Development of Software Platforms for Annotation and Dereplication of Peptidic Natural Products

Ricart Altimiras, Emma

UNIVERSITÉ DE GENÈVE

Département d'informatique

FACULTÉ DES SCIENCES
Docteur Frédérique Lisacek

Development of Software Platforms for Annotation and Dereplication of Peptidic Natural Products

THÈSE

présentée à la Faculté des sciences de l'Université de Genève pour obtenir le grade de Docteur ès sciences, mention bioinformatique

par

Emma RICART ALTIMIRAS

de

Calldetenes (Espagne)

Thèse n° 5521

GENÈVE

Atelier d'impression ReproMail

2021

UNIVERSITÉ DE GENÈVE

FACULTÉ DES SCIENCES

DOCTORAT ÈS SCIENCES, MENTION BIOINFORMATIQUE

Thèse de Madame Emma RICART ALTIMIRAS

intitulée :

«Development of Software Platforms for Annotation and Dereplication of Peptidic Natural Products»

La Faculté des sciences, sur le préavis de Madame F. LISACEK, docteure et directrice de thèse (Département d'informatique), Monsieur G. HOPFGARTNER, professeur ordinaire (Département de chimie minérale et analytique), Madame M. PUPIN, docteure (Département d'informatique, Faculté des sciences et technologies, Université de Lille, Lille, France) et Monsieur M. MEDEMA, professeur (Department of Plant Sciences, Bioinformatics, Wageningen University, Wageningen, Netherlands), Monsieur M. MUELLER, docteur (Swiss Institute of Bioinformatrics (SIB), Lausanne, Switzerland), autorise l'impression de la présente thèse, sans exprimer d'opinion sur les propositions qui y sont énoncées.

Genève, le 18 novembre 2020

Thèse - 5521 -

Le Doyen

N.B. - La thèse doit porter la déclaration précédente et remplir les conditions énumérées dans les "Informations relatives aux thèses de doctorat à l'Université de Genève".

to all those who contributed to this thesis even without knowing it

Acknowledgements

First of all I would like to thank Dr. Frédérique Lisacek for her role as supervisor and for giving me the opportunity of doing a PhD, which helped me to grow as a researcher and to develop further skills to pursuit my scientific career. I would also like to thank all the Proteome Informatics Group (PIG) members that have coincide with me during this period of four and a half years. Particularly Dr. Thibault Robin and Julien Mariethoz that have been there since the start until the end, but also Dr. Oliver Horlacher and Dr. Josefina Lascano that kept supporting me despite not being in the group anymore. I am also grateful to Dr. Markus Mueller, that was particularly involved in my thesis when I first started.

During this Phd I visited Lille multiple times in order to meet our collaborators. Thanks Dr. Maude Pupin, Dr. Valérie Leclère and Areski Flissi for attending me during these periods and for your advises regarding my project.

Additionally, I want to thank Prof. Gérard Hopfgartner, Dr. Marnix Medema, Dr. Markus Mueller and Dr. Maude Pupin for accepting being part of my PhD jury.

Support is crucial during a PhD and that is why I would like to mention Yannick Rémy. Thanks for being always there, in the good and the bad moments. With your support you have somehow contributed to this work as well.

****

També m'agradaria dedicar unes paraules en català a la meva família i als meus amics. Gràcies a tots els que heu estat al meu costat durant aquests anys. Després de les meves visites a Catalunya sempre he tornat a Suïssa amb més energies per continuar la tesi. Guifré, gràcies pels teus consells d'informàtica, han estat més útils del que et penses. Jordi i Nani, els meus pares, us vull donar les gràcies perquè aquesta tesi no hagués estat possible sense vosaltres, que sempre m'heu fet costat amb els estudis.

Table of Contents

List of Figures

List of Tables

Abbreviations

BGC	Biosynthetic gene cluster
ChEBI	Chemical Entities of Biological Interest (database)
CID	Collision induced dissociation
DB	Database
FDR	False discovery rate
GNPS	Global Natural Product Social Molecular Networking (database)
KFP	Kendrick formula predictor (software)
KM	Kendrick mass
KMD	Kendrick mass defect
MS	Mass spectrometry
MS\MS, MS2	Tandem mass spectrometry
NGS	Next-generation sequencing
NKM	Nominal Kendrick mass
NP	Natural product
NPAtlas	Natural Products Atlas (database)
NRP	Nonribosomal peptide
NRPS	Nonribosomal peptide synthetase
PIG	Proteome Informatics Group
PKS	Polyketide synthetase
PNP	Peptidic natural product
PSM	Peptide-spectrum match
rBAN	retroBiosynthetic Analysis of Nonribosomal peptides (software)
SM	Secondary metabolite
SPC	Shared peak count
s2m	Smiles2Monomers (software)

Résumé

Les produits naturels peptidiques (PNP) sont des composés chimiques produits par des organismes vivants. Les PNPs impliqués dans le métabolisme secondaire présentent souvent des activités biologiques intéressantes telles que les antibiotiques, les antitumoraux ou les immunosuppresseurs. Par conséquent, l'étude et la découverte de nouveaux PNP peuvent contribuer à remédier au défi de la résistance aux antibiotiques ou à d'autres problèmes de santé qui menacent actuellement le monde. Les techniques à haut débit telles que le séquençage de nouvelle génération (NGS) et la spectrométrie de masse (MS) fournissent des moyens d'analyser les génomes et les peptidomes des organismes. L'interprétation des données obtenues avec ces techniques est simplifiée grâce à l'utilisation d'outils bioinformatiques. Cependant, dans la recherche des PNP, le développement de logiciels pour l'interprétation des données de spectrométrie de masse en tandem (MS / MS) a été limité, en partie, en raison de la fragmentation complexe de ces composés structurellement complexes. Les bases de données chimiques en libre accès dédiées aux PNP sont également limitées. Les plus grandes bases de données, par exemple, sont consacrées aux produits naturels (NP) de façon globales mais contiennent souvent peu d'annotations et des métadonnées. En effet, à notre connaissance, la base de données Norine est la seule ressource exclusivement dédiée aux NP d'origine peptidique et concrètement aux peptides non ribosomaux (PNR). Dans cette thèse, je présente une collection de plateformes bioinformatiques innovatrices dédiées aux PNP dans le but de faciliter l'analyse structurale de ces composés, spécialement pour les expériences MS. Tout d'abord, je présente rBAN, un outil de rétro-biosynthèse principalement dédié à l'annotation des monomères composant les structures chimiques NRP. Il est notable que rBAN pourrait potentiellement être couplé à des outils d'exploration du génome pour l'identification de clusters de gènes biosynthétiques (BGC). Deuxièmement, je présente l'application KFP, un module élémentaire pour la détection des formules chimiques NRP et l'annotation des pics MS, qui implémente une méthode de prédiction basée sur le défaut de masse de Kendrick. Enfin, la thèse culmine avec NRPro, une nouvelle plateforme d'analyse MS / MS des PNPs permettant l'annotation et la dé-

plication automatiques. Ces trois outils ont été intégrés dans Norine dans le but d'enrichir la ressource et d'introduire un nouveau module dédié à la spectrométrie de masse.

Abstract

Peptidic natural products (PNPs) are chemical compounds produced by living organisms. PNPs involved in the secondary metabolism often exhibit interesting bioactivities such as antibiotic, antitumor or immunosupressor. Hence, the study and discovery of new PNPs may contribute to the remediation of the antimicrobial resistance challenge or other important health problems currently threatening the world. High-throughput techniques such as next-generation sequencing (NGS) and mass spectrometry (MS) provide means for the analysis of the genome and peptidome of the producer organisms. The interpretation of the data obtained with such techniques is usually performed with the assistance of bioinformatic tools that simplify the task. However, in PNP research, the development of software for the interpretation of tandem mass spectrometry (MS/MS) data has been limited, in part, because of the intricate fragmentation of these structurally complex compounds. Open-access chemical databases dedicated to PNPs are limited as well. The largest databases, for instance, are devoted to natural products (NPs) in general and often exhibit poor curation and metadata. In fact, to our knowledge, the Norine database is the only resource that is exclusively dedicated to NPs of peptidic origin and concretely to non ribosomal peptides (NRPs). In this thesis I present a collection of innovative bioinformatic platforms dedicated to PNPs with the aim of facilitating the structural analysis of these compounds, especially for MS experiments. First, I introduce rBAN, a retro-biosynthesis tool mainly dedicated to the monomeric annotation of NRP chemical structures. Interestingly, rBAN could potentially be coupled with genome mining tools for the identification of biosynthetic gene clusters (BGCs). Secondly, I present the KFP application, a straightforward program for NRP chemical formula detection and MS peak annotation, which implements a prediction method based on the Kendrick mass defect. Lastly, the thesis culminates with NRPro, a new platform for the MS/MS analysis of PNPs through automatic annotation and dereplication. These three tools were integrated in Norine with the aim of enhancing the resource and introducing a new module dedicated to mass spectrometry.

Chapter 1

Introduction

Humankind has always benefited from the natural resources for the production of energy, food, etc, and the field of medicine is not an exception. In 1928, Alexander Fleming returned to his laboratory when he realized that a culture plate of *Staphylococcus* had been contaminated by a mould. This contamination turned out to be one of the biggest discoveries of the medicine: penicillin [1]. Indeed, this natural antibiotic produced by *Penicillium notatum*, has saved and it is still saving lives all over the world. Since this discovery, many bioactive products have been isolated from microorganisms such as fungi and bacteria. This is reflected in a study undertaken by Newman and Cragg [2], which shows how the 26% of the drugs approved by the US FDA from 1981 to 2014 were natural product or natural products derivatives (Figure 1.1).

The bio-activities exhibited by peptidic natural products (PNPs) are often related to their structure and composition. This is clearly exemplified by certain lipopeptide antibiotics whose fatty acid (FA) chain promotes the penetration into the bacterial membrane [3, 4]. What is more, FA lengths may determine the microbial activity of some peptides: the longer the chain, the higher the bio-activity [4]. Hence, PNP properties can be enhanced by applying chemical modifications to naturally occurring peptides. Indeed, compounds such as oritavancin, a semisynthetic lipoglycopeptide derived from the introduction of a *N*-alkyl-*p*-chlorophenylbenyzl substituent to chloroeremomycin, have been approved by the FDA to treat drug resistant infections [5, 6]. However, engineering new compounds requires deep understanding of the structures, even bio-synthetic pathways of these compounds.

The emergence of high-throughput technologies and bioinformatics has greatly impacted natural product research. Over the last few decades many tools for natural products (NPs) data analysis have been released [7, 8]. In genomics, next gener-

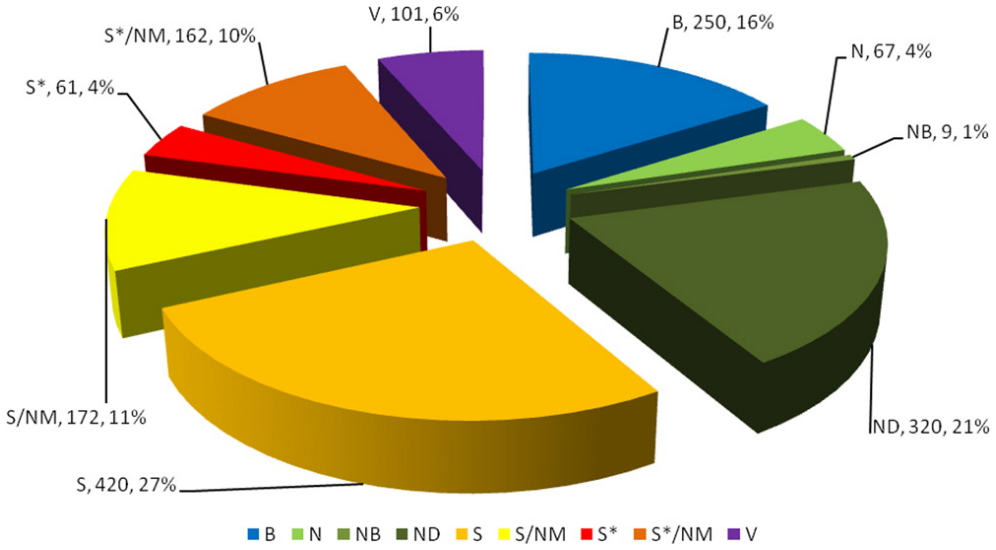


Figure 1.1: All new approved drugs 1981–2014. B: Biological macromolecule, N: Unaltered natural product, NB: Botanical drug (defined mixture), ND: Natural product derivative, S: Synthetic drug, S*: Synthetic drug (NP pharmacophore), V: Vaccine, /NM: Mimic of natural product. Figure courtesy of D. J. Newman and G. M. Cragg [2].

ation sequencing (NGS) has accelerated the analysis of microbial genomes, which provide insight into the potential of a strain to produce certain secondary metabolites. Powerful algorithms have been designed for the identification of biosynthetic gene clusters (BGCs) within the microbial/bacterial genomes [9, 10]. Genome mining tools aid on the detection and interpretation of BGCs but they do not link the identified clusters with structures of known compounds. This gap is filled by algorithms able to reverse the biosynthetic pathways of the compounds in order to identify their initial building blocks [11, 12], a process named retro-biosynthesis. In proteomics, mass spectrometry has become the gold standard technique for high-throughput analysis and has been used in high impact projects such as the first draft of the human proteome [13]. However, the analysis MS/MS spectra from PNPs is much more challenging than that from standard proteogenic peptides. First steps of *dereplication* (identification of known compounds) often involve formula prediction *via* isotope patterns [14, 15, 16, 17]. One of the state-of-the-art dereplication strategies consists in matching the MS/MS spectra against chemical structure databases. This involves modelling the *in silico* fragmentation of the structures and identifying the compound by comparing its theoretical fragments with the experimental counterparts using scoring functions and statistical evaluation. Despite the development of many software for general metabolite analysis [18], hardly any are focused on PNPs. Plus, the few existing PNP-based software present limitations either in the scoring system or in the MS/MS annotation/post-analysis options.

NP databases play an important role in dereplication as well. Unfortunately, the most comprehensive and curated resources [19, 20] are not publicly available. Despite the existence of alternative open-access resources, a common problem within the largest databases is their lack of metadata/annotations [21]. This complicates the retrieval of specific subsets of compounds, in our case those with peptidic origins (PNPs). Alternatively, some databases are specialized in the products of certain strains or in particular classes of natural products. This is the case of Norine, a manually curated database solely devoted to non-ribosomal peptides (NRPs) [22]. To our knowledge, this is the only curated resource entirely dedicated to peptide natural products.

The topic of my thesis arises from the need for **enriching the range of resources devoted to the study of PNPs and addressing some of the limitations of current software**. The algorithms developed in this work are mostly focused on PNP dereplication, either through molecular formula deduction or *via* chemical databases. Retro-biosynthesis is also covered and it is mainly applied to structural annotation rather than genome mining. All the software developed during my thesis has been integrated in Norine in order to include mass spectrometry into the resource. However, before presenting each one of these projects, I provide some background regarding PNPs and the methodologies/software currently existing for their analysis. First, I introduce general concepts regarding NRPs and their biosynthesis together with genome mining and retro-biosynthesis approaches. Then, I focus on the mass spectrometry side of the coin describing the experimental and theoretical fragmentations as well as the different steps and resources for dereplication.

1.1 Non ribosomal peptides

1.1.1 The role of NRPs in medicine

Antimicrobial resistance has become a worldwide problem. In 2014, the World Health Organization (WHO) reported that the treatment of common infections is threatened by an increase of antibiotic resistant strains and many countries have reached alarming levels of resistance [23]. NRPs have shown highly interesting antimicrobial properties that could potentially be used to combat this crisis. A recent finding which caused great excitement in the scientific community was the discovery of the NRP called teixobactin. Teixobactin is produced by *Eleftheria terrae* and binds to lipid II and lipid III of the bacterial cell wall to inhibit biosynthesis [24].

This novel mechanism of action confers antibacterial activity against many Gram-positive drug resistant strains such as *Staphylococcus aureus* or *Mycobacterium tuberculosis*. Furthermore, the properties of teixobactin suggest that it could be the first of a new class of antibiotics with no resistance [25, 26].

Discovering NRPs with antibiotic properties is not uncommon and several NRP-based drugs are already in the market (see Table 1.1). An example of a large family of non ribosomal antibiotics are the β-lactams, which include well-known classes such as the penicillins and cephalosporins. These compounds disrupt the cell wall biosynthesis with the inactivation of the transpeptidation reaction [27]. They have been widely used in medicine and semisynthetic compounds based on their structure have been developed to combat resistant strains [28]. Examples of new drugs derived from cephalosporin are the ceftaroline fosamil acetate and the cetolozane, respectively approved in 2011 and 2014 [2].

Compound	**Class**	**Source**	**Bioactivity**	**Ref.**
Polymyxin B	Polypeptides	*Bacillus polymyxa*	Antibiotic	[29]
Pristinamycin	Depsipeptide	*Streptomyces pristinaespiralis*	Antibiotic	[30]
Gramicidin	Pentadecapeptide	*Bacillus bovis*	Antibiotic	[31]
Bacitracin	Cyclic peptide	*Bacillus subtilis*	Antibiotic	[32]
Capreomycin	Cyclic peptide	*Streptomyces capreolus*	Antibiotic	[33]
Teicoplanin	Glycopeptide	*Actinoplanes teichomyceticus*	Antibiotic	[34]
Vancomycin	Glycopeptide	*Amycolatopsis orientalis*	Antibiotic	[35]
Cephalosporin C	β-lactam	*Acremonium sp.*	Antibiotic	[36]
Oritavancin	-	Semi synthetic	Antibiotic	[37]
Bleomycin	Hybrid peptide	*Streptomyces verticillus*	Antibiotic	[38]
Daptomycin	Lipopeptide	*Streptomyces roseosporus*	Antibiotic	[39]
Cyclosporine A	Cyclic peptide	*Tolypocladium inflatum*	Immunosuppressant	[40]
Actinomycin D	Polypeptide	*Streptomyces sp.*	Antitumor	[41]
Romidepsin	Depsipeptide	*Chromobacterium violaceum*	Antitumor	[42]

Table 1.1: List of NRPs-based drugs in the market. Adapted from S. Agrawal et al. [43].

NRPs also exhibit potential as anticancer agents. This is the case of actinomycin D (Figure 1.3E), a secondary metabolite from *Streptomyces sp*. also known as dactinomycin, that has been used for the treatment of cancer since its approval in 1964 [2]. Actinomycin D inhibits the transcription of the tumoral cells by intercalating into the DNA and blocking the progression of RNA polymerases [44]. The importance of actinomyicn D in medicine is reflected by its presence in the WHO "Model List of Essential Medicines" (2019), where it is recommended for the treatment of gestational trophoblastic neoplasia, rhabdomyosarcoma and nephroblastoma (Wilms tumour) [45]. Another interesting NRP with anticancer properties is the bleomycin (Figure 1.4C). Bleomycin binds with metal ions and forms a complex that in reaction with the oxygen produces highly reactive free radials. These free radicals produce DNA

single or double strand breaks and inhibit the synthesis [46]. It has been applied for the treatment of Hodgkin's lymphoma, testicular cancer, ovarian cancer, and cervical cancer among others [47]. Lastly, trabectedin (*Yondelis*) is an example of a recently approved anti-tumoral NRP-based drug. This marine natural product derives from *Ecteinascidia turbinata* and was introduced in the market in 2007 for the treatment of soft tissue sarcoma [2, 48].

When referring to immunomodulator NRPs, the most relevant example is undoubtebly cyclosporin [49]. This compound produced by *Tolypocladium inflatum* was firstly isolated in 1972 and its approval for medical use revolutionized transplant medicine [50]. The intrinsic immunosupressive properties of the compound make cyclosporin optimal for the treatment of transplant rejection and autoimmune diseases [51, 52]. The mode of action is based on the inhibition of calcineruin, that prevents the activation of the T-cells and overall immune response [49]. Other examples of NRPs with immunomodulating activities are the lipopeptide beauverolide L (Figure 1.4A) and peptides from the edeine family produced by *Bacillus brevis* [53, 54, 55].

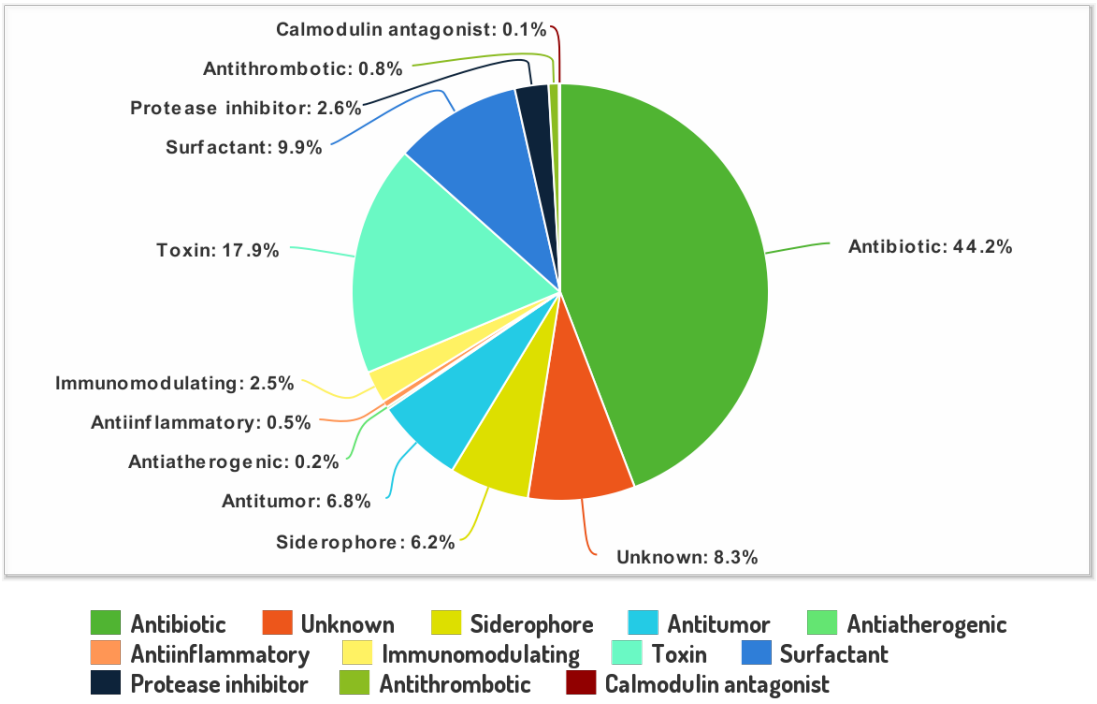


Figure 1.2: Bioactivities of the Norine NRPs.

Toxins, siderophores or surfactants are also present among NRPs. Actually, the range of bioactivities covered by these peptides is considerable and a single NRP may exhibit multiple properties. Just the 1730 compounds in Norine [22] are classified within 12 activity categories. Among them, the antibiotics are predominant (44.2 %), followed by toxins (17.9 %) and surfactants (9.9 %) (Figure 1.2). It is worth-

mentioning that some NRPs and particularly those from marine origin, have shown activity against specific diseases such as HIV, tuberculosis (Figure 1.3D) or malaria [56, 57].

1.1.2 Chemical diversity of NRPs and its impact on the bioactivity

The high degree of diversity encountered in NRPs prevents their classification into a single representative structure. They exist in the form of :

(i) linear chains of diverse lengths, from short sequences of a few monomers (Figure 1.3A), until longer ones with more than 15 monomers (e.g. cephaibols or gramicidins).

(ii) linear sequences with ramifications, often termed "branched" structures. The siderophore cepaciachelin (Figure 1.3B) is an example of them.

(iii) cyclic structures such as those of surfactins (Figure 1.3C), cylcosporins or beauverolides (Figure 1.4A).

(iv) branched cyclic structures. Representative families with this kind of structure include the capreomycins (Figure 1.3D), orfamides and fengycins.

(v) double cyclic structures, as that of the antitumor anctinomycin D (Figure 1.3E).

(vi) complex structures, often corresponding to gylcopeptides such as vancomycin (Figure 1.4B) and teicoplanin (Figure 1.3F).

In terms of composition, while the skeleton of some biological entities such as DNA or proteins can be defined by a small set of repeating structural units, this is not the case of NRPs. Although their subunits are fairly recurrent, the high amount of monomers present in NRPs complicates their listing. Some attempts to define a collection have been made by platforms such as the Pistoia Alliance and Norine. The Pistoia Alliance released a new notation for biopolymers named HELM [58], which is based on their own set of monomers. However, this collection is not specific for NRPs and it also includes monomers from other molecules such as DNA or RNA. Alternatively, the Norine database [22] gathers a unique collection of 543 monomers from NRPs. It includes L- and D-aminoacids, non proteogenic aminoacids, lipids, glycans and chromophores. The combination of these moieties results in lipopeptides, glycopeptides, chromophores and other structures such as peptaibols. The general particularities of these classes and their structure-activity relationships

are presented in the following paragraphs.

Figure 1.3: Biologically relevant NRPs and their structural characteristics.

Lipopeptides are characterized by the presence of a fatty acid moiety within the amino acid chain. When referring to this class of NRPs, one of the most representative families are the surfactins. Surfactins, as suggested by the name, are powerful biosurfactants characterized by a cyclic structure composed of 7 aminoacids (ELLVDLL) and a β-hydroxy fatty acid tail of 12 to 16 carbons (Figure 1.3C). This cyclic structure is involved in the activity of the compound, as reflected by the impact of structural modifications such as linearization, which leads to reduction of surfactin binding [59]. Another relevant structural motif is the FA, that plays an important role for the insertion into the lipid membrane, an action that seems to be enhanced by longer FA chains [60]. Interestingly, the aminoacid composition and the length of the fatty acid do not only depend on the bacterial strain, but also on the culture conditions [61]. For instance, the inclusion of metal ions in cultures of *Bacillus subtilis* stimulates the production of variants with the AME5 monomer and increases the ratio of surfactins with longer fatty acid chains [62]. Together with surfactins, other known families of lipopeptides are the fengycins and iturins, with similar structures and properties.

As already mentioned, **glycopeptides** exhibit some of the most complex structures encountered in NRPs. Mainly represented by the vancomycin antibiotics, their structure presents multiple cycles, sugar motieties and halogenated atoms (Figure 1.4B). Their antibacterial activity has been associated with an affinity for the terminal -D-Ala-D-Ala sequence from the growing cell wall of gram positive bacteria [63, 64]. The dimerization of these compounds plays an important role on binding to the bacterial cell wall [65, 66]. Structural factors that favor dimerization, and thereby increase the antibiotic activity, are the sugar moietes and the chlorine atoms present in some aminoacids [67, 65, 68, 66]. Interestingly, the formation of dimers has not been observed in the teicoplanin antibiotic (Figure 1.3F) [68, 66]. However, the lipophilic tail present in this glycolipopeptide acts as membrane anchor to compensate the lack of dimerization [65, 68, 69].

The peculiarity **chromopeptides** lies in the presence of a chromophore within the aminoacid sequence responsible for their characteristic fluorescence. Probably, the most studied example of chromopeptides are the actinomycins. As already discussed, actinomycin D is currently used for the treatment of certain types of cancer. However, the inherent toxicity of this compound urges the development of structural analogs with the same antitumor properties but reduced toxicity. The primary structure of actinomycin D consists of two cyclic pentapeptide structures

connected by a 2-aminophenoxazin-3-one chromophore (Figure 1.3E). Simple modifications such as the methylation of the amino terminal group of the chromophore [70] or the substitution of aminoacids [71] have shown a reduction of the compound toxicity.

Despite the existence of NRPs with non amino acids moieties, those of peptidic nature are frequent as well. However, the presence of non proteogenic or highly modified amino acids differentiates them from protein-derived peptides. This is well exemplified by the **peptaibol** family. Peptaibols have sequences of 5 to 20 aminoacids and are structurally characterized by three main elements: i) high proportion of α-aminoisobutyric acid (Aib), ii) acetylated N-terminus, iii) an amino alcohol in the C-terminus (Figure 1.3A). The antibiotic activities of these compounds are attributed to the hydrophobic properties given by the rich composition in aliphatic amino acids, which promotes their insertion into the membrane [72, 73]. Similarly to other NRPs, peptaibols are used as scaffold to generate derivatives with higher antimicrobial properties. For instance, the substitution of the amide bonds for triazole rings has been applied to synthesize peptides less susceptible to enzimatic degradation [74, 75].

Finally, it is interesting to mention that another structure occasionally found in some NRPs are the modules with alternating carbonyl and methylene groups ('β-polyketones'). This structural pattern denotes their condition of **non ribosomal peptide-polyketide (NRP-PK) hybrids**. Polyketides are compounds synthesized by polyketide synthases (PKS), multi-modular enzymes different than those responsible of the NRPs biosynthesis, covered in section 1.1.4.

1.1.3 Monomer *bridges*

We use the term monomer *bridges* to refer to those bonds linking NRP monomers. For instance, in the case of proteogenic peptides the monomer bridges are defined by amide groups. NRPs are more complex and instead of a single bond type, multiple functional groups may take this position. Apart from the *classical* peptide bonds, a frequent group acting as a monomer bridge are ester bonds, which "substitute" amide linkages in the so-called depsipeptides (Figure 1.4A). Other fairly recurrent inter-monomer groups include glycosidic bonds, found in NRPs with sugar moieties; thioester and disulfur linkages, usually involving cysteine monomers; or other bonds formed through side-branch linkages such as the aryl-ethers of vancomycin (Figure 1.4B). Indeed, vancomycin is a good example to show the level of complexity reached by NRP linkages, which is not just given by the diversity of bond types, but also by the number of potential links of a single monomer. In van-

comycin, just the Hpg (hydroxyphenylglycine) monomer located in the center of the molecule presents five linkages with other subunits.

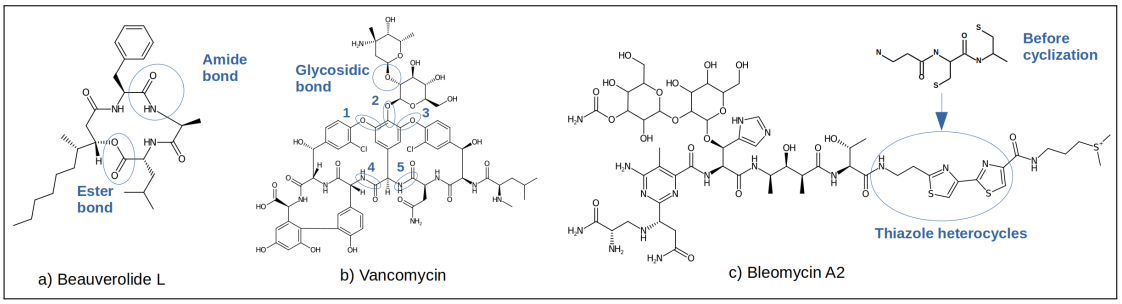


Figure 1.4: Frequent inter-monomer bonds in NRPs.

In addition to the above-mentioned bonds, cyclization reactions occurring during the synthesis of NRPs can lead to the formation of heterocyclic structures between monomer moieties. An example of them are the oxazole and thiazole cycles (Figure 1.4C) often present in NRPs. Note these are some of the frequent NRP linkages, but it is hard to enumerate all the possibilities and, to our knowledge, they have never been exhaustively described. However, in Figure 1.5 I give an overview the monomer bridges observed in Norine compounds. These bonds provide insight into the biosynthesis pathways of NRPs and, as later discussed, they are of high relevance for the development of retro-biosynthesis software.

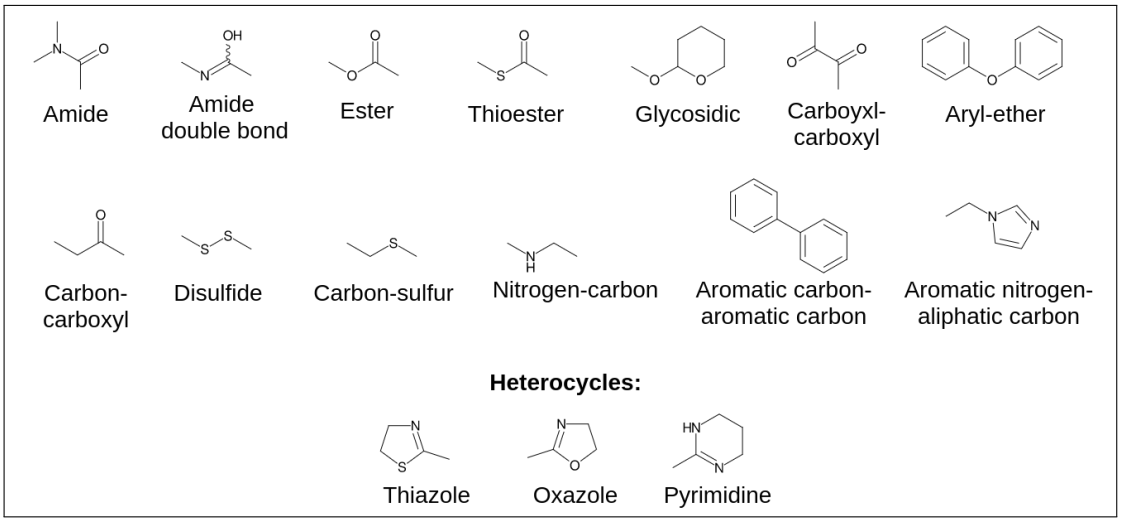


Figure 1.5: Inter-monomer bonds observed in Norine compounds.

1.1.4 NRPs biosynthesis

NRPs are assembled by enzymatic complexes known as non ribosomal peptide synthetases (NRPS). These biosynthetic machineries are constituted of several modules, each one responsible for the incorporation of a specific amino acid in the peptide chain. The modules are classified in initiation, elongation and termination modules, which, as suggested by the name, act on different stages of the building process and contain different catalytic domains. The elongation modules are composed of at least three core domains: i) the adenylation (A) domain, ii) the peptidyl carrier (PCP) domain and iii) the condensation (C) domain (see Figure 1.6).

The **A-domain** regulates the amino acids selection and thereby determines the primary sequence of the peptide bond. The process involves the activation of the amino acids through a reaction with ATP in order to produce the corresponding amino acyl adenylates (Figure 1.61-2). This function has converted the A-domain into the subject of interest of many studies. The examination of the crystal structure from gramicidin S synthetase A uncovered regions from the A-domain that play an important role in the binding with the substrate [76]. More detailed analysis of these regions found specific amino acid fingerprints that, similarly to the codon reading frame used in proteins, resulted in the definition of the *non ribosomal code* [77, 78]. Part of this code was identified by Stachelhaus et al. [78] through the alignment of A-domain sequences from different NRPS, which resulted in the detection of conserved sequences and residues conferring specificity to certain monomers. Then, they proposed specificity-conferring amino acid codes of 10 residues associated with different substrates. These codes were later used for the development of bioinformatic tools predicting A-domain specificities [79, 80, 81]. Machine learning approaches considering the physio-chemical properties of the A domain have also been used for these kind of predictions [82, 83] highlighting the relevance of this domain in the bioinformatics field.

After substrate selection, the amino acyl adenylate is transferred to the **PCP-domain**, which is responsible for the transport of the substrate from one catalytic center to another. A key structure to achieve that it the phosphopantetheinyl arm attached to the domain [85]. The acyl adenylate is covalently tethered to this structure through the formation of a thioester bond [86, 85]. From there, and thanks to the inherent flexibility of the arm, the substrate is brought to the next catalytic center (Figure 1.63). Because of its functionality, the phosphopantetheinyl co-factor is sometimes referred as the *swinging arm* [87]. It is worth noticing that while the aminoacid is fixed to the PCP-domain it can also be modified by other optional domains that will be later introduced.

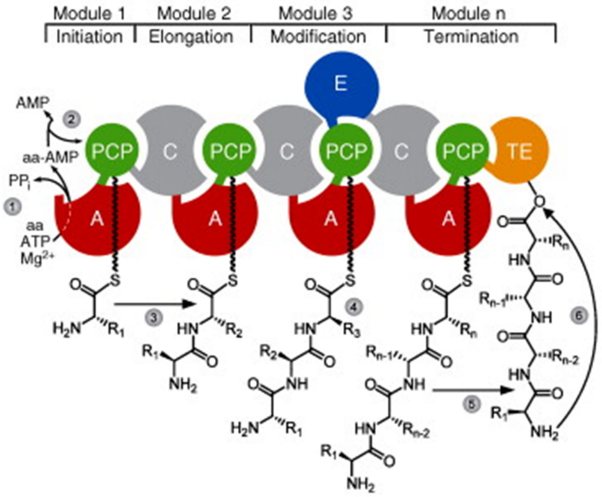


Figure 1.6: Simplified mechanism of NRP synthesis. Courtesy of M. Strieker et al. [84].

The last core domain of the elongation modules is the **C-domain**, which is essential for the extension of the peptide chain because it catalyses the peptide bond formation. To perform this action, two substrates from the adjacent PCP-domains are respectively bound to the acceptor and donor sites from the C-domain (Figure 1.7). The bond formation is achieved by nucleophilic attack of the amino group from the donor amino acid onto the acyl group from the acceptor one [88] (Figure 1.7A). Due to the requirement of two substrates, the condensation domain is usually not found in the first module. The only exception is the starter C-domain, a subclass whose functionality lies on linking the first amino acid with a fatty acid [89]. In fact, starter C-domains are not the only subclass of these enzymes. As already explained, NRPs exhibit monomers with both, L- and D- configurations. Consequently, C-domains can be classified according to the stereochemistry of the amino acids involved in the bond formation. As indicated by their annotations, the $^{L}C_{L}$ domains link monomers with L configuration, while the $^{D}C_{L}$ domains link a monomer with D configuration with another in L form. Finally, it is important to mention that in some modules the C-domain is replaced by the **heterocyclization (Cy) domain**, which catalyzes the bond formation followed by the heterocyclization of cysteine, serine, and threonine monomers [89, 90] (Figure 1.7B). Hence, this module is responsible for the characteristic thiazoline and oxazoline rings found in some NRPs. These structures play an important role on the bioactivity of the peptides as they confer them the ability to chelate metals or interact with proteins, DNA or RNA. Additonally, they can be further modified by either the oxidation (Ox) or reduction (R) domains resulting in thiazole/oxazole rings or thiazolidine/oxazolidine rings respectively [90].

Having introduced the enzymes conforming the elongation modules, only the de-

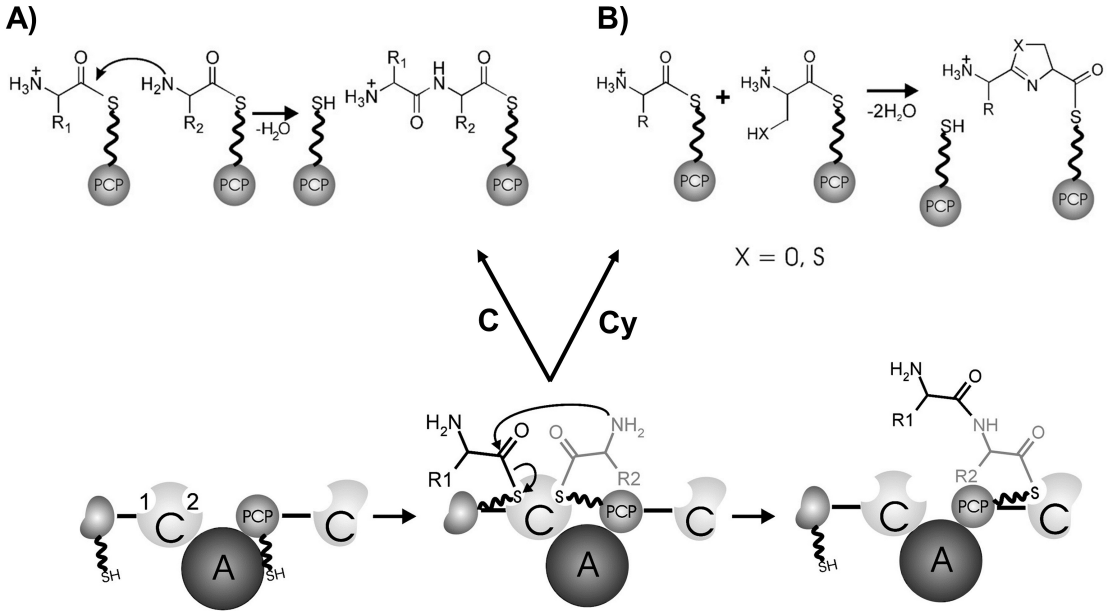


Figure 1.7: C-domain reactions. Adapted from R. Finking and M. A. Marahiel [91].

scription of one core domain remains. This is the **Thioesterase (TE) domain**, located in the last module and which ends the synthesis of the peptide by liberating the product. The release is achieved either by hydrolysis or cyclization, producing a linear or a macrocyclic peptide [92]. Interestingly, the release of the cyclic form seems to be more common, what may be related to the stronger resistance of these structures to proteolytic breakage [91].

Although the core domains are essential to build up NRPs, the synthetases dispose of other optional enzymes known as tailoring domains that highly contribute to the structural diversity of these compounds. The tailoring domains are classified in *cis* and *trans* domains. *Cis* domains are part of the non ribosomal peptide synthetase, while *trans* domains are acting externally.

A well-known tailoring enyzme is the **epimerization (E) domain**, which acts in a *cis* manner and is the main cause of the presence of D-monomers in NRPs. As previously commented, after the establishment of a covalent bond with the PCP-domain, the monomers can be modified. At this specific stage the E domain catalyzes the epimerization of the monomers into D-configuration [93]. Although this is the most common mechanism to introduce D stereoisomers, it can also be accomplished by other enzymes. That is the case of the **Dual Epimerization/Condensation (E/C) domains**, a sub-type of C-domains able to catalyze both, the epimerization and the subsequent condensation of the two residues involved [89, 94]. **D-amino-acid-selective A-domains** are also responsible for the addition of D-monomers. Those

exist, for instance, in the cyclosporin synthetase, whose A-domain incorporates a D-Ala monomer provided by an alanine racemase [95, 96].

Methylation observed in NRPs arises from the **methyltransferase (MT) domains**, which are sub-classified as carbon (C-), nitrogen (N-) or oxigen (O-) methyltransferases depending on the methylated atom. The mechanism of action is common for all the subclasses, consisting in the transfer of a methyl group from S-adenosyl methionine to the targeted atom of the peptide [97, 98]. Normally it occurs in a *cis* acting mode, such in the case of the pyochelin syntethase [99]. However, there are examples where the methylation is *trans*, for example that performed by the N-methyltransferase MtfA in chloroeremomycin synthesis.

Apart from methylated amino acids, formyl groups are also present in NRPs. Those modifications are due to the presence of the **formylation (F) domain** in the initiation modules of the synthetase. Gramicidin for example, is a linear peptide whose first residue in the chain (valine) was found to be formylated by the F-domain in the presence of N10-formyltetrahydrofolate [100].

Chlorine, fluorine and other halogen atoms are also added by an specific enzyme: the **halogenase domain**. Two halogenating enzymes related to NRPs have been discovered so far: i) the flavin-dependent halogenases, responsible of the inclusion of halogen groups in aromatic rings [101] and ii) the non-heme iron-dependent halogenases, that introduce halogens in unactivated aliphatic carbon centers [102].

Finally, another structural entity attached by tailoring domains are the sugars encountered in classes such as the vancomycin antibiotics, which are introduced by **glycotransferases**. Although these enzymes are not encoded by the peptide synthetase genes, they are normally present in the same biosynthetic cluster. In the glycopeptide balhimycin, for instance, the the glycotransferase genes (*bgtfA-C*) are located next to the halogenase (*bhaA*) and oxigenase genes (*oxyA-C*), in a region downstream of the peptide synthetase genes (*'bps*) [103] (Figure 1.8).

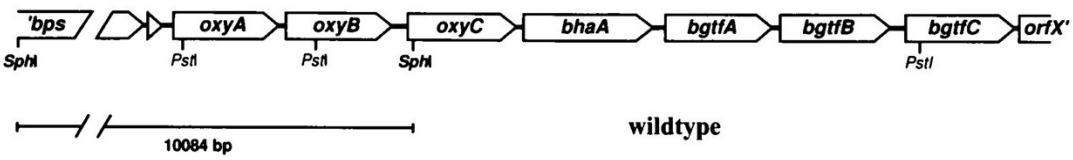


Figure 1.8: Location of the glycotransferase genes (*bgtfA-C*) in the biosynthetic gene cluster of balhimycin. Courtesy of S. Pelzer et al. [103].

1.2 Methods and tools for NRPs biosynthesis analysis and annotation

1.2.1 *In silico* genome mining: the forward path

As observed in the previous section, NRPS domains provide insight into the peptidic product synthesized. Thus, the identification of NRPS gene clusters supports the study of the metabolic pathways and potential NRPs produced by the targeted microorganisms. Here, I briefly describe the main computational strategies for the analysis of secondary metabolite bio-synthetic pathways *via* genome mining and I present multiple software tools and databases dedicated to this field of research. Note that only a part of the available tools are covered in this section, for a broader view of all the existent resources you can refer to the Secondary Metabolite Bioinformatics Portal [104] (SMBP; http://www.secondarymetabolites.org).

Bio-synthetic gene cluster (BGC) identification is the main goal of genome mining software. The most common approaches for BGC detection involve the usage of profile Hidden Markov Models (pHMMs) or other sequence-alignment strategies such as BLAST. In profile-based strategies, pHMMs are generated using signature genes involved in the SM biosynthesis. Then, tools such as HMMer [105] are used to search the profiles within the query dataset. After gene clusters detection, further analysis may be undertaken. For NRPs in particular, the identified domains can be used for substrate structural prediction. As previously described (Section 1.1.4), the active sides of adenylation domains present amino acid signatures that help the prediction of the substrate specificity. Thus, previously characterized A domains can be used for prediction. Although BLAST and HMMer can also be applied to substrate specificity prediction, machine learning approaches such as the support vector machines (SVMs) implemented in **NRPSPredictor2** [83] have shown higher accuracy.

One of the first solutions for detection of secondary metabolite gene clusters was the commercial software **ClustScan** [106]. It provided semi-automatic annotation of gene clusters from PKS and NRPS, as well as product structure prediction. Furthermore, the tool was associated to **ClustScan Database** (CSDB) [106, 107], which contained BGCs identified with the software. However, the original links to these resource are not accessible anymore, that suggests a stop in their maintenance. Alternatively, the **NP.Searcher** [80] algorithm uses subsequent BLAST alignments and an internal database of known A-domain signatures for BGC identification and prediction NRPS/PKS molecular structures. The multiple alignment

steps consist in i) identifying gene clusters by matching key catalytic domains from NRPS/PKS, ii) recognizing A- and AT- domains sequence motifs and matching them against the signature database for specificity prediction, iii) identifying auxiliary domains causing epimerization, reduction or methylation reactions of the substrate and iv) finding tailoring domains that may also affect the substrate through glycosylation, heteorcyclization, etc. This software is specially interesting for the generation of putative structures of NRPs, although new A-domain signature occurrences not present in the database may cause inaccurate predictions.

AntiSMASH [108] is a more comprehensive pipeline for the identification of gene clusters from bacterial, fungal and plant genomes. It is not just limited to the analysis of NRPS/PKS, but it also includes many other signature enzymes such as terpenes, aminoglycosides, bacteriocins and nucleosides, among others. Furthermore, antiSMASH integrates a variety of tools for gene cluster analysis that facilitate the execution of in-depth analysis. For instance, phylogenetic analysis can be performed by classifying the identified genes in Clusters of Orthologous Groups (*smCOG*), which represent SM-specific gene families. Then, phylogenetic trees are generated to show the relatedness between the genes from the same *smCOG*. Another relevant module is ClusterBlast, which identifies homologous gene clusters of the BGC identified. For the analysis of NRPs, antiSMASH integrates NRPSPredictor2 for substrate specificity prediction. Since its release in 2011, antiSMASH has been regularly updated and extended [109, 110, 111, 112]. It provides an interactive GUI (Figure 1.9), cross-links to other services and it has been used in other external pipelines. An example of them is **Pep2Path** [113], that identifies BGC of NRPs and PKs by comparing the substrate specificity predictions from antiSMASH/NRPSPredictor2 with the mass shifts detected in MS/MS spectra of the target compounds.

Other worth-mentioning tools for genome mining include **SMURF** [114], focused on the analysis of fungal genomes, and **PRISM** [115, 116], targeting substrate structural prediction. Indeed, PRISM theoretical structures are used in the *Genomes-to-Natural Products Platform* (GNP) for the identification of novel NRPs and PKs from LC-MS/MS [117]. Additionally, the tool is also integrated in the GARLIC/GRAPE pipeline, an approach to match known chemical structures with their respective BGCs [11].

Apart from BGC identification tools, database resources storing NRP and PK biosynthetic clusters information also play an important role in genome mining. The first approach for collecting BGC data was **ClusterMine360** [118]. This database contains about 300 hand-curated clusters associated with the biosynthesis of more than 200 NRPS/PKS. A similar crowd-sourcing initiative was started in 2015 for

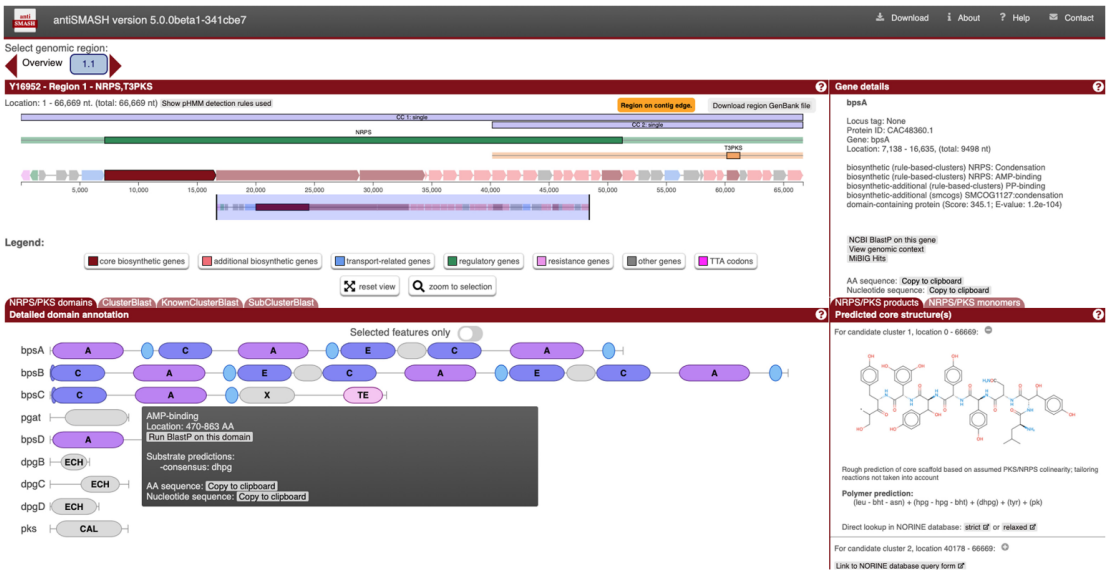


Figure 1.9: antiSMASH interface.

the creation **MIBiG** repository [119, 120]. MIBiG has collected a total of 1923 gene clusters, part of them associated with the biosynthesis of 605 NRPs. The data in MIBiG is manually curated and provides literature references and cross-links to chemical structure databases. Finally, it is important to mention the largest freely available collection of BGCs: the **IMG-ABC** knowledge-base [121, 122, 123]. Both, computationally predicted and manually characterized gene clusters are included in the database. Plus, the users can perform large-scale analysis using the available search, analysis and export tools.

1.2.2 *In silico* retro-biosynthesis for monomer-based annotation.

Genome mining tools identify biosynthetic gene clusters without associating them with already known chemical structures. Yet with the sequenced genomes and elucidated structures from many microorganisms, structures and BGCs are not matched [124] and this matching is not a trivial task. Retro-biosynthesis approaches can solve this problem. In the same way a synthetase provides indications towards its NRP product (forward path), a given product gives insight into its synthetase (reverse path). The goal of retro-biosynthesis is to predict the precursor building blocks (monomers) that generated a specific chemical structure. Then, chemical structures and gene clusters can be linked by comparing the structurally-predicted monomers against the BGC-predicted ones [11]. Alternatively, retro-biosynthesis can be applied for annotation purposes. In this case the monomers obtained with

the retro-biosynthesis should be re-linked as monomer-graphs preserving the original skeleton of the molecule. Monomer-based annotation is not only more human readable, but it also enhances the development of analytical tools. For example, substructure and similarity search algorithms are less costly within monomer-graphs than molecular structures. Furthermore, the moieties obtained are often of biological interest and may be related to the peptide activity [125, 126]. Indeed, retro-synthesis has been since long used to detect functionally important structural motifs suitable for novel compounds design in drug discovery [127].

Computationally, the methods for NRPs retro-biosynthesis involve the conversion of the molecular structures into *in silico* graphs where nodes represent atoms and edges are equivalent to bonds. Then, certain chemical substructures are searched within the graphs. Chemical libraries such as CDK [128, 129, 130] provide useful algorithms for substructure searches in chemical structures using patterns defined in SMARTS [131]. The structure targeted in the substructure search differs depending on the retro-biosynthetic approach employed:

(i) **Retro-biosynthesis *via* bonds mapping** (Figure 1.11B). This strategy mainly consists in the identification and fragmentation of the functional groups linking the monomers. For proteogenic peptides, targeting the amide cleavages would be sufficient for a complete model. However, the complexity of NRPs requires the inclusion of additional functional groups often observed in their structures. Ester, thioester, glycosidic, oxazole or thiazole groups should be considered among others. Selecting a suitable list of cleavages that fits the NRP structures is critical for the development of an optimal model. Hence, prior to the list generation, computational analysis should be combined with manual examination of known NRP structures in order to guarantee the selection of the correct bonds. After disconnecting the selected cleavages, the resulting fragments are matched against a monomer database for identification. Most of the software include an internal database of known building blocks for this purpose, but public collections of monomers such as that present in Norine [22] can also be used. Note that during the biosynthesis, new chemical bonds between monomers are formed, and as a result, the monomer residues within the NRP present some structural differences with respect to those original building blocks stored in the database. Thus, certain retro-biosynthesis reactions must be applied to reverse the chemical modifications occurring in the monomers during substrate assembly. Figure 1.10 illustrates some of these reactions. In the case of a peptide bond, for instance, the condensation reaction that occurs during the bond formation produces the loss of a hydroxyl group and a hydrogen in the C- and N- terminus, respec-

tively. To reverse this effect and obtain the “original” building blocks, the mentioned groups have to be re-added in the corresponding positions. Bond mapping is an effective approach of retro-biosynthesis that uses the structural hints given by the chemical structure to map the monomers. However, it is strongly subjected to the established rules (bond cleavages), what may cause problems if they are not well adjusted to the target compounds.

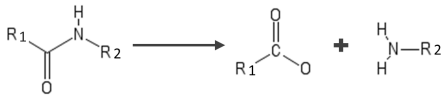
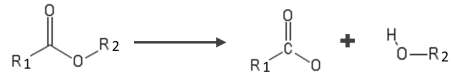
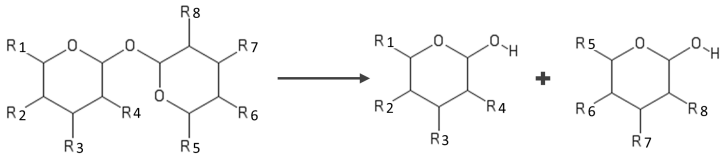
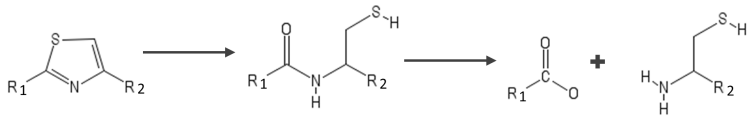
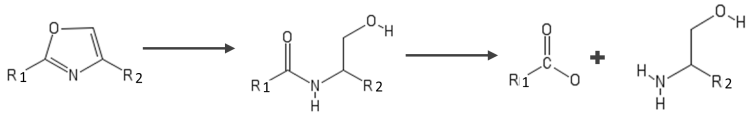
Functional group	Retro-biosynthesis reactions
Amide bonds	
Ester bonds	
Gyclosidic bonds	
Thiazoles	
Oxazoles	

Figure 1.10: Examples of reversed biosynthesis reactions.

(ii) **Retro-biosynthesis *via* monomers mapping** (Figure 1.11A). In this approach the monomers stored in the database are directly mapped within the chemical structure without previous examination of the functional groups contained in the query compound. Contrary to the previous method, chemical modifications are applied to the ”original” monomers and not to residues derived from fragmentation. Hence, the chemical groups that are expected to get lost during bond formation are removed before matching. Note that this approach requires the generation of multiple *residues* per monomer, as the modifications depend on the number of connections formed. For instance, different forms of the same monomer arise if it is placed internally within the structure (loss of H_2O), in the N-terminus (loss of OH) or in the C-terminus (loss of H). The number of possible residues is amplified with the number of potential external links. After computing all possibilities, substructure search is used to map the monomers within the molecules. Some algorithms

prioritize larger fragments [132] while others use combinatorics to find the best fit [12]. The advantage of this approach is that it is not dependent on a set of rules, making it more versatile. However, it is computationally expensive and it omits the information provided by the chemical structure itself. Consequently, the monomers proposed are not always adjusted to the reality.

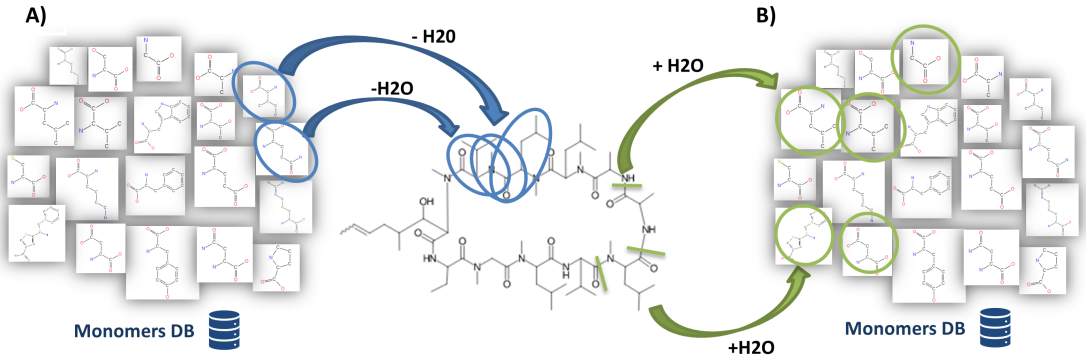


Figure 1.11: Retro-biosynthesis approaches. A) *Via* monomers mapping B) *Via* bonds mapping

CHUCKLES [132] was one of the first algorithms to translate peptide/peptoid chemical structures into monomeric sequences for annotation purposes. The software implements its own monomeric annotation format that emulates the classical protein sequence annotation but amplified for cyclic and branched structures. From the annotation point of view, CHUCKLES stands out for providing both, the *back* (atomic to monomeric) and *forward* (monomeric to atomic) translations. The method relies on a monomer database that can be extended with the addition of new moieties. It relies on the monomer mapping approach previously described and prioritizes the matching of larger monomers before smaller ones without evaluating further solutions. This is rather a simplistic approach that does not always provide the right solution, particularly when working with complex molecules such as NRPs, where the vast amount of monomers and linkages increase the likelihood of matching monomers by chance.

Smiles2Monomers (s2m) [12] was developed for polymers annotation and puts special emphasis on NRPs. Indeed, it is one of the tools integrated in the Norine database and uses its monomers for annotation. Interestingly, s2m provides the results of the retro-biosynthesis in monomeric-graph structures preserving the original linkages of the monomers. The software was tested against chemical structures from two different databases and it annotated most of the compounds. For the retro-biosynthesis, it employs a monomer mapping approach more elaborated than that from CHUCKLES. The process mainly involves two steps: i) mapping of the monomers through substructure search and ii) selection of the best combina-

tion (tiling) that covers the molecule without monomer overlaps. Unfortunately, the combinatorics required for such analysis are computationally expensive and they can only be optimized by searching an approximate solution, which is not always optimal. Another drawback of the approach is that the monomer-graphs generated, miss important information regarding the linkage type and the branch/terminus involved in the connection. This is not a problem for genome mining/drug discovery strategies, but considering that s2m was originally designed for annotation purposes the output graphs should be able to provide the same information than the chemical structures.

GRAPE [11] was presented as a part of a workflow for NRPS/PKS gene cluster analysis. The role of GRAPE in the pipeline is to provide the original building blocks of a given chemical structure (retro-biosynthesis) in order to match them against genome-predicted monomers. The retro-bioynthetic approach applied in GRAPE is *via* bonds mapping. Hence, it deconstructs the chemical structures into the core components by i) cleaving specific functional groups and ii) applying a serie of retro-biosynthesis reactions to the resulting fragments. The predicted components are then matched against an internal database of known building blocks. The results are provided as a sequence of monomers without specifying their linkages in the molecule. This output is adequate for the function of GRAPE in the mentioned application. However, it does not provide enough information to use it for annotation. Similarly to GRAPE, other tools oriented to drug discovery also implement the bond mapping approach. Those include **molBLOCKS** [133], **BRICS** [134] and **RECAP** [127]. However, these approaches are not focused on NRPs neither they attempt to annotate the resulting monomers.

A common feature between all the mentioned tools is the usage of an internal database for monomer annotation. Usually these databases are not large enough to cover all the possible monomers from NRPs, limiting the potential annotations. In chapter 2, I present a retro-biosynthetic approach that solves this problem as well as other limitations from the previously mentioned tools.

1.3 Mass spectrometry of peptidic natural products

1.3.1 Principles, weaknesses and strengths of mass spectrometry

Mass spectrometry (MS) is commonly used for identification and quantification of compounds in fields such as proteomics and metabolomics. Thus, it is not surprising that it became one of the methods of choice for dereplication of natural products [135]. The term *dereplication* is frequently employed in NP research to refer to the identification of known compounds. MS provides means for dereplication through a process that mainly involves the separation of analytes based on their mass to charge ratio (m/z). The three key components of a mass spectrometer are the ionization source, the analyzer and the detector. First, the analytes are ionized choosing the most optimal **ionization source** for the targeted compounds and the experimental set. For PNPs, the usual methods are similar to those used in proteomics, which mainly comprise electrospray ionization (ESI) [136] and matrix-assisted laser desorption/ionization (MALDI) [137], both being soft-ionization techniques and thereby avoiding excessive fragmentation of peptidic compounds. After ionization, the **analyzer** separates the ions on the basis of their m/z. Distinct techniques are employed for the separation depending on the analyzer. The time-of-flight (TOF) uses an electric field to accelerate the ions and measures the time to reach the detector [138]. As heavier ions take longer to arrive, the timing is used to calculate the m/z values of the ions. Other instruments such as the quadrupole mass analyzer [139] and the ion traps [140], use oscillating fields that stabilize/destabilize the trajectory of the ions depending on their m/z values. Finally, the Orbitrap [141] and the Fourier transform ion cyclotron resonance (FTICR) [142] analyzers use a 3-dimensional or a mixed cyclindrical field to place the ions in different orbitals according to their m/z values. Then, the **detection system** measures the image current generated by the ions. The frequencies of the current are directly related to the m/z values of the ions, which are obtained by applying the inverse Fourier transform. Although each instrument has its advantages and disadvantages, the Fourier transform-based mass spectrometeres (FTMS) provide the best mass resolution and accuracy, making them highly interesting for the PNPs analysis [143, 144, 145, 146]. The detection system in the mass spectrometers does not only measure the m/z values but it also records their relative abundance with the aim of providing a spectrum of the sample.

Similarly to proteins, the analysis of PNPs often involves the usage of **tandem mass spectrometry** (MS/MS or MS2). As the name suggests, tandem mass spectrometry consists in coupling two mass analyzers in order to obtain a more exhaustive analysis of the compounds thereby facilitating structural determination. The process includes two subsequent stages of MS analysis connected by a fragmentation step: i) the first analyzer (MS1) separates the ionized analytes by m/z values as previously described, ii) then a specific ion (precursor) is isolated and activated to generate fragments (product ions) that are separated by iii) the second analyzer (MS2) and recorded by the detection system. The resulting MS/MS spectrum represents a *fingerprint* of the precursor that provides an additional layer of information improving the selectivity of the method and the confidence of the identifications. Two MS stages are usually sufficient to obtain enough structural information of the compound analyzed. However, some instruments allow the inclusion of additional stages for further fragmentation (MSn). That is particularly useful to solve ambiguities from the MS2 spectra from cyclic or complex structures [147]. MSn is usually performed with tandem mass spectrometry in time, meaning that instead of using two physically separated mass analyzers in serie (tandem in space), the process takes place in the same instrument but at different times. This configuration can be carried on ion traps and FTICR spectrometers. Typical configurations of tandem in space are the triple quadrupole mass spectrometer (QqQ) and Quadrupole time of flight (QqTOF), where one of the mass analyzers act as a collision cell.

When working with mixtures mass spectrometry is often coupled with **separation methods** such as gas or liquid chromatography (LC), capillary electrophoresis (CE) or ion mobility. These techniques separate the compounds in the mixture facilitating the analysis. The introduction of high pressure in LC columns resulted on a more powerful technique known as high performance liquid chromatography (HPLC). Both, LC and HPLC are commonly used for the analysis of mixtures of natural products and peptides because they optimize the technique and improve identification.

Together with mass spectrometry, the other predominant analytical technique in natural products research is **nuclear magnetic resonance (NMR)** [148]. NMR is mainly used for identification and structural elucidation of compounds, being highly efficient in the determination of complex chemical structures. The main advantages of NMR are its high reproducibility and the minimal requirements of sample preparation [149]. Additionally, it is a non-destructive technique, meaning that the sample is not consumed during the process and can be reused multiple times for further analysis. However, the appearance of chemical shifts between samples due to solvent and pH effects may affect dereplication procedures [150, 151]. The

limited spectral dispersion and the complexity of signal patterns [151] are additional drawbacks that sum up with the inherent low sensitivity of the technique. Improvements in the sensitivity have been achieved with the introduction of high field magnets [152], cryo/micro probes [153, 154] and dynamic nuclear polarization [155], but mass spectrometry is still more sensitive, enabling the detection of secondary metabolites at picomole (pM) and femtomole (fM) levels [149].

The combination of high sensitivity and selectivity offered by **mass spectrometry**, makes of this technique an excellent approach for dereplication of natural products. Although the sample preparation is more demanding than NMR, a small amount of sample (pM-fM) is sufficient to obtain all the structural information stored in a mass spectrum. All the different ionization techniques and mass analyzers available increase the adaptability of the technique enabling the mass spectrometrists to optimize the workflow in accordance to the fragmentation of the target compounds. While this diversity makes it suitable for a wide range of compounds, the lack of common standards for the spectra obtained with different mass analyzers results in a high heterogeneity of the data and complicates the creation of MS/MS databases [151]. Another critical point is the interpretation of the spectra from structural complex molecules such as PNPs. The development of software to support the interpretation and identification of spectra is essential to solve this problem. Different computational strategies exist for the analysis of MS/MS data (Section 1.4), but before introducing them, the principles of the fragmentation process should be understood.

1.3.2 Fragmentation of peptidic natural products

In MS/MS fragmentation, the structural characteristics of the target compounds provide insight into their possible breakages. The peptidic nature of PNPs facilitates the application of the fragmentation rules observed in proteins, mainly characterized by the breakage of the **peptidic backbone**. The nomenclature proposed by Roepstorff and Fohlman [156] and later modified by Biemann [157] has been widely adopted for the annotation of the ion types resulting from the breakage (see Figure 1.12). The labels denote the exact location of the fragmented bond as well as the terminal group that retained the charge. N-terminal fragments are labeled as a, b and c ions while C-terminal fragments are referred as x, y and z types. The annotations also include the number of amino acids in the fragment and are indicated as substcript indices. Although these are the most common breakages, other types of fragments have also been observed. Those include internal and immonium fragments. Internal fragments result from the fragmentation of multiple bonds and

may lack of the characteristic N- and C- terminals. Immonim ions are low mass ions corresponding to single amino acid residues.

Occasionally, some fragments loose certain chemical groups known as **neutral losses**. The most common neutral losses are water and ammonia, observed in the mass spectrum as negative mass shifts of 18.011 Da and 17.027 Da respectively. Water losses normally occur on side chains containing oxygen (Ser, Thr, Asp, etc) and their presence is indicated in the annotation with an asterisk (*b*, y*,a**...). On the other hand, ammonia losses are recurrent in amino acids with nitrogen in the side chain (Arg, Lys, Asn, etc) and are indicated with a *o* super-index (b^o, y^o, a^o...). It should be noted that the patterns described here are typical from proteomics experiments. In natural products MS/MS, neutral losses in fragments without the typical amino acids have also been reported [158, 159]. When mass differences are observed as gains instead of losses, **adducts** are present in the fragment. In some PNPs water adducts are fairly recurrent. Importantly, the masses derived from certain adducts and neutral losses are equivalent to specific ion types. For instance, the loss of CO in a *b* ion has the same mass than the *a* ion from the same fragment and therefore they are not discernible. Further fragmentation may also result in side chain fragmentation (*d,v* and *w* ions), which is particularly useful to differentiate amino acids with the same molecular mass but different side chains such as leucine and isoleucine [160, 161].

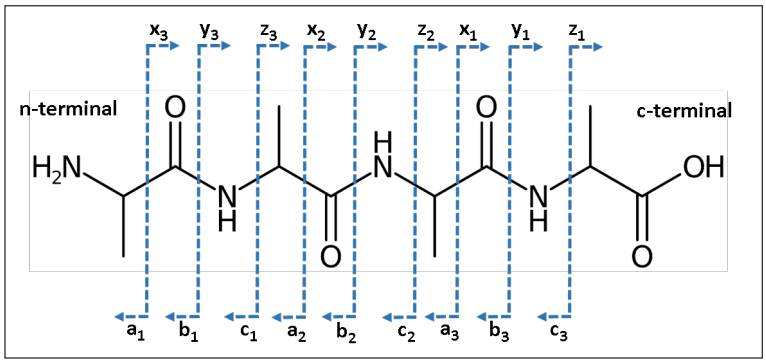


Figure 1.12: Peptides MS/MS fragmentation.

Hypothetically, if PNPs were solely composed of amino acid moieties, the description of their fragmentation would conclude at this point. However, the existence of additional monomers such as glycans extends the possible breakages. The MS/MS **fragmentation of glycans** was exhaustively analyzed by Domon and Costello [162]. Similarly to peptides, the ion types designate the breakage and the remaining terminal groups (see Figure 1.13). In glycans, the A, B and C annotations are used to indicate those fragments that contain the terminal sugar unit (non-reducing end) and the X, Y and Z annotations define the fragments ending with the aglycone (re-

ducing end). Apart from glycans, the presence of fatty acids and depsipeptides also requires to take into consideration the breakage of **ester bonds**. Finally, another aspect to take into account is the non-linearity of the of PNPs, which means that the fragmentation may involve the breakage of multiple bonds.

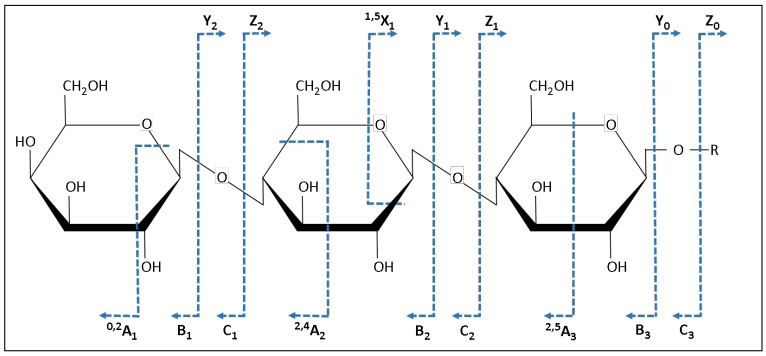


Figure 1.13: Glycans MS/MS fragmentation.

As previously introduced, MS/MS is performed isolating a precursor ion and inducing its fragmentation. Mass spectrometrists dispose of many **activation techniques** to obtain the desired fragmentation. The most employed for the analysis of peptides is collision induced dissociation (CID) [163, 164], which consists in colliding the ions with an inert gas (helium, nitrogen, argon...) to induce their dissociation. Spectra obtained with CID fragmentation are characterized by an abundance of b and y ion types as these are the main ions generated using this approach. a ions are also present with less abundance and neutral losses of ammonia and water are frequently observed [165]. A technique similar to CID but more recently introduced is higher energy collisional dissociation (HCD) [166]. HCD is available in most of the ion trap instruments and performs the fragmentation in a separate collision cell. The higher activation energy of the method results in a predominance of y ions and solves the low m/z values truncation problem observed in ion traps with CID fragmentation that complicates the identification of short fragments [167]. Contrary to CID and HCD, electron-based activation methods such as electron-capture dissociation (ECD) [168] and electron transfer dissociation (ETD) [169] cleave the N-Cα backbone generating c and z ion types. The method employed in both techniques consists in inducing the fragmentation of multiply protonated peptides by low-energy electron donation. ECD and ETD are particularly useful for the localization of PTMs because, contrary to CID, the modifications are still preserved after fragmentation [170]. For glycopeptides, for instance, CID/HCD is useful to identify the glycan, while ETD facilitates the identification of the glycosilation site as it preserves the bond between the glycan and the peptide [171]. Another difference between CID and ECD/ETD, is that the former is more effective for peptides in low charge states while the latter are more optimal for the fragmentation of high

charge state ions [167]. An electron-based technique for singly protonated peptides is electron induced dissociation (EID) [172] that cleaves CC and NC bonds generating *a, x, c* and *z* ions. Although this technique is less used than the previously mentioned methods, it has already been applied for the characterization of some PNPs [173]. In addition to EID, other alternative activation techniques are emerging. Photoactivation-based approaches based on the absorption of photons such as ultraviolet photodissociation (UVPD) [174, 175] have demonstrated to be efficient on the cleavage of disulfide bonds [176] and to promote radical directed dissociation enabling the identification of D amino acids [177, 178]. These structural characteristics are often encountered in PNPs and thereby it is highly promising for the field. Other methods include surface induced dissociation (SID) [179, 180], metastable atom-activated dissociation (MAD) [181], charge transfer dissociation (CTD) [182] or those specialized for deprotonated peptides such as electron detachment dissociation (EDD) [183, 184], negative electron transfer dissociation (NETD) [185] or electron photodetachment (EPD) [186, 187].

1.4 Computational methods for dereplication

1.4.1 Dereplication methods and databases

Dereplicating complex compounds such as PNPs requires elaborated strategies. As exemplified by machine learning approaches, algorithms in bioinformatics are often based on the usage of previous data for future predictions. Similarly, dereplication tools use the spectral/structural data from compounds previously characterized to identify the PNP associated with a given spectra. H. Mohimani and P. A. Pevzner [7] described three main methods for dereplication depending on the data resource: i) dereplication *via* chemical databases ii) dereplication *via* spectral libraries, iii) dereplication *via* spectral networks.

Dereplication *via* chemical databases

The key strategy in software based on chemical databases is the development of *in silico* fragmentation algorithms that virtually fragment the structures retrieved from the database and compare the theoretical fragments with the experimental m/z values. Unfortunately, most of the largest NP databases are commercial (see Table 1.2). From those, the **Dictionary of Natural Products** (DNP) [19] is one of the most complete and curated resources, providing the chemical, physical and structural information from over 230,000 compounds, including peptides. Another

well-known commercial database is **AntiBase** [20], focused on bioactive natural products and including data from more than 40,000 compounds produced by microorganisms and higher fungi. The chemical properties of the compounds are extracted from literature and manually validated, but it has not been updated since 2014 [21]. **MarinLit** [188] is another highly curated database initiated in 1970 in the University of Canterbury (New Zealand). The collecting efforts resulted in a database with more 29,000 marine natural products. Note that AntiMarin [189], a database frequently mentioned in NPs literature, was the merge of AntiBase and MarinLit, but it is no longer accessible [21]. Other commercial databases include **SciFinder** [190], **RÖMPP Online** [191] or **Reaxys** [192].

When working with public resources, the use of databases dedicated to general chemical compounds or metabolites is acceptable although it should optimally be accompanied with the application of taxonomic filters to isolate NPs. One of the largest freely available database for chemical compounds is **PubChem** [193]. Despite its large size, including 103 millions of compounds, the estimated amount of NPs is reduced to approximately 3,500 entries. Nonetheless, this amount is probably underestimated due to the difficulty of identifying NPs using the labeling provided by the database. **ChemSpider** [194] is another prominent database with 34M of structures, but similarly to PubChem, the numbers significantly decrease when filtering the NPs, resulting in about 9,700 NPs. An interesting resource for dereplication is **ChEBI** [195], as it focuses on chemical entities of biological interest and includes over 15,000 easily identifiable and downloadable NPs. The identification of NPs is also simplified using **ZINC** [196], a database for virtual screening that contains 230M of commercially available compounds clearly classified in catalogs. The subset of NPs in this database spans over 85,000 structures.

Public databases specialized in natural products also exists, but many of them have issues such as the lack of manual curation, limited or inexistent download options, metadata or maintenance [21, 8]. **Supernatural II** [197], for instance, is an extensive database with more than 300,000 NPs and their 2D structures, physio-chemical properties and predicted toxicity. However, the database is not downloadable neither provides programmatic access, preventing its coupling with dereplication software. Plus, accuracy problems claiming the presence of non-natural products have been reported [21]. Another large NPs database with more than 200,000 compounds is **UNPD** (Universal Natural Products Database) [198]. The main problem of this resource is that it is no longer accessible using the original link provided in its publication. Alternatively, a copy of the chemical structures is still reachable through ISDB (In-silico MS/MS DataBase) [199]. Other resources with lower compounds coverage provide higher level of curation and metadata. **NPASS** (Natural

Product Activity and Species Source) [200] integrates the structural and chemical properties of more than 30,000 compounds, together with their experimentally-determined activity data and source organisms. NPs can be browsed according to chemical taxonomy, revealing the presence of 356 potential PNPs in the category of *peptidomimetics*. Another interesting database that has recently changed license to become open access is **The Natural Products Atlas** (NPAtlas) [201]. NPAtlas particularly suits PNP dereplication because it is focused on fungi and bacteria, the major producers of biologically relevant PNPs. It provides more than 20,000 NPs with extensive metadata including the structure, source organism, isolation and total syntheses. Additionally, NPAtlas was developed following the FAIR principles (Findable, Accessible, Interoperable and Reusable) and is connected to other NP resources such as MIBiG and GNPS. Other open-access databases are **NPEdia** (Natural Products Encyclopedia) [202], focused on NPs from plants and microoganisms, and **3DMET** [203], that puts special attention on the 3D structures of NPs. However, non of these resources are downloadable. Importantly, a large part of NP databases are specialized on the source organism or the geographic location of the compounds. Examples of these resources are **KNApSaCK** [204], dedicated to plants NPs; **StreptomeDB** [205], which provides a collection of NPs from Streptomyces genus (known for being the source of many bioactive compounds); **AfroDB** [206], containing NPs from African medicinal plants; or the **NuBBEDB** [207] (Nuclei of Bioassays, Ecophysiology and Biosynthesis of Natural Products Database), focused on Brazil-isolated NPs.

Note that all the mentioned resources are either focused on chemical compounds, metabolites or general/specific NPs, but not particularly on PNPs. Currently the only database devoted to natural products of peptidic origin and concretely NRPs is **Norine** [22]. Although Norine is highly interesting for its manually curated data and its specialization on peptides, the small size of the database (1730 PNPs) prevents its usage for dereplication unless complemented with additional databases. Ideally, a complete collection for dereplication should include PNP entries from multiple databases. Tools for automated chemical classification such as ClassyFire [208] can be used to filter out non peptidic records. Alternatively, retro-bioynthetic software can also be used for filtering because they provide the monomeric composition of the compounds, which is useful for classification.

Dereplication *via* spectral libraries

Dereplication *via* spectral libraries involves the comparison of the spectra stored in the database with the query spectra. Similarly to the previously described structural databases, commercial spectral libraries also play an important role in NP

Database	#NPs	Open access	Easily accessible[1]	Description
SciFinder	>300,000	✗	-	Chemical compounds database
Supernatural II	>300,000	✓	✗	General NPs database
DNP	>230,000	✗	✓	General NPs database
Reaxys	>220,000	✗	-	Chemical compounds database
UNPD	170,602	✓	✓	General NPs database
ZINC	85,198	✓	✓	Commercially available chemicals database
AntiBase	>40,000	✗	✓	Bioactive NPs database
NPASS	>30,858	✓	✓	General NPs database
MarinLit	>29,000	✗	✓	Marine NPs database
NPAtlas	>20,035	✓	✓	Bacteria and fungi NPs database
3DMET	18,248	✓	✗	NPs database focused on 3D structures
NPEdia	18,016	✓	✗	General NPs database
ChEBI	15,736	✓	✓	*Small* chemical compounds database
KNApSaCK	10,265	✓	✗	Plant metabolites database
ChemSpider	9,732	✓	✓	Chemical compounds database
StreptomeDB	6,415	✓	✓	NPs produced by streptomycetes
PubChem	3,529	✓	✓	Chemical compounds database
NuBBEDB	2,215	✓	✓	NPs from Brazil
AfroDB	954	✓	✓	NPs from African medicinal plants

Table 1.2: Chemical databases for PNPs dereplication. Most of the information in this table was retrieved from the study performed by M. Sorokina and C. Steinbeck [21].

dereplication (Table 1.3). This is exemplified by the high level of manual curation and data quality offered by **NIST** (National Insitute of Standards and Technology) [209]. NIST17 contains a subset of tandem spectra specially dedicated to biologically active peptides, which comprehends 90,244 spectra of 6,803 precursor ions from 1,904 peptides. The Wiley Registry of Tandem Mass Spectral data, **MsforID** [210], also provides a collection of spectra generated with electron ionization (EI). In 2016, this spectral library gathered 20,000 spectra from 1,200 compounds including drugs, pesticides and other small metabolites.

Regarding open-access spectral libraries, **MassBank** [211] is a well-established database that collects EI, MS/MS and MSn data from small chemical compounds (<3,000Da). The resource is distributed in different platforms including the Massbank of North America (MoNa) [212], the European Massbank [213] and the Japanese MSSJ Massbank [214]. Among them, **MoNa** is the platform integrating data from more sources and it currently contains over 200,000 spectra. It consists of a centralized repository based on crowd-sourcing and auto-curation. Interestingly, dereplication is favored by the provision of programmatic access and downloading options (JSON and MSP). Aside from Massbank resources, another known database is **METLIN** [215]. Created and regularly maintained since 2003, METLIN is the most extensive database for QTOF MS/MS spectra. It is mainly focused on metabolites and it contains spectral data from more than 22,000 compounds. However,

it does not provide bulk download options. Another worth-mentioning resource is **mzCloud** [216], particularly useful for the analysis of of MSn spectra. Currently, it contains MS/MS data from 9,829 compounds arranged into spectral trees. The accessibility to the database is a bit restrictive as it either requires a desktop installation or the usage of Safari or Internet Explorer. Plus, it is not compatible with Linux machines.

Publicly available databases with focus on NPs exist as well. The **Global Natural Product Social Molecular Networking** (GNPS) [217] was initiated in 2016 and it rapidly became a popular resource for dereplication of NPs. GNPS is an open-access knowledge based spectral library for sharing MS/MS data within the natural products community. It currently provides spectra for 12,694 NPs as well as their chemical structures and a set of MS/MS analysis tools. Instead of providing experimental MS/MS, other resources such as **ISDB** offer *in silico* predicted spectra. ISDB [199] contains about 170,000 MS/MS predictions from the NPs in UNPD. Finally, spectral libraries dedicated to certain organisms are also available online. Those include **ReSpect** (RIKEN MSn spectral database for phytochemical) [218], for plant metabolites; **HMDB** (Human Metabolome Database) [219], focused on the human metabolome; and **YMDB** (Yeast Metabolome Database) [220], specialized on small metabolites produced by *Saccharomyces cerevisiae*.

Database	#Compounds	Open access	Easily accessible[1]	Description
UNPD-ISDB	170,602	✓	✓	NPs *in silico* predicted spectra
MoNA	>80,000	✓	✓	Autocurated repository of spectra
NIST17	>30,000	✗	✓	Highly curated database of chemicals
METLIN	>22,000	✓	✗	Relevant for QTOF MS/MS spectra
Massbank	14,382	✓	✓	Community spectral database
GNPS	12,694	✓	✓	Community NPs knowledge base
mzCloud	9,829	✓	✗	MS/MS spectra arranged in trees
ResPect	9,017	✓	✓	MSn spectra for plant metabolomics
HMDB	2,265	✓	✓	Metabolites of the human body
MSforID	1,200	✗	✓	Small molecules MS/MS spectra

Table 1.3: Spectral libraries for PNPs dereplication. The number of compounds comprehend those with MS2 spectra. Note that not all the databases provide statistics of the resource. For those cases the number of compounds was retrieved from literature and may not be up-to-date.

Dereplication *via* spectral libraries is particularly advantageous for the identification of compounds presenting unusual breakages difficult to predict using *in silico* fragmentation models. It is more accurate than dereplication *via* chemical databases. However, compound coverage is limited to those previously ana-

[1]The criteria used to classify a resource as *easily accessible* is that it provides programmatic access to the data or bulk download options.

lyzed, annotated and stored in the spectral libraries, which are often smaller than structural databases. Note that tables 1.2 and 1.3 are not comparable in terms of database size as Table 1.2 is more specific, providing the number of NPs and not the total number of compounds. Probably, the proportion of NPs in spectral libraries is rather small, as reflected in the resources providing such information. For instance, mzCloud reports just 1740 NPs or medicines within its compounds. That is still an acceptable amount when compared with the quantity of PNPs. MoNA exhibits the results of applying ClassyFire to the database compounds and the category of *peptidomimetics* solely includes 14 records. Mohimani and PA Pevzner [7], already mentioned such issue by reporting the presence of only 81 PNPs in a set of 1607 annotated spectra from GNPS. With the aim of solving these limitations, the storage of theoretical MS/MS is becoming popular, either for the creation of *in silico* predicted spectral libraries (ISDB) or for the extension of standard libraries, an approach used by METLIN and HMDB.

Dereplication *via* spectral networks

Dereplication *via* spectral networks complements the *classical* approaches previously described in order to provide additional information of the identifications. This approach has gained popularity due to its potential for variant dereplication (identification of variants from known compounds). The method consists in matching and scoring all the target MS/MS spectra with each other. Then, a network is created where nodes are MS/MS spectra linked by edges according to their similarity scores. Clusters grouping similar spectra are formed. Taking into consideration that multiple NP variants are usually produced by the same source organism, these clusters may represent compounds from the same family. Some of the nodes can be annotated using standard dereplication methods. Then, the spectral connections between annotated and unannotated nodes are used to dereplicate unidentified spectra. Several aspects can be evaluated when comparing connected nodes, but the mass shifts between them usually provide valuable information about the potential modification distinguishing the two compounds. Note that the key point of this method is the production of an overall view of all the identifications. Thereby, spectral networks are particularly useful when working with extensive data sets and compounds with common origins. Unfortunately, large collections of PNP MS/MS data are difficult to obtain, complicating the implementation of this approach.

1.4.2 Molecular formula identification

Identifying the molecular formula of the precursor ion is not strictly necessary for dereplication. However, it provides important information that can be used to discard some candidate structures within the mass range tolerance. Various tools and methods for molecular formula identification have been developed over the years [221, 14, 222]. Classical strategies usually require a previous step known as mass decomposition. Decomposing consists in calculating all the molecular formulas that match with the monoisotopic peak mass using a limited set of chemical elements. As the number of possible formula can be quite high, especially for compounds above 500 Da, a secondary step is required. The selection of the "best" candidate is usually performed i) via isotope patterns or ii) using the *Seven Golden Rules*. Besides these classical methods, in this section, I also present an approach based on the Kendrick mass defect (KMD). KMD-based formula identification is more commonly used in petroleomics, but in Chapter 3 this method was adapted for PNP analysis.

Formula identification *via* isotope patterns

Most of the chemical elements in nature have multiple isotopes that vary in mass and abundance. Carbon-12, for instance, has a natural abundance between 98.84-99.04% while the abundance range of carbon-13 is between 0.96-1.16% [223]. In mass spectrometry, the monoisotopic mass of a molecular formula is calculated using the masses of the most abundant isotope of each element. When examining a mass spectra, monoisotopic peaks are often followed by other peaks separated by a distance of 1 Da. Frequently, these peaks correspond to the same fragment ion but with different isotope variants and together with the monoisotopic peak conform what is known as *isotope pattern*. The shape (intensities) of an isotope pattern mainly depends on the number of elements and their relative abundances. For example, a large compound with a high content of carbons, has more chances to contain a carbon-13 than a small compound with less carbons. Hence, given a molecular formula it is possible to predict the expected isotope pattern. Calculating isotope patterns consists in computing the expected m/z values and intensities of a given molecular formula taking into consideration the chemical elements present in the composition, their amount, and the natural abundances of their isotope variants. Software such as Emass [15], SIRIUS [14], BRAIN [16] and Fourier [17] provide isotope pattern simulation. As different molecular formulas present different isotope patterns, the comparison of the predicted and measured patterns provides means for the identification of the right candidate (see Figure 1.14). Thus, identifying molecular formula *via* isotope patterns consists in i) calculating the theoretical

isotope patterns of all the candidates obtained by decomposition, and ii) ranking them in accordance to their similarity with the experimental isotope pattern. The candidate/s showing the highest similarity are proposed as the potential molecular formulas of the compound.

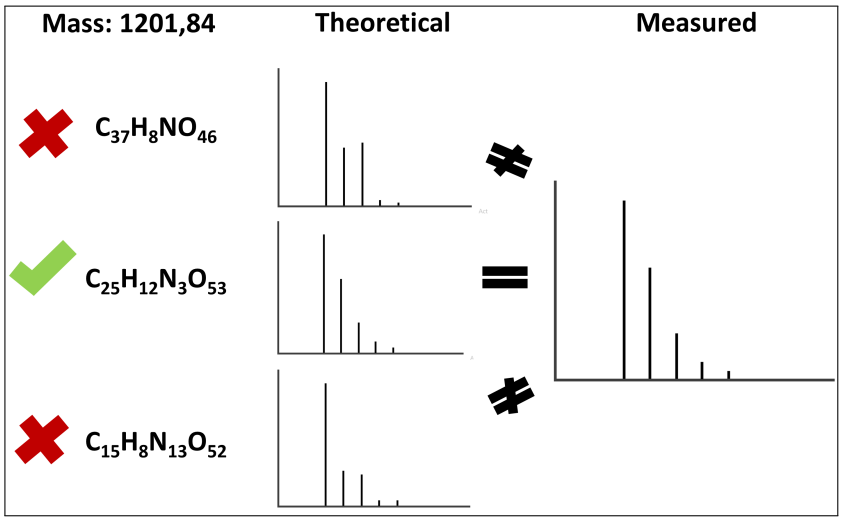


Figure 1.14: Simulation of isotope patterns for molecular formula assignment.

The *Seven Golden Rules*

The *Seven Golden Rules* is a method proposed by Fiehn and Kind [222] to filter out molecular formula candidates. According to this method, predicted chemical formula should always follow these seven rules and those instances not complying should be discarded. The rules are enumerated as follows:

1. *Restrictions for element numbers*. Each type of chemical element is limited by a maximum value depending on the element and the molecular formula mass. For instance, a maximum of 126 hydrogen are allowed in molecular formulas within the mass range of 500 to 1000 Da.
2. *LEWIS and SENIOR check*. Check that the molecular formula obeys LEWIS and SENIOR rules.
3. *Isotopic pattern filter*.
4. *Hydrogen / Carbon element ratio check*. The ratios of hydrogen/carbon atoms have to be within the common range. The 99.7% of the compounds have ratios H/C within 0.2-3.1.
5. *Heteroatom ratio check*. The ratios between carbon and other hetero-atoms (N, O, P, S, Br, Cl, F, Si) have to be within the common ranges as well.

6. *Element probability check*. Despite passing the heteroatom ratio check, multiple heavy element counts can still be present in the formula. The authors provide numerical restrictions of the elements NOPS to remove unlikely formula such as $C_{26}H_{28}N_{17}OP_3S_8$.

7. *TMS (trimethylsilyl) check*. If chemical derivatization has been applied, TMS groups (C_3H_8Si) have to be removed from the formula.

Specific compound classes may not follow some of these rules. Thereby, the software developed with the *Seven Golden Rules* allows the exclusion of certain rules such as the LEWIS or H/C ratio check. Note that all the heuristic values and ratios are specified in tables of the original publication [222].

The Kendrick Mass Defect

As established by IUPAC, the mass scale of the chemical elements takes as reference carbon-12, with an atomic mass of exactly 12 Da and zero mass defect [224]. In 1963, Edward Kendrick proposed a new mass scale based on CH_2 instead of C in order to facilitate the discrimination of homologous compound series characterized by CH_2 repetitions [225]. The approach consists in the calculation of the Kendrick mass (KM) and Kendrick mass defect (KMD) as follows:

$$KM = IUPAC\ mass \times \frac{14}{14.01565} \tag{1.1}$$

$$Nominal\ Kendrick\ Mass\ (NKM) = round(KM) \tag{1.2}$$

$$KMD = NKM - KM \tag{1.3}$$

The key point of these calculations lies in the fact that the KMD value (– 0.5 < KMD < + 0.5) is identical for compounds belonging to homologous series that only differ in the number of base units. This is easily exemplified with hydrocarbon compounds as they are mainly composed of CH_2 units. As shown in Table 1.4, all the selected compounds have the same KMD value of -0.0134.

Applied in mass spectrometry, these calculations can be used for vector-based molecular formula prediction using KMD plots. KMD plots consists of 2D-representations of KMD (y-axis) as a function of NKM (x-axis). Each point in the plot represents a molecular formula and they are lined-up horizontally when originating from homologous series (see Figure 1.15). The closest points in these horizontal rows are

	KM	NKM	KMD
Heptane (C_7H_{16})	100.0134	100	-0.0134
Octane (C_8H_{18})	114.0134	114	-0.0134
Nonane(C_9H_{20})	128.0134	128	-0.0134

Table 1.4: KM, NKM and KMD values of hydrocarbon compound series.

separated by values of 14, corresponding to the NKM of CH_2. Using this logic, only the molecular formula of a single point needs to be known in order to predict the others by simply calculating the point distances. A chemical database of the targeted compounds can be used to generate KMD plots that serve as a basis of the prediction.

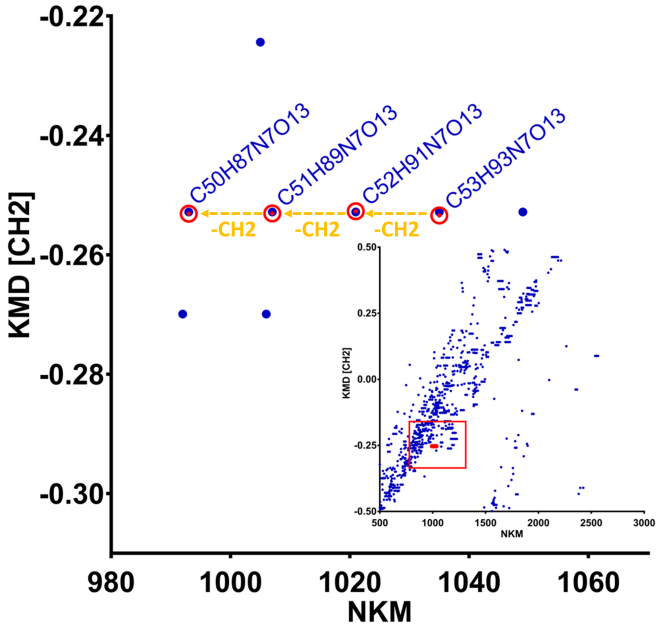


Figure 1.15: Kendrick Mass Defect plot. Compounds differing by CH_2 form an horizontal row. Edited from Chapter 3.

Due to the high amount of CH_2 moieties in petroleum components, this approach is particularly used in petroleomics [226]. However, Sato et al. proposed a modification of the KM formula in order to apply it in other fields, in their case for polymers [227]. It consists in using the mass of a different unit of interest (R) in accordance to the repeating moietes found in the target compounds (Equation 1.4). The equation 1.5 shows an example of the KM using C_3H_6O as the unit of interest.

$$KM(R) = observed\ mass \times \frac{nominal\ mass\,(R)}{exact\ mass\,(R)} \tag{1.4}$$

$$KM(C_3H_6O) = observed\ mass \times \frac{58}{58.04187} \tag{1.5}$$

In Chapter 3, I present a bioinformatic tool for the prediction of molecular formula using KM defect. To our knowledge, it is the first tool applying this approach for NRPs formula identification.

1.4.3 *In silico* fragmentation methods

Theoretical fragmentation is a crucial step for MS/MS identification *via* chemical databases (Section 1.4.1). The *in silico* fragmentation of database candidate structures supports the identification of the compound that fits the best with the query spectra. The two main strategies for PNPs *in silico* fragmentation are: i) monomer-based fragmentation and ii) rule-based fragmentation. One interesting implementation of these methods involves the generation of fragmentation graphs that can be used as well for scoring purposes. In addition to these strategies, in the following paragraphs I also introduce some of the techniques used for dereplication of small metabolites in order to get a general overview of the existing methodologies.

PNP fragmentation methods

Monomer-based fragmentation mainly differs from other methods by using monomer sequences instead of chemical structures to predict the expected fragment ions and their corresponding masses. This approach is widely used in proteomics, as the linearity and composition of proteinogenic peptides reduces the fragmentation algorithm to a calculation of the fragment ions obtained from iterative cuts of the amino acid sequence. However, and despite the shared amino acid composition of PNPs and proteinogenic peptides, PNPs present additional structural characteristics that increase the complexity of the approach. Two data requirements are essential for the implementation of this method:

(i) A *monomer database* with their respective masses. Apart from the 20 proteinogenic amino acids, more than 504 PNP monomers have been reported so far and should be taken into consideration for the monomeric fragmentation. Databases containing monomer collections include Norine, HELM or KEGG. However, these resources do not necessarily cover the vast diversity of PNP moieties, likely missing some monomers. Thus, manual introduction of missing entities should not be discarded.

(ii) A *monomer-graph database*. The presence of cycles and other complex structures in PNPs impossibilities the usage of sequence data such as that employed in proteomics. An additional layer of information providing the structural skeleton of the molecule is required instead. Graphs are suitable to serve this purpose as they can easily represent the linkages between monomers. Thus, a graph for monomer fragmentation should include a set of nodes representing the monomers and a set of edges representing their linking bonds. Ideally, the edges should be i) labeled according to their bond type (peptide, ester…) and ii) directed to indicate the source/target monomers in reference to structural characteristics such as the monomer containing the carbonyl group and the monomer containing the nitrogen (see an example of a monomer-graph in Figure 1.16B). Knowing these properties is important in order to apply the expected hydrogen rearrangements occurring during the breakage. Such graph structures can be generated from chemical data (SMILES) using retro-biosynthesis tools such as rBAN (Chapter 2). Although Smiles2Monomers [12] also provides graphs, the lack of labeled/directed edges makes it less optimal for in silico MS/MS fragmentation.

After creating these resources, the fragmentation itself can be implemented. The process involves iteratively disconnecting the edges of the candidate monomer-graphs to simulate the breakages during the fragmentation. The masses of the monomers are used to calculate the expected m/z values. A potential problem associated with the monomer-based strategy is related to the possible presence of internal peptide bonds -or other breakable bonds- inside the monomers. That is the case, for instance, of N-acetylated (Ac-Aib) or N-formylated (NFo-Gln) monomers from Norine. Fragmentation of compounds containing these monomers would be underestimated as these internal cleavages would not be considered. This problem is fixable by associating specific neutral losses/adducts with certain monomers, but it can be rather complicated and increases the already extensive data pre-processing step.

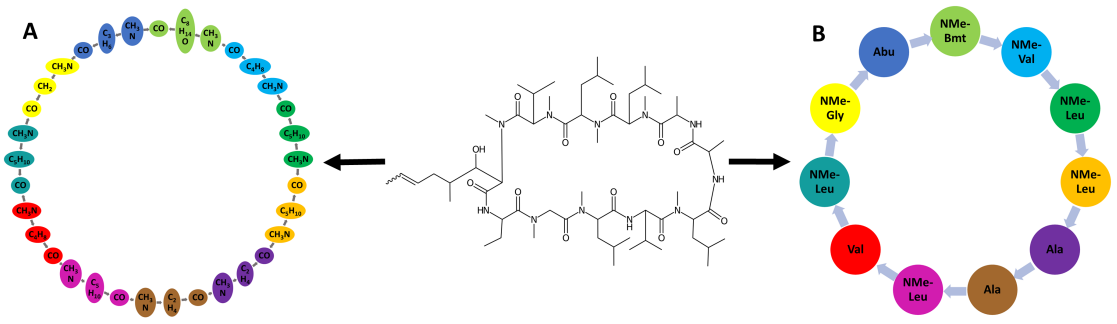


Figure 1.16: Rule-based graph (A) and monomer-based graph (B). Example with Cyclosporin A.

Despite the disadvantages of this technique, it is less computationally expensive than other methods such as combinatorial fragmentation, later covered in this section. Plus, targeting specific bonds avoids over-fragmentation. One of the most interesting points of this approach is that it facilitates the automatic peak annotation using monomers. Many experts prefer monomeric annotation over chemical formula as it provides further structural information, is more human-readable and easier to associate with the part of the molecule fragmented. Hence, monomer-based fragmentation is particularly useful for annotation tools where manually examination of spectra is expected. Indeed, known software implementing automatic peak annotation such as mMass [158] or Cyclobranch [228] employ this approach. Going one step further, the provision of monomeric sequences facilitates the coupling of the dereplication results with genome mining tools for cluster identification.

Rule-based fragmentation consists in cleaving specific bonds that are likely to fragment according to known fragmentation rules of the target compounds. Thus, it is particularly suitable for compounds with characteristic fragmentation patterns previously reported. As already described, the monomeric composition of PNPs includes subunits of well-studied compound classes such as proteogenic peptides, lipids and glycans. The knowledge in these fields can be used for the construction of an optimal rule-based fragmentation model for PNPs. Although rules can always be extended, the fragmentation should at least contain peptide, ester and glycosidic cleavages. Computationally, mapping these cleavages requires the previous conversion of the compounds into chemical graphs. Then, substructure search is employed to identify the patterns associated with the targeted bonds within the chemical structure. Note that, despite pursuing different goals, the mapping algorithms are similar to those employed in retro-biosynthesis *via* bond fragmentation (Section 1.2.2). Chemical libraries such as CDK [128, 130] may help for the implementation of these algorithms as they provide useful code for substructure search using SMARTS annotation as well as tools for the conversion of chemical structures into chemical graphs.

Substructure search is computationally expensive. To avoid its execution *on the fly*, structures can be pre-processed in order to generate *ready-to-use* graphs. Computing these graphs involves breaking the chemical structures through all the targeted bonds, generating the molecular formula of the resulting fragments and creating a new graph preserving the original linkages (see example of a rule-based graph in Figure 1.16A). Then, the new graph only contains the potential cleavages (edges) and the molecular formula between them (nodes). This strategy avoids repeating substructure search, removes the redundant (non-targeted) bonds and speeds up

the process. Similarly to monomeric-graphs, introducing the labeling and direction of the edges would strongly benefit the fragmentation allowing the calculation of hydrogen rearrangements.

The main advantage of this method is that has proven to be highly effective to simulate the fragmentation of peptidic compounds such as PNPs. Indeed, some of the most used dereplication software (iSNAP [229], Dereplicator [230], Dereplicator+ [231]) employ rule-based fragmentation. Furthermore, compared to monomer-based fragmentation, it does not present problems associated with the skipping of internal monomer bonds and is less demanding in terms of data resources and pre-processing. On the other hand, this technique does not provide the benefits of monomer annotations. Plus, as it is based on a previously established set of rules, it is challenged by compounds with unusual fragmentation patterns. Nonetheless, this problem can be mitigated over time by the addition of new rules.

Regardless of the fragmentation method used, when working with graphs, a common computational method to simulate MSn fragmentation is the generation of **fragmentation graphs**. A fragmentation graph consists in a hierarchical structure of fragments equivalent to that expected from MSn experiments. Each child fragment in the tree results from disconnecting one of the edges from its parent fragment. That generates all the fragments of the molecule in a structure whose levels are equivalent to the number of disconnections with respect to the original graph (Figure 1.17). Depth limits are established to avoid excessive fragmentation that would only increase the computational expense and generate unlikely fragments. For instance, in MS/MS CID fragmentation, fragment ions of more than three cuts are rare. Interestingly, apart from providing means of fragmentation, the hierarchical information obtained with fragmentation graph can be used for scoring purposes. In Dereplicator+, Mohimani et al. used fragmentation graphs to match the spectra against the tree nodes but only considering those fragments whose parent had already been observed in the spectra. Hence, this approach assumes that a match is more reliable when its presence is supported by further evidence (parent fragments). Additionally, in the same study, fragmentation graphs were also used for the generation of a decoy database.

Other fragmentation methods in metabolomics

Unlike PNPs, the lack of building blocks defining the fragmentation of small metabolites has led to the development of complex fragmentation models. These models could be equally applicable to PNPs with unusual fragmentation patterns and that is why I present them in the following paragraphs.

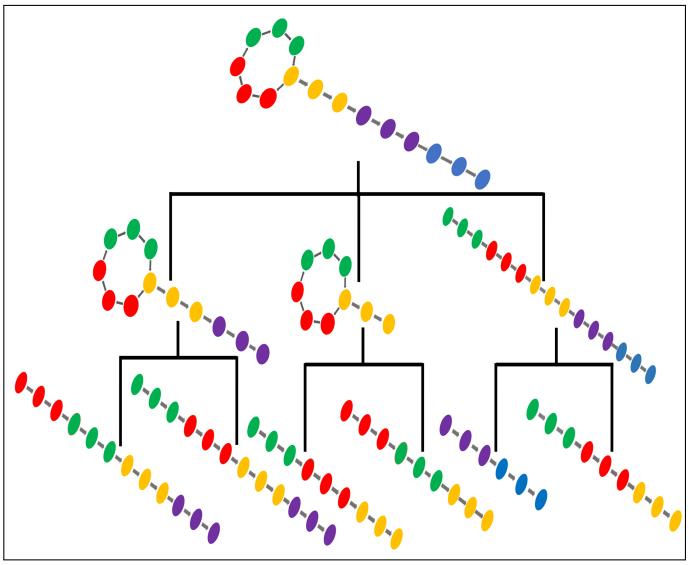


Figure 1.17: Illustration of a fragmentation graph.

Combinatorial fragmentation relies on systematically breaking all the bonds of the target molecule in order to enumerate the whole range of possible combinations that may explain the spectra. Cost functions are used to assign different scores to the cleavages in accordance to their "breakability". Properties such as the type of bond, standard bond energies or bond-dissociation energies are used for the implementation of cost functions. Then, peak annotation is performed prioritizing the fragments with minimal cost. The main problem of this approach is the difficulty of creating realistic cost functions. Even highly elaborated functions may be less efficient than other simpler approaches [232]. Non-optimal cost functions together with the vast amount of combinations provided by this approach can lead to wrong peak annotations, reducing the precision of the method. This problem and the computational expense of combinatorial fragmentation are particularly amplified with large PNPs. A heuristic method presenting a fast implementation of combinatorial fragmentation is Metfrag [233].

Another fragmentation strategy commonly used in metabolomics is the generation of **fragmentation trees**. Fragmentation trees were firstly introduced by Böcker and Rasch in 2008 [234]. Fragmentation trees are hierarchical representations of the molecular formulas annotated in the query spectra. Each node contains a molecular formula explaining a peak in the spectrum and the edges are the neutral losses between them. The precursor ion formula is either obtained using isotopic patterns or computing a fragmentation tree for each possible molecular formula. Combinatorial optimization is used to identify the tree that best explains the query spectra. Then, the alignment of fragmentation trees and the application of similarity score functions is useful for clustering similar compounds. An example of

software that uses fragmentation trees to search analogous compounds is SIRIUS. Unfortunately, the number of fragmentation trees that can explain a spectra increases significantly with the size of the compounds. It has been already reported that this problem is specially notable with compounds above 500 Da [232], complicating its implementation with certain PNPs. Note that fragmentation trees should not be confused with the fragmentation graphs mentioned above.

Machine learning can also be applied to the prediction of structural features. For the development of machine-learning methods, large spectral sets of reference compound are used to train spectral classifiers. For each reference compound, two data structures are provided to the spectral classifier: i) a feature vector and ii) the expected responses. The feature vector is a set of numerical features characterizing the spectra that may be useful for the prediction. Those could include the highest intensity peak, the precursor, the number of peaks, etc. The expected responses would consist in a series of yes/no answers regarding chemical properties and substructures of the compound. For instance, whether it contains or not an aromatic ring, an amine group, etc. Note that scoring or probabilities can also be used instead of yes/no answers. Multiple machine learning classification methods can serve this purpose, including regression methods, neural networks or random forest [232]. Once the model has been constructed, processing a query spectra should only involve the generation of a feature vector equal to those employed in the learning step. Using this vector, the classifier should be able to provide a fingerprint of the target molecule defining its substructures and chemical properties. This approach was first described by Venkataraghavan, McLafferty and van Lear in 1969 [235]. Since then, many studies are based on this idea. Heinonen et al [236], for instance, developed a method using fingerprint prediction to identify the molecular structure of the compounds by chemical database search. Machine learning could certainly be applied for PNPs dereplication. However, it requires large amounts of training data exclusively composed of reliably annotated PNP spectra. Such collection is extremely difficult to obtain. In fact, to our knowledge, it did not exist until the recent publication of NPS [237], a scoring system for PNPs based on a set of MS/MS data with the mentioned characteristics. More details about this scoring are given in the following section.

1.4.4 Dereplication software and their scoring/statistical evaluation strategies

Once the theoretical fragments of the PNP candidates have been computed, the identification of the “correct” peptide-spectrum matches (PSMs) involves i) the as-

signment of similarity scores and ii) the statistical evaluation of the scores (p-values). Multiple identification strategies have been adopted by different dereplication software. They all share similar principles with the classical methods employed in proteomics. In this section, I present software for dereplication of PNPs together with their scoring and statistical significance implementations. But first, I briefly introduce some basic concepts about scoring/statistical significance assignment in proteomics.

Proteomics background

In proteomics, many raw scoring functions such as the dot-product, XCorr [238], HyperScore [239], or Andromeda [240], have been developed to measure the similarity between experimental and theoretical MS/MS values. Among them, the dot-product (DP) score represents a simple and fast approach adequate for most of the spectra. The method simply consists in converting the theoretical values and the experimental spectrum into vectors and computing their dot product (see Equation 1.6). Sometimes, it is also normalized by calculating the cosine similarity between the two vectors (Equation 1.7). Note that when theoretical intensities are not calculated (*bar-code* theoretical spectra), the dot product is equivalent to the sum of the matched experimental intensities. Another common and simple approach that does not consider theoretical intensities is the shared peak count (SPC) function. As described by the name, it represents the number of shared peaks between the theoretical and experimental spectra.

$$DP(\vec{A},\vec{B}) = \vec{A}\cdot\vec{B} = \sum_{i=1}^{n} \vec{a}_i \vec{b}_i \tag{1.6}$$

$$\cos\theta(\vec{A},\vec{B}) = \frac{\vec{A}\cdot\vec{B}}{\|\vec{A}\|\,\|\vec{B}\|} = \frac{\sum_{i=1}^{n} \vec{a}_i \vec{b}_i}{\sqrt{\sum_{i=1}^{n} \vec{a}_i^2 \sum_{i=1}^{n} \vec{b}_i^2}} \tag{1.7}$$

where A is the vector of theoretical intensities and B the vector of experimental intensities.

Although scoring gives insight into the resemblance between the theoretical and experimental values, it does not provide enough information to decide whether the match is correct or not. In order to determine the probability of having a correct match, a statistical significance have to be assigned to the score by calculating its p-value. A p-value of a score S defines the probability of obtaining a score equal

or greater than S by chance. A widely used technique for p-value calculation in proteomics is the target-decoy approach (TDA) [241]. As its name suggests, the TDA requires the usage of a target and a decoy databases. The target database refers to the source containing the target sequences. The decoy database is used to know the possibilities of obtaining matches by chance and it serves as a model of the null hypothesis. In proteomics, decoy databases are constructed by reversing or shuffling the amino-acid sequences from the original (target) database [241]. Thus, the sequences stored in the decoy database are not expected to be identified and they will represent incorrect matches. Once the decoy database has been generated, getting the p-value of a score S simply involves matching the spectra against the decoy and calculating the percentage of decoy PSMs with an equal or greater score than S.

P-value calculation provides the statistical significance at a single-identification level. However, when working with large data sets, the likelihood of obtaining incorrect identifications increases. For a group of identifications, it is more interesting to estimate the overall proportion of wrong identifications by multiple testing correction. A well-known method for multiple testing correction is the estimation of the false discovery rate (FDR) [242]. Given a S score threshold, the FDR is calculated as the ratio between the number of decoy PSMs and target PSMs with score equal or above S (Equation 1.8). For instance, if establishing a threshold of 3 the number of decoy PSMs is 26 and the number of target PSMs is 887, we would have a 2.9% FDR.

$$FDR = \frac{\#\,decoy\,PSMs}{\#\,target\,PSMs} \tag{1.8}$$

PNP dereplication software

In 2012, Ibrahim et al. presented an informatics search algorithm for natural products (iSNAP) discovery [229]. **iSNAP** implements "dereplication *via* chemical databases" (see Section 1.4.1) using an internal database of 1107 NRPs which includes chemical structures from Norine [22], PubChem [193], the *Journal of Antibiotics*, the *Journal of Natural Products* and the KEGG peptide databases [243]. The method described in the publication performs the cleavage of amide bonds (maximum 2 at a time) for the generation of y, b and a ions. However, the interface of the software also allows the fragmentation of ester, thioether and gylcosidic bonds. The hypothetical fragments are then compared with the query spectra and the resulting PSMs are evaluated using a specific scoring system which involves the calculation of three scores: the raw score and the *P1* and *P2* scores. The raw score calculates

the similarity between the theoretical m/z values and the query spectra employing the following equation:

$$Raw\ score = \sum_{each\ matched\ peak\ m_i} log_{10}(200 \times relative\ intensity\ of\ m_i) \quad (1.9)$$

Peaks with a relative intensity lower than 0.5% are previously removed so they are not considered in the scoring. A factor of 200 (1/0.5%) is introduced in order to avoid that peaks with an intensity above 0.5 contribute negatively to the scoring. The raw score is then used to calculate the *P1* and *P2* scores. The *P1* score represents a normalized version of the raw score. Normalization results from the calculation of the raw scores from the database compounds within a mass range from 0 to [M]+100 in reference to the parent mass (M). Note that the mass range is restricted because larger peptides produce more fragments, likely to have higher scores. Then, a gamma distribution is estimated from the raw scores. The distribution is used to calculate the statistical significance of the candidate compounds compared to the other compounds in the database. This is simply done by placing the candidate row score in the distribution and calculating the area under the right side of the curve (p-value). Finally, the *P1* score is obtained applying the following equation: *P1 = −10log10 (p-value)*.

The second score implemented in iSNAP (*P2* score) evaluates the statistical significance of the query spectra compared to *decoy* spectra. For the generation of the *decoy* spectra the authors were inspired by the cross-correlation score in SEQUEST [244], an algorithm widely used in proteomics. The approach consists in shifting the m/z values of the query spectra to obtain artificially generated MS/MS data. Given a MS/MS spectrum with a mass spectral range from *m1* to *m2*, the artificial counterparts are generated by applying an integer shift (Δm) to each m/z peak (x) in order to generate a decoy m/z value (dm) as follows:

if $x + \Delta m \leq m_2$ **then**
  $dm = x + \Delta m$;
else if $x + \Delta m > m_2$ **then**
  $dm = x + \Delta m - m_2 + m_1$

Algorithm 1: iSNAP decoy shift assignment

The shift values start from 1 to *m2-m1*, allowing the computation of multiple decoy spectra. Once the decoy data has been generated, the raw scores between the decoy spectra and the candidate structure are calculated. Employing the same method previously described, a gamma distribution is estimated from the decoy raw scores

and it serves as the basis to evaluate the statistical significance (p-value) of the score obtained with the experimental spectra. The *P2* score is calculated applying the same formula than before: *P2 = –10log10 (p-value)*.

Overall, iSNAP is quite complete in terms of fragmentation and scoring, but it only allows the analysis of a single spectra. Furthermore, the server does not seem to be maintained anymore. Actually, no results can be obtained loading the example file provided in iSNAP web application (https://magarveylab.ca/analogue/#!/search).

Integrated in GNPS [217], **Dereplicator** [230] is currently one of the most comprehensive tools for PNP identification. It provides high-throughput analysis, variable dereplication *via* spectral networks and visualization of the MS/MS annotations. The theoretical fragmentation of Dereplicator mainly cleaves peptide bonds (N-C-O) to split the molecules and construct monomer-graphs. Additionally, C-C-O cleavages are also included in the model to take into consideration structures such as thiazole and oxazole funcional groups. Then, the algorithm calculates all the possible masses resulting from the disconnection of either one or two bonds. For scoring, Dereplicator was originally combining PepNovo [245], an algorithim for the prediction of fragment ion intensities, with MS-GF+ (Mass Spectrometry Generating Function) [246], a dot product-based scoring system. Despite the relevance of MS-GF+ in proteomics, the method is rather oriented to linear peptides with proteogenic aminoacids. Consequently, this scoring system was later replaced by other approaches [237]. First, a simple SPC function was used to just calculate the shared peaks but recently, NPS [237], a new method specifically developed for PNPs, was introduced in the pipeline. This scoring is based on statistically-learned fragmentation patterns extracted from a set of reliable PSMs. It takes into consideration ion intensities and the type of ions generating them. Using the training MS/MS data, NPS discretizes the intensity values into intensity binds and identifies the ion types (b/y ion types, neutral loss fragments, double/triply charged species) associated with the peaks. This information is used to learn the probabilities of obtaining an intensity bin b given a certain ion type. For instance, the probability of a b/y ion type generating the most intense peak in a spectrum is calculated by dividing the number of PSMs obeying this condition (highest peak being a b/y ion) by the total number of PSMs. These probabilities are then used to calculate the NPScore as follows:

$$NPScore(S,T) = \sum_{(b,ion)\in Shared(S,T)} log\frac{Prob(b|ion)}{Prob(b|\emptyset)} + \sum_{(b,ion)\in Missing(S,T)} log\frac{Prob(0|ion)}{Prob(0|\emptyset)} \quad (1.10)$$

where S = experimental spectrum, T = theoretical spectrum, b = intensity bin, *ion* = ion type, $\emptyset$= dummy ion type (noise peak), *0*= missing experimental peak.

Prob (b | ion) is the probability that an experimental peak with intensity bin b comes from a theoretical peak of ion type *ion*. The first summand defines gains for the shared peaks, while the second summand penalizes theoretical peaks not matched (missing in the experimental spectra).

For the evaluation of the statistical significance of the scores, Dereplicator implements the MS-DPR algorithm [247]. MS-DPR was designed for the estimation of p-values from linear and non-linear peptides. The algorithm uses a Markov chain approach to approximate a tail of the score distribution from a given spectrum and a candidate peptide. Then, the p-value is estimated from the distribution and a target-decoy database approach is used to calculate the FDR of the identifications. For the generation of the decoy database, the building blocks of the PNP monomer-graphs used for fragmentation are shuffled. Lastly, Dereplicator also offers variant dereplication *via* spectral networks (Section 1.4.1). The spectral network analysis performed in the publication of Dereplicator showed characteristic mass shifts of 14 Da (CH_2), 17 Da (NH_3), 18 Da (H_2O), 28 Da (C_2H_4; CO), 30 Da (CH_2O), 42Da (C_2H_2O) and 113 Da, facilitating the identification of new PNP family members.

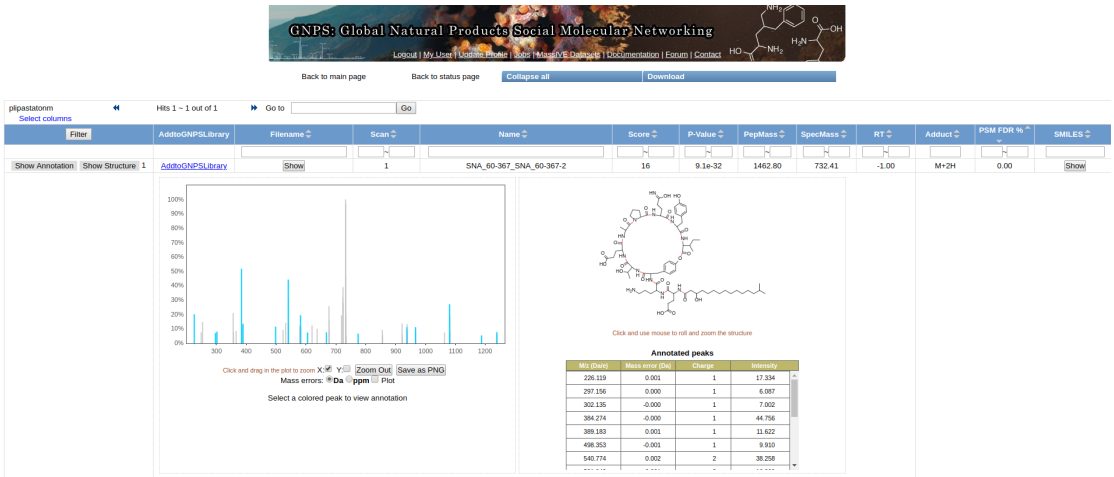


Figure 1.18: Dereplicator interface.

The web application interface of Dereplicator provides the visualization of the identified chemical structure as well as the MS/MS spectra with peak annotations (Figure 1.18). Although it is interesting and uncommon that a high-throughput tool presents options for manual MS/MS examination, the annotations are simplistic and no edition/filtering options are provided. A command line version of Dereplicator is also available as part of the NPDtools package (http://cab.spbu.ru/software/npdtools/).

Cyclobranch [228] is mainly focused on *de novo* sequencing of NRPs and polyketide siderophores, but it also provides MS/MS search against an internal database of NRP structures. Cyclobranch implements a monomer-based fragmentation (Section 1.4.3) that makes use of selectable building block collections integrated in the software. In the first publication of Cyclobranch, the software included a collection of the 19 proteinogenic amino acids, another with 33 monomers from the tested compounds and a 287 monomers library representing the building blocks of NRPs. Later, an additional library of ketide moieties and metals was introduced [248]. Differently than the previous approaches, Cyclobranch does not take chemical structures as starting point. In the *de novo* sequencing mode, monomer-graphs are constructed based on the experimental m/z distances and the building blocks in the database. Alternatively, to match known NRP structures, Cyclobranch contains a database of monomer-graph structures annotated with a custom nomenclature. Regardless of the selected mode, multiple NRP candidates are generated and the theoretical masses resulting from their fragmentation are evaluated using the *S1* and *S2* scoring functions. The *S1* is a simple SPC function, while *S2* is a sum of all the relative intensities of the matched peaks, similar to the raw score implemented in iSNAP. Cyclobranch does not provide statistical evaluation of the scores, but is understandable as their database for MS/MS search just contains 1041 NRPs from Norine (version date 07/2013) [22]. This substantially limits the dereplication capacity of the tool. However, from an annotation point of view, Cyclobranch offers useful features for structural characterization of NRPs. The sequence annotations are displayed through an interface that includes the monomer-graph and spectrum visualizations (Figure 1.19). Options not implemented in other software such as neutral losses addition or isotope patterns [249] calculation are included in Cyclobranch.

Algorithms for general metabolite identification can also be useful for the analysis of PNPs. The same team that developed Dereplicator, later introduced **Dereplicator+** [231] that covers a wider range of natural products such as polyketides, terpenes or benzenoids, among others. Dereplicator+ generalizes the fragmentation of Dereplicator by targeting less specific bonds usually broken in tandem mass

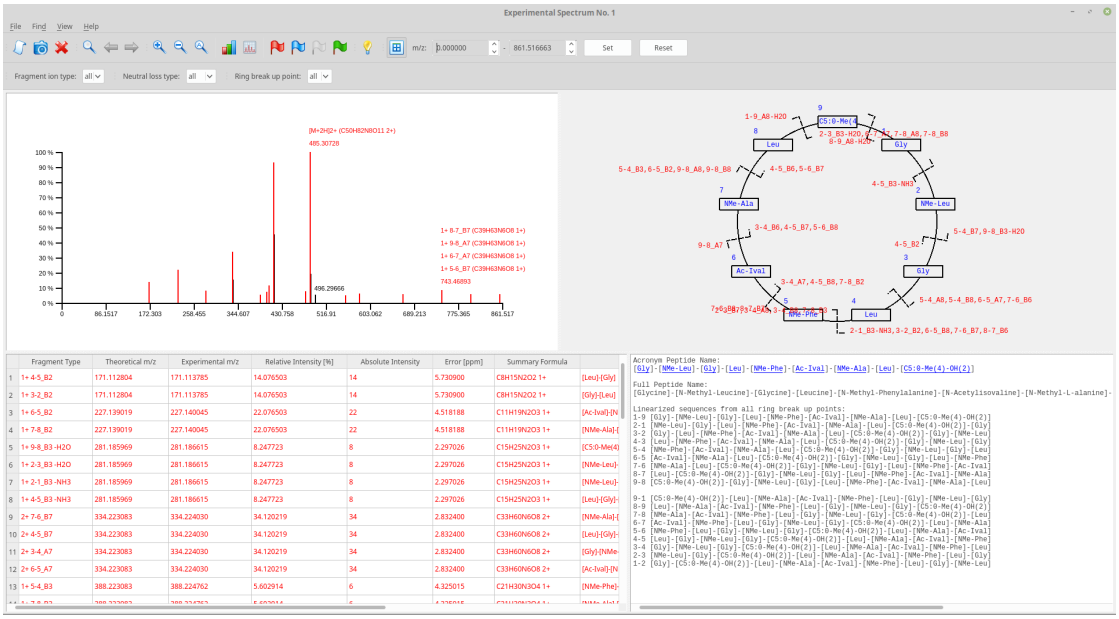


Figure 1.19: CycloBranch interface.

spectrometry such as N-C, O-C and C-C bonds. Additionally, the fragmentation in Dereplicator+ takes into consideration the sequential process of fragmentation by implementing fragmentation graphs (Section 1.4.3). Indeed, fragmentation graphs play a crucial role in Dereplicator as they are used to construct the decoy database for FDR estimation. The scoring implemented in this software mainly consists in a SPC function performed by matching the fragmentation graphs with the spectra. However, the matching is restricted to nodes in the fragmentation tree whose direct ancestors have already been annotated. Dereplicator does not calculate theoretical intensities and peak presence/abscence follows a Poisson binomial distribution that Mohimani et al. used for the calculation of the p-values. Note that the Poisson distribution associates each event with a success probability. As some mass regions tend to present more peaks than others, the authors used the experimental spectra peaks distribution to generate the success probabilities.

In terms of functionality, the main difference between Derepicator and Dereplicator+ is that, apart from PNPs, the latter can analyse a whole range of other metabolites. However, with the generalization of the approach, specific fragmentation patterns for PNPs are not further applied. That is the case of all software for general metabolites, including tools such as **MS-FINDER** [250, 251] or **SIRIUS+CSI:FingerID** [14, 252, 253]. Additionally, Dereplicator+ does not offer the same visual features than Dereplicator (chemical structure depiction, annotated MS/MS spectra...), limiting manual analysis. Both tools are still available in GNPS and they are presented as two different software.

1.5 Objectives and Thesis Overview

The work described in this thesis started in 2015 as a collaboration between the Proteome Informatics Group (PIG) from the University of Geneva (UNIGE) and the Norine team (CRIStAL group) from the University of Lille. The project's idea raised from the observation of a re-emerging trend towards NPs research as a mean to combat antimicrobial resistance and to develop other medically interesting drugs. Exploiting the power of bioinformatics in such field was the intention of this collaboration that involved two groups dedicated to software development and data analysis. Among all the possible applications of bioinformatics for NPs research, the MS/MS expertise of the PIG group combined with the Norine knowledge of NRPs motivated us to focus on the dereplication and annotation of PNPs. Furthermore, we realized that the existent tools in such field presented some limitations that we could potentially address. Hence, the **main goal of this thesis is to develop bioinformatic tools for the annotation/dereplication of PNPs and integrate them as a new platform for MS/MS analysis into Norine**. These tools are individually described in the next three chapters of the manuscript, while chapter 5 includes the last Norine update, already incorporating part of the developed applications. Figure 1.20 illustrates the overview of the thesis.

The identification of chemical substructures as a mean to predict the retro-biosynthesis of NRPs is covered in **chapter 2**, where I present rBAN (retroBiosynthetic Analysis of Nonribosomal peptides). The rBAN software is useful for automatic annotation of PNP monomers and provides highly informative edge-labeled monomer graphs. Indeed, this functionality is later used by the dereplication tool presented in chapter 4. Probably, one of the most innovative aspects of the retro-biosynthesis algorithm behind rBAN, is the inclusion of a *discovery mode*. This mode enables the detection of monomers not present in the internal database used by the software. Instead, unknown substructures are automatically searched in PubChem and suggested as new building blocks.

As a first step for dereplication, in **chapter 3**, I introduce a method for PNPs molecular formula prediction based on the KMD. Similarly to petroleomics, the chemical repeated units found in the composition of NRPs makes them amenable for the application of this technique. This new approach represents an alternative and fast method for NRPs molecular formula detection. The KMD prediction is evaluated by detecting the formula of a mixture of surfactins analyzed with high resolution mass spectrometry. The web application implementing this approach is the Kendrick Formula Predictor (KFP).

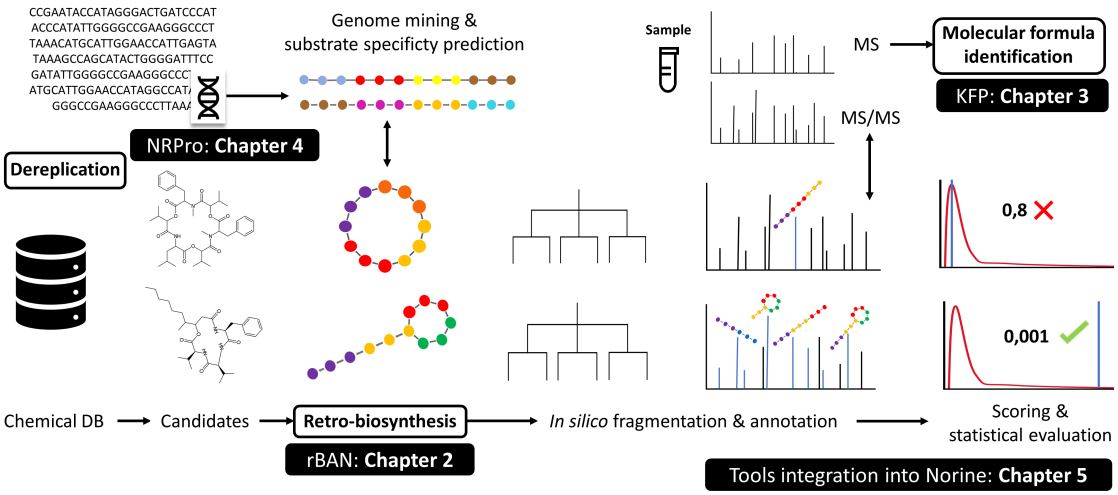


Figure 1.20: Thesis overview. Illustration of the tools covered in the following chapters and the links between them. KFP (chapter 3) and NRPro (chapter 4) are mainly used for dereplication purposes while rBAN (chapter 2) aids on the annotations of NRPro and could potentially be coupled with genome mining software in a peptidogenomics approach. All these tools are integrated in Norine (chapter 5).

A more comprehensive tool for dereplication and annotation of PNPs MS/MS spectra is presented in **chapter 4** under the name of NRPro. NRPro uses an extended version of rBAN to create chemical graphs with monomeric annotations that are used for the *in silico* fragmentation of the structures. Then, the PSMs are scored and a target-decoy database approach is implemented for the statistical evaluation. In contrast to many dereplication software, NRPro provides results in a highly interactive and intuitive interface that allows manual examination and edition of the automatic annotations. Furthermore, it includes many options such as the inclusion of neutral losses or search of spectra with adducts.

Lastly, **chapter 5** covers the last update of the Norine database, which integrates the tools described in the previous chapters. Additionally, a pipeline used for the automatic and massive extension of Norine database is presented. With the aim of maintaining the level of curation that characterizes Norine, the workflow incorporates multiple quality checks before the submission of new entries. One of these steps involves the usage of rBAN.

Chapter 2

rBAN

2.1 Overview

The retro-biosynthesis process detects the original monomers that gave rise to a given chemical structure, which is useful for annotation purposes or to establish the relationship between the chemical structure and its BGC. Before the beginning of this thesis, the Norine team developed the retro-biosynthesis tool Smiles2Monomer (s2m) in order to enhance the curation of monomer graphs in the database. In spite of its innovative algorithm, s2m showed some weaknesses, such as occasional erroneous predictions with not particularly complex molecules. These shortcomings prompted us to develop a new software based on a different strategy. The idea was to define a list of NRP-specific functional groups in order to apply what we define as retro-biosynthesis *via* bonds mapping (see Section 1.2.2). This is computationally less expensive than the approach employed in s2m and could potentially solve some of its limitations. That is how rBAN, the tool introduced in this chapter, was born. Note that a similar strategy was already used in GRAPE, also described in Section 1.2.2, but GRAPE does not provide monomer graphs. As both, GRAPE and s2m were published before rBAN, the study presented in this section includes a benchmark against these two software. Furthermore, we also demonstrate the curation abilities of rBAN by improving some of the annotations in Norine.

Ricart *et al. J Cheminform* (2019) 11:13
https://doi.org/10.1186/s13321-019-0335-x

Journal of Cheminformatics

RESEARCH ARTICLE Open Access

rBAN: retro-biosynthetic analysis of nonribosomal peptides

Emma Ricart[1,2*], Valérie Leclère[3], Areski Flissi[4,5], Markus Mueller[6], Maude Pupin[4,5] and Frédérique Lisacek[1,2,7]

Abstract

Proteinogenic and non-proteinogenic amino acids, fatty acids or glycans are some of the main building blocks of nonribsosomal peptides (NRPs) and as such may give insight into the origin, biosynthesis and bioactivities of their constitutive peptides. Hence, the structural representation of NRPs using monomers provides a biologically interesting skeleton of these secondary metabolites. Databases dedicated to NRPs such as Norine, already integrate monomer-based annotations in order to facilitate the development of structural analysis tools. In this paper, we present rBAN (retro-biosynthetic analysis of nonribosomal peptides), a new computational tool designed to predict the monomeric graph of NRPs from their atomic structure in SMILES format. This prediction is achieved through the "in silico" fragmentation of a chemical structure and matching the resulting fragments against the monomers of Norine for identification. Structures containing monomers not yet recorded in Norine, are processed in a "discovery mode" that uses the RESTful service from PubChem to search the unidentified substructures and suggest new monomers. rBAN was integrated in a pipeline for the curation of Norine data in which it was used to check the correspondence between the monomeric graphs annotated in Norine and SMILES-predicted graphs. The process concluded with the validation of the 97.26% of the records in Norine, a two-fold extension of its SMILES data and the introduction of 11 new monomers suggested in the discovery mode. The accuracy, robustness and high-performance of rBAN were demonstrated in benchmarking it against other tools with the same functionality: Smiles2Monomers and GRAPE.

Keywords: Peptide, Monomer, Retro-biosynthesis, Fragmentation, Structure analysis, Natural product, Curation, Substructure search

Introduction

Natural products are a well-recognized source for drug discovery due to their wide range of antibiotic, antitumor or immunosuppressant activities. Indeed, 26% of the drugs approved by the US FDA from 1981 to 2014 were natural products or natural products derivatives [1]. An important part of those are nonribosomal peptides (NRPs) considered as *secondary metabolites* and found in bacteria and fungi. In these organisms, NRPs are assembled by large enzymatic systems into complex structures from building blocks such as non-proteinogenic amino acids, fatty acids or carbohydrates. Significant portions of the bacterial and fungal genome are devoted to the production of these compounds. Therefore, genome mining tools such as GARLIC [2] and antiSMASH [3] have been developed to automatically identify secondary metabolite biosynthesis gene clusters. However, these tools are not able to distinguish between clusters of already known compounds and clusters uncovering new natural products. A possible approach to solve this problem is to perform the retro-biosynthesis of these compounds obtaining their constituent monomers and align them with the monomers of the predicted clusters [2, 4, 5]. A few methods predicting the retrosynthesis of a compound from its chemical structure have been described. To begin with, CHUCKLES [6] can convert a chemical structure into a monomer-based sequence by matching a set of monomers against the target structure. The monomers are previously sorted by descending size and the matching is done sequentially. The main limitations of this method are: (i) larger monomers are given the

*Correspondence: Emma.Ricart@sib.swiss
[1] Proteome Informatics Group, SIB Swiss Institute of Bioinformatics, CMU, Rue Michel-Servet 1, 1211 Geneva, Switzerland
Full list of author information is available at the end of the article

priority and (ii) monomers with more than three external connections are not handled. This approach is efficient with regular peptides, but not for NRPs. Other methods such as RECAP (Retrosynthetic Combinatorial Analysis Procedure) [7], BRICS (Breaking retrosynthetically interesting chemical substructures) [8] or molBLOCKS [9] use fragmentation rules to obtain drug-like chemical entities. However, these methods are focused on the discovery of structural motifs for drug design and they make no attempt to annotate the target compounds by identifying the resulting fragments. Moreover, their fragmentation rules are derived from common chemical reactions, lacking specificity for particular compounds such as NRPs.

In recent years, two new tools specifically designed to target NRPs have been published. The first one, Smiles-2Monomers (s2m) [10] maps the monomers of a database within an atomic structure and selects the best combination (tiling) that covers the whole molecule with non-overlapping monomers. This approach is algorithmically elegant but computationally expensive. As a result, the best tiling is obtained as an approximate solution and the optimal mapping is not always found, sometimes leading to uncovered regions in the molecule. A second solution is implemented in GRAPE (Generalized Retro-biosynthetic Assembly Prediction Engine) [2] as the theoretical deconstruction of NRPs and Polyketides (PKs) by applying specific retro-biosynthetic reactions. The obtained fragments are then matched against a monomer library integrated in the software. A sequence of monomers is given as a result, but the original monomer linkages are lost. Both, GRAPE and s2m rely on their monomer database, which is a limitation for the analysis of peptides containing new monomers.

Part of the interest in developing retro-biosynthesis tools arises from the benefit of a monomeric representation. Chemical structure databases dedicate an important part of their resources in data curation, analysis and visualization. The complex structure of NRPs often results in too dense and unclear atomic representations. A monomeric format, as with peptide sequence annotation, reduces the complexity of the layout providing the same information in a more intelligible way and facilitates the implementation of substructure and similarity search algorithms [11, 12]. Furthermore, this format is biologically meaningful as the monomers provide direct insights into the peptide activity and origin [11, 13, 14]. These substructures bring essential information to understand the biosynthesis of the peptide and, given their bioactivities, they are interesting data for structure-based drug design studies.

The convenience of the monomeric method is reflected in the emergence of new monomer-based notation formats. Examples of that are the recent languages named HELM (Hierarchical Editing Language for Macromolecules) [15, 16] and SCSR (Self-Contained Sequence Representation) [17], which provide concise annotation of complex biopolymers in a component-based approach. Some databases devoted to bioactive peptides have also chosen this format to represent their data. This is the case of Norine [18, 19], which is entirely dedicated to NRPs and uses monomer graphs for structure depiction and analysis. Indeed, all the structural analysis tools integrated in Norine are monomer-based [10–12, 14], proving the advantages of the approach. Another example is the BIRD (Biologically Interesting molecule Reference Dictionary) [20] project from PDB (Protein Data Bank) [21]. Here, the peptide-like inhibitor and antibiotic molecules are represented as polymers with sequence information or as single components. BIRD is the result of a remediation work in which part of the PDB entries were reviewed in order to improve their representation and facilitate their identification and analysis. This kind of processes require a long and tedious effort that could be accelerated using bioinformatics tools. Hence, the usage of retro-biosynthesis software is decisive to improve these curation tasks by providing automatic annotation and assuring conciseness between the atomic and monomeric annotations. Additionally, the "in silico" retro-biosynthesis can also be applied to validate already annotated entries by checking the correspondence between the existing and the predicted annotations. A practice that would also spot potentially erroneous entries.

In this article, we introduce rBAN, a new tool simulating the retro-biosynthesis of NRPs. The main strategy of the software is to perform the fragmentation of a molecule by breaking it through a set of pattern bonds and matching the resulting fragments to a monomer database (Fig. 1). The matching process was specifically designed to allow tautomer's identification, a feature that was already presented in the s2m tool and named light matching. However, the two approaches are slightly different: the light matching of s2m omits all the implicit hydrogens and bond orders to match the monomer, while rBAN only omits the position of the double/triple bonds, making the method more restrictive and decreasing the likelihood of obtaining false positives. rBAN also introduces the "discovery mode" option that is applied when a monomer cannot be matched in the custom database. In this mode, missing substructure(s) can be automatically searched in PubChem [22] to suggest a new monomer. This feature reduces the dependence to the database, providing more flexibility than the retro-biosynthetic approaches previously presented. Finally, the results are presented in a directed graph format that includes the bond types linking the monomers. To our knowledge, no

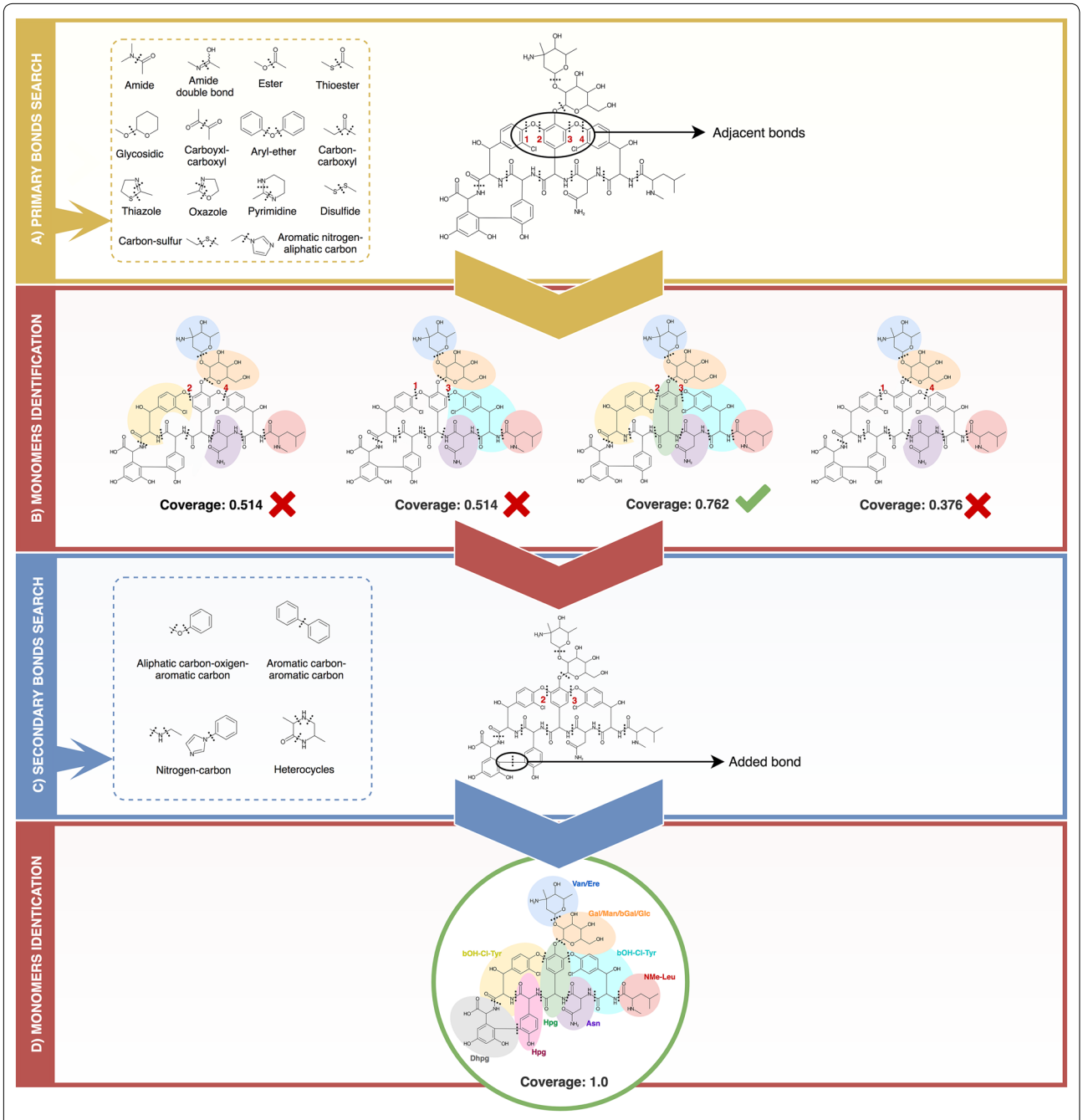


Fig. 1 Example of Vancomycin processing. **A** First, the primary bonds mapping searches the most common bonds between NRP monomers within the molecule. This process results in the mapping of two pairs of adjacent bonds that cannot be targeted simultaneously since it would isolate some atoms. To avoid that all the possible combinations only including one of the neighboring bonds are computed. **B** Then, rBAN retrieves the substructures resulting from each combination and it matches them against the monomer database. A coverage score is given to each combination based on the number of atoms that could be annotated. **C** In this case, any of the results has a full coverage, so the algorithm proceeds to the secondary bonds search of the structure with the highest score. **D** The breakage of a carbon-carbon bond results in the full mapping of the peptide

other tool provides the bond type annotation though it can be highly relevant for its integration into structural analysis pipelines. rBAN is presented in two formats, as an executable jar and as a web application, the latter being a simplified version of the software. We used rBAN

for the curation of the Norine database and benchmarked it against s2m and GRAPE.

Methods

rBAN was developed in Java using the Chemistry Development Kit (CDK) library. Given an input file with chemical structures in SMILES (Simplified molecular-input line-entry system) [23], the tool uses CDK to map the target bonds by substructure search and Norine monomer database to identify the corresponding monomers. The overall process of the software architecture is described in the Fig. 2.

Data preprocessing

NRPs monomer data

Norine is dedicated to NRPs and was used to retrieve the monomer dataset in order to guarantee consistency between the target compounds and their building blocks. This dataset consists of 534 manually-annotated monomers extracted from the compositions of the NRPs in the database. Hence, the dataset is limited to the monomers present in Norine peptides and it may not be sufficient when used for the identification of fragments of new NRPs. To solve this issue, we developed an algorithm that suggests new monomers by adding modifications to the existing ones. In order to add a biological value to the predicted structures, the modifications were selected in accordance to some of the enzymatic reactions occurring in the NRP biosynthesis [24, 25] (see Additional file 1: Table S1). For instance, a methyl group is added in the amino side of each monomer in order to mimic the action of the methyltransferase (MT) domain. Finally, preprocessing is also used to identify monomers with identical chemical graphs (isomers) and group them as a single entry (the tool does not include isomer discrimination). The PubChem PUG-REST service is used to include the PubChem IDs of the monomers.

Software architecture

1. *Primary bond search* NRP monomers are usually connected through certain types of bonds, the most common being amino and ester. Therefore, mapping these bonds is the first step of monomer identification. We rely on a graph isomorphism algorithm provided by CDK to search the substructures of the bonds within the chemical graph of the target compound. The complete list of bond types included in the search (Fig. 1A) was manually constructed based on observations and literature [25–27]. Smiles Arbitrary Target Specification (SMARTS) [28] is the language used to describe the molecular patterns of the bonds since it provides higher flexibility than

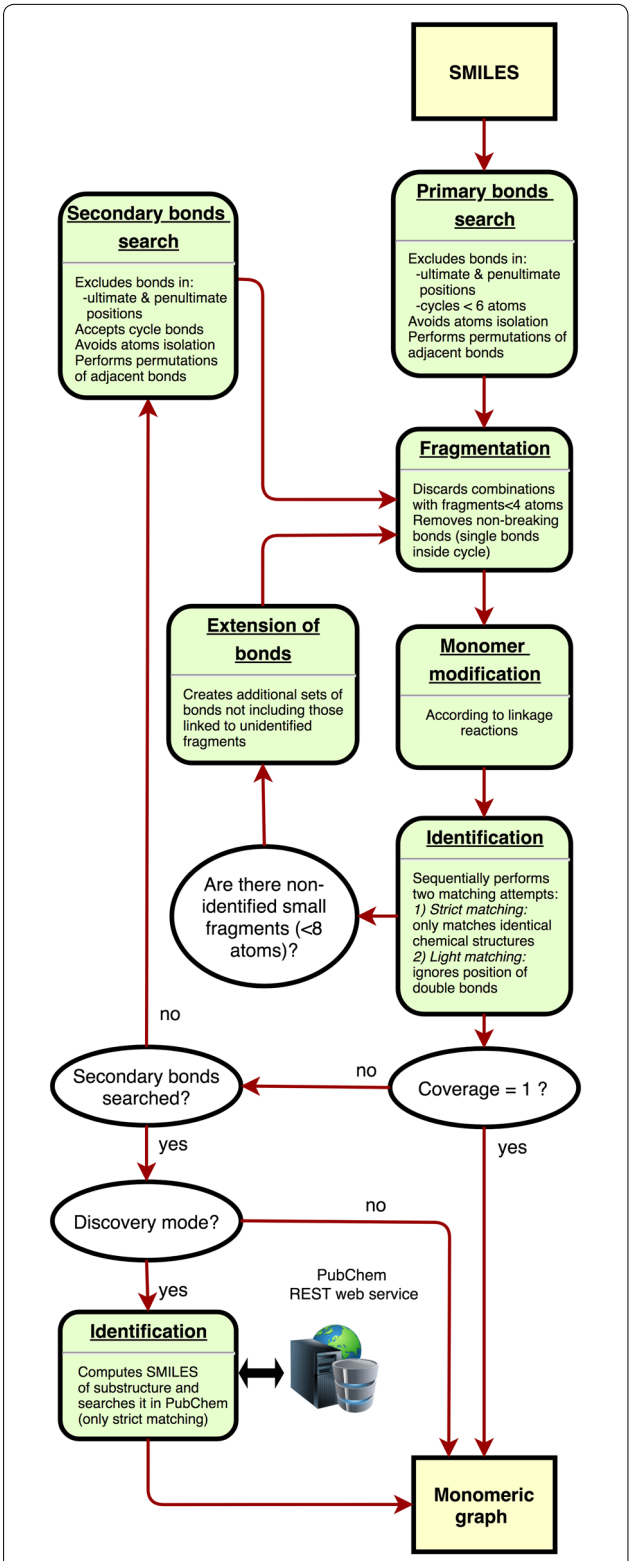


Fig. 2 Software architecture workflow. This flowchart describes the series of steps for processing structures with rBAN

SMILES. During this step, all the bonds matching the target patterns will be selected unless they are positioned on terminal branches of the chemical structure (ultimate or penultimate positions). These bonds are excluded in order to avoid single atom isolation. Bonds pertaining to cycles of less than six atoms are also removed, although they will be evaluated later in the pipeline. The only exception to that rule are the oxazole and thiazole heterocycles, as they are highly abundant in NRPs [29].

Once mapped, the bonds between adjacent positions cannot be simultaneously targeted (single atom isolation problem). This issue is solved by computing multiple combinations, each combination only including one of the neighboring bonds. To do so, the adjacent bonds are grouped in different sets and a recursive algorithm computes the Cartesian product of these sets to generate all possible permutations. Note that to reduce the number of combinations and maximize the number of targeted bonds not all the adjacent bonds are included in this calculation, but only those whose simultaneous breakage implies the isolation of single atoms or pairs (Fig. 3). In a similar way, the presence of an amino or an ester bond in the set also limits combinatorics as they are prioritized due to their predominance as NRP links.

2. *Fragmentation and identification* The bonds mapped in the primary search are used as breaking points to obtain the fragments of the molecule. This is done using a breadth-first search algorithm to iterate through the chemical graph and compute the resulting fragments from those breakages. This action is performed for each permutation of bonds provided, producing several sets of fragments that will be matched against the monomer database. Prior to this matching, the fragments are slightly modified in order to compute their expected structure outside the molecule –when not linked- thereby generating structures equivalent to those stored in the monomer database. The modifications applied are in accordance with the linkage patterns observed for each type of bond. For instance, a hydroxyl group is added to the formyl-ended fragment derived from a peptide bond breakage in order to obtain the "original" carboxyl-terminus structure of the monomer (see Additional file 1: Table S2). Once these modifications are applied, the monomers are matched against the database in order to identify them. Two different matching attempts are sequentially executed: the strict and the light matching. The strict matching will be only successful if the graph of the fragment is identical to the graph from the database. It checks the atom connectivities, the atom types and the bond orders. If a structure cannot be "strictly" matched, rBAN proceeds to light matching, which allows changes in the position of the double/triple bonds facilitating tautomer identification. Failure to match fragments can be due to the fragmentation of inner bonds in a monomer. Hence, when a fragment is not identified, the algorithm repeats the matching process by removing each of its linking bonds consecutively (Fig. 4). This process is limited to small-medium fragments (less than 8 atoms) because of their higher chances of being part of a monomer; such restriction also avoids an excessive amount of combinations. When a whole set of fragments has been matched, it is assigned with a score indicating the number of annotated atoms

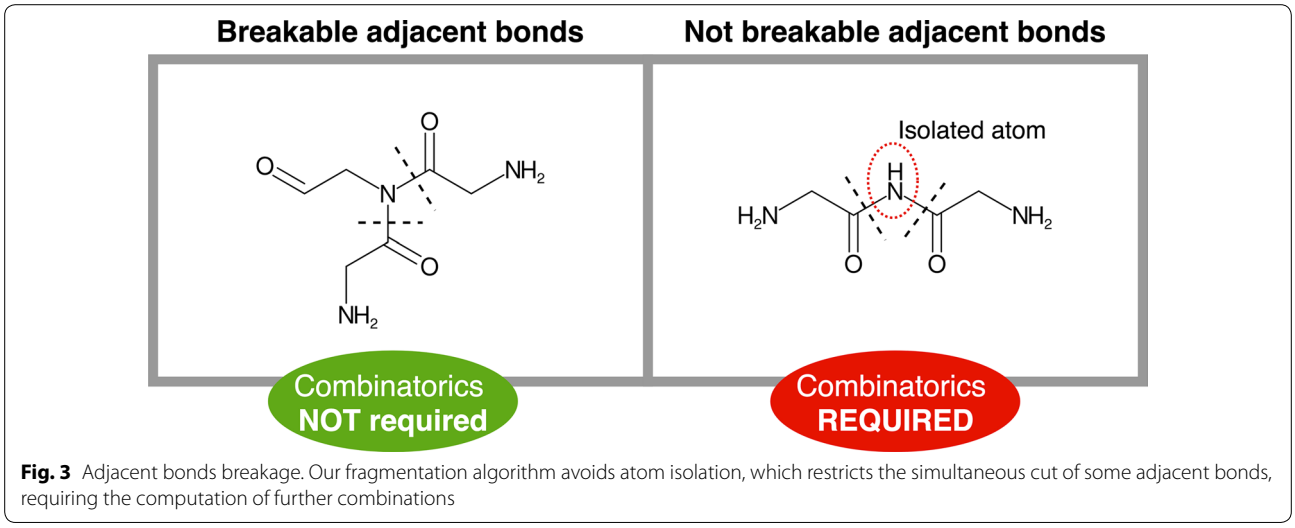


Fig. 3 Adjacent bonds breakage. Our fragmentation algorithm avoids atom isolation, which restricts the simultaneous cut of some adjacent bonds, requiring the computation of further combinations

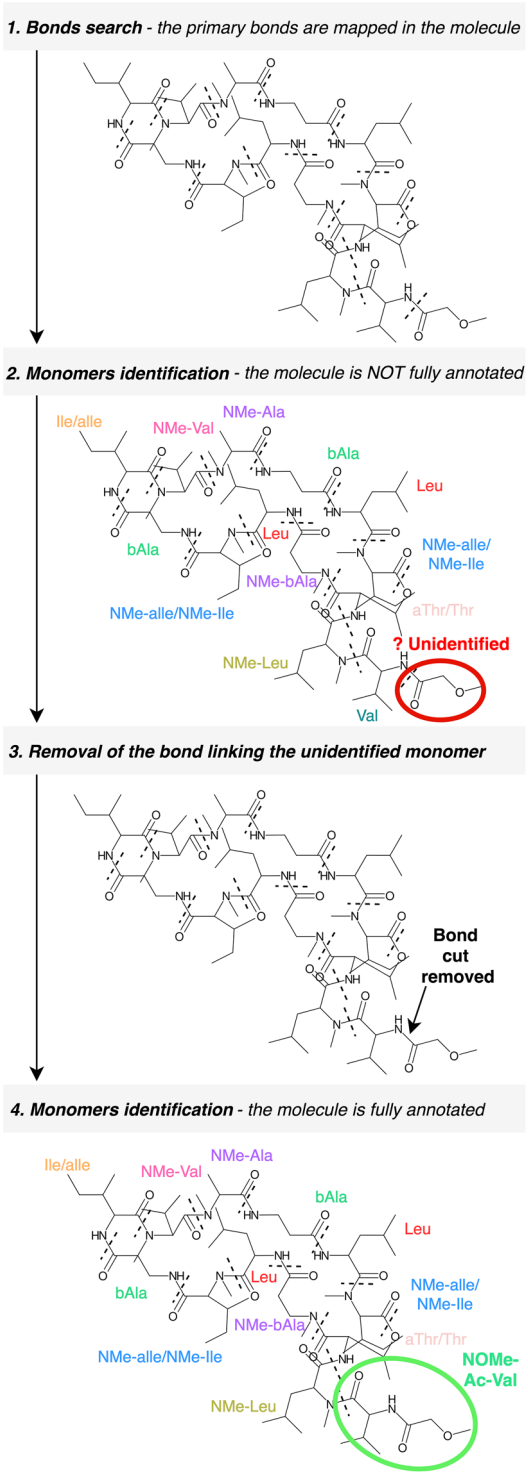


Fig. 4 Identification of monomers containing inner bonds. Some monomer bonds are sometimes fragmented by the algorithm. To handle these cases, when a small region cannot be identified, rBAN repeats the matching process after removing the bond linked to the unidentified substructure (example with *Theonellapeptolide* Ie)

versus the total number of atoms in the molecule (coverage). The next steps in the pipeline depend on these scores. If any of the fragment sets has a score of 1, ergo all the monomers have been identified, the algorithm proceeds to the monomer graph creation. Otherwise, the secondary bonds search is applied to the sets with the highest score (Fig. 1B).

3. *Secondary bond search* Some bond types such as the carbon-carbon linkages are not common as a bridge between NRP monomers and breaking them in the initial step would lead to unnecessary and excessive fragmentation. This is why they are considered as secondary bonds and their mapping is restricted to the fragments that have not been identified. The secondary bond collection comprises less common bonds and non-specific heterocycles (Fig. 1C). Specific heterocycles such as the oxazoles and thiazoles are covered in the primary search, since their cyclisation patterns are well-known [27, 30]. Yet the existence of a wide range of cyclisation forms complicates the individual targeting of the remaining heterocycles. For this reason, we use a general approach that provides several breakage possibilities instead of a single solution. The algorithm performing this task implements substructure search to identify the heterocycles and combinatorics to return the permutations of cycle bonds that break the fragment without leaving isolated atoms. After the secondary bond search, the fragmentation and identification step is repeated. If the full score is still not reached and the monomer discovery mode is activated, rBAN moves to the next step.
4. *Monomer discovery* The unidentified substructures may represent missing monomers in the database. In these cases, the CDK library is used to generate the SMILES of the unknown chemical structure that serves as a parameter for an automatic PubChem search. The substructures successfully identified are annotated using their PubChem name and suggested in the results as new monomers for the Norine database. The information is presented in a JSON file where the compounds containing the suggested monomer are also listed. Graphical results are also provided. For each new monomer, rBAN creates a folder with the depictions of the peptides where the substructure occurs.
5. *Monomer graph serialization* The monomeric structure consists of a directed graph with a set of nodes represented by the predicted fragments and a set of edges symbolizing their linking bonds. To build this graph, the monomers are reconnected using the association between the broken bonds and the resulting fragments. The edges are labeled specify-

ing the type of bond and their direction is chosen based on the type of atoms in the bond. The monomer associated with the carbon atom of the bond is set as the source while the monomer containing the heteroatom is set as the target. For instance, in case of a peptide bond, the monomer with the carboxylic side would be the source while the monomer with the amino side, the target. The graph is serialized in a JSON file also containing the atomic graph of the peptide that associates each atom with the monomer containing it (see Additional file 2). If the theoretical annotation (Norine graph) is given as an input, the output graph will also contain a "correctness" value. This value results from the division of the number of correctly annotated atoms (associated with the expected monomer) by the total number of atoms in the molecule. The graphical depiction of the chemical structures with the labeled monomers is also implemented as an option.

Results and discussion

The Norine database provides structural data of NRPs in both atomic and monomeric formats. The monomer annotation is essential to obtain the correctness of rBAN predictions and for this reason Norine was chosen as the main resource to evaluate the software.

Norine Database Curation and Extension

In Norine, the SMILES (atomic structure) and the monomer graphs (monomeric structure) are sometimes extracted from different resources. To guarantee the conciseness between the two representations and thereby validating the SMILES from Norine, rBAN was run to compare the SMILES-predicted monomeric graph with the Norine annotated graph. When the theoretical and the predicted graphs are identical, then the result is considered as correct and the corresponding SMILES is validated. From the 256 peptides that are described in SMILES, rBAN could validate 249 (97.26%) (Fig. 5a1). The non-validated peptides were manually inspected and errors in their SMILES were identified. Hence, the lack of validation was attributed to wrong input data and not to a wrong mapping of the software. In fact, the software helped to spot these wrong SMILES that were later corrected/removed from the database. An example is Enniatin F, whose monomeric annotation did not match the structure given by the SMILES (Fig. 5b).

As already mentioned, in the previous version of Norine only 256 entries (21.56% of the total) contained the structural information in the SMILES format. In order to increase this count, we used the PubChem PUG-REST Service to perform automatic searches, retrieve the missing SMILES and validate them using rBAN. The only available parameters for the PubChem searches were the name of the compound, which lacks specificity, and the PubChem link provided by Norine, that is rarely present and occasionally wrong. Hence, the validation step becomes essential to reduce the uncertainty of the search and provide more reliable results. From the 403 SMILES retrieved from PubChem, 242 were validated using rBAN (Fig. 5a2). These SMILES were added to the database generating a two-fold increase in Norine SMILES data. The non-validated entries were considered as false positives due to the uncertainty of the search.

Monomers discovery

As already mentioned, the non-validated entries can be due to a wrong annotation either in the SMILES or in the monomeric graph. In the latter case, peptides may contain monomers not present or wrongly annotated in Norine. Thus, rBAN was run in discovery mode to identify new monomers. The software suggested 61 new building blocks. Some of these predictions could be wrong due to mistakes in the input SMILES or wrong mapping of the software. Hence, a manual inspection was required before their addition into the database. To increase confidence, only the monomers present in more than one compound were evaluated.

From the 18 monomers examined, eleven were correct suggestions (Table 1). *N*-Formyl-Lysine was the most commonly found monomer, missing in Norine because CO is currently defined as a monomer in the database and occurs in several NRP graphs. In contrast, rBAN considers CO as formylation and not a monomer therefore suggested a new formylated monomer. Most of the other new entities correspond to monomers that were not properly annotated in the monomeric graph. Such is the case of the "C4:1(3)–OH(2)" monomer that should be beta-Vinyllactic acid (C5:1(4)–OH(2)) (see Additional file 1: Fig. S1). Other cases encompass a missing monomer in the monomeric graph or an incorrect SMILES of the known monomer. All the corrections were made in accordance to the literature associated with the corresponding compounds.

Seven of the monomers suggested by rBAN were rejected (find them in Additional file 1: Fig. S3) because the manual inspection of their corresponding peptides revealed that their SMILES rather than their monomeric graph created the problem. The peptidic structures (SMILES) of these records contained errors or did not even correspond to the right molecule due to the ambiguous PubChem searches previously performed.

The eleven new monomers were added in the Norine database along with the correction of the wrong

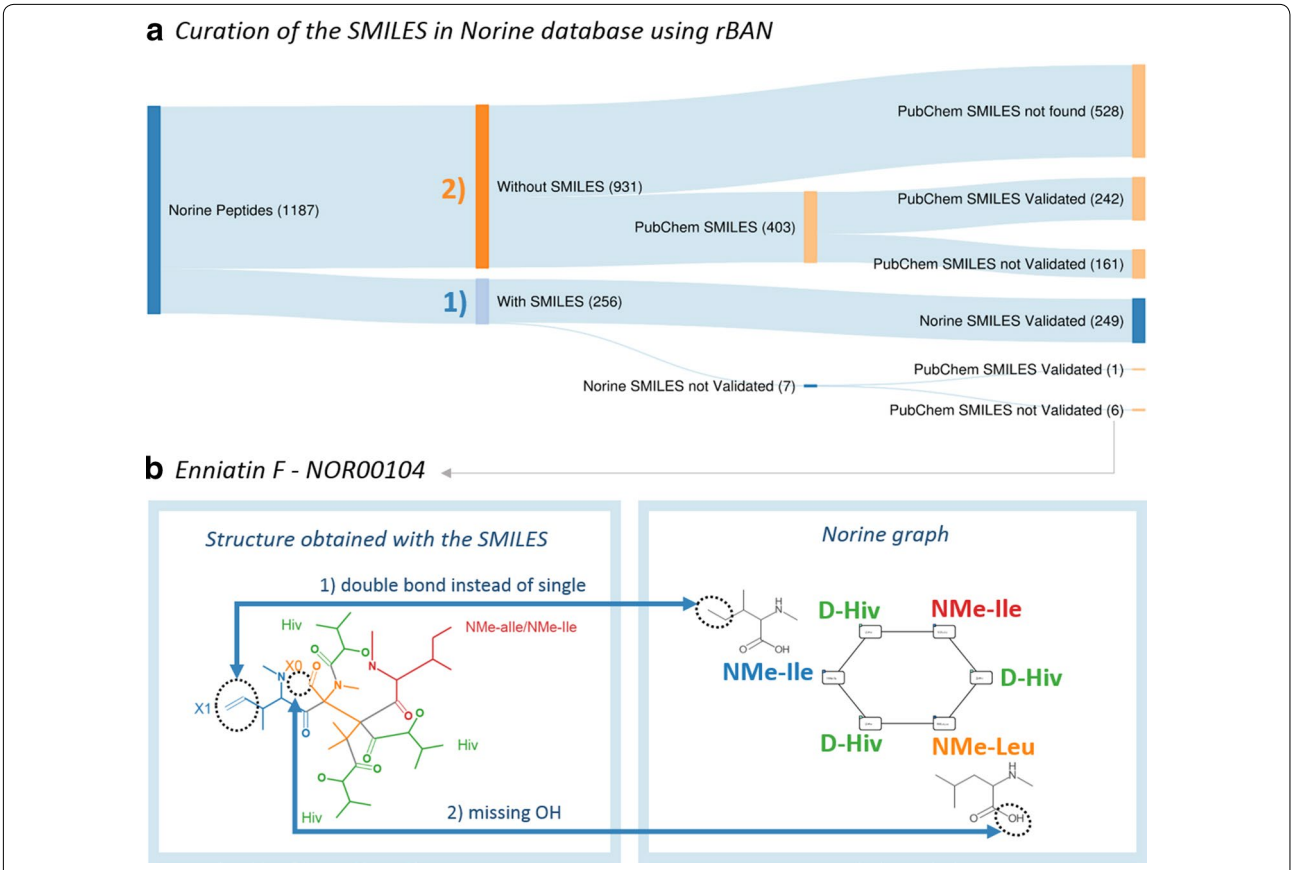


Fig. 5 Norine curation. **a** The curation involves two main steps: (1) Automatic verification and correction of the SMILES in Norine. rBAN validated 249 (97.26%) SMILES and identified seven potential erroneous SMILES. Retrieving the PubChem SMILES from the non-validated entries enabled the correction of the SMILES of Motuporin (NOR00825). The manual inspection of the remaining entries concluded with the confirmation of six wrong SMILES. (2) Automatic addition of SMILES retrieved from PubChem. From the 403 SMILES retrieved from PubChem, 242 were validated using rBAN. The 161 not validated are likely to be false positives due to the ambiguity of the PubChem searches performed. **b** Enniatin F belongs to the set of non-validated peptides. rBAN failed to validate this peptide due to differences between the molecular and monomeric annotations. The monomeric graph is circular and contains *N*-Methyl-Isoleucine while the SMILES encodes a linear structure with dehydro-*N*-Methyl-Isoleucine(1). Additionally, rBAN could not identify what is supposed to be a *N*-Methyl-Leucine because it misses a hydroxyl group (2)

annotations, either in the monomeric graphs or in the SMILES of the compounds (find examples in Additional file 1: Fig. S1). This step was essential for evaluating the increase of the validated data. In the end, 11 added new monomers along with the correction of wrong annotations boosted the count from 492 to 526 validated entries.

Benchmarking

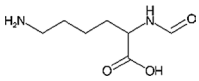
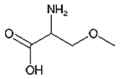
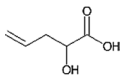
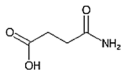
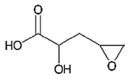
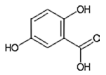
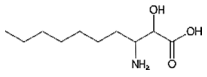
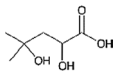
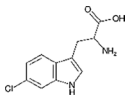
rBAN was compared against two tools with similar functionality: s2m and GRAPE. The benchmarking was performed on a PC computer with an Intel/Core i5-5300U CPU at 2.3 GHz with 4 GB of RAM allocated to the Java Virtual Machine.

rBAN vs s2m

Within the retro-biosynthetic tools targeting NRPs, s2m is the closest to rBAN as it produces the same output: a monomeric graph. Yet the two approaches substantially differ in their features and algorithmic approaches set to handle the issues raised by mapping the molecules. These involve among others, the monomer search, the light matching or the heterocycles treatment (see Table 2). In order to compare both tools, we analyzed their results, their robustness and their computational performance. The benchmark in the following sections was performed running s2m in the light matching mode to allow tautomer identification and obtain results comparable to those of rBAN.

Results comparison s2m was run to validate the same SMILES data previously used in the curation protocol of the Norine database. Out of the 659 peptidic structures retrieved from Norine and PubChem, s2m validated 445. Although the same process with rBAN resulted in a higher amount of validations (492), the comparison singled out

Table 1 Monomers correctly suggested by rBAN

Norine code	PubChemID	IUPAC name	Structure	Compounds	Reason of the missing monomer	Refs.
NFo-Lys	12679627	6-amino-2-formamidohexanoic acid		NOR00261, NOR00262, NOR00263, NOR00264 NOR00266, NOR00267, NOR00269, NOR00270 NOR00271, NOR00272, NOR00274, NOR00275 NOR00276, NOR00277, NOR00278, NOR00580	"CO" monomer in graphs	[32]
D-3OMe-Ala	97963	2-amino-3-methoxypropanoic acid		NOR00422, NOR00423, NOR00424, NOR00425 NOR00588	Wrong SMILES of D-3OMe-Ala monomer	[33]
C5:1(4)-OH(2)	172026	2-hydroxypent-4-enoic acid		NOR00064, NOR00066, NOR00068, NOR00071 NOR00073	Wrong monomer in graphs: C4:1(3)-OH(2) -> C5:1(4)-OH(2)	[34]
N-Suc	12522	4-amino-4-oxobutanoic acid		NOR00160,NOR00166, NOR00903	Missing monomer in graphs	[35, 36]
C5:0-OH(2)-Ep(4)	54305979	2-hydroxy-3-(oxiran-2-yl) propanoic acid		NOR00086, NOR00087	Wrong monomer in graphs: C4:0-OH(2)-Ep(3) -> C5:0-OH(2)-Ep(4)	[34]
Gen	3469	2,5-dihydroxybenzoic acid		NOR00489, NOR00598	Wrong monomer in graphs: 2,3-diOH-Bz -> Gen	[37, 38]
C10:0-OH(2)-NH2(3)	57484230	3-amino-2-hydroxydecanoic acid		NOR01134, NOR01135	Wrong monomer in graphs: Adda -> C10:0-OH(2)-NH2(3)	[39]
iC6:0-OH(2.4)	55300467	2,4-dihydroxy-4-methylpentanoic acid		NOR00078, NOR00077	Wrong monomer in graphs: iC5:0-OH(2.3) -> iC6:0-OH(2.4)	[34]
Isovaleric_acid	10430	3-methylbutanoic acid		NOR00477	Wrong monomer in graph: Hiv -> Isovaleric_acid	[40]
D-Cl-Trp	65259	2-amino-3-(6-chloro-1H-indol-3-yl)propanoic acid		NOR00554	Wrong SMILES of D-Cl-Trp monomer	[41]

Among the suggested monomers, *N*-Formyl-Lysine is the most abundant. rBAN considers CO as a formylation, therefore suggests a new formylated monomer instead of using the "CO" monomer currently present in Norine. A second new entity present in five compounds is D-3OMe-Ala. In this case the monomer name is correct but not the SMILES associated with it. Most of the other suggestions are due to the monomers wrongly annotated in the graph that should be substituted with a new substructure. There is also one case (N-Suc) where the monomer was directly missing in the graph. All these corrections were manually evaluated to confirm the agreement with the literature

five entries that were only verified by s2m (Fig. 6a). These entries were reviewed to identify the reasons why rBAN could not validate them . However, manual inspection only confirmed the validity of a single record, as the rest was not properly matching their monomeric counterparts and turned out to be false positives of s2m. Among these structures, Ennitatin F (Fig. 5b) that was reported earlier as cyclic and containing NMe-Leucine and NMe-Isoleucine. s2m maps these monomers in the structure yet the NMe-Leucine is missing a hydroxyl group while the NMe-Isoleucine has an additional double bond. These artefacts are related to the method of precomputation and light matching in s2m. Prior to the analysis, the precomputation of s2m generates for each monomer all the possible residues that may occur due to the loss of functional groups during the linkage with other monomers. These residues are the substructures that will be mapped by the software to identify the monomers. This strategy loses the association between the linkage and the loss of the functional group. That leads to wrong matches when the implicit hydrogens are not considered (as set in the light matching mode). This is the case of NMe-Leucine that is matched although it misses the hydroxyl group of the carboxyl end, which would be the expected structure if it was

Table 2 Comparison rBAN versus s2m

	rBAN	Smiles2Monomers
a) Monomers mapping	Based on molecule fragmentation through common monomer linking bonds	Based on mapping of monomers and selection of best tiling
b) Light matching	Positions of double/triple bonds are ignored	Implicit hydrogens and bond order are ignored
c) Heterocycles treatment	Accounts for NRP cyclisation patterns initiating oxazoles and thiasoles formation	Does not include any rule/pattern for heterocycles
d) Presence of new monomers	Unmatched regions left unannotated and potentially identified in discovery mode	Matches the most similar monomers in a given database and leaves out uncovered atoms
e) Graph serialization	Labelled edges with bond type and directed in accordance to functional groups in each side	Unlabelled edges

a) To map the monomers rBAN fragments the molecule and matches the results against the monomer database. S2m computes the combinations of monomers that fit in the molecule. b) To enable tautomer identification during the matching process rBAN omits the positions of the double bonds in the monomer, but it keeps considering those, becoming more restrictive than its analog mode in s2m, in which neither the implicit hydrogens nor the bonds order are taken into account. c) Characteristic NRP structural patterns such as heterocycles are specifically targeted in rBAN but not in s2m. d) When a region cannot be matched because of the absence of the monomer in the database, rBAN leaves the whole region unannotated (with the option of recurring to the discovery mode), while s2m tries to match the most similar monomer even if this is a wrong match and it implies leaving unannotated atoms. e) The monomers graph from rBAN has the edges labeled specifying the type of bond and its direction. s2m does not provide bond labels

linked to another monomer, but is wrong when it is terminal in the molecule (see Additional file 1: Fig. S2). The three other false positives of s2m show similar problems. The fifth entry is Kermamide K, the only true positive in the set. It was not validated by rBAN because this software does not consider CO as a monomer.

The manual evaluation of the 52 peptides uniquely validated by rBAN confirmed their validity and uncovered some structural patterns that were optimally handled by rBAN and not by s2m. The fragmentation model implemented in rBAN properly annotated large peptide structures whose monomeric composition was not revealed by the tiling algorithm of s2m. Similarly, the annotation of NRPs containing thiazoles and oxazole heterocycles was successfully carried out using rBAN, while the results of s2m did not match the monomer graph. Another pattern also observed in the rBAN-validated entries was the presence of monomers whose hydrated and dehydrated forms coexist in the monomer database. The restrictive light matching of rBAN succeeded in distinguishing them, while the light matching approach of s2m led to wrong monomer assignments. Finally, to complete the picture of correctness, we computed the distribution of correctness values from each software (Fig. 6b). Both tools showed a similar distribution though slightly shifted. rBAN generates more highly scored peptides (0.9–1) and less with correctness close to 0.

Robustness comparison The existence of several combinations of monomers mapping the same peptide substructure increases the complexity of the problem. Hence, the extension of the monomer database can easily influence the mapping of a molecule and could lead to the appearance of wrong annotations that were previously correct. The robustness of the two software was tested while extending the monomer database and evaluating its impact on the results. An additional set of monomers was obtained using the PubChem Classification browser to retrieve the chemical entities defined as non-proteinogenic amino acids (ChEBI Ontology). Components with a molecular mass higher than 450 g/mol were discarded, as they greatly exceeded the average monomer size. Chemical structures already present in the monomer database were also discarded to avoid repetitions. A total of 550 monomers were sequentially added in order to test the response of both software to different extensions of the database (Fig. 6c). rBAN correctly annotated the same amount of entries (492) for all the database sizes. Note that the number of correct annotations could not be improved because the Norine graphs were not modified to include the new monomers so maintaining the same correctness was the best that could be expected, proving the robustness of the software. On the other hand, s2m correct results dropped from 445 to 435 with the addition of 100 new monomers, although the rest of the extensions was steadily handled, only dropping by two in the extension to 450 monomers.

Computational performance comparison For the evaluation of the computational performance, the timing was limited to the analysis and did not account for the generation of images. We registered the performance of each software varying the number of input peptides from 100 to 600. To obtain the average performance each measurement was repeated five times. rBAN was significantly faster than s2m (Fig. 6d). As expected, computing time increased with the number of peptides and the difference between the two software remained 4 and 5-fold. Although this trend is likely to be confirmed, these measurements may change with a different set of peptides, as

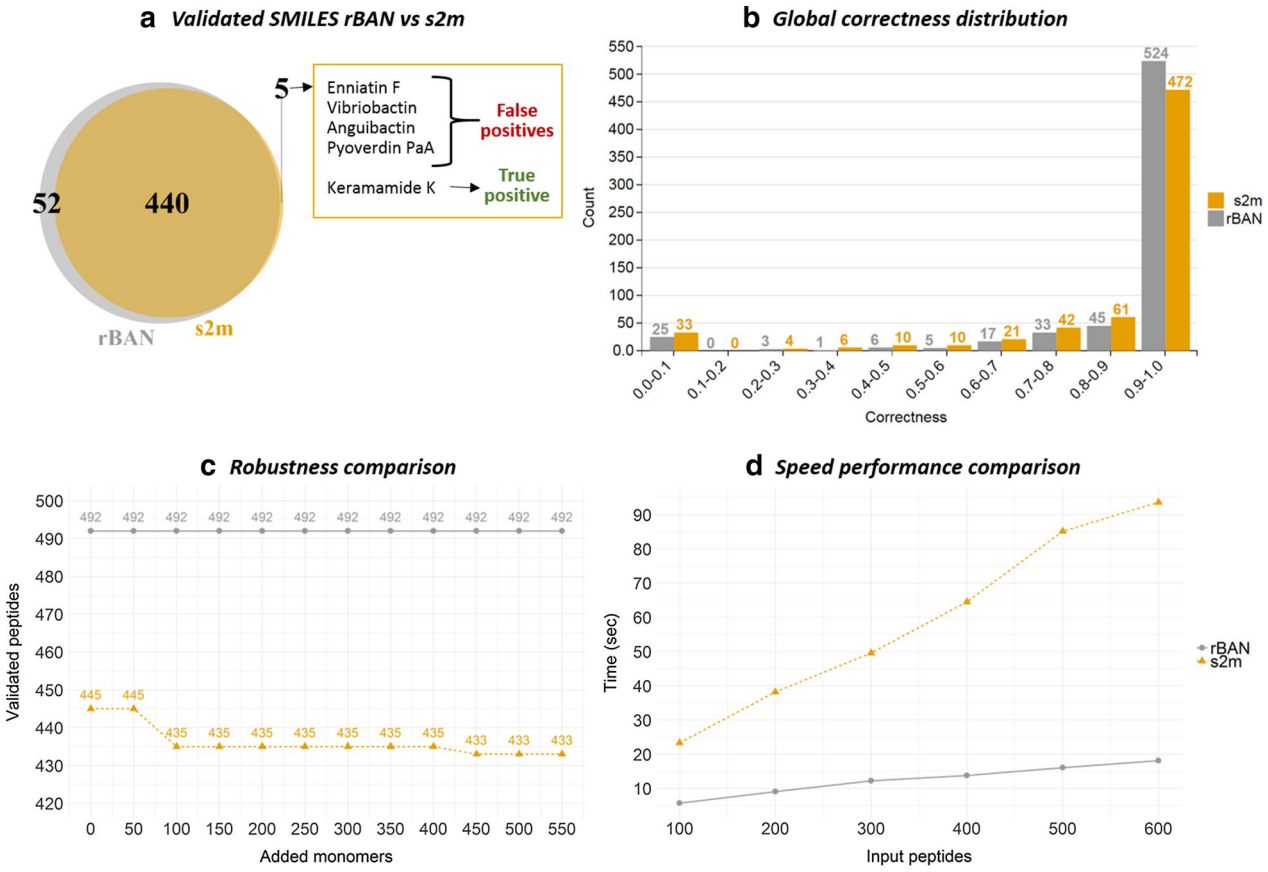


Fig. 6 Benchmarking rBAN versus s2m. **a** Both software were used to validate the SMILES data by comparing the Norine monomer graphs with the SMILES-based predicted graphs. rBAN could validate more peptides than s2m and four of the entries uniquely validated by s2m turned out to be false positives of the software. The manual examination of the entries uniquely validated by rBAN revealed a better capacity of the tool to annotate large structures and peptides containing heterocycles and tautomers. **b** The global distribution of the correctness do not show substantial differences between the two software but it proves that rBAN does not only have more correct peptides, but also less peptides with correctness values close to zero. **c** The monomer database was extended with new chemical entities to evaluate its effects on the peptide mapping. The results of rBAN remained unchanged proving its robustness, while the extension of the monomer database affected mapping in s2m. **d** The computational performance was evaluated with different amounts of input peptides. In all cases rBAN outperformed s2m, being between four and five times faster

computing time depends on the complexity of the chemical structures analyzed. Note that with rBAN, using the discovery mode feature would also change the performance results as the computation time increases due to the RESTful HTTP requests performed to retrieve data from PubChem.

rBAN vs GRAPE

As already mentioned, GRAPE is another tool for the retro-biosynthesis of NRPs and polyketides (PKs). However, the annotations provided by this software differ from those of rBAN as (1) they are based on a different monomer library and (2) the modifications are annotated separately from the monomers. These differences make the comparison of correctness difficult and that explains why the benchmark was limited to the analysis of coverage (ratio between annotated atoms and total number of atoms in the molecule). The same set of SMILES without the wrong entries previously identified was used to test GRAPE. Out of 653 peptide structures, GRAPE fully annotated 468, while rBAN reaches 560 annotated entries, 492 of them being correct. In fact, from these results it is possible to indirectly compare the correctness of the two software. Only the peptides with a full coverage can have full correctness. Hence, assuming that all the annotations from GRAPE are correct (468), the result is still lower than the number of correctly annotated peptides from rBAN (492). The whole distribution of coverage shows how GRAPE tends to leave less peptides with low coverage (Fig. 7). Nevertheless, the annotations of the 18 peptides with zero coverage in rBAN were manually checked for GRAPE. As it turned out, their monomer fragments were categorized as "unknown". Finally, the computational performance was evaluated using the

same data. rBAN analysed the 653 peptides in an average time of 26.94 s, while the same process with GRAPE resulted in an average time of 81.34 min.

Web implementation A web application interface was designed and integrated into Norine as an additional tool for the database curation. With the aim of providing a simple and user-friendly interface, the online version of rBAN is limited to the analysis of a single peptide. It only requires an input SMILES and it automatically depicts the peptide structure with the labeled monomers. Optionally, the Norine graph annotation can be introduced in order to obtain the graph correctness. The generated image can be downloaded in svg or png formats. Apart from the visual results, the serialized monomer graph is also provided as a json file. The discovery mode is still not available in the current web service version.

Conclusions

The usage of rBAN for Norine curation ended with the validation of 97% of the entries and the introduction of 242 SMILES and 11 monomers in the database. These results prove the ability of the algorithm to deduce effectively the monomeric graph of an NRP from its SMILES. The comparison with s2m and GRAPE clearly favored rBAN, which annotates more entries and with a higher perfomance. We also demonstrated the efficacy of the monomer discovery mode for the correction/addition of monomers. Furthermore, rBAN automatically produces graphs where the edges are labeled with the bond types linking the monomers. The current monomeric graphs in Norine do not contain this information, which is useful for the development of automatic substructure search. In the end, rBAN was integrated in Norine as a complementary tool for the future curation of the database. rBAN is accessible as a web service in Norine (http://bioinfo.cristal.univ-lille.fr/rban) and ExPASy (https://web.expasy.org/rban). The jar is publicly available on bitbucket (https://bitbucket.org/sib-pig/rban/downloads/).

Limitations and perspectives

The main limitation of the method is its dependence to the defined fragmentation rules. Hence, it fails mapping natural products following different patterns such as Polyketides (PKs). The introduction of new rules based on PK biosynthesis patterns would solve this issue and would extend the range of secondary metabolites covered. The software currently provides the results in a JSON format but returning the graphs in specific annotation formats such as HELM or SCSR is planned in order to improve the usability of the tool. Finally, the current slow performance of the discovery mode will be addressed by trying alternative programmatic access to PubChem data or by downloading a part of the PubChem database to our local server. In future versions of the software it would also be interesting to include a modification database and implement an optional mapping where the monomers and their modifications are annotated independently.

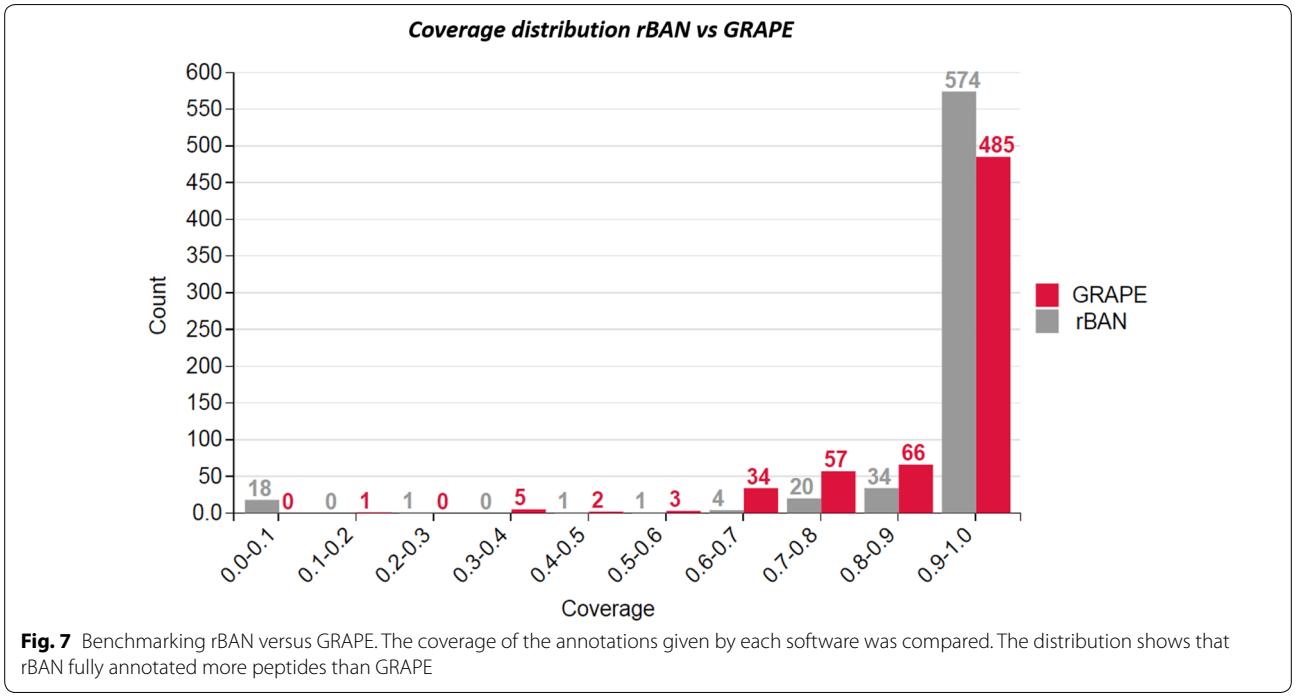


Fig. 7 Benchmarking rBAN versus GRAPE. The coverage of the annotations given by each software was compared. The distribution shows that rBAN fully annotated more peptides than GRAPE

Additional files

Additional file 1. The file contains further details of the rBAN implementation and additional information of the analysis performed in the paper.

Additional file 2. Monomeric graph of Vancomycin. Example of a monomeric graph provided by rBAN.

Authors' contributions

Conceptualization, ER and FL; methodology, ER and MP; software, ER; validation, VL, AF, MP, MM; writing—original draft preparation, ER and FL; writing—review and editing, MM, VL, MP,FL; supervision, MP and FL; funding acquisition, VL, MP and FL. All authors read and approved the final manuscript.

Author details

[1] Proteome Informatics Group, SIB Swiss Institute of Bioinformatics, CMU, Rue Michel-Servet 1, 1211 Geneva, Switzerland. [2] Computer Science Department, University of Geneva, Geneva, Switzerland. [3] EA 7394-ICV- Institut Charles Viollette, University of Lille, INRA, ISA, University of Artois, Univ. Littoral Côte d'Opale, 59000 Lille, France. [4] UMR 9189- CRIStAL- Centre de Recherche en Informatique Signal et Automatique de Lille, University of Lille, CNRS, Centrale Lille, 59000 Lille, France. [5] Bonsai Team, Inria-Lille Nord Europe, 9655 Villeneuve d'Ascq Cedex, France. [6] Vital-IT Group, SIB Swiss Institute of Bioinformatics, Amphipole Building, Quartier Sorge, 1015 Lausanne, Switzerland. [7] Section of Biology, University of Geneva, Geneva, Switzerland.

Acknowledgements

We thank the IT group of the SIB Swiss Institute of Bioinformatics for their help on the integration of rBAN on the ExPASy server.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

The interface of the software is available in the following webservers: https://web.expasy.org/rban, http://bioinfo.cristal.univ-lille.fr/rban. A JAR file with more functionalities than the web application is available in Bitbucket: Project name: rBAN. Project home page: https://bitbucket.org/sib-pig/rban/downloads. Archived version: BitBucket. Operating system(s): Platform independent. Programming language: Java. Other requirements: Java 1.8 or higher. License: Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International Public License. The datasets supporting the conclusions of this article are included within the article (and its additional files).

Funding

SIB Fellowship programme and the European Union funding through the INTERREG Va FWVL SmartBioControl/Bioscreen Project.

Publisher's Note

Received: 1 December 2018 Accepted: 31 January 2019
Published online: 08 February 2019

References

1. Newman DJ, Cragg GM (2016) Natural products as sources of new drugs from 1981 to 2014. J Nat Prod 79:629–661
2. Dejong CA, Chen GM, Li H et al (2016) Polyketide and nonribosomal peptide retro-biosynthesis and global gene cluster matching. Nat Chem Biol 12:1007
3. Medema MH, Blin K, Cimermancic P et al (2011) antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. Nucleic Acids Res 39:W339–W346
4. Harwani D, Begani J, Lakhani J (2018) Genes to metabolites and metabolites to genes approaches to predict biosynthetic pathways in microbes for natural product discovery. In: Choudhary DK, Kumar M, Prasad R, Kumar V (eds) In silico approach for sustainable agriculture. Springer, Berlin, pp 1–16
5. Blin K, Kim HU, Medema MH, Weber T (2017) Recent development of antiSMASH and other computational approaches to mine secondary metabolite biosynthetic gene clusters. Brief Bioinform. https://doi.org/10.1093/bib/bbx146
6. Siani MA, Weininger D, Blaney JM (1994) CHUCKLES: a method for representing and searching peptide and peptoid sequences on both monomer and atomic levels. J Chem Inf Comput Sci 34:588–593
7. Lewell XQ, Judd DB, Watson SP, Hann MM (1998) Recap retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. J Chem Inf Comput Sci 38:511–522
8. Degen J, Wegscheid-Gerlach C, Zaliani A, Rarey M (2008) On the Art of Compiling and Using'Drug-Like'Chemical Fragment Spaces. ChemMedChem 3:1503–1507
9. Ghersi D, Singh M (2014) molBLOCKS: decomposing small molecule sets and uncovering enriched fragments. Bioinformatics 30:2081–2083
10. Dufresne Y, Noé L, Leclère V, Pupin M (2015) Smiles2Monomers: a link between chemical and biological structures for polymers. J Cheminform 7:62
11. Abdo A, Caboche S, Leclère V et al (2012) A new fingerprint to predict nonribosomal peptides activity. J Comput Aided Mol Des 26:1187–1194
12. Caboche S, Pupin M, Leclère V et al (2009) Structural pattern matching of nonribosomal peptides. BMC Struct Biol 9:15
13. Caboche S, Leclère V, Pupin M et al (2010) Diversity of monomers in nonribosomal peptides: towards the prediction of origin and biological activity. J Bacteriol 192:5143–5150
14. Abdo A, Leclère V, Jacques P et al (2014) Prediction of new bioactive molecules using a bayesian belief network. J Chem Inf Model 54:30–36
15. Zhang T, Li H, Xi H et al (2012) HELM: a hierarchical notation language for complex biomolecule structure representation. J Chem Inf Model 52:2796–2806. https://doi.org/10.1021/ci3001925
16. Milton J, Zhang T, Bellamy C et al (2017) HELM software for biopolymers. J Chem Inf Model 57:1233–1239
17. Chen WL, Leland BA, Durant JL et al (2011) Self-contained sequence representation: bridging the gap between bioinformatics and cheminformatics. J Chem Inf Model 51:2186–2208
18. Caboche S, Pupin M, Leclère V et al (2007) NORINE: a database of nonribosomal peptides. Nucleic Acids Res 36:D326–D331
19. Flissi A, Dufresne Y, Michalik J et al (2015) Norine, the knowledgebase dedicated to non-ribosomal peptides, is now open to crowdsourcing. Nucleic Acids Res 44:D1113–D1118
20. Dutta S, Dimitropoulos D, Feng Z et al (2014) Improving the representation of peptide-like inhibitor and antibiotic molecules in the Protein Data Bank. Biopolymers 101:659–668
21. Berman HM, Westbrook J, Feng Z et al (2000) The protein data bank. Nucleic Acids Res 28:235–242
22. Kim S, Thiessen PA, Bolton EE et al (2015) PubChem substance and compound databases. Nucleic Acids Res 44:D1202–D1213
23. Weininger D (1988) SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Comput Sci 28:31–36
24. Felnagle EA, Jackson EE, Chan YA et al (2008) Nonribosomal peptide synthetases involved in the production of medically relevant natural products. Mol Pharm 5:191–211
25. Condurso HL, Bruner SD (2012) Structure and noncanonical chemistry of nonribosomal peptide biosynthetic machinery. Natural product reports 29:1099–1110
26. Giessen TW, Marahiel MA (2012) Ribosome-independent biosynthesis of biologically active peptides: application of synthetic biology to generate structural diversity. FEBS Lett 586:2065–2075
27. Bloudoff K, Schmeing TM (2017) Structural and functional aspects of the nonribosomal peptide synthetase condensation domain superfamily: discovery, dissection and diversity. Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics 1865:1587–1604

28. Daylight Theory: SMARTS—a language for describing molecular patterns. http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html. Accessed 29 Nov 2018
29. Walsh CT, Nolan EM (2008) Morphing peptide backbones into heterocycles. Proc Natl Acad Sci 105:5655–5656
30. Bloudoff K, Fage CD, Marahiel MA, Schmeing TM (2017) Structural and mutational analysis of the nonribosomal peptide synthetase heterocyclization domain provides insight into catalysis. Proc Natl Acad Sci 114:95–100
31. Crone WJK, Leeper FJ, Truman AW (2012) Identification and characterisation of the gene cluster for the anti-MRSA antibiotic bottromycin: expanding the biosynthetic diversity of ribosomal peptides. Chem Sci 3:3516–3521. https://doi.org/10.1039/C2SC21190D
32. Itou Y, Suzuki S, Ishida K, Murakami M (1999) Anabaenopeptins G and H, potent carboxypeptidase A inhibitors from the cyanobacterium Oscillatoria agardhii (NIES-595). Bioorg Med Chem Lett 9:1243–1246
33. Ford PW, Gustafson KR, McKee TC et al (1999) Papuamides A-D, HIV-inhibitory and cytotoxic depsipeptides from the sponges *Theonella mirabilis* and *Theonella swinhoei* collected in papua New Guinea. J Am Chem Soc 121:5899–5909
34. Pedras MSC, Zaharia LI, Ward DE (2002) The destruxins: synthesis, biosynthesis, biotransformation, and biological activity. Phytochemistry 59:579–596
35. Teintze M, Leong J (1981) Structure of pseudobactin A, a second siderophore from plant growth promoting Pseudomonas B10. Biochemistry 20:6457–6462
36. Atkinson RA, Salah El Din ALM, Kieffer B et al (1998) Bacterial iron transport: 1H NMR determination of the three-dimensional structure of the gallium complex of pyoverdin G4R, the peptidic siderophore of Pseudomonas putida G4R. Biochemistry 37:15965–15973
37. Chill L, Kashman Y, Schleyer M (1997) Oriamide, a new cytotoxic cyclic peptide containing a novel amino acid from the marine sponge *Theonella* sp. Tetrahedron 53:16147–16152
38. Fusetani N, Nakao Y, Matsunaga S (1991) Nazumamide A, a thrombin-inhibitory tetrapeptide, from a marine sponge, *Theonella* sp. Tetrahedron Lett 32:7073–7074
39. Sano T, Takagi H, Morrison LF et al (2005) Leucine aminopeptidase M inhibitors, cyanostatin A and B, isolated from cyanobacterial water blooms in Scotland. Phytochemistry 66:543–548
40. Nakao Y, Oku N, Matsunaga S, Fusetani N (1998) Cyclotheonamides E2 and E3, new potent serine protease inhibitors from the marine sponge of the genus *Theonella*. J Nat Prod 61:667–670
41. Schmidt EW, Faulkner DJ (1998) Microsclerodermins C–E, antifungal cyclic peptides from the lithistid marine sponges *Theonella* sp. and Microscleroderma sp. Tetrahedron 54:3043–3056

rBAN: retro-biosynthentic analysis of non ribosomal peptides

Supplementary tables and figures

Table 1. Precomputation modifications.

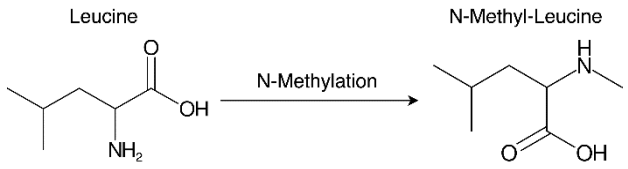
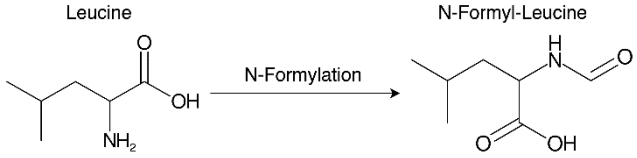
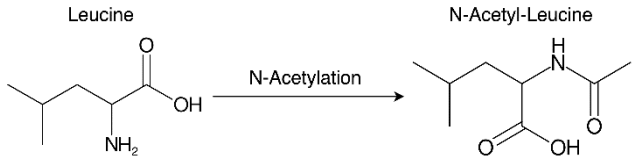
Modification	Example
N-Methylation	
N-Formylation	
N-Acetylation	
Amino group transfer	

Table 2. Modifications of the fragments before the matching. The examples shown below represent the “in silico” process undertaken by rBAN to predict the original structure of the monomers. Note that they do not symbolise chemical equations.

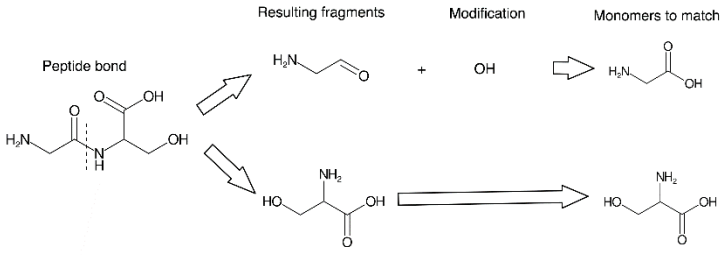
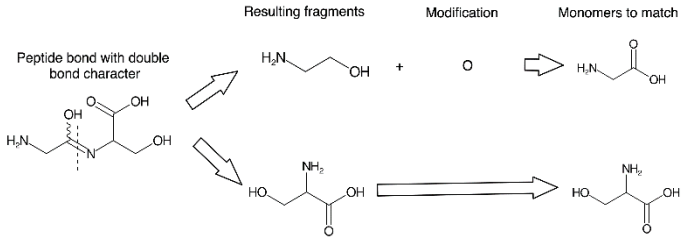
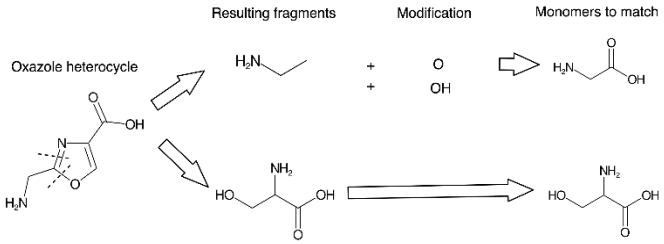
Modification	Targets bonds	Example
Addition of hydroxyl group	Peptidic Ester Thioether Carbon-Carboxyl Carboxyl-Carboxyl Glycosidic	
Addition of oxygen	Peptide bonds with double bond character	
Addition of oxygen and hydroxyl group	Heterocycle bonds	

Table 3. Suggested monomers NOT added in Norine (False positives).

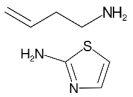
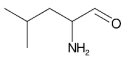
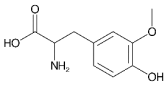
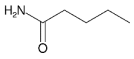
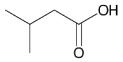
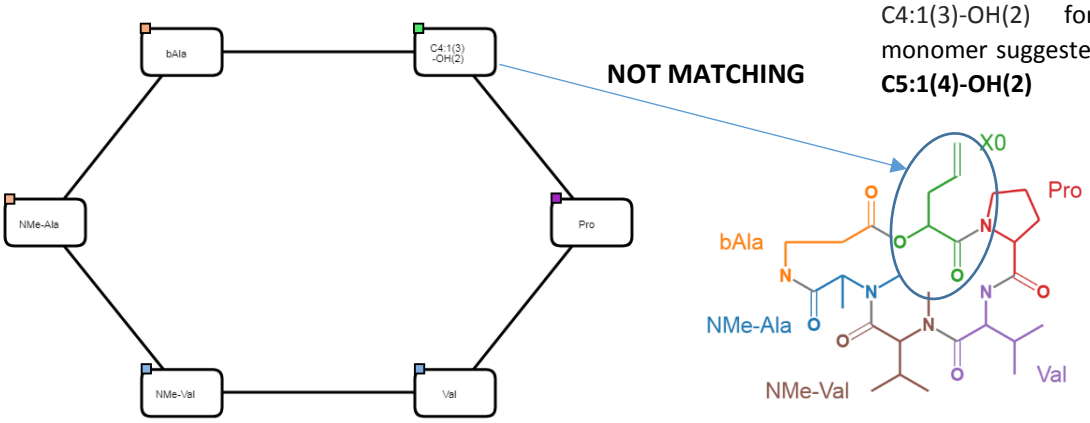
Name	PubChemID	Structure	Compounds	Reason of false positive	References
3-Buten-1-amine and 2-AMINOTHIAZOLE	443732 2155		NOR00008, NOR00015,NOR00079, NOR00115 NOR00120, NOR00122, NOR00127, NOR00130 NOR00135, NOR00144, NOR00145, NOR00154 NOR00157, NOR00159, NOR00750, NOR00817	Wrong SMILES molecules	[1], [2],[3],[4],[5]
2-Amino-4-methyl-pentanal	4473091		NOR00937, NOR00938, NOR00939, NOR00940	Wrong SMILES molecules	[6],[7],[8]
dihydropyrrolecarbaldehyde	18721951		NOR00596,NOR00597	Wrong SMILES and wrong monomeric graph	[9]
3-Methoxytyrosine	1670		NOR00424, NOR00425	Wrong SMILES molecules	[10]
Pentamide	12298		NOR00756, NOR01090	Wrong SMILES molecule (NOR00756) and wrong software annotation (NOR01090)	[11],[12]
Isovaleric acid	10430		NOR00437	Wrong SMILES and wrong monomeric graph	[13]
1,2-Dimethylpropylamine	11731		NOR00406, NOR00407	Wrong SMILES molecules	[14]

Fig. 1. Examples of Norine curation.

a) Correction of the monomeric graph of Destruxin A2 (NOR00068)

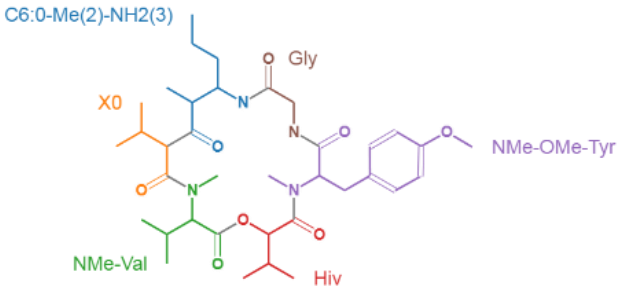


Correction:

Substitution of the monomer C4:1(3)-OH(2) for the new monomer suggested by rBAN: **C5:1(4)-OH(2)**

b) Correction of the SMILES of Guinamide D (NOR00437)

BEFORE



AFTER

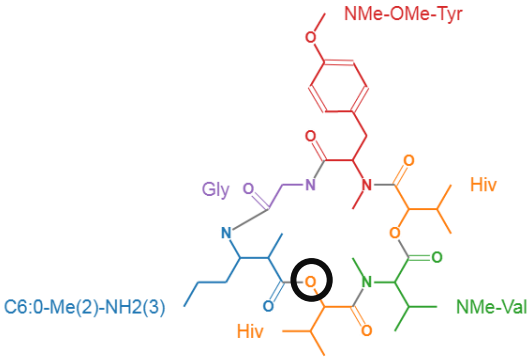
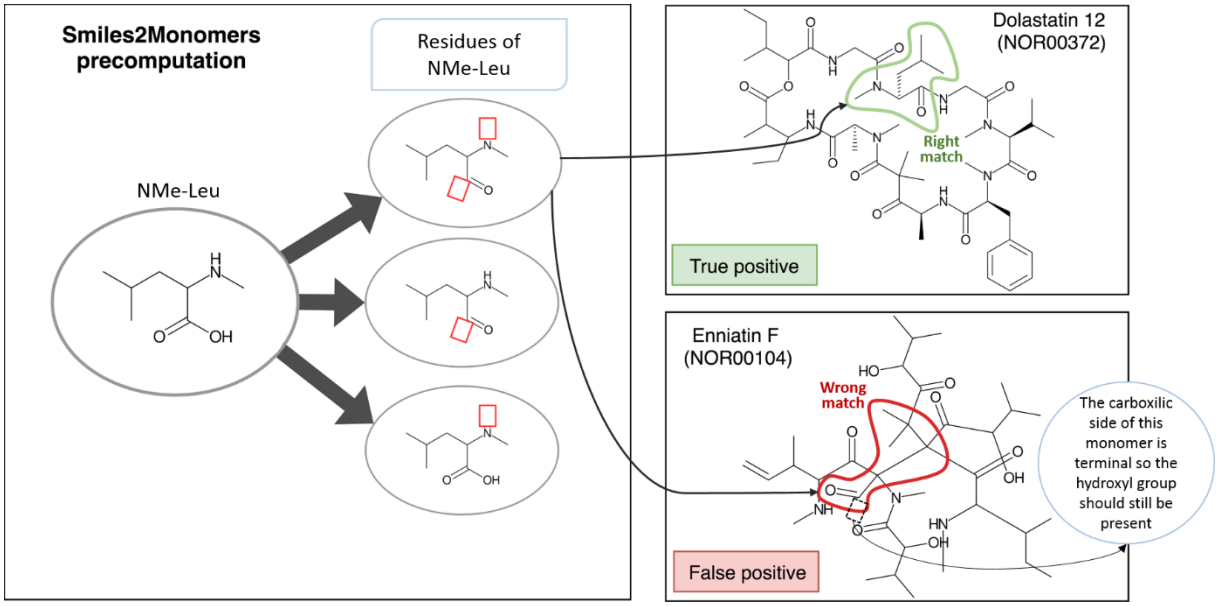


Fig. 2. Smiles2Monomers precomputation problem.



1. Psurek A, Neusüss C, Pelzing M, Scriba GK (2005) Analysis of the lipophilic peptaibol alamethicin by nonaqueous capillary electrophoresis-electrospray ionization-mass spectrometry. Electrophoresis 26:4368–4378

2. Zaharia IL, Gai Y, Zhou Y, Ward DE (2001) In planta sequential hydroxylation and glycosylation of a fungal phytotoxin: avoiding cell death and overcoming the fungal invader. Proc Natl Acad Sci 98:747–752

3. Namikoshi M, Yuan M, Sivonen K, et al (1998) Seven new microcystins possessing two L-glutamic acid units, isolated from Anabaena sp. strain 186. Chem Res Toxicol 11:143–149

4. Mitchell RE, Young H (1985) N-coronafacoyl-L-isoleucine and N-coronafacoyl-L-alloisoleucine, potential biosynthetic intermediates of the phytotoxin coronatine. Phytochemistry 24:2716–2717

5. Beattie KA, Kaya K, Sano T, Codd GA (1998) Three dehydrobutyrine-containing microcystins from Nostoc. Phytochemistry 47:1289–1292

6. Berek I, Becker A, Schröder H, et al (2009) Ampullosporin A, a peptaibol from Sepedonium ampullosporum HKI-0053 with neuroleptic-like activity. Behav Brain Res 203:232–239

7. Kronen M, Görls H, Nguyen H-H, et al (2003) Crystal structure and conformational analysis of ampullosporin A. J Pept Sci Off Publ Eur Pept Soc 9:729–744

8. Kronen M, Kleinwaechter P, Schlegel B, et al (2001) Ampullosporins B, C, D, E1, E2, E3 and E4 from Sepedonium ampullosporum HKI-0053: structures and biological activities. J Antibiot (Tokyo) 54:175–178

9. Gunasekera SP, Pomponi SA, McCarthy PJ (1994) Discobahamins A and B, new peptides from the Bahamian deep water marine sponge Discodermia sp. J Nat Prod 57:79–83

10. Ford PW, Gustafson KR, McKee TC, et al (1999) Papuamides A- D, HIV-Inhibitory and Cytotoxic Depsipeptides from the Sponges Theonella mirabilis and Theonella swinhoei Collected in Papua New Guinea. J Am Chem Soc 121:5899–5909

11. Laird DW, LaBarbera DV, Feng X, et al (2007) Halogenated cyclic peptides isolated from the sponge Corticium sp. J Nat Prod 70:741–746

12. Rinehart Jr KL, Gaudioso LA, Moore ML, et al (1981) Structures of eleven zervamicin and two emerimicin peptide antibiotics studied by fast atom bombardment mass spectrometry. J Am Chem Soc 103:6517–6520

13. Tan LT, Sitachitta N, Gerwick WH (2003) The Guineamides, Novel Cyclic Depsipeptides from a Papua New Guinea Collection of the Marine Cyanobacterium Lyngbya majuscula. J Nat Prod 66:764–771

14. Bringmann G, Lang G, Steffens S, Schaumann K (2004) Petrosifungins A and B, Novel Cyclodepsipeptides from a Sponge-Derived Strain of Penicillium brevicompactum. J Nat Prod 67:311–315

2.2 Concluding Remarks

We have proved the curation capabilities of the retro-biosynthesis analysis implemented in rBAN. Indeed, as later reflected in chapter 5, rBAN has been further used for Norine curation after its publication. rBAN also outperforms the two most similar tools, GRAPE and s2m, in multiple aspects: i) it shows a higher annotation coverage than GRAPE and ii) higher accuracy, robustness and speed than s2m. The reception of the tool seems positive as after one year of its publication the tool has been cited by seven other studies and 383 users have been recorded in the online version of rBAN.

Chapter 3

KFP

3.1 Overview

Molecular formula prediction is commonly used as a first step of dereplication. While most of the algorithms for molecular formula detection rely on isotope patterns, we realized that we could take advantage of the fact that we were working with a specific type of compounds whose composition is based on the repetition of certain chemical groups. This characteristic enables the application of the KMD that has been mainly used in petroleomics so far. In this chapter, I introduce our adaptation of the KMD for NRP formula detection as well as its software implementation named KFP. The approach consists in using the Norine database to create the KMD plot from which molecular formula can be predicted (see Section 1.4.2). In NRPs, the CH_2 unit often distinguishes different members of the same family. Hence, in the following study we based the KMD on the CH_2 pattern and applied the method to predict the formula of surfactin variants analysed with high resolution mass spectrometry. The KFP tool implementing such approach is presented in a web application format and given a mass to charge ratio, the software suggests a list of candidate chemical formulas. It is supported by a graphical representation of the KMD plot to visualize the corresponding process.

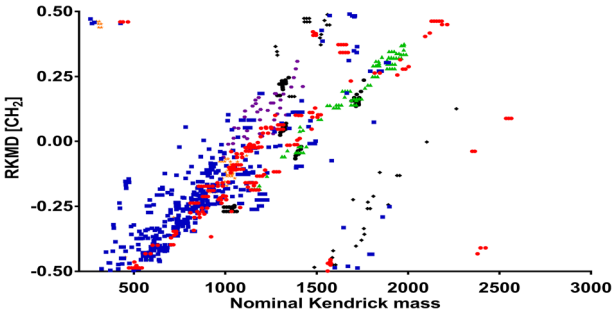
J. Am. Soc. Mass Spectrom. (2019)
DOI: 10.1007/s13361-019-02314-3

RESEARCH ARTICLE

Kendrick Mass Defect Approach Combined to NORINE Database for Molecular Formula Assignment of Nonribosomal Peptides

Mickaël Chevalier,[1] Emma Ricart,[2] Emeline Hanozin,[3] Maude Pupin,[4,5] Philippe Jacques,[6] Nicolas Smargiasso,[3] Edwin De Pauw,[3] Frédérique Lisacek,[2] Valérie Leclère,[1] Christophe Flahaut[1]

[1]Univ. Lille, INRA, ISA, Univ. Artois, Univ. Littoral Côte d'Opale, EA 7394-Institut Charles Viollette (ICV), F-59000, Lille, France
[2]Proteome informatics Group, SIB Swiss Institute of Bioinformatics (SIB), and Computer Science Department, University of Geneva, Geneva, Switzerland
[3]Mass Spectrometry Laboratory, Molecular Systems - MolSys Research Unit, University of Liège, Liège, Belgium
[4]Univ. Lille, CNRS, Centrale Lille, UMR 9189 - CRIStAL - Centre de Recherche en Informatique Signal et Automatique de Lille, F-59000, Lille, France
[5]Inria-Lille Nord Europe, Bonsai team, F-59655, Villeneuve d'Ascq Cedex, France
[6]TERRA Research Centre, Microbial Processes and Interactions (MiPI), Gembloux Agro-Bio Tech University of Liège, B-5030, Gembloux, Belgium



Abstract. The identification of known (dereplication) or unknown nonribosomal peptides (NRPs) produced by microorganisms is a time consuming, expensive, and challenging task where mass spectrometry and nuclear magnetic resonance play a key role. The first step of the identification process always involves the establishment of a molecular formula. Unfortunately, the number of potential molecular formulae increases significantly with higher molecular masses and the lower precision of their measurements. In the present article, we demonstrate that molecular formula assignment can be achieved by a combined approach using the regular Kendrick mass defect (RKMD) and NORINE, the reference curated database of NRPs. We observed that irrespective of the molecular formula, the addition and subtraction of a given atom or atom group always leads to the same RKMD variation and nominal Kendrick mass (NKM). Graphically, these variations translated into a vector mesh can be used to connect an unknown molecule to a known NRP of the NORINE database and establish its molecular formula. We explain and illustrate this concept through the high-resolution mass spectrometry analysis of a commercially available mixture composed of four surfactins. The Kendrick approach enriched with the NORINE database content is a fast, useful, and easy-to-use tool for molecular mass assignment of known and unknown NRP structures.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s13361-019-02314-3) contains supplementary material, which is available to authorized users.

Correspondence to: Christophe Flahaut; *e-mail:* christophe.flahaut@univ–artois.fr

Published online: 28 October 2019

Keywords: Kendrick map, Mass defect, Molecular formula, Nonribosomal peptides, NORINE
Abbreviations *ESI*, Electrospray ionization; *FT*, Fourier transform; *ICR*, Ion cyclotron resonance; *KMD*, Kendrick mass defect; *NRPs*, Nonribosomal peptides; *NRPS*, Nonribosomal peptide synthetases; *NKM*, Nominal Kendrick mass; *MALDI*, Matrix-assisted laser desorption/ionization

Received: 12 January 2019/Revised: 3 July 2019/Accepted: 10 August 2019

Introduction

Nonribosomal peptides (NRPs) are secondary metabolites usually produced by microorganisms. They represent very large families of natural products with a peptidic moiety. Belonging to the class of peptide secondary metabolites, NRPs are organic molecules that are not directly involved in the growth of an organism. Their absence is not lethal but may impact the survival, appearance, or growth of the microorganism in a given ecological niche. NRP production provides an advantage to those microorganisms that synthesize these molecules by boosting competitiveness. In contrast with ribosomal peptides, the molecular structure of NRPs cannot be directly deduced from the genome because their biosynthesis does not result from the translation of mRNA. In fact, NRP synthesis is performed via large enzymatic complexes called nonribosomal peptide synthetases (NRPS) and produces linear, semi-cyclic, cyclic, or branched polymeric structures of masses ranging from 200 to 3000 Da.

Most NRPs are metabolites including both a peptide core and a nonpeptidic moiety. They can be modified during or post-synthesis (N-formylated, N-methylated, acetylated, glycosylated, reduced, oxidized) increasing their structural biodiversity. Currently, there are more than 500 monomers, among them proteogenic and non-proteogenic amino acids, but also aliphatic chains, chromophores, and many others are known and referenced in the NORINE database that gathers more than 1187 NRP curated structures [1, 2]. NRPs display an extremely broad range of biological activities and pharmacological properties ranging from anti-bacterial, anti-inflammatory, surfactants, or siderophores iron chelatant (siderophores). Hence, the interest of identifying new NRPs and developing effective screening tools is high, considering potential applications in many fields as health, cosmetic, agrofood, or biocontrol.

In the "omics" cascade [3] (genomics, transcriptomics, proteomics, lipidomics, glycomics…), metabolomics and metabonomics [4] designate the comprehensive, dynamic, qualitative, and quantitative study of all the small molecules (≤ to about 1500 Da) in biological samples [5, 6]. Therein, metabolomics encompasses the study of secondary metabolites such as NRPs. However, the mass range and structure of NRPs do not fully qualify for processing with any of the metabolomics analytical workflows. This specificity warrants the definition of "NRPomics" as the systematic study of NRPs that entails the comprehensive, dynamic, qualitative, and quantitative characterization of NRPs present in environmental or biological samples.

Concomitantly to the massive and non-controversial use of ultraviolet-visible (UV) and infrared (IR) spectrophotometry, mass spectrometry (MS) (all coupled to separative techniques) is also a frequently used analytical method for secondary metabolite characterization [7, 8]. In this regard, high-resolution mass spectrometry (HRMS) such as very high field Fourier transform ion cyclotron resonance (FT-ICR) technologies allows sub-ppm measurements for the computer-assisted deduction of molecular formulae [9]. This can be achieved computationally with software that usually relies on detecting the isotopic pattern, the protonated or alkali metal adducts, and the state of charge of the molecule. However, some of the most popular mass spectrometers, based on Orbitrap and hybrid Q-TOF technologies, do not have the necessary resolving power and mass accuracy to establish a molecular formula to ions of given *m/z* [10]. Nonetheless, the molecular formula is a first piece of information contributing to the identification of a compound. Overall, when the mass exceeds 500 Da, several possibilities of candidate molecular formulae co-exist and the greater the measured mass, the greater the number of possibilities. As a result, a range of strategies and algorithms has been developed. Kind and Fiehn were among the first to propose a set of tools for the calculation of elementary composition called "the seven golden rules" [11]. This toolset relies on a combination defined by Senior and Lewis, of rules of elementary ratios for the CHONPS elements (respecting the valence of atoms) and rules of isotopic abundance. This software is coded in Visual Basic, usable from Excel, and is freely available. In metabolomics [12, 13] and more generally in chemistry [14, 15], one of the strategies for obtaining a molecular formula consists in using the isotopic profile. Most MS software (Sirius [16] and Brain [17]) can simulate MS signals with respect to both molecular formula and a characteristic terrestrial isotope composition [18] while taking into account the resolving power of the generating device (Chemcalc) [19]. Such an approach significantly reduces the number of candidates and eliminates over 90% of incorrect molecular formulae for masses greater than 1000 Da [16].

Long before this software era, Edward Kendrick had proposed an elegant mathematic method based on the determination of a mass defect (now commonly named Kendrick mass defect (KMD)) to facilitate the discrimination between homologous compounds having different numbers of same base units. Briefly, the notion of mass defect of a single element or chemical compound is calculated as the difference between the exact mass of the corresponding isotope and its nominal mass which is the simple addition of the number of protons and neutrons in a given formula or elemental isotope [20]. By convention, carbon-12 (^{12}C) has been defined [18] as the element with zero mass defect, and therefore, its atomic mass

is 12 Da while the hydrogen (^{1}H) has an atomic mass of 1.00783 for a nominal mass of 1, and hence a mass defect of 0.00783.

The nominal Kendrick mass (NKM) uses an atom group as a building block (or base unit) while applying the principle of the ^{12}C IUPAC definition. NKM refers to setting the mass of an isotope of a specific molecular group rounded to the nearest integer. Typically, for the $^{12}C_1{}^{1}H_2$ building block, the NKM is 14. Therefore, the Kendrick mass (KM) of a compound is:

KM = IUPAC protonated mass × (NKM building block (if $^{12}C_1{}^{1}H_2$ = 14) / exact IUPAC mass building block (if $^{12}C_1{}^{1}H_2$ = 14.01565)).

The KM can be extrapolated to all other building blocks and their isotopes (e.g., $^{12}C_1{}^{1}H_1{}^{2}H_1$; $^{12}C_1{}^{2}H_2$; $^{13}C_1{}^{1}H_2$, $^{13}C_1{}^{1}H_1{}^{2}H_1$; …$^{12}C_2{}^{1}H_1{}^{16}O_1$). For clarity, beyond this point, Kendrick mass will refer to monoisotopic mass.

By analogy to the IUPAC mass defect notion, the Kendrick mass defect (KMD) is defined as the delta between KM and NKM and varies between − 0.5 < KMD < + 0.5. This mass filtering method is best illustrated using 2D-plots representing the value of KMD as a function of NKM [21]; each point of the 2D-plot represents a unique monoisotopic molecular formula, and the molecules differing by one building block are correlated horizontally.

In signal processing, the effect that causes different signals to become indistinguishable is referred to as *aliasing*. It is an undesirable phenomenon that is usually avoided by applying appropriate filters. This term is used here to describe the possible overlap of points in the upper and lower regions of the KMD/NKM 2D-plot. Aliasing can be prevented by plotting regular KMD (RKMD) instead of KMD since RKMD is the result of a numerical shift of KMD that is computed to fit spectral width [22]. This method of mass filtering has been successfully applied to the analysis of compound complex mixtures in petroleomics [23], in polymer chemistry [24] or to the evaluation of water treatment [25]. To our knowledge, it is applied here for the first time to the analysis of microbial metabolites and compounds such as NRPs. Recently, the concept of resolution-enhanced KMD plot taking advantage of the whole spectral width has been proposed for a better separation of the different ion series [24].

In this paper, we demonstrate that combining the Kendrick approach with information stored in the NORINE database is relevant for secondary metabolite identification, using HRMS data of commercially available NRPs. This method is easy-to-use, fast, and useful for dereplication, as well as screening and potential discovery of new bioactive molecules. Firstly, we calculated the theoretical monoisotopic masses of all NORINE compounds using the molecular formulae referenced in the database. Secondly, we ran classical software to compute NKM and RKMD values from the theoretical monoisotopic masses of NORINE compounds. Thirdly, we developed a dedicated tool that generates an interactive RKMD/NKM 2D-plot (i.e., Kendrick-based NORINE map) and performs prediction from *m/z* values. Finally, the approach was validated with accurate masses of commercially available NRPs measured by FT-ICR-MS. These experimental masses were processed and plotted on the Kendrick-based NORINE map to identify their molecular formulae. As the results matched expected compositions, the approach holds promise for identifying new high value compounds in different fields (i.e., health, cosmetic, agrofood, and biocontrol).

Material and Methods

Ultrahigh-Resolution Mass Spectrometry (HRMS)

FT-ICR MS Analysis MS analyses were performed using a Bruker 9.4 Tesla SolariX FT-ICR MS equipped with an ESI/MALDI Dual Ion Source including Smartbeam™ II laser (Bruker Daltonics, Bremen, Germany). Mass spectra were externally calibrated in positive mode using a solution of phosphoric acid (Sigma) in 50/50 (*v*/v) ACN/H_2O at 0.1%. A commercially available surfactin mixture (Sigma) was dissolved at 1 mM in 50/50 (*v*/v) ACN/H_2O, co-crystallized with the matrix solution (10 mg/mL of α-cyano-4-hydroxycinnamic acid in (50/49.9/0.1 (v/v/v) acetonitrile (ACN)/H_2O/trifloroacetic acid (TFA)) onto a polished steel MALDI target (Bruker Daltonics) and dried at room temperature. MALDI mass spectra were acquired in positive ion mode from 100 laser shots. The laser power was set to 100% with a frequency of 1000 Hz. For broadband detection mode analyses, mass range was set to *m/z* 72.2–3500 and time of flight value was 2 ms. Q1 mass was fixed at *m/z* 1200. Ion cooling time was set to 0.01 s. For narrowband detection mode analyses, center mass was set to *m/z* 1046 ± 13.9. Monoisotopic masses from acquired mass spectra were labeled using DataAnalysis 4.0 software (Bruker Daltonics) with the FTMS peak-picking algorithm with default parameters.

Calculation of Theoretical Monoisotopic Masses

The theoretical monoisotopic and theoretical protonated monoisotopic ([M + H^+]) masses of compounds were calculated from molecular formulae referenced in the NORINE database (http://bioinfo.lifl.fr/norine) using the IUPAC 2013 index and Chemcalc free software (www.chemcalc.org). The "molecular formula finder" tool (http://www.chemcalc.org/mf_finder) of the Chemcalc software suite was run to generate a list of molecular formulae, with the following parameters: range, C0–100 H0–100 N0–20 O0–20; limit the results by unsaturations, unsaturations allowed from 0 to 999; and mass error of 0.001 Da.

Calculation of Kendrick Mass (KM); Nominal Kendrick Mass (NKM) and Regular Kendrick Mass Defect (RKMD)

The Kendrick mass (KM) related to the CH_2 pattern is calculated from the molecular formula of NORINE-referenced compounds and from the experimentally measured masses. In agreement with petroleum or polymer analysis, dealiasing was performed by shifting the KM value by −0.28 (supplemental data 1, eq. (1)). The dealiased KM value rounded to the

nearest integer defines the nominal Kendrick mass (NKM)—(supplemental data 1, eq. (2)). NKM subtracted from the Kendrick mass defines the regular Kendrick mass defect (RKMD)—(supplemental data 1, eq. (3)). The RKMD and NKM values (calculated for each known and curated molecular formulae in the NORINE database) were finally plotted to generate the RKMD/NKM 2D-plot with no aliasing.

Variation of RKMD (ΔRKMD), NKM (ΔNKM) and Kendrick Trigonometric Mesh

The difference between the respective RKMD and NKM values of two points of the RKMD/NKM 2D-plot defines the corresponding RKMD and NKM variations (Δ). More precisely, the addition or subtraction of an atom or atom group will always generate the same RKMD variation (ΔRKMD) and the same nominal Kendrick mass variation (ΔNKM). By definition, if CH_2 is added to a reference point in the 2D-plot then ΔRKMD = 0 and the ΔNKM value forms a horizontal line. In all other cases, any two points close to a reference point are the apexes of a right-angled triangle connected by ΔNKM and ΔRKMD values and whose hypotenuse is the line linking the two close points. The hypotenuse value (V) is calculated from the Pythagorean Theorem (supplemental data 1, eq. (5)). The theta (θ) angle value at the reference point is the result of trigonometrical equations using cosine (supplemental data 1, eqs. (6) and (7)). The Kendrick trigonometric mesh of a point in the 2D-plot is the set of vectors of length V connecting that point to all surrounding points and forming a θ angle with respect to the horizontal line defined by ΔRKMD = 0.

Results

Assignment of Monoisotopic Mass for all Compounds Annotated in NORINE

The NORINE database is composed of 1187 entries. Each entry contains manually curated information (structure, activity, family, producing organisms) collected from the scientific literature and related to a single NRP. The molecular formulae of all NRPs were extracted from the NORINE database and used as input of Chemcalc to calculate their IUPAC theoretical monoisotopic mass. These masses were incorporated in NORINE and the IUPAC theoretical protonated monoisotopic masses were used to create the Kendrick-based NORINE map, corresponding to the RKMD/NKM 2D-plot.

KMD Approach Applied to Surfactin Variants

Variants of surfactins such as [Ala4] nC14, [Ile7] nC14, iC15 and [Val7] iC15 have been used as a proof of concept as well as to demonstrate the advantages of the KMD approach (Figure 1). The exact theoretical monoisotopic masses of these surfactins are 1021.667491263 (C52H91N7O13) for [Val7] iC15, [Ala4] nC14, and [Ile7] nC14 and 1035.683136097 (C53H93N7O13) for iC15. The structural difference affects the peptide core (e.g., iC15/[Val7] iC15) or the aliphatic chain (e.g., iC15/[Xaa] nC14) of the lipopeptides. Surfactins with the same mass display the same isotopic distribution (supplemental data 2A). For compounds with such molecular mass, the measurement accuracy of HRMS provides a monoisotopic $[M+H]^+$ of 1022.67476 at best. Using the corresponding (uncharged) monoisotopic mass (1021.66749) and a mass accuracy of less than 1 ppm, software like Chemcalc that correlates a monoisotopic mass with a molecular formula outputs seven distinct candidate molecular formulae (as illustrated in supplementary data 2B). The 2D-plot representing the RKMD as a function of NKM (RKMD/NKM 2D-plot in relation to the CH_2 building block) of the seven possible molecular formulae shows that the corresponding seven points are vertically aligned (supplemental data 2C).

The NRPs are divided in six structurally related classes: lipopeptides, peptides, peptaibols, glycopeptides, chromopeptides, and peptides with a polyketide-NRP moiety. For example, lipopeptides can involve the same peptide moiety but differ by the length of fatty acid chains. These shared structural properties are an asset for the identification of unknown compounds. Molecules of the same family are characterized by a CH_2-based structural correlation (no RKMD variation) and therefore horizontally aligned in the RKMD/NKM 2D-plot.

Kendrick Vector Mesh

The RKMD/NKM 2D-plot can be extended to all NRPs (i.e., all molecular formulae) extracted from NORINE. Therefore, whatever the compound of interest, the addition or deletion of an atom or an atom group (e.g., ± CH_2, ± O, ± N…) generates the same RKMD variation (ΔRKMD) and the same nominal Kendrick mass variation (ΔNKM). The addition or the loss (in the molecular formula) of a nitrogen atom (black line) causes a ΔNKM of 14 but also a ΔRKMD of 0.0126 whatever the molecule (Figure 2a). Note that the subtraction (or addition) of an atom or atom group corresponds to a decrease (or increase) of the NKM value (see Figure 2a). Two close points of the RKMD/NKM 2D-plot are related to each other through a right-angled triangle (except for a CH_2 variation that forms, by definition, only a horizontal line (blue line) since ΔRKMD = 0) defined by the ΔNKM value line, the ΔRKMD value line, and the line joining the two points that forms the hypotenuse. For example, irrespective of the compound, the addition of one oxygen atom (red straight line, Figure 2a) to a molecular formula corresponds trigonometrically to two points, forming the hypotenuse of a right-angle triangle and a θ angle (value = 55°) with respect to the horizontal line. In contrast, the subtraction of one oxygen atom (inverse red straight line, Figure 2a) from a molecular formula corresponds trigonometrically to two points forming the hypotenuse of a right-angle triangle and the same θ angle (value = 55°) with respect to the horizontal line. In the end, due to the calculation mode of the NKM, molecular formulae differing by an atom or atom group (other than CH_2) are diagonally correlated. This is illustrated in Figure 2b where

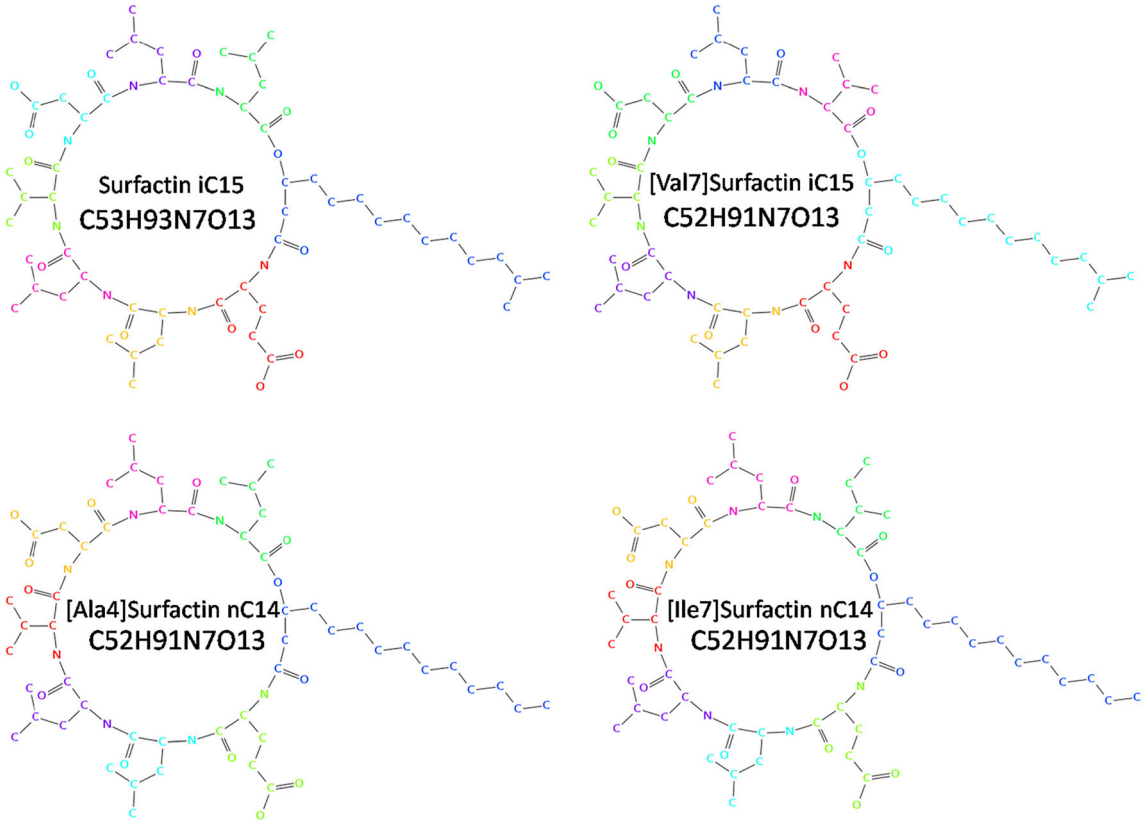


Figure 1. Chemical structure of four variants of surfactins used as models: [Ala4] nC14, [Ile7] nC14, iC15, and [Val7] iC15

the addition or subtraction of one oxygen, one hydroxyl group, and one hydrogen are displayed in red, green, and dark, respectively. The addition or the subtraction of one hydroxyl group (green straight line) corresponds to two points of the RKMD/NKM 2D-plot (θ angle = 44° with respect to the horizontal axis) while the addition or the subtraction of one nitrogen (black straight line) corresponds to two distinct points (θ angle = 42° with respect to the horizontal axis). The RKMD/NKM 2D-plot is being based on a CH_2 building block, the addition or subtraction of CH_2 does not change RKMD (ΔRKMD = 0) and changes NKM by 14 Da which translates into a horizontal correlation on the plot. Therefore, the positioning of known NRPs on the RKMD/NKM 2D-plot based on molecular formulae provides a new and fast means of identifying new microbial secondary metabolites, especially NRPs.

Creation of Kendrick-Based NORINE Map

NORINE is recognized as the unique reference database of curated information relative to NRPs and, as such, provides a good coverage of the possible molecular formulae of NRPs. From these formulae, the theoretical protonated monoisotopic masses were calculated and plotted on the RKMD/NKM 2D-plot (Figure 3) with respect to a CH_2 mass defect. This 2D-plot was called the Kendrick-based NORINE map. The molecular mass of NRPs ranges from 200 to 3000 Da. The vast majority lies between 500 and 1500 Da and forms one cloud of dense points and one of scattered points. The six classes of molecules that were listed earlier (lipopeptides, peptides, peptaibols, glycopeptides, chromopeptides, and peptides with a polyketide-NRP moiety) are depicted using different colors, showing that lipopeptides (red circles), peptides (blue squares), and glycopeptides (violet circles) are globally distributed over the entire RKMD range. The polyketide-NRP hybrids (orange stars) are distributed on the RKMD axis from − 0.25 to 0.00 for a NKM close to 1000 Da. Peptaibol (green triangles) cover the mass range 1000 to 2000 Da for a RKMD from − 0.25 to 0.40. Chromopeptides (in black) are localized between 1000 and 1500 Da for a RKMD from 0.00 to 0.25. This distribution is influenced by the variation of class sizes. Peptide and lipopeptide classes represent 42.2% and 25.1% of the NORINE compounds, respectively. As expected, applying resolution-enhanced KMD improves the separation of the different ion series over the whole spectral width but does not provide a clear-cut definition of NRP classes.

Proof of Concept Based on the Surfactin Family NRPs

Monoisotopic masses from acquired mass spectra were labeled and plotted on both the complete Kendrick-based NORINE (Figure 4a) and the depleted Kendrick-based NORINE maps where surfactins of interest had been erased (Figure 4b). Four monoisotopic masses were extracted from the HRMS spectrum (supplemental data 3), each having the following [NKM; RKMD] coordinates on the RKMD/NKM 2D-plot: [994; −

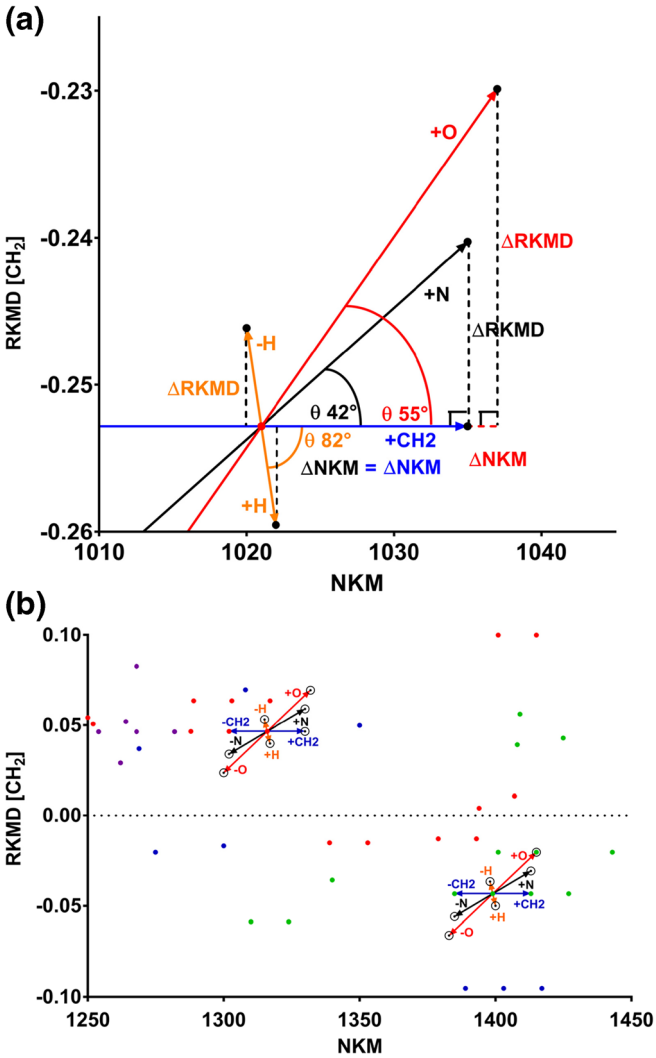


Figure 2. RKMD/NKM 2D-plot correlated to the CH_2 building block (**a**) illustrating the vector (defined by the theta angle and length of the vector) obtained for each increment of a given atom or atom group and (**b**) the vector mesh around the trigonometric circle enabling the connection between the molecular formulae of known compounds to those of other compounds

0.2531], [1008; − 0.2531], [1022; − 0.2527], and [1036; − 0.2534], respectively, as shown in Figure 4a and b (red points). As expected, the four surfactins perfectly match coordinates of the NORINE compounds whose molecular formulae are C50H87N7O13, C51H89N7O13, C52H91N7O13, and C53H93N7O13, respectively (Figure 4a, red circles around blue points). Obviously, these molecular formulae match those of surfactins where variations express different fatty acid chain lengths. Conversely, none of the coordinates of the four tested surfactins matches any point in the map of the NORINE database depleted of surfactin family members. As a result, these molecular formulae remain unknown only showing differences in one $^{12}C_1{}^1H_1$ group with known compounds. Nonetheless, a single point of the depleted NORINE-based Kendrick map can be reached via the trigonometric mesh. The vectors defined by the hypotenuse and θ angle value pairs connect

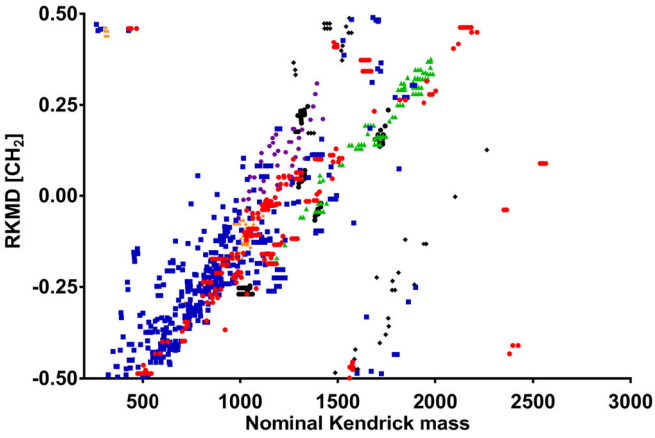


Figure 3. RKMD/NKM 2D-plot (Kendrick-based NORINE map) of all NRPs referenced in the NORINE database. The molecule classes are depicted using colors: lipopeptides (red circles), peptides (blue squares), polyketide-NRP hybrids (orange stars), chromopeptides (in black), glycopeptides (violet circles), and peptaibols (green triangles)

unknown to known points via the difference in numbers of atoms. For example, from the known C53H94N8O12 (central blue point circled in blue) three steps shown by three vectors are required (addition of one oxygen (^{16}O) in red, subtraction of one nitrogen (^{14}N) in black and subtraction of one hydrogen (^1H) in orange) to reach the unknown compound point labeled [1036; − 0.2534] (red point circled in red). Therefore, the putative molecular formula of the [1036; − 0.2534] coordinates is C53H93N7O13. In the end, the correlation between members of this family arises from the atomic difference of $^{12}C_1{}^1H_1$, as illustrated in Figure 4b (blue arrows). Note that several paths may connect two points but they all lead to the same molecular formula. The assignment of molecular formulae to experimental masses was automated.

Software Application

A web application (Supplemental data 4) implementing the method described above was developed in Java (back-end) and Javascript, HTML and CSS (front-end). Given a mass-to-charge ratio, it suggests a set of chemical formulae predicted using the Kendrick approach. Additional input parameters include the mass defect and the database (pic.webApp).The mass defect determines the precision of the search while the database option increases the search space with an additional set of masses retrieved from PubChem [26]. This data set was carefully selected using the ontology search provided by the PubChem Classification Browser with the aim of obtaining NRPs and NRP-like compounds. The predicted formulae are presented in a table specifying the mass defect of each formula and linked to a graphical representation of the RKMD/NKM 2D-plot. As shown in the Supplemental data 4, the point and the vector used for each prediction are highlighted and a color code specifies the origin of the data, as indicated in the legend of the plot (note that the green "All" box refers to the masses present in both databases). The software successfully predicted

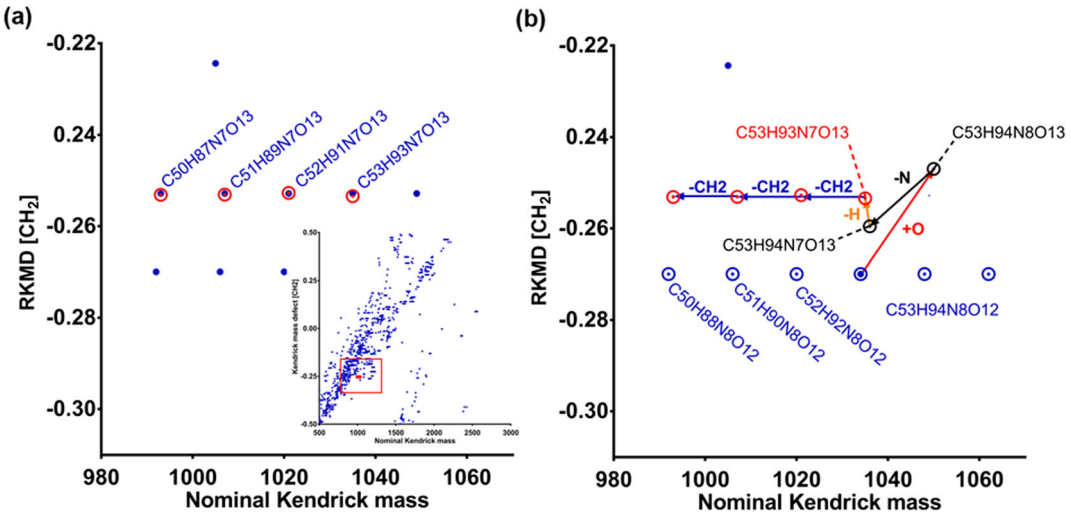


Figure 4. RKMD/NKM 2D-plot displaying the position of points corresponding to the four commercially available surfactins (**a**) on the complete Kendrick-based NORINE map and (**b**) a depleted Kendrick-based NORINE map where the surfactins of interest have been erased. In (**a**), the positioning of the four surfactins on the map perfectly matches the molecular formulae of known compounds of the NORINE database while, in (B), there is no match. Molecular formulae are deduced step-by-step using the Kendrick vector mesh. As illustrated, the vector path connecting the C53H94N8O12 molecular formula of the NORINE known compound to the unknown compound goes through the addition of one oxygen and the subtraction of one nitrogen and one hydrogen. Therefore, the molecular formula of the unknown compound is C54H93N7O12

the molecular formulae of all members of the surfactin family previously described. The tool can be run on-line with any list of mass-to-charge ratios at the following URL: http://bioinfo.cristal.univ-lille.fr/kendrick-webapp/

Discussion

At present, the identification of known (aka, dereplication) or unknown NRPs produced by a microorganism still remains a time consuming, expensive, and challenging task [27] regardless of the application field. The first step of the structural elucidation process often consists in measuring as exactly as possible (less than 1 ppm) the molecular mass of compounds to deduce their molecular formulae from the measured monoisotopic mass. Obviously, a molecular formula does not lead to solving the structure but it is the starting point that guides the gradual elucidation of the structure. But, for compounds of molecular mass > 1000 Da, the relative imprecision of high-resolution mass spectrometers prevents the computer-assisted deduction of a unique molecular formula. Similarly, the use of isotopic distribution and of the Seven golden [11] rules eliminates 90% of molecular formula candidates but does not lead to only one [16]. However, the reduction of the number of molecular formula candidates brings a definite advantage to dereplication and accelerates structure elucidation.

In the Kendrick approach, all compounds, whose molecular formulae differ from one or several $^{12}C_1{}^1H_2$, share the same zero-value RKMD (the compounds are aligned on a horizontal line). When compounds are horizontally correlated by one or more $^{12}C_1{}^1H_2$, structural kinship cannot be easily ascertained as true or false. In addition, compounds that differ by the same molecular formula increment (e.g., addition or deletion of one nitrogen, one oxygen, one hydroxyl group, a sodium or potassium adduct…) have the same ΔRKMD and the same ΔNKM. Of course, when an atom or an atom group is added (or subtracted) to a given molecular formula, the NKM value increases (or decreases). The addition or subtraction of a given atom is represented by a vector joining two formulae and forming a θ angle with the horizontal axis. From a given point (molecular formula) of the plot, a mesh of vectors that connects this point to the close surrounding points can be determined. This vector mesh allows connecting one unknown to a known molecular formula to support the identification of new compounds. Therefore, plotting scaled RKMD versus NKM (RKMD/NKM 2D-plot) offers two advantages: firstly, it highlights the horizontal alignment of compounds related to the reference atomic group (i.e., $^{12}C_1{}^1H_2$), such as homologs or members of a same lipopeptide family; secondly, the superposition of the vector mesh on a given point of the RKMD/NKM 2D-plot connects two close molecular formulae. In our example, this plot led to the identification of surfactin variants as [Ala4] nC14, iC15, [Val7] iC15, and [Ile7] nC14. In a nutshell, the RKMD/NKM 2D-plot supports the rapid correlation of close compounds on the basis of their difference in atomic composition.

The general KMD approach is widely used for the study and structural characterization of chemical polymers such as petroleum compounds [23–25] that only vary by the number of CH_2 groups. In these studies, the determination of molecular formulae is based on the identification of horizontally correlated compound series in the RKMD/NKM 2D-plot, followed by molecular regression. This process is based on the isotopic profiles of low mass compounds and implemented in MS

software that outputs the formula, but it is not directly applicable to the determination of molecular formulae of NRPs, whose structures are built from more than 500 monomers [1]. As seen in the case of surfactins, this diversity of building blocks limits the existence of structural variants in a family and, consequently, of low mass compounds. Therefore, the combination of a chemical or biochemical database (such as PubChem, ChemSpider [26, 28]) with the RKMD/NKM 2D-plot is suggested as a simple and efficient way to annotate NRPs. The molecular formulae of known compounds can be plotted and serve as reference points to deduce the molecular formulae of neighboring points. To our knowledge, such a combination has not been published before. We refined the method further by using NORINE, the unique NRP database that gathers almost 1200 compounds collected from the literature and manually curated. Information about NRPs in NORINE is extensive and includes molecular formulae that we extracted and plotted on the RKMD/NKM 2D-plot to serve as annotations.

In this article, we demonstrated the interest of such a combination with the HRMS analysis of a commercially available surfactin mixture, the calculation of RKMD and NKM values from the monoisotopic mass of each surfactin, and their location on two distinct RKMD/NKM 2D-plots. This example illustrates the ability of this method to generate a single molecular formula from high-resolution mass spectrometry.

Conclusion

Dereplication plays an essential role in the discovery process of new NRPs. Compounds produced by microorganisms are primarily subjected to HRMS analysis (with or without previous separation) to measure as exactly as possible the molecular mass and the isotopic distribution of compounds. The molecular formula of produced compounds can be deduced from the *m/z* (z = 1) using an approach that combines regular Kendrick mass defect calculations with knowledge stored in the NORINE database. The known compounds of the NORINE database then support the RKMD/NKM 2D-plot annotation. In the end, the *m/z* of a compound is represented as coordinates in the RKMD/NKM 2D-plot and the corresponding molecular formula can be deduced, either by matching the coordinates of a known NORINE compound or by determining a vector connection to a neighboring point.

The web tool implementing the method provides a user-friendly and interactive interface, and its predictive function can benefit the NRPomics community. Furthermore, the optional usage of PubChem data demonstrates that the Kendrick map approach can be extended to enhance the quality of prediction.

Acknowledgements

This work has been carried out in the framework of Alibiotech project which is financed by the European Union, French State, and the French Region of Hauts-de-France. Authors would like to thank the European Union funding through the INTERREG Va FWVL BioScreen/SmartBioControl Project. This work has been carried out thanks to the support provided by the Ministries of Europe and Foreign Affairs (MEAE) and Higher Education, Research and Innovation (MESRI) through the program Hubert Curien, Germaine de Staël. Emma Ricart is supported by the SIB Swiss Institute of Bioinformatics Fellowship program. We are grateful to Dr. Areski Flissi for his technical assistance and help in updating the NORINE database.

References

1. Caboche, S., Pupin, M., Leclère, V., Fontaine, A., Jacques, P., Kucherov, G.: NORINE: a database of nonribosomal peptides. Nucleic Acids Res. **36**, D326–D331 (2007)
2. Flissi, A., Dufresne, Y., Michalik, J., Tonon, L., Janot, S., Noé, L., Jacques, P., Leclére, V., Pupin, M.: Norine, the knowledgebase dedicated to non-ribosomal peptides, is now open to crowdsourcing. Nucleic Acids Res. **44**, D1113–D1118 (2016)
3. Schrader, M., Schulz-Knappe, P., Fricker, L.D.: Historical perspective of peptidomics. EuPA Open Proteomics. **3**, 171–182 (2014)
4. Nicholson, J.K., Lindon, J.C.: Metabonomics. **455**, 1054–1056 (2008)
5. Nicholson, J.K., Lindon, J.C., Holmes, E.: "Metabonomics": understanding the metabolic responses of living systems to path physiological stimul via multivariate statistical analysis of biological NMR spectroscopic data. xenobiotica. **29**, 1181–1189 (1999)
6. Fiehn, O.: Combining geonomics, metabolome analysis, and biochemical modelling to understand metabolic networks. Comp. Funct. Genom. **2**, 155–168 (2001)
7. Hubert, J., Nuzillard, J.M., Renault, J.H.: Dereplication strategies in natural product research: how many tools and methodologies behind the same concept? Phytochem. Rev. **16**, 55–95 (2017)
8. Marston, A., Hostettmann, K.: Natural product analysis over the last decades. Planta Med. **75**, 672–682 (2009)
9. Cho, Y., Ahmed, A., Annana, I., Kim, S.: Developments in FT-ICR MS instrumentation, ionization techniques, and data interpretation methods for petroleomics. Mass Spectrom. Rev. 221–235 (2014). https://doi.org/10.1002/mas.21438
10. Glish, G.L., Burinsky, D.J.: Hybrid mass spectrometers for tandem mass spectrometry. J. Am. Soc. Mass Spectrom. **19**, 161–172 (2008)
11. Kind, T., Fiehn, O.: Seven golden rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. BMC Bioinformatics. **8**, 105 (2007)
12. Rogers, S., Scheltema, R.A., Girolami, M., Breitling, R.: Probabilistic assignment of formulas to mass peaks in metabolomics experiments. Bioinformatics. **25**, 512–518 (2009)
13. Werner, E., Heilier, J.-F., Ducruix, C., Ezan, E., Junot, C., Tabet, J.-C.: Mass spectrometry for the identification of the discriminating signals from metabolomics: current status and future trends. J. Chromatogr. B Anal. Technol. Biomed. Life Sci. **871**, 143–163 (2008)
14. Grange, A.H., Genicola, F.A., Sovocool, G.W.: Utility of three types of mass spectrometers for determining elemental compositions of ions formed from chromatographically separated compounds. Rapid Commun. Mass Spectrom. **16**, 2356–2369 (2002)
15. Grange, A.H., Winnik, W., Ferguson, P.L., Sovocool, G.W.: Using a triple-quadrupole mass spectrometer in accurate mass mode and an ion correlation program to identify compounds. Rapid Commun. Mass Spectrom. **19**, 2699–2715 (2005)
16. Böcker, S., Letzel, M.C., Lipták, Z., Pervukhin, A.: SIRIUS: decomposing isotope patterns for metabolite identification. Bioinformatics. **25**, 218–224 (2009)
17. Dittwald, P., Burzykowski, T., Valkenborg, D., Gambin, A.: BRAIN: a universal tool for high-throughput calculations of the isotopic distribution for mass spectrometry. Anal. Chem. (2013). https://doi.org/10.1021/ac303439m
18. Meija, J., Coplen, T.B., Berglund, M., Brand, W.A., De Bièvre, P., Gröning, M., Holden, N.E., Irrgeher, J., Loss, R.D., Walczyk, T., Prohaska, T.: Atomic Weights of the Elements 2013 (IUPAC Technical Report) (2016). https://doi.org/10.1515/pac-2015-0305

19. Patiny, L., Borel, A.: ChemCalc: a building block for tomorrow ' s chemical infrastructure. J. Chem. Inf. Model. 1–21 (2013). https://doi.org/10.1021/ci300563h
20. Sleno, L.: The use of mass defect in modern mass spectrometry. 226–236 (2012). https://doi.org/10.1002/jms.2953
21. Kendrick, E.: A mass scale based on CH 2 = 14.0000 for high resolution mass spectrometry of organic compounds. Anal. Chem. **35**, 2146–2154 (1963)
22. Fouquet, T.N.J., Cody, R.B., Ozeki, Y., Kitagawa, S., Ohtani, H., Sato, H.: On the Kendrick mass defect plots of multiply charged polymer ions: splits, misalignments, and how to correct them. J. Am. Soc. Mass Spectrom. 1–16 (2018). https://doi.org/10.1007/s13361-018-1972-4
23. Roach, P.J., Laskin, J.: And, Laskin, a.: higher-order mass defect analysis for mass spectra of complex organic mixtures. Anal. Chem. **83**, 4924–4929 (2011)
24. Fouquet, T., Sato, H.: Improving the resolution of Kendrick mass defect analysis for polymer ions with fractional base units. Mass Spectrom. **6**, A0055–A0055 (2017)
25. Ohno, T., Parr, T.B., Gruselle, M.-C.I., Fernandez, I.J., Sleighter, R.L., Hatcher, P.G.: Molecular composition and biodegradability of soil organic matter: a case study comparing two New England forest types. Environ. Sci. Technol. **48**, 7729–7236 (2014)
26. Kim, S., Thiessen, P.A., Bolton, E.E., Chen, J., Fu, G., Gindulyte, A., Han, L., He, J., He, S., Shoemaker, B.A., Wang, J., Yu, B., Zhang, J., Bryant, S.H.: PubChem substance and compound databases. Nucleic Acids Res. **44**, D1202–D1213 (2016)
27. Yang, J.Y., Sanchez, L.M., Rath, C.M., Liu, X., Boudreau, P.D., Bruns, N., Glukhov, E., Wodtke, A., De Felicio, R., Fenner, A., Wong, W.R., Linington, R.G., Zhang, L., Debonsi, H.M., Gerwick, W.H., Dorrestein, P.C.: Molecular networking as a dereplication strategy. J. Nat. Prod. **76**, 1686–1699 (2013)
28. Editorial: ChemSpider – a tool for Natural Products research, (2015). https://doi.org/10.1039/c5np90022k

Supplemental data 1

Calculation of Kendrick Mass (KM); nominal Kendrick mass (NKM) and regular Kendrick mass defect (RKMD)

The dealiased Kendrick mass (dKM) related to the CH_2 pattern is calculated from the molecular formula of NORINE-referenced compounds and from the experimentally measured masses using the following equation:

$$dKM = \left(IUPAC\ protonated\ monoisotopic\ mass\ from\ NORINE\ or\ experimentally\ measured\ {}^{m}/_{z} \times \left(\frac{CH_2\ nominal\ mass}{CH_2\ exact\ mass} \right) \right) - 0.28 \quad (1)$$

The dealiased KM value is then rounded to the nearest whole number (as mentioned in the equation below) and defines the nominal Kendrick mass (NKM):

$$Nominal\ Kendrick\ mass\ (NKM) = \ dKM\ rounded\ to\ the\ nearest\ whole\ number \quad (2)$$

The NKM is then subtracted from Kendrick mass to obtain the regular Kendrick mass defect (RKMD).

$$Regular\ Kendrick\ mass\ defect(RKMD) = \ KNM - dKM \quad (3)$$

The value of the RKMD and the NKM (calculated for each known and curated molecular formulae extracted from the NORINE database) were finally plotted to generate the RKMD/NKM 2D-plot.

Variation of KMD (ΔKMD), variation of NKM (ΔNKM) and Kendrick trigonometric mesh

The variations (Δ) of RKMD and NKM between two points of the RKMD/NKM 2D-plot are defined as the difference between the RKMD and NKM values of each point. The addition or subtraction of an atom or an atom group always generates the same value of the regular

Kendrick mass defect variation (ΔRKMD) and the same nominal Kendrick mass variation (ΔNKM). Therefore, except for a CH_2 variation that forms, by definition only a horizontal line, two close points are related to each other through a right-angled triangle where the ΔNKM value forms a side of the right-angled triangle, the ΔRKMD value forms the other side and finally, the line linking the two points forms the hypotenuse. The hypotenuse value is therefore calculated from the Pythagorean Theorem using the following equation:

$$Hypotenuse = \sqrt{(\Delta RKMD^2 + \Delta NKM^2)} \quad (4)$$

The angle values at the reference point are calculated according to the equations below:

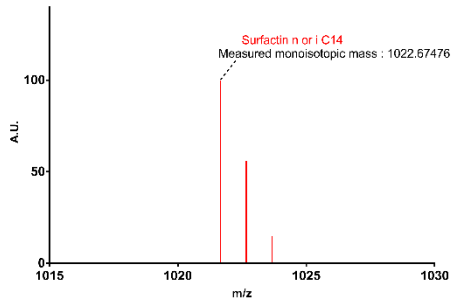
$$\text{Cos}\ (\Delta\ atom\ or\ atom\ group) = \frac{\Delta NKM\ (atom\ or\ atom\ group)}{Hypothenuse\ (atom\ or\ atom\ group)} \quad (5)$$

$$\theta\ (\Delta\ atom\ or\ atom\ group) = \cos^{-1}(\Delta\ atom\ or\ atom\ group) \quad (6)$$

In fine, at the RKMD/NKM 2D-plot level, the Kendrick trigonometric mesh is defined as all portions of straight lines defined by an angle and the hypotenuse value, linking a central point to all surrounding points, defining a vector.

Supplemental data 2: A) Theoretical isotopic distribution of n or iC14 surfactins. B) Table summarizing **i**) the molecular formulae, **ii**) the theoretical monoisotopic masses, **iii**) the mass differences (ppm) between the potential molecular formulae and the theoretical monoisotopic mass of n or iC14 surfactins, **iv**) the Kendrick mass, **i**) the KNM and **iv**) the KMD for the seven distinct candidate molecular formulae proposed by Chemcalc (mass difference of less than 1 ppm. 2C) 2D-plot representing the KMD as a function of NKM (RKMD/NKM 2D-plot in relation to the CH_2 building block) of the seven possible molecular formulae.

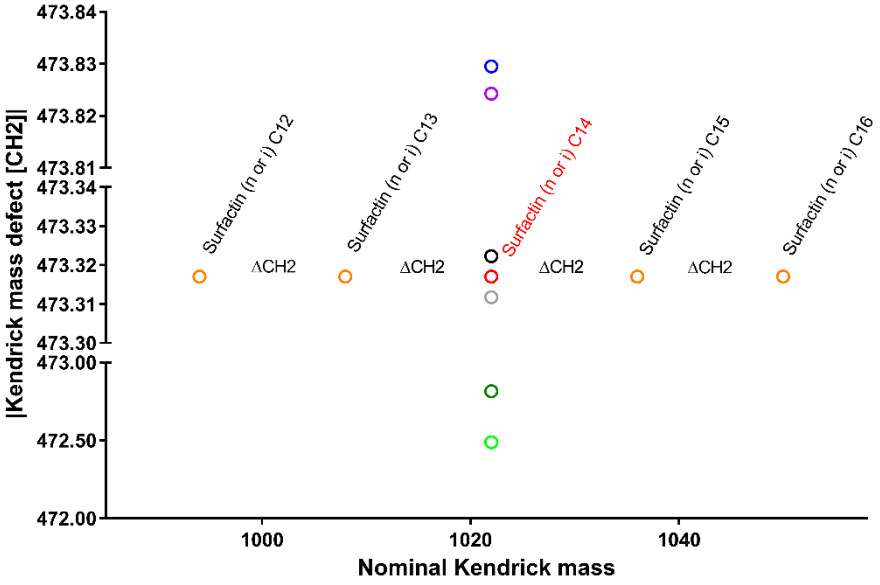
A



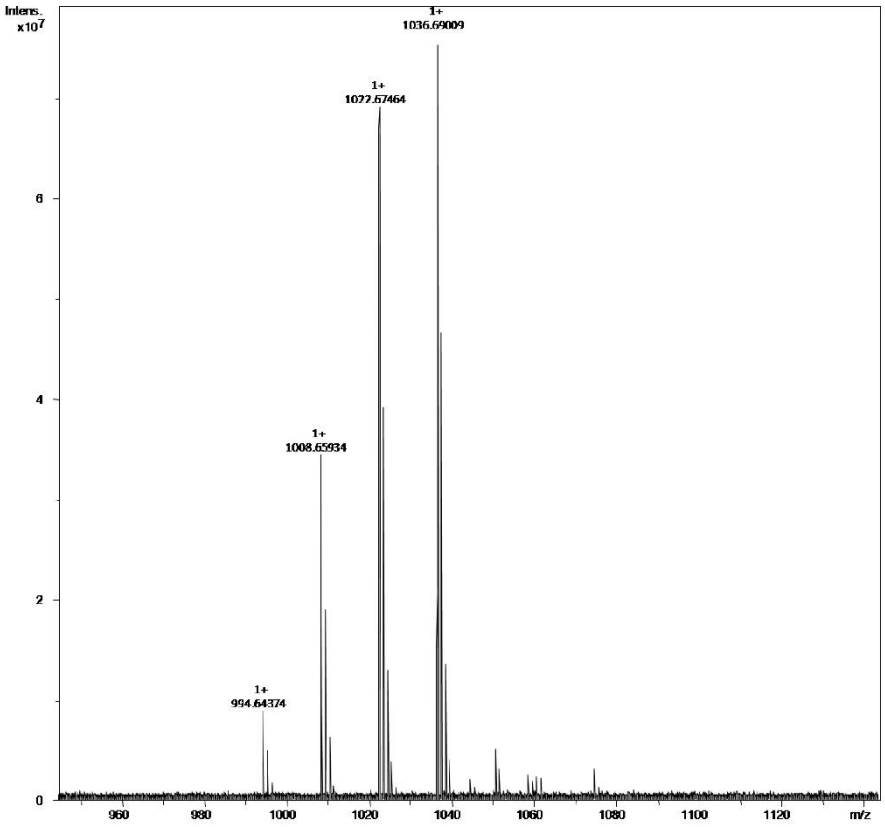
B

#	Molecular formula	Theorical monoisotopic mass	Δ (ppm)	Kendrick mass (KM)	Kendrick nominal mass	Kendrick mass defect (KMD)
1	C65H83N9O2	1021.66697	0.506	1020.52617	1021	473.82952
2	C66H89N2O7	1021.66698	0.501	1020.52618	1021	473.82427
3	C51H85N14O8	1021.66748	0.009	1020.52668	1021	473.32229
4	C52H91N7O13	1021.66749	0.004	1020.52668	1021	473.31704
5	C53H97O18	1021.66749	0.001	1020.52669	1021	473.31178
6	C37H87N19O14	1021.66799	0.488	1020.527185	1021	472.81505
7	C67H85N6O3	1021.66832	0.808	1020.527512	1021	472.48835

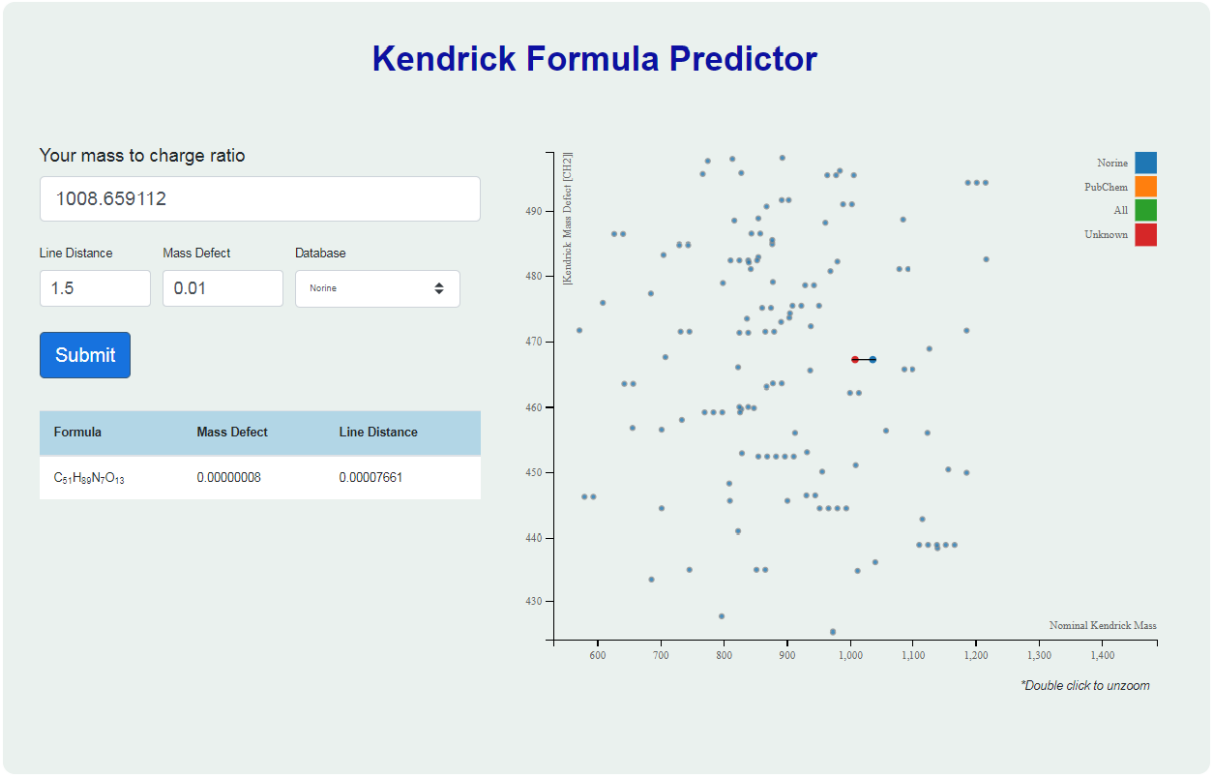
C



Supplemental data 3: MALDI FT-mass spectra of a commercially available surfactin mixture dissolved at 1 mM in 50/50 (v/v) ACN/H2O, co-crystallized with the matrix solution (10 mg/mL of α-cyano-4-hydroxycinnamic acid in (50/49.9/0.1 (v/v/v) acetonitrile (ACN)/H2O/trifloroacetic acid (TFA)) onto a polish steel MALDI target (Bruker Daltonics) and dried at room temperature. See M&M for the details related to mass measurements.



Supplemental data 4: web application interface (http://bioinfo.cristal.univ-lille.fr/kendrick-webapp/)



3.2 Concluding Remarks

Applying the KMD to NRP formula deduction has proved to be useful for all the tested surfactins. These results are highly promising and the current version of the tool is convenient for manual analysis of single spectra. The simplicity and fast performance of such an approach compared to the classical isotope pattern method will facilitate the processing of high-throughput data. However, further testing is required before automating it in order to work with larger amounts of MS data.

Chapter 4

NRPro

4.1 Overview

Tandem mass spectrometry has been since long used for the identification and characterization of PNPs. The development of software to support the dereplication process resulted in the release of tools such as iSNAP, Dereplicator or Cyclobranch. Apart from the individual drawback of each one of these options, we realized that none of them was providing a comprehensive toolset that integrates statistically validated identification as well as user friendly features to explore and customize the results. That is what drove us to the development of NRPro, the tool presented in this chapter. NRPro is a highly interactive dereplication platform with a strong focus on spectral annotation. The software is presented together with a new nomenclature for PNPs which is used for the automatic annotation of the MS/MS spectra. The fragmentation algorithm is able to process any type of PNP structure and takes into consideration amide, ester and glycosidic bonds as well as multiple fragment ion type forms (y/b, x/a and c/z), neutral losses and adducts. The scoring involves the computation of fragmentation trees and the statistical significance of the matches is based on a target-decoy approach. The graphical interface is one of the strong points of NRPro, including a depiction of the candidate molecule, the experimental spectra and a table with the annotations. Visual effects are used to connect these components in order to facilitate the association between each peak and the corresponding fragment ion. In the following study we use Dereplicator and Cyclobranch (tools described in Section 1.4.4) to analyse and compare the dereplication and annotation capacities of NRPro.

analytical chemistry

pubs.acs.org/ac Article

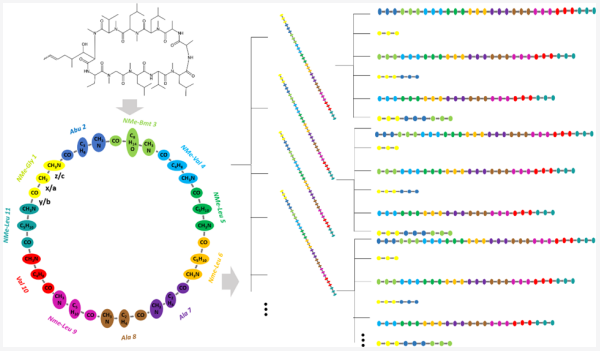
Automatic Annotation and Dereplication of Tandem Mass Spectra of Peptidic Natural Products

Emma Ricart, Maude Pupin, Markus Müller, and Frédérique Lisacek*

Cite This: https://dx.doi.org/10.1021/acs.analchem.0c03208 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The various bioactivity types and potencies of peptidic natural products (PNPs) are of high interest for the development of new drugs. In particular, the intrinsic antibiotic properties of PNPs appear essential to combat antimicrobial resistance that is currently threatening the world. The first steps in dereplication and characterization of PNPs often involve tandem mass spectrometry (MS/MS). However, such structurally complex peptides challenge the interpretation of MS/MS results. Only a few software solutions are dedicated to PNP analysis but with a mutually exclusive focus on dereplication or annotation. Hence, key functionalities such as automatic peak annotation or statistically validated scoring systems to support the characterization/identification processes are missing. Here, we present NRPro, a new MS/MS analysis platform that overcomes some limitations of the existing software and provides a comprehensive toolset for both automatic annotation and dereplication of PNPs.



In the last decades, antimicrobial resistance (AMR) has become a growing health problem being the cause of severe infections, complications, and increased mortality in a worldwide scale.[1−4] The need for new treatments to combat resistant strains puts bioactive natural products into the spotlight of the scientific community as a potential source of new drug development.[5,6] Since the discovery of penicillin, natural product research has uncovered many other peptides with powerful antimicrobial properties.[7] This is exemplified by the recent finding of teixobactin, a lipopeptide produced by *Eleftheria terrae* that has shown antibacterial activity against many Gram-positive drug resistant strains.[8] The properties of teixobactin suggest that it may be the first of a new class of antibiotics without resistance, creating great expectations for future research and urging the re-emergence of peptidic natural products (PNPs) to fight AMR.

The high sensitivity and selectivity of mass spectrometry have converted it into a method of choice for the analysis of biological samples in fields such as proteomics and metabolomics. Advances in tandem mass spectrometry have been accompanied with the development of analytical software to support the identification of natural compounds.[9,10] *In silico* models mimicking the fragmentation of the molecules during collision in the mass spectrometer take into consideration the characteristics of the targeted compounds.[11] In metabolomics, these algorithms rely on chemical structure databases to retrieve the molecule skeleton and virtually break its chemical bonds, either using systematic bond disconnection or considering fragmentation patterns and thereby cleaving specific bonds. In proteomics, the amino acid composition and linearity of proteinogenic peptides reduces the complexity of the theoretical fragmentation, not requiring the usage of chemical structures but peptidic sequences instead. Besides the shared amino acidic composition of proteins and PNPs, the higher structural diversity of the latter resembles that of small metabolites. PNPs are not always linear but they often form macrorings and branches. Highly modified amino acids and nonamino acid monomers including glycans or fatty acids are also present in their structures. The unique structural characteristics and molecular weight range of these compounds substantially reduce the efficiency of proteomics- and metabolomics-related software for their analysis. An efficient fragmentation model for PNPs should integrate peptide fragmentation rules in the structure-breakage approach typically used for metabolites. Additionally, the rules should be extended to not only include peptide bonds but also other linkages such as glycosidic and ester bonds.

There is a broad variety of tools for analyzing MS/MS data[9,10] but only a few are dedicated to PNP dereplication (identification of known compounds). Some of the most

Received: July 28, 2020
Accepted: November 9, 2020

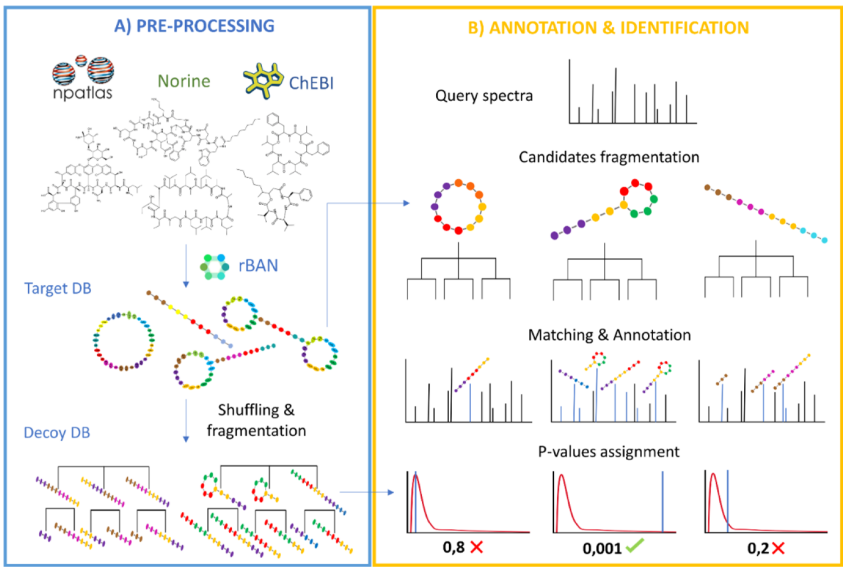


Figure 1. NRPro workflow. (A) Chemical structures are preprocessed with rBAN to generate fragment graphs for the target and decoy DBs. (B) Fragmentation graphs of candidates are computed by disconnecting the graphs through user-specified bonds.

established natural product dereplication tools are those integrated in the platform GNPS (Global Natural Product Social Molecular Networking).[12] Among them, Dereplicator was specifically developed for the identification of PNPs.[13] The interesting features of Dereplicator are the ability to process large datasets and the inclusion of *p*-values for evaluating the peptide spectrum matches (PSMs). The review of spectral annotations through a web interface is another attractive feature of the tool. However, the theoretical fragmentation of Dereplicator is limited to fragments of 2-cuts and does not include ester and glycosidic or x/a ions. Dereplicator+ is a later version of the software with an extended fragmentation model.[14] Nonetheless, this version implements a generic fragmentation for metabolites, not strictly fitting the PNP fragmentation patterns. It also lacks the annotation interface previously available in Dereplicator. Outside GNPS, iSNAP software covers the main bond types found in PNPs and provides two different scores for evaluating the statistical significance of the results, but it can only process a single spectrum per run.[15]

Despite the pros and cons of each individual tool, a recurrent drawback of dereplication software is the lack of interactivity to support users in validating results and improving postanalysis. Remarkable efforts are made for the development of optimal scoring systems, but less work is invested into facilitating expert validation by providing proper peak annotations and user-friendly interfaces. The range of spectral analysis options such as neutral losses (highly observed in PNPs), adducts, or deisotoping features is often limited as well. Thus, in-depth analysis for the structural characterization of PNPs requires additional annotation software. This aspect is covered by mMass, a popular tool that annotates cyclic peptide spectra but whose maintenance was recently ended.[16] Alternatively, Cyclobranch provides many options for the annotation of PNP spectra.[17] Contrary to dereplication tools, the usage of annotation software implies a considerable amount of manual work sometimes resulting in a time-consuming, error-prone, and complicated task. The idea of combining the high efficiency and identification capacity from dereplication software with the extensive annotations and GUIs from annotation software resulted in the solution introduced here, named NRPro. NRPro considers all previously mentioned cleavages to generate fragmentation graphs that are matched against experimental spectra. As in other dereplication tools, statistical significance is evaluated *via* a target-decoy strategy. The PSMs are displayed in an interactive web−application interface supporting postprocessing and edition. NRPro identifications were validated and compared to Dereplicator and Cyclobranch, two reference tools for dereplication and annotation of PNPs.

MATERIALS AND METHODS

Algorithm Architecture and Workflow. NRPro follows a classical web application architecture. On the server side, most of the business logic is implemented in Java. MongoDB[18,19] is used for storage and querying. The client-side was build using HTML, CSS, and JavaScript in an AngularJS framework. The Chemistry Development Kit (CDK),[20,21] MzJava,[22] and JGraphT[23] were used as third-party libraries. Prior to the analysis, preprocessing of the data is required to convert the chemical structures into chemical graphs and for the creation of the target and decoy databases (Figure 1A). The dereplication and annotation workflow involves the following steps: (i) generation of candidate fragmentation graphs, (ii) spectral matching and annotation, and (iii) statistical significance assignment (Figure 1B). These stages are further described in the following sections.

Construction of Target and Decoy Databases (Preprocessing). Structural data in the SMILES format[24] was retrieved from NORINE (version date: 11/12/2018),[25] ChEBI (version date: 04/01/2019),[26] and NPAtlas (v. 2019_08)[27] for the creation of a custom database. Disconnected compounds and charged molecules were discarded. Records with equivalent chemical structures were merged without stereoisomer discrimination. With the aim of recruiting peptide-like structures, a substructure search algorithm from CDK was applied to select those candidates containing at least a single amide or ester bond. Duplicated

https://dx.doi.org/10.1021/acs.analchem.0c03208
Anal. Chem. XXXX, XXX, XXX−XXX

compounds with the same structure but originating from different databases were identified using a graph isomorphism algorithm and merged into a single entry in which annotations from respective sources were preserved. The process concluded with a collection of 26,882 compounds.

rBAN[28] was used to convert peptidic structures into graphs where a node corresponds to a fragment (chemical composition) and an edge to a breakable bond (see Supporting Information). Importantly, the edge labels and direction contain essential information on the molecule fragmentation: (i) the labels specify the bond type and the ion generated when breaking them (y/b, x/a, z/c) and (ii) the direction indicates the position of the fragment containing the carbonyl (or carbon for glycosidic bonds) group. This information is used when applying the protonation rules of the fragment ion resulting from cleavage. For the generation of the decoy database, the original structure of the graphs was maintained but nodes were randomly shuffled. While the breakage of the target graphs is performed on the fly, decoy fragmentation graphs are precomputed to improve the performance of the algorithm. All graphs (target) and fragmentation graphs (decoy) are serialized in the JSON format and imported into MongoDB. Additionally, peptide atomic graphs are also imported for further use in the interface.

Computation of Fragmentation Graphs. A fragmentation graph is generated for each candidate peptide in order to simulate experimental fragmentation. The process consists in iteratively disconnecting the edges of the peptide graph to create a hierarchical structure of fragments (tree) that will be used for scoring (Figure S1B). Each cut results in either one or two child fragments that are introduced in the graph. The depth of the graph was restricted to four to account for the unlikelihood of further fragmentation and to avoid unnecessary computational expenses. Fragments with a mass lower than the first peak of the query spectra are not computed. Amide and ester bonds produce y/b and x/a ions that are systematically fragmented, whereas z/c ions are only cut upon user request. The protonation and hydrogen rearrangements of each type of ions contribute to the calculation of masses. Thus, the mass of y and c ions includes a proton and a hydrogen mass gain. Glycosidic bonds were fragmented in accordance to the model proposed by Domon and Costello[29] but only taking into consideration Y/B and Z/C breakages. For each fragment, masses corresponding to neutral losses (NH_3, H_2O, and C_2H_4) and adducts (NH_3, H_2O, NH, and CH_3), as specified by the user, are respectively subtracted or added to the fragment mass. In case of multiple charged precursors, the m/z of the fragment is calculated for each possible charge state. Accounting for all of these parameters generates multiple annotations for every fragment in the graph. Each annotation is given a score ($annot_{score}$) according to the ion type and the number/type of neutral losses (Figure S2A). This score is a penalty that reflects the "oddity" of the fragment. For instance, a CH_3 neutral loss being less frequent than a water loss will be assigned a higher score. These scores are used in the matching step where a threshold of two is established as the limit for considering a fragment ion as "expected".

Spectral Annotation, Deisotoping, and Scoring. The fragments derived from the graph are matched against experimental spectra (i) to annotate peaks and (ii) to score each PSM. However, matching in these two processes differs significantly. For spectral annotation, all the graph fragments are matched but scoring is stricter. Two conditions must be met in order to score a fragment:

1. The target fragment should have a direct parent already matched and scored in order to increase confidence in the match. Obviously, confidence will raise if the peak of the parent fragment has already been identified in the spectrum. Note that this condition is not applied for the first-level children as high collision energies can result in MS/MS spectra lacking a precursor mass.
2. The $annot_{score}$ should be lower than 2 (see scores assignment in Figure S2A) to guarantee that a fragment ion is expected and not highly modified.

Taking these conditions into account, matching is performed starting from the precursor peptide and traversing the graph using a breath-first algorithm. Fragments matched within the established m/z tolerance will be either scored and annotated or only annotated if they do not fulfil the required conditions. The matching of the m/z to the spectra was performed using binary search. Since isotope matching affects the scoring system and may lead to wrong identification, an optional deisotoping step was introduced in the software (see Supporting Information). Finally, the raw score of a PSM is calculated by summing all the relative intensities divided by the $annot_{score}$ + 1 (Figure S2B). Thus, ions with a low $annot_{score}$ such as unmodified b and y ions ($annot_{score}$ = 0) contribute more to the score.

Statistical Significance Calculation. The fragmentation graphs of the decoy database are used for calculating the statistical significance of PSMs. Fragmentation graphs of decoy compounds within a mass range from 0 to +40 Da with respect to the candidate precursor mass are matched against the experimental spectra. The raw scores resulting from those matches are then used for the estimation of the null hypothesis distribution and the subsequent p-value calculation. The loose range of 40 Da was empirically established in order to (i) guarantee an abundant collection of scores for the distribution and (ii) avoid the introduction of bias resulting from the scoring of large peptides. The scores are then fitted to a Weibull distribution estimated using the maximum likelihood method provided in the SSJ (Stochastic Simulation in Java) library.[30,31] The Weibull distribution was the best fit to the distribution obtained when testing all spectra against the decoy fragmentation graphs within the mass range of 0 to +40 Da (Figure S3). The p-value for a PSM is obtained placing its score in the distribution and calculating the area to the right and under the curve (exceedance frequency). The result indicates the significance of the PSM compared to the random structures of the decoy. Note that candidates with less than four matched peaks are automatically considered as lacking evidence, hence reported as nonsignificant.

RESULTS AND DISCUSSION

MS/MS Analysis Platform with a New Nomenclature System. Figure 1 shows the architecture and methods implemented in NRPro for the analysis of tandem mass spectra. The software represents a new MS/MS analysis platform available in the NORINE database and on the ExPASy[32] server. Features not included in other dereplication software such as the processing of z/c ion fragments and the annotation/editing of the results are key to NRPro. With the aim of annotating any type of PNP structure, a new nomenclature system was adopted in the software (Figure

https://dx.doi.org/10.1021/acs.analchem.0c03208
Anal. Chem. XXXX, XXX, XXX−XXX

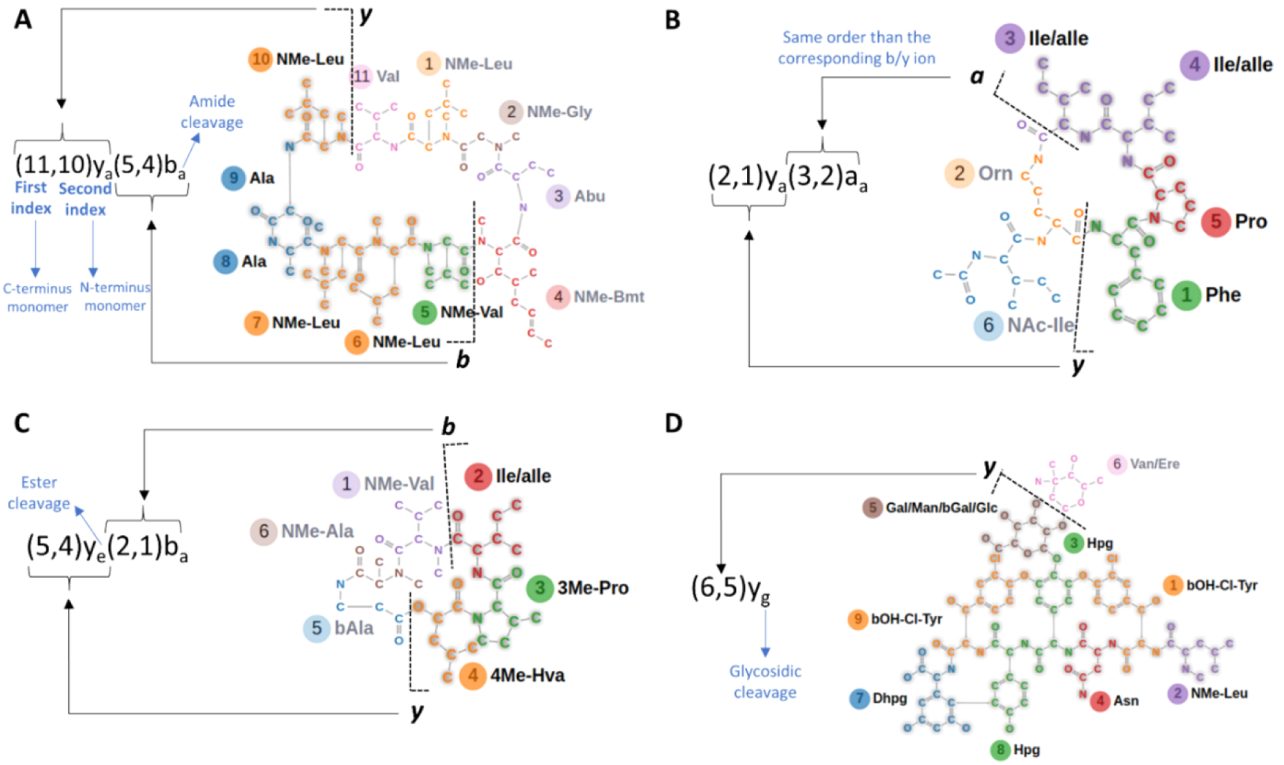


Figure 2. Illustrations of the nomenclature (automatic annotations) used in NRPro. (A) y/b ion in cyclosporin A, (B) y/a ion in pseudacyclin A, (C) y/b ion with ester cleavage in roseotoxin A, (D) lycosidic cleavage in vancomycin.

2). Each multiple cut fragment is labeled with the individual annotation of each cut (cleavage). A single cut annotation contains: (i) the two indexes of the monomers involved in the fragmentation, followed by (ii) the letter associated with the fragmented ion type (according to Roepstorff/Fohlman[33] and Biemann[34] for peptide bonds and Domon and Costello[29] for glycosidic bonds) and (iii) a subscript letter specifying the bond type ("a" for amide; "e" for ester; and "g" for glycosidic). For instance, as shown in Figure 2A, $(11,10)y_a(5,4)b_a$ represents a fragment ion arising from a y cut between monomers 10 and 11 and a b cut between 4 and 5 (both amide bonds). Note that ester bonds are substitutes for amide bonds in depsipeptides, which led to adopt the same nomenclature[33,34] to describe their breakages (Figure 2C). Evidently, these annotations require a visual aid providing the monomer indexes. Both the chemical structure and the monomeric structure of the molecule are simultaneously depicted in the interface (Figure S4). Other notable features of NRPro include multiple annotations per peak, deisotoping, manual edition, as well as filtering and exporting options (Table S1). Importantly, the results of a query are stored for a time period of one month, making both the automatic results and manual editions sharable *via* NRPro URLs. Creating permanent URLs is an option.

The two main functionalities of the software, namely, the automatic annotation and the dereplication, are evaluated in the following sections. The annotations are examined through manual comparison with Cyclobranch and Dereplicator while the dereplication capacity of NRPro is evaluated by analyzing a GNPS dataset.

Validation and Benchmarking of the MS/MS Automatic Annotations. A set of ten freely available tandem mass spectra collected from GNPS (Table S2) was used to manually evaluate and benchmark NRPro automatic annotations. The selected data contained MS/MS spectra from linear, cyclic, and branched cyclic peptides, each category exemplified by two compounds, as well as four spectra of peptides with higher complexity. Cyclobranch and Dereplicator were chosen for the benchmark due to their shared functionalities with our software. The method used for the annotation analysis involved two steps: (i) first running the tools with the minimum number of parameters and (ii) second, running them after adding the most common neutral losses (NH_3 and H_2O) and adducts (H_2O). Dereplicator was excluded from the second analysis because it does not include these options.

Linear peptides. The MS/MS spectra of gramicidin C (M + H) and leucinostatin B (M + 2H) were used to test the annotation of linear peptides. Gramicidin C (Figures 3A and S5A) was successfully identified and annotated by all the software. Besides the higher number of matched peaks from NRPro −23 compared to 16 (Cyclobranch) and 15 (Dereplicator), our tool did not show significant difference regarding intensity coverage. All tools covered about half the total intensity in the spectra. However, the NRPro deisotoping option identified many isotopes that contributed to the intensity but were not considered as annotated, leading to underestimating coverage results. Manual examination confirmed the matching of most of the high intensity peaks which mainly corresponded to b and y ions, the latter with lower intensities. An a ion was annotated by NRPro and Cyclobranch but not by Dereplicator. The other extra annotations given by NRPro corresponded to internal ions that do not greatly impact the intensity coverage. These results differed considerably from those obtained for leucinostatin B (Figures 3B and

https://dx.doi.org/10.1021/acs.analchem.0c03208
Anal. Chem. XXXX, XXX, XXX−XXX

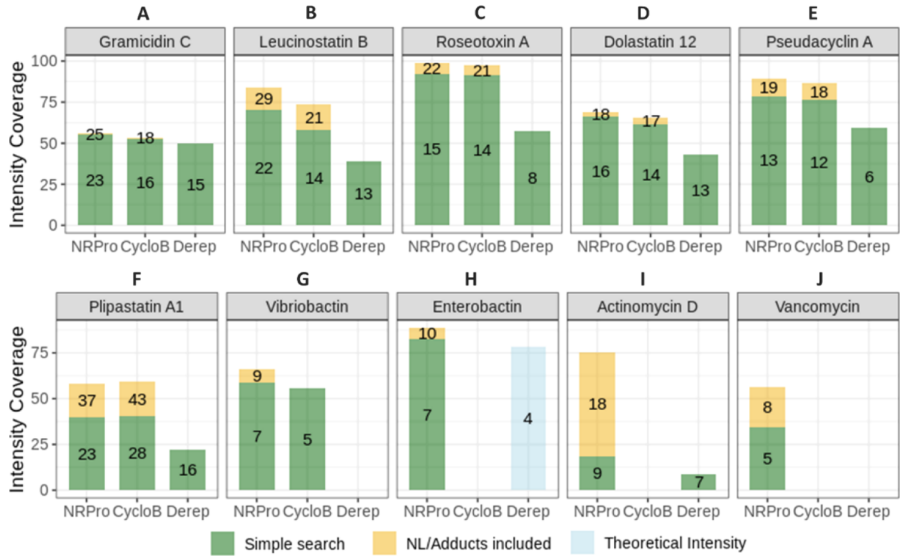


Figure 3. Annotations benchmark. The percentage of annotated intensity from NRPro, Cyclobranch, and Dereplicator in the MS/MS spectra of ten structurally diverse PNPs. The labeled numbers represent the peak count.

S5B), where the internal ions annotated by NRPro were responsible of the higher coverage of our tool compared to the others. Annotations were also examined after enabling the labeling of neutral losses and adducts in the search. While these settings barely affected the results of gramicidin C, the percentage of annotated intensity in leucionstatin B increased by 14−15% in both software. Most of these annotations are due to water losses, and their impact on the coverage exemplifies their high occurrence in some PNPs. It should be pointed out that leucinostatin B could not be identified through database search in Cyclobranch. Alternatively, the annotations of this compound were obtained in a mode provided by Cyclobranch that consists in manually introducing its monomer sequence and comparing the theoretical fragments with the MS/MS spectra. Hence, obtaining the annotations with Cyclobranch was more time consuming than with the other two software, which provided the results automatically in a few seconds.

Cyclic peptides. The roseotoxin A (M + H) and dolastatin 12 (M + 2H) depsipeptides were used to test cyclic structures with ester bonds. All the software identified the spectra of the former compound proposing either roseotoxin A or analogous structures as candidates. Dereplicator showed again lower coverage results (Figure 3C) due to the unannotated a ions. On the other hand, NRPro and Cyclobranch had almost identical results, only differing by a single extra annotation from NRPro corresponding to an a ion of the linearized precursor (precursor-CO). Interestingly, this peak was already reported,[35] as well as seven other peaks that corroborated automatic annotations (Figure S5C). In contrast, dolastatin 12 (Figures 3D and S5D) was successfully identified by NRPro and Dereplicator but not Cyclobranch, whose annotations were again obtained by manual introduction of the sequence. The highest peak corresponded to the doubly charged precursor that Dereplicator did not match. Importantly, fragments resulting from the ester bond breakage were identified in both tested compounds proving the ability of NRPro to annotate these characteristic ions. NRPro and Cyclobranch results did not show significant differences neither in the minimal setting mode nor after the inclusion of additional parameters.

Cyclic-Branched Peptides. For the evaluation of cyclic-branched structures, we used tandem mass spectra of pseudacyclin A (M + H) and the plipastatin A1 lipopeptide (M + 2H). For pseudacyclin A (Figures 3E and S5E), NRPro and Cyclobranch results were very similar but the examination of the single differing peak revealed an interesting case. The additional peak matched by NRPro at m/z 472.3 was annotated as a fragment ion involving the loss of an acetyl modification in an isoleucine. This loss was identified because of the theoretical breakage of the amide bond between the N-terminus of the isoleucine and the acetyl modification, liberating the latter group. Note that this option is not considered in the monomer-based fragmentation model of Cyclobranch, in which the internal breakage of monomers with amide bonds can only be handled with the inclusion of neutral losses or annotating the modifications separately. In this particular case, the proximity of the peak to 471.3 suggests that it could be an isotope and a possible false positive in NRPro. The rest of the peaks was compared with a characterization of the peptide already reported,[36] and all the shared peaks (9) showed the same annotation. The missing matches in Dereplicator corresponded to internal fragments, a ions and the precursor.

The analysis of pliplastatin A1 (Figures 3F and S5F) takes advantage of the structural equivalence between fengycin and plipastatin so that the MS/MS characterization of the former was used to evaluate the results regarding the latter. As in the previous case, NRPro and Cyclobranch matched more peaks than Dereplicator. The annotation of high intensity peaks was identical for the two software, matching characteristic ion peaks at m/z 1080 and 966 already reported and used as fingerprints for the identification of fengycin.[37−39] These two peaks correspond to the compound missing part of the branch (*FA-Glu*) and the missing whole branch (*FA-Glu-Orn*), respectively. Other matches of high intensity peaks correspond to fragment ions of the branch itself or internal ions resulting from three breakages. These examples demonstrate the

https://dx.doi.org/10.1021/acs.analchem.0c03208
Anal. Chem. XXXX, XXX, XXX−XXX

usability of NRPro for branched-cyclic structures, highly common in lipopeptides. Finally, the addition of neutral losses and adducts had a notable effect in plipastatin A1, for which the software identified peaks with NH_3 and H_2O losses, increasing the matches from 23 to 37, corresponding to a 18% increase of the covered intensity.

Linear-Branched Peptides. The compound selected for the analysis of linear-branched structures was vibriobactin (M + H; Figures 3G and S5G), a small linear peptide with a short branch of a single monomer (*diOH-Bz*). NRPro was the only software that automatically identified the molecule. Cyclobranch required the introduction of the sequence and Dereplicator did not output any candidate. The comparison with Cyclobranch revealed that the higher coverage in NRPro was due to fragment ions resulting from more than a single cut. This difference became more noticeable after the addition of neutral losses because NPRro identified two x ions with ammonia loss not matched by Cyclobranch. Note that this spectrum had only a few peaks including isotopes, which if removed, would increase the covered intensity from 66 to 81%.

Complex peptides. The examination of molecules with higher complexity was limited to NRPro and Dereplicator because Cyclobranch cannot process them. The first in this category was enterobactin (M + H; Figures 3H and S5H), a depsipeptide consisting of a central ring attached to three short branches of a single monomer. Both, NRPro and Dereplicator identified the peptide. Only four peaks were matched with Dereplicator leading to an unreliable identification that automatically disables the integrated peak annotation visual aid. Coverage could only be estimated from the number of matches. Based on the previous behavior of Dereplicator, it was assumed that the a ions, 3-cut fragments, and the precursor were missed only to yield four peak matches. If this theoretical coverage was correctly predicted, the two software tools matched all the high intensity peaks and respectively covered the 78% (Dereplicator) and 82% (NRPro) of the total intensity in the spectra. Furthermore, the higher peaks corresponded to fragment ions resulting from breakages of the ester bonds in the ring again confirming their ability to process complex depsipeptides. Although Dereplicator seems to correctly fragment the peptide, the lack of annotations made the interpretation of the results more difficult.

Another structurally complex peptide is actinomycin D (M + H; Figures 3I and S5I), composed of two cycles linked by a chromophore. Both software successfully identified the compound. However, the annotations only covered 18% (NRPro) and 8% (Dereplicator) of the spectral intensity. A significant change was observed when NRPro was run including adducts and neutral losses in the settings, increasing the coverage up to 75% due to the presence of water adducts in the fragments. As before, these results could not be compared to Dereplicator as the adduct/neutral loss option is not included. This case clearly confirms that enabling modifications is an essential feature of a complete and flexible annotation tool for PNPs. Furthermore, it shows the capacity of NRPro to successfully annotate peptides with several cycles.

The last peptide to be analyzed was the well-known vancomycin glycopeptide (M + H; Figures 3J and S5J), an excellent example of the structural complexity found in PNPs. Although the fragmentation of vancomycin is limited, NRPro identified the spectra and properly labeled the two main peaks corresponding to the loss of vancosamine (m/z 1305) and Glucose (m/z 1143), already reported in other studies,[40,41] as well as their respective a ions. Additionally, another a-ion fragment matching the loss of vancosamine and aspargine at m/z 1145 was also identified. Interestingly, the loss of asparagine was also previously reported.[41] These results not only prove the usefulness of NRPro for complex structures but also validate glycopeptide fragmentation. Dereplicator failed to identify the compound.

Verification of the Annotations from Surfactins with Sodium Adducts. Surfactins are cyclic lipopeptides often presenting sodium adducts in their spectra. They have been widely characterized, facilitating their spectral interpretation. Thus, [M + Na] spectra from surfactin C13, surfactin C14, and surfactin C15 were retrieved from GNPS in order to evaluate sodium adduct annotations. The length of the fatty acid chain (indicated by the number of carbons) is the only structural change between these peptides. According to the literature, the ring opening from the surfactin is produced through the breakage of the ester bond attached to the fatty acid.[42] Consequently, the linearized form of the peptide has the following sequence: *β-OHFA-Glu-Leu-Leu-Val-Asp-Leu-Leu-Val*. The fragments automatically annotated by NRPro resulted from the cleavage of the mentioned ester bond, confirming the given sequence. In the case of surfactin C13 [M + Na], for instance, the relevant ions included a series of aliphatic fragments at m/z 590, 689, 786, 804, 917,and 1012 and another series of peptidic ions at m/z 463, 481, 594, and 707. As expected, the same peptidic ion series was matched in the other two surfactins. Equivalent aliphatic fragments were identified as well but with a difference of 14 Da derived from the extra carbon in the lipid chains (Figure S6). The annotated ions coincided with those already defined in other studies,[38,42] verifying the reliability of the annotations provided by the software. Most of these fragment ions presented either neutral losses or water adducts, which had to be included in the search to obtain a complete spectral annotation.

Verification of c/z Ion Annotations from Actinomycin D. Up to this point, mainly b/y fragments have been annotated in the analyzed spectra, but NRPro can also annotate c/z ions, which are still only occasionally reported in the literature. Nonetheless, we used the data of an extensive study performed by Wills and O'Connor[43] that encompasses the characterization of actinomycin D with electron induced dissociation and describes c and z ion annotations, for the construction of a peak list with constant intensities representing the spectra. NRPro automatic annotations were compared with those manually annotated by the authors (Figure S7). NRPro annotated 17 out of the 26 reported peaks. Among those, 15 fragment ions were the same as those described in the article and seven corresponded to z or c ions. Some of them were unusual internal fragments resulting from the simultaneous cleavage of a and z fragmentation termini. Other verified peaks included an x-ion and several fragments derived from the breakage of the ester bond in the ring. Most of the unannotated peaks corresponded to 4-cut fragments, which are currently not considered due to their low expectancy and higher computational expense. Nonetheless, the previously described annotations covered most of the high intensity peaks, proving the usability of the software for the annotation of c/z ions. Although the examination of further data would provide stronger evidence of annotation correctness, these results tend to indicate that the theoretical fragmentation is successfully emulating the c/z ion breakages. Otherwise, the

https://dx.doi.org/10.1021/acs.analchem.0c03208
Anal. Chem. XXXX, XXX, XXX−XXX

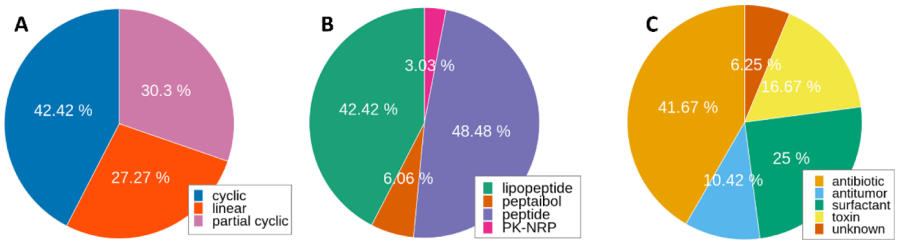


Figure 4. Variability of the PNPs identified within the NORINE subset. (A) PNP structures, (B) PNP categories, and (C) PNP bioactivities.

postanalysis features from NRPro can be used to adjust unexpected annotations.

Dereplication of Data from GNPS MS/MS Spectral Library. An optimal evaluation of NRPro dereplication required a set of spectra (i) mainly composed of MS/MS of PNPs, (ii) sufficiently large to warrant statistical analysis, and (iii) structurally diverse to validate the identification of different peptide types. A freely available set meeting these criteria was retrieved from GNPS and consisted of 352 spectra (M + H and M + 2H) previously identified with Dereplicator (Suppl., Experimental spectra). The parent and fragment mass tolerance was set to 0.02 to reflect the minimal mass shift for precursor identification, equivalent to the product ion tolerance employed in the Dereplicator analysis. Although most compounds annotated by Dereplicator were already in our target database, to guarantee their occurrence, their SMILES and PubChem[44] IDs were introduced/merged with the other database structures. In this way, false positives arise from possible weaknesses of the fragmentation/scoring but not from their absence in the database. NRPro suggested a list of candidates for each spectrum and those with the lowest *p*-values were considered as the top hits (Supporting Information, File 1).

Evaluation of the False Discovery Rate of the Identifications. A common method for estimating the false discovery rate (FDR) in proteomics is a ratio calculation between the number of identifications in the decoy (false positives) and target (true positives) databases. With a similar aim, all spectra of the GNPS dataset were scored against the target and decoy databases. Interestingly, no PSM of the *p*-value lower than 10^{-7} was identified in the decoy database (Figure S8) resulting in an FDR of 0%. With this value of FDR, 85 (24.14%) out of 352 PNPs were identified. Other options with a higher number of identifications showed optimal FDRs as well. Setting the threshold at 10^{-6} resulted in ∼11% increase in the number of identifications (125) and an FDR of 2.4%; just three peptides were identified in the decoy database. Another acceptable option is the *p*-value of 10^{-5}, which results in the identification of 169 PNPs in the target database (48.01%) and 6 in the decoy; in terms of false discovery, it represents a 3.55% FDR. Note that these FDR calculations are more conservative than those in other approaches based on PSM counts instead of PNPs. Furthermore, since the size of the spectral set is small, a single decoy identification has a strong impact on the FDR.

A *p*-value of 10^{-5} was used for the manual examination of identifications associated with NORINE annotations that span compound structure, category, and activities. Due to the low coverage of NORINE (1730 peptides), only 18 PNPs out the 169 confidentially dereplicated ($<10^{-5}$) compounds had NORINE annotations (Table S3). Despite the small size of the set, compounds with multiple structures, compositions, and origins were identified. Such variability is further studied in the following sections.

Validation of the Identifications. Metadata associated with the analyzed spectral set contain information relative to identifications with Dereplicator and as such could ease validation. In particular, the occurrence of SMILES of the peptides identified by Dereplicator in the header of each MS/MS spectra was very helpful. CDK was used to convert these SMILES into a chemical graph to be compared with the peptide identified by NRPro, using a graph isomorphism algorithm. One of the proposed NRPro candidates was always identical to that proposed by Dereplicator. Thus, only the top-scored PNPs (first candidates) were used to calculate the percentage of coincidence between the results of the two software. Out of the 352 identifications, 311 PNPs were listed by both software as first candidates (Supporting Information, File 2), that is, an 88.35% of coincidence. Being more restrictive, when only the 169 spectra with *p*-values lower than 10^{-5} (FDR 3.55%) were considered, the overlap increased up to 93.49% (158 PNPs), which raises confidence in the results even if the comparison remains based on predicted PNP identifications. The eleven remaining spectra (6.51%) with distinct identifications were manually examined to uncover the differences between the candidates proposed by each software. Interestingly, seven of these spectra matched similar linear PNPs from *Trichoderma* sp. In all cases, the PNPs proposed by both software were members of the same family and structurally close. However, the matching of NRPro favored different family members than those proposed by Dereplicator. Comparing the identifications of each software revealed in some cases similar matching results (with the *CCMSLIB00000* prefix:*577602, 577613, 577769, 577583*) between the two candidates of each respective software, while other cases in which the NRPro candidate that matched higher intensity peaks and a larger proportion of b/y ions, (with the *CCMSLIB00000* prefix:*577623, 577654, 577850*) gave NRPro a clear advantage (Figure S9). Three of the other nonoverlapping spectra (with the *CCMSLIB00000* prefix:*577779, 577743, 577840*) matched structurally close lipopeptides with common theoretical peaks. The most uncertain NRPro identification corresponded to the *CCMSLIB00000577747* spectra that were attributed to [*Phe3 N-MeVal5*]-destruxin B instead of the lipopeptide arenamide C proposed by Dereplicator and positioned as the third candidate by NRPro despite the matching of the second highest peak not annotated in the case of destruxin B. The magnitude of this issue would be greater if NRPro was providing a single identification. However, multiple candidate suggestions combined with the provision of a GUI for their manual analysis facilitate the manual resolution of difficult cases such as these.

https://dx.doi.org/10.1021/acs.analchem.0c03208
Anal. Chem. XXXX, XXX, XXX−XXX

Evaluation of the Data Variability. The variability of the identifications was analyzed to evaluate the capacity of the software to process compounds with different properties and structures. The NORINE annotations referred to in previous analyses were used to automatically extract the structures, categories, and activities of the dereplicated PNPs. As already observed, the subset of entries with NORINE annotation was limited. Thus, instead of applying a filter based on *p*-values, confidence in the matches was obtained by selecting those entries identified by both Dereplicator and NRPro. Out of the 311 overlapping identifications, 33 included NORINE annotation. In this subset, the predominant structures corresponded to cyclic peptides (42%), but the abundance of linear (27%) and partial cyclic (30%) structures was also remarkable (Figure 4A). These structures matched with those described in the Dereplicator analysis (linear, cyclic and branch cyclic) and confirmed the ability of NRPro to dereplicate different structural types. Other structural annotations such as double cyclic and branched peptides were not identified in the subset but actinomycins were spotted within the whole dataset. This indicates that the lack of double cyclic structures is probably attributable to the small portion of records with NORINE annotation. Most of the identifications corresponded to peptides (48%) or lipopeptides (42%), but a small percentage belonged to peptaibols and polyketide non-ribosomal peptide (PK-NRP) hybrids (Figure 4B). Thus, the presence of monomers such as fatty acids or α-aminoisobutyric acid (highly frequent in peptaibols) within these compound chemical structures did not affect the dereplication process. No glycopeptide was identified. To determine whether the lack of glycopeptides was due to their absence in the data or to a possible NRPro dereplication problem, all molecular structures (SMILES) associated with the spectra were analyzed with ClassyFire.[45] No ontologies indicating the presence of glycans were suggested nor glycopeptides were mentioned in the publication of Dereplicator, in agreement with the results of NRPro.

Bioactivities were also analyzed. PNPs are known for their antibiotic properties, and unsurprisingly, 41.67% of identifications matches this category (Figure 4C). Surfactants, toxins, and anti-tumors were also present. These results demonstrate the clinical relevance of the compounds targeted by NRPro. Lastly, the phylogenetic examination of PNP producer organisms revealed eukaryotic and bacterial origins. Eukaryotic origins are mostly found in ascomycota fungi (16) but also in marine organisms of the gastropod class (3) and green algae (1). The bacterial microorganisms span different genus/species of cyanobacteria (10) as well as the streptomyces (4), pseudomonas (4), and bacillus (3) genera (Figure S10). These results are in accordance with the origin of data described in Dereplicator. Bioactivities were also analyzed. PNPs are known for their antibiotic properties, and unsurprisingly, 41.67% of identifications match this category (Figure 4C). Surfactants, toxins, and antitumors were also present. These results demonstrate the clinical relevance of the compounds targeted by NRPro. Lastly, the phylogenetic examination of PNP producer organisms revealed eukaryotic and bacterial origins. The eukaryotic origins are mostly found in ascomycota fungi (16) but also in marine organisms of the gastropod class (3) and green algae (1). The bacterial microorganisms span different genus/species of cyanobacteria (10) as well as the streptomyces (4), pseudomonas (4), and bacillus (3) genera (Figure S10). These results are in accordance with the origin of data described in Dereplicator.

From a mass spectrometry point of view, the fragments that match peaks are of high interest. In order to evaluate the fragment ion type and bond cleavages in the matches, scored fragments ($annot_{score}$ = <2) of one and two cuts were examined (Figure 5). Missing y and x ions in 2-cut fragments expectedly

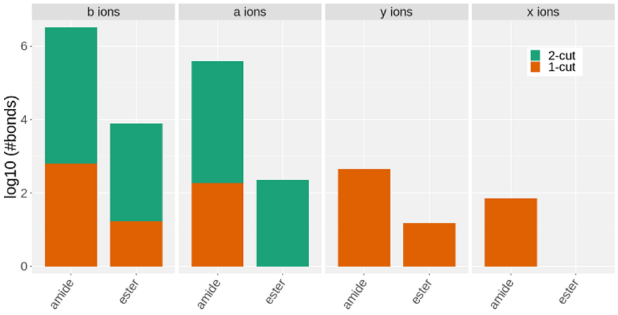


Figure 5. Type of fragment ions scored.

reflected the lack of the terminal carboxyl group. In 1-cut fragments, b ions were slightly more abundant (646 instances) than y ions (470 instances). a and x ion types presented 190 and 71 occurrences, respectively. 7.7% of the matched peaks corresponded to ester cleavages, probably originated from depsipeptides. Note that these results are limited to the peaks used in the scoring calculation.

CONCLUSIONS

The evaluation of NRPro dereplication revealed that the software is able to identify 169 PNPs in a dataset of 352 spectra with an FDR of 3.55. 93.49% of these identifications matched with those provided by Dereplicator. Identifications unique to NRPro pointed out the difficulty of differentiating spectra from structurally close compounds, often members of the same peptide family. In future versions, the introduction of an additional *p*-value evaluating the uncertainty of these cases would better define the significance of the match. With regards to automatic annotations, NRPro annotates as many or more peaks than other tested software, particularly in the case of complex peptides. Manual examination of results revealed an abundant presence of isotopes in some spectra thereby justifying the relevance of the deisotoping option. Additional features such as *z*/c ions and the processing of MS/MS with adducts showed convincing results with the tested compounds. Prospects for NRPro include the implementation of a high-throughput batch mode and the extension of the database to increase identifications. Additionally, the monomeric nomenclature implemented in NRPro could be coupled with genome mining tools for the identification of secondary metabolite biosynthetic gene clusters.[46,47] We expect NRPro to positively impact the natural product community by strengthening results of MS/MS analysis. The tool is available as a web application in NORINE (https://bioinfo.cristal.univ-lille.fr/nrpro) and on the ExPASy server of the SIB Swiss Institute of Bioinformatics (https://web.expasy.org/nrpro). A manual of instructions is provided in the Supporting Information. For transparency, we made the code available in Bitbucket (https://bitbucket.org/sib-pig/nrpro-public/src/master/), but it is currently not executable as it requires the usage of the database stored in the server.

https://dx.doi.org/10.1021/acs.analchem.0c03208
Anal. Chem. XXXX, XXX, XXX−XXX

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.analchem.0c03208.

Methods, figures, and tables (PDF)
NRPro manual (PDF)
PNPs identified by NRPro in the GNPS spectra (ZIP)
PNPs identified by NRPro and Dereplicator in the GNPS spectra (ZIP)
352 GNPS MS/MS spectra (ZIP)

■ AUTHOR INFORMATION

Corresponding Author

Frédérique Lisacek − *Proteome Informatics Group, SIB Swiss Institute of Bioinformatics, CMU, Geneva 1211, Switzerland; Computer Science Department and Section of Biology, University of Geneva, Geneva 1227, Switzerland;* orcid.org/0000-0002-0948-4537; Email: frederique.lisacek@sib.swiss

Authors

Emma Ricart − *Proteome Informatics Group, SIB Swiss Institute of Bioinformatics, CMU, Geneva 1211, Switzerland; Computer Science Department, University of Geneva, Geneva 1227, Switzerland;* orcid.org/0000-0001-9086-7051

Maude Pupin − *University Lille, CNRS, Centrale Lille, UMR 9189−CRIStAL−Centre de Recherche en Informatique Signal et Automatique de Lille, Lille F-59000, France*

Markus Müller − *Proteome Informatics Group, SIB Swiss Institute of Bioinformatics, CMU, Geneva 1211, Switzerland; Vital-IT Group, SIB Swiss Institute of Bioinformatics, Lausanne 1015, Switzerland*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.analchem.0c03208

Author Contributions

Conceptualization, methodology, and software, E.R.; validation, M.M., M.P., and F.L.; writing—original draft preparation, E.R.; writing—review and editing, M.M., M.P., and F.L.; supervision, F.L. All authors read and approved the final manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank Valérie Leclère for her advises in microbiology and Areski Flissi for the integration of NRPro into NORINE. This work was supported by the SIB Swiss Institute of Bioinformatics Fellowship program. The ExPASy portal is maintained by SIB and hosted at the Vital-IT Competency Centre.

■ REFERENCES

(1) O'neill, J. I. M. *Antimicrobial Resistance: Tackling a Crisis for the Health and Wealth of Nations*; Review on Antimicrobial Resistance, 2014; Vol. *20*; pp 1−16.

(2) Roca, I.; Akova, M.; Baquero, F.; Carlet, J.; Cavaleri, M.; Coenen, S.; Cohen, J.; Findlay, D.; Gyssens, I.; Heure, O. E.; Kahlmeter, G.; Kruse, H.; Laxminarayan, R.; Liébana, E.; López-Cerero, L.; MacGowan, A.; Martins, M.; Rodríguez-Baño, J.; Rolain, J.-M.; Segovia, C.; Sigauque, B.; Tacconelli, E.; Wellington, E.; Vila, J. *New Microbes New Infect.* **2015**, *6*, 22−29.

(3) Aslam, B.; Wang, W.; Arshad, M. I.; Khurshid, M.; Muzammil, S.; Rasool, M. H.; Nisar, M. A.; Alvi, R. F.; Aslam, M. A.; Qamar, M. U.; Salamat, M. K. F.; Baloch, Z. *Infect. Drug Resist.* **2018**, *11*, 1645.

(4) Organization, W. H. *Antimicrobial Resistance: Global Report on Surveillance*; World Health Organization, 2014.

(5) Igarashi, M. *J. Antibiot.* **2019**, *72*, 890−898.

(6) Moloney, M. G. *Trends Pharmacol. Sci.* **2016**, *37*, 689−701.

(7) Newman, D. J.; Cragg, G. M. *J. Nat. Prod.* **2016**, *79*, 629−661.

(8) Ling, L. L.; Schneider, T.; Peoples, A. J.; Spoering, A. L.; Engels, I.; Conlon, B. P.; Mueller, A.; Schäberle, T. F.; 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. *Nature* **2015**, *517*, 455−459.

(9) Spicer, R.; Salek, R. M.; Moreno, P.; Cañueto, D.; Steinbeck, C. *Metabolomics* **2017**, *13*, 106.

(10) Perez-Riverol, Y.; Wang, R.; Hermjakob, H.; Müller, M.; Vesada, V.; Vizcaíno, J. A. *Biochim. Biophys. Acta, Proteins Proteomics* **2014**, *1844*, 63−76.

(11) Horlacher, O.; Jin, C.; Alocci, D.; Mariethoz, J.; Müller, M.; Karlsson, N. G.; Lisacek, F. *Anal. Chem.* **2017**, *89*, 10932−10940.

(12) 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.; Bandeira, N. *Nat. Biotechnol.* **2016**, *34*, 828−837.

(13) Mohimani, H.; Gurevich, A.; Mikheenko, A.; Garg, N.; Nothias, L.-F.; Ninomiya, A.; Takada, K.; Dorrestein, P. C.; Pevzner, P. A. *Nat. Chem. Biol.* **2017**, *13*, 30.

(14) Mohimani, H.; Gurevich, A.; Shlemov, A.; Mikheenko, A.; Korobeynikov, A.; Cao, L.; Shcherbin, E.; Nothias, L.-F.; Dorrestein, P. C.; Pevzner, P. A. *Nat. Commun.* **2018**, *9*, 4035.

(15) Ibrahim, A.; Yang, L.; Johnston, C.; Liu, X.; Ma, B.; Magarvey, N. A. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, 19196−19201.

(16) Niedermeyer, T. H. J.; Strohalm, M. *PLoS One* **2012**, *7*, No. e44913.

(17) Novák, J.; Lemr, K.; Schug, K. A.; Havlíček, V. *J. Am. Soc. Mass Spectrom.* **2015**, *26*, 1780−1786.

(18) Chodorow, K. *MongoDB: The Definitive Guide: Powerful and Scalable Data Storage*; O'Reilly Media, Inc., 2013.

(19) MongoDB. https://www.mongodb.com (accessed June 2, 2020).

(20) Steinbeck, C.; Han, Y.; Kuhn, S.; Horlacher, O.; Luttmann, E.; Willighagen, E. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 493−500.

(21) Willighagen, E. L.; Mayfield, J. W.; Alvarsson, J.; Berg, A.; Carlsson, L.; Jeliazkova, N.; Kuhn, S.; Pluskal, T.; Rojas-Chertó, M.; Spjuth, O. *J. Cheminf.* **2017**, *9*, 33.

(22) Horlacher, O.; Nikitin, F.; Alocci, D.; Mariethoz, J.; Müller, M.; Lisacek, F. *J. Proteomics* **2015**, *129*, 63−70.

(23) Michail, D.; Kinable, J.; Naveh, B.; Sichi, J. V. JGraphT−A Java Library for Graph Data Structures and Algorithms. **2019**, arXiv:1904.08355. arXiv preprint.

(24) Weininger, D. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.

(25) Flissi, A.; Ricart, E.; Campart, C.; Chevalier, M.; Dufresne, Y.; Michalik, J.; Jacques, P.; Flahaut, C.; Lisacek, F.; Leclère, V.; Pupin, M. *Nucleic Acids Res.* **2020**, *48*, D465−D469.

(26) Hastings, J.; Owen, G.; Dekker, A.; Ennis, M.; Kale, N.; Muthukrishnan, V.; Turner, S.; Swainston, N.; Mendes, P.; Steinbeck, C. *Nucleic Acids Res.* **2016**, *44*, D1214−D1219.

(27) Van Santen, J. A.; Jacob, G.; Singh, A. L.; Aniebok, V.; Balunas, M. J.; Bunsko, D.; Neto, F. C.; Castaño-Espriu, L.; Chang, C.; Clark, T. N.; Cleary Little, J. L.; Delgadillo, D. A.; Dorrestein, P. C.; Duncan, K. R.; Egan, J. M.; Galey, M. M.; Haeckl, F. P. J.; Hua, A.; Hughes, A. H.; Iskakova, D.; Khadilkar, A.; Lee, J.-H.; Lee, S.; LeGrow, N.; Liu, D. Y.; Macho, J. M.; McCaughey, C. S.; Medema, M. H.; Neupane, R. P.; O'Donnell, T. J.; Paula, J. S.; Sanchez, L. M.; Shaikh, A. F.; Soldatou, S.; Terlouw, B. R.; Tran, T. A.; Valentine, M.; van der Hooft, J. J. J.; Vo, D. A.; Wang, M.; Wilson, D.; Zink, K. E.; Linington, R. G. *ACS Cent. Sci.* **2019**, *5*, 1824−1833.

(28) Ricart, E.; Leclère, V.; Flissi, A.; Mueller, M.; Pupin, M.; Lisacek, F. *J. Cheminf.* **2019**, *11*, 13.

(29) Domon, B.; Costello, C. E. *Glycoconjugate J.* **1988**, *5*, 397−409.

https://dx.doi.org/10.1021/acs.analchem.0c03208
Anal. Chem. XXXX, XXX, XXX−XXX

(30) L'Ecuyer, P. *SSJ: Stochastic Simulation in Java*, Software Library; DIRO, Université de Montréal, 2016.
(31) L'Ecuyer, P.; Meliani, L.; Vaucher, J. SSJ: A Framework for Stochastic Simulation in Java. In *Proceedings of the 2002 Winter Simulation Conference*; Yücesan, E., Chen, C.-H., Snowdon, J. L., Charnes, J. M., Eds.; IEEE Press, 2002; pp 234−242.
(32) Artimo, P.; Jonnalagedda, M.; Arnold, K.; Baratin, D.; Csardi, G.; De Castro, E.; Duvaud, S.; Flegel, V.; Fortier, A.; Gasteiger, E.; Grosdidier, A.; Hernandez, C.; Ioannidis, V.; Kuznetsov, D.; Liechti, R.; Moretti, S.; Mostaguir, K.; Redaschi, N.; Rossier, G.; Xenarios, I.; Stockinger, H. *Nucleic Acids Res.* **2012**, *40*, W597−W603.
(33) Roepstorff, P.; Fohlman, J. *Biomed. Mass Spectrom.* **1984**, *11*, 601.
(34) Biemann, K. *Biomed. Environ. Mass Spectrom.* **1988**, *16*, 99−111.
(35) Jegorov, A.; Paizs, B.; Žabka, M.; Kuzma, M.; Havlíček, V.; Giannakopulos, A. E.; Derrick, P. J. *Eur. J. Mass Spectrom.* **2003**, *9*, 105−116.
(36) Pavlaskova, K.; Nedved, J.; Kuzma, M.; Zabka, M.; Sulc, M.; Sklenar, J.; Novak, P.; Benada, O.; Kofronova, O.; Hajduch, M.; et al. *J. Nat. Prod.* **2010**, *73*, 1027−1032.
(37) Bie, X.; Lu, Z.; Lu, F. *J. Microbiol. Methods* **2009**, *79*, 272−278.
(38) Pecci, Y.; Rivardo, F.; Martinotti, M. G.; Allegrone, G. *J. Mass Spectrom.* **2010**, *45*, 772−778.
(39) Li, X.-Y.; Mao, Z.-C.; Wang, Y.-H.; Wu, Y.-X.; He, Y.-Q.; Long, C.-L. *J. Mol. Microbiol. Biotechnol.* **2012**, *22*, 83−93.
(40) Cao, M.; Feng, Y.; Zhang, Y.; Kang, W.; Lian, K.; Ai, L. *Sci. Rep.* **2018**, *8*, 15471.
(41) Diana, J.; Visky, D.; Hoogmartens, J.; Van Schepdael, A.; Adams, E. *Rapid Commun. Mass Spectrom.* **2006**, *20*, 685−693.
(42) Yang, H.; Li, X.; Li, X.; Yu, H.; Shen, Z. *Anal. Bioanal. Chem.* **2015**, *407*, 2529−2542.
(43) Wills, R. H.; O'Connor, P. B. *J. Am. Soc. Mass Spectrom.* **2013**, *25*, 186−195.
(44) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B. *Nucleic Acids Res.* **2019**, *47*, D1102−D1109.
(45) Feunang, Y. D.; Eisner, R.; Knox, C.; Chepelev, L.; Hastings, J.; Owen, G.; Fahy, E.; Steinbeck, C.; Subramanian, S.; Bolton, E. *J. Cheminf.* **2016**, *8*, 61.
(46) Medema, M. H.; Blin, K.; Cimermancic, P.; de Jager, V.; Zakrzewski, P.; Fischbach, M. A.; Weber, T.; Takano, E.; Breitling, R. *Nucleic Acids Res.* **2011**, *39*, W339−W346.
(47) Dejong, C. A.; Chen, G. M.; Li, H.; Johnston, C. W.; Edwards, M. R.; Rees, P. N.; Skinnider, M. A.; Webster, A. L.; Magarvey, N. A. *Nat. Chem. Biol.* **2016**, *12*, 1007−1014.

https://dx.doi.org/10.1021/acs.analchem.0c03208
Anal. Chem. XXXX, XXX, XXX−XXX

Supporting Information

Automatic annotation and dereplication of tandem mass spectra of peptidic natural products

Emma Ricart[a,b], Maude Pupin[c], Markus Mueller[a,d], Frédérique Lisacek[a,b,e]

[a] Proteome Informatics Group, SIB Swiss Institute of Bioinformatics, CMU, Rue Michel-Servet 1, 1211, Geneva, Switzerland

[b] Computer Science Department, University of Geneva, Geneva, Switzerland

[c] Univ. Lille, CNRS, Centrale Lille, UMR 9189 - CRIStAL - Centre de Recherche en Informatique Signal et Automatique de Lille, F-59000 Lille, France

[d] Vital-IT Group, SIB Swiss Institute of Bioinformatics, Amphipole Building, Quartier Sorge, 1015, Lausanne, Switzerland

[e] Section of Biology, University of Geneva, Geneva, Switzerland

*Corresponding author: *frederique.lisacek@sib.swiss*

Table of contents

Graph generation with rBAN

The original rBAN algorithm generates these graphs by breaking specific bonds and matching the disconnected substructures against the Norine monomers in order to annotate them. The monomeric composition of the peptides is essential for the automatic annotation of the spectra. However, the MS/MS and monomeric fragmentations do not always coincide. Thus, an additional module dedicated to MS/MS fragmentation was added to rBAN. The algorithm was adjusted to perform two sequential searches: i) targeting monomer-linking bonds (the original approach) ii) targeting possible MS/MS breakage of amide, ester and glycosidic bonds. This approach resulted in graphs where all potentially breakable bonds are identified and where the nodes are labeled according to the Norine monomer they belong to (Figure -S1A).

Deisotoping process

NRPro scans the spectra starting from the lowest *m/z* value and searches for an isotopic cluster defined as containing at least two peaks separated by mass distances of 1.003355/charge ± mass tolerance, as follows:

1. Distances corresponding to the highest charged state are searched.
2. If the pattern is not identified, the charge is decremented until a cluster is found or after reaching the minimum charge.
3. When a cluster is found, previously assigned compositions (annotations) are used to calculate theoretical intensities.
4. Experimental intensities are normalized and compared to their theoretical counterparts with a 5% tolerance.
5. Peaks with matching intensities are labeled as isotopes and removed from the list of scored peaks not to interfere with scoring.
6. If the monoisotopic peak has several compositions, those not matching with the isotopic pattern are discarded.

Figure -S1. Illustration of the fragmentation from cyclosporin A. A) Graph structure created using rBAN. B) Theoretical fragmentation graph.

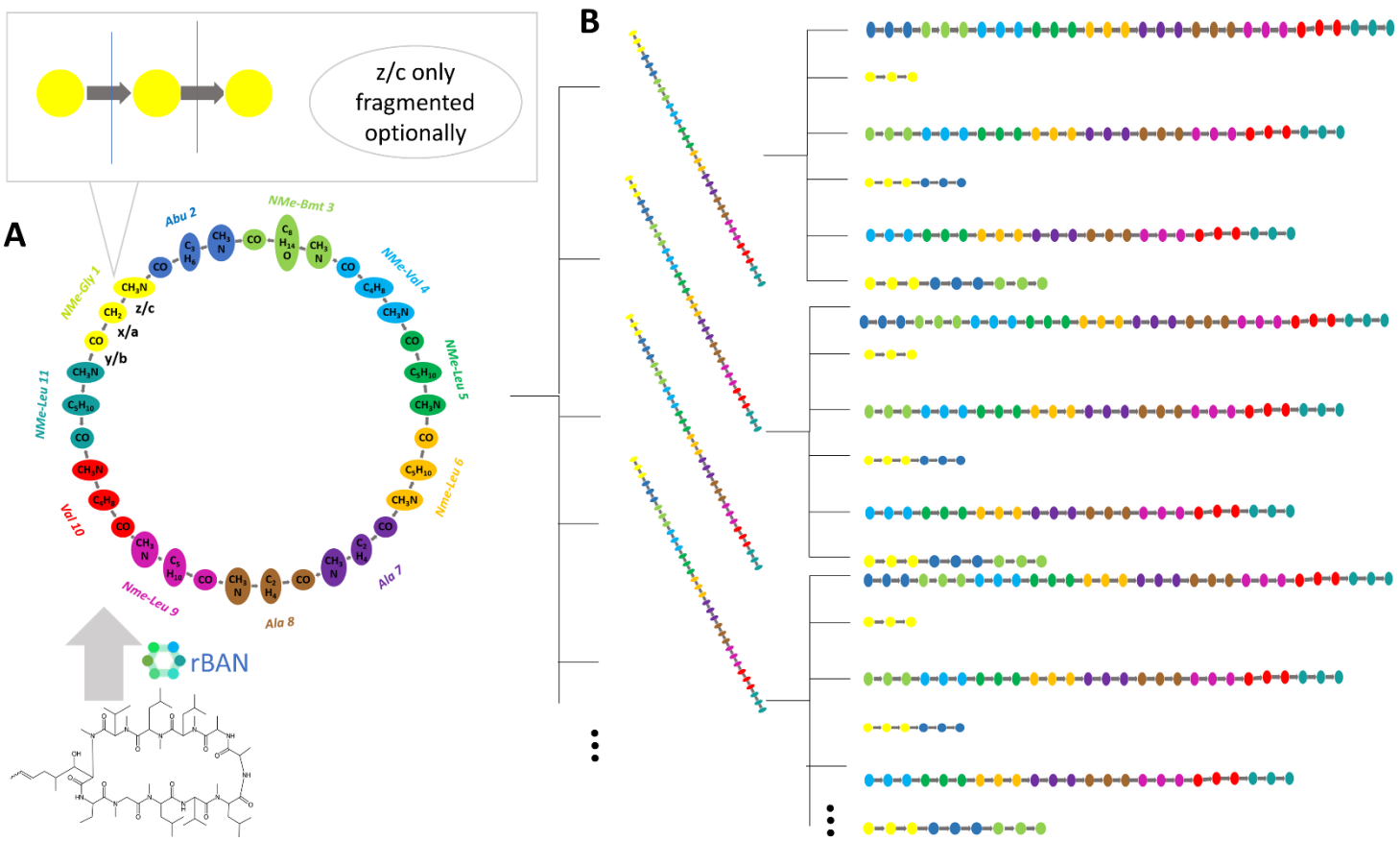


Figure -S2. Annotation score assignment (A) and raw score calculation (B). The path that would follow a [*y,b*] fragment ion with a water loss is illustrated with the orange mark. Note that *z/c* fragment ions are only calculated when specified by the user. When this option is activated, they are expected and therefore not penalized by the algorithm, in contrast with *x* and *a* ions.

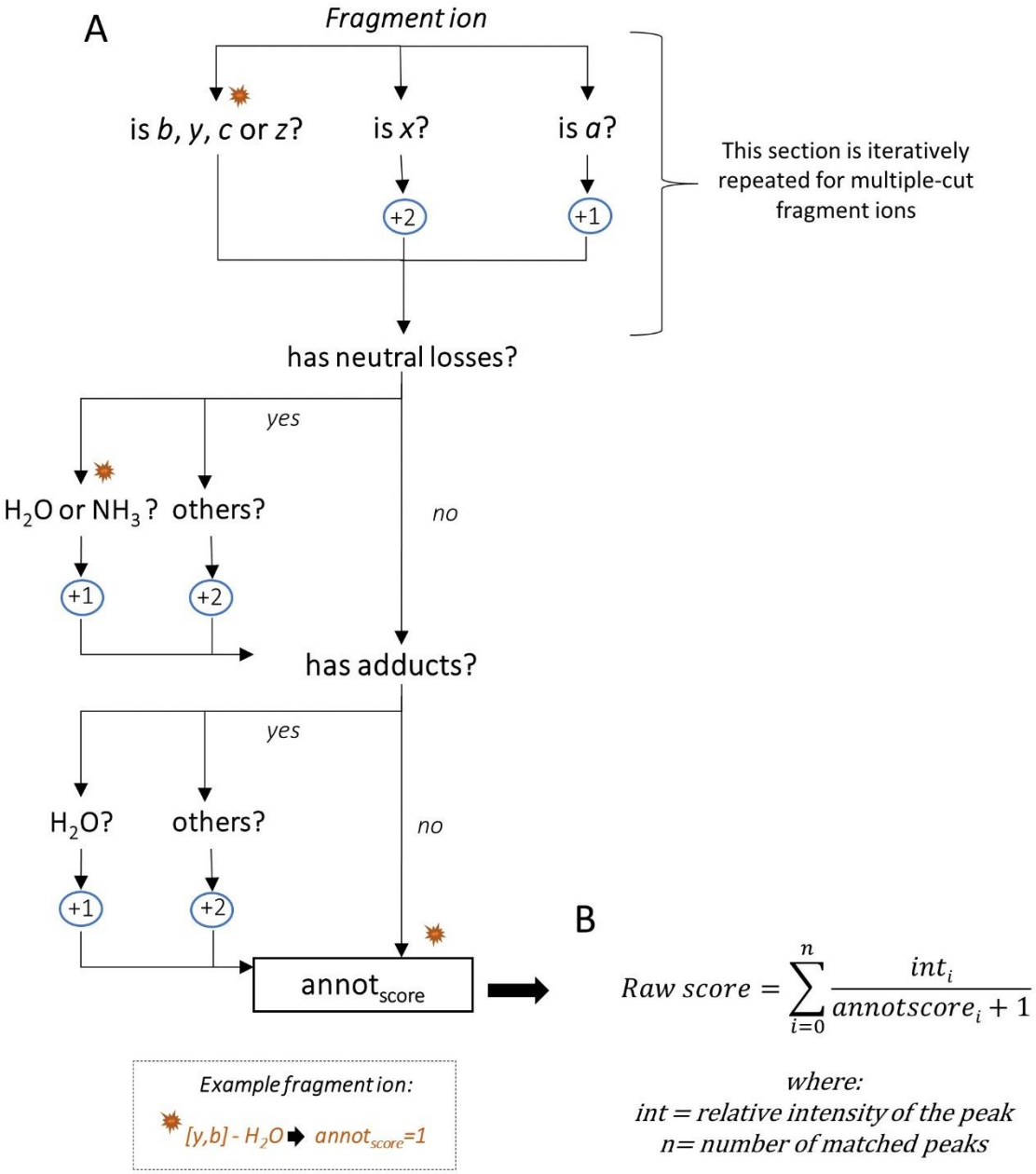
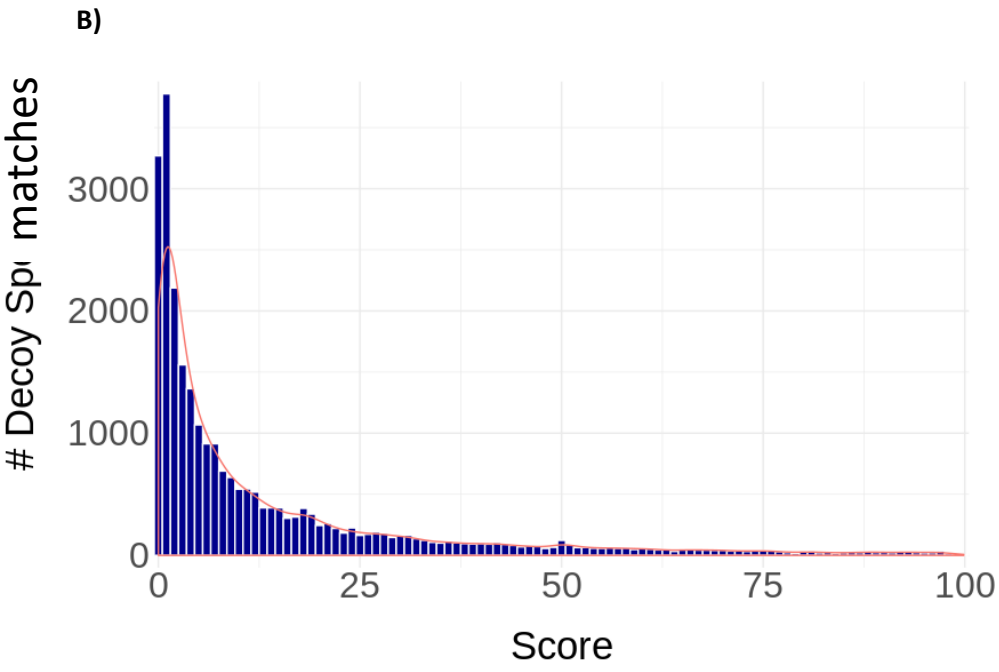
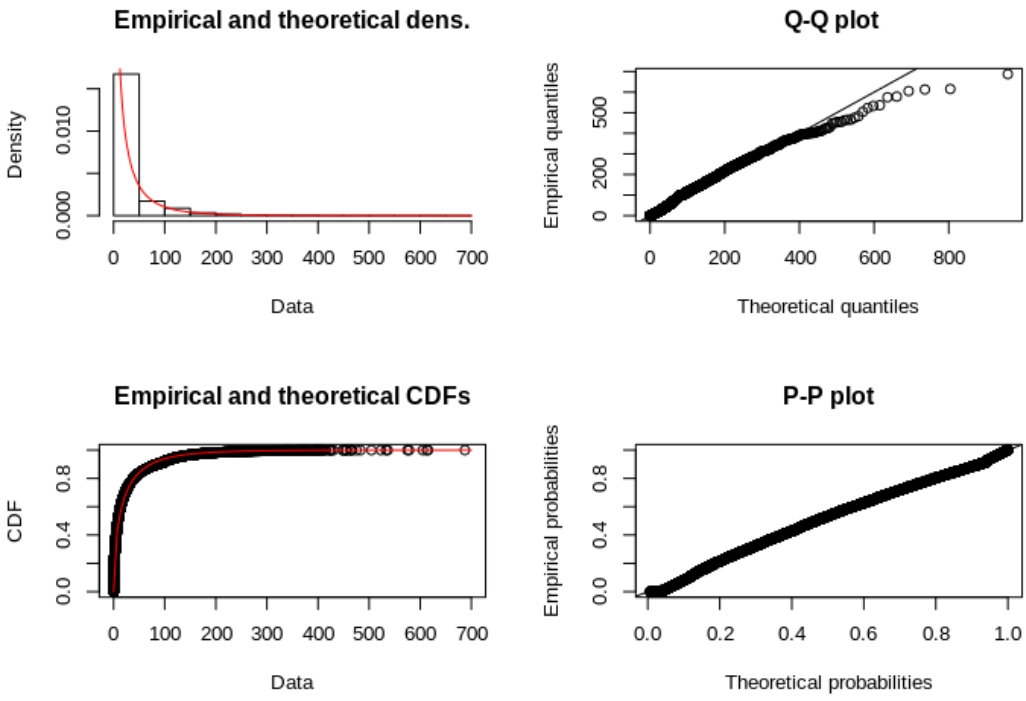


Figure -S3. Decoy distribution. A) The score bar plot obtained after matching each MS/MS from GNPS against the decoy database compounds within a mass range of 0 to +40Da, shows a Weibull distribution. B) Correlations of the theoretical and empirical data after the estimation of a Weibull distribution. C) Comparison of p-values distribution in decoy and target databases.

A)





C)

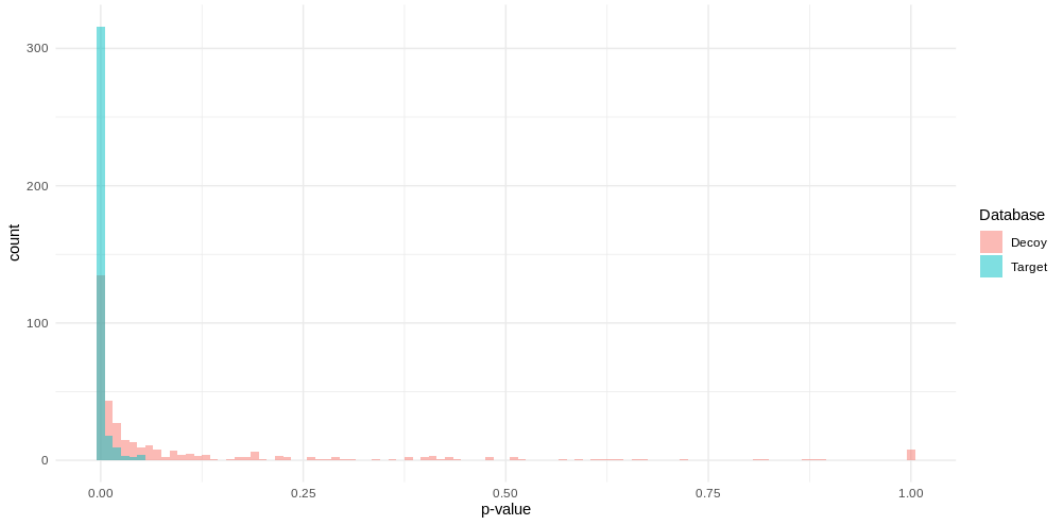


Figure -S4. NRPro interface. After identification, the software provides features for the manual inspection of each match.

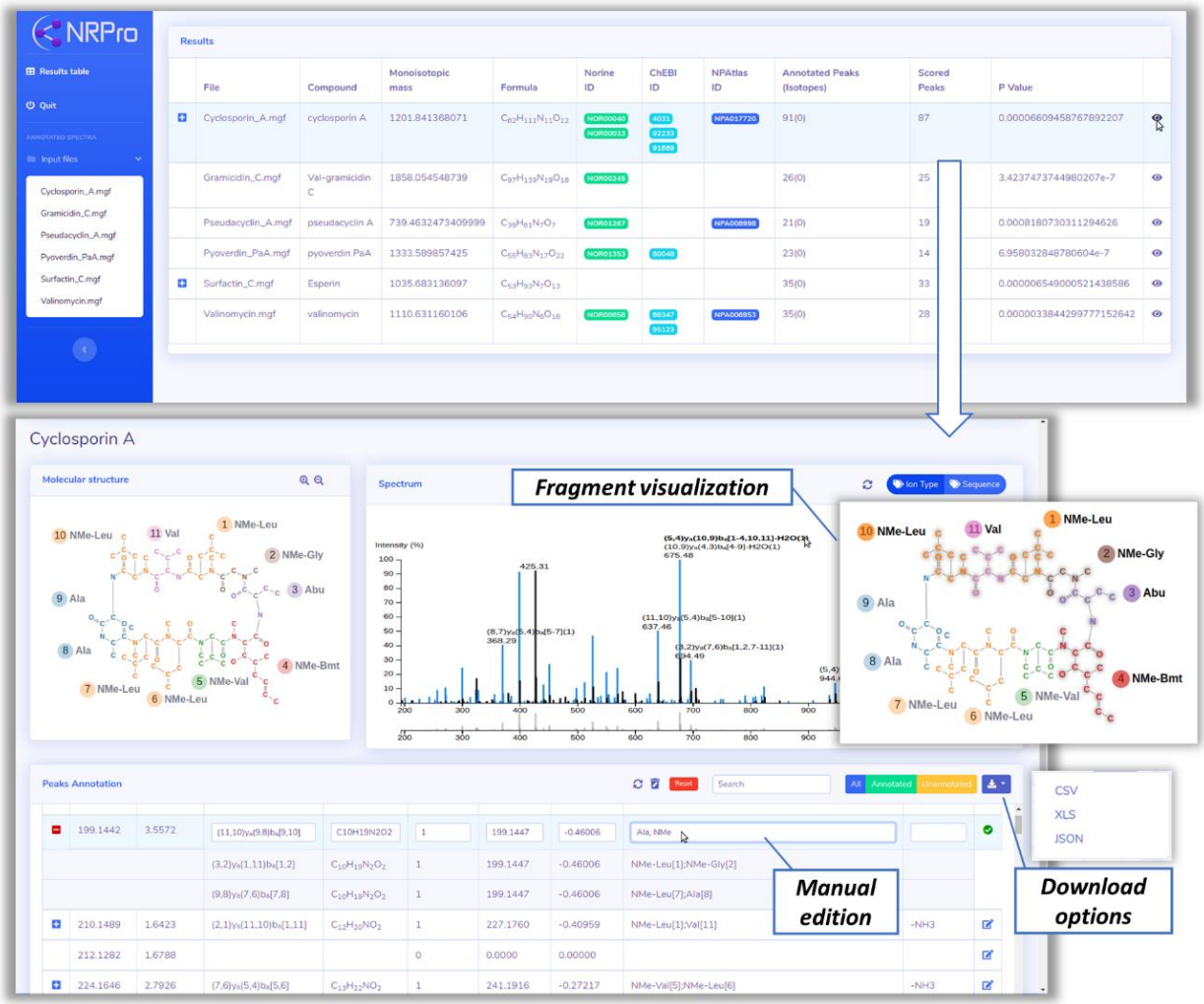


Table -S1. NRPro features benchmark. Note that Cyclobranch is rich in annotation features and Dereplicator is focused on the MS/MS identification, while NRPro provides options from both sections.

		NRPro	Cyclobranch	Dereplicator
MS/MS Identification	Identifies x and a ions	✓	✓	
	Identifies c and z ions	✓	✓	
	Allows MS/MS with Na and K adducts	✓		✓
	Analyses complex peptides*	✓		✓
	Provides analogs search			✓
	Provides a scoring system including a decoy DB	✓		✓
Annotation	Includes monomeric annotations	✓	✓	
	Includes molecular visualization	✓		✓
	Provides multiple annotations per peak	✓	✓	
	Allows manual introduction of the sequence		✓	
	Provides manual edition	✓	✓	
Other	Number of possible spectra analyzed	<20	1	>100
	Default database size	~26000	~1000	>30000

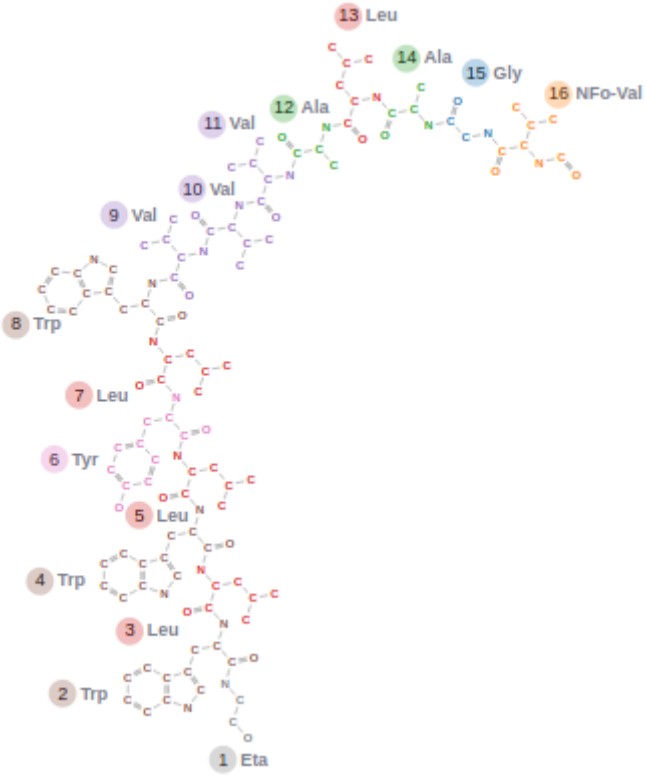
**Complex peptides are those with multiple cycles and branches*

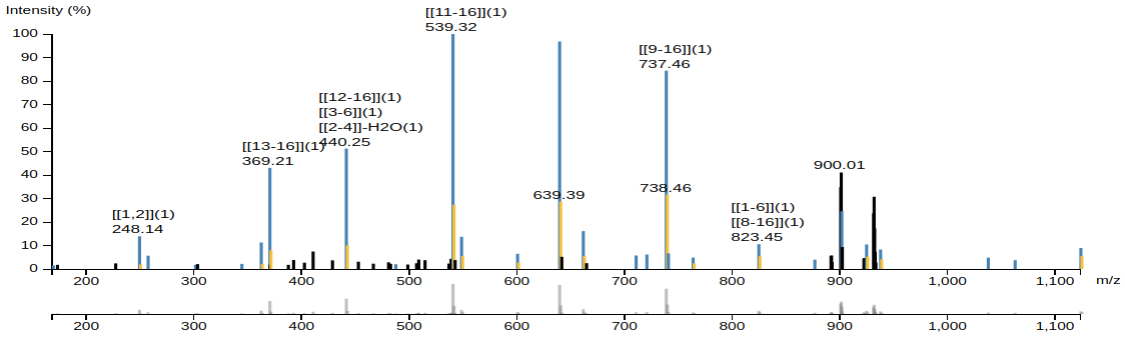
Table -S2. Data collected for the benchmarking of the annotations. To facilitate benchmarking, MS/MS spectra previously used in the publication of each software (see "Publication" column) was included in the set. Compounds with known MS/MS characterization in literature were also prioritized. All the spectra used are freely available in GNPS.

Compound	m/z	*Structure*	*GNPS identifier*	*Int filter*	*Publication*
Gramicidin C (M+H)	1859	Linear	CCMSLIB00000531490	-	Cyclobranch[1]
Leucinostatin B(M+2H)	602	Linear	CCMSLIB00000577504	5%	Dereplicator[2]
Roseotoxin A(M+H)	608	Cyclic	CCMSLIB00000531484	-	Cyclobranch[1]
Dolastatin 12(M+2H)	485	Cyclic	CCMSLIB00000577492	5%	Dereplicator[2]
Pseudacyclin A(M+H)	740	Cyclic-branched	CCMSLIB00000531485	-	Cyclobranch[1]
Plipastatin A1(M+2H)	732	Cyclic-branched	CCMSLIB00000577493	5%	Dereplicator[2]
Vibriobactin(M+H)	706	Linear-branched	CCMSLIB00005435753	3%	-
Enterobactin(M+H)	670	Complex	CCMSLIB00005435752	1%	-
Actinomycin D(M+H)	1255	Complex	CCMSLIB00000479204	-	-
Vancomysin(M+H)	1448	Complex	CCMSLIB00000075312	1%	-

Figure -S5 Benchmarking of the automatic annotations. A fragment ion tolerance of 0.01 was used for the matching of the MS/MS peaks.

A) Gramicidin C (M+H)



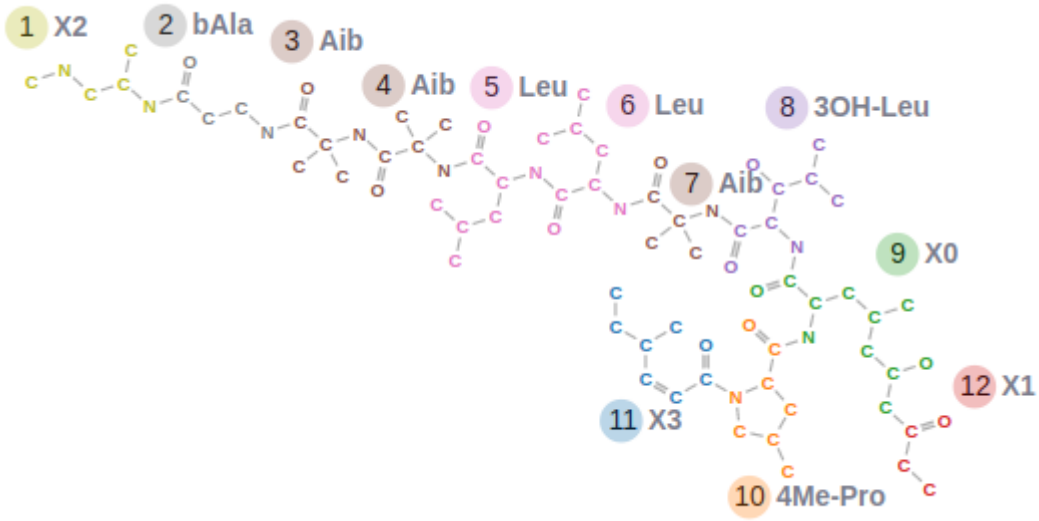


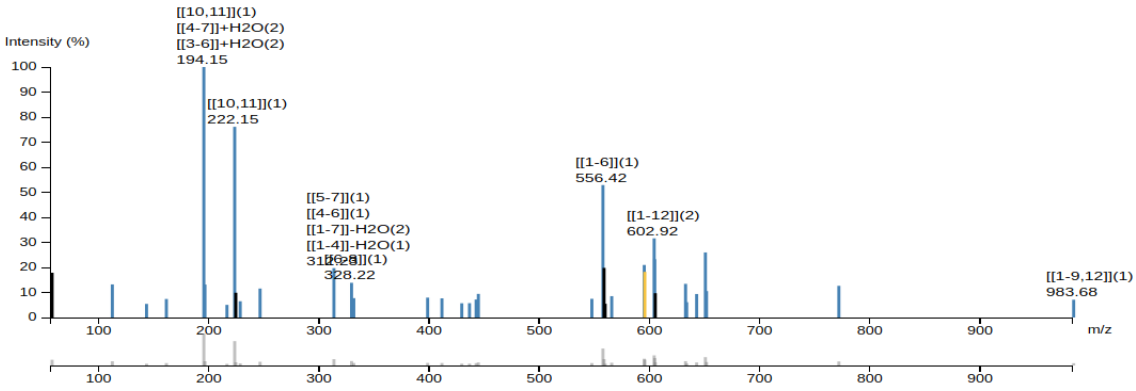
m/z	Intensity	Ion	NRPro	Cyclobranch	Dereplicator
248,1393	13,9038	(3,2)y_a[1,2]			
256,1292	5,6483	(14,13)b_a[14-16]			
300,1707	1,7882	(9,8)y_a(7,6)b_a[7,8]			
343,2131	2,1787	(4,3)y_a[1-3]			
361,2234	11,2716	(4,3)y_a[1-3]			
369,2132	43,0517	(13,12)b_a[13-16]			
440,2503	51,236	(12,11)b_a[12-16]			
486,2496	2,0225	(5,4)y_a(2,1)b_a[2-4]			

539,3188	100	(11,10)b_a[11-16]			
547,3027	13,7024	(5,4)y_a[1-4]			
599,3342	6,4771	(6,5)y_a(2,1)b_a[2-5]			
638,3873	96,8199	(10,9)b_a[10-16]			
660,387	16,1254	(6,5)y_a[1-5]			
709,461	5,7513	(9,8)a_a[9-16]			
719,4454	6,1461	(9,8)b_a[9-16]			
737,4558	84,3991	(9,8)b_a[9-16]			
762,398	4,84	(9,8)y_a(4,3)b_a[4-8]			
823,4506	10,5667	(7,6)y_a[1-6]			
875,4813	3,9844	(9,8)y_a(3,2)b_a[3-8]			
900,5118	24,5823	(12,11)x_a(5,4)b_a[5-12]			
923,5354	10,4299	(8,7)b_a[8-16]			
936,5344	8,2947	(8,7)y_a[1-7]			
1036,6188	4,8233	(7,6)b_a[7-16]			
1061,5619	3,7673	(9,8)y_a(2,1)b_a[2-8]			
1122,6145	8,9413	(9,8)y_a[1-8]			

Simple match
Match with neutral losses
Unmatched

B) Leucinostatin B (M+2H)



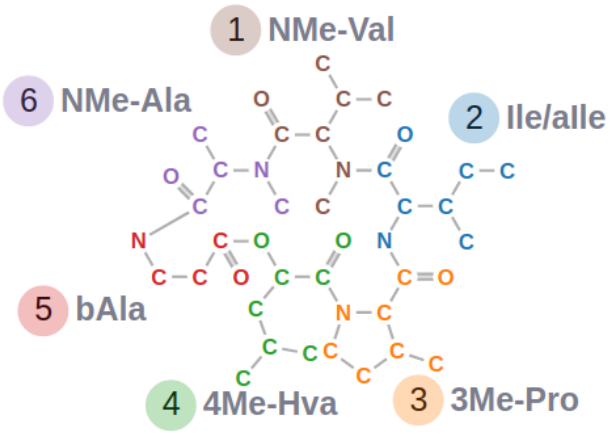


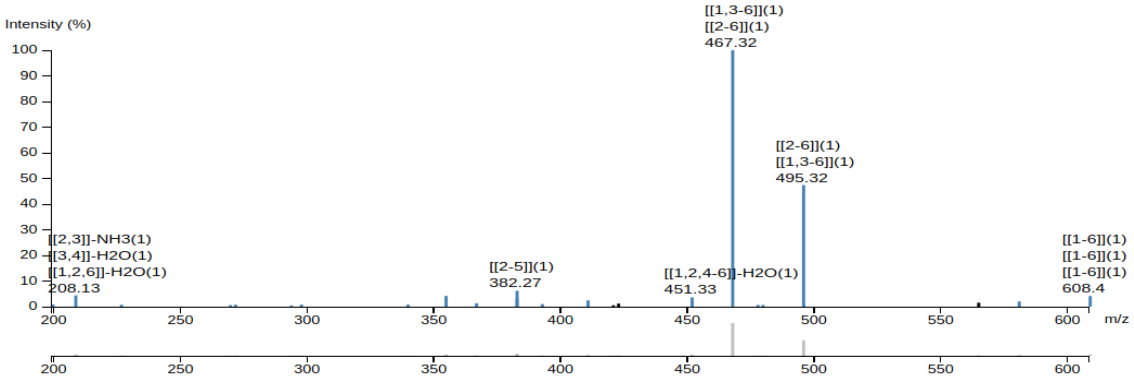
m/z	Intensity	Ion	Charge	NRPro	Cyclobranch	Dereplicator
111,0816	13,2134	(11,10)b_a[11]	1			
142,135	5,4337	(3,2)y_a[1,2]	1			
160,1439	7,3968	(3,2)y_a[1,2]	1			
194,1546	100	(10,9)a_a[10,11]	1			
215,1399	5,0447	(9,8)y_a(7,6)b_a[7,8]	1			
222,1481	76,1259	(10,9)b_a[10,11]	1			
227,1791	6,4912	(7,6)y_a(5,4)b_a[5,6]	1			
245,1981	11,5602	(4,3)y_a[1-3]	1			
312,2278	19,7107	(8,7)y_a(5,4)b_a[5-7]	1			
328,2243	13,8577	(9,8)y_a(6,5)b_a[6-8]	1			
330,2489	7,7311	(5,4)y_a[1-4]	1			
397,2814	7,9438	(7,6)y_a(3,2)b_a[3-6]	1			
410,267	7,6582	(10,9)y_a(7,6)b_a[7-9,12]	1			
428,2813	5,6889	(10,9)y_a(7,6)b_a[7-9,12]	1			
435,2807	5,7315	(9,8)b_a[9-12]	1			
441,3121	7,1841	(9,8)y_a(5,4)b_a[5-8]	1			

443,3322	9,4329	(6,5)y_a[1-5]	1			
546,3495	7,4637	(8,7)b_a[8-12]	1			
556,4153	52,884	(7,6)y_a[1-6]	1			
564,3682	8,5273	(8,7)b_a[8-12]	1			
593,9125	20,9749	pre[1-12]	2			
594,4139	18,2216	pre[1-12]	2			
602,9171	31,5991	pre[1-12]	2			
603,4161	23,2724	(7,6)a_a[7-12]	1			
631,4039	13,4322	(7,6)b_a[7-12]	1			
641,471	9,3843	(8,7)y_a[1-7]	1			
649,4137	25,9709	(7,6)b_a[7-12]	1			
650,4178	10,5452	(8,7)x_a[1-8]	1			
770,554	12,6481	(9,8)y_a[1-8]	1			
983,6796	7,099	(10,9)y_a[1-9,12]	1			

Simple match
Match with neutral losses
Not matched
Isotope match

C) Roseotoxin A (M+H)





m/z	Intensity	Ion	NRPro	Cyclobranch	Dereplicator	Literature[3]
199,1441	0,8591	(2,1)y_a(6,5)b_a[1,6]				Confirmed
208,1332	4,4066	(4,3)y_a(2,1)b_a[2,3]				Confirmed
226,1438	0,7713	(5,4)y_e(3,2)b_a[3,4]				
269,186	0,6209	(6,5)y_a(3,2)a_a[3-5]				
271,1653	0,7951	(1,6)y_a(4,3)b_a[4-6]				
293,2224	0,5084	(4,3)y_a(1,6)a_a[1-3]				
297,1809	0,8126	(6,5)y_a(3,2)b_a[3-5]				
339,2279	0,8242	(5,4)y_e(2,1)b_a[2-4]				
354,2387	4,1906	(1,6)y_a(3,2)a_a[3-6]				Confirmed
366,2387	1,3458	(2,1)y_a(4,3)b_a[1,4-6]				
382,2337	3,096	(1,6)y_a(3,2)b_a[3-6]				Confirmed
382,27	6,2033	(6,5)y_a(2,1)a_a[2-5]				
392,2543	1,0019	(6,5)y_a(2,1)b_a[2-5]				
410,2649	2,4486	(6,5)y_a(2,1)b_a[2-5]				Confirmed
451,3279	3,6216	(3,2)y_a(4,3)a_a[1,2,4-6]				
467,3227	100	(2,1)y_a(3,2)a_a[1,3-6]				Confirmed
477,3069	0,6846	(1,6)y_a(2,1)b_a[2-6]				
479,3224	0,6692	(3,2)y_a(4,3)b_a[1,2,4-6]				
495,3176	47,4041	(1,6)y_a(2,1)b_a[2-6]				Confirmed

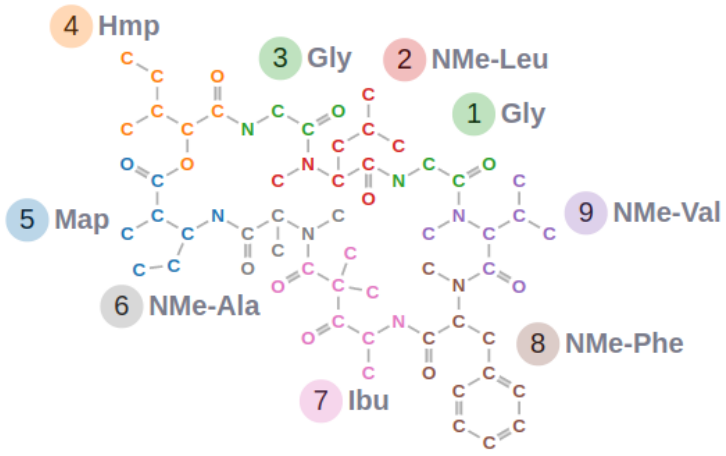
495,3538	0,6043	(6,5)y_a(1,6)a_a[1-5]				
580,4067	2,0384	(2,1)y_a(2,1)a_a[1-6]				Confirmed
608,4018	4,1474	pre[1-6]				

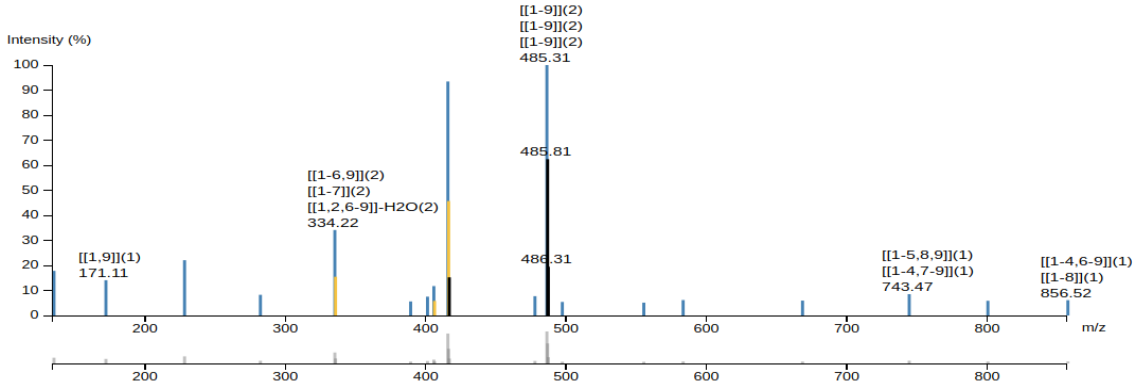
Simple match

Match with neutral losses

Not matched

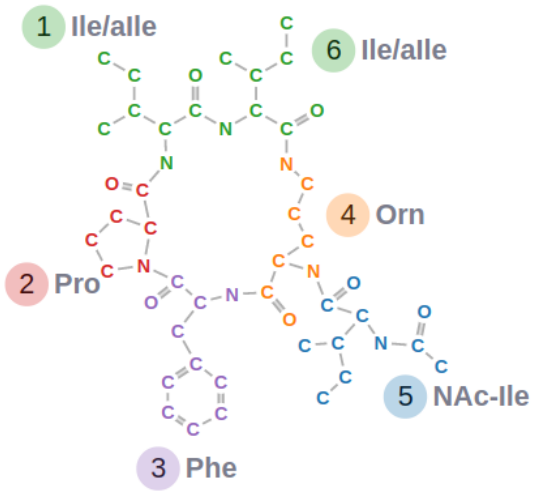
D) Dolastatin 12 (M+2H)

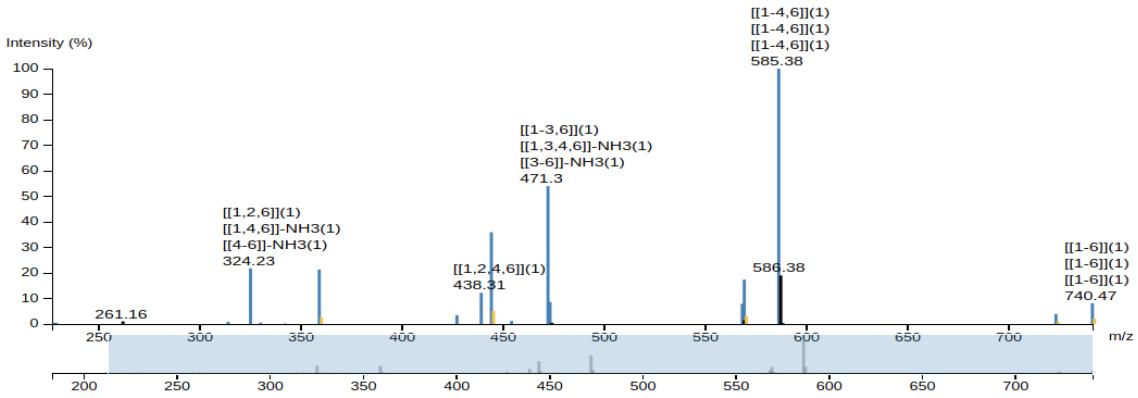




m/z	Intensity	Ion	Charge	NRPro	Cyclobranch	Dereplicator
134,0969	17,8142	(9,8)y_a(8,7)a_a[8]	1			
171,1138	14,0765	(2,1)y_a(9,8)b_a[1,9]	1			
227,14	22,0765	(8,7)y_a(6,5)b_a[6,7]	1			
281,1866	8,2477	(3,2)y_a(9,8)b_a[1,2,9]	1			
334,224	34,1202	(7,6)y_a(9,8)b_a[1-6,9]	2			
388,2248	5,6029	(9,8)y_a(6,5)b_a[6-8]	1			
400,2538	7,5046	(9,8)y_a(2,1)b_a[2-8]	2			
404,7631	11,7377	(8,7)y_a(9,8)b_a[1-7,9]	2			
405,2644	5,8214		2			
414,7674	93,4426	(7,6)y_a(8,7)b_a[1-6,8,9]	2			
476,7948	7,7086	pre[1-9]	2			
485,3073	100	pre[1-9]	2			
496,2967	5,4062	(5,4)x_e(9,8)b_a[1-5,9]	1			
554,355	5,1148	(7,6)y_a(1,9)b_a[1-6]	1			
582,3862	6,1566	(6,5)y_a(9,8)b_a[1-5,9]	1			
667,4398	5,9381	(7,6)y_a(9,8)b_a[1-6,9]	1			
743,4689	8,51	(6,5)y_a(8,7)b_a[1-5,8,9]	1			
799,499	5,8652	(9,8)y_a(2,1)b_a[2-8]	1			
856,5167	6,1129	(5,4)y_e(6,5)b_a[1-4,6-9]	1			

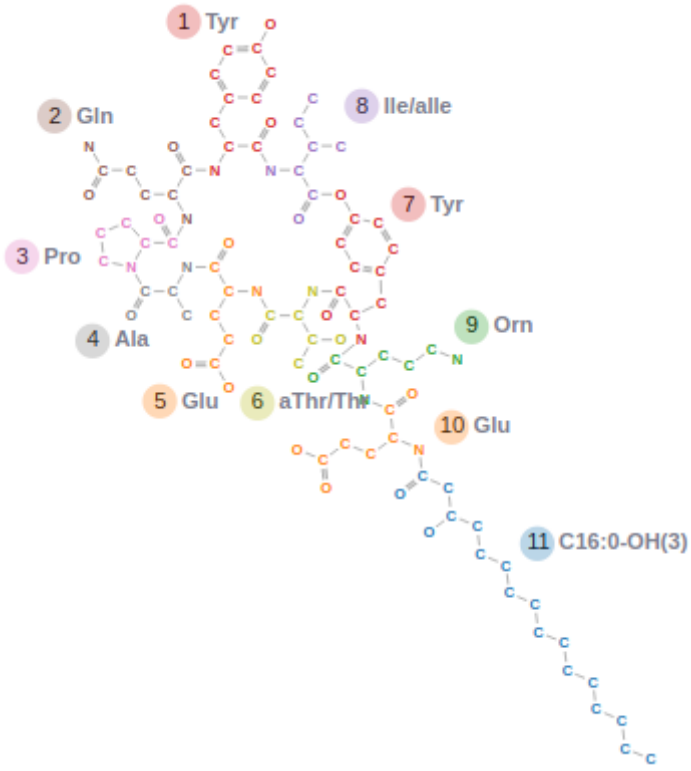
E) Pseudacyclin A (M+H)

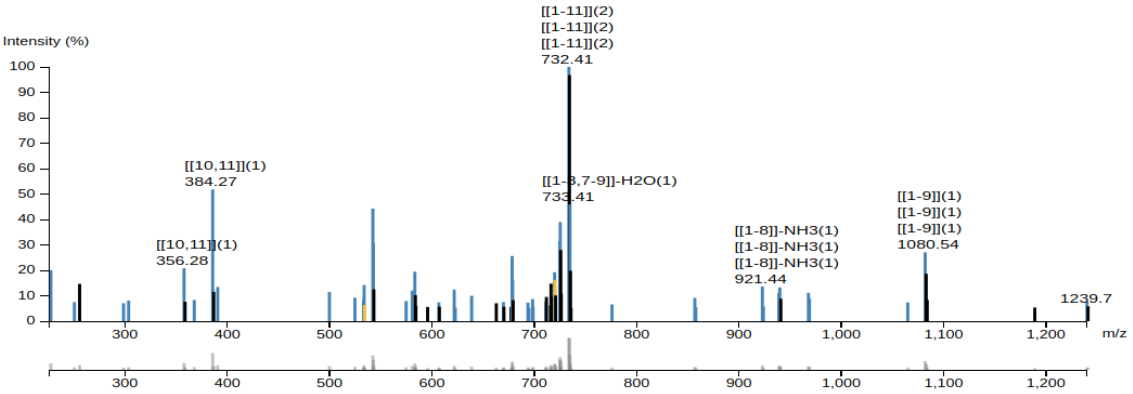




m/z	Intensity	Ion	NRPro	Cyclobranch	Dereplicator	Literature[4]
183,1492	0,5387	(3,2)y_a(1,6)a_a[1,2]				Confirmed
211,1441	1,8732	(3,2)y_a(1,6)b_a[1,2]				Confirmed
227,1754	0,5042	(2,1)y_a(6,4)b_a[1,6]				
228,1706	0,4622	(6,4)y_a(5,5)y_a(4,3)b_a[4,5]				
313,1911	0,8531	(4,3)y_a(1,6)a_a[1-3]				
324,2282	21,7768	(3,2)y_a(6,4)b_a[1,2,6]				Confirmed
341,2546	0,1914	(5,4)y_a(2,1)y_a(4,3)b_a[1,4,6]				
358,2125	21,3079	(4,3)y_a(1,6)b_a[1-3]				Confirmed
426,2751	3,4428	(4,3)y_a(6,4)a_a[1-3,6]				
438,3074	12,2356	(3,2)y_a(5,4)y_a(4,3)b_a[1,2,4,6]				
443,3016	35,9432	(4,3)y_a(6,4)a_a[1-3,6]				Confirmed
453,2861	1,2025	(4,3)y_a(6,4)b_a[1-3,6]				
471,2965	54,0112	(4,3)y_a(6,4)b_a[1-3,6]				Confirmed
472,2992	8,6054	(6,4)y_a(5,5)y_a(2,1)b_a[2-5]				
567,3653	7,926	(5,4)y_a[1-4,6]				
568,3492	17,3956	(5,4)y_a[1-4,6]				
585,3759	100	(5,4)y_a[1-4,6]				Confirmed
722,4599	3,9034	pre[1-6]				Confirmed
740,4705	8,1302	pre[1-6]				Confirmed

F) Plipastatin A1 (M+2H)



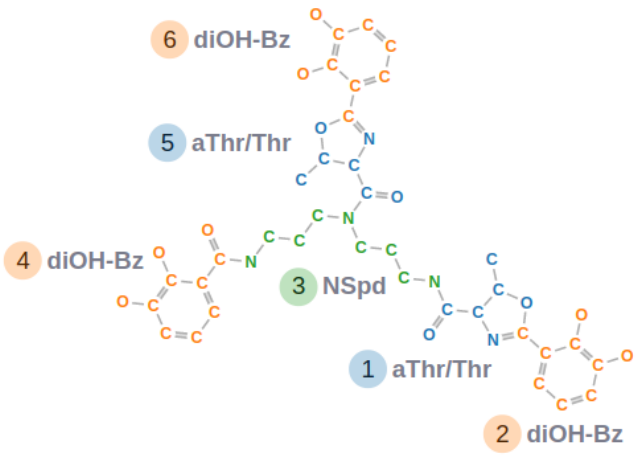


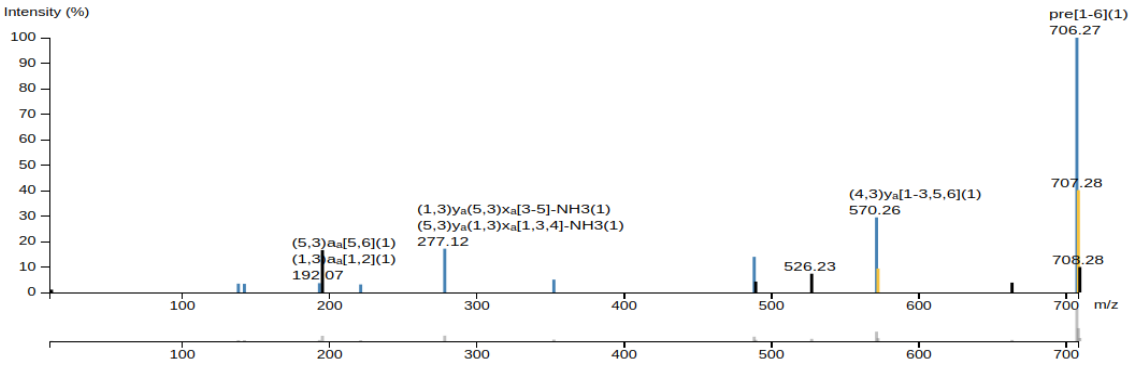
m/z	Intensity	Ion	Charge	NRPro	Cyclobranch	Dereplicator
226,1193	20,059	(4,3)y_a(2,1)b_a[2,3]	1			
249,1596	7,5251	(9,7)y_a(1,8)y_a(7,6)a_a[7,8]	1			
297,1558	7,0435	(5,4)y_a(2,1)b_a[2-4]	1			
302,1346	8,1031	(7,6)y_a(4,3)b_a[4-6]	1			
356,2791	20,8657	(10,9)a_a[10,11]	1			
366,2646	8,3559	(10,9)b_a[10,11]	1			
384,2741	51,791	(10,9)b_a[10,11]	1			
389,1833	13,4489	(4,3)y_a(1,8)b_a[1-3]	1			
498,3526	11,4683	(9,7)b_a[9-11]	1			

523,2581	9,2409	(9,7)y_a(2,1)y_a(6,5)b_a[1,6-8]	1			
531,7703	8,5486	(10,9)y_a[1-9]	2			
532,2627	14,2075	(10,9)y_a[1-9]	2			
540,7736	44,2719	(10,9)y_a[1-9]	2			
541,2722	30,7266	(9,7)y_a(2,1)y_a(6,5)b_a[1,6-8]	1			
573,3281	7,9104	(4,3)y_a(7,6)b_a[1-3,7-11]	2			
579,275	12,0763	(6,5)y_a(1,8)a_a[1-5]	1			
581,843	19,451	(4,3)y_a(7,6)b_a[1-3,7-11]	2			
582,341	10,2161					
594,3318	5,5927					
605,2965	7,3867	(11,10)y_a[1-10]	2			
620,3129	12,4014	(4,3)y_a(9,7)y_a(7,6)a_a[1-3,7,8]	1			
637,3359	9,9633	(10,9)y_a(2,1)y_a(6,5)b_a[1,6-9]	1			
661,4101	7,0435					
668,3828	7,483	(2,1)y_a(3,2)b_a[1,3-11]	2			
675,8671	5,6107					
676,8896	25,6336	(5,4)y_a(6,5)b_a[1-4,6-11]	2			
692,3619	7,2783	(10,9)y_a(1,8)y_a(4,3)b_a[4-9]	1			
693,3542	5,2134	(1,8)y_a(9,7)y_a(3,2)b_a[3-8]	1			
696,8928	8,7051	(4,3)y_a(5,4)b_a[1-3,5-11]	2			
697,3847	5,5024					
709,9015	7,477	(2,1)y_a(2,1)a_a[1-11]	2			
718,4102	19,2342	(2,1)y_a(2,1)a_a[1-11]	2			
723,4042	31,5032	pre[1-11]	2			
732,41	100	pre[1-11]	2			
733,4116	45,8913	(10,9)y_a(4,3)y_a(7,6)a_a[1-3,7-9]	1			
774,4996	6,586	(1,8)y_a(7,6)b_a[7-11]	1			
855,4249	9,1746	(10,9)y_a(2,1)y_a(4,3)b_a[1,4-9]	1			
921,4376	13,6476	(9,7)y_a(2,1)y_a(2,1)a_a[1-8]	1			
922,4406	5,7913					
937,569	11,065	(2,1)y_a(7,6)b_a[1,7-11]	1			
939,468	8,7894					
966,4601	11,1312	(9,7)y_a[1-8]	1			
967,4581	8,8676					
1063,5224	7,3746					
1080,5396	27,0724	(10,9)y_a[1-9]	1			
1187,6566	5,376					
1238,7001	7,6877	(2,1)y_a(4,3)b_a[1,4-11]	1			

Simple match
Match with neutral losses
Not matched

G) Vibriobactin (M+H)

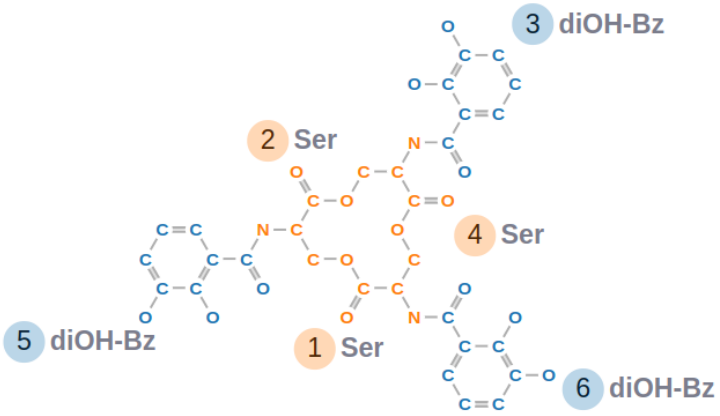


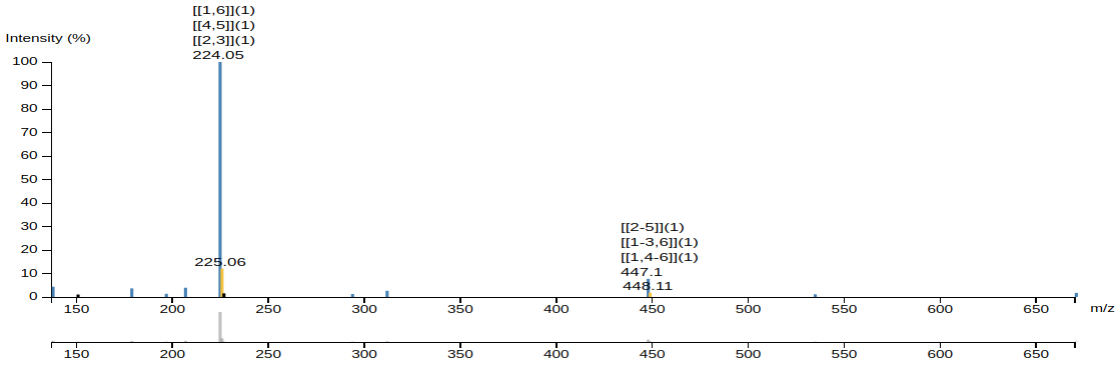


m/z	Intensity	Ion	NRPro	Cyclobranch
137,0228	3,4526	(4,3)b_a[4]		
141,1018	3,3831	(1,3)y_a(4,3)y_a(5,3)x_a[3,5]		
192,0652	3,6518	(5,3)a_a[5,6]		
220,0603	3,1435	(5,3)b_a[5,6]		
277,1183	17,1768	(1,3)y_a(5,3)x_a[3-5]		
351,203	5,045	(4,3)y_a(5,3)y_a[1-3]		
487,219	13,9864	(5,3)y_a[1-4]		
570,2567	29,4619	(4,3)y_a[1-3,5,6]		
706,2726	100	pre[1-6]		

Simple match
Match with neutral losses
Not matched

H) Enterobactin (M+H)

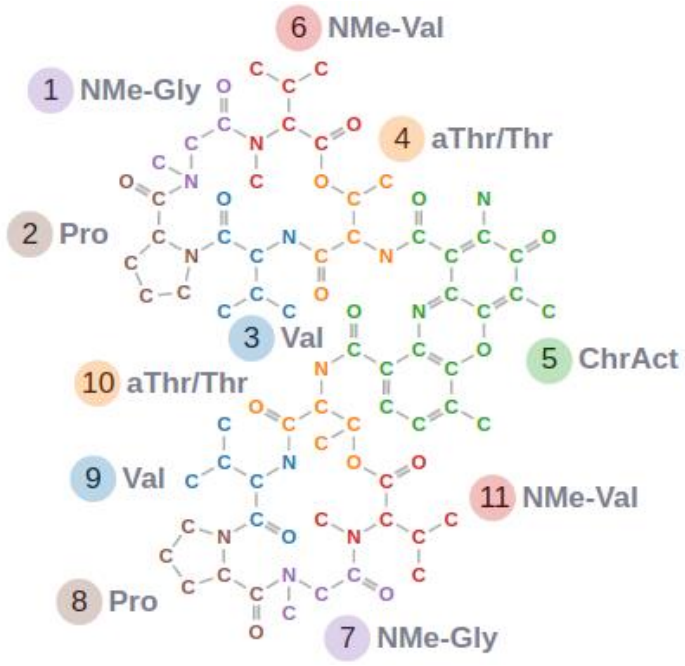


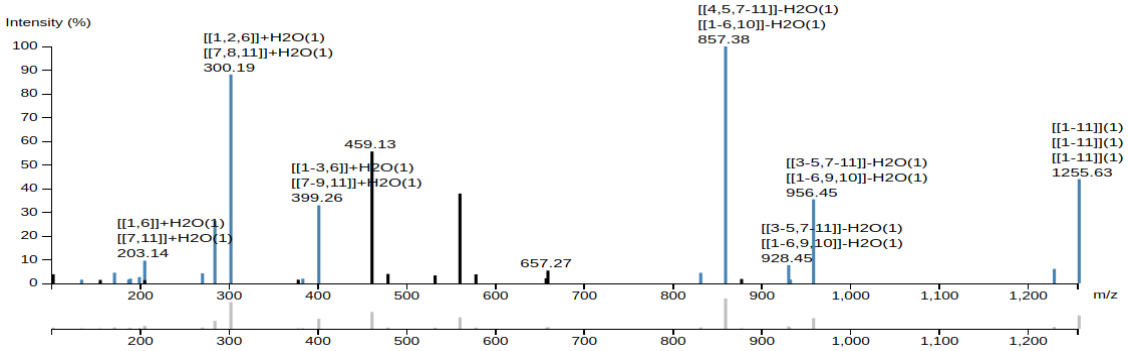


m/z	Intensity	Ion	NRPro	Dereplicator
137,0221	4,3845	(6,1)b_a[6]		
178,0493	3,6674	(1,4)y_e(4,2)a_e[4,5]		
196,0594	1,3601	(1,4)y_e(4,2)a_e[4,5]		
206,044	3,94	(2,1)y_e(1,4)b_e[1,6]		
224,0547	100	(2,1)y_e(1,4)b_e[1,6]		
293,0763	1,2858	(4,2)y_e(6,1)y_a(1,4)b_e[1-3]		
311,0869	2,6394	(4,2)y_e(6,1)y_a(1,4)b_e[1-3]		
447,1027	7,7163	(1,4)y_e(2,1)b_e[2-5]		
534,1357	1,1556	(6,1)y_a[1-5]		
670,1512	1,7538	pre[1-6]		

Simple match
Match with neutral losses
Theoretical match
Theoretical unmatch

I)Actinomycin D (M+H)



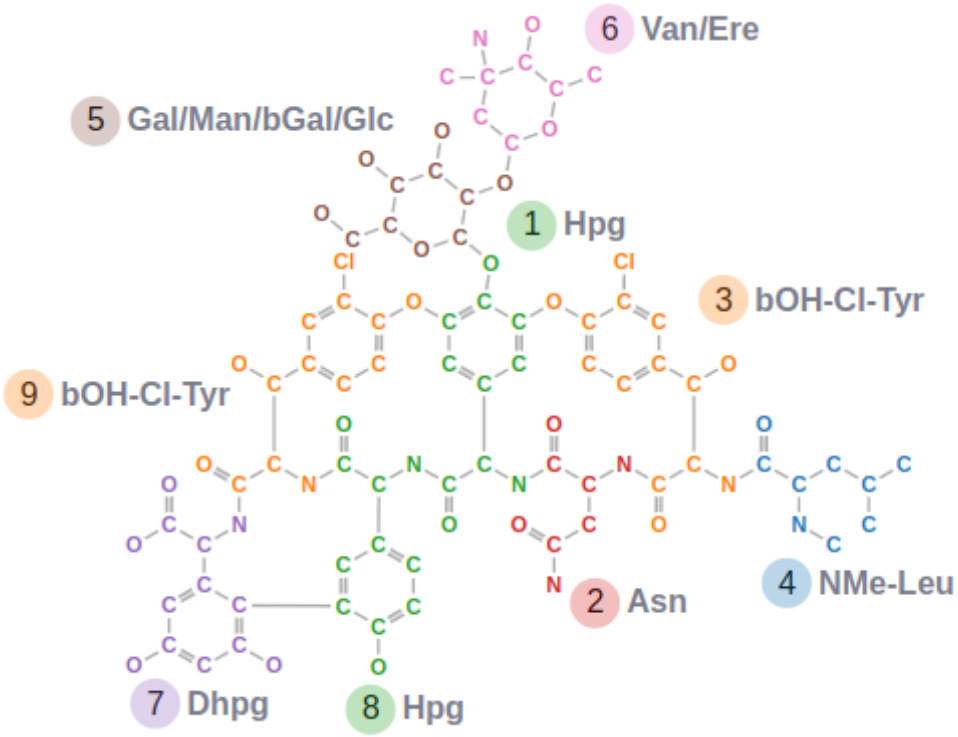


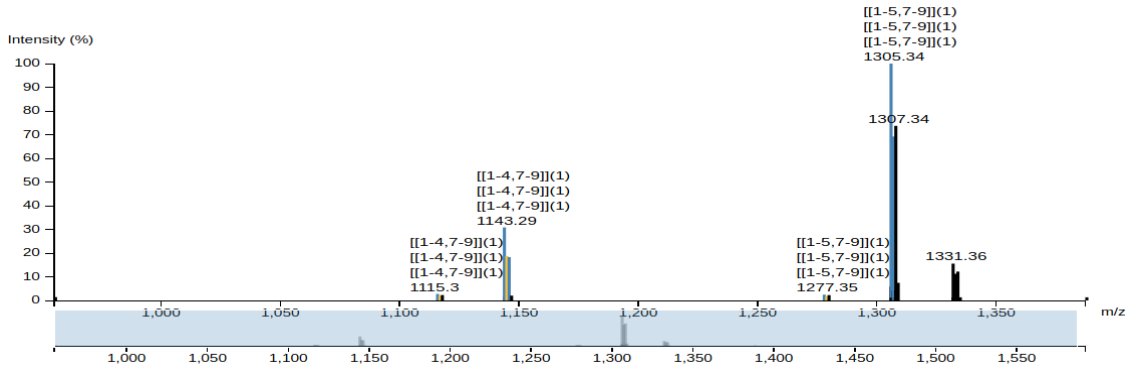
m/z	Intensity	Ion	NRPro	Dereplicator
132,1001	1,5701	(1,6)y_a(6,4)b_e[6]		
169,0962	4,524	(9,8)y_a(7,11)b_a[7,8]		
185,1284	1,6419	(2,1)y_a(6,4)b_e[1,6]		
187,1059	2,0341	(9,8)y_a(7,11)b_a[7,8]		
197,1286	2,6825	(4,3)y_a(2,1)b_a[2,3]		
203,1386	9,5842	(2,1)y_a(6,4)b_e[1,6]		
268,1644	4,2518	(10,9)y_a(7,11)b_a[7-9]		
282,1812	26,458	(3,2)y_a(6,4)b_e[1,2,6]		
300,1917	88,1716	(3,2)y_a(6,4)b_e[1,2,6]		
381,2461	1,9877	(4,3)y_a(6,4)b_e[1-3,6]		
399,2597	32,9792	(4,3)y_a(6,4)b_e[1-3,6]		
829,3876	4,4769	(6,4)y_e(4,3)a_a[4,5,7-11]		
857,3826	100	(6,4)y_e(4,3)b_a[4,5,7-11]		
875,3838	1,86	(6,4)y_e(4,3)b_a[4,5,7-11]		

928,4522	7,7209	(6,4)y_e(3,2)a_a[3-5,7-11]		
956,4492	35,4853	(6,4)y_e(3,2)b_a[3-5,7-11]		
1227,6365	6,1141	(2,1)y_a(2,1)a_a[1-11]		
1255,6337	43,9874	pre[1-11]		

Simple match
Match with neutral losses
Not matched

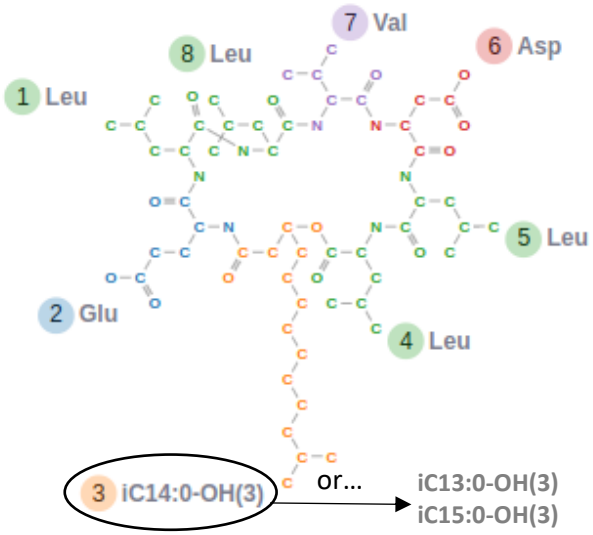
J) Vancomycin (M+H)





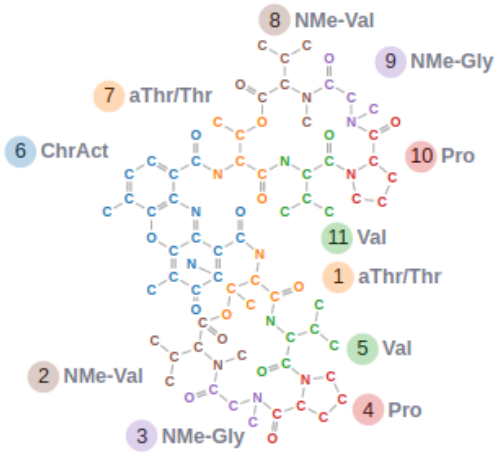
m/z	Intensity	Ion	NRPro
1115,2959	2,7365	(5,1)y_g(9,7)y_a(9,7)a_a[1-4,7-9]	
1143,2295	1,6655	(6,5)z_g(4,3)y_a[1-3,5,7-9]	
1143,292	30,8257	(5,1)y_g[1-4,7-9]	
1145,2897	18,274	(6,5)y_g(2,1)y_a(3,2)a_a[1,3-5,7-9]	
1146,2924	2.0293	(6,5)z_g(2,1)y_a(3,2)a_a[1,3-5,7-9]	
1277,3497	2,4274	(6,5)y_g(9,7)y_a(9,7)a_a[1-5,7-9]	
1305,3448	100	(6,5)y_g[1-5,7-9]	
1306.3484	69,2167	(6,5)z_g[1-5,7-9]	

Figure -S6. Characterization of surfactins C13, C14 and C15 using MS/MS spectra with sodium adducts. GNPS identifiers: CCMSLIB00000006889, CCMSLIB00000006894, CCMSLIB00000006898. The parent and fragment ion tolerance used was of 0.1 Da. H_2O neutral losses and adducts were included.



	Surfactin C13 (m/z 1030)	Surfactin C14 (m/z 1044)	Surfactin C15 (m/z 1058)	Fragment ion	Sequence
Peptidic series	463	463	463	(8,7)y_a(4,3)b_e[4-7](1)	Leu[4];Leu[5];Asp[6];Val[7]
	481	481	481	(8,7)y_a(4,3)b_e[4-7]+H2O(1)	Leu[4];Leu[5];Asp[6];Val[7]
	594	594	594	(1,8)y_a(4,3)b_e[4-8]+H2O(1)	Leu[4];Leu[5];Asp[6];Val[7];Leu[8]
	707	707	707	(2,1)y_a(4,3)b_e[1,4-8]+H2O(1)	Leu[1];Leu[4];Leu[5];Asp[6];Val[7];Leu[8]
Aliphatic series	590	604	618	(4,3)y_e(8,7)b_a[1-3,8](1)	Leu[1];Glu[2];FA[3];Leu[8]
	689	703	717	(4,3)y_e(7,6)b_a[1-3,7,8](1)	Leu[1];Glu[2]; FA[3];Val[7];Leu[8]
	786	800	814	(4,3)y_e(6,5)b_a[1-3,6-8]-H2O(1)	Leu[1];Glu[2]; FA[3];Asp[6];Val[7];Leu[8]
	804	818	832	(4,3)y_e(6,5)b_a[1-3,6-8](1)	Leu[1];Glu[2]; FA[3];Asp[6];Val[7];Leu[8]
	917	931	945	(4,3)y_e(5,4)b_a[1-3,5-8](1)	Leu[1];Glu[2]; FA[3];Leu[5];Asp[6];Val[7];Leu[8]
	1012	1026	1040	pre[1-8]-H2O(1)	Leu[1];Glu[2]; FA[3];Leu[4];Leu[5];Asp[6];Val[7];Leu[8]

Figure -S7. Annotations of the EID spectra from actinomycin D. The fragment ions are compared with those reported in the reference study[5].



m/z	Ion
169,0971	(11,10)y_a(9,8)b_a[9,10]
282,1812	(11,10)y_a(8,7)b_e[8-10]
300,1917	(11,10)y_a(8,7)c_e[7-10]
354,2386	(7,11)z_a(8,7)a_e[8-11]+H2O
381,2497	(7,11)y_a(8,7)b_e[8-11]
399,2601	(1,5)y_a(2,1)c_e[1-5]
431,1352	
459,13	
508,2765	(6,7)x_a[6-11]
558,1984	
576,1984	
657,2668	
675,2769	
776,3611	(8,7)z_e(8,7)b_e(6,1)b_a[6-11]+H2O
803,3727	
829,3875	(8,7)y_e(7,11)a_a[1-7]-H2O
857,3828	(8,7)y_e(7,11)b_a[1-7]-H2O
875,3929	(8,7)y_e(7,11)b_a[1-7]
928,4568	(8,7)y_e(11,10)a_a[1-7,11]-H2O
956,4512	(8,7)y_e(11,10)b_a[1-7,11]-H2O
974,4615	(8,7)y_e(11,10)b_a[1-7,11]
1057,4987	
1124,5415	(8,7)y_e(9,8)b_a[1-7,9-11]-H2O
1127,5417	(10,9)z_a(11,10)b_a[1-9,11]
1156,5681	(11,10)y_a(7,11)b_a[1-10]
1183,6151	

Different annotation than the reference
Not annotated
Same annotation than the reference

Figure -S8. Number of PNPs identified in the decoy and target databases using different p-value cut offs. At a p-value of 10^{-5} the number of identified target peptides is 169 while that of decoy peptides is 6. That translates into a 3.55% FDR.

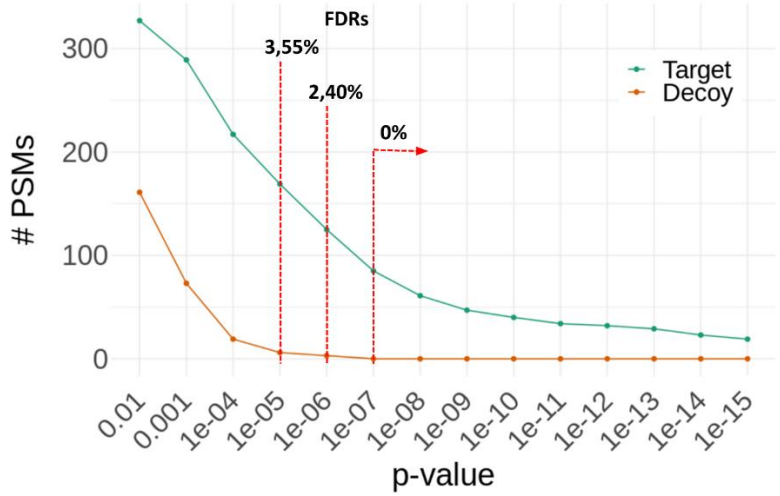


Table -S3. Dereplicated PNPs present in Norine and with a p-value minor than 10^{-5} (FDR 3.55%)

Compound	Norine ID	Structure	Category	Activities	Origin[a]	#Scored/#Annot[b]	P-value
Heptaibin	NOR01001	linear	peptaibol	antibiotic	Fungus	10/19	1,00E-16
Apramide B	NOR00728	linear	lipopeptide	unknown	Bacterium	7/7	1,00E-16
Carmabin B	NOR00433	linear	lipopeptide	surfactant	Bacterium	10/10	1,11E-14
Orfamide C	NOR01257	partial cyclic	lipopeptide	antibiotic,surfactant	Bacterium	33/36	1,28E-12
Apramide A	NOR00409	linear	lipopeptide	unknown	Bacterium	15/15	2,17E-11
Kahalalide G	NOR00396	linear	peptide	unknown	Animal	8/9	4,81E-10
Emerimicin IV	NOR00986	linear	peptaibol	antibiotic	Fungus	67/73	4,03E-08
Massetolide G	NOR00884	partial cyclic	lipopeptide	antibiotic,surfactant	Bacterium	35/37	1,85E-07
Massetolide C	NOR00883	partial cyclic	lipopeptide	antibiotic,surfactant	Bacterium	17/17	2,10E-07
Majusculamide C	NOR00621	cyclic	peptide	antibiotic,antitumor	Bacterium	43/47	2,80E-07
Orfamide B	NOR01256	partial cyclic	lipopeptide	antibiotic,surfactant	Bacterium	15/16	4,36E-07
Destruxin E1	NOR00087	cyclic	peptide	toxin	Fungus	32/36	4,41E-07
Dolastatin 15	NOR00370	linear	peptide	antitumor	Animal	7/7	4,78E-07
Massetolide E	NOR00335	partial cyclic	lipopeptide	antibiotic,surfactant	Bacterium	6/7	1,12E-06
Dolastatin 10	NOR00369	linear	peptide	antitumor	Animal	14/15	2,38E-06
Beauvericin D	NOR00254	cyclic	peptide	antibiotic	Fungus	20/20	3,90E-06
Isariin	NOR00648	cyclic	peptide	antibiotic	Fungus	28/31	4,43E-06
Carmabin A	NOR00432	linear	lipopeptide	surfactant	Bacterium	12/12	6,16E-06

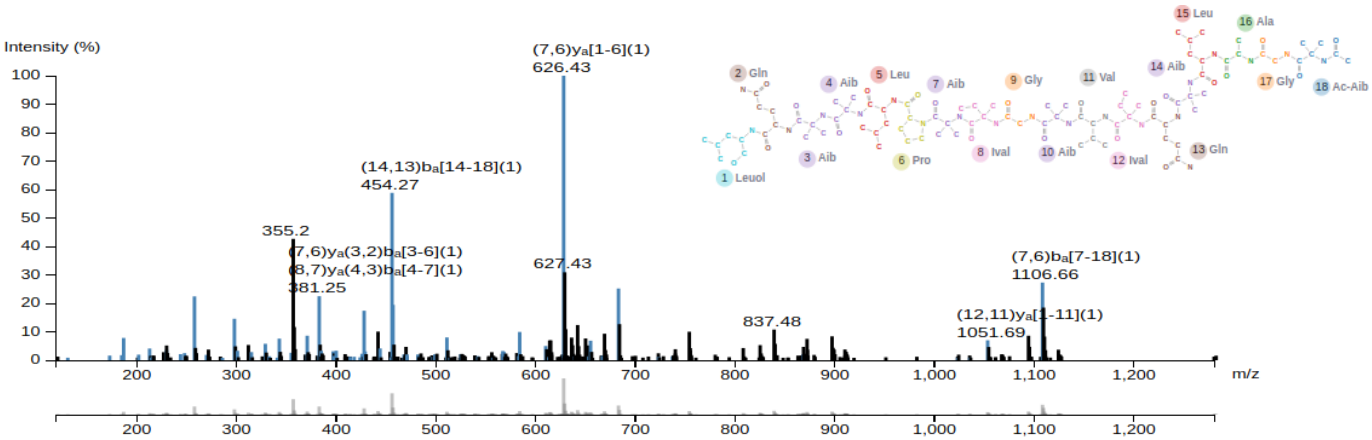
The parent and fragment ion tolerance used in the search were of 0.02. The annotations are automatically extracted from Norine database.

[a]The origin is retrieved from NPAtlas and not Norine. Some were manually introduced.

[b]Note that in NRPro the annotated peaks are not always used for scoring. Here, both the number of scored and annotated peaks are shown.

Figure -S9. Example of an uncommon match between NRPro and Dereplicator. The GNPS spectra CCMSLIB00000577654 was identified as (A) thricovirin II 5 by NRPro (p-value 7.706^{-7}) while Dereplicator suggested (B) thricorzin HA-5 (p-value 1.49^{-18}), which appeared as the 5th candidate (p-value 3.16^{-4}) of NRPro.

A)



B)

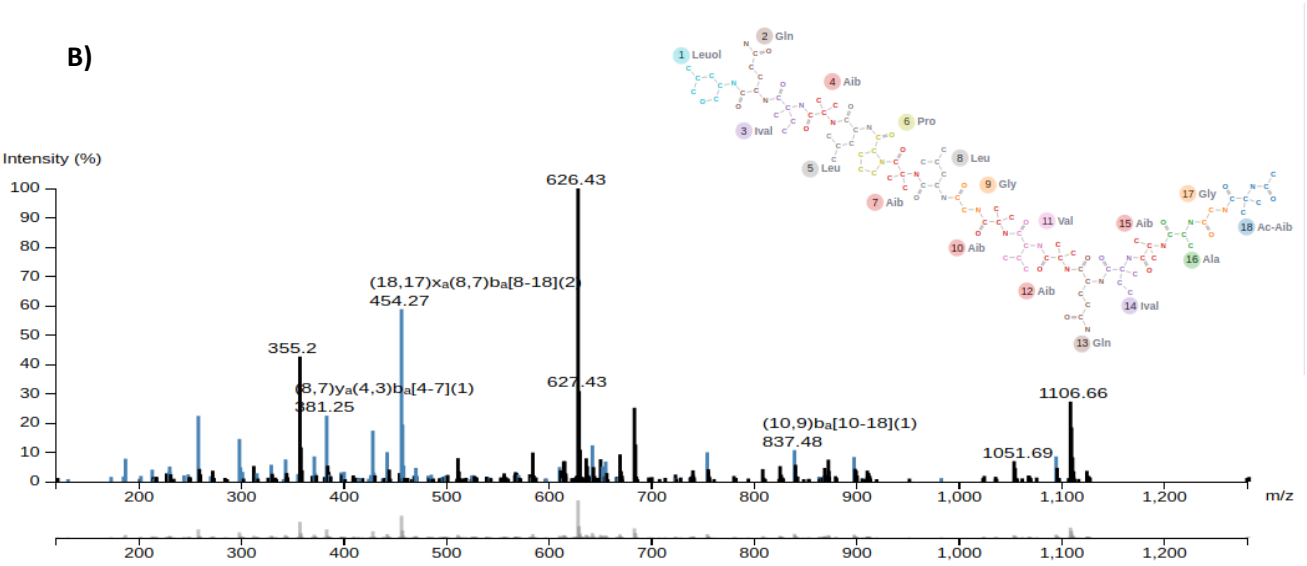
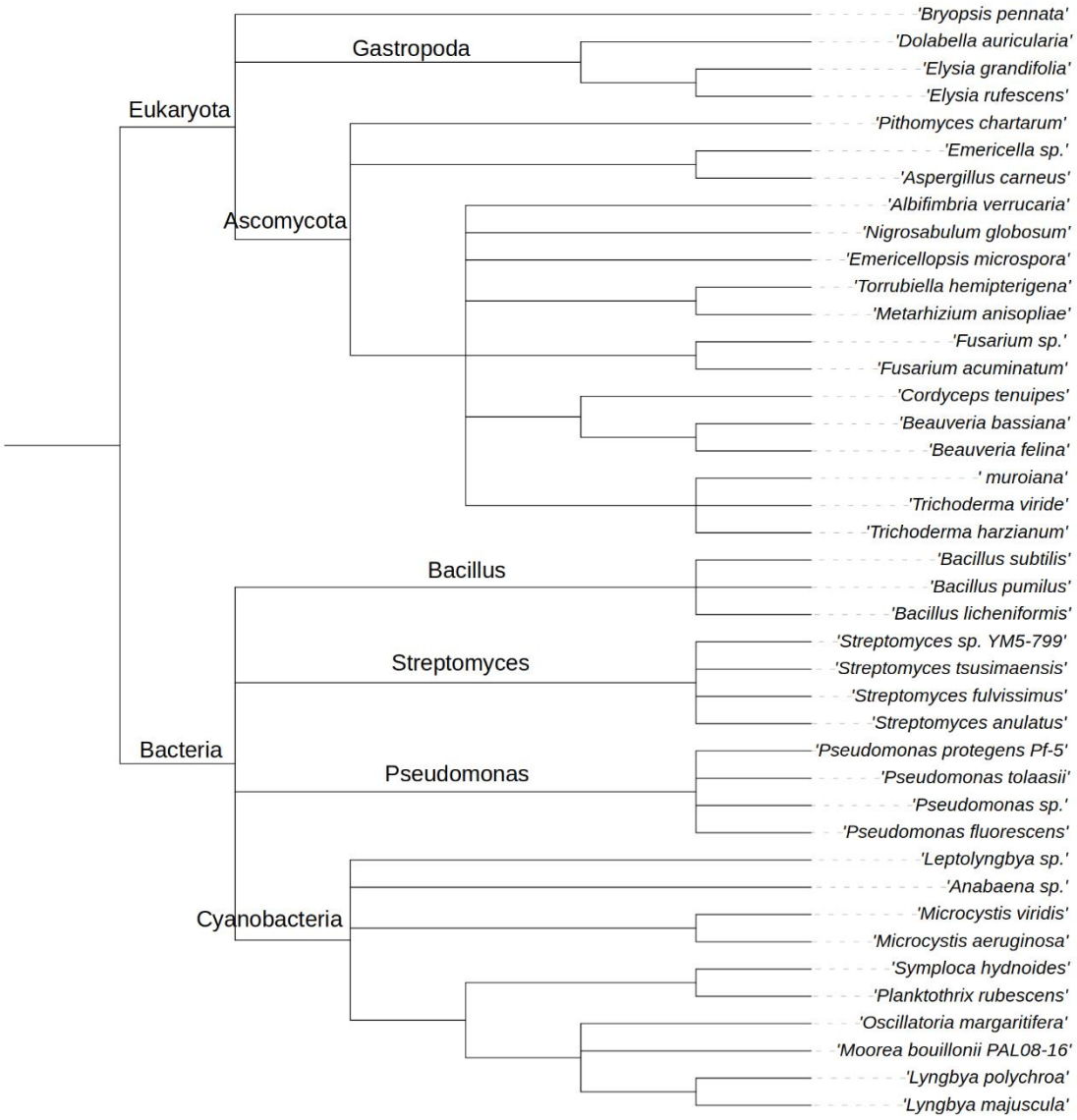


Figure -S10. Phylogenetic tree of the producer organisms from the set of 33 PNP identifications with Norine annotations. Taxonomies not identified by the NCBI taxonomy database[6,7] are not included. Note that there are more than 33 microorganisms because a PNP may have multiple origins.



References

(1) Novák, J.; Lemr, K.; Schug, K. A.; Havlíček, V. CycloBranch: De Novo Sequencing of Nonribosomal Peptides from Accurate Product Ion Mass Spectra. *Journal of The American Society for Mass Spectrometry* **2015**, *26* (10), 1780–1786.

(2) 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* **2017**, *13* (1), 30.

(3) Jegorov, A.; Paizs, B.; Žabka, M.; Kuzma, M.; Havlíček, V.; Giannakopulos, A. E.; Derrick, P. J. Profiling of Cyclic Hexadepsipeptides Roseotoxins Synthesized in Vitro and in Vivo: A Combined Tandem Mass Spectrometry and Quantum Chemical Study. *European Journal of Mass Spectrometry* **2003**, *9* (2), 105–116.

(4) Pavlaskova, K.; Nedved, J.; Kuzma, M.; Zabka, M.; Sulc, M.; Sklenar, J.; Novak, P.; Benada, O.; Kofronova, O.; Hajduch, M.; others. Characterization of Pseudacyclins A- E, a Suite of Cyclic Peptides Produced by Pseudallescheria Boydii. *Journal of natural products* **2010**, *73* (6), 1027–1032.

(5) Wills, R. H.; O'connor, P. B. Structural Characterization of Actinomycin D Using Multiple Ion Isolation and Electron Induced Dissociation. *Journal of The American Society for Mass Spectrometry* **2013**, *25* (2), 186–195.

(6) Sayers, E. W.; Barrett, T.; Benson, D. A.; Bolton, E.; Bryant, S. H.; Canese, K.; Chetvernin, V.; Church, D. M.; DiCuccio, M.; Federhen, S. Database Resources of the National Center for Biotechnology Information. *Nucleic acids research* **2010**, *39* (suppl_1), D38–D51.

(7) Benson, D. A.; Karsch-Mizrachi, I.; Lipman, D. J.; Ostell, J.; Sayers, E. W. GenBank. *Nucleic acids research* **2011**, *39* (Database issue), D32.

4.2 Concluding Remarks

With NRPro, we bring a tool to assist experts and simplify the identification and characterization tasks. Tests have demonstrated that in most cases, NRPro is able to annotate a higher number of experimental peaks than Dereplicator and Cyclobranch. The difference being especially notable when annotating spectra of complex peptides. The identifications of NRPro showed a 93.49% of overlap with the those of Dereplicator, validating its dereplication capacities. Considering these results, the extra features not present in other tools and the highly interactive interface, we expect a good reception by the natural product community.

Chapter 5

Norine

5.1 Overview

Norine is a unique resource exclusively dedicated to NRPs. As reflected in the tight relationship between the tools developed in this thesis and Norine, I have collaborated with the Norine team since the beginning of this project. The work presented in this chapter aimed to update the Norine database while exploiting the functionalities of my tools and introduce them as a part of the Norine project. The update was achieved thanks to a new pipeline for automatically sourcing and enhancing the annotations in Norine. Data from MIBiG, BIRD and StreptomeDB were used to include new entries or modify/complete the existing ones. Quality has always been a priority in Norine and such automation may compromise it. Hence, multiple algorithms are used for filtering and validating the new annotations. Among them, rBAN plays an important role to report inconsistencies between new SMILES and the Norine monomer graphs associated with the same molecule. In order to present our tools as a part of Norine, the main characteristics of rBAN and KFP are exposed in the publication.

Nucleic Acids Research, 2019 **1**
doi: 10.1093/nar/gkz1000

Norine: update of the nonribosomal peptide resource

Areski Flissi[1,*], Emma Ricart[2,3], Clémentine Campart[1], Mickael Chevalier[4], Yoann Dufresne[1], Juraj Michalik[1,5], Philippe Jacques[6], Christophe Flahaut[4], Frédérique Lisacek[2,3,7], Valérie Leclère[4] and Maude Pupin[1,*]

[1]Univ. Lille, CNRS, Centrale Lille, UMR 9189 - CRIStAL - Centre de Recherche en Informatique Signal et Automatique de Lille, F-59000 Lille, France, [2]Proteome Informatics Group, SIB Swiss Institute of Bioinformatics, CMU, Rue Michel-Servet 1, 1211 Geneva, Switzerland, [3]Computer Science Department, University of Geneva, CUI, 7 route de Drize, 1227 Carouge, Switzerland, [4]Univ. Lille, INRA, ISA, Univ. Artois, Univ. Littoral Côte d'Opale, EA 7394-ICV- Institut Charles Viollette, F-59000 Lille, France, [5]bilille, CNRS, cité scientifique, F-59650 Villeneuve d'Ascq, France, [6]TERRA Teaching and Research Centre, Microbial Processes and Interactions, Gembloux Agro-Bio Tech, University of Liège, Avenue de la Faculté d'Agronomie, B5030 Gembloux, Belgium and [7]Section of Biology, University of Geneva, Sciences III, 30 quai Ernest-Ansermet, 1211 Geneva, Switzerland

Received September 15, 2019; Revised October 15, 2019; Editorial Decision October 16, 2019; Accepted October 22, 2019

ABSTRACT

Norine, the unique resource dedicated to nonribosomal peptides (NRPs), is now updated with a new pipeline to automate massive sourcing and enhance annotation. External databases are mined to extract NRPs that are not yet in Norine. To maintain a high data quality, successive filters are applied to automatically validate the NRP annotations and only validated data is inserted in the database. External databases were also used to complete annotations of NRPs already in Norine. Besides, annotation consistency inside Norine and between Norine and external sources have reported annotation errors. Some can be corrected automatically, while others need manual curation. This new approach led to the insertion of 539 new NRPs and the addition or correction of annotations of nearly all Norine entries. Two new tools to analyse the chemical structures of NRPs (rBAN) and to infer a molecular formula from the mass-to-charge ratio of an NRP (Kendrick Formula Predictor) were also integrated. Norine is freely accessible from the following URL: https://bioinfo.cristal.univ-lille.fr/norine/

INTRODUCTION

Norine has been and remains the unique resource dedicated to nonribosomal peptides (NRPs) (1). These secondary metabolites are produced by bacteria and fungi and display a diverse spectrum of biological activity. They are called peptides because they are composed minimally of amino acids connected by peptide bonds and because their length is between two and 26 building blocks. In fact, >500 different building blocks, called monomers, are observed in these peptides, such as derivatives of the proteinogenic amino acids, rare amino acids, fatty acids or carbohydrates. In addition, various types of bonds connect their monomers such as disulfide or phenolic bonds. Some monomers can connect with up to five other monomers, making cycles or branches in the structure of the NRPs. This structural diversity leads to multiple biological functions, which can be further developed to yield pharmaceuticals, biocontrol agents, biocosmetics and bio-cleansing. These molecules have in common that they are synthesized by nonribosomal peptide synthetases. As their name suggests, they are not synthesized following DNA transcription through translation by ribosomes. Nonribosomal peptides synthetases form huge enzymatic complexes that select amino acids or other monomers and connect them with several types of bonds (2).

Norine is a platform dedicated to these compounds. The database stores only natural nonribosomal peptides and it is complemented with analysis tools. In recent years, Norine has been extended and improved to address the needs of distinct scientific communities (mainly biologists and biochemists but also pharmacists among others). In particular, we developed a new pipeline in order to (i) massively add new NRPs and new annotations from external databases and (ii) enhance the quality of these annotations via automatic validation procedures. We also developed new tools processing NRP chemical structures as well as mass spectra.

Furthermore, the web interface has been upgraded to ease the retrieval and the reading of the data. A powerful

*To whom correspondence should be addressed. Tel: +33 3 28 77 85 55; Email: maude.pupin@univ-lille.fr
Correspondence may also be addressed to Areski Flissi. Email: areski.flissi@univ-lille.fr

query builder now allows combining multiple criteria to optimize the database search. A query is built dynamically by the user that can search for any term in any NRP annotation. In the case of annotations with a limited number of possible terms, the term list is displayed and one or several values can be selected. In other cases, an auto-complete feature is provided. Several criteria can be combined using Boolean operators such as AND, OR and AND NOT.

The NRP description page has been refactored. The different categories of annotations are accessible by tabs, allowing more detailed annotation display. For example, the producing organisms are located in a taxonomic tree and a schema of the chemical structure identifies the monomers by different colors generated by in-house tools from the SMILES (Simplified Molecular Input Line Entry Specification) codes.

The pipeline to automate massive sourcing and enhance annotation as well as the new tools are described in the following sections.

AUTOMATIC AND MASSIVE SOURCING

In 2015, Norine was opened to crowd-sourcing through the creation of MyNorine (1) a tool that grants the scientific community access to Norine for submitting new NRPs or suggesting modifications of existing entries. NRP annotation is undoubtedly best achieved by experts. Submissions are manually verified and validated by the Norine team to ensure correctness and quality of data. About 50 contributors registered to MyNorine up to now, and ∼90 new NRPs or modification of annotations have been submitted and validated.

Our next aim was to automatically and massively fill the Norine database with new NRPs or to refine annotations using external resources, without quality loss. To reach this goal, we have developed a new pipeline (see Figure 1 and a detailed description in the supplementary material). Three main databases were targeted for NRP sourcing:

- MIBiG (3) (*Minimum Information about a Biosynthetic Gene Cluster*) is a repository of known biosynthetic gene clusters of secondary metabolites. MIBiG provides community standards for annotations and metadata on biosynthetic gene clusters and their molecular products. A tarball with all entries in raw JSON format is freely available for download.
- BIRD (4) (*Biologically Interesting Molecule Reference Dictionary*) is a resource that provides information about biologically interesting peptide-like antibiotics and inhibitor molecules in the PDB (5) archive. The entire BIRD resource can be downloaded from the wwPDB (World Wide Protein Data Bank) server in CIF format.
- StreptomeDB (6) is a database of molecules produced by bacteria in the *Streptomyces* genus, well known for their prolific production of NRPs. Again, StreptomeDB provides a link to download a file containing all entries.

In the following paragraphs, we describe the pipeline through which data is fetched and selected for insertion in the Norine database. This pipeline is composed of Python scripts that are sequentially executed. For each database, the last version is downloaded and the files are checked before executing the update process. The scripts parse all entries to extract the putative nonribosomal peptides from the external sources. Keywords distinguish the NRPs from the other types of secondary metabolites stored in the external entries. Then, several checks verify if these NRPs are already in Norine or not (see annotation quality enhancement 1 section and supplementary material for more details). Moreover, the *source* field is automatically instantiated with the name of the source database, the *status* field is set to *unreviewed* value and a link to the source entry is recorded. These annotations associated with each newly and automatically added NRP guarantee data traceability.

MIBiG is the first external resource that is parsed. At this point in time, the archive contains ∼1800 entries. Thus, the first script parses all JSON entries and fills the database with the following annotations about the new NRPs: peptide name, family, synonyms, structure type, accession number of the corresponding MIBiG entry and PubChem ID when available, in order to reference MIBiG and PubChem (7) in the corresponding NRP entry of Norine. Finally, the SMILES are fetched for these new Norine entries but not necessarily added. The selection process is detailed in the annotation quality enhancement section 1. As MIBiG is more focused on gene clusters, some NRP structures are incomplete and do not reach Norine quality criteria.

A second script handles entries of BIRD. The CIF files contain information about the chemistry, the biology and structures of ∼1300 molecules. As for MIBiG entries, the script parses these files for finding new NRPs and annotations. The BIRD repository already references 12 Norine entries and, with the script, the cross-links between Norine and BIRD could be extended up to 171 entries. Classic annotation (name, type, synonyms, SMILES, etc.) is completed with formula, molecular weight and monomer composition generated by using an in-house conversion table. Indeed, Norine relies on its own monomer notation which can differ from that of PDB.

The addition of new NRPs from StreptomeDB follows the same process. Then, a dedicated script completes general NRP annotations with chemical activity, producing organisms and references to PubMed IDs (https://www.ncbi.nlm.nih.gov/pubmed/).

Once these scripts are executed, additional annotations are extracted from other sources. When PubMed IDs are fetched, the E-Utilities (*Entrez Programming Utilities*) API gives access to the entire references associated with NRPs (both existing and new ones) such as article title, full journal name and DOI. In the same way, the PubChem IDs give access to the SMILES of the compound. Also, links and chemical annotations from ChEMBL (8), a manually curated database of bioactive molecules with drug-like properties, are retrieved using a REST service (9). Unfortunately, not all SMILES of the three databases could be compared to the SMILES of Norine, because most NRPs entries in Norine did not contain SMILES information. In order to solve this discrepancy, all NRP entries lacking SMILES data were manually searched by name in chemical databases to retrieve the missing information. The execution of the pipeline led to the insertion of 539 new NRPs in Norine. Table 1 shows the contribution of the main external databases

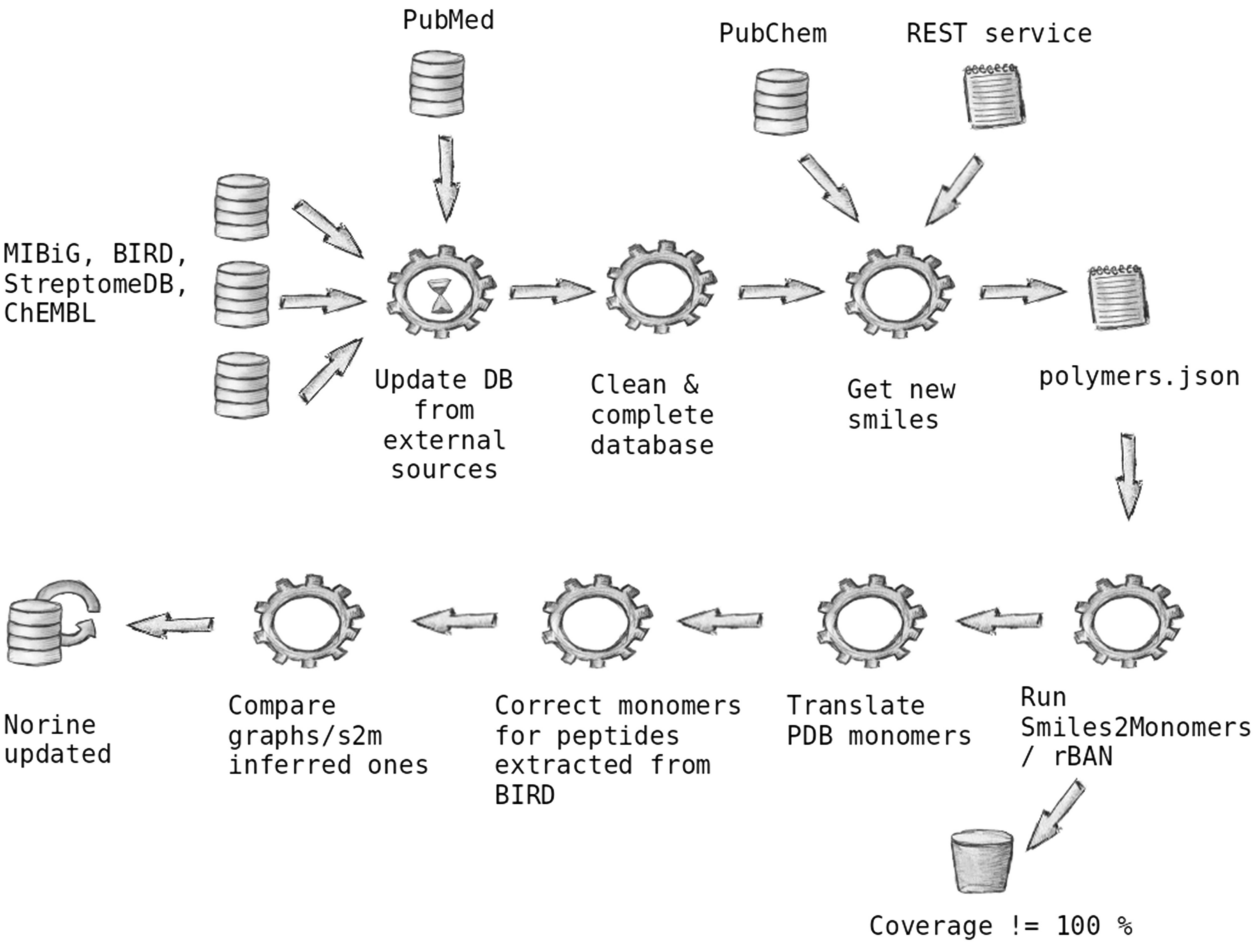


Figure 1. Global view of the update process of the Norine database.

and the 666 external links that have been updated or added. Also, about 527 SMILES and 393 references were updated or added.

ANNOTATIONS QUALITY ENHANCEMENT

The quality of NRP annotation in Norine has been and remains our top priority. Since its creation in 2006 (10), Norine NRP curation has relied on manual extraction from scientific literature as well as meticulous validation prior to insertion in the database. Of course, some errors or incorrect information may occur but these are removed through regular checks. As cited in the first section, the MyNorine tool has been created to boost precision. Nonetheless, automatic massive sourcing from external databases is needed to boost the number of entries, but it also increases the risk of introducing incorrect data. That is the reason why a strict validation pipeline was created. By strict validation, we mean that no entry is added if validation filters fail. Figure 1 shows some validation filters used during the execution process of the pipeline. The following section provides details of some of these filters.

First of all, we obviously verify that any external NRP or particular annotation (*i.e.* synonym, reference, access code, etc.) is not already present in Norine before adding it. If a peptide name is already in Norine, missing annotations for this NRP are added. When a graph (monomer composition of the NRP) is available, the Norine and external graphs are compared. If differences are detected, a failure report is generated. Various reports are created during the process to highlight inconsistencies or errors. These reports are intended to be manually analysed to curate the data. A second verification filter targets the new SMILES that were missing for many NRPs of Norine and that are fetched from external databases. For that purpose, we use *Smiles2Monomers* (`s2m`) (11) and *rBAN* (12), in-house tools that check the consistency of the graph. They infer the monomeric structure of an NRP from its SMILES with two distinct strategies (see next section for description of *rBAN*). Thus, the new SMILES is only added if the inferred graph is the same as the given NRP graph. `s2m` and *rBAN* are used to detect and correct potential errors in the SMILES or monomeric structures recorded in Norine. More than 50 structures were corrected in that way.

Another filter compares all SMILES to check if two NRPs are identical but registered with different names. SMILES are canonized prior to this comparison. Other filters not detailed in this article were developed to enhance

Table 1. New data for Norine

-	MIBiG	BIRD	StreptomeDB	ChEMBL
New NRPs	293	171	75	-
External links (NRPs updated/added)	19/307	- /171	2/80	49/38
Associated references (NRPs updated/added)	20/10	- /1	8/286	65/3

the quality of annotation in order to rebuild the taxonomy tree of the producing organisms, remove duplicates, convert PDB monomers notation, etc.

NEW SOFTWARE

Norine has extended its range of NRP-dedicated software by including two additional tools.

rBAN (12) (*retroBiosynthetic Analysis of Nonribosomal peptides*) infers the monomeric structure of a NRP from a SMILES code. This process can also be qualified as simulating the retro-biosynthesis of NRPs. The first step in *rBAN* is the fragmentation of a molecule that is broken following a set of pattern bonds. Then, the resulting fragments are matched to Norine monomers. The tool can be run in a 'discovery mode' when a monomer cannot be matched to Norine. Then, missing substructure(s) are searched in PubChem so as to suggest potential new monomers. All results are displayed in a directed graph format highlighting the bond types between monomers. *rBAN* addresses some of the limitations of `s2m` cited above. URL: http://bioinfo.cristal.univ-lille.fr/rban

The *Kendrick Formula Predictor* is a tool that uses the Kendrick mass defect (KMD) (13) to predict the chemical formula from the mass-to-charge ratio of a NRP. The required data for the development of the method is extracted from Norine and PubChem. The software was tested with high resolution mass spectrometry data from the surfactin family and the results confirmed the capacity of the tool to successfully predict NRP molecular formulae. Note that this is the first tool in Norine specifically dedicated to the mass spectrometry of NRPs. URL: https://bioinfo.cristal.univ-lille.fr/kendrick-webapp/

Both tools have contributed to the curation/extension of the Norine database. A pipeline combining PubChem and *rBAN* led to the validation of 97.26% of the records in Norine, a two-fold extension of its SMILES and the introduction of 11 new monomers using the discovery mode (12). Kendrick Formula Predictor was used to add missing molecular formulas in Norine increasing the percentage of entries containing this information to 95%. From these formulas, the exact mono-isotopic masses of the peptides were calculated.

CONCLUSION

This paper describes a substantial update of the Norine database. Almost 500 new NRPs and hundreds of annotations for existing Norine NRPs (SMILES, chemical formulas, synonyms, references, external links, etc.) have been added. The quality of annotation was significantly enhanced. To achieve this goal, we developed a pipeline that relies on three main databases that potentially contain NRPs, namely, MIBiG, BIRD and StreptomeDB, and other resources such as PubMed, PubChem or ChEMBL to complement the fetched data. It should be noted that the status of these new NRPs is tagged as *unreviewed*. Our pipeline checks the data before insertion into the database. For example the monomeric composition is inferred from a SMILES using tools such as *rBAN* or `s2m`. If in doubt, a failure report is generated during the execution process of the pipeline to facilitate manual verification by experts. In that sense, the process also enhances the quality of manual annotation that is in turn validated or not by automatic checking. Finally, Norine benefits from two complementary data sources: expert annotations input through the MyNorine tool and automatic annotations produced with the pipeline. In a virtuous circle, data is entered manually by experts, is verified and possibly completed and corrected automatically. Alternatively, data is entered automatically by the pipeline, is verified and possibly completed and corrected manually. The history of all changes for each NRP is kept and easily available from the NRP description page for traceability.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank all scientists that have contributed to Norine by inserting new peptides or new annotations related to an already known NRP.
AF is the main developer of Norine and supervises other developers. E.R. has developed rBAN, the Kendrick Formula Predictor, and parts of the pipeline in relation to peptide structures and also participates in data correction. C.C. participates in the development of Norine. M.C. participates in the design of the Kendrick Formula Predictor, and in data correction. Y.D. and J.M. participated in the development of the pipeline and Y.D. in data correction. P.J., C.F., F.L., V.L. and M.P. supervise Norine database and tool design as well as data curation.

FUNDING

CPER Alibiotech project and the INTERREG V France-Wallonie-Vlaanderen Project SmartBioControl/BioScreen; Institut Français de Bioinformatique [ANR-11-INBS-0013 to J.M.]; SIB Swiss Institute of Bioinformatics Fellowship Programme (to E.R.); Mobility between the CRIStAL and PIG teams is supported by the Germaine de Stael French-Swiss cooperation programme

[39517NH]. Funding for open access charge: University of Lille (SmartBioControl/BioScreen project).
Conflict of interest statement. None declared.

REFERENCES

1. Flissi,A., Dufresne,Y., Michalik,J., Tonon,L., Janot,S., Noé,L., Jacques,P., Leclère,V. and Pupin,M. (2016) Norine, the knowledgebase dedicated to non-ribosomal peptides, is now open to crowdsourcing. *Nucleic Acids Res.*, **44**, D1113–D1118.
2. Süssmuth,R.D. and Mainz,A. (2017) Nonribosomal peptide synthesis-principles and prospects. *Angew. Chem. Int. Ed.*, **56**, 3770–3821.
3. Medema,M.H., Kottmann,R., Yilmaz,P., Cummings,M., Biggins,J.B., Blin,K., de Bruijn,I., Chooi,Y.H., Claesen,J., Coates,R.C. *et al.* (2015) Minimum information about a biosynthetic gene cluster. *Nat. Chem. Biol.*, **11**, 625–631.
4. Young,J.Y., Feng,Z., Dimitropoulos,D., Sala,R., Westbrook,J., Zhuravleva,M., Shao,C., Quesada,M., Peisach,E. and Berman,H.M. (2013) Chemical annotation of small and peptide-like molecules at the protein data Bank. *Database*, **2013**, bat079.
5. Burley,S.K., Berman,H.M., Bhikadiya,C., Bi,C., Chen,L., Costanzo,L.D., Christie,C., Duarte,J.M., Dutta,S., Feng,Z. *et al.* (2019) Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res.*, **47**, D520–D528.
6. Klementz,D., Döring,K., Lucas,X., Telukunta,K.K., Erxleben,A., Deubel,D., Erber,A., Santillana,I., Thomas,O.S., Bechthold,A. *et al.* (2016) StreptomeDB 2.0-an extended resource of natural products produced by streptomycetes. *Nucleic Acids Res.*, **44**, D509–D514.
7. Kim,S., Chen,J., Cheng,T., Gindulyte,A., He,J., He,S., Li,Q., Shoemaker,B.A., Thiessen,P.A., Yu,B. *et al.* (2019) PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res.*, **47**, D1102–D1109.
8. Gaulton,A., Hersey,A., Nowotka,M., Bento,A.P., Chambers,J., Mendez,D., Mutowo,P., Atkinson,F., Bellis,L.J., Cibriàn-Uhalte,E. *et al.* (2017) The ChEMBL database in 2017. *Nucleic Acids Res.*, **45**, D945–D954.
9. Davies,M., Nowotka,M., Papadatos,G., Dedman,N., Gaulton,A., Atkinson,F., Bellis,L. and Overington,J.P. (2015) ChEMBL web services: streamlining access to drug discovery data and utilities. *Nucleic Acids Res.*, **43**, W612–W620.
10. Caboche,S., Pupin,M., Leclère,V., Fontaine,A., Jacques,P. and Kucherov,G. (2008) NORINE: a database of nonribosomal peptides. *Nucleic Acids Res.*, **36**, D326–D331.
11. Dufresne,Y., Noé,L., Leclère,V. and Pupin,M. (2015) Smiles2Monomers: a link between chemical and biological structures for polymers. *J. Cheminform.*, **7**, 62.
12. Ricart,E., Leclère,V., Flissi,A., Mueller,M., Pupin,M. and Lisacek,F. (2019) rBAN: retro-biosynthetic analysis of nonribosomal peptides. *J. Cheminform.*, **11**, 13.
13. Kendrick,E. (1963) A mass scale based on CH2 = 14.0000 for high resolution mass spectrometry of organic compounds. *Anal. Chem.*, **35**, 2146–2154.

5.2 Concluding Remarks

Fourteen years after its first release, Norine is still being maintained and keeps with the quality standards that have always characterized it. While the previous update introduced a crowdsourcing platform named *MyNorine*, the latest extension of the database is performed through a massive sourcing pipeline. Despite the risks of automation, special attention has been given to validate the new data, report inconsistencies and monitor traceability in order to guarantee good quality annotations. The effort has resulted in the introduction of 539 new NRPs in Norine and the addition/correction of annotations from practically all the entries in the resource. Note that at the time of this publication only rBAN and KFP had been integrated in Norine and that is why these are the only tools mentioned in the paper. NRPro was added later on and we expect it to contribute to the quality of the resource as well.

Chapter 6

Discussion

I have described the tools developed throughout the course of this thesis and their individual goals, but the general achievements have not been discussed yet. In the first section of this chapter, I give an overview of the overall accomplishments of the thesis and the problems encountered before reaching the goals. Subsequently, I dedicate the second section on the technical discussion of the computational technologies employed for software development.

6.1 Challenges and Achievements

By the start of this thesis I was new in the field of natural products and my main goal was to construct a MS/MS analysis tool for PNPs. However, when I got further familiarized with the topic I realized that multiple issues needed to be solved and secondary projects emerged, broadening the scope of the project. Here, I present the achievements and the challenges faced during the development of the tools. Figure 6.1 illustrates the contribution of each of the tools in the overall achievements.

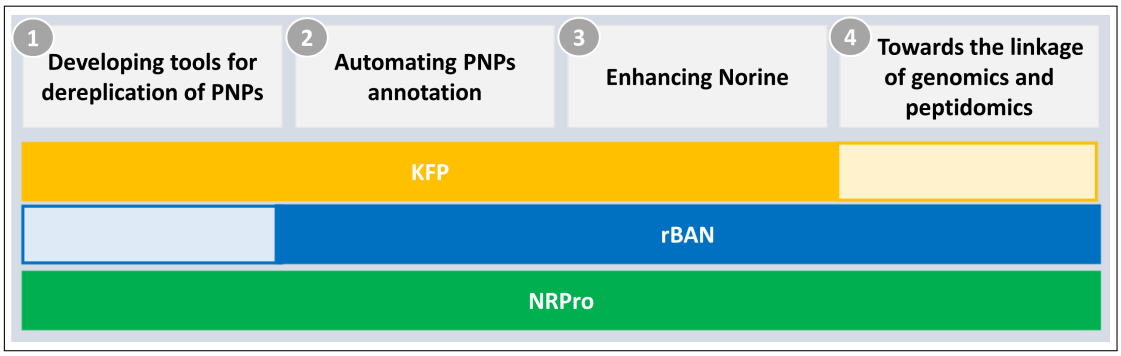


Figure 6.1: Contribution of the tools in the thesis achievements.

6.1.1 Developing tools for dereplication of PNPs

Despite the biological relevance of PNPs, their intricate structures complicate the implementation of software devoted to these compounds. In this thesis I faced several challenges to develop the two dereplication tools presented: NRPro and KFP.

With NRPro, one of the first decisions to be made was the choice of a dereplication method. As PNPs are poorly represented in spectral libraries, dereplication *via* chemical databases was the favoured choice. Our collaboration with the Norine team as well as the high level of curation and peptide nature of the resource influenced the selection of this database. However, the number of PNPs covered by Norine is not high enough for a dereplication tool, so it was extended with ChEBI and NPAtlas. The combination of these three resources resulted in a total of 26882 compounds that enabled the identification of all the tested spectra. Dereplication *via* chemical databases requires the implementation of *in silico* fragmentation. The question here was whether to use a proteomics approach and just calculate the expected masses through monomer-linking cleavages or to target certain functional groups without monomer mapping. The first option would have just required the usage of s2m or rBAN to create breakable monomer-graphs, but the fragmentation would have been underrated due to internal cleavages within the monomers. We did not want to loose the monomer information either. Hence, to preserve as much information as possible, the algorithm of rBAN was extended to perform two sequential searches: i) first, to annotate the monomers, ii) secondly, to create fragment graphs in accordance to the MS/MS expected breakages. This endows NRPro with molecular and monomeric annotation and visualization, that are not present in other dereplication software.

Another question raised during the development of NRPro was the format of the tool. It was essential to provide features that differentiated it from similar software. Depending on the end-goal of MS/MS analytical software, the scoring and statistical significance are more or less elaborated. Software focused on the annotation of MS/MS is commonly used by experts for manual characterization of PNPs. In these cases the knowledge of the expert is the prime concern and the tool is intended to facilitate the analysis. Thus, the scoring is simple or possibly absent. On the other hand, high-throughput software deliver automatic results so much effort was spent on building an optimal scoring system, but not on user assistance. With NRPro we decided to fill the gap between these two options. In terms of scoring, NRPro is not just limited to a simple raw score calculation such as that of Cyclobranch, but it also introduces statistical significance using the classical decoy-target

approach widely accepted in the proteomics. However, we do not pretend to compete with highly elaborated scores. Instead, our intention was to provide the user with means of applying his/her own expertise to verify the correctness of a given PSM. That is why NRPro displays multiple candidate options instead of a single match and why visualization and user friendliness have been prioritized during its design. Another feature that reflects this *hybrid* approach in between an annotation and a dereplication software, are the neutral losses and ion type options. Note that, some of them, are actually not considered in the scoring. However they can be optionally activated for their automatic annotation in the final spectra so that the user can explore these options. With these features we provide a highly interactive and unique resource that was, until now, missing in the natural product community.

A "simpler" alternative for dereplication or at least, for the first steps of dereplication, is the deduction of the chemical formula. This is the main functionality of the KFP tool, which applies the KMD instead of the prediction of isotope patterns typically used in such algorithms. Compared to NRPro, the KFP tool was simpler to develop as it just consisted of applying an already well-known method to the field of PNPs. The software is still in its initial stages and although it is useful for manual interpretation of spectra, I would currently define it as a proof of concept because high-throughput have yet not been performed. Nonetheless, the high performance and simplicity of the method makes it ideal for future big data implementations.

A common problem encountered during the development of both tools was the lack high-throughput MS/MS data of NRPs/PNPs. Initially, this data was supposed to be supplied by one of our collaborators but it was finally never produced. Freely available data for PNPs is also hard to obtain and it is of high importance for building and training bioinformatic models. Fortunately, initiatives such as GNPS have gathered some datasets that can serve this purpose, but compared to fields such as proteomics there is still much work to be done towards the creation of online MS/MS resources dedicated to PNPs.

6.1.2 Automating PNPs annotation

Data collection and curation are the main tasks for database development and maintenance. A lot of time and resources are dedicated to the submission, revision and edition of annotations and metadata. In general, the best annotations result from manual work of data curators. However, this is time consuming and prone to human errors. The usage of bioinformatic tools for automatic annotation may not be the definite solution but it can assist experts and reduce their duties. Despite

the individual goals of each of my tools, they all share the provision of automatic annotations.

Structural annotations of PNPs can be automatically generated using rBAN. To design rBAN I had to familiarise with the biosynthesis reactions of NRPs and, although retro-biosynthesis may seem intuitive, many problems arised during the algorithm implementation. For example, the presence of adjacent cleavages that becomes a combinatorial problem because simultaneous cuts are impossible or the cleavage of bonds close to terminal sides that may leave unannotated atoms. Although monomer-graphs represent the skeleton of the molecule in a simpler form than the molecular structure, the conversion may compromise the structural information. Thus, another challenge was to provide monomer-graph outputs with the same structural details than chemical structures. To do so, directed and labeled edges were included in the graphs. In this way, rBAN graphs provide more information than those obtained with s2m, also designed for annotation purposes, and definitely more than tools just oriented towards retro-biosynthesis such as GRAPE, where the monomer linkages are lost. I also noticed that one of the main limitations in other software was the strong dependency between annotations and the associated monomer database. That is how I thought of a *discovery mode* that allows the detection of new substructures. Currently, mainly two types of structural information are missing in rBAN: i) the distinction between L- and D- amino acid forms and ii) the specification of the linking branch or atom. By fixing these issues we would produce fully-informative monomer graphs, facilitating the conversion of the results into HELM, a language that, as already mentioned in Section 1.1.2, is devoted to the annotation of complex biopolymers in a component-based format. Despite of these potential improvements, the current version of the software already proved to be useful for database curation by validating the 97% of the entries in Norine.

In mass spectrometry, annotation also plays an important role for structure elucidation of PNPs and for the creation of spectral libraries. Collecting annotated MS/MS spectra may enhance the understanding of PNP fragmentation patterns and can lead to the development of statistical models for intensity prediction such as those of NPS [237]. From the tools developed in this thesis, KFP aids on the annotation of MS spectra while NRPro provides automatic MS/MS annotations. Although both tools are helpful for experimental scientists, there is a substantial difference between the outputs from both software. KFP annotations simply consists of the molecular formula of the compounds, while NRPro goes beyond that by suggesting the whole structure of the fragment ions. Furthermore, in NRPro, I put special attention on the design of a concise nomenclature for all type of PNPs and

when combined with the molecular and monomeric visualizations makes interpretation easy for the human eye. As already commented, the integration of rBAN in NRPro was essential to provide the monomer annotations. Monomer annotations are not just advantageous from a visual point of view, but they can be useful to identify patterns such as moieties more prone to fragmentation or monomers with frequent neutral losses. The annotation capacities of NRPro have been demonstrated in its comparison with other tools, where it has shown a higher coverage of annotated peaks.

6.1.3 Enhancing Norine curation and introducing mass spectrometry into the resource

Since its appearance in 2006 Norine has become a reference database for NRPs. It currently stores 1730 NRPs with extensive metadata including the bioactivity, class, structure, source organisms and taxonomy of the peptides, among other information. A strong focus is given to the monomeric annotation of NRPs, providing monomer graphs of the peptide entries which are constructed based on the 544 building blocks included in the database. This highly curated content and the shared scientific goals with the Norine team led to the establishment of a fruitful collaboration that has brought benefits to both sides. From our part, we are aware that Norine attracts the same scientists than our tools, so their integration into the database increased the impact of our software on the natural product community. From their side, this collaboration aided on the curation and extension of Norine. rBAN, for instance, addressed some of the problems of s2m and helped extending Norine, resulting in a two-fold increase of the SMILES in the database as well as the introduction of 11 new monomers. Furthermore, the tool plays an important role in the last Norine update as it is involved in the validation of new SMILES.

In spite of the importance of mass spectrometry for structural characterization of NRPs, this field had yet not been covered in Norine. The introduction of the tools KFP and NRPro enriched the database by providing a new module devoted to mass spectrometry. The development of KFP already brought some changes in the database such as the introduction of monoisotopic masses and some missing molecular formulas. NRPro was recently introduced and its impact is still not measurable. However, this dereplication tool opens many prospects towards the extension of Norine as an MS/MS resource. Taking into consideration the crowd-sourcing nature of Norine, NRPro could include MS/MS submission options in order to create a spectral library associated with the resource. This initiative could even contribute

to the extension of Norine by enabling the submission of MS/MS spectra from new peptides, always followed by manual inspection and potential revision. Although Norine is relatively small, the key point of the database is the manual curation, so it is likely to result in a resource delivering high quality spectral annotations.

6.1.4 Towards the linkage of genomics and peptidomics

As discussed in section 1.2.1, next-generation sequencing and genome mining tools have accelerated the detection BGCs, unveiling the potential for an organism to produce bioactive compounds. However, BGC identification software do not distinguish between clusters of known and unknown compounds, what would serve to optimize the discovery of new natural products. Here is where rBAN could be useful. Combining the results of rBAN with the substrate specificity predictions of a BGC detection tool would improve the identification of known BGCs. It would consists of a similar approach to that employed in the GRAPE/GARLIC pipeline where the monomers from the retro-biosynthesis of GRAPE are aligned against the genomically predicted counterparts given by GARLIC [11]. In the publication of rBAN, our software showed higher coverage and performance than GRAPE, what would result in a faster workflow and could potentially enhance the alignment results. Going one step further, we could take advantage of rBAN being already integrated in NRPro to extend this approach towards peptidogenomics. The dereplication results from NRPro include the monomer annotations of the identified compound so they could simply be aligned against the substrate specificity predictions. This would resemble the approach undertaken by Pep2path [113], although this one is based on mass shift identification in the spectra instead of using peptides *in silico* fragmentation. Note that the ideas discussed in this last section are just prospectives, but they have emerged from to the development of rBAN and NRPro, that opened the way to these new directions.

6.2 Technical discussion

The outcomes of the software developed have been the main focus of discussion so far. However, the technical aspects and technologies chosen for the computational implementation impact important aspects such as the performance, maintenance or visual output of the software. Hence, this section is devoted to the description and evaluation of the strategies employed.

6.2.1 RESTful API

A web application format was chosen for the implementation of the three bioinformatic tools covered in the thesis. The format choice was mainly based on the advantages regarding accessibility: web applications only require a browser and an internet connection, avoiding constraints imposed by computer systems and software installation. The design of the tools follows a classic web architecture consisting of a front-end (client) that includes the HTML, Javascript and CSS code and a back-end (server) developed in Java (Figure 6.2). A key element for such applications is the communication between the two ends. In my tools, the server-client communication is achieved through a **RESTful API**. REST (Representational State Transfer) is a software architecture style that defines standards to be used in order to enhance the data transfer process. Since its release in 1999, REST has become the choice of many web services due to its simplicity and higher performance than SOAP (Simple Object Access Protocol), its predecessor. A RESTful API is characterized for the usage of:

- *Media types*: describing the format of the transmitted resources “image/jpeg”, ”text/html”, “application/json”.
- *URIs*: providing accessibility to those resources (see examples in Figure 6.3).
- *HTTP methods*: defining the action to be performed. The most used are GET, to just obtain resource representation; POST, to create new resources; PUT to update/create resources; and DELETE to remove resources.

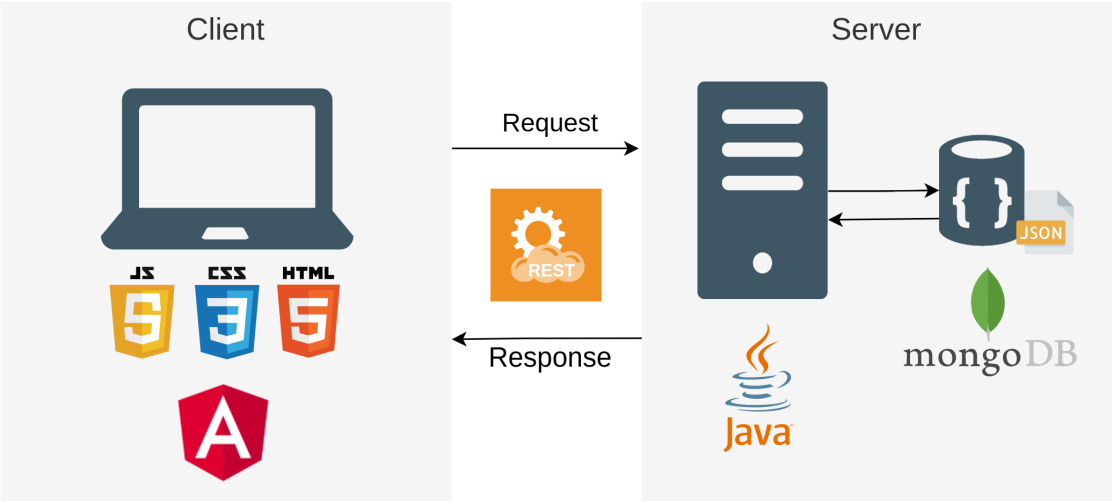


Figure 6.2: Architecture, languages and tools used for the development of the presented software.

In rBAN, for instance, the data introduced by the user (name of the compound, SMILES and Norine graph) is transferred to the server as a JSON object using a POST request. Then, the server applies the business logic required for the conversion of the SMILES into a monomeric graph and returns the output as an *application/json* media type. The front-end uses the returned data to depict the chemical structure as an SVG. Jersey/JAX-RS is used for the creation of the Java RESTful API. Norine also provides REST services for database access. Data from NRPs and monomers can automatically be retrieved through specific URIs (Figure 6.3), which facilitated the coupling with our software.

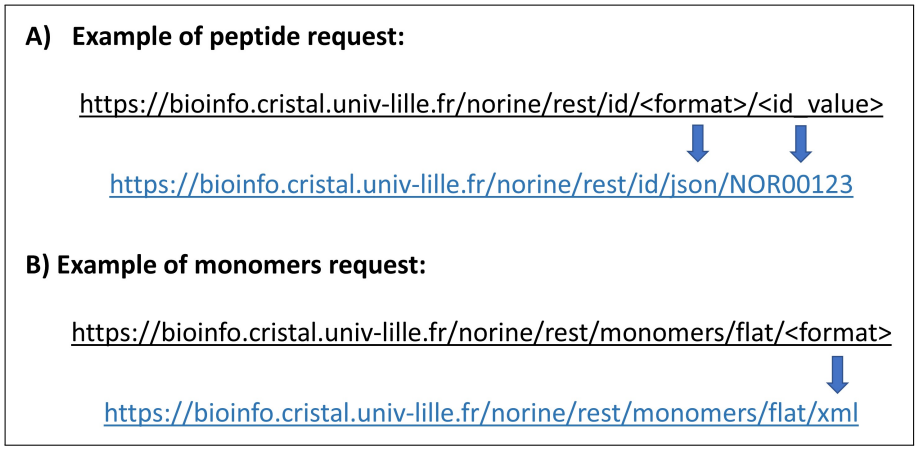


Figure 6.3: Examples of URIs associated to the Norine RESTful API. A) Returns a JSON file with information regarding the peptide with ID: NOR00123. B) Returns a XML file with all the Norine monomer.

One of the main advantages of REST is its scalability, obtained thanks to the stateless nature of the requests. The term *stateless* implies that actions involving the manipulation of the resources (HTTP requests) should be independent of each other and avoid partial updates by always providing the final state of the resource. Furthermore, session-related information (states) should not be stored on the server side. These rules are applied in the software design, which guarantees that in case the number of the resource users increases, the API will easily scale up, as the stateless behaviour will facilitate the introduction of multiple servers and will avoid the need of implementing synchronization tasks.

6.2.2 Back-end development and database construction

As already mentioned, the back-end of the applications is implemented in **Java**. Choosing this programming language was convenient in order to maintain the standards of the PIG group, which adopts this language for most of the bioinformatic

tools developed. Following the same guidelines facilitates the interaction between group members and the future maintenance of the tool. Plus, the simplicity and object-oriented design of Java results in the creation of modular and reusable code that facilitates error solving and refactoring.

Several application servers exist for building Java-based web applications. For the deployment of my tools I used **Apache Tomcat**, which benefits from a light footprint and high flexibility. It is important to mention that, technically, Tomcat is not considered an application server. Java Application servers should provide some infrastructure and functionalities defined in Java EE specifications, that actually Tomcat does not fulfill. Hence, Tomcat should be referred to as a web server or servlet container. Yet not following Java EE stipulations, Tomcat is still widely used because it represents a lightweight option whose features can be extended using third-party dependencies. It is very convenient for simple applications such as those developed in this thesis that do not require many extra features included in Java EE application servers such as Glassfish or WildFly. In this way, one can choose the modules to be included, and adjust them to the needs of the application. For instance, for the implementation of the RESTful web services, I had to insert the Jersey/JAX-RS dependencies.

In the case of NRPro, the back-end also integrates a database resulting from the merge of NPAtlas, Norine and ChEBI. The process of building such resource involves the application of filters, quality checks and the right choice of NP databases. Among all the publicly available databases, our selection was made prioritizing resources providing highly curated data, extensive annotations/metadata and with a strong presence of natural products and particularly PNPs. These requirements are fully met by Norine and, although NPAtlas and ChEBI are not entirely dedicated to peptides, the records were filtered by selecting just those entities containing at least one amide or ester bond. This filtering is performed using rBAN, as the substructure search algorithms required for functional groups detection are already implemented in the tool. Other filters applied involved the removal of charged and/or disconnected compounds. Probably the most important step for the generation of our database was the optimal merging of records from different databases. Note that if this process is not totally effective, the redundancy in the database increases due to the presence of repeated entries. The idea was to merge compounds with identical structures in a single entry with the respective annotations of each database. To implement that process, the structures of the compounds with similar masses were submitted to a graph isomoprohism search from CDK for comparison. That process is computationally expensive but it is performed once only in the pre-processing step limiting the impact on the performance. When multiple en-

tries with the same structure were identified, they were merged into a single object known as *NRProCompound* (Figure 6.4), with the following fields:

- *ID*: custom identification
- *Name*: selected from one of the merged records. Currently, the algorithm prioritizes the name given by Norine, then ChEBI and finally NPAtlas.
- *Synonyms*: the names not used in the previous field.
- *External IDs*: the original IDs. They are conserved in order to maintain the association with the original resources.
- *SMILES, Monoisotopic mass and Formula*
- *Monomeric graph*: computed with rBAN.
- *Size, Coverage and Correctness*: fields given by rBAN related to the monomeric graph: number of monomers (size), the percentage of molecule covered (coverage) and the correctness of the annotations when compared to the respective Norine graphs (correctness).
- *Atomic graph*: computed with rBAN and containing the annotations associated with the monomeric graph.
- *Category, Activities and Structure type*: retrieved from Norine.
- *Organism*: retrieved from Norine and NPAtlas.
- *Origin*: obtained from NPAtlas.
- *Associated spectra*: for those entries with MS/MS spectra.

Note that from the mass spectometry point of view the most relevant annotations are those related to the structural features of the compound, specifically the SMILES, monoisotopic mass, formula and monomeric/atomic graphs. For this reason, to ensure consistency between these fields, they are recalculated from the given SMILES even though some of them are already present in the respective databases. As all the entries are merged automatically and not all of them are fully annotated, a manual check would be interesting to verify the correctness of the merge and fill the empty fields.

In addition to the target *NRProCompound*s, a decoy collection was created for the calculation of the statistical significance. As described in chapter 4, the decoy compounds were created by shuffling the monomeric graphs of the target peptides. Additionally, and contrary to the target database, the decoy collection is not composed of monomeric graphs, but rather of fragmentation graphs. The fragmenta-

Figure 6.4: *NRProCompound* example represented in JSON format.

tion graphs are hierarchical structures of MS/MS fragments emulating the experimental breakages of a mass spectrometry experiment. For the target compounds, these breakages are computed *on-the-fly*, but for the decoy counterparts they are pre-calculated.

The database used to host the described collections was **MongoDB**, a NoSQL document-based DB. One of the main advantages of MongoDB is its JSON-like storage format (Figure 6.4), which turns out more natural for developers working with objects than the traditional row/column approach. The data is organized in collections instead of tables and the collections created for NRPro are the following:

- *Target*: gathers 26882 instances of the *NRProCompound* object.

- *Decoys*: as they consists of fragmentation graphs, the results depend on the selected fragmentation settings. Hence, decoys are divided in multiple collections according to the fragmentation: decoy_byax_H, decoy_byax_Na, decoy_byax_K, decoy_all_H, decoy_all_Na, decoy_all_K. The naming obeys the following pattern: decoy_<ionTypesIncluded>_<adduct>. Thus, collections named "byax" consider b/y and a/x ions, while those described as "all" include z/c ions as well. Each collection assembles 26882 fragmentation trees corresponding to the decoy compounds derived from the original *NRProCompound* structures.

- *Spectra and Requests* : created for the temporal storage of spectra and session-related data of the web application. In MongoDB, this temporality is implemented through the "time to live" (TTL) indexes, which enable the configuration of the collections in order to establish an expiring time before removal. In our case, the time was settled to one week.

Several aspects influenced the choice of a non-relational database instead of the classical SQL option. First of all, in NRPro, links between collections are not necessary, so data relations are not primordial. Secondly, working with relational databases requires defining schemes that strictly fit with the data, reducing the flexibility and not allowing the introduction of unstructured data, as needed in our case. Finally, having in mind possible extensions of the database and a future high-throughput implementation of NRPro, it was interesting to use a technology that can handle big data and horizontal scaling. Compared to the other non-relational options, a document type database such as MongoDB is specially suitable for managing the hierarchical and complex structures with nested objects used in NRPro (e.g., monomeric and fragmentation graphs).

6.2.3 Front-end development

While the back-end side provides all the business logic of our applications, the front-end incorporates the visualization and interactivity needed to reach an optimal user experience. It is quite common in science to invest most of the time and resources to the development of complex analytical algorithms and not on the design of user-friendly interfaces. Although the development of interfaces may seem superfluous, it is a crucial point to warrant the usage of the tools. Especially, and such is our case, when they are destined to experimental scientists, which are rarely familiar with the command-line interface.

As mentioned before, the interfaces developed are based on **HTML 5**, **CSS** and **JavaScript**. The interactivity is given by JavaScript and particularly, the **D3.js** library, which allows the manipulation of documents (DOM) to generate dynamic graphs; for instance, the molecular structure found in rBAN and NRPro (Figure 6.5) consists of a D3 force layout with dragging behavior, enabling the atoms movement. This feature is important for the user who can modify the graphical representation if the automatic depiction is not visually clear. The spectrum viewer of NRPro as well as the Kendrick plot were also implemented using d3.js.

Another relevant aspect of the interface is the layout of the application. Here, the **Bootstrap** framework played an important role, especially for the development of NRPro. The NRPro design originates from a Bootstrap 4 theme that was modified and customized according to our needs. Among other features, Bootstrap enables the creation of responsive web applications thanks to a grid system adaptable to different screen sizes and devices. Figure 6.5 shows the differences between the interface of NRPro in a full-screen mode (1920x1200) and after resizing at 944x1200: observe that the grid changes from two to one column. Additionally, Bootstrap includes reusable components such as *navbars*, *dropdowns* or *buttons* with modern and customizable layouts highly useful for interface design.

While the architecture of Kendrick and rBAN do not require the usage of any Javascript front-end framework, the higher complexity of NRPro urged resorting to **AngularJS** in order to provide a cleaner code with good design patterns. Back to the start of my thesis, I had no experience in web development, which strongly conditioned the choice of AngularJS. One of the reasons to choose this framework, was that I was already familiar with the Model-View-Controller (MVC) design pattern employed in AngularJS. Furthermore, AngularJS is mature and well-documented, with plenty of tutorials and forums online. However, with the release of Angular 2, AngularJS lost support and its latest version is currently under long-term support (LTS) until 2021. Angular 2 is not just a simple upgrade of AngularJS but rather

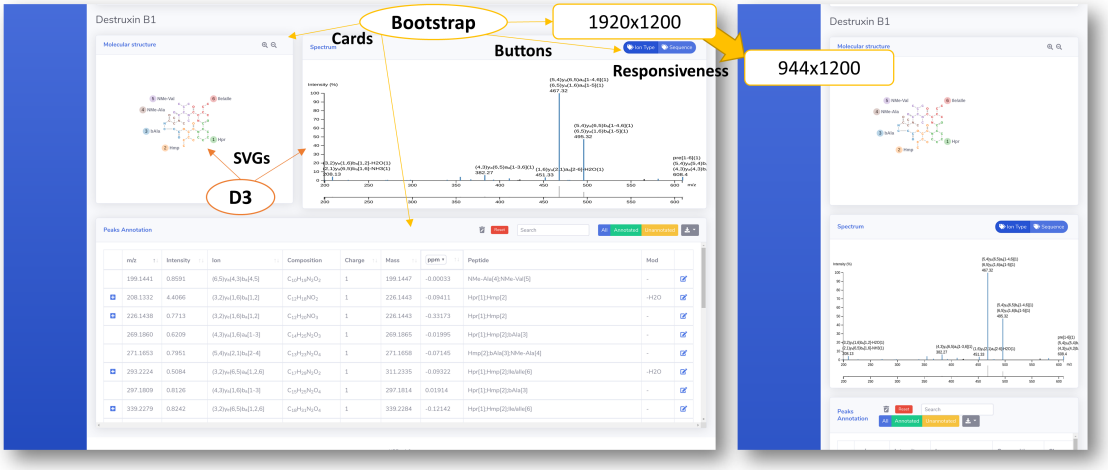


Figure 6.5: Illustration of the responsive design of NRPro. D3 and Bootstrap are used for the creation of the graphs and the layout.

a complete rewrite (see some of the differences in Table 6.1). In Angular 2 and its successive versions (the last one being Angular 9), the MVC model was substituted for components and directives, modularizing the code and increasing its reusability and scalability. Another major change was the usage of Typescript instead of Javacript, providing static typing and enhancing performance. The increased performance is also related to improvement in data-binding and dependency injection. Despite all of these advantages, the migration from AngularJS to Angular 2 is not an easy task and would require a learning period as well as substantial code refactoring. For this reason, AngularJS remains as the framework of NRPro and although it is possible to keep it like that, moving to Angular 2 is recommended for the maintenance and scalability of the code. Alternatively, other component-based frameworks/libraries such as React or Vue could also be considered if we are looking for more light-weight options.

	AngularJS	**Angular 2+ (Angular)**
Architecture	Model-View-Controller	Components and directives
Language	JavaScript	TypeScript
Performance	Lower	Higher
Rendering	Client side only	Server and client side
Tooling support	No	Angular command-line interface (CLI)
Mobile support	No	Yes

Table 6.1: Comparison of AngularJS and Angular 2+.

On the other hand, and despite the advantages of Angular 2, it is important to emphasize that AngularJS has been highly useful for the implementation of some

functionalities. A good example is the routing system used to define the URLs and HTML views of NRPro. As observed in Figure 6.6, the *form view* holds the root path and after submission the page uploads the *dashboard view* and *results table view*. Note that the *dashboard view* is the main layout, but it is never rendered alone. Instead, one of the nested views is always injected inside it. IDs created by the server together with file names and pre-defined strings are used to generate the URLs. The usage of routing is fundamental to enhance the navigation within single-page applications (SPAs) as it keeps the pages history and enables the usage of the "back" button that otherwise would redirect the user outside the application. It also allows sharing links and using bookmarks.

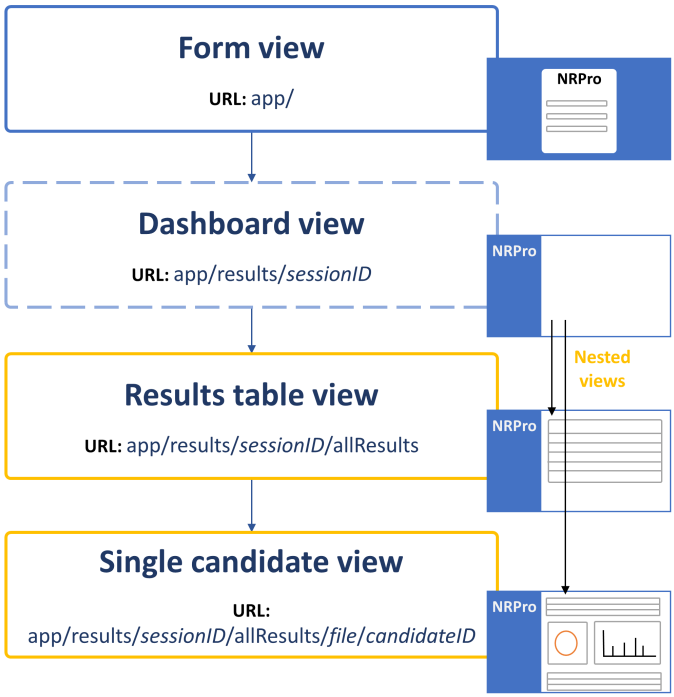


Figure 6.6: Diagram of the routing system implemented in NRPro.

Chapter 7

Conclusions and Outlook

To conclude, the aim of this last chapter is to provide a brief overview of the most relevant characteristics of each tool, what they bring to the scientific community and the prospects that we have for them. Hence, the chapter is divided in three sections dedicated to rBAN, KFP and NRPro.

7.1 rBAN

Among the tools that I developed, rBAN is the most versatile and the one with more potential applications. Just in this thesis, rBAN has been used to complement two other pipelines: the Norine update and NRPro. In the Norine update, rBAN was used as a quality control of the graphs, corroborating its curation capacities that had already been demonstrated when published. In NRPro, the usability of rBAN for mass spectrometry analysis is demonstrated with the creation of fragmentation graphs in the pre-processing step, facilitating the computation of the theoretical fragmentation of NRPro. Note that, originally, the MS/MS fragmentation functionality was not present in rBAN but a slight modification of the cleavage rules allowed the introduction of such module, showing the ease of adapting the tool for other applications. Importantly, the extension of the code did not affect the monomer annotation module, preserving the original functionality of the software. Indeed, as observed in NRPro, the results of the two modules are mergeable in order to create MS/MS fragmentation graphs with monomer annotations. Compared to similar tools rBAN stands out for its high performance, robustness and accuracy. Plus, the feature of the discovery mode reduces the dependency between rBAN and its internal monomer database, a limitation not handled by other software.

We have multiple prospects for rBAN. The tool can process hundreds or even thousands of structures when executed through the command line yet, the web appli-

cation of rBAN is limited to the analysis of a single compound. Increasing the capacity of the online service would help the users with less computational experience to perform larger analyses. Additionally, it would be interesting to enable the choice of the target cleavages. In this way, the users could adapt the rules in accordance to the structural characteristics of their molecules. Another modification that I have in mind is the addition of an alternative way of performing the fragmentation. In the current version, the monomers are annotated together with their modifications. For instance, N-methylated glycine is annotated as a single monomer, NMe-Gly, following the Norine nomenclature. There are no standards for defining whether the monomer and its modifications should be annotated together or separately, so both forms are accepted together with their pros and cons. For this reason, the best solution is to enable both options for the users to choose the optimal for their purposes. Plus, having the split-modification version could enhance the coupling of rBAN with genome mining software. Such an application would be highly beneficial for the discovery of new BGCs and, as already discussed, taking into consideration the link between rBAN and NRPro, the approach could be redirected towards peptidogenomics.

7.2 KFP

KFP in its current state supports the manual interpretation of NRP mass spectra. The geometrical relationship between ion series including vector angles provides a way of exploring further differences in the composition apart from the CH_2 pattern given by the horizontal line. The study conducted with surfactin compounds was used as a proof of concept for the approach, but the method has more potential that still needs to be explored. The tool is currently based on the CH_2 building blocks but NRPs also contain nitrogen and halogenated atoms that could be included in the model as well. Actually, it could be up to the user to choose the desired pattern. Additionally, enabling massive formula prediction would highly increase the value of tool and would not imply large processing times due to the fast performance of the approach. However, before allowing such level of automation, the method should be tested with a dataset large enough to compute some statistics about its reliability. Finally, I would like add more NRPs to the database used for the creation of the KMD plot in order to increase the likelihood of identifying unknown formulas. Such task should be performed carefully and verifying that all the records are truly NRPs, as it is important that the KMD plot solely contains formulas of the target molecules. rBAN could be used as a filter to select the right entries.

7.3 NRPro

NRPro is the most comprehensive tool developed during this thesis, which includes automatic annotations and dereplication features in a single platform. Differently than most dereplication tools, the software considers the user expertise while, at the same time, the chemists can benefit from the computational capacities related to fragmentation and scoring. The front-end offers interactive features to enhance the role of the user and assist on the task of reviewing the results through clear annotations, filters, edition and exporting options. To our knowledge, it is the first tool offering a simultaneous visualization of the monomers and the molecular structure of the candidate compounds. Plus, the *highlighting* feature included in the depiction of the molecule simplifies the association between the matched peaks and the corresponding fragment ions and facilitates the understanding of the annotations. The proposal of multiple candidates instead of a single solution represents another feature that empowers the user and reduces the importance of the scoring by allowing the revision of alternative compounds that the chemist may want to take into consideration as well and that could have been erroneously underscored. The back-end contains all the business logic related to the *in silico* fragmentation of the compounds, which is quite extensive as it considers all types of fragment ions and typical cleavages found in NRPs. In terms of storage, the usage of MongoDB guarantees the scalability of the database, currently gathering records from Norine, NPAtlas and ChEBI. In the publication of NRPro, we show the remarkable performance of the automatic annotations, that matched more peaks than software with strong focus on annotation such as Cyclobranch; and the dereplication capacities of NRPro, comparable to the identifications from software with highly elaborated scoring such as Dereplicator.

Despite these achievements, the current state of NRPro is that of a first version of a software that leaves much room for improvement. The front-end is interactive and user-friendly but, as commented in the discussions, AngularJS may not be the most optimal technology for long-term maintenance. Hence, it would be interesting to move the front-end code to Angular 9 or another component-based framework. One of the main priorities for future versions of the software is the implementation of a batch mode. Allowing the analysis of larger datasets would represent an added-value to the software and would attract big data scientists. Additionally, collecting further MS/MS data would help the enhancement of the scoring, which is currently based on a set of just 356 spectra. Similarly, extending the database with other natural product resources such as NPASS could improve the identifications. Other prospects are related to potential interactions with other software developed in the

thesis. KFP could be used as a filter to discard compounds with unexpected molecular formulas, while rBAN, could bridge the identified compound with the genome. An illustration of this hypothetical pipeline is depicted in Figure 7.1.

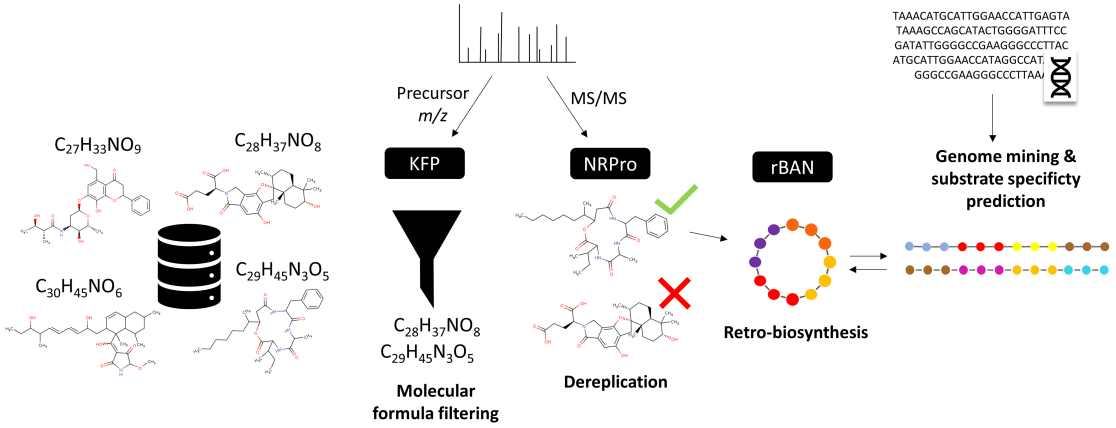


Figure 7.1: Prospective pipeline illustrating the integration of KFP, NRPro and rBAN in a single platform for peptidogenomic analysis.

7.4 Final thoughts

The work described in this thesis involved the development of tools for the study of PNPs that, among other achievements, contributed to Norine extension and curation. rBAN is the most flexible in terms of application, as it is not only useful for retro-biosynthesis, but also for curation, structural annotation and fragmentation. KFP offers a simple and fast alternative to the classical isotope-pattern prediction for formula detection. NRPro differs from most of the pre-existing dereplication software, as it empowers the user with an active role and incorporates many features for annotation and identification in a single platform. Despite having different functionalities, all the presented tools are part of a larger project and could be easily linked together. Indeed, rBAN is already part of NRPro and so could be the KFP tool. This connectivity demonstrates the compactness of the project, that has been all oriented to the analysis of the fragmentation and annotation of PNPs. I expect that the introduction of these tools into Norine does not put an end to the project, but it rather opens a new path towards a larger MS/MS analysis platform. That would not just imply software enhancement and maintenance, but also the extension of functionalities to encompass new areas such as peptidogenomics. The initial steps have been taken to set a clear direction and we have now to keep on moving forward.

Appendices

Appendix A

Supporting Information NRPro

A.1 PNPs identified by NRPro in the GNPS spectra

Table -S4: PNPs identified by NRPro in the GNPS spectra. Only first candidates are shown. Parent and fragment ion tolerance: 0.02.

Spectrum ID	Compound	Structure	Category	Activities	Origin	#Scored/#Annot	P-value
CCMSLIB00000577522	Acremostatin B					34/37	1.0E-16
CCMSLIB00000577573	Leucinostatin F					20/22	1.0E-16
CCMSLIB00000577587	Lyngbyastatin 3				Bacterium	24/24	1.0E-16
CCMSLIB00000577616	SNA-60-367-14				Bacterium	10/14	1.0E-16
CCMSLIB00000577637	Cephaibol-C					17/22	1.0E-16
CCMSLIB00000577675	Almiramide C				Bacterium	8/8	1.0E-16
CCMSLIB00000577677	Pepstatin B				Bacterium	15/15	1.0E-16
CCMSLIB00000577680	WS-7338-B					9/9	1.0E-16
CCMSLIB00000577716	Leucinostatin S					8/10	1.0E-16
CCMSLIB00000577726	Efrapeptin J					6/8	1.0E-16
CCMSLIB00000577748	Heptaibin	linear	peptaibol	antibiotic	Fungus	10/19	1.0E-16
CCMSLIB00000577761	Trichovirin I-7A					7/8	1.0E-16
CCMSLIB00000577777	Apramide B	linear	lipopeptide	unknown	Bacterium	7/7	1.0E-16
CCMSLIB00000577821	A-1437-M					11/15	1.0E-16
CCMSLIB00000577660	Bacillomycin D2					10/10	1.11E-16
CCMSLIB00000577755	Veraguamide J				Bacterium	12/12	1.11E-16
CCMSLIB00000577623	Trichokonin-VII				Fungus	23/33	3.33E-16
CCMSLIB00000577506	Leucinostatin D					22/23	4.44E-16
CCMSLIB00000577704	JBIR-114					14/16	6.661E-16
CCMSLIB00000577789	Efrapeptin-C					9/13	2.442E-15
CCMSLIB00000577597	Cordyheptapeptide A				Fungus	17/17	2.775E-15
CCMSLIB00000577556	Acremostatin A					48/50	3.108E-15
CCMSLIB00000577590	PHB					12/12	8.104E-15
CCMSLIB00000577666	Trichokindin-IIa				Fungus	18/33	1.088E-14
CCMSLIB00000577766	Carmabin B	linear	lipopeptide	surfactant	Bacterium	10/10	1.11E-14
CCMSLIB00000577575	Trikoningin-KB-I					31/31	2.875E-14
CCMSLIB00000577701	WS-7338-D					12/12	4.962E-14
CCMSLIB00000577512	SNA-60-367-19				Bacterium	28/30	7.194E-14
CCMSLIB00000577580	OHB					12/12	7.36E-14
CCMSLIB00000577689	SNA-60-367-21					11/14	1.001E-13

Spectrum ID	Compound	Structure	Category	Activities	Origin	#Scored/#Annot	P-value
CCMSLIB00000577717	Almiramide A				Bacterium	9/10	1.495E-13
CCMSLIB00000577714	Dibenarthin				Bacterium	11/11	4.541E-13
CCMSLIB00000577549	Orfamide C	partial cyclic	lipopeptide	antibiotic,surfactant	Bacterium	33/36	1.279E-12
CCMSLIB00000577723	Cotteslosin B				Fungus	16/16	2.002E-12
CCMSLIB00000577676	Trichokindin-VI					10/12	1.18E-11
CCMSLIB00000577678	Apramide A	linear	lipopeptide	unknown	Bacterium	15/15	2.172E-11
CCMSLIB00000577524	Efrapeptin-F					23/27	2.251E-11
CCMSLIB00000577643	Pepstatin BU					14/14	3.71E-11
CCMSLIB00000577772	RHM3					5/5	4.258E-11
CCMSLIB00000577535	Leucinostatin B2					29/29	6.399E-11
CCMSLIB00000577705	Almiramide B				Bacterium	10/10	1.102E-10
CCMSLIB00000577583	Trichokonin-IX				Fungus	19/25	1.742E-10
CCMSLIB00000577547	Aspergillicin B				Fungus	36/39	2.181E-10
CCMSLIB00000577491	Hormothamnin A					78/90	2.816E-10
CCMSLIB00000577733	Kahalalide G	linear	peptide	unknown		8/9	4.814E-10
CCMSLIB00000577730	RHM2				Fungus	7/9	8.142E-10
CCMSLIB00000577609	Trichokindin-IVa					32/41	9.608E-10
CCMSLIB00000577615	Cycloaspeptide A				Fungus	31/39	1.008E-9
CCMSLIB00000577617	Grividomycin-III					12/13	1.081E-9
CCMSLIB00000577588	Aspergillicin E				Fungus	20/20	1.145E-9
CCMSLIB00000577624	Pumilacidin B				Bacterium	13/13	1.217E-9
CCMSLIB00000577532	Leucinostatin A2					29/29	1.33E-9
CCMSLIB00000577521	Leucinostatin K					39/39	2.399E-9
CCMSLIB00000577769	Trichorovin-IIb				Fungus	8/9	3.051E-9
CCMSLIB00000577671	Veraguamide E				Bacterium	19/20	3.405E-9
CCMSLIB00000577790	Dragonamide E				Bacterium	9/9	4.013E-9
CCMSLIB00000577734	Beauvericin G2					16/16	5.697E-9
CCMSLIB00000577792	Apramide C				Bacterium	9/10	6.083E-9
CCMSLIB00000577851	Amonabactin-P-750					6/6	8.006E-9
CCMSLIB00000577788	SNA-60-367-23				Bacterium	6/7	9.324E-9

Spectrum ID	Compound	Structure	Category	Activities	Origin	#Scored/#Annot	P-value
CCMSLIB00000577797	Destruxin D					10/11	9.834E-9
CCMSLIB00000577636	WS-9320-A					20/23	1.023E-8
CCMSLIB00000577692	Ilamycin B1					8/8	1.718E-8
CCMSLIB00000577720	RHM1				Fungus	7/7	1.883E-8
CCMSLIB00000577640	[Val7]-Surfactin C13ai					10/10	1.895E-8
CCMSLIB00000577760	Oxachelin				Bacterium	14/15	1.9E-8
CCMSLIB00000577724	Tyrocidine-B					13/13	1.941E-8
CCMSLIB00000577787	[Val7]-Surfactin C14i monomethyl ester					21/26	2.063E-8
CCMSLIB00000577653	Trichotoxin-A					50/75	2.472E-8
CCMSLIB00000577518	Leucinostatin R					25/26	2.505E-8
CCMSLIB00000577725	Fungisporin					10/10	2.589E-8
CCMSLIB00000577674	Veraguamide C				Bacterium	12/13	2.725E-8
CCMSLIB00000577731	Veraguamide K				Bacterium	18/18	3.021E-8
CCMSLIB00000577740	Pitipeptolide B				Bacterium	17/18	3.485E-8
CCMSLIB00000577641	Aselacin-C					21/34	3.633E-8
CCMSLIB00000577513	Emerimicin IV	linear	peptaibol	antibiotic	Fungus	67/73	4.035E-8
CCMSLIB00000577764	Veraguamide H					11/11	4.326E-8
CCMSLIB00000577654	Trichovirin II 5				Fungus	66/92	6.063E-8
CCMSLIB00000577850	Harzianin HC XIII	linear	peptaibol	antibiotic		8/9	6.842E-8
CCMSLIB00000577679	Integramide B					22/25	6.983E-8
CCMSLIB00000577763	W-493-A					21/22	7.472E-8
CCMSLIB00000577610	"APD I component a					19/19	8.643E-8
CCMSLIB00000577795	Cocosamide A				Bacterium	11/14	8.794E-8
CCMSLIB00000577595	WS-7338-A					26/28	9.121E-8
CCMSLIB00000577836	WIN-66306					8/8	9.236E-8
CCMSLIB00000577713	Destruxin D1					21/24	1.111E-7
CCMSLIB00000577708	Scopularide A					11/11	1.237E-7
CCMSLIB00000577659	A-Substance Ib					15/15	1.526E-7
CCMSLIB00000577585	Carriebowmide				Bacterium	24/26	1.809E-7
CCMSLIB00000577559	Massetolide G	partial cyclic	lipopeptide	antibiotic,surfactant	Bacterium	35/37	1.852E-7

Spectrum ID	Compound	Structure	Category	Activities	Origin	#Scored/#Annot	P-value
CCMSLIB00000577696	Pepstatin A					20/20	1.873E-7
CCMSLIB00000577571	Leucinostatin V					24/26	1.96E-7
CCMSLIB00000577762	Brunsvicamide C				Bacterium	11/11	2.016E-7
CCMSLIB00000577584	Massetolide C	partial cyclic	lipopeptide	antibiotic,surfactant	Bacterium	17/17	2.101E-7
CCMSLIB00000577538	Cotteslosin A				Fungus	31/31	2.204E-7
CCMSLIB00000577796	Lyngbyastatin 1				Bacterium	17/17	2.411E-7
CCMSLIB00000577516	Majusculamide C	cyclic	peptide	antibiotic,antitumor	Bacterium	43/47	2.805E-7
CCMSLIB00000577830	Versicoloritide A				Fungus	7/8	2.826E-7
CCMSLIB00000577861	Kulomo'opunalide-1					12/12	3.365E-7
CCMSLIB00000577668	Veraguamide G				Bacterium	23/26	3.459E-7
CCMSLIB00000577599	JBIR-113					24/25	3.477E-7
CCMSLIB00000577672	Cocosamide B				Bacterium	16/16	3.602E-7
CCMSLIB00000577567	Bacillopeptin-B					57/78	3.62E-7
CCMSLIB00000577840	Massetolide H	partial cyclic	lipopeptide	antibiotic,surfactant	Bacterium	64/76	3.74E-7
CCMSLIB00000577614	Anthranicine					18/18	3.768E-7
CCMSLIB00000577833	Aspergillicin C				Fungus	7/10	4.174E-7
CCMSLIB00000577601	None					31/31	4.291E-7
CCMSLIB00000577831	E-Dehydroapratoxin A					8/8	4.307E-7
CCMSLIB00000577591	Orfamide B	partial cyclic	lipopeptide	antibiotic,surfactant	Bacterium	15/16	4.364E-7
CCMSLIB00000577838	Brunsvicamide A				Bacterium	13/13	4.378E-7
CCMSLIB00000577626	Destruxin E1	cyclic	peptide	toxin		32/36	4.414E-7
CCMSLIB00000577684	Veraguamide A				Bacterium	18/18	4.423E-7
CCMSLIB00000577773	Wewakazole					16/16	4.618E-7
CCMSLIB00000577744	Dolastatin 15	linear	peptide	antitumor		7/7	4.783E-7
CCMSLIB00000577854	SW-163B				Bacterium	12/16	5.361E-7
CCMSLIB00000577816	Dragomabin				Bacterium	13/13	5.582E-7
CCMSLIB00000577780	Majusculamide D					14/17	6.133E-7
CCMSLIB00000577551	Hypomurocin B-1				Fungus	98/117	6.19E-7
CCMSLIB00000577776	Cyclomarin A				Bacterium	6/6	6.436E-7
CCMSLIB00000577818	Glumamycin					10/13	6.589E-7

Spectrum ID	Compound	Structure	Category	Activities	Origin	#Scored/#Annot	P-value
CCMSLIB00000577742	Dragonamide					9/9	7.098E-7
CCMSLIB00000577855	Harzianin HC-15					7/7	7.301E-7
CCMSLIB00000577499	SNA-60-367-17				Bacterium	108/119	8.15E-7
CCMSLIB00000577581	[Val7]-Surfactin C15ai dimethyl ester					30/35	8.47E-7
CCMSLIB00000577525	Bacillomycin D2					46/55	9.509E-7
CCMSLIB00000577537	Ilamycin B2					17/18	1.038E-6
CCMSLIB00000577864	Massetolide E	partial cyclic	lipopeptide	antibiotic,surfactant	Bacterium	6/7	1.124E-6
CCMSLIB00000577605	Lichenysin-G1a					11/11	1.301E-6
CCMSLIB00000577799	Cyclosporin-Z					9/9	1.306E-6
CCMSLIB00000577693	Emericellamide B				Fungus	17/18	1.462E-6
CCMSLIB00000577545	Lipopeptide NO				Bacterium	28/28	1.553E-6
CCMSLIB00000577634	Veraguamide I					24/26	1.561E-6
CCMSLIB00000577625	Fijimycin A					11/11	2.214E-6
CCMSLIB00000577606	Viscosinamide					32/34	2.218E-6
CCMSLIB00000577602	Trichosporin-Bs-e					23/36	2.273E-6
CCMSLIB00000577849	Dragonamide B				Bacterium	8/8	2.329E-6
CCMSLIB00000577646	Dolastatin 10	linear	peptide	antitumor		14/15	2.384E-6
CCMSLIB00000577569	Pumilacidin E					29/30	2.449E-6
CCMSLIB00000577541	Trikoningin-KB-II					56/58	2.945E-6
CCMSLIB00000577779	Tiahuramide A				Bacterium	27/30	2.99E-6
CCMSLIB00000577721	Pleofungin A				Fungus	19/22	3.098E-6
CCMSLIB00000577858	Protodestruxin					13/15	3.309E-6
CCMSLIB00000577562	Pumilacidin C					26/26	3.393E-6
CCMSLIB00000577743	Emericellamide D					21/21	3.718E-6
CCMSLIB00000577656	Beauvericin D	cyclic	peptide	antibiotic	Fungus	20/20	3.9E-6
CCMSLIB00000577655	Isariin E					22/22	3.952E-6
CCMSLIB00000577598	Isariin G1					26/26	4.07E-6
CCMSLIB00000577533	Efrapeptin H					38/46	4.107E-6
CCMSLIB00000577613	Neoatroviridin A				Fungus	45/66	4.215E-6
CCMSLIB00000577843	Tyrocidin B1					13/14	4.337E-6

Spectrum ID	Compound	Structure	Category	Activities	Origin	#Scored/#Annot	P-value
CCMSLIB00000577604	[§-Me-Pro] destruxin E chlorohydrin					25/29	4.426E-6
CCMSLIB00000577635	Isariin	cyclic	peptide	antibiotic		28/31	4.435E-6
CCMSLIB00000577592	Destruxin-Ed(1)					19/19	4.557E-6
CCMSLIB00000577673	Brunsvicamide B				Bacterium	16/16	4.748E-6
CCMSLIB00000577629	Sporidesmolide-I					32/33	5.063E-6
CCMSLIB00000577791	Kulomo'opunalide-2					24/24	5.081E-6
CCMSLIB00000577747	[Phe3 N-MeVal5]-Destruxin B					18/19	5.405E-6
CCMSLIB00000577560	Trichorzin HA-3				Fungus	95/123	5.427E-6
CCMSLIB00000577589	N-methylsansalvamide				Fungus	28/29	5.586E-6
CCMSLIB00000577526	Bacillomycin D4					22/23	6.001E-6
CCMSLIB00000577775	Carmabin A	linear	lipopeptide	surfactant	Bacterium	12/12	6.168E-6
CCMSLIB00000577737	Enniatin M1				Bacterium	10/10	6.437E-6
CCMSLIB00000577639	Etamycin VI-2				Bacterium	12/12	6.961E-6
CCMSLIB00000577579	Pitiprolamide				Bacterium	47/52	7.973E-6
CCMSLIB00000577645	Pumilacidin D					34/35	8.019E-6
CCMSLIB00000577669	Sansalvamide				Fungus	19/20	8.346E-6
CCMSLIB00000577554	Aspergillicin A				Fungus	47/57	8.692E-6
CCMSLIB00000577670	Cyclosporin-R					12/12	9.32E-6
CCMSLIB00000577741	Tumescenamide B				Bacterium	19/20	9.908E-6
CCMSLIB00000577794	Symplostatin 3				Bacterium	12/14	1.197E-5
CCMSLIB00000577750	Tiahuramide C				Bacterium	20/22	1.263E-5
CCMSLIB00000577687	Pumilacidin A					17/18	1.294E-5
CCMSLIB00000577546	Massetolide A	partial cyclic	lipopeptide	antibiotic,surfactant	Bacterium	43/43	1.535E-5
CCMSLIB00000577869	Trichorovin-IIb				Fungus	13/13	1.552E-5
CCMSLIB00000577815	Planktocyclin				Bacterium	14/18	1.612E-5
CCMSLIB00000577860	Desmethyldestruxin-A					11/12	1.754E-5
CCMSLIB00000577715	Radamicin					23/24	2.02E-5
CCMSLIB00000577706	Dolastatin D					24/25	2.141E-5
CCMSLIB00000577517	Trichogin-A-IV					61/67	2.164E-5
CCMSLIB00000577756	Plipastatin A2					25/33	2.233E-5

Spectrum ID	Compound	Structure	Category	Activities	Origin	#Scored/#Annot	P-value
CCMSLIB00000577572	Dolastatin 14					24/24	2.396E-5
CCMSLIB00000577745	Emericellamide A				Fungus	8/8	2.414E-5
CCMSLIB00000577823	Tiahuramide B				Bacterium	21/23	2.768E-5
CCMSLIB00000577719	Microcystin-FR					16/18	2.776E-5
CCMSLIB00000577814	Beauvericin J				Fungus	13/13	3.209E-5
CCMSLIB00000577642	Sporidesmolide II					33/35	3.34E-5
CCMSLIB00000577576	Scopularide A					25/25	3.405E-5
CCMSLIB00000577709	Sporidesmolide V				Fungus	20/20	3.439E-5
CCMSLIB00000577542	Suzukacillin-B					63/108	3.557E-5
CCMSLIB00000577638	[Ala(2) Val(11)]Cyclosporin					15/15	3.623E-5
CCMSLIB00000577630	Destruxin B1	cyclic	peptide	toxin		29/32	3.743E-5
CCMSLIB00000577622	Pitipeptolide E				Bacterium	41/46	3.934E-5
CCMSLIB00000577857	Trichorovin-Xa				Fungus	5/5	3.936E-5
CCMSLIB00000577586	Mojavenisn A					19/19	3.968E-5
CCMSLIB00000577555	Massetolide F	partial cyclic	lipopeptide	antibiotic,surfactant	Bacterium	65/68	4.136E-5
CCMSLIB00000577651	Isariin D	cyclic	peptide	antibiotic	Fungus	21/22	4.285E-5
CCMSLIB00000577859	Trichodermide E				Fungus	9/10	4.491E-5
CCMSLIB00000577536	[Val7]-Surfactin C14i dimethyl ester					54/62	4.5E-5
CCMSLIB00000577729	Heptaibin	linear	peptaibol	antibiotic	Fungus	9/10	5.196E-5
CCMSLIB00000577826	Sporidesmolide-III				Fungus	25/26	5.309E-5
CCMSLIB00000577702	LP-237-F7					51/62	5.879E-5
CCMSLIB00000577824	Enniatin-B3				Fungus	12/12	6.02E-5
CCMSLIB00000577697	Pepstatin AC					13/13	6.237E-5
CCMSLIB00000577801	Trichorovin-IXa				Fungus	16/19	6.814E-5
CCMSLIB00000577837	Apratoxin G				Bacterium	8/8	6.863E-5
CCMSLIB00000577804	Harzianin HC-13					12/13	7.002E-5
CCMSLIB00000577612	Palmyramide A				Bacterium	24/25	7.335E-5
CCMSLIB00000577498	Orfamide A	partial cyclic	lipopeptide	antibiotic,surfactant	Bacterium	63/71	7.55E-5
CCMSLIB00000577758	Montanastatin	cyclic	peptide	toxin,antitumor		13/13	8.055E-5
CCMSLIB00000577565	[Val7]-Surfactin C14i monomethyl ester					46/51	8.115E-5

Spectrum ID	Compound	Structure	Category	Activities	Origin	#Scored/#Annot	P-value
CCMSLIB00000577633	Destruxin A1	cyclic	peptide	toxin		23/29	8.471E-5
CCMSLIB00000577802	Destruxin-E diol					11/11	8.516E-5
CCMSLIB00000577694	Hypomurocin B-2				Fungus	85/134	9.006E-5
CCMSLIB00000577594	Cyclosporin-V					17/17	9.24E-5
CCMSLIB00000577514	Lichenysin-G5a				Bacterium	27/28	9.336E-5
CCMSLIB00000577736	Peptaivirin A					31/47	9.716E-5
CCMSLIB00000577600	Hypomurocin B-2				Fungus	100/128	9.895E-5
CCMSLIB00000577509	Ibu-epidemethoxylyngbystatin 3					68/86	1.004E-4
CCMSLIB00000577718	Trichorovin-IIIb					14/17	1.012E-4
CCMSLIB00000577539	Lichenysin-G3					21/21	1.032E-4
CCMSLIB00000577505	Leucinostatin A					64/74	1.049E-4
CCMSLIB00000577822	[D-Asp3 ADMAdda5]microcystine-LR					40/53	1.074E-4
CCMSLIB00000577853	Preneocarzinostatin					30/39	1.127E-4
CCMSLIB00000577497	Desmethoxymajusculamide C				Bacterium	74/84	1.133E-4
CCMSLIB00000577819	MK 1688					16/16	1.287E-4
CCMSLIB00000577543	Lichenysin A					26/26	1.415E-4
CCMSLIB00000577862	Beauvericin	cyclic	peptide	antibiotic	Fungus	16/18	1.439E-4
CCMSLIB00000577578	Trichovirin II 2a				Fungus	102/138	1.45E-4
CCMSLIB00000577688	Beauvericin-B					18/18	1.507E-4
CCMSLIB00000577663	Isariin C2					28/29	1.573E-4
CCMSLIB00000577749	Enniatin B1	cyclic	peptide	antibiotic,toxin	Fungus	10/10	1.583E-4
CCMSLIB00000577611	Microcystin LR	cyclic	PK-NRP	toxin		35/45	1.708E-4
CCMSLIB00000577785	Carriebowmide sulfon					13/13	1.802E-4
CCMSLIB00000577722	Nordolastatin G					12/13	1.847E-4
CCMSLIB00000577520	Halobacillin				Bacterium	46/50	1.93E-4
CCMSLIB00000577493	Plipastatin A1					83/88	2.056E-4
CCMSLIB00000577608	[Phe3 N-MeVal5]-Destruxin B					35/41	2.059E-4
CCMSLIB00000577658	Kurstakin 4				Bacterium	18/18	2.089E-4
CCMSLIB00000577700	Hypomurocin A-2				Fungus	28/34	2.129E-4
CCMSLIB00000577519	[Val7]-Surfactin C15ai dimethyl ester					34/36	2.138E-4

Spectrum ID	Compound	Structure	Category	Activities	Origin	#Scored/#Annot	P-value
CCMSLIB00000577664	Plipastatin B1					17/18	2.175E-4
CCMSLIB00000577511	[Val7]-Surfactin C15ai					39/41	2.224E-4
CCMSLIB00000577534	Esperin					72/85	2.39E-4
CCMSLIB00000577681	Beauvericin-A					20/20	2.499E-4
CCMSLIB00000577627	[Val7]-Surfactin C15ai dimethyl ester					38/52	2.546E-4
CCMSLIB00000577566	[Val7]-Surfactin C14i dimethyl ester					28/33	2.619E-4
CCMSLIB00000577557	Bacitracin F				Bacterium	69/98	2.632E-4
CCMSLIB00000577619	Trichorozin-II				Fungus	30/31	2.686E-4
CCMSLIB00000577767	Coibamide A				Bacterium	14/15	2.784E-4
CCMSLIB00000577757	Veraguamide D				Bacterium	15/16	2.894E-4
CCMSLIB00000577727	Leucinostatin H					6/6	2.912E-4
CCMSLIB00000577574	[Leu7]-Surfactin C14i monomethyl ester					25/27	3.018E-4
CCMSLIB00000577561	Peptidolipin NA					20/20	3.132E-4
CCMSLIB00000577710	Trichoaureocin 3					14/31	3.212E-4
CCMSLIB00000577503	Surfactin-D					31/31	3.229E-4
CCMSLIB00000577631	Trichodermide D				Fungus	32/32	3.238E-4
CCMSLIB00000577492	Dolastatin 12	cyclic	peptide	antitumor		98/113	3.28E-4
CCMSLIB00000577746	Valinomycin	cyclic	peptide	antibiotic	Bacterium	29/29	3.371E-4
CCMSLIB00000577808	Phakellistatin 8	cyclic	peptide	antitumor		13/17	3.382E-4
CCMSLIB00000577563	A C0					263/452	3.39E-4
CCMSLIB00000577620	Enniatin-B2				Fungus	15/15	3.459E-4
CCMSLIB00000577527	Acremostatin C					18/20	3.586E-4
CCMSLIB00000577793	Pseudodesmin B				Bacterium	7/7	3.594E-4
CCMSLIB00000577650	Plipastatin B2					38/44	3.849E-4
CCMSLIB00000577661	SNA-60-367-4				Bacterium	44/49	3.953E-4
CCMSLIB00000577657	Psuedodestruxin C				Fungus	26/28	3.981E-4
CCMSLIB00000577662	Enniatin-B2				Fungus	22/22	4.016E-4
CCMSLIB00000577621	Bacitracin B2	other	peptide	antibiotic	Bacterium	61/88	4.133E-4
CCMSLIB00000577577	[Leu7]-Surfactin C13ai dimethyl ester					24/27	4.298E-4
CCMSLIB00000577703	Enniatin J1				Fungus	21/21	4.457E-4

Spectrum ID	Compound	Structure	Category	Activities	Origin	#Scored/#Annot	P-value
CCMSLIB00000577782	Massetolide H	partial cyclic	lipopeptide	antibiotic,surfactant	Bacterium	8/9	4.585E-4
CCMSLIB00000577552	Lichenysin-G9a				Bacterium	18/18	4.613E-4
CCMSLIB00000577652	Symplostatin 1				Bacterium	18/21	4.73E-4
CCMSLIB00000577529	Tumescenamide A				Bacterium	31/31	4.764E-4
CCMSLIB00000577647	Integramide A					55/61	4.998E-4
CCMSLIB00000577596	Bacitracin B2	other	peptide	antibiotic	Bacterium	63/89	5.335E-4
CCMSLIB00000577707	Verrucamide C				Fungus	26/30	5.398E-4
CCMSLIB00000577504	Leucinostatin B				Fungus	68/77	5.942E-4
CCMSLIB00000577686	Laxaphycin E				Fungus	28/31	6.375E-4
CCMSLIB00000577632	CNC457.1424					19/24	6.574E-4
CCMSLIB00000577683	[Val7]-Surfactin C14i					10/10	7.326E-4
CCMSLIB00000577553	Ayfivin					96/133	8.34E-4
CCMSLIB00000577548	Puwainaphycin C				Bacterium	135/194	8.35E-4
CCMSLIB00000577515	Neotelomycin					82/115	8.756E-4
CCMSLIB00000577768	Actinomycin X0beta					25/28	9.124E-4
CCMSLIB00000577682	Enniatin I	cyclic	peptide	antibiotic,toxin	Fungus	12/12	9.127E-4
CCMSLIB00000577531	Hypomurocin B-3b				Fungus	132/149	9.172E-4
CCMSLIB00000577528	Isocyclosporin D					25/27	9.867E-4
CCMSLIB00000577618	Trichorozin-IV				Fungus	30/30	9.868E-4
CCMSLIB00000577685	[Ile2 4 7]Surfactin					14/14	0.0011
CCMSLIB00000577846	Symplostatin 2				Bacterium	14/17	0.0011
CCMSLIB00000577771	Enniatin L	cyclic	peptide	antibiotic,toxin	Bacterium	11/11	0.0012
CCMSLIB00000577805	Trichovirin I-3C					6/8	0.0012
CCMSLIB00000577832	Bassianolide				Fungus	12/12	0.0012
CCMSLIB00000577834	Sch-218157					9/10	0.0012
CCMSLIB00000577866	Ahpatinin F				Bacterium	7/8	0.0012
CCMSLIB00000577667	Enniatin-A1					20/20	0.0013
CCMSLIB00000577690	BZR-cotoxin II					23/28	0.0013
CCMSLIB00000577698	Trichokindin-IIb					106/156	0.0013
CCMSLIB00000577781	Verrucamide C				Fungus	50/58	0.0013

Spectrum ID	Compound	Structure	Category	Activities	Origin	#Scored/#Annot	P-value
CCMSLIB00000577803	Cyclosporin-I					16/16	0.0014
CCMSLIB00000577759	Pseudodestruxin B				Fungus	14/14	0.0015
CCMSLIB00000577691	[Leu-7]surfactin				Bacterium	17/17	0.0016
CCMSLIB00000577695	Leucinostatin N					6/6	0.0016
CCMSLIB00000577806	Tolaasin I	partial cyclic	lipopeptide	antibiotic,toxin,surfactant	Bacterium	8/10	0.0016
CCMSLIB00000577868	Clavariopsin B				Fungus	20/23	0.0016
CCMSLIB00000577644	Enniatin-B4					23/23	0.0019
CCMSLIB00000577825	Suzukacillin-B					20/49	0.0021
CCMSLIB00000577753	Boletusin					27/55	0.0022
CCMSLIB00000577828	Brevigellin				Fungus	4/7	0.0022
CCMSLIB00000577628	Cyclosporin-P					27/27	0.0029
CCMSLIB00000577783	Somamide B				Bacterium	22/34	0.0038
CCMSLIB00000577494	Cyclosporin-L					53/55	0.0041
CCMSLIB00000577496	Cyclosporin-E					56/61	0.0043
CCMSLIB00000577550	Trichopolyn-I					17/19	0.0044
CCMSLIB00000577564	[MeVal]5-cyclosporin					20/22	0.0044
CCMSLIB00000577593	[8'-Hydroxy-MeBmf]1-cyclosporin					18/19	0.0056
CCMSLIB00000577839	Stenothricin component IV					69/125	0.006
CCMSLIB00000577540	Cyclosporin-C					39/46	0.0062
CCMSLIB00000577835	Tyr-Ala-Gly-Phe-Leu-Arg					12/13	0.0062
CCMSLIB00000577530	Cyclosporin-C					34/37	0.0071
CCMSLIB00000577648	Laxaphycin B2				Bacterium	28/34	0.0075
CCMSLIB00000577502	Cyclosporin-U					43/47	0.0079
CCMSLIB00000577848	Telomycin					8/10	0.008
CCMSLIB00000577508	Cyclosporin-U					38/42	0.0084
CCMSLIB00000577500	Cyclosporin-C					63/68	0.0086
CCMSLIB00000577812	Yanucamide A				Bacterium	4/4	0.0096
CCMSLIB00000577786	Puwainaphycin D					86/117	0.0103
CCMSLIB00000577607	Actinomycin X2					19/27	0.0108
CCMSLIB00000577507	Cyclosporin-Y					41/50	0.0111

Spectrum ID	Compound	Structure	Category	Activities	Origin	#Scored/#Annot	P-value
CCMSLIB00000577865	Hymenamide K					52/78	0.0111
CCMSLIB00000577582	Cyclosporin-A					17/17	0.0121
CCMSLIB00000577501	Cyclosporin-U					45/50	0.0131
CCMSLIB00000577711	Hypomurocin B-5				Fungus	129/165	0.0144
CCMSLIB00000577699	Stendomycin					9/10	0.0152
CCMSLIB00000577568	Actinomycin D-CDM-WM-IM					48/51	0.0166
CCMSLIB00000577495	Cyclosporin-U					54/60	0.0167
CCMSLIB00000577510	Cyclosporin-X					61/72	0.0174
CCMSLIB00000577754	Stenothricin component III					91/168	0.0178
CCMSLIB00000577728	Actinomycin VI	double cyclic	chromopeptide	antibiotic,antitumor		19/19	0.018
CCMSLIB00000577770	A C3					257/482	0.0182
CCMSLIB00000577784	SCH-378199					66/90	0.0216
CCMSLIB00000577739	Trichovirin II 2c				Fungus	41/69	0.0228
CCMSLIB00000577712	Ilamycin					10/10	0.0229
CCMSLIB00000577751	Actinomycin K2c					29/31	0.0251
CCMSLIB00000577665	Actinomycin Au6a					28/29	0.0254
CCMSLIB00000577817	Lyngbyazothrin B				Bacterium	12/12	0.0346
CCMSLIB00000577847	Cyclosporin-K					21/24	0.0389
CCMSLIB00000577810	Aureobasidin-T3				Fungus	14/16	0.0443
CCMSLIB00000577738	Laxaphycin D				Fungus	19/22	0.0459
CCMSLIB00000577813	Maltacine B2b				Bacterium	53/117	0.0461
CCMSLIB00000577732	Antamanide					16/20	0.0498

A.2 PNPs identified by NRPro and Dereplicator in the GNPS spectra

Table -S5: PNPs identified by NRPro and Dereplicator in the GNPS spectra. The "rang" column specifies the position of the candidate in NRPro. Parent and fragment ion tolerance: 0.02.

Spectrum ID	Compound	Structure	Category	Activities	Origin	#Scored/#Annot	Rang	P-value
CCMSLIB00000577522	Acremostatin B					34/37	1	1.0E-16
CCMSLIB00000577573	Leucinostatin F					20/22	1	1.0E-16
CCMSLIB00000577587	Lyngbyastatin 3				Bacterium	24/24	1	1.0E-16
CCMSLIB00000577616	SNA-60-367-14				Bacterium	10/14	1	1.0E-16
CCMSLIB00000577637	Cephaibol-C					17/22	1	1.0E-16
CCMSLIB00000577675	Almiramide C				Bacterium	8/8	1	1.0E-16
CCMSLIB00000577677	Pepstatin B				Bacterium	15/15	1	1.0E-16
CCMSLIB00000577680	WS-7338-B					9/9	1	1.0E-16
CCMSLIB00000577716	Leucinostatin S					8/10	1	1.0E-16
CCMSLIB00000577726	Efrapeptin J					6/8	1	1.0E-16
CCMSLIB00000577748	Heptaibin	linear	peptaibol	antibiotic	Fungus	10/19	1	1.0E-16
CCMSLIB00000577761	Trichovirin I-7A					7/8	1	1.0E-16
CCMSLIB00000577777	Apramide B	linear	lipopeptide	unknown	Bacterium	7/7	1	1.0E-16
CCMSLIB00000577821	A-1437-M					11/15	1	1.0E-16
CCMSLIB00000577660	Bacillomycin D2					10/10	1	1.11E-16
CCMSLIB00000577755	Veraguamide J				Bacterium	12/12	1	1.11E-16
CCMSLIB00000577506	Leucinostatin D					22/23	1	4.44E-16
CCMSLIB00000577704	JBIR-114					14/16	1	6.661E-16
CCMSLIB00000577789	Efrapeptin-C					9/13	1	2.442E-15
CCMSLIB00000577597	Cordyheptapeptide A				Fungus	17/17	1	2.775E-15
CCMSLIB00000577556	Acremostatin A					48/50	1	3.108E-15
CCMSLIB00000577590	PHB					12/12	1	8.104E-15
CCMSLIB00000577666	Trichokindin-IIa				Fungus	18/33	1	1.088E-14
CCMSLIB00000577766	Carmabin B	linear	lipopeptide	surfactant	Bacterium	10/10	1	1.11E-14
CCMSLIB00000577575	Trikoningin-KB-I					31/31	1	2.875E-14
CCMSLIB00000577701	WS-7338-D					12/12	1	4.962E-14
CCMSLIB00000577512	SNA-60-367-19				Bacterium	28/30	1	7.194E-14
CCMSLIB00000577580	OHB					12/12	1	7.36E-14
CCMSLIB00000577689	SNA-60-367-21					11/14	1	1.001E-13
CCMSLIB00000577717	Almiramide A				Bacterium	9/10	1	1.495E-13
CCMSLIB00000577714	Dibenarthin				Bacterium	11/11	1	4.541E-13
CCMSLIB00000577549	Orfamide C	partial cyclic	lipopeptide	antibiotic,surfactant	Bacterium	33/36	1	1.279E-12
CCMSLIB00000577723	Cotteslosin B				Fungus	16/16	1	2.002E-12
CCMSLIB00000577676	Trichokindin-VI					10/12	1	1.18E-11
CCMSLIB00000577678	Apramide A	linear	lipopeptide	unknown	Bacterium	15/15	1	2.172E-11
CCMSLIB00000577524	Efrapeptin-F					23/27	1	2.251E-11
CCMSLIB00000577643	Pepstatin BU					14/14	1	3.71E-11

Spectrum ID	Compound	Structure	Category	Activities	Origin	#Scored/#Annot	Rang	P-value
CCMSLIB00000577772	RHM3					5/5	1	4.258E-11
CCMSLIB00000577535	Leucinostatin B2					29/29	1	6.399E-11
CCMSLIB00000577705	Almiramide B				Bacterium	10/10	1	1.102E-10
CCMSLIB00000577547	Aspergillicin B				Fungus	36/39	1	2.181E-10
CCMSLIB00000577491	Hormothamnin A					78/90	1	2.816E-10
CCMSLIB00000577583	Trichocellin-A-VI				Fungus	20/28	2	3.447E-10
CCMSLIB00000577733	Kahalalide G	linear	peptide	unknown		8/9	1	4.814E-10
CCMSLIB00000577730	RHM2				Fungus	7/9	1	8.142E-10
CCMSLIB00000577609	Trichokindin-IVa					32/41	1	9.608E-10
CCMSLIB00000577615	Cycloaspeptide A				Fungus	31/39	1	1.008E-9
CCMSLIB00000577617	Grividomycin-III					12/13	1	1.081E-9
CCMSLIB00000577588	Aspergillicin E				Fungus	20/20	1	1.145E-9
CCMSLIB00000577624	Pumilacidin B				Bacterium	13/13	1	1.217E-9
CCMSLIB00000577532	Leucinostatin A2					29/29	1	1.33E-9
CCMSLIB00000577521	Leucinostatin K					39/39	1	2.399E-9
CCMSLIB00000577671	Veraguamide E				Bacterium	19/20	1	3.405E-9
CCMSLIB00000577790	Dragonamide E				Bacterium	9/9	1	4.013E-9
CCMSLIB00000577734	Beauvericin G2					16/16	1	5.697E-9
CCMSLIB00000577792	Apramide C				Bacterium	9/10	1	6.083E-9
CCMSLIB00000577851	Amonabactin-P-750					6/6	1	8.006E-9
CCMSLIB00000577788	SNA-60-367-23				Bacterium	6/7	1	9.324E-9
CCMSLIB00000577797	Destruxin D					10/11	1	9.834E-9
CCMSLIB00000577636	WS-9320-A					20/23	1	1.023E-8
CCMSLIB00000577692	Ilamycin B1					8/8	1	1.718E-8
CCMSLIB00000577720	RHM1				Fungus	7/7	1	1.883E-8
CCMSLIB00000577640	[Val7]-Surfactin C13ai					10/10	1	1.895E-8
CCMSLIB00000577760	Oxachelin				Bacterium	14/15	1	1.9E-8
CCMSLIB00000577724	Tyrocidine-B					13/13	1	1.941E-8
CCMSLIB00000577787	[Val7]-Surfactin C14i monomethyl ester					21/26	1	2.063E-8
CCMSLIB00000577769	Trichorovin-Ia				Fungus	8/9	2	2.279E-8
CCMSLIB00000577653	Trichotoxin-A					50/75	1	2.472E-8
CCMSLIB00000577518	Leucinostatin R					25/26	1	2.505E-8
CCMSLIB00000577725	Fungisporin					10/10	1	2.589E-8
CCMSLIB00000577674	Veraguamide C				Bacterium	12/13	1	2.725E-8
CCMSLIB00000577623	Trichosporin-B-III-a					18/32	2	2.999E-8
CCMSLIB00000577731	Veraguamide K				Bacterium	18/18	1	3.021E-8
CCMSLIB00000577740	Pitipeptolide B				Bacterium	17/18	1	3.485E-8

Spectrum ID	Compound	Structure	Category	Activities	Origin	#Scored/#Annot	Rang	P-value
CCMSLIB00000577641	Aselacin-C					21/34	1	3.633E-8
CCMSLIB00000577513	Emerimicin IV	linear	peptaibol	antibiotic	Fungus	67/73	1	4.035E-8
CCMSLIB00000577764	Veraguamide H					11/11	1	4.326E-8
CCMSLIB00000577679	Integramide B					22/25	1	6.983E-8
CCMSLIB00000577763	W-493-A					21/22	1	7.472E-8
CCMSLIB00000577610	"APD I component a					19/19	1	8.643E-8
CCMSLIB00000577795	Cocosamide A				Bacterium	11/14	1	8.794E-8
CCMSLIB00000577595	WS-7338-A					26/28	1	9.121E-8
CCMSLIB00000577836	WIN-66306					8/8	1	9.236E-8
CCMSLIB00000577713	Destruxin D1					21/24	1	1.111E-7
CCMSLIB00000577708	Scopularide A					11/11	1	1.237E-7
CCMSLIB00000577659	A-Substance Ib					15/15	1	1.526E-7
CCMSLIB00000577585	Carriebowmide				Bacterium	24/26	1	1.809E-7
CCMSLIB00000577559	Massetolide G	partial cyclic	lipopeptide	antibiotic,surfactant	Bacterium	35/37	1	1.852E-7
CCMSLIB00000577696	Pepstatin A					20/20	1	1.873E-7
CCMSLIB00000577571	Leucinostatin V					24/26	1	1.96E-7
CCMSLIB00000577762	Brunsvicamide C				Bacterium	11/11	1	2.016E-7
CCMSLIB00000577584	Massetolide C	partial cyclic	lipopeptide	antibiotic,surfactant	Bacterium	17/17	1	2.101E-7
CCMSLIB00000577538	Cotteslosin A				Fungus	31/31	1	2.204E-7
CCMSLIB00000577796	Lyngbyastatin 1				Bacterium	17/17	1	2.411E-7
CCMSLIB00000577516	Majusculamide C	cyclic	peptide	antibiotic,antitumor	Bacterium	43/47	1	2.805E-7
CCMSLIB00000577830	Versicoloritide A				Fungus	7/8	1	2.826E-7
CCMSLIB00000577861	Kulomo'opunalide-1					12/12	1	3.365E-7
CCMSLIB00000577668	Veraguamide G				Bacterium	23/26	1	3.459E-7
CCMSLIB00000577599	JBIR-113					24/25	1	3.477E-7
CCMSLIB00000577672	Cocosamide B				Bacterium	16/16	1	3.602E-7
CCMSLIB00000577567	Bacillopeptin-B					57/78	1	3.62E-7
CCMSLIB00000577614	Anthranicine					18/18	1	3.768E-7
CCMSLIB00000577833	Aspergillicin C				Fungus	7/10	1	4.174E-7
CCMSLIB00000577601	None					31/31	1	4.291E-7
CCMSLIB00000577831	E-Dehydroapratoxin A					8/8	1	4.307E-7
CCMSLIB00000577591	Orfamide B	partial cyclic	lipopeptide	antibiotic,surfactant	Bacterium	15/16	1	4.364E-7
CCMSLIB00000577838	Brunsvicamide A				Bacterium	13/13	1	4.378E-7
CCMSLIB00000577626	Destruxin E1	cyclic	peptide	toxin		32/36	1	4.414E-7
CCMSLIB00000577684	Veraguamide A				Bacterium	18/18	1	4.423E-7
CCMSLIB00000577773	Wewakazole					16/16	1	4.618E-7
CCMSLIB00000577744	Dolastatin 15	linear	peptide	antitumor		7/7	1	4.783E-7

Spectrum ID	Compound	Structure	Category	Activities	Origin	#Scored/#Annot	Rang	P-value
CCMSLIB00000577854	SW-163B				Bacterium	12/16	1	5.361E-7
CCMSLIB00000577816	Dragomabin				Bacterium	13/13	1	5.582E-7
CCMSLIB00000577780	Majusculamide D					14/17	1	6.133E-7
CCMSLIB00000577551	Hypomurocin B-1				Fungus	98/117	1	6.19E-7
CCMSLIB00000577776	Cyclomarin A				Bacterium	6/6	1	6.436E-7
CCMSLIB00000577818	Glumamycin					10/13	1	6.589E-7
CCMSLIB00000577742	Dragonamide					9/9	1	7.098E-7
CCMSLIB00000577855	Harzianin HC-15					7/7	1	7.301E-7
CCMSLIB00000577499	SNA-60-367-17				Bacterium	108/119	1	8.15E-7
CCMSLIB00000577581	[Val7]-Surfactin C15ai dimethyl ester					30/35	1	8.47E-7
CCMSLIB00000577525	Bacillomycin D2					46/55	1	9.509E-7
CCMSLIB00000577537	Ilamycin B2					17/18	1	1.038E-6
CCMSLIB00000577864	Massetolide E	partial cyclic	lipopeptide	antibiotic,surfactant	Bacterium	6/7	1	1.124E-6
CCMSLIB00000577605	Lichenysin-G1a					11/11	1	1.301E-6
CCMSLIB00000577799	Cyclosporin-Z					9/9	1	1.306E-6
CCMSLIB00000577693	Emericellamide B				Fungus	17/18	1	1.462E-6
CCMSLIB00000577545	Lipopeptide NO				Bacterium	28/28	1	1.553E-6
CCMSLIB00000577634	Veraguamide I					24/26	1	1.561E-6
CCMSLIB00000577625	Fijimycin A					11/11	1	2.214E-6
CCMSLIB00000577606	Viscosinamide					32/34	1	2.218E-6
CCMSLIB00000577849	Dragonamide B				Bacterium	8/8	1	2.329E-6
CCMSLIB00000577602	Trichosporin-B-IIIc					24/35	2	2.367E-6
CCMSLIB00000577646	Dolastatin 10	linear	peptide	antitumor		14/15	1	2.384E-6
CCMSLIB00000577569	Pumilacidin E					29/30	1	2.449E-6
CCMSLIB00000577541	Trikoningin-KB-II					56/58	1	2.945E-6
CCMSLIB00000577721	Pleofungin A				Fungus	19/22	1	3.098E-6
CCMSLIB00000577858	Protodestruxin					13/15	1	3.309E-6
CCMSLIB00000577562	Pumilacidin C					26/26	1	3.393E-6
CCMSLIB00000577656	Beauvericin D	cyclic	peptide	antibiotic	Fungus	20/20	1	3.9E-6
CCMSLIB00000577655	Isariin E					22/22	1	3.952E-6
CCMSLIB00000577598	Isariin G1					26/26	1	4.07E-6
CCMSLIB00000577533	Efrapeptin H					38/46	1	4.107E-6
CCMSLIB00000577843	Tyrocidin B1					13/14	1	4.337E-6
CCMSLIB00000577604	[§-Me-Pro] destruxin E chlorohydrin					25/29	1	4.426E-6
CCMSLIB00000577635	Isariin	cyclic	peptide	antibiotic		28/31	1	4.435E-6
CCMSLIB00000577592	Destruxin-Ed(1)					19/19	1	4.557E-6
CCMSLIB00000577673	Brunsvicamide B				Bacterium	16/16	1	4.748E-6

Spectrum ID	Compound	Structure	Category	Activities	Origin	#Scored/#Annot	Rang	P-value
CCMSLIB00000577629	Sporidesmolide-I					32/33	1	5.063E-6
CCMSLIB00000577791	Kulomo'opunalide-2					24/24	1	5.081E-6
CCMSLIB00000577613	Trichovirin II 5				Fungus	50/67	2	5.13E-6
CCMSLIB00000577560	Trichorzin HA-3				Fungus	95/123	1	5.427E-6
CCMSLIB00000577589	N-methylsansalvamide				Fungus	28/29	1	5.586E-6
CCMSLIB00000577526	Bacillomycin D4					22/23	1	6.001E-6
CCMSLIB00000577775	Carmabin A	linear	lipopeptide	surfactant	Bacterium	12/12	1	6.168E-6
CCMSLIB00000577737	Enniatin M1				Bacterium	10/10	1	6.437E-6
CCMSLIB00000577639	Etamycin VI-2				Bacterium	12/12	1	6.961E-6
CCMSLIB00000577779	Hantupeptin A				Bacterium	26/30	2	6.962E-6
CCMSLIB00000577579	Pitiprolamide				Bacterium	47/52	1	7.973E-6
CCMSLIB00000577645	Pumilacidin D					34/35	1	8.019E-6
CCMSLIB00000577669	Sansalvamide				Fungus	19/20	1	8.346E-6
CCMSLIB00000577554	Aspergillicin A				Fungus	47/57	1	8.692E-6
CCMSLIB00000577670	Cyclosporin-R					12/12	1	9.32E-6
CCMSLIB00000577741	Tumescenamide B				Bacterium	19/20	1	9.908E-6
CCMSLIB00000577747	Arenamide C				Fungus,Bacterium	10/13	4	1.182E-5
CCMSLIB00000577794	Symplostatin 3				Bacterium	12/14	1	1.197E-5
CCMSLIB00000577687	Pumilacidin A					17/18	1	1.294E-5
CCMSLIB00000577546	Massetolide A	partial cyclic	lipopeptide	antibiotic,surfactant	Bacterium	43/43	1	1.535E-5
CCMSLIB00000577869	Trichorovin-IIb				Fungus	13/13	1	1.552E-5
CCMSLIB00000577815	Planktocyclin				Bacterium	14/18	1	1.612E-5
CCMSLIB00000577860	Desmethyldestruxin-A					11/12	1	1.754E-5
CCMSLIB00000577715	Radamicin					23/24	1	2.02E-5
CCMSLIB00000577706	Dolastatin D					24/25	1	2.141E-5
CCMSLIB00000577517	Trichogin-A-IV					61/67	1	2.164E-5
CCMSLIB00000577756	Plipastatin A2					25/33	1	2.233E-5
CCMSLIB00000577572	Dolastatin 14					24/24	1	2.396E-5
CCMSLIB00000577745	Emericellamide A				Fungus	8/8	1	2.414E-5
CCMSLIB00000577743	Isariin-II					18/18	3	2.569E-5
CCMSLIB00000577719	Microcystin-FR					16/18	1	2.776E-5
CCMSLIB00000577814	Beauvericin J				Fungus	13/13	1	3.209E-5
CCMSLIB00000577840	Massetolide B					35/49	2	3.23E-5
CCMSLIB00000577750	Hantupeptin C				Bacterium	17/17	2	3.257E-5
CCMSLIB00000577642	Sporidesmolide II					33/35	1	3.34E-5
CCMSLIB00000577576	Scopularide A					25/25	1	3.405E-5
CCMSLIB00000577709	Sporidesmolide V				Fungus	20/20	1	3.439E-5

Spectrum ID	Compound	Structure	Category	Activities	Origin	#Scored/#Annot	Rang	P-value
CCMSLIB00000577638	[Ala(2) Val(11)]Cyclosporin					15/15	1	3.623E-5
CCMSLIB00000577630	Destruxin B1	cyclic	peptide	toxin		29/32	1	3.743E-5
CCMSLIB00000577622	Pitipeptolide E				Bacterium	41/46	1	3.934E-5
CCMSLIB00000577857	Trichorovin-Xa				Fungus	5/5	1	3.936E-5
CCMSLIB00000577586	Mojavenisn A					19/19	1	3.968E-5
CCMSLIB00000577555	Massetolide F	partial cyclic	lipopeptide	antibiotic,surfactant	Bacterium	65/68	1	4.136E-5
CCMSLIB00000577651	Isariin D	cyclic	peptide	antibiotic	Fungus	21/22	1	4.285E-5
CCMSLIB00000577826	Sporidesmolide-III				Fungus	25/26	1	5.309E-5
CCMSLIB00000577702	LP-237-F7					51/62	1	5.879E-5
CCMSLIB00000577824	Enniatin-B3				Fungus	12/12	1	6.02E-5
CCMSLIB00000577697	Pepstatin AC					13/13	1	6.237E-5
CCMSLIB00000577536	Surfactin C1					47/49	2	6.238E-5
CCMSLIB00000577823	Hantupeptin B				Bacterium	20/21	2	6.67E-5
CCMSLIB00000577801	Trichorovin-IXa				Fungus	16/19	1	6.814E-5
CCMSLIB00000577837	Apratoxin G				Bacterium	8/8	1	6.863E-5
CCMSLIB00000577804	Harzianin HC-13					12/13	1	7.002E-5
CCMSLIB00000577612	Palmyramide A				Bacterium	24/25	1	7.335E-5
CCMSLIB00000577498	Orfamide A	partial cyclic	lipopeptide	antibiotic,surfactant	Bacterium	63/71	1	7.55E-5
CCMSLIB00000577758	Montanastatin	cyclic	peptide	toxin,antitumor		13/13	1	8.055E-5
CCMSLIB00000577633	Destruxin A1	cyclic	peptide	toxin		23/29	1	8.471E-5
CCMSLIB00000577802	Destruxin-E diol					11/11	1	8.516E-5
CCMSLIB00000577694	Hypomurocin B-2				Fungus	85/134	1	9.006E-5
CCMSLIB00000577594	Cyclosporin-V					17/17	1	9.24E-5
CCMSLIB00000577514	Lichenysin-G5a				Bacterium	27/28	1	9.336E-5
CCMSLIB00000577736	Peptaivirin A					31/47	1	9.716E-5
CCMSLIB00000577509	Ibu-epidemethoxylyngbystatin 3					68/86	1	1.004E-4
CCMSLIB00000577718	Trichorovin-IIIb					14/17	1	1.012E-4
CCMSLIB00000577539	Lichenysin-G3					21/21	1	1.032E-4
CCMSLIB00000577565	Bacirtsin-2					41/41	2	1.042E-4
CCMSLIB00000577505	Leucinostatin A					64/74	1	1.049E-4
CCMSLIB00000577822	[D-Asp3 ADMAdda5]microcystine-LR					40/53	1	1.074E-4
CCMSLIB00000577853	Preneocarzinostatin					30/39	1	1.127E-4
CCMSLIB00000577497	Desmethoxymajusculamide C				Bacterium	74/84	1	1.133E-4
CCMSLIB00000577819	MK 1688					16/16	1	1.287E-4
CCMSLIB00000577850	SPF-5506-A4				Fungus	8/9	3	1.379E-4
CCMSLIB00000577543	Lichenysin A					26/26	1	1.415E-4
CCMSLIB00000577862	Beauvericin	cyclic	peptide	antibiotic	Fungus	16/18	1	1.439E-4

Spectrum ID	Compound	Structure	Category	Activities	Origin	#Scored/#Annot	Rang	P-value
CCMSLIB00000577688	Beauvericin-B					18/18	1	1.507E-4
CCMSLIB00000577663	Isariin C2					28/29	1	1.573E-4
CCMSLIB00000577611	Microcystin LR	cyclic	PK-NRP	toxin		35/45	1	1.708E-4
CCMSLIB00000577578	Trichovirin II 1a				Fungus	97/129	2	1.748E-4
CCMSLIB00000577785	Carriebowmide sulfon					13/13	1	1.802E-4
CCMSLIB00000577722	Nordolastatin G					12/13	1	1.847E-4
CCMSLIB00000577654	Trichorzin HA-5				Fungus	65/95	5	1.854E-4
CCMSLIB00000577729	Stilbellin I					7/11	2	1.864E-4
CCMSLIB00000577520	Halobacillin				Bacterium	46/50	1	1.93E-4
CCMSLIB00000577493	Plipastatin A1					83/88	1	2.056E-4
CCMSLIB00000577608	[Phe3 N-MeVal5]-Destruxin B					35/41	1	2.059E-4
CCMSLIB00000577658	Kurstakin 4				Bacterium	18/18	1	2.089E-4
CCMSLIB00000577700	Hypomurocin A-2				Fungus	28/34	1	2.129E-4
CCMSLIB00000577664	Plipastatin B1					17/18	1	2.175E-4
CCMSLIB00000577511	[Val7]-Surfactin C15ai					39/41	1	2.224E-4
CCMSLIB00000577600	Trichorzin MA-1				Fungus	115/138	2	2.386E-4
CCMSLIB00000577534	Esperin					72/85	1	2.39E-4
CCMSLIB00000577519	Pumilacidin F					30/31	2	2.43E-4
CCMSLIB00000577681	Beauvericin-A					20/20	1	2.499E-4
CCMSLIB00000577566	[Val7]-Surfactin C14i dimethyl ester					28/33	1	2.619E-4
CCMSLIB00000577557	Bacitracin F				Bacterium	69/98	1	2.632E-4
CCMSLIB00000577627	[Leu7]-Surfactin C14i dimethyl ester					36/46	2	2.639E-4
CCMSLIB00000577619	Trichorozin-II				Fungus	30/31	1	2.686E-4
CCMSLIB00000577767	Coibamide A				Bacterium	14/15	1	2.784E-4
CCMSLIB00000577757	Veraguamide D				Bacterium	15/16	1	2.894E-4
CCMSLIB00000577727	Leucinostatin H					6/6	1	2.912E-4
CCMSLIB00000577574	[Leu7]-Surfactin C14i monomethyl ester					25/27	1	3.018E-4
CCMSLIB00000577561	Peptidolipin NA					20/20	1	3.132E-4
CCMSLIB00000577710	Trichoaureocin 3					14/31	1	3.212E-4
CCMSLIB00000577503	Surfactin-D					31/31	1	3.229E-4
CCMSLIB00000577492	Dolastatin 12	cyclic	peptide	antitumor		98/113	1	3.28E-4
CCMSLIB00000577746	Valinomycin	cyclic	peptide	antibiotic	Bacterium	29/29	1	3.371E-4
CCMSLIB00000577563	A C0					263/452	1	3.39E-4
CCMSLIB00000577527	Acremostatin C					18/20	1	3.586E-4
CCMSLIB00000577793	Pseudodesmin B				Bacterium	7/7	1	3.594E-4
CCMSLIB00000577650	Plipastatin B2					38/44	1	3.849E-4
CCMSLIB00000577661	SNA-60-367-4				Bacterium	44/49	1	3.953E-4

Spectrum ID	Compound	Structure	Category	Activities	Origin	#Scored/#Annot	Rang	P-value
CCMSLIB00000577657	Psuedodestruxin C				Fungus	26/28	1	3.981E-4
CCMSLIB00000577662	Enniatin-B2				Fungus	22/22	1	4.016E-4
CCMSLIB00000577577	[Leu7]-Surfactin C13ai dimethyl ester					24/27	1	4.298E-4
CCMSLIB00000577703	Enniatin J1				Fungus	21/21	1	4.457E-4
CCMSLIB00000577782	Massetolide H	partial cyclic	lipopeptide	antibiotic,surfactant	Bacterium	8/9	1	4.585E-4
CCMSLIB00000577552	Lichenysin-G9a				Bacterium	18/18	1	4.613E-4
CCMSLIB00000577808	Sch-217048				Fungus	17/19	2	4.616E-4
CCMSLIB00000577652	Symplostatin 1				Bacterium	18/21	1	4.73E-4
CCMSLIB00000577529	Tumescenamide A				Bacterium	31/31	1	4.764E-4
CCMSLIB00000577647	Integramide A					55/61	1	4.998E-4
CCMSLIB00000577707	Verrucamide C				Fungus	26/30	1	5.398E-4
CCMSLIB00000577504	Leucinostatin B				Fungus	68/77	1	5.942E-4
CCMSLIB00000577686	Laxaphycin E				Fungus	28/31	1	6.375E-4
CCMSLIB00000577620	Enniatin J2				Fungus	19/19	3	6.46E-4
CCMSLIB00000577632	CNC457.1424					19/24	1	6.574E-4
CCMSLIB00000577596	Bacitracin B1	other	peptide	antibiotic	Bacterium	61/95	3	6.615E-4
CCMSLIB00000577683	[Val7]-Surfactin C14i					10/10	1	7.326E-4
CCMSLIB00000577631	Trichorozin-I					22/22	5	7.461E-4
CCMSLIB00000577553	Ayfivin					96/133	1	8.34E-4
CCMSLIB00000577548	Puwainaphycin C				Bacterium	135/194	1	8.35E-4
CCMSLIB00000577621	Bacitracin B3	other	peptide	antibiotic	Bacterium	55/81	3	8.596E-4
CCMSLIB00000577768	Actinomycin X0beta					25/28	1	9.124E-4
CCMSLIB00000577515	Telomycin					81/109	2	9.127E-4
CCMSLIB00000577682	Enniatin I	cyclic	peptide	antibiotic,toxin	Fungus	12/12	1	9.127E-4
CCMSLIB00000577531	Hypomurocin B-3b				Fungus	132/149	1	9.172E-4
CCMSLIB00000577528	Isocyclosporin D					25/27	1	9.867E-4
CCMSLIB00000577618	Trichorozin-IV				Fungus	30/30	1	9.868E-4
CCMSLIB00000577685	[Ile2 4 7]Surfactin					14/14	1	0.0011
CCMSLIB00000577846	Symplostatin 2				Bacterium	14/17	1	0.0011
CCMSLIB00000577771	Enniatin L	cyclic	peptide	antibiotic,toxin	Bacterium	11/11	1	0.0012
CCMSLIB00000577805	Trichovirin I-3C					6/8	1	0.0012
CCMSLIB00000577832	Bassianolide				Fungus	12/12	1	0.0012
CCMSLIB00000577834	Sch-218157					9/10	1	0.0012
CCMSLIB00000577866	Ahpatinin F				Bacterium	7/8	1	0.0012
CCMSLIB00000577667	Enniatin-A1					20/20	1	0.0013
CCMSLIB00000577690	BZR-cotoxin II					23/28	1	0.0013
CCMSLIB00000577698	Trichokindin-IIb					106/156	1	0.0013

Spectrum ID	Compound	Structure	Category	Activities	Origin	#Scored/#Annot	Rang	P-value
CCMSLIB00000577803	Cyclosporin-I					16/16	1	0.0014
CCMSLIB00000577759	Pseudodestruxin B				Fungus	14/14	1	0.0015
CCMSLIB00000577695	Leucinostatin N					6/6	1	0.0016
CCMSLIB00000577806	Tolaasin I	partial cyclic	lipopeptide	antibiotic,toxin,surfactant	Bacterium	8/10	1	0.0016
CCMSLIB00000577868	Clavariopsin B				Fungus	20/23	1	0.0016
CCMSLIB00000577691	Bamylocin A				Bacterium	13/13	5	0.0017
CCMSLIB00000577749	Pimaydolide					10/10	5	0.0017
CCMSLIB00000577644	Enniatin-B4					23/23	1	0.0019
CCMSLIB00000577825	Suzukacillin-B					20/49	1	0.0021
CCMSLIB00000577753	Boletusin					27/55	1	0.0022
CCMSLIB00000577542	Trichoaureocin 4					69/105	2	0.0023
CCMSLIB00000577628	Cyclosporin-P					27/27	1	0.0029
CCMSLIB00000577781	Verrucamide B				Fungus	45/53	2	0.0033
CCMSLIB00000577783	Somamide B				Bacterium	22/34	1	0.0038
CCMSLIB00000577494	Cyclosporin-L					53/55	1	0.0041
CCMSLIB00000577496	Cyclosporin-E					56/61	1	0.0043
CCMSLIB00000577550	Trichopolyn-I					17/19	1	0.0044
CCMSLIB00000577564	[MeVal]5-cyclosporin					20/22	1	0.0044
CCMSLIB00000577593	[8'-Hydroxy-MeBmf]1-cyclosporin					18/19	1	0.0056
CCMSLIB00000577839	Stenothricin component IV					69/125	1	0.006
CCMSLIB00000577835	Tyr-Ala-Gly-Phe-Leu-Arg					12/13	1	0.0062
CCMSLIB00000577828	Chymostatinol A					6/6	2	0.0064
CCMSLIB00000577540	G-Hydroxy-Meleu4-cyclosporin					38/45	3	0.0065
CCMSLIB00000577530	FR-901459					34/35	3	0.0072
CCMSLIB00000577648	Laxaphycin B2				Bacterium	28/34	1	0.0075
CCMSLIB00000577502	Cyclosporin-U					43/47	1	0.0079
CCMSLIB00000577508	[Leu(4)] Cyclosporin A					37/40	2	0.0085
CCMSLIB00000577500	Cyclosporin-C					63/68	1	0.0086
CCMSLIB00000577848	Neotelomycin					7/12	3	0.009
CCMSLIB00000577786	Puwainaphycin D					86/117	1	0.0103
CCMSLIB00000577607	Actinomycin X2					19/27	1	0.0108
CCMSLIB00000577507	Cyclosporin-Y					41/50	1	0.0111
CCMSLIB00000577865	Hymenamide K					52/78	1	0.0111
CCMSLIB00000577582	Cyclosporin-A					17/17	1	0.0121
CCMSLIB00000577501	Cyclosporin-T					45/46	2	0.0133
CCMSLIB00000577812	Mer-N5075-A				Bacterium	4/4	2	0.0149
CCMSLIB00000577699	Stendomycin					9/10	1	0.0152

Spectrum ID	Compound	Structure	Category	Activities	Origin	#Scored/#Annot	Rang	P-value
CCMSLIB00000577568	Actinomycin D-CDM-WM-IM					48/51	1	0.0166
CCMSLIB00000577495	Cyclosporin-N-B					49/52	4	0.0169
CCMSLIB00000577510	Cyclosporin-X					61/72	1	0.0174
CCMSLIB00000577754	Stenothricin component III					91/168	1	0.0178
CCMSLIB00000577770	A C3					257/482	1	0.0182
CCMSLIB00000577711	Trichorzin MA-2				Fungus	88/141	5	0.0189
CCMSLIB00000577859	Trichorovin-IVa				Fungus	7/8	8	0.0211
CCMSLIB00000577784	SCH-378199					66/90	1	0.0216
CCMSLIB00000577728	Actinomycin K1c					19/19	2	0.0225
CCMSLIB00000577739	Trichovirin II 2c				Fungus	41/69	1	0.0228
CCMSLIB00000577712	Ilamycin					10/10	1	0.0229
CCMSLIB00000577751	Actinomycin K2c					29/31	1	0.0251
CCMSLIB00000577665	Actinomycin Au6a					28/29	1	0.0254
CCMSLIB00000577817	Lyngbyazothrin B				Bacterium	12/12	1	0.0346
CCMSLIB00000577847	Cyclosporin-K					21/24	1	0.0389
CCMSLIB00000577810	Aureobasidin-T3				Fungus	14/16	1	0.0443
CCMSLIB00000577738	Laxaphycin D				Fungus	19/22	1	0.0459
CCMSLIB00000577813	Maltacine B2b				Bacterium	53/117	1	0.0461
CCMSLIB00000577732	Antamanide					16/20	1	0.0498

A.3 NRPro Manual

NRPro

User manual

Version 1.0

INDEX

I. SOFTWARE OVERVIEW

NRPro is a software dedicated to the MS/MS analysis of peptidic natural products (PNPs). It is available in a web application software in the Norine database (https://bioinfo.lifl.fr/nrpro/) and on the Expasy server (https://web.expasy.org/nrpro/). The software is intended to assist users in two tasks:

- **i)** **Dereplication:** The MS/MS spectra uploaded by the user are evaluated in order to identify the potential peptide candidates. This is done through theoretical fragment comparison against a chemical database.
- **ii)** **Annotation:** The fragment ion peak annotations of each peptide candidate suggested by NRPro are visualized in the interface and can be manually edited by the user.

II. VIEWS

NRPro consists of a single-page application (SPA) composed of two main views:

- **i)** **The form view (Figure 1A):** allows the user to submit the spectral files and to choose some of the search parameters such as parent ion tolerance or neutral losses (see Section III-Submission form).
- **ii)** **The results view (Figure 1B,C):** once the compounds have been identified the user can navigate through the results. First, a table with all identifications is shown (Figure 1B). Additionally, the table provides access to the spectral annotations of each compound (Figure 1C). Further details are given in Section IV-Results navigation.

Note that the help button ("?") in the form view points to a page in which a short video shows all the software views and instructs the user on how to use NRPro.

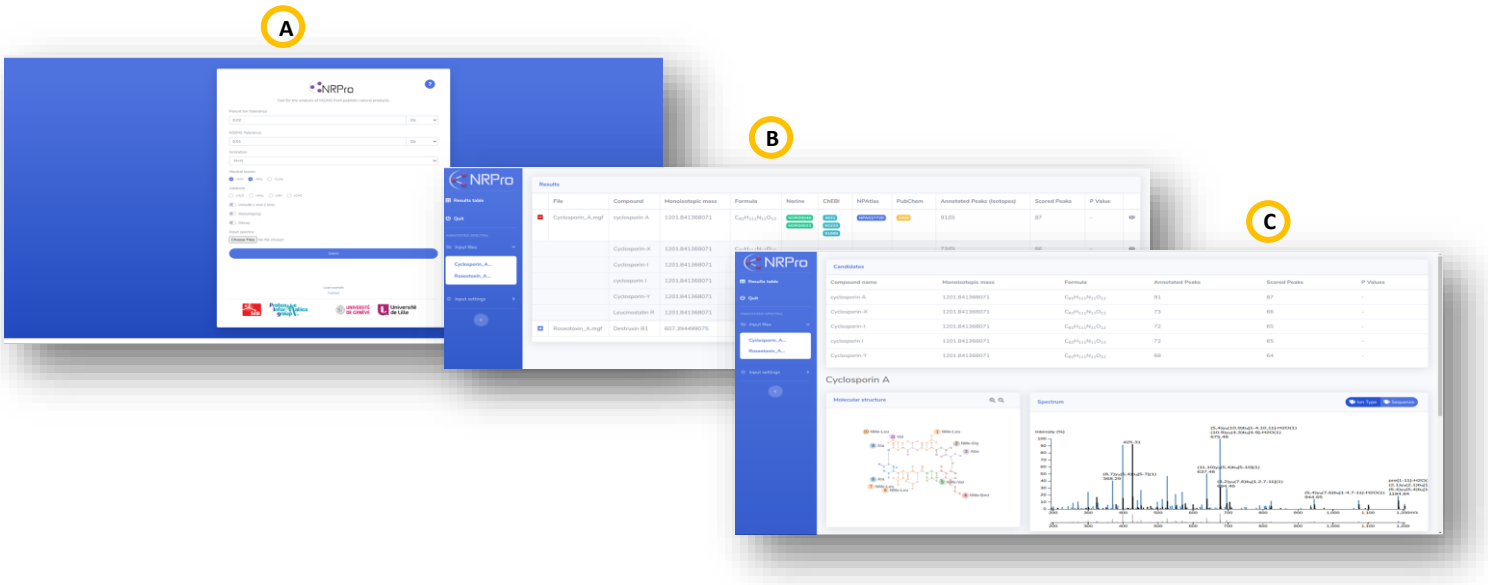


Figure 1. NRPro navigation views. A) Form view B) Identifications view C) Spectral annotations view.

III. SUBMISSION FORM

The submission form is the first view that the user encounters. As already mentioned, it allows the user to submit data and to choose a selection of search parameters. Figure 2 shows the interface of the form: each field labeled with a number is described in the following paragraphs.

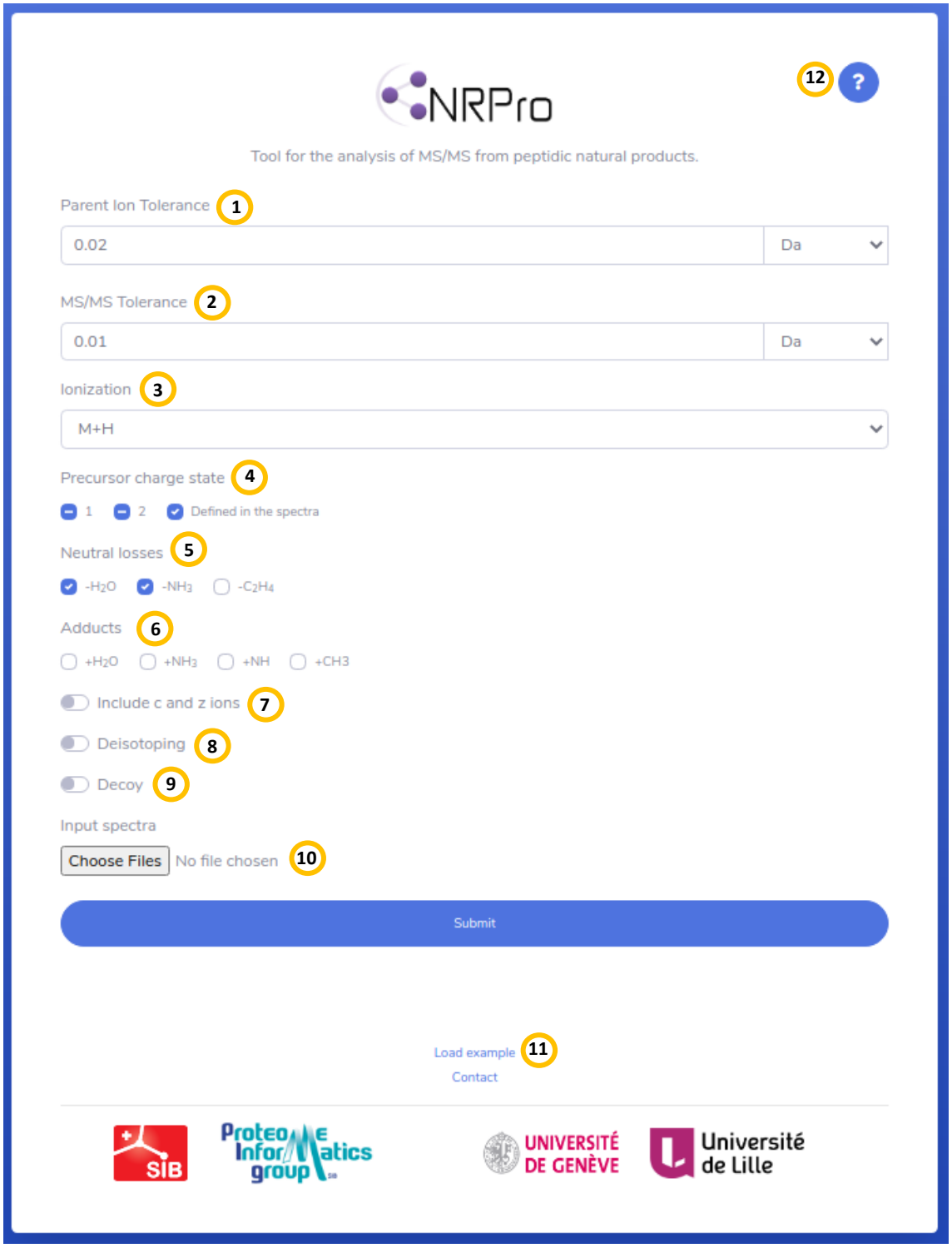


Figure 2. Submission form interface.

1, 2. Parent and MS/MS tolerance

Mass tolerance should be chosen according to the mass instrument employed. The maximal mass tolerance values are 1 Da and/or 50 ppm.

3. Ionization

The presence of K+ and Na+ adducts in the spectrum should be reported in the submission. Thus, differently ionized species (collections of both K+ and Na+ spectra) cannot be simultaneously submitted.

4. Precursor charge state

This field should be only specified if the charge state is not present in the headers of the submitted MS/MS spectra. We recommend having this information in the spectrum to allow the simultaneous submission of spectra with different charge states.

5, 6. Neutral losses and adducts

Select the expected neutral losses or adducts in the spectrum. Note that only ammonia and water neutral losses will be used for the scoring. Hence, the choice of additional neutral losses or adducts does not affect the dereplication process. They will be just reported in the peak annotations.

7. Inclusion of c/z ions

This option should be checked if the user wishes NRPro to consider *c* and *z* fragment ions as well. Note that it only refers to those fragments arising from amide and ester cleavages. In the case of glycosidic cleavages this type of ions is considered by default.

8.Deisotoping

In the visualization of the spectrum, the user can identify isotopic peaks colored in yellow. Note that isotopes are not considered for the scoring.

9.Decoy

The decoy activation generates *p-values* associated with the identifications. Otherwise the candidates are scored without statistical significance. Note that activating the decoy slows down the dereplication process.

10. Input spectra

Multiple files can be introduced with a maximum of 20 spectra. The software accepts the **MGF format** spectra with the following characteristics:

- A single spectrum per file. Multiple spectra in the same file are NOT allowed.

- The precursor *m/z* should be specified in the header with the *PEPMASS* tag (see Figure 3).
- The charge must be specified if it is not done during the form submission (4. Precursor charge state).
- Additional fields such as InCHIKey or SMILES are not required. If present, they are ignored.
- The file should meet the MGF standards, starting with the tag *BEGIN IONS* and finishing with *END IONS*. Find an example in Figure 3.

Note that NRPro can also read **mzXML files**, but this format has not been exhaustively tested. Importantly, the file extension is required in the name of the files (*.mgf* or *.mzXML*).

```
BEGIN IONS
PEPMASS=1448.4385
CHARGE=1
955.1716308594  1.3976086376
1115.2958984375 2.7365397699
1116.2989501953 2.0103579774
1117.2938232422 2.2869768348
1143.2294921875 1.6655460764
1143.2919921875 30.8256904668
1144.2955322266 18.7609825979
1145.2896728516 18.2739953280
1146.2923583984 2.0292950760
1277.3497314453 2.4274453202
1278.3526611328 1.7848258406
1279.3483886719 2.2475892309
1305.1820068359 1.2563756174
1305.2055664062 1.1082483028
1305.2667236328 5.8351812267
1305.3448486328 100.0000000000
1305.4414062500 1.1138192007
1306.1817626953 1.0460676101
1306.2064208984 1.0604953010
1306.2685546875 4.1744764342
1306.3483886719 69.2167383297
1307.1751708984 1.1199757927
1307.2687988281 4.2188279299
1307.3430175781 73.7067505643
1308.3460693359 7.4359896232
1331.2790527344 1.2333808993
1331.3616943359 15.5410036162
1332.3648681641 11.2536297111
1333.2808837891 1.0715574337
1333.3598632812 12.1764503849
1334.3637695312 1.3203185919
1387.3854980469 1.3179814646
1592.4593505859 1.0192420652
END IONS
```

Figure 3. Example of an MGF file of vancomycin (*m/z* 1448.43) meeting NRPro criteria.

11. Load example

Clicking on this button will automatically load three files and the optimal parameters to process them. Then, the user can click on the submission button to generate the results. This option is particularly recommended for new users testing the software for the first time.

12. Help (?)

Includes the main instructions on how to use the software as well as a **tutorial video** to show the main features of the software.

IV. POTENTIAL ERRORS DURING FORM SUBMISSION

Problems in the submission may occur for multiple reasons. NRPro error messages are targeted to assist the user in solving the problems. In this section we detail the different errors that may occur.

Format-related errors

ERROR: Multiple spectra in the file/s. Please make sure that each file contains a single spectrum.

ERROR: Unreadable file/s. Please make sure that the format is accepted and specified in the file extension.

These errors occur when the file/s are not in the correct format. That can be caused by:

- Multiple spectra inside a single file.
- Not supported formats. Only Mgf and mzXML are accepted.
- Inconsistencies in the format.
- Missing format extension (.mgf or .mzXML) in the name of the file.

Header-related errors

ERROR: Multiply charged spectra with K/Na adducts are not allowed.

ERROR: Spectra with more than 2 charges are not allowed.

ERROR: Unknown charge. Please verify that the precursor charge is specified in the spectral data.

ERROR: Unknown precursor mass. Please verify that the precursor mass is specified in the spectral data.

Either unacceptable or missing information may cause these errors:

- The precursor *m/z* or the charge are missing in the header of the files. In the case of the charge that can be also fixed by specifying it in the form parameters.
- The charge states of the peptides are not allowed. Note that currently, peptides with more than two charges are not accepted. Neither +2K or +2Na spectra.

Other errors

ERROR: Maximum number of files exceeded. Please introduce a maximum of 20 files.

This error is self-explanatory and will occur when trying to submit more than 20 files.

Unexpected error. The compound/s could not be identified.

The "Unexpected error" is due to a problem during the dereplication process and may have multiple causes. Trying to submit the spectra after choosing different parameters in the form, may solve it.

Sorry... it seems that the server is not responding. Try it later.

This error spots a problem with the server connection. It can be related to maintenance and that is why trying later is suggested. If the problem persists, please contact us.

V. INTERFACE NAVIGATION

Identifications view

The first results obtained from processing the data are the compound identifications. Those are represented in a table as shown in Figure 4. This table is described in the following paragraphs.

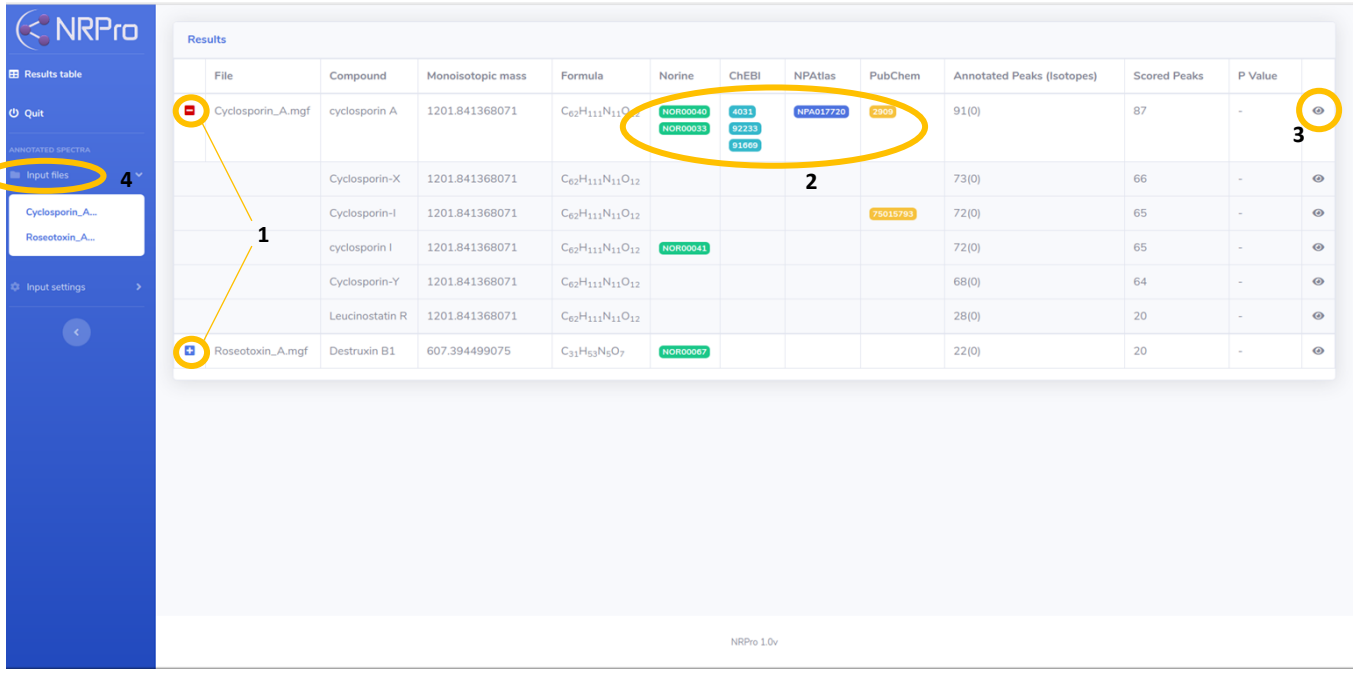


Figure 4. NRPro identification example. This is the first view after results submission.

The table of identifications provides the potential candidates for each spectrum (file) and their characteristics: name, monoisotopic mass, formula, identifiers (Norine, ChEBI, NPAtlas and PubChem), number of annotated peaks, number of scored peaks and p-values. The annotated peaks are not always the same as the scored peaks because some peaks with unusual neutral losses or adducts are not used for the scoring. The p-values will only appear in the table if the "Decoy" option was selected in the submitted form. The labels used in Figure 4 point to the following characteristics of the table:

1. **Extensibility:** The +/ – buttons are used to extend/shrink the table. It is useful to inspect all the candidates for the spectrum. When closed, it only shows the first candidate.
2. **Links:** the identifiers are clickable as they provide the link to the respective database resources.
3. **"Eye":** Provides access to the spectral annotations of the compounds, which is the view presented in the following section.
4. **Input files:** Represents an alternative way to navigate to the spectral annotations of each file (same function than the "eye").

Spectral annotations view

The spectral annotation view provides the fragment ion peak annotations of each candidate and visual support to assist the user on the characterization of the compounds. In this section we define the features used in this view.

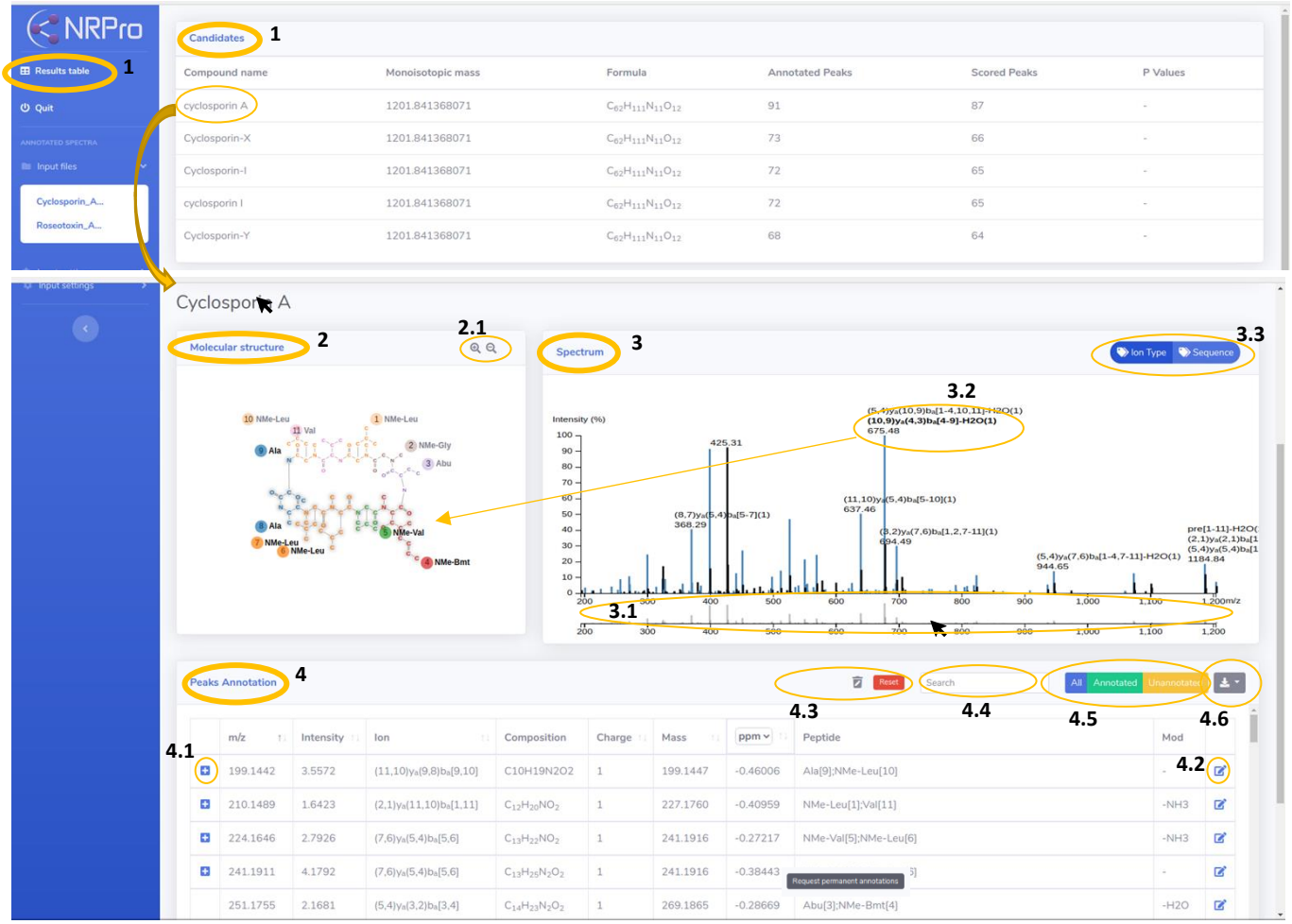


Figure 5. NRPro spectral annotations example.

As observed in Figure 5, the view is divided in four main sections:

1. **Candidates:** This table displays the candidates found by NRPro. The user can click on the candidate for inspection in order to update the sections 2,3 and 4 that will show its spectral annotations. In the example of Figure 4, the annotations of cyclosporin A are displayed.

2. **Molecular structure:** Representation of the candidate molecular structure. It also includes the monomer information that was obtained using rBAN[1]. The depiction is interactive so the user can move the atoms with the mouse in case of overlapping regions. Zooming is possible with the mouse middle button or the labeled icons (**2.1**).

3. **Spectrum:** Representation and color code of the queried (experimental) spectrum. Matched (annotated) peaks are blue, isotope peaks are yellow and black peaks are unmatched (unannotated). Some relevant features of the spectrum are:

3.1 The brushing area: useful to zoom in and out (two clicks).

3.2 Peak labeling: all the matched peaks are labeled. The labels of small intensity peaks will appear when zooming. As shown in Figure 4, the user can see the fragment ion corresponding to a given annotation highlighted in the molecule through mouse over.

3.3 Nomenclature options: the preferred nomenclature used to label the peaks can be selected by the user. Further details about the nomenclature are given in Section VI-Nomenclature.

4. **Peaks annotation:** This table contains all the spectral annotations and has multiple features:
   - **4.1 Extensible:** The +/ – buttons are used to extend/shrink the table. This feature is used to show several annotations of a single peak, because a peak may be explained by multiple fragment ions.
   - **4.2 Editable:** Clicking on this button allows the user to submit his/her own annotations. The annotations are stored in the server to allow future usage.
   - **4.3 Request permanent annotations:** By default, the URL of a query is available and shareable for a period of a month. This button contains the instructions to make the link permanent.
   - **4.4 Searchable:** This option allows the user to do fast typed-in searches in the table.
   - **4.5 Filterable:** By default, all the peaks appear in the table regardless of their annotations. These buttons allow the user to select only the matched or the unmatched peaks.
   - **4.6 Downloadable:** The table can be downloaded in CSV, XLS and JSON formats.

VI. NOMENCLATURE

NRPro relies on its own nomenclature format. Two options are offered to annotate the peak fragments: the “sequence” or “ion type” formats (**3.3**).

Sequence format

This format simply consists in listing the indexes of the monomers present in the fragment ion. For instance, the fragment *[5-7]* represents an ion fragment composed of the monomers with indexes 5,6 and 7. The monomer indexes are not always sequential. It is then possible to find fragments such as *[1-4,7-11]* that is read as a fragment containing the monomers: 1,2,3,4 and 7,8,9,10,11. The charge and additional modifications are written separately. For example the fragment *[5-7]-H2O(1)* includes a

neutral loss of water and a single charge. Note that this format is limited because it does not provide the type of breakage, which does not fully define the resulting fragment ion.

Ion type format

This format goes beyond the simple sequence definition and considers the type of bonds that are broken and the fragment ion that is formed (Figure 6). The fragments are labeled with the individual annotation of each cut (cleavage). A single cut annotation contains: i) the two indexes of the monomers involved in the fragmentation, followed by ii) the letter associated with the fragmented ion type (according to Roepstorff/Fohlman[2] and Biemann[3] for peptide bonds and Domon and Costello[4] for glycosidic bonds) and iii) a subscript letter specifying the bond type ("a" for amide; "e" for ester; "g" for glycosidic). For instance, as shown in Figure 6A, *(11,10)y_a(5,4)b_a* represents a fragment ion arising from a *y* cut between monomers 10 and 11 and a *b* cut between 4 and 5 (both amide bonds). Note that ester bonds are substitutes for amide bonds in depsipeptides, which led to adopt the same nomenclature to describe their breakage (Figure 6C).

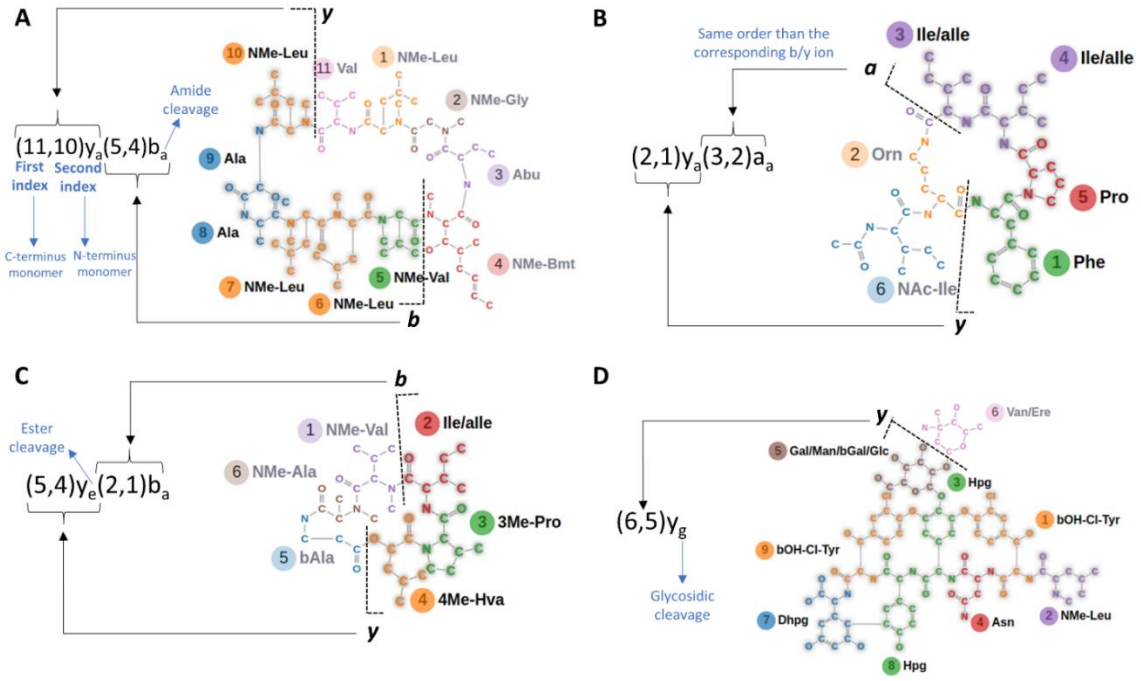


Figure 6. NRPro nomenclature. The highlighted fragment ions are annotated by defining each cleavage involved. The order of the monomer indexes is in accordance to the N- and C-termini. A) y/b ion in cyclosporin A, B) y/a ion in pseudacyclin A, C) y/b ion with ester cleavage in roseotoxin A, D) Glycosidic cleavage in vancomycin.

VII. RELEVANT FEATURES

-SHARABILITY: The URLs of NRPro are shareable between users for a period of one month. In order to increase this period or generate permanent links the users should contact: helpdesk@expasy.org.

-EDITABILITY: Manual editions are allowed for all the users having access to the links and they are stored on the server in order to make them sharable/reusable. When sharing, **please beware of modifying the annotations of other users as they are overwritten**.

-BLOCKING EDITABILITY: Currently this option is not automatically available but if users wish to share the URL in their publications, they can contact us in order to block annotations: helpdesk@expasy.org.

VIII. BROWSER COMPATIBILITY

Currently NRPro runs the best with Google Chrome. Other browsers can be used but they have not been exhaustively tested.

IX. REFERENCES

(1) Ricart, E.; Leclère, V.; Flissi, A.; Mueller, M.; Pupin, M.; Lisacek, F. RBAN: Retro-Biosynthetic Analysis of Nonribosomal Peptides. *Journal of cheminformatics* **2019**, *11* (1), 13.

(2) Roepstorff, P.; Fohlman, J. Proposal for a Common Nomenclature for Sequence Ions in Mass Spectra of Peptides. *Biomedical mass spectrometry* **1984**.

(3) Biemann, K. Contributions of Mass Spectrometry to Peptide and Protein Structure. *Biomedical & environmental mass spectrometry* **1988**, *16* (1–12), 99–111.

(4) Domon, B.; Costello, C. E. A Systematic Nomenclature for Carbohydrate Fragmentations in FAB-MS/MS Spectra of Glycoconjugates. *Glycoconjugate journal* **1988**, *5* (4), 397–409.

Bibliography

[1] Fleming, A. On the antibacterial action of cultures of a penicillium, with special reference to their use in the isolation of b. influenzae. *Reviews of infectious diseases* 129–139 (1980).

[2] 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).

[3] Deris, Z. Z. *et al.* Probing the penetration of antimicrobial polymyxin lipopeptides into gram-negative bacteria. *Bioconjugate chemistry* **25**, 750–760 (2014).

[4] Balleza, D., Alessandrini, A. & García, M. J. B. Role of lipid composition, physicochemical interactions, and membrane mechanics in the molecular actions of microbial cyclic lipopeptides. *The Journal of membrane biology* **252**, 131–157 (2019).

[5] Markham, A. Oritavancin: first global approval. *Drugs* **74**, 1823–1828 (2014).

[6] Brade, K. D., Rybak, J. M. & Rybak, M. J. Oritavancin: a new lipoglycopeptide antibiotic in the treatment of gram-positive infections. *Infectious diseases and therapy* **5**, 1–15 (2016).

[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] Mohamed, A., Nguyen, C. H. & Mamitsuka, H. Current status and prospects of computational resources for natural product dereplication: a review. *Briefings in bioinformatics* **17**, 309–321 (2016).

[9] Chavali, A. K. & Rhee, S. Y. Bioinformatics tools for the identification of gene clusters that biosynthesize specialized metabolites. *Briefings in bioinformatics* **19**, 1022–1034 (2018).

[10] Blin, K., Kim, H. U., Medema, M. H. & Weber, T. Recent development of antismash and other computational approaches to mine secondary metabolite biosynthetic gene clusters. *Briefings in bioinformatics* **20**, 1103–1113 (2019).

[11] Dejong, C. A. *et al.* Polyketide and nonribosomal peptide retro-biosynthesis and global gene cluster matching. *Nature chemical biology* **12**, 1007 (2016).

[12] Dufresne, Y., Noé, L., Leclère, V. & Pupin, M. Smiles2monomers: a link between chemical and biological structures for polymers. *Journal of cheminformatics* **7**, 1–11 (2015).

[13] Wilhelm, M. *et al.* Mass-spectrometry-based draft of the human proteome. *Nature* **509**, 582–587 (2014).

[14] Böcker, S., Letzel, M. C., Lipták, Z. & Pervukhin, A. Sirius: decomposing isotope patterns for metabolite identification. *Bioinformatics* **25**, 218–224 (2009).

[15] Rockwood, A. L. & Haimi, P. Efficient calculation of accurate masses of isotopic peaks. *Journal of the American Society for Mass Spectrometry* **17**, 415–419 (2006).

[16] Claesen, J., Dittwald, P., Burzykowski, T. & Valkenborg, D. An efficient method to calculate the aggregated isotopic distribution and exact center-masses. *Journal of The American Society for Mass Spectrometry* **23**, 753–763 (2012).

[17] Fernandez-de Cossio Diaz, J. & Fernandez-de Cossio, J. Computation of isotopic peak center-mass distribution by fourier transform. *Analytical Chemistry* **84**, 7052–7056 (2012).

[18] Spicer, R., Salek, R. M., Moreno, P., Cañueto, D. & Steinbeck, C. Navigating freely-available software tools for metabolomics analysis. *Metabolomics* **13**, 106 (2017).

[19] Dictionary of natural products 28.2. http://dnp.chemnetbase.com/faces/chemical/ChemicalSearch.xhtml. Accessed 14 Apr 2020.

[20] Antibase. https://application.wiley-vch.de/stmdata/antibase.php. Accessed 14 Apr 2020.

[21] Sorokina, M. & Steinbeck, C. Review on natural products databases: where to find data in 2020. *Journal of Cheminformatics* **12**, 1–51 (2020).

[22] Flissi, A. *et al.* Norine: update of the nonribosomal peptide resource. *Nucleic acids research* **48**, D465–D469 (2020).

[23] Organization, W. H. *et al. Antimicrobial resistance: global report on surveillance* (World Health Organization, 2014).

[24] Ling, L. L. *et al.* A new antibiotic kills pathogens without detectable resistance. *Nature* **517**, 455 (2015).

[25] Piddock, L. J. Teixobactin, the first of a new class of antibiotics discovered by ichip technology? *Journal of Antimicrobial Chemotherapy* **70**, 2679–2680 (2015).

[26] Rawal, T. & Butani, S. Teixobactin: A powerful tool for combating resistant strains. *Indian Journal of Pharmaceutical Sciences* **78**, 697–700 (2017).

[27] Walsh, C. *et al. Antibiotics: actions, origins, resistance.* (American Society for Microbiology (ASM), 2003).

[28] Demain, A. L. & Elander, R. P. The β-lactam antibiotics: past, present, and future. *Antonie Van Leeuwenhoek* **75**, 5–19 (1999).

[29] Paulus, H. & Gray, E. The biosynthesis of polymyxin b by growing cultures of bacillus polymyxa (1964).

[30] de Crécy-Lagard, V. *et al.* Pristinamycin i biosynthesis in streptomyces pristinaespiralis: molecular characterization of the first two structural peptide synthetase genes. *Journal of Bacteriology* **179**, 705–713 (1997).

[31] Kleinkauf, H. & von Döhren, H. The nonribosomal peptide biosynthetic system—on the origins of structural diversity of peptides, cyclopeptides and related compounds. *Antonie van Leeuwenhoek* **67**, 229–242 (1995).

[32] Johnson, B. A., Anker, H. & Meleney, F. L. Bacitracin: a new antibiotic produced by a member of the b. subtilis group. *Science* **102**, 376–377 (1945).

[33] Stark, W., Higgens, C., Wolfe, R., Hoehn, M. & McGuire, J. Capreomycin, a new antimycobacterial agent produced by streptomyces capreolus sp. n. *Antimicrobial Agents and Chemotherapy* **1962**, 596–606 (1962).

[34] Somma, S., Gastaldo, L. & Corti, A. Teicoplanin, a new antibiotic from actinoplanes teichomyceticus nov. sp. *Antimicrobial agents and chemotherapy* **26**, 917–923 (1984).

[35] van Wageningen, A. A. *et al.* Sequencing and analysis of genes involved in the biosynthesis of a vancomycin group antibiotic. *Chemistry & biology* **5**, 155–162 (1998).

[36] Abraham, E. & Newton, G. The structure of cephalosporin c. *Biochemical Journal* **79**, 377 (1961).

[37] Domenech, O. *et al.* Interactions of oritavancin, a new lipoglycopeptide derived from vancomycin, with phospholipid bilayers: effect on membrane permeability and nanoscale lipid membrane organization. *Biochimica et Biophysica Acta (BBA)-Biomembranes* **1788**, 1832–1840 (2009).

[38] Umezawa, H., Maeda, K., Takeuchi, T. & Okami, Y. New antibiotics, bleomycin a and b. *The Journal of antibiotics* **19**, 200–209 (1966).

[39] Miao, V. *et al.* Daptomycin biosynthesis in streptomyces roseosporus: cloning and analysis of the gene cluster and revision of peptide stereochemistry. *Microbiology* **151**, 1507–1523 (2005).

[40] Murthy, M. R., Mohan, E. & Sadhukhan, A. Cyclosporin-a production by tolypocladium inflatum using solid state fermentation. *Process Biochemistry* **34**, 269–280 (1999).

[41] Waksman, S. A. & Woodruff, H. B. Bacteriostatic and bactericidal substances produced by a soil actinomyces. *Proceedings of the society for Experimental Biology and Medicine* **45**, 609–614 (1940).

[42] Ueda, H. *et al.* Fr901228, a novel antitumor bicyclic depsipeptide produced by chromobacterium violaceum no. 968. *The Journal of antibiotics* **47**, 301–310 (1994).

[43] Agrawal, S., Adholeya, A. & Deshmukh, S. K. The pharmacological potential of non-ribosomal peptides from marine sponge and tunicates. *Frontiers in pharmacology* **7**, 333 (2016).

[44] Sobell, H. M. Actinomycin and dna transcription. *Proceedings of the National Academy of Sciences* **82**, 5328–5331 (1985).

[45] Organization, W. H. *et al.* World health organization model list of essential medicines: 21st list 2019. Tech. Rep., World Health Organization (2019).

[46] Dorr, R. T. Bleomycin pharmacology: mechanism of action and resistance, and clinical pharmacokinetics. In *Seminars in oncology*, vol. 19, 3–8 (1992).

[47] Sikic, B. I., Rozencweig, M. & Carter, S. K. *Bleomycin chemotherapy* (Elsevier, 2016).

[48] Gordon, E. M., Sankhala, K. K., Chawla, N. & Chawla, S. P. Trabectedin for soft tissue sarcoma: current status and future perspectives. *Advances in therapy* **33**, 1055–1071 (2016).

[49] Faulds, D., Goa, K. L. & Benfield, P. Cyclosporin. *Drugs* **45**, 953–1040 (1993).

[50] Tedesco, D. & Haragsim, L. Cyclosporine: a review. *Journal of transplantation* **2012** (2012).

[51] Morris, P. The impact of cyclosporin a on transplantation. *Advances in surgery (Chicago)* **17**, 99–127 (1984).

[52] Borel, J. & Hiestand, P. Immunomodulation: particular perspectives. In *Transplantation proceedings*, vol. 31, 1464–1471 (Citeseer, 1999).

[53] Vilcinskas, A., Jegorov, A., Landa, Z., Götz, P. & Matha, V. Effects of beauverolide l and cyclosporin a on humoral and cellular immune response of the greater wax moth, galleria mellonella. *Comparative Biochemistry and Physiology Part C: Pharmacology, Toxicology and Endocrinology* **122**, 83–92 (1999).

[54] Borowski, J., Jakoniuk, P., Jablońska, W. & Borowski, E. Effect of edeine on primary immunologic response in mice. *Archivum immunologiae et therapiae experimentalis* **23**, 195–199 (1975).

[55] Czajgucki, Z., Zimecki, M. & Andruszkiewicz, R. The immunoregulatory effects of edeine analogues in mice. *Cellular and Molecular Biology Letters* **12**, 149–161 (2007).

[56] Heifets, L., Simon, J. & Pham, V. Capreomycin is active against nonreplicating m. tuberculosis. *Annals of clinical microbiology and antimicrobials* **4**, 6 (2005).

[57] Mayer, A. M., Rodríguez, A. D., Berlinck, R. G. & Fusetani, N. Marine pharmacology in 2007–8: Marine compounds with antibacterial, anticoagulant, antifungal, anti-inflammatory, antimalarial, antiprotozoal, antituberculosis, and antiviral activities; affecting the immune and nervous system, and other miscellaneous mechanisms of action. *Comparative Biochemistry and Physiology Part C: Toxicology & Pharmacology* **153**, 191–222 (2011).

[58] Zhang, T., Li, H., Xi, H., Stanton, R. V. & Rotstein, S. H. Helm: a hierarchical notation language for complex biomolecule structure representation (2012).

[59] Seydlová, G., Čabala, R. & Svobodová, J. Surfactin-novel solutions for global issues. *Biomedical engineering, trends, research and technologies* **13**, 305–330 (2011).

[60] Razafindralambo, H., Dufour, S., Paquot, M. & Deleu, M. Thermodynamic studies of the binding interactions of surfactin analogues to lipid vesicles: Application of isothermal titration calorimetry. *Journal of thermal analysis and calorimetry* **95**, 817–821 (2009).

[61] Meena, K. R. & Kanwar, S. S. Lipopeptides as the antifungal and antibacterial agents: applications in food safety and therapeutics. *BioMed research international* **2015** (2015).

[62] Bartal, A. *et al.* Effects of different cultivation parameters on the production of surfactin variants by a bacillus subtilis strain. *Molecules* **23**, 2675 (2018).

[63] Sheldrick, G. M., Jones, P. G., Kennard, O., Williams, D. H. & Smith, G. A. Structure of vancomycin and its complex with acetyl-d-alanyl-d-alanine. *Nature* **271**, 223 (1978).

[64] Nieto, M. & Perkins, H. R. Modifications of the acyl-d-alanyl-d-alanine terminus affecting complex-formation with vancomycin. *Biochemical Journal* **123**, 789 (1971).

[65] Van Bambeke, F., Van Laethem, Y., Courvalin, P. & Tulkens, P. M. Glycopeptide antibiotics. *Drugs* **64**, 913–936 (2004).

[66] Mackay, J. P., Gerhard, U., Beauregard, D. A., Maplestone, R. A. & Williams, D. H. Dissection of the contributions toward dimerization of glycopeptide antibiotics. *Journal of the American Chemical Society* **116**, 4573–4580 (1994).

[67] Gerhard, U., Mackay, J. P., Maplestone, R. A. & Williams, D. H. The role of the sugar and chlorine substituents in the dimerization of vancomycin antibiotics. *Journal of the American Chemical Society* **115**, 232–237 (1993).

[68] Beauregard, D. A., Williams, D. H., Gwynn, M. N. & Knowles, D. Dimerization and membrane anchors in extracellular targeting of vancomycin group antibiotics. *Antimicrobial agents and chemotherapy* **39**, 781–785 (1995).

[69] Mackay, J. P. *et al.* Glycopeptide antibiotic activity and the possible role of dimerization: a model for biological signaling. *Journal of the American Chemical Society* **116**, 4581–4590 (1994).

[70] Chen, Y. *et al.* Methylated actinomycin d, a novel actinomycin d analog induces apoptosis in hepg2 cells through fas-and mitochondria-mediated pathways. *Molecular carcinogenesis* **52**, 983–996 (2013).

[71] Zhang, B.-z. *et al.* In vitro and in vivo antitumor effects of novel actinomycin d analogs with amino acid substituted in the cyclic depsipeptides. *Peptides* **31**, 568–573 (2010).

[72] Oh, S. U., Yun, B. S., Lee, S. J. & Yoo, I. D. Structures and biological activities of novel antibiotic peptaibols neoatroviridins ad from trichoderma atroviride f80317. *Journal of microbiology and biotechnology* **15**, 384–387 (2005).

[73] Daniel, J. F. d. S. & Rodrigues Filho, E. Peptaibols of trichoderma. *Natural product reports* **24**, 1128–1141 (2007).

[74] Das, S. *et al.* Enhancing the antimicrobial activity of alamethicin f50/5 by incorporating n-terminal hydrophobic triazole substituents. *Chemistry–A European Journal* **23**, 17964–17972 (2017).

[75] Ben Haj Salah, K., Legrand, B., Das, S., Martinez, J. & Inguimbert, N. Straightforward strategy to substitute amide bonds by 1, 2, 3-triazoles in peptaibols analogs using aibψ [tz]-xaa dipeptides. *Peptide Science* **104**, 611–621 (2015).

[76] Conti, E., Stachelhaus, T., Marahiel, M. A. & Brick, P. Structural basis for the activation of phenylalanine in the non-ribosomal biosynthesis of gramicidin s. *The EMBO journal* **16**, 4174–4183 (1997).

[77] Du, L., Sánchez, C., Chen, M., Edwards, D. J. & Shen, B. The biosynthetic gene cluster for the antitumor drug bleomycin from streptomyces verticillus atcc15003 supporting functional interactions between nonribosomal peptide synthetases and a polyketide synthase. *Chemistry & biology* **7**, 623–642 (2000).

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

[79] Ansari, M. Z., Yadav, G., Gokhale, R. S. & Mohanty, D. Nrps-pks: a knowledge-based resource for analysis of nrps/pks megasynthases. *Nucleic acids research* **32**, W405–W413 (2004).

[80] Li, M. H., Ung, P. M., Zajkowski, J., Garneau-Tsodikova, S. & Sherman, D. H. Automated genome mining for natural products. *BMC bioinformatics* **10**, 185 (2009).

[81] Bachmann, B. O. & Ravel, J. Methods for in silico prediction of microbial polyketide and nonribosomal peptide biosynthetic pathways from dna sequence data. *Methods in enzymology* **458**, 181–217 (2009).

[82] 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).

[83] Röttig, M. *et al.* Nrpspredictor2—a web server for predicting nrps adenylation domain specificity. *Nucleic acids research* **39**, W362–W367 (2011).

[84] Strieker, M., Tanović, A. & Marahiel, M. A. Nonribosomal peptide synthetases: structures and dynamics. *Current opinion in structural biology* **20**, 234–240 (2010).

[85] Stachelhaus, T., Hüser, A. & Marahiel, M. A. Biochemical characterization of peptidyl carrier protein (pcp), the thiolation domain of multifunctional peptide synthetases. *Chemistry & biology* **3**, 913–921 (1996).

[86] Ehmann, D. E., Shaw-Reid, C. A., Losey, H. C. & Walsh, C. T. The entf and ente adenylation domains of escherichia coli enterobactin synthetase: sequestration and selectivity in acyl-amp transfers to thiolation domain cosubstrates. *Proceedings of the National Academy of Sciences* **97**, 2509–2514 (2000).

[87] Marahiel, M. A. Working outside the protein-synthesis rules: insights into non-ribosomal peptide synthesis. *Journal of peptide science: an official publication of the European Peptide Society* **15**, 799–807 (2009).

[88] Sieber, S. A. & Marahiel, M. A. Molecular mechanisms underlying nonribosomal peptide synthesis: approaches to new antibiotics. *Chemical reviews* **105**, 715–738 (2005).

[89] Rausch, C., Hoof, I., Weber, T., Wohlleben, W. & Huson, D. H. Phylogenetic analysis of condensation domains in nrps sheds light on their functional evolution. *BMC evolutionary biology* **7**, 78 (2007).

[90] Hur, G. H., Vickery, C. R. & Burkart, M. D. Explorations of catalytic domains in non-ribosomal peptide synthetase enzymology. *Natural product reports* **29**, 1074–1098 (2012).

[91] Finking, R. & Marahiel, M. A. Biosynthesis of nonribosomal peptides. *Annu. Rev. Microbiol.* **58**, 453–488 (2004).

[92] Samel, S. A., Wagner, B., Marahiel, M. A. & Essen, L.-O. The thioesterase domain of the fengycin biosynthesis cluster: a structural base for the macrocyclization of a non-ribosomal lipopeptide. *Journal of molecular biology* **359**, 876–889 (2006).

[93] Linne, U., Doekel, S. & Marahiel, M. A. Portability of epimerization domain and role of peptidyl carrier protein on epimerization activity in nonribosomal peptide synthetases. *Biochemistry* **40**, 15824–15834 (2001).

[94] Balibar, C. J., Vaillancourt, F. H. & Walsh, C. T. Generation of d amino acid residues in assembly of arthrofactin by dual condensation/epimerization domains. *Chemistry & biology* **12**, 1189–1200 (2005).

[95] Hoffmann, K., Schneider-Scherzer, E., Kleinkauf, H. & Zocher, R. Purification and characterization of eucaryotic alanine racemase acting as key enzyme in cyclosporin biosynthesis. *Journal of Biological Chemistry* **269**, 12710–12714 (1994).

[96] Weber, G., Schörgendorfer, K., Schneider-Scherzer, E. & Leitner, E. The peptide synthetase catalyzing cyclosporine production in tolypocladium niveum is encoded by a giant 45.8-kilobase open reading frame. *Current genetics* **26**, 120–125 (1994).

[97] Miller, D. J. *et al.* Crystal complexes of a predicted s-adenosylmethionine-dependent methyltransferase reveal a typical adomet binding domain and a substrate recognition domain. *Protein Science* **12**, 1432–1442 (2003).

[98] Martin, J. L. & McMillan, F. M. Sam (dependent) i am: the s-adenosylmethionine-dependent methyltransferase fold. *Current opinion in structural biology* **12**, 783–793 (2002).

[99] Patel, H. M. & Walsh, C. T. In vitro reconstitution of the pseudomonas aeruginosa nonribosomal peptide synthesis of pyochelin: characterization of backbone tailoring thiazoline reductase and n-methyltransferase activities. *Biochemistry* **40**, 9023–9031 (2001).

[100] Schoenafinger, G., Schracke, N., Linne, U. & Marahiel, M. A. Formylation domain: an essential modifying enzyme for the nonribosomal biosynthesis of linear gramicidin. *Journal of the American Chemical Society* **128**, 7406–7407 (2006).

[101] Dorrestein, P. C., Yeh, E., Garneau-Tsodikova, S., Kelleher, N. L. & Walsh, C. T. Dichlorination of a pyrrolyl-s-carrier protein by fadh2-dependent halogenase plta during pyoluteorin biosynthesis. *Proceedings of the National Academy of Sciences* **102**, 13843–13848 (2005).

[102] Galonić, D. P., Vaillancourt, F. H. & Walsh, C. T. Halogenation of unactivated carbon centers in natural product biosynthesis: trichlorination of leucine during barbamide biosynthesis. *Journal of the American Chemical Society* **128**, 3900–3901 (2006).

[103] Pelzer, S. *et al.* Identification and analysis of the balhimycin biosynthetic gene cluster and its use for manipulating glycopeptide biosynthesis in amycolatopsis mediterranei dsm5908. *Antimicrobial agents and chemotherapy* **43**, 1565–1573 (1999).

[104] Weber, T. & Kim, H. U. The secondary metabolite bioinformatics portal: Computational tools to facilitate synthetic biology of secondary metabolite production. *Synthetic and Systems Biotechnology* **1**, 69–79 (2016).

[105] Eddy, S. R. Accelerated profile hmm searches. *PLoS computational biology* **7** (2011).

[106] Starcevic, A. *et al.* Clustscan: an integrated program package for the semi-automatic annotation of modular biosynthetic gene clusters and in silico prediction of novel chemical structures. *Nucleic acids research* **36**, 6882–6892 (2008).

[107] Diminic, J. *et al.* Databases of the thiotemplate modular systems (csdb) and their in silico recombinants (r-csdb). *Journal of industrial microbiology & biotechnology* **40**, 653–659 (2013).

[108] Blin, K. *et al.* antismash 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic acids research* **47**, W81–W87 (2019).

[109] Medema, M. H. *et al.* antismash: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic acids research* **39**, W339–W346 (2011).

[110] Blin, K. *et al.* antismash 2.0—a versatile platform for genome mining of secondary metabolite producers. *Nucleic acids research* **41**, W204–W212 (2013).

[111] Weber, T. *et al.* antismash 3.0—a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic acids research* **43**, W237–W243 (2015).

[112] Blin, K. *et al.* antismash 4.0—improvements in chemistry prediction and gene cluster boundary identification. *Nucleic acids research* **45**, W36–W41 (2017).

[113] Medema, M. H. *et al.* Pep2path: automated mass spectrometry-guided genome mining of peptidic natural products. *PLoS computational biology* **10** (2014).

[114] Khaldi, N. *et al.* Smurf: genomic mapping of fungal secondary metabolite clusters. *Fungal Genetics and Biology* **47**, 736–741 (2010).

[115] Skinnider, M. A. *et al.* Genomes to natural products prediction informatics for secondary metabolomes (prism). *Nucleic acids research* **43**, 9645–9662 (2015).

[116] 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).

[117] Johnston, C. W. *et al.* An automated genomes-to-natural products platform (gnp) for the discovery of modular natural products. *Nature communications* **6**, 1–11 (2015).

[118] Conway, K. R. & Boddy, C. N. Clustermine360: a database of microbial pks/nrps biosynthesis. *Nucleic acids research* **41**, D402–D407 (2012).

[119] Medema, M. H. *et al.* Minimum information about a biosynthetic gene cluster. *Nature chemical biology* **11**, 625–631 (2015).

[120] Kautsar, S. A. *et al.* Mibig 2.0: a repository for biosynthetic gene clusters of known function. *Nucleic acids research* **48**, D454–D458 (2020).

[121] Hadjithomas, M. *et al.* Img-abc: a knowledge base to fuel discovery of biosynthetic gene clusters and novel secondary metabolites. *MBio* **6**, e00932–15 (2015).

[122] Hadjithomas, M. *et al.* Img-abc: new features for bacterial secondary metabolism analysis and targeted biosynthetic gene cluster discovery in thousands of microbial genomes. *Nucleic acids research* **45**, D560–D565 (2017).

[123] Palaniappan, K. *et al.* Img-abc v. 5.0: an update to the img/atlas of biosynthetic gene clusters knowledgebase. *Nucleic acids research* **48**, D422–D430 (2020).

[124] Walsh, C. T. & Fischbach, M. A. Natural products version 2.0: connecting genes to molecules. *Journal of the American Chemical Society* **132**, 2469–2493 (2010).

[125] 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).

[126] Abdo, A., Caboche, S., Leclère, V., Jacques, P. & Pupin, M. A new fingerprint to predict nonribosomal peptides activity. *Journal of computer-aided molecular design* **26**, 1187–1194 (2012).

[127] Lewell, X. Q., Judd, D. B., Watson, S. P. & Hann, M. M. Recap retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combina-

torial chemistry. *Journal of chemical information and computer sciences* **38**, 511–522 (1998).

[128] Steinbeck, C. *et al.* The chemistry development kit (cdk): An open-source java library for chemo-and bioinformatics. *Journal of chemical information and computer sciences* **43**, 493–500 (2003).

[129] Steinbeck, C. *et al.* Recent developments of the chemistry development kit (cdk)-an open-source java library for chemo-and bioinformatics. *Current pharmaceutical design* **12**, 2111–2120 (2006).

[130] Willighagen, E. L. *et al.* The chemistry development kit (cdk) v2. 0: atom typing, depiction, molecular formulas, and substructure searching. *Journal of cheminformatics* **9**, 33 (2017).

[131] Inc, D. Daylight theory: Smarts-a language for describing molecular patterns; 2018.

[132] Siani, M. A., Weininger, D. & Blaney, J. M. Chuckles: a method for representing and searching peptide and peptoid sequences on both monomer and atomic levels. *Journal of chemical information and computer sciences* **34**, 588–593 (1994).

[133] Ghersi, D. & Singh, M. molblocks: decomposing small molecule sets and uncovering enriched fragments. *Bioinformatics* **30**, 2081–2083 (2014).

[134] Degen, J., Wegscheid-Gerlach, C., Zaliani, A. & Rarey, M. On the art of compiling and using'drug-like'chemical fragment spaces. *ChemMedChem: Chemistry Enabling Drug Discovery* **3**, 1503–1507 (2008).

[135] Bouslimani, A., Sanchez, L. M., Garg, N. & Dorrestein, P. C. Mass spectrometry of natural products: current, emerging and future technologies. *Natural product reports* **31**, 718–729 (2014).

[136] Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F. & Whitehouse, C. M. Electrospray ionization for mass spectrometry of large biomolecules. *Science* **246**, 64–71 (1989).

[137] Karas, M. & Hillenkamp, F. Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Analytical chemistry* **60**, 2299–2301 (1988).

[138] Wolff, M. & Stephens, W. A pulsed mass spectrometer with time dispersion. *Review of Scientific Instruments* **24**, 616–617 (1953).

[139] Wolfgang, P. & Helmut, S. Apparatus for separating charged particles of different specific charges (1960). US Patent 2,939,952.

[140] Paul, W. & Steinwedel, H. A new mass spectrometer without a magnetic field. *Zeitschrift fuer Naturforschung (West Germany) Divided into Z. Nautrforsch., A, and Z. Naturforsch., B: Anorg. Chem., Org. Chem., Biochem., Biophys.,* **8** (1953).

[141] Makarov, A. Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis. *Analytical chemistry* **72**, 1156–1162 (2000).

[142] Comisarow, M. B. & Marshall, A. G. Resolution-enhanced fourier transform ion cyclotron resonance spectroscopy. *The Journal of Chemical Physics* **62**, 293–295 (1975).

[143] Park, K. H. *et al.* Simultaneous molecular formula determinations of natural compounds in a plant extract using 15 t fourier transform ion cyclotron resonance mass spectrometry. *Plant methods* **9**, 15 (2013).

[144] Motti, C. A., Freckelton, M. L., Tapiolas, D. M. & Willis, R. H. Fticr-ms and lc-uv/ms-spe-nmr applications for the rapid dereplication of a crude extract from the sponge ianthella flabelliformis. *Journal of natural products* **72**, 290–294 (2009).

[145] Le Ven, J. *et al.* Identification of the environmental neurotoxins annonaceous acetogenins in an annona cherimolia mill. alcoholic beverage using hplc-esi-ltq-orbitrap. *Journal of agricultural and food chemistry* **62**, 8696–8704 (2014).

[146] Le Ven, J. *et al.* Comprehensive characterization of annonaceous acetogenins within a complex extract by hplc-esi-ltq-orbitrap® using post-column lithium infusion. *Journal of Mass Spectrometry* **47**, 1500–1509 (2012).

[147] Williams, S. M. & Brodbelt, J. S. Ms n characterization of protonated cyclic peptides and metal complexes. *Journal of the American Society for Mass Spectrometry* **15**, 1039–1054 (2004).

[148] Rabi, I. I., Zacharias, J. R., Millman, S. & Kusch, P. A new method of measuring nuclear magnetic moment. *Physical Review* **53**, 318 (1938).

[149] Bjerrum, J. T. & Bjerrum. *Metabonomics* (Springer, 2015).

[150] Pauli, G. F., Kuczkowiak, U. & Nahrstedt, A. Solvent effects in the structure dereplication of caffeoyl quinic acids. *Magnetic Resonance in Chemistry* **37**, 827–836 (1999).

[151] Hubert, J., Nuzillard, J.-M. & Renault, J.-H. Dereplication strategies in natural product research: How many tools and methodologies behind the same concept? *Phytochemistry reviews* **16**, 55–95 (2017).

[152] Gruetter, R. *et al.* Resolution improvements inin vivo1h nmr spectra with increased magnetic field strength. *Journal of magnetic resonance* **135**, 260–264 (1998).

[153] Keun, H. C. *et al.* Cryogenic probe 13c nmr spectroscopy of urine for metabonomic studies. *Analytical chemistry* **74**, 4588–4593 (2002).

[154] Grimes, J. H. & O'Connell, T. M. The application of micro-coil nmr probe technology to metabolomics of urine and serum. *Journal of biomolecular NMR* **49**, 297–305 (2011).

[155] Ardenkjær-Larsen, J. H. *et al.* Increase in signal-to-noise ratio of> 10,000 times in liquid-state nmr. *Proceedings of the National Academy of Sciences* **100**, 10158–10163 (2003).

[156] Roepstorff, P. & Fohlman, J. Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomedical mass spectrometry* **11**, 601–601 (1984).

[157] Biemann, K. Contributions of mass spectrometry to peptide and protein structure. *Biomedical & environmental mass spectrometry* **16**, 99–111 (1988).

[158] Niedermeyer, T. H. & Strohalm, M. mmass as a software tool for the annotation of cyclic peptide tandem mass spectra. *PloS one* **7**, e44913 (2012).

[159] Liu, W.-T. *et al.* Interpretation of tandem mass spectra obtained from cyclic nonribosomal peptides. *Analytical chemistry* **81**, 4200–4209 (2009).

[160] Johnson, R. S., Martin, S. A., Biemann, K., Stults, J. T. & Watson, J. T. Novel fragmentation process of peptides by collision-induced decomposition in a tandem mass spectrometer: differentiation of leucine and isoleucine. *Analytical chemistry* **59**, 2621–2625 (1987).

[161] Zhokhov, S. S., Kovalyov, S. V., Samgina, T. Y. & Lebedev, A. T. An ethcd-based method for discrimination of leucine and isoleucine residues in tryptic peptides. *Journal of The American Society for Mass Spectrometry* **28**, 1600–1611 (2017).

[162] Domon, B. & Costello, C. E. A systematic nomenclature for carbohydrate fragmentations in fab-ms/ms spectra of glycoconjugates. *Glycoconjugate journal* **5**, 397–409 (1988).

[163] Cooks, R. G. Special feature: Historical. collision-induced dissociation: Readings and commentary. *Journal of Mass Spectrometry* **30**, 1215–1221 (1995).

[164] Wells, J. M. & McLuckey, S. A. Collision-induced dissociation (cid) of peptides and proteins. *Methods in enzymology* **402**, 148–185 (2005).

[165] Paizs, B. & Suhai, S. Fragmentation pathways of protonated peptides. *Mass spectrometry reviews* **24**, 508–548 (2005).

[166] Olsen, J. V. *et al.* Higher-energy c-trap dissociation for peptide modification analysis. *Nature methods* **4**, 709 (2007).

[167] Brodbelt, J. S. Ion activation methods for peptides and proteins. *Analytical chemistry* **88**, 30–51 (2015).

[168] Zubarev, R. A., Kelleher, N. L. & McLafferty, F. W. Electron capture dissociation of multiply charged protein cations. a nonergodic process. *Journal of the American Chemical Society* **120**, 3265–3266 (1998).

[169] Syka, J. E., Coon, J. J., Schroeder, M. J., Shabanowitz, J. & Hunt, D. F. Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proceedings of the National Academy of Sciences* **101**, 9528–9533 (2004).

[170] Quan, L. & Liu, M. Cid, etd and hcd fragmentation to study protein post-translational modifications. *Mod Chem Appl* **1**, 1–5 (2013).

[171] Scott, N. E. *et al.* Simultaneous glycan-peptide characterization using hydrophilic interaction chromatography and parallel fragmentation by cid, higher energy collisional dissociation, and electron transfer dissociation ms applied to the n-linked glycoproteome of campylobacter jejuni. *Molecular & cellular proteomics* **10**, M000031–MCP201 (2011).

[172] Fung, Y. E., Adams, C. M. & Zubarev, R. A. Electron ionization dissociation of singly and multiply charged peptides. *Journal of the American Chemical Society* **131**, 9977–9985 (2009).

[173] Wills, R. H. & O'connor, P. B. Structural characterization of actinomycin d using multiple ion isolation and electron induced dissociation. *Journal of The American Society for Mass Spectrometry* **25**, 186–195 (2013).

[174] Barbacci, D. C. & Russell, D. H. Sequence and side-chain specific photofragment (193 nm) ions from protonated substance p by matrix-assisted laser desorption ionization time-of-flight mass spectrometry. *Journal of the American Society for Mass Spectrometry* **10**, 1038–1040 (1999).

[175] Thompson, M. S., Cui, W. & Reilly, J. P. Factors that impact the vacuum ultraviolet photofragmentation of peptide ions. *Journal of the American Society for Mass Spectrometry* **18**, 1439–1452 (2007).

[176] Agarwal, A., Diedrich, J. K. & Julian, R. R. Direct elucidation of disulfide bond partners using ultraviolet photodissociation mass spectrometry. *Analytical chemistry* **83**, 6455–6458 (2011).

[177] Tao, Y., Quebbemann, N. R. & Julian, R. R. Discriminating d-amino acid-containing peptide epimers by radical-directed dissociation mass spectrometry. *Analytical chemistry* **84**, 6814–6820 (2012).

[178] Tao, Y. & Julian, R. R. Identification of amino acid epimerization and isomerization in crystallin proteins by tandem lc-ms. *Analytical chemistry* **86**, 9733–9741 (2014).

[179] McCormack, A. L., Jones, J. L. & Wysocki, V. H. Surface-induced dissociation of multiply protonated peptides. *Journal of the American Society for Mass Spectrometry* **3**, 859–862 (1992).

[180] Wysocki, V. H., Joyce, K. E., Jones, C. M. & Beardsley, R. L. Surface-induced dissociation of small molecules, peptides, and non-covalent protein complexes. *Journal of the American Society for Mass Spectrometry* **19**, 190–208 (2008).

[181] Berkout, V. D. Fragmentation of protonated peptide ions via interaction with metastable atoms. *Analytical chemistry* **78**, 3055–3061 (2006).

[182] Hoffmann, W. D. & Jackson, G. P. Charge transfer dissociation (ctd) mass spectrometry of peptide cations using kiloelectronvolt helium cations. *Journal of The American Society for Mass Spectrometry* **25**, 1939–1943 (2014).

[183] Budnik, B. A., Haselmann, K. F. & Zubarev, R. A. Electron detachment dissociation of peptide di-anions: an electron–hole recombination phenomenon. *Chemical Physics Letters* **342**, 299–302 (2001).

[184] Kjeldsen, F. *et al.* Cα c backbone fragmentation dominates in electron detachment dissociation of gas-phase polypeptide polyanions. *Chemistry–A European Journal* **11**, 1803–1812 (2005).

[185] Coon, J. J., Shabanowitz, J., Hunt, D. F. & Syka, J. E. Electron transfer dissociation of peptide anions. *Journal of the American Society for Mass Spectrometry* **16**, 880–882 (2005).

[186] Antoine, R. *et al.* Photo-induced formation of radical anion peptides. electron photodetachment dissociation experiments. *Rapid Communications in Mass Spectrometry: An International Journal Devoted to the Rapid Dissemination of Up-to-the-Minute Research in Mass Spectrometry* **21**, 265–268 (2007).

[187] Larraillet, V., Antoine, R., Dugourd, P. & Lemoine, J. Activated-electron photodetachment dissociation for the structural characterization of protein polyanions. *Analytical chemistry* **81**, 8410–8416 (2009).

[188] Marinlit. http://pubs.rsc.org/marinlit/, note = Accessed 14 Apr 2020.

[189] Blunt, J., Munro, M. & Laatsch, H. Antimarin database (2006).

[190] Gabrielson, S. W. Scifinder. *Journal of the Medical Library Association: JMLA* **106**, 588 (2018).

[191] RÖmpp online. https://roempp.thieme.de/home/keywordoftheweek, note = Accessed 15 Apr 2020.

[192] Reaxys. https://www.reaxys.com/, note = Accessed 15 Apr 2020.

[193] Kim, S. *et al.* Pubchem 2019 update: improved access to chemical data. *Nucleic acids research* **47**, D1102–D1109 (2019).

[194] Pence, H. E. & Williams, A. Chemspider: an online chemical information resource (2010).

[195] Hastings, J. *et al.* Chebi in 2016: Improved services and an expanding collection of metabolites. *Nucleic acids research* **44**, D1214–D1219 (2016).

[196] Sterling, T. & Irwin, J. J. Zinc 15–ligand discovery for everyone. *Journal of chemical information and modeling* **55**, 2324–2337 (2015).

[197] Banerjee, P. *et al.* Super natural ii—a database of natural products. *Nucleic acids research* **43**, D935–D939 (2015).

[198] Gu, J. *et al.* Use of natural products as chemical library for drug discovery and network pharmacology. *PloS one* **8** (2013).

[199] Allard, P.-M. *et al.* Integration of molecular networking and in-silico ms/ms fragmentation for natural products dereplication. *Analytical chemistry* **88**, 3317–3323 (2016).

[200] Zeng, X. *et al.* Npass: natural product activity and species source database for natural product research, discovery and tool development. *Nucleic acids research* **46**, D1217–D1222 (2018).

[201] Van Santen, J. A. *et al.* The natural products atlas: an open access knowledge base for microbial natural products discovery. *ACS Central Science* **5**, 1824–1833 (2019).

[202] Tomiki, T. *et al.* Riken natural products encyclopedia (riken npedia), a chemical database of riken natural products depository (riken npdepo). *J Comput Aid Chem* **7**, 157–162 (2006).

[203] Maeda, M. H. & Kondo, K. Three-dimensional structure database of natural metabolites (3dmet): a novel database of curated 3d structures. *Journal of chemical information and modeling* **53**, 527–533 (2013).

[204] Nakamura, K. *et al.* Knapsack-3d: a three-dimensional structure database of plant metabolites. *Plant and Cell Physiology* **54**, e4–e4 (2013).

[205] Klementz, D. *et al.* Streptomedb 2.0—an extended resource of natural products produced by streptomycetes. *Nucleic acids research* **44**, D509–D514 (2016).

[206] Ntie-Kang, F. *et al.* Afrodb: a select highly potent and diverse natural product library from african medicinal plants. *PLoS One* **8** (2013).

[207] Pilon, A. C. *et al.* Nubbe db: an updated database to uncover chemical and biological information from brazilian biodiversity. *Scientific reports* **7**, 1–12 (2017).

[208] Feunang, Y. D. *et al.* Classyfire: automated chemical classification with a comprehensive, computable taxonomy. *Journal of cheminformatics* **8**, 61 (2016).

[209] Johnson, S. G. Nist standard reference database 1a v17. https://mona.fiehnlab.ucdavis.edu/. Accessed 15 Apr 2020.

[210] Msforid. https://msforid.com/. Accessed 15 Apr 2020.

[211] Horai, H. *et al.* Massbank: a public repository for sharing mass spectral data for life sciences. *Journal of mass spectrometry* **45**, 703–714 (2010).

[212] Massbank of north america (mona). https://mona.fiehnlab.ucdavis.edu/. Accessed 15 Apr 2020.

[213] Massbank | european massbank (norman massbank) mass spectral database. https://massbank.eu/MassBank/. Accessed 15 Apr 2020.

[214] Massbank | mssj massbank mass spectral database. http://www.massbank.jp/. Accessed 15 Apr 2020.

[215] Guijas, C. *et al.* Metlin: a technology platform for identifying knowns and unknowns. *Analytical chemistry* **90**, 3156–3164 (2018).

[216] mzcloud. https://www.mzcloud.org/. Accessed 15 Apr 2020.

[217] Wang, M. *et al.* Sharing and community curation of mass spectrometry data with global natural products social molecular networking. *Nature biotechnology* **34**, 828–837 (2016).

[218] Sawada, Y. *et al.* Riken tandem mass spectral database (respect) for phytochemicals: a plant-specific ms/ms-based data resource and database. *Phytochemistry* **82**, 38–45 (2012).

[219] Wishart, D. S. *et al.* Hmdb 4.0: the human metabolome database for 2018. *Nucleic acids research* **46**, D608–D617 (2018).

[220] Ramirez-Gaona, M. *et al.* Ymdb 2.0: a significantly expanded version of the yeast metabolome database. *Nucleic acids research* **45**, D440–D445 (2017).

[221] Pluskal, T., Uehara, T. & Yanagida, M. Highly accurate chemical formula prediction tool utilizing high-resolution mass spectra, ms/ms fragmentation, heuristic rules, and isotope pattern matching. *Analytical chemistry* **84**, 4396–4403 (2012).

[222] Kind, T. & Fiehn, O. Seven golden rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC bioinformatics* **8**, 105 (2007).

[223] Meija, J. *et al.* Isotopic compositions of the elements 2013 (iupac technical report). *Pure and Applied Chemistry* **88**, 293–306 (2016).

[224] Meija, J. *et al.* Atomic weights of the elements 2013 (iupac technical report). *Pure and Applied Chemistry* **88**, 265–291 (2016).

[225] Kendrick, E. A mass scale based on ch2= 14.0000 for high resolution mass spectrometry of organic compounds. *Analytical Chemistry* **35**, 2146–2154 (1963).

[226] Marshall, A. G. & Rodgers, R. P. Petroleomics: the next grand challenge for chemical analysis. *Accounts of chemical research* **37**, 53–59 (2004).

[227] Sato, H., Nakamura, S., Teramoto, K. & Sato, T. Structural characterization of polymers by maldi spiral-tof mass spectrometry combined with kendrick

mass defect analysis. *Journal of The American Society for Mass Spectrometry* **25**, 1346–1355 (2014).

[228] Novák, J., Lemr, K., Schug, K. A. & Havlíček, V. Cyclobranch: de novo sequencing of nonribosomal peptides from accurate product ion mass spectra. *Journal of The American Society for Mass Spectrometry* **26**, 1780–1786 (2015).

[229] Ibrahim, A. *et al.* Dereplicating nonribosomal peptides using an informatic search algorithm for natural products (isnap) discovery. *Proceedings of the National Academy of Sciences* **109**, 19196–19201 (2012).

[230] Mohimani, H. *et al.* Dereplication of peptidic natural products through database search of mass spectra. *Nature chemical biology* **13**, 30 (2017).

[231] Mohimani, H. *et al.* Dereplication of microbial metabolites through database search of mass spectra. *Nature communications* **9**, 1–12 (2018).

[232] Hufsky, F., Scheubert, K. & Böcker, S. Computational mass spectrometry for small-molecule fragmentation. *TrAC Trends in Analytical Chemistry* **53**, 41–48 (2014).

[233] Wolf, S., Schmidt, S., Müller-Hannemann, M. & Neumann, S. In silico fragmentation for computer assisted identification of metabolite mass spectra. *BMC bioinformatics* **11**, 148 (2010).

[234] Böcker, S. & Rasche, F. Towards de novo identification of metabolites by analyzing tandem mass spectra. *Bioinformatics* **24**, i49–i55 (2008).

[235] Venkataraghavan, R., McLafferty, F. & Van Lear, G. Computer-aided interpretation of mass spectra. *Organic Mass Spectrometry* **2**, 1–15 (1969).

[236] Heinonen, M., Shen, H., Zamboni, N. & Rousu, J. Metabolite identification and molecular fingerprint prediction through machine learning. *Bioinformatics* **28**, 2333–2341 (2012).

[237] Tagirdzhanov, A. M., Shlemov, A. & Gurevich, A. Nps: scoring and evaluating the statistical significance of peptidic natural product–spectrum matches. *Bioinformatics* **35**, i315–i323 (2019).

[238] Yates, J. R., Eng, J. K., McCormack, A. L. & Schieltz, D. Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Analytical chemistry* **67**, 1426–1436 (1995).

[239] Fenyö, D. & Beavis, R. C. A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Analytical chemistry* **75**, 768–774 (2003).

[240] Cox, J. *et al.* Andromeda: a peptide search engine integrated into the maxquant environment. *Journal of proteome research* **10**, 1794–1805 (2011).

[241] Käll, L., Storey, J. D., MacCoss, M. J. & Noble, W. S. Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *Journal of proteome research* **7**, 29–34 (2008).

[242] Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* **57**, 289–300 (1995).

[243] Kanehisa, M. & Goto, S. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research* **28**, 27–30 (2000).

[244] Razumovskaya, J. *et al.* A computational method for assessing peptide-identification reliability in tandem mass spectrometry analysis with sequest. *Proteomics* **4**, 961–969 (2004).

[245] Frank, A. M. Predicting intensity ranks of peptide fragment ions. *Journal of proteome research* **8**, 2226–2240 (2009).

[246] Kim, S. & Pevzner, P. A. Ms-gf+ makes progress towards a universal database search tool for proteomics. *Nature communications* **5**, 5277 (2014).

[247] 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).

[248] Novák, J. *et al.* Batch-processing of imaging or liquid-chromatography mass spectrometry datasets and de novo sequencing of polyketide siderophores. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics* **1865**, 768–775 (2017).

[249] Novák, J., Škríba, A., Zápal, J., Kuzma, M. & Havlíček, V. Cyclobranch: An open tool for fine isotope structures in conventional and product ion mass spectra. *Journal of mass spectrometry* **53**, 1097–1103 (2018).

[250] Tsugawa, H. *et al.* Hydrogen rearrangement rules: computational ms/ms fragmentation and structure elucidation using ms-finder software. *Analytical chemistry* **88**, 7946–7958 (2016).