Title

Peptidogenomics approaches to study peptidic molecules from unsequenced microbes and that observed from imaging mass spectrometry

Author

Wu, Cheng-Hsuan

Publication Date

2012

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA, SAN DIEGO

Peptidogenomics approaches to study peptidic molecules from unsequenced microbes and that observed from imaging mass spectrometry.

A thesis submitted in partial satisfaction of the requirements for the degree Master of Science

in

Chemistry

by

Cheng-Hsuan Wu

Committee in charge:

Professor Pieter C. Dorrestein, chair
Professor Michael Burkart
Professor Nuno Bandeira

2012

The thesis of Cheng-Hsuan Wu is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California, San Diego

2012

DEDICATION

Thanks to UCSD, for everything.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to take this chance to acknowledge the contribution of those individuals without whom this dissertation would not have been possible.

Foremost, I wish to express my deepest appreciation to my committee chair, Pieter C. Dorrestein, who guided me equipping with an analytic technique, mass spectrometry, the knowledge of natural products. I also thank him for providing me enormous assistance in research and opportunity to sharp my presentation. In addition, I want to thank all Dorrestein lab members, Mike Meehan, Wilna Moree, Theodore Alexandrov, Vanessa Phelan, Chris Rath, Wei-ting, Jeramie Watrous, Jane Yang, Roland Kersten, Cheng-Chih Hsu, Don Nguyen, Yu-Ling Yang for giving useful suggestion all the time. It is definite my pleasure that I can work with them and develop myself.

In addition, I would like to thank all committee members, Dr. Pieter Dorrestein, Dr. Nuno Banderia, Dr. Michael Burkart for sparing time to review this thesis and gave valuable comments. Also, I would like to thank department of Chemistry and Biochemistry enriched me by providing various classes and widen my vision with lots of interested seminars, and was grateful to all the staffs helping me as possible as

they can.

Last but not least, I would like to thank my families for giving me support to finish my master degree here. Also, I would like to thank all the friends at UCSD as they colored my master life here.

Chapter 1, in part, is currently being prepared for submission for publication of the material. It is a joint effort between thesis author, Don Nyugen; Nuno Bandeira; Pieter C. Dorrestein The dissertation author was the primary investigator and author of this material

VITA

2008 Bachelor of Science, in Biological Science

National Sun Yat-Sen University, Taiwan.

2012 Master of Science, in Chemistry and Biochemistry

University of California, San Diego.

Publications

1. Inter-kingdom metabolic transformations captured by microbial imaging mass spectrometry. Wilna J. Moree, Vanessa V. Phelan, **Cheng-Hsuan Wu**, Nuno Bandeira, Dale S. Cornett, Brendan M. Duggan, Pieter C. Dorrestein; PNAS (2012)**.**

2. MS/MS networking guided discovery of molecule and gene cluster families. Don D. Nuygen[1,2], **Cheng-Hsuan Wu[1,2]**, Wilna J. Moree[1,2], Anne Lamsa[3], Marnix Medeema[4], Marystella Aparicio[5], Librada Atencio[5], Javier Ballesteros[5], Joel Sanchez[5], Ronnie Gavilán[5], Jaime Martinez[5], Jeramie Watrous[2], Xiling Zhao[2], Vanessa V. Phelan[2], Corine van der Wiel[2,6], Roland Kersten[7], Harold Gross[8], Rene de Mot[9], Elizabeth A. Shank[10] , Pep Charusanti[11], Brendan Duggan[6], Joe P. Noel[12], Bradley S. Moore[7], Nuno Bandeira[6, 13, 14], Bernhard Ø. Palsson[11], Kit Pogliano[3], Marcelino Gutiérrez*,[5], Pieter C. Dorrestein*[,2,6,7, 15]

   (Manuscript in preparation, co-first author)

ABSTRACT OF THE THESIS

Peptidogenomics approaches to study peptidic molecules from unsequenced microbes and that observed from imaging mass spectrometry.

by

Cheng-Hsuan Wu

Master of Science in Chemistry

University of California, San Diego, 2012

Professor Pieter C. Dorrestein, Chair

Genome mining has become an invaluable approach in aiding the discovery of peptide natural products; a large and diverse group of potentially bioactive molecules. Here we propose nanoDESI based MS/MS networking to map the searchable molecular universe of a large number of strains. The organization of the molecular universe via MS/MS networking enables molecular family guided genome mining to connect NRPS molecular families produced by unsequenced organisms to candidate gene cluster families found

in publicly available genome databases. As proof-of-principle, we collected nanoDESI MS/MS data on 42 bacilli and 17 pseudomonads and used molecular MS/MS networking to identify the peptidic products through generation of amino acid sequence tags from the MS/MS data. Furthermore, we use these studied dataset to identify unknown molecules from unsequenced microbes.

Chapter 1

MS/MS networking guided genome mining unsequenced bacteria through association with publicly available genome.

1.1 Introduction:

Tens of thousands of sequenced genomes or rough drafts of genomes are beginning to be available at this time and is set to grow to the millions in the next few decades. This wealth of sequence data that is emerging from the current and future generation sequencing has the potential to be utilized for the discovery of small bioactive molecules through genome mining.[1-10] Genome mining is a process where small molecules are discovered by “predicting” what will be genetically encoded. However, the process of mining genetically encoded small molecules is not keeping pace with the rate by which genomes are being obtained. In general, genome mining is still done one gene cluster at a time and requires many man-years of effort per annotation of a molecule. Therefore with the current implementation of genome mining, it will be impossible to keep pace with sequencing of genomes. The time and significant expertise current genome mining requires makes genome mining very expensive. To perform genome mining and discover all the molecules that are encoded by gene clusters found on the genome, we speculate that this easily costs a million dollars if not tens of millions of dollars. To compound the problem, it will never be possible to sequence

every genome. To highlight the staggering numbers and the magnitude of the sequencing problem, a gram of soil contains 10^9 and a milliliter of seawater contains 10^6 microbes.[11] Every human carries around 100 trillion microbes with very little overlap in the microbes we carry.[12-16] There are an estimated 10^{30} prokaryotes on this planet not including other microbes such as fungi.[11] Therefore there will always be more microbes on this planet than we can sequence. To keep pace with current genome sequencing and given that not all genomes will be sequenced, alternative approaches towards genome mining must be developed that not only take advantage of the sequenced resources available but also make it more efficient to perform genome mining on a more global scale. Furthermore we must also begin to think about developing genome mining strategies to enable the molecular analysis of unsequenced organisms using sequence information that is already available. Such methods will then also significantly reduce the cost of genome mining. Here we forward such a mass spectrometry based strategy that enables the genome mining of small molecules from unsequenced organisms using the genomic resources available in sequence repositories through utilization of MS/MS networking.

The MS/MS network based genome mining approach presented in this paper takes a more global approach than is currently the norm. This paper builds on many advances that have happened over the past decade. First, the enormous increased capacity of modern DNA sequencing technologies and the resulting data that has been deposited in public databases.[17-20] Second, our understanding of biosynthetic pathways and the function of specific enzymes found in gene clusters, especially for complex peptides made by non-ribosomal synthetase (NRPS) biosynthetic pathways that biosynthesize small molecules has dramatically increased.[21-39] This knowledge has translated into significant improvements with respect to the accuracy with which portions of the biosynthetic machineries are predicted. Finally, the last decade has seen very significant advances in mass spectrometry with respect to ion sources and sensitivity of the instruments themselves.[40-47] Atmospheric ionization methods, in combination with significant improvements in sensitivity and mass accuracy of mass spectrometry instrumentation, enables the detection of molecules directly from different surfaces.[48-52] Using the atmospheric method, nanoDESI, the molecular characterization of microbial colonies directly from agar surfaces has become

possible.[52] Here nanoDESI is used in this study to detect peptides produced via non-ribosomal peptide biosynthesis from unsequenced, and for validation, representative sequenced *Pseudomonas* and *Bacillus* strains, (Table 2) that are then subjected to molecular networking followed by peptidogenomic based genome mining to find their gene cluster families (GCF) and their molecular families (MF). GCFs are defined as gene clusters that exhibit similar gene cluster organization and a high degree sequence similarity and where the A-domain specificity is minimally altered. MFs are defined in this paper as a series of related molecules based on their fragmentation behavior that translates to structural similarity. We targeted our MS/MS network based genome mining of NRPS systems to demonstrate that genome mining can be accomplished for unsequenced organisms because NRPS derived molecules are, in our opinion, due to extensive biosynthetic studies in the past few decades, the lowest hanging fruit with respect to genome mining. It is, however, expected that through creative adaptation of the approach or related approaches, other small molecule classes such as isoprenoids, polyketides, oligosaccharides, glycolipids, lipids and other natural products can be mined as well.

For a detailed description how biology creates peptides without a ribosome, one should consult several detailed reviews from the literature.[53-58] In short, NRPS derived molecules are generally considered to be a part of the fourth branch of central dogma, the parvome.[59-60] Important molecules such as the antibiotics penicillin, vancomycin, daptomycin or immune modulators such as cyclosporine and rapamycin are examples of molecules that are made by the NRPS biosynthetic paradigm. Unlike ribosomally and post-translationally encoded peptides such as thiopeptides, lasso peptides and lantipeptides, NPRS derived peptides are produced by protein machineries. These protein machineries build the peptides in an assembly-line fashion from greater than 500 different amino acid building blocks. The non-ribosomal peptide synthetases themselves have specific domains (A-domain) that are responsible for the activation of an amino acid that covalently gets tethers to a flexible arm consisting of a phosphopantetheinyl post-translational modification on a thiolation domain (T-domain). The last required domain to build the peptides are condensation domains (C-domain) that are responsible for the formation of an amide linkage between two adjacent amino acids. Sometimes other tailoring domains, such as

halogenases, methyltransferases or hydroxylases, can be found within the assembly lines as well that further build in changes into the peptides. This domain organization gives predictive power to the backbone core of the peptides that are made by these machineries. Based on the crystal structure of GrsA, a phenylalanine loading A-domain from the gramacidin biosynthetic pathway, it was recognized that the 10-amino acid shell surrounding the activated amino acid tends to have residues similar in polarity to the amino acid that is loaded.[61-62] Using empirical knowledge from many known A-domains, it has become possible to use multiple sequence alignments to find the 10-amino acids lining the pocket of the active site and compare these to other documented 10 amino acid codes. This gives predictability to many A-domains with respect to the amino acids they load, even when the molecules are not known ahead of time. Such predictions have now been automated and integrated into informatic tools where a sequence is uploaded, A-domains are automatically identified, and predictions of A-domain specificity generated.[63-70] This predictable biosynthetic knowledge of NPRS systems and their associated informatics tools were used in this paper to connect MS/MS peptide signatures found in the MS/MS network universe. After

establishing the method with 59 Bacilli and Pseudomonads that enabled the correlation of surfactin/ lichenysin, Kurstakin, Bacitracin, iturin, thanamycin and viscosin/WLIP/MassetolideMFs[71-74] to their respective GCFs (figure 1.2), the given *Bacillus-Pseudomonas* networking was used to dereplicate NRP metabolites from Pseudomonas isolated from salmon egg.

1.2 Result and Discussion:

Genome mining of unsequenced microbes through association with sequenced genomes was accomplished via a 4-step process (outlined in Figure 1.1). The first step is obtaining fragmentation data for the molecules produced by these microbes for analysis by molecular MS/MS networking and in effect mapping the searchable molecular universe of these organisms. For this purpose we chose nanoDESI mass spectrometry as the ionization method as nanoDESI enables ionization of molecules directly from colonies grown on agar surfaces in Petri-dishes without any sample preparation[52], but LC-MS/MS or direct infusion MS/MS data could have been utilized as well. Because our nanoDESI is interfaced with an ion trap, it was also possible to fragment all the ions that are

detected when they are ionized with nanoDESI. We subjected 59 different strains of bacteria to nanoDESI analysis (Table 1); there were 42 bacilli and 17 pseudomonads and the resulting MS2 spectra were networked and visualized with Cytoscape (Figure 1a, step 2 and Figure 1b).[52] Such organization into networks enables the mapping of spectral similarity and viva infra structural similarity mapping of identical and related molecules based on their MS/MS signatures. An MS/MS cluster where many nodes are connected by edges indicates many related molecules observed, while an MS/MS cluster with few nodes can be a unique set of molecules with few alternative forms that result in unique spectra. Furthermore MS/MS networking enables the visualization of groups of unique spectral signatures that indicate the molecules are distinct from the other molecules in a given data set. Both the pseudomonads and bacilli were combined into one MS/MS network in order to map their searchable molecular universe as it was anticipated that none of the signals that come from non-ribosomal peptides would overlap between the two organisms as there are no non-ribosomal peptides that have been described in the literature that are found in both Bacilli and Pseudomonads.[70-74] The merging of the data itself enables the

removal of all signals that overlap, including any signals derived from their growth media as they are not of interest for further analysis. Combining the data from all organisms assists in the peptidogenomic based genome mining as only one MS/MS node needs to be matched to its corresponding genome, which can then be related to the surrounding nodes in the MS/MS cluster so that not every MS/MS spectra has to be individually correlated to candidate gene clusters, even when they originate from different organisms. Finally, the vast majority of NRPSs isolated from these genera contain proteinogenic amino acids, thus simplifying the peptidogenomic analysis. Combining the MS/MS data from bacilli and the pseudomonads resulted in about 22% (972/4311 nodes) overlap in signals in the data. Sources for common data from both pseudomonads and bacilli are primary metabolism, the nanoDESI solvent, and the growth media. The majority of spectra (78%) are unique to either the bacilli or the pseudomonads (Figure 1b). There are 121 MS/MS clusters of sufficiently unique fragmentation patterns that contain 3 or more nodes. These MS/MS clusters are individual molecular families (MFs). Although there are non-peptidic molecules that are observed in the MS/MS network, such as the rhamnolipids, quinolones, and phenazines[74] (Figure 1.2), the

goal of this work was to provide a proof-of-principle to correlate non-ribosomal peptides and their candidate gene clusters to demonstrate that genome mining can be accomplished with MS/MS networking and unsequenced organisms with sequencing info in the public databases through GCF and MF matching.

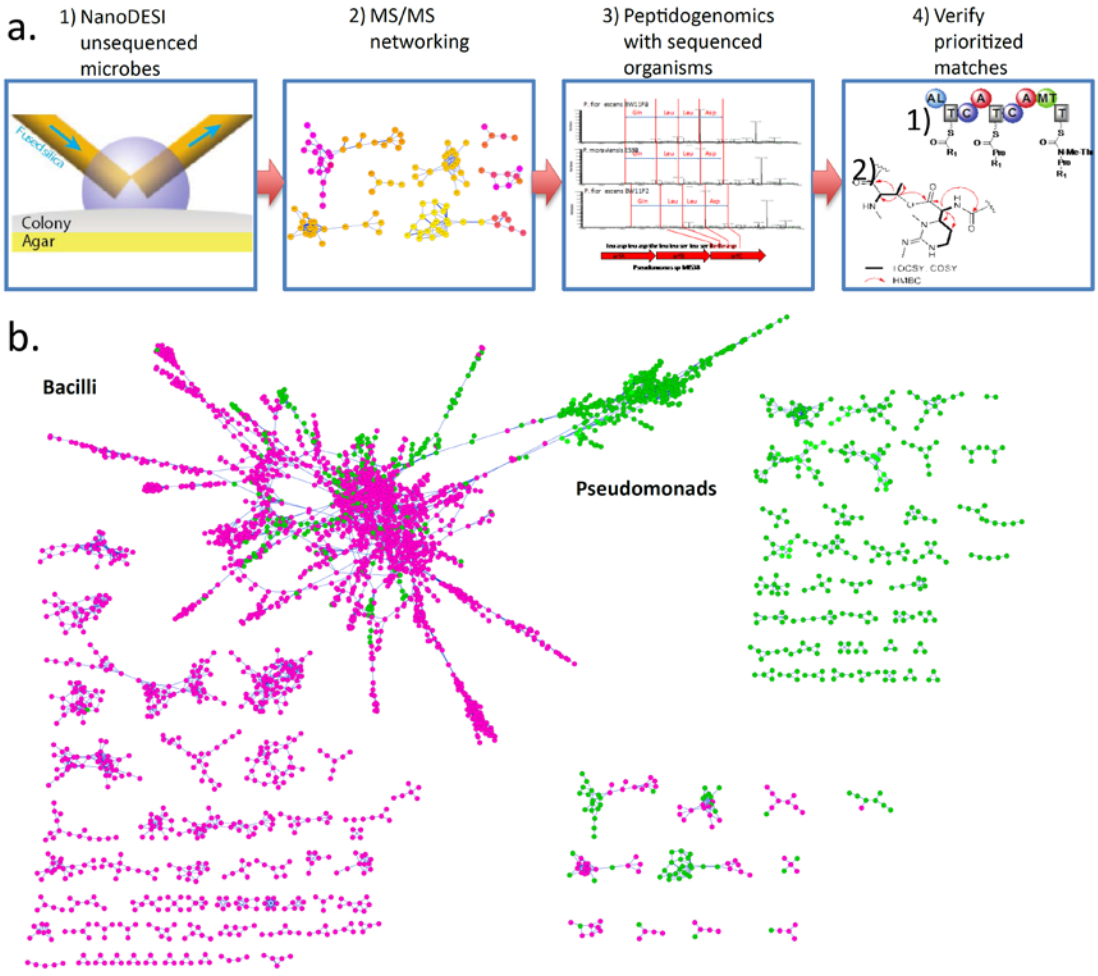

Figure 1.1: NanoDESI based MS/MS networking guided genome mining of nonribosomal peptides from 42 *Bacilli* and 17 *Pseudomonas*. a) Step 1:

Collection metabolites from live colonies by NanoDESI. Step 2: Generation of molecular networking through Cytoscape. Step 3: Sequence tag generated from MS2 raw file and prediction of all NRPS gene clusters from all publicly available genome database. Step 4: Biosynthesis pathway to verify interested molecules. b) Molecular networking of 42 *Bacilli* (pink) and 17 Pseudomonas (green).

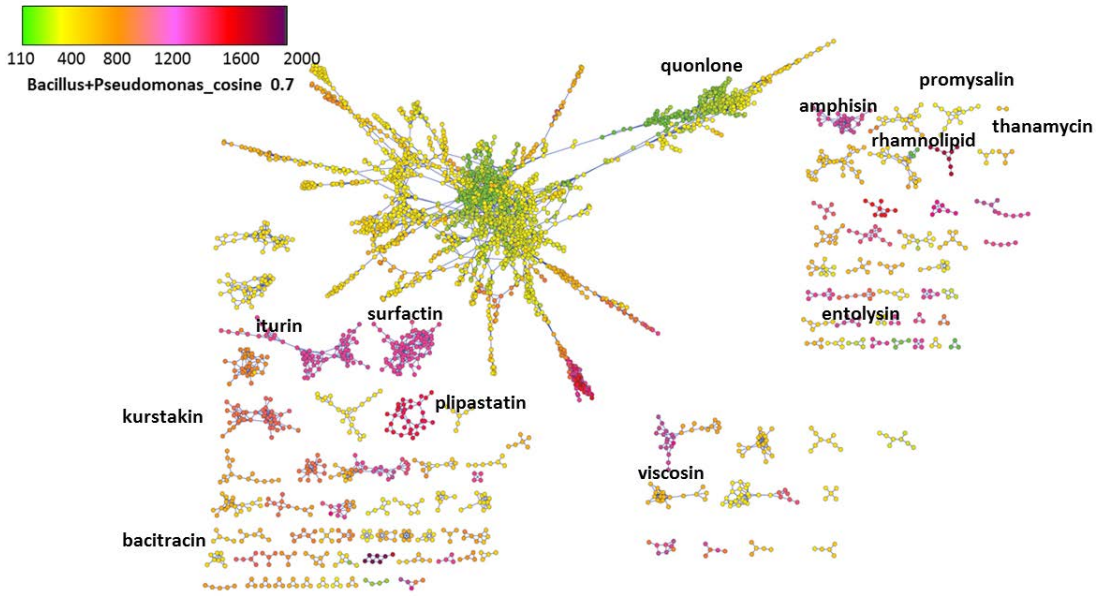

Figure 1.2 8 GCF-MF were identified from *Bacillus* and *Pseudomonas* networking and identified molecules. 4GCF-MF from *Bacillus*: Kurstakin, Surfactin, Iturin, Bacitracin. 4GCF-MF from Pseudomonas: Viscosin, Amphisin, Entolysin, Thanamycin. Identifed molecules: quinolone, promysalin, rhamnolipid, pliplastatin

Once the MS/MS network has been generated in step 2, we examined the

raw MS/MS data looking for mass shifts between adjacent ions corresponding to an amino acid sequence tag that would indicate that a particular MS/MS cluster is peptidic in nature (Table 1.2). The peptidic clusters are then subjected to peptidogenomic analysis (Figure 1.1, step 3). We limited this search space to only proteinogenic amino acids (with and without an N-methyl group) because manual annotation of spectra with more than 500 possible unique amino acids that have been identified to be incorporated into non-ribosomally produced peptides, is a nearly impossible task, especially with low resolution MS/MS data. Furthermore the vast majority of NRPS derived molecules from *Bacilli* and *Pseudomonads* have several proteinogenic amino acids incorporated. Specific algorithms can be developed in the future that will overcome this limitation especially in conjunction with high-resolution data. Instead of carrying out peptidogenomics analysis on a single organism as previously described, the adenylation domain specificity of all bacilli and pseudomonad derived DNA sequencing information in the public databases was predicted using a batch form of antiSMASH ref as well as manual curation (Table 1.2). Using both amino acid MS/MS signatures with the predicted amino acid specificity, candidate matches of MS/MS signatures to gene clusters

or families of gene clusters were obtained. At this stage, the iterative process of looking examining and matching MS data to the gene clusters was carried out because this information is needed for peptidogenomics analysis to correlate MFs and GCFs.[5] To improve our confidence in the peptidogenomics analysis we looked for several correlations. 1) Can we find additional amino acids that correlate to the A-domain specificity predictions that were missed when generating the initial sequence tag from the MS/MS data? 2) Does the biosynthetic gene cluster contain tailoring domains in the NRPS or biosynthetic enzymes to make non-proteinogenic amino acids, and can those amino acids be found in the MS/MS data? 3) Did we observe patterns of mass shifts such as +/- 14 Da indicative of different set of amino acid substitutions (e.g. Gly vs Ala), methylations, or different fatty acid chain lengths that are common to non-ribosomal peptides? 4) Does the size of the molecule match up to the size of the gene cluster (e.g. an NRPS with a 20 A-domains is unlikely to encode for a molecule that is ~1000 Da)? 5) can we dereplicate the molecule to a gene cluster or to a family of related gene clusters that encodes for that molecule but where the observed molecule does not match up perfectly to the known molecules. 6) Are

the genes and/or molecules known to be produced by this genus or species of bacteria? 7) The biosynthetic pathway is drawn with as much information at hand to ensure the mass spectrometry data correlates with the gene cluster and that the spectra matches the identity of the molecule. Using these correlations, it is possible to state that a potential GCF and MF match has been found. When the molecule or gene cluster is very important based on biological prioritization, the GCF-MF correlation will need to be confirmed via other means. Below are two examples of such GCF-MF correlation that were obtained and are described in more detail.

Mapping of the searchable molecular universe of the 59 organisms we analyzed with MS/MS networking revealed a large constellation of 78 nodes ranging from 1002 to 1116 Da and was only found within the bacilli data (Figure 1.3a). The data incorporated in the nodes came from 23 different data sets (Table 1.3). Generation of sequence tags using only proteinogenic amino acids revealed a 113 Da and a 113-113 Da pair of signatures characteristic of peptides (Figure 1.3b). For the purposes of this study, the longest consecutive sequence tag was used to carry out peptidogenomics because longer tags are less likely to detect

incorrect identifications. Future algorithms with high-resolution data will enable one to take in account all the tags that are generated including non-proteinogenic amino acids. For this MS/MS cluster a search tag of 113-113, corresponding to Leu-Leu, Leu-Ile, Ile-Leu or Ile-Ile, was used to search all of the predicted NRPS sequences obtained from A-domain specificity predictions of bacilli and pseudomonas in publicly available sequencing information. This sequence tag matched to sequence tags from *B. subtilis, B. amyoliquefaciens, B. atrophaeus, and B. lichenformis* but not to sequence tags from the pseudomonad predictions. This gene cluster family included the *B. subtilis* surfactin and *B. lichenformis* lichenysin synthetases. Comparing the gene cluster matches from the Bacilli revealed that all of the gene clusters had related A-domain specificity and similar gene cluster organization; over 80% sequence similarity (Figure 1.3d). This is in agreement with the known structures of surfactin and lichenysin. At least 17 lichenysins and 53 surfactins are described in the literature that have different fatty acid lengths and/or different amino acids in the backbone of the molecule, as many A-domains tend to be promiscuous resulting in the production of a family of molecules. Thus, the surfactin/lichenysin family-gene cluster family was identified.

We also added some sequenced strains, such as *Bacillus subtilis 3610* to our proof-of principle experiment, making it possible to verify from the fragmentation alone if the MF contained surfactin. Indeed the surfactin fragmentation data from *B. subtilis 3610* is found in this cluster. Using this approach, 4 candidate GCF-MF pairs were identified from bacilli (Figure 1.3-1.10).

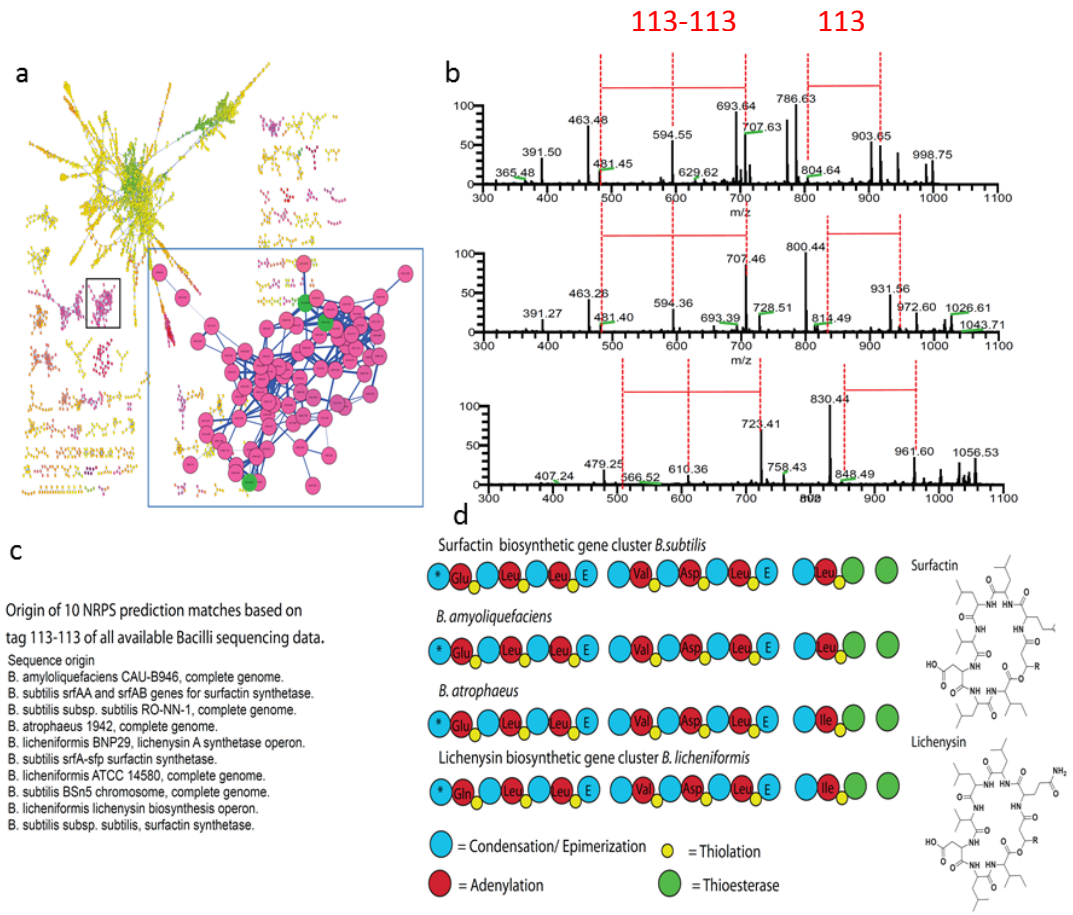

Figure 1.3: Molecular networking of *Bacillus* and Pseudomonas and identification of surfactin GCF-MF. a) Investigation of surfactin molecules from Bacilli MS/MS networking b) Random selection of MS/MS spectra for generation of sequence tag. c) Genome mining all public genomic data from NCBI by ANTISMASH. d) Matching generated tag to all predicted gene cluster from Bacilli genome.

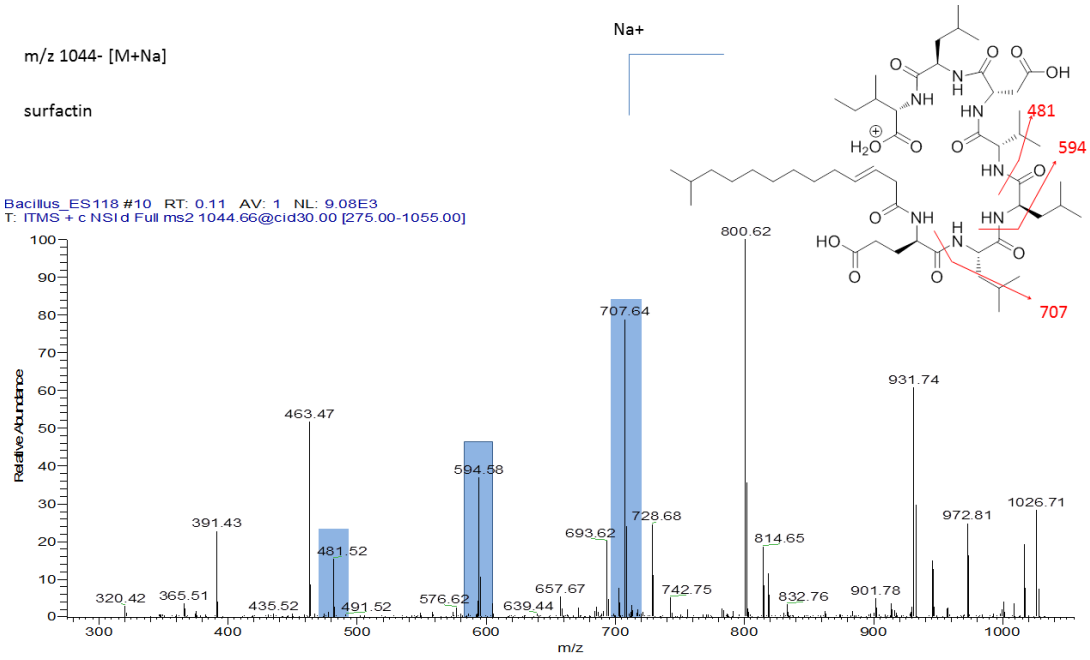

Figure 1.4: characterization of surfactin from *Bacillus subtilis*

.

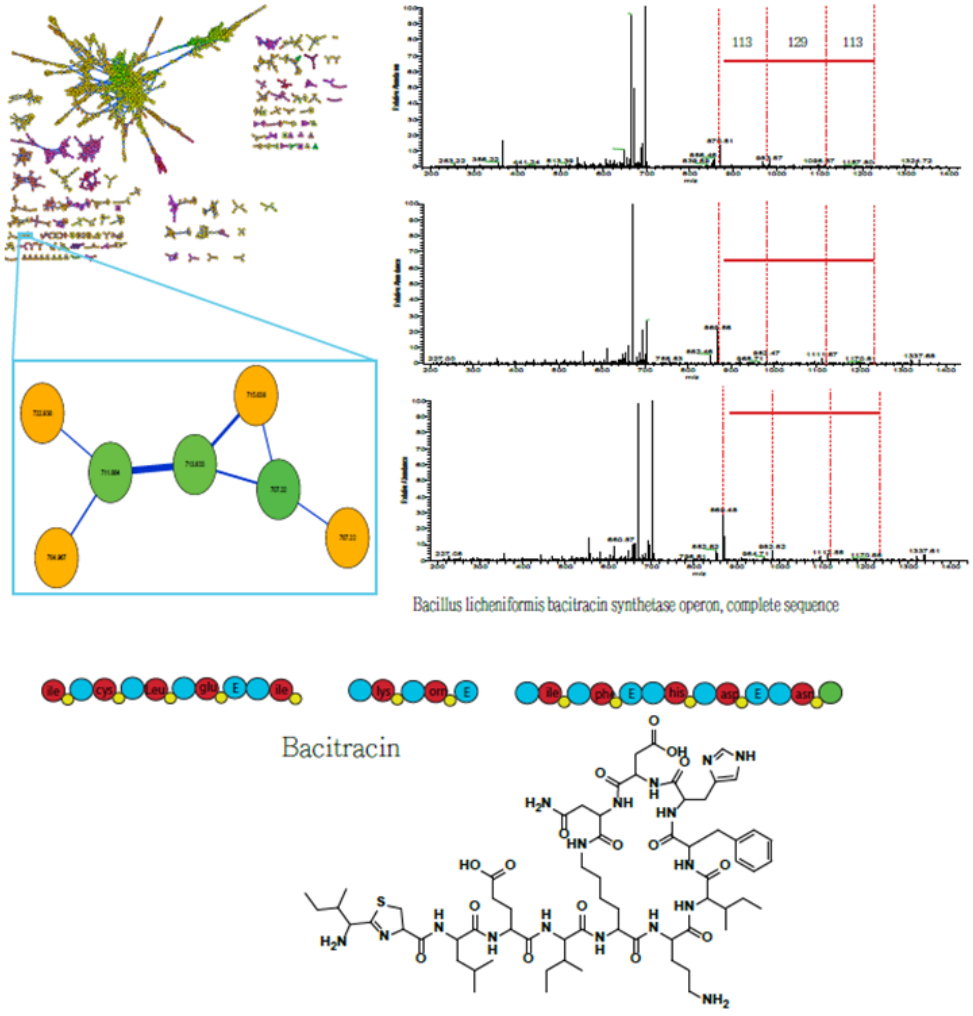

Figure 1.5: Identification of Bacitracin molecule family from *Bacilli* MS/MS networking

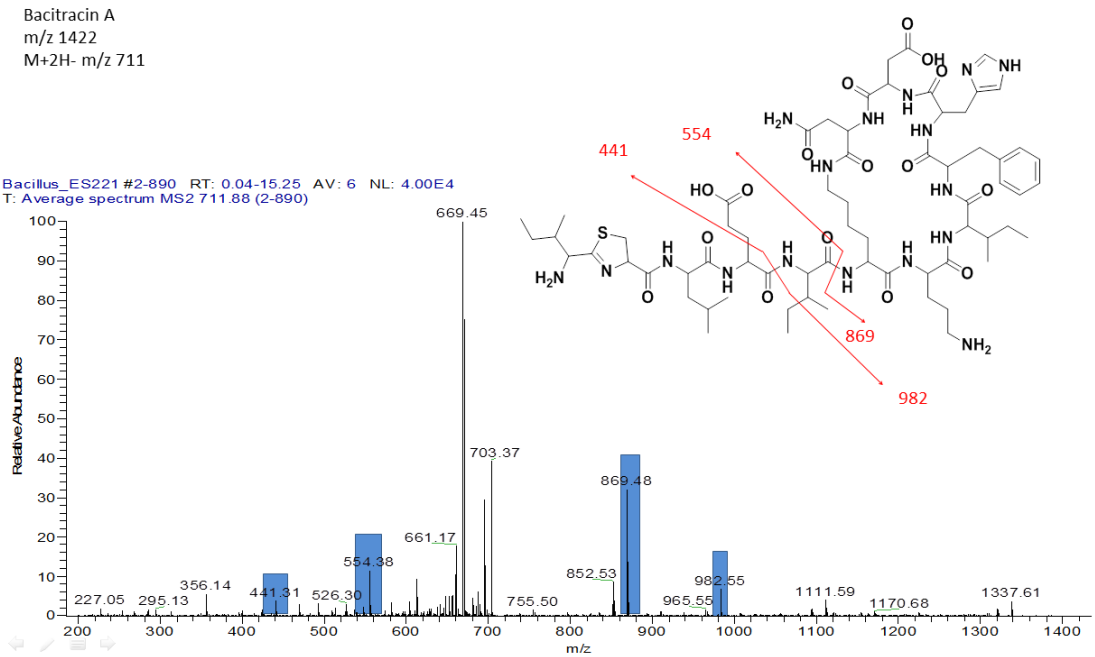

Figure 1.6: Characterization of Bacitracin A from *B. licheniformis ES221*.

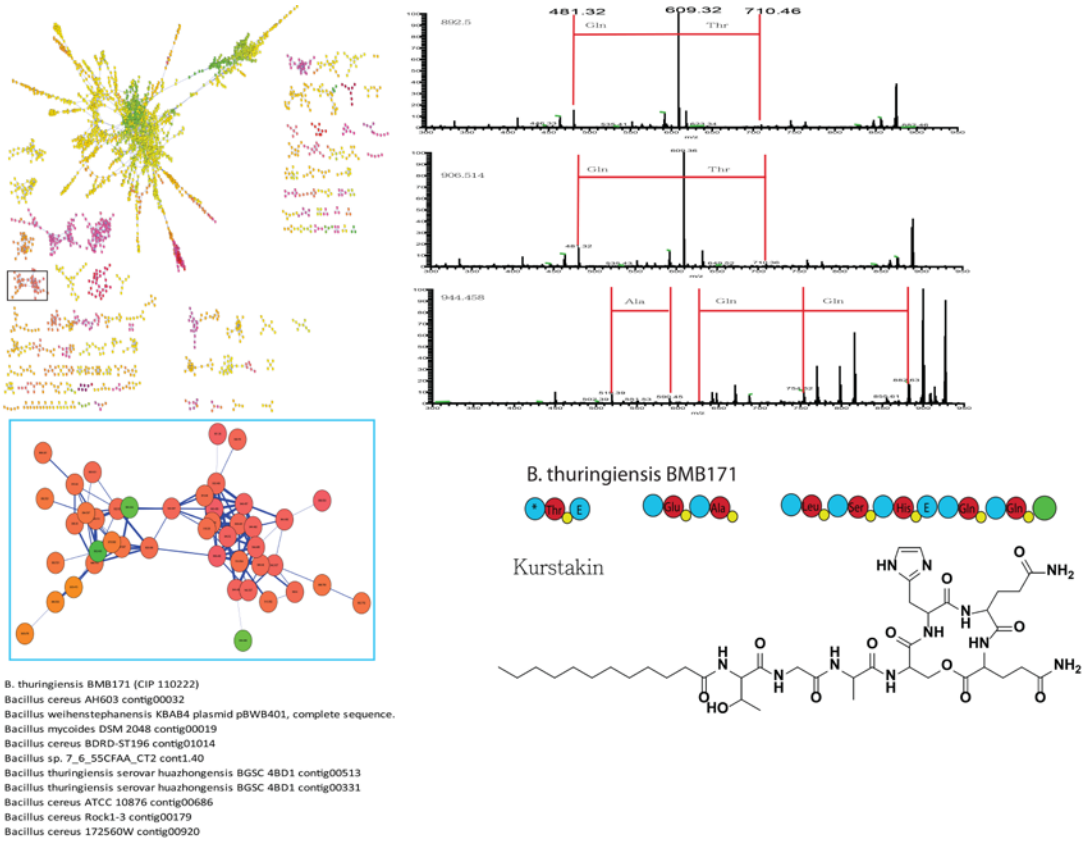

Figure 1.7: Identification of GCF-MF of Kurstakin from *Bacilli*.

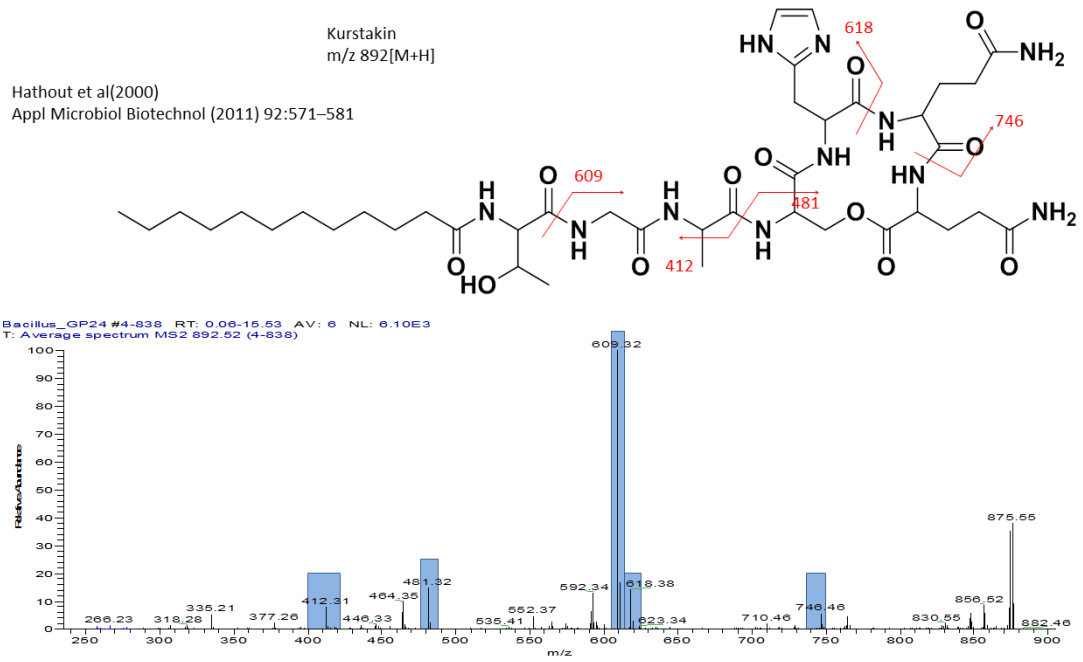

Figure 1.8: Characterization of Kurstakin from *Bacillus cereus GP24.*

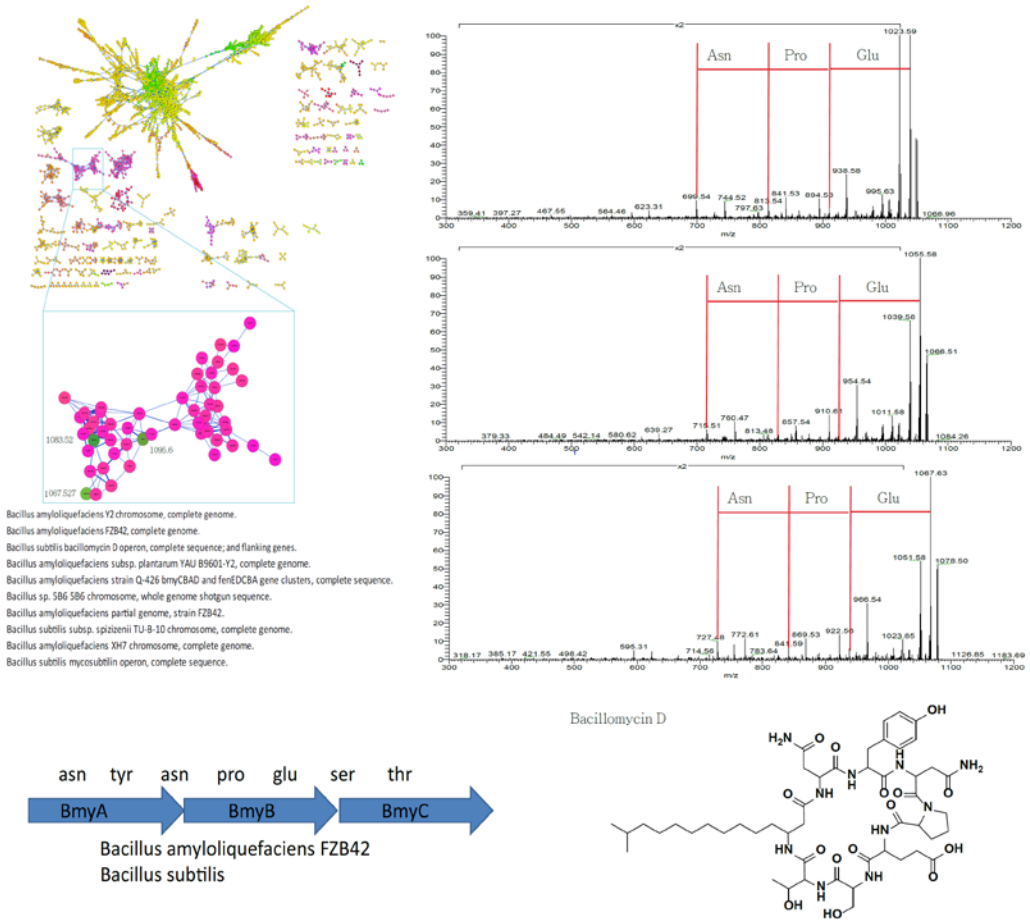

Figure 1.9: Iturin GCF-MF from *Bacilli* networking.

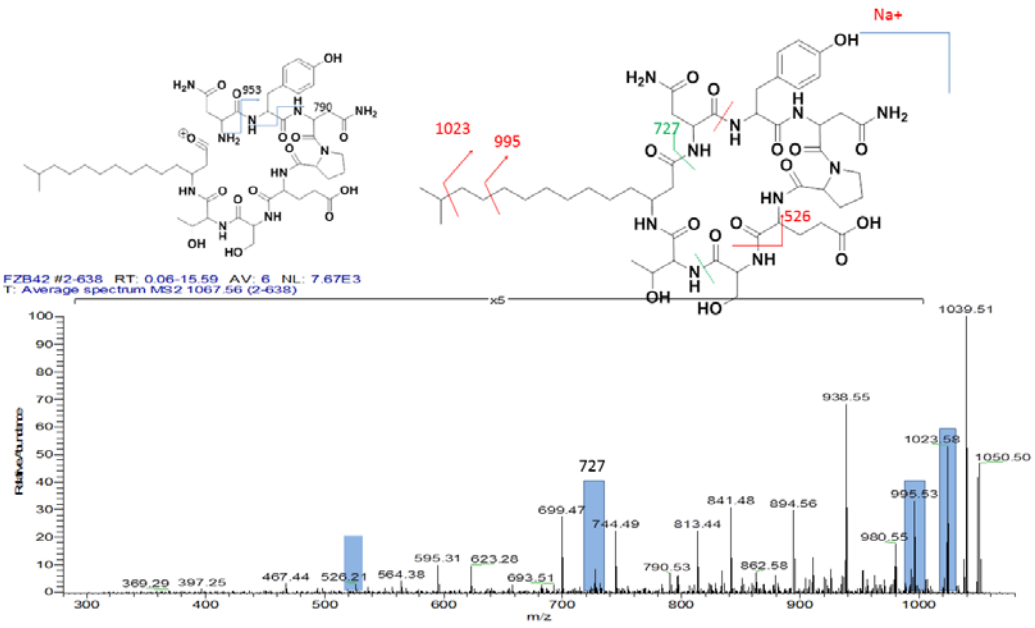

Figure 1.10: Characterization of Bacillomycin D from *B. amyloliquefaciens FZB42.*

Similar results were obtained for pseudomonads. For example there was a 17-node MS/MS cluster ranging from 1133 to 1193 Da that revealed a tag of 87-113-87-113 corresponding to Ser-Leu-Ser-Leu, Ser-Ile-Ser-Leu, Ser-Leu-Ser-Ile, Ser-Ile-Ser-Ile, and reverse sequences (Figure 1.11). This sequence tag from MS/MS data obtained from unsequenced *P. tolaasii, P. putida* and *P. aurantiaca* matched to the gene cluster families only from the Pseudomonads and not the Bacilli. The matches included the Viscosin, WLIP, and Massetolide gene clusters. Therefore both the gene clusters and molecules that

were identified from a GCF-MF pair belong to the viscosin/WLIP and massetolide family of molecules. We confirmed the viscosin cluster by adding MS/MS data from P.putida RW10S2, a known WLIP producer.

4 candidate gene cluster family-molecule pairs were identified from pseudomonads (Figure 1.11-1.15). The two examples above provided the proof-of-principle that genome mining to molecule families and expanding this to all MS/MS clusters with peptidic signatures revealed that 8 out of 107 MS/MS signatures could be correlated to gene cluster families (Figure 1.3).

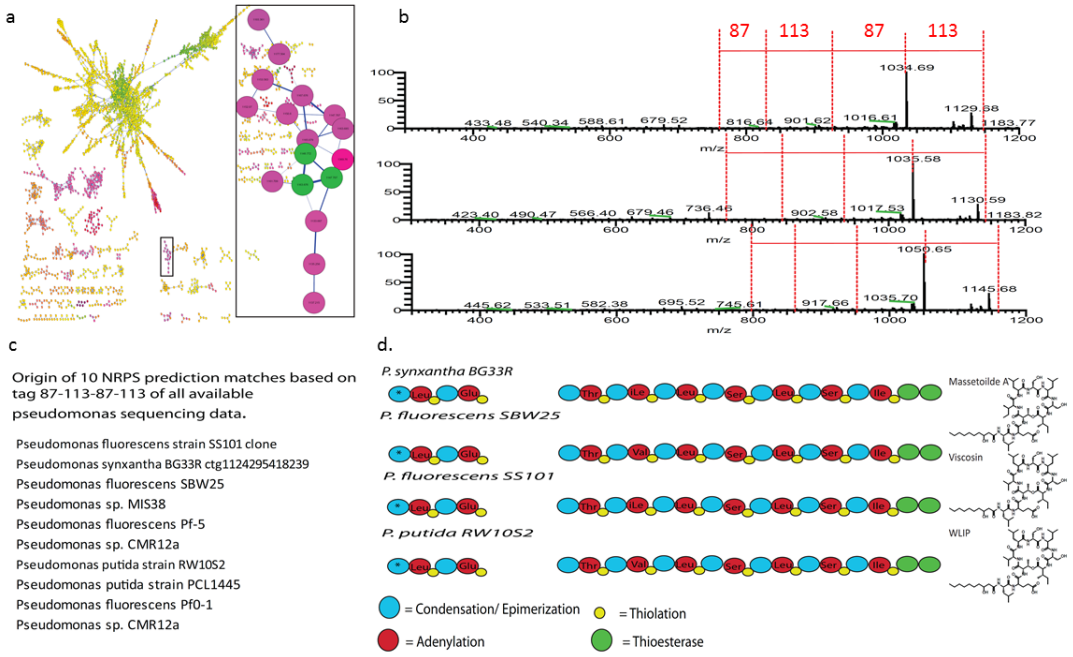

Figure 1.11: Proof of principle from Pseudomonad and idenfication of GCF-MF of viscosin family.

a) Investigation of molecular clusters from Pseudomonad MS/MS networking.

b) Random selection of three spectra and generation of sequence tag: 87-113-87-113

c) Genome mining of all public genome sequence from Pseudomonas

d) Matching sequence tag to all predicted gene clusters from Pseudomonas

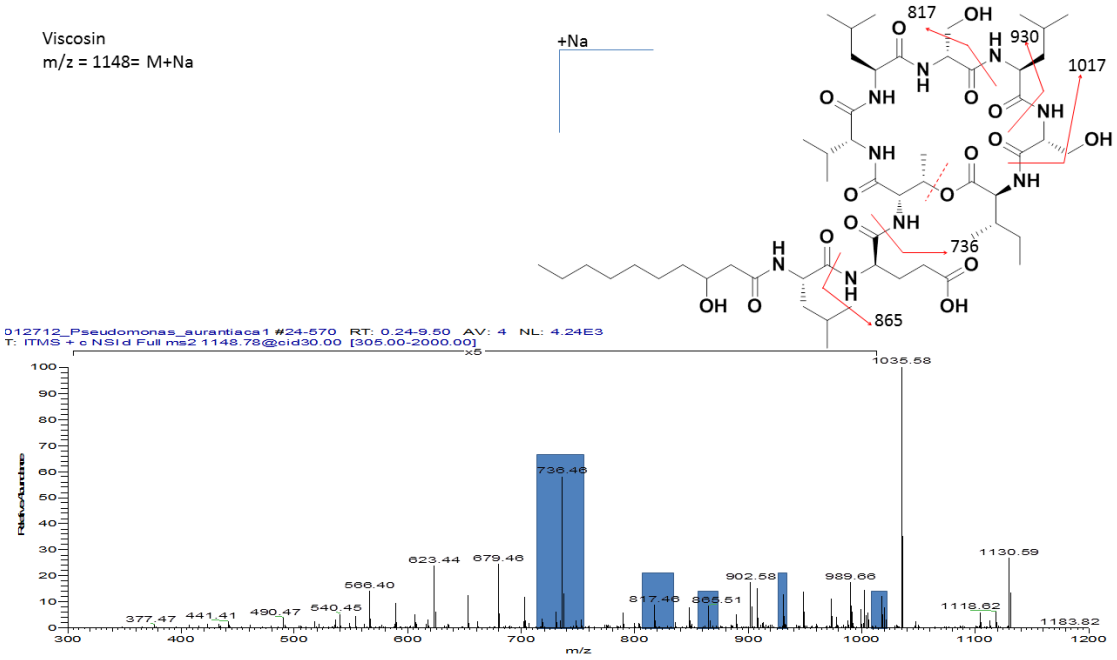

Figure 1.12: Characterization of viscosin from P. aurantiaca.

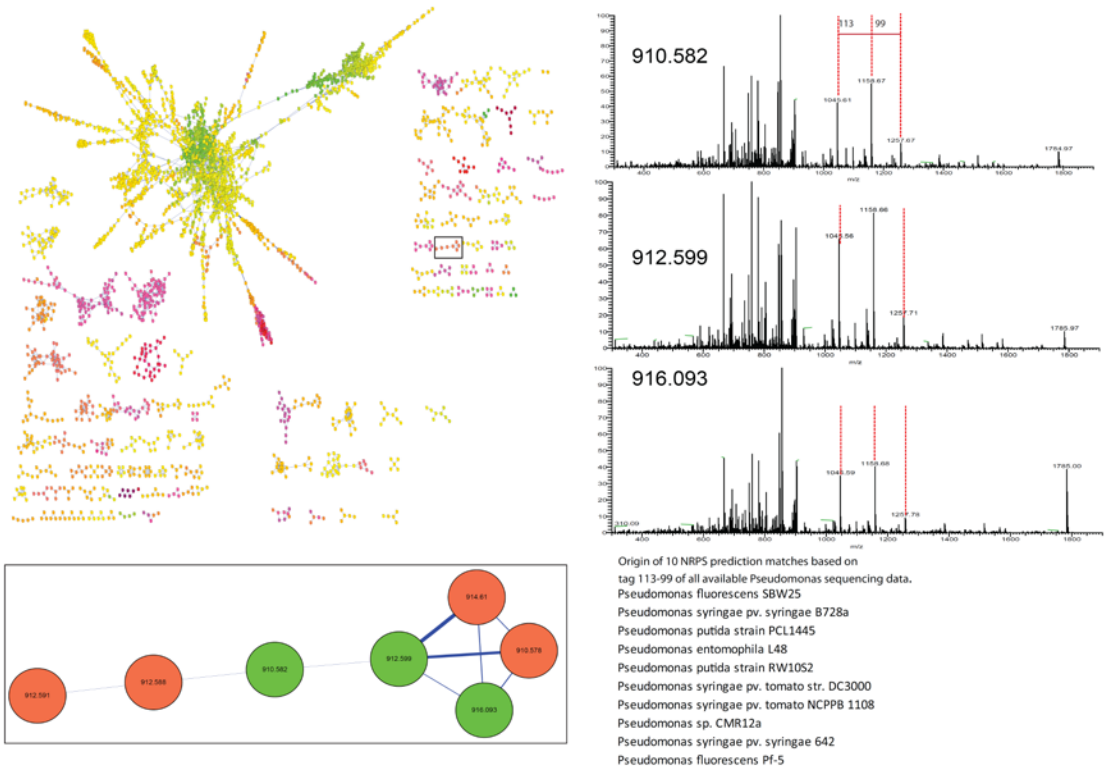

Figure 1.13: Putative GCF-MF of entolysin from Pseudomonas MS/MS networking.

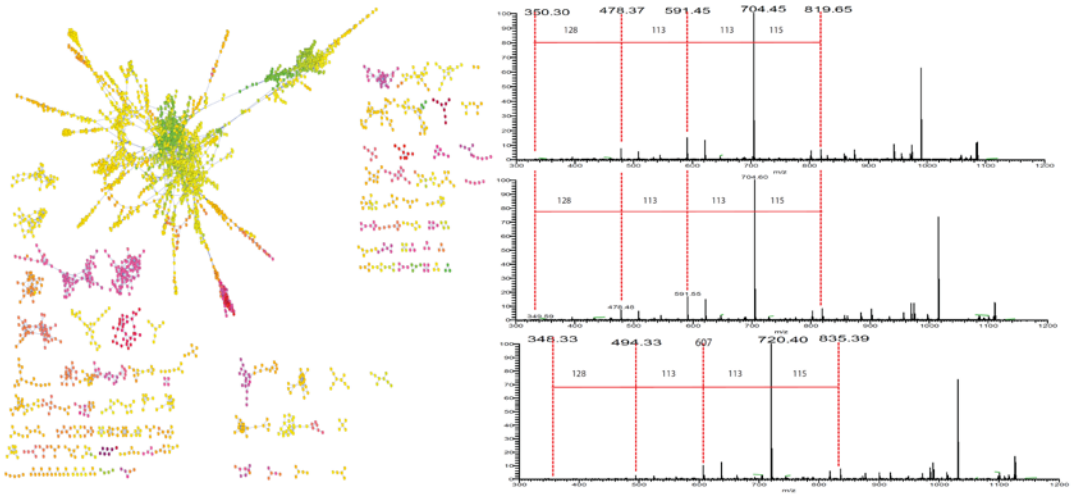

Figure 1.14: putative GCF-MF of amphisin from Pseudomonas MS/MS networking.

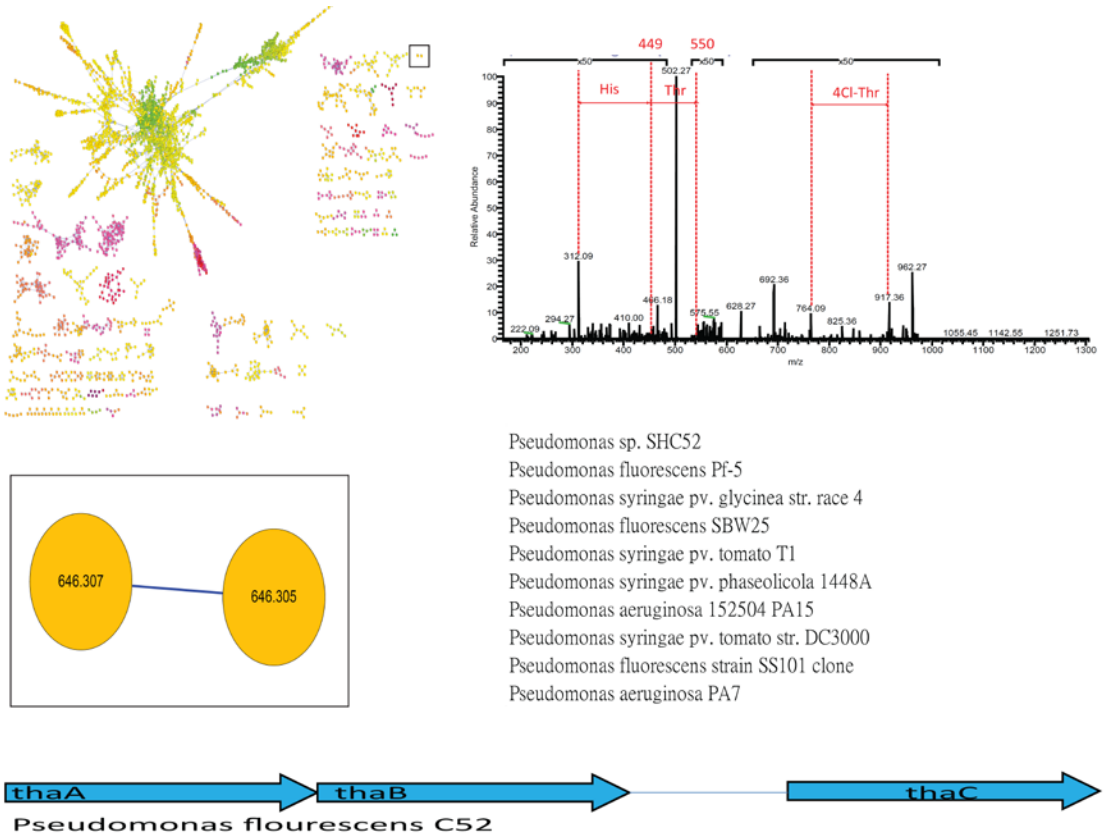

Figure 1.15: Thanamycin GCF-MF from Pseudomonas networking.

When making such correlations there are several outcomes. First, the most confident outcome that perfectly matches both the gene cluster and mass spectrometry data to a known molecule-gene cluster pair already described in the literature. Second, when a molecule is successfully dereplicated to a gene cluster

or family of gene clusters. First, the most confident matches have a perfect match of both the gene cluster and MS data to a known molecule-gene cluster pair described in the literature. Second, when a gene cluster has been found for a molecule and the MS and gene cluster or gene cluster family analysis is in agreement with the molecule to which it was dereplicated to. Third, when a family of gene clusters has been found for a molecule and the MS, MS/MS data, and gene cluster family analysis is in agreement with the molecule to which it was dereplicated to. Finally, there is the possibility to match the gene cluster to a molecule that could not be associated with a molecule already described in the literature using the dereplication strategies we employed, and therefore may be a new molecular entity or the match could may be incorrect. Table 2 highlights the results from this study and the level of confidence. How do we know that these correlations are correct? At this stage we do not know for sure unless additional confirmation is provided. There are many avenues to verify a match and such verifications are not limited to the description here but additional verification include matching the MS/MS data to MS/MS data published in the literature, comparative in silico dereplication to databases such as NORINE, isolation and

NMR analysis of the molecule, creating knock-out's of the gene cluster, or (partially) sequence the cluster to verify that the related gene cluster is indeed present on the genome of the unsequenced organism. Each of these methods, with exception of literature matching, are time consuming, require experts in operating the instrumentation (NMR), interpretation (NMR, sequence analysis), molecular biology (sequence analysis), or using algorithms (sequence analysis) that are not always available in a given lab environment and therefore become costly. It is our opinion that because the costs for further analysis are so significant, that these approaches should only be performed when the molecules is of great interest. MS/MS networking based peptidogenomics analysis can assist in the prioritization of the most interesting molecules. Should additional verification fail, then it is an incorrect identification. Thus far we have not seen such a scenario with our data. Using literature matching to known pseudomonads and bacilli data the gene cluster families of sufactin/lichenmycin, viscosin, thanamycin, kurstakin, bacitracin were successfully paired with their respective molecular families. GCF-MF pairs that could not be verified by the literature should be considered a putative GCF-MF pair at this time. GCF-MF pairs that could not be verified by the

literature should be considered a putative GCF-MF pair at this time. The others, where a literature match was not found, should be considered a putative GCF-MF pair at this time. Having found 8 gene cluster family-molecule family pairs support the notion that molecular data can be correlated to sequencing data in the public domain and that this provides significant structural information. Even though the above analysis was not an exhaustive search, the data provided the demonstration-of-principle that the methodology works and provides the foundation for future algorithmic development.[26]

An application of the networking is dereplication. We take five pseudomonas strains that isolated from salmon eggs as example. Salomon, common name of fish of family Salmonidae, not only lives around North Atlantic and Pacific Oceans but also produced in aquaculture for many countries. Basically, the life cycle of salmon is that they are born in fresh water, migrate to sea, and then return to fresh water. During this cycle, they might come across lots of harsh environment so that only small amount of salmon return to fresh water and reproduce next generation. One of the survival strategies is that remaining salmon produce large quantities of eggs so as to produce offspring. However, not all eggs are able to hatch out into a

tiny salmon because some eggs might be infected by microbes. Among these pathogenic bacteria, pseudomonas fluorescens was the predominant one that associates with salmon egg. In this work, we collected metabolites from live colonies of *749P17, 501, 734P3, 746P18, 746P2* which evolving from common ancestor and are closed species to *Pseudomonas fluorescens SS101* and *Pseudomonas synxantha BG33R*, viscosin and massteolide producers that belong to viscosin molecular family. Therefore, we used MALDI-TOF to detect metabolites collected from live colonies for the five strains as we did for *Bacillus* and *Pseudomonas* so as to observe chemical profiling and confirm no cross containment happened in the NanoDESI-MS (figure 1.15-1.29). Since many molecules have been identified from *Bacillus-Pseudomonas* networking, rhamnolipid, phenazine, quinolone, thanamycin, entylosin, amphisin, viscosin molecule families are known in the networking, we apply this information to help us identify metabolites from these five strains. Consequently, it is viscosin molecule family that clusters with main metabolites, m/z 1100-1200, as shown in figure 1.11. The result suggests similar prediction as phylogenetic tree. Once molecular networking becomes more mature, we can use to study closed species

and compare similarity or difference in order to all metabolites besides Nonribosomal peptides.

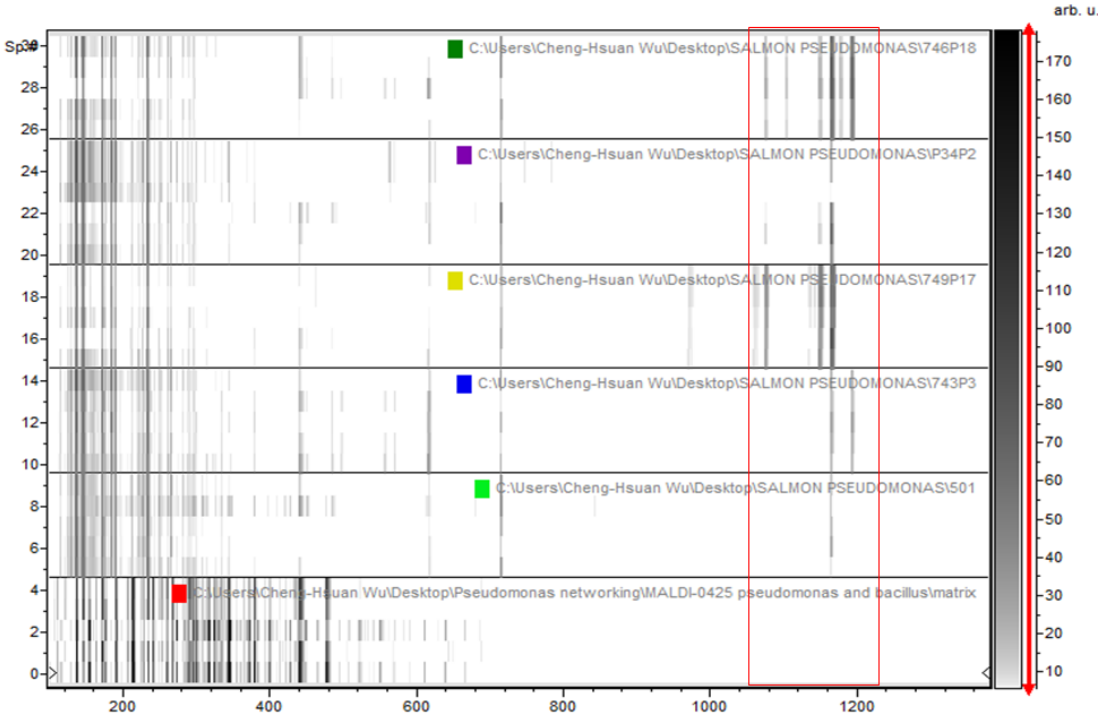

Figure 1.16: Heat map profiling of *501, 746P18, P34P2, 749P17, P43P3*: It shows that they have dominant peaks around m/z 1100-1200.

Psudomonas heat map Profile

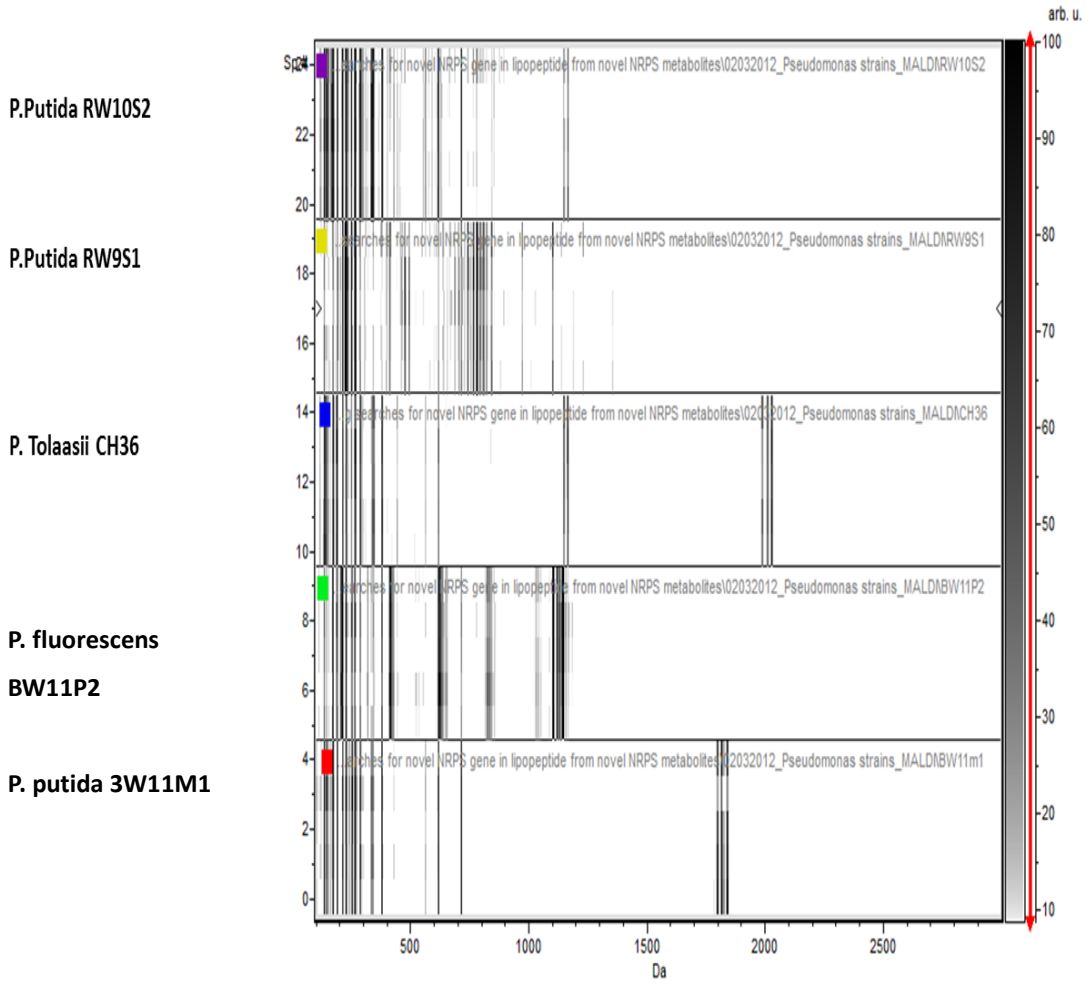

Fig 1.17: Heat map profiling of *Pseudomonas putida RW10S2, Pseudomonas putida RW9S1, Pseudomonas tolaasii CH36, Pseudomonas fluorescens BW11P2, Pseudomonas putida BW11M1.*

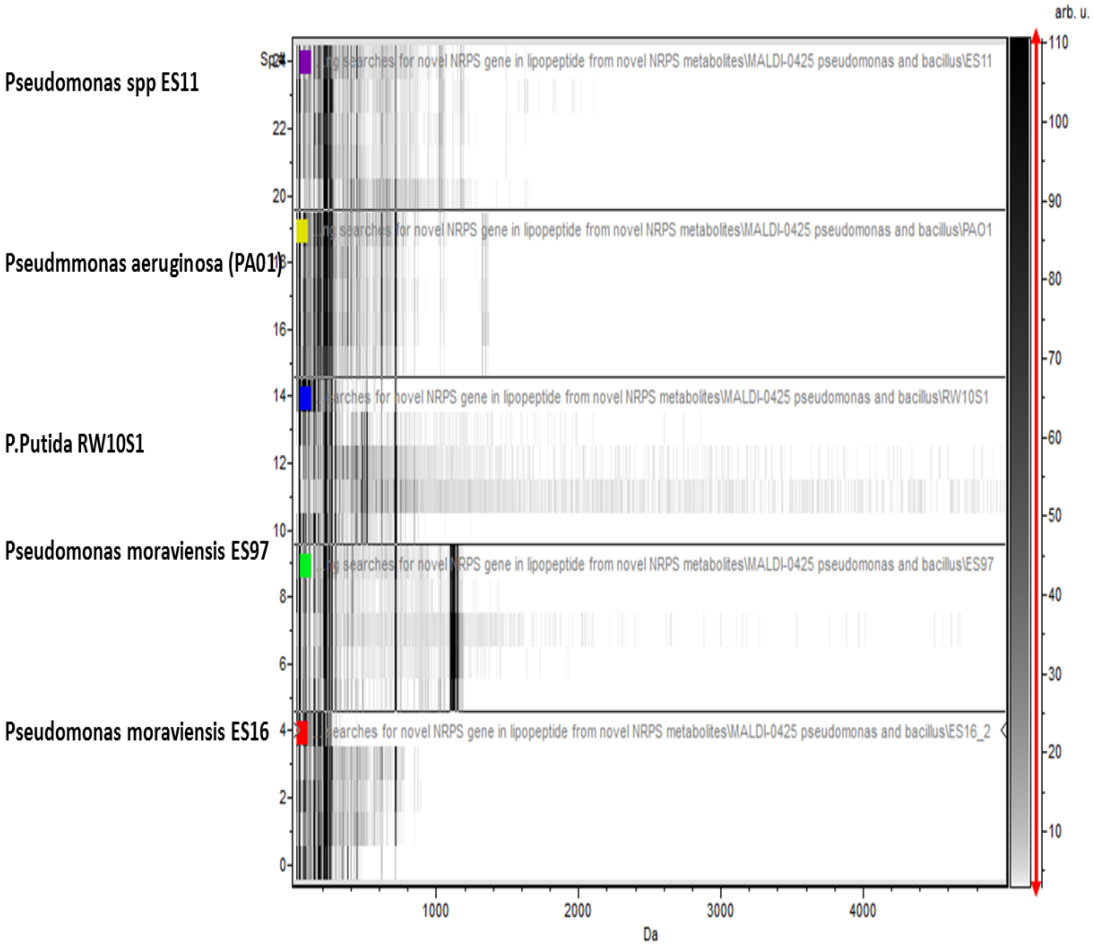

Fig 1.18: Heat map profiling of *Pseudomonas spp ES11, Pseudomonas aeruginosa (PAO1), Pseudomonas putida RW10S1, Pseudomonas moraciensis ES97, pseudomonas moraviensis ES16*.

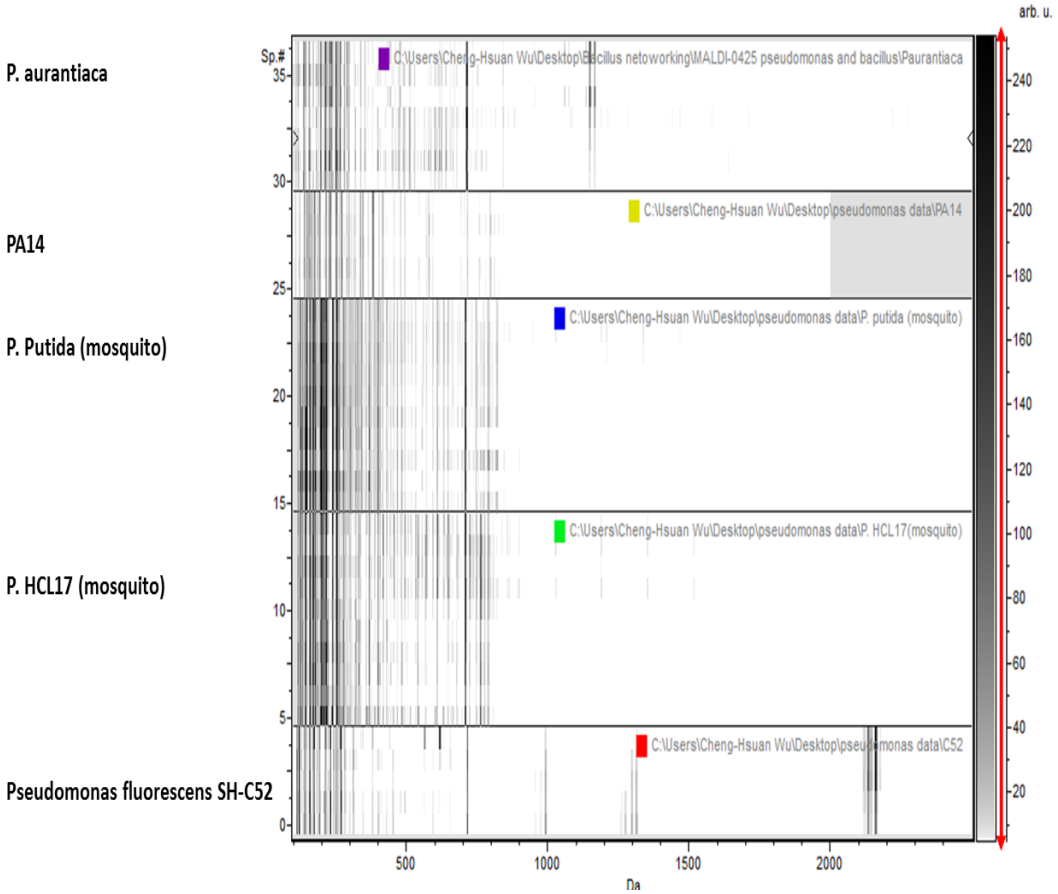

Fig 1.19: Heat map profiling of *Pseudomonas aurantiaca* , *Pseudomonas aeruginosa (PA14), Pseudomonas putida putida* (from mosquito guts), *Pseudomonas HCL17* (from mosquito guts), *pseudomonas fluorescens SH-C52*

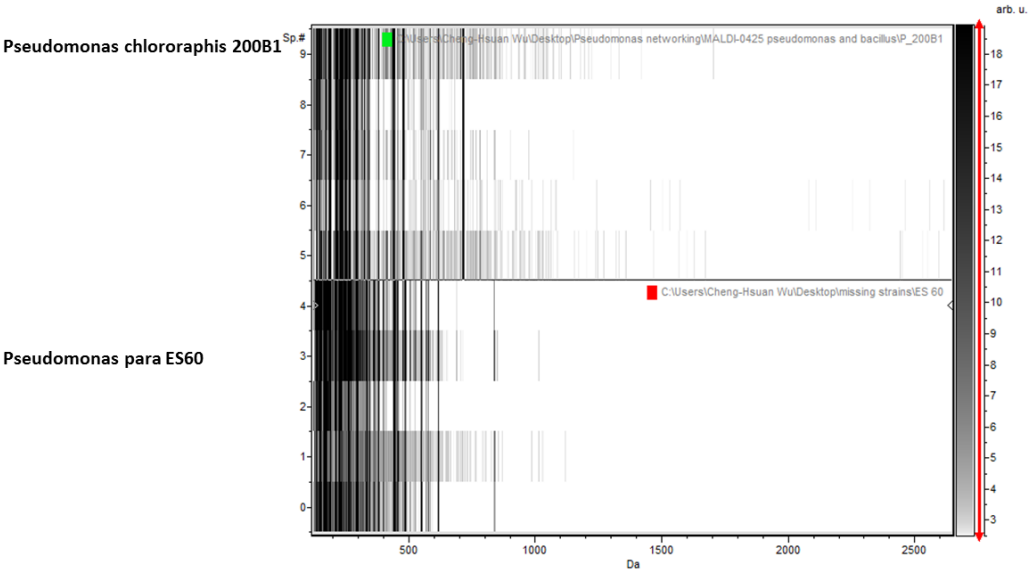

Fig 1.20: Heat map profiling of *Pseudomonas chloroaphis 200B1* , *Pseudomonas para ES60*.

Heat map profile of *Bacillus*

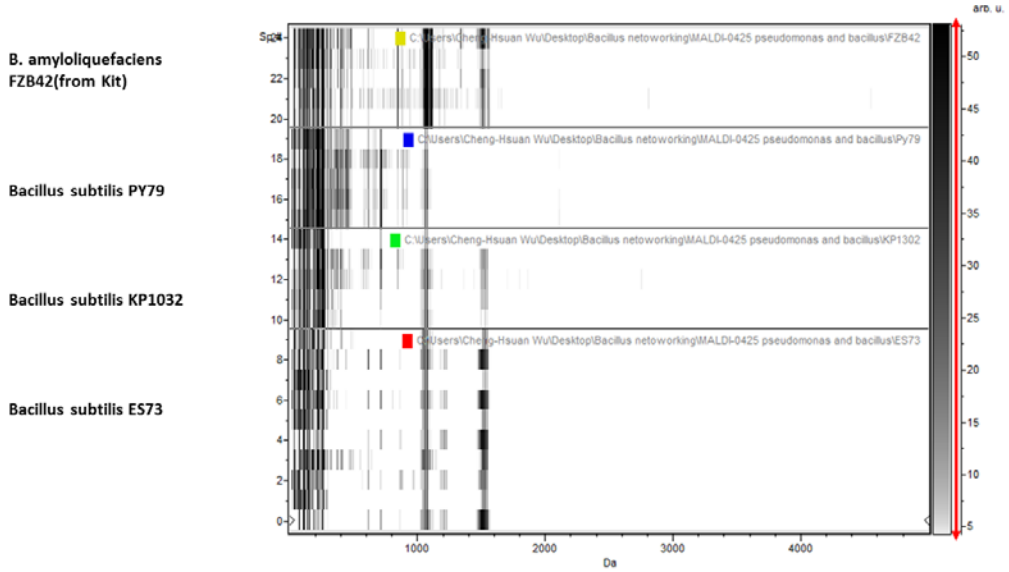

Fig 1.21: Heat profiling of *B. amyloliquefaciens FZB42, Bacillus subtilis PY79, Bacillus subtilis KP1032, Bacillus subtilis ES73*.

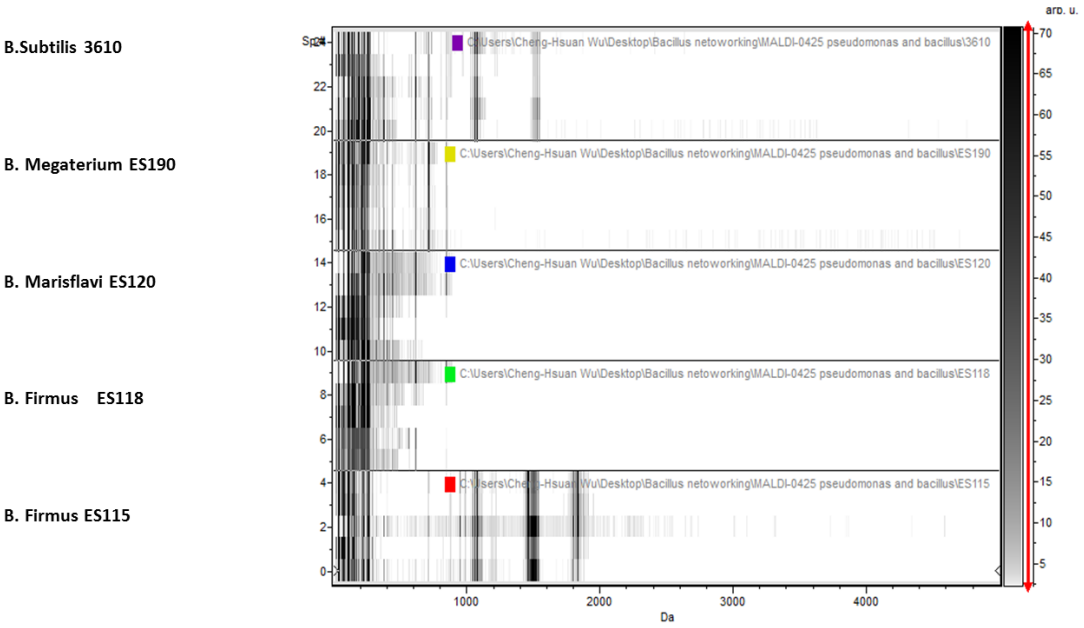

Figure 1.22: Heat profiling of *Bacillus subtilis 3610, Bacillus Megaterium ES190, Bacillus marisflavi ES120, Bacillus firmus ES118, B. firmus ES115.*

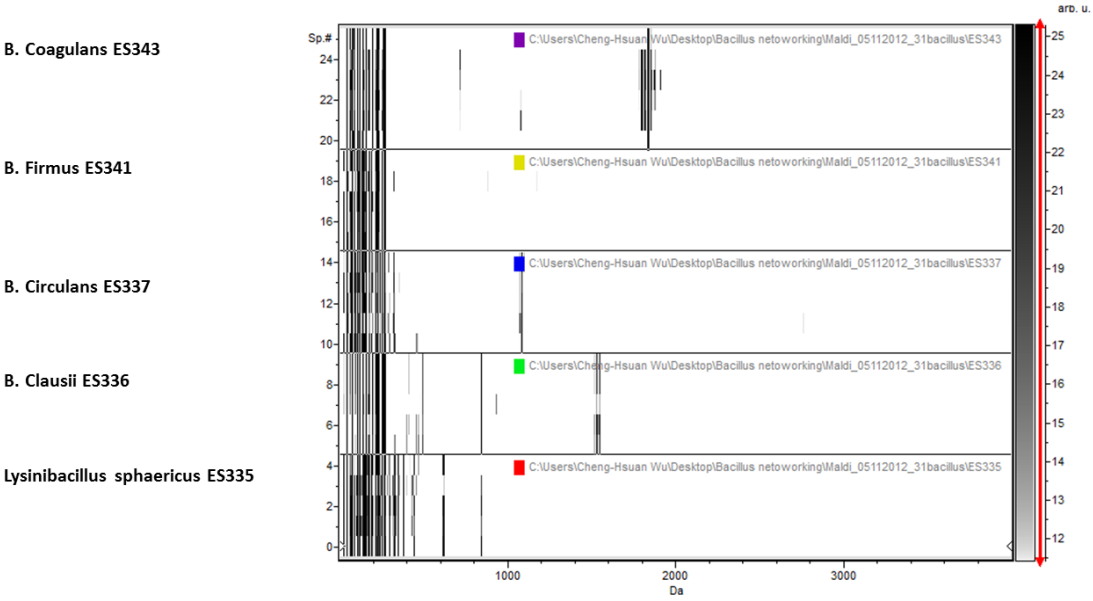

Figure 1.23: Heat profiling of *Bacillus coagulans ES343, Bacillus firmus ES341, Bacillus circulans ES337, Bacillus clausii ES335, Lysinibacillus sphaericus ES335.*

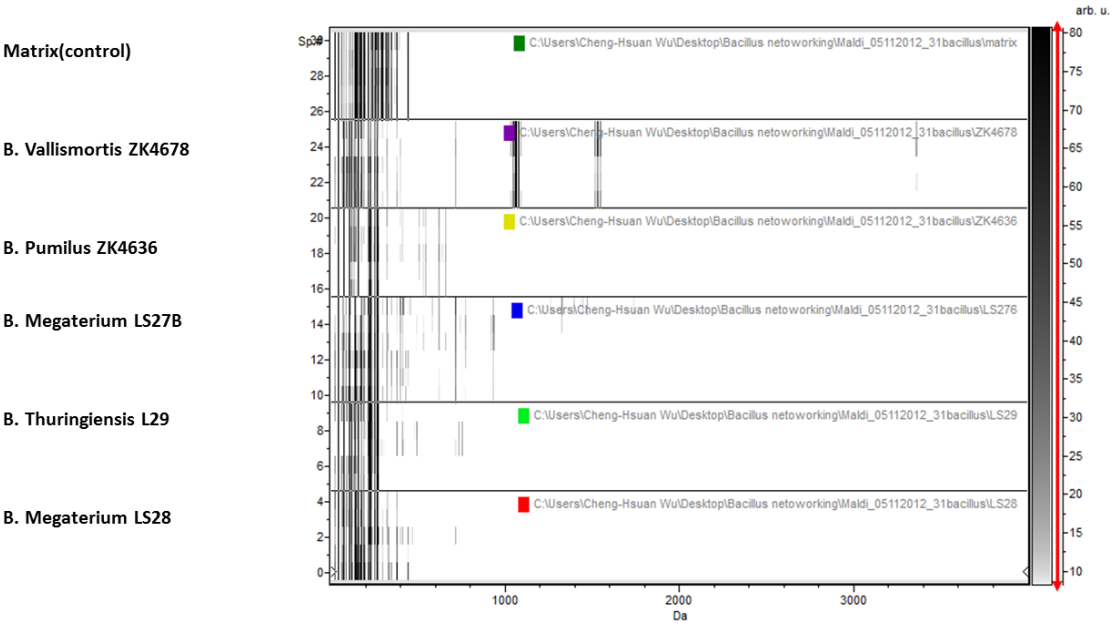

Figure 1.24: Heat profiling of Control , *Bacillus vallismortis ZK4678, Bacillus pumilus ZK4636, Bacillus megaterium LS27B, Bacillus Thuringiensis LS29, Bacillus Megaterium LS28.*

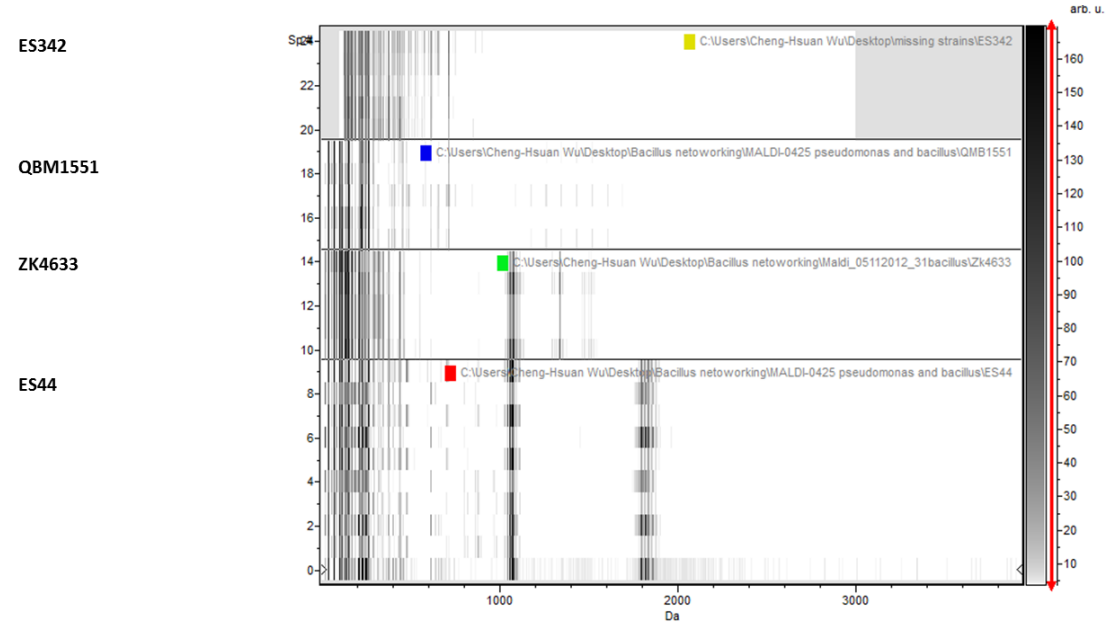

Figure 1.25: Heat profiling of Control , *B. lentus, B. megaterium QBM1551, B. amyloliquefaciens ZK4633, B. licheniformis ES44*.

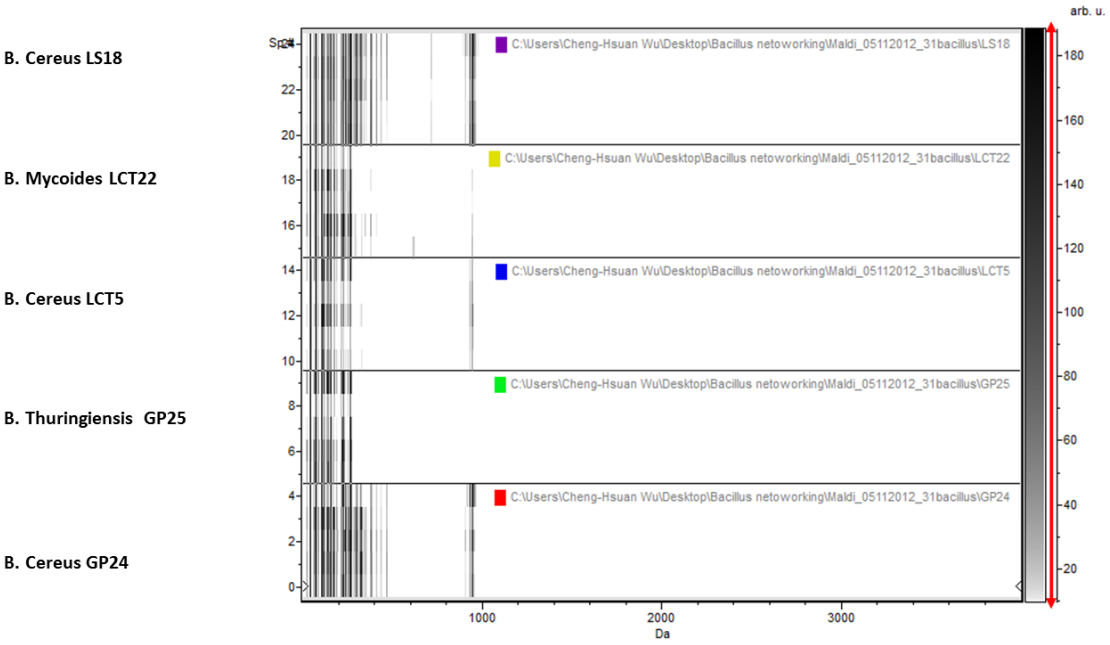

Figure 1.26: Heat profiling of *B. cereus LS18* , *B. mycoides LCT22, B. cereus LCT5, B. thuringiensis GP25, B. cereus GP24*.

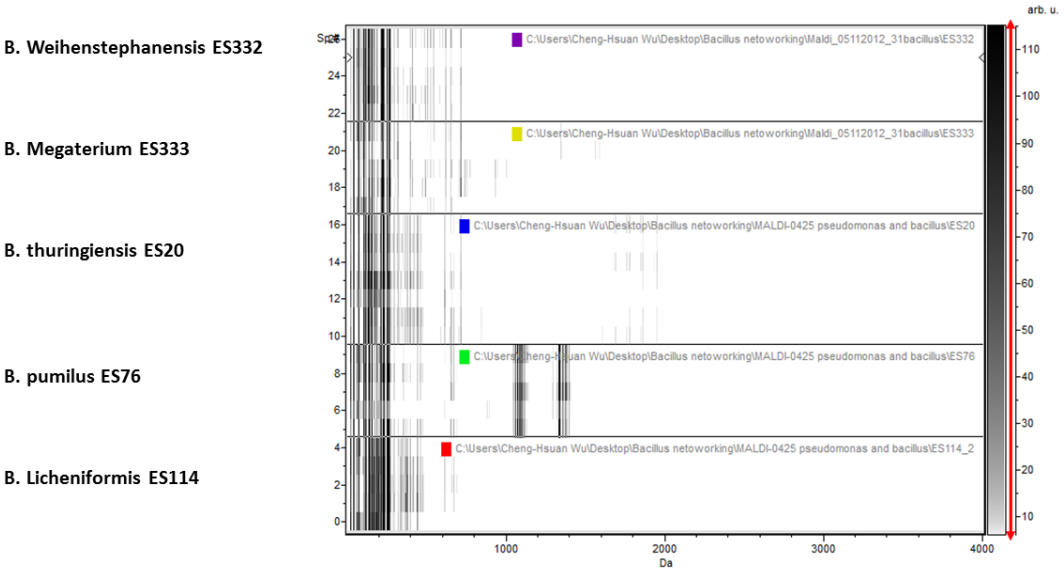

Figure 1.27: Heat profiling of *B. weihenstephanensis ES332, B. megaterium ES333, B. thuringiensis ES20, B. pumilus ES76, B. licheniformis ES114.*

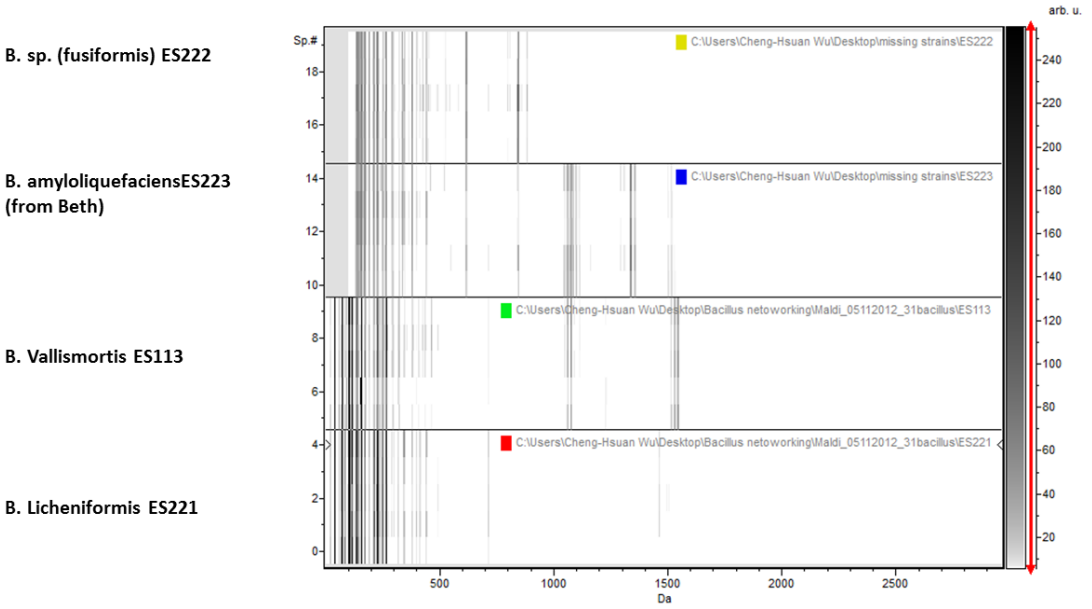

Figure 1.28: Heat profiling of *B. sp. (fusiformis) ES222, B. amyloliquefaciens FZB42 ES223 (from Beth), B. vallismortis ES113, B. licheniformis ES221.*

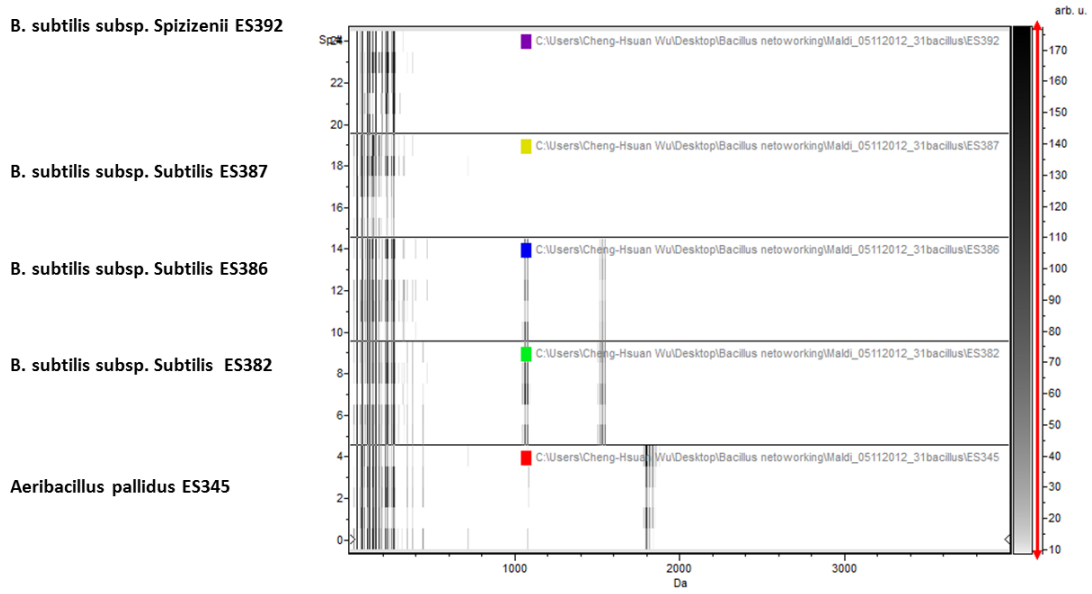

Figure 1.29: Heat profiling of *B. subtilis subsp. Spizizenii ES392, B. subtilis subsp. Subtilis ES387, B. subtilis subsp. Subtilis ES386, B. subtilis subsp.*

***Subtilis ES382, Aeribacillus pallidus ES345*.**

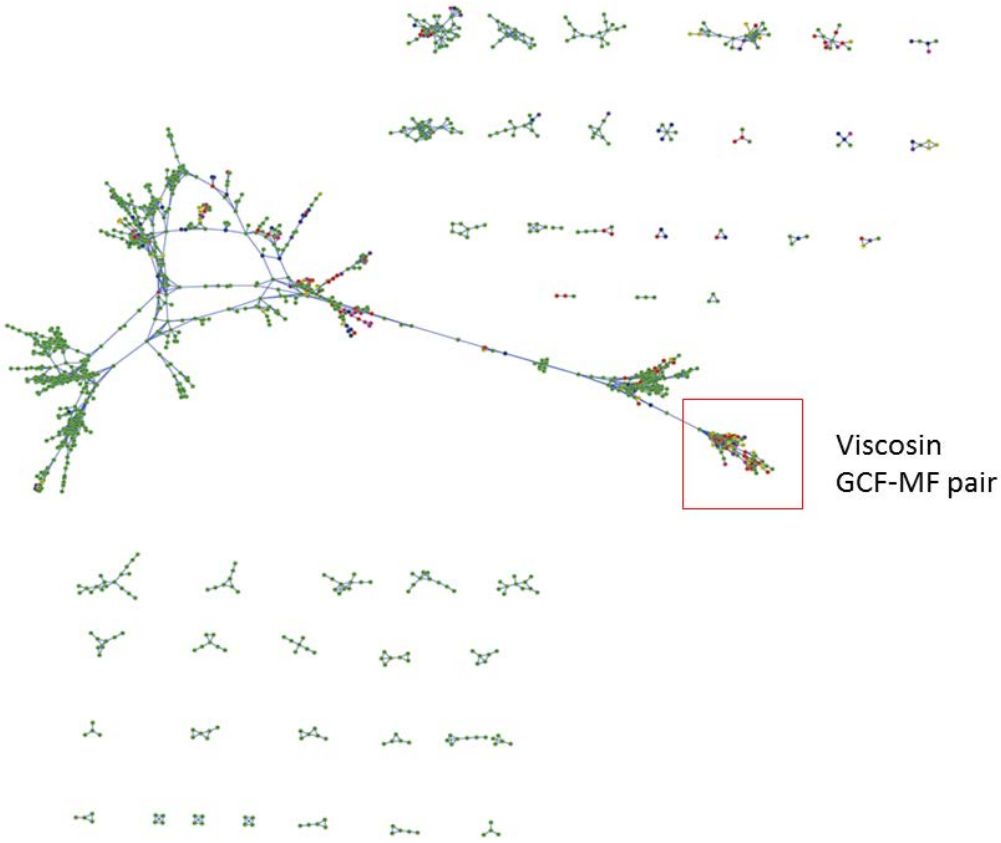

Figure 1.30: Identification of molecules from salmon egg pseudomonas through 17-Pseudomonas networking dataset.

1.3 Conclusion:

One of the major bottlenecks in genome mining is that in order to connect a molecule to their biosynthetic signature it takes significant amount of time and therefore is costly. Perhaps such an approach or related approaches may become the first step in the molecular characterization of unsequenced microbes. NanoDESI based MS/MS networking, in combination with peptidogenomics of unsequenced organisms, reveals that there are gene cluster-molecule families that exist and that this can be subjected to peptidogenomic based genome mining using sequencing information available in the public databases. It is anticipated that through an adjustment of the protocol, where metagenomic data is used as the sequencing data that is searched, there is the potential that one can apply related approaches to carry out genome mining of molecules directly analyzed from environmental and personal microbial communities such as the ones found on our skin, in our gut, in soil, coral reefs or on roots of plants-something not commonly attempted with today's genome mining technologies.

1.4 Future direction:

With integration of peptidogenomic and molecular networking, 8 GCF-MF pairs were set up according to this norm. However, there are still two sequence tags that are unable to match up any gene cluster. This indicates the two molecular clusters could be new metabolites or no available genomic data (Table 1.4). In the future, we should follow up to characterize them and further investigate to build GCF-MF pair.

1.5 experimental section

Bacterial strains and culture conditions.

All of the bacterial strains used in this study are listed in the table. In this table, we have 42 *Bacilli* and 18 *Pseudomonas* strains. Some of them are sequenced organism, some are partial sequenced, and most of them are unsequenced strains. LB broth was purchased from fisher sciencist. ISP2 agar medium was prepared with 2g yeast, 5g malt, 2g dextrose, 10g agar/ 500ml. All strains were cultured in the LB both for 24hr in the 28℃ shake and then transferred to ISP2 agar medium to grow for 48hr in 30℃ incubator. (See table 1.1)

Sample preparation for Matrix-assisted laser ionization (MALDI)

After 48hr all strains grow on the agar medium individually. Enough amounts (approximately 1ul) of metabolites were scraped from live colony of each strains and transferred to MSP 96 MALDI anchor plate. These metabolites were then covered with 1 µL saturated matrix solution, 35mg/ml-cyano-4hydroxycinamic acid, and 15mg DHB saturated 78% acetonitrile and 1% formic acid (universal matrix) until formation of proper crystal. The MALDI plate was inserted into an Autoflex Bruker Daltonics mass spectrometer. Data was recorded in refractron positive mode. With the help of ClinproTool, we are able to process collected data and analyze chemical profile of each strain simultaneously.

NanoDESI MS

Transfer 2 ul LB liquid medium of liquid culture of each strain to ISP2 agar medium and innoculate four colonies (0.5ul for each) for individual strain in every IP2 plates for 48hr. The solvent for NanoDESI was Acetonitrile : H2O 65:35 with 0.05% Formic acid. NanoDESI MS collect data-dependent data from the every

strains around 10 to 20 min according to condition of each strains.

Generate molecular networking by cytoscape

The generation of MS/MS networking (42 *Bacilli* and 17 *Pseudomonas*) was clustered base on the procedure described by Watrous et al. All algorithms assumed fragment mass tolerance of 0.3Da and the cosine cut-off of the networking is 0.7. In addition, two plugins were used to process the crude networking in order to remove unnecessary nodes. One plug-in, FM3, was used to modify the look of networking so that more separated clusters can be visualized. The other one, Hider slider, has a function to delete unnecessary nodes so that we can visualize data only come from pseudomonas or Bacillus alone.

Bioinformatic tools

We gathered almost available genome sequence from Nation Center for Biotechnology Information (NCBI), DOE Joint Genome institute (JGI) and PseudoDB. Targeted nucleotide sequence were subject to ANTISMASH, NP Searcher, NRPS Predictor2 to gain amino acid sequence of those targeted gene cluster

Table 1.1 : List of *pseudomonas* and *bacillus* in the study

Pseudomonas				
Sequenced	unsequenced	Name	source	from
	X	Pseudomonas aurantiaca	plant/soil	Harold Gross
X(did not show in the networking)		Psuedomonas putida RW10S2	plant/soil	René De Mot
	X	Pseudomonas tolaasii CH36	plant/soil	René De Mot
partial		Pseudomonas putida RW10S1	plant/soil	René De Mot
	X	Pseudomonas fluorescens BW11P2	plant/soil	René De Mot
	X	Pseudomonas putida BW11M1	plant/soil	René De Mot

Table 1.1 : continued

Sequenced	unsequenced	Name	source	from
	X	*Pseudomonas florescens RW9S1*	plant/soil	René De Mot
X		*Pseudomonas aeruginosa(MPAO1)*	human	Venassa
X		*Pseudomonas aeruginosa(MPA14)*	human	Venassa
	X	*Pseudomonas chlororaphis 200B1*	plant/soil	Kit's lab
	X	*Pseudomonas moraviensis ES97*	plant/soil	Kit's lab
	X	*Pseudomonas moraviensis ES16*	plant/soil	Kit's lab
	X	*Pseudomonas spp ES11*	plant/soil	Kit's lab

Table 1.1 : continued

Sequenced	Sequenced	Sequenced	Sequenced	Sequenced
	X	*Pseudomonas para ES60*	plant/soil	kit's lab
partial		*Pseudomonas flourescens C52*	plant/soil	Jeramie
	X	*Pseudomonas fluorescens Tn5*	plant/soil	charlottle
	X	*Psuedomonas putida*	mosquito	
	X	*Pseudomonas HCL17*	mosquito	

Bacilli table

Sequenced	unsequenced	Name		source	from
	X	*QMB1551*	*B. megaterium*	soil	Kit' lab
	X	*ES-190*	*B. megaterium*	soil	Kit' lab

Table 1.1: continued

Sequenced	unsequenced	*Name*	*source*	from	Sequenced
X		*FZB42*	*B. amyloliquefaciens*	soil	Kit' lab
	X	*ES-20*	*B. thuringiensis*	soil	Kit' lab
	X	*ES-118*	*B. firmus*	soil	Kit' lab
	X	*ES-115*	*B. firmus*	soil	Kit' lab
	X	*ES-73*	*B. subtilis*	soil	Kit' lab
X		*3610*	*B. subtilis*	soil	Kit' lab
	X	*KP1302*	*B. subtilis*	soil	Kit' lab
X		*PY79*	*B. subtilis*	soil	Kit' lab
	X	*ES-114*	*B. licheniformis*	soil	Kit' lab
	X	*ES-44*	*B. licheniformis*	soil	Kit' lab
	X	*ES-120*	*B. marisflavi*	soil	Kit' lab
	X	*ES-76*	*B. pumilus*	soil	Kit' lab
	X	*ES113*	*B. vallismortis*	soil	Beth

Table 1.1: continued

Sequenced	unsequenced	*Name*	*source*	from	Sequenced
	X	*ES221*	*B. licheniformis*	soil	Beth
	X	*ES222*	*B. sp. (fusiformis)*	soil	Beth
X		*ES223*	*B. amyloliquefaciens FZB42*	soil	Beth
	X	*ES332*	*B. lihenstephanensis*	soil	Beth
	X	*ES333*	*B. megaterium*	soil	Beth
	X	*ES335*	*Lysinibacillus sphaericus*	soil	Beth
	X	*ES336*	*B. clausii*	soil	Beth
	X	*ES337*	*B. circulans*	soil	Beth
	X	*ES341*	*B. firmus*	soil	Beth
	X	*ES342*	*B. lentus*	soil	Beth

Table 1.1 : continued

Sequenced	unsequenced	Name	source	from	Sequenced
	X	*ES343*	*B. coagulans*	soil	Beth
	X	*ES345*	*Aeribacillus pallidus*	soil	Beth
	X	*ES382*	*B. subtilis subsp. subtilis*	soil	Beth
	X	*ES386*	*B. subtilis subsp. Subtilis*	soil	Beth
	X	*ES387*	*B. subtilis subsp. subtilis*	soil	Beth
	X	*ES392*	*B. subtilis subsp. spizizenii*	soil	Beth
	X	*ZK4633*	*B. amyloliquefaciens*	soil	Beth
	X	*ZK4636*	*B. pumilus*	soil	Beth

Table 1.1 : continued

Sequenced	unsequenced	Name	source	from	Sequenced
	X	*ZK4678*	*B. vallismortis*	soil	Beth
	X	*GP24*	*B. cereus*	soil	Beth
	X	*GP25*	*B. thuringiensis*	soil	Beth
	X	*LCT22*	*B. mycoides*	soil	Beth
	X	*LCT5*	*B. cereus*	soil	Beth
	X	*LS18*	*B. cereus*	soil	Beth
	X	*LS27B*	*B. megaterium*	soil	Beth
	X	*LS28*	*B. megaterium*	soil	Beth
	X	*LS29*	*B. thuringiensis*	soil	Beth

Table 1.2 : Mass shift of proteinogenic amino acids used for sequence tag identification

Pseudomonas				
mass shift	tag	organism	accession number	predicted sequence
87-113-87-113	ser_leu_ser_leu	*Pseudomonas fluorescens strain SS101 clone 2*	EU199081.2	thr_ile_leu_ser_leu_ser_ile
		Pseudomonas synxantha BG33R ctg1124295418239	NZ_AHPP01000001.1	thr_ile_leu_ser_leu_ser_ile

Table 1.2 : continued

mass shift	tag	organism	accession number	predicted sequence
		Pseudomonas fluorescens SBW25	NC_012660.1	lys_orn_ser_thr_val_leu_ser_leu_ser_ile_ser_lys_gly_orn
		Pseudomonas sp. MIS38	AB107223.1	leu_asp_thr_leu_leu_ser_leu_ser_ile_ile_asp
		Pseudomonas fluorescens Pf-5	NC_004129.6	abu_thr_ile_leu_ser_leu_asp_leu_leu_ser_hyv-d
		Pseudomonas sp. CMR12a	JQ309921.1	leu_leu_ser_hyv-d_thr_hyv-d_leu_ser_leu_asp
		Pseudomonas putida strain RW10S2	JN982333.1	thr_val_leu_ser_leu_ser_ile

Table 1.2 : continued

mass shift	tag	organism	accession number	predicted sequence
		Pseudomonas putida strain PCL1445	DQ151887.2	leu_asp_leu_leu_gln_ser_val_leu_ser_leu_val_ser
		Pseudomonas fluorescens Pf0-1	NC_007492.2	leu_leu_ser_ile_gln_ile_leu_gln_ser_leu_asp
		Pseudomonas sp. CMR12a	JQ309920.1	thr_thr_ile_lys_dab_trp_leu_val_val_gln_leu_val_thr_pro_ser_leu_val_gln

Table 1.2 : continued

mass shift	tag	organism	accession number	predicted sequence
113-99	val_leu	*Pseudomonas fluorescens SBW25*	NC_012660.1	leu_ser_ile_lys_orn_ser_thr_val_leu_ser_ser_lys_gly_orn
		Pseudomonas syringae pv. syringae B728a	NC_007005.1	ser_ser_dab_dab_arg_phe_thr_asp_thr_thr_pro_val_leu_ala_ala_ala_val_thr_ala_val_dab_dab_tyr_ala_ala_thr_thr_ser_ala_thr_ala
		Pseudomonas putida strain PCL1445	DQ151887.2	leu_asp_leu_val_ser_leu_leu_gln_ser_val_leu_ser

Table 1.2 : continued

mass shift	tag	organism	accession number	predicted sequence
		Pseudomonas entomophila L48	NC_008027.1	val_leu_ser_ile_gln_val_leu_gln_val_leu_gln_ser
		Pseudomonas putida strain RW10S2	JN982333.1	leu_ser_ile_thr_val_leu_ser
		Pseudomonas syringae pv. tomato str. DC3000	NC_004578.1	leu_leu_gln_leu_thr_val_leu_leu

Table 1.2 : continued

mass shift	tag	organism	accession number	predicted sequence
		Pseudomonas syringae pv. tomato NCPPB 1108	NZ_ADGA01000001.1	leu_thr_val_leu_leu
		Pseudomonas sp. CMR12a	JQ309921.1	thr_hyv-d_leu_ser_leu_asp_leu_leu_ser_hyv-d
		Pseudomonas syringae pv. syringae 642	NZ_ADGB01000036.1	ile_leu_leu
		Pseudomonas fluorescens Pf-5	NC_004129.6	abu_leu_leu_ser_hyv-d_leu_asp_thr_ile_leu_ser

Table 1.2 : continued

mass shift	tag	organism	accession number	predicted sequence
128-113-113-115	asp_leu_leu_gln	*Pseudomonas fluorescens Pf0-1*	NC_007492.2	gln_ile_leu_gln_ser_leu_asp_leu_leu_ser_ile
		Pseudomonas fluorescens Pf-5	NC_004129.6	abu_thr_ile_leu_ser_leu_asp_leu_leu_ser_hyv-d
		Pseudomonas putida strain PCL1445	DQ151887.2	leu_asp_leu_leu_gln_ser_val_leu_ser_leu_val_ser
		Pseudomonas sp. CMR12a	JQ309921.1	thr_hyv-d_leu_ser_leu_asp_leu_leu_ser_hyv-d

Table 1.2 : continued

mass shift	tag	organism	accession number	predicted sequence
		Pseudomonas sp. MIS38	AB107223.1	leu_ser_ile_ile_asp_leu _asp_thr_leu_leu_ser
		Pseudomonas syringae pv. tomato str. DC3000	NC_004578.1	leu_thr_val_leu_leu_leu _leu_gln
		Pseudomonas syringae pv. syringae B728a	NC_007005.1	leu_leu_gln_leu_thr_ile _leu_leu

Table 1.2 : continued

mass shift	tag	organism	accession number	predicted sequence
		Pseudomonas syringae pv. tomato NCPPB 1108	NZ_ADGA01000001.1	leu_thr_val_leu_leu
		Pseudomonas syringae pv. syringae 642	NZ_ADGB01000121.1	leu_leu_gln_leu
		Pseudomonas syringae pv. tabaci ATCC 11528	AEAP01000030.1.1	leu_thr_ile_leu_leu
137-101	his_thr	*Pseudomonas sp. SHC52*	HQ888764.1	ser_orn_asp_lys_his_thr_thr_asp_thr

Table 1.2 : continued

mass shift	tag	organism	accession number	predicted sequence
		Pseudomonas fluorescens Pf-5	NC_004129.6	leu_asp_thr_ile_leu_ser_leu_leu_ser_hyv-d_abu
		Pseudomonas syringae pv. glycinea str. race 4	AEGH01000062.1	glu_arg_ser_lys_asn_thr_thr_ser_asn_ser
		Pseudomonas fluorescens SBW25	NC_012660.1	leu_ser_ile_lys_orn_ser_thr_val_leu_ser_ser_lys_gly_orn
		Pseudomonas syringae pv. tomato T1	NZ_ABSM01000024.1	asp_ser_thr_ser_asn_thr_lys_glu_arg_ser

Table 1.2 : continued

mass shift	tag	organism	accession number	predicted sequence
		Pseudomonas syringae pv. phaseolicola 1448A	NC_005773.3	glu_arg_ser_lys_asn_thr_thr_ser_asn_ser
		Pseudomonas aeruginosa 152504	AEVW01000121.1	
		Pseudomonas syringae pv. tomato str. DC3000	NC_004578.1	leu_leu_gln_leu_thr_val_leu_leu

Table 1.2 : continued

mass shift	tag	organism	accession number	predicted sequence
		Pseudomonas fluorescens strain SS101 clone 2	EU199081.2	leu_ser_ile_thr_ile_leu_ser
		Pseudomonas aeruginosa PA7	NC_009656.1	gly_thr_ser_orn_ser_orn_lys

Bacilli

tag	sequence	Organism	accession number	predicted sequence
113-113	leu_leu	*Bacillus amyloliquefaciens CAU-B946*	NC_016784.1	glu_leu_leu_val_asp_leu_leu
		Bacillus subtilis	D13262.1	val_asp_leu_glu_leu_leu

Table 1.2 : continued

mass hift	tag	organism	accession number	predicted sequence
		Bacillus subtilis subsp. subtilis RO-NN-1	NC_017195.1	val_asp_leu_leu
		Bacillus atrophaeus 1942	NC_014639.1	val_asp_leu_ile_glu_leu_leu
		Bacillus licheniformis BNP29	AJ005061.1	ile_gln_leu_leu_val_asp_leu
		B.subtilis	X70356.1	glu_leu_leu_val_asp_leu_leu

Table 1.2 : continued

mass shift	tag	organism	accession number	predicted sequence
		Bacillus licheniformis ATCC 14580	NC_006270.3	ile_val_asp_leu_gln_leu_leu
		Bacillus subtilis BSn5	NC_014976.1	glu_leu_leu_leu
		Bacillus licheniformis	U95370.1	gln_leu_leu_ile_val_asp_leu
		Bacillus subtilis subsp. Subtilis	JQ073775.1	leu_val_asp_leu_glu_leu_leu

Table 1.2 : continued

mass shift	tag	organism	accession number	predicted sequence
		Bacillus thuringiensis BMB171	NC_014171.1	Thr_Gln _Ala _Ser _His Gln _Gln
128-101-71	thr_gln	*Bacillus thuringiensis serovar huazhongensis BGSC 4BD1*	NZ_ACNI01000052.1	gln_ala_thr_ser_pro_glu_glu
		Bacillus cereus ATCC 10876	NZ_ACLT01000053.1	gln_ala_thr_ser_pro_glu_glu
		Bacillus cereus Rock1-3	NZ_ACMG01000150.1	gln_ala_thr_ser_leu_glu_glu

Table 1.2 : continued

mass shift	tag	organism	accession number	predicted sequence
		Bacillus cereus BDRD-ST196	NZ_ACMD01000179.1	gln_ala_thr_ser_pro_thr_glu
		Bacillus cereus 172560W	NZ_ACLV01000048.1	gln_ala_thr_ser_pro_glu_glu
		Bacillus cereus AH603	NZ_ACMP01000212.1	gln_ala_thr_ser_pro_thr_glu
		Bacillus cereus AH1134	NZ_ABDA02000007.1	gln_ala_thr_ser_pro_glu_glu
		Bacillus sp. 7_6_55CFAA_CT2	NZ_ACWE01000040.1	ser_pro_glu_gln_ala_thr_glu

Table 1.2: continued

mas shift	tag	organism	accession number	predicted sequence
		Bacillus cereus Rock1-15	NZ_ACMH01000131.1	gln_ala_thr_ser_pro_glu_glu
		Bacillus cereus m1550	NZ_ACMA01000045.1	gln_ala_thr_ser_pro_glu_glu
114-97-129	asn_pro_leu	*Bacillus amyloliquefaciens Y2*	NC_017912.1	ser_thr_tyr_asn_pro_glu_asn_4-hppa_ile_tyr_val_tyr_thr
		Bacillus amyloliquefaciens FZB42	NC_009725.1	ser_thr_tyr_asn_pro_glu_asn_4-hppa_ile_pro_glu_tyr_glu_val_tyr_thr_glu_orn
		Bacillus subtilis	AY137375.1	ser_thr_tyr_asn_pro_glu_asn

Table 1.2: continued

mass shift	tag	organism	accession number	predicted sequence
		Bacillus amyloliquefaciens subsp. plantarum YAU B9601-Y2	NC_017061.1	ser_thr_tyr_asn_pro_glu_asn_4-hppa_ile_pro_glu_tyr_val_glu_tyr_thr_glu_orn
		Bacillus amyloliquefaciens	JQ271536.1	ser_thr_tyr_asn_pro_glu_asn_ile_pro_glu_tyr_glu_val_tyr_thr_glu_orn
		Bacillus sp. 5B6 5B6	NZ_AJST01000001.1	ser_thr_tyr_asn_pro_glu_asn_4-hppa_ile_pro_glu_tyr_glu_val_tyr_thr_glu_orn

Table 1.2: continued

mass shift	tag	organism	accession number	predicted sequence
		Bacillus amyloliquefaciens FZB42.	AJ576102.1	ser_thr_tyr_asn_pro_glu_asn_4-hppa_ile_pro_glu_tyr_glu_val_tyr_thr_glu_orn
		Bacillus subtilis subsp. Spizizenii TU-B-10	NC_016047.1	asn_tyr_asn_gln_pro_4-hppa_ser_asn
		Bacillus amyloliquefaciens XH7	NC_017191.1	4-hppa_asn_ser_tyr_asn_gln_pro_asn
		Bacillus subtilis mycosubtilin	AF184956.1	4-hppa_ser_asn_tyr_asn_gln_pro_asn

Table 1.3 : MS/MS networking.contains molecules families.

Cluster name	m/z range	Organism found in cluster	Number of organism in cluster
Iturin	1025-1137 Da	ES113	11
		ZK4678	
		FZB42	
		ES223	
		ES73	
		ZK4633	
		ES118	
		ES345	
		ES382	
		3610	
		KP1032	
Surfactin	1002-1116Da	ES118	23
		3610	
		PY79	
		Zk4633	
		FZB42	
		KP1032	
		ES115	
		ES44	
		ES120	
		ES382	
		ES386	
		ES387	
		ES118	

Table 1.3: continued			
Cluster name	m/z range	Organism found in cluster	Number of organism in cluster
		ES73	
		ES113	
		ES221	
		ES223	
		ZK4678	
		ES44	
		ES120	
		ES343	
		ES345	
		ES76	
Kurstakin	835-960Da	GP24	7
		LCT5	
		LS18	
		ES118	
		ES333	
		ES336	
		ES343	
Bacitracin	704-722Da	ES115	3
		ES221	
		ES120	
Plipastatin	1441-1530Da	3610	9
		ES73	
		ES221	
		ES113	
		ES118	
		ES336	
		ES73	
		KP1032	

Table 1.3: continued			
Cluster name	m/z range	Organism found in cluster	Number of organism in cluster
		ES115	
Viscosin	1133-1369Da	P. tolaasii CH36	2
		P. aurantiaca	
Amphisin	1008-1148 Da	P.fluorescens BW11P2	
		ES97	
Rhamnolipid	499-867Da	P. aeruginosa 14	4
		P. fluorescens SH-C52	
		P. aeruginosa O1	
		P. HCL17(mosquitos)	
Promysalin	445-489 Da	*P. putida RW10S1*	1
Thanamycin	646Da	*P. fluorescens SH-C52*	1
Tolaasin	980-1008Da	*P. tolaasin CH36*	1

Table 1.4: Heat map guided bacilli and Pseudomonads and sequence tag

Number	Organism	Masses in Heat Map	Sequence Tag Y/N	Sequence Tag
Pseudomonas				
1	*Pseudomonas aurantiaca*	700-800Da	N	
		1100Da	Y	113-87-113-87
2	*Psuedomonas putida RW10S2*	750Da		
		1100Da	Y	113-87-113-87

Table 1.4 : continued				
Cluster name	m/z range	Organism found in cluster	Number of organism in cluster	Cluster name
3	*Pseudomonas tolaasii CH36*	1100Da	Y	113-87-113-87
		2000Da	Y	99-99-113-87-99
4	*Pseudomonas putida RW10S1*	750Da	N	N/A
5	*Pseudomonas fluorescens BW11P2*	700Da	N	
		800Da	N	
		1000-1100 Da	Y	128-113-113-115
6	*Pseudomonas putida BW11M1*	700Da	N	
		1800Da	Y	113-99
Number	Organism	Masses in Heat Map	Sequence Tag Y/N	Sequence Tag
7	*Pseudomonas florescens RW9S1*	700-800Da	N	
		900Da	N	
		1100Da	N	
8	*Pseudomonas aeruginosa(MPAO1)*	1100Da	N	

Table 1.4 : continued				
Cluster name	m/z range	Organism found in cluster	Number of organism in cluster	Cluster name
9	*Pseudomonas aeruginosa(MPA 14)*	600-800Da	N	
10	*Pseudomonas chlororaphis 200B1*	700Da	N	
11	*Pseudomonas moraviensis ES97*	700Da	N	
		1100Da	Y	128-113-113-115
12	*Pseudomonas moraviensis ES16*	N/A	N	
13	*Pseudomonas spp ES11*	1000Da	N	
14	*Pseudomonas para ES60*	800Da	N	
15	*Pseudomonas flourescens C52*	700Da	N	
		1000Da	Y	99-99
		1300Da	N	137-101
		2100Da	N	
16	*Pseudomonas fluorescens Tn5*	no heat map	Y	71-71-99
17	*Psuedomonas putida*	700Da	N	

Table 1.4 : continued				
Number	Organism	Masses in Heat Map	Sequence Tag Y/N	Sequence Tag
18	*Pseudomonas HCL17*	700Da	N	
Bacilli				
19	*QMB1551*	1000-1800 Da	N	N
20	*ES-190*	800Da	N	N
21	FZB42(bmyA, trpAB, & pks1KS1 mts)	800Da		N
		1000Da, 1200Da	Y	113-113
		1500Da	Y	orn
22	ES-20	800-900Da		
		1800-2000 Da		N
23	ES-118	N/A		
24	ES-115	1000Da		113-113
		1500Da	Y	Orn
		1500Da	Y	113-129-113
		1800Da	N	
25	ES-73	1000Da	Y	113-113
		1500Da	Y	Orn
		1500Da	Y	113-129-113
26	3610	800Da		
		1000Da	Y	113-113
		1500Da	Y	Orn
27	KP1302	800Da		
		1000Da	Y	113-113

Table 1.4 : continued				
Number	Organism	Masses in Heat Map	Sequence Tag Y/N	Sequence Tag
28	PY79	800Da		
		1000Da	Y	113-113
29	ES-114	looks like only matrix		
30	ES-44	1000Da	Y	113-113
		1800Da		
31	ES-120	800Da	N	
		900Da		
		N/A	Y	113-129-113
32	ES-76	800Da	N	
		1000Da	Y	113-113
33	ES113	1000Da	Y	113-113
		1500Da	Y	Orn
34	ES221	1500Da	Y	Orn
			Y	113-129-113
35	ES222	800Da	N	
36	ES223(from Beth)	1000Da	Y	113-113
		1200Da	N	
		1500Da	Y	Orn
37	ES332	700Da		
38	ES333	800Da		
39	ES335	~825	N	
40	ES336	826.41	Y	gly_ala_val
		~900	N	
		1500Da	Y	Orn
41	ES337	~1150	N	

Table 1.4 : continued				
Number	Organism	Masses in Heat Map	Sequence Tag Y/N	Sequence Tag
42	ES341	N/A		
43	ES342	700Da	N	
		800Da	N	
44	ES343	1000Da	Y	Gln-thr
		1800Da	N	
45	ES345	1000Da	Y	113-113
		1800Da	N	
46	ES382	1000Da	Y	113-113
		1500Da		
47	ES386	1058	Y	ala_thr // ile/leu_ile/leu
		1074	Y	ile/leu_met
		~1500	don't see	
48	ES387	none		
49	ES392	none		
50	ZK4633	1000Da	Y	113-113
		1200Da	N	
		1500Da	Y	Orn
51	ZK4636	N/A		
52	ZK4678	750Da	N	
		1000Da	Y	113-113
		1031	Y	leu-leu
		1044	Y	leu-leu
		1060	Y	leu-leu
		1074	Y	leu-leu

Table 1.4 : continued				
Number	Organism	Masses in Heat Map	Sequence Tag Y/N	Sequence Tag
		1500Da	N	
		3400Da	N	
53	GP24	900Da	Y	Gln-thr
		892	Y	glu_gln/lys_his
		906	Y	glu_gln/lys // gln/lys
54	GP25	N/A		
55	LCT22	900Da	Y	Gln-thr
		892	Y	glu_gly_his
		906	Y	glu_gln/lys
		928	Y	glu_gln/lys
		944	Y	glu_gln/lys
56	LCT5	900Da		Gln-thr
57	LS18	900Da		Gln-thr
58	LS27B	900Da	N	
		1300Da	N	
59	LS28	750Da	N	
60	LS29	750Da	Y	113(or 87)-99-113-113
		714	Y	ile/leu_val_ile/leu_ile/leu
		752	N	

Chapter 1, in part, is currently being prepared for submission for publication of the material. It is a joint effort between thesis author, Don Nyugen; Nuno Bandeira; Pieter C. Dorrestein The dissertation author was the primary investigator and author of this material.

Chapter 2

Capture of induced molecules by imaging mass spectrometry.

2.1 Introduction:

In history, penicillin, therapeutic antibiotic, was first drug discovered by Alexander Fleming from the interaction between Staphylococcus aureus and Pencillicum rubens around 1900 and further developed into medical purpose. Afterword, this norm, bacteria interaction, became a predominant approach to discover natural products for application. This idea is also applied for biocontrol so as to facilitate agriculture economics. In general, pathogenic microorganisms giving side-effect for plant promotion are main threatens to crop production and infecting plants.

In order to control such pathogenic microorganisms, high frequency and high dosage of agrochemicals have been used more and more common than before because pathogenic microorganisms owing well-resistant ability survive in the environment and severely affect plant health. Therefore, chemical agents became less efficient, which makes people to come up with new methodologies to overcome such situation, including planting disease free seeds to avoid being infected by pathogenic microorganism or using biological control approach, a more promising way of decreasing the use of chemical agents in agriculture, to

control plant pathogens. In the ecosystem, not all microorganisms are plant pathogens but some are beneficial ones facilitate plant growth, which called plant growth promoting bacteria (PGPB). Some PGPB colonize around root surfaces to help absorb nutrient, some living in root interior form endophytic population, and some of them are able to move not only from the root cortex to vascular system but also prosper in stem, leaves. Generally speaking, some PGPB interact with host plants to promote plants growth. Some of these bacteria are capable of secreting secondary metabolites to against pathogenic microbes.[77] For instance, sugarcane is a vital cash crop in many countries, such as USA, Thailand, and Australia. Because of its economic value, increasing the yield or productivity of per unit area is a good way to reduce cost. [75-76] However, sugarcane infected and damaged by fungus are not able to produce so much sugar as expected. Red-rot disease of sugarcane resulted from a fungus, *Colletotrichum falcatum* is a major notorious plant pathogen that threats to growth of sugarcane so that the disease is subject to widely investigation. In addition, pseudomonas strains were known for production of antibiotics, phenzine, viscosinamide, pyoluteorin, and so on. A novel strain *Pseudomonas aurantiaca PB-St2*, isolated from the stem surface, are

able to secret some antimicrobial compounds, such as phenazine, viscosin, PCA and also show ability to inhibit Colletotrichum falcatum.[7] [5-76] Therefore, we selected these two microorganisms isolated from sugarcane to demonstrate how interplay happen by MALDI imaging mass spectrometry.

MALDI imaging mass spectrometry have been applied to study biological samples, cells, tissues, organs. It is the advantage that its ability to visualize the distribution of large amount of different analytes. Our lab firstly developed a mass spectrometry-based approach on microbes so as to visualize spatial distribution of molecules and observe metabolic exchange so as to show communication between investigated microbes. Examples included observation of chemical interplay between *Bacillus subtilis* and *Streptomyces coelicolor*, or *Pseudomonas aeruginosa* and *Aspergillus fumigatus*. Therefore, we applied imaging mass spectrometry (MALDI-TOF-IMS) to study interaction between *P. aurantiaca vs C. falcatum* through time course from 8hr to 96hr and eventually observed up-regulated ions, m/z 1183, 1189, 1205, 1211,1227. Further purification of these ions were investigated by tandem mass spectrometry. This work demonstrates application of MALDI-IMS in discovering possible metabolites coming from

biocontrol agent to suppress plant pathogens and investigate interested ions.

2.2 Discussion and Result

It has been reported that P. aurantiaca has the ability to inhibit the fungal, Colletotrichum falcatum, and control the red-rot disease, so we want to understand what kind of compounds works to against Colletotrichum falcatum. Interestingly, several up-regulated molecules have been discovered via imaging mass spectrometry (IMS). Start with 8hr, three peaks, phenazine (m/z 226), viscosin (M+Na, m/z 1148), Viscosin (M+K, m/z 1164) were observed (figure 2.1). Furthermore, at 12hr, m/z 1183, 1189, 1205, 1211, 1227 showed up around the interface without observation of same peak in the control, which suggested that those ions are up-regulated ions (Figure 2.1).

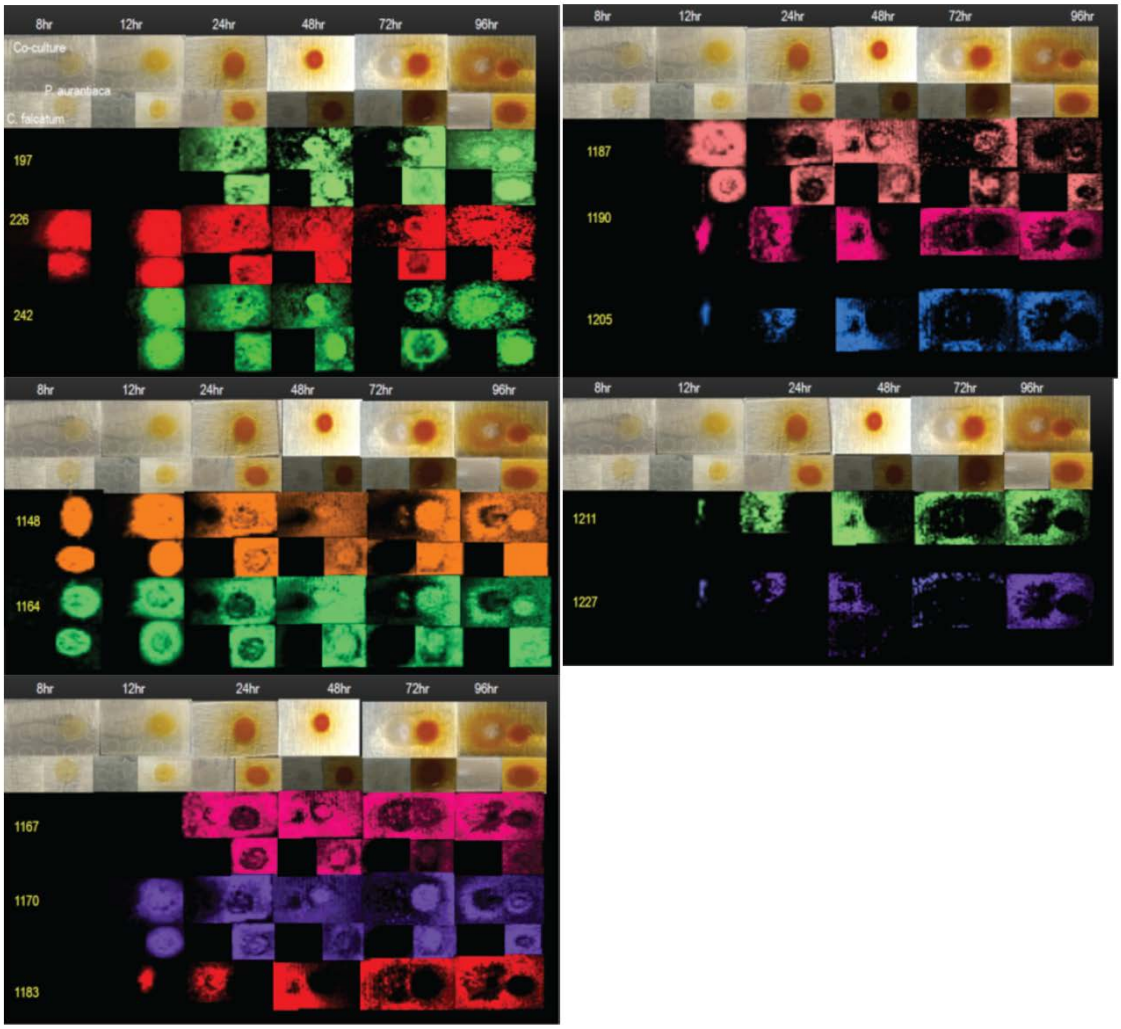

Figure 2.1: Imaging mass spectrometry of *P. aurantiaca vs C. falcatum* from 8hr to 96hr. m/z 226, 1148, 1164 appeared at 8hr indicated they might induce up-regulated ions, m/z 1183, 1189, 1205, 1211, 1227 as induced ions show up at 12hr. Top panel represent interaction between C. falcatum and P.aurantiaca: Left is C. falcatum, right side is P. aurantiaca. Left bottom: C. falcatum alone, Right bottom: P. aurantiaca alone. m/z 1183, 1189, 1205, 1211, 1227 come out as induced molecules. Bioassay C. falcatum with viscosin induced few up-regulated molecues, which suggested ions derived from fungus.

These induced ions come out at 12hr indicates some molecules might play important role in inducing up-regulated ions before 12hr. we assume phenazine or viscosin might play vital role in inducing up-regulated ions. According to literature, lipopeptides from plant-associated pseudomonas have variety of functions, including formation of biofilm, defense against microorganisms, or virulent factors, swarming motility. However, it is hard to tell these ions are fungus-derived or bacteria-derived so that we have two assumptions. 1) If induced ions are fungus-derived, we are able to see induced molecules by treating *Colletotrichum falcatum* with viscosin. 2) If induced ions are bacteria-derived, we are not able to observe any induced molecules from fungus by spotting viscosin. Here, we purified viscosin from *Psuedomonas aurantiaca* by HPLC and then *Colletotrichum falcatum* was treated with 300ug viscosin. It was few induced molecules, m/z 1183, m/z 1189, m/z 1211, that come out from bioassay (Figure 2.2), which matches to both assumptions. In addition, from timecourse data, those ions start coming from the interface of fungus side and then distribute throughout the whole plate, which suggests these molecules may come from *Colletotrichum falcatum* to prevent affecting by *P. aurantiaca*. Put together, it is suggested that these induced

molecules derived from fungus.

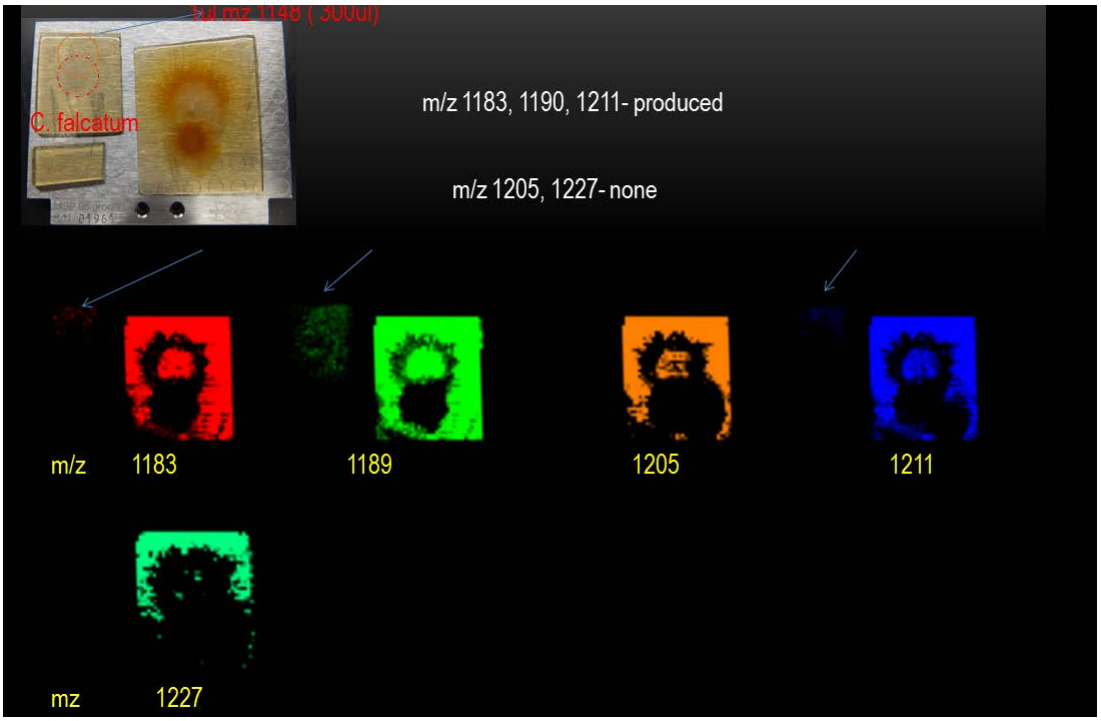

Fig 2.2: Discovery induced molecules by spotting around 300ul viscosin with *C. falcatum.* Induced molecules come along with this order: m/z 1189> m/z 1211> m/z 1183

We focus on m/z 1189 (induced ion) as the amount of it is relatively larger than others based on gel filtration and bioassay (figure 2.2). In order to confirm that we target at right induced molecules, MS/MS fragmentation of m/z 1189 gathered from methanol crude extract (ion trap) and TOF-TOF from co-culture sample and correlated to each other. The fragmentation patterns are similar, suggesting m/z 1189 is the interested ions (figure 2.3).

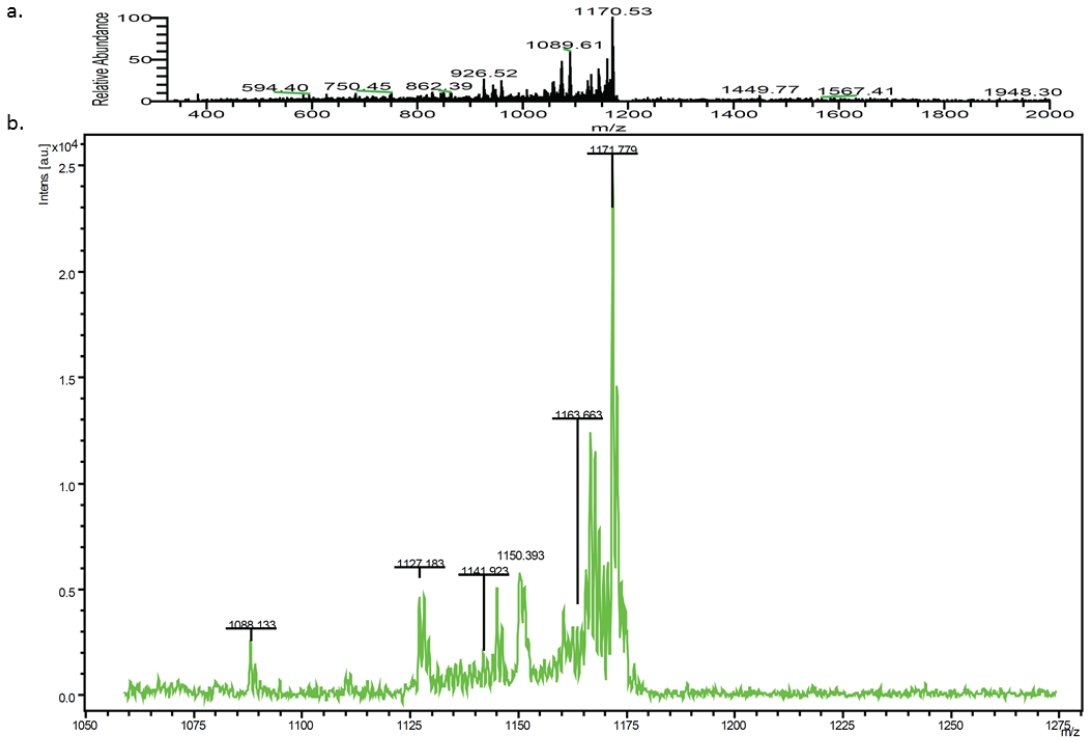

Figure 2.3: In comparison of m/z 1189 (induced molecule) from methanol extract and that on co-culture sample. a) MS2 of methanol crude extract. b) Detected induced molecule on interaction samples.

Methanol crude extract was firstly cleaned by gel filtration and further purified by HPLC. First run of HPLC, in retention time= 17, 18, 19, observation of m/z 1189 (M+K), 1173 (M+Na), low intensity of m/z 1151(M+H) and other ions exist in these three fractions. Then we collect these fractions for second run of HPLC and

extend gradient time. m/z 1151 increase and primary locate in retention time = 20, 21, 22, which suggests salts are removed from adduct form and signal became [M+H] = m/z 1151 that is subject to tandem mass spectrometry (Figure 2.4, Table 2.1).

Table 2.1: Factions contain induced after purification: gel filtration, 1st HPLC, 2nd HPLC

Purify method/ factions	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
Gel filtration							X	X	X																	
HPLC Run first run																	x	x	x							

Table 2.1: continued																									
HPLC 2nd run																			x	x	x				

Figure 2.4: Purification induced molecule by HPLC (2nd): induced molecules (m/z 1151) locate in fraction 20, 21, 22.

Analysis of MS/MS spectrum indicates the induced molecule is peptidic compound as we can found proteinomic sequence. A partial sequence tag, 137-113-97-97-129, was obtained from mass spectrum. So far, there are not many available genomes containing *Colletotrichum falcatum* .It is necessary to sequence *Colletotrichum falcatum* so as to get available genome for genome mining (Figure 2.5). In addition, since this molecule has peptide character, so De novo sequencing of it is also in favor of harvestuing more structure information (Figure 2.6).

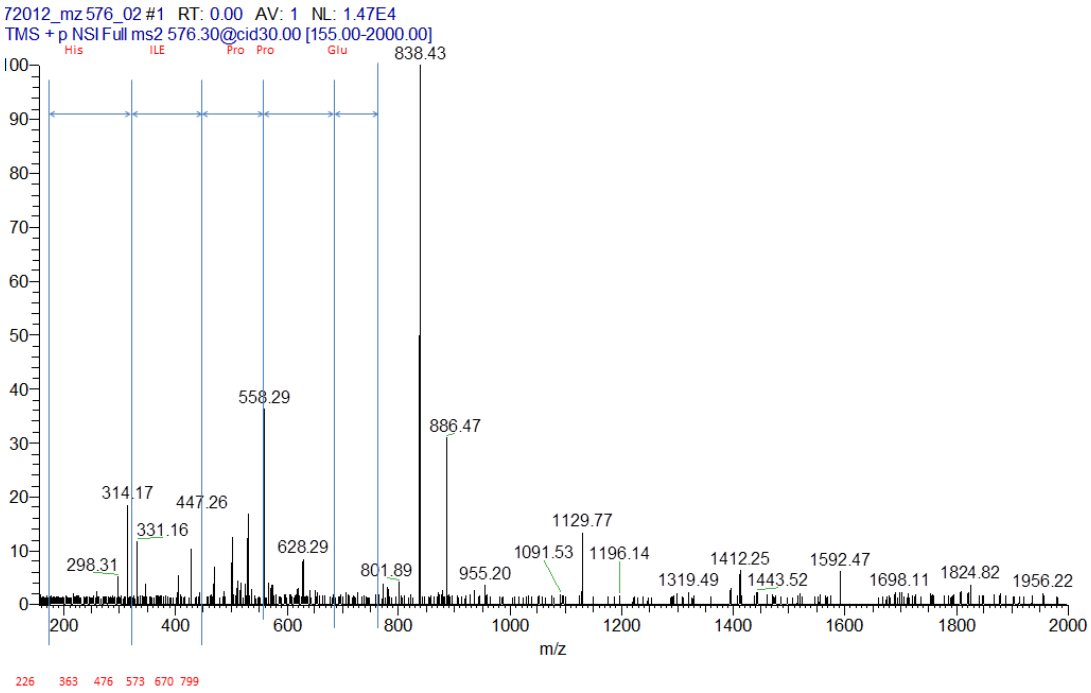

Figure 2.5: MS2 of m/z 1151 (M+2H): the spectrum contains 137-113-97-129

2.3 Material and methods

2.3.1 Preparation of *P.aurantiaca* and *Colletotrichum falcatum*

LB broth was purchased from fisher sciencist. ISP2 agar medium was prepared with 2g yeast, 5g malt, 2g dextrose, 10g agar/ 500ml. *P. aurantiaca* and *Colletotrichum falcatum* were cultured in the LB both for 24hr in the 28℃ shake and then transferred to ISP2 agar medium to grow for from 8hr to 96hr in 30℃ incubator. (The OD of *Pseudomonas aurantiaca* was 1.706 and that of *Colletotrichum falcatum* was 0.567.)

2.3.2 IMS these two strains from 8hr to 96hr

Cut a thin layer of growth medium that containing interesting bacteria and transfer to MALDI plate. Cover the transferred sample with universal matrix until formation of even matrix on the top of samples. Place the prepared sample into over in 37℃ oven to dry the samples. Finally, subject the dried samples to Microflex (RP mode) to monitor the interaction from 8hr to 96hr.

2.3.3 Methanol extraction from co-culture assay and gel filtration

400 ml methanol was used to extract the small amount of induced molecules from a hundred of interaction petri dish. The interaction extraction liquid was

dried by rotavapor and then redissolved in 3 ml methanol. Loading the 3ml crude extract onto the top of silica gel column chromatology (sephadex column) is to separate induced molecules from crude extract. The fraction with m/z 1189 was collected in order to subject to HPLC and get purified compounds.

2.3.4 Sephadex column filtration and HPLC

After gel filtration, we took fraction having m/z 1189 to run HPLC.

First run, in retention time= 17, 18, 19, we observed m/z 1189, 1173, low intensity of m/z 1151 with other signals in these three fractions. Then we collect these fractions to do HPLC for second with extend gradient time so m/z 1151 increase and primary locate in retention time = 20, 21, 22, which suggests salts are removed from adduct form and signal becomes [M+H] = m/z 1151. 50 ul of crude fraction was subjected into HPLC connected with 25X10mm 5ul column. Fraction collector collected fractions every 1 min. (HPLC gradient solvent water: Acetonitrile. Gradient :0 min 5% (ACN), 5min 5%, 60min 95%, 65 min 95%, 70min 5 %)

2.3.5 Tandem mass spectrometry

1 ul samples were diluted to 10 ul in methanol and then infused by nano-electrospray ionization with Biversa Nanomate. (Pressure: 0.3 psi, spray voltage: 1.4kV). The data were recorded by Ion trap, and FT-MS (Thermo-Electron) running with Tune Plus and Xcalibur software. The interested ions was isolated in the FT and fragmented by CID (3 m/z) with 30 collision energy. The MS/MS scans were run with QualBrowser software to analyze the recorded data.

2.4 Future direction:

Peptidogenomic will be applied to build connection between genotype and chemotype and help to characterize possible structure of m/z 1151. Once we have available genome information, we can use same approach,association of unsequenced bacteria with publicly genome sequence, described in chapter 1.

References

1. Tim T. Binnewies, Yair Motro, Peter F. Hallin, Ole Lund, David Dunn, Tom La, David J. Hampson, Matthew Bellgard, Trudy M. Wassenaar, David W. Ussery. Ten years of bacterial genome sequencing: comparative genomics based discoveries. Functional & Integrative Genomics. 2006, (6): 165-185.

2. Gregory L. Challis. Genome Mining for Novel Natural Product Discovery. Journal of Medicinal Chemistry. 2008, 51: 2618-2628.

3. Gregory L. Challis. Mining microbial genomes for new natural products and biosynthetic pathways. Microbiology. 2008, 154: 1555–1569.

4. Sylvie Lautru, Robert J. Deeth, Lianne M. Bailey, Gregory L. Challis. Discovery of a new peptide natural product by Streptomyces Coelicolor genome mining. Nature Chemical Biology. 2005, 1 (5): 265-269.

5. Roland D. Kersten, Yu-Liang Yang, Yuquan Xu, Peter Cimermancic, Sang-Jip Nam, William Fenical, Michael A. Fischbach, Bradley S. Moore, and Pieter C. Dorrestein. A mass spectrometry–guided genome mining approach for natural product peptidogenomics. Nature Chemical Biology. 2011, 7 (11): 794-802.

6. Christophe Corre, Gregory L. Challis. New natural product biosynthetic chemistry discovered by genome mining. Natural Product Reports. 2009, 26 (8): 977-986.

7. Irene de Bruijn, Maarten J D de Kock, Meng Yang, Pieter de Waard, Teris A van Beek, Jos M Raaijmakers. Genome-based discovery, structure prediction and functional analysis of cyclic lipopeptide antibiotics in Pseudomonas species. Molecular Microbiology. 2007 63(2): 417–428.

8. Michael HT Li, Peter MU Ung, James Zajkowski, Sylvie Garneau-Tsodikova and David H Sherman. Automated genome mining for natural products. BMC Bioinformatics. 2009 10:185.

9. Helge B. Bode, Rolf Müller. The impact of bacterial genomics on natural product research. Angew Chem Int Ed Engl. 2005 Oct 28;44(42):6828-46.

10. Mikhail O. Maksimov, Istan Pelczer, A. James Link. Precursor-centric genome-mining approach for lasso peptide discovery, Proc Natl Acad Sci U S A. 2012 Sep 18;109(38):15223-8.

11. William B Whitman, David C Coleman, William J Wiebe. Prokaryotes: The unseen majority. Proceedings of the National Academy of Sciences, USA. 1998, 95: 6578-6583.

12. Dwayne C Savage. Microbial Ecology of the Gastrointestinal Tract. Annual Review of Microbiology. 1977, 31:107-133.

13. R. E. Ley, D. A. Peterson, and J. I. Gordon, “Ecological and evolutionary forces shaping microbial diversity in the human intestine,” Cell, vol. 124, no. 4, pp. 837–848, 2006.

14. K. Chiller, B. A. Selkin, and G. J. Murakawa, “Skin microflora and bacterial infections of the skin,” Journal of Investigative Dermatology Symposium Proceedings, vol. 6, no. 3, pp. 170–174, 2001.

15. Andrew S. Neish. The gut microflora and intestinal epithelial cells: a continuing dialogue. Microbes and Infection. 2002, 4: 309-317.

16. Usha Vyas, Natarajan Ranganathan. Probiotics, Prebiotics, and Synbiotics: Gut and Beyond. Gastroenterology Research and Practice. 2012.

17. Jay Shendure, Hanlee Ji. Next-generation DNA sequencing. Nature Biotechnology. 2008, 26: 1135-1145.

18. Kircher M, Kelso. High-throughput DNA sequencing--concepts and limitations. Bioessays. 2010 32 (6): 524-536.

19. Claudine Me´digue, Ivan Moszer. Annotation, comparison and databases for hundreds of bacterial genomes. Research in Microbiology. 2007, 158: 724-736.

20. Victor M. Markowitz, I-Min A. Chen, Krishna Palaniappan, Ken Chu, Ernest Szeto, Yuri Grechkin, Anna Ratner, Biju Jacob, Jinghua Huang, Peter Williams, Marcel Huntemann, Iain Anderson, Konstantinos Mavromatis, Natalia N. Ivanova, Nikos C. Kyrpides. IMG: the integrated microbial genomes database and comparative analysis system. Nucleic Acids Res. 2012, 40(D1): D115–D122.

21. Gene H. Hur , Christopher R. Vickery and Michael D. Burkart, Explorations of catalytic domains in non-ribosomal peptide synthetase enzymology, Nat. Prod. Rep., 2012,29, 1074-1098.

22. Robert Finking and Mohamed A. Marahiel, BIOSYNTHESIS OF NONRIBOSOMAL PEPTIDES, Annual Review of Microbiology Vol. 58: 453-488.

23. Stefano Donadio, Paolo Monciardini and Margherita Sosio, Polyketide synthases and nonribosomal peptide synthetases: the emerging view from bacterial genomics, Nat. Prod. Rep., 2007, 24, 1073–1109.

24. Ségolène Caboche, Maude Pupin, Valérie Leclère, Arnaud Fontaine, Philippe Jacques, and Gregory Kucherov, NORINE: a database of

nonribosomal peptides, Nucleic Acids Res. 2008 January; 36(Database issue): D326–D331.

25. B. Shen, L. Du, C. Sanchez, D.J. Edwards, M. Chen, and J.M. Murrell, The biosynthetic gene cluster for the anticancer drug bleomycin from Streptomyces verticillus ATCC15003 as a model for hybrid peptide–polyketide natural product biosynthesis, Journal of Industrial Microbiology & Biotechnology (2001) 27, pp. 378–385.

26. Morgan A. Wyatt, Wenliang Wang, Christelle M. Roux, Federico C. Beasley, David E. Heinrichs, Paul M. Dunman, Nathan A. Magarvey, Staphylococcus aureus Nonribosomal Peptide Secondary Metabolites Regulate Virulence, science(2010) , VOL 329.

27. Mossialos D, Ochsner U, Baysse C, Chablain P, Pirnay JP, Koedam N, Budzikiewicz H, Fernández DU, Schäfer M, Ravel J, Cornelis P, Identification of new, conserved, non-ribosomal peptide synthetases from fluorescent pseudomonads involved in the biosynthesis of the siderophore pyoverdine, Mol Microbiol. 2002 Sep;45(6):1673-85.

28. Masayuki Inoue, Naoki Shinohara, Shintaro Tanabe, Tomoaki Takahashi, Ken Okura, Hiroaki Itoh, Yuki Mizoguchi, Maiko Iida, Nayoung Lee & Shigeru Matsuoka, Total synthesis of the large non-ribosomal peptide polytheonamide B, Nature Chemistry (2010), 280–285.

29. Michael Zimmermann, Michael A. Fischbach, A Family of Pyrazinone Natural Products from a Conserved Nonribosomal Peptide Synthetase in Staphylococcus aureus, Chem Biol. (2010) Sep 24;17(9):925-30.

30. Michael A. Marques, Diane M. Citron, and Clay C. Wang, Development of Tyrocidine A Analogues with Improved Antibacterial Activity, Bioorg Med Chem. 2007 November 1; 15(21): 6667–6677.

31. Lars Robbel, Mohamed A. Marahiel. Daptomycin, a Bacterial Lipopeptide Synthesized by a Nonribosomal Machinery. Journal of Biological Chemistry. 285: 27501-27508.

32. Christian Rausch, Tilmann Weber1, Oliver Kohlbacher, Wolfgang Wohlleben1 and Daniel H. Huson, Specificity prediction of adenylation domains in nonribosomal peptide synthetases (NRPS) using transductive support vector machines (TSVMs), Nucleic Acids Research, 2005, Vol. 33, No. 18 5799–5808

33. Jan Grünewald and Mohamed A. Marahiel, Chemoenzymatic and Template-Directed Synthesis of Bioactive Macrocyclic Peptides, Microbiol Mol Biol Rev. 2006 March; 70(1): 121–146.

34. Ng J, Bandeira N, Liu WT, Ghassemian M, Simmons TL, Gerwick WH, Linington R, Dorrestein PC, Pevzner PA , Dereplication and de Novo sequencing of nonribosomal peptides, Nat Methods. 2009 Aug;6(8):596-9.

35. Chen Y, Ntai I, Ju KS, Unger M, Zamdborg L, Robinson SJ, Doroghazi JR, Labeda DP, Metcalf WW, Kelleher NL, A proteomic survey of nonribosomal peptide and polyketide biosynthesis in actinobacteria, J Proteome Res. 2012 Jan 1;11(1):85-94.

36. Mohimani H, Liu WT, Yang YL, Gaudêncio SP, Fenical W, Dorrestein PC, Pevzner PA, Multiplex de novo sequencing of peptide antibiotics, J Comput Biol. 2011 Nov;18(11):1371-81.

37. Sascha Doekel, Marie-Françoise Coëffet-Le Gal, Jian-Qiao Gu, Min Chu, Richard H. Baltz and Paul Brian, Non-ribosomal peptide synthetase module fusions to produce derivatives of daptomycin in Streptomyces roseosporus, Microbiology September 2008 vol. 154 no. 9 2872-2880.

38. Rounge TB, Rohrlack T, Nederbragt AJ, Kristensen T, Jakobsen KS, A genome-wide analysis of nonribosomal peptide synthetase gene clusters and their peptides in a Planktothrix rubescens strain, BMC Genomics. 2009 Aug 25;10:396.

39. Leslie Hicks, Paul Weinreb, Dirk Konz, Mohamed A. Marahiel, Christopher T. Walsh, Neil L. Kelleher, Fourier-transform mass spectrometry for detection of thioester-bound intermediates in unfractionated proteolytic mixtures of 80 and 191 kDa portions of Bacitracin A synthetase, Analytica Chimica Acta 496 (2003) 217–224.

40. Xinyu Wang, Huijuan Chen, Jeonghoon Lee, Peter T.A. Reilly. Increasing the trapping mass range to m/z = 109—A major step toward high resolution mass analysis of intact RNA, DNA and viruses. International Journal of Mass Spectrometry. 2012, 328-329: 28-35.

41. Eduard Denisov, Eugen Damoc, Oliver Lange, Alexander Makarov. Orbitrap mass spectrometry with resolving powers above 1,000,000. International Journal of Mass Spectrometry. 2012, 325-327: 80-85.

42. Jeramie D. Watrous & Pieter C. Dorrestein, Imaging mass spectrometry in microbiology, Nature Reviews Microbiology 9, 683-694.

43. Edwin Lasonder, Yasushi Ishihama, Jens S. Andersen, Adriaan M. W. Vermunt, Arnab Pain, Robert W. Sauerwein, Wijnand M. C. Eling, Neil Hall, Andrew P. Waters, Hendrik G. Stunnenberg & Matthias Mann, Analysis of the Plasmodium falciparum proteome by high-accuracy mass spectrometry, Nature 419, 537-542

44. Joseph Zaia, Mass Spectrometry and Glycomics, OMICS. 2010 August; 14(4): 401–418.

45. Herring KD, Oppenheimer SR, Caprioli RM, Direct tissue analysis by matrix-assisted laser desorption ionization mass spectrometry: application to kidney biology, Semin Nephrol. 2007 Nov;27(6):597-608
.
46. Marvin LF, Roberts MA, Fay LB, Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry in clinical chemistry, Clin Chim Acta. 2003 Nov;337(1-2):11-21.

47. Martin A. Claydon, Simon N. Davey, Valeria Edwards-Jones & Derek B. Gordon, The rapid identification of intact microorganisms using mass spectrometry, Nature Biotechnology 14, 1584 – 1586.

48. M. Constapel, M. Schellentra¨ger, O. J. Schmitz, S. Ga¨b, K. J. Brockmann, R. Giese and Th. Benter, Atmospheric-pressure laser ionization: a novel ionization method for liquid chromatography/mass spectrometry, Rapid Commun. Mass Spectrom. 2005; 19: 326–336.

49. Gary J. Van Berkel,∗ Sofie P. Pasilis and Olga Ovchinnikova, Established and emerging atmospheric pressure surface sampling/ionization techniques for mass spectrometry, J. Mass Spectrom. 2008; 43: 1161–1180.

50. Sarah Trimpin, Ellen D. Inutan, Thushani N. Herath, and Charles N. McEwen, Laserspray Ionization, a New Atmospheric Pressure MALDI Method for Producing Highly Charged Gas-phase Ions of Peptides and Proteins Directly from Solid Solutions, Molecular & Cellular Proteomics 9:362–367, 2010.

51. Patrick J. Roach, Julia Laskin, and Alexander Laskin, Molecular Characterization of Organic Aerosols Using

Nanospray-Desorption/Electrospray Ionization-Mass Spectrometry, Anal. Chem., 2010, 82 (19), pp 7979–7986.

52. Jeramie Watrous, Patrick Roach, Theodore Alexandrov, Brandi S. Heath, Jane Y. Yang, Roland D. Kersten, Menno van der Voort, Kit Pogliano, Harald Gross, Jos M. Raaijmakers, Bradley S. Moore, Julia Laskin, Nuno Bandeira, and Pieter C. Dorrestein. Mass spectral molecular networking of living microbial colonies. Proceedings of the National Academy of Sciences, USA. 2012.

53. Sylvie Lautru and Gregory L. Challis, Substrate recognition by nonribosomal peptide synthetase multi-enzymes, Microbiology (2004), 150, 1629–1636.
54. Du L, Lou L, PKS and NRPS release mechanisms, Nat Prod Rep. 2010 Feb;27(2):255-78. Epub 2009 Dec 4.

55. Lauren B. Pickens, Yi Tang, and Yit-Heng Chooi, Metabolic Engineering for the Production of Natural Products, Annu. Rev. Chem. Biomol. Eng. 2011. 2:211–36.

56. Robert Finking and Mohamed A. Marahiel, BIOSYNTHESIS OF NONRIBOSOMAL PEPTIDES, Annual Review of Microbiology Vol. 58: 453-488.

57. Niran Roongsawang, Kenji Washio and Masaaki Morikawa, Diversity of Nonribosomal Peptide Synthetases Involved in the Biosynthesis of Lipopeptide Biosurfactants, Int. J. Mol. Sci. 2011, 12, 141-172.

58. Michael A. Fischbach and Christopher T. Walsh, Assembly-Line Enzymology for Polyketide and Nonribosomal Peptide Antibiotics: Logic, Machinery, and Mechanisms, Chem. Rev. 2006, 106, 3468-3496.

59. Davies J, How to discover new antibiotics: harvesting the parvome, Curr Opin Chem Biol. 2011 Feb;15(1):5-10.

60. Julian Davies and Katherine S. Ryan, Introducing the Parvome: Bioactive Compounds in the Microbial World, ACS Chem. Biol., 2012, 7 (2), pp 252–259.

61. Gregory L Challis, Jacques Ravel and Craig A Townsend, Predictive, structure-based model of amino acid recognition by nonribosomal peptide synthetase adenylation domains, Chemistry & Biology 2000, 7:211–224.

62. J Krätzschmar, M Krause, and M A Marahiel, Gramicidin S biosynthesis operon containing the structural genes grsA and grsB has an open reading frame encoding a protein homologous to fatty acid thioesterases, J Bacteriol. 1989 October; 171(10): 5422–5429.

63. Marnix H. Medema, Kai Blin, Peter Cimermancic, Victor de Jager, Piotr Zakrzewski, Michael A. Fischbach, Tilmann Weber, Rainer Breitling & Eriko Takano (2011). antiSMASH: Rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters. Nucleic Acids Research 39: W339-W346.

64. Marc Röttig, Marnix H. Medema, Kai Blin, Tilmann Weber, Christian Rausch, NRPSpredictor2—a web server for predicting NRPS adenylation domain specificity, Nucleic Acids Research, 2011, 1–6.

65. Christian Rausch, Ilka Hoof, Tilmann Weber, Wolfgang Wohlleben and Daniel H Huson, Phylogenetic analysis of condensation domains in NRPS sheds light on their functional evolution, BMC Evolutionary Biology 2007, 7:78.

66. Prieto C , Garcia-Estrada C , Lorenzana D and Martin JF, NRPSsp: Non-Ribosomal Peptide Synthase substrate predictor, Bioinformatics Advance Access published November 29, 2011.

67. Mohd. Zeeshan Ansari, Gitanjali Yadav, Rajesh S. Gokhale, and Debasisa Mohanty, NRPS-PKS: a knowledge-based resource for analysis of NRPS/PKS megasynthases, Nucleic Acids Res. 2004 July 1; 32(Web Server issue): W405–W413.

68. Holger Jenke-Kodama and Elke Dittmann, Bioinformatic perspectives on NRPS/PKS megasynthases: Advances and challenges, Nat. Prod. Rep., 2009, 26, 874–883.

69. Ansari MZ, Sharma J, Gokhale RS, Mohanty D, In silico analysis of methyltransferase domains involved in biosynthesis of secondary metabolites, BMC Bioinformatics. 2008 Oct 25;9:454.

70. Harald Gross and Joyce E. Loper, Genomics of secondary metabolite production by Pseudomonas spp, Nat. Prod. Rep., 2009,26, 1408-1446.

71. Sansinenea E, Ortiz A, Secondary metabolites of soil Bacillus spp, Biotechnol Lett. 2011 Aug;33(8):1523-38. Epub 2011 Apr 29.

72. Jos M. Raaijmakers, Irene de Bruijn, Ole Nybroe & Marc Ongena, Natural functions of lipopeptides fromBacillus andPseudomonas: more than surfactants and antibiotics, FEMS Microbiol Rev 34 (2010) 1037–1062.

73. Ongena M, Jacques P, Bacillus lipopeptides: versatile weapons for plant disease biocontrol, Trends Microbiol. 2008 Mar;16(3):115-25.

74. Soberón-Chávez G, Lépine F, Déziel E, Production of rhamnolipids by Pseudomonas aeruginosa, Appl Microbiol Biotechnol. 2005 Oct;68(6):718-25.

75. Characterization of a Phenazine and Hexanoyl Homoserine Lactone Producing Pseudomonas aurantiaca Strain PB-St2, Isolated from Sugarcane Stem, Mehnaz, Samina, Deeba Noreen Baig, Farrukh Jamil, Brian Weselowski, and George Lazarovits, J. Microbiol. Biotechnol. (2009), 19(12) 1688–1694.

76. Genetic and Phenotypic Diversity of Plant Growth Promoting Rhizobacteria Isolated from Sugarcane Plants Growing in Pakistan, Mehnaz, Samina, Deeba N. Baig, and George Lazarovits, Microbiol. Biotechnol. (2010), 20(12), 1614–1623.

77. Comparative Genomics of Plant-Associated *Pseudomonas* spp.: Insights into Diversity and Inheritance of Traits Involved in Multitrophic Interactions, Joyce E. Loper, PLOS Genetics, 2012.

78. How to write a Master's Thesis, Yconne N. Bui, 1412957109, 2009