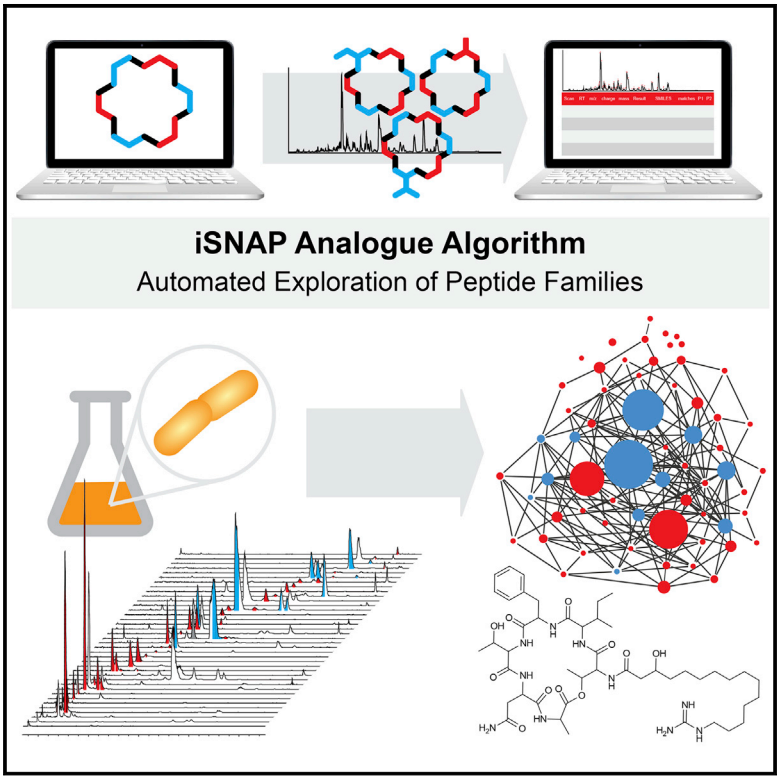
Resource

Chemistry & Biology

Exploration of Nonribosomal Peptide Families with an Automated Informatic Search Algorithm

Graphical Abstract



Highlights

- An informatic tool for automated discovery of peptide natural product variants
- Elaboration of extensive natural product families from microbial extracts
- Discovery of over 70 novel variants from undeveloped scaffolds
- Discovery of novel glycosylated arylomycin variants with improved activity

Authors

Lian Yang, Ashraf Ibrahim, Chad W. Johnston, Michael A. Skinnider, Bin Ma, Nathan A. Magarvey

Correspondence

magarv@mcmaster.ca

In Brief

Yang et al. report an automated untargeted informatic strategy to screen for variants in peptide natural product families. This discovery strategy led to the identification of over 70 novel unreported variants and one having greater potency.

Yang et al., 2015, Chemistry & Biology 22, 1–11
September 17, 2015
http://dx.doi.org/10.1016/j.chembiol.2015.08.008

CellPress

Please cite this article in press as: Yang et al., Exploration of Nonribosomal Peptide Families with an Automated Informatic Search Algorithm, Chemistry & Biology (2015), http://dx.doi.org/10.1016/j.chembiol.2015.08.008

Chemistry & Biology
Resource

Exploration of Nonribosomal Peptide Families with an Automated Informatic Search Algorithm

Lian Yang,[1,4] Ashraf Ibrahim,[2,4] Chad W. Johnston,[3,4] Michael A. Skinnider,[3] Bin Ma,[1] and Nathan A. Magarvey[2,3,*]
[1]The David R. Cheriton School of Computer Science, University of Waterloo, Waterloo, ON N2L 3G1, Canada
[2]Department of Chemistry and Chemical Biology, McMaster University, Hamilton, ON L8N 3Z5, Canada
[3]The Michael G. DeGroote Institute for Infectious Disease Research, Department of Biochemistry and Biomedical Sciences, McMaster University, Hamilton, ON L8N 3Z5, Canada
[4]Co-first author
*Correspondence: magarv@mcmaster.ca
http://dx.doi.org/10.1016/j.chembiol.2015.08.008

SUMMARY

Microbial natural products are some of the most important pharmaceutical agents and possess unparalleled chemical diversity. Here we present an untargeted metabolomics algorithm that builds on our validated iSNAP platform to rapidly identify families of peptide natural products. By utilizing known or in silico-dereplicated seed structures, this algorithm screens tandem mass spectrometry data to elaborate extensive molecular families within crude microbial culture extracts with high confidence and statistical significance. Analysis of peptide natural product producers revealed an abundance of unreported congeners, revealing one of the largest families of natural products described to date, as well as a novel variant with greater potency. These findings demonstrate the effectiveness of the iSNAP platform as an accurate tool for rapidly profiling large families of nonribosomal peptides.

INTRODUCTION

Small-molecule natural products have long served as a valuable source of pharmaceuticals, providing molecular scaffolds useful in treating an ever-expanding range of pathologies. Microbial natural products, particularly the polyketides and nonribosomal peptides, have proved particularly useful as a source of antibacterial drugs and scaffolds, making up roughly two-thirds of clinical antibacterials since 1981 (Newman and Cragg, 2007). Polyketides and nonribosomal peptides are produced by modular assembly line-like enzymes (PKSs and NRPSs) that frequently display promiscuity in substrate selection and chemical tailoring reactions, giving rise to molecular families based on a set scaffold. By utilizing this promiscuity, a single assembly line can deploy a library of molecules that can possess divergent affinities for a given target, or even different activities (Yu et al., 2012). As the forces that drive the evolution of these molecules in the environment are typically not directed toward clinical needs, minor analogs can often be identified with substantially improved clinical utility, including drugs and leads such as mannopeptimycin (He et al., 2002), rhizoxin (Scherlach et al., 2006), epothilone (Chou et al., 1998), pneumocandin (Balkovec et al., 2014), and burkholdine (Lin et al., 2012). Similarly, minor differences in homologous assembly lines (observable in sequenced genomes and metagenomes) can produce related natural products with superior activity (Wang et al., 2011). In each instance, variations in substrate- or monomer-selection promiscuity by NRPS/PKS enzymes can lead to the generation of natural biosynthetic libraries with a range of target affinities and biological effects. Traditionally, natural product discovery efforts have led to the most abundant members being targeted through bioassay-guided fractionation, as bioactivity is concentration dependent. Such limited sampling can lead to an underestimation of a scaffold's potential as a therapeutic agent, as superior minor variants may go undiscovered. Reanalysis and reengagement of overlooked natural product scaffolds will likely facilitate the discovery of analogs with improved pharmaceutical properties, providing new clinically useful structures and scaffolds for further development.

Traditional natural products chemistry approaches based on bioactivity-guided isolation and rigorous structure elucidation techniques have provided the vast majority of pharmaceutically relevant natural products. Unfortunately, these techniques are biased toward abundant compounds and necessitate large amounts of material, expertise, and time, often requiring months of work for the identification, isolation, and elucidation of a single natural product. Reanalyzing microbial culture extracts for overlooked molecules will require techniques that can provide systems-level analysis and define both structures and retention times to prioritize and facilitate subsequent isolation efforts. This requirement is complicated by the diversity of natural product families, whose exotic molecular scaffolds confound the development of automated analyses, and whose frequent isobaric species prevent accurate and automated structure elucidation. While genome sequencing data can provide predictive value in guiding discovery efforts (Kersten et al., 2011), obtaining sufficiently well-assembled biosynthetic gene clusters for accurate structure predictions can be time consuming, impeding discovery and development efforts. Pure, untargeted metabolomic approaches that make use of liquid chromatography-coupled mass spectrometry (LC-MS) remain a powerful means of rapidly assessing chemical potential (Winnikoff et al., 2014; Yang et al., 2013; Hou et al., 2012), and are uniquely

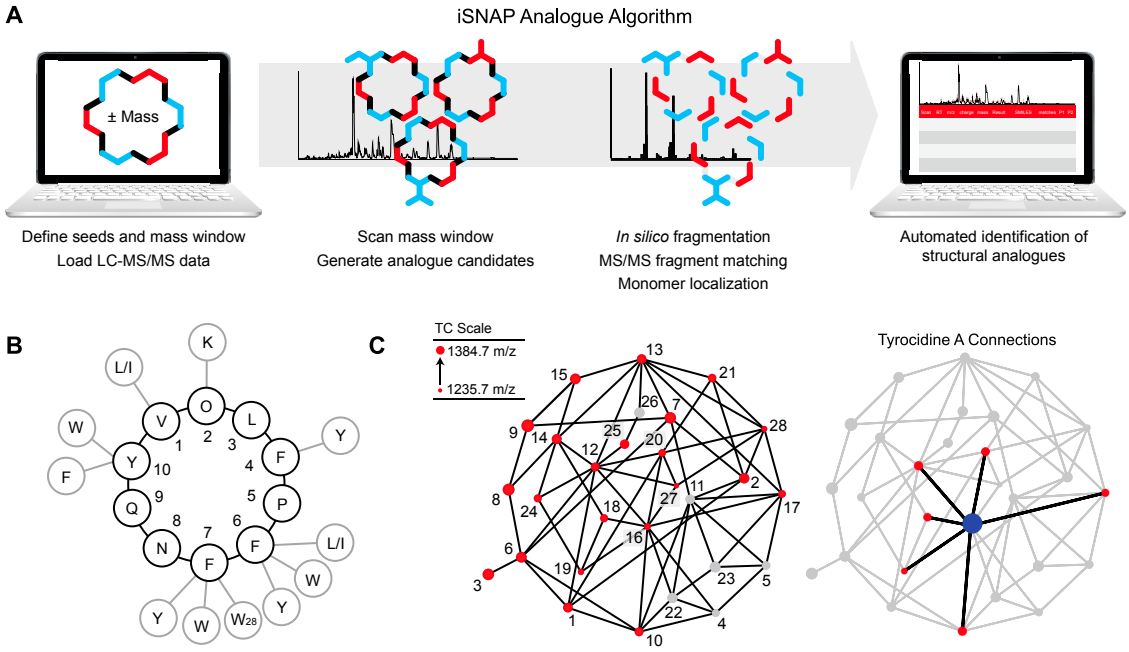


Figure 1. iSNAP Analog Discovery Algorithm Can Map Families of Natural Peptides

(A) Schematic workflow of the automated iSNAP analog algorithm. After uploading an LC-MS/MS data file (.mzXML), users are required to define seed structures of interest—either predefined or following initial iSNAP dereplication analysis—in addition to a mass window around these seeds. Following dereplication, the algorithm searches the defined mass window and generates candidate analog structures. Hypothetical candidates are fragmented in silico to facilitate MS/MS fragment matching and monomer localization, driving the automated identification of structural analogs from the initial seed.

(B) Amino acid composition of tyrocidine A (tyrocidine 16; black) and various substitutions found in the remaining 27 known members of the tyrocidine family of cyclic peptide natural products (gray).

(C) Tyrocidine single-monomer substitution networks. Structures of tyrocidines 1–28 dereplicated by iSNAP (red) or known from previous literature (gray) are connected to one another through single amino acid substitutions (left). Through iSNAP analog analysis, single seeds can be used to access related structures, demonstrated using tyrocidine A (right).

positioned to facilitate high-throughput, information-rich analysis of complicated microbial extract libraries with high sensitivity. Mirroring advances in LC-MS-based proteomics, a number of targeted metabolomic approaches have been explored for the detection of peptide natural products (Kersten et al., 2011; Mohimani et al., 2011, 2014a, 2014b), which reliably fragment along amide bonds but frequently possess complicated architectures. Following pioneering work in 2009 defining a de novo means of sequencing cyclic peptides (Ng et al., 2009), a number of approaches with varying degrees of automation have worked toward rapid, accurate detection of complex peptide sequences, first from standards (Mohimani et al., 2011) and then from microbial culture extracts (Kersten et al., 2011; Mohimani et al., 2014a, 2014b). In 2012 we defined iSNAP, a novel algorithm capable of detecting known peptide natural products from complex culture extracts (Ibrahim et al., 2012). Taking advantage of a comprehensive in-house database of known peptide natural product structures, this automated "dereplication" technology screens LC-coupled tandem MS (MS/MS) data and applies a series of statistical processes to identify known compounds based on matching peaks between real and in silico MS/MS fragments. Here, we present a new analog module for the iSNAP platform (available at http://magarveylab.ca/analog/), facilitating both dereplication and highly accurate analog identification, which is capable of discerning between isobaric species and automatically providing structures for novel, superior analogs from nanograms of material.

RESULTS

iSNAP Analog Search Algorithm

iSNAP analog search is an algorithm designed for the discovery of novel analogs of known peptide natural products from tandem mass spectral data. It is built on the iSNAP platform to extend the platform's capability of elucidating families of peptide natural products. As it is known that families of peptide natural products can co-exist in microbial culture extracts, the dereplication of a peptide is often a good indication of the existence of analogs that have similar structures. In line with this observation, the algorithm is designed to utilize dereplicated peptides as seed structures to guide the search for analogs with similar structures.

The analog search algorithm takes tandem mass spectral data and a list of seed structures as inputs. It analyzes each tandem mass spectrum individually in the following steps: (1) construct analog candidates that are one monomer different from seed structures; (2) match the spectrum to hypothetical spectra of analog candidates; (3) evaluate the matches by calculating p-value-derived scores, P1 and P2; (4) report the identified analog, if the best match has scores above specified thresholds (Figures 1A and S1). The algorithm reports the seed structure, the

monomer site of difference, and the mass difference for each identified analog.

Construction of Analog Candidates

Assuming the parent mass of a tandem mass spectrum is M and the mass values of seed $S_1, S_2, \ldots, S_n$ are $m(S_1), m(S_2), \ldots, m(S_n)$, the algorithm first compares M with the mass of each seed. A seed is selected to generate analog candidates if the mass difference is smaller than a user-specified threshold M_T. We denote the selected seeds as $\hat{S}_1, \hat{S}_2, \ldots, \hat{S}_m$, and we have

$$\hat{S} \in \{S_i \mid |M - m(S_i)| < M_T\}, \quad \text{where} \quad i = 1, \ldots n.$$

For each selected seed $\hat{S}$, the algorithm annotates amide bonds and ester bonds in its structure. Monomer blocks between two adjacent bonds are therefore detected and numbered as $R_1, R_2, \ldots, R_r$. The algorithm assumes that the true structure for the spectrum is an analog that can be constructed from the seed by modifying one monomer. The modified monomer accounts for the mass difference between the seed and the spectrum parent mass. As such, analog candidates are generated by adding the mass difference to each and every monomer block in the seed.

$$\begin{aligned} \hat{S}_1' &\leftarrow R_1', R_2, \ldots, R_r \\ \hat{S}_2' &\leftarrow R_1, R'_2, \ldots, R_r \\ &\vdots \\ \hat{S}_r' &\leftarrow R_1, R_2, \ldots, R'_r \end{aligned} \quad \text{where} \quad m(R_i') = m(R_i) + M_\Delta.$$

For each generated analog candidate, there is one monomer R_i with altered mass value to make up for the mass difference M_Δ. This ensures the mass of every analog candidate matches with the spectrum parent mass M. The process only involves mass calculation, and the program does not attempt to make structural interpretation for the modified monomer, as this is challenging to decipher using MS/MS alone. Therefore the program does not need a monomer database, such as Norine, for generating analogs, providing an unbiased method for identifying substitutions and monomers.

Evaluation of Analog Candidates

Having generated analog candidates from selected input seeds, a scoring mechanism is needed to evaluate the significance of matching between a spectrum and an analog candidate. In this work, we adopted the scoring system validated in the original iSNAP database search algorithm. A match between a spectrum and analog candidate is evaluated with three scoring metrics: raw score, P1 score, and P2 score.

Raw score is a basic spectral-matching score. The algorithm generates hypothetical spectral fragments of the analog candidate using the original iSNAP platform. The mass-to-charge ratios of these hypothetical fragments are matched to the spectrum. The raw score is then calculated as the logarithmic sum of the peak intensity of all the matched peaks.

P1 score is a statistical normalization of the raw score. By scoring the spectrum with all compounds in the dereplication database, we acquire a raw score distribution of random matches. For an analog candidate, a p value is calculated by its raw score on this distribution. As P1 score is subjected to the composition of the database, having similar structures in a database would slightly skew the distribution and affect the p value. A P2 score was introduced to alleviate this issue. By shifting all peaks in the spectrum, a list of false spectra is created to match with the compound, generating a distribution for p-value calculation. The calculation of P2 score is independent of other database structures. Both P1 and P2 score are converted from p values for better readability using the formula P1 = −10log10(p value). A higher P1 score indicates lower probability to have a random structure matching better with the spectrum. A higher P2 score indicates lower probability to have a random spectrum matching better with the analog candidate.

For each spectrum, P1 score and P2 score are calculated for every analog candidate. Analog candidates are deemed significant if having both P1 score and P2 score higher than the specified thresholds. By default, the thresholds for P1 and P2 are set to 27 and 24, respectively, as empirically determined in our previous work. The thresholds can be adjusted by the user to give more flexibility in result filtering.

Elaborating the Tyrocidine Family

To test whether this new algorithm would be capable of accurate analog detection, we chose to analyze crude extracts of *Bacillus parabrevis*, which produces the tyrocidines, one of the most diverse and well-annotated natural product complexes known. The tyrocidine family of cyclic decapeptides comprises 28 known structural variants (Tang et al., 1992) made up of amino acid substitutions within the peptide core (Figure 1B), with relative abundances ranging from ~1% to 100%. An examination of the chemical space of the tyrocidines (Figure 1C) reveals the single-monomer interconnectivities or "network" of the tyrocidine scaffold, highlighting the correlations between the structures in terms of monomer position and mass difference. As a first step, we optimized our LC conditions with a focus on the use of high-efficiency core-shell columns. This proved useful in resolving many of the co-eluting peaks, allowed for greater confidence of low-abundance compounds, and avoided excess artifact hits within the ion trap (Figure S2). Following established fermentation and extraction conditions (Tang et al., 1992), LC-MS/MS analysis was performed followed by automated metabolite dereplication by iSNAP, to validate the number of tyrocidine compounds within the crude extract. Over a 37-min interval, iSNAP dereplicated 21 of the 28 reported structures (Figure 1C and Table S1). While several of the low-abundance tyrocidines were not detected, this is likely a result of minor discrepancies in culture conditions affecting metabolite production or analytes being below the intensity threshold of the automated MS/MS settings, as manual investigation also failed to reveal the missing, previously reported structures. Having identified 21 tyrocidines within the fermentation culture, we used these dereplicated spectra as references for evaluating analog identification and testing the monomer localizations within the tyrocidine single-monomer substitution network (Figure 1C). The LC-MS/MS data file was rescreened 21 times with the iSNAP analog search algorithm, using each dereplicated structure as the seed. For each seed structure processed, only the analog identifications corresponding to the reference spectra are evaluated. By comparing the analog identifications with the dereplicated structures, we can determine whether a correct identification is

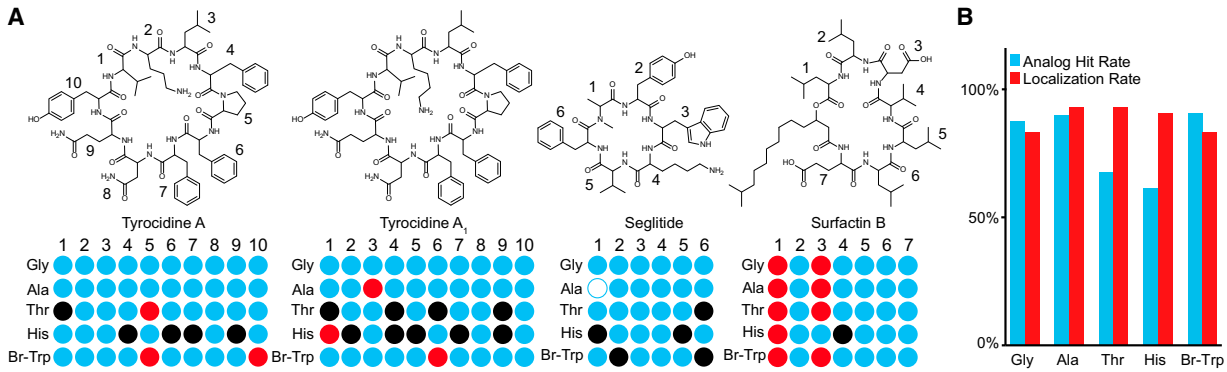


Figure 2. Analog Identification and Localization

(A) In silico candidate analog screening with predefined monomer substitutions. To evaluate analog identification and correct monomer localization rates, hypothetical seeds of cyclic peptides were substituted with glycine, alanine, threonine, histidine, or 5-Br-tryptophan at each position and used to identify their corresponding standard within LC-MS/MS data. Positively identified analog candidates with correct localization are shown in cyan, analog candidates with incorrect localization are shown in red, and non-matching seed structures (unidentified) are shown in black.

(B) Analog identification and localization rates from in silico candidate screening of 17 dereplicated tyrocidines. A total of 17 tyrocidine compounds have been evaluated, representing candidates with relative abundances of ~1%–100%. A total of 170 candidate seed structures are evaluated for each substituted monomer, representing a total of 850 analog screens. The analog hit rate represents the total number of analog seed structures correctly matched to an MS/MS spectrum (localized and mislocalized), scored above the P1 and P2 cutoffs, and divided by the total number of matches possible. The localization rate is the total number of positive localizations divided by the total number of analog matches.

made. As we had expected, the iSNAP analog program correctly matched each MS/MS spectra with its corresponding analog seed structure (Figures 1C and S3). As an example, using tyrocidine A (TC#16) as the seed, the analog search made correct identifications on the MS/MS spectra of tyrocidines #10, 12, 17, 18, and 19, with mass differences and localization correctly identified; +39 *m/z* at monomer 7, +39 *m/z* at monomer 6, +14 *m/z* at monomer 2, +23 *m/z* at monomer 10, and −16 *m/z* at monomer 10, respectively. By making use of MS/MS data, the iSNAP analog program was capable of reliably distinguishing between isobaric analogs: (1) TC #10 and 12, difference in one site modification Phe_6-Trp_7 and $Tryp_6$-Phe_7, respectively; (2) TC #18 (Phe_6-Phe_7,Trp_{10}) and TC-24 (Trp_6-Phe_7,Phe_{10}). These reliably accessed networks suggest that through iterative dereplication and analog steps, it is possible to access all known tyrocidines following the dereplication of a single structure (Figures 1C and S3). These examples demonstrate the utility of the iSNAP analog program, having elaborated the tyrocidine family of cyclic peptides and correctly localized sites of variation within the scaffold.

In Silico Investigation of Analog Identification and Localization Rate

We next sought to establish a controlled testing scenario in which we could evaluate, with confidence, the effectiveness of iSNAP's ability in making sensitive analog identifications, and determine the accuracy with which the algorithm can localize structural differences. More specifically, the analog identification and localizations rates can provide a performance indicator or a measure of how the algorithm may perform when evaluating LC-MS/MS datasets in a real analog searching scenario.

In this experiment, we evaluate LC-MS/MS and MS/MS datasets using the iSNAP analog search algorithm to first search and dereplicate any known NRPS structures. We then artificially create, in silico, new analog seed structures of the dereplicated knowns, by making alterations to the NRPS scaffold, one monomer site at a time. In this testing scenario, the iSNAP analog search algorithm should be capable of correctly identifying all of the artificial in silico seed structures as being analogous to the known NRPSs. Moreover, it should be capable of accurately localizing the position of the altered sites within the NRPS scaffold. By evaluating a series of NRPS structures from crude microbial extracts we can realize, with high confidence, iSNAP's potential as being a sensitive and accurate discovery tool within this controlled setting.

For this experiment, we first manually created a series of in silico seed structures for 17 of the dereplicated tyrocidines by substituting a new monomer block at each of the ten residue positions. The selected tyrocidines represent those with low and high abundance (Figures 2, S4, and S5), avoiding any bias toward those with higher-quality MS/MS spectra. We next incorporated a range of monomers that are not present within the tyrocidine family scaffold. As the tyrocidine family scaffold comprises only 12 amino acid building blocks (Figure 1B), we selected five different amino acids as in silico substitutions for the artificial seed structures, representing increasing mass-to-charge values: glycine, alanine, threonine, histidine, and a 5-Br-tryptophan. We selected the new monomers to avoid any bias toward common fragmentation losses or conserved sequences as well as any bias associated with small or large mass differences that might be more readily matched within the MS/MS spectra of tyrocidines.

For the 17 dereplicated tyrocidines, these newly created and artificial seed structures represent a total of 850 in silico variants being evaluated, with 50 per structure (10 positions × 5 substitutions), and 170 variants globally per substitution (Figures 2, S4, and S5; Table S3). Using the constructed seed structures, we can then use the iSNAP analog search algorithm and examine the analog identifications made to the dereplicated tyrocidine spectra. From these results, analog identification rates and monomer localization rates can be established, providing a

diagnostic for the algorithm's performance. The analog identification rate for the in silico variants is calculated by the total number of analog seed structures correctly matched to an MS/MS spectrum (localized and mislocalized), scored above the P1 and P2 cutoffs, and divided by the total number of matches possible. The monomer localization rate is calculated as the total number of positive localizations divided by the total number of analog matches. As we expected, variations in the analog identification rate could be seen across the substituted amino acids, resulting in a global analog identification rate of 79.41% (675/850) for the tyrocidine family (Figures 2 and S5; Tables S2 and S3). On an individual basis, the analog identification rates are 87.65%, 90.00%, 67.65%, 61.18%, and 90.59% for glycine, alanine, threonine, histidine, and 5-Br-tryptophan, respectively. Rewardingly, of the analogs identified the global monomer localization rate is more than 88% (595/675), with rates of 83.22%, 92.81%, 93.04%, 90.38%, and 83.12%, for glycine, alanine, threonine, histidine, and 5-Br-tryptophan, respectively. To expand on these initial findings, we further probed the molecular scaffolds of several other cyclic peptides (WS9326a, seglitide, surfactin B) to evaluate analog identification rates and monomer localization rates, using the five substituted monomers as in silico seeds. In the case of WS9326a, seglitide, and surfactin, analog identification rates were ~73%, 87%, and ~98%, with monomer localization rates of ~59%, 100%, and 87%, respectively (Figures 2 and S5). These findings demonstrate the effectiveness of the iSNAP analog algorithm in identifying analog variants and localizing structural differences with high accuracy.

Iterative Analog Analysis Maps the Loloatins, Rare Tyrocidine-like Cyclic Peptides

Building on our tyrocidine network findings, we sought to identify small-molecule products of a tyrocidine-like NRPS in the genome of the previously uninvestigated isolate *Brevibacillus laterosporus* (DSM 25). *B. laterosporus* was cultured, extracted, and subjected to LC-MS/MS analysis, providing an.mzXML data file that could be analyzed by the iSNAP algorithm to identify peptide natural products. Rather than the well-studied tyrocidines, iSNAP LC-MS/MS analysis revealed the loloatin-related natural products, which possess two constitutive substitutions that differentiate them from the tyrocidines (Gerard et al., 1999). These peptides possess considerably improved activity relative to the tyrocidines, against multidrug-resistant Gram-positive and Gram-negative bacteria (Gerard et al., 1999). In contrast to the well-studied tyrocidines, only four loloatin structures have been previously reported, and of these we only observed production of loloatin A. To identify novel variants, we screened the culture extract using the four known structures (loloatins A–D) as seeds for iterative analog profiling. Following each round of analysis, analog identifications were confirmed by manual MS/MS examination and annotation (Appendix 1), with correctly identified analogs forwarded as seeds for the next round of discovery. After four rounds of iterative iSNAP analog search, we identified a total of 33 new loloatins (Figure 3), including 22 which have been identified as unique structures, along with nine structures which are presumably isomeric analogs, and two structures that could not be confirmed by manual MS/MS annotation due to poor-quality MS/MS spectra. Of these 33 identified loloatins, 25 demonstrated correct monomer localizations (Appendix 1). Only three MS/MS scans were detected as false positives throughout the analysis, and with low P scores (average FP P1/P2 of 24.5/23.4; global P1/P2 of 61.4/30.1). By iterating our analog approach, we successfully identified an entire family of loloatins comprising 22 distinct, novel variants, revealing an extensive network similar to that of the tyrocidine family.

Automated Analoging Reveals a Massive Family of Lipodepsipeptides

Next, we chose to examine a more modified peptide architecture, one incorporating acyl and ester moieties. We focused our efforts on the LI-F0 series of lipodepsipeptides from *Paenibacillus polymyxa*. Originally described in 1987 (Kurusu et al., 1987), these nonribosomal hexapeptides possess a rare guanidinylated acyl tail, and demonstrate considerable variability in the incorporation of hydrophobic amino acids (Val/Ile and Val/Ile/Phe/Tyr) at positions 2 and 3. These substitutions, along with variable Asn/Gln incorporation at position 5, have led to the elucidation of 12 reported structures, although 16 distinct molecules are theoretically possible with modifications at these sites. These structures have not yet been discovered through earlier MS/MS works (Kuroda et al., 2001) or synthetic biology efforts (Han et al., 2012), likely a result of their low abundance. To assess whether the iSNAP analog search algorithm could display sufficient sensitivity to identify these proposed analogs and reveal any unforeseen variants, we cultured 6 l of *P. polymyxa* and extracted the supernatant for iSNAP analysis. As we had expected, each of the 12 known LI-F0 structures was reliably dereplicated, with retention times relatively similar to those in published works (Kurusu et al., 1987) and MS/MS fragmentation patterns consistent with the LI-F0 scaffold (Figure 4, Appendix 2). Surprisingly, the iSNAP analog search also revealed the presence of an entire suite of unreported LI-F0 variants (Figure 4). iSNAP analysis of the *P. polymyxa* crude extract revealed 39 novel variants, detected over 42.4 min (Figure 4), making this, to our knowledge, the largest natural product complex ever discovered from a single organism, with more than 50 structures identified (Figure 4; Appendix 2). During the analysis no other secondary metabolites were detected, while 10 of 4,339 scans (average FP P1/P2 of 24.8/23.4; global P1/P2 of 29.0/36.5) were false positives and attributed to artifacts, as revealed by manual inspection (Appendix 2). Furthermore, seven compounds were correctly identified as LI-F0 antibiotics by iSNAP but were incorrectly annotated due to convoluted fragmentation or insufficient abundance; three were assigned by manual MS/MS annotation. This complex series of structures appears to arise from several recurring variations, including amino acid substitutions at every position (such as Ser/Thr1, Val/Ile/Leu2, Val/Ile/Leu/Phe/Tyr3, Ser/Thr4, Asn/Gln/Glu5, Ala/Gly6), linearization of the macrocycle ester, loss of the terminal alanine, or addition of a second C-terminal monomer (Ala or Gly). The four analogs that complete the original set arising from combinatorialization of monomers at positions 2, 3, and 5 were also observed, with preliminary quantification efforts demonstrating an isolatable yield of 1 μg/l. The obscurity of these variants is evident in both the LC-MS/MS chromatogram and the corresponding LI-F0 family network (Figure 4C),

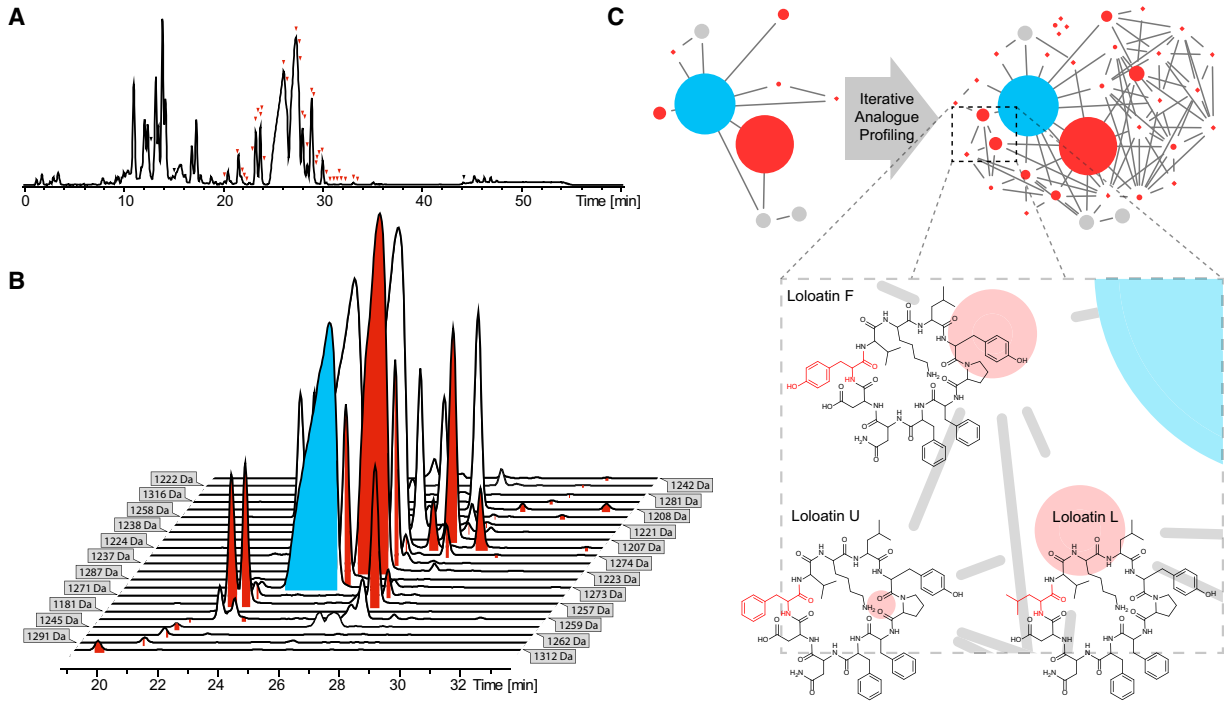


Figure 3. Iterative Informatic Exploration of Loloatin Chemical Space

(A) Informatic detection of a large family of loloatin natural products from a crude extract of *B. laterosporus* (DSM 25). Red arrowheads denote unique loloatins identified by iSNAP. Black arrowheads denote natural products that were falsely identified.

(B) Extracted ion chromatograms of loloatin natural products. Each MS/MS scan determined by iSNAP to contain new loloatin structures is colored red, while MS/MS scans containing known loloatins are colored cyan.

(C) Expansion of known loloatin chemical space through four rounds of iterative iSNAP analog identification. Loloatin species are represented as nodes sized by abundance and connected by single-monomer substitutions. Known loloatins identified in this work are shown in cyan, while new loloatins identified in this work are shown in red. Previously discovered loloatins that were not observed in this work are shown in gray. The magnified section of the analog network depicts structural alterations linking three new loloatins.

thus underscoring the analytical and discovery value of informatic search approaches in revealing previously undiscovered structures.

Identification and Elucidation of a Minor Arylomycin Variant with Improved Bioactivity

Following our lipodepsipeptide screening, we next chose to investigate glycosylated natural products for new, minor variants that may yield superior bioactivity. We analyzed a series of extracts from environmental Actinomycetes to identify candidate natural products, and detected a suite of glycosylated arylomycins (Figure 5A), which had previously been reported by researchers at Eli Lily as potent inhibitors of type 1 signal peptidase with antibacterial and β-lactam-sensitizing activity (Kulanthaivel et al., 2004). iSNAP analog analysis of these partially cyclic, glycosylated lipopeptides identified a total of six novel variants, with alterations in acyl tail length and phenylglycine dihydroxylation. With these new findings, we then reinvestigated our crude extracts for additional analogs using the novel structures as seeds. Prior to running the iSNAP analog search algorithm, we first dereplicated the novel structures to further validate and confirm the annotations. Next, we expanded the analog mass tolerance window to facilitate the discovery of more divergent analogs. During this second round of study, the iSNAP analog search identified an additional four congeners, corresponding to aglycones of the original series (Schimana et al., 2002), which could be confirmed by multistage mass spectrometry (MS^n) analysis (Figure 5; Appendix 3). Previous SAR studies on this antibacterial complex identified that decreasing acyl tail lengths led to increased activity against Gram-negative bacteria such as *Escherichia coli* (Kulanthaivel et al., 2004). To assess whether the shorter acyl tails observed in our novel minor variants led to improved activity, we isolated one of these structures (molecular weight [MW] 1,000 Da; isolated at ~17 μg/l) for bioactivity testing and structure elucidation alongside the previously described parent structure (MW 1,014 Da; isolated at ~ 220 μg/l). High-resolution mass spectrometry (HRMS) measurement of the novel variant matched our expectation as glycosylated arylomycin, having a shorter acyl tail ($[M + H]^+$: $C_{49}H_{73}N_6O_{16}$; 1.198 ppm error; Appendix 4). However, comparison of the ^1H, ^{13}C-heteronuclear multiple-bond correlation, ^1H, ^1H-correlation spectroscopy, and ^1H, ^1H-total correlation spectroscopy experiments between the 1,014-Da parent and the 1,000-Da analog revealed differences in the N-methylated N-terminal amino acid, which is not readily seen within the MS/MS spectra. In contrast to the N-methyl serine observed in the parent structure, the minor novel variant possesses both an N-methylated threonine and shorter acyl tail (C12 versus C14). With the exception of the chemical

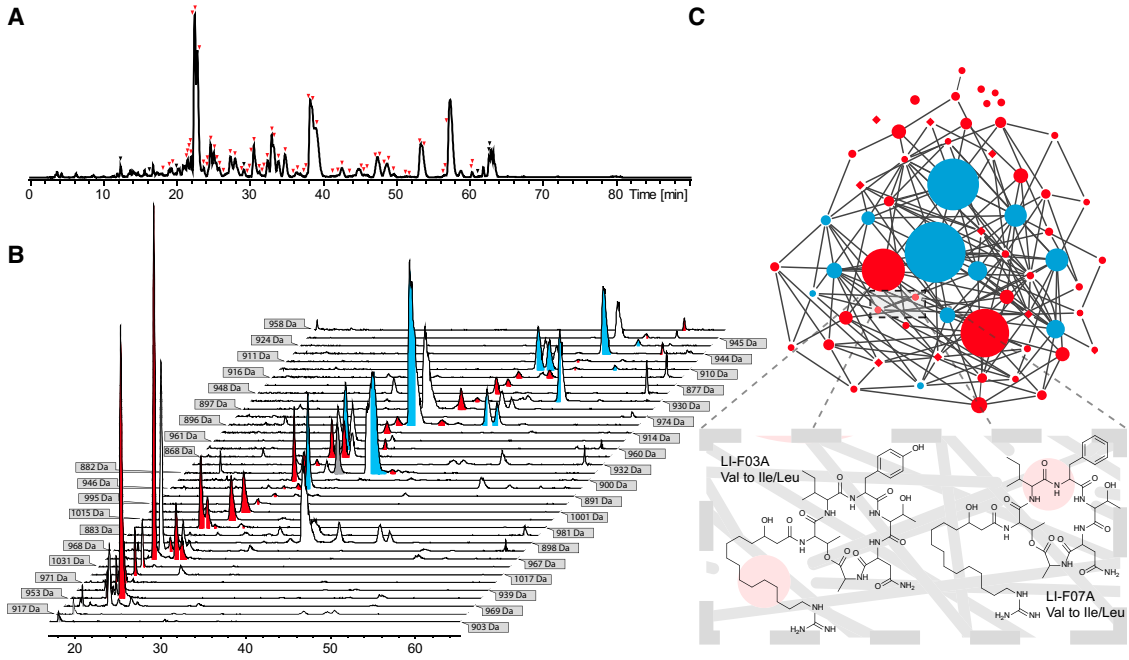


Figure 4. iSNAP-Driven Discovery of a Massive Family of LI-F0 Series Natural Products

(A) Informatic detection of a massive family of LI-F0 series lipodepsipeptides from a crude extract of *P. polymyxa* (ATCC no. 21,830). Red arrowheads denote unique LI-F0 series products identified by iSNAP. Black arrowheads denote natural products that were falsely identified.

(B) Extracted ion chromatograms of LI-F0 series molecules. Each MS/MS scan determined by iSNAP to contain new LI-F0 series structures is colored red, while MS/MS scans containing known LI-F0 series molecules are colored cyan. Scans containing molecules incorrectly identified as LI-F0 series compounds are shown in gray.

(C) Expansion of known LI-F0 series chemical space through iSNAP analog identification. LI-F0 species are represented as nodes sized by abundance and connected by single-monomer substitutions. Known LI-F0 series molecules identified in this work are shown in cyan, while new LI-F0 series molecules identified in this work are shown in red. The magnified section of the analog network depicts structural alterations linking two new LI-F0 series structures.

shifts associated with the substituted amino acid, our findings are consistent with the fully resolved and elucidated parent structure (Appendix 4). To assess whether this minor variant possessed superior activity to the more abundant parent compound, we selected a sensitive *E. coli* test strain, similar to that used originally by researchers at Eli Lily (Kulanthaivel et al., 2004). Microdilution minimum inhibitory concentration (MIC) assays of our novel arylomycin variants against *E. coli* revealed that, consistent with previous findings (Kulanthaivel et al., 2004), our short-tail variant possessed modestly greater activity compared with the parent compound, with MIC values of 0.6 and 0.4 μg/ml, respectively.

DISCUSSION

During the golden age of natural product antibiotic discovery, researchers identified a staggering number of unique chemical scaffolds, including several that have endured as modern therapeutic agents (Newman and Cragg, 2007). However, the vast majority of these bioactive molecules were discarded, as apparent chemical diversity provided more efficacious or easily developed leads, such as glycopeptides and macrolide antibiotics. Daptomycin (originally LY 146032) was one such discarded scaffold that went through years of genomic and metabolomic engineering to emerge as an immeasurably useful antibiotic (Eisenstein et al., 2010). As resistance to conventional antibiotic scaffolds continues to rise, reinvestigation of overlooked natural products may prove to be a valuable method for meeting clinical demand.

Microbial natural products are thought to be exceptionally bioactive because they possess evolved chemical scaffolds that provide fitness benefits to the producer organism within their native environment. While these evolved molecules provide excellent leads for drug development, properties that are useful in a natural environment are often at odds with our desires for their use as pharmaceutical agents. Given this disparity, there is not necessarily a direct correlation between the abundance of a given congener and its activity toward a pathogen of interest (He et al., 2002; Balkovec et al., 2014; Lin et al., 2012; Gerard et al., 1999; Kurusu et al., 1987). Given the intrinsic promiscuity of many biosynthetic enzymes and the resultant chemical diversity observed in natural product extracts, metabolomic investigations of undeveloped scaffolds can lead to the discovery of new variants with improved pharmaceutical potential. Fast and reliable automated methods are now sorely needed to mine large extract libraries and expand the chemical space of clinically promising pharmacophores.

LC-MS/MS is currently the benchmark standard for rapidly profiling complex natural product extracts, and MS/MS can be used to mine vast quantities of rich data to identify chemical entities. During the past decade a number of methodologies have been developed that make use of de novo MS/MS sequencing

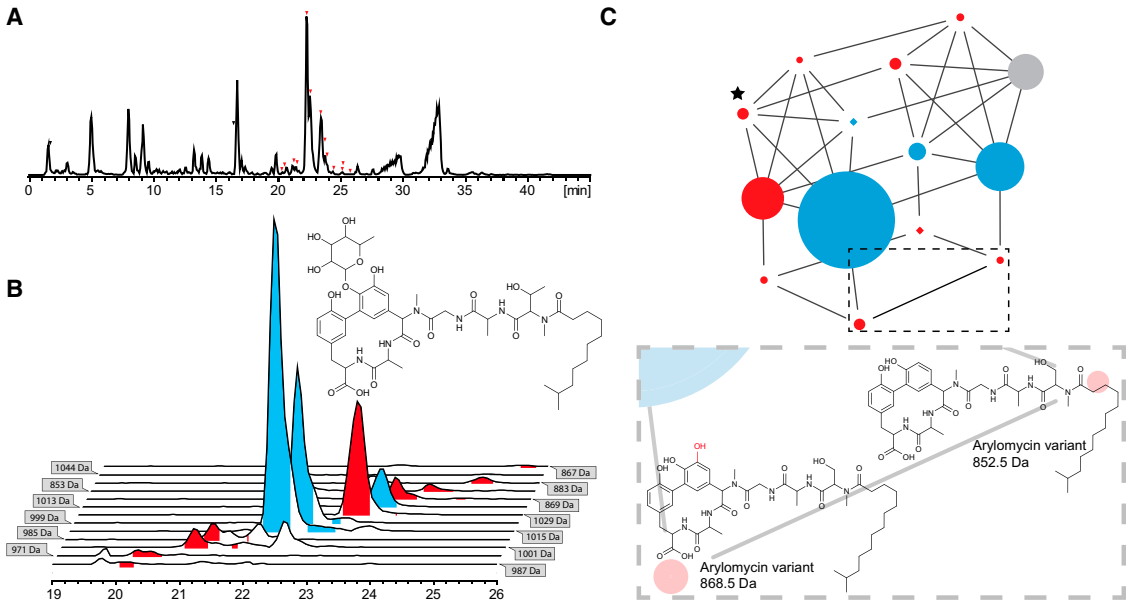


Figure 5. Expansion of Glycosylated Arylomycin Chemical Space Facilitates the Discovery of an Analog with Improved Bioactivity

(A) Informatic detection of a family of arylomycin natural products from a crude extract of environmental Actinomycete isolate NAM12. Red arrowheads denote unique arylomycins identified by iSNAP. Black arrowheads denote natural products that were falsely identified.

(B) Extracted ion chromatograms of arylomycins. MS/MS scans determined by iSNAP to contain new arylomycin structures based on the dereplicated glycosylated arylomycin scaffold are shown in red, and scans containing known glycosylated arylomycins are shown in cyan. A new analog (1001 Da [M + H]; inset) was isolated and its structure elucidated by HRMS, 1D NMR, and 2D NMR experiments.

(C) Expansion of known arylomycin chemical space through two rounds of iterative iSNAP analog identification. Arylomycin species are represented as nodes sized by abundance and connected by single-monomer substitutions. Known glycosylated arylomycins identified in this work are shown in cyan, while previously unobserved glycosylated arylomycins and corresponding aglycones identified in this work are shown in red. Previously discovered glycosylated arylomycins that were not observed in this work are shown in gray. The new, more active arylomycin variant is denoted with a star. The magnified section of the analog network depicts structural alterations linking two arylomycin aglycones.

techniques (Ng et al., 2009; Medema et al., 2014) and genomic data (Kersten et al., 2011; Mohimani et al., 2014a, 2014b), in efforts to guide the discovery of ribosomal and nonribosomal peptides. In the case of iSNAP, by making use of a comprehensive library of peptide natural product structures, we can utilize our validated statistical algorithms (Ibrahim et al., 2012) to expand the known chemical space of key natural products identified in an untargeted and automated manner. Because the iSNAP analog program is supplied as a user-friendly Web application, it provides a more straightforward and directed means of identifying the structures and locations of families of peptide natural products using LC-MS/MS data from crude microbial extracts. This pure metabolomics approach capitalizes on the extensive knowledge base of natural product chemistry and is not necessarily tied to genomic data, which is not always available for extant microbial extract libraries.

We have demonstrated that the iSNAP analog function can provide consistent and accurate results using low-resolution LC-MS/MS data of crude extracts containing low concentrations of natural products. This is in contrast to previously published de novo sequencing approaches that required pure, cyclic peptide standards that were directly infused for high-resolution MS^n analysis (Ng et al., 2009). While de novo sequencing represented a pioneering achievement compared with earlier works, the challenges of this technique remain, including issues with mixtures of "direct sequence ions" and "nondirect sequence ions" in cyclic peptides, where multiple ring-opening events can occur and complicate sequencing. While early programs could be designed for analysis of pure standards of cyclic peptides (Ng et al., 2009; Mohimani et al., 2011), they were not applicable to complex mixtures containing other architectures of nonribosomal peptides. In addition, while foundational early work demonstrated effective use of informatic search strategies to correctly identify structures from large databases (Mohimani et al., 2011), peptide standards were typically pretreated with chemical reducing agents to increase fragmentation, an approach that is not amenable to metabolomics studies of natural product extracts.

One important application of the iSNAP analog algorithm is the rapid analysis of peptide natural product mixtures and the identification of related molecular families. Visualization of natural product families can also be performed using molecular networking approaches (Watrous et al., 2012; Nguyen et al., 2013) that cluster observed ions, based on similarities in their mass spectral fragmentation patterns. Fragment commonalities are scored using cosine vectors (Frank et al., 2008) that ease the clustering of related spectra, but do not provide a means to identify individual compounds with high specificity or without extraneous analysis. Unlike molecular networking approaches, the iSNAP analog function provides results with retention times and confidence scores for individual scans, and indicates peptide identities without the need for genetic knockouts (Watrous

et al., 2012) or direct comparison with known compounds (Yang et al., 2013) or known producers (Nguyen et al., 2013). In contrast to molecular networking approaches, iSNAP also does not merge or discard seemingly identical spectra (Watrous et al., 2012; Frank et al., 2008), and thus allows for the detection of distinct isobaric species at different retention times, allowing users to take full advantage of optimized chromatography. The iSNAP analog program also does not require manual annotation of MS/MS data to assign amino acids (Kersten et al., 2011; Medema et al., 2014) or monomer modifications (Watrous et al., 2012), and has been designed to handle proteinogenic and non-proteinogenic amino acids, allowing it to function as an information-rich and rapid means of detecting families of nonribosomal peptides. These differences allow the iSNAP analog program to define important information about specific peptides—as well as general trends about peptide families—with speed and accuracy superior to previously published methods. Despite this, there remain limitations to the iSNAP analog algorithm and MS-based approaches in general. First, although MS can provide a wealth of information about the nature and identity of molecules or monomers units, more comprehensive spectroscopic techniques such as nuclear magnetic resonance (NMR) are still required to define exact structural features. Second, sensitivity of LC-MS-based detection is often tied to the ability of a given molecule to be ionized, meaning that some molecules are observed more easily than others. We have conclusively demonstrated that the iSNAP analog algorithm can map peptide families, but the identification of new analogs is currently limited by alterations to a single-monomer site, so low-abundance analogs with multiple independent substitutions remain challenging to correctly detect and define. In regard to the scoring scheme, P1 and P2 scores served their purpose as quality indicators for individual matches, but are not corrected for multiple testing. The algorithm can be improved if it also reports an estimated false discovery rate for identification results. In addition, this approach is currently limited to peptidic natural products, which possess reliable and predictable potential fragmentation patterns, providing sufficient data for the highly effective analog detection and elucidation presented here.

In this work, we investigated the tyrocidine family of cyclic peptides and demonstrated that the iSNAP analog search algorithm correctly identified all analog compounds for each dereplicated tyrocidine. Through an in silico screening approach, we have shown that our discovery method is highly sensitive in analog identification and highly accurate in site-specific monomer localization. We have also demonstrated iSNAP's discovery potential by identifying more than 125 peptide structures, of which 70 are novel variants, from a series of cyclic, lipo-, and glycopeptides. While the extensive genetic diversity observed in microbial life offers a glimpse of new, promising leads for natural product discovery, we have demonstrated here that the chemical diversity of lone biosynthetic assembly lines can provide new, improved variations of desired scaffolds that parallel this boundless potential.

SIGNIFICANCE

In this work, we present the iSNAP analog method as an effective strategy for the untargeted discovery of new nonribosomal peptides from crude microbial extracts. This approach does not require the aid of genomic sequence information, HRMS systems, or prior knowledge of the sample for positive analog identifications. iSNAP's automated analog processes have been applied to several potent antimicrobial producers, leading to the discovery of more than 70 novel unreported peptide variants including one with improved potency, all without the use of bioassay-guided isolation. HRMS, MS^n analysis, and 1D/2D NMR measurements of isolated variants further establishes iSNAP's analog capabilities as a true discovery tool.

EXPERIMENTAL PROCEDURES

General

LC-MS/MS data acquisition was obtained on a Bruker Amazon-X ion-trap mass spectrometer coupled to a Dionex Ultimate 3000 high-performance liquid chromatography (HPLC) system running under Hystar 3.2 control with Trap control 7.0 and Chromeleon 6.2. Spectra were generated using electrospray ionization and under collision-induced dissociation (N_2 nebulizer gas: 25 psi; He dry gas: 7.5 psi; temperature: 250°C). Automated MS^n acquisitions were performed from 450 to 1600 *m/z* for tyrocidines, across a scan range of 100–2,000 *m/z* using the Enhanced Resolution setting. MS/MS analysis parameters included isolation width at n = 4, precursor ions at n = 10, threshold cutoff from 400 to 600,000, active exclusion at n = 4 spectra over 20 s, and CID fragmentation set to 1.25 V across a voltage sweep of 25%–200%. Bruker raw data files were converted to .mzXML format using Bruker conversion software, CompassXport, prior to iSNAP Analog analysis (http://www.ionsource.com/functional_reviews/CompassXport/CompassXport.htm). For analytical flow rates a UV/MS flow splitter of 10:1 was used. LC-MS spectral analysis was performed using Compass DataAnalysis 4.1 (Bruker). HRMS measurements were performed in positive electrospray ionization using a Bruker maXis 4G UHR-TOF mass spectrometer coupled to a Dionex Ultimate 3000 HPLC system running Hystar 3.2 control and under standardized LC conditions, with calibrations done using sodium formate. 1D and 2D NMR measurements were acquired using a Bruker Avance III 700-MHz NMR spectrometer equipped with a 5-mm QNP cryoprobe, operating at 700.17 MHz for ^1H NMR and 176.08 MHz for ^{13}C NMR. Chemical shifts were referenced to the internal solvent peaks: 3.31 ppm (^1H) and 49.00 ppm (^{13}C) for CD_3CD. Surfactin (S-3523) and seglitide (S-1316) standards were purchased from Sigma-Aldrich.

Microbial Strains

B. laterosporus and *B. parabrevis* were obtained from the German Resource Center for Biological Material (DSMZ, DSM no. 25 and 362, respectively). *P. polymyxa* was obtained from the American Type Culture Collection (ATCC; no. 21,830). Environmental Actinomycete NAM12 was isolated from an environmental soil sample collection performed at McMaster University. *B. laterosporus*, *B. parabrevis*, and *P. polymyxa* were maintained on Luria-Bertani agar plates at 30°C. NAM12 was maintained on Bennett's agar plates at 30°C.

Cytoscape Depiction of Tyrocidine, Loloatin, LI-F0, and Arylomycin Chemical Space

Cytoscape plots were assembled using Cytoscape 3.0.1, imported from a manually curated network file of for each peptide, connected through single amino acid substitutions. Nodes were automatically distributed by preset preferred layout and manually adjusted to account for subsequent resizing. An exported .pdf file was manipulated using Adobe Illustrator CS6, resizing nodes according to their relative abundance, and in the case of the tyrocidines by relative size.

In Silico Scanning Analysis

Compound structures were modified in ChemBioDraw Ultra 13.0 to substitute an a glycine, alanine, threonine, histidine, and Br-tryptophan at each of the ten monomer positions for the tyrocidines, six positions for seglitide, seven for surfactin B, and six for WS9326a. The structures were converted to SMILES

codes and compiled into an uploadable text file for analysis. The LC-MS/MS chromatograms of the tyrocidines, WS9326a, seglitide, and surfactin were then analyzed by iSNAP analog using each of the monomer-substituted analog candidates as individual seeds via the uploaded text files. Reports were analyzed manually to identify whether an analog candidate of the monomer-substituted structures was matched to the MS/MS scans containing the appropriate peptide structure.

Additional experimental procedures can be found in the Supplemental Information.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, five figures, three tables, and four appendices and can be found with this article online at http://dx.doi.org/10.1016/j.chembiol.2015.08.008.

AUTHOR CONTRIBUTIONS

L.Y. and A.I. conceived of and designed the iSNAP analog algorithm. L.Y. wrote the software. M.A.S. designed the user interface. L.Y., A.I., and C.W.J. performed software testing and method validation. A.I. and C.W.J. performed experiments and analyzed data. C.W.J. cultured strains for loloatins, LI-F0 antibiotics, and arylomycins, isolated compounds, performed structural analysis, and characterized new compounds. L.Y., A.I., C.W.J., B.M., and N.A.M. contributed to study design. L.Y., A.I., C.W.J., B.M., and N.A.M wrote the manuscript.

ACKNOWLEDGMENTS

This work was funded through Natural Sciences and Engineering Research Council (NSERC) of Canada Discovery grants (RGPIN 371576-2009; 101997-2006), an NSERC Strategic grant (STPGP385235-09), and a Canadian Foundation for Innovation grant (2010M00022). C.W.J. is funded through a CIHR Doctoral Research Award. N.A.M. is funded by a CIHR New Investigator Award.

Received: April 25, 2015
Revised: July 30, 2015
Accepted: August 10, 2015
Published: September 10, 2015

REFERENCES

Balkovec, J.M., Hughes, D.L., Masurekar, P.S., Sable, C.A., Schwartz, R.E., and Singh, S.B. (2014). Discovery and development of first in class antifungal caspofungin (CANCIDAS®)—a case study. Nat. Prod. Rep. *15*, 15–34.

Chou, T.C., Zhang, X.G., Balog, A., Su, D.S., Meng, D., Savin, K., Bertino, J.R., and Danishefsky, S.J. (1998). Desoxyepothilone B: an efficacious microtubule-targeted antitumor agent with a promising in vivo profile relative to epothilone B. Proc. Natl. Acad. Sci. USA *95*, 9642–9647.

Eisenstein, B.I., Oleson, F.B., Jr., and Baltz, R.H. (2010). Daptomycin: from the mountain to the clinic, with essential help from Francis Tally, MD. Clin. Infect. Dis. *50*, S10–S15.

Frank, A.M., Bandeira, N., Shen, Z., Tanner, S., Briggs, S.P., Smith, R.D., and Pevzner, P.A. (2008). Clustering millions of tandem mass spectra. J. Proteome. Res. *7*, 113–122.

Gerard, J.M., Haden, P., Kelly, M.T., and Andersen, R.J. (1999). Loloatins A-D, cyclic decapeptide antibiotics produced in culture by a tropical marine bacterium. J. Nat. Prod. *62*, 80–85.

Han, J.W., Kim, E.Y., Lee, J.M., Kim, Y.S., Bang, E., and Kim, B.S. (2012). Site-directed modification of the adenylation domain of the fusaricidin nonribosomal peptide synthetase for enhanced production of fusaricidin analogs. Biotechnol. Lett. *34*, 1327–1334.

He, H., Williamson, R.T., Shen, B., Graziani, E.I., Yang, H.Y., Sakya, S.M., Petersen, P.J., and Carter, G.T. (2002). Mannopeptimycins, novel antibacterial glycopeptides from *Streptomyces hygroscopicus*, LL-AC98. J. Am. Chem. Soc. *124*, 9729–9736.

Hou, Y., Braun, D.R., Michel, C.R., Klassen, J.L., Adnani, N., Wyche, T.P., and Bugni, T.S. (2012). Microbial strain prioritization using metabolomics tools for the discovery of natural products. Anal. Chem. *84*, 4277–4283.

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

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.

Kulanthaivel, P., Kreuzman, A.J., Strege, M.A., Belvo, M.D., Smitka, T.A., Clemens, M., Swartling, J.R., Minton, K.L., Zheng, F., Angleton, E.L., et al. (2004). Novel lipoglycopeptides as inhibitors of bacterial signal peptidase I. J. Biol. Chem. *279*, 36250–36258.

Kuroda, J., Fukai, T., and Nomura, T. (2001). Collision-induced dissociation of ring-opened cyclic depsipeptides with a guanidino group by electrospray ionization/ion trap mass spectrometry. J. Mass. Spectrom. *36*, 30–37.

Kurusu, K., Ohba, K., Arai, T., and Fukushima, K. (1987). New peptide antibiotics LI-F03, F04, F05, F07, and F08, produced by *Bacillus polymyxa*. I. Isolation and characterization. J. Antibiot. *40*, 1506–1514.

Lin, Z., Falkinham, J.O., Tawfik, K.A., Jeffs, P., Bray, B., Dubay, G., Cox, J.E., and Schmidt, E.W. (2012). Burkholdines from *Burkholderia ambifaria*: antifungal agents and possible virulence factors. J. Nat. Prod. *75*, 1518–1523.

Medema, M.H., Paalvast, Y., Nguyen, D.D., Melnik, A., Dorrestein, P.C., Takano, E., and Breitling, R. (2014). Pep2Path: automated mass spectrometry-guided genome mining of peptidic natural products. PLoS Comput. Biol. *10*, e1003822.

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. (2011). Cycloquest: Identification of cyclopeptides via database search of their mass spectra against genome databases. J. Proteome. Res. *10*, 4505–4512.

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.

Newman, D.J., and Cragg, G.M. (2007). Natural products as sources of drugs over the last 25 years. J. Nat. Prod. *70*, 461–477.

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., Wu, C.H., Moree, W.J., Lamsa, A., Medema, M.H., Zhao, X., Gavilan, R.G., Aparicio, M., Atencio, L., Jackson, C., et al. (2013). MS/MS networking guided analysis of molecule and gene cluster families. Proc. Natl. Acad. Sci. USA *110*, E2611–E2620.

Scherlach, K., Partida-Martinez, L.P., Dahse, H.M., and Hertweck, C. (2006). Antimitotic rhizoxin derivatives from a cultured bacterial endosymbiont of the rice pathogenic fungus *Rhizopus microsporus*. J. Am. Chem. Soc. *128*, 11529–11536.

Schimana, J., Gebhardt, K., Höltzel, A., Schmid, D.G., Süssmuth, R., Müller, J., Pukall, R., and Fiedler, H.P. (2002). Arylomycins A and B, new biaryl-bridged lipopeptide antibiotics produced by *Streptomyces* sp. Tü 6075. I. Taxonomy, fermentation, isolation and biological activities. J. Antibiot. *55*, 565–570.

Tang, X.J., Thibault, P., and Boyd, R.K. (1992). Characterisation of the tyrocidine and gramicidin fractions of the tyrothricin complex from *Bacillus brevis* using liquid chromatography and mass spectrometry. Int. J. Mass. Spectrom. Ion. Process. *122*, 153–179.

Wang, C., Henkes, L.M., Doughty, L.B., He, M., Wang, D., Meyer-Almes, F.J., and Cheng, Y.Q. (2011). Thailandepsins: bacterial products with potent histone deacetylase inhibitory activities and broad-spectrum antiproliferative activities. J. Nat. Prod. *74*, 2031–2038.

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.

Winnikoff, J.R., Glukhov, E., Watrous, J., Dorrestein, P.C., and Gerwick, W.H. (2014). Quantitative molecular networking to profile marine cyanobacterial metabolomes. J. Antibiot. *67*, 105–112.

Yang, J.Y., Sanchez, L.M., Rath, C.M., Liu, X., Boudreau, P.D., Bruns, N., Glukhov, E., Wodtke, A., de Felicio, R., Fenner, A., et al. (2013). Molecular networking as a dereplication strategy. J. Nat. Prod. *76*, 1686–1699.

Yu, Z., Vodanovic-Jankovic, S., Kron, M., and Shen, B. (2012). New WS9326A congeners from *Streptomyces* sp. 9078 inhibiting *Brugia malayi* asparaginyl-tRNA synthetase. Org. Lett. *14*, 4946–4949.