© American Society for Mass Spectrometry, 2015

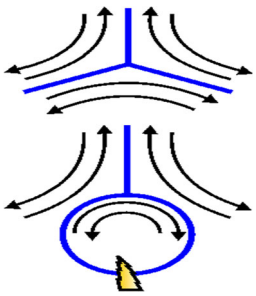
J. Am. Soc. Mass Spectrom. (2015) 26:1780–1786
DOI: 10.1007/s13361-015-1211-1

RESEARCH ARTICLE

CycloBranch: De Novo Sequencing of Nonribosomal Peptides from Accurate Product Ion Mass Spectra

Jiří Novák,[1] Karel Lemr,[1,2] Kevin A. Schug,[3] Vladimír Havlíček[1,2]

[1]Institute of Microbiology ASCR, v.v.i., Videnska 1083, CZ 14220, Prague 4, Czech Republic
[2]Regional Centre of Advanced Technologies and Materials, Department of Analytical Chemistry, Faculty of Science, Palacky University, 17 listopadu 12, 771 46, Olomouc, Czech Republic
[3]Department of Chemistry and Biochemistry, The University of Texas at Arlington, Arlington, TX 76019-0065, USA



Abstract. Nonribosomal peptides have a wide range of biological and medical applications. Their identification by tandem mass spectrometry remains a challenging task. A new open-source de novo peptide identification engine CycloBranch was developed and successfully applied in identification or detailed characterization of 11 linear, cyclic, branched, and branch-cyclic peptides. CycloBranch is based on annotated building block databases the size of which is defined by the user according to ribosomal or nonribosomal peptide origin. The current number of involved nonisobaric and isobaric building blocks is 287 and 521, respectively. Contrary to all other peptide sequencing tools utilizing either peptide libraries or peptide fragment libraries, CycloBranch represents a true de novo sequencing engine developed for accurate mass spectrometric data. It is a stand-alone and cross-platform application with a graphical and user-friendly interface; it supports mzML, mzXML, mgf, txt, and baf file formats and can be run in parallel on multiple threads. It can be downloaded for free from http://ms.biomed.cas.cz/cyclobranch/, where the User's manual and video tutorials can be found.

Keywords: De novo sequencing, Nonribosomal peptides, Linear, Cyclic, Branched, Branch-cyclic

Received: 15 December 2014/Revised: 1 June 2015/Accepted: 8 June 2015/Published Online: 21 July 2015

Introduction

Nonribosomal peptides (NRPs) are most commonly produced by bacteria and fungi using nonribosomal peptide synthetases [1]. The NRPs building blocks include hundreds of monomers, including proteinogenic and non-proteinogenic amino acids, hydroxy acids, residues having N-terminally attached fatty acid chains, N- and C-methylated residues, N-formylated residues, and many others. The popularity of NRPs stems from their diverse biological activities [2]. The structures of NRPs have been reviewed elsewhere and include linear, cyclic, branched, and branch-cyclic NRPs [3]. The structure characterization of ribosomal peptides by mass spectrometry approaches involves targeted analysis with peptide/protein or genome database searches [4, 5], applying de novo peptide sequencing algorithms [6], and/or sequence-tag methods [7]. Extending the building block databases, some of these methods can also be applied to NRPs.

Commercial software tools for fast dereplication of natural products were reviewed recently [8]. AntiMarin is a commercial database containing about 50,000 compounds from marine and terrestrial microorganisms and represents a merger of AntiBase and MarinLit databases. In addition, there are many other useful repositories (Dictionary of natural products, SciFinder, Beilstein commander, KEGG, Metlin, Human metabolome database, and a Norine database of NRPs [9]). The latter one is used by NRP-Dereplication [10] as well as iSNAP [11]. The iSNAP is not just a dereplication tool but it also provides correct identification (if the peptide is present in the database) or excellent similarity search (if an analogous peptide has previously been reported). It searches against a fragment database generated in silico from 1107 compounds that were extracted from Norine, Pubchem, *Journal of Antibiotics*, *Journal of Natural Products*, and the KEGG peptide databases. Its performance is hence limited to dereplication of already

Electronic supplementary material The online version of this article (doi:10.1007/s13361-015-1211-1) contains supplementary material, which is available to authorized users.

Correspondence to: Jiří Novák; *e-mail:* jiri.novak@biomed.cas.cz, Vladimír Havlíček; *e-mail:* vlhavlic@biomed.cas.cz

known fragments. NRPquest [12] is a tool for identification of cyclic and branch-cyclic NRPs that performs modification and mutation-tolerant searches of experimental spectra against a database of theoretical spectra of putative NRPs.

Many tools for de novo sequencing of linear but ribosomal peptides have been proposed [13]. For example, PEAKS is one of the most widely used commercial tools [14], and Lutefisk [15] and PepNovo [16] are popular open-source tools. While Lutefisk is lacking a graphical user interface (GUI), DeNovoGUI [17] has been provided for PepNovo.

A few methods have been proposed for de novo identification of cyclic NRPs from low-resolution mass spectrometry data. NRP-Tagging [10] is a tool that works with MS^3 data. A method based on multistage mass spectrometry has also been proposed [18]. However, the length of a peptide sequence must be predicted before the search. Both de novo tools report the lists of masses of building blocks on the output instead of NRP sequences. A method based on multiplex de novo sequencing has also been proposed providing much better results [19]. In addition to library matching, there are multiple annotation tools [20–22].

In our previous work, we introduced a hypertext preprocessor (PHP) script Cyclone for identification and de novo sequencing of cyclic NRPs [23]. It involved a simple text interface and required manual conversion of NRP sequence candidates from one PHP script to another. The work with peak lists was not user friendly and the search algorithm had multiple limitations. In this paper, we present CycloBranch—a stand-alone, cross-platform, and open-source de novo NRP sequence identification engine. This software has an intuitive GUI and its applications were extended from cyclic peptides to linear, branched, and branch-cyclic peptides. To the best of our knowledge, CycloBranch represents the first true de novo searching engine working both for ribosomal and nonribosomal peptides of various molecular structures and is independent of any intact peptide or fragment ion database. The application can be downloaded for free from http://ms.biomed.cas.cz/cyclobranch/ and includes a nonribosomal annotated building block database comprising 287 residues with unique elemental compositions. Including the isobaric isomers, the overall number of building blocks is 521. In addition to de novo sequencing, the database search of MS^2 spectra against an in-house collection of NRPs and Norine is supported. The identification of compounds in MS spectra can also be performed via database search.

Materials and Methods

Mass Spectra, Software, and Building Block Databases

Product ion mass spectra were collected on 12 T SolariX FTICR mass spectrometer (Bruker Daltonics, Billerica, MA, USA) equipped with a dual ESI/MALDI source. Standard peptides were infused in picomolar aqueous solutions by a linear syringe pump. Gramicidin C, Substance P, surfactin C, and valinomycin were from Sigma Aldrich, Czech Republic. The remaining standards were from an in-house peptide collection. The testing set of peptides represented 11 compounds: two linear peptides (gramicidin C and Substance P), five cyclic NRPs (beauverolide I, roseotoxin A, cyclosporin A, surfactin C, and valinomycin), two branched peptides [linearized pseudacyclin A and the synthetic peptide *N*-acetyl-ESL(KNFI)DQYG$_{NH2}$, referenced as T-peptide below], and two branch-cyclic NRPs (pseudacyclin A and pyoverdin Pa A).

Standard mass/charge error in all product ion mass spectra was better than 2 ppm. CycloBranch was implemented in C++ and utilizes integral databases with up to 287 nonredundant building blocks. All monomers in the database were annotated by reference numbers publicly available through the chemical databases ChemSpider, PubChem, Protein Data Bank (PDB), and Norine. Three databases of building blocks were used for software evaluation—a database D_{19} of proteinogenic amino acids (19 building blocks where leucine and isoleucine were not distinguished), a database D_{33} (all building blocks involved in 11 testing peptides), or D_{287} (the complete nonredundant nonribosomal database) [23]. Data processing was performed on a desktop personal computer with the processor Intel Core i7-3770 (3.4GHz), 8 GB RAM, and OS Windows 7 (64-bit). CycloBranch supports several datafile formats—mzML and mzXML (requires OpenMS 1.11 installed) [24], mgf (Mascot generic format), txt [containing a tab-separated mass-to-charge (*m/z*) ratio and intensity on each line], and baf (a native file format of Bruker Daltonics requiring prior installation of CompassXPort 3.0). A list of detected NRP sequence candidates can be exported as a comma-separated values (csv) file or as a web page. A configuration of the engine can be stored into a file (*.ini). Sample files are distributed with the engine. CycloBranch can run in parallel on multiple threads.

Software Algorithm

The algorithm for the de novo identification of NRPs covers the four main steps: (1) The construction of a de novo graph from an experimental spectrum using the database of building blocks; (2) the detection of a set of peptide sequence candidates by traversing the graph; (3) the comparison of theoretical spectra of candidates with the experimental spectrum using a selected scoring function; and (4) the reporting of NRP sequence candidates.

The graph construction is similar to a common de novo approach, but the database with hundreds of monomers is used instead of a set of genuine proteinogenic amino acids. The experimental *m/z* values are represented by nodes in the graph with two additional nodes applied. The first one corresponds to proton *m/z* value as a starting point of a *b*-series or the *m/z* value of H_3O^+ in case of a *y*-series. The last node corresponds to *m/z* value of a precursor ion (Figure 1).

Proton IUPAC mass is also used as a starting point for cyclic NRPs; both starting points are applicable for linear, branched, and branch-cyclic NRPs. An *edge* is inserted if a difference between any two *m/z* values fits to a mass of any existing

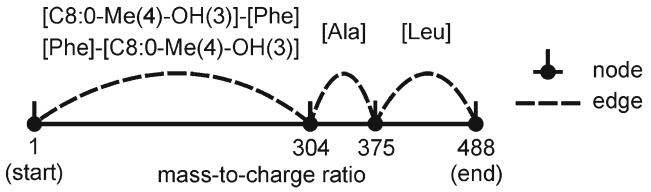


Figure 1. A de novo graph created from the experimental spectrum of *beauverolide I* (*C8:0-Me(4)-OH(3)* stands for 3-hydroxy-4-methyloctanoic acid). N-terminal b_i ion series is shown with b_1 ion missing in the spectrum

building block or their combination. The sequence of edges corresponds to linear or cyclic peptide candidates. Edges that occur because of branching can cause ambiguities in the case of branched and branch-cyclic NRPs. Thus, multiple NRP sequence candidates are generated from the sequence of edges (for details see CycloBranch manual). In the next step, theoretical mass spectra generated for NRP sequence candidates are compared with the experimental spectrum using scoring functions S_1 and S_2, which reflect the number of matched peaks and the sum of relative intensities, respectively. Finally, candidates with the best scores are reported.

Peptide Fragmentation

The generation of theoretical mass spectra depends on the type of NRP. With linear NRPs the standard N-terminal b_i and a_i as well as C-terminal y_i-ions are generated identically to linear ribosomal peptides [25]. In cyclic peptides up to k possible ring primary ring opening sites exist for a cyclic NRP with k building blocks. The initial ring cleavage may create up to k different linear ions ${}^{i\text{-}j}b_k$ (*i-j* stands for positions of building blocks between which the ring is primarily opened) [23, 26]. For example, the ions ${}^{4\text{-}3}b_4$, ${}^{3\text{-}2}b_4$, ${}^{2\text{-}1}b_4$, and ${}^{1\text{-}4}b_4$ exist when $k = 4$. Any ion ${}^{i\text{-}j}b_k$ may undergo a further fragmentation, so the spectrum of a cyclic peptide usually represents the superposition of up to k spectra corresponding to different linear sequences [10]. As the peptide is cyclic and lacking its C-terminus, *b*-ions (or other N-terminal ions) are exclusively observed [20]. Sometimes, a *b*-ion may cyclize by head-to-tail mechanism; its reopening between other two monomers and further fragmentation provides *b*-ions corresponding to a scrambled sequence of building blocks [27]. Thus, k series of b-ions must be generated in a theoretical spectrum of a cyclic NRP and ions with scrambled sequences should be considered.

By definition, a branched peptide is represented by a core and a single lateral branch. The core can be longer and has always one N-terminus and one C-terminus. The branch can potentially be shorter and is either N- or C-terminated. Thus, two series of *b*-ions (or other N-terminal ions) and two series of *y*-ions (or other C-terminal ions) may be present (Figure 2a, b). The four series of fragment ions can be generated in a theoretical spectrum of a branched NRP when a terminal modification of the branch is detected. Since modifications are commonly defined for certain termini, one can determine if the branch is N- or C-terminated. All six nonredundant series (Figure 2a, b) must be generated when the branch is not modified.

The branch-cyclic NRPs contain a ring with a lateral branch (Figure 2c, d). Their theoretical spectra are generated similarly to cyclic peptides but their ring cleavages provide branched fragment ions instead of linear ones. A theoretical spectrum of a branch-cyclic peptide is a superposition of up to *k'-2* theoretical spectra (where k' stands for the number of building blocks forming the ring) corresponding to *k'-2* different branched sequences and two spectra corresponding to linear sequences (peptide cleavage from both sides of the branch may occur). Each branched fragment gives six theoretical series of b-ions (N-terminated branch) or four series of *b*-ions and two series of *y*-ions (C-terminated branch). In analogy, each linear fragment theoretically provides two series of *b*-ions or one series of *b*-ions and one series of *y*-ions, respectively.

Results and Discussion

CycloBranch (v. 1.0.1216) has been tested on a set of 11 linear, cyclic, branched, and branch-cyclic peptides. With the linear pentadecapeptide gramicidin C, the search for a sequence tag took less than 1 s and the correct LAVVVWL tag was displayed at the positions 1–2 for D_{33} database. The same result was returned using the D_{287} unrestricted database with the same S_1 scoring criterion (Table 1). For the linear undecapeptide Substance P, the correct tag QFFGLM$_{NH2}$ was reported as the top hit in less than 1 s using D_{33}. In addition, it was reported at the third position in 9 s using D_{287} and S_2. Note that Substance P is ribosomally encoded; thus, the spectrum can also be searched against a restricted database of proteinogenic amino acids D_{19}.

Beauverolide I, a tetradepsipeptide cyclo([C8:0-Me(4)-OH(3)]-[Phe]-[Ala]-[Leu]), was identified by CycloBranch in less than 1 s using both D_{33} and D_{287}. The sequence has been reported among peptide sequence candidates at the positions 1–2 of a result list of NRP sequence candidates. Beauverolide can also be easily identified using D_{33} by Cyclone if the number of (four) monomers in the peptide was predicted [23].

Roseotoxin A is a hexadepsipeptide cyclo([C5:0-Me(4)-OH(2)]-[Me-Pro]-[Ile]-[NMe-Val]-[NMe-Ala]-[bAla]) where C5:0-Me(4)-OH(2) stands for 2-hydroxy-4-methylpentanoic acid, Me-Pro for 3-methylproline, NMe-Val for N-methyl-valine, NMe-Ala for *N*-methyl-alanine, and bAla for beta alanine. The correct assignment was returned as the top hit in less than 1 s by CycloBranch using D_{33}. It appeared as the second hit using D_{287} and S_2 in 2 s. Since a small monomer ethanolamine (43 Da) is also present in D_{287}, CycloBranch reported another candidate containing this block as the top hit. If a user predicted the number of monomers, Cyclone returned 216 NRP sequence candidates using D_{33}. With that software interface, the user had to retype manually all 216 sequences to another PHP script in order to obtain results consistent with what we can obtain using CycloBranch.

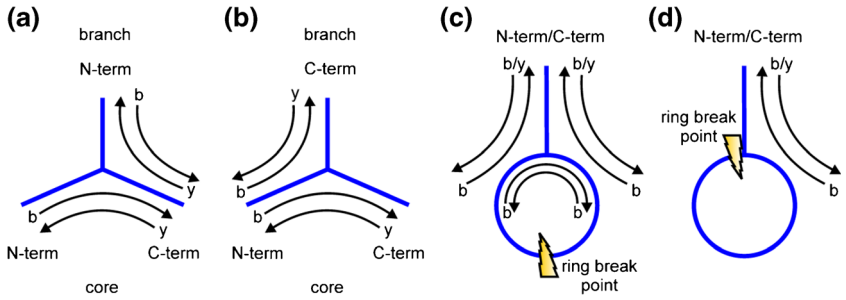


Figure 2. **(a)**, **(b)** Ion series originating from the protonated molecule of a branched peptide; theoretical fragment ion series generated because of a ring opening of a branch-cyclic NRP – **(c)** a branched fragment, **(d)** a linear fragment

Cyclosporin A is a cyclic undecapeptide cyclo([MeBmt]-[Abu]-[Sar]-[Me-Leu]-[Val]-[Me-Leu]-[Ala]-[Ala]-[Me-Leu]-[Me-Leu]-[NMe-Val]) where MeBmt stands for *N*-methyl-butenylthreonine, Abu for 2-aminobutanoic acid, Sar for sarcosine, and Me-Leu for *N*-methylleucine. Sequence tag [Val]-[Me-Leu]-[Ala]-[Ala]-[Me-Leu]-[Me-Leu]-[NMe-Val] was identified as the second hit in 6 s using D_{33}. It was identified at the fifth position in 25 s using D_{287}. S_2 was applied in both cases.

We also tested a two-phase identification of cyclosporin A on D_{33}. In the first phase, the minimum threshold of relative intensity was set up to 5% and the identification of sequence tags was enabled (number of b-ions and dehydrated b-ions was used as a scoring function). The tag [Me-Leu]-[Ala]-[Ala]-[Me-Leu]-[Me-Leu]-[NMe-Val] was reported as a top hit. In the second phase, the relative intensity threshold was lowered, the identification of tags was disabled, and only NRP sequence candidates having the previously detected tag were processed. The correct sequence of cyclosporin was reported as a top hit (see Tutorials 2 and 3 in Supplemental material).

Surfactin C is an octadepsipeptide cyclo([C14:0-Me(13)-OH(3)]-[Glu]-[Leu]-[Leu]-[Val]-[Asp]-[Leu]-[Leu]) where C14:0-Me(13)-OH(3) stands for 3-hydroxy-13-methyltetradecanoic acid. It was reported at the positions 2–3 in less than 1 s for D_{33} (Figures 3 and 4) or at the same positions in 1 s for D_{287} and S_1.

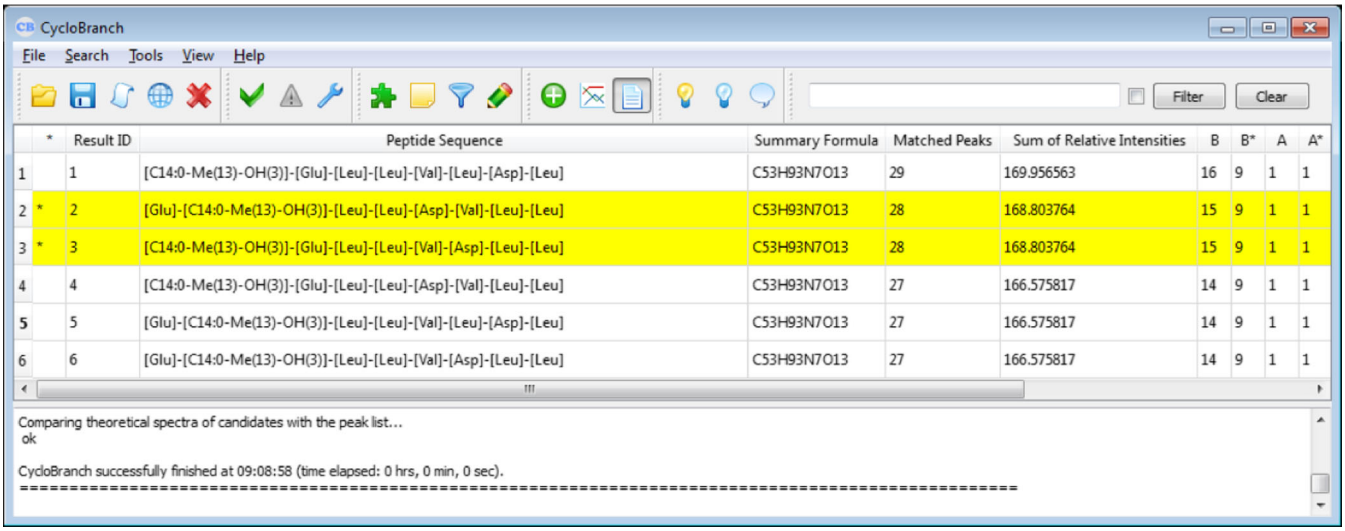
Valinomycin is a dodecadepsipeptide cyclo([Val]-[Lac]-[Val]-[Hiv]-[Val]-[Lac]-[Val]-[Hiv]-[Val]-[Lac]-[Val]-[Hiv]) where Lac stands for lactic acid and Hiv for 2-hydroxyisovaleric acid. A tag [Val]-[Hiv]-[Val]-[Lac]-[Val]-[Hiv] was reported among sequence tag candidates at the positions 1–2 in less than 1 s using D_{33} and S_2. Cyclosporin, surfactin, and valinomycin could not be identified on D_{33} or on D_{287} by Cyclone as they contain more than six monomers.

Linearized pseudacyclin A is a branched peptide with two N-termini and one C-terminus (Figure S-1a in the

Table 1. NRPs Identified or Sequence Tag Determination by CycloBranch

Peptide type	Peptide	Matched peaks (S_1)	Rank S_1	Sum of RI of matched peaks (S_2)	Rank S_2	D_n	Time
Linear	*Gramicidin C* (7 of 15 blocks tag)	13	1–2	436.1	1–2	D_{33}	<1 s
			1–2		3–4	D_{287}	<1 s
	Substance P (6 of 11 blocks tag)	27	1	881.8	1	D_{33}	<1 s
			3–5		3	D_{287}	9 s
Cyclic	*Beauverolide I* (4 blocks)	10	1–2	290.5	1–2	D_{33}	<1 s
			1–2		1–2	D_{287}	<1 s
	Roseotoxin A (6 blocks)	17	1	180.3	1	D_{33}	<1 s
			1–15		2	D_{287}	2 s
	Cyclosporin A (7 of 11 blocks tag)	32	10–24	666.4	2	D_{33}	6 s
			>100		5	D_{287}	25 s
	Surfactin C (8 blocks)	28	2–3	168.8	2–3	D_{33}	<1 s
			2–3		21–22	D_{287}	1 s
	Valinomycin (6 of 12 blocks tag)	26	1–4	626.6	1–2	D_{33}	<1 s
Branched	*Linearized pseudacyclin A* (6 blocks)	10	1–4	202.1	1–4	D_{33}	<1 s
			1–24		25–48	D_{287}	40 s
	T-peptide (11 blocks)*	23	1–24	370.9	1–24	D_{19}	31 s
Branch-cyclic	*Pseudacyclin A* (6 blocks)	17	1	290.8	2	D_{33}	<1 s
			1		2	D_{287}	<1 s
	Pyoverdin Pa A (10 blocks)	25	1–36	295.7	1–12	D_{33}	150 s
	Pyoverdin Pa A (3 of 10 blocks tag)	26	1	294.8	1	D_{33}	<1 s
			10–11		6	D_{287}	1 s

*For description of two-phase identification see the text.
Rank = an order of a correct peptide sequence in an output list of peptide sequence candidates; D_n = database of n building blocks; Time = average time of identification (one thread used for comparisons of theoretical spectra with an experimental spectrum); RI = relative intensities of matched peaks.



	*	Result ID	Peptide Sequence	Summary Formula	Matched Peaks	Sum of Relative Intensities	B	B*	A	A*
1		1	[C14:0-Me(13)-OH(3)]-[Glu]-[Leu]-[Leu]-[Val]-[Leu]-[Asp]-[Leu]	C53H93N7O13	29	169.956563	16	9	1	1
2	*	2	[Glu]-[C14:0-Me(13)-OH(3)]-[Leu]-[Leu]-[Asp]-[Val]-[Leu]-[Leu]	C53H93N7O13	28	168.803764	15	9	1	1
3	*	3	[C14:0-Me(13)-OH(3)]-[Glu]-[Leu]-[Leu]-[Val]-[Asp]-[Leu]-[Leu]	C53H93N7O13	28	168.803764	15	9	1	1
4		4	[C14:0-Me(13)-OH(3)]-[Glu]-[Leu]-[Leu]-[Asp]-[Val]-[Leu]-[Leu]	C53H93N7O13	27	166.575817	14	9	1	1
5		5	[Glu]-[C14:0-Me(13)-OH(3)]-[Leu]-[Leu]-[Val]-[Leu]-[Asp]-[Leu]	C53H93N7O13	27	166.575817	14	9	1	1
6		6	[Glu]-[C14:0-Me(13)-OH(3)]-[Leu]-[Leu]-[Val]-[Asp]-[Leu]-[Leu]	C53H93N7O13	27	166.575817	14	9	1	1

Figure 3. Initial graphic CycloBranch output indicating a list of NRP sequence candidates. The experimental spectrum of *surfactin C* was searched against a D_{33} database

Supplemental Section). When the acquired spectrum was analyzed by CycloBranch, the sequence of this compound was reported among candidates at the positions 1–4 or 1–24 in result sets in less than 1 s using D_{33} or in 40 s using D_{287}, respectively (S_1 only).

T-peptide is also a branched peptide with two N-termini and one C-terminus (Figure S-1b). Since the peptide is synthetic and composed exclusively from proteinogenic amino acids, the spectrum was searched against D_{19}. The correct peptide sequence was reported among sequence candidates at the positions 1–144 in 2 min and 4 s. A high number of false positive candidates was caused by all series being incomplete, especially by the missing ions b_1, b_2, and b_9 in both N-terminal series, and by the missing ions y_1 and y_9 in both C-terminal series (Figure 5).

We also performed a two-phase identification of T-peptide on D_{19}. In the first phase, the peptide was considered as a linear one and the detection of sequence tags was enabled. Two tags

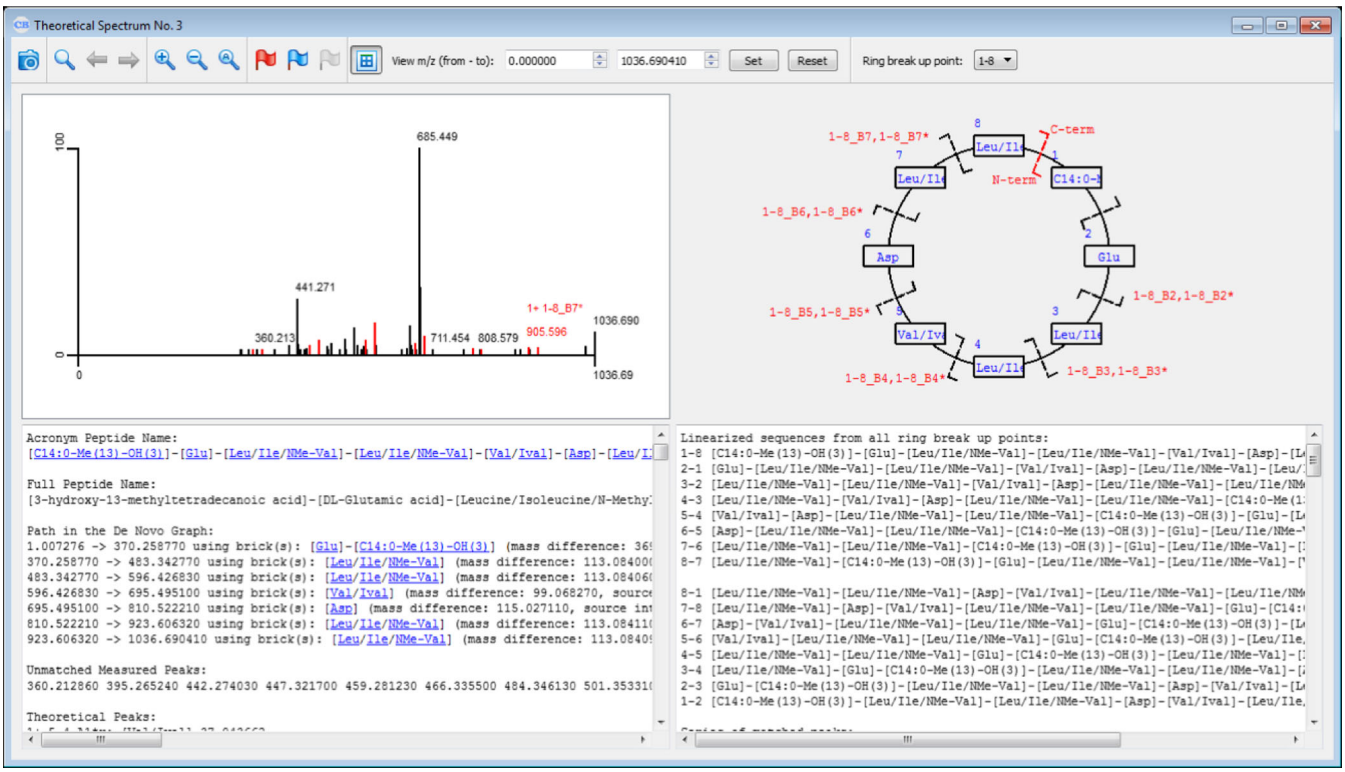


Figure 4. Examining the hits retrieved by CycloBranch. The theoretical *surfactin C* spectrum was compared with its experimental one

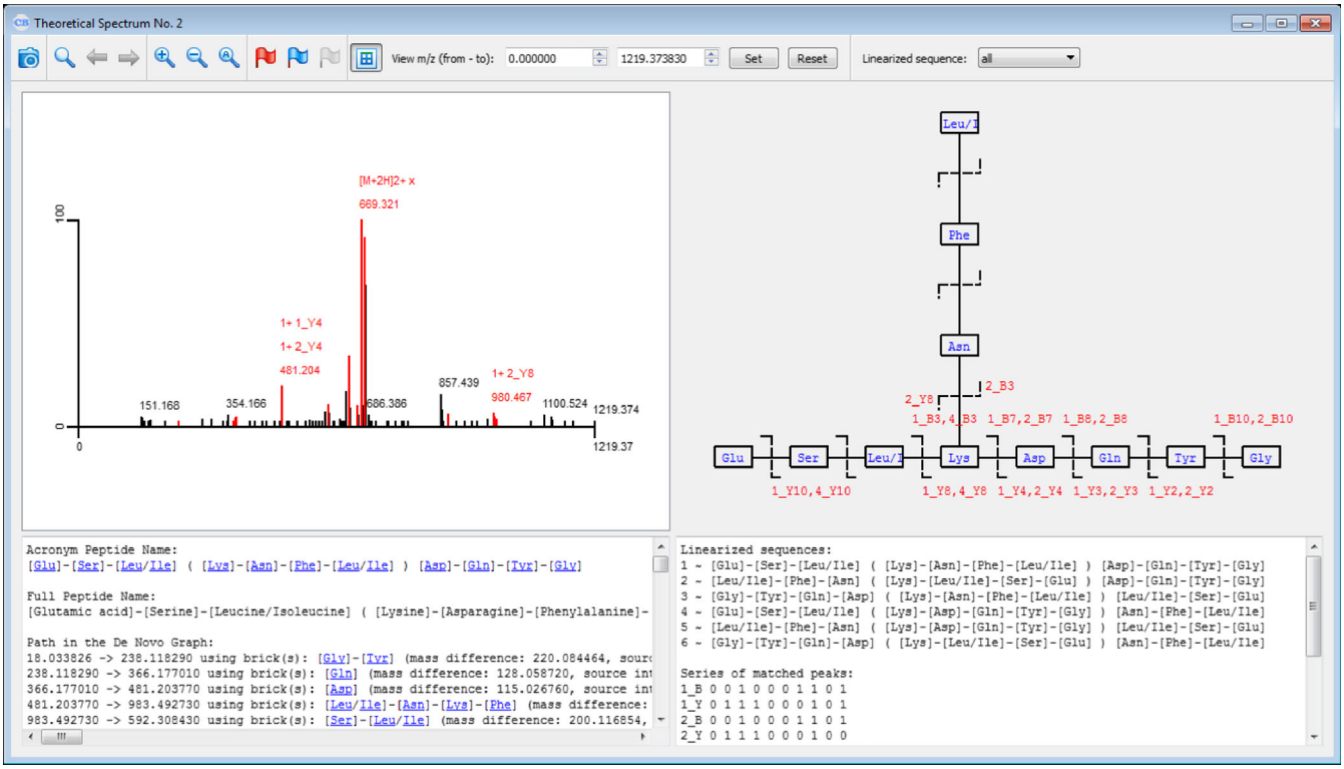


Figure 5. The theoretical *T-peptide* spectrum was compared with its experimental one

[Asp]-[Gln] and [Ser] were reported. In the second phase, the peptide was considered as a branched one, the detection of sequence tags was disabled and only NRP sequence candidates having both tags were processed. CycloBranch reported 24 NRP sequence candidates having the equal backbone *N*-acetyl-EXX(KXXX)DQYG$_{NH2}$ in 31 s (see Tutorial 4 in Supplemental material).

Pseudacyclin A, a hexapeptide cyclo([Phe]-[Pro]-[Ile]-[Ile]([Orn]-[N-Ac-Ile])), contains five building blocks in a ring and one building block [N-Ac-Ile], which forms a side chain (Orn stands for ornithine, N-Ac-Ile for *N*-acetylisoleucine). It has not been identified by CycloBranch as a branch-cyclic peptide due to predominant elimination of [N-Ac-Ile] during the mass spectrometric analysis. Since the branch is short, pseudacyclin data were treated as for a cyclic peptide. CycloBranch reported its sequence as the top hit using both databases D_{33} and D_{287} in less than 1 s (S_1). Similarly to roseotoxin, pseudacyclin can be identified on D_{33} by Cyclone but the number of monomers has to be predicted. Cyclone returns 1240 NRP sequence candidates. The user has to retype manually all 1240 candidates to another script to obtain the same results that are obtained with CycloBranch in a single and automated approach.

Pyoverdin Pa A is a decapeptide [Suc]-[ChrPaA]-[Ser]-[Arg]-[Ser]-[Fo-OH-Orn]-cyclo([Lys]-[Fo-OH-Orn]-[Thr]-[Thr]) where Suc, ChrPaA, and Fo-OH-Orn stand for succinic acid, pyoverdin Pa A chromophore, and N6-formyl-hydroxyornithine, respectively (Figure S-2). The compound was not identified by CycloBranch as a branch-cyclic peptide, as no *b*-ions arising from the ring opening were observed. Since the branch was long, it was treated as a linear peptide having the C-terminus cyclized (i.e., 18 Da have to be subtracted from C-terminal fragment ions masses). The correct NRP sequence was reported among candidates at the positions 1–12 in 150 s using D_{33} and S_2. A tag [Arg]-[Ser]-[Fo-OH-Orn] was reported as a top hit using D_{33} or at the sixth position using D_{287} and S_2 in 1 s.

Conclusion

The stand-alone and open-source de novo peptide identification engine CycloBranch was effectively utilized for identification or sequence tag determination of NRPs from accurate product ion mass spectra. It represents the first and true de novo engine working for nonribosomal peptides. It supports sequencing of linear, branched, and branch-cyclic NRPs as well as cyclic peptides. NRP sequence tags were provided in the output even when a building block was not present in a database or the spectrum contained incomplete fragment ion series. The parts of an unknown structure, if not covered by database building blocks or their combinations, are returned as exact monoisotopic masses in suggested sequence tags. CycloBranch has a graphical and user-friendly interface and it can run in parallel on multiple threads.

The remaining challenge is the automated detection of the peptide type that is to be identified. Although distinguishing of a linear and a cyclic peptide spectrum has been proposed for

ribosomal peptides [5], distinguishing various types of NRPs is still a nontrivial task. It may be advantageous to run the engine repeatedly for different types of NRPs. For example, if *y*-ions are not observed, multiple overlapping *b*-ion series and scrambled fragment ions occur, a peptide likely contains a cycle. Otherwise, it may correspond to a linear or a branched NRP. The tool is designed to be extended for the identification of other types of NRPs (e.g., bicyclic or multiply branched) and structures containing other building blocks (e.g., saccharides, nucleotides). It is worth noting that de novo engine for top-down sequencing of proteins is also desperately needed by the proteomic community.

Acknowledgment

The authors acknowledge the major direct support from the Czech Science Foundation (P206/12/1150). Access to instrumental and other facilities was also supported by EU (Operational Program Prague – Competitiveness project CZ.2.16/3.1.00/24023, Ministry of Education, Youth, and Sports of the Czech Republic (LH14064, NPU LO1509 and NPU LO1305) and IMIC Institutional Research Concept RVO61388971.

References

1. Strieker, M., Tanović, A., Marahiel, M.A.: Nonribosomal peptide synthetases: structures and dynamics. Curr. Opin. Struct. Biol. **20**, 234–240 (2010)
2. Caboche, S., Leclere, V., Pupin, M., Kucherov, G., Jacques, P.: Diversity of monomers in nonribosomal peptides: towards the prediction of origin and biological activity. J. Bacteriol. **192**, 5143–5150 (2010)
3. Adamska, A., Janecka, A.: Endless peptides - circular forms in nature. Curr. Med. Chem. **22**, 352–359 (2015)
4. Novák, J., Sachsenberg, T., Hoksza, D., Skopal, T., Kohlbacher, O.: On comparison of SimTandem with state-of-the-art peptide identification tools, efficiency of precursor mass filter, and dealing with variable modifications. J. Int. Bioinform. **10**, 228 (2013)
5. Mohimani, H., Liu, W.T., Mylne, J.S., Poth, A.G., Colgrave, M.L., Tran, D., Selsted, M.E., Dorrestein, P.C., Pevzner, P.A.: Cycloquest: identification of cyclopeptides via database search of their mass spectra against genome databases. J. Proteome Res. **10**, 4505–4512 (2011)
6. Dančík, V., Addona, T.A., Clauser, K.R., Vath, J.E., Pevzner, P.A.: De novo peptide sequencing via tandem mass spectrometry. J. Comput. Biol. **6**, 327–342 (1999)
7. Mann, M., Wilm, M.: Error-tolerant identification of peptides in sequence databases by peptide sequence tags. Anal. Chem. **66**, 4390–4399 (1994)
8. Klitgaard, A., Iversen, A., Andersen, M.R., Larsen, T.O., Frisvad, J.C., Nielsen, K.F.: Aggressive dereplication using UHPLC-DAD-QTOF: screening extracts for up to 3000 fungal secondary metabolites. Anal. Bioanal. Chem. **406**, 1933–1943 (2014)
9. Caboche, S., Pupin, M., Leclère, V., Fontaine, A., Jacques, P., Kucherov, G.: NORINE: a database of nonribosomal peptides. Nucleic Acids Res. **36**, D326–D331 (2008)
10. Ng, J., Bandeira, N., Liu, W.T., Ghassemian, M., Simmons, T.L., Gerwick, W.H., Linington, R., Dorrestein, P.C., Pevzner, P.A.: Dereplication and de novo sequencing of nonribosomal peptides. Nat. Methods **6**, 596–599 (2009)
11. Ibrahim, A., Yang, L., Johnston, C., Liu, X., Ma, B., Magarvey, N.A.: Dereplicating nonribosomal peptides using an informatic search algorithm for natural products (iSNAP) discovery. Proc. Natl. Acad. Sci. U. S. A. **109**, 19196–19201 (2012)
12. Mohimani, H., Liu, W.T., Kersten, R.D., Moore, B.S., Dorrestein, P.C., Pevzner, P.A.: NRPquest: coupling mass spectrometry and genome mining for nonribosomal peptide discovery. J. Nat. Prod. **77**, 1902–1909 (2014)
13. Allmer, J.: Algorithms for the de novo sequencing of peptides from tandem mass spectra. Expert. Revi. Proteom. **8**, 645–657 (2011)
14. Ma, B., Zhang, K., Hendrie, C., Liang, C., Li, M., Doherty-Kirby, A., Lajoie, G.: PEAKS: Powerful software for peptide de novo sequencing by tandem mass spectrometry. Rapid Commun. Mass Spectrom. **17**, 2337–2342 (2003)
15. Taylor, J.A., Johnson, R.S.: Sequence database searches via de Novo peptide sequencing by tandem mass spectrometry. Rapid Commun. Mass Spectrom. **11**, 1067–1075 (1997)
16. Frank, A., Pevzner, P.: PepNovo: de novo peptide sequencing via probabilistic network modeling. Anal. Chem. **77**, 964–973 (2005)
17. Muth, T., Weilnböck, L., Rapp, E., Huber, C.G., Martens, L., Vaudel, M., Barsnes, H.: DeNovoGUI: an open source graphical user interface for de novo sequencing of tandem mass spectra. J. Proteome Res. **13**, 1143–1146 (2014)
18. Mohimani, H., Yang, Y.L., Liu, W.T., Hsieh, P.W., Dorrestein, P.C., Pevzner, P.A.: Sequencing cyclic peptides by multistage mass spectrometry. Proteomics **11**, 3642–3650 (2011)
19. Mohimani, H., Liu, W.T., Yang, Y.L., Gaudêncio, S.P., Fenical, W., Dorrestein, P.C., Pevzner, P.A.: Multiplex de novo sequencing of peptide antibiotics. J. Comput. Biol. **18**, 1371–1381 (2011)
20. Liu, W.T., Ng, J., Meluzzi, D., Bandeira, N., Gutierrez, M., Simmons, T.L., Schultz, A.W., Linington, R.G., Moore, B.S., Gerwick, W.H., Pevzner, P.A., Dorrestein, P.C.: Interpretation of tandem mass spectra obtained from cyclic nonribosomal peptides. Anal. Chem. **81**, 4200–4209 (2009)
21. Niedermeyer, T.H.J., Strohalm, M.: mMass as a Software tool for the annotation of cyclic peptide tandem mass spectra. *PLoS ONE* 7 (2012).
22. Jagannath, S., Sabareesh, V.: Peptide Fragment Ion Analyser (PFIA): a simple and versatile tool for the interpretation of tandem mass spectrometric data and de novo sequencing of peptides. Rapid Commun. Mass Spectrom. **21**, 3033–3038 (2007)
23. Kavan, D., Kuzma, M., Lemr, K., Schug, K.A., Havlicek, V.: CYCLONE—a utility for de novo sequencing of microbial cyclic peptides. J. Am. Soc. Mass Spectrom. **24**, 1177–1184 (2013)
24. Sturm, M., Bertsch, A., Gröpl, C., Hildebrandt, A., Hussong, R., Lange, E., Pfeifer, N., Schulz-Trieglaff, O., Zerck, A., Reinert, K., Kohlbacher, O.: OpenMS—an open-source software framework for mass spectrometry. BMC Bioinformatics **9** (2008)
25. Steen, H., Mann, M.: The ABC's (and XYZ's) of peptide sequencing. Nat. Rev. Mol. Cell Biol. **5**, 699–711 (2004)
26. Ngoka, L.C.M., Gross, M.L.: A nomenclature system for labeling cyclic peptide fragments. J. Am. Soc. Mass Spectrom. **10**, 360–363 (1999)
27. Bleiholder, C., Osburn, S., Williams, T.D., Suhai, S., Van Stipdonk, M., Harrison, A.G., Paizs, B.: Sequence-scrambling fragmentation pathways of protonated peptides. J. Am. Chem. Soc. **130**, 17774–17789 (2008)