Nonribosomal Peptide Analog Identification with Tandem Mass Spectrometry

by

Shiwei Li

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Computer Science

Waterloo, Ontario, Canada, 2014

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Shiwei Li

Abstract

Nonribosomal peptides (NRP) are a class of peptide secondary metabolites, usually produced by microorganisms like bacteria and fungi. NRP antibiotics, cytostatics, and immunosuppressants are in commercial use. In pharmacological studies, novel NRPs are often promising substances for new drug development. To discover novel NRPs with limited resources from microbial fermentations, a significant process is to identify known NRPs and their analogs in an early stage and exclude them from further investigation. This so-called "dereplication" step ensures less resource wasted in the subsequent experiments. Tandem mass spectrometry has been routinely used for NRP dereplication. Other researchers have developed software to identify known NRPs with a database. However, only a rather small part of NRPs are discovered by now and identifying analog of these NRPs is still occupying much resources and hindering the throughput of novel NRP discovery.

In this thesis, we review the nature of nonribosomal peptides and investigate the challenges in computationally solving the analog finding problem. After that, a program called NRP Analog Finder is introduced as an automated method to identify NRPs and their analogs with tandem mass spectrometry. It is designed to identify mixtures of NRP compounds from LC-MS/MS of complex extract; find structural analogs that differ from an identified known NRP compound with at most two monomers; localize the modified residues; and determine how much mass is changed at each modification site. NRP analog finder is tested to be an effective tool for mass spectrometry based NRP analog identification.

Acknowledgements

I would first like to thank my supervisor, Dr. Bin Ma, for directing the NRP analog finder project. I acquired my bachelor degree in Electronic Engineering and thus had a comparatively weak foundation in computer science. It is Dr. Bin Ma who guided me to become a qualified master student in computer science. I feel privileged to have such a great advisor.

I would like to thank Lian Yang for assisting me in the project. This project relies much on his software iSNAP and he helps me get familiar with NRP identification and the principle of the software. Without him, there will not be such project.

I would like to thank my research collaborators in McMaster University, Dr. Nathan Magarvey, for visioning the utilization of informatics in natural product discovery, ensuring the project to be useful in practical lab research from a biochemist's perspective and providing actual data to test this program.

I would like express my gratefulness to my colleagues in the bioinformatics research group, Lin He, Xiaofei Zhao and Xiao-bo Li, for all the help they gave me.

I would like to thank my parents, for providing the best possible education and encouraging me to persist when I feel low. I would not be myself without their support. I would like to thank my girlfriend, Lin Wang, for the love she gave me.

Dedication

The thesis is dedicated to my parents for their unconditional love.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

Nature provides man with an almost infinite number of unique and effective molecules. Among these resourceful molecules produced by nature, a lot have been proved to have diverse bioactivities, such as antibiotics, immune suppressants, toxins, and etc. [1]. All of these suggest pharmaceutical potentials. According to Newman's review [2], natural product and/or natural product structures play a highly significant role in the drug discovery and development process, thus influence the design of small molecules. Approximately 50% of all new drug approvals in the past 30 years in US either come from natural products and their variants, or are semi-synthetics (synthesized using natural compounds as starting materials) (Figure 0.1). And in all countries, only less than 30% of newly approved drugs are totally synthetic (Figure 0.2).

Among these natural products, nonribosomal peptides (NRP) are a class of peptide secondary metabolites, usually produced by microorganisms like bacteria and fungi. NRPs are also found in higher organisms, but are thought to be made by bacteria inside these organisms. As secondary metabolites, NRPs are not directly necessary in an organism's life, but often play important role in the organism's continuing existence in adverse situations and interspecies defense [3]. Thus, NRPs are a very diverse family of natural products with an extremely broad range of biological activities and pharmacological properties.

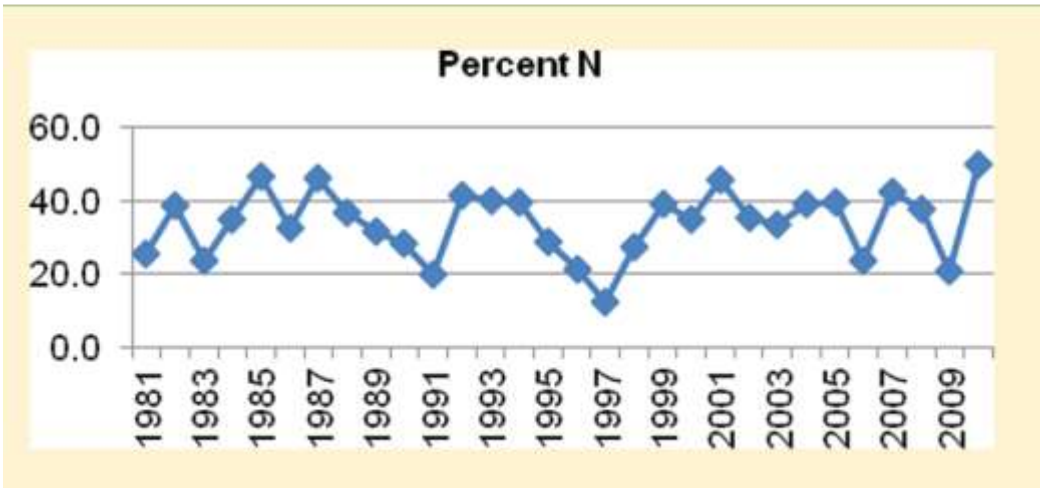


Figure 0.1 Natural products as source of new drugs in US from 1981 to 2010 [2].

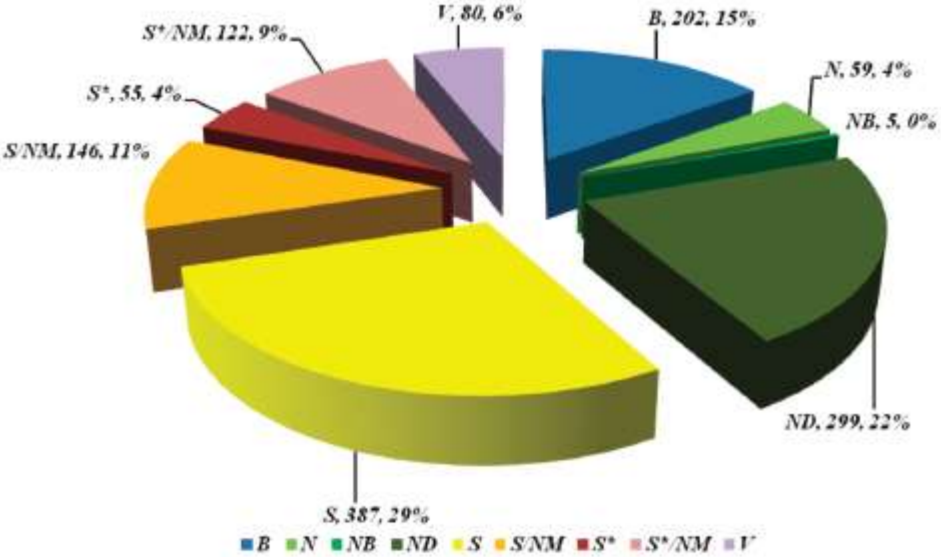


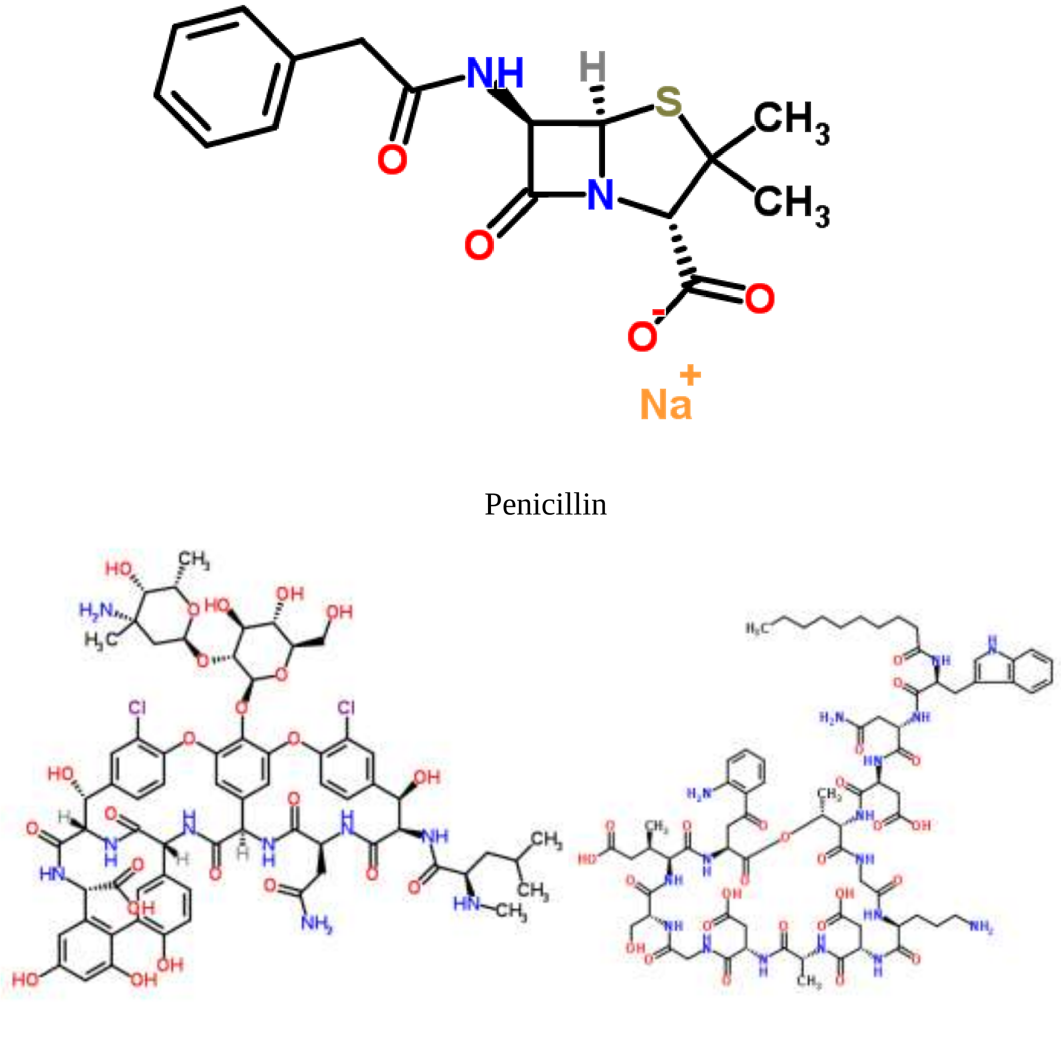
Figure 0.2 All new drugs in all countries from 1981 to 2010 [2]. “B” Biological; “N” Natural product. “NB” Natural product “Botanical”. “ND” Derived from a natural product and is usually a semisynthetic modification. “S” Totally synthetic drug. “S*” Made by total synthesis, but the pharmacophore is/was from a natural product. “V” Vaccine.

As such, nonribosomal peptides always provide an attractive platform to search for therapeutic agents and agrochemicals [4]. Proliferation of drug-resistant bacteria makes the demand for powerful antibiotics substantial. And lots of nonribosomal peptides have been discovered to be used as therapeutic agents. Here are some well-known examples. The famous antibiotic, penicillin (Figure 0.3), discovered in 1928 by Sir Alexander Fleming, was historically significant because they were the first drugs that were effective against many previously serious diseases. Vancomycin [5] (Figure 0.3) is a NRP discovered and isolated in 1953. This group of NRPs was once used as the strongest weapon of human beings against bacteria and pathogenic organisms. With the proliferation of drug-resistant bacteria, more novel drugs have been developed, such as daptomycin [6] (Figure 0.3), which is also originated from NRPs. At present the need for more and more effective antibiotics is indeed urgent and NRPs provide us with plenty of promising candidates due to various particularities and a large structural diversity.

Vast numbers of other kinds of drugs have also been developed from NRPs. For antibiotics drugs, besides penicillin and vancomycin, more NRP compounds have been developed or synthesized to be approved as novel drugs, such as polymyxins [7] and gramicidins [8] (Figure 0.3). NRPs are also widely used as anticancer agents in clinical treatment of various types of cancers, such as epothilone [9], bleomycin [10] and their analogs. In the immunosuppressant category, cyclosporine-A [11] and rapamycin [12] (Figure 0.4) were isolated to be novel drugs, widely used in order to reduce the risk of rejection in organ transplantation. According to a French researcher Caboche [13], around 5% of 205 families of NRPs are currently clinically approved drugs.

With many NRPs discovered and developed to be new drugs, more research is done in order to find new compounds with bioactivity [14]. To identify novel NRPs, researchers grow microbial strains under various conditions, and often stressed with heat and ethanol shock in order to provide the situation, in which nonribosomal peptide synthetase (NRPS) may take place. Fragments of proteins are separated for bioactivity by liquid chromatograph (LC). Then researchers can get structural information of compounds, mainly mass and charge information with tandem mass spectrometry (MS/MS) in bioactive fractions. As long as there could be promising compounds in certain compounds, they are then purified to be processed with nuclear magnetic resonance spectroscopy (NMR) in order to allow structural confirmation.

Unfortunately, most natural NRPs have complicated structure with nonstandard amino acids, thus are notoriously difficult to sequence. Moreover, the dominant

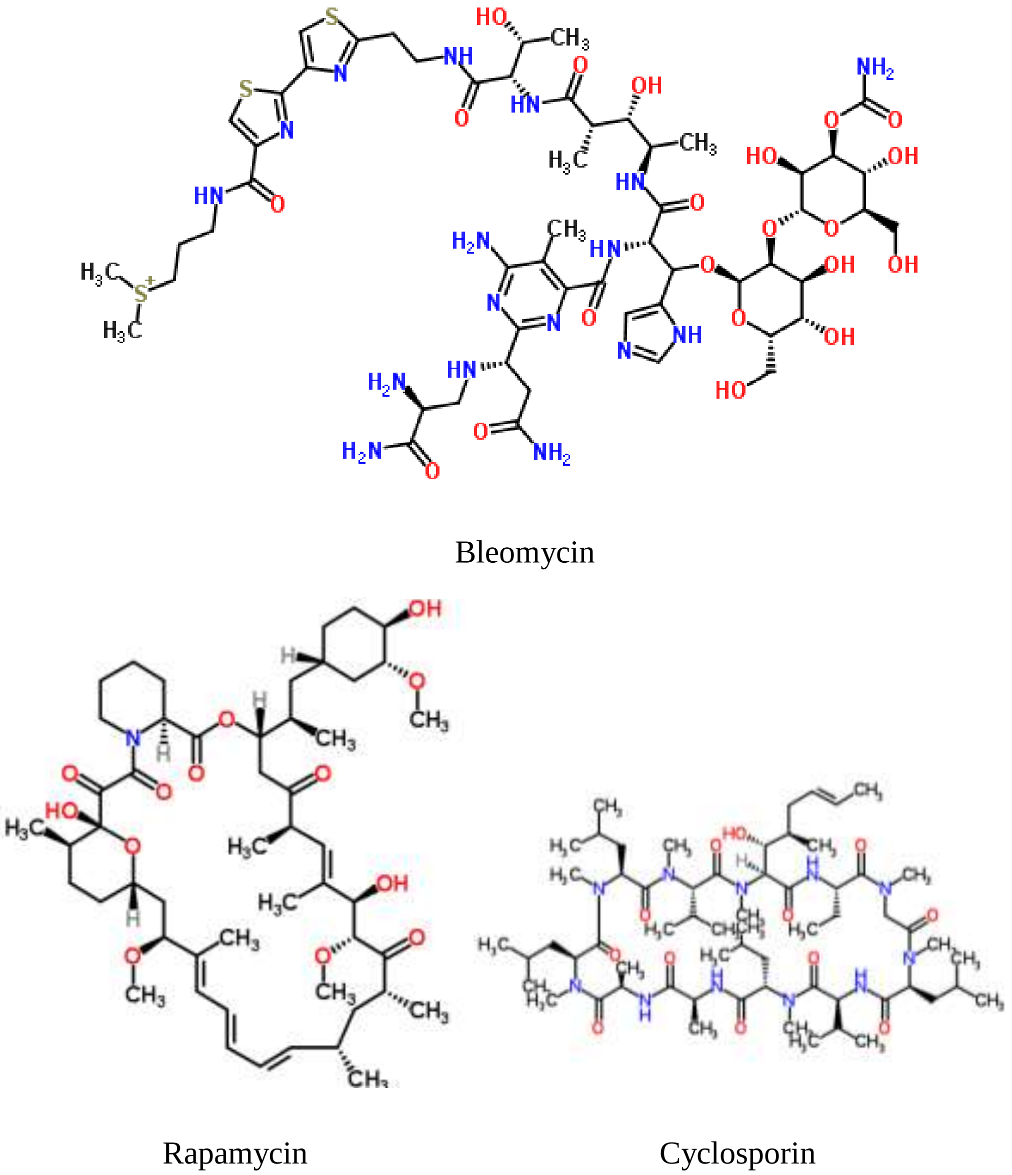


Penicillin

Vancomycin

Daptomycin

Figure 0.3 Structures of NRPs as approved drugs.



Bleomycin

Rapamycin

Cyclosporin

Figure 0.4 Structures of NRPs as approved drugs.

technique, NMR, for sequencing antibiotics requires large amounts of highly purified materials. In addition, NRP discovery increasingly results in reconfirmation of known compounds, which is quite a waste of precious resources. Thus it is necessary to identify known NRPs in an early stage and exclude them from the subsequent NMR process. This process is also referred to as *dereplication* [4].

Structural confirmation with NMR consumes expensive reagents and highly purified material. A good dereplication process can help researchers save time and resources. Besides, additional structural information of experimental spectra is useful for further investigation. We want to further improve the performance of NRP dereplication and focus limited resources on more promising novel NRPs. Thus using computer technology to identify analogs of known NRPs before applying more complicated and expensive technology seems to be a good option.

1.2 **Problem Definition**

The building blocks of an NRP structure are the NRP residues, also referred to as *monomers*. There are several hundred observed types of residues that can be used by the bacteria to build an NRP. These residues are connected by the residue bonds and usually form a branching structure. By treating residues as vertices and the bonds as edges, such a branching structure can be represented as a graph in computer. Figure 0.5 shows the chemical structure of an NRP and its corresponding graph representation. The graph representation of an NRP is usually not too complicated, containing no more than one cycle and a few branches. But more complicated structures are possible.

When such a structure is measured in an MS/MS spectrometer, the structure can be broken into pieces. Each resulting fragment is a connected subgraph of the NRP's structural graph. If the mass (molecular weight) of each residue is known, then the mass of each possible fragment can be computed as the total residue mass in that fragment, plus or minus some commonly known mass offsets (such as -18 Da for the loss of a water molecule). Each possible fragment potentially forms a signal peak in the MS/MS spectrum. Thus, by examining the mass values of the peaks in the spectrum, it is possible to derive the structural information of the measured NRP molecule.

Former researchers (Yang et al [3]) have developed iSNAP software for NRP identification. The software essentially searches in an NRP structure database, and picks the structure of which the predicated spectrum matches the input spectrum the best. Some later extensions of the software also made it possible to consider small

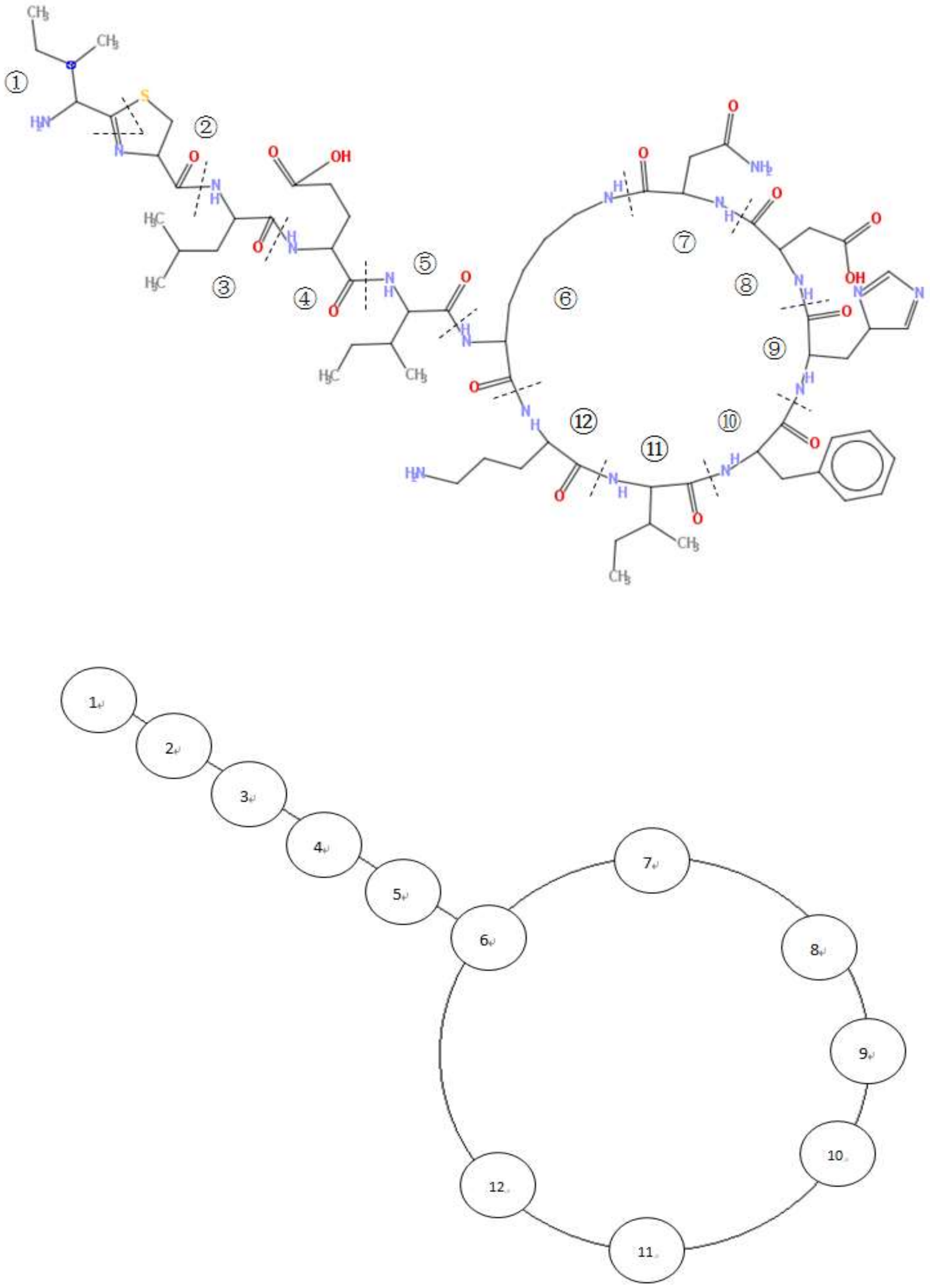


Figure 0.5 Structure of bactracin A and its graph. As is shown in the figure, the graph represents the structure of bactracin A, 12 vertices represent 12 residues and edges the bonds between different residues. Every vertex i has a mass value m_i. Particularly for bacitracin A, they are 98.166, 101.12, 113.158, 129.114, 113.158, 127.164, 114.146, 113.158, 119.174,137.139, 115.087, 114.103.

variations of the structures in the database. For example, by using a given structure as the “seed”, iSNAP can identify a structure that is at most one residue difference from the seed, and matches a given spectrum the best. Throughout this thesis, the given structure is called the *seed structure* or *seed NRP*, and the other structures that are similar to the seed are called *analogs* of the seed. For example in Figure 0.6, each molecule is an analog of any other molecule.

However, it is often insufficient to identify only the analogs with only one residue difference. Since NRP modifications are common, the real NRP analogs often differ from the seed by two or more residues. Thus, it would be useful to identify analogs with two or more residue differences. One immediate difficulty for the analog identification is the imperfect quality of the spectrum. Although in theory each fragment of the structure can produce a peak in the spectrum, it is rarely the case in reality. In a real spectrum, a significant portion of the theoretically predicted peaks are missing from the experimental spectrum. Moreover, many peaks in the experimental spectrum are unexplainable by the fragments of the NRP. These peaks can be due to contaminants, noise, and more complex fragmentation pathways. Thus, the one residue difference achieved by iSNAP seems to be already at the limit of the analog identification with the spectrum of the analog structure.

In the thesis, we propose to include additional information to make the remote analog identification possible. More specifically, in addition to the seed structure and the analog spectrum, we further require the spectrum of the seed structure as the input. This way, by comparing the two spectra of the seed structure and the analog structure, important differences can be found. Then the algorithm will be able to make more confident derivations by focusing on the differences of the two spectra. By doing so, the irrelevant peaks (such as the peaks caused by a common contaminant) can be removed; and the less useful fragments (such as those do not produce peaks in both spectra) are excluded from the analysis. Given two MS/MS spectra of a known NRP structure (the seed) and an unknown analog NRP structure, the main purpose of the thesis is to develop algorithm to compute the unknown analog NRP structure.

1.3 Requirements

Ever since 1970, when Paulien Hogeweg coined the term “Bioinformatics”, computational technology has been used more and more frequently in dealing with biological problem. Using these methods to analyze mass spectrometry data is not a piece of news either. These algorithms help researchers link MS/MS data to certain already known peptides or lead them to discovery of new ones. Former

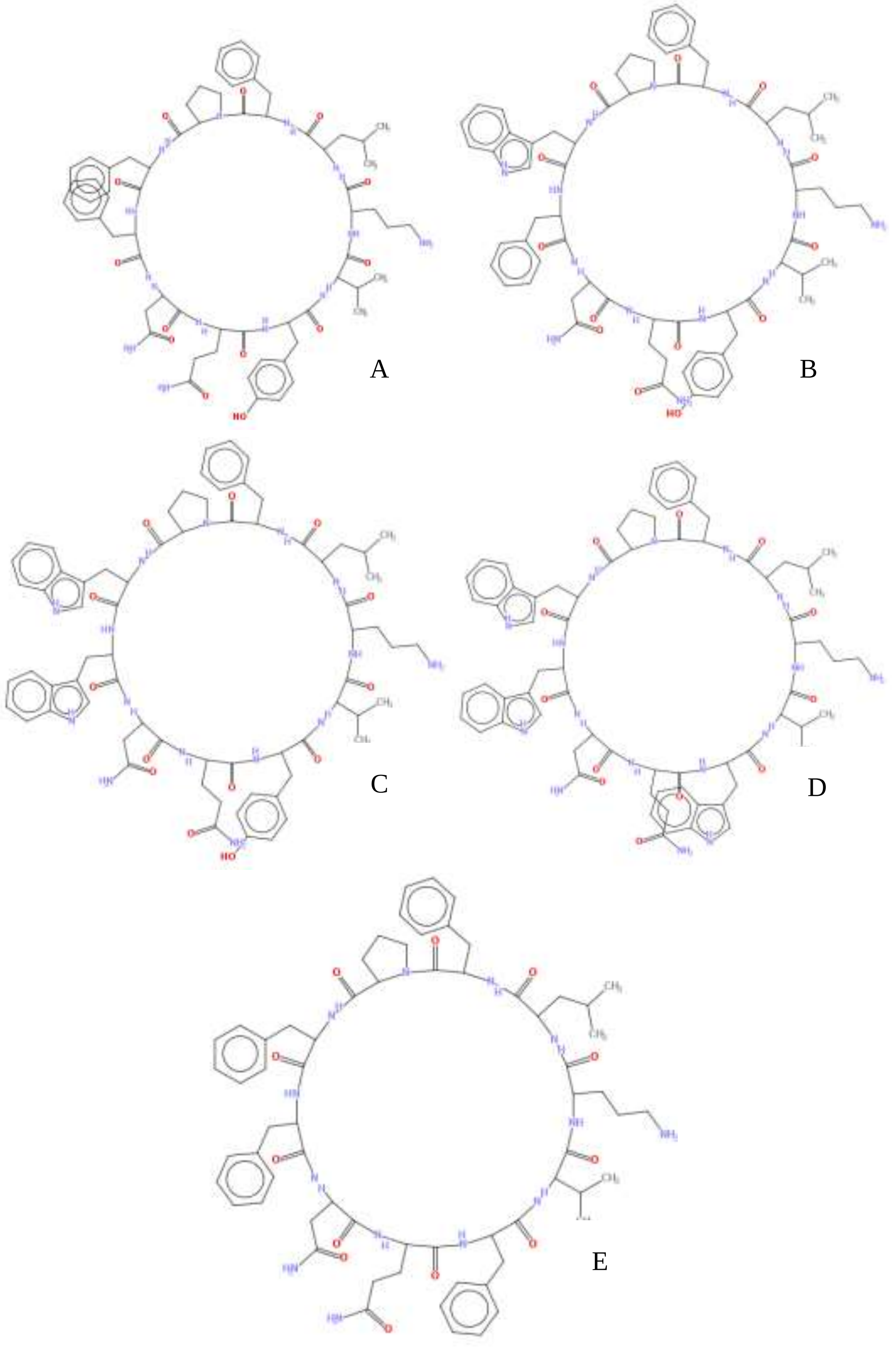


Figure 0.6 Structures of tyrocidine A, B, C, D, E.

work has already shown that the workflow could be adopted to NRPs (Figure 0.7) [5].

On the basis of former work, the primary objective is to develop a NRP analog identification algorithm which can practically identify NRP analog spectra and compute their structure with tandem mass spectral information and structural information of the seed.

The specified requirements which need to be satisfied practically are as follows:

- **Usefulness**
  - The algorithm should be able to identify NRPs and their analogs from either purified sample or complex compounds.
  - Identifications results should be interpretable and represented by certain statistical scores. The scores should help users decide whether the sample contains an already known NRP or an analog or a novel NRP in order to improve the dereplication process.
  - Actual spectra of good quality should be picked to be used for future identification.

- **Correctness**
  - Different analogs of seed molecule should be distinguishably identified. The algorithm should also tell the site of the residue that is modified as well as the mass difference caused by the modification.
  - The output result should have a relatively low false positive and false negative rate.
  - Novel NRP with different structures should not be identified as NRPs already known or their analogs.

However, as it is difficult to get high quality spectra and fairly purified samples, analogs with more than two different sites are nearly impractical to be identified based on the existing spectrum quality. Thus we only aimed to identify analogs with at most two different monomers.

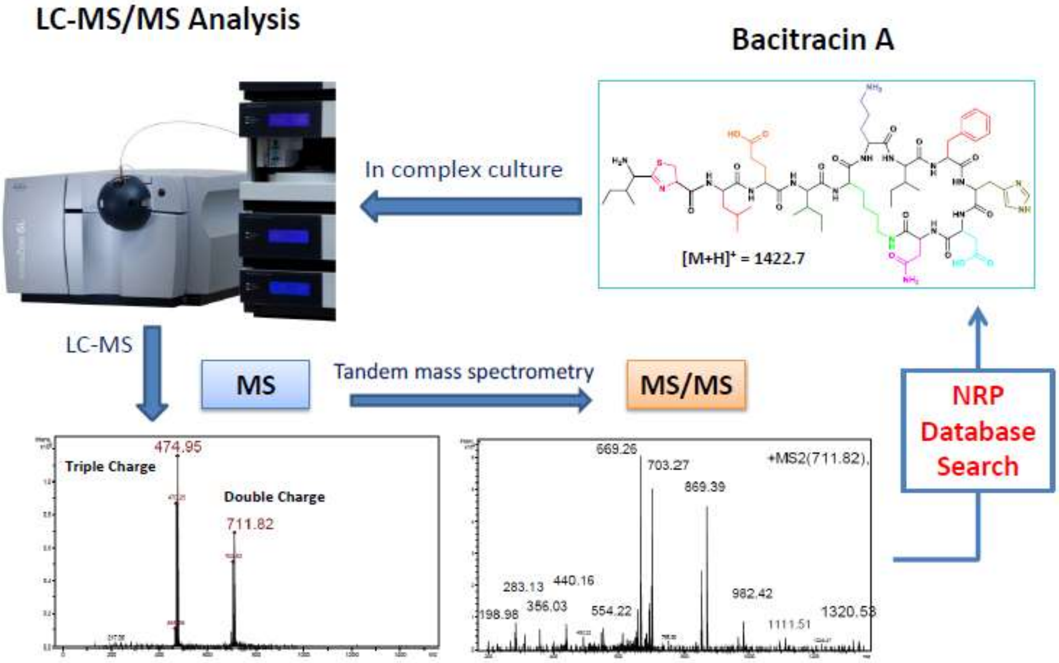


Figure 0.7 Workflow of MS/MS based NRPs dereplication. Microbial fermentation with bioactivity is analyzed using an LC-MS system. With data dependent acquisition or pre-determined acquisition window, ionized compounds in the complex mixture are selected and then fragmented in tandem mass spectrometry. For each selected precursor ion, an MS/MS spectrum is generated with detected fragment ions. In such an experiment, the number of MS/MS spectra can be hundreds. Software is needed to compare those MS/MS spectra with a database of discovered NRPs, so that known NRPs in the fermentation can be identified and excluded from further studies [3].

1.4 Thesis Overview

The remaining of the thesis is organized as follows.

In Chapter 2, we briefly review the characteristics of NRPs. And then the main challenges to be overcome to develop our NRP analog identification algorithm are discussed. They are followed by a brief review of the work of former researchers. In Chapter 3, the algorithm of this thesis is introduced. After describing data structure we used in this thesis, we include how to find the possible mass difference values of the analog and potential modified monomers. After that the scoring scheme is discussed. Chapter 4 provides the experiment results that illustrate the usefulness and the correctness of the algorithm. The next chapter, chapter 5 presents some ideas on the future work. At last, conclusions of this thesis are given in Chapter 6.

Chapter 2

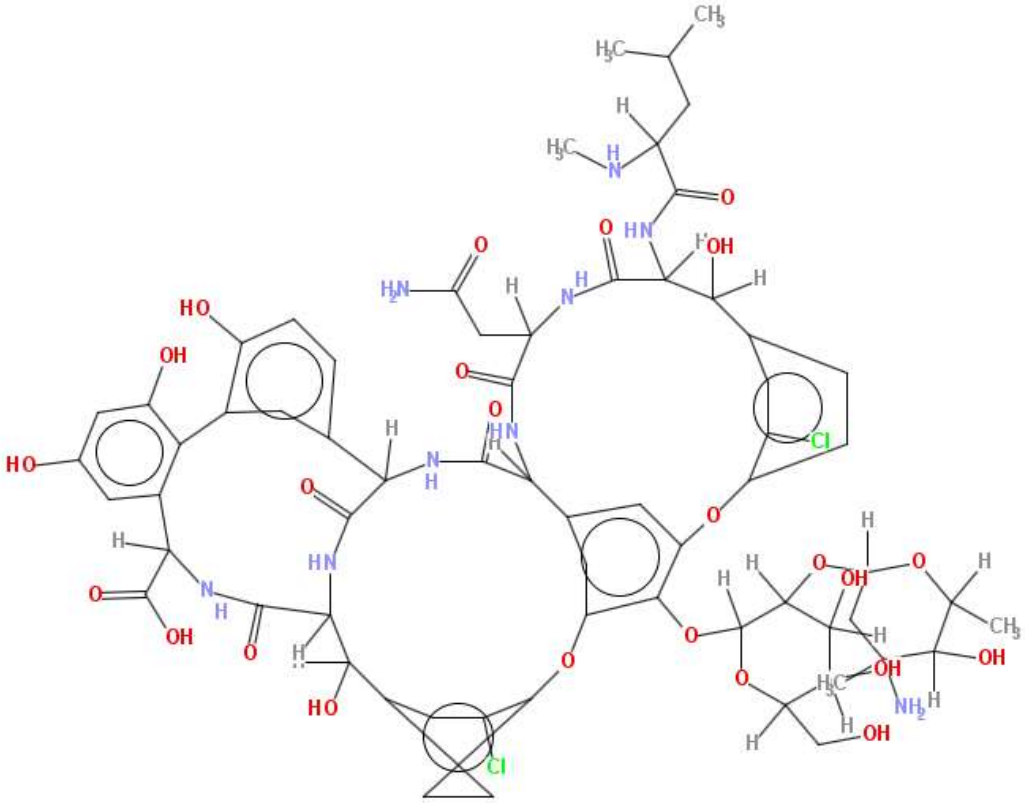
Background

2.1 Nonribosomal Peptides

Nonribosomal peptides are a class of peptide secondary metabolites, which are synthesized by NRPSs. Unlike the ribosomal peptides, they are independent of messenger RNA (mRNA). In the synthesis of ribosomal peptide, a certain gene is transcribed to mRNA which comprises a series of codons. These mRNAs are bound by ribosomes, and dictate to the ribosome the sequence of the amino acids needed to make the protein. Transfer RNA (tRNA) recognizes codons and brings in the corresponding amino acid. The ribosome traverses each codon (3 nucleotides) of the mRNA, pairing it with the appropriate amino acid. 20 different proteinogenic amino acids then form a peptide. Although lariat structures have been observed by former scientists [17], in general, the chain of translated amino acids is then assembled to form a linear peptide.

For NRPs, things are quite different.

The most straightforward difference comes in structure. Unlike the linear ribosomal peptides, NRPs have much more diverse structures, often having a non-linear peptide backbone which is cyclic, branched, or a combination of the two [18]. Vancomycin is an example of molecule with a very complex structure (Figure 0.1). Amino acids in this molecule are connected to each other to form several cyclic structures, which is rare, if will be discovered, in ribosome peptides. Other NRPs, like bacitracin-A, have both cyclic structure and linear branch.



Vancomycin

Figure 0.1 Structure of vancomycin (multiple cyclic structure).

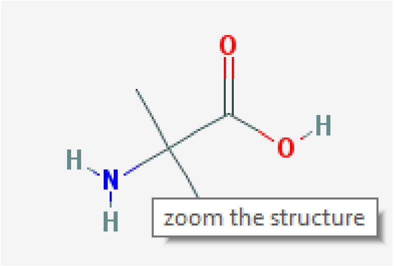


Figure 0.2 Example of non-proteinogenic amino acid in NRPs.

What makes NRPs more complex is their monomer composition. For ribosomal peptides, there are mainly 20 proteinogenic amino acids, except a few exceptions before post-translation modification. However, NRPs may consist of monomers out of those proteinogenic amino acids coded for by DNA. Former researchers (Caboche et al [13]) have proposed a NRPs database called Norine in 2007. Norine contains 1164 peptides by July 2013, while 528 monomers are documented. In addition, NRPs can carry modifications like N-methyl and N-formyl groups, or are glycosylated, acylated, halogenated, or hydroxylated and many other types of modification, which are more commonly observed with NRPs [19] . All of the above give NRP larger diversity.

With the discovery of some NRPs, when there was no such a concept then, researchers began to pay their attention to the synthesis of these NRPs. Comparatively comprehensive knowledge about NRPs synthesis was acquired this century. NRPs are synthesized by one or more specialized NRPSs. NRPSs are multi-enzyme complexes. They vary in size, with their masses ranging from several *KDa* to *MDa*. These NRPSs are organized in modules and each module consists of several domains with different functions. The modules are initiation module, elongation module and termination module. They function during the different stages, which have the same names of those modules. What is also different from the synthesis of ribosomal peptides is that NRPS also determine the types of monomers. Then the NRPs often undergo cyclization and modified such as glycosylation, acylation, or hydroxylatioin [19] . All of these explain the complexity of the structure of NRPs

The biosynthesis of NRPs shares characteristics with polyketide synthetases (PKS). In some databases these two are even put under a common category. Due to these structural and mechanistic similarities, some NRPs may contain polyketide components [20]. Such a situation is mainly spotted in secondary metabolites, which makes the structure of NRPs more complex.

2.2 Challenges for NRP Analog Identification

As discussed above, NRPs are so different in structure from its ribosomal counterpart, thus difficulties in NRP identification are also different.

For traditional peptide identification, in order to interpret hundreds of thousands of MS/MS spectra, researchers have begun to use computational technology in peptide identification for nearly two decades. Ever since 1994, several pieces of software have been developed and put into use commercially. In general, peptides identification algorithms fall into two classes, database search and *de novo* sequencing search. Database search can be further classified into sequence database search and spectral library search. And these algorithms help researchers correlate tandem spectra and peptides in order to identify linear peptides from mass spectra. Some famous examples are SEQUEST developed by Yates et al [21], PEAKS DB developed by Ma et al [22].

Database search, particularly sequence database search, is more popular and considered to produce higher quality results in most cases at present. This kind of algorithms identifies peptides by comparing the tandem mass spectra against a database containing all amino acid sequences assumed to be present. The spectra are generated by the algorithm on the basis of the assumed sequence and certain fragmentation rules. And then the software gives a score that shows how well this experimental spectrum is matched by a hypothetical spectrum. As long as the score is better than a decent threshold, the experimental spectrum is considered to be generated by certain sequence in the database. SEQUEST, Mascot [23], PEAKS DB and X!Tandem [24] are all well-known examples which apply sequence database search algorithms for peptide identification.

Spectral library search has been used in mass spectra identification as early as 1980s [25]. But until recent years, it began to show practical use with the development of millions of MS/MS spectra. Spectral library search is, to some extent, similar to sequence database search. What is different from the former algorithm is that, instead of using hypothetical spectra generated from peptide sequence in the database, it matches the experimental spectra with a library of actual spectra. Generally speaking, spectral library search can process more spectra than sequence database search in given time because there is no need to fragment the peptide computationally and generate the hypothetical spectra. However, the shortcoming of spectral library search for the time being is also obvious. The sizes and availability of spectral libraries is not satisfactory. Moreover, it cannot be used to identify a novel peptide. However, with more and higher quality spectra being acquired by researchers, this approach is showing a more and more promising future.

Although database search yields decent results in most cases and can identify large number of peptides from large quantities of data. This method requires a well-built database beforehand. Normally, novel peptides can hardly find a match in the database, even those with unexpected modifications or mutations can fail the algorithm. For NRPs, this kind of situation may appear more often than ribosomal peptides because of their complex structure and modifications.

De novo sequencing for mass spectrometry is a kind of algorithm which directly analyzes peak information in the spectra and typically is performed without prior knowledge of the sequence. It yields a peptide with highest matching score composed with the 20 proteinogenic amino acids. *De novo* sequencing does not have to match the experimental spectra to any spectra or peptide in any database, thus it can be used to identify modified peptides with mutations as well as to discover totally novel peptides. Compared to database search, *de novo* sequencing is not as popular, but can be used to confirm and expand upon results from database searches. The most commonly used *de novo* sequencing software is PEAKS [26].

Traditional algorithms have been developed for some time and are becoming more and more mature. However, it is not practical to apply these algorithms directly to identify NRPs owing to the complex structures of NRPs. However, NRP identification also shares some similarities with ribosomal peptide identification. It makes us believe that we could identify NRP analogs with certain algorithms similar to traditional ones. However, some particular problems need to be discussed before a proper identification algorithm being given.

Traditionally, each amino acid is stored as a single letter code. Every time when we need to fragment the molecule, we just use substrings to represent the fragments. However, the number of types of monomers which appeared in NRPs is more than five hundred and more may be discovered in future. Also, these monomers may form a cyclic or branching structure instead of a linear structure. This makes it impossible to represent a peptide with a string of twenty letters as what we do with ribosomal peptides. Another method to represent the peptide and its fragments is needed.

Besides, NRPs can be composed of more than five hundred types of monomers, rather than the 20 amino acids for its ribosomal counterpart. Taking the enormous number of monomers and non-linear structure, to reconstruct a peptide sequence with the monomers in the database is a serious problem. Applying algorithms like *de novo* sequencing to identify NRPs can have a considerably large search space, thus implies the requirement for both high quality spectrum and the search time.

What's more, for traditional linear ribosomal peptides, the peptides are mostly dissociated at one amide bond at a time and yield two fragment ions. For cyclic structure, a common molecular structure in NRPs, this kind of fragmentation done on different locations of the cyclic structure can yield only one linear fragment ion, which does not provide much information to derive the structure. Thus in order to get enough peaks information, we have to generate fragments with further dissociation on these linear sequence.

Another issue to be discussed is the availability of NRP database. For spectral library search, there are currently only a small amount of spectra data from comparatively pure nonribosomal molecules, which makes spectral library search not a practical option for NRP identification at present.

Meanwhile structural database for NRPs are still developing. NORINE, the database of NRPs now have more than 1100 molecules documented and the number keeps increasing. It has been freely available since 2008, which makes it easier to identify NRPs with a database.

Thus with all the challenges discussed above, we have already made a conclusion that compared with library database search and *de novo* sequencing, sequence database search is a more suitable method to identify NRP spectra currently. Hence it is more appropriate to satisfy the requirement of NRP dereplication. However, as the experimental spectra used are normally not pure, additional spectral information is needed to identify the analogs of certain NRPs in order to increase the accuracy. Moreover, an adequate scoring scheme is in need to tell whether the input experimental spectrum can be explained by an analog structure we compute or not.

2.3 **Related Works**

Despite that NRP and NRP analog identification is still in its infancy, there are already some former researches who have laid some basis for this thesis research. We will review some of these works and introduce some particular details of iSNAP which is very useful and helps a lot in this thesis. The limitations of these researches will also be discussed.

2.3.1 Interpretations of Spectra of Cyclic NRPS

Liu et al developed a program as well as a user friendly web interface called MS-CPA [27] in 2009 which readily annotates a mass spectrum resulting from the collision induced dissociation of cyclic peptides. MS-CPA is capable of direct annotating the actual input cyclic peptide MS spectra. This program has been proved to be capable of annotating seglitides and tyrocidines. It was also used to confirm the sequence of two newly discovered NRPs, desmethoxymajusculamide-C (DMMC) and dudawalamide-A, both from marine product.

In addition to this program, they also discovered that more than 28% of the ion intensity remained unexplained by fragments generated from the corresponding NRP. These data were acquired with high resolution. They further alternated combinations of amino acids that would result from peptide residues rearrangements. 10% of the total ion intensity could be explained by a rearrangement of the amino acid sequence in the cyclic peptide backbone. This kind of abnormal fragmentation behavior is called *non-direct sequence* (NDS). By these NDS behavior included, the explained intensity increased from 71.5% to 82.1%.

Liu et al developed MS-CPA, which could interpret tandem spectra, and revealed some unknown facts about the fragmentation behavior of cyclic NRPs. However, this program might be a good method to confirm the identified spectra but not an ideal way to do NRP analog identification. It required a NRP structure as already known information, which is not the case in our project. Also, they only used this program to match NRPs with a cyclic backbone without more complex structure, like branch structure. However complex structures are quite common in NRPs.

2.3.2 *De Novo* Sequencing of NRPs

Alex et al introduced their research in *de novo* sequencing of cyclic NRPs in their paper in 2009 [28]. A *de novo* sequencing algorithm using MS3 spectra was introduced and successfully identified certain types of NRPs. It was rather challenging given the fact that there were hundreds of types of monomers in NRPs. For NRPs with a cyclic structure, the amide bonds in the cycle could all be disconnected. In the stage of MS/MS, the cycle of each molecule of NRPs was disconnected at just one site with collision energy controlled. Since the molecule had no other structure but a cycle, MS/MS generated different linearized versions of the original NRP, and not any other fragments (Figure 0.3). Then the researchers further fragmented these linear molecules in MS3 stage. The spectra of MS3 contained more peaks information, which was the combination of ions of different linear fragments.

In order to identify the sequence with these peaks information, they first used spectral auto-convolution to find a list of significant peak values in the MS3 spectrum. After this auto-convolution, the mass shift values which gave several largest outputs were believed to correspond to the mass of monomers in the NRPs. With these mass values, researchers could get a list of monomers that formed this peptide. Thus this list of monomers acted as the role of the 20 proteinogenic amino acids in ribosomal peptides case. The sequence was built without using all the monomers but a shorter list and the complexity of the algorithm became much lower.

The algorithms could sequence purified tyrocidines, cyclosporines and surfactins. Although it was a great algorithm to sequence NRPs, like MS-CPA, this algorithm was applied to only perfect cyclic molecule, which made it impractical for most NRPs. Because the quality of the spectra could not be guaranteed, using this algorithm to identify NRPs analogs might not work in a high-throughput setting. However, using peaks information to find mass of monomers could provide a good idea for our research.

2.3.3 NRP Identification Software iSNAP

Yang et al proposed a NRP identification software called iSNAP [3], which is available freely online since 2012. iSNAP was designed to be a high through-put NRP dereplication algorithm. It was capable of handling LC-MS/MS as well as MS/MS data and was proved to be able to identify NRPs such as Kutzneride, Di-bromokutzneride and Tyrocidines A, B, C, D, E.

iSNAP was composed of three parts. The first one was the structural nonribosomal database. NRPs molecules were represented with SMILES (Simplified Molecular Input Line Specification) code [16]. SMILES stands for *Simplified Molecular Input Line Specification*, which is a standard encoding method that represents non-linear molecules with linear strings. Amide bonds of these molecules were broken to yield a list of fragments, and then created hypothetical spectra.

The next part was the scoring scheme. When experimental spectra were inputted, the algorithm matched the experimental spectrum with some spectra in the database by comparing ion mass. The score of this match was calculated based on the relative intensity of certain peaks of the experimental spectra. With additional normalized scores as well as appropriate threshold value, the algorithm could correctly identify NRPs in low or moderate quality samples.

The last component of iSNAP was for NRP analog search. The identification output of the database search served as seeds in this part. The algorithms analyzed the input experimental spectra again to find analogs of seed NRPs. It could also use user specified seeds.

iSNAP did well in identifying NRPs with comparatively high speed. It could also identify NRP analogs. However, the analog identification part could only search peptides with only one different monomer from the seed NRP. This was because the algorithm just shifted mass of each monomer in seed NRP by mass difference of the two precursors. However, for analogs with two or more different monomers, it is impractical to enumerate all mass shift values. Not only will the running speed of the algorithm become too slow, but the spectrum also does not contain enough information to accurately determine two modifications from the seed structure. Thus a new algorithm is in need to identify NRP analogs with more than one different monomer.

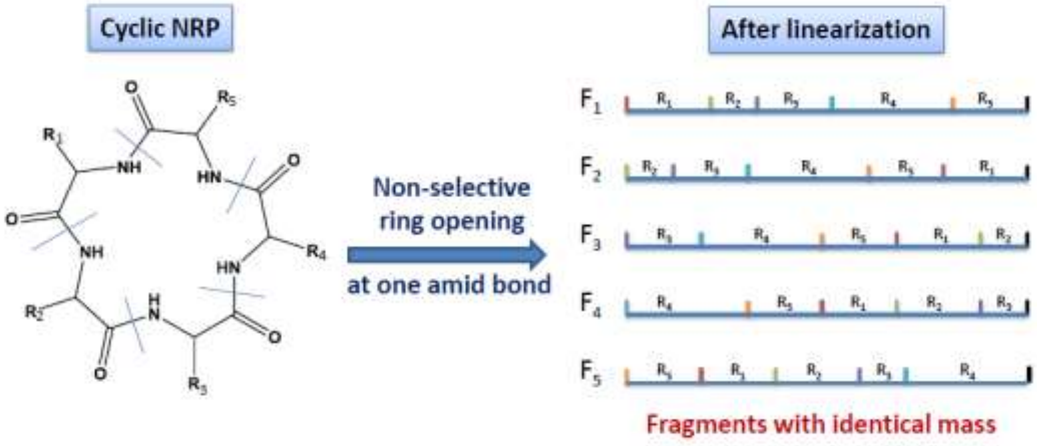


Figure 0.3 Linearization of a cyclic NRP molecule [3].

Chapter 3

Nonribosomal Peptides Analog Identification

3.1 Problem Definition

Recall that an analog NRP structure is a structure that differs from a known structure (the seed structure) with very few residues. In our thesis, we are given two MS/MS spectra, one from a known NRP structure (the seed) and the other possibly from an unknown analog NRP structure. The main purpose of the thesis is to develop algorithm to compute the analog NRP structure, or label the spectrum as not generated from an analog.

Let S be a given spectrum and M the total peptide mass of its corresponding ion. S is represented by a peak list. Each peak (m_i, h_i) corresponds to a fragment ion, where m_i represents the mass to charge (m/z) value of the peak and h_i is its intensity. Over the whole spectrum h_i could vary a lot. A peak at m_i is called a significant peak, if its corresponding h_i is no less than 0.5% of the largest h_j in the spectrum.

Structures of seed peptides are represented by its SMILES code [16]. Each building block of nonribosomal peptides as well as part or whole molecule can be represented by a unique SMILES code. These SMILES codes are further parsed and represented by a graph G (Figure 0.5) of residues and bonds. Each vertex v_i represents a single residue which has mass $m(v_i)$. Each edge e_{ij} represents bonds between residues v_i and v_j. Let Φ be the set of connected subgraphs of G, every subgraph in Φ represents a possible fragment of the NRP. In the process of fragmentation, possible mass offset of hydrogen and other common neutral losses (water, ammonium and carbon monoxide), make m/z value of the actual ion slightly different from the fragment peptide. For each NRP structure G, by considering all these mass offsets, we can get m/z value of all possible ions and generate a peak list $\{m_i\}$. This is called the hypothetical spectrum $S(G)$ of the analog structure G. In theory, this hypothetical spectrum should match the experimental spectrum of G relatively well.

Conversely, given an experimental spectrum S of an NRP, in theory one can try to compute its structure G by enumerating all possible structures and finding the one that maximizes the matching score between $S(G)$ and S. However, as discussed in the introduction section, the noise and missing of peaks in the experimental spectrum makes such a process difficult. The experimental spectrum S may not have enough information to confidently determine G from all NRP structures.

To achieve the goal of NRP structure determination, in this thesis we make two additional assumptions. First, we require the unknown NRP structure to be an analog of a known seed structure G_{seed}. This requirement effectively reduces the search space of the unknown structure, and therefore makes it more likely to use the noisy experimental spectrum to determine the unknown structure. Meanwhile, since there are a significant number of NRP structures that are analogs of a known structure, the solution to the problem under this assumption is still widely useful.

Our second requirement is to have the experimental spectrum of the known seed structure available. Denote the seed structure and the unknown analog structure by G_{seed} and G_{analog}, respectively. Denote the experimental spectra of the two structures by S_{seed} and S_{analog}, respectively. Since G_{seed} and G_{analog} differ by only one or two residues, their spectra S_{seed} and S_{analog} share a lot of peaks, while differ at a few critical peaks that can be used to derive the residue differences. The availability of S_{seed} in addition to S_{analog} can help highlight these critical peaks and help us to design more accurate scoring functions and simpler algorithms to compute G_{analog}. This requirement is very reasonable practically. In fact, a major application of our algorithm is to identify the analogs from the data after the seed structure has been identified from the same dataset. In such an application, the spectrum of the seed structure is naturally known without any additional experiments.

We leave the discussion of the scoring function in later sections, and assume that such a scoring function $score(G_{seed}, S_{seed}, G_{analog}, S_{analog})$ is made available to us. Then our main problem can be defined as follows:

> **Analog Identification Problem**: Given a spectrum S_{analog}, a seed structure G_{seed}, and the spectrum S_{seed} of G_{seed}, compute an analog structure G_{analog} such that $score(G_{seed}, S_{seed}, G_{analog}, S_{analog})$ is maximized.

To achieve this goal, we have to deal with two main tasks: generating potential analog peptide graph G_{analog}, and developing the scoring function $score(G_{seed}, S_{seed}, G_{analog}, S_{analog})$. With both tasks solved, we can solve the main problem by enumerating each potential G_{analog}, computing its score, and reporting the one with the highest score.

G_{analog} and G_{seed} share the same structure and the only difference is the corresponding $m(v_i)$ of v_i. Thus in order to compute G_{analog}, we have to decide the modified sites v_i and the corresponding mass shift value $\Delta mass$. We calculate v_i and possible $\Delta mass$ by shifting S_{seed} to match S_{analog}. Let $\Omega = \{\Delta mass_1, \ldots, \Delta mass_{|\Omega|}\}$ be the set of $\Delta mass$, each $\Delta mass_i$ has a matching score sc_i. Denote the mass of seed NRP and NRP corresponding to S_{analog} by M_{seed} and M_{input}, respectively.We wanted to find combinations of $\Delta mass_i$ with higher sc_i such that the elements in each combination add up to the difference between M_{seed} and M_{input}. This is a combinatorial optimization problem. Obviously, $\Delta mass$ can be both positive and negative, which leads to large number of possible combinations. Hence it is impractical to do exhaustive search. Thus, we need to develop a combinatorial algorithm to find the optimal combination of $\Delta mass_i$.

After generating all potential G_{analog}, we also need to develop a decent scoring scheme $score(G_{seed}, S_{seed}, G_{analog}, S_{analog})$. The scoring function should tell apart better G_{analog} from random ones and yield satisfactory output. Under this scoring scheme, a higher score means a higher probability that the input spectrum is generated from an analog of the seed and can be explained by our G_{analog}.

In practice, we are usually given a large dataset with hundreds to thousands of spectra for unknown structures. Additionally, we are given a list of seed structures and their experimental spectra. However, the experiments do not tell us whether a spectrum is from an analog of the given seed structures, and if yes, which seed structure it is. Although one can apply the algorithm for the above Analog Identification Problem to each combination of spectrum and the seed, such an exhaustive approach would be too time consuming. So, we additionally need to develop a faster filtration algorithm to find the potential analog spectrum – seed structure pairs from the large dataset provided to us.

3.2 Algorithm Overview

This nonribosomal peptide analog identification algorithm is proposed as a computational tool to identify NRP analogs of seed structure from spectra set $\{S_{analog}\}$ generated in lab and compute optimal G_{analog} to best explain the experimental spectra. Our algorithm makes use of the result of iSNAP software, which was previously developed by Yang et al [3].

The structure of this algorithm is shown in Figure 0.1.

The first part is potential NRP analog filter for each seed structure. In this part, all experimental spectra $\{S_{analog}\}$ are provided as input. With the filtration, part of $\{S_{analog}\}$ that have higher possibility to be generated by a G_{analog} of the corresponding seed molecule will be picked out and transmitted to the following analog identification part. We denote these spectra as $\{S_{analog}\}$

The task of the next part of this algorithm is to generate a list of possible G_{analog}, $\{G_{analog}\}$ for every S_{analog}. In this part, a list of S_{analog} picked by the filter is provided as input. For every S_{analog} in this list, by solving a combinatorial optimization problem, the algorithm yields a list of v_i as potential modified monomers ($\{v_i\}_{modified}$), a list of possible $\Delta mass$ combination ($\{[\Delta mass_i]\}$), where $[\Delta mass_i]$ is a combination of possible $\Delta mass$. With $\{v_i\}_{modified}$, $\{[\Delta mass_i]\}$ and G_{seed}, the algorithm could generate $\{G_{analog}\}$.

The last part is the analog matching algorithm. We evaluate the matches between $S(G_{analog})$ of each G_{analog} in $\{G_{analog}\}$ with $S(G_{seed})$, S_{analog} and S_{seed}. With decent scoring scheme and appropriate threshold value, a G_{analog} is finalized to explain the S_{analog} or we label this S_{analog} as not generated from an analog of our seed NRP. Identification of NRP analogs is fulfilled.

3.3 Nonribosomal Peptides Data

In this thesis, both structure and spectrum of seed NRPs are in need. However, there is no such spectral library available currently. We got actual experimental spectra S_{analog} collected by research collaborators from Nathan Magarvey Lab at McMaster University. All the data used in this thesis have been previously published in [3]. Then, these S_{input} went through iSNAP as the input. Normally, there are tens of spectra of the same NRP in one group of experimental spectra. After iSNAP identified part of them, we manually picked spectra with higher quality as our S_{seed}

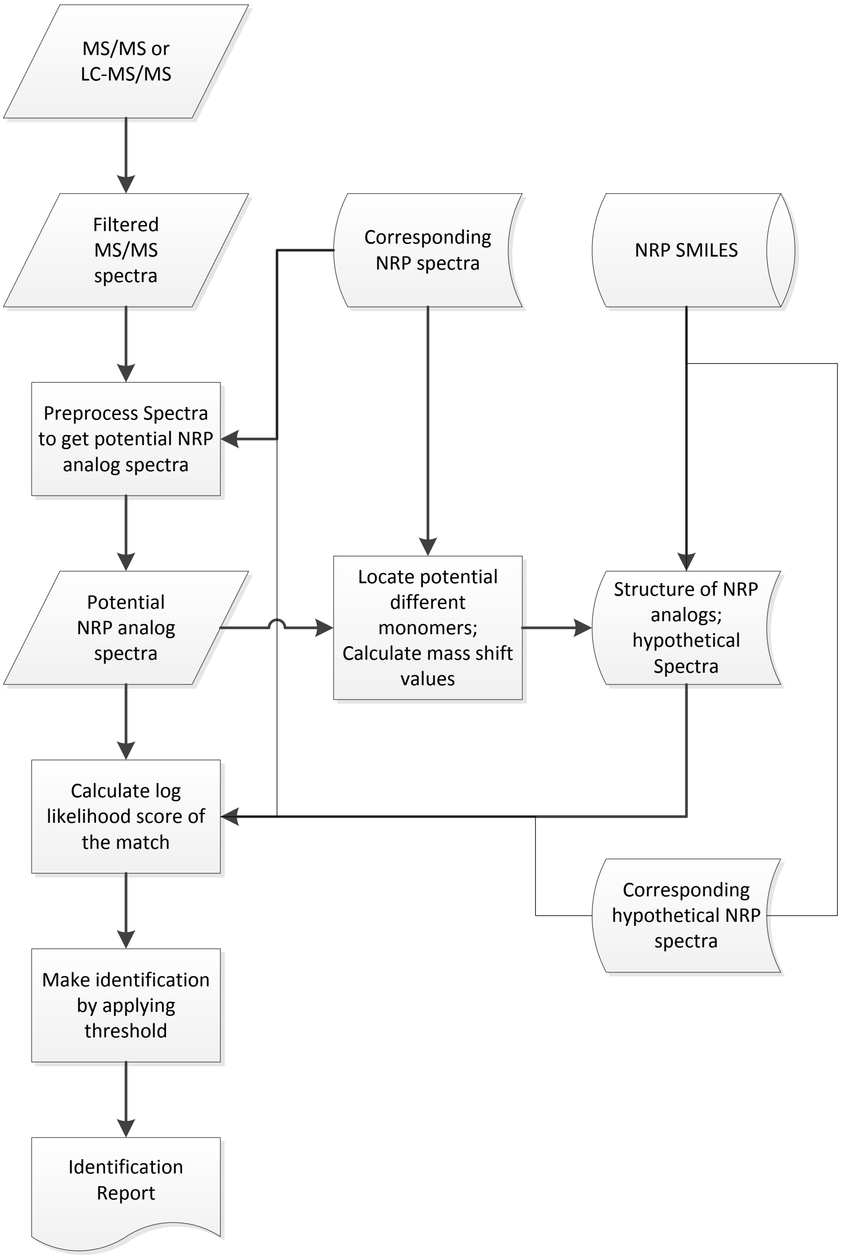
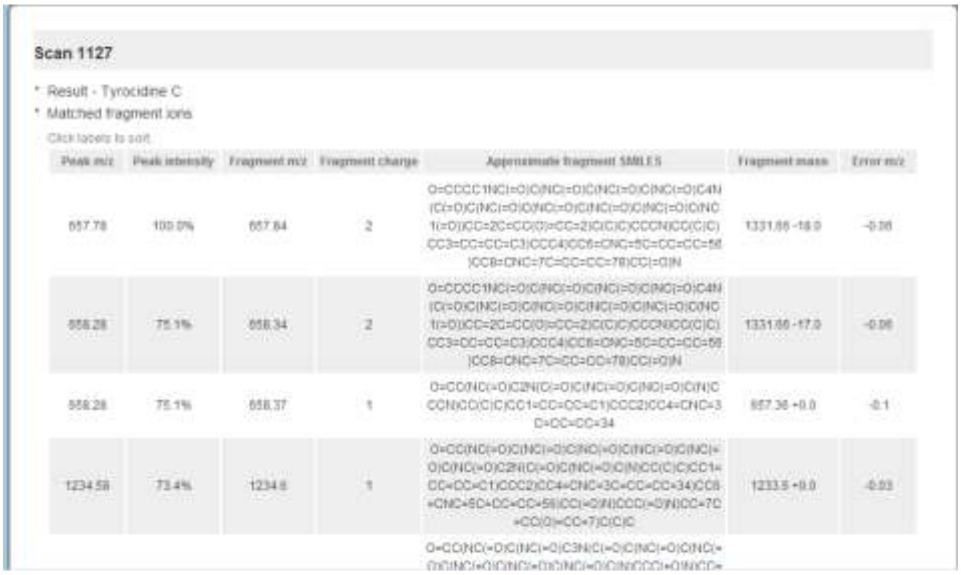

MS/MS or LC-MS/MS
Filtered MS/MS spectra
Corresponding NRP spectra
NRP SMILES
Preprocess Spectra to get potential NRP analog spectra
Potential NRP analog spectra
Locate potential different monomers; Calculate mass shift values
Structure of NRP analogs; hypothetical Spectra
Calculate log likelihood score of the match
Corresponding hypothetical NRP spectra
Make identification by applying threshold
Identification Report

Figure 0.1 Workflow of NRP analog identification. MS/MS spectra first go through preprocessing and a short list of spectra is selected as possible analog spectra. For every input spectrum, after calculating modified sites and mass shift values, analog structure and their corresponding hypothetical spectra are generated. Then each hypothetical analog spectrum will be matched with the input spectrum, seed spectrum and hypothetical spectrum of seed NRP. The structure with highest score will be our final output, thus the identification is done.

by checking spectra with PEAKS studio software. This is an important step because a well selected seed spectrum is essential for the accuracy of identification. These S_{seed} should have enough identified pairs of peaks (m_i, h_i), and as few unidentified significant peaks as possible. Thus the possibility that spectrum of two or more NRPs are contained in our single S_{seed} is relatively low.

The output of iSNAP (Figure 0.2) also includes the annotation of the experimental spectra, which means some of the peaks are identified as fragments of the NRP molecule. NRP structures, both the whole peptides and their fragments are represented by SMILES code. The SMILEs code of seed molecules and their fragments are formatted in csv files as input of our algorithm to provide structural information.

The SMILES code is firstly parsed and converted to a graph using atom-bond model using Chemistry Development Kit (CDK) [29], which stores atoms and bonds as vertices and edges. In this thesis, structures in individual residue are not considered in our identification algorithm, thus we simplify the graph and use each vertex v_i to represent a single residue. Each v_i has mass of the residue $m(v_i)$ and each edge e_{ij} represents bonds between residues v_i and v_j, then we get G_{seed}. With G_{seed}, our algorithm could handle NRP with linear, cyclic and cyclic-branching components.



Scan 1127

- Result - Tyrocidine C
- Matched fragment ions

Click labels to sort.

Peak m/z	Peak intensity	Fragment m/z	Fragment charge	Approximate fragment SMILES	Fragment mass	Error m/z
657.78	100.0%	657.84	2	O=CCCC1NC(=O)C(NC(=O)C(NC(=O)C(NC(=O)C4N(C(=O)C(NC(=O)C(NC(=O)C(NC(=O)C(NC(=O)C(NC1(=O))CC=2C=CC(O)=CC=2)C(C)C)CCCN)CC(C)C)CC3=CC=CC=C3)CCC4)CC6=CNC=5C=CC=CC=56)CC8=CNC=7C=CC=CC=78)CC(=O)N	1331.66 -18.0	-0.06
658.28	75.1%	658.34	2	O=CCCC1NC(=O)C(NC(=O)C(NC(=O)C(NC(=O)C4N(C(=O)C(NC(=O)C(NC(=O)C(NC(=O)C(NC(=O)C(NC1(=O))CC=2C=CC(O)=CC=2)C(C)C)CCCN)CC(C)C)CC3=CC=CC=C3)CCC4)CC6=CNC=5C=CC=CC=56)CC8=CNC=7C=CC=CC=78)CC(=O)N	1331.66 -17.0	-0.06
858.28	75.1%	858.37	1	O=CC(NC(=O)C2N(C(=O)C(NC(=O)C(NC(=O)C(N)CCN)CC(C)C)CC1=CC=CC=C1)CCC2)CC4=CNC=3C=CC=CC=34	857.36 +0.0	-0.1
1234.58	73.4%	1234.6	1	O=CC(NC(=O)C(NC(=O)C(NC(=O)C(NC(=O)C(NC(=O)C(NC(=O)C2N(C(=O)C(NC(=O)C(N)CC(C)C)CC1=CC=CC=C1)CCC2)CC4=CNC=3C=CC=CC=34)CC6=CNC=5C=CC=CC=56)CC(=O)N)CCC(=O)N)CC=7C=CC(O)=CC=7)C(C)C	1233.6 +0.0	-0.03
				O=CC(NC(=O)C(NC(=O)C3N(C(=O)C(NC(=O)C(NC(=O)C(NC(=O)C(NC(=O)C(NC(=O)C(N)CCC(=O)N)CC=		

Figure 0.2 Output report of iSNAP.

3.4 Potential NRP Analog Filter

The first part of our algorithm is potential NRP analog filter of S_{analog}.

Identifying analogs with more than one different building block gives considerably large upper bound for the size of $\{[\Delta mass_i]\}$, because as long as the modified $m(v_i)$ is positive, this could be a possible option. Larger size of $\{[\Delta mass_i]\}$ means more possible G_{analog}. For each G_{analog}, we have to do in *silico* fragmentation to generate connected subgraphs of G_{analog}. And this process could occupy much running time. It is too time consuming to run identification process over all S_{analog}. Thus, we need a preprocessing of the raw data to shorten the list of candidate S_{analog} to identify in order to identify the $\{S_{analog}\}$ in a shorter time.

Let $M_{diff} = |M_{input} - M_{seed}|$ be the mass difference between seed NRP and input ion, G_{input} be the actual molecular structure of S_{analog}. If we have a large M_{diff}, the actual G_{input} and G_{seed} may have too many different $m(v_i)$ or even different number of vertices. Both of above situations are not situations included in our NRP analog identification and will be tagged as not an analog. Thus we eliminate spectra whose M_{diff} is beyond a certain threshold Δm defined by the user. In our thesis we set default value of Δm as 200*Da*. Hence for each S_{analog} to be considered, we have $M_{diff} < 200$.

In this thesis, we assume that seed NRP and its analogs should have a similar way of fragmentation. In the collision-induced dissociation [30], the protonated NRPs and their analogs break the amide bonds of the same position within the gas phase of an MS/MS experiment. Thus subgraphs of G_{analog} and G_{seed} should share the same structure. When all v_i with different $m(v_i)$ are contained in the subgraph, the mass of the corresponding fragments differ from its counterpart by M_{diff}.

We shift S_{analog} by adding every m_i in S_{analog} by M_{diff}. Then significant peaks in S_{analog} are matched to significant peaks in S_{seed}. It is obvious that those S_{analog} whose corresponding G_{analog} represents an analog have more matched m_i. A score for this math between shifted S_{analog} and S_{seed} will be generated to evaluate similarity between the spectra (Figure 0.3). Here, we have m_i be m/z of one peak in S_{analog}, m_j be m/z of one peak in S_{seed}. Let $m_{diff} = |m_i - m_j|$, if $m_{diff} \leq 0.1$, then we call this a match. This mass error tolerance (0.1) is empirical decided to allow for system errors as well as random errors owing to the poor accuracy of certain mass spectrometer. Lower error tolerance makes the list of potential NRP analogs too short, while a too high tolerance results in too many matches and can hardly tell the difference between a real analog from a false one. Then we give matching scores for every match. The matching score contains two parts. The first part is the number of matched m_i and we mark it with n. The other part is the summary of log value of relative intensity of the each experimental spectrum.

$$Score = \sum_{each\ matched\ peak\ m_i} log_{10}(200 * \frac{h_i}{h_{max}}) \qquad (0\text{-}1)$$

Where h_{max} is the largest h_i in current S_{analog}. The factor 200 in the above formula is decided because in this case matched m_i with $h_i/h_{max} = 0.5\%$ yields a score of 0.

Threshold for this preprocess is set empirically in this thesis. In order to pass the filter process, an S_{analog} should have at least ten percent of the significant m_i matched with significant m_j in S_{seed}. The matching score should be larger than 20. S_{analog} generated from a G_{analog} is believed to have more matched significant peaks. Thus those S_{analog} which satisfy our filtering requirement are picked out to form $\{S_{analog}\}$ and identified by the other parts of the algorithm.

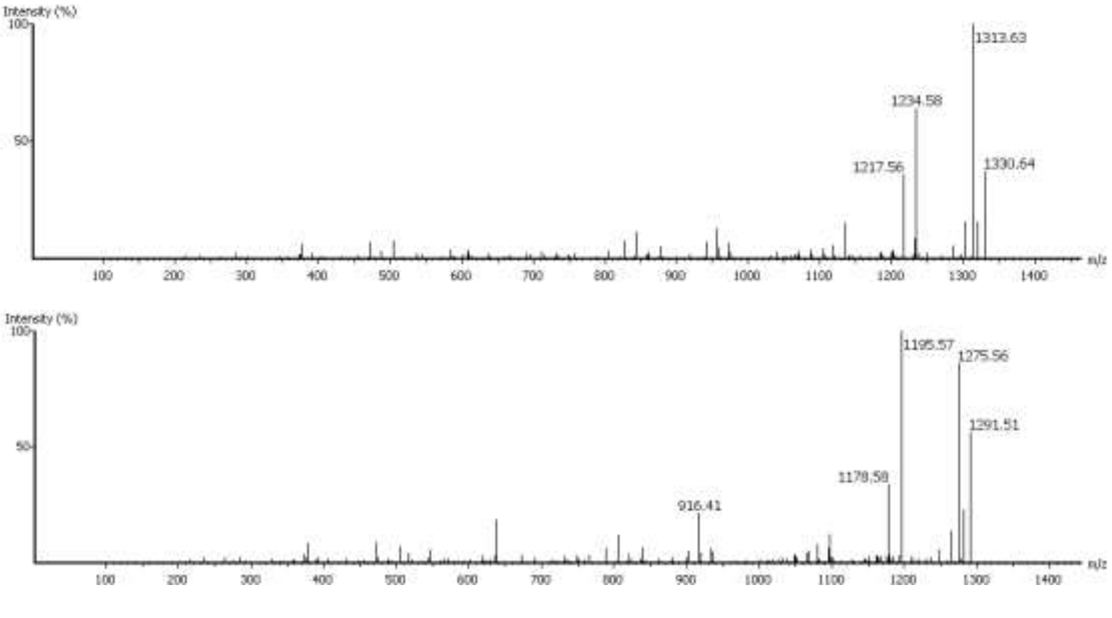


Figure 0.3 Spectra of tyrocidine C (S_{seed}) and its "potential" analog tyrocidine B (S_{analog}). When we add 39 (M_{diff}) to S_{analog}, we will get several pairs of matched significant peaks.

Each S_{seed} have a short list of S_{analog} which is considered to be a spectrum possible generated from its analog. With this filtration procedure, the running time of the subsequent analysis can be reduced considerably.

3.5 NRP Potential Analog Structure Generator

After we get $\{S_{analog}\}$ for every S_{seed}, a generator is used to generate all possible G_{analog} for each S_{analog} in the list by modifying G_{seed}. This generator mainly consists of four components. The objective of the first part is to find $\{v_i\}_{modified}$ in G_{seed}. The second part is to calculate a list of $\Delta mass$. The third part is to get combination of $\{v_i\}_{modified}$ and $[\Delta mass_i]$. The task of the last part is to generate G_{analog} and $S(G_{analog})$.

3.5.1 Modification Sites Locator

For modification sites locator, we have a certain S_{analog}, S_{seed} and G_{seed} as input, what we need is to calculate $\{v_i\}_{modified}$.

It is known that G_{seed} and our G_{analog} have the same structure and at most two different $m(v_i)$. Let $\{G_{seedsub}\}$ and $\{G_{analogsub}\}$ be the set of connected subgraphs of G_{seed} and our G_{analog} respectively. Naturally, for part of the subgraphs in $\{G_{seedsub}\}$ and $\{G_{analogsub}\}$, that do not contain those different v_i, the corresponding fragment and ions as well as their corresponding peaks m_i should be the same. Thus it is very likely that these m_i appear in both S_{analog} and S_{seed}. Consequently, the corresponding subgraphs of these m_i less likely contain v_i with different $m(v_i)$.

In the iSNAP identification report of our seed NRP, we have SMILES code of some m_i. These SMILES code can be parsed into a subgraph of G_{seed}. Thus we will know which v_i is covered by this m_i. After we get all m_i in both S_{seed} and S_{analog}, we should know the number of times each v_i appears in all matched m_i, referred as C_i. Apparently a larger C_i indicates v_i is more likely contained in those shared subgraphs. Thus we select several v_i with smallest C_i to form $\{v_i\}_{modified}$ as the list of potential modified monomers.

However, in some cases of our experiment, different C_i differ very little from each other for all the v_i. Thus it is not conclusive to say which v_i has modified $m(v_i)$ and which does not. Hence when the differences between C_i are not obvious, we can simply include all v_i in $\{v_i\}_{modified}$.

3.5.2 ΔMass Calculator

After $\{v_i\}_{modified}$ is generated, the next part of the algorithm is to calculate $\{[\mathit{\Delta} mass_i]\}$ with S_{analog} and S_{seed} as input.

In the collision-induced dissociation, the seed NRP molecules and their analogs may break the amide bonds of the same position within the gas phase of an MS/MS experiment because of their similar structure. Suppose we have v_i in G_{seed} and its counterpart v_i' in G_{analog}, and $\Delta m = m(v_i) - m(v_i')$, $|\Delta m| \geq 1$, there should be some connected subgraphs of G_{seed} and their counterpart of G_{analog}, which only contain this one v_i with different mass. Thus if the ions generated by these subgraphs have their corresponding m_i and m_j respectively in S_{seed} and S_{analog}, we should have $\Delta m = m_i - m_j$. Hence, if we add $\mathit{\Delta} m$ to S_{analog}, those m_j should be matched with an m_i in S_{seed}. Ideally every element in $[\mathit{\Delta} mass_i]$ should be covered by all possible Δm. Inevitably, some random Δm may also result in some match. However for significant m_i and m_j, they have higher possibility to find their matches when $\Delta m = m(v_i) - m(v_i')$.

Therefore, in this step we just add $\mathit{\Delta} m$ to S_{analog}, where $\mathit{\Delta} m$ is a integer and $-t < \mathit{\Delta} m < t$, t is a mass shift range set by the user. The default value of t is 200 *Da*. Then for each $\mathit{\Delta} m$, we get a S_{analog}', further we calculate matching score sc between S_{analog}' and S_{seed} following (3-1).

Then we have this score for all $\mathit{\Delta} m$, and an example score of tyrocidine B and a potential tyrocidine D is shown in Figure 0.4.

In order to get the final $\{[\mathit{\Delta} mass_i]\}$, we firstly group all $\mathit{\Delta} m$ into two different groups. The first group $\{\mathit{\Delta} m^L\}$ contains a few $\mathit{\Delta} m$ with larger sc, and the second group $\{\mathit{\Delta} m^S\}$ consists of the majority of $\mathit{\Delta} m$ with smaller sc.

Then, since we only consider no more than two different $m(v_i)$, we do 1-combination and 2-combination with repetition of the first group and get two lists of combinations of $\mathit{\Delta} m^L$. The elements in the first list $\mathit{\Delta} m^{L1}$ are combinations with only one $\mathit{\Delta} m^L$, and the elements in the second lists $\mathit{\Delta} m^{L2}$ contains two $\mathit{\Delta} m^L$. For each combination in the two lists, we calculate the total of the mass change in the combination and denote it as $\mathit{\Delta} m_{sum}$. If $\mathit{\Delta} m_{sum}$ equals to $\Delta M = M_{seed} - M_{analog}$, we just add this combination to our $\{[\mathit{\Delta} mass_i]\}$ and calculate $sc_{sum} = \sum sc^L$, where sc^L is corresponding score for those $\mathit{\Delta} m^L$. After we do this for all combinations, we get the largest sc_{sum} and we mark it as sc_{max} or we get nothing in $\{[\mathit{\Delta} mass_i]\}$

The last part of ΔMass Calculator goes as follows. For every element in the 1-combination list, we find an element in $\{\Delta m^S\}$, when $\Delta m^S + \Delta m^{L1} = \Delta M$. Then we add their corresponding matching score. If the score is larger than 20% of sc_{max}, we also add this combination to $\{[\Delta mass_i]\}$. Hence $\{[\Delta mass_i]\}$ is properly generated by ΔMass Calculator. The whole algorithm of ΔMass Calculator is shown in Figure 3.5.

3.5.3 Combination of Sites and Mass Lists

The input of this part is $\{v_i\}_{modified}$ and $\{[\Delta mass_i]\}$. And what we want is to generate a list of combinations for $\Delta mass_i$ and v_i, which is referred as $\{[(v_i, \Delta mass_i)]\}$. Every element in this list is a list of paired $(v_i, \Delta mass_i)$, and each pair consists of vertex v_i (i is the position of the vertex in G_{seed}) and mass shift value $\Delta mass_i$ of v_i.

For every $[\Delta mass_i]$, let n ($n \in \{1,2\}$) be the size of $[\Delta mass_i]$. In order to generate all possible $[(v_i, \Delta mass_i)]$, the first step is to get permutation of $[\Delta mass_i]$. Thus each $[\Delta mass_i]$ will have one (only one element in $[\Delta mass_i]$ or two equal $\Delta mass_i$) or two permutations. Then we calculate all n-combination of

$\{v_i\}_m$ $_{ied}$. Let $[\Delta mass_i, \Delta mass_j]$ be a permutation of $[\Delta mass_i]$, and

$[v_m, v_n]$ be one of its

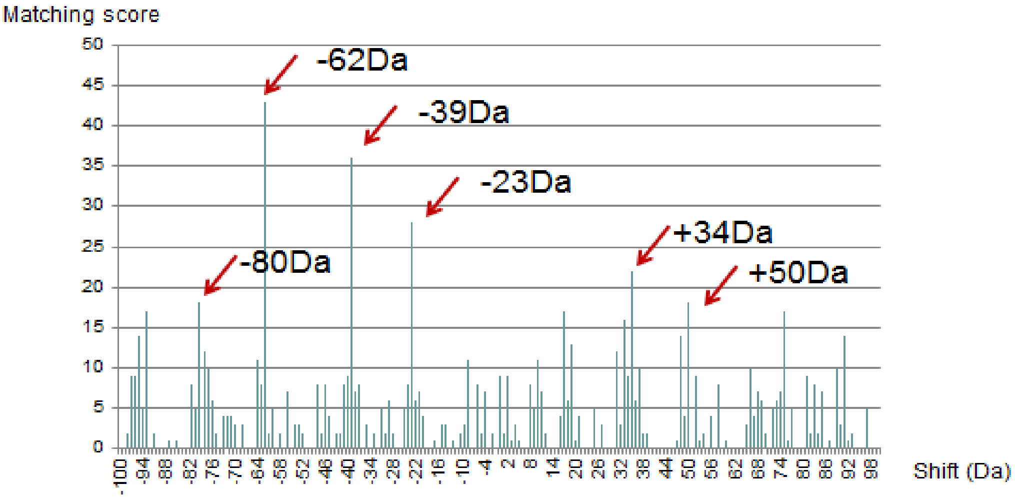


Figure 0.4 Possible $\Delta mass_i$ for tyrocidine B and its analog tyrocidine D. ΔM between the seed and experimental molecule (tyrocidine D in this example) is -62. ΔM consists of two

modified sites, and the corresponding values are -39 and -23. In this example, the program could correctly compute $\{[\Delta mass_i]\}$.

```
Input: Spectra $S_{analog}$, $S_{seed}$ and their ion mass $M_{analog}$, $M_{seed}$.
Output: a list of mass shift value $[\Delta mass_i]$ or an empty list.
  1: For every integer $-t < \Delta m < t$, let it be $\Delta m_k (0 < k < 2t)$:
      1.1 Add $\Delta m_k$ to every $m_i$ in $S_{analog}$ to get $S_{analog}^{k}$.
      1.2 Calculate matching score $sc_k$ between $S_{analog}^{k}$ and $S_{seed}$.
  2: Group all $\Delta m_k$ according to their corresponding $sc_k$, $\{\Delta m^L\}$ for $\Delta m_k$ with higher $sc_k$ and $\{\Delta m^S\}$ for $\Delta m_k$ with lower $sc_k$.
  3: Do 1-combination $\Delta m^{L1}$ and 2-combination $\Delta m^{L2}$ with repetition for $\{\Delta m^L\}$.
  4: For every combination $\{\Delta m_{candidate}\}$ in $\{\Delta m^{L1}\}$ and $\{\Delta m^{L2}\}$, if $\Delta M = M_{seed} - M_{analog} = \Delta m_{sum} = \sum \Delta m_{candidate}$:
        4.1 Add $\{\Delta m_{candidate}\}$ to $\{[\Delta mass_i]\}$.
        4.2 Calculate $sc_{sum}$.
  5: Get largest $sc_{sum}$ and mark it $sc_{max}$.
  6: For every element $\Delta m_k^{L1}$ in $\{\Delta m^{L1}\}$, find an $\Delta m^S$ in $\{\Delta m^S\}$, let $\Delta m^S + \Delta m_k^{L1} = \Delta M$:
        6.1 Calculate $sc_{sum}$ for this $\Delta m_k^{L1}$ and $\Delta m^S$.
        6.2 If $sc_{sum} > 0.2 \times sc_{max}$, add $\{\Delta m_k^{L1}, \Delta m^S\}$ to $\{[\Delta mass_i]\}$.
  7: Output $\{[\Delta mass_i]\}$ or an empty list.
```

Figure 0.5 Algorithm of ΔMass Calculator.

corresponding 2-combination of $\{v_i\}_{modified}$. We get $[(v_m, \Delta mass_i), (v_n, \Delta mass_j)]$ and add this list of paired combination to $\{[(v_i, \Delta mass_i)]\}$. Similar list will be generated when we consider other combinations of $\{v_i\}_{modified}$ for $[\Delta mass_i, \Delta mass_j]$, and other permutation of $[\Delta mass_i]$ if there is. After we deal with every $[\Delta mass_i]$ in $\{[\Delta mass_i]\}$, we should get a complete list of $[(v_i, \Delta mass_i)]$ to generate all possible G_{analog}.

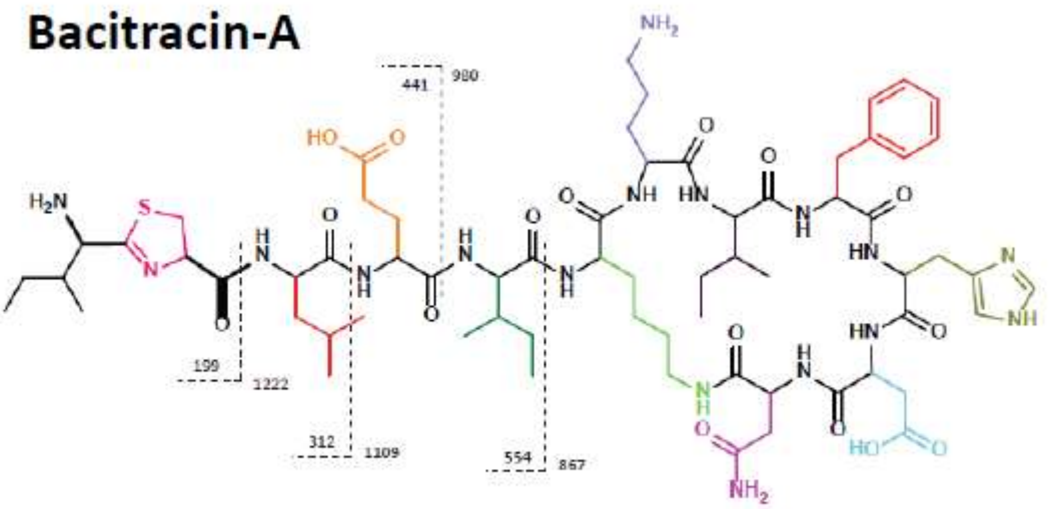
3.5.4 Potential Analog Spectra Generator

For this Potential Analog Spectra Generator, we have $\{[(v_i, \Delta mass_i)]\}$ and G_{seed} as our input and the objective is to generate all possible G_{analog} and its corresponding hypothetical spectrum $S(G_{analog})$.

For every $[(v_i, \Delta mass_i)]$, by subtracting $\Delta mass_i$ from $m(v_i)$ of G_{seed}, we could generate a G_{analog}. For example, if we have a $[(v_i, \Delta mass_i)]$ = $[(v_3, \Delta mass_3), (v_5, \Delta mass_5)]$, we just subtract $\Delta mass_3$ from $m(v_3)$ and subtract $\Delta mass_5$ from $m(v_5)$. Then this altered G_{seed} is one of our possible G_{analog}. After applying this to all $[(v_i, \Delta mass_i)]$, we get the complete list of possible G_{analog}.

For each G_{analog}, we generate all connected subgraphs and further give hypothetical spectrum $S(G_{analog})$. The method to generate this $S(G_{analog})$ is developed on the basis of work by former researchers, Yang et al (Figure 0.6). Each amide bond of these molecules, which is represented by an edge in G_{analog}, is tagged as a potential site for cleavage. By exhaustively fragmentation of the structures and taking possible mass offset into account, the algorithm enumerates all possible hypothetical peaks m_i and forms hypothetical spectrum $S(G_{analog})$. We also generate hypothetical spectrum $S(G_{seed})$.

In order to generate proper hypothetical spectrum for our evaluation algorithm, we made a minor modification of former work. In iSNAP, in order to simulate a real experimental spectrum, if we have two peaks m_i, m_j and $|m_i - m_j| < 0.01$, then it simply delete m_j from $S(G_{seed})$. In our algorithm, we keep both m_i and m_j. Thus, with this modification, $S(G_{analog})$ and $S(G_{seed})$ have the same length. Suppose m_i is the ith peak of $S(G_{seed})$, m_i' is the ith peak of $S(G_{analog})$. In our algorithm, m_i is generated from a subgraph $G_{analogsub}$ of G_{analog}, then m_i' should be generated from a subgraph $G_{seedsub}$ of G_{seed}. And $G_{analogsub}$ have the same structure as $G_{seedsub}$. If these subgraphs do not contain v_i in $[(v_i, \Delta mass_i)]$, then we have m_i = m_j, otherwise $m_j - m_i = \sum \Delta mass_i$, for every $v_i \in [(v_i, \Delta mass_i)]$ and $v_i \in G_{analogsub}$. Then we say m_i and m_i' are each other's counterpart.



O=C(CCC(C(NC(C(CC)C)C(NC1C(NC(CCCN)C(NC(C(CC)C)C(NC(CC2=CC=CC=C2)C(NC(CC3=CNC=N3)C(NC(CC(O)=O)C(NC(CC(N)=O)C(NCCCC1)=O)=O)=O)=O)=O)=O)=O)=O)NC(C(CC(C)C)NC(C4N=C(C(N)C(CC)C)SC4)=O)=O)O

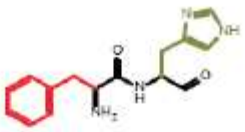
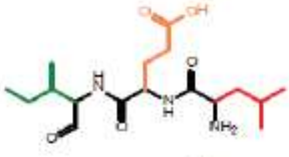
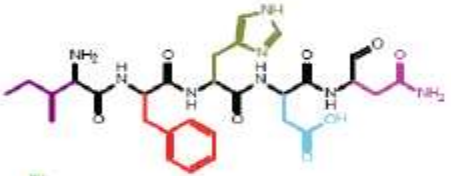
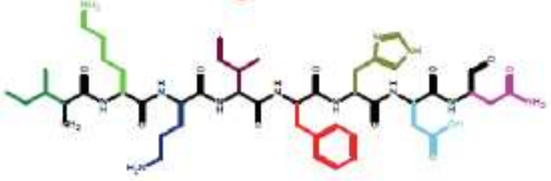
Fragment #	SMILES	Molecular structure	Molecular weight
15	O=CC(NC(=O)C(N)CC1=CC=CC=C1)CC=2N=CNC=2		286.14
16	O=CC(NC(=O)C(NC(=O)C(N)CC(C)C)CCC(=O)O)C(C)CC		357.23
24	O=CC(NC(=O)C(NC(=O)C(NC(=O)C(NC(=O)C(N)C(C)CC)CC1=CC=CC=C1)CC=2N=CNC=2)CC(=O)O)CC(=O)N		628.29
43	O=CC(NC(=O)C(NC(=O)C(NC(=O)C(NC(=O)C(NC(=O)C(NC(=O)C(NC(=O)C(N)C(C)CC)CCCCN)CCCN)C(C)CC)CC1=CC=CC=C1)CC=2N=CNC=2)CC(=O)O)CC(=O)N		983.56

Figure 0.6 Fragmentations of bacitracin-A [3].

3.6 Match Evaluation

The final step of our NRP analog identification algorithm is to evaluate matches between S_{input}, $S(G_{analog})$, S_{seed} and $S(G_{seed})$. In order to yield a good and convincing identification output, to compute an appropriate G_{analog} to explain S_{analog}, we develop a scoring function $score(S(G_{seed}), S_{seed}, S(G_{analog}), S_{analog})$ and relative scoring scheme.

3.6.1 Scoring Scheme

In this section, we introduce our scoring function. It takes $S(G_{seed})$, S_{seed}, $S(G_{analog})$, S_{analog} as input, and gives matching score for G_{analog} and S_{analog} as output.

For m_i in $S(G_{analog})$, and its counterpart m_i' in $S(G_{seed})$, we look for their match m_j ($|m_i - m_j| \leq 0.1$) and m_j' in S_{analog} and S_{seed} respectively.

When we try to find matches for an m_i and its counterpart m_i', we could run into four situations in total:

- Case 1: Both m_j in S_{analog} and an m_j' in S_{seed}.
- Case 2: An m_j in S_{analog} but no m_j' in S_{seed}.
- Case 3: An m_j' in S_{seed}, but no m_j in S_{analog}.
- Case 4: Neither m_j in S_{analog}, nor m_j' in S_{seed}.

As introduced before, NRP molecules and their analogs are similar in structure. Thus they tend to break the amide bonds of the same position within the gas phase of an MS/MS experiment and should have similar ions. Moreover, S_{analog} of a real analog should have more matches with $S(G_{analog})$. Thus, when G_{analog} is or is not the actual spectrum for S_{analog}, the distribution pattern of the four cases above should be different.

In this thesis we use log likelihood ratio to calculate the matching score. Every time when we run into one of the three situations, we calculate the score for m_i and its counterpart m_i' with log likelihood ratio. Matching score of the whole spectra is the sum of scores for all m_i and m_i':

$$Score = \sum_{j=1}^{4} n_j \log \frac{p_j}{q_j} \qquad (3\text{-}2)$$

Where n_j is the number of m_i in case j, $(j = 1,2,3,4)$. p_j is the probability that an m_i belongs to case j for a real pair of G_{analog} and S_{analog}, and q_j is the same probability when G_{analog} is randomly chosen.

Naturally, case 1 and case 2 happen more often in matches involving actual analog while case 3 and case 4 happen more often in random ones. Thus in (3-2), case 1 and case 2 yield positive score while case 3 and case 4 give negative score. Matches with higher score means they have closer distribution pattern with correct G_{analog} situation. Hence the G_{analog} we compute can best explain the S_{analog} and this scoring function works.

Before we apply this scoring scheme, we need to calculate the probabilities of case 1, 2 3 and 4. We calculate these probabilities of correct G_{analog} using spectrum of tyrocidine C as S_{seed}, hypothetical spectrum generated from G_{seed} (structure of tyrocidine C) as $S(G_{seed})$, spectrum of tyrocidine B as S_{analog}, hypothetical spectrum generated from structure of tyrocidine B as $G(S_{analog})$. We calculate these probabilities of random G_{analog} by using different $\{[(v_i, \Delta mass_i)]\}$ (adding $[\Delta mass_i]$ to other $m(v_i)$ or adding other possible $[\Delta mass_i]$). Probabilities of these four cases for correct G_{analog} and random G_{analog} are shown in Table 0.1.

Then we could calculate matching score following (3-2) for all $G(S_{analog})$ generated from G_{analog} in the list. Then the algorithm picks G_{analog}, whose $G(S_{analog})$ yields the highest matching score. This G_{analog} is the best analog structure computed which can explain S_{analog}. The algorithm gives this matching score as $Score_{analog}$.

3.6.2 Result Filtering

Sometimes, an S_{analog} that passes filtering process may not be generated from an analog of the seed NRP, like some random S_{analog} with resourceful m_i. For these S_{analog} even the G_{analog} with highest matching score is not good enough to explain it. Fortunately, this kind of S_{analog} only yields low matching score, no matter what our G_{analog} looks like. Thus we apply a filtering function before final output to improve identification result.

We provide a threshold to judge whether a G_{analog} is decent enough to be reported as the structure generating S_{analog}. The threshold is acquired empirically by running experiments. Precisely, if $Score_{analog} > 0$, the G_{analog} picked by our scoring scheme will be outputted, otherwise this S_{analog} will be labeled as not generated from an analog of the seed NRP.

An example of matching score between a seed NRP (tyrocidine C), and an S_{input} of its potential analog (tyrocidine B) is shown below in Figure 0.7.

	Case 1	Case 2	Case 3	Case 4
Correct G_{analog}	0.301	0.103	0.108	0.481
Random G_{analog}	0.245	0.082	0.171	0.501

Table 0.1 Probabilities of the four cases. We use tyrocidine C as our seed and fourteen different spectra of tyrocidine B as our S_{analog}. $S(G_{seed})$ and $S(G_{analog})$. We have 3696 pairs of peaks for correct G_{analog}. The number of pairs for the four cases is 1142, 380, 398, 1776 respectively. To get random G_{analog}, we added the mass shift value to the other monomers, one at a time. Then we get 33264 pairs of peaks in total. The number of pairs for the three cases is 8164, 2733, 5696, 16671 respectively.

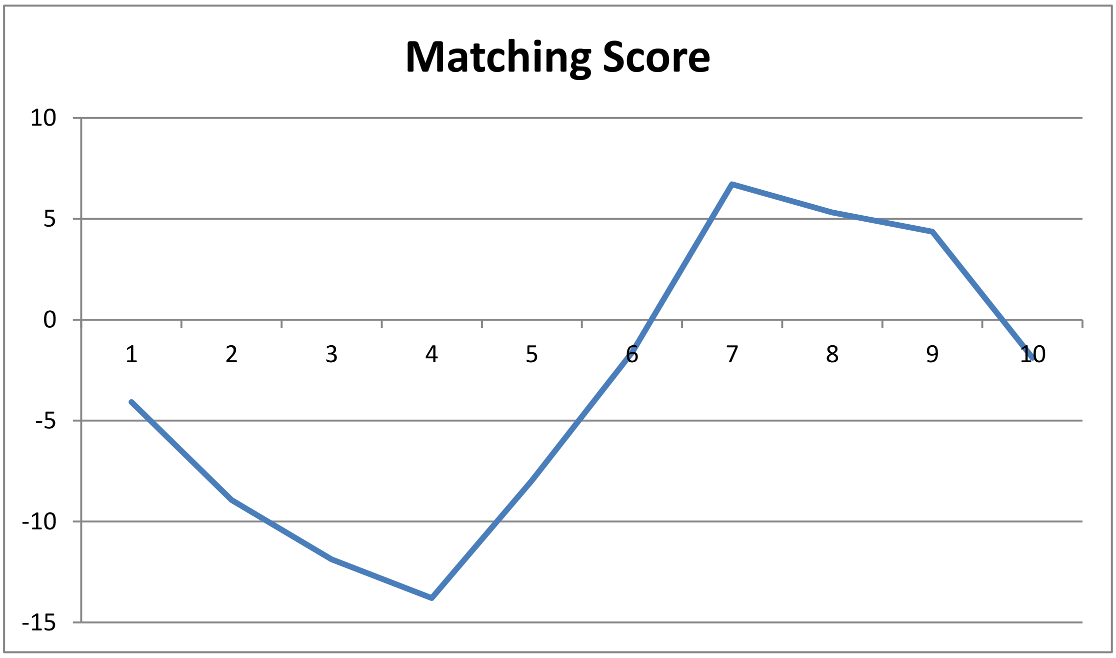


Figure 0.7 Matching score of tyrocidine C and its potential analog with different sites modified. $\Delta M =$ 39, and $\{[\Delta mass_i]\} = \{[39]\}$. G_{seed} consists of 10 vertices, thus we just subtract 39 from $m(v_i)$ (i=1,...,10), one vertex at a time. Then we generate 10 similar G_{analog} with only differences of particular $m(v_i)$. The real structure of tyrocidine B $G_{tyrocidineB}$ differs from G_{seed} by 39 at $m(v_7)$. G_{analog} with highest score differs from G_{seed} at $m(v_7)$ which agrees with what we expect.

Chapter 4

Experiments and Results

4.1 Overview

In this chapter, we will present five experiments on NRP analog identification algorithm. The first two experiments were designed to make sure that the individual parts of the algorithm work. The other three experiments tested the algorithm's ability to identify analogs under different conditions.

Experiment I was designed to test the effectiveness of filtration part of the algorithm, to check whether it could pick out those S_{input} which were more likely to be generated from an analog NRP of a seed. Lots of S_{input} cannot be explained by a G_{analog} due to poor quality of the scans. The identification part of this algorithm spends much time in building G_{analog} and generating $S(G_{analog})$ even for these S_{input}. In order to avoid the waste of resources, it is essential for the filtration part to distinguish S_{analog} which have higher possibility to be explained by a G_{analog} from those random S_{input}. However, if we set the filtration requirement too strict, too many meaningful S_{analog} might be discarded. This experiment tested the effectiveness of filtering function. Also by altering the threshold we acquired an appropriate threshold to filter the list of S_{input}.

The goal of Experiment II was to test the scoring function of the algorithm. Scoring function plays an essential part in making a correct identification. It should be able to output the correct G_{analog}. It also should be capable to discard S_{analog} that cannot be decently explained by a G_{analog}. This experiment tested the scoring function with S_{analog} of a true G_{analog} and random S_{analog}. For S_{analog} of a true

G_{analog}, we modified $(v_i, \Delta mass_i)$ to generate different G_{analog}. With the output, we examined whether the true G_{analog} was distinguished.

After these two experiments, we designed three other experiments to demonstrate the effectiveness of the algorithm as a whole.

In Experiment III, we only identified S_{analog} generated from G_{analog} which only had one different $m(v_i)$. *Bacillus sp.* [31] were cultured to produce tyrocidines, which was a series of bioactive cyclic NRPs with similar structures. Five of them (Figure 0.6) could be identified by iSNAP, and were selected to be seed structure. The NRP analog identification algorithm was used to interrogate the LC-MS/MS spectra (S_{input}) of the microbial culture, and to identify a series of G_{analog} with only one different $m(v_i)$ from G_{seed}.

In both Experiment IV and V, we demonstrated identification of S_{analog} which can be explained by G_{analog} with no more than two different $m(v_i)$. In Experiment IV, we only provided the program with S_{analog}, S_{seed}, and G_{seed}. In Experiment V, additional $[\Delta mass_i]$ was given. The reason why we provided $[\Delta mass_i]$ information was that sometimes G_{analog} and G_{seed} differed from each other on two consecutive vertices (tyrocidine A and C), and this made it very difficult for the program to get the right $[\Delta mass_i]$. However, by analyzing the structure of the molecule and possible modified monomers, researchers could reasonably guess some $\Delta mass$ in addition to those generated by the ΔMass calculator of the algorithm. We tested whether the program can yield better performance with additional information and the ability to distinguish similar G_{analog}.

The experimental spectra used in this thesis were generated and shared by research collaborators in Nathan Magarvey lab in McMaster University, with a Bruker amaZon-X ion-trap instrument and electro-spray ionization source. The mass spectrometer was coupled to a Dionex Ultimate 3000 HPLC system to perform LC-MS/MS analysis [3]. All datasets were previously published in [3].

4.2 Experiment I – Preprocessing

This experiment was designed to validate the preprocessing part, which is also the filtering part of the algorithm. The preprocessing function should be able to output a list of S_{analog} more likely generated from a G_{analog} from the whole set of input spectra. By doing this it could reduce running time and avoid the waste of resource. To do this experiment, we selected tyrocidine C as our seed NRP. S_{input} was matched and filtered by S_{seed}. It was expected that as many as S_{analog} generated from an analog of tyrocidine C would be chosen while other random S_{input} would be discarded.

We altered the filtration threshold from 10 to 30 with step size of 2. The number of selected S_{analog} and discarded S_{input} outputted as well as total number of each kind is shown in Table 0.1. We observed those S_{analog} in PEAKS studio software as the judging criteria and compared the output with the identification report of iSNAP. It is noticed that the false positive and false negative rate of the filtering output is acceptable. It illustrates that this filtration part is useful in reducing running time of the program and help decide a proper threshold value.

To further prove the usefulness of our filtration function, we also applied this it on other seed NRPs, tyrocidine A, B, D and E. For each different molecule, the output was recorded and shown in Figure 0.1 and Figure 0.2. Combining information provided by Figure 0.1 and Figure 0.2, we choose 20 to be the threshold for preprocessing part of the algorithm, on the tradeoff of false negative and false positive rate. This threshold value may not be ideal for all seed molecules spectra, however will not be too much different. With proper threshold, most S_{analog} generated from a G_{analog} were selected and random S_{input} were discarded.

4.3 Experiment II – Scoring Function

This experiment aimed to prove that the scoring scheme could be trustworthy to distinguish true G_{analog} from false one. It should be capable of giving higher score for real G_{analog} than G_{analog} with different $[(v_i, \Delta mass_i)]$. It should also give relatively low score for a random spectrum regardless the G_{analog}. To evaluate the scoring scheme, we matched two S_{analog} with S_{seed} (tyrocidine C). One was generated from its analog, tyrocidine B, and the other was a random spectrum.

The two S_{analog} were both selected by filtration function of the algorithm, and the program gave the same list of $\Delta mass_i$, only one element 39. In order to demonstrate our experiment thoroughly, we included all vertices in the $\{v_i\}_{modified}$. Thus we generated 10 different G_{analog} for each S_{analog}. Their corresponding matching scores are shown in Table 0.2. It is noticed that matching scores of S_{analog} representing an analog of the seed is greatly higher than random S_{analog}, even in the case where $[(v_i, \Delta mass_i)]$ was wrong. We can also see correct combination of $[(v_i, \Delta mass_i)]$ yield highest score. This illustrates that the scoring scheme is practical in distinguishing true G_{analog} from the fake ones.

We also tested the scoring scheme with different S_{seed} but the same G_{seed}.

Threshold value	Filtered Scan	Tyrocidines (Analog)	Not Tyrocidines
10	182	113	69
12	173	110	63
14	164	108	56
16	158	107	51
18	155	107	48
20	147	105	42
22	145	103	42
24	144	102	42
26	141	100	41
28	138	98	40
30	137	98	39

Table 0.1 Summary of filtered S_{input} for tyrocidine C as seed. The threshold values goes from 10 to 30 in this experiment. Total number of S_{input} generated from tyrocidine NRPs is 146. As the threshold value becomes larger, number of filtered S_{input} gets smaller, as well as false positive ones. As is shown in the table, the highest percentage of S_{input} generated from an analog filtered appears when threshold equals to 20.

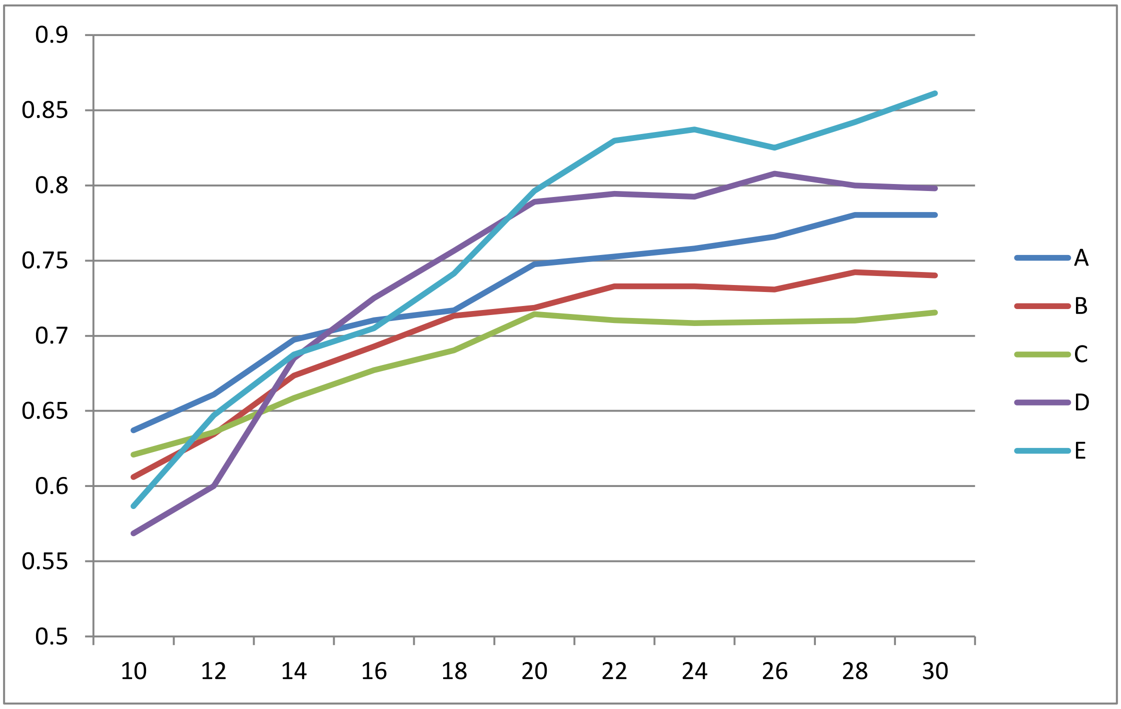


Figure 0.1 Percentage of NRP analog spectra in the filtered scans. With the increase of threshold value, percentage of spectra finally identified in filtered S_{input} keeps increasing generally. However, as we can see in the figure, this increasing trend becomes less obvious when the threshold reaches 20 for every G_{seed} except tyrocidine E. For tyrocidine C, this percentage even becomes lower, which means that the false positive rate becomes higher.

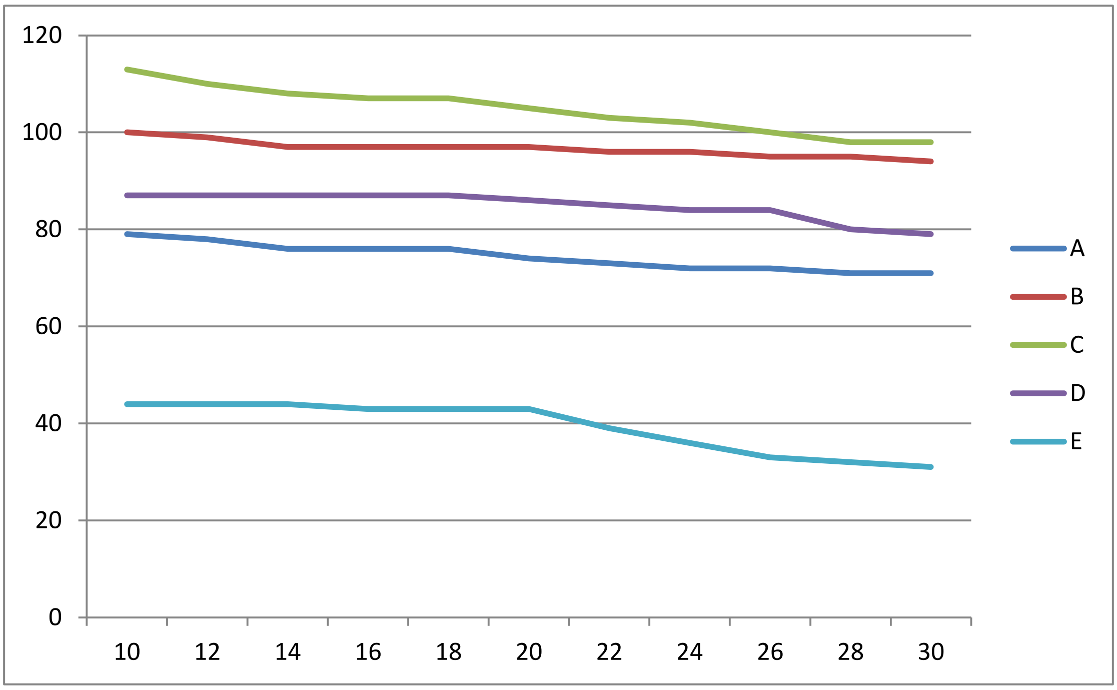


Figure 0.2 Number of S_{analog} selected by filtering function. As is shown in Figure 4.2, number of S_{analog} selected by filtering function becomes lower with the increase of threshold value. The descending trend stays nearly at the same pace from 10 to 30.

v_i	Analog	Random
1	-4.0733	-17.6411
2	-8.9321	-18.4400
3	-11.8701	-13.4764
4	-13.7910	-13.4764
5	-7.9717	-17.4796
6	-1.6198	-23.1373
7	6.7095	-22.7662
8	5.3213	-23.8881
9	4.3609	-19.2956
10	-1.8861	-16.9470

Table 0.2 Matching score between S_{seed} (spectrum of tyrocidine C) and S_{analog} of its analog (tyrocidine B) and a random S_{analog}. v_i is the different vertex between G_{seed} and G_{analog}. And the actual $G_{tyrocidineB}$ differs from G_{seed} at v_7. It is noticed that G_{analog} with the correct $[(v_i, \Delta mass_i)]$ has the highest score. We can also see that matching scores for random S_{analog} is lower than S_{analog} generated from analog. Hence this scoring system is effective.

This experiment showed whether a different S_{seed} affected the scoring scheme much. As is shown in Figure 0.3, the scoring function is still trustworthy. Although there is a little difference between different S_{seed}, it is quite reasonable because different S_{seed} may have different matched m_i in the process of identification.

4.4 Experiment III – Analog Identification with One Different Monomer

This experiment was designed to test the ability of the algorithm to identify S_{analog} possible generated from an analog NRP of the seed and further compute corresponding G_{analog}. As mentioned before, we used five different molecules in our experiment, which are all in tyrocidine family. NRPs in the same family can be generated at once in a biosynthesis pathway in the same microbial fermentation, and all the products are structurally similar. The molecules we used belong to tyrocidine family. It contains 28 cyclic NRPs as far as known by now [31]. They only differed from each other at a few monomers (Figure 0.6). Thus they are each other's analogs. However, only five of them, what we used as seed, tyrocidine A to E had been characterized. This situation provided an appropriate task for NRP analog identifier to function. In addition to these five known NRPs, there were probably S_{analog} of other molecules in tyrocidine family. In order to be a practical NRP analog identification program, the algorithm should be able to identify S_{analog} of these similar NRPs and their analogs in the same LC-MS/MS dataset.

The spectra were provided by our co-researchers in McMaster University, publicly accessible on iSNAP website. In Nathan Magarvey lab, bacitracin strain *Bacillus sp.* [31] was cultured as the source of tyrocidines. During the process of screening the microbial cultured for the tyrocidine family NRP compounds, an assay was used to detect antibiotic agents from crude fermentation extracts. Then they used HPLC to separate the crude extract (Figure 0.4), and fragmented the extract for antibacterial testing. A bioluminescent strain of staphylococcus was used as a bioactivity indicator [3]. Bioactivity screening of the extract is shown below in Figure 0.5. High bioactivity ensured the abundance of tyrocidine molecules, which made a decent material for our experiment. All datasets have been published in [3].

Analogs with only one different monomer were very structural similar to seed NRP. Hence their spectra were similar to seed molecule spectra, which made them easier to be identified. It is the fundamental requirement for NRP analog identifier to be capable of correctly identifying S_{analog} generated from these analog NRPs and

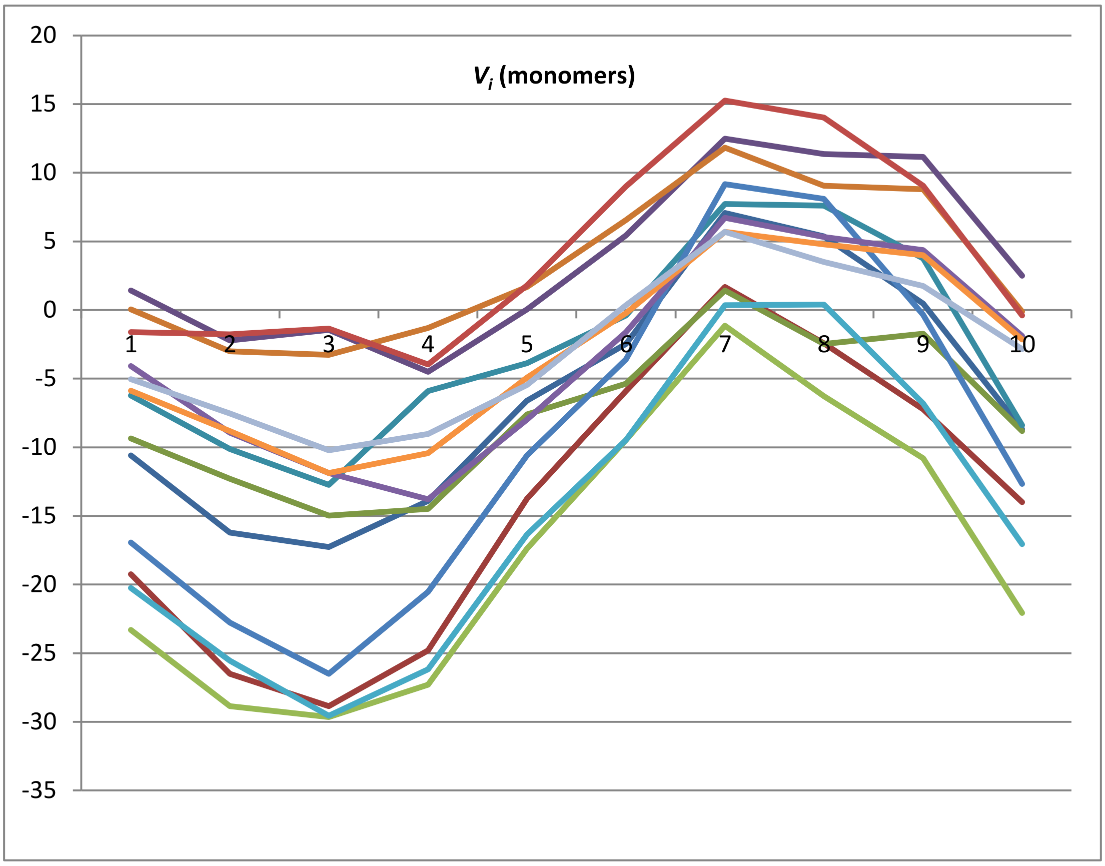


Figure 0.3 Matching score for different S_{seed} of tyrocidine C and its analog (tyrocidine B). As is shown in the figure, matching score for different S_{seed} but the same G_{seed} can vary, from below the threshold 0 to above 10. This is due to the various qualities of the spectra. Thus choosing a decent S_{seed} is very essential. Meanwhile, all S_{seed} except one have their highest score when different $m(v_i)$ is located at v_7. This agrees with correct modified sites between tyrocidine C and B. This testifies the scoring scheme is practical with a decent S_{seed} provided.

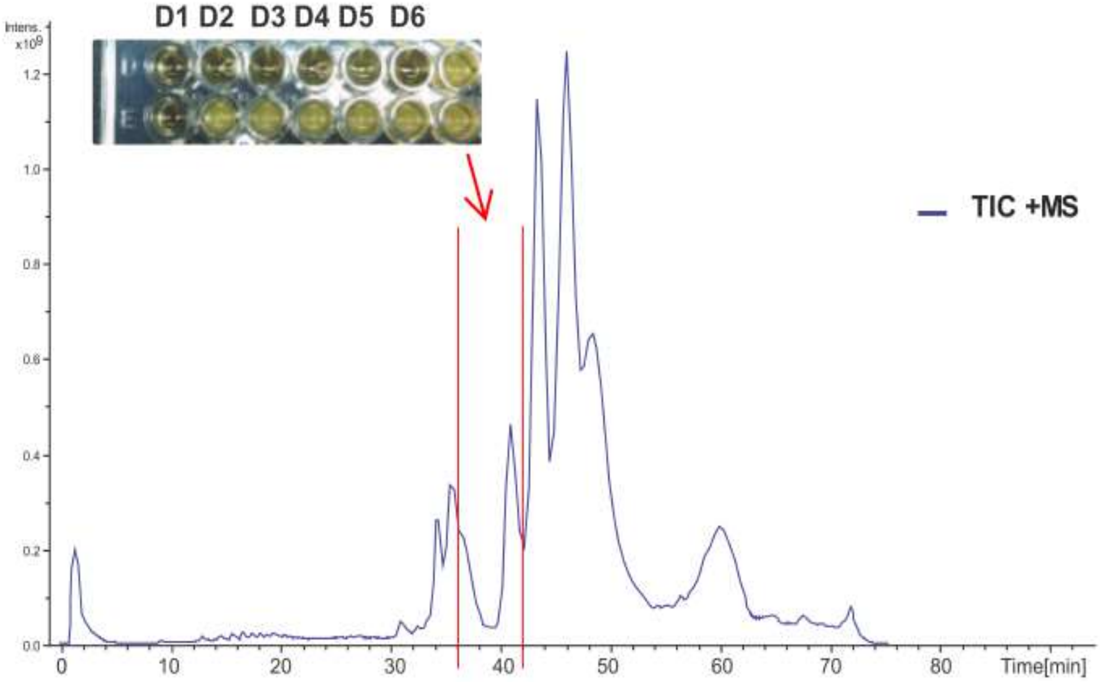


Figure 0.4 Liquid chromatogram of the Bacillus sp. extract. The bioactive fractions D1-D6 corresponds to the retention time of 36 - 42min. The LC chart shows the fermentation is a complex mixture with various compounds [3].

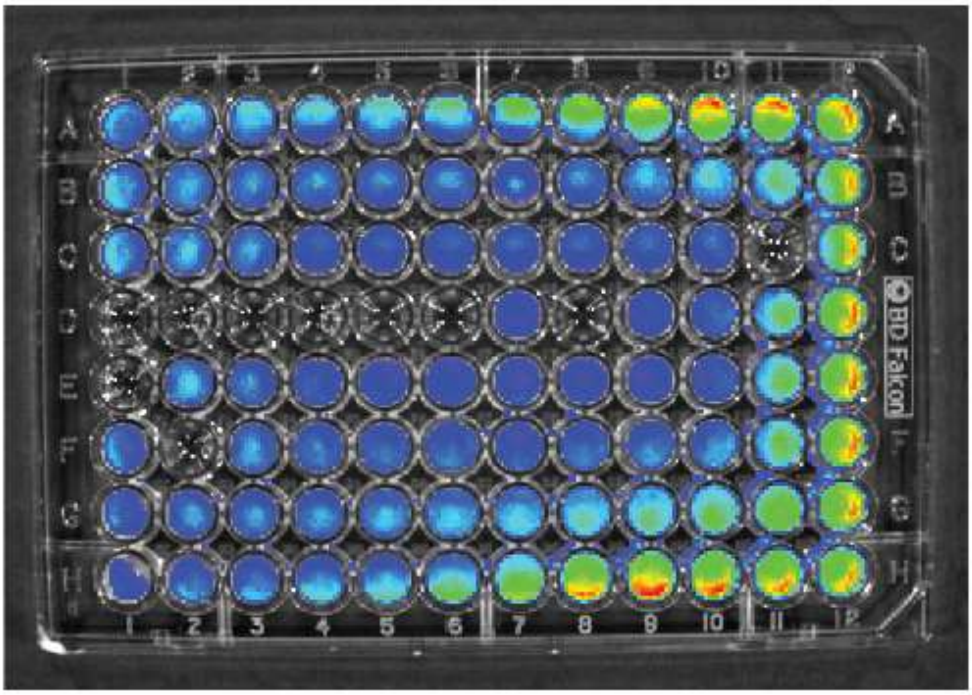


Figure 0.5 Bioactivity screening of LC fractions of Bacillus sp. extract. A bioluminescent strain of staphylococcus was used as the bioactivity indicator. Grey wells (C11, D1-6, D8, E1 and F2) indicate that the fractions are antibacterial and have killed the staphylococcus [3].

computing the right G_{analog} in more than 95% cases, giving the possibility that there might be some really tricky S_{analog} in the experimental data. We applied NRP analog identifier to interrogate the LC-MS/MS of the microbial culture and compared our identification result to iSNAP output to verify it. We still used tyrocidine C as our seed NRP. Tyrocidine B and D differ from tyrocidine C with only one $m(v_i)$, which made them our target analog. The identification report is shown in Figure 0.6 and Table 0.3.

Combining report in Figure 0.6 and Table 0.3, NRP analog identifier has been proved as an effective computational tool to identify NRP analog spectra with one modified site from seed molecule. For the other seed NRPs except for tyrocidine C, their identification report will be covered in the following Experiment IV and Experiment V.

4.5 Experiment IV – Analog Identification with More than one Different Residue

In order to further use our algorithm as NRP analog identifier, we applied it to identify S_{analog} correlates with G_{analog} with at most two different $m(v_i)$. The same S_{analog} in Experiment III were used. All seed molecules, tyrocidine A, B, C, D and E, were used. We only focused on S_{analog} of G_{analog} with no more than two different $m(v_i)$ for the following reasons. First of all, G_{analog} identified with three or more different $m(v_i)$ were not trustworthy. The matching score was generally lower than G_{analog} with one or two different $m(v_i)$. Thus the identification output is of little use. In addition, time complexity of the algorithm became extremely higher when we allowed three or more $m(v_i)$ to be modified at once. In our experiment, size of $\{[\Delta mass_i]\}$ roughly became 10 times larger when we have three different $m(v_i)$ comparing to two. Corresponding $\{v_i\}_{modified}$ is eight times of the size. They together make running time of the program intolerantly long. Hence, we limited the number of different residues to two.

In order to demonstrate the ability of the algorithm to identify analog S_{analog} and compute corresponding G_{analog} with two different $m(v_i)$, we mainly focused on tyrocidine B and D. Their structures are shown in Figure 0.7. Tyrocidine A and C were also covered in this experiment. However, they differ from each other at two consecutive monomers. Such fact makes it difficult for the program to find correct $\{[\Delta mass_i]\}$. Their identification report will be discussed in detail in Experiment V.

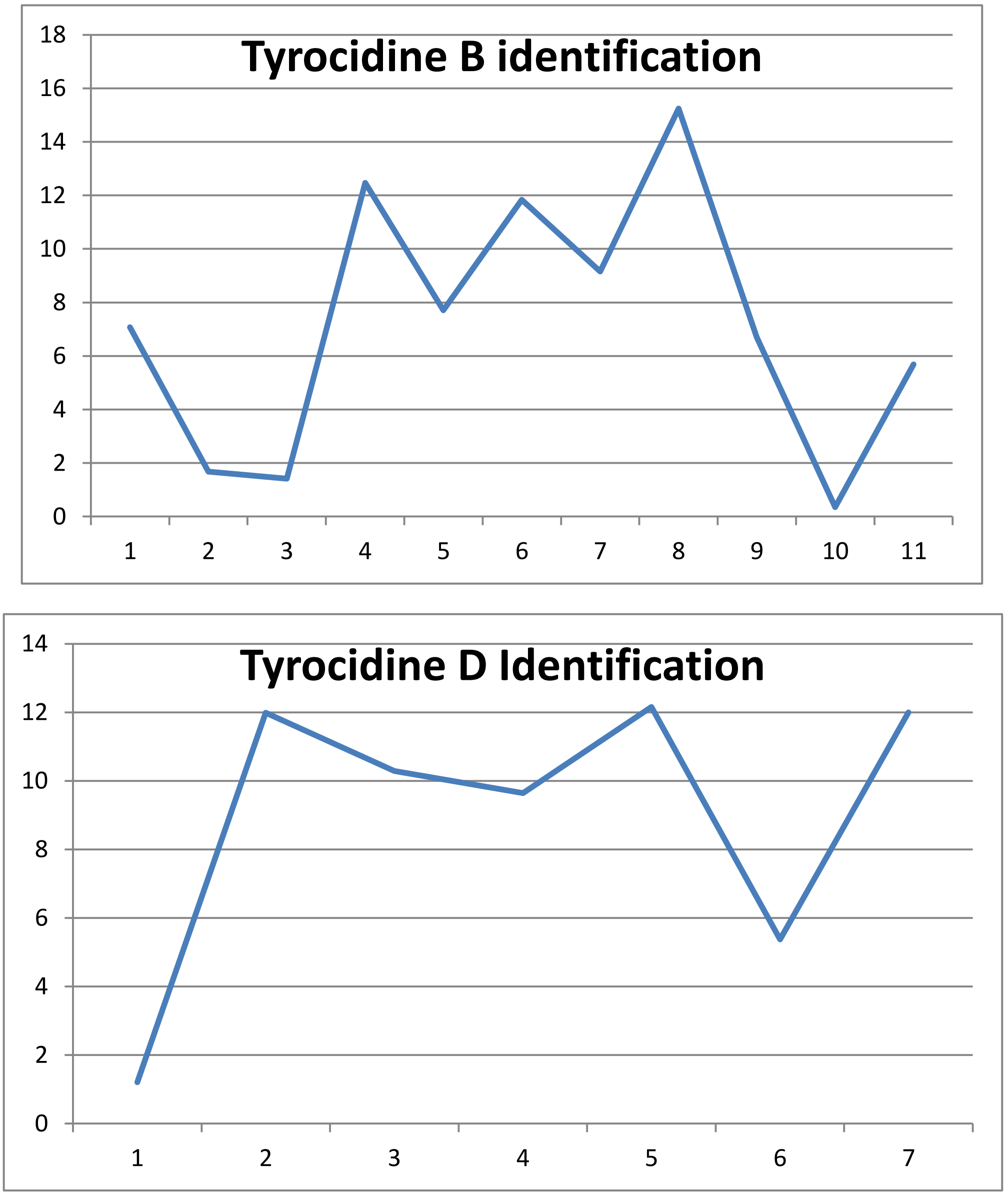


Figure 0.6 Identification result for tyrocidine B and D with tyrocidine C as the seed NRP. In iSNAP, 13 spectra of tyrocidine B and 10 spectra of tyrocidine D could be identified. We interrogated S_{analog} with our algorithm, 11 spectra of tyrocidine B and 7 spectra of tyrocidine D could be identified as tyrocidine C's analog and their G_{analog} are correctly given.

Tyrocidine B scan number	Modified Monomer (mass = 39)	Tyrocidine D scan number	Modified Monomer (mass = -23)
1162	7	1025#	
1163	7	1179	<
1167	7	1182	4
1168	7	1187	4
1172	7	1188	4
1173	7	1192	No mass
1177	7	1197	4
1178	7	1202	No mass
1183	7	1208	No mass
1184	7	1214	No mass
1189	7	1185*	4
1190	No mass	1193*	4
1200	No mass	1198*	4
1195*	8		

Table 0.3 Comparison with iSNAP database search identification report. Scan number like 1195* means this S_{analog} was not identified by iSNAP but through manual observation with PEAKS, it is an analog spectrum. "<" in the mass cell mean score is lower than threshold. "No mass" means no identification because of the incorrect $\{[\Delta mass_i]\}$. Scan number with a "#" means that iSNAP report a false positive identification.

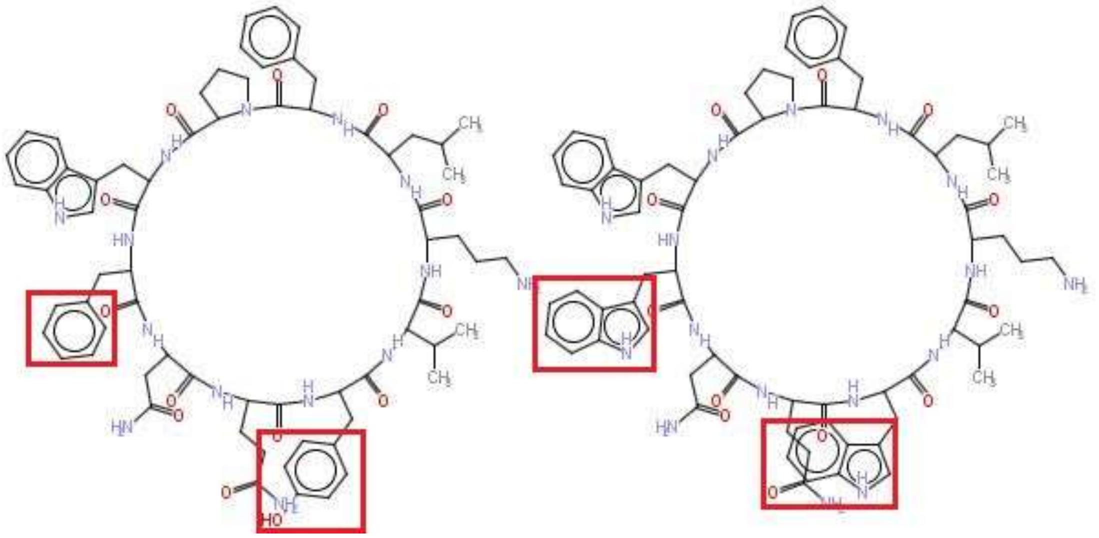


Figure 0.7 Structure comparison of tyrocidine B and D. The residues which are marked with red rectangles are the v_i with different $m(v_i)$. The two monomers were represented by v_4 and v_7. Residue at position v_4 has a Δmass of 23 and the other one at v_7 has a $\Delta mass$ of 39. Tyrocidine D has the heavier monomer at both positions. Unlike tyrocidine A and C, the two different vertices are not consecutive, thus it will be a good pair of molecules to test the algorithm's ability to identify G_{analog} with more than one modified residues.

As shown in Table 0.4 and Figure 0.8, S_{analog} whose corresponding G_{analog} differs from G_{seed} with one $m(v_i)$ were identified correctly in most case. Identification of G_{analog} with two different $m(v_i)$ was a little complicated. For tyrocidine A and C, none of G_{analog} was correctly outputted in the report. For tyrocidine B and D, only half of the S_{analog} filtered had correct G_{analog} outputted.

However, an uncharacterized molecule with M_{input} of 1293 had 8 scans identified correctly when we used tyrocidine D as seed. The identification report showed it differed from tyrocidine D at two monomers. Such phenomenon is not hard to understand. The two different $m(v_i)$s in tyrocidine A and C were consecutive and there were two vertices between them in tyrocidine D and B, while for tyrocidine D and the uncharacterized tyrocidine there were three between them. The larger number of vertices between the two different $m(v_i)$ made it easier to tell the right $[\Delta mass_i]$. This is because these $\Delta mass_i$ can be reflected by enough different m_i in S_{analog}. With correct $[\Delta mass_i]$, the actual G_{analog} could be generated.

4.6 Experiment V – NRP Analog Identification with Additional Mass Information

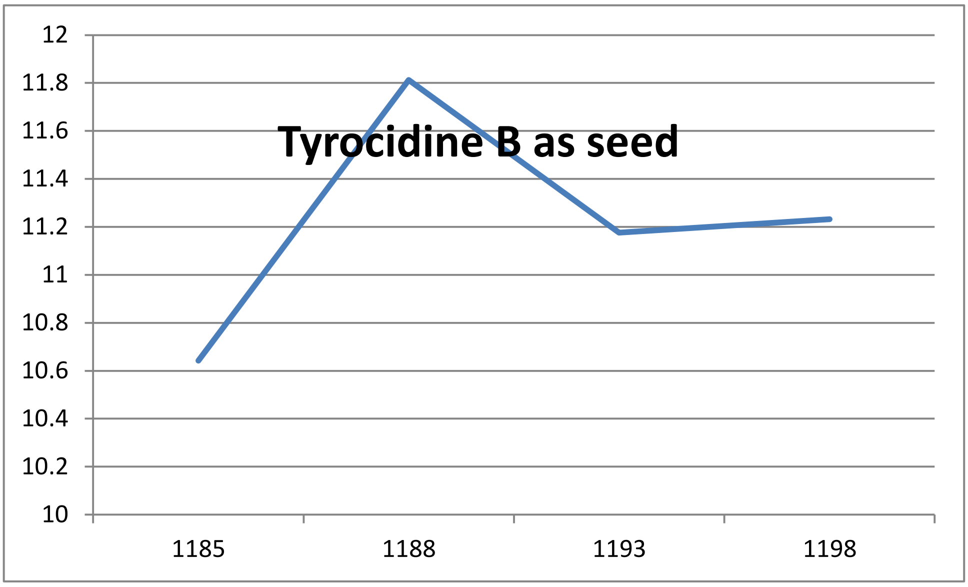
As stated in experiment IV, analogs with two different residues was hard to identify due to lack of correct $[\Delta mass_i]$, especially for two consecutive modified $m(v_i)$. Additionally, for a few cases of analog with only one different $m(v_i)$, they also lacked $m(v_i)$. Thus, we designed experiment V to test the NRP analog identifier's effectiveness when provided with additional $\Delta mass$.

We added these $\Delta mass$ on the basis of the following principles. First of all, ΔM (mass difference of the two NRPs) was added into the list if the original list does not contain it. This was done because ΔM was the mass difference of certain monomer if there is only one modified site. We also added other $\Delta mass$ based on other information. As shown in Figure 0.6, tyrocidine A, B, C, D, E were structurally similar. Their monomers differ from each other by only three values, 39, 23 and 16, which could be positive or negative. These three $\Delta mass$ were included in ΔM of the five ions, and they were also reported in our analog identification output. Thus, we could add $\Delta mass$ which appeared in our matches with high scores. Hence we could get these additional $\Delta mass$ by running this program twice. In the first time, we ran this program on S_{analog} with the same mass only once for each seed NRP to get these additional $\Delta mass$. In the second time, we added these $\Delta mass$ to our

$[\Delta mass_i]$ and applied this identification program to all S_{analog}.

Seed	A	B	C	D	E
Filtered Spectra	99	135	147	109	54
Not tyrocidine Spectra	25	37	41	23	11
A	#	7	0	#	1
B	7	#	11	5	0
C	0	10	#	8	#
D	#	4	7	#	#
E	1	3	#	#	#
Molecule of mass 1332	2	10	9	8	3
Molecule of mass 1293	9	10	2	8	8

Table 0.4 NRP analog identification report. In iSNAP, number of tyrocidine A, B, C, D and E identified was 10, 13, 10, 9, and 6 respectively. As is shown in the table, with only spectral information provided, most part of S_{analog} can be identified. In addition, S_{analog} unidentified by iSNAP but observed in PEAKS were identified. Besides, two novel analogs were identified by NRP analog identifier.



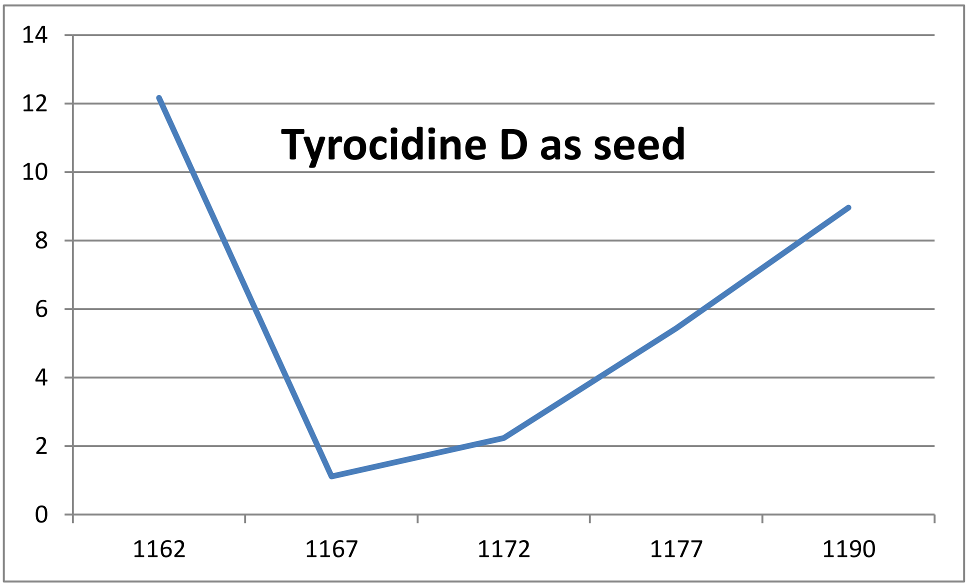


Figure 0.8 Matching score between tyrocidine D and B. We ran the program to identify analogs with two modified residues. S_{analog} (spectra of tyrocidine B when tyrocidine D is the seed and vice versa) which passed filtering process were twice the size of the final identified S_{analog}. Only one of them was discarded because of low matching score, while the other spectra were discarded due to lack of decent $[\Delta mass_i]$.

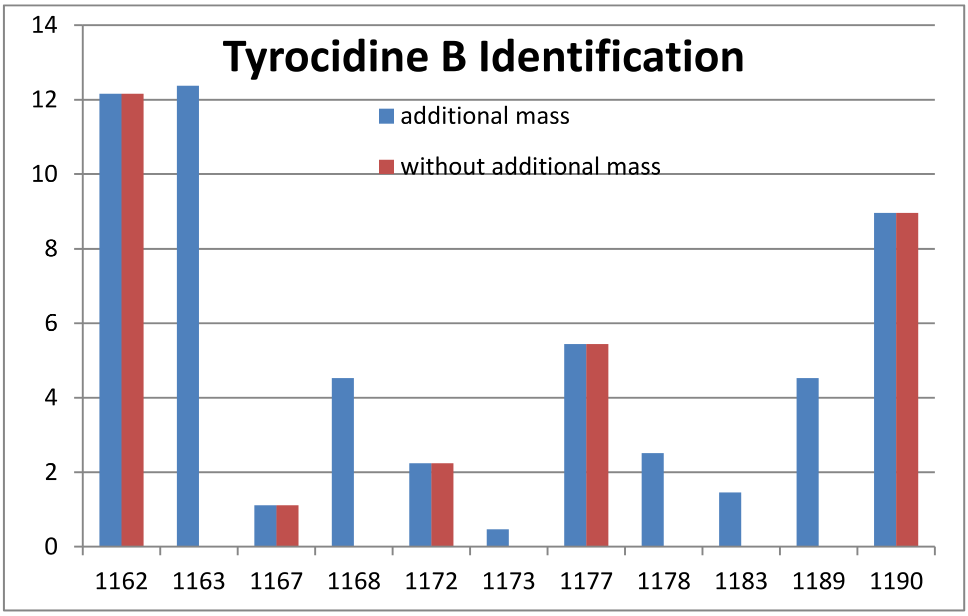
Seed	A	B	C	D	E
Filtered Scan	99	135	147	109	54
Not tyrocidine Scan	25	37	41	23	11
A	#	7	9	#	2
B	11	#	11	11	4
C	9	13	#	13	#
D	#	7	11	#	#
E	1	3	#	#	#
Molecule of mass 1332	9	10	9	8	3
Molecule of mass 1293	10	10	5	8	8

Table 0.5 NRP analog identification report with additional $\Delta mass$. Compared to Table 0.4, identification output of analogs with two different $m(v_i)$ improved a lot.

Identification report of the all S_{analog} was shown in Table 0.5. We recorded identification of tyrocidine B and D with each other one as seed in Figure 0.9. Identification report of tyrocidine A and C was shown in Figure 0.10.

As shown in Experiment IV and Experiment V, there were unknown molecules identified by NRP analog identifier. Two of them were included in Table 4.4 and Table 4.5 because we believed they were analogs of seed NRPs. As was stated before, there are more than 28 tyrocidines in the family. The majority of them have mass around 1300Da [31]. The identified structure satisfied this characteristic and report of different seed molecules agreed with each other. Their $\Delta mass_i$ appeared in the other five characterized tyrocidines and could be structurally explainable at corresponding v_i. Thus although we could not conclude whether they were tyrocidines or belonged to different family, they were analogs of seed NRPs without question.

The experiment shows that NRP analog identifier has the capability in identifying analog structures of seed molecule within a complex fermentation with no more than two monomers. It can be used as an analog dereplication tool as well as novel analog discovery tool. NRPs in the same family can be generated at once in a biosynthesis pathway in the same microbial fermentation, and all the products are structurally similar. Even if we only know a few or even one NRP in that family, we can run NRP analog identifier several rounds with previously recognized NRPs as seed to identify more NRPs. Identification of different seed can also verify each other's findings. Although conclusive characterization of the novel NRP analog structure still needs further experimental techniques (such as NMR), identification report of NRP analog identifier gives a list of well-selected candidates and references to start with.



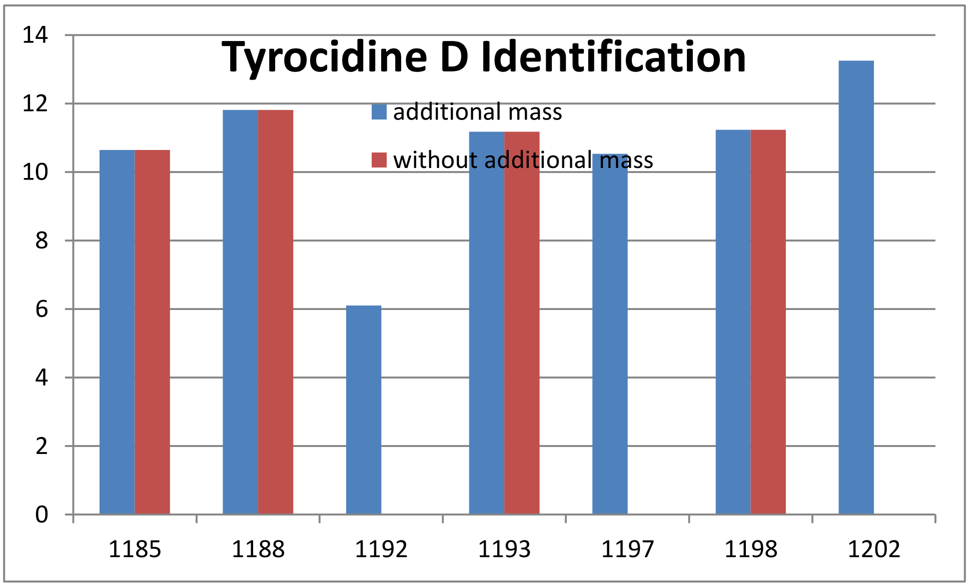
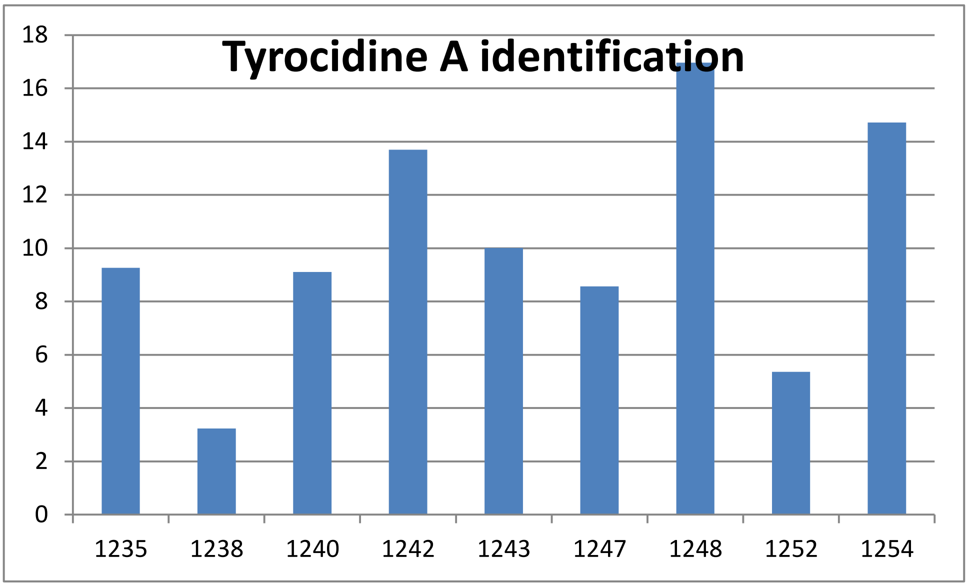


Figure 0.9 Comparison of identification output with and without additional $\Delta mass$. It can be noticed that with additional $\Delta mass$ added, NRP analog identifier could identify 11 S_{analog} of tyrocidine B and 7 S_{analog} of tyrocidine D with each other as seed molecule. While only 5 and 4 could be correctly identified without additional $\Delta mass$. Matching scores of spectra identified before and after adding $\Delta mass$ don't differ much. It illustrated that these unidentified spectra were unidentified only because of the lack of correct $\Delta mass$. Thus additional $\Delta mass$ could improve the performance of NRP analog identifier.



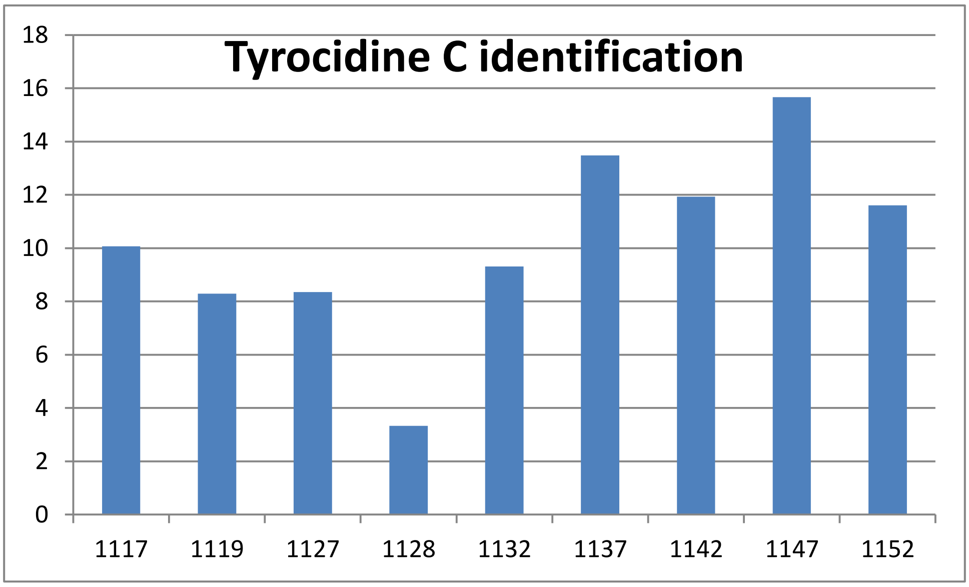


Figure 0.10 Identification of tyrocidine A and C. Without additional $\Delta mass$, NRP analog identifier could identify none S_{analog} generated from tyrocidine A and C correctly. With additional $\Delta mass$, it could identify nine S_{analog} of tyrocidine As and Cs respectively. The matching score is acceptable. Compared to 10 identified by iSNAP, this result is satisfactory. This identification result further proved the capability of NRP analog identifier to identify NRP analogs with more than one different $m(v_i)$ with additional $\Delta mass$.

Chapter 5

Future Work

5.1 NRP Analog Identification with a much Larger Seed NRP Dataset

As demonstrated in experiments in chapter 4, NRP analog identifier is able to correctly identify NRP analogs of the seed molecule with no more than two modified sites with our current seed NRP dataset. However, in our experiment, we only have five different tyrocidines as our seed NRP. Thus we could just match every experimental spectrum with all seed NRPs and yield a decent identification report.

In our experiment, we ran the program on a Lenovo personal laptop, with Intel i7-3630 2.4GHz CPU and 8 GB memory. It took about 35 minutes to identify analogs without additional $\Delta mass$. In order to improve the identification accuracy, we ran the program on part of the S_{input} to get potential $\Delta mass$ first and then on the whole set. Running time increased to around 50.

Without question, the number of available seed NRP spectra will become larger and larger in future. Thus, we may have hundreds of seed NRP structures and their corresponding spectra instead of five in future. Matching all seed NRPs with every input spectrum will lead to relatively long running time. Hence in order to reduce the running time of the program, we must improve the algorithm to reduce the time complexity of the algorithm.

Moreover, all seed NRPs used in our experiment come from tyrocidine family, and their spectra and the input spectra to be identified were generated in the same experiment. Thus there is possibility that they share some similarities that can help this algorithm work. With the development of datasets of NRP spectra, we will have seed spectra from various NRP families. Also, spectra of seed NRPs and the unknown possible analog NRPs may not be generated in the same experiment. With more and more data acquired, the scoring scheme may also need to be modified. All of these may affect the effectiveness of this NRP analog identification algorithm, thus we may need to improve the algorithm to deal with these differences in future.

However, such large seed NRP dataset is not available right now. With the development of seed NRP spectra library, future work is needed to adapt this NRP analog identification algorithm to a much larger seed NRP dataset.

Chapter 6

Conclusion

In this thesis, we defined a mathematical problem in NRP analog identification. Given a mass spectrum of a known NRP structure (the seed) and another mass spectrum of an unknown possible analog NRP structure, the problem is to either compute the unknown analog structure, or determine that the second spectrum is not from an analog of the known NRP structure. To solve this problem, we present a new algorithm and a software tool NRP analog identifier in this thesis.

The algorithm utilizes both structural and spectral information of seed molecules to identify NRP analog spectra. The algorithm first calculates combination of modified sites and potential combination of mass shift values. Then several analog structures and their corresponding hypothetical spectra as well as hypothetical spectrum of seed NRP are generated. After that matching score is calculated by log likelihood ratio. At last an analog structure is output or if matching score is below a threshold, the spectrum is labeled as not from an analog.

Through several different experiments, NRP analog identifier was proved to be an effective program in computing analog structure with no more than two modified residues. With additional mass shift information, the algorithm can correctly identify analogs even if the two modifications are on two adjacent monomers. Hence, NRP analog identifier is a useful automated tool in NRP dereplication and novel NRP discovery.

Appendix A

Appendix

A.1 Acknowledgment

NRP Analog Identifier uses following software packages as libraries.

- The Chemistry Development Kit (CDK) [29], under LGPL license.
- iSNAP

References

[1] David J Newman and Gordon M Cragg. Natural products as sources of new drugs over the last 25 years, Journal of Natural Products, 70 (3): 461-77, March 2007.

[2] David J Newman and Gordon M. Cragg. Natural products as sources of new drugs over the 30 years from 1981 to 2010, *Journal of natural products* 75.3 : 311-335, 2012.

[3] Lian Yang. Nonribosomal Peptide Identification with Tandem Mass Spectrometry by Searching Structural Database, 2012.

[4] Thirlway, Jenny, Richard Lewis, Laura Nunns, Majid Al Nakeeb, Matthew Styles, Anna‐Winona Struck, Colin P. Smith, and Jason Micklefield, Introduction of a Non‐Natural Amino Acid into a Nonribosomal Peptide Antibiotic by Modification of Adenylation Domain Specificity, Angewandte Chemie International Edition 51.29 (2012): 7181-7184.

[5] A MA van Wageningen, PN Kirkpatrick, DH Williams, and BR Harris. Sequencing and analysis of genes involved in the biosynthesis of a vancomycin group antibiotic, Chemistry & Biology, 5(3):155-162, 1998.

[6] Peter Kirkpatrick, Aarti Raja, Jason LaBonte, and John Lebbos. Daptomycin. Nature Reviews, Drug Discovery, 2(12):943-944, December 2003.

[7] D.R. Storm, K.S. Rosenthal, and P.E. Swanson. Polymyxin and related peptide antibiotics, Annual Review of Biochemistry, 46(1):723-763, 1977.

[8] Kazuko Hori, Yoshihiro Yamamoto, Toshitsugu Kurotsu, Masayuki Kanda, Setsuko Miura, Kaori Okamura, Junichi Furuyama, and Yoshitaka Saito. Molecular cloning and nucleotide sequence of the gramicidin S synthetase 1 gene, Journal of Biochemistry, 106(4):639-645, 1989.

[9] I Moln_ar, T Schupp, M Ono, R Zirkle, M Milnamow, B Nowak-Thompson, N Engel, C Toupet, A Stratmann, D D Cyr, J Gorlach, J M Mayo, A Hu, S Goff, J Schmid, and J M Ligon. The biosynthetic gene cluster for the microtubule-stabilizing agents epothilones A and B from Sorangium cellulosum So ce90, Chemistry & Biology, 7(2):97-109, February 2000.

[10] H Umezawa, K Maeda, T Takeuchi, and Y Okami. New antibiotics, bleomycin A and B, Journal of Antibiotics (Tokyo), 19(5):200-209, 1966.

[11] Stähelin, H. F. The history of cyclosporin A (Sandimmune®) revisited: Another point of view, Experientia 52.1: 5-13, 1996.

[12] Cllaud Vezin, Alicia Kudelski, and S N Sehgal. Rapamycin (AY-22,989), a new antifungal antibiotic, The Journal of Antibiotics, 28(10):721-726, 1975.

[13] Caboche, Ségolène, Maude Pupin, Valérie Leclère, Arnaud Fontaine, Philippe Jacques, and Gregory Kucherov. NORINE: a database of nonribosomal peptides, Nucleic acids research 36, no. suppl 1: D326-D331, 2008.

[14] James B McAlpine. Advances in the understanding and use of the genomic base of microbial secondary metabolite biosynthesis for the discovery of new natural products, Journal of natural products, 72(3):566-72, March 2009.

[15] Ibrahim, Ashraf, Lian Yang, Chad Johnston, Xiaowen Liu, Bin Ma, and Nathan A. Magarvey, Dereplicating nonribosomal peptides using an informatic search algorithm for natural products (iSNAP) discovery, Proceedings of the National Academy of Sciences 109.47: 19196-19201, 2012.

[16] D Weininger. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules, Journal of Chemical Information and Computer Sciences, 28(1):31-36, 1988.

[17] Pons, Miquel, Miguel Feliz, M. Antònia Molins, and Ernest Giralt, Conformational analysis of bacitracin A, a naturally occurring lariat, Biopolymers 31.6: 605-612, 1991.

[18] Robert Finking and Mohamed a Marahiel. Biosynthesis of nonribosomal peptides, Annual Review of Microbiology, 58:453-88, 2004.

[19] Dirk Schwarzer, Robert Finking, and Mohamed a. Marahiel. Nonribosomal peptides: from genes to products, Natural Product Reports, 20(3):275, 2003.

[20] Michael a Fischbach and Christopher T Walsh. Assembly-line enzymology for polyketide and nonribosomal Peptide antibiotics: logic, machinery, and mechanisms, Chemical Reviews, 106(8):3468-96, 2006.

[21] Jimmy K Eng, Ashley L Mccormack, and John R Yates. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database, Journal of the American Society for Mass Spectrometry, 5(11):976-989, 1994.

[22] Jing Zhang, Lei Xin, Baozhen Shan, Weiwu Chen, Mingjie Xie, Denis Yuen, Weiming Zhang, Zefeng Zhang, Gilles a Lajoie, and Bin Ma. PEAKS DB: De Novo sequencing assisted database search for sensitive and accurate peptide identification, Molecular & Cellular Proteomics, 11(4), December 2011.

[23] David N. Perkins, Darryl J. C. Pappin, David M. Creasy, and Jogn S. Cottrell. Probability-based protein identification by searching sequence databases using mass spectrometry data, Electrophoresis, 20(18):3551-67, 1999.

[24] The Global Proteome Machine Organization, X! Tandem Project, The Global Proteome Machine Organization. Retrieved 2009-10-21.

[25] Domokos, L., D. Hennberg, and B. Weimann. Computer-aided identification of compounds by comparison of mass spectra, Analytica Chimica Acta 165: 61-74, 1984.

[26] Bin Ma, Kaizhong Zhang, Christopher Hendrie, Chengzhi Liang, Ming Li, Amanda Doherty-Kirby, and Gilles Lajoie. PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry, Rapid Communications in Mass Spectrometry, 17(20):2337-42, 2003.

[27] Wei-Ting Liu, Julio Ng, Dario Meluzzi, Nuno Bandeira, Marcelino Gutierrez, Thomas L Simmons, Andrew W Schultz, Roger G Linington, Bradley S Moore, William H Gerwick, Pavel a Pevzner, and Pieter C Dorrestein. Interpretation of tandem mass spectra obtained from cyclic nonribosomal peptides, Analytical Chemistry, 81(11):4200-9, 2009.

[28] Alex G Harrison, Alex B Young, Christian Bleiholder, Sandor Suhai, and B_ela Paizs. Scrambling of sequence information in collision-induced dissociation of peptides, Journal of the American Chemical Society, 128(32):10364-5, 2006.

[29] Christoph Steinbeck, Yongquan Han, Stefan Kuhn, Oliver Horlacher, Edgar Luttmann, and Egon Willighagen. The Chemistry Development Kit (CDK): an opensource Java library for Chemo- and Bioinformatics, Journal of Chemical Information and Computer Sciences, 43(2):493-500, 2003.

[30] Mitchell J. Wells and S A McLuckey. Collision-induced dissociation (CID) of peptides and proteins, Methods in Enzymology Biological Mass Spectrometry, 402:148-185, 2005.

[31] XJ Tang, Pierre Thibault, Robert K Boyd, Elsevier Science Publishers B V, and K Boyd. Characterisation of the tyrocidine and gramicidin fractions of the tyrothricin complex from Bacillus brevis using liquid chromatography and mass spectrometry, International Journal of Mass Spectrometry and Ion Processes, 122(3487):153-179, 1992.