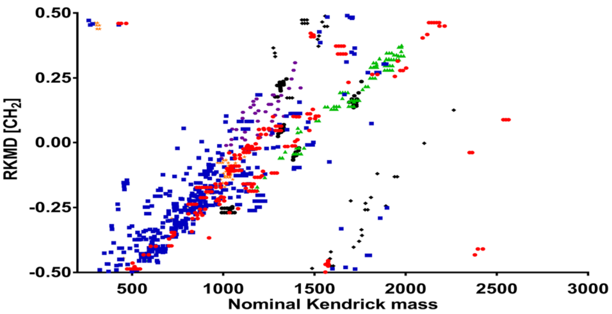
J. Am. Soc. Mass Spectrom. (2019) 30:2608–2616
DOI: 10.1007/s13361-019-02314-3

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

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 /Published Online: 28 October 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 $-0.5 < \text{KMD} < +0.5$. 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

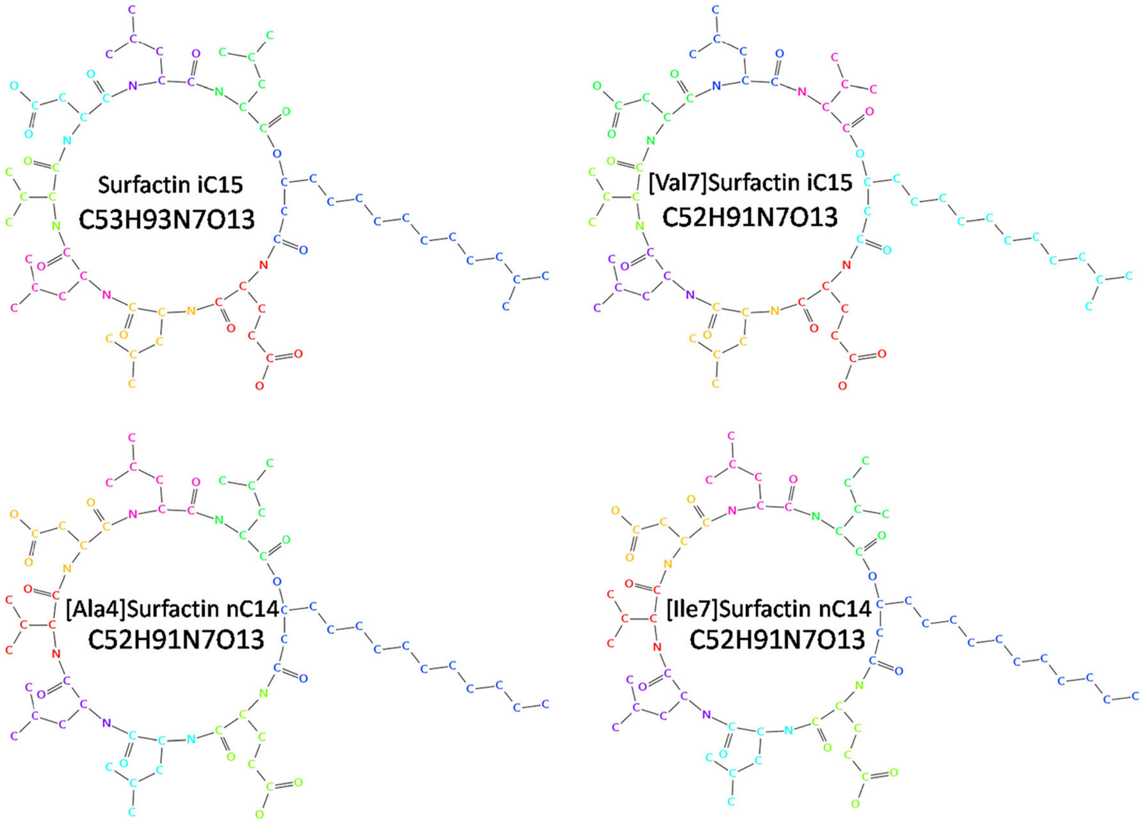


Figure 1. Chemical structure of four variants of surfactins used as models: [Ala4] nC14, [Ile7] nC14, iC15, and [Val7] iC15

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; −

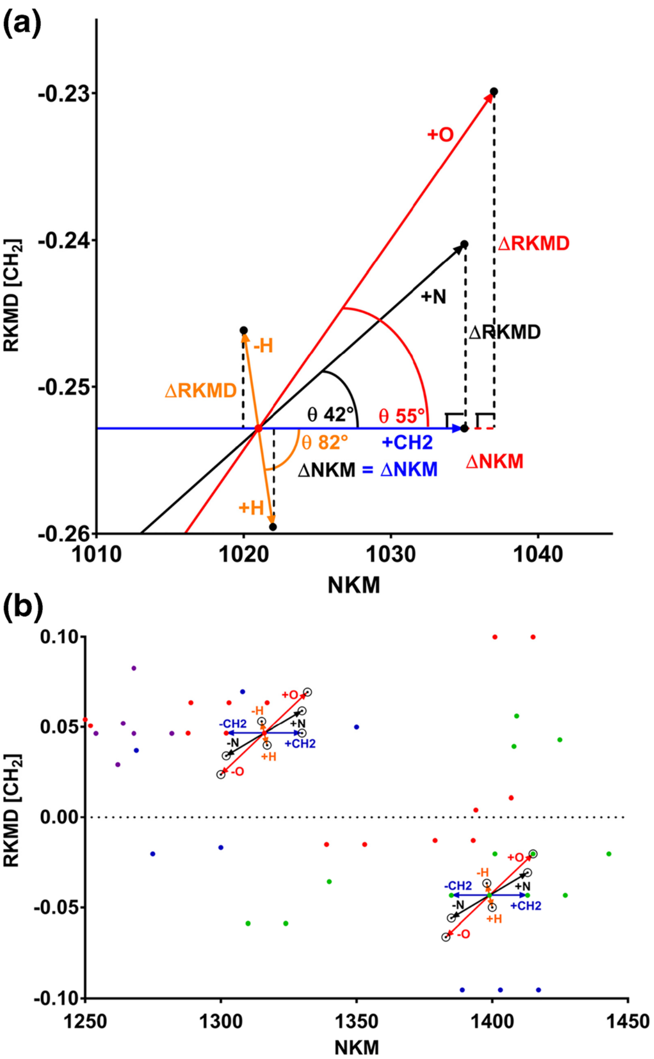


Figure 2. RKMD/NKM 2D-plot correlated to the CH$_2$ building block (**a**) illustrating the vector (defined by the theta angle and length of the vector) obtained for each increment of a given atom or atom group and (**b**) the vector mesh around the trigonometric circle enabling the connection between 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 $^{12}\mathrm{C}_1{}^{1}\mathrm{H}_1$ 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 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 (^{1}H) 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 $^{12}\mathrm{C}_1{}^{1}\mathrm{H}_1$, 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.

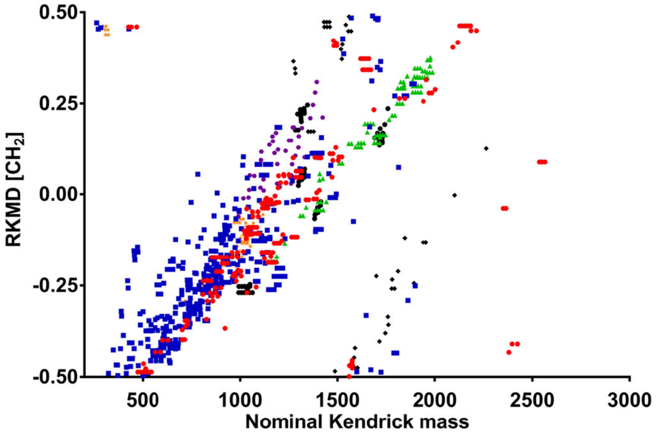


Figure 3. RKMD/NKM 2D-plot (Kendrick-based NORINE map) of all NRPs referenced in the NORINE database. The molecule classes are depicted using colors: lipopeptides (red circles), peptides (blue squares), polyketide-NRP hybrids (orange stars), chromopeptides (in black), glycopeptides (violet circles),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

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

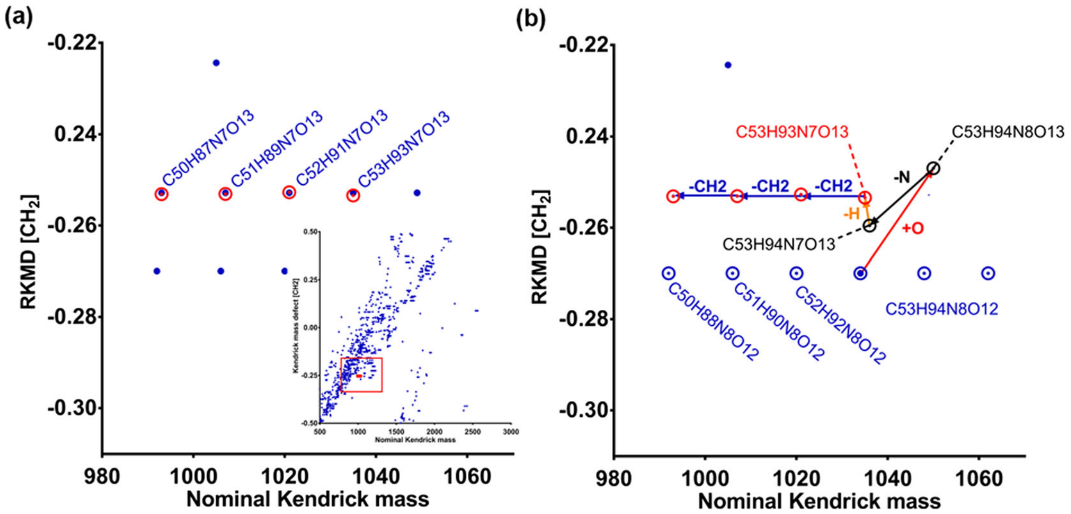


Figure 4. RKMD/NKM 2D-plot displaying the position of points corresponding to the four commercially available surfactins (**a**) on the complete Kendrick-based NORINE map and (**b**) a depleted Kendrick-based NORINE map where the surfactins of interest have been erased. In (**a**), the positioning of 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 genomics, 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