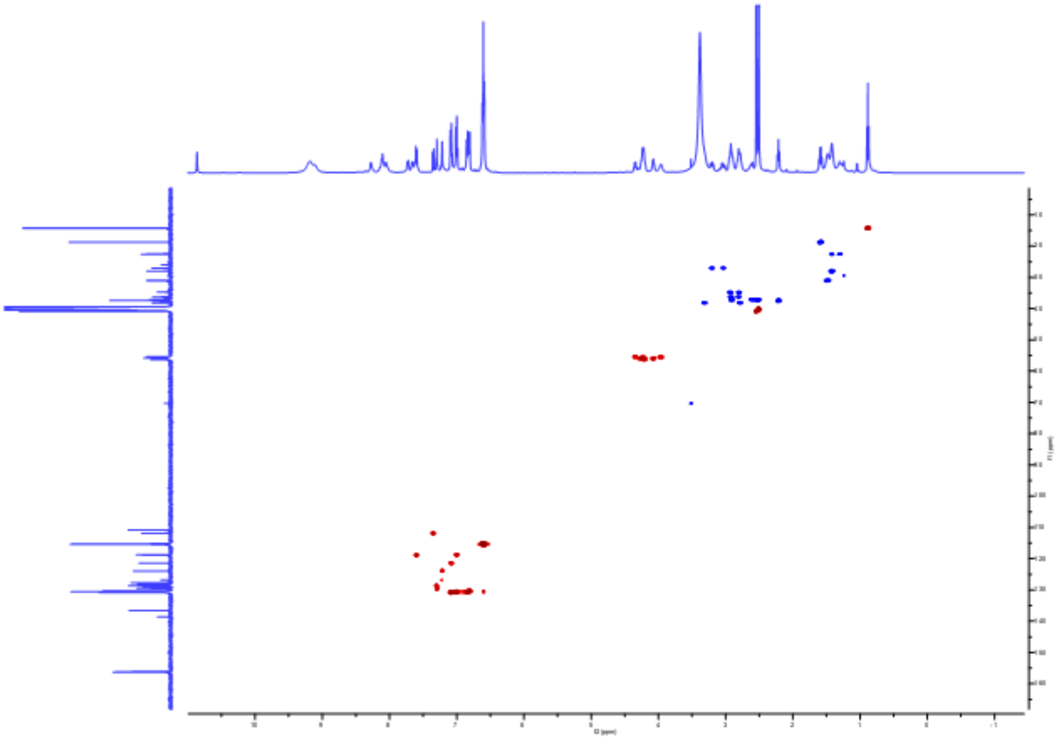


Figure 2.16. **HSQC spectrum of compound PRT-1037.**

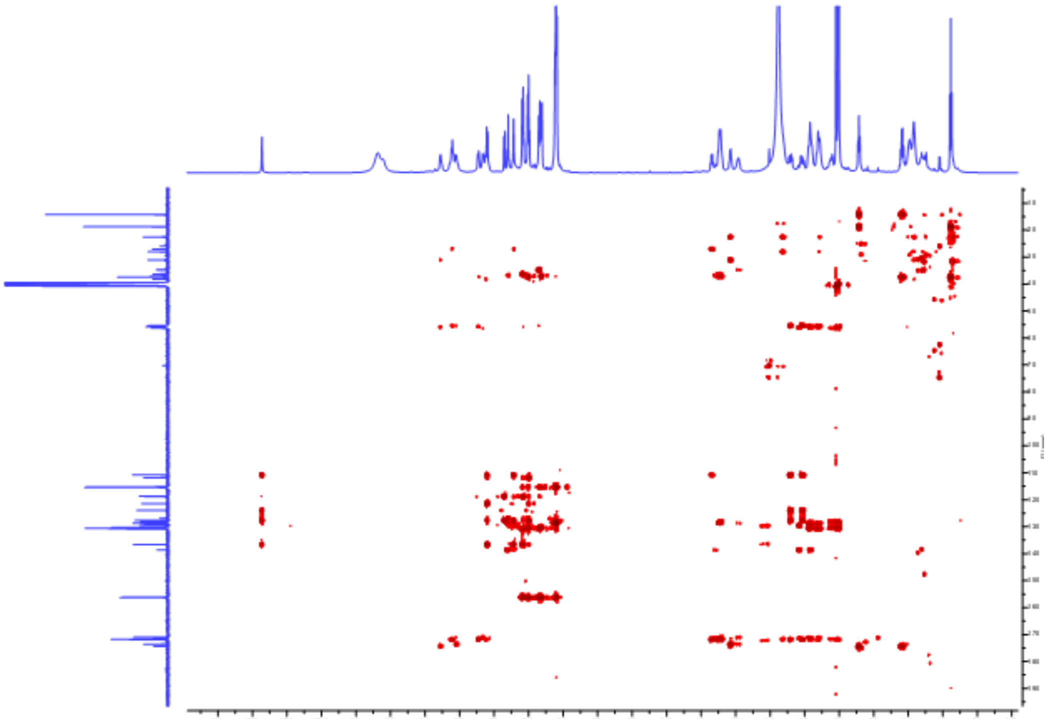


Figure 2.17. HMBC spectrum of compound PRT-1037.

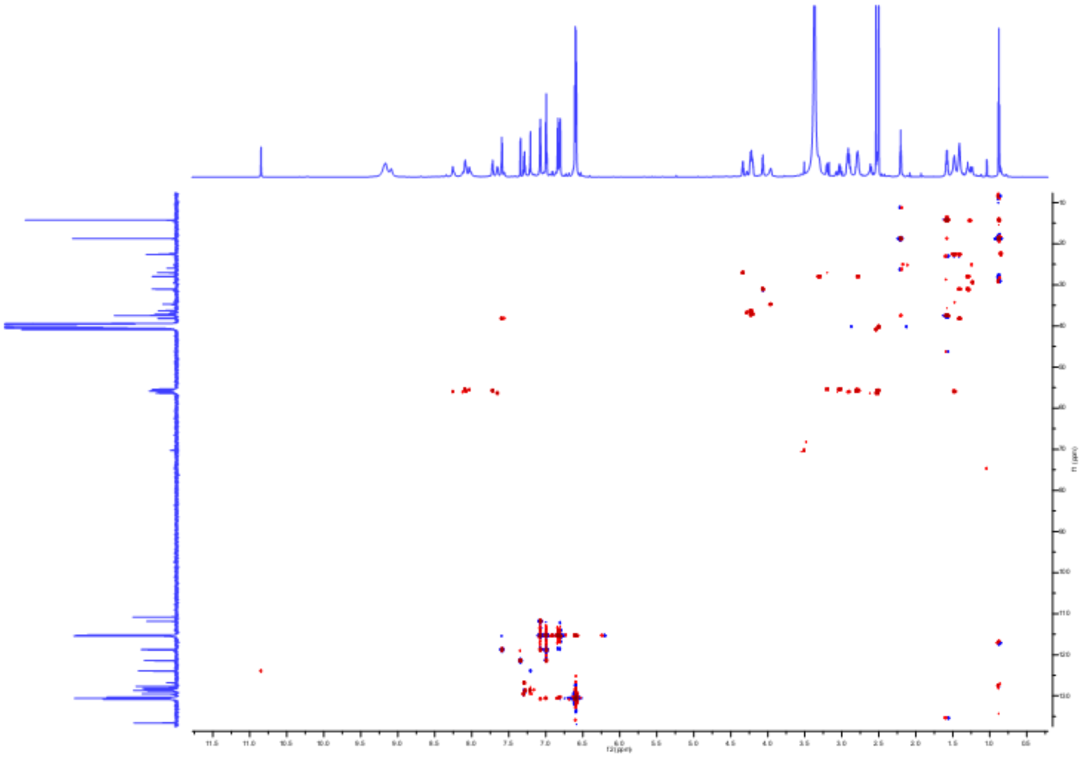


Figure 2.18. HSQC-COSY spectrum of compound PRT-1037.

Table 2.4. ^{1}H (500 MHz) and ^{13}C (125 MHz) NMR spectroscopic data for PRT-1037 in DMSO-d_6 (δ in ppm and J in Hz).

no.	PRT-1037	
	δ_C, type	δ_H (mult., J)
1	14.25	0.88 (d, 7.4)
2	18.76	1.59 (dq, 14.6, 7.2)
3	37.45	2.21 (t, 7.2)
4	174.41	
5		8.27 (br s)
6	56.00	4.08 (m)
7	31.04	1.48 (m)
8	22.58	1.42 (br s)
9	27.98	1.42 (overlap)
10	38.17	3.31 (m)
		2.78 (m)
11		7.62 (br s)
12	171.73	
13	55.39	4.35 (m)
14	27.05	3.20 (m)
		3.03 (dd, 14.7, 9.0)
15	110.82	
16	123.95	7.22 (d, 1.9)
17		10.86 (br s)
18	136.59	
19	111.85	7.59 (d, 8.0)
20	118.82	7.00 (t, 7.2)
21	121.43	7.08 (t, 7.2)
22	118.74	7.35 (d, 8.0)
23	127.67	
24		8.11 (br s)
25	171.83	
26	55.83	4.22 (m)
27	37.12	2.60 (m)
		2.53 (overlap)
28	128.14	
29	130.32	6.81 (d, 8.5)
30	115.28	6.60 (d, 8.5)
31	156.27	
32	115.28	6.60 (d, 8.5)
33	130.32	6.81 (d, 8.5)
34		8.11 (d, 6.6)
35	171.68	
36	56.35	4.22 (overlap)
37	37.26	2.92 (overlap)
		2.80 (overlap)
38	128.69	
39	130.58	6.81 (overlap)
40	115.42	6.61 (overlap)
41	156.07	
42	115.42	6.61 (overlap)
43	130.58	6.81 (overlap)
44		8.11 (overlap)
45	171.92	
46	55.76	4.22 (overlap)
47	36.3	2.92 (overlap)
		2.80 (overlap)
48	128.67	
49	130.80	6.81 (overlap)
50	115.21	6.61 (overlap)
51	156.20	
52	115.21	6.61 (overlap)
53	130.80	6.81 (overlap)
54		8.05 (d, 6.6)
55	171.06	
56	55.53	3.96 (m)
57	34.70	2.92 (overlap)
		2.80 (overlap)
58	128.48	
59	130.61	6.81 (overlap)
60	115.32	6.61 (overlap)
61	156.29	
62	115.32	6.61 (overlap)
63	130.61	6.81 (overlap)
64		8.05 (overlap)
65	173.61	

Besides PAA, other starter acyl units are isovaleric acid (in PRT-1012; NRPminer prediction 99.06+*Leu*; see Figure 2.6.f) and butyric acid (in PRT-1037; see Figure 2.6.e). Figure 2.10 describes labelling data and mass spectra for the identified protegomycins in *X. doucetiae*. The isolated derivatives PRT-1037 and PRT-1021 were tested against various protozoa and showed a weak activity against *Trypanosoma brucei rhodesiense* (IC_{50} [mg/L] 79 and 53) and *Plasmodium falciparum* (IC_{50} [mg/L] >50 and 33) with no toxicity against L6 rat myoblast cells (IC_{50} [mg/L] both >100). Figures 2.19 and 2.20 present further information about protegeomycin BGC and NRPs.

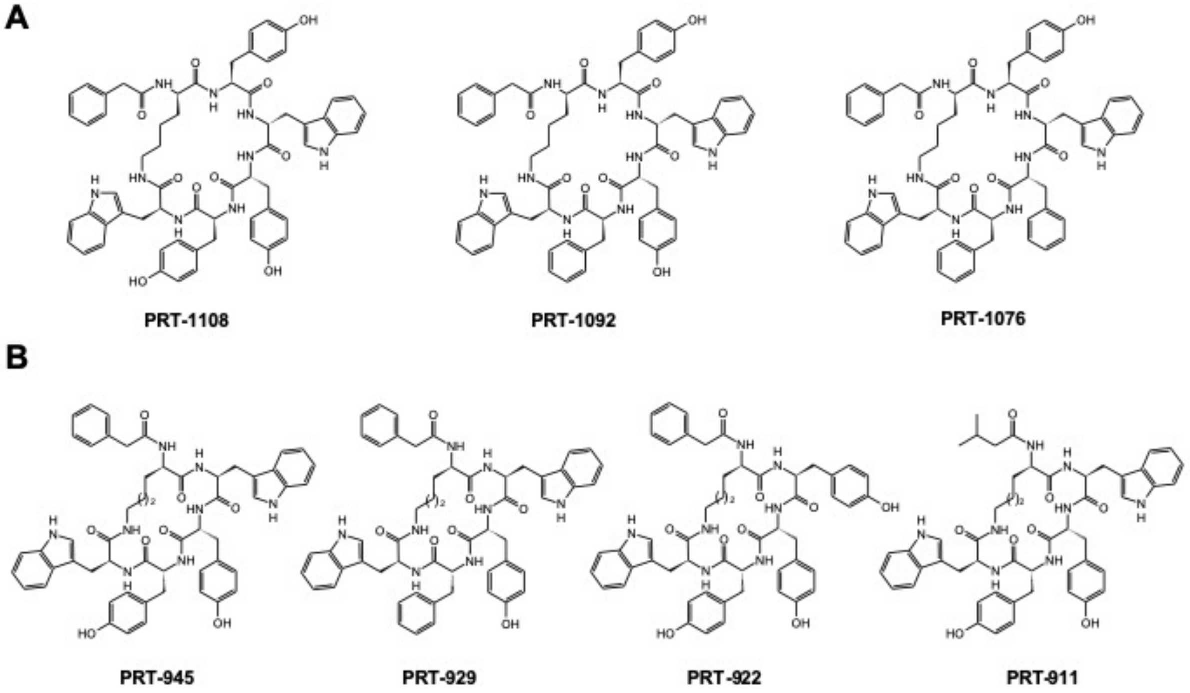


Figure 2.19. Predicted structures of PRT derivatives produced by *Xenorhabdus* sp. 30TX1 and *X. proinarii*. (A) Predicted structures for PRT derivatives produced by *Xenorhabdus* sp. 30TX1 including amino acid configuration as found in *X. doucetiae*. (**B**) Predicted structures for PRT derivatives produced by *X. poinarii* including amino acid configuration as concluded from the presence of epimerization domains in the corresponding NRPS PrtAB.

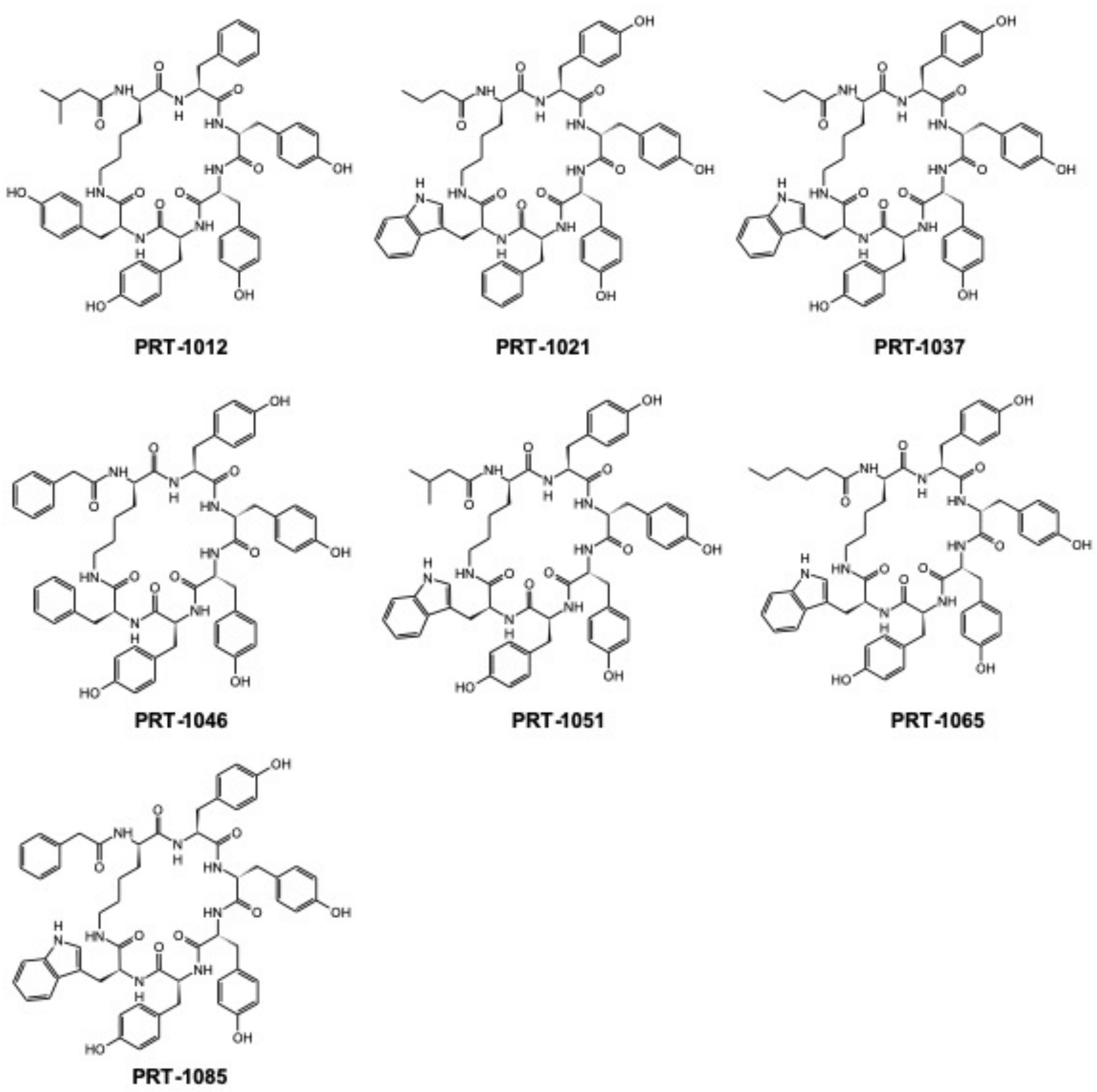
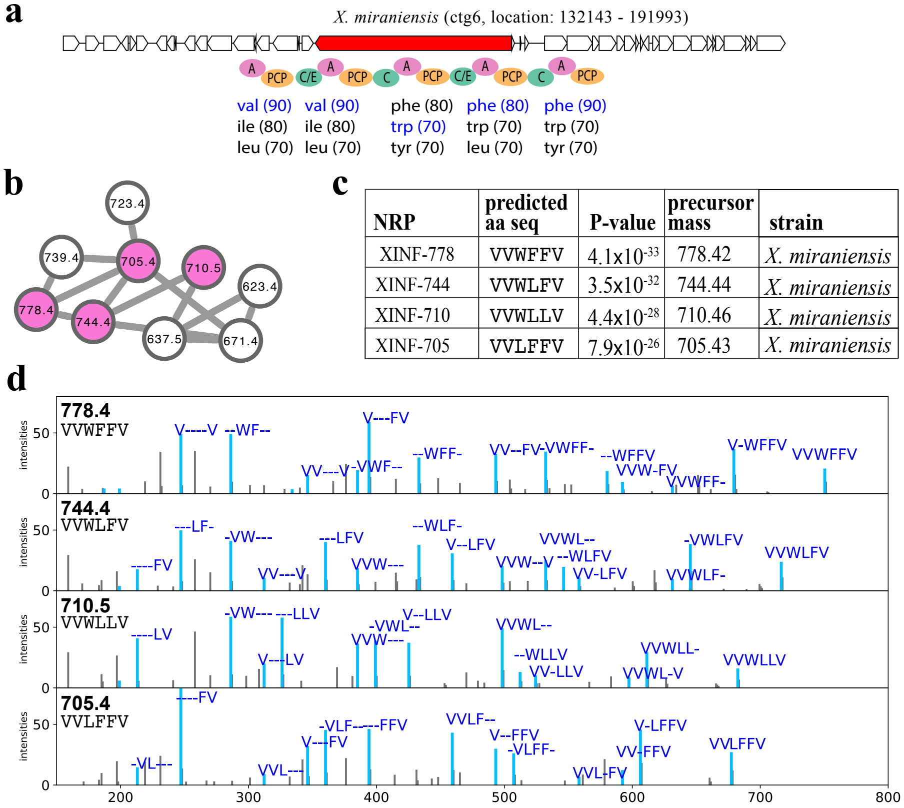


Figure 2.20. Structures for PRT derivatives produced by *X. doucetiae*. The structure are shown including amino acid configuration as concludes from the presence of epimerization domains in the corresponding NRPSs PrtAB.

Novel xenoinformycin (XINF) NRP family in the XPF dataset. NRPminer matched four spectra representing four cyclic NRPs from *X. miraniensis* spectral dataset to a novel BGC (figure 2.21). NRPminer reported a modification with total mass of 99.068 for all the four identified NRPs, which matches the valine mass. We hypothesize that one of the valine-specific adenylation domains is responsible for the activation of two consecutive valine units, suggesting an iterative use of the Val-incorporating module (similar to stuttering observed in polyketide synthases[43,67]) but this is yet to be experimentally verified. Interestingly, the predicted xenoinformycin producing NRPS XinfS is highly similar to the widespread NRPS GxpS found in *Xenorhabdus* and *Photorhabdus*, responsible for the GameXPeptide production[30,68]. While both XinfS and GxpS have five modules, XinfS has a C-domain instead of the usual C/E-domain in the last module, suggesting a different configuration of the amino acid *Phe* or *Leu* (corresponding to the second last A-domain on their NRPSs), respectively.

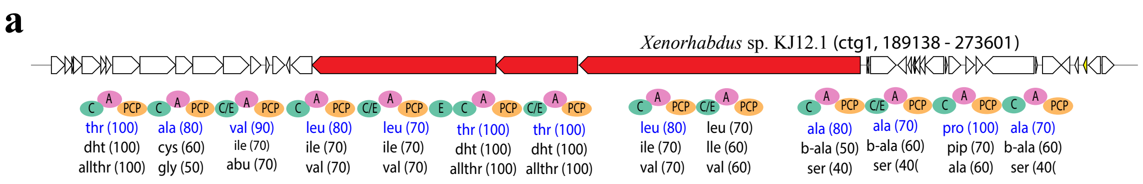

a
X. miraniensis (ctg6, location: 132143 - 191993)
A
PCP
C/E
A
PCP
C
A
PCP
C/E
A
PCP
C
A
PCP
val (90) ile (80) leu (70)
val (90) ile (80) leu (70)
phe (80) trp (70) tyr (70)
phe (80) trp (70) leu (70)
phe (90) trp (70) tyr (70)
b
723.4
739.4
705.4
710.5
623.4
778.4
744.4
637.5
671.4
c
d
778.4
VVWFFV
744.4
VVWLFV
710.5
VVWLLV
705.4
VVLFFV
intensities
50
0
200
300
400
500
600
700
800

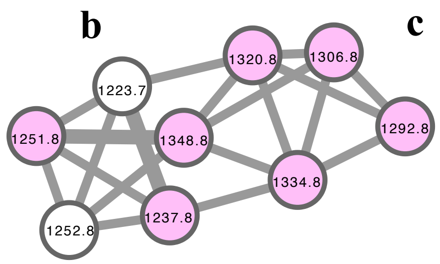
NRP	predicted aa seq	P-value	precursor mass	strain
XINF-778	VVWFFV	$4.1x10^{-33}$	778.42	*X. miraniensis*
XINF-744	VVWLFV	$3.5x10^{-32}$	744.44	*X. miraniensis*
XINF-710	VVWLLV	$4.4x10^{-28}$	710.46	*X. miraniensis*
XINF-705	VVLFFV	$7.9x10^{-26}$	705.43	*X. miraniensis*

Figure 2.21. Novel xenoinformycin NRP family. (***a***) The BGC generating the NRP in *X. miraniensis* along with NRPS genes (shown in red) and the A-, C-, PCP-, and C/E-domains appearing on the corresponding NRPS. The rest of the genes in the corresponding contigs are shown in white. Three highest-scoring amino acids for each A-domain in this BGC (according to NRPSpredictor2[13] predictions) are shown below the corresponding A-domains. Amino acids appearing in the NRP VVWFF identified by NRPminer (with the lowest p-value) are shown in blue. (***b***) Spectral network formed by the spectra that originate from NRPs in the xenoinformycin family. A node is colored if the corresponding spectrum forms a statistically significant PSM (with p-value threshold 10^{-15}) and not colored otherwise. (***c***) Sequences of the identified NRPs in the xenoinformycin family (with the lowest p-value among all spectra originating from the same NRP). XINF represents xenoinformycin. (***d***) For each identified NRP, an annotated spectrum forming a PSM with the lowest p-value is shown.

Novel xenoamicin-like (XAM) NRP family in the XPF dataset. NRPminer discovered a novel NRP family that includes eight distinct NRPs, along with their BGC (Figure 2.22). While the matched BGC for this family is evolutionary related to the xenoamicin BGC[69] and both BGCs include 13 A-domains, 7 out of 13 amino acids in XAM differ from the corresponding amino acids in xenoamicin A (Figure 2.23). We named this novel class of xenoamicins class III. Interestingly, the occurrence of XAM-1237 and XAM-1251 suggest a loss of Pro in their structure indicating another possibility of NRP diversification, namely module skipping as previously observed in other NRPSs[67,70,71].

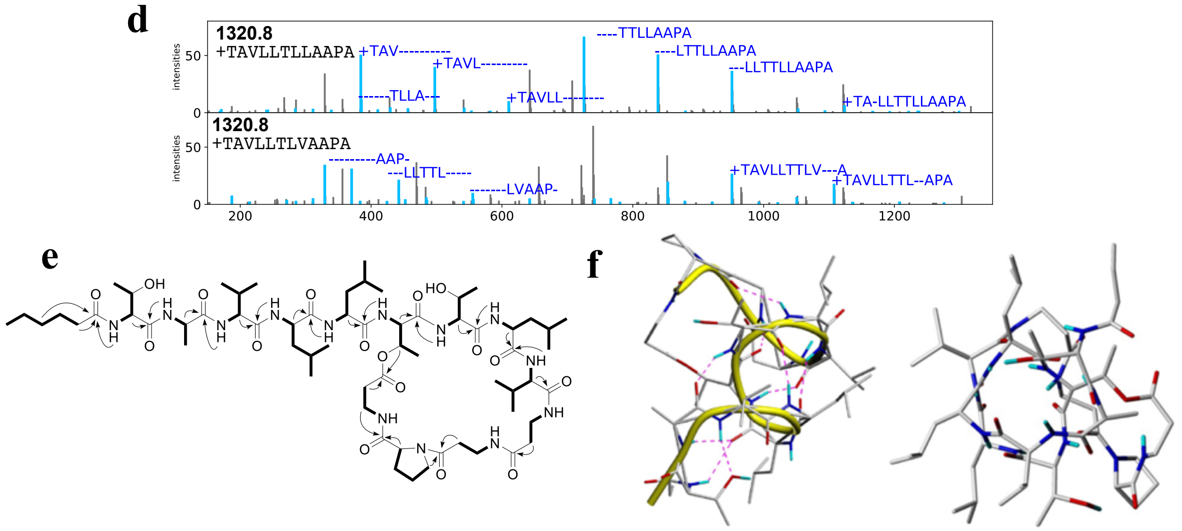
Figure 2.22. Novel xenoamicin-like NRP family. (a) The BGCs generating the NRP in *Xenorhabdus* sp. KJ12 along with NRPS genes (shown in red) and A-, C-, PCP-, and E-domains in these NRPSs. The rest of the genes in the corresponding contigs are shown in white. Three highest-scoring amino acids for each A-domain in these BGCs (according to NRPSpredictor2[13] predictions) are shown below the corresponding A-domains. Amino acids appearing in the NRP [+99.06]TAVLLTTLLAAPA identified by NRPminer (with the lowest p-value) are shown in blue. (b) Spectral network formed by the spectra that originate from NRPs in the XAM family. (c) Sequences of the identified NRPs in this family (with the lowest p-value among all spectra originating from the same NRP). (d) For each strain, an annotated spectrum representing the lowest p-value is shown. The spectra were annotated based on predicted NRPs [+99.06]TAVLLTTLLAAPA and [+99.06] TAVLLTTLVAAPA from top to bottom. The "+" sign represents the addition of [+99.06]. Figure 2.26 and 2.27 show the annotated spectra for the other NRPs shown in part (c). (e) NMR-based correlations of XAM-1320 (*m/z* 1320.8 $[M+H]^+$) produced by *Xenorhabdus* KJ12.1 (Table 2.5) HSQC-TOCSY (bold lines) and key ROESY correlations (arrows) are shown. (f) 3D structure of XAM-1320 derived from 121 ROE-derived distance constraints (Table 2.6), molecular dynamics and energy minimization. Peptide backbone is visualized with a yellow bar (left). Predicted hydrogen bonds stabilizing the β-helix are shown as dashed lines. View from above at the pore formed by XAM-1320. (right) NRPminer identified this NRP with p-value 8.4×10^{-50}.


a
Xenorhabdus sp. KJ12.1 (ctg1, 189138 - 273601)
C A PCP C A PCP C/E A PCP C A PCP C/E A PCP E C A PCP C/E A PCP
thr (100) dht (100) allthr (100)
ala (80) cys (60) gly (50)
val (90) ile (70) abu (70)
leu (80) ile (70) val (70)
leu (70) ile (70) val (70)
thr (100) dht (100) allthr (100)
thr (100) dht (100) allthr (100)
C A PCP C/E A PCP
leu (80) ile (70) val (70)
leu (70) lle (60) val (60)
C A PCP C/E A PCP C A PCP C A PCP
ala (80) b-ala (50) ser (40)
ala (70) b-ala (60) ser (40(
pro (100) pip (70) ala (60)
ala (70) b-ala (60) ser (40(



c

NRP	predicted *aa* seq	P-value	precursor mass	strains
XAM-1237	[+99.06]TAVLLTTLLAA-A	2.8x10^{-33}	1237.8	*KJ12.1, KK7.4*
XAM-1251	[+113.07]TAVLLTTLLAA-A	6.5x10^{-31}	1251.8	*KJ12.1*
XAM-1292	[+71.04]TGVLLTTLVAAPA	4.0x10^{-22}	1292.8	*KJ12.1, KK7.4, stockiae*
XAM-1306	[+99.06]TAVVLTTLVAAPA	9.0x10^{-22}	1306.8	*KJ12.1, KK7.4, stockiae*
XAM-1320	[+99.06]TAVLLTTLVAAPA	8.4x10^{-50}	1320.8	*KJ12.1, KK7.4, stockiae*
XAM-1334	[+99.06]TAVLLTTLLAAPA	2.8x10^{-45}	1334.8	*KJ12.1, KK7.4, stockiae*
XAM-1348	[+113.07]TAVLLTTLLAAPA	2.1x10^{-19}	1348.8	*KJ12.1, KK7.4, stockiae*



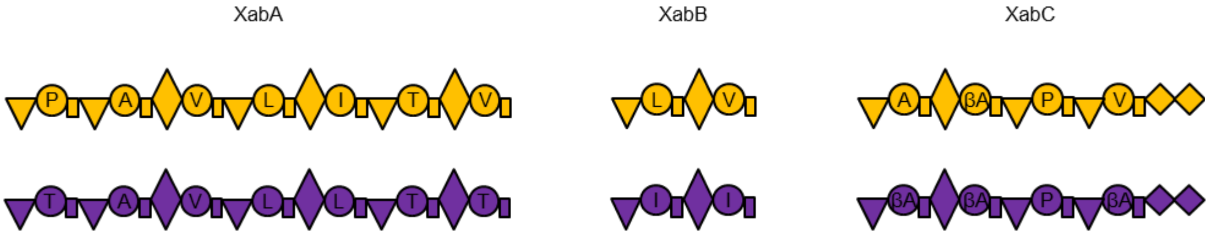


Figure 2.23. General NRPS structure of xenoamicin XabABC in *X. doucetiae* (yellow) and *Xenorhabdus* KJ12.1 (violet). Amino acid specificities are assigned for all A-domains. For domain assignment the following symbols are used: A (large circles), T (rectangle), C (triangle), C/E (diamond), TE-TE (two C-terminal small diamonds).

We confirmed the sum formula of XAM-1320 and XAM-1334 by feeding (Figures 2.24 and 2.25) and MS-MS experiments (figure 2.26 and Appendix 2) and were also able to isolate the major derivative XAM-1320 from *Xenorhabdus* KJ12.1 and to elucidate its structure by NMR including its 3D solution structure (Table 2.5 and Table 2.6) that confirms its helical structure from the alternating D/L configurations (confirmed by the advanced Marfey's analysis; Figure 2.27 and Appendix 2) throughout the peptide chain from the presence of C/E domains, except for the C-terminal part shown in Figure 2.22. XAM-1320 was also tested against protozoa and showed a good activity against *T. brucei rhodesiense* (IC_{50} [mg/L] 3.9) but much lower activity against *Trypanosoma cruci*, *Plasmodium falciparum* and rat L6 cells (IC_{50} [mg/L] 25.5, 56.2 and 46.0, respectively). Figure 2.26 provides information about the isolation and structure elucidation of XAM-1320, XAM-1278, XAM-1292, and XAM-1348 that differed in the starter acyl unit and the following amino acid (Ala or Gly) as pictured in Figure 2.28.

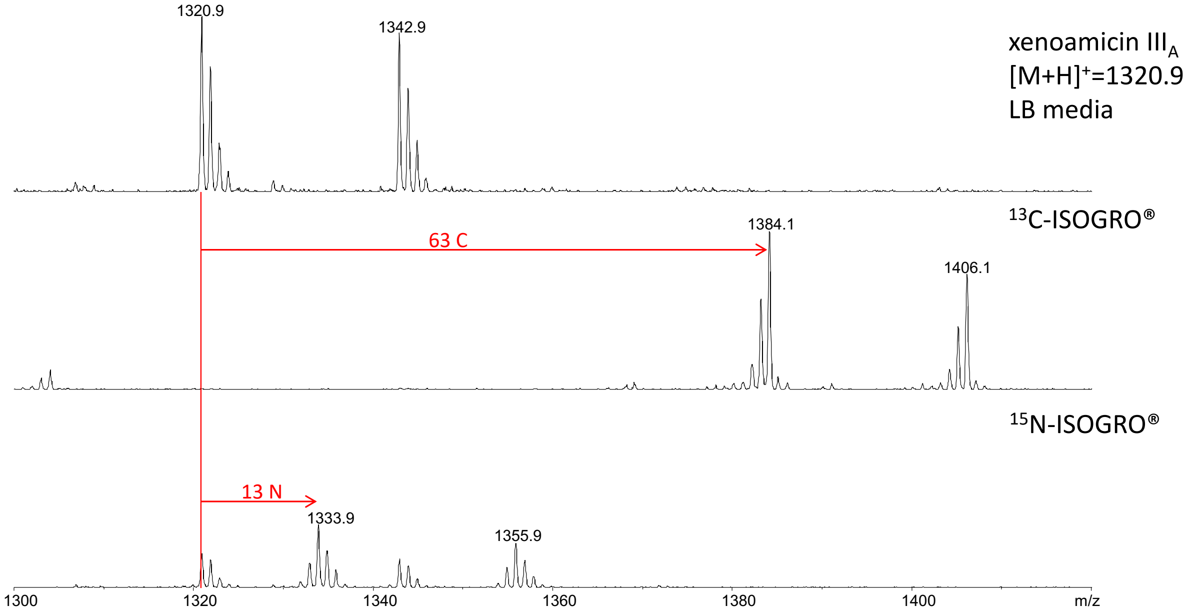


Figure 2.24. Determination of the number of carbon and nitrogen atoms in XAM-1320 by cultivation of *Xenorhabdus* KJ12.1 in LB medium, ^{13}C labelled or ^{15}N labelled ISOGRO® medium and the following mass shift detected by mass spectrometry.

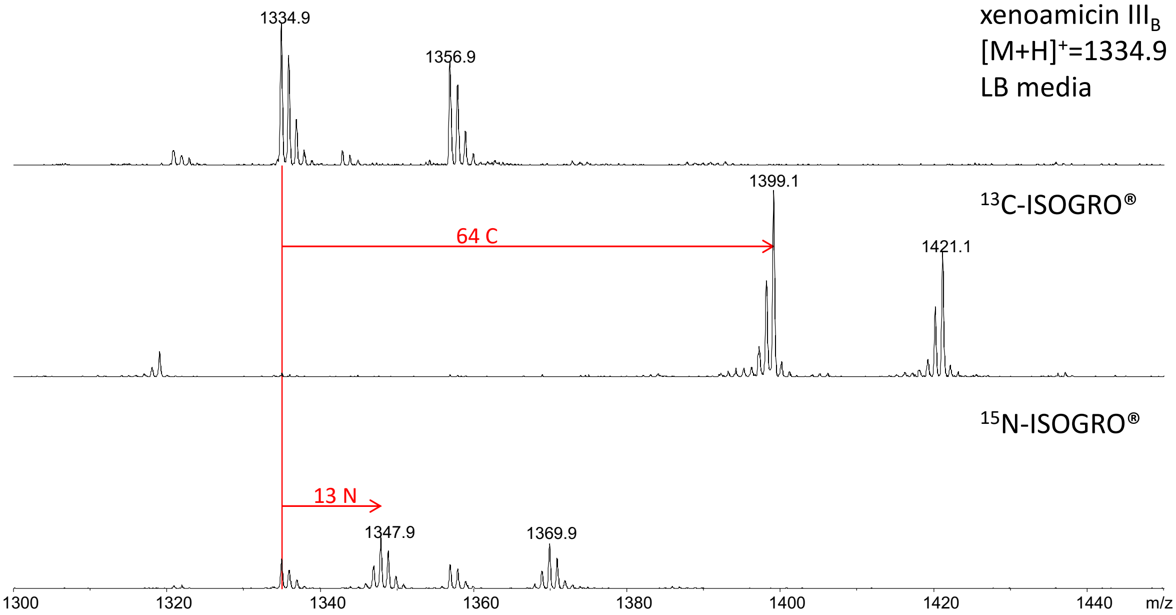


Figure 2.25. Determination of the number of carbon and nitrogen atoms in XAM-1334 by cultivation of *Xenorhabdus* KJ12.1 in LB medium, ^{13}C labelled or ^{15}N labelled ISOGRO® medium and the following mass shift detected by mass spectrometry.

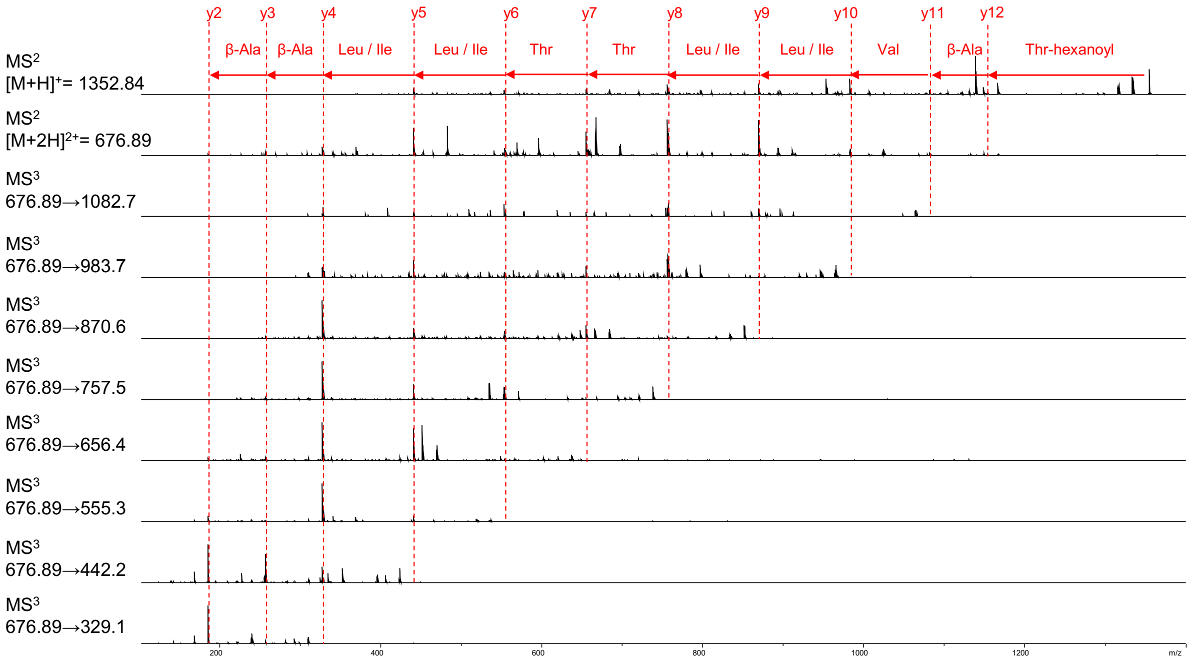


Figure 2.26. MS^2 and MS^3 spectra of linearized XAM-1334. The complete serial of y-ions could be assigned in MS^3 spectra from the double charged xenoamicin ion (m/z = 676.9 $[M+2H]^{2+}$).

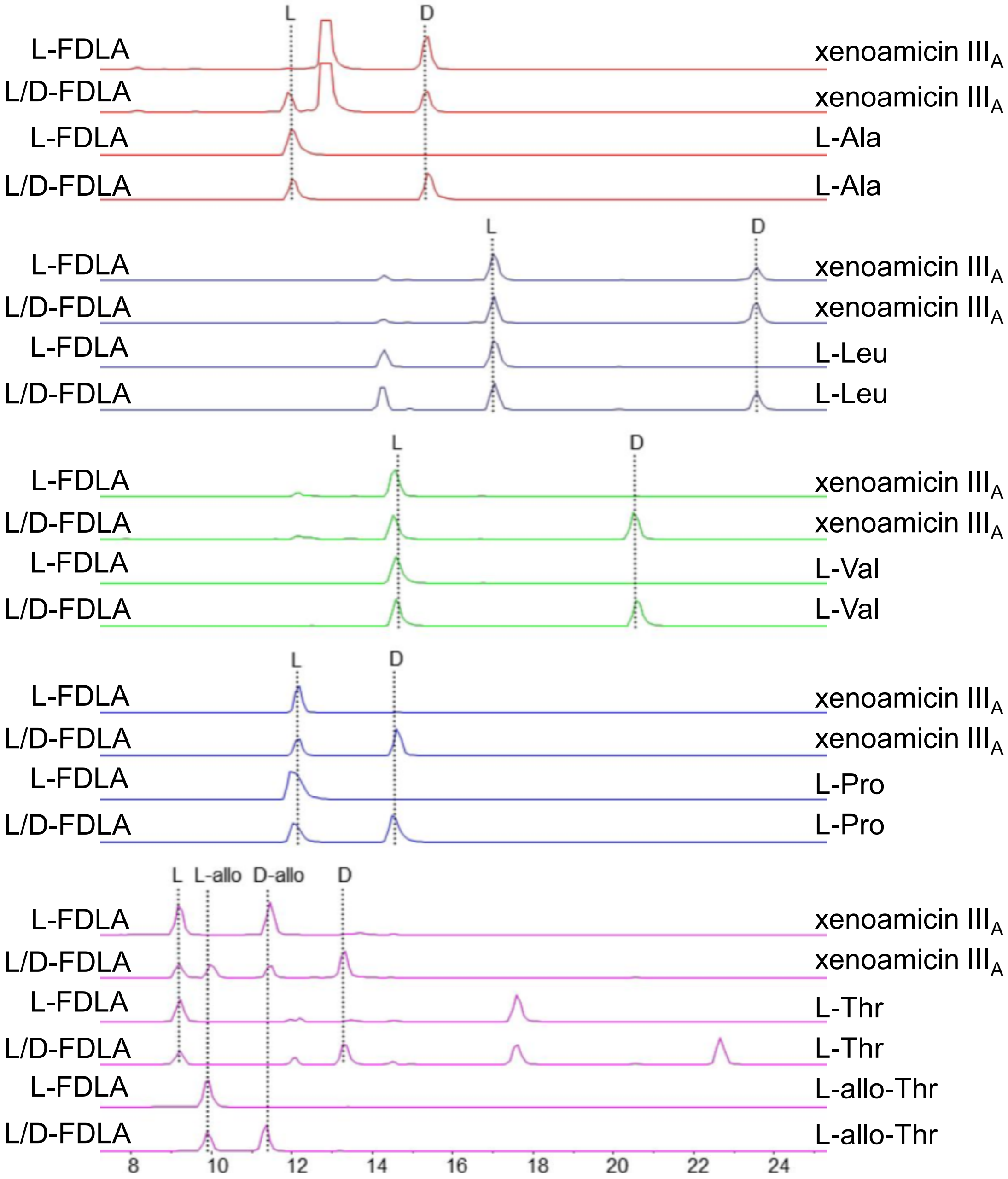


Figure 2.27. Determination of the absolute configuration of amino acids in XAM-1320 (XAM-III$_A$) by the advanced Marfey's method. The single amino acids were measured in the positive mode. The following m/z ratios ([M+H]$^+$) were used to detect the amino acids: alanine 384, leucine 426, valine 412, proline 410, threonine 414. For every amino acid the references are also shown.

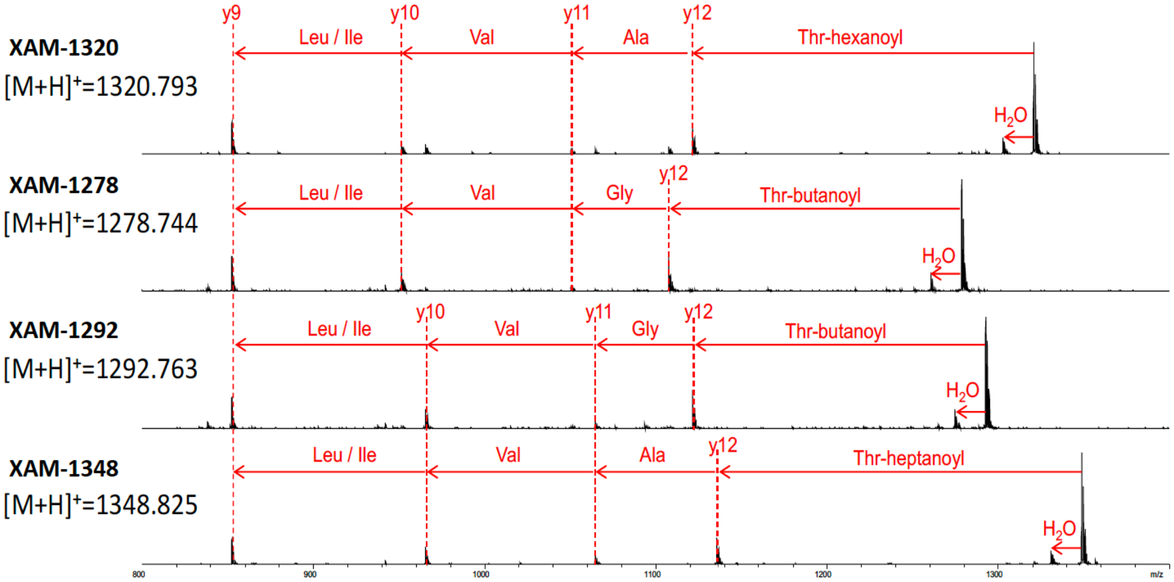


Figure 2.28. MS^2 spectra of derivatives according to subclass xenoamicin III. Compounds 14 (*m/z* = 1278.744 [M+H$^+$], 15 (*m/z* = 1292.763 [M+H$^+$] and 16 (*m/z* = 1348.825 [M+H$^+$] differ to multiple of 14 Da from compound 12. Mass differences could be localised between y12 and y10 ions.

Table 2.5. NMR spectroscopic data (600 MHz (^{1}H), 125 MHz (^{13}C) in $CDCl_3$) of XAM-1320; δ in ppm; HM, hexanoyl moiety.

Spin Sys.	Pos.	δ_C	δ_H
1-HM	C=O	173.73	
	α	36.12	2.23
	β	31.64	1.28
	γ	25.46	1.66
	δ	22.50	1.29
	ε	14.00	0.88
2-Thr	C=O	172.09	
	NH		8.38
	α	58.41	4.74
	β	67.43	4.78
	γ	19.14	1.18
3-Ala	C=O	172.48	
	NH		7.10
	α	47.25	4.66
	α	14.87	1.41
4-Val	C=O	172.44	
	NH		7.51
	α	59.32	4.26
	β	30.60	1.94
	γ1	19.10	1.00
	γ2	19.10	0.94
5-Leu	C=O	169.95	
	NH		8.50
	α	50.31	4.44
	β	39.28	1.70
	β		1.56
	γ	24.84	1.53
	δ1	22.68	0.89
	δ2	22.28	0.88
6-Leu	C=O	173.20	
	NH		7.48
	α	51.48	5.01
	β	39.09	1.65
	γ	24.97	1.64
	δ1	22.34	0.97
	δ2	21.95	0.90
7-Thr	C=O	169.44	
	NH		8.64
	α	69.34	5.26
	β	57.76	4.33
	γ	17.37	1.25
8-Thr	C=O	170.65	
	NH		8.91
	α	65.84	4.12
	β	61.22	4.52
	γ	20.00	1.28
9-Leu	C=O	170.61	
	NH		7.46
	α	51.28	4.76
	β	41.18	1.72
	β		1.49
	γ	24.88	1.53
	δ1	22.64	0.93
	δ2	22.56	0.98
	α	60.71	3.76
	β	29.43	2.07
	γ1	19.21	0.86
	γ2	19.00	0.89
11-β-Ala	C=O	172.09	
	NH		6.49
	α	37.00	3.81
	α		3.24
	β	35.66	3.24
	β		2.25
12-β-Ala	C=O	172.27	
	NH		7.53
	α	33.89	4.06
	α		3.52
	β	35.26	2.71
	β		2.47
13-Pro	C=O	172.83	
	α	60.67	4.57
	β	29.64	2.25
	β		2.07
	γ	24.33	2.00
	γ		1.93
	δ	47.46	3.68
	δ		3.40
14-β-Ala	C=O	170.71	
	NH		6.56
	α	36.31	3.58
	α		3.35
	β	33.81	2.78
	β		2.34

Table 2.6. ROE list with upper and lower distance restraint limits (90%, 110%) including pseudoatom correction from experimentally determined distance for 3D modelling of XAM-1320. Average distance and average violation of single distance restraints over ten conformations from the final MD trajectory (after energy minimization) are shown.

ROEs					
ATOM1	**ATOM2**	**LOWER**	**UPPER**	**AV_DIST**	**AV_VIOL**
8-THR NH	8-THR γ	4	5.8	4.45	0
8-THR NH	3-ALA γ	4	5.8	4.01	0
8-THR NH	8-THR β	2.5	3.1	2.92	0
4-VAL α	8-THR NH	2.8	3.4	2.86	0
7-THR α	8-THR NH	2.2	2.7	2.42	0
8-THR α	8-THR NH	2.9	3.5	2.71	0.19
7-THR β	8-THR NH	3.5	4.3	3.82	0
10-VAL NH	10-VAL β	2.6	3.2	2.41	0.19
10-VAL α	10-VAL NH	3	3.7	2.88	0.12
9-LEU α	10-VAL NH	2.2	2.6	2.44	0
6-LEU α	10-VAL NH	3.1	3.8	4.33	0.53
7-THR NH	7-THR γ	3.6	5.5	4.2	0
7-THR α	7-THR NH	2.87	3.5	2.8	0.07
6-LEU α	7-THR NH	2	2.5	2.52	0.02
7-THR NH	7-THR β	2.8	3.4	2.85	0
5-LEU NH	8-THR α	3.6	4.4	4.83	0.43
4-VAL α	5-LEU NH	2.1	2.6	2.35	0
5-LEU NH	7-THR α	3.3	4.1	4.15	0.05
5-LEU α	5-LEU NH	2.8	3.4	2.94	0
2-THR NH	2-THR γ	2.5	4	3.63	0
2-THR NH	Acyl α	2.4	3.9	2.87	0
2-THR NH	Acyl β	2.6	4	3.43	0
2-THR NH	3-ALA NH	2.3	2.8	2.81	0.01
2-THR NH	4-VAL NH	3.3	4.1	4.27	0.17
7-THR α	12-ALA NH	3.6	4.4	5.07	0.68
8-THR α	12-ALA NH	3.2	3.9	4.21	0.31
5-LEU NH	8-THR NH	4.2	5.2	4.36	0
10-VAL NH	10-VAL γ	3.1	4.8	3.63	0
7-THR β	12-ALA NH	3.7	4.5	3.38	0.32
4-VAL NH	3-ALA γ	4.3	6.2	4.5	0
4-VAL NH	4-VAL β	2.7	3.3	3.1	0
4-VAL α	4-VAL NH	2.7	3.4	2.84	0
4-VAL NH	3-ALA α	2.1	2.6	2.7	0.1
2-THR α	4-VAL NH	3.6	4.5	4.26	0
2-THR β	4-VAL NH	3.3	4.1	2.89	0.41
4-VAL NH	3-ALA NH	3.4	4.2	2.97	0.43
6-LEU NH	6-LEU β	2.4	3.9	2.45	0
5-LEU α	6-LEU NH	2.1	2.6	2.26	0
6-LEU α	6-LEU NH	2.7	3.4	3.01	0
9-LEU NH	8-THR γ	3.7	5.5	4.35	0
2-THR β	9-LEU NH	3.8	4.6	4.74	0.14
8-THR α	9-LEU NH	2	2.5	2.34	0
9-LEU α	9-LEU NH	2.7	3.3	3.06	0
3-ALA NH	7-THR γ	3.5	5.3	3.23	0.27
3-ALA NH	Acyl α	3.4	5.1	3.64	0
3-ALA NH	7-THR α	3	3.7	2.76	0.24
3-ALA α	3-ALA NH	2.7	3.4	3.01	0

Table 2.6. ROE list with upper and lower distance restraint limits (90%, 110%) including pseudoatom correction from experimentally determined distance for 3D modelling of XAM-1320. Average distance and average violation of single distance restraints over ten conformations from the final MD trajectory (after energy minimization) are shown, Continued.

ROEs					
ATOM1	**ATOM2**	**LOWER**	**UPPER**	**AV_DIST**	**AV_VIOL**
2-THR α	3-ALA NH	3	3.6	3.7	0.1
2-THR β	3-ALA NH	3.5	4.2	4.22	0.01
7-THR β	7-THR γ	2.4	2.9	2.49	0
7-THR β	3-ALA γ	4	5.9	5.75	0
7-THR β	Acyl α	3.1	4.7	5.09	0.39
6-LEU α	10-VAL β	2.5	3	3.42	0.42
9-LEU α	9-LEU δ	2.6	4.2	4.67	0.47
9-LEU α	9-LEU δ	2.6	4.2	3.17	0
2-THR β	2-THR γ	2.2	3.7	2.47	0
2-THR α	2-THR γ	2.3	3.8	2.88	0
3-ALA α	3-ALA γ	2.1	3.6	2.66	0
13-PRO α	8-THR γ	2.6	4.2	3.58	0
13-PRO α	13-PRO δ	2.6	3.2	3.48	0.28
13-PRO α	13-PRO δ	3.6	4.5	4.15	0
13-PRO α	13-PRO β	2.6	3.2	2.24	0.36
8-THR α	8-THR γ	2.3	3.9	2.82	0
5-LEU α	5-LEU δ	2.4	3.9	3.59	0
5-LEU α	5-LEU γ	2.6	3.15	2.51	0.09
7-THR α	7-THR γ	2.4	3.9	3.09	0
7-THR α	3-ALA γ	2.6	4.2	4.19	0
4-VAL β	7-THR α	3.6	4.5	5.01	0.51
4-VAL α	4-VAL β	2.7	3.3	2.38	0.32
4-VAL α	8-THR β	2.1	2.6	2.34	0
4-VAL α	4-VAL γ	2.5	3.1	3.7	0.6
4-VAL α	4-VAL γ	2.5	3.1	3.16	0.06
8-THR β	4-VAL γ	2.2	3.7	3.36	0
8-THR β	4-VAL γ	3.1	4.8	5.15	0.35
8-THR β	8-THR γ	3	4.6	2.71	0.29
10-VAL α	10-VAL γ	2.5	4.1	3.44	0
10-VAL α	10-VAL γ	2.5	4.1	3.04	0
10-VAL α	10-VAL β	2.8	3.4	3	0
12-ALA α1	12-ALA NH	2.4	3	2.27	0.13
12-ALA α1	12-ALA β1	2.5	3	2.58	0
12-ALA α2	12-ALA β2	2.4	3	2.57	0
12-ALA α2	12-ALA β1	3.3	4.1	3.16	0.14
12-ALA β1	13-PRO δ	2.6	3.2	3.46	0.26
12-ALA β1	13-PRO δ	2.2	2.7	2.48	0
12-ALA β2	13-PRO δ	2.3	2.8	2.56	0
12-ALA β2	13-PRO δ	2.8	3.5	2.61	0.19

Table 2.6. ROE list with upper and lower distance restraint limits (90%, 110%) including pseudoatom correction from experimentally determined distance for 3D modelling of XAM-1320. Average distance and average violation of single distance restraints over ten conformations from the final MD trajectory (after energy minimization) are shown, Continued.

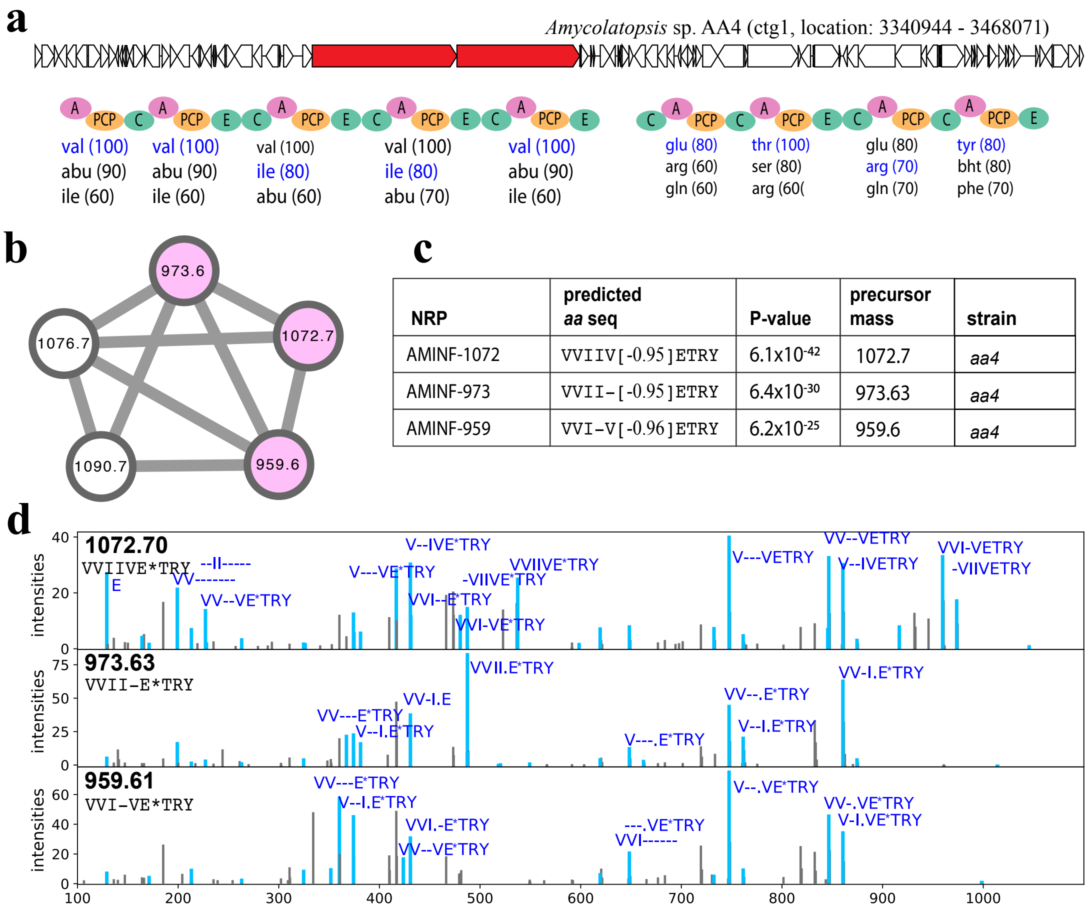
ROEs					
ATOM1	**ATOM2**	**LOWER**	**UPPER**	**AV_DIST**	**AV_VIOL**
9-LEU NH	12-ALA α2	2.6	3.3	2.88	0
8-THR NH	14-ALA α	3.8	5.5	3.58	0.22
8-THR NH	14-ALA β	3.9	5.7	4.89	0
7-Thr NH	14-ALA α	3.6	5.3	4.47	0
5-LEU NH	5-LEU β	3	3.9	2.43	0.57
5-LEU NH	5-LEU β	3.2	3.9	3.74	0
5-LEU NH	9-LEU α	3.1	3.8	3.9	0.1
4-VAL NH	4-VAL γ	3	3.7	2.76	0.24
4-VAL NH	4-VAL γ	3.6	4.4	4.64	0.24
9-LEU NH	9-LEU β	3.1	3.8	2.53	0.57
9-LEU NH	9-LEU β	3.3	4	3.78	0
6-LEU α	6-LEU β	2.6	4.3	2.59	0.01
9-LEU α	9-LEU β	2.6	4.1	2.53	0.07
5-LEU α	5-LEU β	2.4	3.9	2.82	0
9-LEU α	10-VAL γ	3.2	4.9	4.23	0
7-THR β	12-ALA α1	3.4	4.2	4.36	0.16
7-THR β	12-ALA β1	2	2.5	2.26	0
13-PRO α	14-ALA β	4.3	6.1	6.27	0.17
8-THR α	14-ALA α	3.1	4.8	4.42	0
7-THR α	14-ALA α	3.5	5.12	5.26	0.14
7-THR α	12-ALA β2	3	3.7	4.45	0.75
11-ALA NH	10-VAL γ	3.4	4.2	3.68	0
14-ALA NH	13-PRO γ	3.5	5.2	3.41	0.09
13-PRO α	14-ALA NH	2.9	3.6	3.51	0
14-ALA NH	13-PRO δ	2.9	3.6	3.07	0
8-THR α	14-ALA NH	3.1	3.8	4.06	0.26
Average Restraint Violation: 0.114 Average RMS RestrViolation: 0.116					

Novel aminformatide NRP family produced by *Amycolaptosis sp. aa4* in the SoilActi dataset. Table 2.7 presents the number of NRP-producing BGCs and the number of putative core NRPs generated by NRPminer for each analyzed genome in SoilActi (before and after filtering).

Table 2.7. The number of predicted core NRPs before and after filtering for the genomes of the 20 soil-dwelling Actinobacteria strains in SoilActi. The columns show the number of NRP-producing BGCs (column "#NRP-producing BGC") along with the number core NRPs generated by the canonical and non-canonical assembly lines for each genome before and after filtering by NRPminer using *OrfDel* option. Column "removing no ORFs" shows the number of core NRPs generated from the canonical assembly lines before and after filtering. For example, in case of *S. albus* genome, NRPminer produces 102,852,968,758 core NRPs before filtering, while after filtering only 2,368 core NRPs are retained. Column "removing one ORF" shows the number of core NRPs generated from all non-canonical assembly lines resulted from removing A-domains encoded by one ORF on the corresponding BGC, before and after filtering with NRPminer. Column "removing two ORFs" shows this figure for non-canonical assembly lines generated by removing A-domains encoded by two ORFs. Column "total" shows the total number of core NRPs before and after filtering across all considered assembly lines for each organism. The strains corresponding to the datasets yielding the novel NRPs in SoilActi are shown in blue.

		#unique core NRPs before / after filtering generated by different assembly lines			
strain	**#NRP-producing BGCs**	**removing no ORFs**	**removing one ORF**	**removing two ORFs**	**total**
SCNY228	3	2,369/102,852,968,758	5,759/1,537,478,841	7,483/4023,756	15,611/104,394,471,355
albus	3	3,189/25,713,264,922	5,788/473,652,036	7,471/2237,220	16,460/2,618,9154,178
CNS654	5	1,560/21,499,085,734	3,870/87,589,011	2,331/45,216	7,761/21,586,719,961
griseoflav	7	3,235/17,916,143,265	6,431/75,146,556	2,484/45,695	12,150/17,991,335,516
hygro	5	3,753/79,748,772	12,887/27,905,444	11,964/5481,248	28,604/113,135,464
15998	3	2,436/19,088,674	8,084/49,356,874	19,156/43,902,448	29,676/112,347,996
coelicolor	3	1,191/787,524	1,693/75,438	91/819	2,975/863,781
lividan	2	1,032/262,476	2,662/178,686	1,572/31644	5,266/472,806
ghana	2	1,666/115,488	5,516/246,416	6,137/146728	13,319/508,632
kutzneria	9	4,983/47,046	9,866/73,172	5,748/53050	20,597/173,268
aa4	2	1,381/111,780	798/2,554	103/351	2,282/114,685
CNB091	3	960/29,448	603/16,976	290/3124	1,853/49,548
cattleya	4	1,300/23,068	1,475/6,165	77/225	2,852/29,458
11379	4	1,643/6,853	2,800/11,961	1,632/6,882	6,075/25,696
griseoflav	2	2,173/15,488	1,240/3,016	368/368	3,771/18,872
tu6071	4	1,674/10,638	0/0	0/0	1,674/10,638
pristin	2	864/864	279/279	0/0	1,143/1,143
afghan	0	252/252	288/288	72/72	612/612
e14	1	240/240	0/0	0/0	240/240
viridochromoges	1	36/36	0/0	0/0	36/36

NRPminer identified 11 PSMs (representing three NRPs) when searching the **SoilActi** spectral dataset against *Amycolaptosis sp.* aa4 genome (Figure 2.29). Previously, another NRP family, siderophore amychelin, and its corresponding BGC was reported from this organism[72]. Using the NRPSpreidctor2[13]-predicted amino acids NRPminer predicted a modification of ~0.95 Da on the Glu in aminoformatide-1072 VVII[E-1.0]TRY. Since NRPSpredictor2 is the least sensitive in recognizing Lys (as compared to other amino acids)[13], we hypothesize that this amino acid is in fact a Lys as we have seen in the case of protegomycins (with Lys), but this is yet to be determined.



NRP	predicted *aa* seq	P-value	precursor mass	strain
AMINF-1072	VVIIV[-0.95]ETRY	6.1×10^{-42}	1072.7	*aa4*
AMINF-973	VVII-[-0.95]ETRY	6.4×10^{-30}	973.63	*aa4*
AMINF-959	VVI-V[-0.96]ETRY	6.2×10^{-25}	959.6	*aa4*

Figure 2.29. Novel aminformatide (AMINF) NRP family discovered by NRPminer in the *SoilActi* dataset. (*a*) The BGC generating the core NRP in *Amycolatopsis* sp. AA4 along with NRPS genes (shown in red) and the A-, C-, PCP, and E-domains appearing in the corresponding NRPS. The rest of the genes in the corresponding contigs are shown in white. Three highest-scoring amino acids for each A-domain in this BGC (according to NRPSpredictor2[13] predictions) are shown below the corresponding A-domains. Amino acids appearing in the NRP VVIVETRY identified by NRPminer (with the lowest p-value) are shown in blue. (*b*) Spectral network formed by spectra that originate from the AMINF NRPs. A node is colored if the corresponding spectrum forms a statistically significant PSM and not colored otherwise. (*c*) Sequences of the NRPs identified by NRPminer in the aminformatide family (with the lowest p-value among all PSMs originating from the same NRP). NRPminer predicted a PAM with loss of ~0.96 Da on E, represented by E*. AMINF represents aminformatide. (*d*) For each identified NRP, an annotated spectrum representing the lowest p-value is shown.

Identifying lugdunin NRP family in the SkinStaph dataset. Antibiotics lugdunins[7] represent the only NRP family reported in the human commensal microbiota. NRPminer matched nine spectra representing three NRPs from a single family in the *spectra*$_{SkinStaph}$ dataset against

Staphylococcus lugdunensin genome. In addition to the two known cyclic variants of lugdunin, NRPminer also discovered a novel lugdunin variant with precursor mass 801.52 (Figure 2.30). Due to a +18.01Da mass difference, NRPminer predicted a linear structure for this new variant that represents the linear version of the known one. Since NRPminer predicts sequence VWLVVVt for the linear lugdunin, with the breakage between valine and Cys-derived thiazolidine, we hypothesize that this is a naturally occurring linear derivative in the lugdunin family. Lugdunins, synthesized by a non-canonical assembly line, were predicted using the non-canonical assembly line feature of NRPminer.

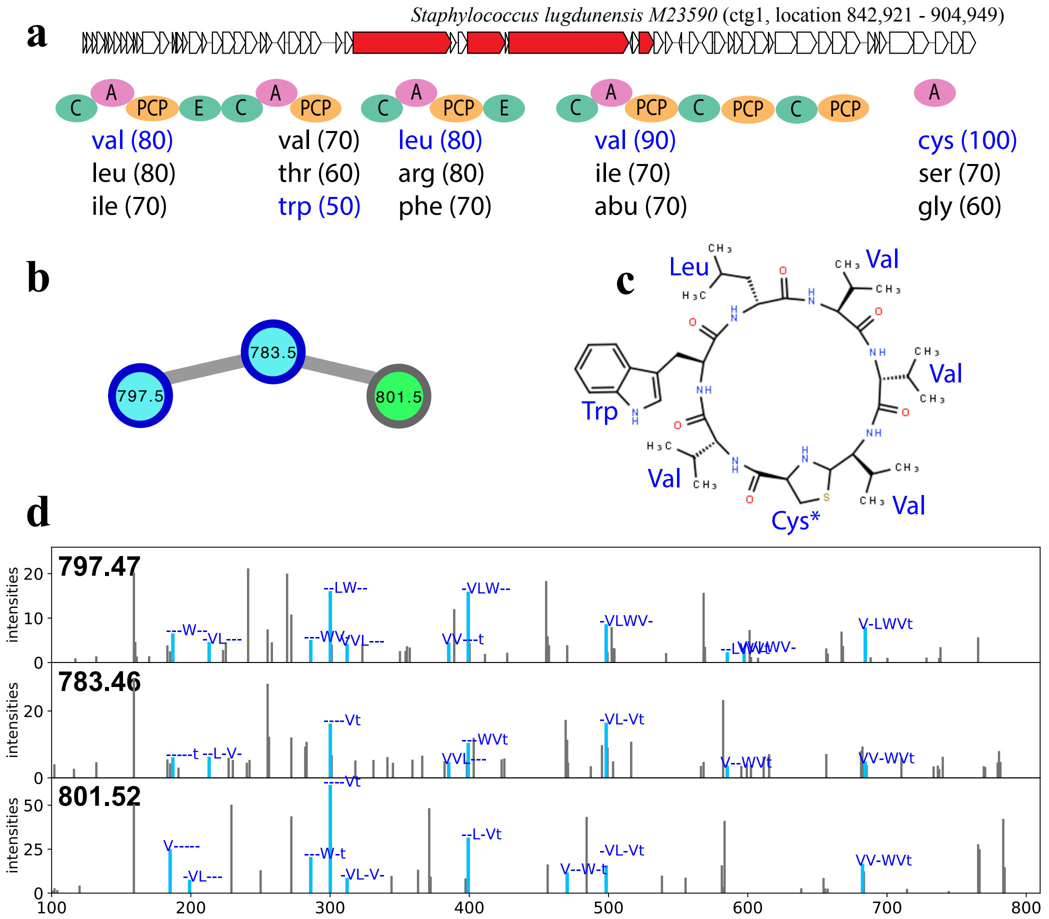


Figure 2.30. Lugdunin NRP family matched by NRPminer in the SkinStaph dataset. (*a*) The BGC generating the core NRP in *S. lugdunensin* along with NRPS genes (shown in red) and the A-, C-, PCP-, and E-domains appearing in the corresponding NRPS. The rest of the genes in the corresponding contigs are shown in white. Three highest-scoring amino acids for each A-domain in this BGC (according to NRPSpredictor2[13] predictions) are shown below the corresponding A-domains. Amino acids appearing in the NRP VYLVV identified by NRPminer (with the lowest p-value) are shown in blue. The "Cys*" represent Cys-derived thiazolidine in the lugdunin structure. (*b*) Spectral network formed by spectra that originate from the NRPs in the lugdunin family. The known lugdunin NRPs are shown in blue, while the green node represents the novel variant identified by NRPminer. (*c*) Structure of a known lugdunin synthesized by a non-canonical assembly line. (*d*) For each matched NRP, an annotated spectrum of a PSM yielding the lowest p-values (2.7×10^{-21}, 3.6×10^{-15}, and 7.5×10^{-15} from top to bottom) are shown.

Identifying novel lipopeptides in the TinyEarth dataset. Our NRPminer analysis of the TinyEarth dataset generated 498 PSMs representing 31 NRPs from three families using the 200

Da threshold for PAM identification. Table 2.8 provides information about the NRPminer-generated PSMs representing these three NRP families.

Table 2.8. PSMs identified by NRPminer in the TinyEarth dataset representing the known NRP families. For each NRP family, the information about the PSM with the lowest p-value among all PSMs corresponding to the spectra representing the NRPs in that family, is listed. The column “matched genome” shows the name of the organism whose BGCs generated the putative NRP structure corresponding to the listed PSM and the column "BGC position” presents the contig and the starting and ending nucleotide position of the BGC in that contig. Columns “precursor mass” and “charge” list the precursor mass and the charge state of the matched spectra.

NRP family name	matched genome	BGC position	p-value	precursor mass	charge
Surfactin	*Bacillus amyloliquefaciens* sp. GZYCT-4-2	ctg1: 416695 - 482102	1.6×10^{-46}	1036.7	1
Plipastatin	*Bacillus amyloliquefaciens* sp. GZYCT-4-2	ctg1: 2727818 - 2749701	7.0×10^{-55}	731.4	2
Arthrofactin	*Pseudomonas baetica* sp. 04-6(1)	ctg1: 3,566,169 - 3,642,017	2.7×10^{-39}	1354.8	2

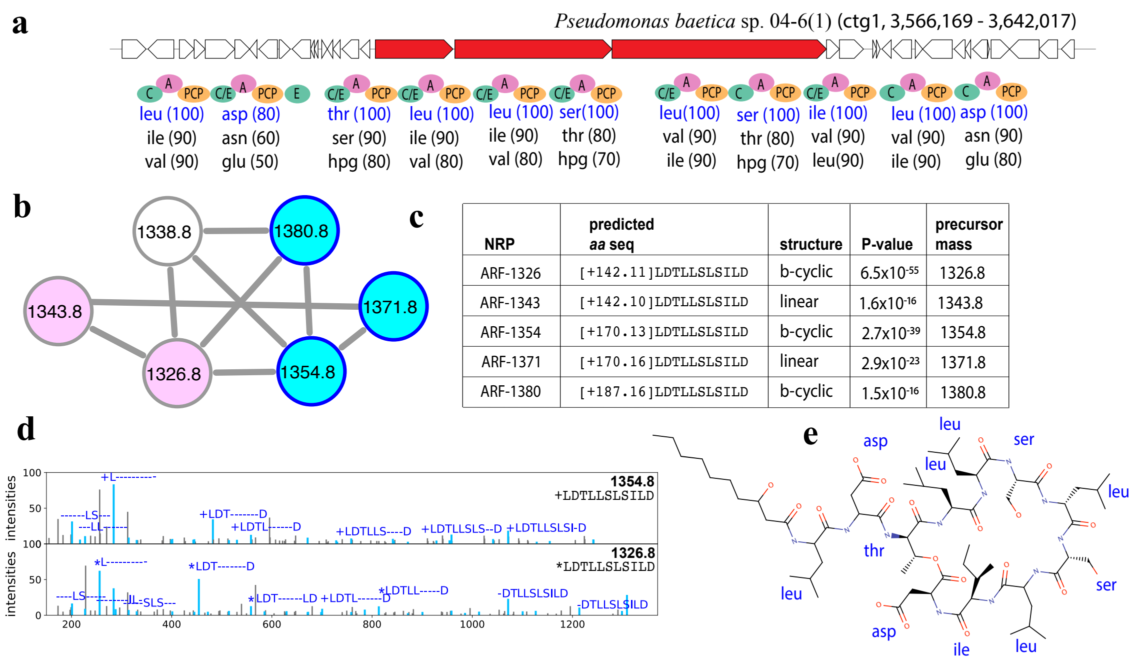
Bacillus derived surfactins[73] and plipastatin[74] are bioactive lipopeptide with wide variety of activities. Surfactins are reported to have anti-viral[75,76], anti-tumor[77], anti-fungal[78] and anti-microbial[79] functions[80–83] and plipastatins have known anti-fungal activities[84]. In the analysis of *Bacillus amyloliquefaciens* sp. GZYCT-4-2, NRPminer correctly reported all known surfactins (17 NRPs) and plipastatins (9 NRPs) identified in this dataset (PSMs listed in Table 2.9).

Table 2.9. NRPminer-generated PSMs representing all known surfactins[85] and plipastatins[74,86] identified in *spectra*$_{TinyEarth}$ dataset. For each known NRP, the PSM with the lowest p-value among all PSMs corresponding to the spectra generated from that NRP, is listed. The columns "core NRP aa sequence" and "structure" presents the core NRP and the backbone structure of each variant identified in TinyEarth dataset. Column "precursor mass" and "charge" lists the precursor mass and the charge state of the matched spectra.

NRP family name	core NRP aa sequence	structure	precursor mass	p-value	charge
Surfactins	`ELLVDLL`	cyclic	966.5	2.5×10^{-23}	1
	`ELLVDLL`	cyclic	980.6	3.4×10^{-30}	1
	`ELLVDLL`	cyclic	994.7	4.0×10^{-35}	1
	`ELLVDLL`	cyclic	1008.7	2.9×10^{-45}	1
	`ELLVDLL`	linear	1012.7	2.1×10^{-17}	1
	`ELLIDLL`	cyclic	1022.7	1.5×10^{-41}	1
	`ELLVDLL`	linear	1026.7	3.3×10^{-19}	1
	`ELLVDLL`	cyclic	1029.7	8.1×10^{-20}	1
	`ELLIDLL`	cyclic	1036.7	1.6×10^{-46}	1
	`ELLVDLL`	linear	1040.7	9.0×10^{-19}	1
	`ELLVDLL`	cyclic	1044.7	9.2×10^{-16}	1
	`ELLVDLL`	cyclic	1050.7	2.1×10^{-28}	1
	`ELLVDLL`	linear	1054.7	6.8×10^{-16}	1
	`ELLIDLL`	cyclic	1057.7	7.8×10^{-24}	1
	`ELLVDLL`	cyclic	1064.7	2.0×10^{-41}	1
	`ELLVDLL`	linear	1068.7	6.4×10^{-31}	1
	`ELLVDLL`	cyclic	1071.7	3.9×10^{-22}	1
Plipastatins	`EOYTEAPQYI`	cyclic	718.4	9.6×10^{-33}	2
	`EOYTEAPQYI`	cyclic	724.4	2.9×10^{-30}	2
	`EOYTEAPQYI`	cyclic	725.4	2.5×10^{-38}	2
	`EOYTEAPQYI`	cyclic	731.4	7.0×10^{-55}	2
	`EOYTEAPQYI`	cyclic	732.4	2.3×10^{-37}	2
	`EOYTEVPQYI`	cyclic	739.4	3.2×10^{-49}	2
	`EOYTEVPQYI`	cyclic	746.4	5.5×10^{-43}	2
	`EOYTEVPQYI`	cyclic	753.4	6.1×10^{-42}	2
	`EOYTEVPQYI`	cyclic	760.4	2.6×10^{-21}	2

Moreover, NRPminer search of *spectra*$_{TinyEarth}$ against putative NRP structures generated from *Pseudomonas baetica* sp. 04-6(1) genome, identified 63 PSMs representing the arthrofactins

(ARF) NRP family (Figure 2.31). NRPminer identified the known branch-cyclic arthrofactins[87] that only differ in the fatty acid tail (namely ARF-1354 and ARF-1380) and a known linear arthrofactin ARF-1372 (the linear version of ARF-1354). Furthermore, it identified two novel arthrofactins: ARF-1326 (predicted to only differ in its side chain from the known branch-cyclic ARF-1354 shown in Figure 2.31.e) and ARF-1343 (predicted to be the linear version of the putative ARF-1326). NRPminer missed one known NRP family identified in $spectra_{TinyEarth}$ (xantholysins[88]) since the xantholysin BGC was split among multiple contigs in the *Pseudomonas plecoglossicida* sp. YNA158 genome assembly

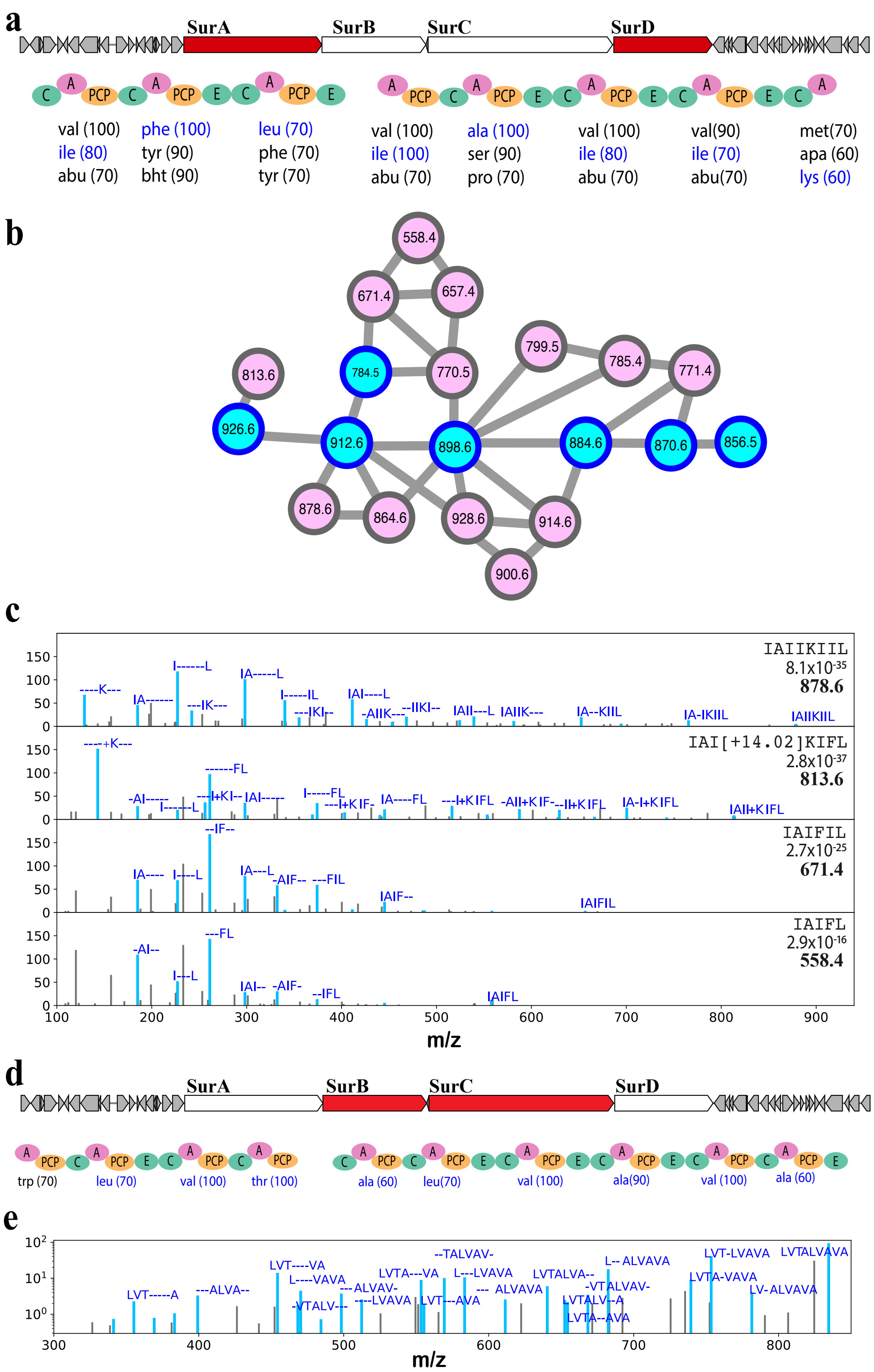

a
Pseudomonas baetica sp. 04-6(1) (ctg1, 3,566,169 - 3,642,017)
C A PCP C/E A PCP E
leu (100) ile (90) val (90)
asp (80) asn (60) glu (50)
C/E A PCP C/E A PCP C/E A PCP C/E A PCP
thr (100) ser (90) hpg (80)
leu (100) ile (90) val (80)
leu (100) ile (90) val (80)
ser(100) thr (80) hpg (70)
C/E A PCP C A PCP C/E A PCP C A PCP C A PCP
leu(100) val (90) ile (90)
ser (100) thr (80) hpg (70)
ile (100) val (90) leu(90)
leu (100) val (90) ile (90)
asp (100) asn (90) glu (80)
b
1338.8
1380.8
1343.8
1371.8
1326.8
1354.8
c
d
1354.8
+LDTLLSLSILD
1326.8
*LDTLLSLSILD
intensities
intensities
e
asp
leu
leu
ser
leu
thr
ser
leu
asp
ile
leu

NRP	predicted *aa* seq	structure	P-value	precursor mass
ARF-1326	[+142.11]LDTLLSLSILD	b-cyclic	$6.5x10^{-55}$	1326.8
ARF-1343	[+142.10]LDTLLSLSILD	linear	$1.6x10^{-16}$	1343.8
ARF-1354	[+170.13]LDTLLSLSILD	b-cyclic	$2.7x10^{-39}$	1354.8
ARF-1371	[+170.16]LDTLLSLSILD	linear	$2.9x10^{-23}$	1371.8
ARF-1380	[+187.16]LDTLLSLSILD	b-cyclic	$1.5x10^{-16}$	1380.8

Figure 2.31. Arthrofactin (ARF) NRP family. (*a*) The BGCs generating the NRP in *Pseudomonas baetica* sp. 04-6(1) along with the NRPS genes (shown in red) and A-, C-, C/E-, PCP-, and E-domains in these NRPSs. The rest of the genes in the corresponding contigs are shown in white. Three highest-scoring amino acids for each A-domain in these BGCs (according to NRPSpredictor2[13] predictions) are shown below the corresponding A-domains. Amino acids appearing in the known NRP ARF-1354 with amino acid sequence [+170.13]LDTLLSLSILD are shown in blue. (*b*) Spectral network formed by the spectra that originate from NRPs in the ARF family. The known arthrofactins are shown in blue, while the purples nodes represent the novel variants identified by NRPminer. all identified athrofactins share the same core NRP LDTLLSLSILD. (*c*) Sequences of the identified NRPs in this family (with the lowest p-value among all spectra originating from the same NRP). Column "structure" shows if the predicted structure for the identified NRPs is linear or branch-cyclic (shown by b-cyclic). (*d*) Two annotated spectra representing the PSMs (with the lowest p-values among spectra originating from the same NRPs) corresponding to ARF-1354 and 1326. The two spectra were annotated based on predicted NRPs [+170.13]LDTLLSLSILD (PSM p-value $2.7×10^{-39}$) and [+142.11]LDTLLSLSILD (PSM p-value $6.5×10^{-55}$), from top to bottom. The "+" and "*" signs represent the addition of [+170.13] and [+142.11], respectively. (*e*) The 2D structure of known arthrofactin ARF-1354[87]. NRPminer identified this NRP with p-value $2.7×10^{-39}$.

Identifying novel surugamides in the SoilActi dataset. NRPminer identified 183 spectra representing 25 NRPs when searching *spectra*$_{SoilActi}$ against *S. albus* J10174 genome, hence extending the set of known surugamide variants from 8 to 21 (Figure 2.32 and Table 2.8). Spectral network analysis revealed that these spectra originated from two NRP families. VarQuest search of this spectral dataset against PNPdatabase[41] identified only 14 of these 21 NRPs. The remarkable diversity of surugamide NRPs, that range in length from 5 to 10 amino acids, is explained by the non-canonical assembly lines[29,41]. In addition to the surugamides synthesized by the SurA-SurD pair, NRPminer also discovered a novel Surugamide G synthesized by the SurB-SurC pair (Figure 2.32.d). In comparison with surugamide F from *Streptomyces albus*[31], this NPR lacks the N-terminal tryptophan. Surugamide F was not identified in the spectral dataset from *Streptomyces albus*.

Figure 2.32. Known and novel surugamide variants identified by NRPminer in the *SoilActi* dataset. Suragamide BGC contains four successive genes, namely SurA, SurB, SurC, and SurD with five, four, six, and three A-domains, respectively. SurA and SurD synthesize cyclic surugamides A-D using a non-canonical assembly line, while SurB and SurC synthesize a linear surugamide F. (a) Surugamide BGC from *S. albus* with SurA and SurD highlighted in red, while SurB and SurC are shown in white. In the middle, A-, C-, PCP-, and E-domains appearing in the corresponding NRPS are shown. Three highest-scoring amino acids for each A-domain in this NRPS (according to NRPSpredictor2[13] predictions) are shown below the corresponding A-domains. Amino acids appearing in surugamide A (IFLIAIIK) are shown in blue. (b) Spectral network formed by spectra that originated from cyclic surugamides (corresponding to the NRPS shown in part a) including the seven known cyclic surugamides. The known cyclic surugamides are shown in blue, while the purples nodes represent the novel cyclic variants identified by NRPminer. (c) NRPminer predicted novel cyclic surugamides with eight, seven, six, and five amino acids. For each length, the annotated spectrum representing the lowest p-value (among all PSMs corresponding to the identified novel surugamides with that length) is presented. Amino acid sequence, p-value, and precursor mass of each PSM is shown in the top right corner. Annotated peaks are shown in blue. The spectra were annotated based on predicted NRPs IAIIKIIL, IAIKIFL, IAIFIL, IAIFL, from top to bottom. The "+" sign represent the addition of [+14.02Da]. Table 2.9 shows the predicted amino acids and p-values for all NRPs represented by the nodes in part b. (d) Surugamide BGC from *S. albus* with SurB and SurC highlighted in red, while SurA and SurD are shown in white. In the middle, A-, C-, PCP-, and E-domains appearing in the corresponding NRPS are shown. The highest-scoring amino acids for each A-domain in this NRPS (according to NRPSpredictor2[13] predictions) are shown below the corresponding A-domains. Amino acids appearing in the novel surugamide G (LVTALVAVA) are shown in blue. The amino acid shown in black did not appear in the predicted surugamide G. (e) Annotated spectrum representing the novel surugamide G (synthetized by the NRPS shown in part d) with the lowest p-value among all spectra representing this NRP (p-value=$5.0\times \mathbf{10^{-46}}$). Annotated peaks are shown in blue.



a
SurA
SurB
SurC
SurD
C A PCP C A PCP E C A PCP E
A PCP C A PCP E C A PCP E C A PCP E C A
val (100)
ile (80)
abu (70)
phe (100)
tyr (90)
bht (90)
leu (70)
phe (70)
tyr (70)
val (100)
ile (100)
abu (70)
ala (100)
ser (90)
pro (70)
val (100)
ile (80)
abu (70)
val(90)
ile (70)
abu(70)
met(70)
apa (60)
lys (60)
b
558.4
671.4
657.4
784.5
770.5
799.5
785.4
771.4
813.6
926.6
912.6
898.6
884.6
870.6
856.5
878.6
864.6
928.6
914.6
900.6
c
IAIIKIIL
8.1x10-35
878.6
----K---
IA------
I------L
---IK---
IA-----L
I-----IL
---IKI--
IAI----L
-AIIK---
--IIKI--
IAII---L
IAIIK---
IA--KIIL
IA-IKIIL
IAIIKIIL
IAI[+14.02]KIFL
2.8x10-37
813.6
---+K---
-AI-----
I------L
---I+KI--
IAI-----
------FL
I-----FL
---I+KIF-
IA----FL
---I+KIFL
-AII+KIF-
--II+KIFL
IA-I+KIFL
IAII+KIFL
IAIFIL
2.7x10-25
671.4
IA----
I----L
--IF--
IA---L
-AIF--
---FIL
IAIF--
IAIFIL
IAIFL
2.9x10-16
558.4
-AI--
I---L
---FL
IAI--
-AIF-
--IFL
IAIFL
0
50
100
150
100
200
300
400
500
600
700
800
900
m/z
d
SurA
SurB
SurC
SurD
A PCP C A PCP E C A PCP C A PCP
C A PCP C A PCP E C A PCP E C A PCP E C A PCP C A PCP E
trp (70)
leu (70)
val (100)
thr (100)
ala (60)
leu(70)
val (100)
ala(90)
val (100)
ala (60)
e
10^0
10^1
10^2
LVT-----A
---ALVA--
LVT----VA
L----VAVA
-VTALV---
---ALVAV-
----LVAVA
LVTA---VA
--TALVAV-
LVT---AVA
L---LVAVA
--- ALVAVA
LVTALVA--
-VTALVAV-
LVTALV--A
LVTA--AVA
L-- ALVAVA
LVTA-VAVA
LVT-LVAVA
LV-ALVAVA
LVTALVAVA
300
400
500
600
700
800
m/z

Table 2.10. Amino acid sequences of the 19 NRPs identified by NRPminer appearing in spectral network presented in Figure 2.23.b (with the lowest p-value among the PSMs corresponding to all spectra originating from the same NRP). The known surugamide variants are shown in green. The column "predicted aa sequence" shows the sequence of corresponding NRPs as predicted by NRPminer. The "[+14]" represents addition of [+14.01Da] and "[+28]" represents addition of [+28.03Da]. Column "precursor mass" shows the precursor mass of the matched spectra and the column "p-vale" presents the p-value of the corresponding PSMs.

predicted aa Sequence	precursor mass	p-value
IAI---FL	558.37	9.2×10^{-16}
IAV--IFL	657.44	4.9×10^{-19}
IAI--IFL	671.45	3.1×10^{-32}
IAII-IFL	770.52	3.1×10^{-27}
IAV-KVFL	771.52	1.3×10^{-44}
IAII-IFL	784.54	3.5×10^{-20}
IAV-KIFL	785.53	8.1×10^{-47}
IAI-KIFL	799.55	6.4×10^{-43}
IAI-[+14]KIFL	813.56	5.6×10^{-50}
VAVVKVFL	856.57	4.9×10^{-45}
IAIVKIIL	864.63	4.1×10^{-55}
IAVVKVFL	870.59	8.7×10^{-73}
IAIIKIIL	878.65	1.4×10^{-27}
IAVVKIFL	884.60	2.6×10^{-59}
IAIVKIFL	898.62	3.3×10^{-67}
IAIIKIFL	912.63	6.9×10^{-65}
IAIVKIYL	914.61	3.5×10^{-43}
IAII[+14]KIFL	926.65	1.3×10^{-56}
IAII[+28]KIYL	928.63	1.9×10^{-56}

2.4. DISCUSSION

We developed the scalable and modification-tolerant NRPminer tool for automated NRP discovery by integrating genomics and metabolomics data. We used NRPminer to match multiple publicly available spectral datasets against 241 genomes from RefSeq[60] and genome online database (GOLD)[64]. NRPminer identified 55 known NRPs (13 families) whose BGCs have been identified previously, without having any prior knowledge of them (Figure 2.3, Figure 2.4, Figure 2.30, and Figure 2.31, and Table 2.3 and Table 2.9). Furthermore, NRPminer identified the BGC for an *orphan* NRP family (xentrivalpeptides) with previously unknown BGC (Figure 2.5). In addition to the known NRPs, NRPminer reported 121 novel NRPs from a diverse set of microbial organisms. Remarkably, NRPminer identified four completely novel NRP families (representing 25 novel NRPs), three in the XPF dataset (Figure 2.6, Figure 2.21, and Figure 2.22)and one in the SoilActi dataset (Figure 2.29), illustrating that it can match large spectral datasets against multiple bacterial genomes for discovering novel NRPs that evaded identification using previous methods. We further validated two of the novel families predicted by NRPminer using NMR and demonstrated their anti-parasite activities.

Existing peptidogenomics approaches are too slow (and often memory-intensive) to conduct searches of large MS datasets against many genomes. Moreover, these approaches are limited to NRPs synthesized by canonical assembly lines and without PAMs, which limits the power of these methods for discovering novel NRPs. NRPminer is the first peptidogenomics tool that efficiently filters core NRPs based on their specificity scores without losing sensitivity and enables searching millions of spectra against thousands of microbial genomes. Furthermore, NRPminer can identify NRPs with non-canonical assembly lines of different types (e.g., surugamides, xenoinformycin and lugdunin) and PAMs (e.g. surfactins, arthrofactins, plipastatins, protegomycins, and PAX peptides).

Majority of the spectral datasets in GNPS are currently not accompanied by genomics/metagenomics data. To address this limitation, NRPminer can search a spectral dataset against all genomes from RefSeq[60] or GOLD databases[64] within a user-defined taxonomic clade. This one-vs-all mode enables analysis of spectral datasets that are not paired with genomic/metagenomic data by searching them against multiple genomes. This mode, that relies on the scalability of NRPminer, enabled NRPminer to identify the lugdunin family (by searching the SkinStaph spectral dataset) even though the paired genome sequence from the same strain was not available.

In contrast to the previous peptidogenomics approaches, NRPminer is robust against errors in specificity prediction in genome mining tools and can efficiently identify mature NRPs with PAMs. This feature was crucial for finding the novel protegomycins that include a PAM (lipid chain) and a mis-prediction (Phe instead of Lys), as well as for identifying the lipopeptide biosurfactant in the TinyEarth dataset. While NRPminer is a powerful tool for discovering novel NRPs it can only succeed if the genome mining algorithms successfully identify an NRP-encoding BGC and predicts the correct amino acids for nearly all A-domains. (see Methods section). One of the bottlenecks of genome mining methods for NRP discovery is the lack of training data for many non-standard amino acids from under-explored taxonomic clades. We anticipate that more NRPs will be discovered using automated methods, and these discoveries will increase the number of A-domain with known specificity, which in turn will pave the path toward the development of more accurate machine learning techniques for A-domains specificity prediction.

In case of *metagenomic* datasets, NRPminer's one-vs-all function allows for searching the spectral dataset generated from a sample against all the *metagenomic assemblies* generated from that same sample. However, the success of genome mining crucially depends on capturing the entire BGCs in a single contig during genome assembly. NRPS BGCs are long (average length ~60 kb[45]) and repetitive (made up of multiple highly similar domains), making it difficult to assemble them into a single contig. Meleshko *et al.*[45], recently developed the biosyntheticSPAdes tool for BGC reconstruction in short-read isolate assemblies, but at the same time acknowledged that short-reads metagenome assemblies are not adequate for full-length BGC identification. Even with biosyntheticSPAdes[45], it remains difficult to capture long and repetitive BGCs within a single contig. The one-vs-all approach can also mitigate this deficiency in cases where the BGCs corresponding to an NRP is poorly assembled, however, a *related* BGC from another genome, in that sample or another, is assembled in full. Furthermore, we note that with the advent of long-read sequencing technologies, more contiguous microbial genome assemblies are becoming available[84], increasing the power of NRPminer. Kolmogorov *et al.*[85] recently demonstrated that long-read sequencing is effective in identifying NRP-producing BGCs in metagenomic samples.

NRPminer only considers methylation and epimerization tailoring enzymes in the BGCs and does not recognize any other modification enzymes that modify NRPs, such as glycosylation and acylation[86]. These modifications can only be predicted as blind modifications using the modification tolerant search of their corresponding spectral datasets against the input genomes.

Currently, NRPminer identified ~1% of spectra of isolated microbes as NRPs. However, ~99% of spectra in these datasets remain unidentified, representing the dark matter of metabolomics. These spectra could represent primary metabolites (e.g. amino acids), other classes of secondary metabolites (e.g. RiPPs, polyketides, lipids, terpenes, etc.), media contaminations, and lower intensity/quality spectra that are difficult to identify. Thus, further advances in experimental and computational mass spectrometry are needed toward a comprehensive illumination of the dark matter of metabolomics.

2.5. METHODS

Outline of the NRPminer algorithm. NRPminer expands on the existing tools for automated NRP discovery[28,38] by utilizing new algorithms that enable high-throughput analysis and handle non-canonical assembly lines and PAMs. Below we describe various steps of the NRPminer pipeline:

(***a***) **Predicting NRPS BGCs in (meta)genome sequences by genome mining.** NRPminer uses antiSMASH[15] to identify the NRP-producing BGCs in the assembled genome. Given a genome (or a set of contigs), antiSMASH uses hidden Markov models to find NRP-producing BGCs. The NRPminer software package also includes biosyntheticSPAdes[45], a specialized short-read BGC assembler.

(***b***) **Predicting putative amino acids for each A-domain in the identified BGCs.** NRPminer uses NRPSpredictor2[13] to predict putative amino acids for each position in an NRP. Given an A-domain, NRPSpredictor2 uses support vector machines (trained on a set of A-domains with known specificities) to predict the amino acids that are likely to be recruited by this A-domain. NRPSpredictor2 provides a specificity for each predicted amino acid that is based on the similarity between the analyzed A-domain and the previously characterized A-domains[14,16]. NRPminer uses NRPSpredictor2[13] predictions to calculate the specificity scores for each predicted amino acid.

During NRP synthetase, the Adenylation domains (A-domains) recognize and activate the specific amino acid that will be appended to the growing peptide chain by other NRPS enzymes. Conti *et al.*[89] showed that some residues at certain positions on each A-domain are critical for substrate activation and bonding; they reported 10 such positions. Stachelhaus *et al.*[90] showed that for each A-domain *AD*, the residues at these decisive 10 positions can be extracted to form a specificity-conferring code called *non-ribosomal code* of *AD*. They demonstrated that the specificity of an uncharacterized A-domain can be inferred based on the sequence similarity of its *non-ribosomal code* to those of the A-domains with known specificities[90].

Given an input A-domain *AD*, NRPSpredictor2[13] first compares the sequence of the non-ribosomal code of *AD* to those of the already characterized A-domains in the NRPSpredictor2[13] database. Afterwards, for each amino acid *a*, NRPSpredictor2[13] reports the *Stachelhaus score* of (specificity of) *a* for A-domain *AD*, that is (the integer value of) the percentage of sequence identity between the *nonribosomal code* of *AD* and that of the most similar A-domain within NRPSpredictor2[13] search space that encodes for *a*.

Furthermore, Rausch *et al.*[91], expanded the set of specificity-conferring positions on A-domains to 34 residue positions and proposed a predictive model trained on residues at these 34 positions (instead of just the 10 included in *Stachelhaus code*) to provide further specificity predictions[13]. Given an A-domain, they used a Support Vector Machine (SVM) method trained on previously annotated A-domains. For each input A-domain, this approach[91] predicts three sets of amino acids in three different hierarchical levels based on the physio-chemical properties of the predicted amino acids: *large clusters*[91] (each *large cluster* is at most 8 amino acids), *small clusters*[91] (each *small cluster* is at most three amino acids), and *single amino acid prediction* (the single amino acid most likely to be activated by the given A-domain), as described by Rausch *et al.*[91] For a given A-domain *AD*, we use the terms *large cluster*, *small cluster*, and *single prediction* of *AD* to describe the sets of amino acids predicted at each of these hierarchical levels. While Rausch *et al.*[91] demonstrated that their approach reports better specificity predictions for less commonly observed A-domains, they also showed that integrating their score with the sequence similarity approach described by Stachelhaus *et al.*[90] results in the highest accuracy[91].

Similar to the approach used by NRP2Path[38], NRPminer combines the two predictions provided by NRPSpredictor2[13]. Given an A-domain *AD* and an amino acid *a*, NRPminer defines the *SVM score* of *a* for AD to be 100 if *a* matches the *single amino acid* prediction, 90 if *a* appears in the *small cluster* predictions, and 80 if *a* appears in the *large cluster*. If *a* does not appear in any of these sets, NRPminer defines the *SVM score* of *a* for *AD* to be 0. The total number of amino acids per A-domain with SVM score above 0 is at most 12 (considering all three sets of amino acids). For a given A-domain *AD*, NRPminer only considers amino acids with a predicted Stachelhaus score>50 and a predicted SVM score>0 for *AD*. Finally, NRPminer defines the *specificity* (or *NRPSpredictor2*) *score* of *a* for *AD* as the mean of *Stachelhaus* and *SVM* scores of *a* for *AD*.

(*c*) Generating multiple NRPS assembly lines. NRPminer generates multiple NRPS assembly lines by allowing for the option to either delete an entire ORFs, referred to as "*orfDel*" or duplicate A-domains encoded by an ORF, referred to as "*orfDup*" (Figure 2.1). For example, for surugamide BGC with four ORFs (shown in yellow in Figure 2.33.a), with "*orfDel*" option, NRPminer generates six NRP assembly lines formed by two ORFs (Figure 2.33.b), four assembly lines formed by three ORFs, and one canonical assembly line formed by all four ORFs. Using this approach, NRPminer discovered the non-canonical assembly lines for synthesizing surugamide and lugdunin NRP families in SoilActi and SkinStaph datasets, respectively (Figures 2.32 and 2.30).

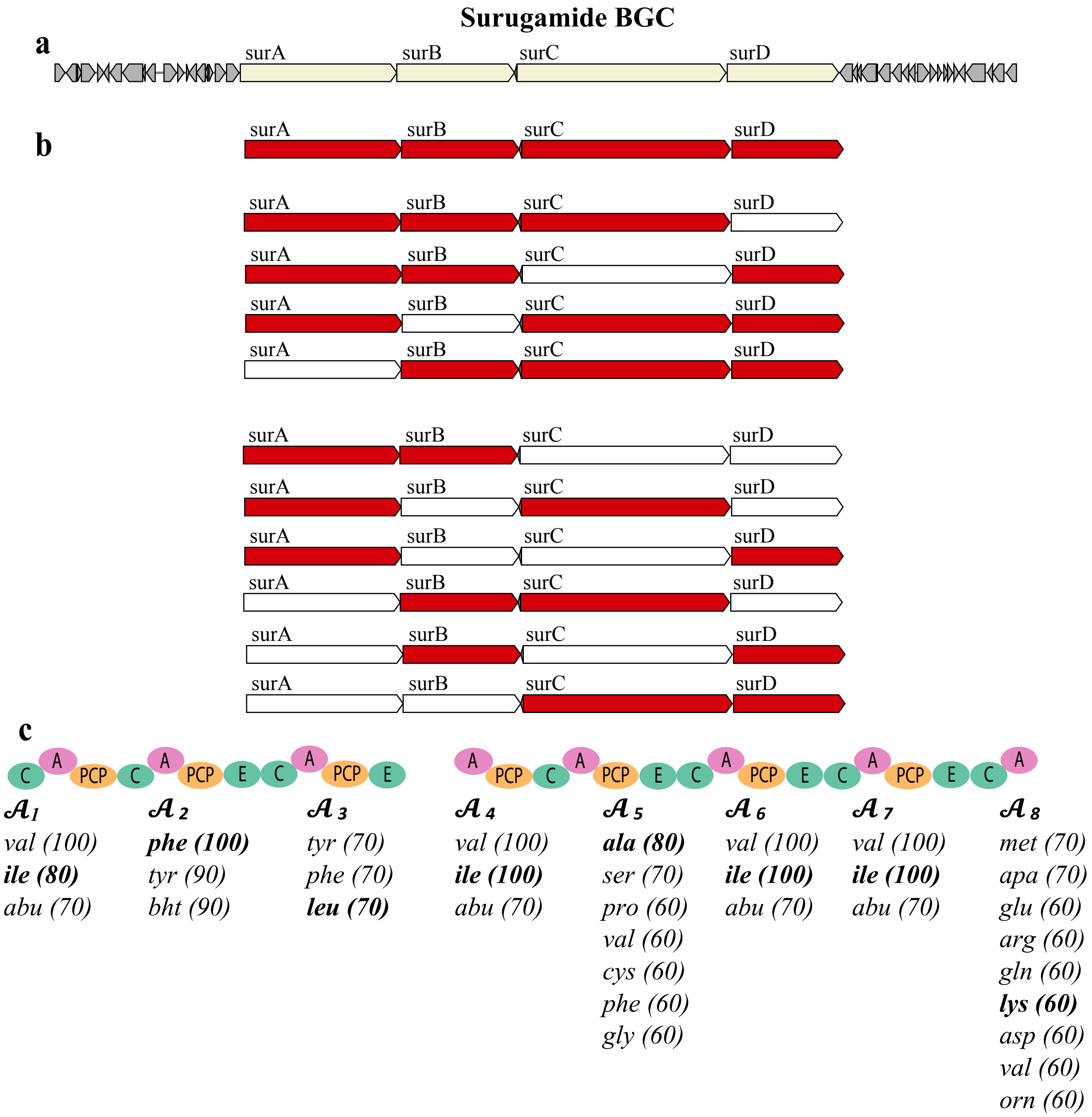


Figure 2.33. Surugamide BGC and the surugamide assembly line formed by the *SurA* and *SurD* genes. (a) Surugamide BGC with four ORFs shown in yellow. **(b)** 11 assembly lines formed by deletion of zero, one and two ORFs (shown in red). NRPminer in the *OrfDel* mode explores all assembly lines generated by removing up to two ORFs. **(c)** The NRPS assembly line that synthesizes cyclic surugamides (formed by the SurA and SurD genes). At least three highest-scoring amino acids (along with their NRPSpredictor2[13] scores) are shown below each A-domain in this assembly line. Amino acids appearing in surugamide A are shown in bold. NRPminer considers all amino acids with the same score as the score of the third highest-scoring amino acid as illustrated in the case of the fifth and the eighth A-domains.

Given a BGC, an *assembly line* refers to a sequence of NRPS modules in this BGC that together assemble the core NRP. NRPminer represents an assembly line as the sequence of A-domains appearing in its NRP modules and allows a user to explore various assembly lines using *OrfDel* and *OrfDup* options. Each portion of an NRPS that is encoded by a single ORF is an *NRPS subunit*. With *OrfDel* option, NRPminer considers skipping up to two entire NRPS subunits. Figure 2.33.b illustrates the assembly lines generated from surugamide BGC by deleting A-domains appearing on zero, one, and two NRPS subunits, out of the four NRPS subunits encoded by the four ORFs appearing in this BGC. We represent an NRPS assembly line as a sequence of sets of amino acids, $\mathcal{A}_1, \ldots, \mathcal{A}_k$ where each $\mathcal{A}_i$ represents the set of amino acids predicted for the *i*-th A-domain of this assembly line along with their specificity scores. Figure 2.33.c illustrates that for surugamide NRPS assembly line formed by SurA and SurD genes, $\mathcal{A}_1$ = {*val*, *ile*, *abu*}, $\mathcal{A}_2$ = {*phe, tyr, bht*}, etc.

Given an NRPS assembly line with *k* A-domains and the corresponding sets $\mathcal{A}_1, \ldots, \mathcal{A}_k$, the set of all possible core NRPs for this assembly line is given by the cartesian product $\mathcal{A}_1 \times \ldots \times \mathcal{A}_k$. In case of the Surugamide BGC (Figure 2.33.a), there are 45,927 possible core NRPs for the assembly line formed by SurA and SurD genes (2.33.c) and a total of 3,927,949,830 assembly lines for all 11 possible assembly-line for the surugamide BGC Figure 2.33.d.

In the default "*orfDel*" setting, NRPminer considers all assembly lines formed by deleting up to two ORFs. With "*orfDup*" option, NRPminer generates non-canonical assembly lines that tandemly duplicate all A-domains appearing in a single ORF. For example, Figure 2.34 describe that using this mode for lugdunin BGC with four ORFs, NRPminer generates one canonical assembly line formed by all four ORFs appearing once, four assembly lines where one of the ORFs appears two times, and four assembly lines where one of the ORFs appears three times. NRPminer considers all assembly lines made up of at least three and at most 20 NRPS modules.

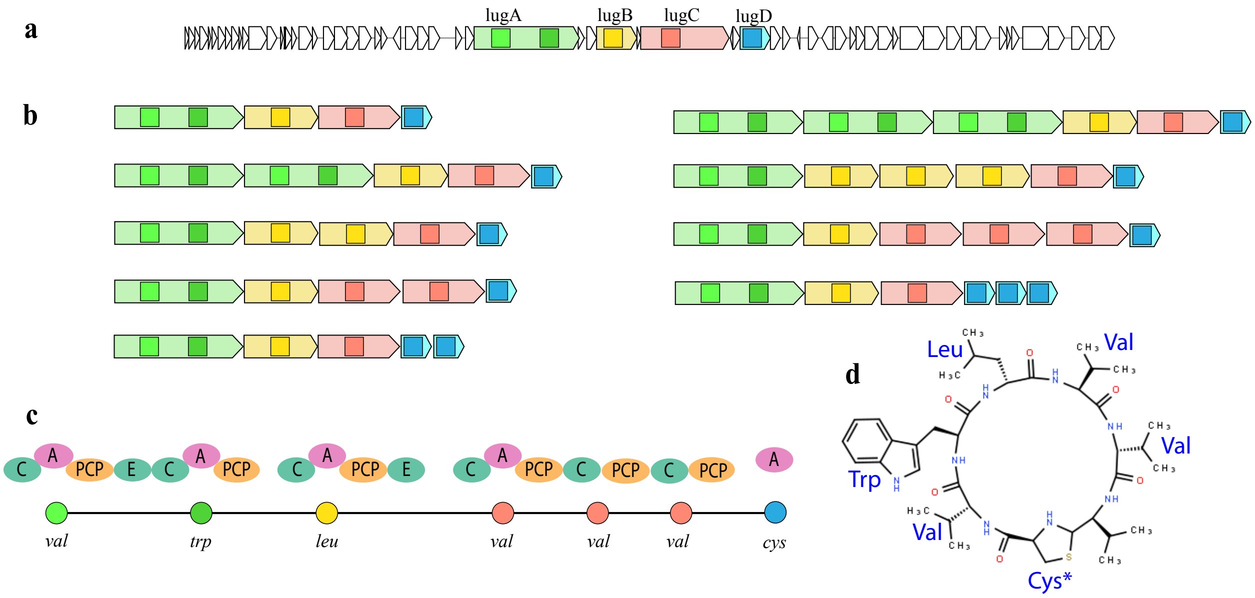


Figure 2.34. Lugdunin BGC and the assembly lines formed by NRPminer using the *OrfDup* option. (a) Lugdunin BGC with the four ORFs shown in different colors. The squares represent the A-domains. (b) Assembly lines formed by duplication of a single NRPS subunit (corresponding to each ORF) zero, one and two times are pictured. NRPminer explores all assembly lines generated by duplicating each ORF up to two times when the "*OrfDup*" option is selected. (c) The NRPS assembly lined (with A-, C-, PCP-, and E-domains pictured) appearing in the NRPS that synthesizes lugdunin, where one *val*-specific A-domain loads three amino acids (*valines*) to the growing peptide. Amino acids corresponding to lugdunin structure are shown below each A-domain. Circles represent amino acids (different amino acids are shown by different colors). (d) Cyclic structure of lugdunin with the amino acids highlighted in blue. The "Cys*" represent Cys-derived thiazolidine in lugdunin structure.

(*d*) Filtering the core NRPs based on their specificity scores. Table 2.2 and Table 2.7 illustrate that some BGC-rich genomes give rise to trillions of putative core NRPs. NRPminer uses the specificity scores of amino acids in each core NRP to select a smaller set of core NRPs for downstream analyses. Given an assembly line $\mathcal{A}_1, \ldots, \mathcal{A}_k$, for each $a \in \mathcal{A}_i$ $(i=1,\ldots,k)$, NRPminer first divides the specificity score of a by the maximum specificity score observed across all amino acids in $\mathcal{A}_i$; we refer to the integer value of the percentage of this number as the *normalized specificity score* of a. Table 2.11 show the normalized specificity scores of the amino acids predicted for the assembly line of cyclic surugamides (corresponding to SurA and SurD genes). We define the score of a core NRP to be the sum of the normalized scores of its amino acids.

NRPminer algorithm for filtering core NRPs.

NRPminer uses a dynamic programming algorithm to efficiently find *N* highest-scoring core NRPs for further analyses (the default value is *N*=1000), which enables peptidogenomics analysis of BGCs with many A-domains. presents the number of core NRPs generated from the assembly line formed by SurA and SurD genes, based on their scores. In total, 14,345 core NRPs from the original 3,927,949,830 core NRPs of the 11 assembly lines of surugamide BGC (listed in Figure 2.33.b) are retained.

Given an NRPS assembly line $A=A_1,...,A_n$, where A_i is the set of amino acids predicted for the *i*th A-domain of *A,* for every $a \in A_i$ (*i*=1,...,*n*), let $SpecificityScore_{A_i}(a)$ be the *specificity score* of *a* for the *i*th A-domain of *A* as described in Supplementary Note 3. Then, for each integer $1 \leq i \leq n$ and $a \in A_i$, we define *normalized specificity score* of *a* for *i*th A-domain of *A*, denoted by $S_A(i,a)$**,** to be the nearest integer to the following value:

$$\frac{SpecificityScore_{A_i}(a)}{\max_{b \in A_i} SpecificityScore_{A_i}(b)} \times 100.$$

We use this scoring function (instead of *SpecificityScore*) to reduce the bias towards the more frequently observed A-domains that usually result in higher specificity scores compared to the less commonly observed ones, which do not have closely related A-domains in NRPSpredictor2 training datasets[13]. Consider the assembly line of cyclic surugamides A-D shown in Figure 2.33.c (corresponding to SurA-SurD gene pairs in surugamide BGC) which is made up of eight A-domains, we refer to this assembly line by $SurugamideAL$. Table 2.11 presents the values of $S_{SurugamideAL}$ for integers $1 \leq i \leq 8$ and (at least) the three amino acids with the highest normalized specificity scores for each A-domain in this assembly line.

Table 2.11. Predicted amino acids for the eight A-domains appearing on cyclic surugamides A-D assembly line *SurugamideAL*. A_i represents the set of amino acids predicted for the *i*th A-domain in $SurugamideAL$. For each A_i at least three amino acids with the highest normalized specificity scores (listed in parentheses) are presented. Amino acids appearing in surugamide A (IFLIAIIK) are shown in bold. NRPminer considers all amino acids with the same normalized specificity score, as illustrated in the case of the fifth and the eighth A-domains.

A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8
val (100) ***ile (80)*** *abu (70)*	***phe (100)*** *tyr (90)* *bht (90)*	*tyr (100)* *phe (100)* ***leu (100)***	*val (100)* ***ile (100)*** *abu (70)*	***ala (100)*** *ser (87)* *pro (75)* *val (75)* *cys (75)* *phe (75)* *gly (75)*	*val (100)* ***ile (100)*** *abu (70)*	*val (100)* ***ile (100)*** *abu (70)*	*met (100)* *apa (100)* *glu (86)* *arg (86)* *gln (86)* ***lys (86)*** *asp (86)* *val (86)* *orn (86)*

Given $A = A_1,...,A_n$ we call the set of all core NRPs generated by the cartesian product $A_1 \times ... \times A_n$ as *the core NRPs of A*. For each core NRP of *A*, $a_1a_2...a_n$, we define the *adenylation score* of $a_1a_2...a_n$, denoted by $Score_A(a_1a_2...a_n)$, to be the sum of the normalized specificity scores of all of its amino acids:

$$Score_A(a_1a_2...a_n) = \sum_{i=1}^{n} S_A(i,a_i).$$

Therefore, given assembly line *SurugamideAL* and core NRP, *P*=IAIIKIFL (the core NRP corresponding to surugamide A), $Score_{SurugamideAL}(P)$=80+100+100+100+100+100+100 +86=766. Note that, for any assembly line *A*, the maximum value of $Score_A$ denoted by $maxScore_A = \sum_{i=1}^{n} max_{a_i \in A_i} \boldsymbol{S_A}(i, a_i) = 100n$.

For many organisms, the total number of possible core NRPs is prohibitively large, making it infeasible to conduct search against massive spectral repositories. Currently, even the fastest state-of-the-art spectral search methods are slow for searching millions of input spectra against databases with over 10^5 peptides in a modification-tolerant manner as the runtime grows exceedingly large when the database size grows[41]. Tables 2.7 and 2.12 show that for 24 (22) out for 27 organisms in XPF dataset and 9 (7) out of 20 organisms in *SoilActi* dataset, the total number of core NRPs exceed 10^5 (10^6). Therefore, to enable scalable peptidogenomics for NRP discovery, for each constructed assembly line NRPminer, selects a set of candidate core NRPs. To do so, NRPminer starts by finding the number of core NRPs of A according to their adenylation scores (Problem 1) and then it uses these numbers for generating all core NRPs of A with adenylation scores higher than a threshold (Problem 2).

Problem 1. Given $A=A_1,\ldots,A_n$ and a positive integer s, find the of number of all core NRPs of A with adenylation score equal to s.

Let $k = \max_{i\in\{1,\ldots,n\}}(|A_i|)$ where $|A_i|$ shows the number of amino acids in A_i. For any positive integers i and s satisfying,$1 \leq i \leq n$ and $s \leq maxScore_A$, let $numCoreNRPs_A(i,s)$ denote the number of core NRPs, of assembly line $A_1,\ldots,A_i$ with $Score_{A_1,\ldots,A_i}$ equal to s. Let $numCoreNRPs_A(0,s) = 0$ for any positive integer s, and $numCoreNRPs_A(i,s) = 0$ for any integer $s < 0$, across all possible values of i. Then, for any positive integers i and s satisfying $1 \leq i \leq n$ and $0<s\leq maxScore_A$, we have,

$$numCoreNRPs_A(i,s)= \sum_{a_i\in A_i} numCoreNRPs_A(i-1, s-S_A(i,a_i)). \quad (1)$$

Using the recursive formula (1), NRPminer calculates $numCoreNRPs_A$ using parametric dynamic programming in a bottom-up manner: NRPminer first, computes $numCoreNRPs_A(1,s)$, for all positive integers $s\leq maxScore_A$. then proceeds to $numCoreNRPs_A(2,s)$ for all such s, and so on, computing $numCoreNRPs_A(n,s)$ for all such $0<s$. Using this approach, for each value of i

and s, NRPminer computes $numCoreNRPs_A(i,s)$ by summing over at most k values. Therefore, NRPminer calculates all values of $numCoreNRPs_A$ with time complexity $O(k \times n \times maxScore_A)$.

Given a positive integer $N<10^5$, let $score_{A,N}$ be the greatest integer $s' \leq maxScore_A$ such that,

$$N \leq \sum\nolimits_{s' \leq s \leq maxScore} numCoreNRPs_A(n,s) \text{ .}$$

Then, we define,

$$thresholdScore_A(N) = \begin{cases} score_N & \text{if } score_N < score_{10^5} \\ score_N - 1 & \text{if } score_N = score_{10^5} \end{cases} . \quad (2)$$

NRPminer selects, $candidateCoreNRPs_A(N)$, defined as the set of all core NRPs of A, with adenylation score at least $thresholdScore_A(N)$. NRPminer selects core NRPs $candidateCoreNRPs_A(N)$ for downstream spectral analyses. Using this approach, NRPminer is guaranteed to be scalable as at most 10^5 candidate core NRPS are explored per assembly line.

Table 2.12 Table 2.12 presents the values of $numCoreNRPs_{SurugamideAL}(8,s)$ for various values of s. Note that, this table presents the number of core NRP only for a single assembly line, *SurugamideAL,* corresponding to cyclic surugamides (surugamide A-D). In total, 14,345 core NRPs were retained from the original 3,927,949,830 core NRPs of the 11 assembly lines of surugamide's BGC.

Table 2.12. Number of core NRPs of *SurugamideAL* (assembly line corresponding to cyclic surugamides A-D) according to their adenylation scores. Only values of s with non-zero number of cores and corresponding to the top 1000 high-scoring core NRPs are shown.

$\boldsymbol{s}$	800	790	788	786	780	778	776	774	772	**total**
$\boldsymbol{numCoreNRPs_{SurugamideAL}(8,s)}$	24	48	24	192	24	48	384	192	168	1104

Problem 2. Given an assembly line A and a positive integer N, generate $candidateCoreNRPs_A(N)$, defined as all core NRPs of A with adenylation scores at least $thresholdScore_A(N)$.

NRPminer follows a graph-theoretic approach to quickly generate $candidateCoreNRPs_A(N)$ by using the computed values of *numCoreNRPs*. Let *G(A)* be the acyclic directed graph with nodes corresponding to pairs of positive integers $i \leq n$ and $s \leq maxScore_A$, such that $numCoreNRPs_A(i,s) > 0$, denoted by $v_{i,s}$. For every node $v_{i,s}$ (i=1,...,n) and every $a \in A_i$ such that $numCoreNRPs_A(i-1, s-S_A(i,a)) > 0,$ there exists a directed edge from $v_{i-1,s-S_A(i,a)}$ to $v_{i,s}$. Let *Source* be $v_{0,0}$ and let *Sink* be the set of all nodes $v_{n,s}$ such that $thresholdScore_A(N) \leq s$. We call each directed path in *G(A)* from Source to the nodes in Sink as a *candidate path* of *G(A)*.

Each candidate path of *G(A),* corresponds to a distinct core NRP of *A* with adenylation score at least $thresholdScore_A(N)$ and vice versa. Therefore, the problem of finding all core NRPs of *A* with adenylation score at least $thresholdScore_A(N)$, corresponds to the problem of finding all candidate paths of *G*(*A*). While enumerating all paths with *n* nodes in a directed acyclic graph can grow exponentially large (as there can be exponentially many such paths), but due to our choice of $thresholdScore_A(N)$, the number of candidate paths of *G(A)* is bound by 10^5 (or *N* if $score_N = score_{10^5}$). NRPminer uses the default value *N*=1000. Moreover, $n \leq 20$, (only assembly lines made up of up to 20 A-domains are considered) and $k \leq 12$.

(e) Identifying domains corresponding to known modifications and incorporating them in the core NRPs. NRPminer searches each BGC for methylation domains (PF08242) and accounts for the possible methylations on corresponding residues for all resulting core NRPs (corresponding to +14.01Da mass shift). NRPminer also searches each BGC for epimerization domains (as well as dual condensation-epimerization domains) that provide information about the structure of the final NRP (D- or L-amino acids).

(f) Generating linear, cyclic, and branch-cyclic backbone structures for each core NRP. NRPminer generates linear and cyclic structures for all core NRPs. Similar to NRPquest[28], whenever NRPminer finds a cytochrome P450 domain, it also generates branched-cyclic NRPs by considering a side-chain bond between any pair of residues in the peptide.

(g) Modification-tolerant search of spectra against the constructed backbone structures. Similar to PSMs in proteomics, a PSM in peptidogenomics is scored based on similarities between the theoretical spectrum of the peptide and the mass spectrum[41] (See Supplementary Note 8). The *standard search* of a spectrum against a peptide database refers to finding a peptide in the database that forms a highest-scoring PSM with this spectrum. Similarly, the *modification-tolerant search* of a spectrum against the peptide database refers to finding a variant of a peptide in the database that forms a highest-scoring PSM with this spectrum. In the case of NRPs, it is crucial to conduct modification-tolerant search in a blind mode in order to account for unanticipated PAMs in the mature NRP. For example, NRPminer identified PAX-peptides family and their corresponding BGC in X. *nematophila* ATCC 19061 in the XPF dataset even though these NRPs include lipid side-chains that are not predictable via genome mining. As another example, NRPminer identified lugdunin in the SkinStaph dataset that contains an unusual Cys-derived thiazolidine modification[7].

Existing peptidogenomics methods utilize a brute-force approach for modification-tolerant search, by creating a database of all possible unanticipated modification[28]. For example, given a spectrum and a core NRP structure with *n* amino acids, these methods consider a modification of mass δ on all possible amino acids in the NRP, where δ is the mass difference between the spectrum and the NRP. Gurevich *et al.* developed the VarQuest[15] tool for modification-tolerant search of large spectral datasets against databases of peptidic natural products that is two orders of magnitude faster than the brute-force approach. NRPminer utilizes VarQuest for identification of PAMs with masses up to *MaxMass* with the default value *MaxMass*=150 Da (See below for further information). This approach also allows NRPminer to identify loss or addition of an amino acid (for amino acids with molecular mass up to *MaxMass* Da). Note that, similar to identification of post-assembly modifications in linear proteomics[28], MS-based methods for NRP discovery are limited to finding modification masses and cannot provide information about the exact chemistry of the identified modifications.

Forming Peptide-Spectrum-Matches (PSMs) and Calculating PSM Scores. Peptide-Spectrum-Matches (PSMs) and their PSM scores are described by Gurevich *et al.*[41]. Given a peptide *P* (with any backbone structure), the *graph of P* is defined as a graph with nodes corresponding to amino acids in *P* and edges corresponding to *generalized peptide bonds* as described in Mohimani *et al.*[92]. The mass of this graph (referred to as Mass(*P*)) is defined as the total mass of its amino acids and *TheoreticalSpectrum*(*P*) is defined as the set of masses (theoretical peaks) of all connected components of the graph of *P* resulting from removal of two-cuts (e.g. a pair of edges in the ring potion of a cyclic and branch-cyclic PNPs) or a bridge (e.g. a bond in a linear peptide, or branch of a branch-cyclic peptide)[92]. Note that each such removal results in two peaks (fragments) with a total mass equal to Mass(*P*).

Given a peptide *P* and a spectrum *S*, the *PSM score* of the PSM formed between *P* and *S* (or the shared peak count score), denoted by *SPCScore*(*P*,*S*), is defined as the number of peaks shared between *TheoreticalSpectrum*(*P*) and *S*. Two peaks are shared if their masses are within a threshold ε (0.02 Da for high-resolution spectra). We compute the PSM score only if the precursor mass of the spectrum, denoted as Mass(S), matches Mass(P) with error up to Δ (0.02 Da for high-resolution data).

If (A_1, …, A_n) is the list of amino acid masses in a peptide P, we define *Variant*(P,i,δ) as (A_1,…, $A_\mathrm{i} + \delta$, …, A_n), where P and *Variant*(P,i,δ) have the same topology and $A_i + \delta \geq 0$. *VariableScore*(P,S) is defined as:

$$max(SPCScore(Variant(P,i,\omega),S)),$$

where ω is Mass(P) − Mass(S) and i varies from 1 to n (n stands for the number of amino acids in the peptide P)[41]. We define *a variant of peptide P derived from a spectrum S* as *Variant*(P,i,ω) of peptide P, that maximizes *SPCScore*(*Variant*(P,i,ω),S) across all positions i in P. For simplicity, we refer to this variant as *Variant(P,S)*. Given P and S, VarQuest[41] uses a heuristic approach to efficiently find *Variant(P,S).*

NRPminer uses *VarQuest*[41] to perform modification-tolerant search of the input spectral datasets against the constructed peptide structures generated from selected core NRPs (see the NRPminer step "generating linear, cyclic and branch-cyclic backbone structures for each core NRP " in Figure 2.2 and in Method section). Given a positive number *MaxMass* representing the maximum allowed modification mass (default value of *MaxMass*=150), for each constructed structure P and input spectrum S, if |*Mass(P)-Mass(S)*|≤*MaxMass,* NRPminer uses VarQuest[41] to find the *Variant(P, S)*. In this context, *Variant(P,S)* represents the mature NRP with a single post-assembly modification (PAM) on P that resulted in the mass difference |*Mass(S)-Mass(P)*|. Similar idea has been applied to identification of post-translational modifications in traditional proteomics[50,93].

NRPminer has the *one-vs-one* mode for searching a spectral dataset against the genome corresponding to its producer. Additionally, NRPminer features *one-vs-all* mode that a spectral dataset is searched against all genomes in the corresponding taxonomic clade. One-vs-all is useful in cases when an entire BGC is not assembled in a single contig in the producer's genome, but well-assembled in a related genome. For example, the spectra representing the three protegomycins produced by *Xenorhabdus* sp. 30TX1 did not match any core NRP generated from its genome because the corresponding BGC was not assembled in a single contig in this genome. However, they were identified with statistically significant p-values using the one-vs-all search when these spectra were searched against core NRPs from *X. doucetiae* genome (Figure 2.6) that included an orthologous BGC in a single contig.

In scoring PSMs, NRPminer has a user-adjustable threshold for the accuracy of precursor and products ions, thus improving the accuracy of PSM scoring in the case of modification-tolerant search of high-resolution spectral datasets. This feature improves on NRPquest whose applications are largely limited to low-resolution spectra.

(h) Computing statistical significance of PSMs. NRPminer uses MS-DPR[94] to compute p-values of the identified PSMs. Given a PSM, MS-DPR computes the probability (p-value) that a random peptide has a score greater than or equal to the PSM score. The default p-value threshold (10^{-15}) is chosen based on the previous studies where the p-value cut-off 10^{-15} was necessary for reaching a False Discovery Rate (FDR) below 1% against non-ribosomal peptides[15]. However, the user can change the p-value thresholds (using "--*pvalue*" handle) depending on their study.

NRPminer uses the MS-DPR[94] to compute the statistical significance (p-value) of each identified PSM. Given a PSM(*P,S*) between a NRP *P* and a spectrum *S*, MS-DPR[94] computes the probability (p-value) that a peptide with the same length as *P* forms a PSM with the spectrum *S* with a PSM score that is greater than or equal to the score of PSM(*P,S*). We refer to this probability, as the *p-value* of PSM(*P,S*).

A simple way to estimate the p-value of a PSMs is to use *Monte Carlo simulations* - that is to generate a population of billions of random peptides and estimate the distribution of PSM scores of all peptides against the spectrum *S*. However, this approach becomes prohibitively time-consuming for estimating very low p-values, i.e., when calculating the probabilities of extremely rare events. For example, estimating p-values as low as 10^{-12} requires calculating PSM scores of trillions of randomly generated peptides. Therefore, naïve Monte Carlo is impractical in mass spectrometry experiments where PSMs with p-values as low as 10^{-12} and below are common, including the case of NRP studies[95].

To overcome this challenge, MS-DPR[94] uses a method for evaluating probability of rare events (peptides yielding "high" PSM scores) called *multilevel splitting*. This method, that was originally developed in nuclear physics, rapidly approximates an extreme tail of the probability distribution of PSM scores against a spectrum. It constructs a Markov Chain over a space of PSM scores of millions of random peptides similar to *P* (in molecular weight and length) against *S*. It then uses selection mechanisms that favors the trajectories in this Markov chain deemed likely to lead to high-scoring PSMs. Using this method, MS-DPR[94] dedicates a greater fraction of the computational effort to a portion of the peptide space that leads to higher PSM scores against *S,* and therefore can efficiently estimate the total probability of all peptides with high scores in the constructed Markov chain.

(i) Expanding the set of identified NRPs using spectral networks. Spectral datasets often contain multiple spectra originating from the same compound. NRPminer clusters similar spectra using MS-Cluster[65] and estimates the number of distinct NRPs as the number of clusters. It further constructs the spectral network[58,51] of all identified spectra and estimates the number of distinct NRP families as the number of connected components in this network.

Spectral networks reveal the spectra of related peptides without knowing their amino acid sequences[58]. Nodes in a spectral network correspond to spectra, while edges connect *spectral pairs*, i.e. spectra of peptides differing by a single modification or a mutation. Ideally, each connected component of a spectral network corresponds to a single NRP family[58] representing a set of similar NRPs. In this study, we only report an identified NRP family if at least one NRP in the family is identified with a PSM p-value at least 10^{-20}. NRPminer utilizes spectral networks for expanding the set of identified NRPs.

Availability. NRPminer is available as both a stand-alone tool (https://github.com/bbehsaz/nrpminer) and as a web application via GNPS in silico toolbox. All described datasets are available through the corresponding public repositories.

2.6. ACKNOWLEDGEMENTS

We thank Dr Rob Knight for the fruitful discussion and the help during this project. We thank Louis Felix Nathan for fruitful discussions.

Chapter 2, in full, has been submitted for journal publication by: Bahar Behsaz, Edna Bode, Alexey Gurevich, Yan-Ni Shi, Florian Grundmann, Andrés Mauricio Caraballo-Rodríguez, Amina Bouslimani, Morgan Panitchpakdi, Annabell Linck, Changhui Guan, Julia Oh, Pieter C. Dorrestein, Helge B. Bode, Pavel A. Pevzner, Hosein Mohimani. The dissertation author was the primary author of this work.

Contributions. B.B. designed and implemented NRPminer algorithm and performed the benchmarking included for all datasets in this study. E.B. performed the NMR and bio-activity analyses (described in Appendix 1) for final structure elucidation and characterization of protegomycin NRP family in XPF dataset. Y.S. and F.G. and H.B.B. performed the NMR and bio-activity analyses (described in Appendix 2) for final structure elucidation and characterization of novel xenoamicin-like NRP family in XPF dataset and commented on the results in XPF dataset. J.O., C.G., M.P., and A.B. contributed the SkinStaph sample set for analysi and data generation. A.M.C. commented on the results in SoilActi dataset. H.B.B., P.A.P. and H.M. directed the work.

Funding. The work of B.B. and P.A.P. was supported by the 2018 Grand Challenge Award from the UC San Diego Center for Microbiome Innovation. The work of H.M. was supported by a research fellowship from the Alfred P. Sloan Foundation, a National Institutes of Health New Innovator Award DP2GM137413, and a U.S. Department of Energy award DE-SC0021340. The work in the Bode lab was supported by the LOEWE-Center Translational Biodiversity Genomics (TBG) and the LOEWE Schwerpunkt MegaSyn, both funded by the State of Hesse and an ERC Advanced Grant (SYNPEP, grant agreement number 835108). The work of A.G. was supported by St. Petersburg State University, St. Petersburg, Russia (grant ID PURE 51555639). J.O. was supported by the NIH/NIGMS (1 DP2 GM126893-01) and the Department of Defense (W81XWH1810229). A.M.C.R. and P.C.D. were supported by the National Sciences Foundation grant IOS-1656481 and National Institutes of Health (NIH) subaward 1090607-425868.

Pavel Pevzner is a co-founder, has an equity interest and receives income from Digital Proteomics, LLC. The terms of this arrangement have been reviewed and approved by the University of California, San Diego in accordance with its conflict of interest policies.

2.7. APPENDECIES

Appendix 1: Additional Analyses for Novel Protegomycin Family

Description of Experiment. *X. doucetiae*-Δ*hfq* was constructed as described before[66]. Exchange of the natural promoter against the inducible P*BAD* was performed as described[96]. Briefly, the first 598 base pairs of *prtA* were amplified with primer pEB_317-fw TTTGGGCTAACAGGAGGCTAGCAT_ATGAGAATACCTGAAGGTTCG and PEB_318-rv TCTGCAGAGCTCGAGCATGCACAT_CGTAATGAAACGAGTTCAGG. The resulting fragment was cloned via hot fusion cloning into pCEP-km via homologous arms. The resulting construct pCEP_XdV3_70082-km was transformed into *E. coli* S17-1 λpir resulting in *E. coli* pCEP_XdV3_70082. Conjugation of this strain with *X. doucetiae* wt or *X. doucetiae*-Δ*hfq* was followed by integration of pCEP_XdV3_70082 into the acceptors genome via homologous recombination (zit). In *X. doucetiae*-Δ*hfq*-P*BAD*-*prtA* the production of protegomycin was induced by adding 0.2 % L-arabinose into the fresh inoculated medium[66].

For large scale production of protegomycin, 6 x 1 L LB medium was inoculated with *X. doucetiae*-Δ*hfq*_P*BAD*-*prtA* pre-culture 0.02 %. 2 % Amberlite® XAD-16 adsorber resin was added and the production was induced with 0.2 % L-arabinose. The cultures were constantly shaking at 130 rpm at 30 °C. After 72 h the XAD beads were harvested and protegomycins extracted using 3 L of methanol. The solvent was evaporated, and the crude extract was then used for isolation and analyzation of protegomycin derivatives. Part of crude extraction was purified by preparative HPLC with a gradient mobile from 5% to 95% ACN in H_2O (v/v) in 30 mins followed by semi-preparative HPLC (ACN–H_2O, 35-45% in 30 mins, v/v) to yield PRT-1037 (24.4 mg).

For structure elucidation and determination of incorporated C- and N- atoms and amino acids into protegomycins, cultivation of *X. doucetiae*-Δ*hfq*_P*BAD*-*prtA* and *X. doucetiae*_ P*BAD*-

prtA, induced with 0.2 % L-arabinose was performed in 5 mL LB (^{12}C), ^{13}C- and ^{15}N-isogrow® medium (Sigma Aldrich). The cultures were supplemented with 2 % Amberlite® XAD-16 adsorber resin. To analyze the incorporated amino acids, induced mutants were grown in LB medium supplemented with selected ^{13}C-labeled amino acids with a concentration of 2 mM. After 48 h cultivation at 30 °C, constantly shaking at 200 rpm, Amberlite® XAD-16 beads were harvested and extracted with 5 mL MeOH for 45 min. Samples were taken from the filtered extracts and centrifuged for 15 min at 13.000 rpm for further HPLC-MS analysis (Dionex Ultimate 3000 coupled to a Bruker AmaZon X ion trap). Generated HPLC-MS data were interpreted as described previously[66,97].

Appendix 2: Additional Analyses for Novel Xenoamicin-like Family

Cultivation of strains. *Xenorhabdus* KJ12.1 was routinely cultivated in Luria-Bertani (LB) medium (pH 7.0) at 30°C and 200 rpm on rotary shaker and on LB agar plates at 30°C. Inverse feeding experiments were applied in either ISOGRO® ^{13}C medium, ISOGRO® ^{15}N medium. 50 ml ISOGRO® medium was prepared with ISOGRO® powder (0.5 g), K_2HPO_4 (1.8 g/l), KH_2PO_4 (1.4 g/l), $MgSO_4$ (1 g/l) and $CaCl_2$ (0.01 g/l) solved in water. Feeding experiments in ISOGRO® ^{13}C medium supplemented with ^{12}C amino acids was inoculated with ISOGRO® washed overnight cultures.

Production cultures were grown in LB media containing 2% Amberlite® XAD-16 resin inoculated with 1% overnight culture. Promotor exchange mutants were induced with 0.2% arabinose at the beginning of the cultivation. Resin beads and bacterial cells were harvested by centrifugation after 72 h cultivation time, washed twice with one culture volume methanol. The crude extracts were analysed by means of MALDI-MS and HPLC-MS (Bruker AmaZon).

HPLC based purification. XAM-1320 was isolated by a two-step chromatography. Strain KJ12.1 was cultivated in a BIOSTAT A plus fermenter (Sartorius) equipped with a 2-L vessel in 1.5 L of LB broth at 30 °C for 12 hours. For the inoculation, 1% overnight preculture was used and 2% XAD-16 were added. Additionally, 10 g of glucose and 5 mL Antifoam 204 (Sigma-Aldrich) were added. The fermentation was performed with an aeration of 2.25 vvm, constant stirring at 300 rpm and at pH 7, stabilized by the addition of 0.1 N phosphoric acid or 0.1 N sodium hydroxide. The XAD resin was washed with methanol to get the extract after evaporation. Xenoamicin III A was isolated by a two-step chromatography. In the frist step the extract was fractionated with a 5-95% water/acetonitrile gradient over 15 min on a Luna C_{18} 10 □m 50x50 mm column (Phenomenex). In the second step XAM-1320 was isolated with a 40-60% water-acetonitrile gradient over 19 min on Luna C_{18} 5 □m 30x75 mm column (Phenomenex).

MS analysis. MS analysis was carried out by using an Ultimate 3000 LC system (Dionex) coupled to an AmaZon X electronspray ionization mass spectrometer (Bruker Daltonics). Separation was done on a C18 column (ACQITY UPLC BEH, 1.7 mm, 2.1x50 mm, flow rate 0.4 ml/min, Waters). Acetonitrile/water containing 0.1% formic acid was used as mobile phase. The gradient started with 5% acetonitrile continuous over 2 minutes. Over 0.5 minutes under a linear gradient acetonitrile reaches 40%. Following an equilibration phase over 1.5 minutes with 40% acetonitrile takes place. For separation a linear gradient from 40-95% acetonitrile over 10.5 minutes were used. The gradient ends up with 95% acetonitrile continuous over 1.5 minutes. Collision-induced dissociation (CID) was performed on ion trap in the AmaZon X in positive mode.

HR-ESI-HPLC-MS data were obtained with on a LC-coupled Impact II ESI-TOF spectrometer (Bruker Daltonics).

Advanced Marfey's method. The advanced Marfey's method to determine the configurations of the amino acid residues was performed as described previously[69].

2.8. REREFERENCES

1. Newman, D. J. & Cragg, G. M. Natural Products as Sources of New Drugs from 1981 to 2014. *Journal of Natural Products* **79,** 629–661 (2016).

2. Li, J. W. H. & Vederas, J. C. Drug discovery and natural products: End of an era or an endless frontier? *Science* **325,** 161–165 (2009).

3. Ling, L. L., Schneider, T., Peoples, A. J., Spoering, A. L., Engels, I., Conlon, B. P., Mueller, A., Hughes, D. E., Epstein, S., Jones, M., Lazarides, L., Steadman, V. a, Cohen, D. R., Felix, C. R., Fetterman, K. A., Millett, W. P., Nitti, A. G., Zullo, A. M., Chen, C. & Lewis, K. A new antibiotic kills pathogens without detectable resistance. *Nature* **517,** 455–459 (2015).

4. Harvey, A. L., Edrada-Ebel, R. & Quinn, R. J. The re-emergence of natural products for drug discovery in the genomics era. *Nature Reviews Drug Discovery* **14,** 111–129 (2015).

5. Wang, H., Fewer, D. P., Holm, L., Rouhiainen, L. & Sivonen, K. Atlas of nonribosomal peptide and polyketide biosynthetic pathways reveals common occurrence of nonmodular enzymes. *Proceedings of the National Academy of Sciences of the United States of America* **111,** 9259–9264 (2014).

6. Donia, M. S., Cimermancic, P., Schulze, C. J., Wieland Brown, L. C., Martin, J., Mitreva, M., Clardy, J., Linington, R. G. & Fischbach, M. A. A Systematic Analysis of Biosynthetic Gene Clusters in the Human Microbiome Reveals a Common Family of Antibiotics. *Cell* **158,** 1402–1414 (2014).

7. Zipperer, A., Konnerth, M. C., Laux, C., Berscheid, A., Janek, D., Weidenmaier, C., Burian, M., Schilling, N. A., Slavetinsky, C., Marschal, M., Willmann, M., Kalbacher, H., Schittek, B., Brötz-Oesterhelt, H., Grond, S., Peschel, A. & Krismer, B. Human commensals producing a novel antibiotic impair pathogen colonization. *Nature* **535,** 511–516 (2016).

8. Wilson, M. R., Jiang, Y., Villalta, P. W., Stornetta, A., Boudreau, P. D., Carrá, A., Brennan, C. A., Chun, E., Ngo, L., Samson, L. D., Engelward, B. P., Garrett, W. S., Balbo, S. & Balskus, E. P. The human gut bacterial genotoxin colibactin alkylates DNA. *Science* **363,** eaar7785 (2019).

9. Vizcaino, M. I. & Crawford, J. M. The colibactin warhead crosslinks DNA. *Nature Chemistry* **7,** 411–417 (2015).

10. Marahiel, M. A., Stachelhaus, T. & Mootz, H. D. Modular Peptide Synthetases Involved in Nonribosomal Peptide Synthesis. *Chemical Reviews* **97,** 2651–2674 (1997).

11. Süssmuth, R. D. & Mainz, A. Nonribosomal Peptide Synthesis—Principles and Prospects. *Angewandte Chemie - International Edition* **56,** 3770–3821 (2017).

12. Renier, A., Vivien, E., Cociancich, S., Letourmy, P., Perrier, X., Rott, P. C. & Royer, M. Substrate specificity-conferring regions of the nonribosomal peptide synthetase adenylation domains involved in albicidin pathotoxin biosynthesis are highly conserved within the species Xanthomonas albilineans. *Applied and Environmental Microbiology* **73,** 5523–5530 (2007).

13. Röttig, M., Medema, M. H., Blin, K., Weber, T., Rausch, C. & Kohlbacher, O. NRPSpredictor2 - A web server for predicting NRPS adenylation domain specificity. *Nucleic Acids Research* **39,** 362–367 (2011).

14. Medema, M. H., Blin, K., Cimermancic, P., de Jager, V., Zakrzewski, P., Fischbach, M. A., Weber, T., Takano, E. & Breitling, R. antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic acids research* **39,** 339–346 (2011).

15. Blin, K., Shaw, S., Steinke, K., Villebro, R., Ziemert, N., Lee, S. Y., Medema, M. H. & Weber, T. antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic acids research* **47,** 81–87 (2019).

16. Chevrette, M. G., Aicheler, F., Kohlbacher, O., Currie, C. R. & Medema, M. H. SANDPUMA: Ensemble predictions of nonribosomal peptide chemistry reveal biosynthetic diversity across Actinobacteria. *Bioinformatics* **33,** 3202–3210 (2017).

17. Mori, T., Cahn, J. K. B., Wilson, M. C., Meoded, R. A., Wiebach, V., Martinez, A. F. C., Helfrich, E. J. N., Albersmeier, A., Wibberg, D., Dätwyler, S., Keren, R., Lavy, A., Rückert, C., Ilan, M., Kalinowski, J., Matsunaga, S., Takeyama, H. & Piel, J. Single-bacterial genomics validates rich and varied specialized metabolism of uncultivated Entotheonella sponge symbionts. *Proceedings of the National Academy of Sciences U.S.A.* **33,** 3202–3210 (2018).

18. Hover, B. M., Kim, S. H., Katz, M., Charlop-Powers, Z., Owen, J. G., Ternei, M. A., Maniko, J., Estrela, A. B., Molina, H., Park, S., Perlin, D. S. & Brady, S. F. Culture-independent discovery of the malacidins as calcium-dependent antibiotics with activity against multidrug-resistant Gram-positive pathogens. *Nature Microbiology* **3,** 415–422 (2018).

19. Parkinson, E. I., Tryon, J. H., Goering, A. W., Ju, K.-S., McClure, R. A., Kemball, J. D., Zhukovsky, S., Labeda, D. P., Thomson, R. J., Kelleher, N. L. & Metcalf, W. W. Discovery of the tyrobetaine natural products and their biosynthetic gene cluster via metabologenomics. *ACS Chemical Biology* **13,** 1029–1037 (2018).

20. Khaldi, N., Seifuddin, F. T., Turner, G., Haft, D., Nierman, W. C., Wolfe, K. H. & Fedorova, N. D. SMURF: Genomic mapping of fungal secondary metabolite clusters. *Fungal Genetics and Biology* **47,** 736–741 (2010).

21. Palaniappan, K., Chen, I.-M. A., Chu, K., Ratner, A., Seshadri, R., Kyrpides, N. C., Ivanova, N. N. & Mouncey, N. J. IMG-ABC v. 5.0: an update to the IMG/Atlas of

Biosynthetic Gene Clusters Knowledgebase. *Nucleic acids research* **48,** D422–D430 (2020).

22. Kautsar, S. A., Blin, K., Shaw, S., Navarro-Muñoz, J. C., Terlouw, B. R., van der Hooft, J. J. J., van Santen, J. A., Tracanna, V., Suarez Duran, H. G., Pascal Andreu, V., Selem-Mojica, N., Alanjary, M., Robinson, S. L., Lund, G., Epstein, S. C., Sisto, A. C., Charkoudian, L. K., Collemare, J., Linington, R. G., Weber, T. & Medema, M. H. MIBiG 2.0: a repository for biosynthetic gene clusters of known function. *Nucleic acids research* **48,** D454–D458 (2020).

23. Medema, M. H. Computational Genomics of Specialized Metabolism: from Natural Product Discovery to Microbiome Ecology. *mSystems* **3,** e000182 (2018).

24. Johnston, C. W., Skinnider, M. A., Dejong, C. A., Rees, P. N., Chen, G. M., Walker, C. G., French, S., Brown, E. D., Berdy, J., Liu, D. Y. & Magarvey, N. A. Assembly and clustering of natural antibiotics guides target identification. *Nature Chemical Biology* **12,** 233–239 (2016).

25. Weissman, K. J. The structural biology of biosynthetic megaenzymes. *Nature Chemical Biology* **11,** 660 (2015).

26. Caboche, S., Leclère, V., Pupin, M., Kucherov, G. & Jacques, P. Diversity of monomers in nonribosomal peptides: Towards the prediction of origin and biological activity. *Journal of Bacteriology* **192,** 5143–5150 (2010).

27. Medema, M. H. & Fischbach, M. A. Computational approaches to natural product discovery. *Nature chemical biology* **11,** 639–648 (2015).

28. Mohimani, H., Liu, W.-T., Kersten, R. D., Moore, B. S., Dorrestein, P. C. & Pevzner, P. A. NRPquest: coupling mass spectrometry and genome mining for nonribosomal peptide discovery. *Journal of natural products* **77,** 1902–1909 (2014).

29. Juguet, M., Lautru, S., Francou, F. X., Nezbedová, Š., Leblond, P., Gondry, M. & Pernodet, J. L. An Iterative Nonribosomal Peptide Synthetase Assembles the Pyrrole-Amide Antibiotic Congocidine in Streptomyces ambofaciens. *Chemistry and Biology* **16,** 421–431 (2009).

30. Tobias, N. J., Wolff, H., Djahanschiri, B., Grundmann, F., Kronenwerth, M., Shi, Y. M., Simonyi, S., Grün, P., Shapiro-Ilan, D., Pidot, S. J., Stinear, T. P., Ebersberger, I. & Bode, H. B. Natural product diversity associated with the nematode symbionts Photorhabdus and Xenorhabdus. *Nature Microbiology* **2,** 1676–1685 (2017).

31. Ninomiya, A., Katsuyama, Y., Kuranaga, T., Miyazaki, M., Nogi, Y., Okada, S., Wakimoto, T., Ohnishi, Y., Matsunaga, S. & Takada, K. Biosynthetic Gene Cluster for Surugamide A Encompasses an Unrelated Decapeptide, Surugamide F. *ChemBioChem* **17,** 1709–1712 (2016).

32. Goyal, R. K. & Mattoo, A. K. Multitasking antimicrobial peptides in plant development and host defense against biotic/abiotic stress. *Plant Science* **228,** 135–149 (2014).

33. Reimer, D., Cowles, K. N., Proschak, A., Nollmann, F. I., Dowling, A. J., Kaiser, M., Constant, R. ffrench, Goodrich-Blair, H. & Bode, H. B. Rhabdopeptides as insect-specific virulence factors from entomopathogenic bacteria. *ChemBioChem* **14,** 1991–1997 (2013).

34. Hacker, C., Cai, X., Kegler, C., Zhao, L., Weickhmann, A. K., Wurm, J. P., Bode, H. B. & Wöhnert, J. Structure-based redesign of docking domain interactions modulates the product spectrum of a rhabdopeptide-synthesizing NRPS. *Nature Communications* **9,** 1–11 (2018).

35. Hoyer, K. M., Mahlert, C. & Marahiel, M. A. The Iterative Gramicidin S Thioesterase Catalyzes Peptide Ligation and Cyclization. *Chemistry and Biology* **14,** 13–22 (2007).

36. Li, S., Wu, X., Zhang, L., Shen, Y. & Du, L. Activation of a cryptic gene cluster in lysobacter enzymogenes reveals a module/domain portable mechanism of nonribosomal peptide synthetases in the biosynthesis of pyrrolopyrazines. *Organic Letters* **19,** (2017).

37. Cai, X., Nowak, S., Wesche, F., Bischoff, I., Kaiser, M., Fürst, R. & Bode, H. B. Entomopathogenic bacteria use multiple mechanisms for bioactive peptide library design. *Nature Chemistry* **9,** 379 (2017).

38. Medema, M. H., Paalvast, Y., Nguyen, D. D., Melnik, A., Dorrestein, P. C., Takano, E. & Breitling, R. Pep2Path: Automated Mass Spectrometry-Guided Genome Mining of Peptidic Natural Products. *PLoS Computational Biology* **10,** e1003822 (2014).

39. Moss, N. A., Seiler, G., Leão, T. F., Castro-Falcón, G., Gerwick, L., Hughes, C. C. & Gerwick, W. H. Nature's Combinatorial Biosynthesis Produces Vatiamides A–F. *Angewandte Chemie* **58,** 9027–9031 (2019).

40. Mohimani, H., Gurevich, A., Mikheenko, A., Garg, N., Nothias, L.-F., Ninomiya, A., Takada, K., Dorrestein, P. C. & Pevzner, P. A. Dereplication of peptidic natural products through database search of mass spectra. *Nature Chemical Biology* **13,** 30–37 (2017).

41. Gurevich, A., Mikheenko, A., Shlemov, A., Korobeynikov, A., Mohimani, H. & Pevzner, P. A. Increased diversity of peptidic natural products revealed by modification-tolerant database search of mass spectra. *Nature Microbiology* **3,** 319–327 (2018).

42. Mohimani, H., Wei-Ting Liu, Y.-L. Y., Susana P. Gaudêncio, W. F., Dorrestein, P. C. & Pevzner, P. A. Multiplex de novo sequencing of peptide antibiotics. *Journal of Computational Biology* **18,** 1371–1381 (2011).

43. He, J. & Hertweck, C. Iteration as Programmed Event during Polyketide Assembly; Molecular Analysis of the Aureothin Biosynthesis Gene Cluster. *Chemistry and Biology* **10,** 1225–1232 (2003).

44. Wilkinson, B., Foster, G., Rudd, B. A. M., Taylor, N. L., Blackaby, A. P., Sidebottom, P. J., Cooper, D. J., Dawson, M. J., Buss, A. D., Gaisser, S., Böhm, I. U., Rowe, C. J., Cortés, J., Leadlay, P. F. & Staunton, J. Novel octaketide macrolides related to 6-deoxyerythronolide B provide evidence for iterative operation of the erythromycin polyketide synthase. *Chemistry and Biology* **7,** 111–117 (2000).

45. Meleshko, D., Mohimani, H., Tracanna, V., Hajirasouliha, I., Medema, M. H., Korobeynikov, A. & Pevzner, P. A. BiosyntheticSPAdes: reconstructing biosynthetic gene clusters from assembly graphs. *Genome Research* **29,** 1352–1362 (2019).

46. Kersten, R. D., Yang, Y.-L., Xu, Y., Cimermancic, P., Nam, S.-J., Fenical, W., Fischbach, M. A., Moore, B. S. & Dorrestein, P. C. A mass spectrometry-guided genome mining approach for natural product peptidogenomics. *Nature chemical biology* **7,** 794–802 (2011).

47. Nguyen, D. D., Wu, C. H., Moree, W. J., Lamsa, A., Medema, M. H., Zhao, X., Gavilan, R. G., Aparicio, M., Atencio, L., Jackson, C., Ballesteros, J., Sanchez, J., Watrous, J. D., Phelan, V. V., Van De Wiel, C., Kersten, R. D., Mehnaz, S., De Mot, R., Shank, E. A., Charusanti, P., Nagarajan, H., Duggan, B. M., Moore, B. S., Bandeira, N., Palsson, B., Pogliano, K., Gutieŕrez, M. & Dorrestein, P. C. MS/MS networking guided analysis of molecule and gene cluster families. *Proceedings of the National Academy of Sciences of the United States of America* **110,** E2611–E2620 (2013).

48. Nguyen, D. D., Melnik, A. V., Koyama, N., Lu, X., Schorn, M., Fang, J., Aguinaldo, K., Lincecum, T. L., Ghequire, M. G. K., Carrion, V. J., Cheng, T. L., Duggan, B. M., Malone, J. G., Mauchline, T. H., Sanchez, L. M., Kilpatrick, A. M., Raaijmakers, J. M., De Mot, R., Moore, B. S., Medema, M. H. & Dorrestein, P. C. Indexing the Pseudomonas specialized metabolome enabled the discovery of poaeamide B and the bananamides. *Nature Microbiology* **2,** 1–10 (2016).

49. Behsaz, B., Mohimani, H., Gurevich, A., Prjibelski, A., Fisher, M., Vargas, F., Smarr, L., Dorrestein, P. C., Mylne, J. S. & Pevzner, P. A. De Novo Peptide Sequencing Reveals Many Cyclopeptides in the Human Gut and Other Environments. *Cell Systems* **10,** 99–108 (2020).

50. Tsur, D., Tanner, S., Zandi, E., Bafna, V. & Pevzner, P. A. Identification of post-translational modifications by blind search of mass spectra. *Nature Biotechnology* **23,** 1562–1567 (2005).

51. Wang, M., Carver, J. J., Phelan, V. V, Sanchez, L. M., Garg, N., Peng, Y., Nguyen, D. D., Watrous, J., Kapono, C. A., Luzzatto-Knaan, T., Porto, C., Bouslimani, A., Melnik, A. V, Meehan, M. J., Liu, W.-T., Crüsemann, M., Boudreau, P. D., Esquenazi, E., Sandoval-Calderón, M., Kersten, R. D., Pace, L. A., Quinn, R. A., Duncan, K. R., Hsu, C.-C., Floros, D. J., Gavilan, R. G., Kleigrewe, K., Northen, T., Dutton, R. J., Parrot, D., Carlson, E. E., Aigle, B., Michelsen, C. F., Jelsbak, L., Sohlenkamp, C., Pevzner, P., Edlund, A., McLean, J., Piel, J., Murphy, B. T., Gerwick, L., Liaw, C.-C., Yang, Y.-L., Humpf, H.-U., Maansson, M., Keyzers, R. A., Sims, A. C., Johnson, A. R., Sidebottom,

A. M., Sedio, B. E., Klitgaard, A., Larson, C. B., Boya P, C. A., Torres-Mendoza, D., Gonzalez, D. J., Silva, D. B., Marques, L. M., Demarque, D. P., Pociute, E., O'Neill, E. C., Briand, E., Helfrich, E. J. N., Granatosky, E. A., Glukhov, E., Ryffel, F., Houson, H., Mohimani, H., Kharbush, J. J., Zeng, Y., Vorholt, J. A., Kurita, K. L., Charusanti, P., McPhail, K. L., Nielsen, K. F., Vuong, L., Elfeki, M., Traxler, M. F., Engene, N., Koyama, N., Vining, O. B., Baric, R., Silva, R. R., Mascuch, S. J., Tomasi, S., Jenkins, S., Macherla, V., Hoffman, T., Agarwal, V., Williams, P. G., Dai, J., Neupane, R., Gurr, J., Rodríguez, A. M. C., Lamsa, A., Zhang, C., Dorrestein, K., Duggan, B. M., Almaliti, J., Allard, P.-M., Phapale, P., Nothias, L.-F., Alexandrov, T., Litaudon, M., Wolfender, J.-L., Kyle, J. E., Metz, T. O., Peryea, T., Nguyen, D.-T., VanLeer, D., Shinn, P., Jadhav, A., Müller, R., Waters, K. M., Shi, W., Liu, X., Zhang, L., Knight, R., Jensen, P. R., Palsson, B. Ø., Pogliano, K., Linington, R. G., Gutiérrez, M., Lopes, N. P., Gerwick, W. H., Moore, B. S., Dorrestein, P. C. & Bandeira, N. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nature Biotechnology* **34,** 828–837 (2016).

52. Skinnider, M. A., Merwin, N. J., Johnston, C. W. & Magarvey, N. A. PRISM 3: Expanded prediction of natural product chemical structures from microbial genomes. *Nucleic Acids Research* **45,** W49–W54 (2017).

53. Johnston, C. W., Skinnider, M. A., Wyatt, M. A., Li, X., Ranieri, M. R. M., Yang, L., Zechel, D. L., Ma, B. & Magarvey, N. A. An automated Genomes-to-Natural Products platform (GNP) for the discovery of modular natural products. *Nature Communications* **6,** 1–11 (2015).

54. Tietz, J. I., Schwalen, C. J., Patel, P. S., Maxson, T., Blair, P. M., Tai, H. C., Zakai, U. I. & Mitchell, D. A. A new genome-mining tool redefines the lasso peptide biosynthetic landscape. *Nature Chemical Biology* **13,** 470 (2017).

55. Mohimani, H., Gurevich, A., Shlemov, A., Mikheenko, A., Korobeynikov, A., Cao, L., Shcherbin, E., Nothias, L. F., Dorrestein, P. C. & Pevzner, P. A. Dereplication of microbial metabolites through database search of mass spectra. *Nature Communications* **9,** (2018).

56. Dührkop, K., Shen, H., Meusel, M., Rousu, J. & Böcker, S. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proceedings of the National Academy of Sciences of the United States of America* **112,** 12580–12585 (2015).

57. da Silva, R. R., Wang, M., Nothias, L. F., van der Hooft, J. J. J., Caraballo-Rodríguez, A. M., Fox, E., Balunas, M. J., Klassen, J. L., Lopes, N. P. & Dorrestein, P. C. Propagating annotations of molecular networks using in silico fragmentation. *PLoS Computational Biology* **14,** e1006089 (2018).

58. Bandeira, N., Tsur, D., Frank, A. & Pevzner, P. A. Protein identification by spectral networks analysis. *Proceedings of the National Academy of Sciences U.S.A.* **104,** 6140–6145 (2007).

59. Handelsman, J. Tiny Earth - Studentsourcing Antibiotic Discovery. *Tiny Earth* (2018). at <https://tinyearth.wisc.edu>

60. Kim D., P., Tatiana, T. & Donna R., M. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research* **35,** D61–D65 (2006).

61. Bouslimani, A., Porto, C., Rath, C. M., Wang, M., Guo, Y., Gonzalez, A., Berg-Lyon, D., Ackermann, G., Christensen, G. J. M., Nakatsuji, T., Zhang, L., Borkowski, A. W., Meehan, M. J., Dorrestein, K., Gallo, R. L., Bandeira, N., Knight, R., Alexandrov, T. & Dorrestein, P. C. Molecular cartography of the human skin surface in 3D. *Proceedings of the National Academy of Sciences of the United States of America* **112,** E2120–E2129 (2015).

62. Bouslimani, A., Da Silva, R., Kosciolek, T., Janssen, S., Callewaert, C., Amir, A., Dorrestein, K., Melnik, A. V., Zaramela, L. S., Kim, J. N., Humphrey, G., Schwartz, T., Sanders, K., Brennan, C., Luzzatto-Knaan, T., Ackermann, G., McDonald, D., Zengler, K., Knight, R. & Dorrestein, P. C. The impact of skin care products on skin chemistry and microbiome dynamics. *BMC Biology* **17,** 47 (2019).

63. Mohimani, H., Yang, Y. L., Liu, W. T., Hsieh, P. W., Dorrestein, P. C. & Pevzner, P. A. Sequencing cyclic peptides by multistage mass spectrometry. *Proteomics* **11,** 3642–3650 (2011).

64. Mukherjee, S., Stamatis, D., Bertsch, J., Ovchinnikova, G., Katta, H. Y., Mojica, A., Chen, I. M. A., Kyrpides, N. C. & Reddy, T. B. K. Genomes OnLine database (GOLD) v.7: Updates and new features. *Nucleic Acids Research* **47,** D649–D659 (2019).

65. Frank, A. M., Monroe, M. E., Shah, A. R., Carver, J. J., Bandeira, N., Moore, R. J., Anderson, G. A., Smith, R. D. & Pevzner, P. A. Spectral archives: extending spectral libraries to analyze both identified and unidentified spectra. *Nature Methods* **8,** 587–591 (2011).

66. Bode, E., Heinrich, A. K., Hirschmann, M., Abebew, D., Shi, Y. N., Vo, T. D., Wesche, F., Shi, Y. M., Grün, P., Simonyi, S., Keller, N., Engel, Y., Wenski, S., Bennet, R., Beyer, S., Bischoff, I., Buaya, A., Brandt, S., Cakmak, I., Çimen, H., Eckstein, S., Frank, D., Fürst, R., Gand, M., Geisslinger, G., Hazir, S., Henke, M., Heermann, R., Lecaudey, V., Schäfer, W., Schiffmann, S., Schüffler, A., Schwenk, R., Skaljac, M., Thines, E., Thines, M., Ulshöfer, T., Vilcinskas, A., Wichelhaus, T. A. & Bode, H. B. Promoter Activation in Δhfq Mutants as an Efficient Tool for Specialized Metabolite Production Enabling Direct Bioactivity Testing. *Angewandte Chemie* **131,** 19133–19139 (2019).

67. Moss, S. J., Martin, C. J. & Wilkinson, B. Loss of co-linearity by modular polyketide synthases: A mechanism for the evolution of chemical diversity. *Natural Product Reports* **21,** 575–593 (2004).

68. Nollmann, F. I., Dauth, C., Mulley, G., Kegler, C., Kaiser, M., Waterfield, N. R. & Bode, H. B. Insect-specific production of new GameXPeptides in Photorhabdus luminescens

TTO1, widespread natural products in entomopathogenic bacteria. *ChemBioChem* **16,** 205–208 (2015).

69. Zhou, Q., Grundmann, F., Kaiser, M., Schiell, M., Gaudriault, S., Batzer, A., Kurz, M. & Bode, H. B. Structure and biosynthesis of xenoamicins from entomopathogenic xenorhabdus. *Chemistry - A European Journal* **19,** 16772–16779 (2013).

70. Wenzel, S. C., Meiser, P., Binz, T. M., Mahmud, T. & Müller, R. Nonribosomal peptide biosynthesis: Point mutations and module skipping lead to chemical diversity. *Angewandte Chemie - International Edition* **45,** 2296–22301 (2006).

71. Wenzel, S. C., Kunze, B., Höfle, G., Silakowski, B., Scharfe, M., Blöcker, H. & Müller, R. Structure and biosynthesis of myxochromides S1-3 in Stigmatella aurantiaca: Evidence for an iterative bacterial type I polyketide synthase and for module skipping in nonribosomal peptide biosynthesis. *ChemBioChem* **6,** 375–385 (2005).

72. Seyedsayamdost, M. R., Traxler, M. F., Zheng, S. L., Kolter, R. & Clardy, J. Structure and biosynthesis of amychelin, an unusual mixed-ligand siderophore from amycolatopsis sp. AA4. *Journal of the American Chemical Society* **133,** 11434–11437 (2011).

73. Arima, K., Kakinuma, A. & Tamura, G. Surfactin, a crystalline peptidelipid surfactant produced by Bacillus subtilis: Isolation, characterization and its inhibition of fibrin clot formation. *Biochemical and Biophysical Research Communications* **31,** 488–494 (1968).

74. Nishikiori, T., Naganawa, H., Muraoka, Y., Aoyagi, T. & Umezawa, H. Plipastatins: New inhibitors of phospholipase A2, produced by bacillus cereus BMG302-fF67: II. structure of fatty acid residue and amino acid sequence. *The Journal of Antibiotics* **39,** 745–754 (1986).

75. Vollenbroich, D., Özel, M., Vater, J., Kamp, R. M. & Pauli, G. Mechanism of inactivation of enveloped viruses by the biosurfactant surfactin from Bacillus subtilis. *Biologicals* **25,** 289–297 (1997).

76. Huang, X., Lu, Z., Zhao, H., Bie, X., Lü, F. X. & Yang, S. Antiviral activity of antimicrobial lipopeptide from Bacillus subtilis fmbj against Pseudorabies Virus, Porcine Parvovirus, Newcastle Disease Virus and Infectious Bursal Disease Virus in vitro. *International Journal of Peptide Research and Therapeutics* **12,** 373–377 (2006).

77. Wu, Y. S., Ngai, S. C., Goh, B. H., Chan, K. G., Lee, L. H. & Chuah, L. H. Anticancer activities of surfactin potential application of nanotechnology assisted surfactin delivery. *Frontiers in Pharmacology* **8,** 761-undefined (2017).

78. Sandrin, C., Peypoux, F. & Michel, G. Coproduction of surfactin and iturin A, lipopeptides with surfactant and antifungal properties, by Bacillus subtilis. *Biotechnology and Applied Biochemistry* **12,** 370–375 (1990).

79. Cochrane, S. A. & Vederas, J. C. Lipopeptides from Bacillus and Paenibacillus spp.: A Gold Mine of Antibiotic Candidates. *Medicinal Research Reviews* **36,** 4–31 (2016).

80. Rodrigues, L., Banat, I. M., Teixeira, J. & Oliveira, R. Biosurfactants: Potential applications in medicine. *Journal of Antimicrobial Chemotherapy* **57,** 609–618 (2006).

81. Wang, C. L., Ng, T. B., Yuan, F., Liu, Z. K. & Liu, F. Induction of apoptosis in human leukemia K562 cells by cyclic lipopeptide from Bacillus subtilis natto T-2. *Peptides* **28,** (2007).

82. Agrawal, S., Acharya, D., Adholeya, A., Barrow, C. J. & Deshmukh, S. K. Nonribosomal peptides from marine microbes and their antimicrobial and anticancer potential. *Frontiers in Pharmacology* **21,** (2017).

83. Zhao, H., Zhao, X., Lei, S., Zhang, Y., Shao, D., Jiang, C., Sun, H. & Shi, J. Effect of cell culture models on the evaluation of anticancer activity and mechanism analysis of the potential bioactive compound, iturin A, produced by: Bacillus subtilis. *Food and Function* **10,** 1478–1489 (2019).

84. Gong, A. D., Li, H. P., Yuan, Q. S., Song, X. S., Yao, W., He, W. J., Zhang, J. B. & Liao, Y. C. Antagonistic mechanism of iturin a and plipastatin a from Bacillus amyloliquefaciens S76-3 from wheat spikes against Fusarium graminearum. *PLoS ONE* **10,** e0116871 (2015).

85. Sandrin, C., Peypoux, F. & Michel, G. Coproduction of surfactin and iturin A, lipopeptides with surfactant and antifungal properties, by Bacillus subtilis. *Biotechnology and Applied Biochemistry* **12,** (1990).

86. Gao, L., Han, J., Liu, H., Qu, X., Lu, Z. & Bie, X. Plipastatin and surfactin coproduction by Bacillus subtilis pB2-L and their effects on microorganisms. *Antonie van Leeuwenhoek, International Journal of General and Molecular Microbiology* **110,** 1007–1018 (2017).

87. Lange, A., Sun, H., Pilger, J., Reinscheid, U. M. & Gross, H. Predicting the Structure of Cyclic Lipopeptides by Bioinformatics: Structure Revision of Arthrofactin. *ChemBioChem* **13,** 2671–2675 (2012).

88. Li, W., Rokni-Zadeh, H., De Vleeschouwer, M., Ghequire, M. G. K., Sinnaeve, D., Xie, G. L., Rozenski, J., Madder, A., Martins, J. C. & De Mot, R. The Antimicrobial Compound Xantholysin Defines a New Group of Pseudomonas Cyclic Lipopeptides. *PLoS ONE* **8,** e62946-undefined (2013).

89. Conti, E., Stachelhaus, T., Marahiel, M. A. & Brick, P. Structural basis for the activation of phenylalanine in the non-ribosomal biosynthesis of gramicidin S. *EMBO Journal* **16,** 4174–4183 (1997).

90. Stachelhaus, T., Mootz, H. D. & Marahiel, M. A. The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. *Chemistry and Biology* **6,** 493–505 (1999).

91. Rausch, C., Weber, T., Kohlbacher, O., Wohlleben, W. & Huson, D. H. Specificity prediction of adenylation domains in nonribosomal peptide synthetases (NRPS) using transductive support vector machines (TSVMs). *Nucleic Acids Research* **33,** 5799–5808 (2005).

92. Mohimani, H. & Pevzner, P. A. Dereplication, sequencing and identification of peptidic natural products: from genome mining to peptidogenomics to spectral networks. *Natural product reports* **33,** 73–86 (2016).

93. Tanner, S., Shu, H., Frank, A., Wang, L. C., Zandi, E., Mumby, M., Pevzner, P. A. & Bafna, V. InsPecT: Identification of posttranslationally modified peptides from tandem mass spectra. *Analytical Chemistry* **77,** 4626–4639 (2005).

94. Mohimani, H., Kim, S. & Pevzner, P. A. A new approach to evaluating statistical significance of spectral identifications. *Journal of Proteome Research* **12,** 1560–1568 (2013).

95. Kim, S. & Pevzner, P. A. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nature Communications* **5,** 5277 (2014).

96. Bode, E., Brachmann, A. O., Kegler, C., Simsek, R., Dauth, C., Zhou, Q., Kaiser, M., Klemmt, P. & Bode, H. B. Simple 'on-demand' production of bioactive natural products. *ChemBioChem* **16,** 1115–1119 (2015).

97. Bode, H. B., Reimer, D., Fuchs, S. W., Kirchner, F., Dauth, C., Kegler, C., Lorenzen, W., Brachmann, A. O. & Grün, P. Determination of the absolute configuration of peptide natural products by using stable isotope labeling and mass spectrometry. *Chemistry - A European Journal* **18,** 2342–2348 (2012).

CHAPTER 3.

Long-Read Metagenome Assembly for Reconstructing Biosynthetic Gene Clusters

3.1. ABSTRACT

Non-Ribosomal Peptides (*NRPs*) are biomedically important natural products that include many antibiotics and antitumor agents[1,2]. Search for new NRPs is an important goal since many pathogens have developed resistance against most drugs, including NRP antibiotics of the last resort as daptomycin and vancomycin[3]. Today, little is known about antibiotic NRPs that are produced by bacteria that live in the human gut (rather than doctor-prescribed) and it is unclear whether the continuous exposure to them leads to the development of antibiotic resistance. Identifying NRP-producing *Biosynthetic Gene Clusters* (*BGCs*) in human gut microbiome is critically important for discovering such NRPs. In this chapter, we show that, *long-read metagenomic* assemblies reveal many BGCs that synthesize previously unknown NRPs in the *human gut microbiome* as well as some BGCs encoding for NRPs associated with *colorectal cancer*. We also benchmarked multiple assembly methods and demonstrated that *metaFlye*[4] method for scalable long-read metagenome assembly, improves on other assemblers with respect to identification of BGCs that synthesize NRPs.

3.2. INTRODUCTION

Non-ribosomal peptides. *Non-Ribosomal Peptides* (*NRPs*) are biomedically important natural products that include many antibiotics and antitumor agents[1,2]. Most NRPs are *cyclopeptides* synthesized via *non-ribosomal* (rather than genetic) code and built from over 300 different amino acids (rather than 20 standard proteinogenic amino acids). Search for new NRPs is an important goal since many pathogens have developed resistance against most drugs, including NRP antibiotics of the last resort such as daptomycin and vancomycin[3]. Today, little is known about antibiotic NRPs that are produced by bacteria that live in the human gut (rather than doctor-prescribed) and it is unclear whether the continuous exposure to them leads to the development of antibiotic resistance.

De novo cyclopeptide sequencing tools (that analyze *cyclospecta* arising from cyclopeptides) only succeed in the case of well-fragmented mass spectra that constitute a small fraction of all cyclospectra[5]. As a result, although Behsaz et al., 2020[6] recently demonstrated that there exists a surprisingly large array of still unknown cyclopeptides in the human gut (by identifying cyclospectra), the amino acid sequences of the vast majority of these cyclopeptides remain unknown. NRPs are not directly encoded in the genome and are instead assembled by *Non-Ribosomal Peptide Synthetases* (*NRPSs*). NRPSs are multi-modular proteins that are encoded by a set of chromosomally adjacent genes called *biosynthetic gene clusters* (*BGCs*)[7,8]. Currently, the most promising way to sequence an NRP cyclopeptide is the *peptidogenomics* approach that matches its cyclospectrum with a *Non-Ribosomal Peptide Synthetase (NRPS)* that synthesizes this cyclopeptide using tools like NRPminer[9]. In this chapter, we used the terms NRPS to refer to NRPS-encoding BGCs as we are focusing on DNA data.

Metagenome mining for NRPSs. New NRPSs are usually discovered using various *genome mining* approaches enabled by antiSMASH[10] and other genome mining tools[11]. However, *metagenome mining* is still in infancy since the performance of antiSMASH deteriorates in the case of fragmented metagenomic assemblies[12]. Since NRP-producing BGCs are long (average length ~60 kb) and repetitive (made up of multiple highly similar domains) they are specifically difficult to assemble[13]. Meleshko et al.[14], 2019 recently developed the biosyntheticSPAdes tool for NRPS identification in short-read *isolate* assemblies, but, at the same time, acknowledged that short-reads *metagenome* assemblies are not adequate for full-length BGC assembly in case of longer BGCs. Since genome mining approach fails unless almost an entire NRPS is assembled within a single contig[15], only a small number of NRPs have been discovered via *peptidometagenomics* approach based on joint analysis of short-read metagenomic assemblies and mass spectra[9].

In this Chapter, we demonstrated that long-read assemblies address this limitation and identify many NRPSs in the human gut metagenome. Furthermore, we benchmarked several state-of-the-art long-read assemblers (metaFlye[4], Canu[16], and OPERA-MS[17]) using sequencing data generated from human gut samples by Bertrand et al, 2019[17].

3.3. METHODS

Generating human microbiome assemblies. To evaluate the performance of long-read assemblers on human gut datasets, we extracted the available records from the ENA database generated from a cohort of stool samples[17] (project ID: PRJEB29152) and excluded three samples where Canu failed (two samples) or metaFlye failed (one sample). Removing these samples resulted in 19 datasets (Table 3.1) with total read lengths varying from 1.6 Gbp to 8.0 Gbp. Each dataset includes nanopore long-read metagenomics reads as well as the Illumina short-read metagenomic reads generated from the same sample[17].

Table 3.1. ENA/NCBI accession numbers for 19 human gut samples used in benchmarking. Sample numbers are given as they appear in the original manuscript by Bertrand et al., 2019[17]. We removed four datasets (out of the original 23) failed by either canu (in three cases) or metaFlye (in one case).

Sample No	Sample ID	ONT reads accession	Illumia reads accession
1	V06-T-0501-S07	ERR3201932	ERR3201927
2	V03-S-0457-S04	ERR3201933	ERR3201920
3	V02-T-1664-S03	ERR3219598	ERR3201913
5	V00-S-0509-S01	ERR3201936	ERR3201908
6	V05-S-0512-S05	ERR3201937	ERR3201930
7	V03-T-0508-S04	ERR3201938	ERR3201914
8	V05-T-0513-S05	ERR3201939	ERR3201928
9	V01-T-0506-S02	ERR3201940	ERR3201921
10	V03-T-0504-S04	ERR3201941	ERR3201909
11	V02-T-1665-S03	ERR3201942	ERR3201916
14	V07-T-0504-S08	ERR3201945	ERR3201931
15	V03-T-0506-S04	ERR3201946	ERR3201923
16	V04-S-0509-S04	ERR3201947	ERR3201912
17	V07-S-0510-S08	ERR3201948	ERR3201917
18	V03-S-1663-S04	ERR3201949	ERR3201911
19	V07-S-0512-S07	ERR3219597	ERR3201925
21	V02-T-0504-S03	ERR3201952	ERR3201910
22	V04-T-0508-S05	ERR3201953	ERR3201915
23	V08-S-0510-S09	ERR3201954	ERR3201924

We used metaFlye[4] and Canu[16] to assemble each dataset separately, followed by polishing with the corresponding Illumina reads using Pilon[18]. metaFlye and Canu assembled 837 and 815 Mbp of sequence in contigs >10 kbp and 152 and 125 Mbp in contigs >1 Mbp, respectively (separate sample statistics are given in Table 3.2. In brief, metaFlye has produced more 90%-complete contigs (14), had a higher rate of contigs validated using 16S rRNA (77 out of 100) and recovered more plasmids (109) and viruses (49), as compared to Canu. OPERA-MS[17] implements a hybrid approach that initially assembles short-read contigs and then uses long reads to scaffold these contigs. This strategy has resulted in longer, but less contiguous assembly with only one 90%-complete contig and only 16 complete 16S rRNA genes (while metaFlye and Canu reconstructed 852 and 1,091 complete 16S rRNA genes, respectively).

Table 3.2. metaFlye and Canu assemblies of 19 human gut samples. Contigs shorter than 10 kbp were filtered out. All assemblies were further polished using Pilon. The NG50 statistic was calculated based on a genome size equal to the minimum of the metaFlye and Canu total assembly lengths. Genes were predicted using Prodigal.

Sample ID	Total read length (Gbp)	Assembly size (Mb)		NG50 (Kb)		Longest contig (Mb)	
		metaFlye	Canu	metaFlye	Canu	metaFlye	Canu
1	3.18	23	29	31	85	1.8	1.7
2	2.66	14	15	130	179	4.8	4.8
3	5.37	92	80	476	179	4.5	4.5
5	4.22	65	63	77	64	3	1.4
6	1.99	5	5	1,172	891	1.2	1.2
7	7.99	62	66	234	279	2.7	2
8	2.53	29	37	69	179	2.2	1.5
9	1	26	30	51	46	0.9	1.1
10	5.86	78	68	127	113	2.4	2.9
11	4.25	72	58	498	420	4.5	5.8
14	4.3	45	41	153	176	2.3	1.7
15	1.63	53	54	58	54	1.4	1.7
16	2.57	13	31	39	146	0.62	1.3
17	4.93	48	43	145	218	1.8	1.9
18	2.52	29	28	56	110	0.59	1
19	5.17	36	37	142	126	3.2	4
21	7.25	25	31	47	24	0.52	0.13
22	2.67	30	28	994	672	2.8	4.9
23	5.23	97	75	119	63	5.2	2.4

Co-assembly of multiple human gut samples. We further used SibeliaZ[19] to analyze the sequence overlap between the samples (Figure 3.1) and found that 159 Mbp (~40%) of the total sequence generated by metaFlye for all 19 samples appears in at least two samples. We therefore performed co-assembly by running metaFlye on the mix of reads from all samples. As there is a large sequence overlap between human gut samples, we co-assembled all of them by running metaFlye on the mix of reads from all samples. Co-assembly is computationally more difficult than assembling each sample separately due to (1) increased strain divergence levels and (2) increased shared sequence content that complicates the assembly graph. Nevertheless, metaFlye co-assembly resulted in 453 Mbp of sequence, which closely matched the amount of nonredundant sequence from assemblies of separate samples. We also attempted to run Canu on the mix of all reads but terminated the pipeline after no substantial progress within a month of running it on a computational server.

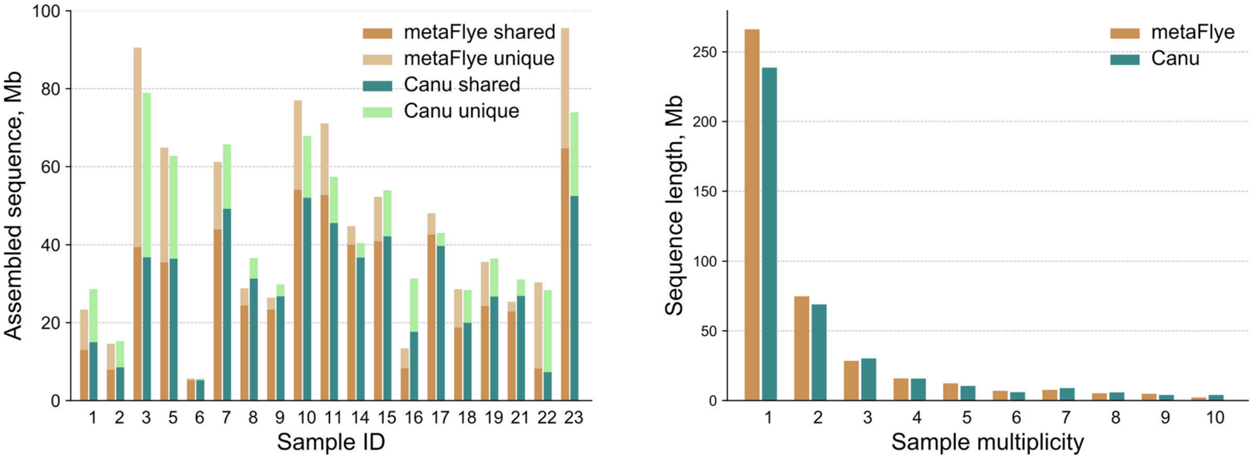


Figure 3.1. Multi-way sequence alignments were computed using SiebliaZ[19]. **(left)** The proportions of unique and shared sequences in each sample. An assembled segment within a sample is called unique if it has no alignments against sequence from any other samples. Otherwise, the segment is shared. **(right)** The total amount of sequence for each multiplicity bin. A sequence fragment belongs to the multiplicity bin X if it is shared by exactly X samples.

Identification of NRPS contigs. Genome mining tools use the previously identified NRPSs to identify NRPSs in a newly sequenced genome. AntiSMASH genome mining tool[10] translates the nucleotide sequence of each contig into amino acid sequences (in all frames) and constructs protein-to-protein alignments against all known NRPSs. It classifies a protein in the translated sequence as *homologous* to a protein from a known NRPS, if they form a sufficiently long alignment with sufficiently high sequence identity. Using a predefined set of NRPS-specific domains/proteins (such as *adenylation domains* that are present in all NRPS), antiSMASH classifies a contig as an *NRPS contig* if it finds a protein homologous to an NRPS-specific domain/protein in this contig. We extend the definition of the NRPS contigs to contigs containing *hybrid NRP-polyketide* BGCs that encode hybrids of NRPs and polyketides. In addition to adenylation domains *(A-domains),* related to NRP biosynthesis, these hybrid BGCs include *acetyltransferases domains* (*AT-domains*) specific to polyketide biosynthesis.

For each NRPS contig, antiSMASH computes a *match* metric which is the percentage of genes within the closest known NRPS that have significant similarity to the genes in this contig [10]. To minimize the effect of false positives in antiSMASH predictions, we ignored an NRPS contig if it has only three or fewer A-domains (antiSMASH is unlikely to identify more than three A-domains in a contig that does not contain an NRPS) unless this contig has more than 50% match to a known NRPS. Since the vast majority of NRPSs are longer than 20 kb, we also ignored NRPS contigs shorter than 20 kb.

3.4. RESULTS

Search for NRPSs in human metagenomes. We searched for NRPSs in Opera-MS, Canu, and metaFlye assemblies of 19 human metagenome datasets generated by Bertrand *et al* [17] as well as the metaFlye co-assembly of all these datasets. Opera-MS, Canu, metaFlye, and metaFlye co-assembly identified 6, 8, 10, and 10 NRPSs, respectively. Since some NRPSs appear in multiple samples in Opera-MS, Canu, and metaFlye assemblies, they may be counted multiple times in this analysis. In contrast, each NRPS identified in the metaFlye co-assembly is unique. Table 3.3 provides information about NRPS contigs identified in the metaFlye co-assembly.

Table 3.3. Information about the NRPS contigs in the metaFlye co-assembly. Each row provides a NRPS contig generated via metaFlye co-assembly method.

contig length (kb)	closest known BGC	matching genes (%)	BGC length	# A-domains in contig/ reference	# AT-domains in contig/ reference	MIBiG ID of known BGC	length of reference BGC (kb)
20	Unknown		20394	4	0		
29	Unknown		29124	5	0		
62	Acinetobactin	82	50973	3 / 2	0 /0	BGC0000294	32
68	Colibactin	95	67552	7 / 7	3 / 3	BGC0000972	55
77	Unknown		61682	4	1		
78	Paenibacterin	60	48993	3 / 13	0 / 0	BGC0000400	53
106	Turnerbactin	30	53145	4 /1	0 / 0	BGC0000451	24
121	Unknown		56532	5	0	-	
213	Myxothiazol	42	55443	8 / 3	0 /6	BGC0001024	43
1060	Unknown		59688	5	0	-	

An NRPS identified within an NRPS contig is classified as *matching* if at least 50% of the genes in a known NRPS match the genes in an NRPS appearing in this contig. Identification of matching NRPSs is important because it enables analysis of NRPS conservation/evolution and connections between the NRPS and the peptide it encodes through the non-ribosomal code. Since antiSMASH aligns translated contigs against proteins in known NRPSs, frame-shift causing errors (i.e., insertions and deletions in metagenomics assemblies) may “hide” similarities between the compared proteins. We thus minimized the effect of frame-shift causing indels by constructing nucleotide-based (rather than amino acid-based) comparison between all matching contigs and the reference BGCs. metaFlye co-assembly identified three matching NRPSs that synthesize acinetobactin[20], colibactin[21], and paenibacterin[22] (Table 3.4). Opera-MS, Canu, and metaFlye (separate) assemblies identified only one of these NRPSs (Opera-MS identified colibactin while Canu and metaFlye (separate) identified acinetobactin).

Table 3.4. Matching NRPSs (acinetobactin, colibactin, and paenibacterin) found in human metagenomes. The column "match (%)" refers to the protein-level match metric, the percentage of matching proteins from a known NRPS (computed by antiSMASH). The column "sequence identity" refers to the nucleotide-level percent identity of the alignment against the reference NRPS. Column "reference BGC aligned (%)" shows the total alignment length (not considering the indels) as the percentage of the length of the corresponding reference NRPS. The last two columns present the number of A-domains and AT-domains in the contig and in the corresponding reference NRPS. The GenBank entry "*Escherichia coli* colibactin polyketide biosynthesis gene cluster and flanking regions, strain IHE3034" (NCBI accession ID AM229678.1) was used as the colibactin NRPS reference. The GenBank entry "*Acinetobacter baumannii* genes involved in acinetobactin biosynthesis and ferric complex transport" (NCBI accession ID AB101202.1) was used as the acinetobacter NRPS reference. The GenBank entry *Ruminococcus obeum A2-162 draft genome* (NCBI accession ID FP929054.1 was used as the paenibacterin reference since it generated the longest alignment 94.2% nucleotide sequence identity.

Assembler	**match (%)**	**contig length (kb)**	**reference BGC aligned (%)**	**sequence identity (%)**	**# A-domains contig / reference**	**# AT-domains contig / reference**
Acinetobactin **- reference BGC (32,436 bp)**						
Canu	65	24	73	94.62	2 / 2	NA
metaFlye	95	62	100	94.75	2 / 2	NA
metaFlye co-assembly	82	62	100	94.29	2 / 2	NA
Colibactin - reference BGC (55,140 bp)						
Opera-MS	95	115	95	99.93	5 / 7	3 / 3
metaFlye co-assembly	95	68	100	99.86	7 / 7	3 / 3
Paenibacterin - reference BGC (52,556 bp)						
metaFlye co-assembly	60	78	93	94.20	3 / 13	0/0

Acinetobactin BGC. Acinetobactin is an NRP with two A-domains produced by *Acinebacteria baumanii*, a multi-drug resistant gram-negative pathogen that causes serious infections of immunocompromised patients[20]. All methods but Opera-MS assembled contigs matching the acinetobactin-encoding BGC. Both Canu and metaFlye assemblies of all samples resulted in the identification of acinetobactin in a single sample #22. metaFlye (separate) and metaFlye (co-assembly) captured acinetobactin BGC within 62 kb long contig each, while Canu captured it within a 24 kb long contig with a lower match score (Table 3.4).

Colibactin BGC. About 20% of humans carry *E. coli* strains containing a BGC that encodes a DNA-damaging compound colibactin (a hybrid NRP-polyketide with seven A-domains and three AT-domains). When some inflammatory condition co-occurs with *E. coli* infection, these strains are able to deliver colibactin to enterocytes and induce gastrointestinal cancer[23]. antiSMASH identified contigs matching colibactin BGC in Opera-MS assembly of a single sample (sample #7) and metaFlye co-assembly, both with a 95% match. Figure 3.2 presents the dot-plot comparison of the identified and reference colibactin BGCs in both assemblies. It shows that Opera-MS assembly resulted in a gapped alignment against the reference BGC as it failed to assemble two identical copies of the cysteine-recruiting A-domain of length ~1,430 bp.

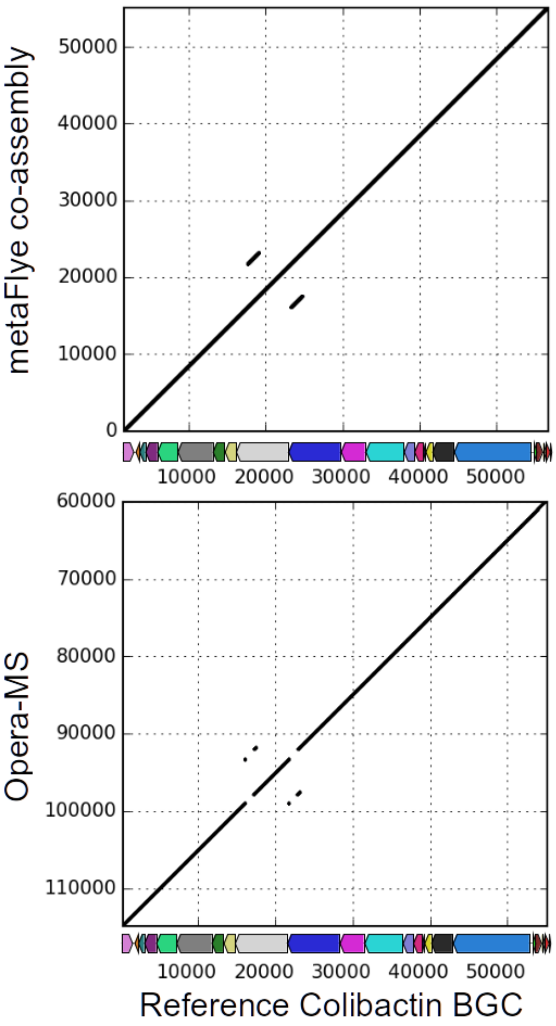


Figure 3.2. Nucleotide alignment of the NRPS contigs matching colibactin-encdoing BGC in assemblies generated by Opera-MS (sample #7) and metaFlye co-assembly against the reference colibactin BGC. Dot plots showing the aligned region based upon the blastn results between the reference colibactin BGC and metaFlye co-assembly of the colibactin contig (top) and Opera-MS assembly of the colibactin contig (bottom). The query sequence is represented on the X-axis and the reference BGCs are represented on the Y-axis. The track under each Y-axis shows the genes and their positions on the reference colibactin BGC where each gene is shown with a different color. The two smaller lines in the top dot-plot highlight the position of a ~1430 bp long repeat in this BGC that corresponds to the identical cysteine-recruiting A-domains appearing on genes *clbG* (positions 16,063 to 17,542 on the reference colibactin BGC) and *clbK* (the positions 21,716 to 23,195 on the reference colibactin BGC). The contig generated by metaFlye co-assembly resolves this repeat, resulting in the full-length BGC captured within a single contig.

Paenibacterin BGC. Meta-Flye co-assembly is the only method that identified an NRPS contig matching the paenibacterin NRPS from *Paenibacillus* sp. *thiaminolyticus* strain OSY-SE **(**60% match score). antiSMASH reported a rather low maximum amino acid identity with proteins in the reference paenibacterin NRPS (39%) indicating a remote evolutionary relationship between the reference strain and one the strain in the analyzed community relevant to the analyzed NRPS contig. We thus searched for a bacterial species with the closest genome (highest sequence identity) by comparing the identified NRPS contig against the entire bacterial nucleotide collection in NCBI. The GenBank entry *Ruminococcus obeum A2-162 draft genome* (NCBI accession ID FP929054.1) generated the longest alignment with 14% coverage of the contig and 94.2% nucleotide sequence identity. Since Ruminococci are extensively studied gut microbes[24], we hypothesize that the identified NRPS contig contains a novel Ruminococci NRPS with remote similarity to the paenibacterin NRPS.

3.5. Discussion

In this chapter, we focused on the question of identifying NRPS-encoding BGCs using long-read human metagenomics datasets and compared this ability them across different assembly methods. We benchmarked OPERA-MS, Canu and metaFlye assemblers and demonstrated that metaFlye co-assembly recovered more known NRP-synthesizing BGCs than the other assemblies (including separate sample assemblies by metaFlye). Majority of the identified contigs does not have a close counterpart in the database of known BGCs (Table 3.3), indicating existence of unknown NRPS BGCs in human microbiome awaiting further characterization.

metaFlye co-assembly was the only method that resolved all repeats in a known NRP-synthesizing BGC that synthesizes a compound colibactin associated with colorectal cancer[21]. As these repeats represent adenylation domains (that define the colibactin structure), identification of the complete BGC, including each domain, is essential for follow-up structure elucidation efforts using peptidogenomics approaches[9,15]. In this chapter we successfully showed how metaFlye, a scalable algorithm specialized for long-read metagenome assembly can deliver assemblies that are suitable for automated NRP discovery in metagenomic samples.

3.6. ACKNOWLEDGEMENTS

Chapter 3, in full, is a reformatted reprint of a part of “metaFlye: scalable long-read metagenome assembly using repeat graphs” as it appears in *Nature Methods* (2020) by Mikhail, Kolmogorov, Derek M. Bickhart, Bahar Behsaz, Alexey Gurevich, Mikhail Rayko, Sung Bong Shin, Kristen Kuhn, Jeffrey Yuan, Evgeny Polevikov, Timothy P. L. Smith & Pavel A. Pevzner. The dissertation author was a primary author of the material included in this chapter.

Contributions. B.B. performed all the BGC identification efforts and the comparison across the methods in this chapter. M.K. assembled all the datasets using the different assembly methods discussed in this chapter.

3.7. REFERENCES

1. Kersten, R. D., Yang, Y.-L., Xu, Y., Cimermancic, P., Nam, S.-J., Fenical, W., Fischbach, M. A., Moore, B. S. & Dorrestein, P. C. A mass spectrometry-guided genome mining approach for natural product peptidogenomics. *Nature chemical biology* **7,** 794–802 (2011).

2. Medema, M. H., Paalvast, Y., Nguyen, D. D., Melnik, A., Dorrestein, P. C., Takano, E. & Breitling, R. Pep2Path: Automated Mass Spectrometry-Guided Genome Mining of Peptidic Natural Products. *PLoS Computational Biology* **10,** e1003822 (2014).

3. Ling, L. L., Schneider, T., Peoples, A. J., Spoering, A. L., Engels, I., Conlon, B. P., Mueller, A., Hughes, D. E., Epstein, S., Jones, M., Lazarides, L., Steadman, V. a, Cohen, D. R., Felix, C. R., Fetterman, K. A., Millett, W. P., Nitti, A. G., Zullo, A. M., Chen, C. & Lewis, K. A new antibiotic kills pathogens without detectable resistance. *Nature* **517,** 455–459 (2015).

4. Kolmogorov, M., Bickhart, D. M., Behsaz, B., Gurevich, A., Rayko, M., Shin, S. B., Kuhn, K., Yuan, J., Polevikov, E., Smith, T. P. L. & others. metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nature Methods* 1–8 (2020).

5. Mohimani, H. & Pevzner, P. A. Dereplication, sequencing and identification of peptidic natural products: from genome mining to peptidogenomics to spectral networks. *Natural product reports* **33,** 73–86 (2016).

6. Behsaz, B., Mohimani, H., Gurevich, A., Prjibelski, A., Fisher, M., Vargas, F., Smarr, L., Dorrestein, P. C., Mylne, J. S. & Pevzner, P. A. De Novo Peptide Sequencing Reveals Many Cyclopeptides in the Human Gut and Other Environments. *Cell Systems* **10,** 99–108 (2020).

7. Marahiel, M. A., Stachelhaus, T. & Mootz, H. D. Modular Peptide Synthetases Involved in Nonribosomal Peptide Synthesis. *Chemical Reviews* **97,** 2651–2674 (1997).

8. Süssmuth, R. D. & Mainz, A. Nonribosomal Peptide Synthesis—Principles and Prospects. *Angewandte Chemie - International Edition* **56,** 3770–3821 (2017).

9. Behsaz, B., Bode, E., Gurevich, A., Shi, Y., Grundmann, F., Mauricio Caraballo-Rodríguez, A., Bouslimani, A., Panitchpakdi, M., Linck, A., Guan, C., Oh, J., Dorrestein, P. C., Bode, H. B., Pevzner, P. A. & Mohimani, H. Integrating Metagenomics and Metabolomics for Scalable Non-Ribosomal Peptide Discovery. *bioRxiv* (2020).

10. Blin, K., Shaw, S., Steinke, K., Villebro, R., Ziemert, N., Lee, S. Y., Medema, M. H. & Weber, T. antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic acids research* **47,** 81–87 (2019).

11. Navarro-Muñoz, J. C., Selem-Mojica, N., Mullowney, M. W., Kautsar, S. A., Tryon, J. H., Parkinson, E. I., De Los Santos, E. L. C., Yeong, M., Cruz-Morales, P., Abubucker, S.,

Roeters, A., Lokhorst, W., Fernandez-Guerra, A., Cappelini, L. T. D., Goering, A. W., Thomson, R. J., Metcalf, W. W., Kelleher, N. L., Barona-Gomez, F. & Medema, M. H. A computational framework to explore large-scale biosynthetic diversity. *Nature Chemical Biology* **16,** 60–68 (2020).

12. Stevenson, L. J., Owen, J. G. & Ackerley, D. F. Metagenome Driven Discovery of Nonribosomal Peptides. *ACS Chemical Biology* **14,** 2115–2126 (2019).

13. Medema, M. H. & Fischbach, M. A. Computational approaches to natural product discovery. *Nature chemical biology* **11,** 639–648 (2015).

14. Meleshko, D., Mohimani, H., Tracanna, V., Hajirasouliha, I., Medema, M. H., Korobeynikov, A. & Pevzner, P. A. BiosyntheticSPAdes: reconstructing biosynthetic gene clusters from assembly graphs. *Genome Research* **29,** 1352–1362 (2019).

15. Mohimani, H., Liu, W.-T., Kersten, R. D., Moore, B. S., Dorrestein, P. C. & Pevzner, P. A. NRPquest: coupling mass spectrometry and genome mining for nonribosomal peptide discovery. *Journal of natural products* **77,** 1902–1909 (2014).

16. Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H. & Phillippy, A. M. Canu: Scalable and accurate long-read assembly via adaptive κ-mer weighting and repeat separation. *Genome Research* **27,** 722–736 (2017).

17. Bertrand, D., Shaw, J., Kalathiyappan, M., Ng, A. H. Q., Kumar, M. S., Li, C., Dvornicic, M., Soldo, J. P., Koh, J. Y., Tong, C., Ng, O. T., Barkham, T., Young, B., Marimuthu, K., Chng, K. R., Sikic, M. & Nagarajan, N. Hybrid metagenomic assembly enables high-resolution analysis of resistance determinants and mobile elements in human microbiomes. *Nature Biotechnology* **37,** 937–944 (2019).

18. Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q., Wortman, J., Young, S. K. & Earl, A. M. Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* **9,** e112963 (2014).

19. Minkin, I. & Medvedev, P. Scalable multiple whole-genome alignment and locally collinear block construction with SibeliaZ. *bioRxiv* (2019).

20. Dijkshoorn, L., Nemec, A. & Seifert, H. An increasing threat in hospitals: Multidrug-resistant Acinetobacter baumannii. *Nature Reviews Microbiology* **5,** 939–951 (2007).

21. Vizcaino, M. I. & Crawford, J. M. The colibactin warhead crosslinks DNA. *Nature Chemistry* **7,** 411–417 (2015).

22. Cochrane, S. A. & Vederas, J. C. Lipopeptides from Bacillus and Paenibacillus spp.: A Gold Mine of Antibiotic Candidates. *Medicinal Research Reviews* **36,** 4–31 (2016).

23. Wilson, M. R., Jiang, Y., Villalta, P. W., Stornetta, A., Boudreau, P. D., Carrá, A., Brennan, C. A., Chun, E., Ngo, L., Samson, L. D., Engelward, B. P., Garrett, W. S.,

Balbo, S. & Balskus, E. P. The human gut bacterial genotoxin colibactin alkylates DNA. *Science* **363,** eaar7785 (2019).

24. Ze, X., Duncan, S. H., Louis, P. & Flint, H. J. Ruminococcus bromii is a keystone species for the degradation of resistant starch in the human colon. *ISME Journal* **6,** 1535–1543 (2012).