ABSTRACT: Nonribosomal peptides (NRPs) such as vancomycin and daptomycin are among the most effective antibiotics. While NRPs are biomedically important, the computational techniques for sequencing these peptides are still in their infancy. The recent emergence of mass spectrometry techniques for NRP analysis (capable of sequencing an NRP from small amounts of nonpurified material) revealed an enormous diversity of NRPs. However, as many NRPs have nonlinear structure (e.g., cyclic or branched-cyclic peptides), the standard de novo sequencing tools (developed for linear peptides) are not applicable to NRP analysis. Here, we introduce the first NRP identification algorithm, NRPquest, that performs mutation-tolerant and modification-tolerant searches of spectral data sets against a database of putative NRPs. In contrast to previous studies aimed at NRP discovery (that usually report very few NRPs), NRPquest revealed nearly a hundred NRPs (including unknown variants of previously known peptides) in a single study. This result indicates that NRPquest can potentially make MS-based NRP identification as robust as the identification of linear peptides in traditional proteomics.
et al., 2009, also described algorithms for dereplication of cyclic peptides, answering the question of whether a spectrum arises from a known peptide in a chemical database such as Norine. They further introduced a variable dereplication algorithm answering the question of whether a spectrum is derived from a peptide that is similar to a known peptide. Ibrahim et al., 2012, further extended the dereplication approach from cyclic to branched-cyclic peptides. However, all of these approaches represent either de novo sequencing algorithms or dereplication searches in chemical databases and do not utilize valuable information about biosynthetic genes in the genomes of the NRP-producing organisms.

In contrast to the dereplication approaches (that use the same chemical database for different organisms), our new NRPquest tool, available at www.cyclo.ucsd.edu, first generates a database of putative NRPs extracted from the genome.

Figure 1. NRPquest pipeline starts with mining the microbial genome for putative NRPs using standard tools such as NRPSpredictor2 and constructing a database of putative NRPs. In the green rectangle, the results of NRPSpredictor2 are illustrated for Bacillus subtilis subsp. subtilis NCIMB 3610. This strain has two NRPS gene clusters, which according to NRPSpredictor2 produce two surfactins (7 amino acids each) and oneiplipastatin (10 amino acids). Adenylation domains are shown in red, condensation domains in blue, PCP domains in green, and thioesterase domains in light blue. Two blind modifications (with arbitrary offsets) are added to each NRP, and different possible structures (linear/cyclic/branched-cyclic) are considered (blue rectangle), resulting in ~134 million modified peptides. The red rectangle illustrates PSMs formed between each spectrum and each putative modified NRP with feasible mass difference. PSMs are scored and their p-values are computed using MS-DPR. MS-DPR approximates the probability distribution of scores of PSMs formed by a random peptide and the spectrum and further derives the p-value as the area under the extreme tail of the distribution. Spectra are further analyzed by spectral networks to enlarge the set of identified statistically significant PSMs. The yellow rectangle illustrates a spectral network of surfactins. The red arrows in the network illustrate how annotations are propagated from a node with low p-value $2.4 \times 10^{-10}$ (precursor $m/z$ 1022.7 Da) to nodes with higher p-values (e.g., a node with precursor $m/z$ 1030.7 Da), thus rescuing these nodes from being discarded as statistically insignificant.
sequence of the organism using the nonribosomal code. Because the nonribosomal code, in contrast to the ribosomal genetic code, is not yet fully understood (and is not as specific as the ribosomal genetic code), it may result in a multitude of peptides predicted from a single NRPS domain, and these putative NRPs may miss the correct NRP variant. Thus, the NRP identification approach “from genome to NRP gene prediction to NRP prediction to spectra matching” is more intricate than the classical proteomics approach “from genome to gene prediction to spectra matching”. NRPquest couples MS and genome mining and represents the first tool that transforms previous approaches for de novo sequencing/dereplication of NRPs into an MS/MS database search approach for the identification of NRPs. We demonstrate that, similarly to proteomics, peptide identification tools for NRP discovery are much more accurate than de novo sequencing tools.

NRPs are organized in modules that are responsible for the incorporation (and, if necessary, modification) of each additional amino acid in the synthesized NRP. Each module consists of several domains with defined functions, separated by short spacer regions of about 15 amino acids. A minimum of three domains are required for each NRPS module, termed the adenylation domain (A-domain), the peptidyl carrier domain (PCP-domain), and the condensation domain (C-domain). The A-domain is responsible for picking the specific amino acid monomers that are to be incorporated into the final product. Hundreds of different A-domain specificities have been classified using the Stachelhaus code, each one recruiting a specific amino acid as a monomer. This allows one to determine the putative sequence of the NRP by looking at the sequential order of A-domains along the assembly line and assigning a specific amino acid to each one. Many genome mining tools have been introduced for identification of NRPS gene clusters and the determination of their adenylation specificities, such as NRPSpredictor,18 ClustScan,19 NP.searcher,20 antiSMASH,21 and NRSPredictor2.22 As the nonribosomal code is not yet fully understood (particularly with respect to unusual nonstandard amino acids) and is promiscuous (adenylation domains can often load multiple amino acids), the accurate determination of the specificities of the adenylation domains remains difficult. Moreover, most NRPs go through postassembly line modifications such as backbone macrocyclization and the addition of fatty acid chains, and genome mining tools currently fail to predict most of these modifications. NRPquest recruits NRPSpredictor2 for the prediction of adenylation specificities and implements a mutation-tolerant and modification-tolerant (blind) MS/MS search that allows for multiple modifications and mutations. This is a difficult computational problem even in the case of linear peptides,23 let alone nonlinear peptides.

As the first tool integrating genomic and mass spectrometric evidence for identification of NRPs, NRPquest promises to greatly increase the number of known peptide—spectrum matches (PSMs) for NRPs; for example, this study alone identified nearly a hundred PSMs. Previous studies emphasized the importance of increasing the number of PSMs (formed by NRPs) for developing rigorous statistical approaches and more adequate scoring functions for cyclic peptide analysis.4,13

### RESULTS AND DISCUSSION

NRPquest uses a sequenced genome and a mass spectral data set as an input and includes the following steps: (i) uses NRP prediction tools for identifying NRPSs in the genome and constructing the database of putative NRPs, (ii) matches a spectral data set against the database of putative NRPs in a blind mode, (iii) computes statistical significance of the resulting cyclic, branched-cyclic, and linear PSMs and further ranks confident PSMs, (iv) constructs a spectral network29 to enlarge the set of identified PSMs and reveal families of related NRPs via spectral network dereplication (Figure 1).

**NRP Database Construction.** The NRPquest software pipeline starts with annotation of the genome of an organism of interest by NRPSpredictor2,22 which attempts to predict a set of all possible monomers for each adenylation domain in the genome (Figure 1). By considering all possible combinations of monomers, we construct a database of all putative NRPs (referred to as NRP database) that can be produced by the organism. NRPquest further searches the genome for methylation domains (PF08242) and accounts for the corresponding modification in the NRP database. If NRPquest finds a methylation domain, it allows these modifications for the corresponding residues. If NRPquest finds a cytochrome P450 domain, it models side chain bonds between any two residues of the peptide, turning a linear peptide into a branched-cyclic peptide. Note that NRPquest also allows all peptides to have cyclic backbones with amino acid side chains. For each amino acid sequence in the database of putative NRPs, NRPquest considers linear, cyclic, and branched-cyclic structures representing these amino acid sequences.

**Blind Search for Modifications.** Each spectrum is matched against each putative peptide in the NRP database using a brute force algorithm that allows for blind modifications. The standard blind PTM search in traditional proteomics with tools (e.g., InsPecT23 and MODa24) typically limits the searches to at most two modifications, since further increase in the number of blind modifications makes the identified peptides less reliable. Since predictions provided by NRPSpredictor2 are typically within two mutations/modifications from the correct peptide, NRPquest also limits searches to at most two blind modifications whose total mass does not exceed 300 Da. This accounts for possible inaccurate adenylation specificity prediction of rare nonstandard amino acids (such as kynurenine from daptomycin), postassembly line modifications (such as modification of homoproline to 4-oxo-homoproline in pristinamycin), and addition of a fatty acid tail. Similarly to identification of linear peptides, MS-based methods for NRP identifications are limited to finding modification masses but do not provide insights into specific chemistry of modifications.

For example, in the case of *Bacillus subtilis*, the genome sequence has two NRPS clusters (surfactin and pilipastatin). NRPSpredictor2 predicts a single NRP for pilipastatin but two NRPs for surfactin (both Glu and Gln are predicted as first amino acid). As a result, we have only three possible NRPs initially predicted for *B. subtilis*. However, after considering different structures (linear, cyclic, branched-cyclic) there are 48 possible structures. If we consider two blind modifications with masses less than 300 Da for each peptide, there would be ~134 million possibilities. However, many of these peptides have masses very different from the spectrum precursor mass, and on average, only ~450 thousand of these putative modified peptides have a precursor mass matching each spectrum within 0.5 Da.

**Statistical Significance.** As there is still no large data set of PSMs formed by nonlinear peptides for automated learning of
fragmentation propensities, previous studies of cyclic peptides used a somewhat primitive scoring based on the "shared peak count." NRPquest scores cyclic PSMs using a previously proposed approach (Figure 1) and, to be consistent with the cyclic case, uses the same "shared peak count" approach for linear peptides. For branched-cyclic peptides, NRPquest simply adds up the scores of linear and cyclic parts (assuming that the linear branch is a modification on the cyclic part).

While methods for evaluating statistical significance of linear PSMs are well developed, until 2013 there were no methods for evaluating statistical significance of nonlinear PSMs. NRPquest calculates the statistical significance of each PSM using the recently introduced MS-DPR algorithm, which works for linear, cyclic, and branched-cyclic peptides (Figure 1). It further reports PSMs with low p-values (Table S1).

**Spectral Networks.** We use spectral networks to enlarge the set of identified NRPs. Spectral networks (also known as molecular networks) analyze multiple spectra to simultaneously sequence related unknown peptides. The advantage of this approach (compared to de novo sequencing of individual spectra) is that finding peptides that simultaneously explain all spectra in a spectral network results in more accurate peptide reconstructions. Most NRPs form families of related peptides, and spectral network analysis can be used to reveal relationships between different spectra without knowing the amino acid sequences corresponding to these spectra.

Given a set of peptides \( P_1, \ldots, P_m \), their peptide network is a graph with \( m \) nodes \( P_1, \ldots, P_m \) and edges connecting two peptides if they differ by a single amino acid substitution or a single modification. Figure 2 shows the peptide network for nine variants of tyrocidine, a well-studied NRP from *Bacillus brevis*. For example, peptide 1 (tyrocidine B1) in this network (red node) is connected to four peptides differing from tyrocidine B1 by a single mutation or modification: tyrocidine A1 (peptide 2), tyrocidine B (peptide 5), tyrocidine C1 (peptide 8), and a previously unreported peptide with mass 1338.7 Da (peptide 9). However, it is not connected to peptides 3, 4, 6, and 7 since they differ from peptide 1 by multiple modifications. Six of these nine tyrocidines (1, 2, 3, 5, 7, 8) are contained in the database of putative NRPs generated by NRPSpredictor (without modifications), and three more differ from these variants by one or two modifications/mutations.

In reality, we are not given peptides \( P_1, \ldots, P_m \) but only their spectra \( S_1, \ldots, S_m \). Nevertheless, one can approximate the peptide network by constructing the spectral network on nodes \( S_1, \ldots, S_m \), where spectra \( S_i \) and \( S_j \) are connected by an edge if they can be aligned against each other. For linear peptides, the spectral alignment algorithm reveals the lion's share of spectra that correspond to peptides differing by a single mutation/modification, and similar approaches are proposed for nonlinear peptides. Figure 2 shows the spectral network of nine tyrocidines and illustrates that it captures all edges from the peptide network (shared edges between peptide and spectral networks are shown by thick lines). While the spectral network in Figure 2a is not identical to the peptide network in Figure 2b, their shared edges usually allow one to interpret the peptides corresponding to the nodes of the spectral network.
using the spectral network dereplication algorithm.\textsuperscript{13} NRPquest requires a \(p\)-value threshold to report reliable PSMs, and with a threshold of \(10^{-10}\), four of nine peptides in this spectral network were identified (Figure 2c).

Below we illustrate how spectral networks allow us to confidently identify less reliable PSMs. The nine PSM peptides found by NRPquest have \(p\)-values ranging from \(4.7 \times 10^{-16}\) (extremely reliable identification) to \(1.6 \times 10^{-8}\) (less reliable identification). However, since NRPquest reports only PSMs with low \(p\)-values (otherwise, many statistically insignificant PSMs would be reported), some of these PSMs are deemed problematic and are not reported as identified peptides. For example, the spectrum corresponding to node 4 (with mass 1371.6 Da) was not identified by NRPquest (its \(p\)-value exceeds the threshold \(10^{-10}\)), but it is connected to spectrum 7, corresponding to the reliably identified peptide tyrocidine C (mass 1348.7 Da), with a low \(p\)-value of \(2.7 \times 10^{-12}\). Thus, peptide 4 differs from tyrocidine C by a modification with mass 23.1 Da. The variable dereplication algorithm\textsuperscript{13} further predicts the site of this modification at the tyrosine residue. This means the change is likely to be a mutation of tyrosine to tryptophan, and the peptide at node 4 is tyrocidine D.\textsuperscript{31}

After constructing the spectral network, we extract their connected components (that correspond to families of related peptides) and further consider two possible scenarios: (i) no node (spectrum) in the connected component has been identified by NRPquest and (ii) at least one node in the connected component has been identified by NRPquest. In the former case, we use the multitag algorithm\textsuperscript{13} for multiplexed de novo sequencing of peptides represented by spectra forming this connected component. In the latter case (like in Figure 2), we use a variable dereplication version of the multitag algorithm.

Variable dereplication\textsuperscript{8} via peptide propagation through a spectral network\textsuperscript{29} allows one to identify peptides with more than two modifications that NRPquest missed during blind searches of individual spectra (as blind searches with more than two modifications become prohibitively time-consuming and increase the number of false PSMs). In the example in Figure 2a, four out of nine tyrocidines were identified (with \(p\)-values below \(10^{-10}\)) even before constructing the spectral network. After peptide propagation through the spectral network, all nine peptides were identified by the multitag algorithm with \(p\)-values below the \(10^{-7}\) threshold.\textsuperscript{13} The case of the tyrocidines is somewhat special because many of them form statistically significant PSMs with exceptionally low \(p\)-values. For other spectral networks, the number of statistically significant PSMs is smaller; for example, in the case of arylomycins, no PSMs were identified as statistically significant before applying spectral networks.

Figure 3 shows the connected components of the spectral networks corresponding to six peptide families representing previously sequenced peptides. Daptomycin\textsuperscript{32} and arylomycin\textsuperscript{33} are identified from \(S.\) \textit{roseosporus}, surfactin\textsuperscript{34} and plipastatin\textsuperscript{35} are identified from \(B.\) \textit{subtilis}, pristinamycin\textsuperscript{36} is identified from \(S.\) \textit{pristinaeptidis}, and tyrocidine\textsuperscript{37,38} is identified from \(B.\) \textit{brevis}. For example, the spectral networks for the tyrocidines consist of nine peptides in the case of charge +1 spectra shown in red (eight peptides in the case of charge +2 spectra shown in green). Six out of these nine peptides have been previously sequenced in various studies,\textsuperscript{38} while others have not been reported in the literature yet (some of them may represent chemical adducts of known peptides). Similarly, only a small fraction of peptides in the spectral networks for daptomycin, arylomycin, surfactin, and plipastatin have been previously identified. The ability to identify known NRPs in a blind experiment and to discover previously unknown variants of known NRPs illustrates the power of NRPquest.

We have described how spectral networks help in identifying PSMs with high \(p\)-values (that would otherwise be discarded) that are adjacent to statistically significant PSMs with low \(p\)-values in the spectral networks. Another contribution of

Figure 3. Spectral networks of six NRP families identified by NRPquest: (a) daptomycin, (b) arylomycin, (c) pristinamycin, (d) plipastatin, (e) surfactin, and (f) tyrocidine (Table S1). Only spectra forming the most statistically significant PSMs are shown. Each node in these spectral networks may represent either a single spectrum or a group of very similar spectra (with similar precursor masses) compressed into a single node to simplify the network (in the latter case, the \(m/z\) of a cluster is the average of \(m/z\) of spectra in the cluster). The thickness of the edges indicates the level of similarity between the nodes in the spectral networks. Two connected components (in the case of plipastatin and tyrocidine) correspond to two different charge states (currently, the spectral alignment algorithm may fail to connect spectra from related peptides with different charges by an edge).
spectral networks is reranking and potentially reusing the entire spectral network of PSMs even if all PSMs in this spectral network have high p-values. Starting from a PSM $P_1, S_1$ and a spectral network with nodes $S_1, ..., S_m$, the dereplication algorithm determines unknown peptides $P_1, ..., P_n$ that form PSMs with spectra $S_1, ..., S_m$ and computes the SpecNetScore:

$$\text{SpecNetScore}(P_1, ..., P_n, S_1, S_m) = \sum_{i=1}^{m} -\log p\text{-value}(P_i, S_i)$$

Afterward, it reports all found PSMs $(P_1, S_1), ..., (P_n, S_m)$ if the SpecNetScore exceeds a threshold.

**Benchmarking NRPquest.** We analyzed bacterial extracts from *Streptomyces roseosporous* NRRL 15998 (SR), *Streptomyces pristinaespiralis* ATCC 25486 (SP), *Bacillus subtilis* subsp. *subtilis* NCIMB 3610 (BS), and *Bacillus brevis* (BB). Table S1 illustrates that peptides with the lowest p-values that were identified by NRPquest in a blind experiment correspond to previously sequenced NRPs and provides confidence that some other PSMs that NRPquest finds may correspond to novel variants of known NRPs. While it is not possible to verify these low-abundance NRP variants by NMR, we were able to verify that these novel compounds are related to known NRPs using spectral networks (Figure 3). We further verified the significance of the resulting PSMs by computing their p-values (Table S1).

NRPquest constructs spectral networks (Figure 3) to improve the identification of individual spectra and to enlarge the set of identified PSMs (Table S2). After annotating spectra in the spectral network, we removed all nodes corresponding to PSMs with high p-values above 0.0001, as some large networks are difficult to visualize. Table S2 illustrates that spectral networks allow one to confidently sequence all spectra in a connected component of a spectral network via peptide propagation as long as even a single spectrum in this component is identified at the previous MS/MS database search step.

In addition to NRPs, peptide natural products include RiPPs. In a separate paper, Mohimani et al., 2013, described RiPPquest, an automated genome mining approach for RiPP identification. As NRPquest and RiPPquest are complementary tools searching completely different protein databases, the sets of peptides they identify do not overlap. Thus, the cases when NRPquest reports RiPPs (or RiPPquest reports NRPs) are not expected to occur in practice.

For example, in the case of *S. roseosporous* NRRL 15998, the spectral data set has 4355 spectra, out of which only six are assigned to NRPs with a p-value smaller than $10^{-18}$. In this particular data set, two RiPPs, called SRO-2212 and SRO-3108, were previously discovered. NRPquest correctly discards spectra of these RiPPs since they have a low score against the previously sequenced NRPs and provides confidence for the NRP variants.

The identification of arylomycin via spectral networks illustrates the power of combining genome mining with spectral networks. Arylomycin is an antibiotic that selectively binds type I signal peptidases (SPases). Despite in vivo activity, and the fact that SPase is conserved and easily accessible as a drug target, arylomycin was dismissed as a drug candidate in the past, as it was perceived to have a narrow spectrum of activity. However, recent studies demonstrated that modified arylomycins are actually broad-spectrum antibiotics and that, if optimized to bind their targets with slightly more affinity, arylomycins would have a spectrum of activity that supports their progression as broad-spectrum therapeutics. Currently, RQQ Pharmaceuticals is proceeding with clinical trials of arylomycin variants. Our study reveals a variety of arylomycins, opening the possibility to examine some of them as potential therapeutic agents. The limited number of fragmentation sites observed in some arylomycin MS/MS spectra makes it difficult to identify them individually. Table S2(A) illustrates that no arylomycin was identified with a p-value below the threshold. However, simultaneous analysis of different variants of arylomycin enables identification of seven arylomycin variants, as Table S2 and Figure 3b illustrate. Note that identification of arylomycins was based solely on their MS/MS and genomic information, and no prior information about known arylomycin peptides was used.

The effort involved in NRP discovery from MS/MS can be divided into three steps. The first step is determining which of the MS/MS spectra represent NRPs. This is a crucial step in natural product discovery because most of the time bacterial extracts also include endogenous peptides, RiPPs, PKs, and other chemical molecules in addition to NRPs. The second step is the correlation of MS/MS spectra to NRPs in the genome. The third step is the annotation of spectra by assigning modifications to individual NRP residues. Our results in Table S1 show that while NRPquest has succeeded in the first two steps, the last step remains problematic and needs a better understanding of NRPs modification enzymes and NRP fragmentation. We hope that bioinformatics methods such as NRPquest will enable this progress by helping in the development of large data sets of annotated NRPS gene clusters.

**Experimental Section.**

**Bacterial Metabolite Extraction.** We analyzed bacterial strains of *Streptomyces roseosporous* NRRL 15998, *Streptomyces pristinaespiralis* ATCC 25486, *Bacillus subtilis* subsp. *subtilis* NCIMB 3610, and *Bacillus brevis* ATCC 8185. Each agar plate was inoculated with each bacterial strain by four parallel streaks. The plates were incubated for 10 days at 28 °C. The agar was sliced into small pieces, then put in a 50 mL centrifuge tube, covered with equal amounts of Milli-Q H$_2$O and nBuOH/MeOH for 12 h at 28 °C, and shaken at 225 rpm. The nBuOH/MeOH layer was collected using a transfer pipet and dried with a rotary evaporator.

**Genome Data Sets.** Genomes of *S. roseosporous* and *S. pristinaespiralis* were recently sequenced at the Broad Institute and are available from the Actinobacteria database Web site. Genomes of *B. subtilis* and *B. brevis* are available from NCBI.

**Spectral Data Sets.** Collision-induced dissociation (CID) MS/MS data sets were collected with or without liquid chromatography (LC)
were converted to .mzXML using the program ReAdW (http://tools.proteomecenter.org).

**REFERENCES**


