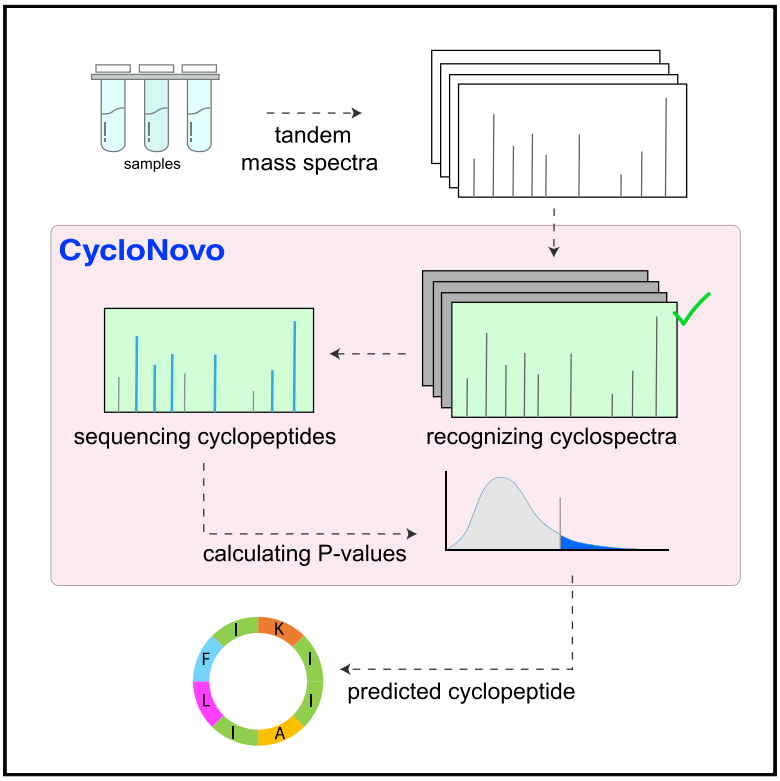
Report

Cell Systems

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

Graphical Abstract



Highlights

- Software for high-throughput *de novo* cyclopeptide analysis of tandem mass spectra
- Reported the first cyclopeptides discovered through fully automated *de novo* sequencing
- Found cyclopeptides from food and human gut that survived human's gastrointestinal tract

Authors

Bahar Behsaz, Hosein Mohimani, Alexey Gurevich, ..., Pieter C. Dorrestein, Joshua S. Mylne, Pavel A. Pevzner

Correspondence

bbehsaz@ucsd.edu

In Brief

We present a computational pipeline for automated analysis and discovery of cyclic and branch cyclic peptides (cyclopeptides). Cyclopeptides represent an important class of bioactive natural products that includes many antibiotics and anti-tumor compounds. Our CycloNovo algorithm for *de novo* cyclopeptide sequencing of spectral datasets reconstructed many previously unreported cyclopeptides (to the best of our knowledge) that were validated with matching transcriptomics and genome or metagenome mining. CycloNovo revealed several cyclopeptides from food that survived the complete human gastrointestinal tract, raising the question of how these compounds might affect the human microbiome.

Behsaz et al., 2020, Cell Systems *10*, 99–108
January 22, 2020 © 2019 Published by Elsevier Inc.
https://doi.org/10.1016/j.cels.2019.11.007

CellPress

Cell Systems
Report

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

Bahar Behsaz,[1,9,*] Hosein Mohimani,[2,8] Alexey Gurevich,[3] Andrey Prjibelski,[3] Mark Fisher,[4] Fernando Vargas,[7] Larry Smarr,[2,5,6] Pieter C. Dorrestein,[6,7] Joshua S. Mylne,[4] and Pavel A. Pevzner[2,3,6]
[1]Bioinformatics and Systems Biology Program, University of California, San Diego, La Jolla, CA, USA
[2]Department of Computer Science and Engineering, University of California, San Diego, La Jolla, CA, USA
[3]Center for Algorithmic Biotechnology, Institute of Translational Biomedicine, St. Petersburg State University, St. Petersburg, Russia
[4]School of Molecular Sciences and the ARC Centre of Excellence in Plant Energy Biology, University of Western Australia, Crawley, WA, Australia
[5]California Institute for Telecommunications and Information Technology, University of California, San Diego, La Jolla, CA, USA
[6]Center for Microbiome Innovation, University of California, San Diego, La Jolla, CA, USA
[7]Department of Pharmacology, University of California, San Diego, La Jolla, CA, USA
[8]Computational Biology Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA
[9]Lead Contact
*Correspondence: bbehsaz@ucsd.edu
https://doi.org/10.1016/j.cels.2019.11.007

SUMMARY

Cyclic and branch cyclic peptides (cyclopeptides) represent a class of bioactive natural products that include many antibiotics and anti-tumor compounds. Despite the recent advances in metabolomics analysis, still little is known about the cyclopeptides in the human gut and their possible interactions due to a lack of computational analysis pipelines that are applicable to such compounds. Here, we introduce CycloNovo, an algorithm for automated *de novo* cyclopeptide analysis and sequencing that employs de Bruijn graphs, the workhorse of DNA sequencing algorithms, to identify cyclopeptides in spectral datasets. CycloNovo reconstructed 32 previously unreported cyclopeptides (to the best of our knowledge) in the human gut and reported over a hundred cyclopeptides in other environments represented by various spectra on Global Natural Products Social Molecular Network (GNPS). https://github.com/bbehsaz/cyclonovo

INTRODUCTION

The golden age of antibiotics was followed by a decline in the pace of antibiotic discovery in the 1990s. However, antibiotics and other natural products are again at the center of attention as exemplified by the recent discovery of teixobactin (Ling et al., 2015). A key prerequisite for the resurgence of natural product research is the development of computational discovery pipelines (Medema and Fischbach, 2015) such as the Global Natural Products Social (GNPS) molecular network (Wang et al., 2016), Dereplicator (Mohimani et al., 2017), and VarQuest (Gurevich et al., 2018). The GNPS project alone has already accumulated over one billion mass spectra, an untapped resource for the discovery of new antibiotics. Currently, however, the GNPS network is mainly used for identifying previously discovered natural products and their analogs. Therefore, developing algorithms that are not bound by the current database of natural products is necessary to truly realize the promise of computational natural product discovery.

This study focuses on *de novo* analysis of cyclopeptides, which includes cyclic and branch-cyclic peptides (might include several branches), an important and large class of bioactive natural products with an unparalleled track record in pharmacology. Many antibiotics as well as anti-tumor agents, immunosuppressors, and toxins are cyclopeptides. The favorable properties of bioactive cyclopeptides such as high affinity and selectivity has made them particularly interesting as drug candidates (Zorzi et al., 2017) among peptidic natural products. Despite this great interest, cyclopeptide sequencing and analysis from tandem mass spectra are extra challenging as the propensity of these molecules to break at all pairs of points in their cyclic backbone gives a far more complex series of ions than in linear peptides. Cyclopeptides are divided into cyclic ribosomally synthesized and post-translationally modified peptides (RiPPs) and cyclic nonribosomal peptides (NRPs). RiPPs are encoded using the genetic code and are built from the 20 proteinogenic amino acids, which, however, are subjected to numerous post-translational modifications. NRPs are not directly inscribed in the genomes and cannot be inferred with traditional DNA sequencing. Instead, they are encoded using "nonribosomal code" (Marahiel et al., 1997) and are built from over 300 different naturally occurring amino acids.

The discovery of the cyclopeptide gramicidin S in 1942 (first antibiotic used for treating soldiers during the World War II) led to two Nobel prizes and has been followed by the discovery of ≈400 families of cyclopeptides (cyclofamilies) in the last 75 years (Gurevich et al., 2018). A relatively small number of known cyclofamilies reflects the experimental and computational challenges in cyclopeptide discovery. The question of how many

cyclofamilies stayed below the radar of previous studies (even though their spectra have already been deposited to public databases) remains open. To answer this question, first we considered the problem of recognizing cyclospectra (tandem mass spectra that originated from cyclopeptides) in large spectral datasets (targeted or untargeted).

Bandeira et al. (2007) introduced the concept of spectral networks that reveal the spectra of related peptides without knowing their amino acid sequences. Nodes in a spectral network correspond to spectra, while edges connect spectral pairs, i.e., spectra of peptides differing by a single modification or a mutation. Ideally, each connected component of a spectral network corresponds to a cyclofamily (Mohimani et al., 2011b) representing a set of similar cyclopeptides (see Figures S1 and S2). Although spectral networks of various GNPS datasets have become the workhorse of the cyclopeptide studies, they typically contain false-positive edges that render the analysis of cyclofamilies challenging (Wang et al., 2016). Moreover, constructing the spectral network of all GNPS spectra remains an open algorithmic problem.

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

Recognition of cyclospectra not only creates the possibility to get an estimate on the number of cyclofamilies for large-scale projects but also makes otherwise time-consuming downstream analyses feasible for such projects. For example, recognized cyclospectra can be matched against NRP and RiPP biosynthetic genes using various genome mining and peptidogenomics tools (Mohimani et al., 2014a, 2014b). These tools typically generate a huge database of putative cyclopeptides mounting to possibly millions of putative peptides (Mohimani et al., 2014b), making it prohibitively time consuming to search large spectral datasets against such databases. This problem is especially aggravated in the case of NRPs where the traditional DNA sequencing approaches do not apply. Genome mining tools for NRP prediction such as NRPSpredictor2 (Röttig et al., 2011) combined with NRPquest (Mohimani et al., 2013) predict a database of putative NRPs based on the nonribosomal code. Because the nonribosomal code is not yet fully understood and is not as specific as the genetic code, these analyses often result in colossal databases of putative error-prone peptides predicted from a single gene cluster and therefore are limited in their power for peptidogenomics analysis (Medema and Fischbach, 2015; Mohimani et al., 2014b). Fast algorithms for recognizing cyclospectra are critical as they greatly reduce the set of spectra that need to be matched against databases of putative cyclopeptides.

In addition to recognizing cyclospectra, CycloNovo *de novo* sequences the recognized cyclospectra. We distinguish between cyclopeptide identification (identifying cyclopeptides by matching their spectra against databases of known cyclopeptides (Mohimani and Pevzner, 2016)) and *de novo* cyclopeptide sequencing (determining the cyclopeptide sequence from a spectrum alone). Although recent studies have made progress toward cyclopeptide identification (Gurevich et al., 2018; Ibrahim et al., 2012; Mohimani et al., 2011a, 2017), previous cyclopeptide sequencing algorithms (Kavan et al., 2013; Mohimani et al., 2011c; Ng et al., 2009; Townsend et al., 2018) are not compatible with the large-scale nature of current metabolomics and mass spectrometry studies. These tools are rarely used (Kersten et al., 2011; Mohimani and Pevzner, 2016) because they are inaccurate, too slow for analyzing high-throughput spectral datasets, limited to the proteinogenic amino acids, and also cannot distinguish cyclopeptides from other compounds. To date, no cyclopeptide has been introduced through fully automated *de novo* sequencing of tandem mass spectra.

To close these gaps, here we present CycloNovo, an algorithm that performs fast cyclopeptide sequencing based on the concept of the de Bruijn graph of a spectrum, a compact representation of putative *k*-mers (strings formed by *k* consecutive amino acids) in an unknown cyclopeptide. Although de Bruijn graphs represent the workhorse of DNA sequencing (Compeau et al., 2011), they have not previously been applied to cyclopeptide sequencing. We demonstrate that using this technique, CycloNovo enables high-throughput analysis of cyclopeptides in large spectral datasets, sequencing many cyclopeptides in diverse samples that include marine, soil, and human gut bacterial communities.

RESULTS

To illustrate how CycloNovo works, we used a spectrum of the cyclopeptide surugamide A (Takada et al., 2013) (referred to as surugamide hereon) with the amino acid sequence AIIKIFLI (Figure 1).

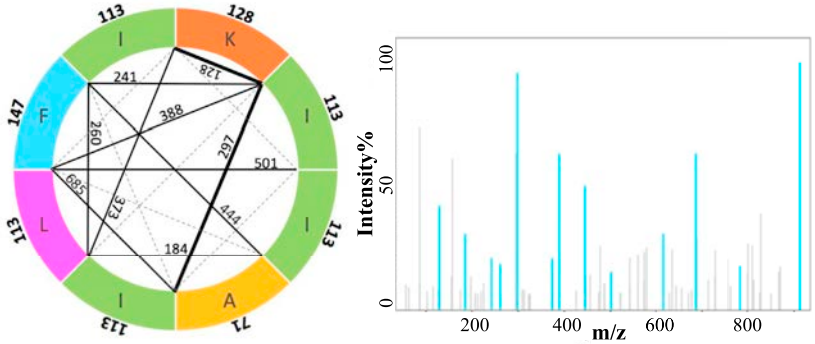
Theoretical and Experimental Spectra

Given an amino acid string, its mass is defined as the sum of masses of its amino acids. Given a cyclopeptide *Peptide*, its theoretical spectrum *TheoreticalSpectrum*(Peptide) is the set of masses of all substrings of *Peptide* (Figure 1). For example, *TheoreticalSpectrum*(AGCD) contains masses of A, G, C, D, AG, GC, CD, DA, AGC, GCD, CDA, DAG, and AGCD. Note that if multiple fragments have the same mass, they contribute a single mass to the theoretical spectrum.

An experimental spectrum is a list of peaks, where each peak is characterized by its intensity and *m*/*z* (*m* and *z* represent the mass and the charge of the ion corresponding to the peak). For simplicity, we represent a preprocessed spectrum as an increasing sequence of numbers Spectrum $=\{s_1, \ldots, s_n\}$, assuming that all peaks in the spectrum have charge 1 and ignoring intensities (see the STAR Methods section about preprocessing spectra). We estimate the *PeptideMass* of the cyclopeptide that generated *Spectrum* based on the precursor mass and the charge of *Spectrum*. We define the symmetric version of *Spectrum* (denoted *Spectrum**) as a spectrum that, in addition to all masses in *Spectrum*, contains *PeptideMass*-*s* for each mass *s* in *Spectrum*.

Scoring Peptide-Spectrum Matches

A mass *s* in a (preprocessed) experimental spectrum *Spectrum* matches a mass *s′* in *TheoreticalSpectrum*(Peptide) if *s* is "equal" to *s′*. By "equal" we mean "approximately equal" with error below the *error threshold* ε (all default values are listed in Table S1). The score between *Peptide* and *Spectrum*, denoted



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

Figure 1. Theoretical and Experimental Spectra of Surugamide

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

score(*Peptide, Spectrum*), is defined as the number of matches between masses in *Spectrum* and masses in *TheoreticalSpectrum*(Peptide). Although CycloNovo uses accurate masses, examples below use nominal masses for simplicity.

Figure 1 illustrates that *score*(*Surugamide*, $Spectrum_{Surugamide}$) = 13. For a linear peptide *Peptide, score*(*Peptide, Spectrum*) is the number of matches between the masses of all linear substrings of *Peptide* and all masses in *Spectrum*. For example, *score*(ILFIK, $Spectrum_{Surugamide}$) = 7 because the theoretical spectrum of the linear peptide ILFIK has 7 shared masses with $Spectrum_{Surugamide}$ corresponding to 7 chords within the ILFIK segment in Figure 1. These chords correspond to the following substrings: K (nominal mass 128), IL (226), IK (241), LF (260), LFI (373), LFIK (501), and ILFIK (614).

Selecting the Set of Amino Acids

Gurevich et al. (2018) combined all currently known peptidic natural products into a single *PNPDatabase*. This database contains 1,257 cyclopeptides (387 cyclofamilies) that we refer to as *CyclopeptideDatabase*. We formed the set of *selected amino acids* (i.e., amino acids that occur in many cyclopeptides) by considering the 25 most frequent amino acids in cyclopeptides from *CyclopeptideDatabase* and extending this list to include all proteinogenic amino acids and some common amino acids in RiPPs (see Table S2 and the STAR Methods for more information). There are 255 out of 1,257 cyclopeptides in the *CyclopeptideDatabase* that include only selected amino acids.

Spectral Convolution

The *convolution* of a spectrum is the set of all pairwise differences between its masses (Ng et al., 2009). Given a mass *a*, the convolution of *Spectrum with offset a* (denoted *convolution*(*Spectrum, a*)) is defined as the number of masses in the convolution equal to *a* (with error up to ε). As shown by Ng et al. (2009), the value *convolution*(*Spectrum, a*) is expected to be high if *a* is the mass of an amino acid in a cyclopeptide that gave rise to *Spectrum*. Thus, offsets with high convolutions reveal the masses of amino acids in an unknown cyclopeptide that gave rise to an experimental spectrum.

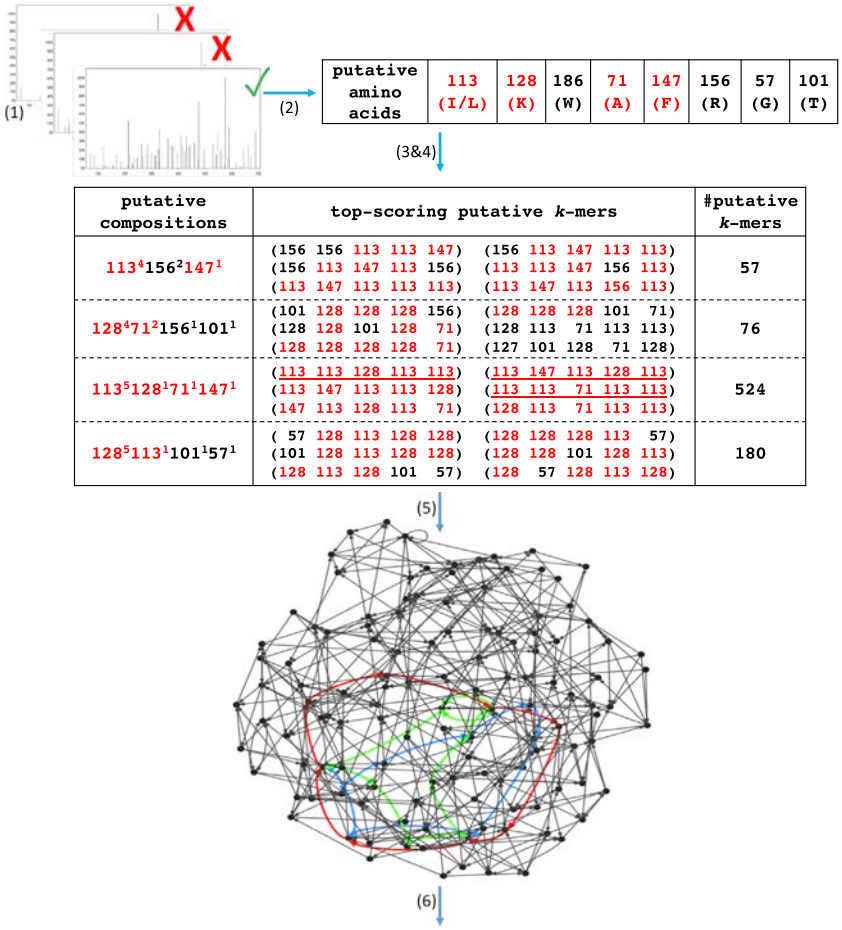
To account for measurement errors, we cluster the masses in the convolution using *single linkage clustering* by combining pairs of masses in a cluster if they are less than ε apart. We define the *cluster mass* as the median mass of its members, and *cluster multiplicity* as the number of elements in the cluster. We call a cluster cyclopeptidic if one of its elements is within ε of the mass of a selected amino acid. Since high-multiplicity clusters reveal amino acids in the unknown cyclopeptide that gave rise to an experimental spectrum, we use them to generate the set of putative amino acids in an unknown cyclopeptide (Ng et al., 2009). See STAR Methods section Analyzing Spectral Convolution for more details.

CycloNovo Outline

Given an experimental spectrum *Spectrum*, the Cyclopeptide Sequencing Problem refers to finding a cyclopeptide *Peptide* that maximizes *score*(*Peptide, Spectrum*). Figure 2 illustrates the CycloNovo pipeline for solving this problem:

Recognizing Cyclospectra

Natural product researchers use Marfey's analysis for inferring the amino acid composition and configuration of an unknown peptide. However, as Marfey's analysis requires a purified peptide and has a number of limitations (Bhushan and Brückner,



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

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

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

Figure 2. CycloNovo Outline Illustrated Using $Spectrum_{Surugamide}$

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

2011), we describe its *in silico* alternative for deriving an *approximate* amino acid composition of a cyclopeptide that gave rise to a given spectrum (see STAR Methods). If applying this approach reveals that a spectrum originated from a cyclopeptide, we classify it as a cyclospectrum.

Predicting Amino Acids in a Cyclopeptide

For each cyclospectrum, CycloNovo predicts the set of putative amino acids in a cyclopeptide that gave rise to this spectrum. CycloNovo considers each cyclopeptidic cluster with multiplicity exceeding the *cyclopeptidic aa threshold* and classifies the selected amino acids corresponding to this cluster as a *putative amino acid* of the cyclopeptide that generated the cyclospectrum. Figure 2 illustrates that CycloNovo classifies amino acids A, I/L, F, K, T, W, R, and G as putative amino acids for $Spectrum_{Surugamide}$ (amino acids occurring in suragamide are shown in bold).

Predicting Amino Acid Composition of a Cyclopeptide

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

Predicting k-mers in a Cyclopeptide

For each *Composition*(*Spectrum*), CycloNovo analyzes all linear *k*-mers formed by amino acids in this composition (the default value k = 5) and scores them against *Spectrum** using linear scoring. It assumes that if a Peptide-Spectrum Match has a high score score(*Peptide, Spectrum*) (a condition that usually holds for well-fragmented spectra), then each linear k-mer in *Peptide* also has a high score (for an appropriately chosen *k*). High-scoring *k*-mers (defined as *k*-mers with scores exceeding the *k-mer score threshold*) represent putative *k*-mers in an unknown cyclopeptide. For example, for *Composition* = $71^1113^5128^1147^1$, there exist 4^5 = 1,024 5-mers and CycloNovo identifies 524 of them as high-scoring 5-mers. We refer to the set of high-scoring *k*-mers as $Kmers_{Composition,k}(Spectrum)$. Figure 2 illustrates that three out of six highest-scoring 5-mers for $Spectrum_{Surugamide}$ are correct,

i.e., represent 5-mers from surugamide. CycloNovo computes the *k-merScore*, the score of the highest-scoring k-mer.

Constructing the de Bruijn Graph of a Spectrum

Given a set $Kmers = Kmers_{Composition,k}(Spectrum)$, CycloNovo constructs the *de Bruijn graph* $DB_{Kmers}(Spectrum)$ (Compeau et al., 2011). Nodes in $DB_{Kmers}(Spectrum)$ correspond to all (*k-1*)-mers from *Kmers* and each directed edge corresponds to a k-mer from *Kmers* and connects its first (*k-1*)-mer with its last (*k-1*)-mer. Each cycle in $DB_{Kmers}(Spectrum)$ spells out a cyclic amino acid sequence. Figure 2 presents the *pruned de Bruijn graph* for the putative composition $71^1113^5128^1147^1$ that is obtained by iterative removal of tips (nodes without outgoing or incoming edges) and single isolated edges from the de Bruijn graph. The composition $113^5128^171^1147^1$ results in a de Bruijn graph with 202 vertices and 524 edges and the pruned de Bruijn graph with 126 vertices and 392 edges (Figure 2). Figure S10 shows the pruned de Bruijn graph for the other compositions.

Generating Cyclopeptide Reconstructions

A cycle in the de Bruijn graph of a spectrum is *feasible* if it spells a cyclopeptide with the mass matching the precursor mass of the spectrum. Using the breadth-first search algorithm, CycloNovo finds all feasible cycles in the de Bruijn graph with a length equal to the number of amino acids in *Composition* (a cycle may traverse the same edge multiple times). Each such cycle spells a putative cyclopeptide and CycloNovo scores each of them against *Spectrum*. Finally, it reports the highest-scoring cyclopeptides along with the p values of their peptide-spectrum matches (PSMs) computed using MS-DPR (Mohimani et al., 2013). The default p value threshold (10^{-15}) is chosen based on the previous studies where the p value cutoff 10^{-15} was necessary for reaching a false discovery rate (FDR) below 1% against *CyclopeptideDatabase* (Gurevich et al., 2018; Mohimani et al., 2017). However, the user can change the p value thresholds depending on their study. See STAR Methods section Computing P Values for *De Novo* Generated PSMs for a brief summary on p value estimations in CycloNovo.

Datasets

We analyzed various spectral datasets obtained from diverse bacterial communities (Tables 1, S5, and S6 and STAR Methods section Information about Spectral Dataset). To benchmark CycloNovo, we also analyzed a plant spectral dataset that had a paired RNA sequencing (RNA-seq) dataset, thus enabling us to validate the CycloNovo reconstructions by matching them against the transcriptome.

The CYCLOLIBRARY dataset contains 81 spectra from 81 distinct cyclopeptides (forming 41 cyclofamilies) that were identified by Dereplicator (Mohimani et al., 2017) after searching the GNPS network against *CyclopeptideDatabase* (Gurevich et al., 2018).

The S.VULGARIS dataset was generated from a single sample collected from seeds of the plant *Senecio vulgaris* (both medicinal and poisonous) (Jayasena et al., 2017) from the *Asteraceae* family. We also analyzed the RNA-seq reads from the same sample (~74 million 100 bp long Illumina reads) (Jayasena et al., 2017), assembled them using rnaSPAdes (Bankevich et al., 2012), and used the assembled transcripts (61.9-Mb total length) and prior knowledge of cyclopeptide processing (Elliott et al., 2014; Fisher et al., 2018; Mylne et al., 2011) to validate the reconstructed cyclopeptides.

The HUMANSTOOL dataset was generated from 65 stool samples of a single person (L.S., co-author of this paper and a contributor to the "quantified self" initiative) collected over a course of four years. This dataset is accompanied by the detailed medical and food metadata (Yazdani et al., 2016) as well as metagenomics reads generated from the same samples (BioProject: PRJEB24161).

The GNPS dataset was formed by combining forty datasets from GNPS (Wang et al., 2016). The $GNPS_{CYANO}$, $GNPS_{PSEUDO}$, and $GNPS_{ACTI}$ datasets represent sub-datasets of the GNPS dataset corresponding to three phyla with extensively analyzed cyclopeptides (*Cyanobacteria*, *Pseudomonas*, and *Actinobacteria*).

Analyzing the CYCLOLIBRARY Dataset

As the cyclopeptides that gave rise to the spectra in the CYCLOLIBRARY dataset are known, we used this dataset to benchmark CycloNovo. We considered a cyclopeptide (or a spectrum) as correctly sequenced if the sequence of the cyclopeptide appeared among reconstructions with three highest-scores. CycloNovo recognized 45 spectra in the CYCLOLIBRARY dataset as cyclospectra and correctly sequenced 38 of these cyclospectra (see Table S7 for more information). For 22 cyclopeptides, the correct sequence of the cyclopeptide was among the highest-scoring reconstructions (Table S7). For each of the remaining 16 correctly sequenced cyclopeptides, Table S7 lists a high-scoring reconstruction and shows that for all those cases, at least one of the high-scoring reconstructions represented a rearranged sequence of amino acids compared to the correct sequence, except for a single cyclopeptide, where CycloNovo failed to predict correct amino acids for the top-scoring reconstruction. Table S8 provides information about the remaining seven cyclospectra in the CYCLOLIBRARY that CycloNovo failed to reconstruct.

Analyzing the S.VULGARIS Dataset

The 23 recognized cyclospectra in this dataset correspond to twelve distinct cyclopeptides. CycloNovo sequenced ten of them with p values below 10^{-15} (Table S9). Nine of ten reconstructed cyclopeptides matched the assembled transcriptome. One reconstructed cyclopeptide (with the highest-scoring reconstruction AFLLADV and score 22), did not match the assembled transcriptome but a suboptimal ALFLGLD reconstruction with score 20 did (see Figure S11).

The ten reconstructed cyclopeptides (nine highest-scoring reconstruction and one suboptimal reconstruction) matched 11 transcripts (some transcripts encode multiple cyclopeptides and some cyclopeptides are encoded by multiple transcripts) that belong to cyclopeptide-encoding *PawS1-Like* genes in various *Asteraceae* species (Fisher et al., 2018; Jayasena et al., 2017) (see Table S10). While three out of 11 identified *PawL1* ORFs and the four cyclopeptides encoded by them (PLP-12 through PLP-15) have been extensively analyzed in recent studies (Fisher et al., 2018; Jayasena et al., 2017), the remaining eight ORFs represented, to the best of our knowledge, previously unknown cyclopeptide-encoding genes in *S. vulgaris*. See Figure S11.

Table 1. Information about Various High-Resolution Spectral Datasets Analyzed by CycloNovo

Dataset	#Spectra	#Spectra after Preprocessing	#Cyclospectra	#Distinct Cyclopeptides/ Cyclofamilies	#Known Cyclopeptides/ Cyclofamilies
CYCLOLIBRARY	81	81	45	45/27	45/27
S. vulgaris	667	212	23	12/9	4/4
Human stool	1,242,178	451,962	703	79/69	7/5
GNPS	51,220,679	27,883,895	12,004	512/213	67/37
$GNPS_{ACTI}$	5,903,921	4,435,893	1,478	116/56	38/24
$GNPS_{CYANO}$	23,582,408	12,118,482	317	74/35	5/4
$GNPS_{PSEUDO}$	697,812	581,012	2,076	120/39	5/2

The number of distinct cyclopeptides and cyclofamilies was estimated using MS-Cluster (Frank et al., 2011) and SpecNets (Wang et al., 2016), respectively. The last column shows the number of known cyclopeptides/cyclofamilies (identified by Dereplicator) in each dataset. For each identified cyclopeptide in the CYCLOLIBRARY dataset, we selected the PSM with the minimum p value (among all PSMs for that cyclopeptide), resulting in a spectral dataset CYCLOLIBRARY with 81 spectra.

Analyzing the HUMANSTOOL Dataset

A Dereplicator search of the HUMANSTOOL dataset against *CyclopeptideDatabase* identified seven PSMs at 0% FDR, namely an antimicrobial orbitide citrusin V found in various *Citrus* species (Belknap et al., 2015; Noh et al., 2015) and cyclolinopeptides A (Kaufmann and Tobschirbel, 1959), B (Morita et al., 1997), C (Morita et al., 1999), D (Morita et al., 1999), H (Morita et al., 1999), and E (Morita et al., 1999). Cyclolinopeptides are bioactive flaxseed orbitides from *Linum usitatissimum*. The individual who provided the HUMANSTOOL sample frequently ate flaxseeds because they contain α-linolenic acids. CycloNovo sequenced six flaxseed cyclopeptides from the *CyclopeptideDatabase* as well cyclolinopeptide P (a recently discovered cyclopeptide (Okinyo-Owiti et al., 2014) that has not been added to *CyclopeptideDatabase* yet) as the highest-scoring reconstructions (see Table S11).

In addition to the eight reconstructed orbitides, CycloNovo reconstructed 32 cyclopeptides in the HUMANSTOOL dataset with p values below 10^{-15} forming 26 cyclofamilies (see Table S11). Figure 3 shows a connected component in the spectral network formed by four previously unreported cyclopeptides (to the best of our knowledge) in the HUMANSTOOL dataset and illustrates that CycloNovo reconstructions are consistent with the spectral network.

The Dereplicator search of all 703 cyclospectra in the HUMANSTOOL dataset against *CyclopeptideDatabase* resulted in a single hit and identified a cyclic lipopeptide massetolide F (Gerard et al., 1997) with a p value of 7.5×10^{-22}. As this compound includes lipid chains not included in the set of selected amino acids, CycloNovo was not able to generate its full-length reconstructions but correctly reconstructed its partial amino acid sequence (see Table S11).

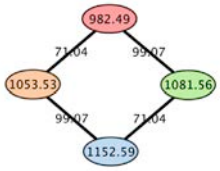
Massetolides are nonribosomal lipopeptides produced by *Pseudomonas fluorescens*, an indigenous member of human and plant microbiota (O'Sullivan and O'Gara, 1992; Scales et al., 2014). Analysis of the metagenome assembly of reads paired with the HUMANSTOOL dataset confirmed that *P. fluorescens* is present in the stool samples where massetolide F was detected. Therefore, massetolide F might be originated from *P. fluorescens* in the human gut (see Table S11), but further investigation is necessary to test this hypothesis.

Analyzing the GNPS Dataset

We analyzed all cyclospectra in the GNPS dataset using MS-Cluster (Frank et al., 2011) and SpecNets (Watrous et al., 2012) with the goal of estimating the number of still unknown cyclopeptides and cyclofamilies originating from spectra already deposited into GNPS. To provide a conservative estimate for the number of cyclopeptides and cyclofamilies, we limited the analysis to clusters with at least three spectra. The 12,004 cyclospectra in the GNPS dataset originated from 512 cyclopeptides and 213 cyclofamilies. A Dereplicator search of these cyclospectra against *CyclopeptideDatabase* identified only 67 cyclopeptides from 37 cyclofamilies (see Figure 4C). For each putative cyclopeptide, we selected a representative spectrum with the highest *k-merScore*, resulting in 512 spectra corresponding to the 512 cyclopeptides. CycloNovo *de novo* sequenced 94 cyclopeptides with p values below 10^{-15} in this set of 512 cyclospectra (Figure S14).

Comparing CycloNovo and Dereplicator

We compared the number of distinct cyclopeptides, including some branch-cyclic peptides, and cyclofamilies revealed by CycloNovo and identified by Dereplicator in searches against the *CyclopeptideDatabase* (Figure 4). As Figure 4 illustrates, even for the extensively studied phyla of *Cyanobacteria* and *Pseudomonas*, only a small fraction of cyclopeptides and cyclofamilies revealed by CycloNovo are currently known. Moreover, CycloNovo revealed many cyclopeptides in known cyclofamilies that were previously unreported (to the best of our knowledge). For example, CycloNovo reconstructed six variants of surugamide by analyzing the $GNPS_{ACTI}$ dataset (see Figures S1 and S2) and revealed the widespread proliferation of the recently described *A-domain skipping* phenomenon (Gurevich et al., 2018; Nguyen et al., 2016), suggesting that it is more prevalent than was previously thought. The A-domain skipping phenomenon defies the traditional view that each A-domain encodes a single amino acid in an NRP according to the nonribosomal code. Genome mining efforts typically rule out such events due to the consecutive arrangements of A-domains in NRP synthetases (see Figure S3). CycloNovo found all the known cyclopeptides identified by Dereplicator in the HUMANSTOOL and S.VULGARIS datasets. For each MassIVE dataset included in



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

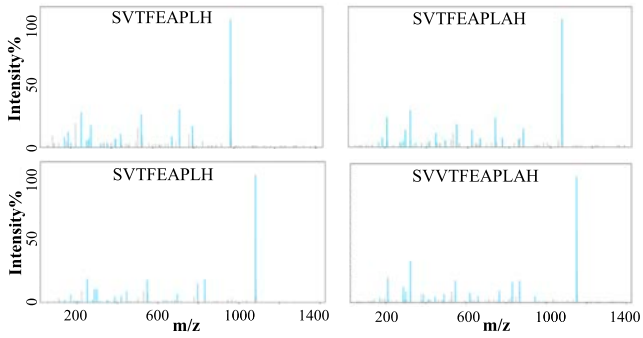


Figure 3. A Previously Unreported Cyclofamily (to the Best of Our Knowledge) Reconstructed by CycloNovo in the HUMANSTOOL Dataset

(Top) Four cyclopeptides reconstructed by CycloNovo form a cyclofamily represented by a connected component in the spectral network of the HUMANSTOOL dataset (label "L" stands for either amino acids L or I). Each node represents a spectrum and two nodes are connected by an edge if their spectral similarity (Wang et al., 2016) exceeds 0.8. The numbers on the edges show the mass shifts between the corresponding spectra. (Middle) the *de novo* reconstructions corresponding to the four spectra forming the spectral network. For each cyclopeptide, the cyclic sequence of the highest-scoring reconstruction along with their scores, number of reconstructions with scores larger or equal to the PSM score (column "#reconstructions with score ≥ PSM score"), and p values are listed. The "dates" column shows the dates when the corresponding samples were taken. Note that the cyclopeptides in this cyclofamily appear on the same dates. (Bottom) the annotated spectra of the four cyclopeptides based on the CycloNovo reconstructions.

the GNPS dataset, Figure 4B and Table S6 presents the number of known cyclopeptides or cyclofamilies identified by Dereplicator and missed by CycloNovo.

DISCUSSION

The advances in shotgun mass spectrometry and the advent of public repositories such as the GNPS molecular network (Mohimani et al., 2017) has created powerful new resources for natural product discovery. But, as the existing computational approaches (Gurevich et al., 2018; Mohimani et al., 2017) are limited to identifying known cyclopeptides and their variants, *de novo* cyclopeptide analysis pipelines are needed to reveal the "dark matter of cyclopeptidomics." To address this bottle neck, we developed CycloNovo, an algorithm for *de novo* recognition and sequencing of cyclopeptides. The *de novo* recognition feature relies on the idea of spectral convolutions and can be used as a stand-alone tool for selection and prioritization of cyclopeptides in large metabolomics datasets for further downstream analyses. After recognition of cyclospectra, we showed how CycloNovo *de novo* sequences those cyclospectra by using a de Bruijn graph representation of them. This representation enables CycloNovo to rapidly find a small set of potential cyclopeptide sequences, by finding cycles in reconstructed graphs, which it then scores against the input spectra.

In this study, we successfully applied CycloNovo to spectral datasets generated from cultured and environmental samples of different origins and phyla and produced in different laboratories and demonstrated that CycloNovo correctly sequenced many known cyclopeptides in a blind mode and reconstructed 400 previously unreported cyclopeptides (to the best of our knowledge) that were validated using (meta)genomic and transcriptomics data. Our study reports the first cyclopeptides found through fully automated *de novo* sequencing of mass spectra. Furthermore, we showed how our *de novo* analysis can provide new insights from metabolomics datasets.

Using the cyclospectra recognition feature, we analyzed ≈51 million spectra from already published datasets on the GNPS molecular network (Gurevich et al., 2018). Our CycloNovo analysis revealed over 400 cyclopeptides that were previously unreported (to the best of our knowledge) in 176 cyclofamilies, while the database search tools identified only 81 (42 cyclofamilies) of known cyclopeptides in this dataset. These findings confirm that a large body of unreported cyclopeptides still exists in already published datasets and on the GNPS network (Mohimani et al., 2017) that is now accessible via CycloNovo for further downstream analysis.

Our CycloNovo analysis of the HUMANSTOOL dataset generated from 65 human stool samples showed that several bioactive cyclopeptides from consumed food remain stable throughout the proteolytic, absorptive, and microbial ecosystem provided by the gastrointestinal system and thus may be interacting with the human microbiome. Our analysis also found cyclospectra originating from the branch-cyclic peptide massetolide produced by an indigenous member of the human microbiota and confirmed by metagenomics analysis. In addition, it revealed a large number of still unknown cyclopeptides in the human gut that are either a part of the human diet or products of the human gut microbiome and provided the composition of some of these peptides including a previously unreported family of cyclopeptides (to the best of our knowledge) present in the stool samples. While cyclopeptides are known for their higher proteolytic stability, their stability levels and reaction to the human intestinal proteases vary (Baeriswyl and Heinis, 2013). Our study is the first demonstration of cyclopeptides surviving gastrointestinal proteases and further highlights the prevalence of bioactive compounds (such as cyclolinopeptides and massetolides) in human gut. Our findings call for subsequent analysis of cyclopeptides as compounds that humans are exposed to in their daily lives through their interaction with the environment or as food.

As with all mass spectrometry-based tools that do not use stable isotope labeling (Bode et al., 2012), CycloNovo is unable to infer stereochemical information and to distinguish leucine from isoleucine. Moreover, CycloNovo is unable to differentiate

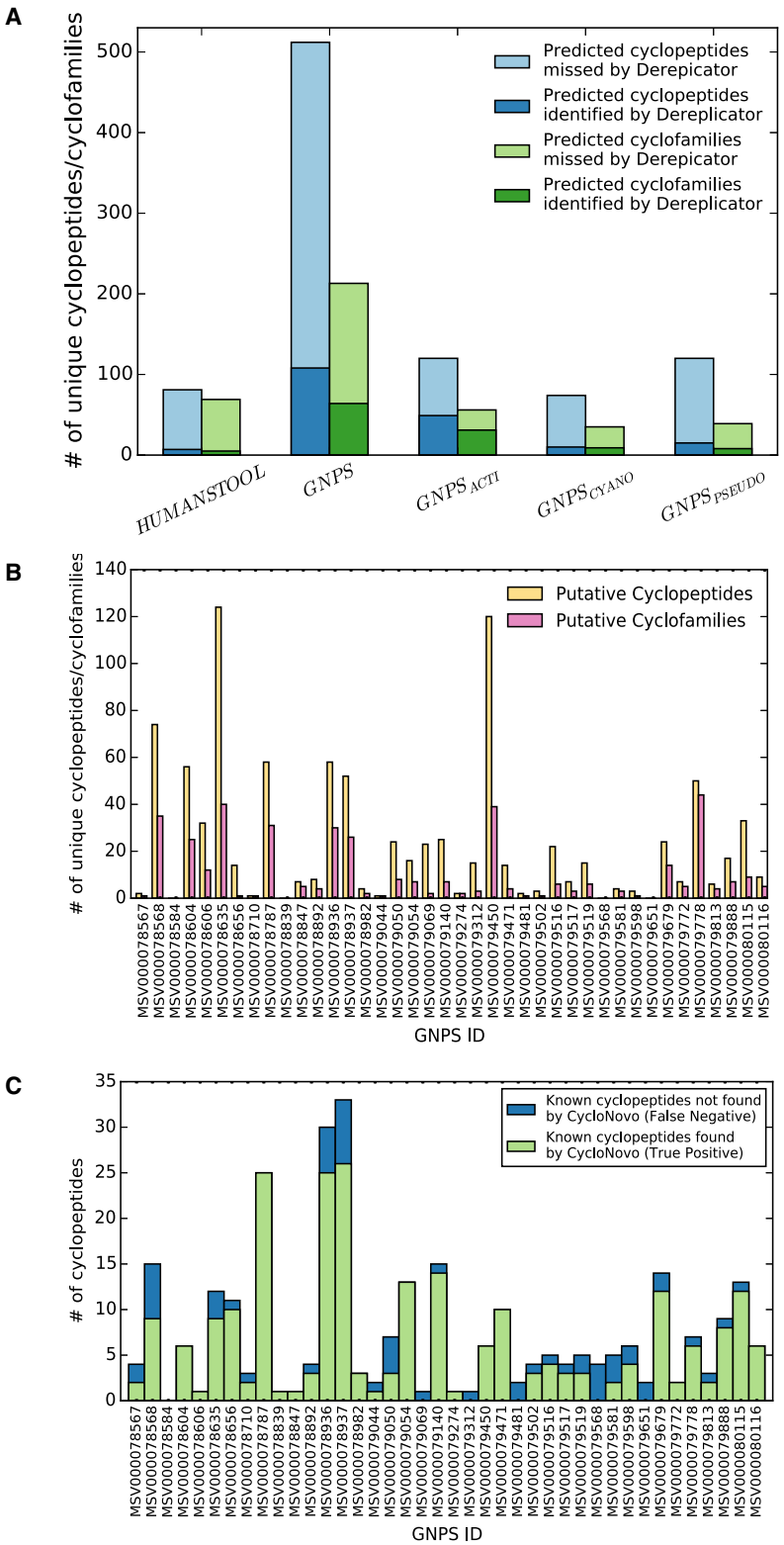


Figure 4. Number of Cyclopeptides Found by CycloNovo across Different Datasets

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

(B) Number of cyclopeptides identified by Dereplicator across all GNPS sub-dataset. Dereplicator identified 81 unique cyclopeptides in the GNPS dataset. Since some cyclopeptides are identified in multiple sub-datasets, the total numbers of identified cyclopeptides across all GNPS sub-datasets (180) exceeds 81.

(C) Number of cyclopeptides (yellow) and cyclofamilies (pink) found by CycloNovo across all GNPS dataset.

between similar cyclopeptides (such as cyclopeptides with some rearranged amino acids) that yield near identical theoretical spectra and are not discernable using our PSM function. However, in these cases, deriving further information about the final structure (for example via nuclear magnetic resonance) is simplified when some partial sequence information is available through CycloNovo predictions. See the STAR Methods section under Limitations of CycloNovo for more information about other possible failure modes of CycloNovo and the limitations of our study.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- LEAD CONTACT AND MATERIALS AVAILABILITY
- METHOD DETAILS
  - Spectral Convolution
  - Recognizing Cyclospectra
  - Estimating the Number of Distinct Cyclopeptides and Cyclofamilies
  - Preprocessing Spectra
  - Selected Amino Acids
  - Analyzing Spectral Convolution
  - Forming CyclopeptideDatabase and CYCLOLIBRARY
  - Computing p Values for *De Novo* Generated Peptide Spectrum Matches
  - Limitations of CycloNovo
- QUANTIFICATION AND STATISTICAL ANALYSIS
- DATA AND CODE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.cels.2019.11.007.

ACKNOWLEDGMENTS

The work of B.B., H.M., and P.A.P. was supported by the US National Institutes of Health (grant 2-P41-GM103484). The work of B.B. was also supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) Graduate Scholarship, Center for Microbiome Innovation (CMI) Grand Challenges Award grant, and CMI Microbial Sciences Graduate Fellowship. The work of A.G. was supported by the Russian Science Foundation (grant 19-16-00049). The work of A.P. was supported by the Russian Science Foundation (grant 19-14-00172). J.S.M. was supported in part by an Australian Research Council Future Fellowship (FT120100013). M.F. was supported by the Australian Government's Research Training Program and a Bruce and Betty Green Postgraduate Research Scholarship. The work of H.M. was supported by a research fellowship from the Alfred P. Sloan Foundation and a National Institutes of Health New Innovator Award (DP2GM137413).

AUTHOR CONTRIBUTIONS

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

DECLARATION OF INTERESTS

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

Received: June 10, 2019
Revised: September 18, 2019
Accepted: November 18, 2019
Published: December 18, 2019

REFERENCES

Arnison, P.G., Bibb, M.J., Bierbaum, G., Bowers, A.A., Bugni, T.S., Bulaj, G., Camarero, J.A., Campopiano, D.J., Challis, G.L., Clardy, J., et al. (2013). Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature. Nat. Prod. Rep. *30*, 108–160.

Baeriswyl, V., and Heinis, C. (2013). Phage selection of cyclic peptide antagonists with increased stability toward intestinal proteases. Protein Eng. Des. Sel. *26*, 81–89.

Bandeira, N., Tsur, D., Frank, A., and Pevzner, P.A. (2007). Protein identification by spectral networks analysis. Proc. Natl. Acad. Sci. USA *104*, 6140–6145.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. *19*, 455–477.

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

Bhushan, R., and Brückner, H. (2011). Use of Marfey's reagent and analogs for chiral amino acid analysis: assessment and applications to natural products and biological systems. J. Chromatogr. B Analyt. Technol. Biomed. Life Sci. *879*, 3148–3161.

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

Compeau, P.E.C., Pevzner, P.A., and Tesler, G. (2011). How to apply de Bruijn graphs to genome assembly. Nat. Biotechnol. *29*, 987–991.

Dančík, V., Addona, T.A., Clauser, K.R., Vath, J.E., and Pevzner, P.A. (1999). De novo peptide sequencing via tandem mass spectrometry. J. Comput. Biol. *6*, 327–342.

Elliott, A.G., Delay, C., Liu, H., Phua, Z., Rosengren, K.J., Benfield, A.H., Panero, J.L., Colgrave, M.L., Jayasena, A.S., Dunse, K.M., et al. (2014). Evolutionary origins of a bioactive peptide buried within Preproalbumin. Plant Cell *26*, 981–995.

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

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

Gerard, J., Lloyd, R., Barsby, T., Haden, P., Kelly, M.T., and Andersen, R.J. (1997). Massetolides A-H, antimycobacterial cyclic depsipeptides produced by two pseudomonads isolated from marine habitats. J. Nat. Prod. *60*, 223–229.

Gurevich, A., Mikheenko, A., Shlemov, A., Korobeynikov, A., Mohimani, H., and Pevzner, P.A. (2018). Increased diversity of peptidic natural products

revealed by modification-tolerant database search of mass spectra. Nat. Microbiol. *3*, 319–327.

Ibrahim, A., Yang, L., Johnston, C., Liu, X., Ma, B., and Magarvey, N.A. (2012). Dereplicating nonribosomal peptides using an informatic search algorithm for natural products (iSNAP) discovery. Proc. Natl. Acad. Sci. U.S.A *109*, 19196–19201.

Jayasena, A.S., Fisher, M.F., Panero, J.L., Secco, D., Bernath-Levin, K., Berkowitz, O., Taylor, N.L., Schilling, E.E., Whelan, J., and Mylne, J.S. (2017). Stepwise evolution of a buried inhibitor peptide over 45 my. Mol. Biol. Evol. *34*, 1505–1516.

Kaufmann, H.P., and Tobschirbel, A. (1959). Über ein oligopeptid aus Leinsamen. Chem. Ber. *92*, 2805–2809.

Kavan, D., Kuzma, M., Lemr, K., Schug, K.A., and Havlicek, V. (2013). CYCLONE - A utility for de novo sequencing of microbial cyclic peptides. J. Am. Soc. Mass Spectrom. *24*, 1177–1184.

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

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

Kim, S., and Pevzner, P.A. (2014). MS-GF+ makes progress towards a universal database search tool for proteomics. Nat. Commun. *5*, 5277.

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

Marahiel, M.A., Stachelhaus, T., and Mootz, H.D. (1997). Modular peptide synthetases involved in nonribosomal peptide synthesis. Chem. Rev. *97*, 2651–2674.

Medema, M.H., and Fischbach, M.A. (2015). Computational approaches to natural product discovery. Nat. Chem. Biol. *11*, 639–648.

Mohimani, H., and Pevzner, P.A. (2016). Dereplication, sequencing and identification of peptidic natural products: from genome mining to peptidogenomics to spectral networks. Nat. Prod. Rep *33*, 73–86.

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

Mohimani, H., Wei-Ting Liu, Y.-L.Y., Gaudêncio, S.P., Dorrestein, W.F., and Pevzner, P.C. (2011b). Multiplex de novo sequencing of peptide antibiotics. J. Comp. Biol. *18*, 1371–1381.

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

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

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

Mohimani, H., Liu, W.T., Kersten, R.D., Moore, B.S., Dorrestein, P.C., and Pevzner, P.A. (2014b). NRPquest: coupling mass spectrometry and genome mining for nonribosomal peptide discovery. J. Nat. Prod. *77*, 1902–1909.

Mohimani, H., Gurevich, A., Mikheenko, A., Garg, N., Nothias, L.F., Ninomiya, A., Takada, K., Dorrestein, P.C., and Pevzner, P.A. (2017). Dereplication of peptidic natural products through database search of mass spectra. Nat. Chem. Biol. *13*, 30–37.

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

Morita, H., Shishido, A., Matsumoto, T., Takeya, K., Itokawa, H., Hirano, T., and Oka, K. (1997). A new immunosuppressive cyclic nonapeptide, cycloinopeptide B from Linum usitatissimum. Bioorg. Med. Chem. Lett. *7*, 1269–1272.

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

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

Nguyen, D.D., Melnik, A.V., Koyama, N., Lu, X., Schorn, M., Fang, J., Aguinaldo, K., Lincecum, T.L., Ghequire, M.G.K., Carrion, V.J., et al. (2016). Indexing the Pseudomonas specialized metabolome enabled the discovery of poaeamide B and the bananamides. Nat. Microbiol. *2*, 16197.

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

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

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

Okinyo-Owiti, D.P., Young, L., Burnett, P.G.G., and Reaney, M.J.T. (2014). New flaxseed orbitides: detection, sequencing, and15N incorporation. Biopolymers *102*, 168–175.

Röttig, M., Medema, M.H., Blin, K., Weber, T., Rausch, C., and Kohlbacher, O. (2011). NRPSpredictor2-A web server for predicting NRPS adenylation domain specificity. Nucleic Acids Res. *39*, W362–W367.

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

Takada, K., Ninomiya, A., Naruse, M., Sun, Y., Miyazaki, M., Nogi, Y., Okada, S., and Matsunaga, S. (2013). Surugamides A – E, cyclic octapeptides with four D-amino acid residues, from a marine Streptomyces sp.: LC-MS-aided inspection of partial hydrolysates for the distinction of D- and L-amino acid residues in the sequence. J. Org. Chem. *78*, 6746–6750.

Townsend, C., Furukawa, A., Schwochert, J., Pye, C.R., Edmondson, Q., and Lokey, R.S. (2018). CycLS: accurate, whole-librarysequencing of cyclic peptides using tandem mass spectrometry. Bioorg. Med. Chem. *26*, 1232–1238.

Wang, M., Wang, J., Carver, J., Pullman, B.S., Cha, S.W., and Bandeira, N. (2018). Assembling the community-scale discoverable human proteome. Cell Syst *7*, 412–421.e5.

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., et al. (2016). Sharing and community curation of mass spectrometry data with global natural products social molecular networking. Nat. Biotechnol. *34*, 828–837.

Watrous, J., Roach, P., Alexandrov, T., Heath, B.S., Yang, J.Y., Kersten, R.D., van der Voort, M., Pogliano, K., Gross, H., Raaijmakers, J.M., et al. (2012). Mass spectral molecular networking of living microbial colonies. Proc. Natl. Acad. Sci. USA *109*, E1743–E1752.

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

Zolg, D.P., Wilhelm, M., Schnatbaum, K., Zerweck, J., Knaute, T., Delanghe, B., Bailey, D.J., Gessulat, S., Ehrlich, H.C., Weininger, M., et al. (2017). Building ProteomeTools based on a complete synthetic human proteome. Nat. Methods *14*, 259–262.

Zorzi, A., Deyle, K., and Heinis, C. (2017). Cyclic peptide therapeutics: past, present and future. Curr. Opin. Chem. Biol. *38*, 24–29.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and Algorithms		
Dereplicator	Mohimani et al., 2017	https://github.com/ablab/npdtools/releases
MS-Cluster	Frank et al., 2011	https://gnps.ucsd.edu/ProteoSAFe/index.jsp?params=%7B%22workflow%22:%22METABOLOMICS-SNETS-V2%22,%22library_on_server%22:%22d.speclibs;%22%7D
Spectral Network	Wang et al., 2016	https://gnps.ucsd.edu/ProteoSAFe/index.jsp?params=%7B%22workflow%22:%22METABOLOMICS-SNETS-V2%22,%22library_on_server%22:%22d.speclibs;%22%7D
SPAdes	Nurk et al., 2017	http://cab.spbu.ru/software/spades/

LEAD CONTACT AND MATERIALS AVAILABILITY

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Bahar Behsaz (bbehsaz@ucsd.edu).

This study did not generate any new materials.

METHOD DETAILS

Spectral Convolution

We represent each spectrum *Spectrum*={s_1, …, s_n} as its *spectral diagram*, the set of $n\times(n-1)/2$ 2-dimensional points (s_i, s_j) for $1 \le i < j \le n$. Given a mass *a*, the convolution of *Spectrum with offset a* (denoted *convolution*(*Spectrum, a*)) is equivalently defined as the number of points in the diagonal (45°) band $y \approx x+a$ in the spectral diagram. Figure S4 presents the spectral diagram of *TheoreticalSpectrum*(AGCD) and reveals that bands corresponding to its amino acids (71, 57, 103, and 115 Da) are the most *populous* (contain a large number of points as compared to other bands), meaning *convolution*(*Spectrum,a*) is high when *a* is the mass of amino acids A, G, C, or D. For example, *TheoreticalSpectrum*(AGCD) includes five pairs of fragment masses ((G,AG), (D,AD), (AGC,GC), (CD,CDA), and (GCD,AGCD)) that are located on the "blue" diagonal $y = x+mass(A)$ in Figure S4.

The spectral diagrams for *TheoreticalSpectrum*(Surugamide) and experimental $Spectrum_{Surugamide}$ highlight four populous diagonal bands $y \approx x+a$, where *a* is the mass of one of four amino acids in surugamide with integer masses 71, 113, 128, and 147 (see STAR Methods). These populous bands in the spectral diagram reveal the masses of amino acids in an unknown cyclopeptide that gave rise to an experimental spectrum.

Figure S4 lists high-multiplicity clusters for $Spectrum_{Surugamide}$ (see Supplementary Note "Analyzing spectral convolution") and shows that many of them have masses that are similar to the masses of amino acids in surugamide. Since populous diagonals (high-multiplicity clusters) in the spectral diagram reveal amino acids in the unknown cyclopeptide that gave rise to an experimental spectrum, we use them to generate the set of putative amino acids (Compeau et al., 2011).

Recognizing Cyclospectra

A cluster in the spectral convolution is called *frequent* if its multiplicity exceeds the *cluster multiplicity threshold* (the default threshold for $Spectrum_{Surugamide}$ is 7). CycloNovo classifies a spectrum as a cyclospectrum if the number of frequent cyclopeptidic clusters in its spectral convolution is at least *minNumberFrequentClusters* (the default value *minNumberFrequentClusters*=2). Since there exist two frequent cyclopeptidic clusters for $Spectrum_{Surugamide}$ (corresponding to amino acids I/L and K), it is classified as cyclopeptidic (Figure S4). In addition to $Spectrum_{Surugamide}$, out of 938 spectra passing the preprocessing step in the small spectral dataset for *Streptomyces CNQ329* that contains $Spectrum_{Surugamide}$, CycloNovo recognized only one cyclospectrum, also originated from surugamide.

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

repeat units form high multiplicity clusters with masses equal to the masses of a selected amino acid, triggering a possibility to misclassify a polymer spectrum as a cyclospectrum.

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

we analyzed two spectral datasets described below:

- LINEARLIBRARY is a set of 105,871 Collision-Induced Dissociation (CID) tandem mass spectra of distinct linear peptides from the MassIVE Knowledge-Based (MassIVE-KB) spectral library (Wang et al., 2018) of linear peptides distilled from all human proteomics data in the MassIVE database. In particular the CID library under MassiveIVE-KB is the annotated spectra from CID datasets generated from a collection of over 330,000 synthetic tryptic peptides representing almost all of the canonical human gene products as a synthetic human proteome collection (Zolg et al., 2017). In this study, pools of synthetic peptides were subjected to LC-MS/MS analysis using CID fragmentation method coupled with ion trap or Orbitrap readouts.
- POLYMERLIBARY is a set of 448 tandem spectra generated from polyethylene glycol (MSV000081544).

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

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

- High multiplicity cyclopeptidic clusters test (distinguishing cyclospectra from spectra of linear peptides): As described in the main text, CycloNovo first selects a spectrum for further analysis if its spectral convolution has at least *minNumberFrequentClusters* frequent cyclopeptidic clusters, i.e., clusters with multiplicities exceeding the cluster multiplicity threshold. Since the cluster multiplicities typically increase with the increase in the length of a peptide, this threshold increases with the increase in the peptide mass. We thus defined the *cluster multiplicity threshold* as $\alpha \times precursorMass + \beta$ (see below for selecting parameters α and β).
- *Polymer test (distinguishing cyclospectra from polymer spectra):* For each cyclospectrum *Spectrum*, CycloNovo analyzes clusters with masses of repeat units observed in background contamination from polyethylene glycol, NaCl, polypropylene glycol, and trimethylsiloxane (44.03, 57.96, 58.04, and 72.04 Da, respectively). We refer to these masses as *polymeric units* (Keller et al., 2008) and refer to clusters with masses equal to polymeric units as *polymer-clusters.* CycloNovo classifies a spectrum as polymeric if there exist at least *minNumberFrequentClusters* polymer-clusters with multiplicities at least the *cluster multiplicity threshold.* Polymeric spectra are filtered out from the set of found cyclospectra.
- *cycloIntensity test:* For each cyclospectrum, CycloNovo considers all frequent cyclopeptidic clusters. For each such cluster of mass *a*, we consider all pairs of masses *x* and *y* in the spectrum contributing to this cluster, i.e., satisfying the condition $y \approx x+a$. The *cyclointensity* of the spectrum, referred to as *cycloIntensity*, is defined as the total intensity of all such peaks (across all frequent cyclopeptidic clusters) divided by the total intensity of all peaks in *Spectrum.* Spectra with cyclointensity below the *cycloiIntensity threshold* are filtered out.
- *k-merScore test:* CycloNovo computes the *k-merScore*, the score of the highest-scoring k-mer that contributes to the de Bruijn graph of the spectrum and filters out cyclospectra with *k-merScore* below the *k-merScore threshold.*

Selecting thresholds for recognizing cyclospectra. To select the default value of *cluster multiplicity threshold*=$\alpha \times$precursorMass+β, we varied parameters α (from 0.005 to 0.02) and β (from -5 to +5) and analyzed all found cyclospectra in the CYCLOLIBRARY, LINEARLIBRARY, and POLYMERLIBARY datasets (Figure S5). Despite its smaller size, CYCLOLIBRARY is the only dataset where CycloNovo recognizes cyclospectra for all analyzed values of α and β. Since α=0.07 and β=-1 yielded the largest number of recognized cyclospectra in CYCLOLIBRARY (46 out of 81) and no cyclospectra in the LINEARLIBRARY and POLYMERLIBRARY datasets (Figure S5), we selected these values as the default parameters.

Figure S6 presents the values of *cycloIntensity* and *k-merScore* for each spectrum in the CYCLOLIBRARY, LINEARLIBRARY, and POLYMERLIBRARY datasets and reveals a separation between the former and the two latter datasets with respect to these two parameters. CycloNovo thus classifies a spectrum as a cyclospectrum if its *cycloIntensity* exceeds the *cycloIntensity* threshold (60%) and its *k-merScore* exceeds the *k-merScore* threshold (5). 45 spectra in the CYCLOLIBRARY datasets, that pass all four tests described above, are classified as cyclospectra.

We also investigated how CycloNovo's ability to recognize a cyclospectrum is affected by the fragmentation quality of the corresponding PSM (measured by the p-value of this PSM). For each spectrum in the CYCLOLIBRARY dataset, we identified the minimum value of the parameter α that leads to classifying this spectrum as cyclopeptidic (for β=-1). Figure S7 illustrates that well-fragmented spectra can be recognized even with more restrictive threshold values (larger values of α).

Estimating the Number of Distinct Cyclopeptides and Cyclofamilies

Spectral datasets often contain multiple spectra originating from the same compound. CycloNovo clusters similar cyclospectra using MS-Cluster (Frank et al., 2011) and estimates for the number of distinct cyclopeptides as the number of constructed clusters. CycloNovo further constructs the spectral network of cyclospectra using SpecNets (Wang et al., 2016) and estimates for the number of distinct cyclofamilies as the number of connected components in this network.

Preprocessing Spectra

Similar to pre-processing practices in proteomics (Frank et al., 2011), CycloNovo filters out low-intensity peaks in each spectrum by retaining at most 5 peaks with the highest intensities in each 50 Da window. CycloNovo further filters out all peaks that are less than 0.05 Da apart from another peak with higher intensity. It further removes spectra with a small number of peaks (less than 20) and spectra with a small precursor mass (less than 500 Da). We subtract the mass of a hydrogen atom from all masses in the spectrum (for simplicity, we assume that each ion is protonated with a single proton).

Selected Amino Acids

Table S2 lists all amino acids selected by CycloNovo in the default mode and the RiPP mode. To curb the run-time and memory footprint of CycloNovo, we limited the set of amino acids to those that appear in at least *fraction* of the peptides in *CyclopeptideDatabase* and the standard amino acids, yielding the 25 most frequent amino acids listed in Table S2 (for the default value *fraction*=0.02). Table S3 illustrates that the runtime and memory footprint rapidly increase with the increase in the number of selected amino acids for $Spectrum_{surugamide}$. The user can change the default number of selected amino acids by specifying the argument '—topaminoacids *N*' to select the *N* most frequent amino acids in *CyclopeptideDatabase* ($N \leq 40$). Moreover, CycloNovo has an additional feature for studying RiPPs (the '—ripps' argument) by considering amino acids that are common in RiPPs (Arnison et al., 2013) even though they do not appear frequently in the entire *CyclopeptideDatabase* (see Table S2 for the list of these amino acids and their frequencies).

Analyzing Spectral Convolution

Figure S8 illustrates that each amino acid in surugamide results in a populous diagonal in the spectral diagram of $Spectrum_{Surugamide}$. For each constructed cluster (diagonal band in the spectral diagram), we consider all pairs of masses in *Spectrum* that contributed to this cluster and form a *band* as the set of these *k* pairs.

We define the *cluster diameter* as the difference between its maximum and minimum elements. Figure S9 presents the band for the cluster with multiplicity 8 and mass 128.09 (diameter 0.03) in the spectral convolution of $Spectrum_{Surugamide}$ and reveals that the 8 elements of this band can be partitioned into 7 groups of closely located points. We are interested in the number of such groups (rather than the raw cluster multiplicities) since experimental spectra often contain *satellite masses* resulting from *neutral losses* and *isotopic peaks.* For example, in addition to the integer mass 242 Da corresponding to the peptide IK, $Spectrum_{Surugamide}$ also contains the integer mass 225 Da corresponding to the loss of NH_3 from this peptide.

Since satellite masses artificially inflate cluster multiplicities, there is a need to reduce biases caused by these masses. We thus define the set of common satellite offsets (1 Da (H), 18 Da (H_2O), 17 Da (NH_3), and 28 Da (CO)) and perform additional single linkage clustering in each populous band by combining pairs of masses in a single cluster if both their *x*-coordinates and *y*-coordinates differ by a satellite offset. We redefine the concept of cluster multiplicity as the number of the resulting clusters in the band (Table S4).

Forming CyclopeptideDatabase and CYCLOLIBRARY

Information about CyclopeptideDatabase. The *PNPDatabase* (Gurevich et al., 2018) combines all known peptidic natural products from various databases. Many peptides in this database are *lipopeptides* containing a lipid chain, e.g., surfactin is a cyclopeptide containing a fatty acid side chain connected to a fully peptidic part via a peptide bond.

We classify a peptide in the *PNPdatabase* as a cyclopeptide if its backbone could be represented as a circular graph (cycle) with nodes corresponding to either a single amino acid or a single lipid tail (i.e. monomers) and edges corresponding to the amide bonds in the peptide structure. 1,257 out of 5,021 peptides in the *PNPDatabase* represent cyclopeptides and form *CyclopeptideDatabase* (note that the *CyclopeptideDatabase* database contains lipopeptides).

Information about the CYCLOLIBRARY dataset. We searched ~130 million GNPS spectra against the *CyclopeptideDatabase* using Dereplicator (Mohimani et al., 2017) and identified 81 distinct cyclopeptides (41 cyclofamilies) corresponding to PSMs with FDR=0% and P-value below 10^{-15}. For each identified cyclopeptide, we selected the PSM with the minimum P-value (among all PSMs identified for this cyclopeptide), resulting in a set of 81 PSMs and hence created a spectral dataset CYCLOLIBRARY with 81 spectra (Table S5). CYCLOLIBRARY includes only 13 cyclopeptides (6 cyclofamilies) that are made up entirely of selected amino acids (Table S5). 34 peptides (25 cyclofamilies) in the CYCLOLIBRARY dataset contain lipid tails and 34 peptides (14 cyclofamilies) contain non-selected amino acids.

Computing p Values for *De Novo* Generated Peptide Spectrum Matches

The PSM score, while informative, is biased with respect to length. Therefore, CycloNovo computes a P-value to evaluate the statistical significance of each individual PSM formed. Given a PSM(*P,S*) between a cyclopeptide *P* and a spectrum *S*, MS-DPR (Mohimani et al., 2013) computes the probability (p-value) that a random peptide forms a PSM with the spectrum *S* with a score that is greater than or equal to the score of PSM(*P,S*). Therefore, the p-value statistics is not biased by length.

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

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

Note that computing p-values of *de novo* reconstructions of linear peptides does not make sense since there exists an efficient dynamic programming algorithm for finding a *linear* peptide with maximum score against a spectrum *S* (Dančík et al., 1999). Indeed, the peptide with maximum score against a spectrum *S,* over the set of all possible peptides, corresponds to the most extreme point in the probability distribution and therefore its P-value is exactly 0.

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

Limitations of CycloNovo

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

a. Cyclopeptide has monomers not included in the initial amino acid list: either they have modifications that are not considered by CycloNovo or they include non-proteinogenic amino acids not considered by CycloNovo.
b. Cyclopeptide includes side chains (or lipid chain). Specifically, CycloNovo often fails to reconstruct branch-cyclic peptides with long side chains (peptidic or non-peptidic) or multiple side chains, where only a small portion of amino acids are appearing in their ring, especially if the total mass of the ring is smaller compared to the total mass of the cyclopeptide. The spectral convolution analysis step of such peptides typically fails because their cyclopeptidic convolution does not meet the threshold set by their peptide mass as the mass representing the ring is much smaller than the intact peptide mass.
c. Cyclopeptide has many fragments with similar masses (hence similar fragments): since the number of unique peaks in the spectrum matching the theoretical spectrum of such cyclopeptide is low, there is insufficient amount of information in their corresponding spectrum. Therefore, the cyclospectrum might fail the spectral convolution or *k-merScore* tests.
d. Cyclopeptide is a hybrid compound and contains large non-peptidic elements, for example polyketide-NRP hybrid or a lipopeptide with large lipid chains (see the explanation B about side chains above)
e. Cyclopeptide includes bonds other than amide bonds (for example disulfide bonds).
f. Cyclopeptide only is represented in spectra with charge states larger than 2: CycloNovo only consider spectra with charge states 1 and 2.

Note that, although the mentioned features hinder the recognition and sequencing of cyclospectra by CycloNovo, CycloNovo may perform well on cyclopeptides with above conditions. We distinguish between *simple* cyclopeptides (like surugamide) and *complex* peptides such as peptides with monomers not included in the initial set of amino acids or the ones with a complex backbone structure such as branch-cyclic peptides, lipopeptides, depsipeptides and others. As mentioned by the reviewer, CycloNovo has two main applications: recognizing a cyclospectrum and sequencing it. With respect to sequencing, CycloNovo can output either a complete (for simple cyclopeptides) or partial cyclopeptide reconstruction (for complex cyclopeptides). Table S5 shows the number of correct *k*-mers predicted by CycloNovo (using only the default set of amino acids) versus the length of the cyclopeptide. This column highlights that CycloNovo reports at least one correct k-mer for many lipopeptides (19 out of 34) although it is not designed for complex compounds. With respect to cyclospectrum recognition, CycloNovo successfully recognized many spectra that originated from complex peptides as cyclospectra (see Table S6).

Although we were able to predict many putative cyclopeptides with extremely low p-values and confirmed several of them with accompanying datasets, but we did not have the resources to experimentally validate our putative cyclopeptides through NMR or cyclopeptide synthesis.

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical methodologies and software used for performing various types analysis in this work are cited where appropriate in the STAR Methods text. The p value calculations were done using MS-DPR software.

DATA AND CODE AVAILABILITY

CycloNovo is available as both a stand-alone tool (https://github.com/bbehsaz/cyclonovo) and as a web application (https://gnps.ucsd.edu/ProteoSAFe/index.jsp?params=%7B%22workflow%22:%22CYCLONOVO%22%7D). All described datasets are available through the corresponding public repositories via provided accession numbers.