UNIVERSITY OF CALIFORNIA, SAN DIEGO

Bioinformatics Methods for Natural Product Discovery

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Electrical Engineering (Communications Theory and Systems)

by

Hosein Mohimani

Committee in charge:

Professor Pavel A. Pevzner, Chair
Professor Alexander Vardy, Co-Chair
Professor Vineet Bafna
Professor Nuno Bandeira
Professor Pieter C. Dorrestein
Professor William Hodgkiss
Professor Paul Siegel

2013

The dissertation of Hosein Mohimani is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Co-Chair

Chair

University of California, San Diego

2013

DEDICATION

To my parents.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Without te help of many others, it would have been impossible for me to get to where I am now. I would first like to thank my advisor, Dr Pavel Pevzner, for teaching me how to be a scientist. What I learned from him is not limited to science, but it covers writing skills, how to express ideas to scientific community, and many more. I'd also like to thank my co-advisor, Dr. Pieter Dorrestein, for motivating me with so many fascinating ideas during my graduate school years. I would also like to thank Dr Vineet Bafna, Dr Nuno Bandeira, and all my labmates during the last five years, in particular Boyko Kakaradov, Kyowon jeong, Mingxun Wang and Sangtae Kim . I also thank my ECE advisor Dr Alexander Vardy, and my committee members Dr Paul Siegel and Dr William Hodgkiss.

Finally, I would not even be in this position without the sacrifices made by my family over the years. I would like to thank my parents, my sister, and my brother.

Chapter 2 is in part adapted from *Hosein Mohimani, Wei-Ting Liu, Yu-Liang Yang, Susana P. Gaudenico, William Fenical, Pieter C. Dorrestein, and Pavel Pevzner, Multiplex De Novo Sequencing of Peptide Antibiotics, Journal of Computational Biology, 2011, 18(11): 1371-1381.* The dissertation author was the primary author of this paper responsible for the research.

Chapter 3 is in part adapted from *Hosein Mohimani, Yu-Liang Yang, Wei-Ting Liu, Pei-Wen Hsieh, Pieter C. Dorrestein, and Pavel Pevzner, Sequencing Cyclic Peptides by Multistage Mass Spectrometry, 2011, Journal of Proteomics, 11(18), 3642-50.* The dissertation author was the primary author of this paper responsible for the research.

Chapter 4 is in part adapted from *Hosein Mohimani, Wei-Ting Liu, Joshua S. Mylne, Aaron G. Poth, Michelle Colgrave, Michael Selsted, Pieter C. Dorrestein, and Pavel A. Pevzner, Cycloquest: Identification of cyclopeptides via database search of their mass spectra against genome databases, 2011, Journal of Proteome Research, 10 (10), pp 4505-4512.* The dissertation author was the primary author of this paper responsible for the research.

Chapter 5 is in part adapted from *Hosein Mohimani, Sangtae Kim, and*

Pavel A. Pevzner, A new approach to evaluating statistical significance of spectral identifications, 2013, Journal of Proteome Research, 12 (4), pp 1560-1568. The dissertation author was the primary author of this paper responsible for the research.

VITA

2008	B.S. in Electrical Engineering and Mathematical Sciences, Sharif University of Technology, Iran
2013	Ph. D. in Electrical Engineering (Communications Theory and Systems), University of California, San Diego, USA

PUBLICATIONS

Hosein Mohimani, Roland Kersten, Wei-Ting Liu, Pieter C. Dorrestein, and Pavel A. Pevzner, NRPquest: Coupling Mass Spectrometry and Genome Mining for Non Ribosomal Peptide Discovery (submitted).

Hosein Mohimani, Roland Kersten, Wei-Ting Liu, Samuel O. Purvine, Si Wu, Heather M. Brewer, Ljiljana Pasa-Tolic, Bradley S. Moore, Pavel A. Pevzner, and Pieter C. Dorrestein, Informatimicin, a novel Streptomyces viridochromogenes lanthipeptide discovered by automated peptidogenomics (in preparation).

Hosein Mohimani, Sangtae Kim, and Pavel A. Pevzner, A New Approach to Evaluating Statistical Significance of Spectral Identifications, Journal of Proteome Research, 2013, 12 (4), pp 15601568

Hosein Mohimani, Wei-Ting Liu, Joshua S. Mylne, Aaron G. Poth, Michelle Colgrave, Michael Selsted, Pieter C. Dorrestein, and Pavel A. Pevzner, Cycloquest: Identification of cyclopeptides via database search of their mass spectra against genome databases, Journal of Proteome Research, 2011, 10 (10), pp 4505-4512.

Hosein Mohimani, Yu-Liang Yang, Wei-Ting Liu, Pei-Wen Hsieh, Pieter C. Dorrestein, and Pavel Pevzner, Sequencing Cyclic Peptides by Multistage Mass Spectrometry, 2011, Journal of Proteomics, 11(18), 3642-50.

Hosein Mohimani, Wei-Ting Liu, Yu-Liang Yang, Susana P. Gaudenico, William Fenical, Pieter C. Dorrestein, and Pavel Pevzner, Multiplex De Novo Sequencing of Peptide Antibiotics, Journal of Computational Biology, 2011, 18(11), 1371-1381

Emily Mevers,Wei-Ting Liu, Niclas Engene, **Hosein Mohimani**, Tara Byrum, Pavel A. Pevzner, Pieter C. Dorrestein, Carmenza Spadafora, and William H. Gerwick, Cytotoxic veraguamides, alkynyl bromide-containing cyclic depsipeptides from the marine cyanobacterium cf. Oscillatoria margaritifera, 2011, Journal of Natural Product, 74 (5), pp 928-936.

ABSTRACT OF THE DISSERTATION

Bioinformatics Methods for Natural Product Discovery

by

Hosein Mohimani

Doctor of Philosophy in Electrical Engineering (Communications Theory and Systems)

University of California, San Diego, 2013

Professor Pavel A. Pevzner, Chair
Professor Alexander Vardy, Co-Chair

Most of new chemical entities introduced as antibacterials over the last decades are derivated from natural products produced by living organisms. Some of the most effective antibiotics are peptidic natural products. The traditional process of natural products discovery is to elucidate strcuture of the compound of interest by chemical assays such as Nuclear Magnetic Resonance. This process is long, laborious, and requires large amounts of highly purified material. Recent advances in mass spectrometry has enabled natural product discovery from picograms of material. In this thesis we propose various computational techniques to aid natural product discovery by computational mass spectrometry.

Chapter 1

Introduction

About 70% of new chemical entities introduced as antibacterials over the last 25 years are derivated from natural products produced by living organisms [1]. Natural products are classified into a variety of chemical classes, including peptidic natural products (PNPs), lipids, and carbohydrates. Various biosynthetic machineries are involved in the production of natural products, e.g. Non-Ribosomal Peptide Synthesize (NRPS) [2, 3], Polyketide Synthesize (PKS) [4], and Post Ribosomal Peptide Synthesize (PRPS) [5, 6]. NRPS, PKS and PRPS synthesize Non-Ribosomal Peptides (NRPs), Polyketides (PKs), and Ribosomally synthesized and Posttranslationally modified Peptides (RiPPs).

The traditional process of natural products discovery is to elucidate strcuture of the compound of interest by chemical assays such as Nuclear Magnetic Resonance and Crystallography, and association of the chemical compound to its biosynthetic gene cluster by genome manipulation. This process is long, laborious, and requires large amounts of highly purified material. Moreover, rather than discovering novel natural products, this process frequently rediscovers known natural products resulting in wasted efforts. Recent advances in mass spectrometry has enabled natural product discovery from picograms of material.

Mass spectrometry has been the method of choice for study of proteins in a high-throughput manner. Advances in instrumentation and software has allowed researchers to take over more ambitious projects, both in terms of scale and complexity of the experiments. Mass spectrometrys versatility lies in the fact that it can detect molecules in small concentrations in a wide range of masses. Mass spectrometry has been utilized to detect biomarkers for diseases, identification and quantification of expressed proteins, identification of modifications on proteins, and aiding gene annotations, just to mention a few application in the biological sciences.

Availability of microbial genome sequences has enabled prediction of approximate structure of natural products that an organism is capable of producing. Genome mining for a natural product refers to using information about the biosynthetic genes (responsible for synthesizing this natural product) to infer information about the natural product itself. Discovery of the Non Ribosomal Peptide (NRP) coelichelin in *Streptomyces coelicolor* was one of the first examples of the discovery of a natural product through genome mining [7, 8]. Since then, genome mining was utilized to discover numerous natural products including NRPs, PKs, and RiPPs. Computational mass spectrometry techniques for discovery of novel natural products can be devided into two categories; methods that utilize genome mining to aid in identification, and genome independent methods.

While availablity of the genome sequence can greatly reduce the effort required for the discovery of natural products, many of natural products are from organisms without genome sequenced. The general approaches for structure elucidation of these chemicals is *dereplication* from publicly available chemical databases [9], *spectral library search* [10, 11, 12, 13, 14], *spectral networks* [15] and *de novo sequencing* [16, 17, 18]. **Chapter 2** and **Chapter 3** describes *multiplex de novo sequencing* and *multistage de novo sequencing*, two novel methods for denovo sequencing of peptidic natural products. **Chpater 4** describes a novel database search of ribosomal cyclopeptides. **Chapter 5** describes *NRPquest* a novel method for discovery of non-ribosomal peptides by mass spectrometry and genome mining. **Figure 1.1** describes these pipelines, and **Table 1.1** compares them.

Table 1.1: Comparing different natural product discovery approaches.

Method	Genome?	Novel?	Non-peptide?	References
Dereplication	No	varinats	Yes	[9, 19, 20, 16, 21]
Spectral libray search	No	No	Yes	[10]
Spectral networks	No	varinats	Yes	[15]
Denovo PNP sequencing	No	Yes	No	[16, 17, 18]
PNP database search	Yes	Yes	No	[22, 23]

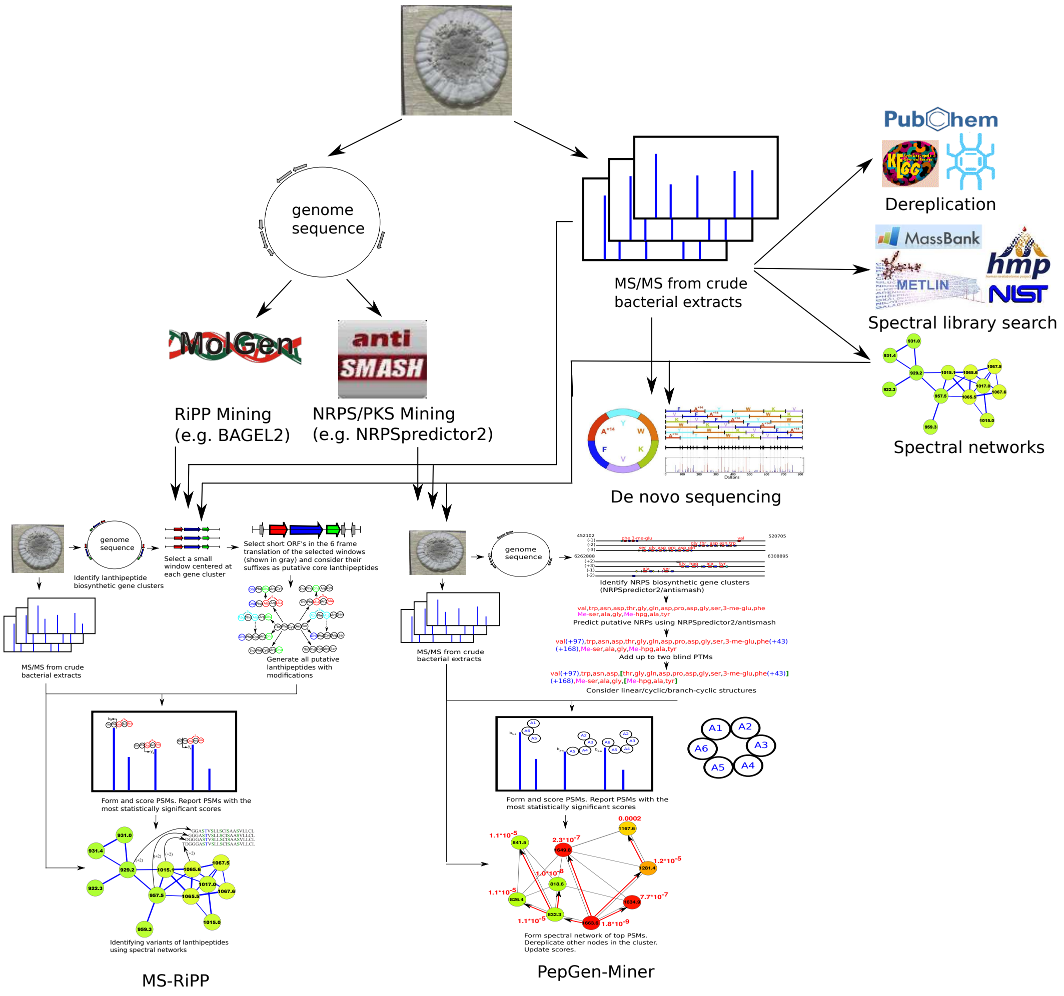

genome sequence
MS/MS from crude bacterial extracts
Dereplication
Spectral library search
Spectral networks
RiPP Mining (e.g. BAGEL2)
NRPS/PKS Mining (e.g. NRPSpredictor2)
De novo sequencing
Daltons
genome sequence
Identify lanthipeptide biosynthetic gene clusters
Select a small window centered at each gene cluster
Select short ORF's in the 6 frame translation of the selected windows (shown in gray) and consider their suffixes as putative core lanthipeptides
MS/MS from crude bacterial extracts
Generate all putative lanthipeptides with modifications
Form and score PSMs. Report PSMs with the most statistically significant scores
931.0
931.4
929.2
922.3
1015.1
1065.6
1067.5
1017.0
1067.6
957.5
1065.5
1015.0
959.3
GGASTVSLLSCISAASVLLCL
GGGASTVSLLSCISAASVLLCL
DGGGASTVSLLSCISAASVLLCL
TDGGGASTVSLLSCISAASVLLCL
Identifying variants of lanthipeptides using spectral networks
MS-RiPP
genome sequence
MS/MS from crude bacterial extracts
Identify NRPS biosynthetic gene clusters (NRPSpredictor2/antismash)
val,trp,asn,asp,thr,gly,gln,asp,pro,asp,gly,ser,3-me-glu,phe Me-ser,ala,gly,Me-hpg,ala,tyr
Predict putative NRPs using NRPSpredictor2/antismash
val(+97),trp,asn,asp,thr,gly,gln,asp,pro,asp,gly,ser,3-me-glu,phe(+43) (+168),Me-ser,ala,gly,Me-hpg,ala,tyr
Add up to two blind PTMs
val(+97),trp,asn,asp,[thr,gly,gln,asp,pro,asp,gly,ser,3-me-glu,phe(+43)] (+168),Me-ser,ala,gly,[Me-hpg,ala,tyr]
Consider linear/cyclic/branch-cyclic structures
A1
A2
A3
A4
A5
A6
Form and score PSMs. Report PSMs with the most statistically significant scores
0.0002
1167.6
1.1*10-5
841.5
2.3*10-7
1649.8
1.2*10-5
1281.4
1.0*10-8
818.6
1.1*10-5
826.4
7.7*10-7
1634.9
832.3
1.1*10-5
1663.6
1.8*10-9
Form spectral network of top PSMs. Dereplicate other nodes in the cluster. Update scores.
PepGen-Miner

Figure 1.1: Computational mass spectrometry pipeline for natural product discovery.

Chapter 2

Multiplex De Novo Sequencing of Peptide Antibiotics

2.1 Introduction

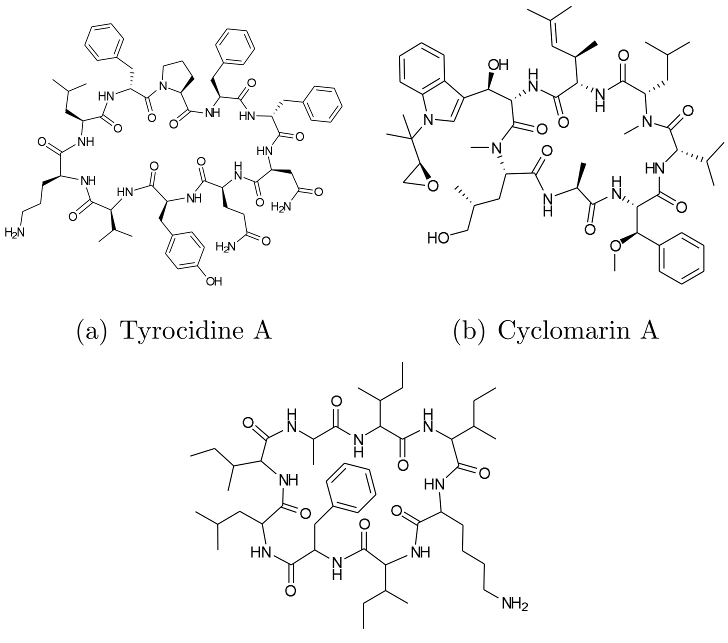
In 1939 Renê Dubos discovered that the peptide fraction *Tyrothricin*, isolated from the soil microbe *Bacillus brevis*, had an ability to inhibit the growth of *Streptococcus pneumoniae*, rendering it harmless. Although discovered 10 years after Penicillin, it was the first mass produced antibiotic deployed in Soviet hospitals in 1943. Unfortunately, the identification of amino acid sequences of cyclic peptides, once a heroic effort, remains difficult today. The dominant technique for sequencing cyclic peptide antibiotics is 2D NMR spectroscopy, which requires large amounts of highly purified materials that, are often nearly impossible to obtain.

Tyrothricin is a classic example of a mixture of related cyclic decapeptides whose sequencing proved to be difficult and took over two decades to complete. By the 1970s, scientists had sequenced 5 compounds, Tyrocidine A-E, from the original mixture. However, these five are not the only peptides produced by *B. brevis* and even today it remains unclear whether *all* of the antibiotics produced by this bacterium have been documented (see reference [24] for a list of 28 known peptides from *B. brevis*).

Figure 2.1(a) shows structure of Tyrocidine A. **Table S1** illustrates that most cyclic decapeptides in the Tyrocidine/Tryptocidine family can be represented as shown (the rounded amino acid masses in daltons are also shown):

$$Val\left\{\begin{matrix}Orn\\Lys\end{matrix}\right\}LeuPhePro\left\{\begin{matrix}Phe\\Trp\end{matrix}\right\}\left\{\begin{matrix}Phe\\Trp\end{matrix}\right\}AsnGln\left\{\begin{matrix}Tyr\\Trp\\Phe\end{matrix}\right\}$$

$$99\ \left\{\begin{matrix}114\\128\end{matrix}\right\}\ 113\ 147\ \ 97\ \left\{\begin{matrix}147\\186\end{matrix}\right\}\ \left\{\begin{matrix}147\\186\end{matrix}\right\}\ 114\ 128\left\{\begin{matrix}163\\186\\147\end{matrix}\right\}$$



(a) Tyrocidine A

(b) Cyclomarin A

(c) Reginamide A

Figure 2.1: Structures of Tyrocidine A (a), Cyclomarin A (b), and Reginamide A (c).

It may come as a surprise that there are no genes in *B. brevis* whose codons encode any of the Tyrocidine peptides! Tyrocidines, similar to many antibiotics such as Vancomycin or Daptomycin, represent cyclic *non-ribosomal* peptides (NRPs) that do not follow the central dogma “DNA produces RNA produces Protein”. They are assembled by nonribosomal peptide synthetases that represent both the mRNA-free template and building machinery for the peptide biosynthesis [25]. Thus, NRPs are not directly inscribed in genomes and cannot be inferred with traditional DNA sequencing. Cyclic NRPs are of great pharmacological importance as they have been optimized by evolution for chemical defense and communication. Cyclic NRPs include antibiotics, antitumor agents, immunosuppressors, toxins, and many peptides with still unknown functions.

Most NRPs are cyclic peptides that contain nonstandard amino acids, increasing the number of possible building blocks from 20 to several hundreds. The now dominant 2D NMR-based methods for NRP characterization are time-consuming, error prone, and requires large amounts of highly purified material. Because NRPs are often produced by difficult to cultivable microorganisms, it may not be possible to get sufficient quantities for 2D structure elucidation, therefore it is important to develop a nmol scale structure elucidation approach [26, 27]. Such methods promise to greatly accelerate cyclic NRP screening and may illuminate a vast resource for the discovery of pharmaceutical agents [28].

The first automated Mass Spectrometry (MS) based approach to sequencing cyclic peptides correctly sequenced 2 out of 6 Tyrocidines analyzed by Ng et al. [16]. While the correct sequences for 4 other Tyrocidines were highly ranked, Ng et al., 2009 [16] came short of identifying them as the *highest-scoring* candidates. Leao et al., [29], 2010, and Liu et al., [30], 2010, recently applied the algorithm from [16] for analyzing new cyclic peptides. In [29], the authors study peptides produced by the cyanobacterium *Oscillatoria sp.* that inhibit the growth of green algae and demonstrated that they function in a synergistic fashion, i.e., mixtures of these analogous peptides are needed to inhibit green algal growth. This observation emphasizes the importance of studying various peptide variants and calls for the development of a technology able to simultaneously sequence *all* peptides produced by a single organism.

Our first attempt to sequence cyclic NRPs from *Oscillatoria sp.* via MS using the algorithm described by Ng et al., [16] was inconclusive. We (Leao et al. 2010 [29]) resorted to purification of the most abundant peptide with the goal to sequence it via 2D NMR (purification of individual NRPs is often difficult since various NRP variants have similar physicochemical properties). This amounted to a large effort that involved applications of various NMR technologies (including HSQC, HMBC, COSY, and NOESY) but still failed to identify some inter-residue dependencies. Applications of both NMR and MS to finally sequence four compounds using NRP-Dereplication algorithm from [16] represented a large and time-consuming effort of a multidisciplinary team. A better approach would be to generate MS/MS spectra of *all* variant NRPs (without the need to purify large amounts of individual peptides) and to *multiplex* sequence them. By multiplex sequencing we mean simultaneous (and synergetic) sequencing of related peptides from their spectra.

Using this approach, we sequenced many known members of the Tyrocidine family as well as some still unknown Tyrocidine variants. Finding new Tyrocidine variants is surprising since this family has been studied for sixty years now. We further sequenced a previously unknown family of NRPs isolated from a bacterial strain that produces natural products with anti-asthma activities (named *Reginamides*). To validate these new sequences (obtained from a single mass spectrometry experiment) we analyzed one of them (named Reginamide A) using (rather time consuming) NMR experiments. The mass spectrometry approach revealed the sequence of masses with molecular composition (C_3H_5NO, $C_6H_{11}NO$, $C_6H_{11}NO$, $C_7H_{12}N_2O_2$, $C_6H_{11}NO$, C_9H_9NO, $C_6H_{11}NO$, $C_6H_{11}NO$) that was matched by NMR as the cyclic peptide AIIKIFLI with structure shown in **Figure 2.1(c)**. We emphasize that NMR confirmation of a compound with a known sequence (derived by MS) is much easier than NMR sequencing of a completely unknown compound. The crux of our approach is the analysis of the entire spectral network [15] of multiple Tyrocidines/Reginamides (**Figure 2.4(b-c)** and **Table 2.2 and 2.3**) rather than analyzing each Tyrocidine/Reginamide isomer separately. The derived sequences of the Reginamides represent the first automated sequencing of a cyclic peptide family *before* NMR and highlights the future role that mass spectrometry may play in sequencing cyclic peptides. MS-CyclicPeptide software is available from the NCRR Center for Computational Mass Spectrometry at http://proteomics.ucsd.edu.

2.2 Results

Spectral datasets. We analyzed Tyrocidine, Cyclomarin, and Reginamide families of cyclic peptides (see Methods section for the detailed description of experimental protocols).

The Cyclomarins represent a family of cyclic heptapeptides with anti-inflammatory activity, isolated from a marine *Streptomyces* strain [31, 32, 33]. The structure of Cyclomarin A is shown in **Figure 2.1(b)**. We sequenced four variants of the Cyclomarins that differ in a single amino acid residue.

The Reginamides represent a newly isolated family of cyclic octapeptides isolated from a marine *Streptomyces* strain that also produces secondary metabolites with anti-asthma activities (*Splenocins*). Multiple variants of Reginamide isomers were sequenced using MS. Due to limited quantities of these cyclic peptides and severe separation challenges, it was only possible to purify one of the variants (named Reginamide A) for validating the derived sequences by NMR. Multi-dimensional NMR analysis confirmed the sequence of Reginamide A, derived by our multiplex sequencing algorithm.

Sequencing of individual peptides. Below we describe an algorithm for sequencing *individual cyclic peptides.* The goal of this algorithm is not improving the method of [16], but rather proposing the ground for multiplex peptide sequencing, something that the algorithm from [16] is not suited for.

Consider the cyclic peptide VOLFPFFNQY (Tyrocidine A) with integer masses (99, 114, 113, 147, 97, 147, 147, 114, 128, 163). We will interchangeably use the standard notation (VOLF...) and the sequence of rounded masses (99, 114, 113, 147, ...) to refer to a peptide. One may partition this peptide into three parts as OLF-PFF-NQYV with integer masses 374, 391 and 504 respectively. In general, a k*-partition* is a decomposition of a peptide P into k subpeptides with integer masses $m_1 \dots m_k$ (we refer to $mass(P) = \sum_{i=1}^{k} m_i$ as the *parent mass* of peptide P). A k*-tag* of a peptide P is an arbitrary partition of $mass(P)$ into k integers. A k-tag of a peptide P is *correct* if it corresponds to masses of a k-subpartition of P, and *incorrect* otherwise. For example, (374, 391, 504) is a correct 3-tag, while (100, 1000, 169) is an incorrect 3-tag of Tyrocidine A.

A (linear) *subtag* of a cyclic k-tag $Tag = (m_1, \cdots, m_k)$ is a (continuos) linear substring $m_i \cdots m_j$ of the k-tag (we assume $m_i \cdots m_j = m_i \cdots m_k m_1 \cdots m_j$ in the case $j < i$). There are $k(k-1)$ subtags of a k-tag. The mass of a subtag is the sum of all elements of the subtag and the length of a subtag is the number of elements in the subtag. We define $\Delta(Tag)$ as the multiset of $k(k-1)$ subtag masses. For a peptide P, the *theoretical spectrum* of P is defined as $\Delta(P)$. For example, the theoretical spectrum of a cyclic peptide $AGPT = (71Da, 57Da, 97Da, 101Da)$ consists of 12 masses (57, 71, 97, 101, 128, 154, 172, 198, 225, 229, 255, and 269).

The problem of sequencing a cyclic peptide from a (complete and noiseless) spectrum corresponds to the *Beltway Problem* [34] and can be stated as follows:

Cyclic Peptide Sequencing Problem.

- *Goal:* Given a spectrum, reconstruct the cyclic peptide[1] that generated this spectrum.
- *Input:* A spectrum S (a set of integers).
- *Output:* A cyclic peptide P, such that $\Delta(P) = S$.

While the Beltway Problem is similar to the well-studied Turnpike Problem [35, 36], the former is more difficult than the latter one [34]. Moreover, de novo sequencing of cyclic peptides is much harder than the (already difficult) Beltway Problem. Indeed, the real spectra are incomplete (missing peaks) and noisy (additional peaks). **Table S2** represents an experimental spectrum of Tyrocidine A and illustrates that while the experimental spectrum captures many masses from the theoretical spectrum (45 out of 90 masses), it also contains 30 other masses (corresponding to noisy peaks and neutral losses). The limited correlation between the theoretical and experimental spectra makes the spectral interpretation difficult.

Given a tag Tag and an experimental spectrum S (represented as a set of integer masses), we define $Score(Tag, S)$ as the number of elements (masses) shared between $\Delta(Tag)$ and S (ignoring multiplicities of elements in

[1] We emphasize that the peptide might have amino acids with arbitrary masses, rather than the 20 standard amino acids.

$\Delta(Tag)$). For example, for the 3-tag $Tag = (374, 391, 504)$ of Tyrocidine A, $Score(Tag, S) = 5$, since the spectrum S contains 5 out of 6 elements in $\Delta(Tag) = (374, 391, 504, 765, 878, 895)$.

The problem of sequencing a cyclic peptide from an incomplete and noisy spectrum can be stated as follows:

Cyclic Peptide Sequencing Problem from Incomplete/Noisy Spectrum.

- *Goal:* Given an incomplete and noisy spectrum, reconstruct the cyclic peptide that generated this spectrum.
- *Input:* A spectrum S (a set of integers) and an integer k (peptide length)
- *Output:* A cyclic peptide P of length k, such that $Spectrum$ and $\Delta(P)$ are as similar as possible, *i.e.* $Score(P, S)$ is maximized among all cyclic peptides of length k.

A tag is *valid* if all its elements are larger than or equal to 57 (minimal mass of an amino acid). A valid $(k+1)$-tag derived from a k-tag Tag by breaking one of its masses into 2 masses is called an *extension* of Tag. For example, a 4-tag (374, 100, 291, 504) is an extension of a 3-tag (374, 391, 504). All possible tag extensions can be found by exhaustive search since for each k-tag $(m_1 \ldots m_k)$ there exist at most $\sum_{i=1}^{k} m_i$ extensions.

Our algorithm for sequencing individual peptides starts from scoring all 2-tags and selecting t top-scoring 2-tags, where t is a parameter. It further iteratively generates a set of all extensions of all top-scoring k-tags, combines all the extensions into a single list, and extracts t top scoring extensions from this list. **Table 2.1(a)** shows the reconstructed 7-tags for the Tyrocidine family and illustrates that the highest-scoring tags are incorrect for most Tyrocidines. However, by *simultaneously* sequencing pairs of spectra of related peptides, one can achieve better results. For the sake of simplicity, we illustrate how our approach works with integer amino acid masses. However, with available high precision mass spectrometry data we are able to derive the elemental composition of each amino acid (see **Text S5**).

Furthermore, we describe an algorithm for combining information from all high scoring tags to generate a *spectral profile* (**Figure 2.2**) that compactly represents all high-scoring tags (similar to sequence logos [37]). Each $Tag = (m_1 \dots m_k)$ with $\sum_{i=1}^{k} m_i = M$ defines an M-dimensional boolean vector $\overrightarrow{Tag}$ with 1s at k positions $\sum_{i=1}^{j} m_i$ for $1 \leq j \leq k$. For example, a tag (3,2,4) defines a vector 001010001. Given a vector $\mathbf{x} = x_1 \dots x_M$, we define its *$i$-shift* as the vector $x_{M-i+1}x_{M-i+2} \dots x_M x_1 \dots x_{M-i}$ and its *reversal* as the vector $x_M x_{M-1} \dots x_2 x_1$. We define the *reversed i-shift* as the reversal of the i-shift. For example, 2-shift of 001010001 is 010010100, and reversed 2-shift is 001010010. Given vectors $\mathbf{x}$ and $\mathbf{y}$, we define $alignment(\mathbf{x}, \mathbf{y})$ as a shift or reversed shift of $\mathbf{x}$ with maximum dot-product with $\mathbf{y}$. For $\mathbf{x} = 001010001$ and $\mathbf{y} = 101000000$, $alignment(\mathbf{x}, \mathbf{y}) = 101000100$.

Our algorithm for constructing the spectral profile (generated from a spectrum with parent mass M) starts from ordering t high-scoring k-tags $Tag_1 \dots Tag_t$ in the decreasing order of their scores and defines T_0 as an M-dimensional vector with all zeros. It proceeds in t steps, at each step aligning the tag Tag_i against the vector T_{i-1}. At step i, it finds $alignment(\overrightarrow{Tag_i}, T_{i-1})$ between $\overrightarrow{Tag_i}$ and T_{i-1} and adds it to T_{i-1} to form $T_i = alignment(\overrightarrow{Tag_i}, T_{i-1}) + T_{i-1}$. After t steps, the algorithm outputs the vector $\frac{T_t}{t}$ as the spectral profile.

For example, for Tyrocidine A, the two 7-tags with the highest scores are Tag_1 = (114, 147, 244, 260, 111, 119, 274) and Tag_2 = (114, 147, 244, 291, 80, 133, 260). After the first step, we form a vector $T_1 = \overrightarrow{Tag_1}$ with 1s at positions 114, 261, 505, 765, 876, 995 and 1269. At the second step, we align $\overrightarrow{Tag_2}$ and T_1 and form a vector T_2 with 1s at positions 765, 995, 796, 1009 and 2s at positions 114, 261, 505, 876, and 1269. Repeating these steps for 100 high-scoring tags for Tyrocidine A results in the spectral profile shown in **Figure 2.2(a)**. **Table S4** provides the annotations of the spectral profiles for Tyrocidine A, B and C.

Sequencing of peptide pairs. We define a *spectral pair* as spectra S and S' of peptides P and P' that differ by a single amino acid. Consider a spectral pair (S, S') and set $\delta = Mass(S') - Mass(S)$. Given a k-tag $Tag = (m_1 \dots m_k)$ of a spectrum S and an offset δ, we define a *corresponding*

k-tag $Tag^i_{S\rightarrow S'} = (m_1 \ldots m_i + \delta \ldots m_k)$ of S' for each $1 \leq i \leq k$. For example for $Tag = (213, 260, 244, 147, 114, 128, 163)$ of Tyrocidine A, $Tag^1_{Tyc\ A\rightarrow Tyc\ A1} = (227, 260, 244, 147, 114, 128, 163)$ is the corresponding tag of Tyrocidine A1. Any k-tag of S corresponds to at most k k-tags of S', and any correct k-tag of S corresponds to (at least) one correct k-tag of S'. Given a k-tag Tag of a spectrum S, define its $PairwiseScore$ as

$$PairwiseScore(Tag, S, S') = \frac{Score(Tag, S) + \max_{1\leq i\leq k} Score(Tag^i_{S\rightarrow S'}, S')}{2}$$

The algorithm for pairwise sequencing of the cyclic peptides is exactly the same as the algorithm for sequencing individual cyclic peptide but instead of using $Score(Tag, S)$ for scoring a single tag, it uses $PairwiseScore(Tag, S, S')$. **Table 2.1(b)** shows that while pairwise sequencing improves on sequencing of individual cyclic peptides, it does not lead to correct reconstructions of all Tyrocidines.

Identifying spectral pairs. While the described algorithm assumed that we know which spectra form spectral pair, i.e. which peptides differ by a single substitution, such an information is not available in *de novo* sequencing applications. The problem of whether spectra of two *linear* peptides form a spectral pair was investigated by Bandeira et al., [15]. In this section we address a more difficult problem of predicting whether the spectra of two *cyclic* peptides form a *spectral pair* based only on their spectra. Our approach extends the dereplication algorithm from [16] by comparing spectra of mutated peptides (rather than comparing a spectrum against a sequence of a mutated peptide) and is based on the observation that related peptides usually have high-scoring corresponding tags. A simple measure of similarity between spectra is the number of (S, S')-shared peaks (see **Table S6**). In the following we introduce $\Delta(S, S')$ distance between spectra, that, in some cases, reveals the similarity between spectra even better than the number of (S, S')-shared peaks. Given a set of k-tags $TagList$ for a spectrum S, we define:

$$MaxScore(TagList, S) = \max_{Tag\in TagList} Score(Tag, S)$$

Given an additional spectrum S', we define:

$$MaxPairwiseScore(TagList, S, S') = \max_{Tag\in TagList} PairwiseScore(Tag, S, S')$$

Finally, given a set of k-tags $TagList$ for a spectrum S and a set of k-tags $TagList'$ for a spectrum S', define $\Delta(TagList, TagList', S, S')$ (or, simply, $\Delta(S, S')$) as the differences between the sum of scores of the best-scoring tags for S and S' and the sum of pairwise scores of the best-scoring tag of S/S' and S'/S pairs:

$$\Delta(S, S') = MaxScore(TagList, S) + MaxScore(TagList', S')$$
$$-MaxPairwiseScore(TagList, S, S') - MaxPairwiseScore(TagList', S', S)$$

It turned out that $\Delta(S, S')$ is a good indicator of whether or not peptides P and P' that produced S and S' are only one amino acid apart. **Table S6** illustrates that all seven spectral pairs of Tyrocidines have Δ less than or equal to five, while for remaining pairs, Δ is greater than or equal to seven, with exception of Tyrocidine A1/C1 pair representing two substitutions at *consecutive* amino acids FF → WW. Such substitutions at consecutive (or closely located) positions are difficult to distinguish from single substitutions. For example, the theoretical spectrum for FF → WW substitutions (each with 39 Da difference in the mass of amino acids) is very similar to the theoretical spectrum of a peptide with a single substitution on either of Phe residues with 78 Da difference.

Spectral Network Construction. Given a set of peptides $P_1, \cdots, P_m$, we define their *spectral network* as a graph with m vertices $P_1, \cdots, P_m$ and edges connecting two peptides if they differ by a single amino acid substitution. In reality, we are not given peptides $P_1, \cdots, P_m$, but only their spectra $S_1, \cdots, S_m$. Nevertheless, one can approximate the spectral network by connecting vertices S_i and S_j if the corresponding peptides are predicted to differ by a single amino acid, i.e. if $\Delta(S, S')$ is less than a threshold. **Figure 2.4(a)** show the *spectral network* of six Tyrocidines analyzed in [16].

Multiplex sequencing of peptide families. We now move from pairwise sequencing to multiplex sequencing of *spectral networks* of (more than two) related cyclic peptides. While we use the notion of spectral networks from [15], the algorithm for sequencing linear peptides from spectral networks (as described in [15]) is not applicable for sequencing cyclic peptides.

In *multiplex sequencing of peptide families*, we are given a set of spectra of peptides of the same length n, without knowing their amino acid sequences, and without knowing which ones form spectral pairs. Sequencing of individual cyclic peptides is capable of generating a set of candidate k-tags, that typically contains a correct tag (at least for k smaller than n). However, sequencing of individual spectra typically fails to bring the correct peptide to the top of the list of high-scoring peptides or even, in some cases, fails to place it in this list. To alleviate this problem, we analyze all spectra in the spectral network and introduce a *multiplex scoring* that utilizes the information from all spectra.

Below we formulate the multiplex sequencing problem. Given a spectral network G of spectra $\mathbf{S} = (S_1, \cdots S_m)$, we call a set of peptides $(P_1, \cdots P_m)$ *G-consistent* if for every two spectra S_i and S_j connected by an edge in G, P_i and P_j differ by a single amino acid.

Multiplex Cyclic Peptide Sequencing Problem.

- *Goal:* Given spectra of related cyclic peptides (of the same length) and their (estimated) spectral network, reconstruct all cyclic peptides that generated this spectra.
- *Input:* Spectra $\mathbf{S} = S_1, \cdots, S_m$, their (estimated) Spectral Network G, and an integer k.
- *Output:* A G-consistent[2] set of peptide $P_1, \cdots, P_m$ (each of length k) that maximizes $\sum_{i=1}^{m} Score(P_i, S_i)$ among all sets of G-consistent peptides of length k.

Let $\mathbf{S} = (S_1, \cdots S_m)$ be a set of spectra of m peptides forming a spectral network and let $\mathbf{Tag} = (Tag_1, \cdots, Tag_m)$ be a *multitag*, which is a set of tags such that Tag_i is a k-tag of spectrum S_i (for $1 \leq i \leq m$). In **Text S1** we describe multiplex scoring of multitags, taking into account dependencies between spectra in the spectral network. This is in contrast to scoring multitags as $\sum_{j=1}^{m} Score(Tag_j, S_j)$

[2] Since we work with estimated (rather than exact) spectral networks, the multiplex cyclic peptide sequencing may not have a solution (i.e. a set of G-consistent peptides does not exist). Given a parameter u, a set of peptides is called (G, u)-consistent if for all but u edges (S_i, S_j), P_i and P_j differ by a single amino acid. The algorithm address finding (G, u)-consistent sets of peptides for a small parameter u.

that is equivalent to independent optimization of individual scores on all individual k-tags. This approach will not give any payoff in comparison to individual spectral sequencing.

$MultiplexScore$ defined in **Text S1** scores a multitag against all spectra in the spectral network. However, generating a correct multitag from m lists of t top-scoring tags in spectra $S_1, \ldots, S_m$ is impractical since (i) the number of candidate multitags (t^m) is large, and (ii) some lists may not contain correct individual tags. We therefore generate candidate multitags from individual tags and score them against all spectra using $MultiplexScore$. **Figure 2.3** describes the algorithm for generating a k-multitag from a single individual k-tag using the spectral network G. Given a candidate individual tag Tag of a spectrum S_u, $1 \leq u \leq m$, our algorithm generates a candidate multitag $\mathbf{multitag}(Tag, u, \mathbf{S}, G) = (Tag_1, \cdots, Tag_m)$, satisfying $Tag_u = Tag$. Note that given a tag $Tag = (m_1, \cdots, m_k)$, the (i, δ)-modification of Tag is defined as $(m_1, \cdots, m_i + \delta, \cdots, m_k)$.

We now define *multiplex score* on an individual tag Tag of a spectrum S_u as follows:

$$MultiplexScore(Tag, u, \mathbf{S}, G) = MultiplexScore(\mathbf{multitag}(Tag, u, \mathbf{S}, G), \mathbf{S}, G)$$

The multiplex sequencing algorithm (i) generates lists of individual tags for each spectrum in the spectral network, (ii) constructs the spectral network G, (iii) selects an individual Tag that maximizes $MultiplexScore(Tag, u, \mathbf{S}, G)$ among all individual tags, and (iv) outputs $\mathbf{multitag}(Tag, u, \mathbf{S}, G)$ as the solution of the multiplex sequencing problem.

Multiplex sequencing algorithm is exactly the same as the individual sequencing algorithm, with the only difference that we use $MultiplexScore$ here, instead of $Score$ (individual sequencing). Again we start with high scoring 2-tags (in $MultiplexScore$ sense), and extend them, keeping t highest scoring tags in each step. **Table 2.1(c)** illustrates that the multiplex sequencing algorithm sequences all six Tyrocidines studied in [16] correctly.

Figure2.2 (b-d) shows spectral profiles for $t = 100$ high scoring tags of multiplex sequencing of Q-TOF spectra of Tyrocidines, Cyclomarins, and Reginamides.

Figure 2.4(b) and **Table 2.2** show spectral network and sequences of Tyrocidines, predicted by multiplex sequencing algorithm (using ESI-IT spectra, see **Text S3** for details). **Figure 2.4(c)** and **Table 2.3** show similar results for Reginamides (see **Text S4** for details).

To analyze Reginamides, the Q-TOF and ESI-IT tandem mass spectrometry data was collected on both ABI QSTAR and ThermoFinnigan LTQ. In both cases, sequencing of Reginamide A resulted in a sequence of integer masses (71, 113, 113, 128, 113, 147, 113, 113). Using accurate FT spectra collected on ThermoFinnigan, we further derived amino acid masses as (71.03729, 113.08406, 113.08405, 128.09500, 113.08404, 147.06849, 113.08397, 113.08402) that pointed to amino acids Ala (71.03711), Ile/Leu (113.08406), Lys (128.09496) and Phe (147.06841) and revealed the elemental composition. These sequences were further confirmed by NMR (see **Text S6**).

2.3 Methods

Generating mass spectra. Q-TOF tandem mass spectrometry data for Tyrocidines, Cyclomarines, and Reginamides were collected on ABI-QSTAR. In addition, ESI-IT tandem mass spectrometry data were collected for Tyrocidines and Reginamides on a Finnigan LTQ-MS. All spectra were filtered as described in [16, 38] by keeping five most intense peaks in each 50 dalton window. All masses were rounded after subtraction of charge mass and multiplication by 0.9995 as described in [39]. High resolution FT spectra of Reginamides were also collected on a Finnigan. Typical mass accuracy of IT instruments are between 0.1 to 1 Da, while typical accuracy of TOF and FT instruments are between 0.01 to 0.1Da, and 0.001 to 0.01Da respectively.

Isolation of Reginamide A. CNT357F5F5 sample was obtained from a cultured marine streptomyces in five 2.8 L Fernbach flasks each containing 1 L of a seawater-based medium and shaken at 230 rpm at 27 ℃. After seven days of cultivation, sterilized XAD-16 resin was added to adsorb the organic products, and the culture and resin were shaken at 215 rpm for 2 hours. The resin was filtered through cheesecloth, washed with deionized water, and eluted with acetone. Pure

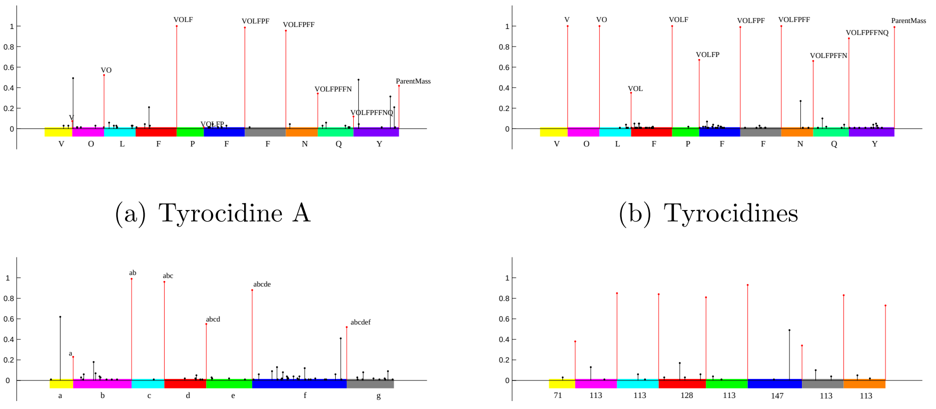
Reginamide A eluted at 12.6 min to give 2.0 mg of pure material.

Generating NMR spectra. CD_3OD and C_5D_5N were purchased from Cambridge Isotope. ^1H NMR, ^{13}C NMR, $^1H-{}^1H$ COSY, $^1H-{}^1H$ TOCSY (mixing time 90 ms), HMBC (^2J or $^3J_{^1H-^{13}C}$ = 7 Hz), HSQC ($^1J_{^1H-^{13}C}$ = 145 Hz), and ROESY (mixing time = 400 ms) spectra were generated on the Bruker (AVANCE III 600) NMR spectrometer with 1.7 mm cryoprobe. All the NMR spectra are provided in the **Supplementary Information**.

Parameter Setting. Text S7 discusses setting of parameters of the algorithm.

2.4 Acknowledgement

SPG thanks Fundao para a Cincia e Tecnologia, Portugal, for a postdoctoral fellowship. This work was supported by US National Institutes of Health grants 1-P41-RR024851-01 and GM086283. This chapter is published at Journal of Computational Biology, 2011, 18(11), 1371-1381.



(a) Tyrocidine A

(b) Tyrocidines

(c) Cyclomarins

(d) Reginamides

Figure 2.2: (a) Spectral profile of 100 highest scoring 7-tags for Tyrocidine A. Intensities of correct peaks account for 68% of total intensity. (b) Spectral profile of 100 highest scoring 10-tags for Tyrocidine A generated by multiplex sequencing of Tyrocidines. Intensities of correct peaks account for 86% of total intensity. (c) Spectral profile of 100 highest scoring 7-tags generated for Cycolmarin A by multiplex sequencing of four Cyclomarins (Cycolmarin A, Cyclomarin C, Dehydro Cyclomarin A and Dehydro Cyclomarin C). For Cyclomarin A, amino acids a, b, c, d, e, f and g stand for Alanine (71Da), β-methoxyphenylalanine (177Da), Valine (99Da), N-methylleucine (127Da), 2-amino-3,5-dimethylhex-4-enoic acid (139Da), N-(1,1-dimethyl-2,3-epoxyprophyl)-β-hydroxytryptophan (286Da) and N-methyl-δ-hydroxyleucine (143Da). In Cyclomarin C, f is replaced by N-prenyl-β-hydroxytryptophan (270Da). Dehydrations also occur on residue f. Intensities of correct peaks accounts for 59% of total intensitites. (d) Spectral profile of 100 top scoring 8-tags of Reginamide A generated by multiplex sequencing of Reginamides. The top scoring 8-tag of Reginamide A, also verified by NMR, is $(71, 113, 113, 128, 113, 147, 113, 113)$. Intensities of correct peaks account for 81% of total intensity.

Table 2.1: Individual (a), pairwise (b) and multiplex (c) de novo sequencing of Tyrocidines. The correct tag is selected from the set of 1000 top-scoring tags (the top scoring correct tag and its rank are shown). **Table S3** shows the process of extensions of top scoring tags of Tyrocidine A from 2-tags to 7-tags. Rank $1 \cdots 7$ for the highest scoring tag of Tyrocidine A1 means that the seven highest scoring tags have equal score, and one of them is the correct tag. Composite masses such as [113+147] for Tyrocidine A mean that the sequencing algorithm returned 260Da instead of 113Da and 147Da corresponding to Leu and Phe. [99 + 114/128] for Tyrocidine A/A1 pair means that the mass 99 + 114 = 213 in the first position of Tyrocidine A is substituted by the mass 99 + 128 = 227 in Tyrocidine A1. Part (c) shows 10-tags resulting from multiplex sequencing of six Tyrocidines (projected to Tyrocidine A). Correct masses are shown in bold. MS stands for Multiplex Score, and WMS stands for weighted Multiplex Score (See **Text S2** for details).

Peptide	The highest-scoring correct 7-tag (among all generated tags)										Rank
Tyc A	[99+	114]	[113+	147]	97	147	147	114	[128+	163]	$384 \cdots 1000$
Tyc A1	[99+	128]	[113+	147]	[97+	147]	147	114	128	163	$1 \cdots 7$
Tyc B	[99+	114]	113	147	97	[147+	186]	114	[128+	163]	$14 \cdots 134$
Tyc B1	99	128	[113+	147]	[97+	186]	147	[114+	128]	163	$2 \cdots 13$
Tyc C	99	114	[113+	147]	[97+	186]	[186+	114]	128	163	$6 \cdots 72$
Tyc C1	99	128	[113+	147]	[97+	186]	186	114	[128+	163]	$4 \cdots 38$

(a) Individual

Pair	The highest-scoring correct 7-tag (among all generated tags)										Rank
Tyc A/A1	[99+	114/128]	[113+	147]	[97+	147]	147	114	128	163	$2 \cdots 5$
Tyc B/B1	99	114/128	[113+	147]	[97+	186]	147	[114+	128]	163	1
Tyc C/C1	99	114/128	[113+	147]	[97+	186]	186	[114+	128]	163	1
Tyc A/B	99	114	[113+	147]	[97+	147/186]	147	[114+	128]	163	$2 \cdots 6$
Tyc B/C	99	114	[113+	147]	[97+	186]	147/186	[114+	128]	163	1
Tyc A1/B1	99	128	[113+	147]	[97+	147/186]	147	[114+	128]	163	$1 \cdots 4$
Tyc B1/C1	99	128	[113+	147]	[97+	186+	147/186]	114	128	163	$43 \cdots 82$

(b) Pairwise

Family	Sequences (10-tags)										MS	WMS	Rank
Tyrocidines	**99**	**114**	**113**	**147**	**97**	**147**	**147**	**114**	**128**	**163**	232	29.14	1
	99	**114**	**113**	**147**	**97**	**147**	**147**	69	173	**163**	228	28.78	2
	99	**114**	141	119	**97**	**147**	**147**	**114**	**128**	**163**	222	28.14	3
	99	**114**	**113**	**147**	**97**	**147**	**147**	**114**	111	180	222	27.85	4

(c) Multiplex

goal: Given spectra of related cyclic peptides (of the same length), sequence of one of them, and their (estimated) spectral network, reconstruct all the cyclic peptides that generated this spectra.
input: Spectra $\mathbf{S} = (S_1, \cdots, S_m)$ of m related cyclic peptide, their (estimated) Spectral Network G, an integer k, a k-tag Tag of S_u for some $1 \leq u \leq m$, a scoring function $Score(Tag, S)$ for individual spectra.
output: an approximate solution **multitag**$(Tag, u, \mathbf{S}, G)$ of constrained multiplex cyclic peptide sequencing problem.

```
for $j = 1$ to $m$ do
    $Tag_j \leftarrow$ null
end for
$Tag_u \leftarrow Tag$
repeat
    $Change \leftarrow 0$
    for all spectral pairs $(S_j, S_r)$ in $E(G)$ do
        $\delta = ParentMass(S_r) - ParentMass(S_j)$
        if $Tag_j \neq$ null and $r \neq u$ then
            for $i = 1$ to $k$ do
                $Tag'_r \leftarrow (i, \delta)$-modified $Tag_j$
                if $Score(Tag'_r, S_r) > Score(Tag_r, S_r)$ then
                    $Tag_r \leftarrow Tag'_r$
                    $Change \leftarrow Change + 1$
                end if
            end for
        end if
    end for
until $Change = 0$
return $(Tag_1, \cdots, Tag_m)$
```

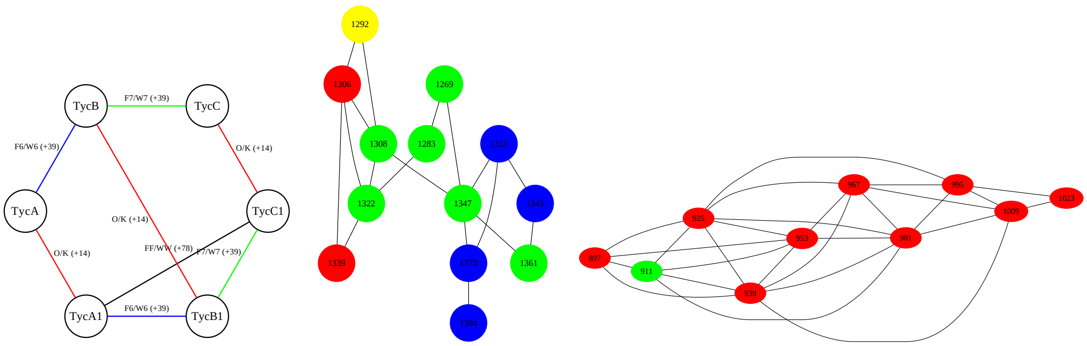
Figure 2.3: Algorithm for generating multitags from a candidate Tag of a spectrum S_u in the spectral network formed by spectra $S_1, \ldots, S_m$ corresponding to the spectral network G. Given a k-tag Tag of the spectrum S_u, the algorithm initializes $Tag_u = Tag$ and $Tag_j = Null$ for all other $1 \leq j \leq m$. We assume that $Score(Null, S_i) = -\infty$ for all $1 \leq i \leq m$. $E(G)$ stands for the edge set of the spectral network G. Since the sum $\sum_{i=1}^{m} Score(Tag_i, S_i)$ is monotonically increasing, the algorithm converges (typically after few iterations).

Table 2.2: Reconstructed peptides from the spectra corresponding to vertices in the spectral network shown in **Figure 2.4(b)**. The spectra were dereplicated using (known) Tyrocidines A, A1, B, B1, C and C1 by applying the **multitag** algorithm described in **Figure 2.3**. Four of the sequences are reported previously (see **Table S12**). For one spectrum with previously reported parent mass, 1292 Da, our reconstruction slightly differs from that of [1].

PM	Tag	Score	Comment
1269	99 114 113 147 97 147 147 114 128 163	21	Tyrocidine A
1283	99 128 113 147 97 147 147 114 128 163	26	Tyrocidine A1
1291	99 114 113 147 97 186 147 97 128 163	18	New
1292	99 114 113 147 97 186 131 114 128 163	22	PM matches Tryptocidine A[1]
1306	99 128 113 147 97 186 147 114 112 163	23	New
1308	99 114 113 147 97 186 147 114 128 163	25	Tyrocidine B
1322	99 128 113 147 97 186 147 114 128 163	32	Tyrocidine B1
1331	99 114 113 147 97 186 147 114 128 186	24	Tryptocidine B[1]
1345	99 128 113 147 97 186 147 114 128 186	27	previously reported[1]
1347	99 114 113 147 97 186 186 114 128 163	24	Tyrocidine C
1361	99 128 113 147 97 186 186 114 128 163	30	Tyrocidine C1
1370	99 114 113 147 97 186 186 114 128 186	26	Tyrocidine D[1]
1384	99 128 113 147 97 186 186 114 128 186	24	previously reported[1]

Table 2.3: Dereplication of Reginamide variants represented by the spectral network in the **Figure 2.4(c))** from the Reginamide A, using **multitag** algorithm.

PM	Peptide	Score
897	71 99 113 128 113 147 113 113	31
911	71 113 113 128 113 147 113 113	31
925	71 113 113 142 113 147 113 113	25
939	71 113 113 156 113 147 113 113	31
953	71 113 113 170 113 147 113 113	29
967	71 113 113 184 113 147 113 113	28
981	113 85 113 184 113 147 113 113	28
995	71 113 113 212 113 147 113 113	24
1009	113 113 113 184 113 147 113 113	26
1023	71 113 113 240 113 147 113 113	20



(a) Six tyrocidines (b) All tyrocidines (c) Reginamides

Figure 2.4: (a) The spectral network of six Tyrocidines analyzed in [16] reveals 7 (correct) spectral pairs differing by a single substitution and one (incorrect) spectral pair (Tyc A1 and Tyc C1) differing by two substitutions. (b) The spectral network of Tyrocidines after clustering similar spectra (see **Text S3** for details). The sequences were dereplicated from Tyrocidines A, A1, B, B1, C and C1 in **Table 2.2** (green node) using the **multitag** algorithm. (c) The spectral network of Reginamides after clustering similar spectra (see **Text S4** for details). The sequences were dereplicated from Reginamide A in **Table 2.3** (green node) using the **multitag** algorithm.

Chapter 3

Sequencing Cyclic Peptides by Multistage Mass Spectrometry

3.1 Introduction

Sequencing cyclic peptides, once a heroic effort, remains difficult today. The dominant technique for sequencing cyclic peptides is 2D nuclear magnetic resonance (NMR) spectroscopy, which requires large amount (miligrams) of highly purified materials that are often nearly impossible to obtain [28]. Tandem mass spectrometry (MS/MS) provides an attractive alternative to NMR since it allows one to sequence a peptide from picograms of non-purified material. However, the algorithms for interpreting mass spectra of cyclic peptides are still in infancy.

In the case when a cyclic peptide is a new variant of a known peptide family (differing from a known peptide by one or two mutations) *dereplication algorithm* presented in [16] can usually resequence the new variant. However, the approach works well when there exist a similar peptide and does not work for a peptide from a previously unknown family or distant homologs from the same family. De novo sequencing by mass spectrometry can be tricky even for linear peptides[1] [41, 42, 40], let alone for cyclic peptides. In the case of linear peptides, mass spectrometrists usually reserve to database search since it is more accurate

[1] De novo sequencing of linear peptides by MS/MS remains difficult and fails to correctly sequence $60 - 70\%$ of all spectra [40].

than de novo sequencing [43, 44]. The recently developed database search approach for spectra of cyclic peptides (Ng et al. [16]) only works if an identical or very close variant is present in a database of cyclic peptides. Since many cyclic peptides are either nonribosomal (and thus are not directly encoded by codons), or are generated by concatenating and cyclization of peptides from different proteins (e.g., θ-defensins [45]), the existing databases of cyclic peptides (e.g. NORINE [46]) are very limited and represent only a small fraction of cyclic peptides present in various organisms. Thus, in difference from linear peptides, de novo sequencing rather than database search represent the primary mode for analysing cyclic peptides.

Two approaches has emerged to improve accuracy of de novo sequencing of linear peptides: multistage mass spectrometry [47, 48] and spectral networks [15]. Both approaches use information about related peptides (either generated during multistage mass spectrometry experiment or naturally present in the sample) to synergistically sequence a peptide of interest. Both multistage mass spectrometry and spectral networks enable an ability to distinguish between C-terminal and N-terminal ion series [48, 49], a major obstacle in interpreting mass spectra [50, 51, 52].

While spectra of linear peptides are characterized by two ion series (N-terminal and C-terminal ions), spectra of cyclic peptides of length k have k ion series (each series correspond to subpeptides starting at position i of a cyclic peptide, $1 \leq i \leq k$). Thus, de novo sequencing of cyclic peptides can be more complex than sequencing of linear peptides. Similar to the case of linear peptides, one can think of two approaches for de novo sequencing of cyclic peptides: multistage mass spectrometry and spectral network analysis. While Ng et al. [16] presented the first automated algorithm for de novo sequencing of individual cyclic peptides, and Mohimani et al., [17] improved on [16] by applying the idea of spectral networks to cyclic peptides, the application of multistage mass spectrometry remains poorly explored for sequencing of cyclic peptides. We show that multistage mass spectrometry improves the quality of de novo sequencing of cyclic peptides as compared to single stage sequencing and illustrate its application to Reginamides, Etamycins, Dianthins and Tyrocidines.

3.2 Materials and methods

Spectral datasets. We analyzed cyclic peptides Reginamides, Tyrocidines, Etamycins and Dianthins using multistage mass spectrometry.

The *Reginamides* represent a newly isolated family of cyclic octapeptides isolated from a marine *Streptomyces* strain that also produces secondary metabolites with anti-asthma activities (Splenocins). Mohimani et al., 2010 [17], sequenced ten variants of Reginamides simultanously. In this paper we analyze the same ten variants of Reginamides using multistage mass spectrometry.

The antibiotic *Tyrothricin*, isolated from the soil microbe *Bacillus brevis* by Rene Dubos in 1939, is a classic example of a mixture of related cyclic decapeptides whose sequencing proved to be difficult and took over two decades to complete. Tang et al., [24] listed 28 known peptides from *B. brevis*. Mohimani et al. [17] showed how to sequence multiple variants of Tyrocidines, and even discover new variants from a single mass spectrometry experiment. In this paper we analyze six variants of Tyrocidines.

Etamycin is an antibiotic isolated from terrestrial actinomycete *S. griseus* alongside the streptogramin A antibiotic, and the two molecules together displayed bactericidal activity against some Gram-positive bacteria [53]. Recently, Etamycin is shown to be active against Methicillin-Resistant Staphylococcus aureus [54]. In this paper we analyse four variants of Etamycins.

Dianthins are cyclic peptides of variable length isolated from plant *Dianthus superbus*, which is used as a traditional Chinese medicine for the treatment of urethritis, carbuncles, and carcinoma [55, 56]. In this study we investigate five known Dianthins (Dianthins B-F) and discover six new variants. While dianthins B-F show some faint sequence similarities with each other, this level of similarity is insufficient for construction of the spectral network of dianthins, thus making the approach from [17] inapplicable.

For each of the above peptides, MS^3 and MS^4 spectra were collected by data dependent acquisition [57] using Thermo Scientific linear ion trap mass spectrometers. Thermo LTQ instrument was configured for the acquisition of up to 20 MS^3 spectra for each MS^2 spectra ($n_2 = 20$) and up to 20 MS^4 spectra for each

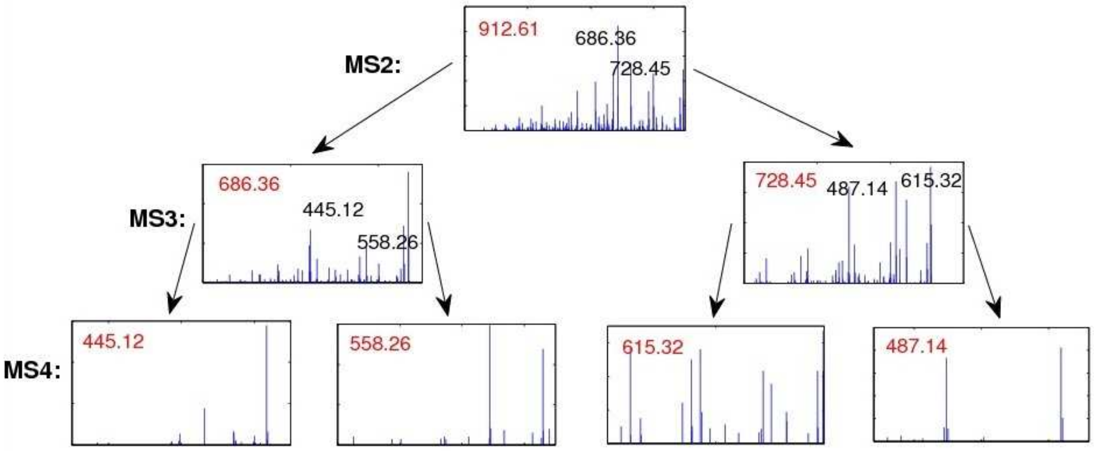
MS^3 spectra ($n_3 = 20$). Therefore, we have a single MS2, 20 MS3, and 400 MS4 spectra for each peptide investigated. **3.1** shows an example of MS^3 and MS^4 spectra acquisition.

Cyclic tags and linear subtags. Consider the cyclic peptide VOLFPFFNQY (Tyrocidine A) with integer masses (99, 114, 113, 147, 97, 147, 147, 114, 128, 163). One may partition this peptide into three parts as OLF-PFF-NQYV with integer masses 374, 391 and 504 respectively. In general, a *k-partition* is a decomposition of a peptide P into k subpeptides with integer masses $m_1 \dots m_k$ (we refer to $mass(P) = \sum_{i=1}^{k} m_i$ as the *parent mass* of peptide P). A *k-tag* of a peptide P is an arbitrary partition of $mass(P)$ into k integers. A k-tag of a peptide P is *correct* if it corresponds to masses of a k-subpartition of P, and *incorrect* otherwise. For example, (374, 391, 504) is a correct 3-tag, while (100, 1000, 169) is an incorrect 3-tag of Tyrocidine A.

A (linear) *subtag* of a cyclic k-tag $(m_1, \cdots, m_k)$ is a (continuos) linear substring $m_i \cdots m_j$ of the k-tag (we assume $m_i \cdots m_j = m_i \cdots m_k m_1 \cdots m_j$ in the case $j < i$). There are $k(k-1)$ subtags of a k-tag. The mass of a subtag is the sum of all elements of the subtag. The length of a subtag is the number of elements in the subtag. For example, $114, 260, 244, 147$ is a subtag of 7-tag $(99, 114, 260, 244, 147, 242, 163)$ of Tyrocidine A with length 4 and mass of 765Da.

For a $Subtag = m_i \cdots m_j$, all the subtags contained in $Subtag$ that either start at m_i or end at m_j are called *children* of $Subtag$ and $Subtag$ is called their *parent*. A subtag of length k has $2(k-1)$ children. For example, subtag $260, 244, 147$ is a *child* of subtag $114, 260, 244, 147$, and $114, 260, 244, 147$ is parent of $260, 244, 147$.

Experimental ion tree. A multistage MS experiment generates multiple spectra $S^1, \cdots, S^t$ of related peptides. The experimental ion tree is a graph with vertices $(S^1, \cdots, S^t)$ where a vertex (spectrum) S^i is connected to a vertex (spectrum) S^j with a directed edge if S^j is a product spectra generated from a peak m in S^i. In this case we set $PrecursorMass(S^j) = m$ and $PrecursorSpectrum(S^j) = S^i$. The spectra of original peptide, S_r, is called the *root of ion tree.* Figure 3.1 illustrates (part of) experimental ion tree of Reginamide



(a)

Figure 3.1: Illustration of experimental ion tree of Reginamide A, a peptide with amino acid sequence AIIKIFLI and mass 912.59 (plus charge). 686.42 is the mass of $AIIKIF$ and $KIFLIA$. 728.47 is the mass of $IKIFLI$ and $IIKIFL$. 445.28 is the mass of $FLIA$. 558.37 is the mass of $IFLIA$. 615.46 is the mass of $IKIFL$, $KIFLI$ and $IIKIF$. 487.40 is the mass of $IFLI$.

A consisting of MS^2, MS^3 and MS^4 spectra.

Peptide Ion Tree Match Score (PITMScore) Assume we are given a CyclicPSM Score $CyclicPSMScore(Tag, Spectrum)$ that assign a score to each pair of cyclic tag and cyclic spectra (e.g. [17]) and a linearPSM Score $LinearPSMScore(Tag, Spectrum)$ that assign a score to each pair of linear tag and linear spectra (e.g. [42]). Then Given a peptide Ion Tree Match (Peptide, IT), PITMScore can be defined as:

$$PITMScore(Tag, IT) = CyclicPSMScore(Tag, S_r) + c_{depth(S)} \cdot \sum_{S \in V(IT) - S_r} linearPSMScore(Tag(S), S)$$

where V(IT) is the vertex set of Ion Tree. Tag is the defined in an upside-down order, from root to leaf, as follows: $Tag(S_r) = Tag$, and $Tag(S)$ is defined as the child of $Tag(precursor(S))$ that satisfies $mass(Tag(S)) = PrecursorMass(S)$ and Null[2] if no such tag exists[3]. $depth(S)$ is the distance of vertex S from the root S_r of the ion tree, and $c_2 \cdots c_n$ are parameters for an ion tree of depth n. Ideally,

[2] PSMScore(Null, .) is defined as zero.

[3] if more than one subtag satisfies the mass constraint, we define $Tag(v)$ as the subtag maximizing $linearPSMScore(Tag(v), spectrum(v))$.

one should learn and optimize these parameters from a larger collection of PITMs. However, due to unavailability of a large training set of PITMs, we simply assume $c_2 = c_3 = \cdots = c_n = 1$.

Now we define the *Multistage Cyclic Peptide Sequencing Problem.*

- *Goal:* Given an experimental ion tree, reconstruct the cyclic peptide (tag) that generates this ion tree.
- *Input:* An experimental ion tree IT, and a parameter k (tag length).
- *Output:* A cyclic tag Tag of length k that maximizes $PITMScore(Tag, IT)$.

To find the tag with maximum score against the given experimental ion tree, we adapt the branch and bound procedure, which is briefly described below.

```
goal: Given an experimental ion tree and a Peptide Ion Tree Match Score, construct a set of high
scoring peptides of length $k$.
input: an experimental ion tree $IT$, a Peptide Ion Tree Match Score $PITMScore$, a peptide length
$k$, and number of returned high scoring tags $t$.
output: the set $T_k$ of $t$ high scoring tags of length $k$.

  Find the set $T_3$ of all high score 3-tags by brute force search.
  for $u = 4$ to $k$ do
      Extend all $k - 1$-tags in $T_{k-1}$ to $k$-tags.
      Select $T_k$ as the $t$ top scoring $k$-tags of this extended set.
  end for
```

Figure 3.2: Finding high scoring k-tags using branch and bound approach.

A tag is *valid* if all its elements are larger than or equal to 57 Da (minimal mass of an amino acid). A valid $(k+1)$-tag derived from a k-tag Tag by breaking one of its masses into 2 masses is called an *extension* of Tag. For example, a 4-tag (374, 100, 291, 504) is an extension of a 3-tag (374, 391, 504). All possible tag extensions can be found by exhaustive search since for each k-tag $(m_1 \ldots m_k)$ there exist at most $\sum_{i=1}^{k} m_i$ extensions.

Our algorithm for sequencing cyclic peptides starts from scoring all 3-tags and selecting t top-scoring 3-tags, where t is a parameter. It further iteratively generates a set of all extensions of all top-scoring k-tags, combines all the extensions into a single list, score each k-tag using $MutiStageScore$, and extracts t top scoring extensions from this list. **Figure 3.2** shows the main steps of our algorithm.

3.3 Results

We tested multistage de novo sequencing on Reginamides, Tyrocidines, Etamycins and Dianthins **Table 3.1**. The multistage approach resulted in sequencing peptides that evade MS^2-sequencing [17, 16]. Previously described reconstructions (whenever available) are shown in **Table 3.2**.

Table 3.3 compares the result of mutistage analysis with the result of single (MS2) spectral analysis[4]. For each peptide $Peptide$, IT is a collection of a single MS^2, 20 MS^3 and 400 MS^4 spectra, each one with 20 highest intensity peaks, and $Spectrum$ is a single MS^2 spectra with 100 highest intensity peaks. We use the shorthands $S = CyclicPSMScore(Peptide, S_r)$, $MS = PITMScore(Peptide, IT)$. p_e is the emprical p-value of score of correct peptide among 1000000 randomly generated valid tags with length and parent mass similar to $Peptide$. Because of the limited number of randomly generated tags, many of empirical p-values are zero, and this makes it difficult to compare single stage and multi stage scores.

As an alernative benchmark we define local p-value, p_l as follows: Construct the set U of all the valid tags generated by substitution of a pair of adjacent masses m_i and m_{i+1} by $m_i + \delta$ and $m_{i+1} - \delta$ for $\delta \neq 0$ and $1 \leq i \leq k$[5]. For example given a integer sequence (99, 114, 113, 147, 97, 147, 147, 114, 128, 163), (96, 117, 113, 147, 97, 147, 147, 114, 128, 163) falls in U (only two adjacent masses have different values), while (99, 114, 113, 145, 97, 149, 147, 114, 128, 163) does not belong to U. p_l is defined as the ratio of tags in U having a score higher than or equal to the score of $Peptide$.

Figure 3.3 ilustrates distribution of both single stage and multi stage scores, on both the whole set of valid tags, and the restricted set U, for regi-

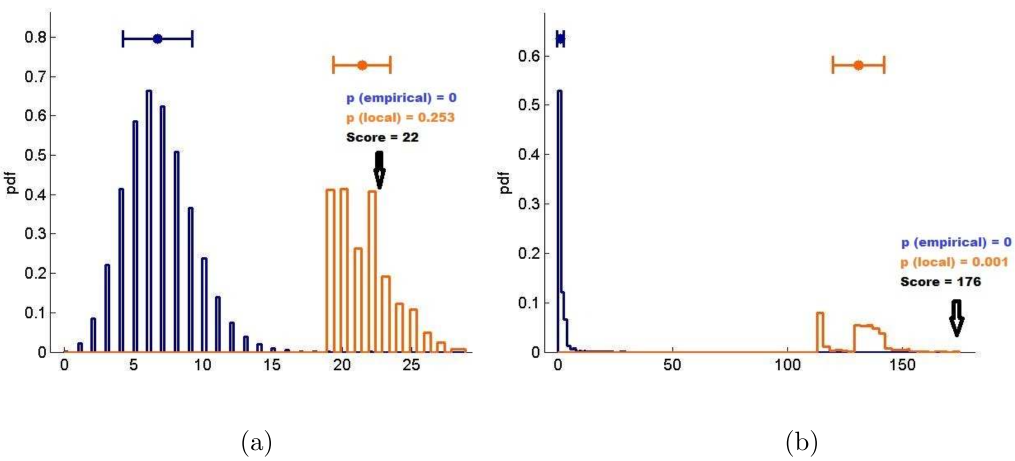
[4] For MS2 spectral analysis, we use the scoring function from [17] for benchmarking in **Table 3.3**.

[5] m_{i+1} is defined equal to $m_{i \ mod \ k \ +1}$.

Table 3.1: Multistage sequencing results. Masses that are verified by NMR are shown in bold. PM stands for Parent Mass of the peptide. For Tyrocidines, MS^2 Time of Flight (TOF) spectra is used in addition to MS^n ion trap (IT) spectra. Rank $1 \cdots 3$ for the highest scoring tag of Reginamide 925 means the three high scoring tags of Reginamide 925 have equal scores, and one of them is the tag shown. Asterisk on 147Da and 113Da means if we exchange these masses, the score wouldnt change. $222-18$ and $147+18$ masses for Etamycin 878 means instead of returning the correct masses 222Da and 147Da, the algorithm has returned 204Da and 165Da (this alternative breakage is also reported in [58]). $\rightleftarrows$ between 128Da and 113Da residues of Reginamide A means the algorithm has made a mistake in the order of those two residues, compared to previous reconstructions.

Peptide	MS^4 reconstruction										PM	rank
Reginamide A	**71**	**113**	**128** $\rightleftarrows$	**113**	**113**	**147**	**113**	**113**			911	$4 \cdots 6$
Reginamide 897	71	113	99	128	113	113	147	113			897	$2 \cdots 3$
Reginamide 925	71	113	99	156	113	147*	113*	113			925	$1 \cdots 3$
Reginamide 939	71	113	113	156	113	147*	113*	113			939	$4 \cdots 6$
Reginamide 953	71	113	170	113	113	147	113	113			953	$3 \cdots 4$
Reginamide 967	71	113	184	113	113	147	113	113			967	$24 \cdots 30$
Reginamide 981	71	113	113	85	226	147	113	113			981	$1 \cdots 2$
Reginamide 995	113	113	331	226	212						995	$3 \cdots 4$
Reginamide 1009	113	113	297	147	113	226					1009	$1 \cdots 5$
Reginamide 1023	113	113	797								1023	$5 \cdots 15$
Tyrocidine A	**99**	**114**	[**113**+	**147**]	[**97**+	**147**]	**147**	**114**	**128**	**163**	1269	$20 \cdots 44$
Tyrocidine A1	**99**	**128**	[**113**+	**147**]	**97**	**147**	**147**	**114**	**128**	**163**	1283	$22 \cdots 49$
Tyrocidine B	**99**	**114**	[**113**+	**147**]	**97**	**186**	**147**	[**114**+	**128**]	**163**	1308	$11 \cdots 19$
Tyrocidine B1	**99**	**128**	[**113**+	**147**]	**97**	**186**	**147**	[**114**+	**128**]	**163**	1322	$37 \cdots 105$
Tyrocidine C	**99**	**114**	[**113**+	**147**]	**97**	**186**	[**186**+	**114**]	**128**	**163**	1347	$67 \cdots 169$
Tyrocidine C1	**99**	**128**	**113**	**147**	**97**	**186**	**186**	**114**	**128**	**163**	1361	$10 \cdots 33$
Etamycin 878	**71**	**141**	**71**	**113**	**113**	**222** − 18	**147** + 18				878	$5 \cdots 8$
Etamycin 864	71	127	71	113	113	222 − 18	147 + 18				864	$1 \cdots 3$
Etamycin 862	71	141	71	97	113	222 − 18	147 + 18				862	$9 \cdots 12$
Etamycin 858	71	141	71	113	113	222 − 18	127 + 18				858	$11 \cdots 12$
Dianthin F	**57**	**97**	**99** $\rightleftarrows$	**147**	**147**						547	$13 \cdots 20$
Dianthin 564	57	113	113	71	97*	113*					564	$6 \cdots 14$
Dianthin E	**113**	**87**	[**147**+	**99**+	**57**+	**97**]					600	$7 \cdots 36$
Dianthin 610	97	99	[97+	57]	113	147					610	$7 \cdots 11$
Dianthin 624	57	97	147	113	97	113					624	$5 \cdots 9$
Dianthin 640	57	113	113	[97+	147]	113					640	$25 \cdots 66$
Dianthin 644	57	97	99	147	147	97					644	1
Dianthin B	**113**	**147**	**398**								658	1
Dianthin 672	113	559									672	$1 \cdots 6$
Dianthin C	**57**	**147** $\leftrightarrows$	**97**	**163**	**99**	**113**					676	$5 \cdots 7$
Dianthin D	**87**	**113**	**97**	**97**	**113**	[**147**+	**57**]				711	$13 \cdots 18$

namide A. Empirical p-value is the ratio of the valid tags with score above the score of correct peptide, and local p-value is the ration of tags in U with score above the score of correct peptide. **Figure 3.3** shows while the empirical p-value can not differentiate between single stage and multi stage scores, local p-value of multi stage score is much lower than single stage score.



(a) (b)

Figure 3.3: (a) Distribution of single stage scores on all the valid tags (shown by blue) and tags in U (shown by orange). (b) Distribution of multi stage scores on all the valid tags (shown by blue) and tags in U (shown by orange). One can see empirical p-value of both single and mutistage scores are zero, but local p-values are positive.

Text S1 describes how to combine information from all high scoring tags to generate a spectral profiles, and **Figure S1** shows a comparison of MS2 and MS4 results using spectral profiles. **Text S2** shows a more comprehensive comparison of single-stage and multi-stage sequencing on synthetic data.

3.4 Discussion

Sequencing cyclic peptides adds two fundamental difficulties to the already challenging task of de novo peptide sequencing: the amino acid masses are not known in advance and the peptides are cyclic rather than linear. Current de novo sequencing algorithms cannot adequately address these difficulties. Using multistage mass spectrometry leads to multiple lower-quality spectra from shorter subpeptides that need to be integrated to reveal the sequence of the cyclic peptide. Although the theoretical problem of an interpretation of a multistage spectrum is difficult, we have shown that a tag-based approach works well in practice.

There is a catch-22 when it comes to using mass spectrometry for interpretation of cyclic peptides. On the one hand, there is hardly any MS data for

Table 3.2: Previous reconstructions for Reginamide A [17], Etamycin 878 [54], Dianthins [55, 56] and Tyrocdines [24]. For Etamycin 878, Reginamide A and Dianthins B and C the sequences are determined by NMR, while for Dianthis D-F the sequence ois determined by ESI-MS2. Orn stands for amino acid Ornithine. Hyp stands for HydroxyProline. Phg stands for Phenylglycine.

Compound	NMR reconstruction									
Reginamide A	Ala	Ile	Ile	Lys	Ile	Phe	Leu	Ile		
Tyrocidine A	Val	Orn	Lue	Phe	Pro	Phe	Phe	Asn	Gln	Tyr
Tyrocidine A1	Val	Lys	Lue	Phe	Pro	Phe	Phe	Asn	Gln	Tyr
Tyrocidine B	Val	Orn	Lue	Phe	Pro	Trp	Phe	Asn	Gln	Tyr
Tyrocidine B1	Val	Lys	Lue	Phe	Pro	Trp	Phe	Asn	Gln	Tyr
Tyrocidine C	Val	Orn	Lue	Phe	Pro	Trp	Trp	Asn	Gln	Tyr
Tyrocidine C1	Val	Lys	Lue	Phe	Pro	Trp	Trp	Asn	Gln	Tyr
Etamycin 878	Ala	MeLeu	N-MeGly	Hyp	Leu	Thr+Hpca	MePhg			
Dianthin B	Ile	Phe	Phe	Pro	Gly	Pro				
Dianthin C	Gly	Pro	Phe	Tyr	Val	Ile				
Dianthin D	Gly	Ser	Leu	Pro	Pro	Ile	Phe			
Dianthin E	Gly	Pro	Ile	Ser	Phe	Val				
Dianthin F	Gly	Pro	Phe	Val	Phe					

Table 3.3: Comparison of Single Stage and MultiStage spectra. *MultiS* refers to multistage score, while *S* refers to single stage score.

Compound	Single Stage (MS^2)			MultiStage (MS^2, MS^3 and $MS4$)		
	S	p	p_l	$MultiS$	p_e	p_l
Reginamide A	22	0	0.253	178	0	0.001
Tyrocidine A	30	0	0.107	45	0	0.017
Tyrocidine A1	30	0	0.080	42	0	0.018
Tyrocidine B	28	0	0.032	50	0	0.041
Tyrocidine B1	27	0	0.153	27	0	0.006
Tyrocidine C	27	0	0.025	26	0	0.035
Tyrocidine C1	32	0	0.006	25	0	0.011
Etamycin 878	22	0	0.014	64	0	0.009
Dianthin F	11	0	0.052	17	0	0.006
Dianthin E	9	0.061	0.079	6	0.001	0.031
Dianthin B	5	0.432	0.249	9	0	0.028
Dianthin C	14	0	0.030	39	0	0.009
Dianthin D	20	0	0.051	40	0	0.008

cyclic peptides because nobody knows how to interpret the spectra automatically, thus providing little incentive for generating large datasets. On the other hand, absence of MS data for cyclic peptides slows down development of algorithms for their interpretation because large MS datasets are needed to develop such algorithms. As has been the case with de novo sequencing of linear peptides, large MS samples can be used to derive elaborate statistical models. Since cyclic peptides are implicated in many biologically important processes (see [30, 29] for the role of cyclic peptides in chemical defense and communication), the time has come to generate large datasets of annotated spectra of cyclic peptides.

3.5 Acknowledgemet

This work was supported by US National Institutes of Health grants 1-P41-RR024851-01 and GM086283. This chapter is published at Journal of Proteomics, 11(18), 3642-50.

Chapter 4

Cycloquest: Identification of cyclopeptides via database search of their mass spectra against genome databases

4.1 Introduction

A growing number of cyclic peptides (cyclopeptides) that are biosynthesized by a ribosomal pathway have been discovered in recent years [59] (**Figure 4.1** and **Figure S1**). The cyclic nature of the backbone renders cyclopeptides impervious to the action of exopeptidases and provides protection in some cases from endoproteases. The cyclic backbone also imparts rigidity on these molecules, which may facilitate conformation-specific interactions with other proteins. A large proportion of cyclopeptides represent biologically important agents, such as antibiotics (e.g. subtilosin A from *Bacillus subtilis* [60, 61], microcin J25 from *Escherichia coli* [62] and Circulin A and B from *Bacillus circulans* [63, 64]), innate immune system peptides (e.g. θ-defensins from *Macaca mulatta* [45]), bacteriocins (e.g. uberolysin from *Streptococcus uberis* [65] and carnocyclin from *Carnobacterium maltaromaticum* [66]), toxins (e.g. amatoxin and virotoxin from *Amanita* fam-

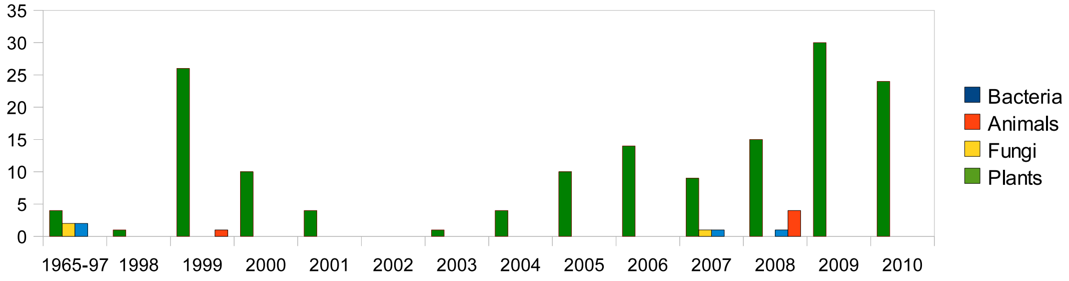


Figure 4.1: Ribosomal cyclopeptides appear in all domains of life. The number of cyclopeptides sequenced in 1965-2010. The majority of known cyclopeptides have been found in plants. The data on cyclopeptides prior to 2008 have been imported from Uniprot, and the data for 2009 and 2010 have been imported from Cybase [75]. The detail of cyclopeptides found before 1997 is shown in **Table S1**

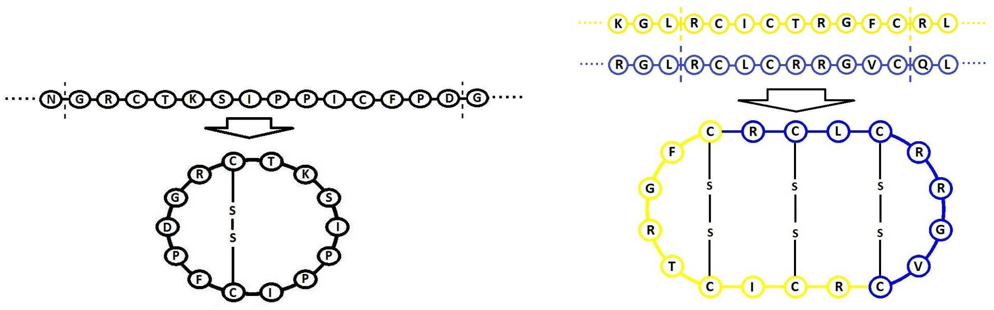
.

ily [67, 68]), protease inhibitors (e.g. SFTI-1 from *Helianthus annuus* [69]), bacterial cannibalism agents (e.g. SKF from *Bacillus subtilis* [70, 30]), agents active in plant defence (e.g. Kalata B1 from *Oldenlandia affinis* [71, 72], Cyclopsychotride A from *Psychotria longipes* [73] and Circulins from *Chassalia parvifolia* [74]) and many others. It seems that the world of ribosomal cyclopeptides is much more diverse than originally anticipated, and their structural diversities are only just beginning to be appreciated[59]. The availability of genomes for many species and our incomplete knowledge of the biosynthetic pathways employed by ribosomally synthesized cyclopeptides encourages us to use genome mining approaches in combination with mass spectrometry to discover novel cyclopeptides.

Sequencing cyclopeptides, once a heroic effort, remains a challenge today. Tandem mass spectrometry (MS/MS) provides an attractive alternative to 2D nuclear magnetic resonance (NMR) spectroscopy, as it can provide access to peptide sequence information from picograms of non-purified material [28]. However, the development of algorithms for the interpretation of mass spectra of cyclopeptides is still in its infancy. Non-ribosomal cyclopeptides are not encoded by nucleotide sequence in a genome through synthesize via mRNA to peptide. Instead, they are biosynthesized by large enzyme modules (nonribosomal peptide-synthetase), where each enzyme module is responsible for incorporating one amino acid subunit. Therefore mass spectrometrists must often conduct *de novo* interpretation of mass spectra[16, 17]. *De novo* peptide sequencing algorithms give promising results for short (up to 10 amino acid) cyclopeptides [17], but often fail to correctly sequence longer peptides.

Currently, peptide sequence tag [76] (PST)-based searches of genomes are the method of choice for sequencing longer ribosomally-synthesized peptides from mass spectrometry data. For example, using imaging mass spectrometry in conjunction with a five amino acid PST (LPHPA) search, Liu *et al.* [30] identified an active metabolites from the *Bacillus subtilis* cannibalism system. This metabolite was identified as a 26 amino acid peptide named sporulation killing factor (SKF). The success of the PST approach was critically dependent on the existence of a long series of consecutive ions with standard amino acid mass differences. The sequence tag is used to search against a database comprising proteins from the organism of interest. In the reported example, the sequence tag (LPHPA) yielded a single match when searched against the *Bacillus subtilis* proteome, however, the same tag could have many more matches if searched against larger proteomes. When a human proteome is queried with the same (LPHPA) for instance, 12 putative peptide matches results. For many species, the complete genome and hence proteome are not known and it is necessary to search against closely related species or larger databases to identify novel peptides. This implies a need for database search tools that can identify cyclopeptides, analogous to Sequest [43] and Mascot [44] for linear peptides. Recently, Colgrave *et al.* [77] proposed a method for the identification of known and novel cyclotides, a class of three-disulfide knotted plant cyclopeptides of 28-37 amino acids, by searching spectra of their linear derivatives against a database of all linearized products for all cyclotides from the Cybase database[75]. However, to date, no such database search method exists for the interrogation of genomes and proteomes.

Although most ribosomal cyclopeptides are formed via a head-to-tail ligation of a single peptide, the θ-defensins are generated by concatenation and cyclization of a pair of peptides from different proteins [45] (**Figure 4.2**). In contrast with linear peptide identification tools such as Sequest and Mascot, a cyclopeptide


N G R C T K S I P P I C F P D G
C T K S I P P I C F P D G R
S S
K G L R C I C T R G F C R L
R G L R C L C R R G V C Q L
C R C L C R R G V C R C I C T R G F
S S S S S S

(a)

(b)

Figure 4.2: Head-to-tail cyclization versus concatenated cyclization. (A) Head to tail cyclization of SFTI-1 from within PawS1. (B) Concatenated cyclization of θ-defensin. Nine amino acid segments of two different proteins (RTD1a and RTD1b) are concatenated.

database search tool must also address concatenation events in addition to the more commonly observed head-to-tail ligation events.

In this paper we present Cycloquest, the first database search algorithm for cyclopeptides. The search strategy is validated using sunflower trypsin inhibitor-1 (SFTI-1) and SFTI-like 1 (SFT-L1) from *Helianthus annuus*, sporulation killing factor (SKF) from *Bacillus subtilis*, and Rhesus θ-defensin (RTD-1) from *Rhesus macaque*. Our Cycloquest software for identifying cyclic peptides from their mass spectra is open source and available at http://proteomics.ucsd.edu .

4.2 Materials and Methods

Spectral datasets

Preparation of MALDI matrix (SFTI-1 and SFT-L1). A saturated solution of α-cyano-4-hydroxycinnamic acid (CHCA; Sigma Aldrich) was prepared by dissolving the matrix in 50% acetonitrile, 0.1% trifluoroacetic acid (TFA) with 5 mM ammonium phosphate to a final concentration of 5 mg/mL. The solution was vortexed thoroughly, sonicated in a water bath for several minutes, and centrifuged at 18,000 x g for 10 minutes at room temperature. The supernatant was used in the preparation of samples for MALDI-TOF MS.

Matrix assisted laser desorption/ionisation time-of-flight mass spectrometry (SFTI-1 and SFT-L1). Stock solutions (1 mg/mL) of sunflower trypsin inhibitor-1 (SFTI-1) or the peptide SFT-L1 were prepared in water and 1 μL mixed directly with the matrix (1:1, v/v). Aliquots (0.6 μL) of the mixtures were spotted on a 192 well plate (Applied Biosystems) and air dried. Mass analysis was carried out in positive ion reflector mode on a 4700 Proteomics Analyzer (Applied Biosystems) using a 200 Hz frequency tripled Nd:YAG laser operating at 355 nm. Fifty spectra at each of twenty randomly selected positions were accumulated per spot between 800 and 5000 Da using an MS positive ion reflectron mode acquisition method. MS/MS spectra were acquired at seven different laser energy settings from 4000-7000 (in increments of 500) and the spectra with optimum fragmentation was used for cyclopeptide sequencing. Calibration of the instrument was carried out using the MSCal1 peptide standard (Sigma Aldrich). Data were analyzed on the accompanying 4000 series Explorer Software.

Electrospray ionization ion trap mass spectrometry (SKF and RTD-1). SKF and RTD-1 were prepared to a concentration of 20 μg/mL in 50:50 methanol:water with 1% acetic acid and were then subjected to electrospray ionization on a Biversa Nanomate (Advion Biosystems, Ithaca, NY) nano-spray source (pressure: 0.3 psi, spray voltage: 1.4-1.8 kV). MS spectra were acquired on a 6.42 T Finnigan LTQ-FTICR MS or a Finnigan LTQ-MS (Thermo-Electron Corporation, San Jose, CA) running Tune Plus software version 1.0 and Xcalibur software version 1.4 SR1. For MS/MS experiment, the instrument was first autotuned on the m/z value of the ion to be fragmented. Then, the ions were isolated by the linear ion trap and fragmented by collision induced dissociation (CID) (isolation window: 3 m/z; collision energy: 30). Hundreds of MS/MS scans were acquired in centroid mode and averaged using QualBrowser software version 1.4 SR1 (Thermo). The Thermo- Finnigan RAW files containing the average spectra were then converted to mzXML file format using the program ReAdW (tools.proteomecenter.org).

Sodium Borohydride treatment of SKF. Dethiolated SKF was prepared by dissolving 1 μg of SKF with 1.5 μg $NaBH_4$ and 1.5 μg $NiCl_2$ in 6.25 μL of 60% $MeOH$. This reaction was incubated at 50, and an additional 1.5 μg of $NaBH_4$

and $NiCl_2$ were added into the reaction 5 and 10 minutes after initiation of the reaction to ensure complete conversion of SKF into dethiolated SKF. The mixture was then centrifuged for 1 min at 14,500 rpm to remove the insoluble particles and then purified by HPLC using an Agilent Eclipse XDB-C18 column running MeCN gradients or by C18 ZipTip (Millipore) following the manufacturers protocol prior to MS analysis.

PFA treatment of RTD-1.

The PFA treatment was performed using a four step process: (1) Peptide sample 0.1-10 μg (equivalent to 50-2000 pmol) was vacuum dried; (2) 19 volumes of 97% formic acid were mixed with 1 volume of hydrogen peroxide and allowed to stand on ice for 60 min; (3) $10\mu L$ of this reagent was added to the dried sample and incubated for 30 min at room temperature; and (4) the resulting solution was vacuum dried and washed three times with 50 μl of ice cold water.

Cycloquest algorithm.

Similar to the MS/MS database search algorithms employed by Sequest and Mascot, our database search consist of four steps: filtering the database (e.g. by parent mass as in Sequest or Mascot, by PST as in InsPecT, etc), constructing the theoretical spectra for candidate peptides, scoring the theoretical spectra against the experimental spectra, and finally, listing the top scoring peptide spectrum matches (PSMs). While the first and the last steps of our method are very similar to Sequest and Mascot, construction of the theoretical spectra and their scoring needed to be redefined for cyclopeptides. Fortunately, we could use the scoring defined in [16, 17] with slight modifications in the second and third steps of the algorithm. Another difference between Cycloquest and major database search algorithms is that Cycloquest uses a non-enzyme search strategy. The reason for this is two-fold. Many cyclic peptides are resistant to enzymatic digestion because of their compact and often disulfide-bonded nature. Additionally, digestion of cyclic peptides may result in formation of peptide fragments too small to analyze and too difficult to confidently identify. An additional step in cyclopeptide identification is to decide whether the spectrum is generated by a cyclic or a linear peptide. We address this additional complication in the Results section.

We defined a (linear) *subpeptide* of a cyclopeptide $Peptide = A_1A_2\cdots A_k$ as a (continuous) linear substring $A_i\cdots A_j$ of the peptide (we assume $A_i\cdots A_j = A_i\cdots A_kA_1\cdots A_j$ in the case $j < i$). There are $k(k-1)$ subpeptides of a peptide of length k. The mass of a subpeptide is the sum of masses of all its amino acids. We define the *theoretical spectrum* of a peptide, $\Delta(Peptide)$, as the multiset of $k(k-1)$ subpeptide masses. For example, the theoretical spectrum of a cyclopeptide AGPT = (71.037 Da, 57.021 Da, 97.052 Da, 101.047 Da) consists of 12 masses (57.021 Da, 71.037 Da, 97.052 Da, 101.047 Da, 128.058 Da, 154.073 Da, 172.084 Da, 198.099 Da, 225.110 Da, 229.105 Da, 255.120 Da, and 269.136 Da). We represented the experimental spectrum as a set of top t high intensity masses from the spectra, where t is a parameter. $CyclicScore_\delta(Peptide, S)$, the number of elements (masses) shared between theoretical spectrum of $Peptide$ and S within tolerance δ was defined (**Text S1**).

Distinguishing cyclic spectra from linear spectra

One of the challenges in identification of cyclopeptides is being able to distinguish between the spectra of linear and cyclic peptides. In this section we describe a method that given a spectrum and a protein database, enables the determination of whether the spectrum was derived from a cyclic or a linear peptide.

In addition to the $CyclicScore_\delta(Peptide, S)$ defined above, given a linear spectrum S and a peptide $Peptide$, we define the $LinearScore_\delta(Peptide, S)$ as the number of masses shared between S and the *linear theoretical spectrum* of $Peptide$ within the accuracy δ, where the linear theoretical spectrum is the set of $k-1$ b-ions and $k-1$ y-ions of $Peptide$ of length k (for CID spectra).

By using the cyclic and linear scores defined above, cyclopeptides can be distinguished from linear peptides based on the normalized score. Normalization is required due to different statistics of linear and cyclic scores (**Figure S2**). Moreover, peptides with different length have different statistics. Therefore, we normalize the score based on structure type (cyclic or linear) and peptide length. The normalized score of a match is equal to the difference of score of that match and average score of all the matches with the same length and peptides mass within 0.5 Da tolerance, over the standard deviation of all such matches.

MS/MS database search for concatenated peptides

The θ-defensin peptides are more difficult to identify than other cyclopeptides, because they are generated by concatenation and cyclization of peptides from two different protein precursors. It is computationally difficult to score the concatenation of each peptide pair over the entire macaque proteome (with 36,424 proteins totalling to 16,143,647 amino acids).

A similar problem arises for linear peptides known as the *fusion peptide identification problem*. While Ng and Pevzner [78] proposed a method for identification of the fusion peptides, their approach is not applicable to cyclopeptides. To address the quadratic growth of the number of generated concatenates, one needs a more efficient filter than the sole parent ion mass.

Many database search methods for linear peptides are *hybrid*, meaning that they attempt to use filters constructed by *de novo* searches for PSTs in order to speed up the database search by filtration using the found PSTs [79, 42, 40, 39]. The following subsection explains our approach for making the database search of concatenated peptides computationally feasible.

While fast implementations of linear peptide database search methods are based on PSTs, we used cycloPSTs (cyclo-Peptide Sequence Tags) to speed up our database search algorithm. Given a cycloPST $CycloPST = A_1A_2 \cdots A_k$ and a parent mass $ParentMass$, we define an artificial peptide $Peptide(CycloPST) = A_1A_2 \cdots A_kA_{k+1}$, where A_{k+1} is an artificial amino acid satisfying

$$mass(A_{k+1}) = ParentMass - mass(A_1) - \cdots - mass(A_k).$$

For example, for $cycloPST$ [156.10,57.02,99.06] corresponding to RGV and $ParentMass = 2086.24$, $Peptide(CycloPST) = [156.10, 57.02, 99.06, 1774.04]$. For a cycloPST $cycloPST$ and an experimental spectrum S, we define

$$CyclicScore_\delta(CycloPST, S) = CyclicScore_\delta(Peptide(CycloPST), S)$$

For example, for cycloPST [RGV] from θ-defensin with parent ion mass 2086.24:

$$CyclicScore_{\delta}([RGV], S) = CyclicScore_{\delta}([156.10, 57.02, 99.06, 1774.04], S)$$

.

Given an experimental spectrum S and a parent mass $ParentMass$, the first step of the algorithm consists of finding high scoring CycloPSTs. However, it is computationally difficult to try all 20^k length k cycloPSTs when k gets large. The strategy used in this study is the application of a branch and bound approach, in which all length three cycloPSTs are extended in each step by one of the 20 possible amino acids, and then a fixed number of high scoring cycloPSTs are selected for the next step. For θ-defensins, we use this approach to retain 1000 high scoring length nine cycloPSTs at each iteration. In this case, the list of 1000 high scoring cycloPSTs contained the correct cycloPST [RC*IC*RRGVC*R], where C* stands for cysteic acid, and all leucines are converted to isoleucines.

Given a high scoring cycloPST $CycloPST = A_1A_2 \cdots A_k$ of length k, we can generally divide it into two parts in $k - 1$ possible ways, i.e. $\{A_1|A_2 \cdots A_k\}$, $\{A_1A_2|A_3 \cdots A_k\}$, $\cdots$, $\{A_1 \cdots A_{k-1}|A_k\}$. For each of these divisions, we search both fragments in the genome and select the pairs of hits that can be extended to a pair of peptides with a total mass close to $ParentMass$. Assuming peptide concatenation is N-terminal to C-terminal (excluding infeasible N-terminal to N-terminal or C-terminal to C-terminal concatenations), we only accept pairs of peptides with matching directions. By concatenating each pair of peptides, we derive a set of candidate peptides which is much smaller than the original set. The final step is scoring all the candidate peptides using the cyclic score defined. **Figure 4.3** shows the steps of algorithm.

4.3 Results

Trypsin Inhibitor and Trypsin Inhibitor-Like peptides

The cyclopeptide SFTI-1 is a potent trypsin inhibitory peptide isolated from sunflower (*Helianthus annuus*) seeds. The peptide is 14 amino acids in length, and features a single disulfide bond and a head-to-tail cyclicized backbone[69]. The cyclic and braced nature of SFTI-1 makes the peptide more resistant to degradation than linear peptides of the same size and for this reason SFTI-1 has been extensively studied in the last decade as a potentially stable peptide-based drug template [80]. In addition to potent trypsin inhibition, SFTI-1 is shown to inhibit matriptase, a serine protease overexpressed in prostate and ovarian tumors, highlighting the importance of fast-tracking cyclic peptide discovery [81, 82]. Recently, Mylne *et al.* [83] reported the identification of a 12 amino acid peptide also isolated from sunflower seeds named SFT-L1 that shares some structural elements with SFTI-1 but lacks the trypsin inhibitory activity. SFTI-1 and SFT-L1 both emerge through proteolytic processing of much larger and functionally unrelated precursor proteins. SFT-L1 was identified through similarity of its precursor PawS2 to PawS1, the precursor of SFTI-1. SFT-L1 was manually sequenced by MS/MS, and its structure was obtained by NMR[83]. In this study, we determined the sequences of these cyclopeptides by searching the six frame translation of the sunflower nucleotide database using MS/MS spectra generated by MALDI-TOF/TOF mass spectrometry.

The lack of a complete sunflower genome required that we use the Expressed Sequence Tag (EST) library of seven *Helianthus* species available at the UC Davis Compositae Genome Project website, consisting of 136,935 cDNAs (totalling 96,493,071 nucleotides). Rather than covering the whole genome, ESTs only cover the RNA coding region of genome. With our interest in ribosomally synthesized cyclopeptides, searching ESTs is entirely suitable.

Both SFTI-1 and SFT-L1 contain a single disulfide bond that interferes with collision-induced dissociation during tandem mass spectrometric analysis.

The disulfide bonds were removed by reduction during sample preparation[1]. The theoretical mass to charge ratio (m/z) of SFTI-1 and SFT-L1 in the native form are 1513.73 and 1203.48 respectively. The theoretical m/z of reduced SFTI-1 and SFT-L1 are 1515.74 (observed 1515.72 Da) and 1205.49 Da (observed 1205.46 Da) respectively. The TOF spectra of SFTI-1 and SFT-L1 were collected with laser energy settings of 4500, yielding optimum fragmentation in each case to allow *de novo* sequencing and database searching. We also analyzed a synthetic linear version of SFTI-1 called SFTI-1[K,S], which corresponds to the peptide SIPPICFPDGRCTK, with reduced mass of 1533.67 Da.

The first step of the database search consisted of filtering the database by parent mass. **Table 4.1, 4.2, 4.3** show the top scoring hits for singly charged MALDI-TOF spectra of the sunflower peptides to the six frame translation of the EST library (assuming 0.5 Da mass accuracy for the parent ion mass). After scoring MS/MS fragments, normalizing scores, and sorting, both SFTI-1[K,S] and SFT-L1 are listed as the best match, while SFTI-1 is the third top match to its spectrum. In addition to the correct peptide sequence, there are some other high scoring hits from each spectrum to the database. These hits are usually computational artifacts. An additional validation step is usually required in order to distinguish the correct sequence from the shortlisted top scoring hits, e.g. by checking if the peptide is within known protein domains, in an ER signal sequence or a non-transcribed region of genomic DNA. For large proteomics datasets, false discovery rate (FDR) of the peptide sequence matches (PSMs) can be estimated to rule out false positives, similar to what occurs in the database matching of MS data to linear peptide.

Sporulation Killing Factor

When bacteria become cannibalistic, a differentiated subpopulation harvests nutrients from their genetically identical siblings to allow continued growth in nutrient-limited conditions[70]. One of the active metabolites in *Bacillus sub-*

[1] After reduction, the peptides can be alkylated to prevent reoxidation of the cysteines. The results for reduced and alkylated peptides are shown in **Table S2**.

Table 4.1: Top score reconstructions for SFT-L1 from a singly charged mass spectrum. Correct reconstructions are shown in bold.

Peptide	Struct	NScore	Score	PME	len
^{1a}**G**^{2}**CIEGSPVCFPD**^{1}**G**^{2}	cyclic	5.4	49	0.0	12
LRCLSVRKCQ	linear	5.1	10	0.2	10
CRLIFSLNHC	linear	5.1	10	0.1	10

a. Superscript numerals indicate alternative cyclization positions[2].

Table 4.2: Top score reconstructions for SFTI-1 from a singly charged mass spectrum. Correct reconstructions are shown in bold.

Peptide	Struct	NScore	Score	PME	len
WRSCVGGHCNIRQ	linear	5.9	11	-0.0	13
QTLIHNNGINCWC	linear	5.2	10	-0.0	13
^{1}**G**^{2}**RCTKSIPPICFPD**^{1}**G**^{2}	cyclic	4.9	44	0.0	14

tilis cannibalism is sporulation killing factor (SKF), a 26 amino acid cyclopeptide that is post-translationally modified with one disulfide and one cysteine thioether bridged to the α-position of a methionine[30]. After breaking the disulfide and thioether bridges, we were able to search for, and identify SKF in the proteome database of *Bacillus subtilis*.

The theoretical mass of SKF (with disulfide and thioether bridges) is 2781.30 Da (a triply charged ion of m/z 928.11 measured by FT-ICR [30] corresponds to a mass of 2781.30 Da, 1.5 ppm error). By sodium borohydride reduction, all the cysteines are reduced to alanine and all the methionines are reduced to homoalanines[3]. Sodium borohydride has no effect on any other standard amino acid. The theoretical mass of sodium borohydride reduced SKF is 2551.45 Da (a triply charged ion at m/z 851.49 measured by FT-ICR [30] was observed, which corresponds to a mass of 2551.44 Da, 2.2 ppm).

We use the proteome database of *Bacillus subtilis* available from UniProt with 4,188 proteins, totalling 1,230,503 amino acids.

Table 4.4 shows the top scoring hits for the electrospray ionization ion

[3] with mass of 85.0527 Da and composition C_4H_7ON

Table 4.3: Top score reconstructions for SFTI-1[K,S] from a singly charged mass spectrum. Correct reconstructions are shown in bold.

Peptide	Struct	NScore	Score	PME	len
SIPPICFPDGRCTK	linear	10.8	21	0.0	14
KYCLLLHRSACNL	linear	5.5	11	0.0	13
CVSSFSFSFSFWC	linear	5.5	11	-0.1	13

trap-generated spectra of the sodium borohydride reduced SKF to the *Bacillus subtilis* proteome database (assuming a 0.01 Da accuracy for the parent ion mass). The correct peptide is listed as the top scoring match.

Table 4.4: Top score reconstructions of SKF from a triply charged mass spectrum. Correct reconstruction is shown in bold.

Peptide	Struct	NScore	Score	PME	len
CMGCWASKSIAMTRVCALPHPAMRAI	cyclic	4.5	167	0.0	26
AKWLLSELNKLEKKERRKDW	cyclic	4.3	96	0.0	20
QSLKDLKGKTVGVQLGSIQEEKGK	cyclic	3.6	124	-0.0	24

Another active metabolites in *Bacillus subtilis* cannibalism is the killing factor (SDP), a 42 amino acid linear peptide that is post-translationally modified with a disulfide bond. We analyzed a triply charged native version of SDP, with triply charged parent mass ion at 1438 Da. Cycloquest identified SDP correctly, as the top hit to the *Bacillus subtilis* proteomic database (**Table 4.5**).

Table 4.5: Top score reconstructions of SDP from a triply charged mass spectrum. Correct reconstruction is shown in bold.

Peptide	Struct	NScore	Score	PME	len
CGLYAVCVAAGYLYVVGVNAVALQTAAAVTTAVWKYVAKYSS	linear	8.1	48	0.2	42
SVFFLWILNFVIGFAFPILLSSVGLSFTFFIFVALGVLA	linear	4.4	28	0.4	39
ELPGDLIARAQDILKELEHSGNKPEVPVQKPQVKEEPAQ	cyclic	4.0	316	0.3	39

θ-defensin

The first cyclopeptide discovered in animals was θ-defensin, an antimicrobial octadecapeptide that is expressed in the leukocytes of the *Macaca mulatta*. Like the previously characterized α- and β-defensin families, θ-defensins

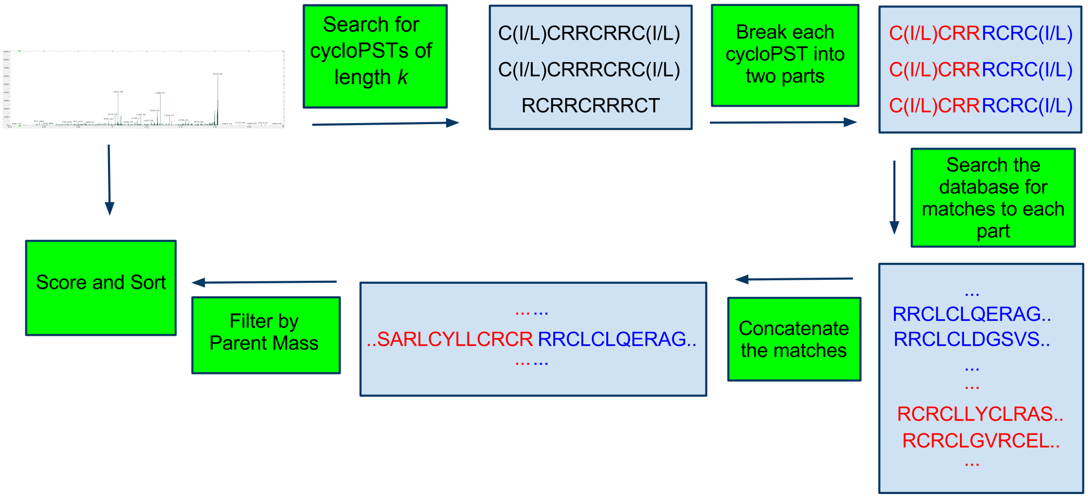


Figure 4.3: Steps of Cycloquest (for concatenated cyclopeptides such as RTD-1).

possess broad spectrum antimicrobial activities against bacteria, fungi, and protect mononuclear cells from infection by HIV-1[84].

We were able to identify the θ-defensin peptide using a doubly charged ion-trap (IT)-generated tandem mass spectrum. The theoretical mass of θ-defensin is 2079.90 Da (a doubly charged ion at m/z 1040.50 was observed, which corresponds to a mass of 2080.00 Da), and after performic acid treatment it increases to 2373.70 Da (a doubly charged ion at m/z 1188.50 was observed using ion-trap, which corresponds to a mass of 2374.00 Da), indicating the presence of three disulfide bonds.

Under PFA treatment, cysteine residues are modified to cysteic acid with a residue mass of 150.99 Da. According to Williams *et. al.* [85] only cysteine residues are affected by the treatment, and the on-target oxidation is not complicated by reactions with H, M, W or Y amino acid containing peptides.

Table 4.6 shows highest scoring hits to the triply charged IT spectra (assuming 0.5 Da mass accuracy for the parent ion mass).

False discovery rate of Cycloquest

In order to calculate false discovery rate, we tested the method on the previously published *Shewanella oneidensis MR-1* spectral data set containing 14.5

Table 4.6: Top score reconstruction for θ-defensin from a doubly charged IT spectrum. Correct reconstruction is shown in bold.

Peptide	struct	NScore	Score	PME	len
RCICTRGFCRCLCRRGVC	cyclic	15.3	160	-0.1	18
CLCRTPCNRCICTRGFCR	cyclic	13.9	149	-0.1	18
CRCRRCRCICTRGFCRL	cyclic	12.2	139	-0.1	17

million spectra. The spectra were acquired on an ion trap MS instrument (LCQ, ThermoFinnigan, San Jose, CA) using ESI. The protocol for acquiring the spectra and identifications from this data set is described in Gupta *et al.* [86]. 28,377 peptides were reliably identified with false discovery rate 5% using InsPecT [79] (spectrum-level false discovery rate (FDR) is 1%). We selected 21,087 tryptic peptides with a net charge of 2, obtained one representative spectra for each of these peptides (most peptides were identified from multiple spectra), and grouped these by the length of their peptide identifications to form a test data set for each length. We will refer to the length of the InsPecT identification of a spectrum as the *spectrum length.*

Our test set is a set of 1,663 spectra with spectrum length 12. We searched this dataset against the *Shewanella* database, and the corresponding decoy database. The classic reverse databases are not good candidates for decoy databases, because the theoretical spectrum of a cyclic peptide PEPTIDE, is exactly equal to the theoretical spectrum of the reverse cyclic peptide EDITPEP. Therefore, instead of using reverse sequences, the decoy is generated by shuffling the odd amino acids a_{2i+1} with the even amino acids a_{2i}, for a protein sequence $a_1, a_2, \cdots, a_n$. For example, the protein sequence PEPTIDE is shuffled to EPT-PDIE. After testing the method in this dataset using a parent ion mass accuracy of 0.5 Da and fragment ion mass accuracy of 0.5 Da, out of 1,663 spectra, the method classified 1595 of them as linear target hits, 25 as cyclic targets, 26 as linear decoys, and 17 as cyclic decoys. **Figure 4.4** shows the number of cyclic targets, linear decoys and cyclic decoys for different number of identifications. It takes about 35 minutes for Cycloquest to search 1663 Shewanella spectra against

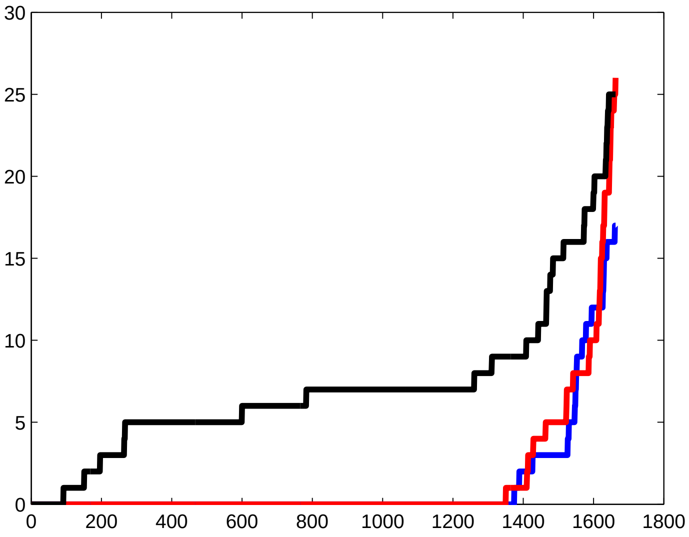


Figure 4.4: The number of decoy hits plotted against the number of identification. The number of cyclic target hits is shown in black, the number of linear decoy hits is shown in red, and the number of cyclic decoy hits is shown in blue.

Shewanella proteome (about 1 spectrum/second) on a 3.00 GHz Core 2 Duo CPU.

While this experiment indicates a small false positive rate for Cycloquest, we are unable to estimate false negative rate due to the unavailability of suitable spectral data sets for cyclopeptides.

4.4 Discussion

While the rate of the cyclopeptide identification has increased in recent years, computational approaches for the identification of cyclopeptides are still in their infancy. As a result, papers reporting new cyclopeptides typically discuss a single family of cyclopeptides per paper. In this study we have analyzed cyclopeptides from three different kingdoms.

We propose Cycloquest as a database search method for the identification of cyclopeptides from mass spectrometric data. The general steps of Cycloquest are similar to Sequest and Mascot. However, the scoring scheme used in Cycloquest is designed specifically for cyclopeptides. We demonstrated the utility of Cycloquest through its application to the sequencing of SFTI-1, a trypsin inhibitor from *Helianthus annuus* and a related peptide. Additionally, Cycloquest sequenced SKF, a bacterial cannibalism factor from *Bacillus subtilis*, and RTD-1, the θ-defensin from Rhesus macaque. Thus, Cycloquest is capable of correctly identifying all four of these cyclopeptides, opening a possibility of sequencing of novel cyclopeptides in future studies.

4.5 Acknowledgement

This chapter is published at Journal of Proteome Research, 2011, 10 (10), pp 4505-4512.

Chapter 5

A new approach to evaluating statistical significance of spectral identifications

5.1 Introduction

The dominant technique for sequencing cyclic peptides is nuclear magnetic resonance (NMR) spectroscopy, which requires large amount (milligrams) of highly purified materials that are often nearly impossible to obtain [28]. Tandem mass spectrometry (MS/MS) provides an attractive alternative to NMR because it allows one to sequence a peptide from picograms of non-purified material. Recently, new algorithms have been developed for interpreting mass spectra of cyclic peptides using de novo sequencing [16, 17, 18] and database search [87].

MS/MS coupled with database search is the most popular method for identification of (linear) peptides. A database search engine selects candidate peptides from a database of protein sequences that match the precursor mass from a mass spectrum. Then for each candidate peptide, the software compares a theoretical MS/MS derived from the peptide to the experimental mass spectrum, and reports a peptide with best score.

In the last decade, much effort has been invested in computing statistical significance of Peptide Spectrum Matches (PSMs). Many of these studies stem from the pioneering paper by Fenyo and Beavis [88] that proposed approximating the statistical significance of PSMs by first modeling the distribution of PSM scores (e.g. by Gumbel distribution [88]) and further using this distribution to compute p-values [89, 90, 91, 92, 93, 94, 95]. Unfortunately, this approximation approach, while useful in many applications, often fails when one has to estimate extremely small p-values typical for mass spectrometry (e.g. PSM p-values of the order 10^{-10} are often required to achieve 1% FDR [96]). Fortunately, the challenge of estimating the probability of extremely rare events has already been addressed by particle physicists in 1950s [97], and communication systems engineers in 1980s [98]. However, the mass spectrometry community has overlooked these fundamental studies (directly relevant to mass spectrometry) resulting in inaccurate p-value estimation in some mass spectrometry studies [99].

In the late 1940s, many top mathematicians worked on the *neutron shielding* problem that was crucial for designing nuclear facilities [100, 101]. In this problem, one has to compute the probability that a neutron, doing a random walk, would pass through a slab, an extremely rare event. Two general methods emerged for evaluating extremely rare events by Monte Carlo random sampling (using computers that became available in mid 1940s); *importance sampling* and *multilevel splitting*. Both were developed for nuclear-physics calculations by Fermi, Harris, Kahn, Metropolis, Ulam, von Neumann, and their colleagues, during the production of the first nuclear bomb [97, 100, 101, 102, 103]. Importance sampling is based on the notion of modifying the underlying probability distribution in such a way that the rare events occur much more frequently. Multilevel splitting uses a selection mechanism to favor the trajectories deemed likely to lead to the rare events of interest. While importance sampling is the most popular rare event simulation method today, the main advantage of the multilevel splitting approach is the fact that it does not need to modify the probabilistic model governing the system. This makes multilevel splitting applicable to any system represented as a black box [101], and specifically applicable to mass spectrometry studies. Kahn and Harris solved the neutron shielding problem using multilevel splitting in 1951 [97].

Later, similar rare event estimation techniques found applications in communication systems [98], financial mathematics [104], air traffic management [105], and chemistry [106]. However, this powerful approach has never been applied to mass spectrometry. This is surprising because there is a clear analogy between statistical significance evaluation in mass spectrometry, and the neutron shielding problem, where a spectrum plays the role of a neutron, a peptide plays the role of a slab, and the rare event "spectrum gets a high score against a peptide" plays the role of an event "neutron passes through a slab".

Currently, the dominant technique for statistical evaluation of a set of PSMs is to compute the False Discovery Rate (FDR) using the Target Decoy Approach (TDA) [107]. TDA is attractive for proteomics studies because it is widely applicable to different instrument platforms and database search algorithms. However, TDA is not applicable to non-linear peptide studies, because in these studies researchers usually work on a few non-linear peptide at a time, whereas TDA is best suited for statistical analysis of large spectral datasets [107]. Even in the case of linear peptides, some popular database search tools are not TDA-compliant [108].

An alternative technique is to compute a p-value for an *individual* PSM [99]. Given a PSM $(Peptide, Spectrum)$ of score t, the p-value of $(Peptide, Spectrum)$ is defined as the fraction of random peptides with a score equal to or exceeding t [99]. Unlike the FDR that is defined on a set of PSMs, the p-value is defined on a single PSM. Therefore computing the p-value is adequate for non-linear peptide studies, where a single or a few non-linear peptides are considered at a time. Since our results can be applied to both cyclic peptides (e.g. surfactin) and branch-cyclic peptides (e.g. daptomycin), we will use the same term 'cyclic' to refer to both cyclic and branch-cyclic peptides.

For cyclic peptide studies, computing p-values offers additional advantages. In studies of peptide natural products, we are given a mixture of spectra of linear and cyclic peptides, from which a small number of spectra of cyclic peptides should be separated and investigated independently. Therefore we need a method that identifies whether a given spectrum represents a linear or a cyclic peptide. This is difficult because different scoring functions are used for linear and cyclic peptides.

Since scores from different scoring functions are not usually comparable [109, 110], we need to convert them into p-values (**Fig. 5.1**) [108].

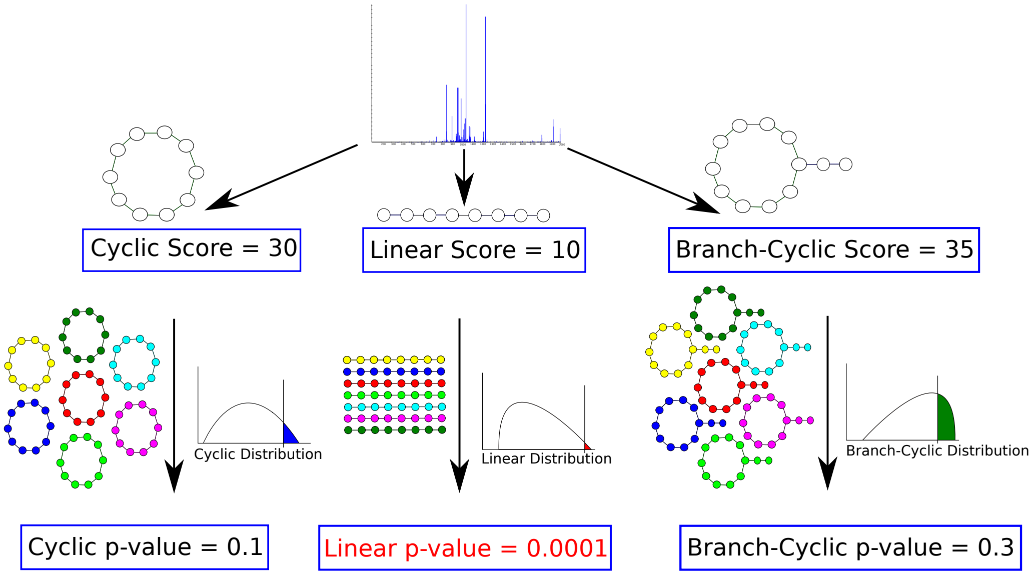


Figure 5.1: Deciding whether a peptide that produced a spectrum is linear, cyclic or branch-cyclic. Given a spectrum with unknown structure, we compute its score under different structure assumptions (e.g. linear/cyclic/branch-cyclic), and derive a p-value for each assumption. If one of the structures result in a very small p-value (e.g. linear structure with p-value of 0.0001), that structure is accepted as the most likely structure.

In the case of linear peptides, Kim *et al.*, 2008 [99] presented a polynomial time algorithm for computing p-values, called MS-GF. However, MS-GF is only applicable to scoring functions that can be represented as a dot-product of vectors, i.e. *additive scoring functions*. Moreover, MS-GF is only applicable to linear peptides, and no one has generalized MS-GF to non-linear peptides yet.

Fenyo and Beavis [88] constructed an empirical score distribution of low-scoring (erroneous) peptide identifications and extrapolated it to evaluate the p-value of high-scoring peptide identifications in the tail of the distribution. Similar approaches are now used in many tools, that provide p-value or E-value of individual PSMs, e.g. OMSSA [111]. However, this approach was demonstrated to be inaccurate [99]. While the pitfalls of such approaches are well recognized in genomics, they remain under-appreciated in proteomics. Waterman and Vingron [112] argued that it is difficult to accurately estimate the extreme tails of a distribution in general, requiring accurate estimation of rare event probability. To do so, one may consider estimating p-values by a Monte-Carlo simulation generating a population of millions of peptides and estimating the probability distribution of scores on this population [113]. This approach becomes time-consuming for estimating extremely low p-values, since it requires calculating scores of billions of randomly generated peptides for accurate estimation of p-values as low as 1 in a billion.

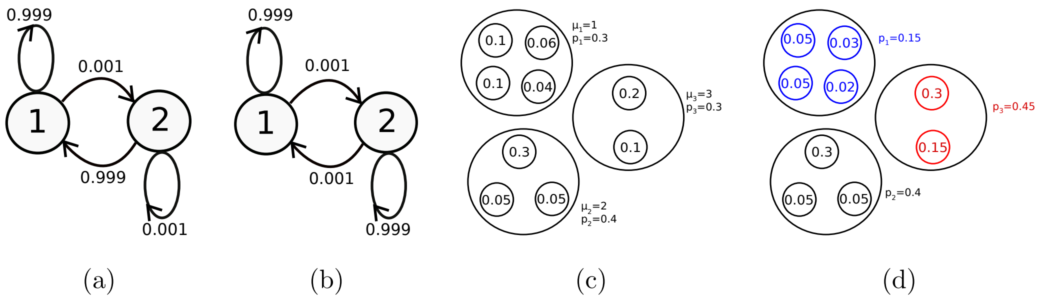
In this paper, we propose MS-DPR (MS-Direct Probability Redistribution), a new method for estimating p-values of PSMs based on rare event probability estimation by multilevel splitting. We show that MS-DPR reports p-values similar to those reported by MS-GF in the case of linear peptides, confirming that it accurately estimates p-values. Furthermore, we show that unlike MS-GF, MS-DPR can compute p-values of PSMs when an arbitrary (non-additive) scoring function is used or when the peptide is non-linear.

5.2 Materials and Method

In contrast to importance sampling, which changes the probability laws driving the model, multilevel splitting [97, 102] constructs a *Markov chain* and uses a selection mechanism to favor the *trajectories* in the Markov chain deemed likely to lead to rare events. Multilevel splitting is composed of three steps. First, decompose the trajectories to the rare events of interest into shorter sub-trajectories whose probability is not so small. Second, encourage the realizations that take these sub-trajectories (leading to the events of interest) by giving them a chance to reproduce by introducing *reproduction probabilities.* Third, discourage the realizations that go in the wrong direction by killing them with some positive *killing probability.* The sub-trajectories are usually delimited by levels. Starting from a given level, the trajectories that do not reach the next level will not reach the rare event, but those that do will split into multiple copies when they reach the next level. Each copy pursues its evolution independently from then on. This creates an artificial drift toward the rare event by favoring the trajectories that go in the right direction. In the end, an unbiased estimator can be recovered by multiplying the contribution of each trajectory by the appropriate weight [97].

While multilevel splitting has wide applicability across diverse fields, it is not clear how to select the reproduction and killing probabilities, and the number of offsprings in mass spectrometry applications. Inspired by Kahn and Harris [97] and proposed by Haraszti and Townsend [114], *Direct Probability Redistribution* (DPR) is a realization of multilevel splitting for estimating the probability of rare states in a Markov chain. Given a Markov chain, DPR implicitly constructs a modified Markov chain where probabilities of states are increased by an arbitrary order of magnitude. For a Markov chain with n states and (unknown) equilibrium probabilities $p_1, \cdots, p_n$, given *oversampling factors* $\mu_1, \cdots, \mu_n$, DPR constructs a Markov chain with (unknown) equilibrium probability $p'_1 = \mu_1 p_1 / \sum \mu_k p_k, \cdots, p'_n = \mu_n p_n / \sum \mu_k p_k$. For example, take a two-state Markov chain with equilibrium probabilities $p_1 = 0.999$ and $p_2 = 0.001$. If we choose $\mu_1 = 1$ and $\mu_2 = 999$, we end up with equilibrium probability $p'_1 = 0.5$ and $p'_2 = 0.5$, illustrated in **Fig. 5.2(A-B)**. If one decides to estimate probability distribution of **Fig. 5.2(a)** by Monte Carlo, thousands of simulations are required (since $p_2 = 0.001$ is small). However, if one tries to estimate probability distribution of **Fig. 5.2(b)**, only a few simulations are sufficient (since $p_1 = p_2 = 0.5$ is not small). This contrast in the number of simulations is the key idea of DPR. Here we descibe how to apply DPR to the problem of estimating probability distribution of PSM scores.

For simplicity, we define a spectrum as a set of integer masses. A peptide of length k is defined as a string of k positive integers $Peptide = m_1 m_2 \cdots m_k$. The mass of the peptide is defined as the sum of all the integers in the string. A score of a PSM (*Peptide*, *Spectrum*) is denoted by $Score(Peptide, Spectrum)$. Note that the proposed algorithm works for an arbitrary set of amino acid alphabets, not only for the alphabet of 20 standard amino acids. Since nonribosomal peptides often contain non-standard amino acids, in this section we consider peptides in the alphabet of all integers. In the Result section we also consider the case of the



(a) (b) (c) (d)

Figure 5.2: (a) Markov chain before performing DPR, with equilibrium probabilities $(0.999, 0.001)$. (b) Markov chain after performing DPR, with equilibrium probabilities $(0.5, 0.5)$. (c) An example of a Markov chain with nine peptides in three score states (d) Probability distribution after performing DPR with oversampling factors $(\mu_1, \mu_2, \mu_3) = (1, 2, 3)$. The states with decrease in probability are shown in blue, and the states with increase in probability are shown in red.

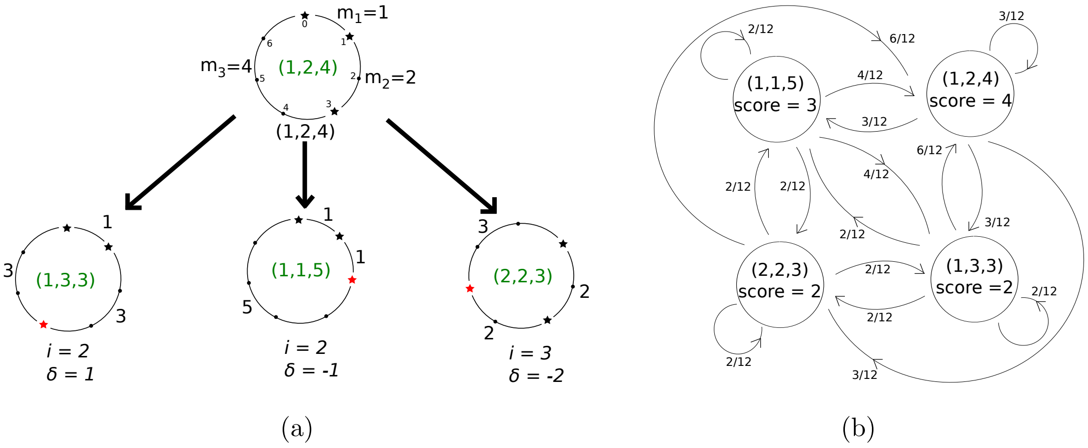
standard 20 amino acid alphabet.

Note that while a linear peptide of length k has a unique representation $m_1, \cdots, m_k$, a cyclic peptide of length k can have up to k equivalent representations. For example, peptide $(3, 7, 1)$ could also be presented as $(7, 1, 3)$ and $(1, 3, 7)$. One can choose an arbitrary representation among these representations, e.g., the representation where the first residue has minimum mass.

Given $Peptide = (m_1, \cdots, m_i, m_{i+1}, \cdots, m_k)$, integer residue index $1 \leq i \leq k$, and integer mass $-m_i < \delta < m_{i+1}$, we define $Peptide(i, \delta)$ as a peptide $(m_1, \cdots, m_i + \delta, m_{i+1} - \delta, \cdots, m_k)$. These peptides are called *sister peptides*. Note that sister peptides have equal lengths and equal (parent) masses, and they share all the amino acid masses but at most two (see **Fig. 5.3(a)**). There are many alternative ways to define the notion of a sister peptide. $RandomTransition(Peptide)$ is a $Peptide(i, \delta)$, where i and δ are integer random variables, i chosen from the uniform distribution on $[1, k]$, and δ chosen from the uniform distribution on $[-m_i, m_{i+1}]$. We define $PeptideSpace$ as the set of all peptides with length k and mass m. Consider the following Markov chain defined on $PeptideSpace$:

$$Peptide_{t+1} = RandomTransition(Peptide_t)$$

where $Peptide_0$ is chosen from $PeptideSpace$ with uniform distribution. Then the problem of finding probability distribution of all scores of peptides from *PeptideSpace* against *Spectrum* is equivalent to finding equilibrium distribution of the above Markov chain. We use the DPR technique to accurately estimate the total probability of all peptides with high scores (rare events) in this Markov chain. **Figure 5.3(b)** illustrates this Markov chain.



(a) (b)

Figure 5.3: (a) Illustration of all sister peptides $(1, 3, 3)$, $(1, 1, 5)$ and $(2, 2, 3)$ for the cyclic peptide $(1, 2, 4)$. (b) Illustration of the Markov chain for cyclic peptides of length 3 and mass 7. We have total of four different cyclic peptides, $(1, 1, 5)$, $(1, 2, 4)$, $(1, 3, 3)$, and $(2, 2, 3)$. Each random mutation is determined by selecting i (three cases), and δ (four cases), giving rise to a total of twelve equiprobable mutations. Transition probabilities between different states of the Markov chain, derived from the uniform mutation probabilities $(1/12)$, are also shown for each edge in the Markov chain.

Assume the set of all feasible scores (called *score states*) is $ScoreSpace = \{1, \ldots, n\}$, with (unknown) probabilities $p_1, \cdots, p_n$. Assume arbitrary oversampling factors $\mu_1, \cdots, \mu_n$ are given. Then the DPR approach provides a way to modify the transition probabilities such that in the equilibrium distribution of the resulting Markov chain, the probability of states with score i are oversampled by a factor μ_i, i.e. $p'_1 = \mu_1 p_1 / \sum \mu_k p_k, \cdots, p'_n = \mu_n p_n / \sum \mu_k p_k$. An example of this procedure is shown in **Fig. 5.2(C-D)**. **Figure 5.4(a)** shows the MS-DPR algorithm, which is a modification of the original DPR algorithm [114].

Glasserman *et. al.*, 1998, [115] show that the optimal choice of $\mu_1, \cdots, \mu_n$

(with respect to reducing the number of trials to achieve the required accuracy for estimation of score distribution) is the one that makes all score states equiprobable, *i.e.* $(\mu_1, \cdots, \mu_n) = (1/p_1, \cdots 1/p_n)$. However, since in practice $p_1, \cdots, p_n$ are unknown beforehand, one needs their rough estimate to efficiently implement DPR. Our idea is to first run the algorithm with $\mu_1 = \cdots = \mu_n = 1$, and obtain a rough estimate of $p_1, \cdots, p_n$. Then we choose $\mu_k = 1/p_k$ in the next iteration. This procedure is summarized in **Fig. 5.4(b)**.

5.3 Results

We used the Standard Protein Mix database consisting of 1.1 million spectra generated from 18 proteins using eight different mass spectrometers [116]. For this study, we considered the charge 2 spectra generated by Thermo Electron LTQ where 1,388 linear peptides of length between 7 and 20 are identified with false discovery rate 2.5% using Sequest [43] and PeptideProphet [117] in the search against the *Haemophilus influenzae* database appended with sequences of the 18 proteins (567,460 residues). For testing MS-DPR on cyclic peptides, we use the dataset from the Cycloquest paper[87], that includes cyclopeptides SFTI-1 and SFT-L1 from *Helianthus annuus*, as well as a linear and a cyclic peptide, SDP and SKF, from *Bacillus subtilits*.

To apply MS-DPR, we first need to define scoring functions for linear and cyclic peptides. Linear theoretical spectrum of a peptide $Peptide = (m_1, \cdots, m_k)$, $LinearSpectrum(Peptide)$, is a set of $k-1$ b-ions and $k-1$ y-ions, where each b-ion is the mass of a prefix of the peptide plus rounded H^+ mass, $m_1 + \cdots + m_{j-1} + 1$, and each y-ion is the mass of a suffix of the peptide plus rounded H^+ and H_2O mass, $m_j + \cdots + m_k + 19$. Similarly to the Cycloquest paper [87], The cyclic theoretical spectrum of the peptide, $CyclicSpectrum(Peptide)$, is defined as the set of masses of its $k(k-1)$ substrings of the peptide, $m_i + \cdots + m_{j-1}$ ($m_i + \cdots + m_k + m_1 + \cdots + m_{j-1}$, if $i \geq j$), illustrated in **Fig. 5.5(a)**. For branch-cyclic peptide $Peptide$, $BranchCyclicSpectrum(Peptide)$ is defined as the union of $LinearSpectrum(Peptide_l)$ and $CyclicSpectrum(Peptide_c)$, where $Peptide_l$ is the linear part of $Peptide$ with cyclic tail assumed as a modification, and $Peptide_c$ is the cyclic part of $Peptide$ with the linear tail assumed as a modification, illustrated in **Fig. 5.5(b)**.

Similarly to the Cycloquest paper [87], $CyclicScore(Peptide, Spectrum)$ and $BranchCyclicScore(Peptide, Spectrum)$ are defined as the number of shared masses between $Spectrum$ with $CyclicSpectrum(Peptide)$ and $BranchCyclicS$ $pectrum(Peptide)$, respectively. For simplicity, score of linear peptide $Peptide$ and a spectrum $Spectrum$, $LinearScore(Peptide, Spectrum)$, is defined as the number of shared masses between $Spectrum$ and $LinearSpectrum(Peptide)$ (In our experiments we will also use advanced MS-GF scores for linear peptides). We emphasize that while we use the same "shared peak count" principle, the resulting scoring functions are very different in the case of linear, cyclic and branch-cyclic peptides.

In addition to the p-value computed by MS-DPR (denoted by p_{DPR}), we also compute the empirical p-value (denoted by p_E), using a Monte Carlo approach by generating millions (or even billions) of random peptides and estimating probability distribution. Moreover, p_{MS-GF} stand for p-value of MS-GF software tool [99] (with 20 standard amino acid assumption), while p_{GF} stands for exact score probabilities computed by the generating function approach [99] for the case of arbitrary masses of amino acids.

Figure shows the evolution of μ and $\mathbf{p}$ in three iterations of MS-DPR. $\mathbf{p} = (p_1, \cdots p_n)$ is the original probability distribution, $\mathbf{p}' = (p'_1, \cdots p'_n)$ is the modified probability distribution, and $\mu = (\mu_1, \cdots \mu_n)$ is the vector of oversampling factors. $\mathbf{p}'$ converges to uniform distribution, and $\mathbf{p}$ converges to the correct distribution p_{GF}.

To evaluate the accuracy of the MS-DPR approach, we used all 1388 identifications from the ISB database. We compared p_{DPR} and p_{GF} (**Fig. 5.7(a)**), under the following assumptions: (i) all integers are considered as possible masses of amino acids (typical assumption for analyzing non-ribosomal peptides in the alphabet of arbitrary amino acid masses [18]), (ii) p-values are computed under the assumption that peptides have fixed known length, and (iii) the shared peak count

Table 5.1: Comparison of theoretical p-value of cyclic PSM $(\mathit{Peptide}, \mathit{CyclicSpectrum}(\mathit{Peptide}))$, with the p-value estimated by MS-DPR with a million simulations.

Peptide	score	theoretical p-value	p_{DPR}
(10, 20, 40)	7	0.0025	0.0021
(10, 20, 40, 80)	13	1.42e-05	1.35e-05
(10, 20, 40, 80, 160)	21	2.59e-08	2.49e-08
(10, 20, 40, 80, 160, 320)	31	1.45e-11	1.09e-11
(10, 20, 40, 80, 160, 320, 640)	43	2.40e-15	6.49e-15
(10, 20, 40, 80, 160, 320, 640, 1280)	57	1.15e-19	2.71e-20

is used as score. A correlation $R^2 = 0.9998$ between the two p-values shows that our method accurately estimates the probability distribution. **Fig. 5.7(b)** shows the comparison with the p-values computed by actual MS-GF software tool for the case of the standard amino acids alphabet [99] (correlation of 0.9990). These small deviations of MS-DPR from the theoretical value are acceptable, as the accuracy of a Monte Carlo algorithm depends on the number of simulations.

To validate MS-DPR for cyclic peptides, we designed the following experiment. For cyclic peptide $\mathit{Peptide} = (10, 20, 40)$, and the spectrum $\mathit{Spectrum} = \mathit{CyclicSpectrum}(\mathit{Peptide}) = (10, 20, 30, 40, 50, 60, 70)$, $\mathit{CyclicScore}(\mathit{Peptide}, \mathit{Spectrum}) = 7$. In this case we have total of $\binom{70}{2}$ peptides of length three with mass 70, and six of them (rotations and reverse rotations of $(10, 20, 40)$), have score 7. Therefore, the exact p-value for score 7 in this case is equal to $6/\binom{70}{2} = 0.0025$, while MS-DPR returns 0.0021. **Table 5.1** shows comparison of theoretical and estimated p-value for some cyclic PSM of variable length.

To validate our approach for cyclic peptides and branch-cyclic peptides in practice, we compared p_{DPR} and p_E for Tyrocidine A and Daptomycin A21978C2 spectra. Tyrocidine A is a cyclic peptide with length 10 and mass 1269.7Da, and Daptomycin A21978C2 is a branch-cyclic peptide with length 14 and mass 1652.8Da. Three different scores are used: $\mathit{CyclicScore}$, $\mathit{MultiStageCyclicScore}$ defined in the multistage de novo sequencing paper [18], and $\mathit{BranchCyclicScore}$.

Figure 5.8 demonstrates that these approaches produce similar results for probabilities higher than 10^{-6}.

To validate efficiency of MS-DPR in identifying whether a spectrum is from a linear, or a cyclic peptide, we compare each spectrum in our dataset against the corresponding proteome. Cycloquest [87], a database search for identification of linear and cyclic peptides from the mass spectra, is used for searching these peptides, and MS-DPR is used to re-rank top scoring PSMs given by Cycloquest. For *Helianthus annuus*, we used the EST database described in the Cycloquest paper [87], for *B. subtilis* we used the genome available from Uniprot, and for ISB dataset, we used the 18 protein sequences. By calculating p-values of all PSMs, the method correctly identifies SFTI-1 and SFT-L1 as cyclic peptides with lowest p-values. SDP and SKF are also identified as linear and cyclic peptides with lowest p-values (**Table 5.2**). Among 1388 linear peptides from ISB dataset, 1358 (97.8%) are correctly identified as linear peptides, and 99.6% of linear peptide identifications have identical sequences with the ones found by the InsPecT database search tool [79]. Note that the standard ISB dataset does not contain any cyclic peptide, and all 31 cyclic PSMs are non-significant (p-values assigned are larger than 0.01). Lets define $p_{lin}(Spectrum)$ as the p-value of the most statistically significant linear PSM of *Spectrum*, and $p_{cyc}(Spectrum)$ as the p-value of the most statistically significant cyclic PSM. **Figure 5.7(c)** shows p_{lin} versus p_{cyc} for SFTI-1, SFT-L1, SKF, SDP and all spectra in ISB dataset. The figure shows that MS-DPR distinguishes cyclic peptides from their linear counterparts.

MS-DPR takes about one second per spectrum in the non-standard amino acid case and about one minute per spectrum in the standard amino acid case with MS-GF score. MS-DPR is specifically designed for computing p-values for cyclic, branch-cyclic and other non-linear peptides, where no alternative tools are available. We do not suggest using MS-DPR for linear peptides in the case of additive scoring function, where fast analytical solution is available. [99]. However, some MS/MS database search tools use non-additive scoring function and compute empirical estimates of p-values or E-values. Since these estimates may be inaccurate [96], MS-DPR may be used for validating or correcting these estimates.

Table 5.2: Top score reconstructions of three cyclic and one linear peptides from the Cycloquest paper (a) Top score reconstruction of SFT-L1 peptide from a singly charged ion-trap spectrum. Correct reconstructions are shown in bold. PME stands for Parent Mass Error. (b) Top score reconstruction of SFTI-1 peptide from a singly charged spectrum. (c) Top score reconstruction of SKF peptide from a triply charged ion-trap spectrum. (d) Top score reconstruction of linear SDP peptide from a triply charged ion-trap spectrum. Note that the previous version of Cycloquest [87] (that lacked the algorithm for computing p-values) was unable to identify SFTI-1.

Peptide	score	p-value	PME	len	struct
GCIEGSPVCFPD	49	5.2e-11	0.036	12	cyclic
ICTQGNCQLEP	13	1.5e-7	0.069	11	linear
LNICCNVEVAQ	11	9.9e-6	0.105	11	linear
GRCTKSIPPICFPD	42	1.7e-7	0.024	14	cyclic
ICKQRVACWKNKG	36	8.7e-7	0.083	13	cyclic
KKCQKEVIENVCL	35	2.2e-6	0.082	13	cyclic
PSTHCWHHGMTHC	35	2.2e-6	-0.137	13	cyclic
PPMTTQCNICSFSS	10	0.00017	-0.092	14	linear
CMGCWASKSIAMTRVCALPHPAMRAI	167	1.7e-15	0.007	26	cyclic
GERTKVAGVKEANKENVKAWLKD	120	7.9e-12	-0.055	23	cyclic
ESLLKAVRSLEADVYHLELKDAA	119	1.1e-11	-0.077	23	cyclic
KEDAEKRVKSNLTLEAIAKAENL	119	1.1e-11	-0.045	23	cyclic
LGVLFIWLVAASIIKWRRFTY	16	6.6e-06	0.040	21	linear
CGLYAVCVAAGYLYVVGVNAVALQTAAAVTTAVWKYVAKYSS	39	9.1e-15	0.24	42	linear
CLLHDPKVLILDEPTNGLDPAGIREIRDHLKKLTRERG	180	3.4e-6	0.37	38	cyclic
YLPQLRGPMMIFTKVGRMSLTCYLLHSIIGTVLFLRY	168	9.9e-6	0.34	37	cyclic

5.4 Discussion

Most of the computational techniques developed in mass spectrometry focus on linear rather than non-linear peptides. Hence, computational mass spectrometry has not benefited the field of natural products yet, where the majority of interesting peptides are cyclic or branch-cyclic. One of the important questions in the field of peptide natural products is how to determine the structure (linear/cyclic/branch-cyclic) and amino acid sequence of a peptide from its spectrum. Since scoring functions for linear, cyclic and branch-cyclic peptides are very different, converting these scores to p-values is the first step toward automated MS-based discovery of peptide natural products.

We presented MS-DPR, a method for estimating statistical significance of PSMs in mass spectrometry. In contrast to existing methods for estimating p-values, MS-DPR can work with arbitrary scoring functions and non-linear peptides. Comparison of p-values estimated by MS-DPR with the p-values given by the generating function approach [99] validated MS-DPR in the case of additive scoring function and linear peptides. While there is no method for computing exact p-value of cyclic PSMs for a comprehensive evaluation of MS-DPR in the case of cyclic peptides, incorporating p-values in the recently developed Cycloquest algorithm [87] improved its performance (e.g. identification of cyclic peptide SFTI-1 missed by Cycloquest in previous study).

In the case of non-linear peptides, we used the shared peak count to score PSMs. While advanced scoring algorithms accounting for peak intensities increase the number of identifications of linear peptides at a given FDR, such scoring methods are not currently available for non-linear peptides. This is partially due to the fact that there are not enough annotated non-linear peptide spectra to train scoring algorithms. Recently, the natural product community has started collecting large scale mass spectrometry datasets. Thus, development of more comprehensive scoring algorithms will be possible in the near future.

While we tested MS-DPR only on linear, cyclic, and branch-cyclic peptides, our method is independent of a specific peptide structure and specific score scheme used. By defining a proper scoring function and random mutation for each peptide structure, MS-DPR can convert the score to an accurate p-value.

Cycloquest web-server reporting MS-DPR p-values is available at http://cyclo.ucsd.edu. The source code for MS-DPR is freely available at http://proteomics.ucsd.edu.

5.5 Acknowledgement

We would like to thank Pieter Dorrestein and Wei-Ting Liu for making the spectral dataset for SKF and SDP available, and Michelle Colgrave, Aaron Poth and Joshua Mylne for making the spectral dataset for SFTI-1 and SFT-L1 available. This work was supported by the U.S. National Institutes of Health 1-P41-RR024851-01. We are grateful to Kira Vyatkina for insightful suggestions. This chapter is published at Journal of Proteome Research, 2013, 12 (4), pp 15601568.

Supporting Information Available: This material is available free of charge via the Internet at http://pubs.acs.org.

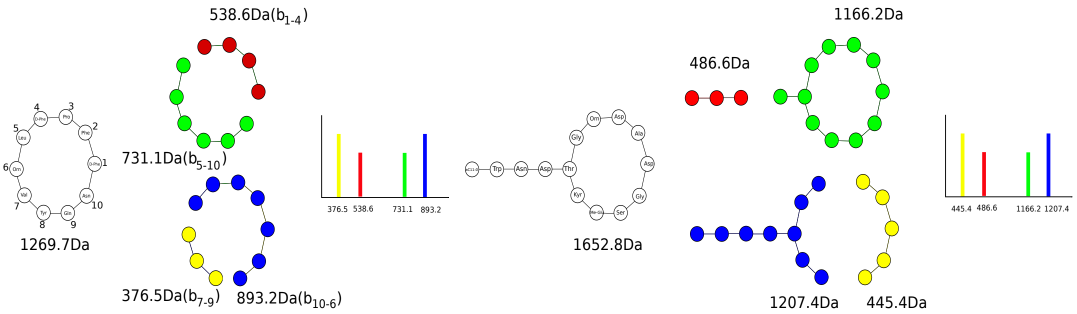
(a)

```
procedure MS-DPR-Iteration($\mu_1, \cdots, \mu_n$)
input: Spectrum $Spectrum$, score function $Score(Peptide) = Score(Peptide, Spectrum)$ with scores in $\{1, \cdots, n\}$ domain, random transition generator $RandomTransition(Peptide)$, number of output peptides $N$, and oversampling factors $\mu_1 \cdots \mu_n$.
output: An estimate of score probability distribution $p_1, \cdots, p_n$ on the score space.
select a random $Peptide_0$ from $PeptideSpace$
$z \leftarrow 0$ and $\mu_{min} \leftarrow min_{k=1,\cdots,n}\mu_k$
SimulateDPR($Peptide_0, \mu_{min}$)

procedure SimulateDPR($Peptide, \Omega$)
    while $z < N$ do
        $Peptide' \leftarrow RandomTransition(Peptide)$
        if $\mu_{Score(Peptide')} < \Omega$
            return
        if $\mu_{Score(Peptide')} > \mu_{Score(Peptide)}$
            $Y \leftarrow \mu_{Score(Peptide')}/\mu_{Score(Peptide)}{}^*$
            for $i = 1$ to $Y - 1$
                choose $\Omega'$ from the uniform distribution on $[\mu_{Score(Peptide)}, \mu_{Score(Peptide')}]$
                $SimulateDPR(Peptide', \Omega')$
            end
        end
        $z \leftarrow z + 1$
        $Peptide_z \leftarrow Peptide'$
    end
    return
end

for $k = 1$ to $n$        $n_k \leftarrow \#\{z|Score(Peptide_z) = k\}$.
    $p'_k \leftarrow n_k/N$.
    $p_k \leftarrow \frac{n_k/\mu_k}{\sum n_i/\mu_i}$.
end
return $(p_1, \cdots, p_n)$
```

(b)

```
procedure MS-DPR($K$)
input : Number of iterations $K^{**}$
output : an estimation of the probability distribution $p_1, \cdots p_n$
$(\mu_1 \cdots \mu_n) \leftarrow (1, \cdots, 1)$
for $j = 1$ to $K$
    $(p_1, \cdots, p_n) \leftarrow$MS-DPR-Iteration($\mu_1, \cdots, \mu_n$).
    $(\mu_1 \cdots \mu_n) \leftarrow (1/p_1, \cdots, 1/p_n)$
end return $(p_1, \cdots, p_n)$
```

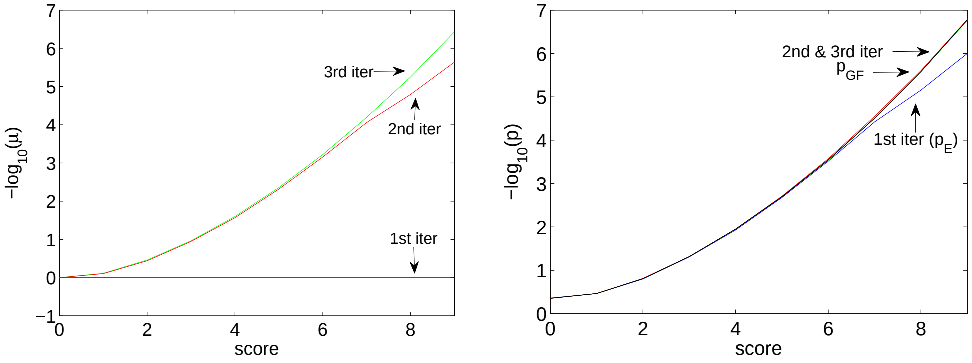
Figure 5.4: (a) MS-DPR-Iteration($\mu_1, \cdots, \mu_n$) algorithm[114] adapted for estimating statistical significance of PSMs. (b) MS-DPR(K) algorithm for estimating the probability distribution of scores.



(a)

(b)

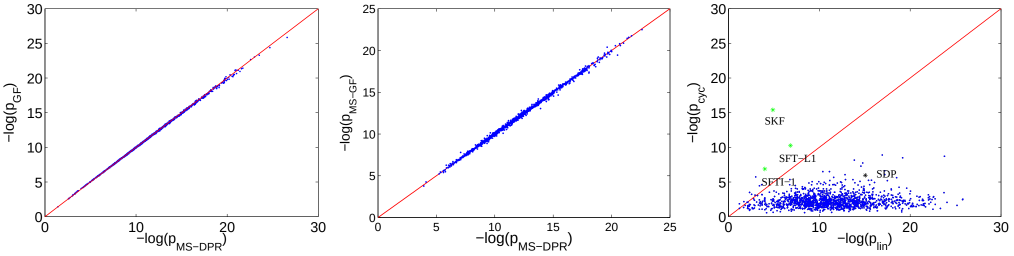
Figure 5.5: (a) Illustration of $CyclicSpectrum(Tyrocidine)$. (b) Illustration of $BranchCyclicSpectrum(Daptomycin)$.



(a)

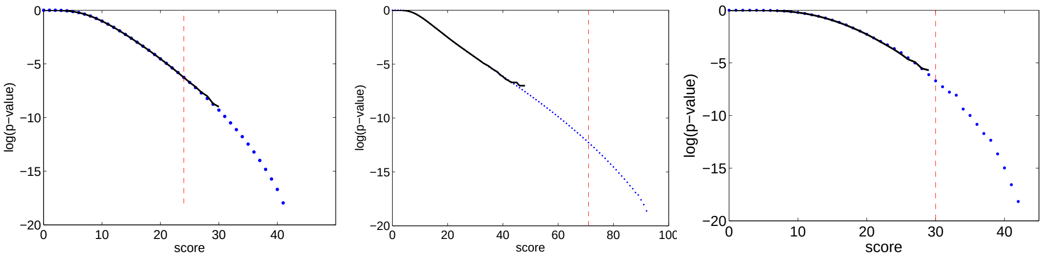
(b)

Figure 5.6: Evolution of (a) μ_k (b) p_k for three iterations of MS-DPR. The analysis is performed for $N = 1,000,000$ simulated peptides of length 7, and a spectrum of peptide KYIPGTK from standard ISB database with parent mass 787. Blue, red and green plot stands for first, second, and third iterations respectively. In part (b) p_{GF} is plotted by black. Note that the blue plot in part (b) corresponds to first iteration of MS-DPR, which simply gives the empirical p-value, p_E. From the second iteration on, p_{DPR} is very similar to p_{GF}.



(a) (b) (c)

Figure 5.7: (a) Comparison of $-log_{10}$ of generating function p-value with MS-DPR p-value for 1388 peptides from ISB database. Red line shows the $x = y$ line. Correlation between the two p-values is 0.9998. Non-standard amino acid model is used, assuming each peptide has a fixed known length, and peak count score. MS-GF approach [99] is modified accordingly, to satisfy these assumptions. (b) Comparison of $-log_{10}$ of the original, publicly available MS-GF p-value with MS-DPR p-value. Correlation between the two p-values is 0.9990 . Standard amino acid model is used, with the variable peptide length assumption and MS-GF score [99]. (c) Comparison of $-log_{10}$ of p_{lin}, versus $-log_{10}$ of p_{cyc} for SFTI-1, SFT-L2, SKF, SDP, and spectra from the ISB dataset. Cyclic peptides SFTI-1, SFT-L2 and SKF are shown as green stars, and linear peptide SDP is shown as a black star. Blue dots show spectra from ISB dataset, and red line shows the $x = y$ line.



(a) (b) (c)

Figure 5.8: (a) Estimating the score distribution for PSMs formed by the cyclic peptide Tyrocidine A (single-stage MS). Solid line shows the distribution of scores of 10^9 peptides that are randomly generated. The dots show the MS-DPR p-values. (b) Similar results for the $MultiStage$ score defined in the multistage de novo sequencing paper [18], for 10^7 peptides. Red dashed lines represent the scores of the correct peptide. The figure shows that MS-DPR p-values and empirical p-values are well correlated. Moreover, the p-value of the correct peptide is lower for multi-stage score $(5e-13)$ single-stage score $(5e-07)$, illustrating the advantage of multi-stage mass spectrometry. MS-DPR enables comparisons between arbitrary scoring functions. (c) Similar results for the score distribution for PSMs formed by the branch-cyclic peptide A21978C2 (single-stage MS).

Bibliography

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

[2] M.A. Marahiel, T. Stachelhaus, and H.D. Mootz. Modular peptide synthetases involved in nonribosomal peptide synthesis. *Nat Prod Rep.*, 7:2651–2674, 1997.

[3] D. Schwarzer and M.A. Finking, R.and Marahiel. Nonribosomal peptides: from genes to products. *Nat Prod Rep.*, 20:275–287, 2003.

[4] H. Y. Ridley, C. Lee and C. Khosla. Evolution of polyketide synthases in bacteria. *Proceedings of the National Academy of Sciences*, 105:4595–4600, 2008.

[5] Oman T.J. and van der Donk W.A. Follow the leader: the use of leader peptides to guide natural product biosynthesis. *Nat Prod Rep.*, 6:9–18, 2010.

[6] Donia M. S. McIntosh J. A. and E.W. Schmidt. Ribosomal peptide natural products: Bridging the ribosomal and nonribosomal worlds. *Nat Prod Rep.*, 26:537–559, 2009.

[7] G. Challis and J. Ravel. Coelichelin, a new peptide siderophore encoded by the streptomyces coelicolor genome: structure prediction from the sequence of its non ribosomal peptide synthetase. *FEMS microbiology letter*, 187(2):111–114, 2000.

[8] S. Lautru, R. Deeth, L. Bailey, and G. Challis. Discovery of a new peptide natural product by streptomyces coelicolor genome mining. *Nat. Chem. Biol.*, 1(5):265–269, 2005.

[9] A.W. Hill and R.J. Mortishire-Smith. Automated assignment of high-resolution collisionally activated dissociation mass spectra using a systematic bond disconnection approach. *Rapid Commun. Mass Spectrom.*, 19:3111–3118, 2005.

[10] J.M. Halket, D. Waterman, A.M. Przyborowska, R.K.P. Patel, P.D. Fraser, and P.M. Bramley. Chemical derivatization and mass spectral libraries in metabolic profiling by gc/ms and lc/ms/ms. *Journal of Experimental Botany*, 56(410):219–243, 2005.

[11] R. Craig, J.C. Cortens, D. Fenyo, and R.C. Beavis. Using annotated peptide mass spectrum libraries for protein identification. *J Proteome Res*, 5:1843–1849, 2006.

[12] H. Lam, E.W. Deutsch, J.S. Eddes, J.K. Eng, N. King, S.E. Stein, and R. Aebersold. Development and validation of a spectral library searching method for peptide identification from ms/ms. *Proteomics*, 7:655667, 2007.

[13] H. Lam and R. Aebersold. Spectral library searching for peptide identification via tandem ms. *Methods Mol Biol*, 604:95–103, 2010.

[14] H. Lam, E.W. Deutsch, and R. Aebersold. Artificial decoy spectral libraries for false discovery rate estimation in spectral library searching in proteomics. *J Proteome Res*, 9:605–610, 2010.

[15] N. Bandeira, D. Tsur, A. Frank, and P.A. Pevzner. Protein identification by spectral networks analysis. *Proc. Natl. Acad. Sci.*, 104(15):6140–5, 2007.

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

[17] H. Mohimani, W.T. Liu, Y. Liang, S. Gaudenico, W. Fenical, P.C. Dorrestein, and P. Pevzner. Multiplex *de novo* sequencing of peptide antibiotics. *J. Comput. Biol.*, 18(11):1371–1381, 2011.

[18] H. Mohimani, Y. Liang, W.T. Liu, P.W. Hsieh, P.C. Dorrestein, and P. Pevzner. Sequencing cyclic peptides by multistage mass spectrometry. *J. Proteomics*, 11(8):3642–3650, 2011.

[19] S. Wolf, S. Schmidt, M. Mller-Hannemann, and S. Neumann. In silico fragmentation for computer assisted identification of metabolite mass spectra. *BMC Bioinformatics*, 11:148, 2010.

[20] R. Gugisch, A. Kerber, A. Kohnert, R. Laue, M. Meringer, C. Rcker, and A. Wassermann. Molgen 5.0, a molecular structure generator. *Submitted to Bentham Science Publishers Ltd. 2012*, 2013.

[21] A. Ibrahim, L. Yang, C. Johnston, X. Liu, B. Ma, and N.A. Magarveya. Dereplicating nonribosomal peptides using an informatic search algorithm for natural products (isnap) discovery. *Proc. Natl. Acad. Sci.*, 109(47):19196–19201, 2012.

[22] H. Mohimani, W.T. Liu, R. Kersten, P.C. Dorrestein, and P.A. Pevzner. Pepgen-miner: Coupling mass-spectrometry and genome mining for non ribosomal peptide discovery. *in preparation*, 2013.

[23] H. Mohimani and P.A. Pevzner. Dereplication of ms/ms spectra from large chemical databases. *in preparation*, 2013.

[24] X.J. Tang, P. Thibault, and R.K. Boyd. Characterization of the tyrocidine and gramicidin fractions of the tyrothricin complex from bacillus brevis using liquid chromatography and mass spectrometry. *Int. J. Mass Spectrom. Ion Processes*, 122:153–179, 1992.

[25] S.A. Sieber and M.A. Marahiel. Molecular mechanisms underlying nonribosomal peptide synthesis: Approaches to new antibiotics. *Chem. Rev.*, 105:715–738, 2005.

[26] T.F. Molinski, D.S. Dalisay, S.L. Lievens, and J.P. Saludes. Drug development from marine natural products. *Nat. Rev. Drug Discovery*, 8(1):69–85, 2009.

[27] T.F. Molinski. Nmr of natural products at the nanomole-scale. *Nat. Prod. Rep.*, 27(3):321–329, 2010.

[28] J.W. Li and J.C. Vederas. Drug discovery and natural products: end of an era or an endless frontier? *Science*, 325(5937):161–165, 2009.

[29] P.N. Leao, A.R. Pereirab, W.T. Liu, J. Ng, P.A. Pevzner, P.C. Dorrestein, G.M. Konig, M. Teresa, S.D. Vasconcelos, V.M. Vasconcelos, and W.H. Gerwick. Synergistic allelochemicals from a freshwater cyanobacterium. *Int. J. Mass Spectrom. Ion Processes*, 107(25):11183–8, 2010.

[30] W.T. Liu, Y.L. Yang, Y. Xu, A. Lamsa, N.M. Haste, J.Y. Yang, J. Ng, D. Gonzalez, C.D. Ellermeier, P.D. Straight, P.A. Pevzner, J. Pogliano, V. Nizet, K. Pogliano, and P.C. Dorrestein. Imaging mass spectrometry of intraspecies metabolic exchange revealed the cannibalistic factors of *bacillus subtilis*. *Proc. Natl. Acad. Sci.*, 107(37):16286–16290, 2010.

[31] W.H. Fenical, R.S. Jacobs, and P.R. Jensen. Cyclic heptapeptide anti-inflammatory agent. *US Patent 5593960*, 1995.

[32] H. Sugiyama, T. Shioiri, and Yokokawa F. Synthesis of four unusual amino acids, constituents of cyclomarin a. *Tetrahedron Letters*, 143:3489–92, 2002.

[33] A.W. Schultz, D.C. Oh, J.R. Carney, R.T. Williamson, D.W. Udwary, P.R. Jensen, S.J. Gould, W. Fenical, and B.S. Moore. Biosynthesis and structures of cyclomarins and cyclomarazines, prenylated cyclic peptides of marine actinobacterial origin. *Tetrahedron Letters*, 13:4507–16, 2008.

[34] S.S. Skiena, W.D. Smith, and P. Lemke. Reconstructing sets from interpoint distances. *The sixth annual symposium on Computational Geometry, Berkeley, California*, pages 332–339, 1990.

[35] S.S. Skiena and G. Sundaram. A partial digest approach to restriction site mapping. *Bulletin of Mathematical Biology*, 56(2):272–294, 1994.

[36] M. Cieliebak, S. Eidenbenz, and P. Penna. Partial digest is hard to solve for erroneous input data. *Theoretical Computer Science*, 349(3):361–381, 2005.

[37] T.D. Schneider and R.M. Stephens. Sequence logos: A new way to display consensus sequences. *Nucleic Acids Res.*, 18(20):6097–6100, 1990.

[38] W.T. Liu, J. Ng, D. Meluzzi, N. Banderia, M. Gutirrez, T.L. Simmons, A.W. Schultz, R.G. Linington, B.S. Moore, W.H. Gerwick, P.A. Pevzner, and Dorrestein P.C. Interpretation of tandem mass spectra obtained from cyclic nonribosomal peptides. *Anal. Chem.*, 81(11):4200–4209, 1990.

[39] S. Kim, N. Gupta, N. Bandeira, and P.A. Pevzner. Spectral dictionaries: Integrating de novo peptide sequencing with database search of tandem mass spectra. *Mol. Cell. Proteomics*, 8(1):53–69, 2009.

[40] A.M. Frank. A ranking-based scoring function for peptide-spectrum matches. *J. Proteome. Res.*, 8(5):2241–2252, 2009.

[41] B. Ma, K. Zhang, C. Hendrie, C. Liang, M. Li, A. Doherty-Kirby, and G. Lajoie. Peaks: powerful software for peptide de novo sequencing by tandem mass spectrometry. *in preparation*, 2013.

[42] A. Frank and P. Pevzner. Pepnovo: De novo peptide sequencing via probabilistic network modeling. *Anal. Chem.*, 77:964–983, 2005.

[43] J.K. Eng, A.L. McCormack, and J.R. Yates. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.*, 5(11):976–989, 1994.

[44] D.N. Perkins, D.J. Pappin, D.M. Creasy, and J.S. Cottrell. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 20(18):3551–3567, 1999.

[45] Y.Q. Tang, J. Yuan, G. Oesapay, K. Oesapay, D. Tran, C.J. Miller, A.J. Ouellette, and M.E. Selsted. A cyclic antimicrobial peptide produced in primate leukocytes by the ligation of two truncated alpha-defensins. *Science*, 286(5439):498–502, 1999.

[46] S. Caboche, M. Pupin, V. Leclre, A. Fontaine, P. Jacques, and G. Kucherov. Norine: a database of nonribosomal peptides. *Nucleic Acids Res.*, 36:D326–D331, 2008.

[47] Z. Zhang and J.S. McElvain. De novo peptide sequencing by two-dimensional fragment correlation mass spectrometry. *Anal. Chem.*, 72:2337–2350, 2008.

[48] N. Bandeira, J. Olsen, M. Mann, and P. Pevzner. Multi-spectra peptide sequencing and its applications to multistage mass spectrometry. *Bioinformatics*, 24:416–423, 2008.

[49] T. Lin and G.L. Glish. C-terminal peptide sequencing via multistage mass spectrometry. *Anal. Chem.*, 70:5162–5, 1998.

[50] D.F. Hunt, J.R. 3rd Yates, J. Shabanowitz, S. Winston, and C.R. Hauer. Protein sequencing by tandem mass spectrometry. *Proc. Nat. Acad. Sci.*, 83:6233–7, 1986.

[51] Jedrzejeski P. Lehmann W. D. Schnoelzer, M. Protease-catalyzed incorporation of 18o into peptide fragments and its application for protein sequencing by electrospray and matrix-assisted laser desorption/ionization mass spectrometry. *Electrophoresis*, 17:945–953, 1996.

[52] Rucknagel P. Kuellertz G. Schierhorn A. Pfeifer, T. A strategy for rapid and efficient sequencing of lys-c peptides by matrix-assisted laser desorption/ionisation time-of-flight mass spectrometry post-source decay. *Rapid Communications in Mass Spectrometry*, 13:362–369, 1999.

[53] C. Garcia-Mendoza. Studies on the mode of action of etamycin (viridogrisein). *Biochim. Biophys. Acta.*, 97:394–396, 1965.

[54] Perera V.R. Maloney K.N. Tran D.N. Jensen P. Fenical W. Nizet V. Haste, N.M. and M.E. Hensler. Activity of the streptogramin antibiotic etamycin against methicillin-resistant staphylococcus aureus. *J. Antibiot.*, 63:219–224, 2010.

[55] Tan N.H. Zhou J. Wang, Y.C. and H.M. Wu. Cyclopeptides from dianthus superbus. *Phytochemistrye*, 49:1453–1456, 2004.

[56] P.W. Hsieh, F.R. Chang, C.C. Wu, K.Y. Wu, C.M. Li, and Y.C. Wu. New cytotoxic cyclic peptides and dianthramide from dianthus superbus. *J. Nat. Prod.*, 67:1522–1527, 2004.

[57] Psb-120: Data dependent analysis for ion trap mass spectrometers. *Product support bulletin of Thermo Scientific linear ion trap mass spectrometers.*

[58] K.P. Bateman, K. Yang, P. Thibault, R.L. White, and L.C. Vining. Inactivation of etamycin by a novel elimination mechanism in streptomyces lividans. *J. Am. Chem. Soc.*, 118:5335–5338, 1996.

[59] J.E. Velasquez and W.A. van der Donk. Genome mining for ribosomally synthesized natural products. *Curr. Opin. Cell Biol.*, 15(1):11–21, 2011.

[60] K. Babasaki, T. Takao, Y. Shimonishi, and K. Kurahashi. Subtilosin a, a new antibiotic peptide produced by bacillus subtilis 168: isolation, structural analysis, and biogenesis. *J. Biochem.*, 98(3):585–603, 1985.

[61] K. Kawulka, T. Sprules, R.T. McKay, P. Mercier, C.M. Diaper, P. Zuber, and J.C. Vederas. Structure of subtilosin A, an antimicrobial peptide from *bacillus subtilis* with unusual posttranslational modifications linking cysteine sulfurs to alpha-carbons of phenylalanine and threonine. *J. Am. Chem. Soc.*, 125(16):4726–4727, 2003.

[62] R.A. Salomon and R.N. Farias. Microcin 25, a novel antimicrobial peptide produced by *escherichia coli*. *J. Bacteriol.*, 174(22):7428–7435, 1992.

[63] K. Fujikawa, Y. Suketa, K. Hayashi, and T. Suzuki. Chemical structure of circulin a. *Cell. Mol. Life Sci.*, 21(6):307–308, 1965.

[64] K. Hayashi, Y. Suketa, and T. Suzuki. Chemical structure of circulin b. *Cell. Mol. Life Sci.*, 24(7):656–657, 1968.

[65] R.E. Wirawan, K.M. Swanson, T. Kleffmann, R.W. Jack, and J.R. Tagg. Uberolysin: a novel cyclic bacteriocin produced by *streptococcus uberis*. *Microbiology*, 153:1619–1630, 2007.

[66] L.A. Martin-Visscher, M.J. van Belkum, S. Garneau-Tsodikova, R.M. Whittal, J. Zheng, L.M. McMullen, and J.C. Vederas. Isolation and characterization of carnocyclin a, a novel circular bacteriocin produced by *carnobacterium maltaromaticum* UAL307. *Appl. Environ. Microbiol.*, 74(15):4756–4763, 2008.

[67] T. Wieland. Poisonous principles of mushrooms of the genus *amanita*. Four-carbon amines acting on the central nervous system and cell-destroying cyclic peptides are produced. *Science*, 159(818):946–952, 1968.

[68] H. Faulstich, A. Buku, H. Bodenmueller, and T. Wieland. Virotoxins: actin-binding cyclic peptides of *amanita virosa* mushrooms. *Biochemistry*, 19(14):3334–3343, 1980.

[69] S. Luckett, R.S. Garcia, J.J. Barker, A.V. Konarev, P.R. Shewry, A.R. Clarke, and R.L. Brady. High-resolution structure of a potent, cyclic proteinase inhibitor from sunflower seeds. *J. Mol. Biol.*, 290(2):525–533, 1999.

[70] J.E. Gonzlez-Pastor, E.C. Hobbs, and R. Losick. Cannibalism by sporulating bacteria. *Science*, 301(5632):510–513, 2003.

[71] L. Gran. On the effect of a polypeptide isolated from "kalata-kalata" (oldenlandia affinis dc) on the oestrogen dominated uterus. *Acta Pharmacol. Toxicol.*, 33(5):400–408, 1973.

[72] O. Saether, Craik D.J., I.D. Campbell, K. Sletten, J. Juul, and Norman D.G. Elucidation of the primary and three-dimensional structure of the uterotonic polypeptide kalata b1. *J. Nat. Prod.*, 34(13):4147–4158, 1995.

[73] K.M. Witherup, M.J. Bogusky, P.S. Anderson, H. Ramjit, R.W. Ransom, T. Wood, and Sardana M. Cyclopsychotride a, a biologically active, 31-residue cyclic peptide isolated from psychotria longipes. *J. Nat. Prod.*, 57(12):1619–1625, 1994.

[74] K.R. Gustafson, R.C. II Sowder, L.E. Henderson, I.C. Parson, Y. Kashman, J.H. Jr. Cardellina, J.B. McMahon, R.W. Jr. Buckheit, L.K. Pannell, and M.R. Boyd. Circulins a and b: novel hiv-inhibitor macrocyclic peptide from tropical tree chassalia parvifolia. *J. Am. Chem. Soc.*, 116(20):9337–9338, 1994.

[75] J.P. Mulvenna, C. Wang, and D.J. Craik. Cybase: a database of cyclic protein sequence and structure. *Nucleic Acids Res.*, 36:192–194, 2006.

[76] M. Mann and Wilm M. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal. Chem.*, 66(24):4390–4399, 1994.

[77] M.L. Colgrave, A.G. Poth, Q. Kaas, and D.J. Craik. A new era for cyclotide sequencing. *Biopolymers*, 94(5):592–601, 2010.

[78] J. Ng and P.A. Pevzner. Cannibalism by sporulating bacteria. *J. Proteome. Res.*, 7(01):89–95, 2007.

[79] S. Tanner, H. Shu, A. Frank, L.C. Wang, E. Zandi, M. Mumby, P.A. Pevzner, and V. Bafna. InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. *Anal. Chem.*, 77(14):4626–4639, 2005.

[80] J.E. Swedberg, L.V. Nigon, J.C. Reid, S.J. de Veer, C.M. Walpole, C.R. Stephens, T.P. Walsh, T.K. Takayama, J.D. Hooper, J.A. Clements, A.M. Buckle, and J.M. Harris. Substrate-guided design of a potent and selective kallikrein-related peptidase inhibitor for kallikrein. *Chem. Biol.*, 16(6):633–646, 2009.

[81] S. Jiang, P. Li, S.L. Lee, C.Y. Lin, Y.Q. Long, M.D. Johnson, R.B. Dickson, and P.P. Roller. Design and synthesis of redox stable analogues of sunflower trypsin inhibitors (SFTI-1) on solid support, potent inhibitors of matriptase. *Org. Lett.*, 9(1):9–12, 2007.

[82] Y. Long, S.L. Lee, C. Lin, I.J. Enyedy, S. Wang, P. Li, R.B. Dickson, and P.P. Roller. Synthesis and evaluation of the sunflower derived trypsin inhibitor as a potent inhibitor of the type ii transmembrane serine protease, matriptase. *Bioorg. Med. Chem.*, 11(18):2515–2519, 2001.

[83] J.S. Mylne, M.L. Colgrave, N.L. Daly, A.H. Chanson, A.G. Elliott, E.J. McCallum, A. Jones, and D.J. Craik. Substrate-guided design of a potent and selective kallikrein-related peptidase inhibitor for kallikrein. *Nat. Chem. Biol.*, doi : 10.1038/nchembio.542, 2011.

[84] N. Venkataraman, A.L. Cole, P. Ruchala, A.J. Waring, R.I. Lehrer, and et al. Reawakening retrocyclins: ancestral human defensins active against HIV-1. *PLoS Biol.*, 7:e95, 2009.

[85] B.J. Williams, W.K. Russell, and D.H. Russell. High-throughput method for on-target performic acid oxidation of maldi-deposited samples. *J. Mass. Spectrom.*, 45(2):157–66, 2010.

[86] N. Gupta, S. Tanner, N. Jaitly, J. Adkins, M. Lipton, R. Edwards, M. Romine, A. Osterman, V. Bafna, R. Smith, and P. Pevzner. Whole proteome analysis of post-translational modifications: applications of mass-spectrometry for proteogenomic annotation. *PLoS Biol.*, 17(9):1362–1377, 2007.

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

[88] D. Fenyo and R. Beavis. A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Anal. Chem.*, 75:768–774, 2003.

[89] R.G. Sadygov, H. Liu, and J. R. Yates. Statistical models for protein validation using tandem mass spectral data and protein amino acid sequence databases. *Anal. Chem.*, 76(6):1664–1671, 2004.

[90] R. Matthiesen, M. B. Trelle, Hjrup P., J. Bunkenborg, and O. N. Jensen. VEMS 3.0: Algorithms and computational tools for tandem mass spectrometry based identification of post-translational modifications in proteins. *J. Prot. Res.*, 4(6):2338–2347, 2005.

[91] D.C. Chamrad, G. Koerting, J. Gobom, H. Thiele, J. Klose, H.E. Meyer, and M. Blueggel. Interpretation of mass spectrometry data for high-throughput proteomics. *Anal. Bioanal. Chem.*, 376(7):1014–1022, 2007.

[92] A. Nesvizhskii, O. Vitek, and R. Aebersold. Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat. Methods*, 4:787–1797, 2007.

[93] A. Nesvizhskii and R. Aebersold. Analysis, statistical validation and dissemination of large-scale proteomics datasets generated by tandem ms. *Drug Discov. Today*, 9(4):173–181, 2004.

[94] Spirin V., Shpunt A., Seebacher J., Gentzel M., Shevchenko A., Gygi S., and Sunyaev S. Assigning spectrum-specific p-values to protein identifications by mass spectrometry. *Bioinformatics.*, 27(8):1128–1134, 2011.

[95] B. Weatherly, J.A. Atwood, T.A. Minning, C. Cavola, R.L. Tarleton, and R. Orlando. A heuristic method for assigning a false-discovery rate for protein identifications from mascot database search results. *Mol. Cell. Proteomics*, 4:762–772, 2005.

[96] S. Kim, N. Mischerikow, N. Bandeira, J.D. Navarro, L. Wich, S. Mohammed, A.J.R. Heck, and P.A. Pevzner. The generating function of cid, etd and cid/etd pairs of tandem mass spectra: Applications to database search. *Mol. Cell. Proteomics*, 9:2840–2852, 2010.

[97] H. Kahn and T.E. Harris. Estimation of particle transmission by random sampling. *National Bureau of Standards Applied Mathematics*, 12:27–30, 1951.

[98] Villen-Altamirano M. and Villen-Altamirano J. RESTART: A method for accelerating rare events simulations. queueing performance and control in atm. *Proceedings of ITC*, 13:71–76, 1991.

[99] S. Kim, N. Gupta, and P. Pevzner. Spectral probabilities and generating functions of tandem mass spectra: A strike against decoy databases. *J. Proteome Res.*, 7(8):3354–3363, 2008.

[100] J.M. Hammersley and D.C. Handscomb. Monte carlo methods. *London*, 1964.

[101] G. Rubino and B. Tuffin. Rare event simulation using monte carlo methods. *Wiley*, 2009.

[102] H. Kahn and A. W. Marshall. Rare event simulation using monte carlo methods. *Oper. Res. Soc. Amer.*, 1953.

[103] H. Kahn. Use of different monte carlo sampling techniques. *RAND corporation*, 1956.

[104] P. Glasserman, P. Heidelberger, and P. Shahabuddin. Asymptotically optimal importance sampling and stratification for pricing path dependent options. *Mathematical Finance*, 9(2):117–152, 1999.

[105] H.A.P. Blom, J. Krystul, G.J. Bakker, M.B. Klompstra, and B.K. Obbink. Free flight collision risk estimation by sequential mc simulation. *Stochastic Hybrid Systems CRC Press*, 2007.

[106] W. Sandmann. Applicability of importance sampling to coupled molecular reactions. *In Proceedings of the 12th International Conference on Applied Stochastic Models and Data Analysis*, 2007.

[107] Elias J.E. and Gygi S.P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods*, 4(3):207–214, 2007.

[108] N. Gupta, N. Bandeira, U. Keich, and Pevzner P. Target-decoy approach and false discovery rate: when things may go wrong. *J. Am. Soc. Mass Spectrom.*, 22:1111–1120, 2011.

[109] Nesvizhskii. A. Survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *J. Prot. Res.*, 73(11):2092–2123, 2010.

[110] Kwon T., Choi H., Vogel C., Nesvizhskii A. I., Marcotte, and E. M. MS-blender : A probabilistic approach for integrating peptide identifications from multiple database search engines. *J. Prot. Res.*, 10(7):2949–2958, 2011.

[111] L.Y. Geer, S.P. Markey, J.A. Kowalak, L. Wagner, M. Xu, D.M. Maynard, X. Yang, W. Shi, and S.H. Bryant. Open mass spectrometry search algorithm. *J. Prot. Res.*, 3(5):958–964, 2004.

[112] M. Waterman and M. Vingron. Rapid and accurate estimates of statistical significance for sequence data base searches. *Proc. Natl. Acad. Sci.*, 91(7):4625–4628, 1994.

[113] Asmussen S. and Glynn P.W. Stochastic simulation: algorithms and analysis. *Springer*, 2007.

[114] Haraszti Z. and Townsend J. K. The theory of direct probability redistribution and its application to rare even simulation. *ACM Trans. Modeling and Computer Simulation*, 9(2):105–140, 1999.

[115] Glasserman P., Heidelberger P., and Shahabuddin P. A large deviations perspective on the efficiency of multilevel splitting. *IEEE Trans. Automat. Contr.*, 43(12):1666–1679, 1998.

[116] J. Klimek, J. S. Eddes, L. Hohmann, J. Jackson, A. Peterson, S. Letarte, P. R. Gafken, J.E. Katz, P. Mallick, H. Lee, A. Schmidt, R. Ossola, J.K. Eng, R. Aebersold, and D.B. Martin. The standard protein mix database: a diverse data set to assist in the production of improved peptide and protein identification software tools. *J. Prot. Res.*, 7:96–103, 2008.

[117] A. Keller, A. Nesvizhskii, E. Kolker, and R. Aebersold. Empirical statistical model to estimate the accuracy of peptide identifications made by ms/ms and database search. *Anal. Chem.*, 74:5383–5392, 2002.