
A bioinformatics approach for conceptual genome mining

Dissertation
zur Erlangung des Grades
des Doktors der Naturwissenschaften
der Naturwissenschaftlich-Technischen Fakultät III
Chemie, Pharmazie, Bio- und Werkstoffwissenschaften
der Universität des Saarlandes

von
Srikanth Duddela
Saarbrücken
2015

Tag des Kolloquiums: 22. März 2016

Dekan: Prof. Dr.-Ing. Dirk Bähre

Berichterstatter: Prof. Dr. R. Müller
Ass. Prof. Dr. M.H. Medema

Vorsitz: Prof. Dr. C. Wittmann

Akad. Mitarbeiter: Dr. Y. Khatri

Diese Arbeit entstand unter der Anleitung von Prof. Dr. Rolf Müller in der Fachrichtung 8.2 Pharmazeutische Biotechnologie der Naturwissenschaftlich-Technischen Fakultät III der Universität des Saarlandes von August 2011 bis December 2015.

Acknowledgement

I am very grateful to many individuals who have supported my work and continually encouraged me through the writing of this dissertation. Thanks for cheering me up in hard times and sharing nice moments in good times. Without their time, attention, encouragement, thoughtful feedback, and patience, I would not have been able to see it through.

At first, I would like to express my deep gratitude to my adviser, Prof. Dr. Rolf Müller for his support and for giving me the opportunity to work in his group. I thank him for his supervision, inspiration and encouragement during the last years.

I would like to thank my co-advisor, Dr. Daniel Krug, for his inspirational and timely advice and constant encouragement over the last several years. I have learned a great deal from his unique perspective on research, his sharp insight on almost any issue, and his personal integrity and expectations of excellence. He has always been patient when explaining the concepts of chemistry, and has shared with me many witty jokes.

Sharath kumar Kondreddi is a wonderful and generous friend who has been through a lot and I admire his positive outlook and his ability to smile in any situation. I would like to convey my deep thanks for his valuable contribution for the successful completion of the projects. It would be been very difficult without your constant support and encouragement.

Thank you so much Nestor Zaburannyi, for your kind help and cooperation particularly during the development of BiosynML for antiSMASH. It would have been very difficult to manage several issues without your help. I am indebted for your selfless act of assistance.

I would like to thank my friends thanks to Martin Slawaski, Siva sankar Lingam, Aravind Pasula and Shyam sundar Rangapuram for supporting me during tough times. Martin, I learnt a lot from you in terms of punctuality, discipline and approach towards a scientific problem. Aravind anna and Siva anna, thanks for your constant encouragement. You were not only my friends but also teachers leading me to the path of success. You were always there when ever I need someone to cheer me up. Shyam anna, you were always there for me whenever I run for some help in academics right from my master studies.

Thanks to my friends Venu, Bhanu, Renuka, Sujith, Praveen, Jyothi, Chaithanya, Sam, Narendra, Sreekanth, Manoj, Hema Shekar and other member of pokermaniacs group. I can never forget our social gathering. They were relaxing, refreshing and motivating. Venu, I can always count on you in any situation, you never displayed any change in behaviour towards me even after achieving a commanding

position in life. Bhanu, Renuka and Sujith, you guys have great sense of humour and your comments and jokes during our gathering made it memorable. We had so much fun and nice time together.

I am deeply indebted to my parents Prasad Rao, Gayathri devi, wife Sravani and sister Yamini for their continuous encouragement, feedback and support all through my life. I thank you for always believing in whatever I did. I hope to make you happy and proud. These are the four people to whom I dedicate my achievements. Last but not least, I thank the Almighty for making my dream come true.

Vorveröffentlichungen der Dissertation

Teile dieser Arbeit wurden vorab mit Genehmigung der Naturwissenschaftlich-Technischen Fakultät III, vertreten durch den Mentor der Arbeit, in folgenden Beiträgen veröffentlicht.

Tilmann Weber, Kai Blin, **Srikanth Duddela**, Daniel Krug, Hyun Uk Kim, Robert Brucocoleri, Sang Yup Lee, Michael A. Fischbach, Rolf Müller, Wolfgang Wohlleben, Rainer Breitling, Eriko Takano and Medema H. Medema. antiSMASH 3.0—a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic Acids Research* (2015); 43:237–243.
doi: 10.1093/nar/gkv437.

Marnix H Medema, Renzo Kottmann, Pelin Yilmaz, Matthew Cummings, John B Biggins, Kai Blin, Irene de Bruijn, Yit Heng Chooi, Jan Claesen, R Cameron Coates, Pablo Cruz-Morales, **Srikanth Duddela**, Stephanie Düsterhus, Daniel J Edwards et al. Minimum Information about a Biosynthetic Gene cluster. *Nature chemical biology* (2015); 11(9):625–31.
doi:10.1038/nchembio.1890

Zusammenfassung

Die fortschreitende Verbesserung von Sequenzieretechnologien ermöglicht den Zugang zu einer stetig wachsenden Zahl von mikrobiellen Genomsequenzen. Gleichzeitig liefern bioinformatische Methoden ein immer besseres Bild des genetischen Potentials der Mikroorganismen für die Produktion von Sekundärmetaboliten. Die vorliegende Arbeit befasst sich mit der Entwicklung von bioinformatischen Werkzeugen um die Entdeckung, die Dereplikation und letztendlich die Charakterisierung von multimodularen Biosynthesewegen in mikrobiellen Genomen zu unterstützen. Kernstück des Ansatzes ist der „konzept-basierte“ Vergleich der Architekturen von komplexen PKS-, NRPS- und hybriden Genclustern, der sich auf Anordnung und Eigenschaften biosynthetischer Domänen stützt anstelle von Sequenzähnlichkeit. Das neu entwickelte Softwarewerkzeug, genannt BiosynML, wurde mit antiSMASH (dem de-facto Standard für die automatische Annotation von Biosynthesewegen) verknüpft und in eine bestehende Forschungsdatenbank (Mxbase) integriert. BiosynML Methoden wurden anhand der Biosynthesewege für 42 bekannte Naturstoffe in 71 myxobakteriellen Genomsequenzen getestet und auf öffentlich zugängliche Genome relevanter Mikroorganismen angewendet. Die Analyse von 1347 Biosynthesewegen aus den Genomen der Myxobakterien, darunter ein derepliziertes Set von 783 Typen, ergab eine nur minimale Überlappung zwischen Unterordnungen und ermöglichte die Abschätzung der Diversität an myxobakteriellen Sekundärmetaboliten-Genclustern.

Abstract

Recent advances in sequencing technology have set the stage for a steadily growing number of microbial whole-genome sequences. At the same time, bioinformatic analysis increasingly sheds light on the genome-encoded capacity of certain microorganisms for the production of secondary metabolites. This work describes the development of a bioinformatic toolkit to underpin discovery and dereplication efforts in a genomics-based workflow aimed at the characterization of multimodular biosynthetic gene clusters from bacterial genomes. Key to the “conceptual genome mining” approach implemented here is the comparison of pathways architectures represented by arrangement and properties of domains in complex PKS-, NRPS- and hybrid pathways rather than resorting to DNA- or protein-level sequence similarity. The new analysis framework named BiosynML toolkit was interfaced to antiSMASH, the de-facto standard for automatic annotation of biosynthetic pathways, and integrated with an existing in-house research database system (Mxbase). BiosynML methods were tested using 42 characterized pathways from 71 myxobacterial genomes and also applied to publicly accessible genomes from relevant microbial taxa. BiosynML tools were ultimately used to create an overview of 1347 pathways of which 783 distinct models were identified. This analysis revealed minimal overlap between suborders and enabled the tentative estimation of myxobacterial secondary metabolite gene cluster richness.

Contents

ACKNOWLEDGEMENT	V
VORVERÖFFENTLICHUNGEN DER DISSERTATION	VII
ZUSAMMENFASSUNG	IX
ABSTRACT	XI
1 INTRODUCTION	1
1.1 NATURAL PRODUCTS FOR DRUG DISCOVERY.....	1
1.2 ASSEMBLY LOGIC OF SECONDARY METABOLISM: MODULAR PATHWAYS	3
1.2.1 <i>Polyketide synthases – PKS</i>	3
1.2.2 <i>Nonribosomal peptide synthetases – NRPS</i>	6
1.2.3 <i>NRPS-PKS hybrid pathways</i>	9
1.3 ROLE OF GENOME MINING FOR NATURAL PRODUCTS DISCOVERY.....	10
1.4 MYXOBACTERIA AS PRODUCERS OF NATURAL PRODUCTS	13
1.5 MYXOBASE / MXBASE SERVER: A COMPREHENSIVE CHEMICAL & BIOLOGICAL DATABASE.....	17
1.6 AIMS AND SCOPE OF THIS WORK: A NEW APPROACH TO THE (MYXOBACTERIAL) GENOME MINING CHALLENGE	19
2 MATERIALS AND METHODS	25
2.1 DNA SEQUENCING AND ASSEMBLY	25
2.2 MXBASE INFRASTRUCTURE.....	26
2.3 GENEIOUS FRAMEWORK	27
2.4 ANTISMASH: ANTIBIOTIC AND SECONDARY METABOLITE ANALYSIS SHELL	28
2.5 DATASETS USED IN THIS STUDY.....	29
2.6 PROGRAMMING LANGUAGE	33
2.6.1 <i>C# and Microsoft Visual Studio</i>	33
2.6.2 <i>Java and NetBeans Platform</i>	33
2.7 APACHE THRIFT	33
2.8 GRAPH VISUALIZATION: ZEDGRAPH.....	34
2.9 MYSQL DATABASE PLATFORM	34
2.10 EXTENSIBLE MARKUP LANGUAGE (XML).....	34
2.11 MUSCLE: ALIGNMENT SOFTWARE	35
2.12 BIOINFORMATICS FUNCTIONS	36

3	RESULTS AND DISCUSSION	38
3.1	BIOINFORMATICS FRAMEWORK FOR CONCEPTUAL GENOME MINING	40
3.1.1	<i>BiosynML Language and container</i>	40
3.1.2	<i>Interfacing BiosynML to the “antibiotics & Secondary Metabolite Analysis Shell” (antiSMASH) ..</i>	44
3.1.3	<i>The BiosynML Geneious plugin</i>	45
3.1.4	<i>BiosynML Editor for manual creation of pathway models</i>	55
3.1.5	<i>Integration of BiosynML with Mxbase.....</i>	56
3.2	THE BIOSYNML ANALYSIS ENGINE	59
3.2.1	<i>Algorithms developed for conceptual genome mining</i>	60
3.2.2	<i>Usage scenarios and comparison of BiosynML methods.....</i>	69
3.2.3	<i>Influence of parameter settings on the outcome of the pathway comparison</i>	81
3.2.4	<i>Pathway query based on signature domains.....</i>	90
3.3	CONCEPTUAL GENOME MINING WITH NATURAL PRODUCTS SOURCES	97
3.3.1	<i>Overview of datasets used in this study</i>	97
3.3.2	<i>Targeted genome mining: identification of similar gene clusters</i>	99
3.3.3	<i>Architectural matching vs. sequence-based comparison</i>	104
3.3.4	<i>Genome annotation and dereplication analysis of biosynthetic gene clusters</i>	113
3.3.5	<i>Exposing the diversity of secondary metabolite pathways in myxobacterial genomes</i>	119
4	CONCLUSION AND OUTLOOK	137
5	LITERATURE CITED	ERROR! BOOKMARK NOT DEFINED.

1 Introduction

1.1 Natural products for drug discovery

For thousands of years, natural products proved to be a rich source of drugs and drug leads, playing a vital role for prevention and cure of various diseases (1). During the last century, the discovery of natural products is filled with stories of numerous lifesaving drugs produced by microorganisms. For example, Sir Alexander Fleming in 1928 discovered penicillin from *Penicillium notatum*, which became world's first industrially produced antibiotic and was widely used to combat infections (2).

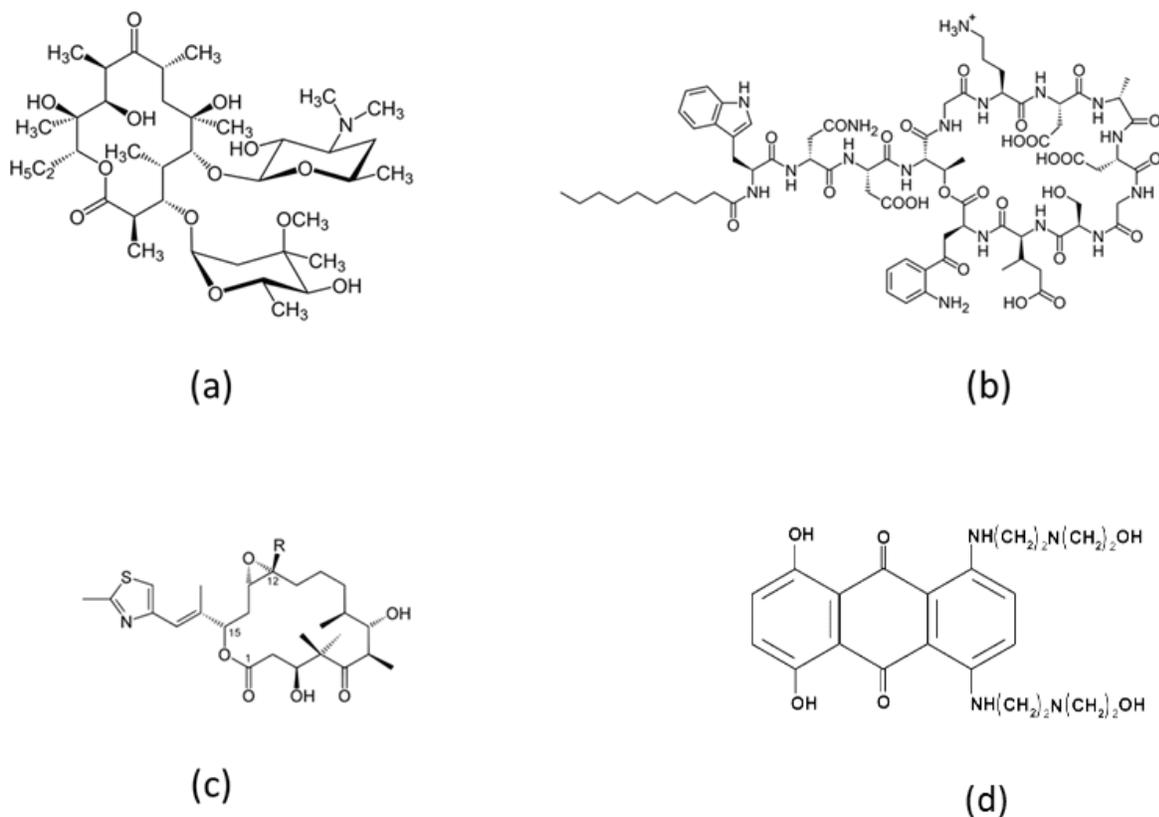


Figure 1.1: Drugs derived from natural products. (a) Erythromycin (macrolide antibiotic drug used to treat different types of infections caused by bacteria), (b) Daptomycin (antibiotic drug used to treat bacterial infections of the skin and underlying tissues) (c) Epothilone (the semisynthetic derivative Ixabepilone is on the market for anticancer treatment) (d) Mitoxantrone (antineoplastic agent).

The majority of the drugs available in the market are derived from natural products obtained from various sources such as plants and microorganisms (3, 4). For example, drugs like penicillin, erythromycin (Figure 1.1 a), daptomycin (Figure 1.1 b) are used clinically against bacterial infections (5–8); drugs like epothilone (Figure 1.1 c), doxorubicin, mitoxantrone (Figure 1.1 d) (9–12) are used against cancer. Other fields of application include natural product-derived drugs used as immune suppressants and to lower the cholesterol levels (13).

Although natural products from microorganisms are known as therapeutic agents for the treatment and prevention of diseases from late 1920s (14). The delay of clinical trials was caused by the lack of proper procedures to produce enough quantities of the pure products. Starting from this stage, an era followed which is characterized by a steady increase in the discovery of novel compounds (15, 16). However, natural products as a source for new medicines were largely abandoned by “big pharma” after the 1970’s mainly for economic reasons, leading to an “innovation gap”: basically no new structural classes of antibiotics were introduced until the 2000’s (17).

Enabled by progress in modern techniques such as next-generation sequencing and high-resolution mass spectrometry, strategies for drug discovery for pharmaceutical applications are currently in a revolutionary period (18). With the availability of automated instrument systems, robots and high-throughput screening (HTS) platforms providing powerful tools for screening large compound libraries in a cost-effective manner (15). Over the last two decades, development in HTS and analytical techniques in combination with genomics-based methods. This triggered new directions of natural product research, including studies of biosynthesis, revealing that genes responsible for biosynthesis of complex secondary metabolites are often located adjacent to each other (clustered) in microbial DNA that encode for polypeptides or proteins. Altogether, genes responsible for synthesis of secondary metabolite are encoded in a large gene cluster producing protein domains with defined functions. These insights into the molecular basis of natural product formation have changed the view of natural product research by enabling the emendation of known structures and prediction of novel compounds based on the gene sequences and generation of unusual compounds by combinatorial biosynthesis.

These methods facilitate the development of new drugs which are needed to control the pathogenic bacteria showing resistance to the effect of antibacterial drugs. These bacteria were reported as “ESKAPE” pathogens (*Enterococcus faecium*, *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Acinetobacter baumannii*, *Pseudomonas aeruginosa*, and *Enterobacter* species) (19). There are only a few natural products that pass through clinical trials and are approved as drugs. This generates a great demand for new compounds to fill the gaps in the pharmaceutical industry. Developing methods as such could ensure a steady stream of new drugs to meet the current needs.

1.2 Assembly logic of secondary metabolism: modular pathways

Secondary metabolites in general are chemical entities produced by but not involved in the normal growth of an organism. These compounds are often thought to be used as defences against competitors; plausible examples are compounds with antibacterial or antifungal activities. Advances in natural products biosynthesis research over the last 10-20 years led to an improved understanding of their biosynthesis including the organization of biosynthetic genes in so called gene clusters found as genomic islands in the chromosome of the producing organism. These gene clusters produce remarkable peptide secondary metabolites belonging to the class of nonribosomal peptide synthetase (NRPS) and polyketide synthase (PKS). As a result, several biosynthetic pathway which are multi enzymatic and multi domain megasynthases were identified. In addition, several bacterial species have been identified as the source of novel natural products using so-called genome mining approaches (20). Since the availability of automated tools for genome annotation, bacterial genomes are nowadays routinely investigated for the presence of PKS and NRPS clusters, exposing a large number of novel gene clusters with currently unknown function.

Compounds derived from PKS and NRPS biosynthetic machinery represent two large families of structurally diverse and complex microbial metabolites that include many potential drug leads. Understanding the 'assembly line' logic behind the formation of these compounds helps to develop strategies for the production improvement and targeted alteration of the metabolites associated with PKS and NRPS gene clusters (20). According to textbook biosynthetic logic, biosynthesis by PKSs and NRPSs is typically accomplished using acyl-coenzyme A monomers and amino acid building blocks in an assembly line fashion. Both of their biosynthesis are assisted by large multimodular proteins in which each enzymatic module catalysis one round of elongation and a variable set of modifications on a growing polyketide- or polypeptide chain, respectively (21, 22). To some extent, this knowledge allows to estimate the structures afforded by PKS and NRPS pathways based on bioinformatic analysis of genome information; however the correct prediction of precise molecular structures of products is not routinely possible.

1.2.1 Polyketide synthases – PKS

Polyketides are a class of secondary metabolites which are structurally complex organic compounds exhibiting wide range of biological properties. Biosynthesis of polyketides is usually accomplished through the decarboxylative condensation of activated dicarboxyl acid units resembling

fatty acid synthesis (FAS) (23, 24). Polyketide synthases (PKSs) are a group of multi-domain enzymes that are responsible for the production of polyketides.

Depending on their molecular architecture and enzymatic properties like the subunit organization and mode of synthesis, PKSs are historically subdivided into three groups (25). Type I PKS consist of multifunctional mega-enzymes organized into modules in which each active site is involved in the enzymatic processing of intermediates along the pathway as exemplified by the 6-deoxyerythromycin B synthase (DEBS) for the biosynthesis of reduced polyketides such as erythromycin A (21) where each successive incorporation of building block which is further reduced by keto reductase (KS) from ketone group to a hydroxyl group (Figure 1.2). They are further divided into iterative PKS and modular PKS. Iterative type I PKS are monomodular exhibiting a set of domains used in a cyclic fashion. Modular PKSs contain several individual modules and usually do not iterate reactions, although exceptions have been reported. Type II PKSs, also work iteratively, featuring each catalytic site on individual mono- or bi-functional proteins as exemplified by the biosynthesis of aromatic polyketides such as tetracenomycin C (26–28). Type I and type II PKS uses acyl carrier protein (ACP) to activate the acyl CoA substrates to transport the polyketide intermediates. Type III PKSs also known as chalcone synthase like PKSs, are homodimers where each monomer catalyzes a complex set of reactions including priming, extension, and cyclization iteratively to form polyketide products extensively utilizing acyl-coenzyme A instead of ACP-bound monomers, as exemplified by the RppA synthase for the biosynthesis of aromatic polyketides such as flaviolin (29–32).

Every PKS module consists of a set of core domains. A loading module generally contains an acyltransferase (AT) domain selecting the appropriate starter unit which frequently is an acetyl-CoA. The starter group is loaded on to the acyl carrier protein (ACP) domain on the starter module (Figure 1.2). The chain is then transferred from the ACP domain of the previous module to the Keto-synthase (KS) domain of the current module. The malonyl group is attached to a thiol of the current ACP domain catalyzed by the current AT domain. This thioester formed with the terminal thiol of a phosphopantetheine moiety that is posttranslationally attached to a serine of the ACP. The phosphopantetheine (ppant) has the function of a flexible arm, which carries the growing chain to the different catalytic domains that act on the biosynthetic intermediate (33).

Furthermore, optional domains such as ketoreductase (KR) which reduces the β -keto group to a β -hydroxy group, oxidation (Ox), dehydratase (DH) which chops off H₂O resulting in the α - β -unsaturated alkene, enoylreductase (ER) which reduces the α - β -double-bond to a single-bond, methyltransferase (MT) domains inducing α -methyl branches, modify the growing polyketide molecule. The last module

typically possesses a thioesterase (TE) domain to catalyse the hydrolysis of the polyketide chain from the final ACP-domain, resulting in the release of the final product from the enzyme (30).

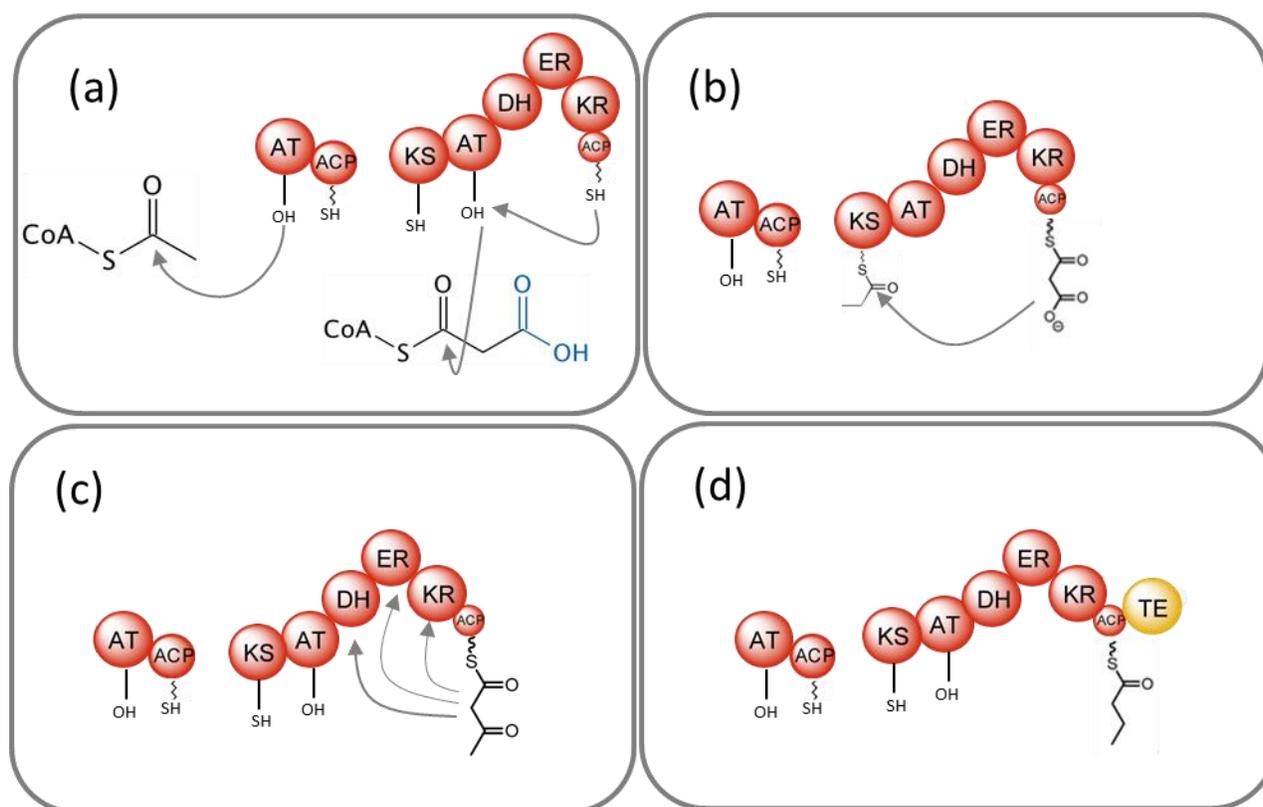
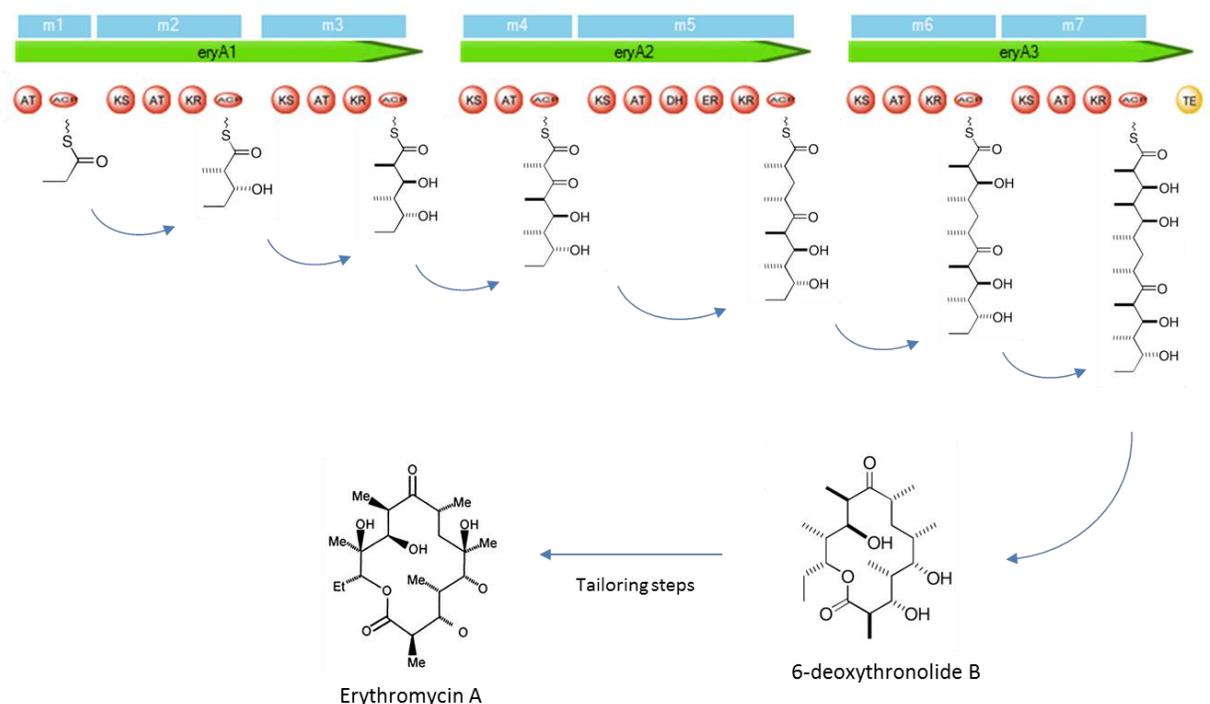


Figure 1.2: Schematic representation of modules involved in initiation and chain extension steps. (a) Selection of acyl-CoA by AT domain for chain building and transfer to ACP. (b) Acylation of starter unit by KS domain by first extension module. (c) Modification of keto group by DH, ER and KR domains. (d) Extension of reduced chain by successive modules and the final product is catalyzed by TE

Some bacterial biosynthetic pathways lack internal AT domains, which are complemented by freestanding AT-domain containing enzymes, so called trans-AT pathways (34). The trans-AT PKSs are highly diverse often containing previously unseen arrangements of enzymatic domains, thus usually resulting in poor biosynthetic assignment of clusters using standard collinearity rules, i.e. the assumption that the order of modules encoded in a biosynthetic pathway is reflected by the order of monomer incorporation or elongation (35). This can be exemplified by the gene cluster for the antibiotic Chlorotoniil, but more and more trans-AT pathway examples are now being reported (36, 37).

Several novel compounds have been assessed by molecular and biochemical studies including experiments such as loading module relocation, point mutations, module deletion, module insertion, domain swapping, domain inactivation and more (21, 38). The biosynthetic genes can be to some extent



transferred between pathways from different organisms has given rise to the thought that horizontal gene transfer (HGT) might also play a role in PKS evolution (39–41).

A common graphical representation for multimodular biosynthetic pathways is the “ball scheme” derived from annotation or bioinformatic prediction, which can be read like an “enzymatic activity string” in order to get an idea of the compound produced (Figure 1.3).

1.2.2 Nonribosomal peptide synthetases – NRPS

A conceptual alternative to peptide synthesis on ribosomes was first demonstrated in 1971 with the nonribosomal peptide synthetases (NRPS) of gramicidin S and tyrocidin (42). Today, these nonribosomal peptide synthetases are accepted as a “factory” for synthesizing many specialized peptides. Similar to PKS, NRPS are also multimodular enzyme organizations comprising a series of functional units, each responsible for the addition of a specific and often non-natural amino acid. The modules of NRPSs contain a minimum of three domains catalyzing a particular set of reactions in the incorporation of a monomer (Figure 1.4)(43–45).

Figure 1.3: Ball scheme representation for biosynthesis of erythromycin from strain NRRL2338 (Saccharopolyspora erythraea) along with the growth of polyketide chain by the addition of ketide group in each step. The blue line denotes the extension of individual modules.

In the loading step, the amino acid is activated by an adenylation (A) domain. A domains catalyze the ATP-dependent reaction, transforming the amino acid into an AMP-charged aminoacyl (46). Then, the activated monomer is loaded onto the peptidyl carrier protein domain (PCP) binding covalently with the thiol of a 4'-phosphopantetheinyl (Ppant) cofactor by forming a thioester, generating an enzyme-bound aminoacylthioester intermediate (33, 47, 48). This step enables the transport of the activated amino acids between the catalytic centers of the NRPS. In order to attach, transport substrates and intermediates, Ppant is post-translationally attached to PCP to a conserved serine residue, thereby converting the PCP domain into its holo form. This reaction is catalysed by a specialized external enzyme 4'-Ppant transferase (49–51). Elongation stage starts with loading of specific amino acid onto PCP of each module. The condensation domain is responsible for forming the amide bond between the amino acids of previous module to that of the current module (47). The extended peptide is now tethered to the PCP of the downstream module (Figure 1.4).

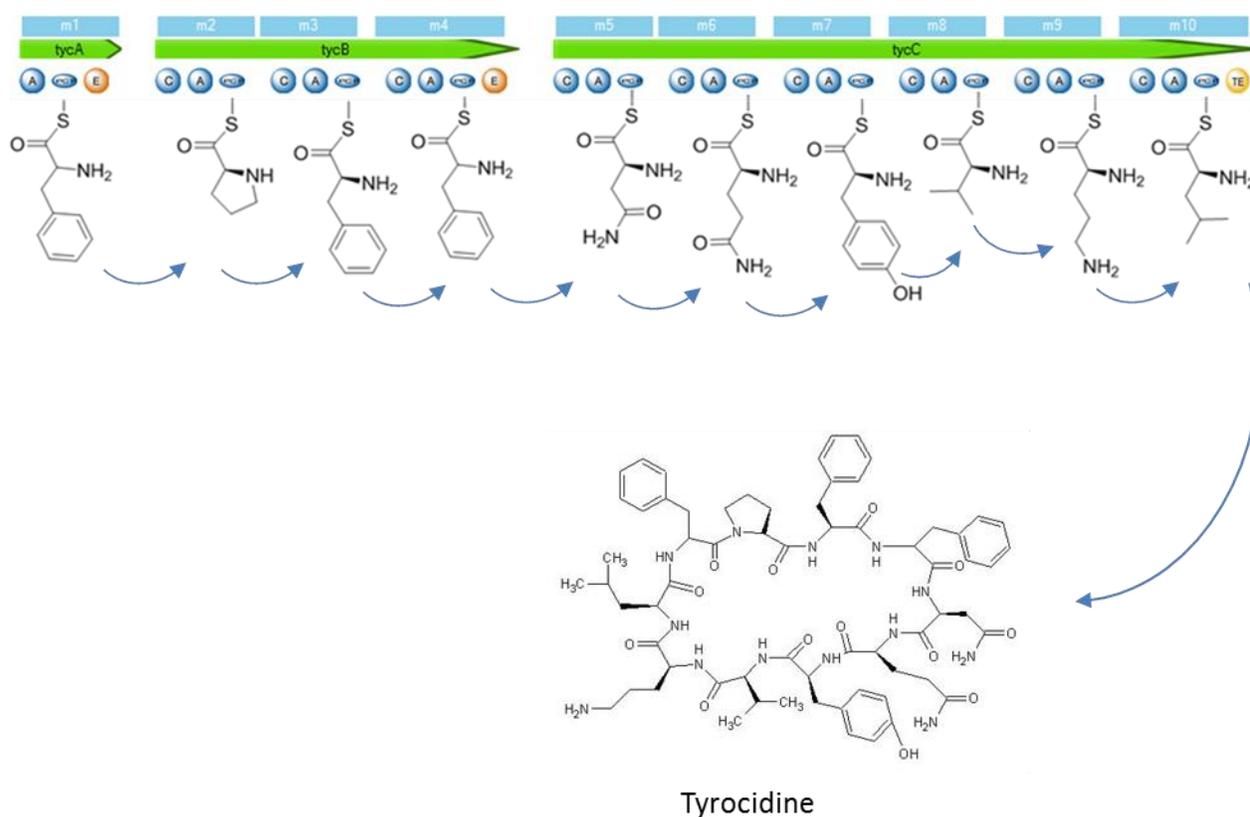


Figure 1.4: Domains and modules involved in the biosynthesis of Tyrocidine, a polypeptide antibiotic.

A module, solely containing A and PCP domains, is sometimes observed as the first module of the NRPS, while a module with thioester domain (TE) is considered as the terminating module responsible for hydrolysis or macrocyclization of the peptide chain from the PCP domain of the previous module, releasing peptide chain from synthetase (52, 53). In linear NRPSs, synthesis is initiated with the loading module, the sequence of modules acts as the template for adding a specific amino acid that dictates the resulting peptide's sequence, finishing it by TE domain releasing the peptide (43). The final-stage intermediate can also be released by cyclisation, as in the case of tyrocidine (Figure 1.4).

Along with these three domains (C, A, PCP), additional optional domains may be present within the modules, which are responsible for residue modifications. For example, epimerization (E) domains lead to formation of D-amino acids, *N*-methyltransferase (N-MT) control the methylation state of the peptide products, heterocyclization (HC) domains from thiazolines from cysteines, oxidation (Ox) domains are responsible for oxidation of thiazolines, reduction (Red) reduces thiazolines or oxazolines to thiazolidines or oxazolidines, respectively. Formylation (F) and heterocyclisation (HC) domains induce cyclization into thiazoline or oxazolines. All these additional domains facilitate NRPSs to synthesize varied number of biologically active peptides which are distinct from peptides synthesized by the ribosome (22, 54).

Nonribosomal peptides not only contain proteinogenic amino acids but also frequently contain non-proteinogenic amino acids. In addition, A-domains may exhibit some degree of substrate promiscuity (exemplified by incorporation of D-Phe or D-Trp in tyrocidines). These traits make it possible to obtain great structural diversity of NRPS-derived natural products, yielding wide-ranging biological activities and pharmacological properties. Nonribosomal peptides are often the backbone for the biosynthesis of antibiotics (e.g. daptomycin, vancomycin) and immunosuppressants (e.g. ciclosporin), of which some prominent ones are used commercially (55–57).

Since the first description of an NRPS assembly line, numerous bioactive molecules synthesized by NRPSs have been discovered and their biosynthetic pathways characterized (58). NORINE is a database extensively dedicated for collecting NRPS structures and linking them to biosynthetic gene clusters, equipped with computational tool for systematic study of these molecules across microorganism species (59). Increase in the number of biosynthetic clusters found in the public databases provides a clear advantage to investigate both chemical and enzymatic properties of those pathways. In addition, directed evolution and combinatorial approaches also contribute valuable information to a better understanding and exploitation of these complex machineries, but also leading to the identification of new metabolites. For example, a peptide siderophore from *Streptomyces coelicolor* named coelichelin

was identified from a NRPS gene cluster sequence through prediction of the structure and biochemical properties (60).

1.2.3 NRPS-PKS hybrid pathways

Although distinct mechanisms of condensation exist in PKS and NRPS from chemical aspects, there are significant similarities between them in terms of structure and functionality, including similarity between the carrier domains of both types which are post-translationally modified by a 4'-phosphopantetheine prosthetic group by a family of 4'-Ppant transferases (50, 51). Many secondary metabolites such as rapamycin (antifungal/immunosuppressive), epothilone (anticancer), bleomycin (antibiotic/antitumor) were found to be synthesized by the fusion of PKS and NRPS synthetases (PKS-NRPS hybrids) (61–63). For this, new strategies are developed to exploit the potential of microorganisms to produce important bioactive compounds. It is the peptide-polyketide metabolites studies that also gained spotlight for their potential utility in the field of combinatorial biosynthesis. Therefore, a pronounced challenge lies ahead in the studies exposing structure-function association and molecular organization features of hybrid pathways so that they may contribute to the production of new peptide-polyketide metabolites through combinatorial biosynthesis.

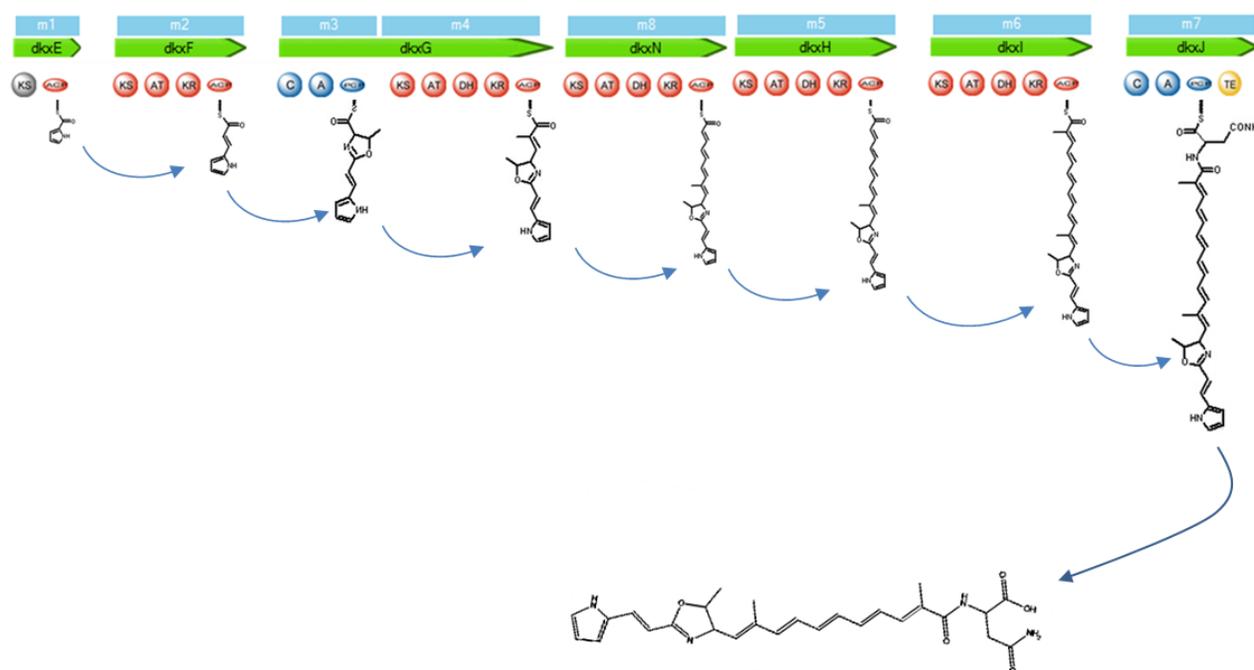


Figure 1.5: Ball scheme representation of domains and modules involved in the biosynthesis of Dkxanthe from strain DK1622 (*Myxococcus xanthus*)

Based on the domains functions, intermodular linkers and similar modular arrangement of NRPS and PKS modules, construction of chimeric PKS-NRPS hybrids is in principle possible (64). An example for a natural occurring NRPS-PKS hybrid pathway from myxobacteria is illustrated in Figure 1.5. In fact, the Dkxanthere molecule was only identified through a “gene to compound” approach following the inactivation of the assembly line in the producing strain *Myxococcus xanthus* DK1622 (65).

1.3 Role of genome mining for natural products discovery

The process of identifying natural products from microbes is usually long and laborious. Isolation of a pure bacterial culture is generally followed by cultivation in laboratory conditions and description of its secondary metabolites via physical and chemical methods, usually following bioactivity profiling and chromatographic purification of active fractions (Figure 1.6). However, some gene clusters from microbes may not be expressed at all when grown *in vitro* using traditional cultivation methods (66–68). An estimation of the “silent” part of biosynthetic potential is therefore required for the majority of bacterial strains. One of the methods used to evaluate the non-expressing clusters employs initial screening through degenerate PCR, however, not all biosynthetic gene types can be identified by this method. As such a wide range of natural products that are encoded in microbial genomes are yet to be discovered. Much faster and comprehensive is the process of sequencing the whole genome of a given organism, which can grant full access to a high volume of genetic information translating into a comprehensive view of the potential for biosynthesis of natural products.

Genome sequencing aims to determine the whole genetic information of an organism, including coding and non-coding sequences of nucleic acids (DNA) (69). Genome sequencing has been initiated since the seventies with the invention of the Sanger sequencing method (70–72).

Since then, this has led to the tremendous growth of various disciplines in the field of bioinformatics. Enabled with efficient techniques a large number of sequencing projects were finished, starting from bacterial genomes to human genomes (73–76). Later, whole genome shotgun sequencing was proven to be more cost effective and is now widely accepted (77, 78). Since the last century, many new high throughput sequencing methods have emerged such as to 454/Roche, Illumina, PacBio, Sequencing by synthesis and Ion semiconductor sequencing etc. (79–82) resulting in the deposition of large quantities of sequencing data from various organisms in public databases. Such data can be utilized to survey genomic information for enzyme encoding genes. This process of exploiting genomic sequence data has intensified rapidly, expanding the knowledge of genetic and biochemical aspects of secondary metabolite biosynthesis especially in microorganisms (83). It became clear that many microbial genomes

contain genes participating in the synthesis of complex bioactive products which are not associated with the known metabolites. In the case of microorganisms, these circumstances were first experienced with complete genome sequences of *Streptomyces coelicolor* A3 and *Streptomyces avermitilis* MA-4680 (66, 67, 84). These examples provide evidence that the availability of genomic sequences can empower the research of natural product producers with the help of bioinformatics approaches (i.e. genome mining).

Genome mining is now used as a tool for discovering important and complex secondary metabolites such as nonribosomal peptides (NRPSs), polyketides (PKSs) and terpenoids. In other words, the term genome mining basically comprises actions to exploit genetically encoded information in the gene clusters of microbial genomes for the discovery of new metabolites. Every family of natural products has a characteristic underlying pattern of proteins which is also evident on genomic level. This provides key for genome mining where conserved protein motifs and consensus sequences are used to identify loci of putative biosynthetic pathways.

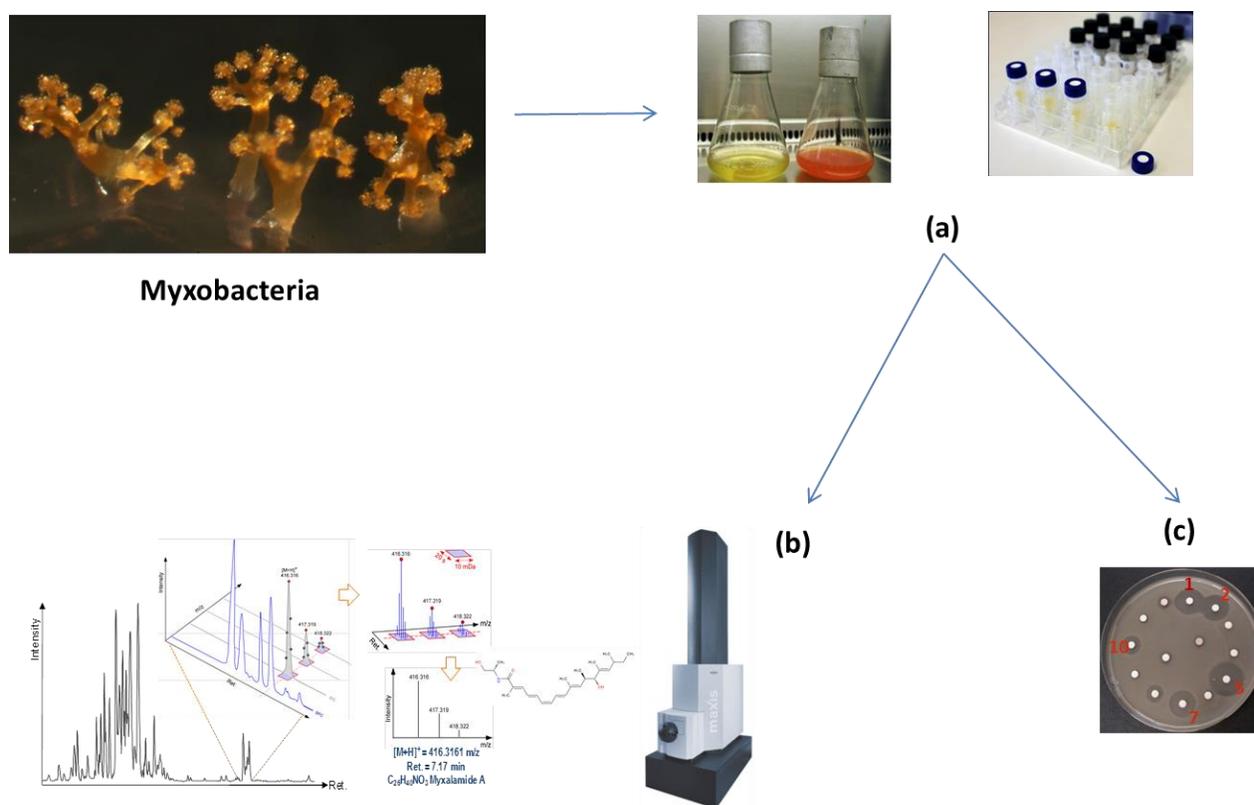


Figure 1.6: Schematic workflow of identifying natural products. (a) Cultivation and extraction of microbes. (b) Chemicals creening using insrtumentation analysis for derplication of known compounds.(c) Biological screening by bioactivity assays and discover novel compounds

With the increasing availability of genomic data from databases and the use of computational comparison tools, enzymes encoding genes that are involved in biosynthesis can be readily identified in

genomes. This information can be used to search for biosynthetic gene loci likely involved in assembly of new bioactive compounds. Several routes can be taken to pursue genome-mining. These include "gene to compound" correlation approaches using gene knockout experiments, statistics-based methods that resemble biomarker discovery, or "compound to gene" strategy where putative biosynthetic assembly lines are inferred from a molecule's structure. The different approaches to discover new natural product chemistry by genome mining are briefly explained in the following. Metabolite profiling is a well-used approach for identification of new metabolites and their corresponding pathways in the genomes. For example, the bioinformatics analysis identified 18 biosynthetic gene clusters from strain DK1622 belonged to myxobacteria. Until 2012 only five compound classes were known but by coupling principal-component analysis (PCA) with pre-processed LC-MS datasets resulted in three new secondary metabolites c506 (myxoprincomide), c844, and c329 (85). Similarly, compounds such as aurafuron A and B from myxobacterium *Stigmatella aurantiaca* (86, 87), the iron chelator coelichelin from *Streptomyces coelicolor* (60) and several other compounds were isolated by a comparative metabolic profiling genome mining approach. These results unarguably demonstrate the importance of genome mining for the identification of new secondary metabolites.

Genome mining is supported by numerous techniques providing hints for the structure prediction or at least estimation of compound class, although statements about associated bioactivity cannot usually be made (88). Moreover, precise structure prediction is also hindered by the fact that new biosynthetic logic is frequently encountered in the biosynthesis of microbial natural products by PKSs and NRPSs machinery (89–91).

However, advances in the field of predictive bioinformatics tools made it possible to classify genes responsible for secondary metabolite production according to their functional specificities. As the prime example, several models have emerged that can predict substrate specificity of adenylation (A) and acyltransferase (AT) domains responsible (Figure 1.2) for building block selection in each module of NRPS and PKS assembly lines.

Substrate specificity predictions along with the collinear enzymatic logic in modular PKS and NRPS pathways (although deviations are possible and notably frequent in myxobacteria) provide powerful tools, which in some cases can predict many of the structural features of the products of novel modular PKS and NRPS systems from their sequences. However, the predictions are not always reliable as many non-linear enzymatic assembly logic exist, such as iterative domain or module usage or module skipping both present in PKS and NRPS pathways which makes it complex for predicting the structure. Furthermore, for pathways where enzymatic action is unknown, prediction of structure is highly error prone.

Several bioinformatics tools are being developed for the prediction of secondary metabolite gene clusters including ClustScan (92), SBSPKS toolbox (93) and CLUSEAN (94) are available. Among all, the most user friendly and convenient tool in this field is currently antiSMASH 3.0 (95). It is a widely used tool that comes close to a de-facto standard for the automated genome-wide analysis of bacteria, fungi, and archaea for biosynthetic gene clusters and offers various modules for analyzing the pertinent pathways. This freely available tool enables rapid annotation studies that can be used as preliminary basis to prioritize species of interest by their biosynthetic potential, as well as helping with gene cluster annotation. However, the antiSMASH (95) could only provide the predicted secondary metabolite pathways that the genome is capable of producing. It does not provide any related information of the predicted biosynthetic pathways in terms of similarity to that of the existing characterized pathways. This leaves the natural product researcher to perform exhaustive search for the identification of known genes clusters from the pool of gene clusters reported. Currently the most widely used for the process of identifying known gene clusters is through sequence based approaches such as BLAST (96). These sequence based tools become complex and probably hard to analyse with increase in amount of sequenced data. In cases, where a biosynthetic pathway is hypothesized by researcher upon observing the novel compound extracted, it is impossible to use these sequence based tools for the identification of similar gene clusters without the availability of sequence information. Hence, new tools are needed which can efficiently analyse and generate dereplication library of biosynthetic pathways with or without the presence of sequence information.

1.4 Myxobacteria as producers of natural products

Natural products are important as leads for drug development, and the traditionally established producers of secondary metabolites include many plants, fungi and various bacteria, like, for example, the long known actinomycetes. More recently the myxobacteria have been added to this list, and are now also considered as an established source for natural products (97, 98).

Myxobacteria are a group of soil-living bacteria which are classified under the delta-subgroup of proteobacteria (99). They were first isolated by Roland Thaxter in 1892 recognizing them as an independent group. Since their discovery, this group of bacteria has fascinated scientists over generations due to their striking characteristics, mainly their distinct life cycle and social behaviour. Additionally, unique is their ability to aggregate into multicellular fruiting bodies when lack of nutrients is encountered (Figure 1.7). They move by gliding over the surface and are found to be ubiquitous in nature (100). Myxobacteria have relatively large genomes when compared to other bacteria. For example, the

circular genome of *Sorangium cellulosum* strain So ce56 contains 13,033,779 base pairs, which makes it one of largest bacterial genomes sequenced to date (99). Myxobacteria produce secondary metabolites with structural elements that are not commonly produced by other microbes such as unusual hybrids of polyketides and non-ribosomally made peptides (86, 101) (Figure 1.7). Many of these myxobacterial compounds also exhibit structural features which were novel at the time of their discovery (100).

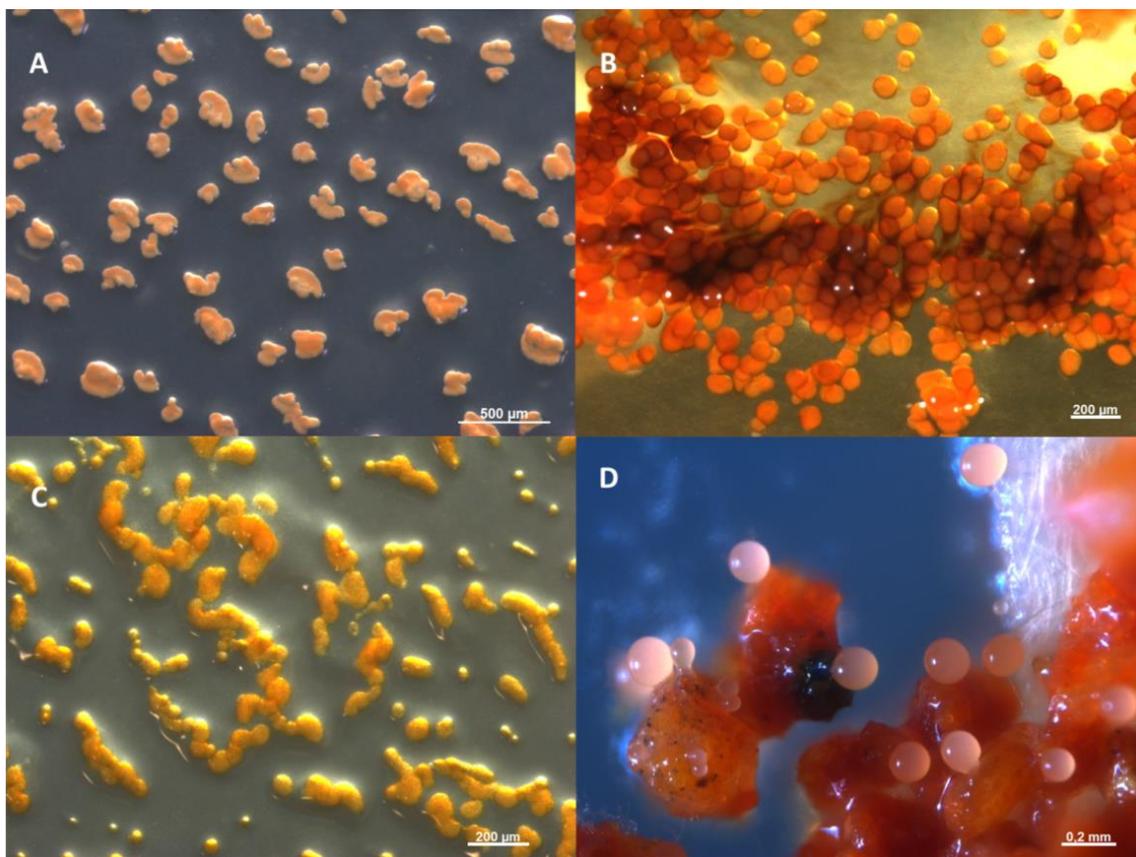


Figure 1.7: Stereophotomicrographs of myxobacterial fruiting bodies. A) *Corallocooccus* sp. on agar. B) *Cystobacter* sp. aggregates of sporangioles C) *Pyxidicoccus* sp. golden yellow color sporangioles D) *Myxococcus* sp. on soil crumb and at the edge of filter paper. Photos courtesy: Ronald Garcia.

Over the last 30 years, myxobacteria have been shown to produce compounds exhibiting a wide range of biological activities. For example, epothilones produced by the myxobacterium *Sorangium cellulosum* strain So ce90, show anti-tumor activity (Figure 1.8 a). Their analogues are used for the treatment of cancer, one such analog is the “ixabepilone”, which has recently been approved by FDA (U.S. Food and Drug Administration) and is now in use for the treatment of breast cancer (102). Other secondary metabolites such as etnangien (Figure 1.8 b), corallopyronin (Figure 1.8 d) and ripostatatin (Figure 1.8 c) from myxobacteria strains *Corallocooccus coralloides* Cc c127 and *Sorangium cellulosum*

Soce377 are observed to possess antibacterial activity due to their function as inhibitors of RNA polymerases (97, 103). Argyrin (Figure 1.8 g), Cystobactamids (Figure 1.8 i) extracted from the *Archangium gephyra*, *Cystobacter sp.* Cbv34 is a new myxobacterial metabolite exhibiting activity against gram-negative pathogens. Soraphen (Figure 1.8 h) is an immunosuppressive cyclic peptide extracted from *Sorangium cellulosum* Soce26, exhibiting the ability to inhibit T-cell independent antibody formation by murine B cells (104–106), while myxochelin (Figure 1.8 e) and nannochelin (Figure 1.8 f) are two similar compounds extracted from the myxobacteria *Angiococcus disciformis* and *Nannocystis exedens*, respectively, which have iron-chelating activity (107, 108).

Over 9,000 myxobacterial strains were isolated to date by workgroups at Helmholtz Center for Infection Research, Braunschweig and at the Helmholtz Institute for Pharmaceutical Research, Saarbruecken. In-house data indicate that these strains yielded around 900 compounds belonging to more than 140 compound families (although numbers cited in the literature are lower).

Myxobacteria are considered as “metabolite factory” with novel types of chemical structures and unique mode of action. However, since these microbes are difficult to isolate and are slow growing (109). In addition, the amount of secondary metabolites in cultures is often low which might be because of product degradation or end product inhibition (110). This might be one of the reasons that there are several unexplored metabolites from the strains of myxobacteria which are often cultivated in the laboratories. With respect to their genome size, it is thought that network is essential to produce chemical substances enhancing the survival and competitiveness of both the individual and the population (111). This can be observed at the genetic level of myxobacteria involved in secondary metabolism. Exceptionally large genome of the myxobacterium *Sorangium cellulosum* Soce56 has around 20 secondary metabolite loci reported, and thus should have more than the reported compounds to be discovered (99). *Mycococcus xanthus* DK1622 has around 18 secondary metabolite gene clusters accounting for around 9% of its genome (112). The high abundance of secondary metabolite biosynthesis-related gene clusters in the genomes of these strains indicates their genetic potential for the production of secondary metabolites which exceeds far more the number of previously reported compound classes from each strain. This finding implies that there is a need for improved analytical methods for the detection, identification and characterization of myxobacterial secondary metabolites.

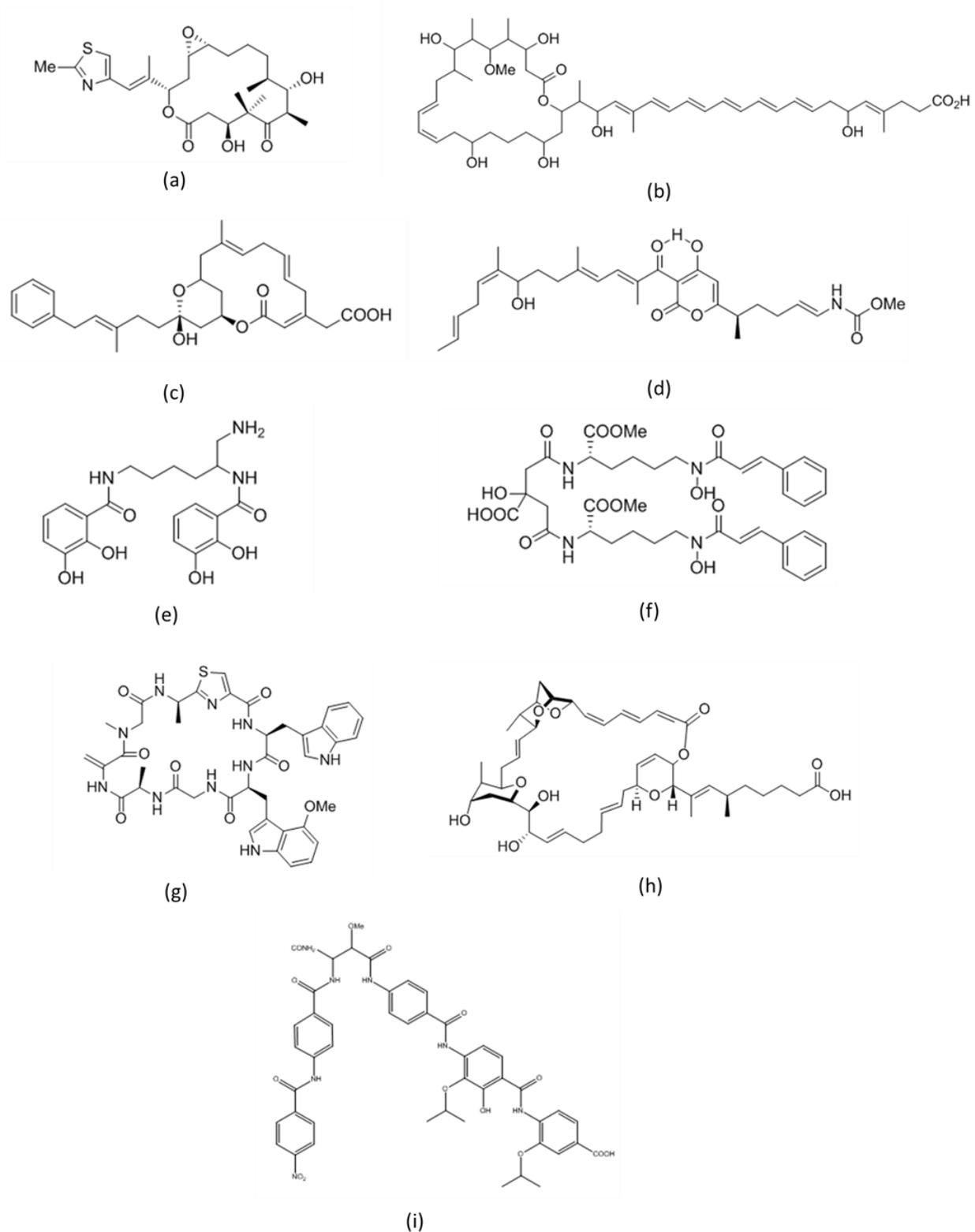


Figure 1.8: (a) Epothilon, drug with anti-tumor activity (b, c, d) Etnangien, Ripostatin and Corallopyronin possess antibacterial activity. (e, f) Myxochelin and Nannochelin are iron-chelating agents (g, h) Argyrin, and Soraphen possess T-cell regulatory abilities, (i) Cystobactamide is a new myxobacterial metabolite exhibiting activity against gram-negative pathogens.

1.5 Myxobase / Mxbase server: a comprehensive chemical & biological database

As the increase of the availability of sophisticated technology, multi-disciplinary research in the field of biology is quickly transforming it into data-rich science generating enormous amounts of information, steadily increasing the need for storing, analyzing and communicating this information. For this, a database project named "*Myxobase*" for collecting and linking chemical and biological information about known myxobacterial natural products and the producing strains has been pursued for five years at the Institute of Pharmaceutical Biotechnology of Saarland University and the Helmholtz Institute for Pharmaceutical Research Saarland (Group of Prof. Dr. Rolf Müller). This is a typical client-server application where the authorized user can add or update the information and the system is multi-user enabled. Contributing to this framework by improving existing-and adding new functionality was also an integral part of the work presented here.

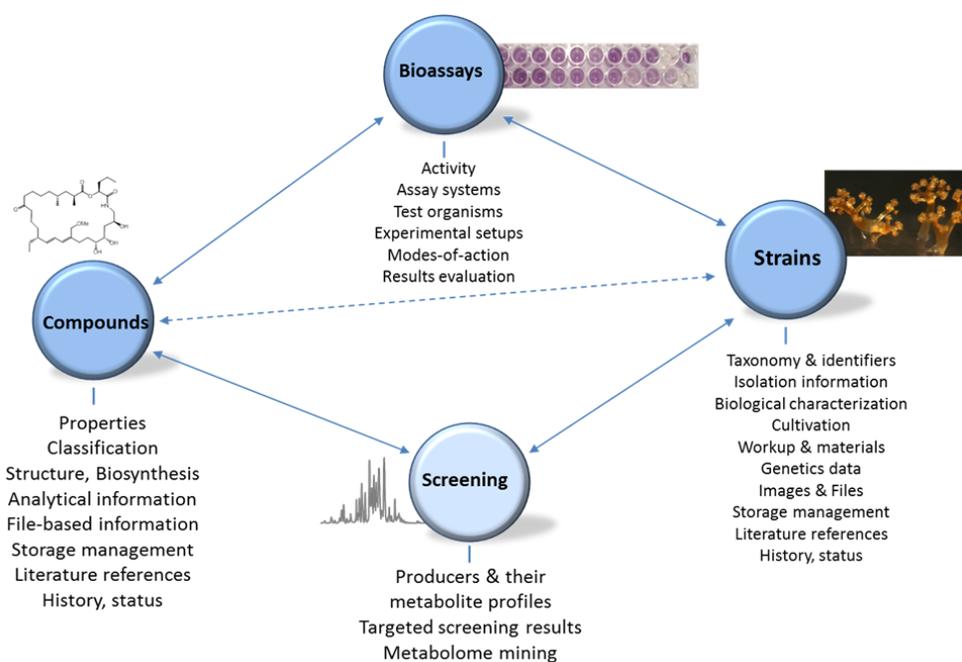


Figure 1.9: Organization and information content of Myxobase

Mxbase is a system for retrieving, storing and analysing of chemical and biological experimental data in a comprehensive multi-part database. Researchers can use the database to expand and improve the discovery process of novel microbial natural products. This also provides functions for extracting only that information needed to answer a specific biological question. *Mxbase* is not only a powerful approach to collaborative knowledge management in a multi-disciplinary lab. It integrates tightly with instruments-based analytical workflows, and the available tool set expands continuously. The *Mxbase*

concept covers diverse techniques needed e.g. for the characterization of new microorganisms, screening and dereplication, chemical structure elucidation, mass spectrometric analysis, bioactivity testing, genome- and metabolome mining (contributed by the present work) and many other methods (Figure 1.9).

The concept behind the *Mxbase* system (Figure 1.10) is to supply a combined framework for knowledge management and analysis workflows centred on the discovery of natural products from large numbers of biological sources. Technically, *Mxbase* can be divided into:

- the client software - *MxbaseExplorer* - which most users interact with for data deposition, information retrieval and specific data analysis procedures
- and the *Mxbase* server, which hosts an extensive database backend and additional analysis functions used by the client software, but also accessible to specialized data mining procedures using standard tools such as KNIME, Matlab and R.

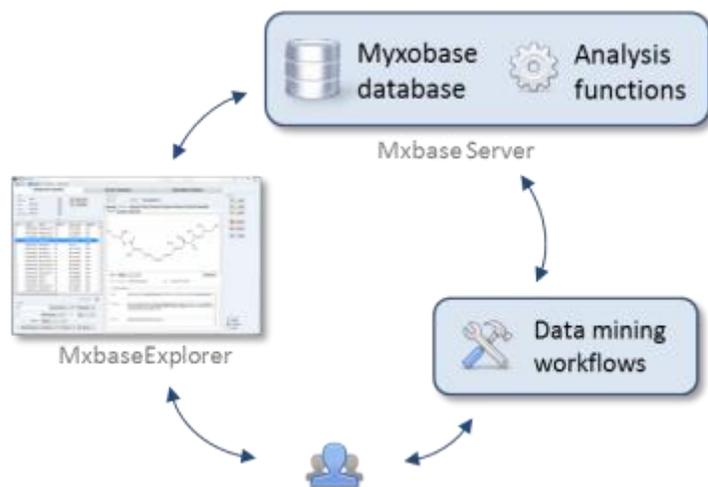


Figure 1.10: Framework of *Mxbase* demonstrating the workflow of various components involved

In the course of this work, important additions to the *Mxbase* project have been made in the fields of genetics data, archiving and presentation of biosynthetic pathways, as well as (most importantly) the management and comparison of biosynthetic pathways datasets from myxobacterial strains.

1.6 Aims and scope of this work: a new approach to the (myxobacterial) genome mining challenge

Initial situation and prerequisites

In light of the latest advances in whole-genome sequencing technologies, enabling the bioinformatic mining of increasing numbers of large myxobacterial genomes for the presence of biosynthetic pathways becomes an important issue for successful genomics-based discovery of new natural products and biochemical processes..

Myxobacterial genomes contain a wide variety of secondary metabolic gene clusters encoding polyketide synthases (PKSs), non-ribosomal peptide synthetases (NRPSs), or PKS/NRPS hybrids, besides other types of pathways. Thus, they offer great potential for the discovery of new bioactive metabolites. Different strains of myxobacteria have already been shown to produce compounds which are of potential clinical value. The rapid decrease in the cost of genome sequencing now allows the detection of hundreds or even thousands of gene clusters encoding the biosynthetic machinery for such compounds. However, laboratory research cannot keep pace with the speed of genomics-based discovery, as the experimental characterization of each gene cluster is often laborious and time-consuming. It is therefore important to prioritize resources for the investigation of pathways. From the point of view of natural products discovery, the most interesting gene clusters are those encoding pathways which are: i) responsible for the production of a compound with known structure exhibiting potent biological activity, or ii) similar to a previously characterized biosynthetic pathway so that an assumption about the molecule class likely produced can be made, or iii) uncommon or even unique pathways likely to produce novel chemistry. As a consequence, effective *in silico* identification and comparison of known and predicted biosynthetic pathways within genomes is essential for the efficient mining of the genomic richness available. Similar to the dereplication of known chemical entities in complex natural product-rich extracts, the dereplication of known biosynthetic concepts (also termed “models” in the following) and the underlying gene clusters has to be achieved in a genomics-based discovery workflow. The methods to be developed for this purpose have to take into consideration that there are different sources for- and ways of generating biosynthetic models (see Figure 1.11)

1. Automated genome analysis, where large numbers of reads obtained from sequencing are assembled through assembler algorithms resulting in contigs or scaffolds or closed genomes. These are commonly submitted to antiSMASH, a software pipeline to predict and annotate secondary metabolite pathways. As it is the quasi-standard in the field, it was one requirement that newly developed tools in this project must be able to interface with the output of the antiSMASH analysis engine.

2. Experimental characterization of biosynthesis, where a biosynthetic model is created from the structure and matched with the *in-silico* interpretation of a confirmed biosynthetic pathway (genes already available). Such verified pathway-compound assignments are at the heart of a biosynthetic model database to be developed for the purpose of biosynthetic gene cluster dereplication.

3. Starting from a newly discovered compounds, where a hypothetical biosynthetic pathway is designed manually for the freshly elucidated structure through sequential biochemical considerations, in a way that could be termed “retro-biosynthetic analysis” (no genetic information available yet). It is an additional requirement, that the toolbox to be developed in this work has to be able to accommodate and operate also with such “sequence-free” biosynthetic pathway models.

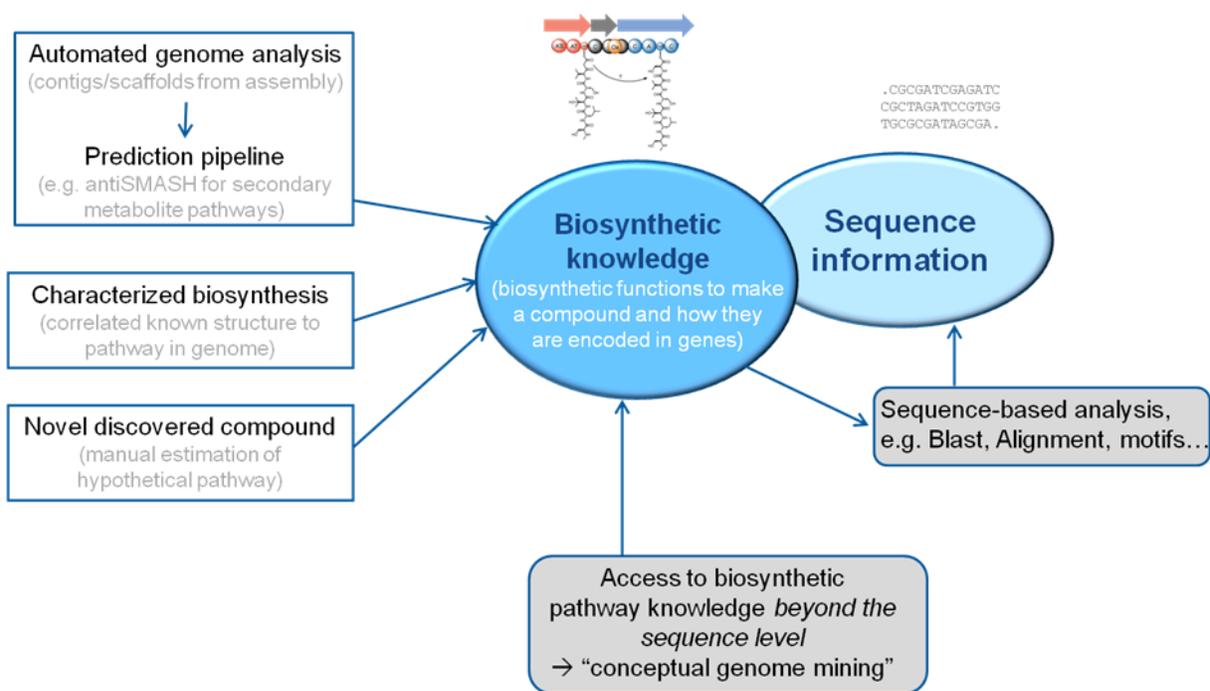


Figure 1.11: Methods of generating biosynthetic models

Challenges and strategy

Current approaches for the comparison of biosynthetic gene clusters are mostly using sequence-based methods. Sequence alignment and comparison tools, such as BLAST and HMM, are the commonly-used, traditional methods to identify homologous genes. These methods assume that sequences exhibiting significant similarity share common ancestry i.e. is homologs. However, the complex and to date not exhaustively understood mechanisms driving the evolution of multi-modular biosynthetic pathways imposes difficulties on sequence-based approaches. Furthermore, the performance of

sequence-based tools can quickly become a serious bottleneck when it comes to compare high numbers of genes/pathways present in elevated numbers of sequenced genomes. For example, if the following myxobacterial genome data availability were to become a reality within few years from now (which is not far-fetched), it is foreseeable that a “real-time” or “on the fly” analysis of this data volume is hardly feasible unless tremendous computational capacity would be spent even for relatively simple analysis scenarios:

- up to 200 myxobacterial whole-genome sequences of 9-18 Mbps
- where each encodes 10 to 30 secondary metabolite biosynthetic pathways
- the gene cluster for each pathway has 1 to >20 genes with up to > 50 domains
- each domain has multiple biochemical properties / predictions inferred from the primary sequence

In addition, much of the valuable information for pathway comparison in light of the produced structures does not lie in the primary sequences and their similarity, but is derived from the “enzymatic activity string” that results from bioinformatics analysis of the genomic input. These data include (but are not limited to) domain types and module composition, operon organization, domain order and predicted substrate specificity of monomer-activating domains.

Thus, the undertaking described in this work was met with a challenge to relate/compare the domain pattern of biosynthetic models with and without connection to genome sequences, linked to various layers of meta-information about building blocks, linkage, and enzymatic modifications that are expected to occur during the process of biosynthesis. These requirements motivated the development of a new approach to the comprehensive biosynthetic concepts-based mining of myxobacterial genomes in this study. Thereby, the genomic, chemical and biosynthetic knowledge which has accumulated in the workgroup over the past fifteen years (reflected largely in myxobase) plus genomic information from publicly accessible sources served as input for tool development and as test dataset for the performance evaluation of developed methods.

Three basic usage scenarios could be envisaged for this bioinformatics framework:

1. Matching a single pathway, e.g. an experimentally characterized secondary metabolite gene cluster or a speculative biosynthetic model from retro-biosynthetic analysis, against the multitude of pathways present in the database (which could themselves be pathways found in auto-annotated genomes, or discrete curated pathways manually deposited in the database). This search aids the identification of potential sources encoding a gene cluster for a newly discovered compound, and provides alternative sources for a gene cluster under investigation.

2. Assigning to each gene cluster found in a newly sequenced genome a plausibly close relative from the database, which helps the dereplication process and basically translates into a sequential procedure as in (1). When a compound produced by a reference gene cluster is already known, the link between gene cluster and putatively produced structure or compound class can be rapidly made.
3. Creating a comprehensive overview of biosynthetic pathway diversity in the database through a “match all against all” approach, where the aim is to construct a comparative map of gene cluster similarity. This multiple relation in turn may be used to support large-scale analysis such as phylogenetic distributions and the study of pathway evolution across entire microbial taxa. The overview obtained also enables the estimation of species richness in terms of biosynthetic pathways and may reveal insights into the degree of “pathway uniqueness” within a given clade of producers.

Implementation and deliverables

The aforementioned concepts and prerequisites served as a guideline for bioinformatic tool development throughout the present work and markedly influenced the directions taken during implementation. Since “everyday usability” of newly established tools and analysis workflows by scientists working in the institute’s research environment was desirable so that the new methods provided can unfold their impact on scientific advancement, the project also comprised significant software engineering efforts. The key components of the implementation and deliverables which were made available in the course of this work are briefly highlighted in the following (see also Figure 1.12).

- Direct interfacing to the output of the antiSMASH annotation pipeline, which required the integration of a new module into antiSMASH to generate structured results representations. The developed functionality has become a standard output module since the release of antiSMASH version 3 (95).
- Development of an xml-based container format named “BiosynML” to shuttle biosynthetic models and all their associated meta data, with or without sequence information, between different parts of the envisaged analysis workflows.
- Visualization and editing tools allowing for manual inspection and curation of annotated pathways by scientists. This was implemented in the form of a “BiosynML plugin” for the widely used Geneious bioinformatics software, allowing making use of an already well-developed and familiar graphical frontend as well as re-using certain standard sequence-based functionality from the Geneious software package. In addition, the plugin enables direct submission and result retrieval

to/from the antiSMASH web service (public or in-house). The BiosynML plugin for Geneious has been released to the general public, coordinated with the publication of antiSMASH version 3 (95).

- Addition of BiosynML import functionality to the in-house research database system Mxbase, including the creation of all required data schemes and software components to develop Mxbase towards full support for management and analysis of genomic data and biosynthetic pathway models.
- The bioinformatic core of the project: development of algorithms for matching and comparing biosynthetic pathways and their implementation within a remote-procedure framework, based on the Apache Thrift technology. This approach enables future re-use of the developed BiosynML analysis engine for various application scenarios, including the possibility for its integration into analysis workflows different from the one established in this study.
- Parametrization, critical evaluation of analytical performance and optimization of developed algorithms using a variety of real-world test cases.
- Implementation of user interface functionality allowing users of Mxbase to conduct analysis using the BiosynML engine, including the formulation and submission of jobs to the analysis queue as well as graphical results representation and evaluation. These components were continuously built into the MxbaseExplorer software (the graphical frontend to the Mxbase system) since version 3.
- Evaluation of the BiosynML concept in terms of its inter-operability with the MIBiG (“minimum information on biosynthetic gene clusters”) initiative (113), a recent community-supported move towards standardizing the basic information needed to describe secondary metabolite biosynthesis pathways. This included the tentative implementation of an adaptor to “bridge” between the information-rich BiosynML containers and the evolving MIBiG database.

Following the development and implementation phase, the newly created tools and workflows were applied to the analysis of biosynthetic pathways from myxobacterial genomes available at the institute, as well as using pathways and genomes from publicly accessible sources. For this purpose a collection of 35 manually curated myxobacterial secondary metabolite gene clusters was compiled and 71 myxobacterial genomes from ongoing sequencing projects were used to populate the test database. Several scenarios for conceptual genome mining were subsequently run in order to test-drive the BiosynML analysis framework and to evaluate analytical performance of the method. Furthermore, the power of the conceptual approach to generate an overview of pathway diversity in the database was explored and the method was compared to conventional sequence-based genome mining strategies.

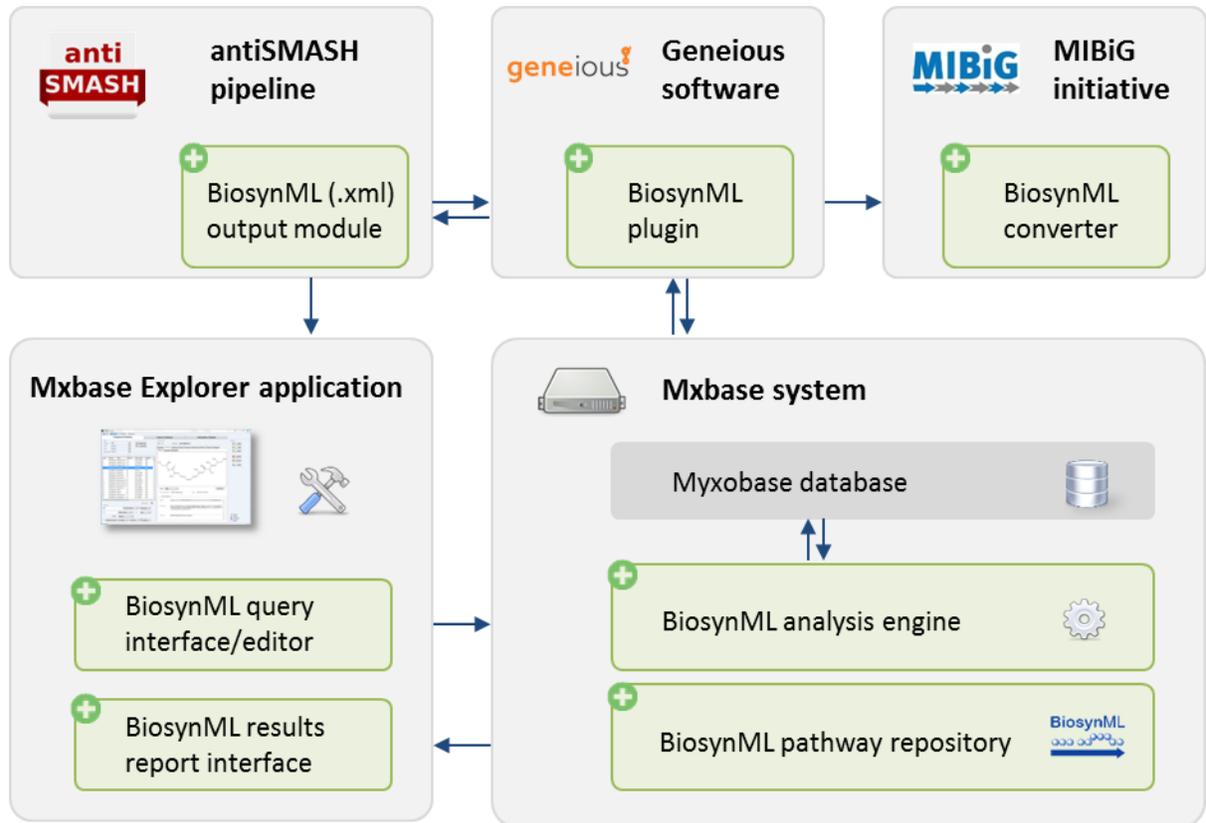


Figure 1.12: Schematic overview of the BiosynML framework and interoperability of its components with publicly available resources (antiSMASH pipeline, Geneious software suite, MIBiG initiative) and in-house technology (Mxbase system, Myxobase). Green boxes highlight contributions from the present work.

2 Materials and methods

2.1 DNA sequencing and assembly

Accurate determination of the primary DNA sequence is crucial in the field of natural product genome mining and discovery of novel compounds. DNA sequencing of bacterial strains used by us included massively parallel next-generation sequencing (NGS) technologies (Solexa/Illumina, Roche/454) and recently developed third generation sequencing technology (TGS) by Pacific Biosciences. The third generation sequencing technologies differ from NGS mostly by the ability to sequence single,

Strain	Technology	Assembly software / algorithm	Outcome, contigs
Chondromyces crocatus MSr (Cm_c5)	PacBio	SMRT Portal 2.1.1 / HGAP 2	1
Corallococcus coralloides MCy6431 (Ccc1071)	PacBio	SMRT Portal 2.2.0 / HGAP 3	1
Pyxidicoccus sp. MCy9557 (SBCy002)	PacBio	SMRT Portal 2.2.0 / HGAP 3	1
Sorangium cellulosum MSr8242 (Soce26)	PacBio	SMRT Portal 2.2.0 / HGAP 3	1
Sorangium cellulosum MSr1566 (Soce836)	PacBio	SMRT Portal 2.2.0 / HGAP 3	1
Sorangium cellulosum MSr8412 (Soce1128)	PacBio	SMRT Portal 2.1.1 / HGAP 2	1
Sorangium cellulosum MSr7282 (SoceGT47)	PacBio	SMRT Portal 2.1.1 / HGAP 2	1
Myxococcus virescens ST200611	454 Roche	Newbler unspecified version / Overlap Layout	987
Myxococcus xanthus MCy8278 (Mxx48)	Illumina	Abyss pe 1.3.6 / K mer	393
Myxococcus xanthus MCy8986 (DK897)	454 Roche	Newbler 2.6 / Overlap Layout	1081
Sorangium cellulosum MSr9369 (Soce26Y2)	Illumina	Abyss pe 1.3.6 / K	350
Sorangium cellulosum MSr1795 (Soce307)	Illumina	Abyss pe 1.3.6 / K mer	320
Sorangium cellulosum MSr8404 (Soce38)	454 Roche	Newbler 2.6 / Overlap Layout Consensus	888
Sorangium cellulosum MSr7234 (Soce377)	Unknown	Unknown	1478
Sorangium cellulosum MSr6597 (Soce1525)	454 Roche	Newbler 2.6 / Overlap Layout Consensus	721

Table 2.1: Assembly algorithms used based on overlay-layout-consensus, k-mer approaches used for strains

unamplified molecules. This fact also significantly reduces skews in data and context-dependence (114).

Commercially-available Single Molecule Real-time Sequencing (SMRT) from Pacific Biosciences can yield read lengths up to several tens of kilobases, a figure outperforming all other available technologies by few orders of magnitude. Another issue specifically inherent to biosynthetic gene clusters is their modular structure, which contributes to difficulties of assembling them because of increased likelihood for repeats (115). Various algorithms were developed which to various extent mitigate the problem of assembling highly repetitive sequences in NGS and TGS data. For this work, we used assembly algorithms based on overlay-layout-consensus, k-mer approaches (MIRA, abyss-pe, HGAP) (see Table 2.1). For genetic sequence annotation, Prokka 1.7 was used.

2.2 Mxbase infrastructure

Mxbase explorer is a user interface which is designed to support efficient handling of the scientific data. The Mxbase explorer is equipped with several functions or methods through which a researcher communicates with a program to achieve desired result. Many of the functions in Mxbase explorer perform processing on the client computer but functions that require more computational power run on the Mxbase server backend via apache thrift service which provides a binary RPC protocol for supporting service invocations, enabling multiple programming language implementations to talk to each other. Thrift is a framework designed to be efficient and available across all platforms and programming languages. Functions like *target analysis report upload* and *export*, *metabolome upload* and *mining* and *search and match functions* of biosynML require increased computational power for execution; in order to optimize speed and usability of these functions they run on Mxbase server which directly process the data and write the information to the appropriate tables in the database. In addition, mxbase server has scripts that notifies information to the all users via email and has a file system that stores data in the database which can be downloaded through mxbase explorer. It also has documents about the different projects that uses mxbase server extensively. With the availability of all the technology the researchers can easily and quickly generate reports on various datasets stored in the myxobase depending upon their interest.

A detailed schematic workflow of various components in the Mxbase infrastructure is illustrated in Figure 2.1.

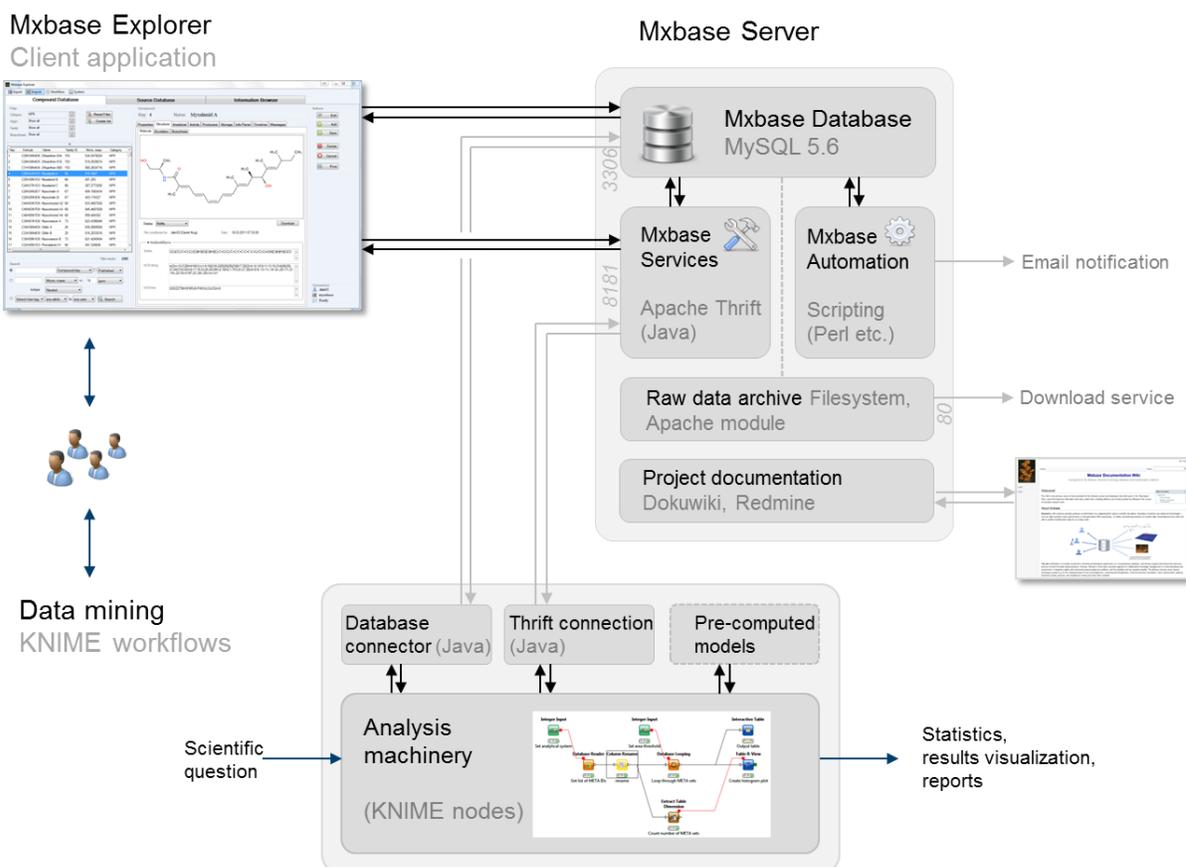


Figure 2.1: Mxbase infrastructure

2.3 Geneious framework

Geneious (Figure 2.2) is a cross-platform bioinformatics software suite developed by Biomatters for search, organize and analyze genomic and protein information via a powerful desktop program. One of the main advantages is its strong focus on user-friendly interface and ease of use in combination with established bioinformatics tools which are “wrapped” into the application (Figure 2.2). It features wide variety of algorithms for sequence alignment, phylogenetic analysis, contig assembly and primer design. In addition, researchers can use this tool to access NCBI and UniProt databases, BLAST, protein structure viewing, automated PubMed searching (116).

In addition, Geneious provides a powerful public API consisting of a set of classes and interfaces necessary for the programmers to develop customized plugins and integrate them into the framework. Exploiting the concept of plugin development kit, a plugin is developed in this study to improve the

quality of the biosynthetic pathways predicted. This development became instrumental for building a department-wide myxo pathways repository.

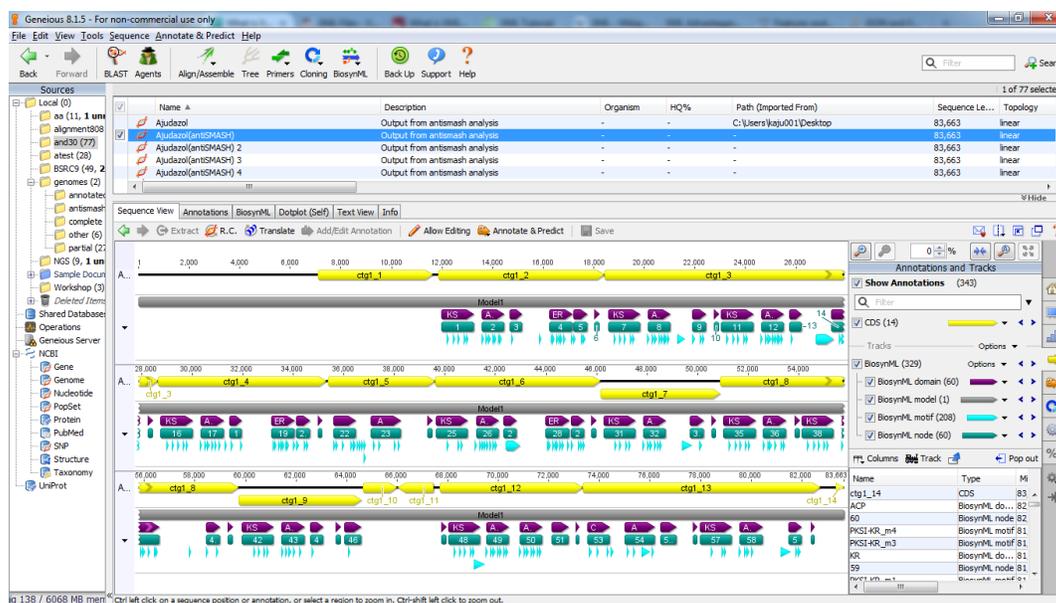


Figure 2.2: Geneious: tool for search, organize and analyze genomic and protein information

2.4 antiSMASH: antibiotic and Secondary Metabolite Analysis Shell

antiSMASH is a tool capable of identifying biosynthetic loci covering a wide range of known secondary metabolite compound classes (polyketides, non-ribosomal peptides, terpenes, aminoglycosides, aminocoumarins, indolocarbazoles, lantibiotics, bacteriocins, nucleosides, beta-lactams, butyrolactones, siderophores, melanins and others). It aligns the identified regions at the gene cluster level to their nearest relatives from a database containing all other known gene clusters, and integrates or cross-links all previously available secondary metabolite specific gene analysis methods in one interactive view. Genes are extracted or predicted from the input nucleotide sequence, and gene clusters are identified with signature gene pHMMs and predicts the biosynthetic pathway. Finally, the output is visualized in an interactive XHTML web page (Figure 2.3) (95).

antiSMASH also provides meta information including substrate specificity of PKS and NRPS which are predicted based on the active sites of the domains such as acyltransferase (AT) and adenylation (A) using various methods. Ketoreductase (KR) domain-based stereochemistry predictions for PKSs are also performed. A final chemical structure of the biosynthetic pathway predicted is generated as a SMILES string.

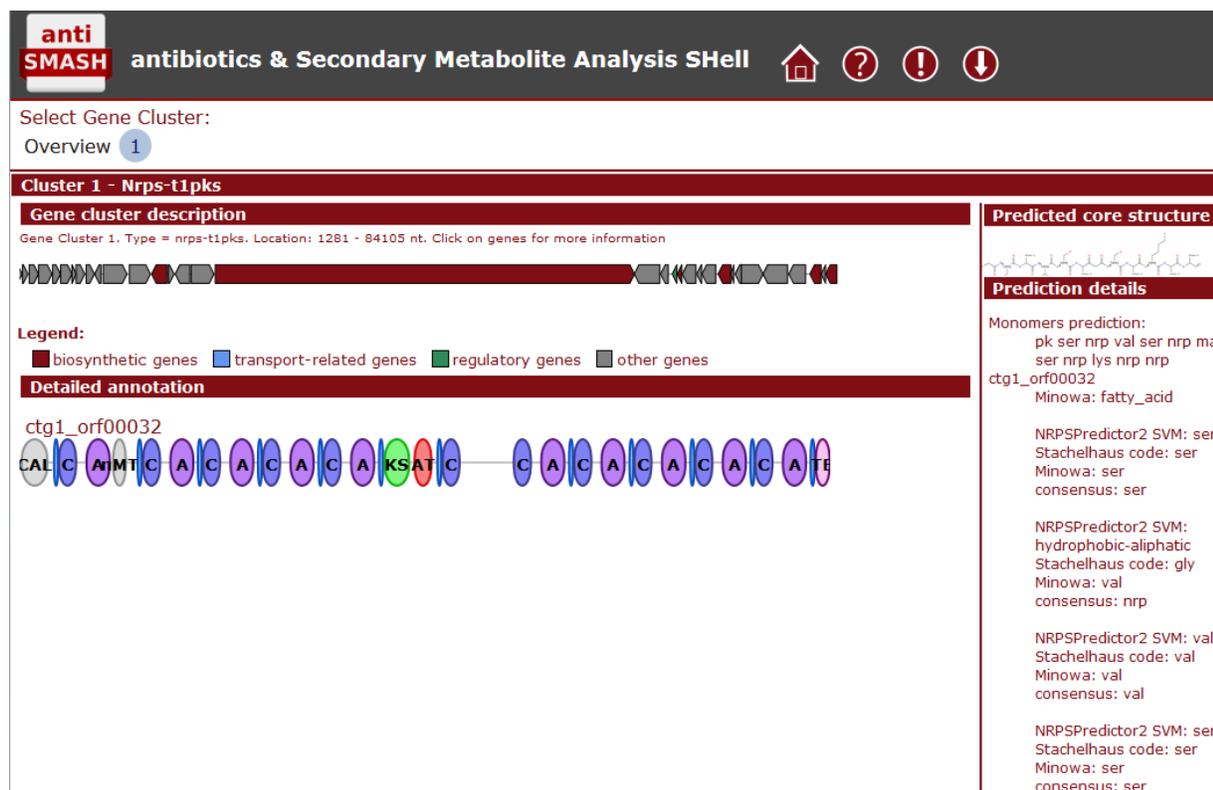


Figure 2.3: antiSMASH output for secondary metabolite gene cluster

2.5 Datasets used in this study

In this work to perform the evaluation test of newly implemented tools, genomes of Actinomycetes (277), Bacillus (77), Cyanobacteria (63), Proteobacteria (47) and Serratia(3) genomes and 42 published clusters (Table 2.4) from myxobacteria are downloaded from NCBI. These genome datasets and clusters are passed through antiSMASH for secondary metabolite cluster prediction which gives biosynml (.xml) file as output. The respective output module was added in the course of this work. The biosynML files of genomes and published clusters from myxobacteria were curated manually using biosynML plugin in Geneious. These datasets (.xml files) were imported into the Myxobase biosynthetic pathway repository and are used as input for targeted search query and genome annotation tools. The published and unpublished genomes (Table 2.2 and Table 2.3) from ongoing projects in the department were also channelled into the analysis.

Organism	#Gene clusters
Chondromyces crocatus Cmc5	28
Corallocooccus coralloides MCy9080	30
Haliangium ochraceum DSM14365	14
Myxococcus fulvus HW1	20
Myxococcus stipitatus MCy9235	26
Myxococcus xanthus DK1622	19
Corallocooccus coralloides Ccc1071	37
Pyxidicooccus unclassified MCy9557	22
Sorangium cellulosum Soce56	19
Sorangium cellulosum So0157	26
Sorangium cellulosum SoceGT47	18
Sorangium cellulosum Soce836	37
Sorangium cellulosum Soce1128	44
Stigmatella aurantiaca DW4/3-1	28
Sorangium cellulosum Soce26	28
unclassified unclassified MNa9516	6
Melittangium lichenicola Mel24	22
unclassified unclassified MCy10635	10
unclassified unclassified MCy11274	3
unclassified unclassified MSr10681	14
unclassified unclassified MSr9506	12
unclassified unclassified MSr10575	15
unclassified unclassified MSr9528	15
Plesiocystis pacifica MNa11107	12
unclassified unclassified MCy11225	15
unclassified unclassified MCy10622	16

Table 2.2: List of complete genomes used in this study

Organism	#Scaffolds	#Gene clusters
Aetherobacter fasciculatus MSr9330	6	20
Aetherobacter fasciculatus MSr9337	3	20
Aetherobacter rufus MSr9331	11	22
Aetherobacter unclassified MSr9329	48	20
Aetherobacter unclassified MSr9335	83	21
Angiococcus disciformis AngGT8	92	40
Archangium unclassified MCy9003	4	31
Bysovorax cruenta Har1=Byc1	14	26
Chondromyces apiculatus Cma2	182	26
Chondromyces catenulatus MSr9030	9	21
Corallocooccus coralloides ST201330	26	40

Cystobacter ferrugineus Cbfe23	155	24
Cystobacter fuscus MCy9118	188	29
Cystobacter fuscus MCy9127	2	30
Cystobacter unclassified MCy9101	36	39
Cystobacter unclassified MCy9104	133	31
Cystobacter velatus Cbv34	20	28
Hyalangium minutum NOCb10	158	17
Minicystis unclassified MSr9310	3	26
Myxococcus fulvus MCy9270	62	30
Myxococcus fulvus Mxf50	141	41
Myxococcus fulvus Mxf65	168	24
Myxococcus unclassified MCy9171	47	24
Myxococcus virescens ST200611	14	23
Myxococcus xanthus DK897	42	21
Myxococcus xanthus MxA47	73	18
Nannocystis exedens Nae478	3	20
Polyangium spumosum Plsm9	138	22
Sorangium cellulosum Soce10	138	13
Sorangium cellulosum Soce1525	14	26
Sorangium cellulosum Soce307	51	19
Sorangium cellulosum Soce377	31	22
Sorangium cellulosum Soce38	17	19
Sorangium cellulosum Soce969	101	23
Stigmatella aurantiaca Sga32	14	28
Stigmatella erecta Pde77	43	26
Archangium gephyra Ar3548	57	40
Corallocooccus coralloides Ccc127	3	25
unclassified unclassified MCy10585	2	42
unclassified unclassified MCy10597	2	41
unclassified unclassified MCy10634	2	7
unclassified unclassified MCy10649	2	33
unclassified unclassified MCy10653	2	43
unclassified unclassified MCy9487	2	69

Table 2.3: List of draft genomes with less than 200 scaffolds used in this study

Cluster	Characterized from strain
Ajudazol	Cmc5 (Chondromyces crocatus)
Althiomycin	DK897 (Myxococcus xanthus)
Ambruticin	So ce10 (Sorangium cellulosum)
Argyrim	Ar8082 (Archangium gephyra)
Aurafuron	DW4/3-1 (Stigmatella aurantiaca)
Carolacton	So ce836 (Sorangium cellulosum)
Chivosazol	So ce56 (Sorangium cellulosum)
Chondrochloren	Cmc5 (Chondromyces crocatus)
Crocacin	Cm c5 (Chondromyces crocatus)
Dawenol	DW4/3-1 (Stigmatella aurantiaca)
DKxanthene	DW4/3-1 (Stigmatella aurantiaca)
Etnangien	So ce56 (Sorangium cellulosum)
Hyafurone	NOCB10 (Hyalangium minutum)
Leupyrrin	So ce690 (Sorangium cellulosum)
Lipothiazole	SoceGT47 (Sorangium cellulosum)
Melithiazol	Me l46 (Melittangium lichenicola)
Microsclerodermin M	MSr9139 (Jahnella unclassified)
Myxalamid	DK1622 (Myxococcus xanthus)
Myxochromid	DK1622 (Myxococcus xanthus)
Myxoprincomide	DK1622 (Myxococcus xanthus)
Myxothiazol	DW4/3-1 (Stigmatella aurantiaca)
Myxoalargin	MCy6431 (Coralloccoccus coralloides)
Myxovirescin	MCy9151 (Myxococcus xanthus)
Pellasoren	So ce38 (Sorangium cellulosum)
Phenalamid	MCy6431 (Coralloccoccus coralloides)
Pyxidienon	MCy9557 (Pyxidicoccus sp. SBCy002)
Rhizopodin	Sga 15 (Stigmatella aurantiaca)
Ripostatin	Soce377 (Sorangium cellulosum)
Sorangicin	So ce12 (Sorangium cellulosum)
Soraphen	So ce26 (Sorangium cellulosum)
Spirangien	So ce90 (Sorangium cellulosum)
Stigmatellin	Sga 15 (Stigmatella aurantiaca)
Thuggacin	Cmc5 (Chondromyces crocatus)
Vioprolide	Cbvi35 (Cystobacter violaceus)
Nannochelin	Nae620 (Nannocystis exedens)

Table 2.4: List of curated gene clusters used in this study

2.6 Programming language

2.6.1 C# and Microsoft Visual Studio

Microsoft Visual Studio (VS) is an integrated development environment (IDE) developed by Microsoft. It is used to develop console and graphical user interface applications along with so-called “Windows Forms” applications, web sites etc. VS can run on all versions of windows and has built-in languages such as Visual Basic .NET (VB.NET) and C sharp (C#). Among these C# is versatile and powerful and is used widely to create applications running on the Windows platform. It is well suited to develop software which features graphical interfaces extensively.

Here, the VS 2013 IDE with C# built in was used for the project. The VS 2013 is licensed under University of Saarland/ MSDN Academic Alliance conditions.

2.6.2 Java and NetBeans Platform

Netbeans is a framework written in java which is programmed for NetBeans Integrated Development Environment (IDE), but can support other languages such as PHP, C/C++ and HTML5. The modular architecture improves the usability of the application by allowing the functions to be reused. We mainly used this application for programming the functions for matching and scoring algorithms as well as for development of the BiosynML plugin for Geneious. These functions mainly run on the Apache Thrift service through remote procedure call framework (RPC). Because of its advanced window management, powerful built-in profiler, excellent integration with Apache Thrift service and easy modular design allowing developers to develop and distribute extensions to the modules.

2.7 Apache Thrift

Apache Thrift is a software framework (originally developed by Facebook, later handed over to the Apache foundation) for creating interoperable scalable cross-language services development. Thrift is composed of efficient protocols and services infrastructure which Facebook use for their back-end services. It combines a software stack with a code generation engine to build services that work efficiently and seamlessly between C++, Java, Python, PHP, Ruby, Erlang, Perl, Haskell, C#, Cocoa, JavaScript, Node.js, Smalltalk, OCaml and Delphi and other languages. Although thrift was originally developed at Facebook, it is now an open source project in the Apache Software Foundation licensed under the Apache 2.0 (117). The main advantage of thrift is the high-performance services that can be

called from multiple languages. The design of thrift makes it as the ideal choice where the speed is a concern while communication between multiple languages is needed and when clients and servers are co-located.

2.8 Graph Visualization: ZedGraph

ZedGraph (<http://sourceforge.net/projects/zedgraph/>) is a set of classes, written in C#, for creating 2D line and bar graphs of arbitrary datasets. The classes provide a high degree of flexibility in terms of personalization of graphs according to need. At the same time, usage of the classes is kept simple by providing default values for all of the graph attributes. The classes include code for choosing appropriate scale ranges and step sizes based on the range of data values being plotted.

ZedGraph also includes a User Control interface, allowing drag and drop editing within the Visual Studio forms editor, plus access from other languages such as C++ and VB. ZedGraph is licensed under the LGPL.

2.9 MySQL Database platform

MySQL is an open source database originally developed by Michael Widenius and now developed further by Oracle as a community project. MySQL is the most popular database which is widely used in many prominent websites. Key features include efficient storage ability but also high performance, high reliability, scalability and ease of use.

In order to access MySQL the user needs to have an account (username and password) on MySQL server. We use a MySQL database (version 5.5.43) as backend for data storage throughout the Myxobase project. The server system was kindly hosted by the ITS department of Universität des Saarlandes.

2.10 Extensible Markup Language (XML)

Extensible Markup Language (XML) is a simple, very flexible text format developed by World Wide Web Consortium (W3C). It is a markup language derived from Standard Generalized Markup Language (SGML), a format which is both human-readable and machine-readable. The design goals of XML are to meet the challenges of large-scale electronic publishing and share structured data across the internet. XML is a robust self-describing or self-defining document that can be stored without schemas

as they contain meta-data. Any XML tag can possess an unlimited number of attributes, containing structured information.

An XML document contains elements, defined by beginning and an ending tag. XML documents must contain a root element under which all elements are contained. XML can also support nested elements, or elements within elements. This ability allows XML to support hierarchical structures where the terms parent, child, and sibling are used to describe the relationships between elements. Element names describe the content of the element, and the structure describes the relationship between the elements. An attribute in xml of elements describes the characteristics of the elements in the beginning tag.

The major advantage of using xml is that there is no fixed set of tags; new tags can be created as they are needed, creating liberty to define a markup language in terms of specific problem set allowing everyone to build their own tag library which suits their needs perfectly. It is not only free style to develop, but also free to develop tools that meet needs exactly based on the user defined structure of tags. The tree structure of XML documents provides better searching and navigation efficiently element by element. XML data is stored in text format making it easier to expand or upgrade without losing data by allowing user to add additional tags.

Altogether, xml is simple, can manage large data by consolidating them in to an xml document in an organized way.

2.11 MUSCLE: alignment software

MUSCLE (multiple sequence comparison by log-expectation) is a multiple sequence alignment program used for protein and nucleotide sequences. The MUSCLE algorithm process an alignment in three stages. First, is the draft progressive, the algorithm concentrates on speed over accuracy and outputs a draft multiple alignment. The second step called as improved progressive, the algorithm reestimates the binary tree used to create the draft alignment using the Kimura distance. In the third step called as refinement, the algorithm improvises the generated in the second step. Due to the speed and accuracy of the MUSCLE, it is widely used ahead of ClustalW as it also gives more robust results depending upon the chosen options. MUSCLE has been made an integral part of several free and commercial softwares such as Geneious, MacVector, Sequencher, MEGA and UGENE (118).

2.12 Bioinformatics functions

It was a fundamental requirement of this work to develop algorithms for the rapid comparison of (potentially large numbers of) biosynthetic models which can answer simple and complex questions. Similarity score is the measure to show how similar two or many sets of pathways are to each other. To find the similarity is to find the comparison between the two or more pathways and grade it after a score system.

Cosine similarity

It is a measure of similarity between two vectors of an inner product space that measures the cosine of the angle between them, where the range of the cosine similarity is [0, 1].

- "1" indicates that x and y have the same domains
- "0" indicates that they share no domains

To assign a numeric score to a document for a query, the model measures the similarity between the query pathway (since query is also just domains and can be converted into a vector) and the target pathway. Typically, the angle (similarity) between two pathways is used as a measure of divergence between the pathway, and cosine of the angle is used as the numeric similarity (119).

The cosine similarity between two vectors x (the target pathway) and y (query pathway) is given by:

$$\text{similarity}(x, y) = \cos(\theta) = \frac{x \cdot y}{\|x\| * \|y\|}$$

The cosine similarity provides a nice metaphor. Cosine similarity gives maximum value when $\theta = 0$ or when the vectors coincide. It gives lowest value when the vectors are independent of each other. This can be seen in the Figure 2.4. The main advantage of Cosine similarity index is their ability to score partial matches irrespective of the order of the domains in the pathways.

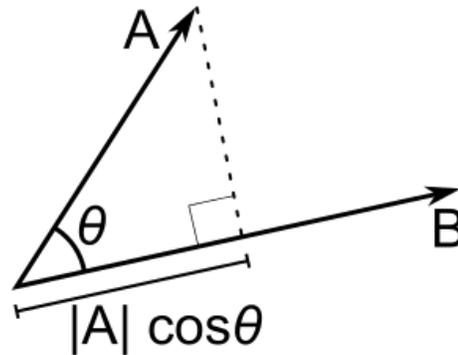


Figure 2.4: The projection of the vector A into the vector B. By Wikipedia.

Smith-Waterman algorithm

The Smith-Waterman algorithm is a dynamic programming method for determining similar regions between nucleotide or protein sequences. The algorithm was first proposed in 1981 by Smith and Waterman and is identifying homologous regions between sequences by searching for optimal local alignments. It is a variant to the idea proposed by Needleman and Wunsch (120) for global alignments. To find the optimal local alignment, instead of looking at the total sequence, the algorithm compares segments of all possible lengths. Based on these calculations, scores or weights are assigned to each character-to-character comparison: positive for exact matches/substitutions, negative for insertions/deletions (121). For a given strings, let x and y are the alphabets iterated over the string. $H(i, j)$ stores the similarity score for the prefixes $x[1, i]$ and $y[1, j]$, W is the gap penalty for insertion or deletions of single characters.

$$H_{ij} = \max\{H_{i-1,j-1} + s(x_i, y_j); H_{i-k,j} - W; H_{i,j-1} - W; 0\}$$

for all $i, 1 \leq i \leq |x|$,

and $j, 1 \leq j \leq |y|$

The Smith-Waterman algorithm starts with the highest values and walks back to the path of the previous high value recursively until it reaches to the least value. Then, the alignment is reconstructed where gaps (insertions or deletions) are placed if a diagonal jump is observed.

EstimateS: Statistical estimation of species richness and shared species from samples

EstimateS (<http://viceroy.eeb.uconn.edu/estimates/index.html>) is a tool that computes variety of biodiversity statistics, estimators, and indices based on biotic sampling data. One such function is the estimation of species richness through non-parametric estimators (Chao1 and Chao2) (122).

3 Results and Discussion

The main objectives of the present work, as outlined in chapter 1.6, can be briefly summarized as follows: firstly, the development of a framework to facilitate the seamless transfer of biosynthetic gene cluster information including extensive meta data between the antiSMASH pipeline and various downstream analysis tools; secondly, the development of new algorithms for the conceptual comparison of genome-encoded secondary metabolite pathways as opposed to primarily sequence-based analysis; and finally the application and critical performance evaluation of newly developed tools for genome-mining with natural product sources, with special focus on myxobacteria. According to these objectives, this chapter reports first the efforts to establish the BiosynML analysis framework including a number of technological and implementation issues (chapter 3.1), followed by the description of BiosynML algorithm development (chapter 3.2).

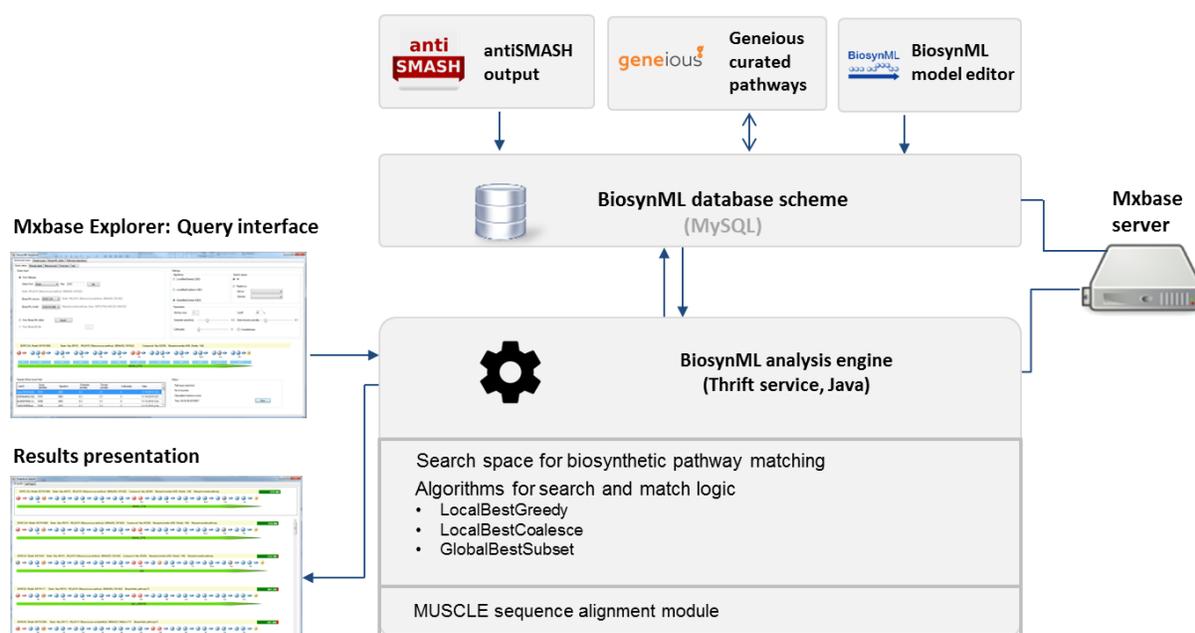


Figure 3.1: Schematic overview summarizing the functions and interoperation of key components forming the BiosynML framework

Datasets accumulated inside Myxobase in the course of this work form the basis for the development and testing of conceptual genome mining functions. Ultimately, application of the complete BiosynML framework to several prototypical analysis scenarios using genomic data from both

in-house and public sources is the subject of chapter 3.3, covering in particular mycobacterial pathways and genomes. The schematic overview below (Figure 3.1) summarizes the functions and interoperation of key components forming the BiosynML framework.

3.1 Bioinformatics Framework for conceptual genome mining

The routine analysis of secondary metabolite producers from various taxa which are sources for secondary metabolite-rich samples in the course of screening for novel compounds, in combination with the analysis of biosynthetic gene clusters encoded in the genomes of these producers, generates an overwhelming amount of biosynthetic pathway-related data. It was therefore a major aim of this project to develop practical tools enabling researchers to take better advantage of the high information content of these data. In particular, existing sequence-based analysis approaches for identifying and comparing biosynthetic pathways (based largely on the output of antiSMASH) should be complemented by an approach which can operate with or without sequence information, allowing anyone to instantly annotate a secondary metabolite cluster on the basis of all biosynthetic model information in the database available at that very moment. This ability should help to better “bridge” the knowledge from chemical, biochemical, biological and genomics research efforts within an integrated and highly interdisciplinary natural products discovery setting. Another important requirement was the ability to conduct targeted queries across all biosynthetic pathway datasets, to facilitate the identification of pathways exhibiting specific domain arrangements along with their properties of interest. It was also a crucial pre-requisite that all tools be implemented in conjunction with the Mxbase system, in order to enable distributed work across several labs and achieve “near-real time” collaboration of researchers concerning the data itself as well as the derived conclusions.

The framework described in the following, reflecting these ideas, evolved gradually throughout this study until it reached the current state. Besides describing this state of art, the following paragraphs will shed light on the most important steps that lead to the development of different components in the framework, collectively named “BiosynML”. This explains best how the current features emerged and the design decisions which have been taken.

3.1.1 BiosynML Language and container

The basic needs for this framework start with the definition of a format and a container for storing biosynthetic models, including the primary sequence data (optional) and associated annotations for secondary metabolite gene cluster(s). A language (syntax) is essential to capture these models and

describe all their properties in great detail, in a systematic and extensible way. At the same time a container (file format) to store these descriptions and transport them between different tools and analysis workflows is required.

Traditionally, the GenBank format has been widely used, as a computer and human readable flat file format used to store sequences and associated annotations where the fields containing different types of information are well-labelled. Although, GenBank has an advantage for being the de-facto standard used by many tools a major difficulty faced is in updating and extending the information. In addition, the GenBank file format was not designed to comprise the potentially verbose output of multiple prediction tools applied to sequence motifs (many of which co-exist in typical secondary metabolite modular pathways). Moreover, it is generally difficult to represent relationships between various data fields in a flat file, which makes it hard to group fields and link information across multiple entries. The area of natural product research is a dynamic field with complex information related to the annotations which is hard to handle by flat files. To overcome this difficulty, an XML format, being a portable and accessible format, was chosen to store the annotated data where the key feature is to use identifiers to enclose the sections of the data. It is also possible to further define data with appropriate tags and attributes which makes it easy for the researcher to identify the nature of the data.

BiosynML: “Biosynthetic Markup Language”

To overcome the lack of a suitable data format for storing and transporting complete biosynthetic pathway models between tools and databases, a new specialized mark-up language, named "BiosynML" was devised at the beginning of this study. BiosynML is an xml dialect, annotating a document in a way that is syntactically distinguishable from the content (text) itself. Information stored in the BiosynML containers can extend significantly beyond the content of GenBank files, where only relatively minimal information on genes and domains is included. Typically, information in xml schemes is stored in a systematic order in the form of hierarchic nodes. BiosynML has 6 elementary nodes:

- Header, stores the basic information of the content like the date, author etc. (Figure 3.2a).
- Model, stores the complete meta information of the pathway like the details of the organism the pathway belongs to, compound information that the pathway involved in and the gene cluster information. The model also has a chemical and a modules layer where the information of the building blocks and the modules containing set of domains which function together are stored,

respectively. A BiosynML file can store multiple independent pathways (models) which belong to the same organism (Figure 3.2b).

- Domainslist, stores all information on biosynthetic domains, such as (observed or predicted) enzymatic activities and substrates used. It also stores the positions in the genes. (Figure 3.2c).
- Genelist, stores the information of the coding sequences involved in the pathway(s). It also accommodates the qualifiers from antiSMASH and GenBank such as gene prediction scores, translated protein sequence etc. (Figure 3.2d). The information for each gene is connected to sequences in Sequencelist.
- Motiflist, stores the information and predictions for signature motifs e.g. as obtained from antiSMASH (Figure 3.2e) or any other analysis tool.
- Sequencelist, optionally stores the primary sequence for a sequence-related biosynthetic model, this could vary from single scaffold to multiple scaffolds or even a whole genome (Figure 3.2f).

Importantly, the BiosynML format is able to capture the full information content available from various analysis tools and is flexible and extensible to adapt for future changes with regards to this information. It is designed for semantic access to this information, i.e. so that specialized analysis modules can easily extract the required information layer. Moreover, the format maintains a certain degree of “human readability”, which is an important aspect for its use in script language-based analysis workflows.



Figure 3.2: BiosynML layers a: header, b: model, c: domains, d: genes, e: motifs, f: sequences

3.1.2 Interfacing BiosynML to the “antibiotics & Secondary Metabolite Analysis Shell” (antiSMASH)

The variety of biological activities observed from (microbial) secondary metabolites - including antibiotics, immunomodulation agents, receptor antagonists and agonists, enzyme inhibitors and antitumor agents – motivates efforts to detect their biosynthetic pathways in the genomes of producers. This information in turn facilitates the experimental elucidation of biosynthetic pathways for secondary metabolites clusters and may also support the finding of new metabolites by genome mining approaches.

To underpin these endeavours, software tools are required to predict and annotate biosynthetic pathways and flexibly link it to the sequence information which is used for more detailed analysis in the course of pathway characterization.

antiSMASH is a tool for the analysis of secondary metabolite gene clusters in bacterial and fungal genomes (95). It is in fact a combination of various secondary metabolite-related prediction tools with high accuracy of identifying individual cluster annotations in a genome; in particular, it integrates several tools for the analysis of distinct types of NRPS- and PKS-related domains, such as NRPSpredictor for A-domains (123).

As part of this work, a new function was appended to the antiSMASH prediction pipeline which exports the results generated by antiSMASH into the information-rich BiosynML format. Based on the functional information of domains from antiSMASH, BiosynML can add extra meta information of domains from an online domain directory (see also 3.1.3) which is at the moment maintained specifically for use with BiosynML (but with official integration of BiosynML into antiSMASH could also be continued as a community repository in the future).

The HTML interface of antiSMASH displays the minimum information of substrate predictions for domains, whereas, BiosynML files store detailed substrate predictions obtained from various prediction modules integrated in antiSMASH. Furthermore, when antiSMASH predicted genes lack domains, BiosynML can add specialized domains from the domain directory based on smCOG analysis and later researchers may choose to delete or modify these domains (Figure 3.3).

The code to generate BiosynML output from antiSMASH results has been integrated into the publicly available antiSMASH package with the release of version 3.0 (95).

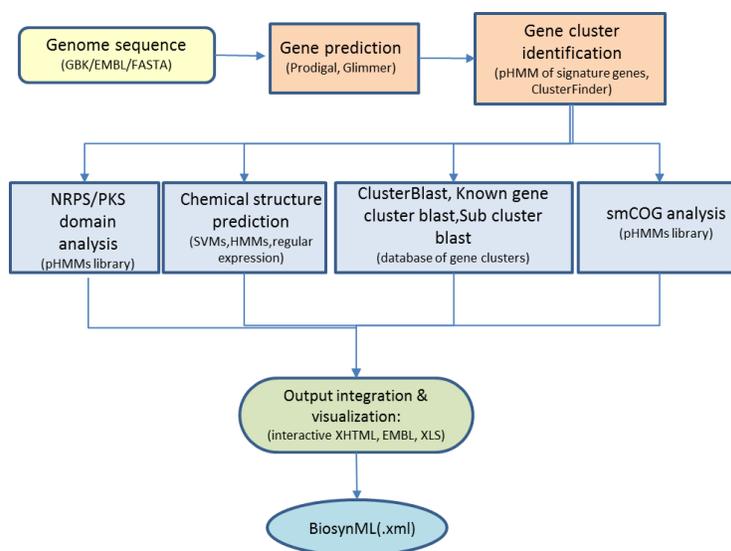


Figure 3.3: Outline of the pipeline for genomic analysis of secondary metabolites (modified from (1))

3.1.3 The BiosynML Geneious plugin

In the course of analysis of secondary metabolite biosynthetic gene clusters there is a constant need to visualize pathways, as well as to add and edit pathway-related meta information and to create an overview of pathways present in one or multiple genome(s). In addition, biosynthetic models need to be submitted to bioinformatics tools such as antiSMASH in order to create annotated models, and the results must be retrieved to feed them back into in-house databases and downstream analysis workflows. These requirements were addressed in the course of this project by the development of a BiosynML plugin for the Geneious software. Geneious is a cross-platform bioinformatics software suite developed by Biomatters for search, organize and analyse genomic and protein information via a desktop program (116). One of the main advantages is its strong focus on user-friendly interface and ease of use along with the seamless integration of a number of published bioinformatics methods. The newly devised BiosynML plugin originating from this work facilitates the creation of detailed biosynthetic pathway annotations, especially for modular gene clusters responsible for the production of microbial secondary metabolites. Typical tasks include refining the automatic predictions obtained by antiSMASH, addition of domains and meta-data based on manual analysis and/or experimental results (curation), such as grouping of domains into functional modules and assignment of biosynthetic building blocks. This information is key to establishing the connection between genes and chemical compounds and also a crucial prerequisite for the design of experiments to investigate the molecular basis for secondary metabolite formation. The core functionality of the BiosynML plugin for Geneious which has been implemented as part of this work is described in the following.

BiosynML Import:

This BiosynML plugin function creates a new sequence object in Geneious using the DNA sequence from the BiosynML sequencelist and creates "CDS" features and a "BiosynML" annotation track. The latter contains qualifiers generated on the sequence according to genelist, motiflist, modulelist, domainlist and nodelist (Figure 3.4). It assigns the information from various tags to the appropriate BiosynML annotation tracks as additional "qualifier" values which help to create annotations conforming to the BiosynML ontology. By observing to the guidelines for Geneious plugin construction, the resulting Geneious object is in principle GenBank-compatible which means, when exported in the widely used .gb format, BiosynML qualifiers can be preserved (albeit without the semantic access which the BiosynML format offers). The plugin provides an additional BiosynML tab in the Geneious user interface where user can add/edit the pathway information.

BiosynML tab consist of several functions to deal with the information imported into Geneious such as an overview table which displays the whole information of the pathway, and a sidebar for the detailed information review (Figure 3.4).

Node ID	Class	Context	Building block	Domain ID	Function	Status	Module label	Domain label	Chemistry	Substrate	Evidence	Gene name	Position
1	linkage	Polyketide		1KS					condensation	Intermediate	Sequence-based p...	93 - 138	
2	activation	Polyketide	Mal	2AT	active	in4	KS	AT	acyl transfer	building block	Structure-based p...	1695 - 2583	
3	modification	Polyketide		3DH	active	in4	DH	DH	dehydration	intermediate	Sequence-based p...	2775 - 3279	
4	modification	Polyketide		4BR	active	in4	BR	BR	enoyl reduction	intermediate	Sequence-based p...	4339 - 5262	
5	modification	Polyketide		5KR	active	in4	KR	KR	ketoreduction	intermediate	Sequence-based p...	5327 - 5844	
6	carrier	Polyketide		6ACP	active	in4	ACP	ACP	thioester	building block	Sequence-based p...	6141 - 6351	
7	linkage	Polyketide		7KS	active	in5	KS	KS	condensation	intermediate	Sequence-based p...	132 - 138	
8	activation	Polyketide	Mal	8AT	active	in5	AT	AT	acyl transfer	building block	Structure-based p...	1752 - 2561	
9	modification	Polyketide		9KR	active	in5	KR	KR	ketoreduction	intermediate	Sequence-based p...	3471 - 4005	
10	carrier	Polyketide		10ACP	active	in5	ACP	ACP	thioester	building block	Sequence-based p...	4325 - 4545	
11	linkage	Polyketide		11KS	active	in5	KS	KS	condensation	intermediate	Sequence-based p...	4652 - 5883	
12	activation	Polyketide	Mal	12AT	active	in5	AT	AT	acyl transfer	building block	Structure-based p...	6204 - 7083	
13	modification	Polyketide		13DH	active	in5	DH	DH	dehydration	intermediate	Sequence-based p...	7296 - 8192	
14	modification	Polyketide		14KR	active	in5	KR	KR	ketoreduction	intermediate	Sequence-based p...	8973 - 9504	
15	carrier	Polyketide		15ACP	active	in5	ACP	ACP	thioester	building block	Sequence-based p...	9825 - 10029	
16	linkage	Polyketide		16KS	active	in7	KS	KS	condensation	intermediate	Sequence-based p...	114 - 1401	
17	activation	Polyketide	PK	17AT	active	in7	AT	AT	acyl transfer	building block	Structure-based p...	1715 - 2607	
18	modification	Polyketide		18DH	active	in7	DH	DH	dehydration	intermediate	Sequence-based p...	2820 - 3124	
19	modification	Polyketide		19BR	active	in7	BR	BR	enoyl reduction	intermediate	Sequence-based p...	4485 - 5424	
20	modification	Polyketide		20KR	active	in7	KR	KR	ketoreduction	intermediate	Sequence-based p...	5481 - 6018	
21	carrier	Polyketide		21ACP	active	in7	ACP	ACP	thioester	building block	Sequence-based p...	6327 - 6531	
22	modification	IPP		22HC	active	in8	HC	HC	condensation het...	intermediate	Sequence-based p...	219 - 1119	
23	activation	IPP	Ser	23A	active	in8	A	A	aldolization	building block	Structure-based p...	1695 - 2504	
24	modification	unspecified		24Ox	active	in8	Ox	Ox	oxidation	intermediate	Sequence-based p...	3153 - 3665	
25	carrier	IPP		25PCP	active	in8	PCP	PCP	thioester	building block	Sequence-based p...	3948 - 4152	
26	linkage	Polyketide		26KS	active	in9	KS	KS	condensation	intermediate	Structure-based p...	12 - 1292	
27	activation	Polyketide	Mnal	27AT	active	in9	AT	AT	acyl transfer	building block	Sequence-based p...	1605 - 2502	
28	modification	Polyketide		28DH	active	in9	DH	DH	dehydration	intermediate	Sequence-based p...	2763 - 3204	
29	modification	Polyketide		29BR	active	in9	BR	BR	enoyl reduction	intermediate	Sequence-based p...	4255 - 5289	
30	modification	Polyketide		30KR	active	in9	KR	KR	ketoreduction	intermediate	Sequence-based p...	5352 - 5889	
31	carrier	Polyketide		31ACP	active	in9	ACP	ACP	thioester	building block	Sequence-based p...	6187 - 6390	
32	activation	Polyketide		32AT	active	in10	AT	AT	acyl transfer	building block	Structure-based p...	1111 - 1389	
33	modification	Polyketide	Mal	33AT	active	in10	AT	AT	acyl transfer	building block	Structure-based p...	1761 - 2589	
34	carrier	Polyketide		34BR	active	in10	BR	BR	ketoreduction	intermediate	Sequence-based p...	3837 - 4071	
35	linkage	Polyketide		35ACP	active	in10	ACP	ACP	thioester	building block	Sequence-based p...	4300 - 4590	
36	linkage	Polyketide		36KS	active	in11	KS	KS	condensation	intermediate	Sequence-based p...	117 - 1392	
37	activation	Polyketide	Mal	37AT	active	in11	AT	AT	acyl transfer	building block	Structure-based p...	1893 - 2589	

Figure 3.4: BiosynML tab: Ball scheme representation (box outlined in green), sidebar (box outlined in blue) and overview table displaying the whole information of the pathway

Ball-scheme representation:

This function uses the BiosynML qualifiers to create a ball-scheme representation of a biosynthesis model, like it is commonly created by researchers and is the de-facto standard for illustration of biosynthetic models in publications. Each ball, representing a biosynthetic function carries its internal node number and displays a functional abbreviation alongside with selected additional data, such as substrate specificity for monomer-incorporating domains (e.g. A, AT). This function graphically enables users to have a quick overview of the domains in a biosynthetic model. A tool tip is provided to view the properties of the each node represented as a ball (Figure 3.4 (green box)) and the details are also accessible in the sidebar.

Sidebar for information review and editing:

The BiosynML plugin sidebar displays the detailed information of a node selected from the gridview where the user can review or modify details. The sidebar consists of panels where the information is displayed in a pre-organized way, e.g. a panel exists carrying the information of model, node, domain, motifs and modules (Figure 3.4 (blue box)).

Creating modules:

Domains in biosynthetic pathways are commonly grouped into modules based on (bio) chemical considerations, in order to declare their joint action to bring about specific substructures of a compound. BiosynML models generated from antiSMASH output initially do not contain any information about modules. To fill the module information, a function has been added to the BiosynML Geneious plugin for the users to create modules by selecting set of nodes in the gridview which the user presumes to function as a module. Furthermore, on module creation the ball scheme representation will also be updated where ball images are grouped representing their participation in modules.

Automatic BiosynML:

If the user intends to generate a BiosynML from an previously existing annotation document e.g. in GenBank format, then the function “Automatic BiosynML” is helpful which will guide the user to generate a BiosynML document. For this, the user has to select the type of annotation in the Geneious frontend which presumably contains domain information in GenBank format, e.g. misc_features, and then the function will search for the matching domains from the annotated features and generate a list

of domains and their properties in the BiosynML tab. If details cannot be auto-assigned by the function, the user can later add those details with the help of functions in the BiosynML tab. Thus, this function assists with the conversion of existing annotations into BiosynML format.

Cleanup:

It is a function when user wish to modify some data in the Geneious sequence view window for e.g. user wish to change the domain length or extending the module region, then this function will reflect the changes in the BiosynML tab and autocorrects the existing data with the changed data. For e.g. if the length of the domain is increased then this function will adjust the length of the node automatically.

Add node/domain:

Biosynthetic pathways predicted from the antiSMASH might have some shortcomings, such as a domain might be missing or it may predict a domain which in reality may not be present. In order to correct such prediction errors, users can add/delete a node/domain and can even manually assign new domain to an existing module (Figure 3.5). This manual refinement of biosynthetic models is commonly carried out in the course of experimental characterization of a biosynthetic pathway.

The screenshot shows a dialog box titled "Add node/domain". It is divided into two main sections: "Node" and "Domain".

- Node section:** Includes a checked checkbox "Add node", a "Context" dropdown menu set to "Polyketide", and a "Class" dropdown menu set to "activation".
- Domain section:** Includes a checked checkbox "Add domain", a "Function" dropdown menu set to "?", a "Subtype" dropdown menu, an "Active" dropdown menu set to "active", a "Comment" text input field, a "Chemistry" dropdown menu set to "acetyl transfer", a "Substrate" dropdown menu set to "building block", a "Label" text input field set to "?", and empty text input fields for "Gene name", "Position in gene", "Protein name", and "Position in protein".

At the bottom of the dialog, there is a checked checkbox "Assign to existing node" and an "Assign module" dropdown menu set to "m1". "Set" and "Cancel" buttons are located at the bottom right of the dialog.

Figure 3.5: Window to add a node/domain

Extract domains:

It is a function to extract annotations of interest to a new document. Users are provided with multiple options in order to extract the regions of interest, such as “Extract currently selected domains” is an option useful when the user likes to extract a subset of domains selected from the grid view in BiosynML tab. Additionally, if user wants to extract the derived protein sequence, then the function will automatically translate the nucleotide sequence into protein and add annotation on to the translated protein sequence. On the other hand, if user wants to extract annotations of interest, for e.g. if one wants to extract all the “AT” domains then user can select the option “Extract domains matching these criteria” and can chose the extraction option to obtain the required subset of the annotation features. The result of the extraction can be done as a sequence list in a single document or can save them as individual documents (Figure 3.6). The extraction function is typically used to submit domain sequences from selected pathways to downstream analysis, e.g. Alignment, tree-building and visualization as phylogenetic trees.

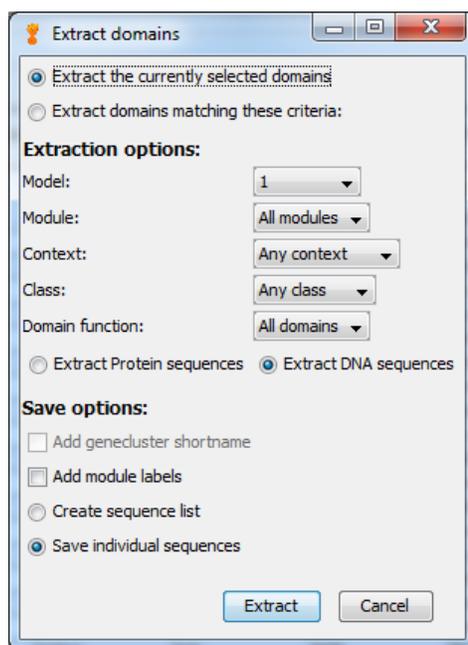


Figure 3.6: Features of the Extract domains function

Add/edit building blocks.

The output from antiSMASH carries predicted information about building blocks presumably incorporated by monomer-activating domains. This information can be manually amended, and in addition information about the building blocks which are actually incorporated (based on experimental

evidence) can be added or edited using the “Add/edit building block” function. This could also be helpful if the BiosynML is generated by “Automatic BiosynML generation function” or to add missing building block information to the nodes (Figure 3.7).

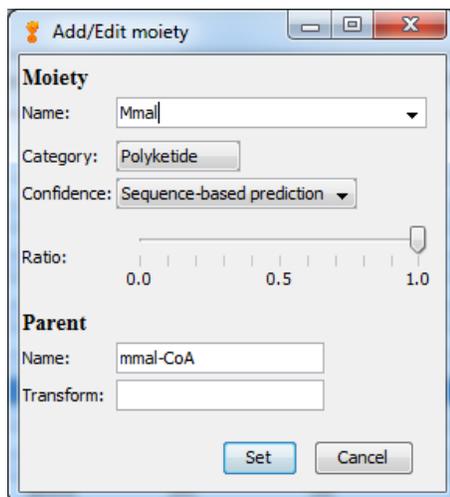


Figure 3.7: Window for building block modifications

Split model:

Several gene clusters which represent separate pathways are sometimes “fused together” by antiSMASH during cluster predictions. This happens due to the simple rationale used by antiSMASH for the definition of independent clusters, which basically uses a 5-20 kb window depending on the type of the cluster. Any detected core gene within this window will be assigned to the same cluster. Split model function is used to manually separate those fused clusters, making BiosynML plugin recognize them as individual clusters.

Pathways overview:

This function is used to summarize biosynthetic clusters contained in a genome in the form of a collection of ball scheme representations. Thus, it creates a visual overview and at the same time enables users to access the detailed information on selected models / nodes (Figure 3.8).

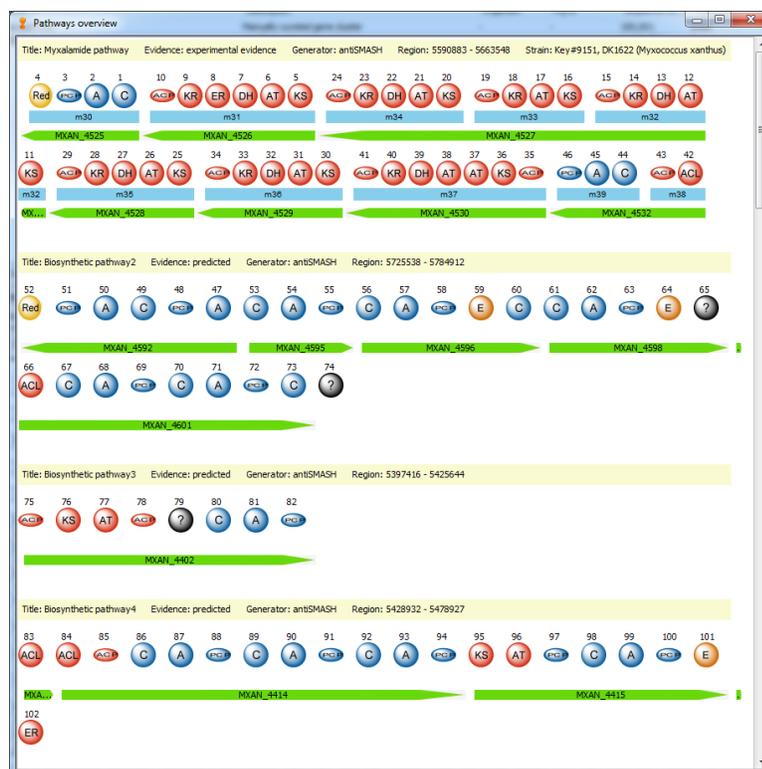


Figure 3.8: Window showing detailed overview of biosynthetic pathways within a genome

Biosynthetic domains directory:

A number of recurring enzymatic domains are involved in multi-step biosynthetic pathways, each with a specific role in precursor selection and activation, linkage of building blocks, or their modification, to name only a few processes. BiosynML uses an online domain directory that supplies the common "vocabulary" for biosynthetic pathway annotation, also providing basic information about common domains. In addition, new submissions of missing or new domains are actively curated and added to domains directory by experts (Figure 3.9).

This helps to maintain controlled vocabulary used in diverse project parts, which is consistency and interoperability especially important for meta information such as name, class, context, function, chemistry, substrate and keywords associated with domains. It prevents ambiguities during data markup of BiosynML file. Thus, the domain directory approach can also serve in the future as a bridge between potentially concurring vocabularies.

Icon	Function	Name	Class	Context	Subtypes	Chemistry	Substrate	Keywords	DLID	Qualif...
?	?	Domain...	Unspecifed	Unspec...		unknown	unknown		27	
A	A	Adenyl...	activation	NRP		adenylation	building block	A,AMP-binding,ade...	8	
ACL	ACL	Acyl-Co...	activation	Polyket...		acyl-CoA ligation	building block	CAL,CoA-ligase	21	
ACP	ACP	Acyl-ca...	carrier	Polyket...		thioester	building block	ACP	1	
ACT	ACT	Acetyl...	activation	Polyket...		acetyl transfer	building block	acetyltransferase	45	
AGP	AGPAT	Acylgly...	activation	Polyket...		acyltransfer	building block	Acylglycerolphosph...	31	
ALY	ALY	Ammoni...	modificat...	Polyket...	PAL,TAL,HAL	non-oxidative deam...	building block	ammoniumlyase,am...	17	
AMT	AMT	aminotr...	modificat...	NRP	class-1/2,class-3,d...	amino group transfer	intermediate	aminotransfer	36	
Amu	Amu	Aminom...	modificat...	NRP	TAM,PAM,LAM	amino group migration	building block	aminomutase	16	
	Any-NRPS	Unspec...	Unspecifed	NRP		other	other		23	
	Any-PKS	Unspec...	Unspecifed	Polyket...		other	other		24	
A-Ox	A-Ox	adenyl...	modificat...	NRP		adenylation and oxi...	building block		49	
AT	AT	Acyltra...	activation	Polyket...		acyl transfer	building block	AT	2	
Aux	AUX	Auxiliar...	modificat...	Unspec...		Unspecifed	intermediate		22	
C	C	Conden...	linkage	NRP	unknown,LCL,DCL,...	condensation	intermediate	DCL,LCL,C,Dual	9	
CLF	CLF	Chain l...	linkage	Polyket...		condensation	intermediate	df,chain length factor	30	
CT	CT	Carbam...	modificat...	NRP		carbamoylation	intermediate	carbamoyl	34	
Cy	Cy	aromat...	modificat...	Polyket...		chromone ring form...	intermediate	chromone,cyclase,a...	48	
CYP	CYP	Cytoch...	modificat...	Unspec...		oxidation	intermediate	cyp,P450	37	
Dck-nrps	Dck-nrps	NRPS D...	docking	NRP	N-term,C-term	protein-protein inte...	Unspecifed	COM,docking	25	
Dck-pks	Dck-pks	PKS Do...	docking	Polyket...	N-term,C-term,tran...	protein-protein inte...	Unspecifed	docking	26	
Dec	Dec	Decarb...	modificat...	Unspec...		decarboxylation	intermediate	decarboxylase	44	
Deh	Deh	Dehydr...	modificat...	Unspec...		dehydrogenation	intermediate	dehydrogenase	42	

Figure 3.9: Overview of domain directory created by BiosynML plugin in Geneious

Building blocks registry:

Assigning biosynthetic building blocks to enzymatic domains that activate small molecules and incorporate them into complex products is an important analysis step when dealing with secondary metabolite pathways. The BiosynML plugin assists this analysis by providing a list of previously observed biosynthetic building blocks, using short codes in agreement with other popular databases from the natural products field (such as e.g. Norine (2)). Structures are generated by using Indigo library using SMILES format. In addition, new submissions of missing or new building blocks are actively curated and added to building blocks repository by experts (Figure 3.10) (59).

Code	Full name	Synonym	Context	Formula	IUPAC	SMILES	Mol. weight	Parent	Transform	BBID	Comment
Ph-Lac	Phenyl-lactate	Phenyl-lact...	other	C9H10O3	2-hydroxy-...	C1=CC=C(...	166.1739			495	
Ph-Ser	phenylserine	beta-Hydro...	other	C9H11NO3	(2S,3S)-2-a...	C1=CC=C(...	181.1885			496	
PMST	propanoyl-...		other	C9H12N2O3S		C(C=CC1=...	228.2682			497	
pOH-Bz	para-hydro...	para-hydro...	other	C7H6O3	4-hydroxyb...	C1=CC(=C...	138.1207	Bz		498	
Pro	Proline	L-proline	NRP	C5H9NO2	(2S)-pyrrol...	C1CC(NC1)...	115.1305	Pro		499	
ProC	proline carb...	prolinamide	NRP	C5H10N2O	(2S)-pyrrol...	C1CC(NC1)...	114.1457			500	
Pro-Thz	proline-thia...		other	C8H10N2O2S		C1=CSC(=...	198.2422			501	
PT	phosphinot...	GLUFOSINATE	NRP	C5H12NO4P	2-amino-4(...	CP(=O)(CC...	181.1268			502	
pTrp	phototrypto...		other	C11H12N2O3	8b-hydroxy...	C1C(NC2C1...)	220.2246			503	
PTTA	4-propanoyl...		other	C14H14N2...		C(O)(=O)C...	290.3376			504	
Put	putrescine	1,4-diamino...	other	C4H12N2	butane-1,4...	C(CCN)CN	88.1515			505	
Pya	pyruvate	Pyroracemic...	other	C3H4O3	2-oxopropa...	CC(=O)C(=...	88.0621			506	
Pyr	pyrrolidone	2-Pyrrolidin...	other	C4H7NO	pyrrolidin-2...	C1CC(=O)N...	85.1045			507	
Rha	L-rhamnose	isodulcit,Rh...	CS	C6H12O5	2,3,4,5-tetr...	CC(C(C(C(C...	164.1565			508	
Ria	L-ristosamine		CS	C6H13NO3	(3R,4R,5S)-...	CC(C(C(C(C...	147.1723			509	
Ser	Serine	beta-Hydro...	NRP	C3H7NO3	(2S)-2-amin...	C(C(C(=O)...)	105.0926	Ser		510	
Serol	Serinol		NRP	C3H9NO2	2-aminopro...	C(C(CO)N)O	91.1091	Ser		511	
Spd	spermidine	1,8-Diamino...	other	C7H19N3	N-(3-amino...	C(CCNCCC...	145.2459			512	
Sta	statine		other	C8H17NO3	4-amino-3-h...	CC(C)CC(C(...)	175.2255			513	
Thr	Threonine	threonin	NRP	C4H9NO3	(2S,3R)-2-a...	CC(C(C(=O)...)	119.1192	Thr		514	
t-Leu	tert-Leu	3-methyl-va...	NRP	C6H13NO2	(2S)-2-amin...	CC(C)(C)C(...)	131.1729	Leu		515	
Trp	Tryptophan	Tryptophan...	NRP	C11H12N2O2	(2S)-2-amin...	C1=CC=C2...	204.2252	Trp		516	

Figure 3.10: Building blocks list generated within the BiosynML plugin adapted from Norine database

Prepare for MiBIG:

The MiBIG initiative aims to establish a community repository of annotated natural product biosynthetic pathways according to the MiBIG standard ("Minimal information on biosynthetic gene clusters"). The BiosynML plugin helps users to prepare their curated pathways for submission into the MiBIG database, by pre-filling many fields in the MiBIG form with the relevant details (113).

Submission and retrieval to/from antiSMASH:

BiosynML plugin functions can be used to manually add and edit biosynthetic pathway information in Geneious, but information-enhanced documents are usually first created by submitting the sequence of a biosynthetic pathway (or an entire genome) online to the antiSMASH annotation engine. The BiosynML plugin handles job submission and retrieval of results, using the BiosynML format as a shuttle between the analysis server and the Geneious client. This works with both the public antiSMASH web

service and an in-house server setup using the standalone variant of antiSMASH (Figure 3.11). The BiosynML antiSMASH submission function basically reflects the options also available in the web interface. The function has become publicly available with the release of antiSMASH v3.0 (95).

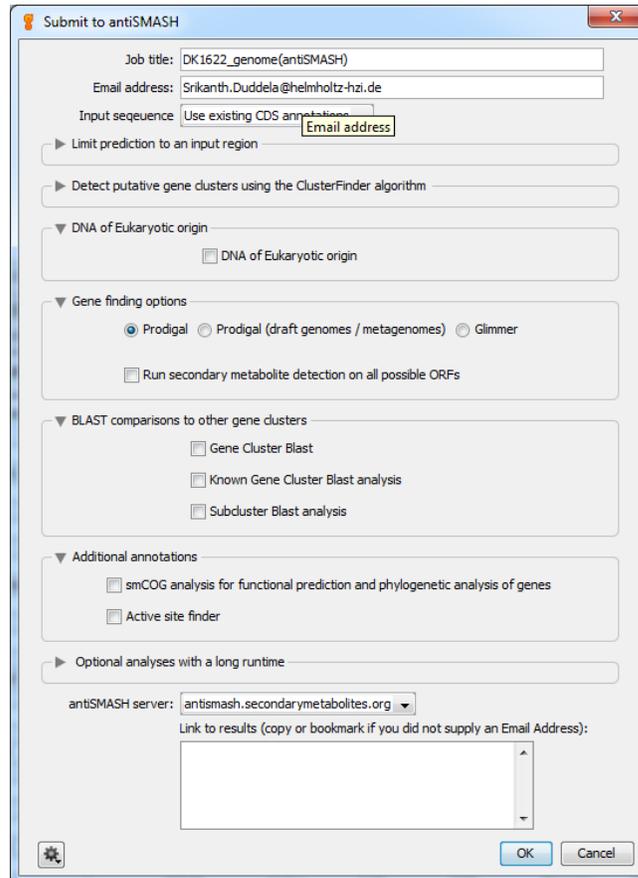


Figure 3.11: antiSMASH job submission and retrieving window

Export BiosynML:

Once the BiosynML content is imported, Geneious provides several methods to manipulate the "BiosynML" qualifiers, or enables the user to manually add such qualifiers to Geneious documents which do not have BiosynML content. After the modification of BiosynML content, the plugin can export the complete information from the Geneious sequence object to the BiosynML format. The export is basically the reverse function as described for the Import (Figure 3.12). This is important so that BiosynML documents edited within Geneious can be transferred into databases, i.e. the Mxbase system used in this project, and also facilitates their use in script-language based analysis workflows.

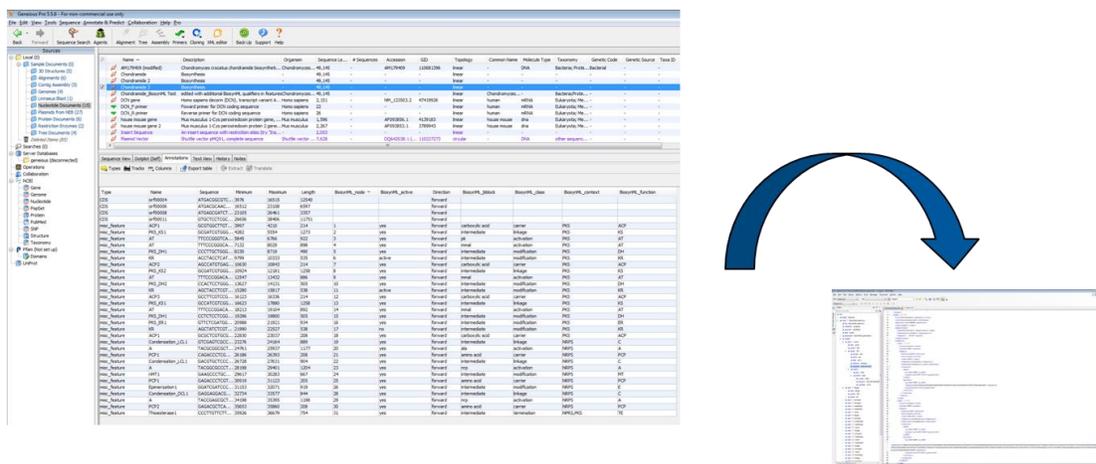


Figure 3.12: Export of BiosynML file from Geneious BiosynML plugin

In summary, the BiosynML Geneious plugin provides all essential functionality for users to interact with biosynthetic models in the course of this project, while interfacing to the antiSMASH analysis pipeline ensures that the output of this de-facto standard tool can be seamlessly channelled into downstream analysis workflows (see also previous Figure 3.1).

3.1.4 BiosynML Editor for manual creation of pathway models

In order to generate hypothetical pathways, in the sense that no sequence is known but a string of enzymatic activities (domains) which may produce a metabolite is deduced from the chemical structure of a known compound, an editor has been implemented where the user can create a biosynthetic model using functions like drag and drop of domains. This editor is typically used by the researcher to create pathway models manually for newly discovered and then structurally elucidated compounds. This allows researchers to store and compare their “estimated” pathways making use of BiosynML functions. In particular, researchers can also submit their pathways directly to the matching function to find the similar pathways that are available in the database (Figure 3.13) even in the absence of sequence information for the strain producing a novel compound.

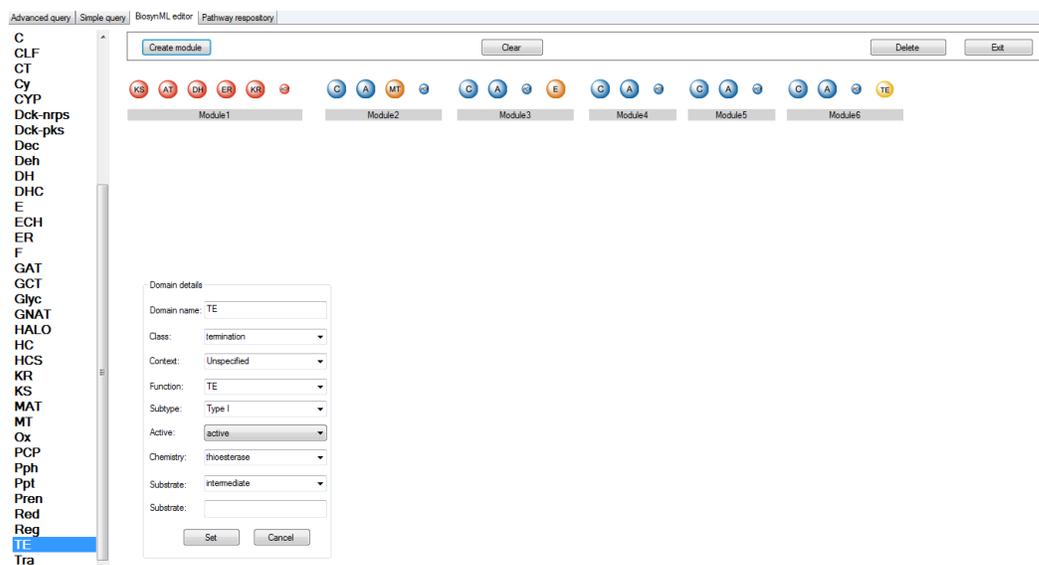


Figure 3.13: BiosynML editor for drawing hypothetical modules of biosynthetic pathway which can be submitted for search and match algorithm

3.1.5 Integration of BiosynML with Mxbase

Import of BiosynML models into Myxobase

Myxobase is the Helmholtz Institute's collection of data that is related to research on myxobacterial strains and compounds. The technical platform for Myxobase is the Mxbase system (see chapter 2.2) which is built on a relational database, making use of the widely popular MySQL technology. It helps to organize all accumulating research results, such as bioactivity screening results according to a strain or compound, thus facilitating to connect these data. Similar to metabolomic data, the BiosynML information on genomes and pathways encoded therein is a large amount that would become unmanageable in spreadsheet form or on hard disk with the increasing amount of genomes that are sequenced. To enable better collaboration on results, a frontend function was added to Mxbase to transfer the information from the BiosynML files (generated by antiSMASH or Geneious plugin) and store all relevant data in the Mxbase database. In order to maintain records for ongoing use, BiosynML identifiers are assigned to strains and compound families which makes it easy to search and retrieve the information and also to generate reports (Figure 3.14). Furthermore, this biosynthetic model repository serves as the basis for more extensive bioinformatics analysis and genome-mining processes, triggered automatically or manually by individual users. Throughout this study, the BiosynML pathway repository

in Myxobase was updated whenever new genomes became available, or when new or existing pathways were characterized.

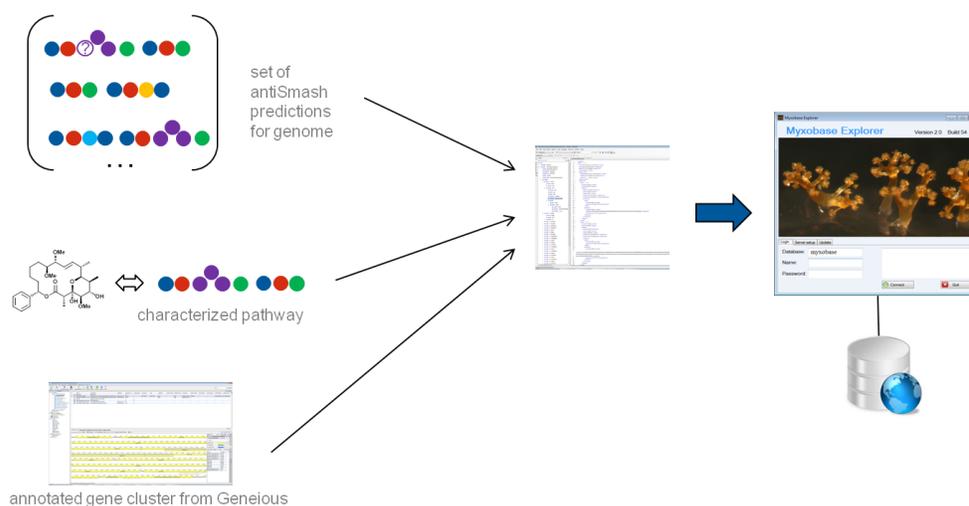


Figure 3.15: A schematic representation of integrating BiosynML into Myxobase

All the features from BiosynML pathway datasets are stored in the Mxbase biosynthetic repository with a unique biosynthetic pathway key linked to strain and compound key. The interface (Figure 3.15) designed for importing has a function that can read the strain and compound information from the .xml file supplied via the Geneious BiosynML plugin after curation of the pathway annotation initially generated by antiSMASH. This information is then assigned to the biosynthetic pathway while it is imported into Myxobase biosynthetic pathway repository.

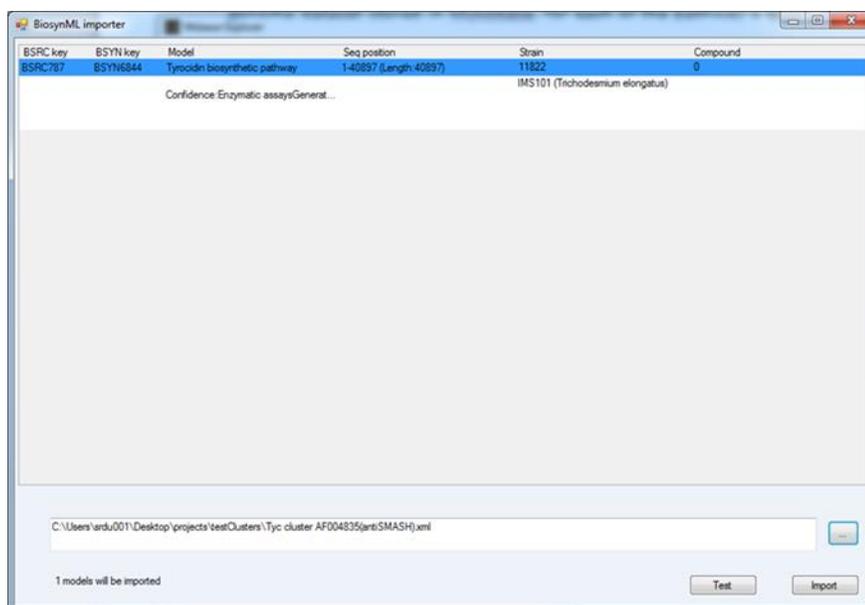


Figure 3.14: BiosynML importer interface for biosynthetic gene clusters (.xml file) integrated in Mxbase Explorer

Myxobacterial pathways overview

The Mxbase explorer is an interface that provides access to the various gene clusters in the BiosynML repository that are linked to strains. This provides a medium for the researcher to have an overview of the clusters that a strain harbours. The interface “pathway repository” generates a detailed report of all the pathways that are available or a detailed report of the pathways with restricted taxonomy (class, order, suborder, family, genus and species) (Figure 3.16). The import of pathways and assignment to a strain will immediately highlight this respective strain as a potential producer of all compounds linked already to gene clusters contained in that report. In addition, a function is implemented in the Mxbase explorer which can extract the information from the BiosynML content stored and generate a ball scheme representation for the biosynthetic model which is similar to the representation created manually by researchers. It enables users to access the detail information on models / nodes (Figure 3.17).

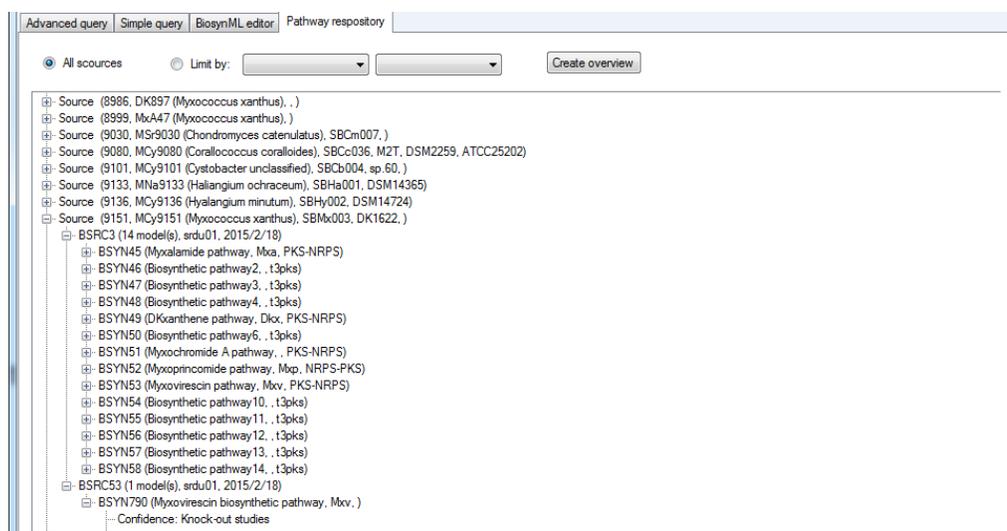


Figure 3.16: Biosynthetic pathway repository overview window showing the list of pathways that is available in the database

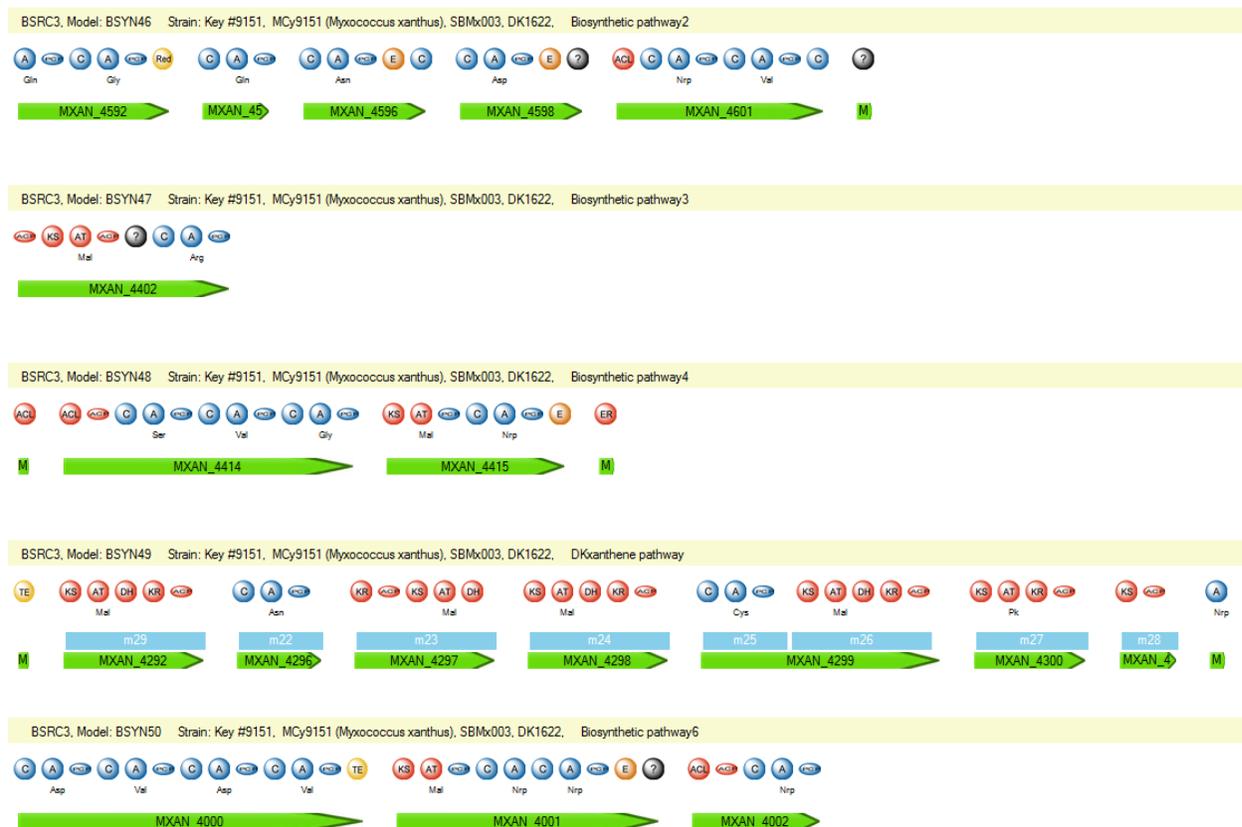


Figure 3.17: The “pathway overview” window showing ball scheme representation of biosynthetic gene clusters

3.2 The BiosynML analysis engine

In the previous sections the establishment of peripheral components of the BiosynML bioinformatic framework for conceptual genome mining and the technical prerequisites for its operation were described. The following part of this work deals with the development of methods and algorithms which are at the heart of the analysis framework, i.e. which form the core analysis engine. The main objective is to create a set of tools allowing to compare biosynthetic models on the basis of conceptual similarity as opposed to primary sequence similarity. Key considerations for these tools include the requirement to match single pathways with a library of well-characterized pathways, to search for pathways showing architectural similarity and rank these accordingly, and to create an overview of biosynthetic model diversity within a potentially extensive database of (predicted and characterized) pathways.

3.2.1 Algorithms developed for conceptual genome mining

A typical modular secondary metabolite biosynthetic pathway of PKS, NRPS or hybrid type (example in Figure 3.18) consists of a set of genes $G = \{g_1, g_2, g_3, \dots, g_n\}$, where each gene encodes several domains with predefined functions. Each domain has properties such as function (overall biochemical role of a domain), substrate (the monomer incorporated or modified by the catalysed reaction) and status (whether the domain is active or inactive).



Figure 3.18: Biosynthetic gene cluster for surfactin biosynthesis in *Bacillus subtilis* strain w168, repeatedly used as an example in the following

We consider the problem of finding the similarity of a given pathway of interest, which we call query pathway Q , against a database of pathways $P = \{P_1, P_2, P_3, \dots, P_n\}$. Here, we call each database pathway P_i as a target pathway. In the course of this project three different methods for comparing and finding the similarity between the query and each of the target pathways were developed. Since each pathway is a sequence of genes, and the distribution of domains doesn't follow a fixed pattern, hence matching should not be based on a symmetric relation i.e., $\text{Sim}(Q, P_i) \neq \text{Sim}(P_i, Q)$. Thus, we match pathways in both ways, i.e., we evaluate $\text{Sim}(Q, P_i)$, $\text{Sim}(P_i, Q)$, which enhances the similarity score between query and target pathways by comparing all possible genes and gene combinations in either of the pathways. We note that the similarity between two genes essentially depends on the domain compositions of the genes compared.

Approach 1: LocalBestGreedy (LBG)

The first method, termed LocalBestGreedy, performs gene-wise comparison of pathways taking into account the domain compositions based on similarity calculations over genes. To find the similarity between the query pathway Q against a database of pathways P , the first step is to find a gene (or a consecutive sequence of genes of a given window size w) in the target pathway having a similar domain composition as the query gene.

We emphasize here, that we compare a single gene from the query with a consecutive sequence of more than one gene from the target pathway; the length of such sequence (or the number of genes considered) is limited by a user-defined maximum window size w_{\max} .

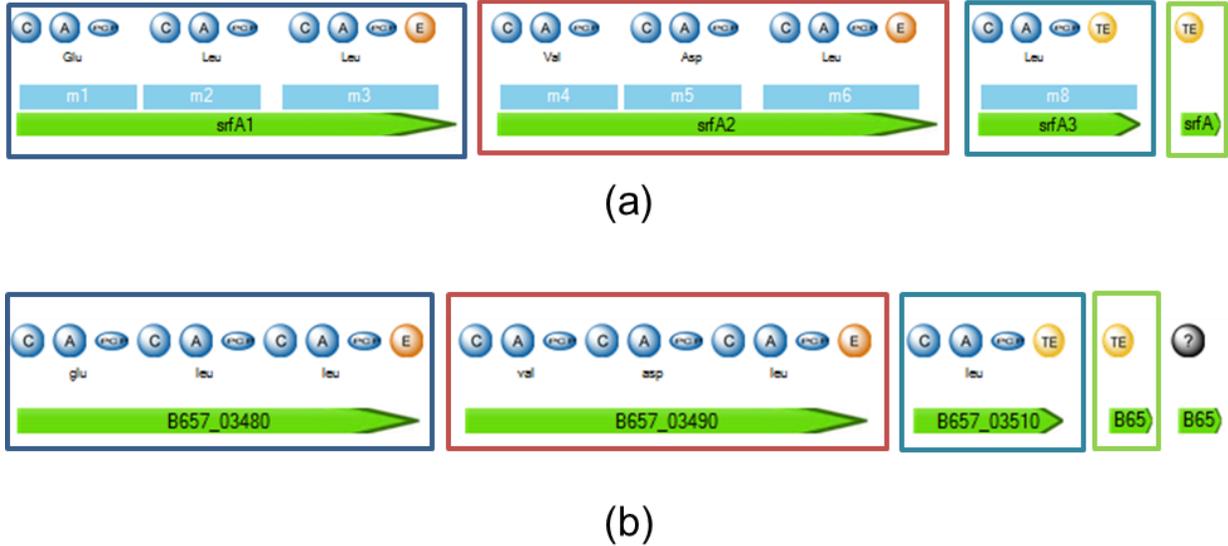


Figure 3.19: Search for genes with similar domain composition using cosine similarity. (a) query pathway from *Bacillus subtilis* strain w168, (b) target pathway from *Bacillus subtilis* strain QB928. The colour of the boxes shows the expected matching of the genes respectively

To make it more precise, let us define G_p^w as a set of gene sequences, where each sequence is obtained by concatenating w consecutive genes from the target pathway P . We now describe how we measure similarity between query gene g_Q and G_p^w . For this, we need a similarity measure between two genes g_Q and g_p^w , which we denote by $\text{Sim}(g_Q, g_p^w)$. We define

$$\text{Sim}(g_Q, G_p^w) = \max_{w \in \{1, 2, \dots, w_{\max}\}} \max_{\{g_p^w \in G_p^w\}} \text{sim}(g_Q, g_p^w)$$

The next question is how to measure the similarity between two genes g_Q and g_p^w . The similarity between the domains of two genes is calculated considering the domain properties, namely function, substrate and status. We call the similarity based on function, substrate and status as functional similarity, substrate similarity and status similarity respectively.

We denote functional similarity, substrate similarity and status similarity as S^F , S^{Sb} , S^{St} respectively between domain compositions of two genes, calculated as

$$S^F(g_Q, g_p^w) = \max(\text{cosineSimilarity}(g_Q, g_p^w))$$

$$S^{Sb}(g_Q, g_P^w) = \max(\text{cosineSimilarity}(g_Q, g_P^w))$$

$$S^{St}(g_Q, g_P^w) = \max(\text{cosineSimilarity}(g_Q, g_P^w))$$

the final scores for any two genes is given as the weighted mean of S^F , S^{Sb} and S^{St} where the coefficients are the weight of function (WF), the weight of substrate (WSb) and the weight of status (WSt)

$$S = (WF * S^F(g_Q, g_P^w)) + (WSb * S^{Sb}(g_Q, g_P^w)) + (WSt * S^{St}(g_Q, g_P^w))$$

$$S(g_Q, g_P^w) = S/3$$

By experimenting with various weight values, we arrived at the following combination: Through manual analysis of the results using various weight values, the best results are obtained at WF = 2.9, WSb = 0.05 and WSt = 0.05. Note that the weight of the domain function is much higher than the other weights. This is because the biosynthetic pathways, predicted through antiSMASH, can have occasionally problems with accurate substrate and status prediction whereas the type of domain is in most cases assigned with high confidence. Moreover, this setting also simply reflects that the basic biochemical role of a domain is its most relevant feature when defining the enzymatic activity string underlying the biosynthesis of a PKS/NPS metabolite.

In Figure 3.19, using cosine similarity, the query genes are mapped to that of the target gene with similar domain composition within a given window size. The bidirectional matching (i.e., matching between and query and target both ways) results in a so-called global set where all the genes in the query are matched to all the genes in the target. High scoring pairs (HSP) from the global set are extracted based on their scores.

The overall quality between query and target is calculated based on the alignment of domains from HSPs found using the Smith–Waterman algorithm (121). More precisely, we count the number of matches and mismatches in each HSP by aligning domains of each property across all the hits. This contributes to the similarity score between query and target pathways.

To enhance the similarity score of domain composition between pathways, the weight of the matches (m) is chosen as 3 and mismatches (mm) weight is chosen as 2 providing 60% weight to the matches and 40% to the mismatches.

Let us define MF as the number of matched functional property, MMF as the number of mismatched functional property, MSt as the number of matched status property and MMSt as the

number of mismatched status property. Similarly, MS_b is the number of matched substrate property and MMS_b is the number of mismatched substrate property.

Then the raw scores are calculated as

$$\begin{aligned} \text{rawFunctionScore (RF)} &= \left(\sum_{i=0}^v MF_i \right) * m - \left(\sum_{i=0}^v MMF_i \right) * mm \\ \text{rawSubstrateScore (RSb)} &= \left(\sum_{i=0}^v MSb_i \right) * m - \left(\sum_{i=0}^v MMSb_i \right) * mm \\ \text{rawstatusScore (RSt)} &= \left(\sum_{i=0}^v MSt_i \right) * m - \left(\sum_{i=0}^v MMSt_i \right) * mm \end{aligned}$$

where $v = 1, 2, 3, \dots, R$, where R is the number of HSPs.

Bit-score is a normalized score, expressed in bits, that estimates the magnitude of the search space to look through before finding a score as good as or better than other one by chance. Bit-score is calculated based on Althshul definition (124),

$$\begin{aligned} \text{FunctionBitscore (BF)} &= \frac{\lambda * RF - \ln(k)}{\ln 2} \\ \text{SubstrateBitscore (BSb)} &= \frac{\lambda * RSb - \ln(k)}{\ln 2} \\ \text{StatusBitscore (BSt)} &= \frac{\lambda * RSt - \ln(k)}{\ln 2} \end{aligned}$$

where $\lambda = 1.39$ and $k = 0.747$ are the constants.

The overall bit-score (BS, for short) is calculated by taking mean of all the bit-scores obtained from properties of domains in the pathways. The weights of the substrate (sb) are defined by user depending on the importance of substrate matched and to reduce the influence of status on the overall score, the weight of status (st) is set at 0.2,

$$BS = BF + (BSb * sb) + (Bst * st)$$

In the context of database searches, the E-value is the number of distinct matches with a score equivalent to or better than BS, that are expected to occur in a database search purely by chance. The lower the E value, the more significant the score is,

$$E = kXYe^{-BS},$$

where X is the length of domains in the query pathway, Y is the size of the database, BS is the bitscore calculated and k is the constant from Althshul definition.

The advantage of the greedy approach over a window of genes (a combination of domains obtained by neighbouring genes within a given window size provided by the user) is that it requires only a small amount of memory and it is much faster to compute. However, since the greedy approach decides the next best step by exploiting only the local information (here, the window size) without considering the global structure of the problem, one cannot guarantee to find an optimal solution. An example illustrating this scenario is shown in Figure 3.20, where the target pathway has similar domain composition but different operon organization and domain distribution, i.e. the domains from the set of query genes are relocated in the target genes, alongside other differences such as predicted substrate specificity. In such cases, the algorithm fails to identify the pathways although the two pathways have overall similar domain composition and could be plausibly regarded as candidates for the production of related molecule classes, based on biochemical considerations. Without providing a proper window size the hit might get a low score and hence it is buried in the other hits among the results. It should be pointed out here that deviations shown between biosynthetic models in Figure 3.20 are not arbitrary and could actually occur to this extent between pathways which are nevertheless similar from a biochemical point of view. Modules could be re-arranged, split across several genes or combined on one gene (e.g. m1, m2, m3; m4, m5, m6) and still bring about the same molecular substructure in the product. Similarly, single-standing domains in the vicinity of multimodular genes could be located differently, and additional domains (possibly of unassigned type) could be present in the compared pathways. Thus, in the context of conceptual genome mining one must demand for increased robustness of the method in order to include in the result set also the pathways exhibiting a considerable extent of deviation in terms of operon organization and the distribution of domains across genes.

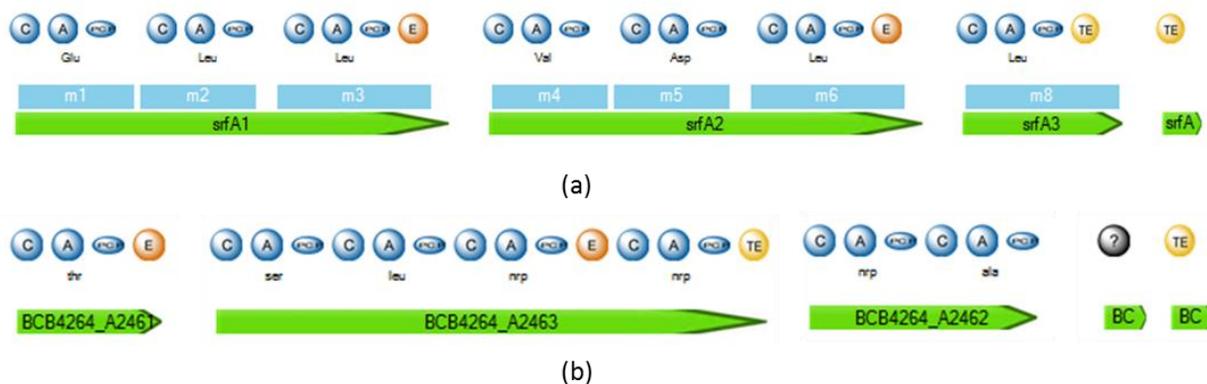


Figure 3.20: An example highlighting the disadvantage of LBG where the algorithm fails as the domains from the query genes are differently located in target genes. Despite the different domain distribution, these pathways could be considered potential conceptual relatives and in principle the chance to achieve matching through the BiosynML comparison engine would thus be desired.

Approach 2: LocalBestCoalesce (LBC)

Note that the first method LocalBestGreedy (LBG, for short) has the disadvantage of missing hits where sets of domains are distantly located/distributed in a target pathway, albeit overall domain composition is similar. To overcome the disadvantage a modified version of LBG, named LBC, was developed that matches domain compositions of pathways iteratively (Figure 3.21).

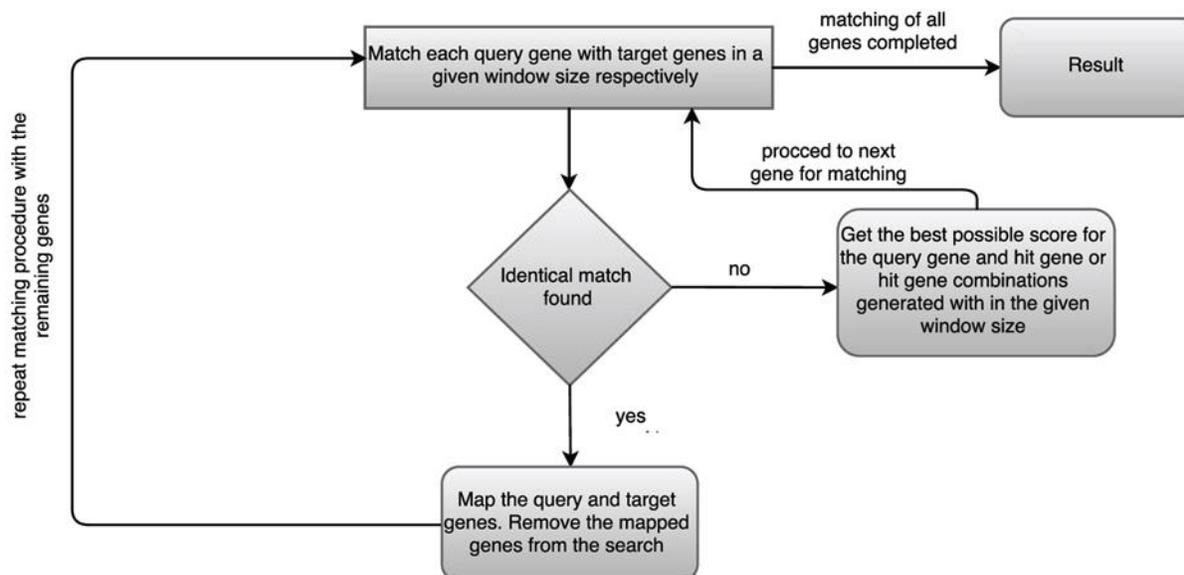


Figure 3.21: A flow chart describing the iterative process of matching genes between query and target pathway where each identical query and hit genes are mapped and removed from the search space

If a hit for domain composition of a query gene is obtained, then the hit and query gene are removed from the search space. The procedure is repeated with the remaining domains of the genes in the pathway. In this way with every iteration, the neighbouring genes will be changed and can lead to discovering pathways with high similarity even though the domains composition of genes are distantly located. Because of the iterative removal and matching, this method is able to identify domain subgroups which are distantly located. However, one main disadvantage is that in order to identify distantly located domain composition, the pathways have to be of similar size. Otherwise, the algorithm fails to find similar pathways which are only partially predicted and are thus missing a number of domains. This may happen especially when working with draft genomes scattered across increased numbers of scaffolds, showing a high probability that predicted biosynthetic models are incomplete. The problem is illustrated with an example in Figure 3.22.

Due to the presence of genes (highlighted in blue box) with mismatching domains in between potential hit target genes in the pathway, LBC fails to match the domains of query gene *srfA1* to the target genes *BCB4264_A2461* and *BCB4264_A2462*, if a proper window size is not provided. For e.g. with a given window size of 2, the query gene *srfA2* and *srfA3* will result as a perfect match with domains of target gene *BCB4264_A2463* in the target pathway. But domain of query gene *srfA1* will not be matched to domains of target genes *BCB4264_A2461* and *BCB4264_A2462* due to the presence of genes with unspecific domain between them. This results in having a low similarity score for the hit pathway despite of having identical domain composition along with additional domains.

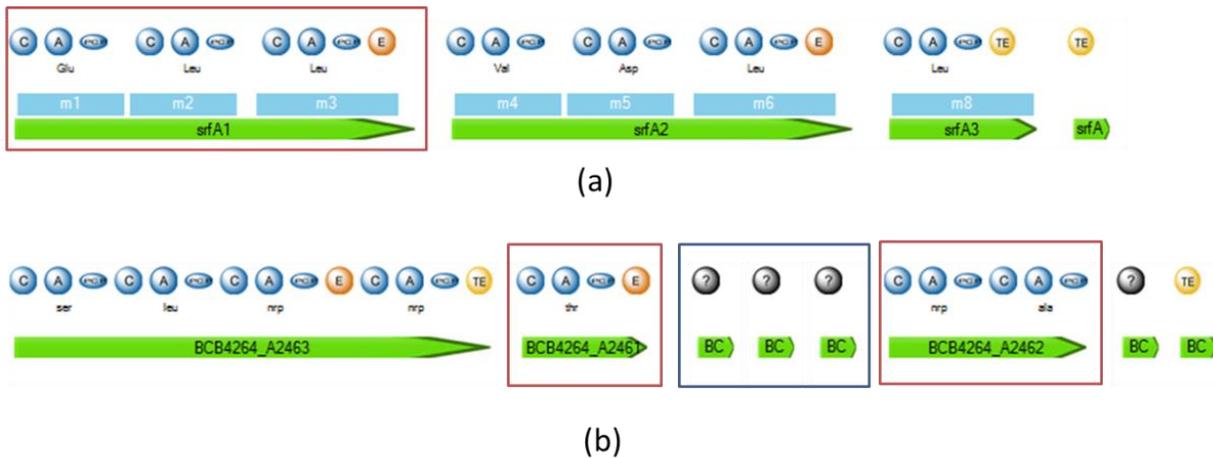


Figure 3.22: An example highlighting the disadvantage of LBC. Due to the presence of genes unidentified domains/genes (highlighted in blue box) in between potential hit target genes in the pathway, LBC cannot efficiently match the domains of query gene *srfA1* those located on target genes *BCB4264_A2461*, *A2462*.

Approach 3: GlobalBestSubset (GBS)

Since the methods LBG and LBC depend on the window size that heavily influences the outcome, both algorithms have problems to locate a pathway with distantly located domains groups. Thus, a new method is needed that is independent of the window size and should be able to identify hits where distinct domain arrangements are potentially disconnected or even distantly located. Considering this requirement, a new algorithm was developed which we call here as GlobalBestSubset (GBS for short). In this approach the concept of set intersection is used (125) which can find pathways that have distantly located domain compositions without depending on the window size.

Let $S(Q, P_i)$ be the similarity between the query pathway Q and one of the pathways P_i , $i = \{1, 2, 3, 4, \dots, R\}$ from the database P containing R pathways, and let $S_g(g_Q, G)$ be the similarity between domains of a gene in query and a list of genes from pathway where $G = \{g_1, g_2, g_3, \dots, g_n\}$.

As described earlier in LBC ad LBG, the similarity between domains of two genes is calculated based on domain properties namely function, substrate and status.

Let S_{jk} be the similarity score between gene g_j of query pathway and gene g_k of target pathway, calculated by intersecting the individual domain properties in the genes compared,

$$S_{jk} = S(g_j^Q, g_k^{Pi}) \quad \text{where } g_j^Q \in Q, g_k^{Pi} \in P_i$$

Again, we refer similarity based on function as functional similarity, substrate as substrate similarity and status as status similarity.

We denote functional similarity, substrate similarity and status similarity as S^F , S^{Sb} , S^{St} respectively between domain compositions of two genes and MF as the number of matched functional property, MMF as the number of mismatched functional property, MSt as the number of matched status property and MMSt as the number of mismatched status property. Similarly, MSb is the number of matched substrate property and MMSb is the number of mismatched substrate property, calculated as,

Let M = length of g_j , N= length of g_k

$$\text{Functional Match (MF)} = |g_j^F \cap g_k^F|$$

$$\text{Functional mismatch (MMF)} = (M - MF) + (N - MF)$$

$$S^F(g_j^F, g_k^F) = \frac{MF}{M} + \frac{MF}{N}$$

$$\text{Substrate Match (MSb)} = |g_j^{Sb} \cap g_k^{Sb}|$$

$$\text{Substrate mismatch (MMSb)} = (MSb - M) + (MSb - N)$$

$$S^{Sb}(g_j^{Sb}, g_k^{Sb}) = \frac{MSb}{M} + \frac{MSb}{N}$$

$$\text{Status Match (MSt)} = |g_j^{St} \cap g_k^{St}|$$

$$\text{Status mismatch (MMSt)} = (MSt - M) + (MSt - N)$$

$$S^{St}(g_j^{St}, g_k^{St}) = \frac{MSt}{M} + \frac{MSt}{N}$$

The final scores for any two genes are given by putting more weight on functional property, which is a linear combination of F, Sb and St with weight of function (WF), weight of substrate (WSb) and weight of status (WSt)

$$S = (WF * S^F(g_j^F, g_k^F)) + (WSb * S^{Sb}(g_j^{Sb}, g_k^{Sb})) + (WSt * S^{St}(g_j^{St}, g_k^{St}))$$

$$S_{jk} = S(g_j^Q, g_k^{Pi}) = S/3$$

This calculation is repeated for domains of each and every gene of the query pathway with all the genes in the target pathway. If there are no identical matches found, then the algorithm proceeds to the next step to find similar composition of domains by combining multiple genes with respect to the domain composition of the query gene.

Let, $G_t = \{g_1, g_2 \dots g_m\}$ be the top hit genes from the target pathway P_i to a gene g_Q in the query pathway and let $S_{g_Q I_k}^k$ be the similarity between g_Q , let I_k be the concatenation of any of the k genes of G_t . Here we vary k from 1 to m where m is chosen as 4 in our experiments to limit the combinatorial explosion in the computational time. This process is repeated for all the genes in the query pathway to find the best hits in the target pathway. Each gene is mapped to the target gene with a similarity score S_{jk} . Finally we take the best of all the scores $\max\{S_{jk}, \max_{I_k} S_{g_Q I_k}^k\}$

The bidirectional matching (i.e., matching between query and target forward and backwards) results in a so-called global set where all the genes in the query are matched to all the genes in the target. High scoring pairs (HSP) from the global set are extracted based on their scores. To get the overall quality of the pathways matched, the bitscore and E-values are obtained by applying the methods described in section 2.2.1.

Pseudocode:

Method *SubsetMatchingOfPathways()*

forward = SubsetsMatching(Source, Target)

reverse = SubsetsMatching(newTarget, newSource)

globalmap = combine (forward, reverse)

bestResultMapping = GetBestFromGlobalMap(globalMap);

Bitscore = CalculateBitScore(bestResultMapping);

end

Method *SubsetsMatching*

Input: Q, T

Output: score

for each $g_j \in Q$

for each $g_k \in T$

compute $S(g_j, g_k)$

end

Let $G_t = \{g_1, g_2 \dots g_m\}$ be the top hit genes

compute $S^k(g_Q, I_k)$, where I_k is concatenation of any of the k genes of G_t

```
Let  $I_{k^*}$  be the best match for  $g_Q$ 
score( $g_Q, I_{k^*}$ ) =  $S^k(g_Q, I_{k^*})$ 
end
end
```

The pathway matching algorithms described earlier in this section showed the results of LBG and LBC, which majorly depend on the windows size setting provided by the user. Although these algorithms are fast, in cases of complex pathway matching where the genes are distantly located, these algorithms fail to identify those particular pathways as a good hit that are supposed to appear on the top of the output with high similarity scores. GBS uses the concept of set intersection and works independent of windows size, which makes it robust compared to other algorithms. Since it considers evaluating all combinations of genes and their domain compositions, there is no chance of missing pathways with genes having highly similar domain composition even when domain subsets are distantly located.

3.2.2 Usage scenarios and comparison of BiosynML methods

The BiosynML methods presented in the previous section 3.2.1 were devised with several application scenarios in mind, as has been outlined already in chapter 1.6. Perhaps the most prototypical task which occurs repeatedly throughout research with secondary metabolite pathways is the comparison of previously characterized (“known”) pathways to predicted genome-encoded biosynthetic models stored in a database using a scoring function, where the scores account for the differences between the predicted and established pathways like missing modules, extra modules, ambiguity of predicted properties, altered positions of the genes and domains and their additional properties (meta information like specificity, active or inactive, etc.). Usually this initial comparison is followed by expert review and by additional manual analysis which may later extend to the complex genome context. The pathways in the database are considered for in depth comparisons based on the genes, domains and their arrangements, predicted substrates etc. A similar approach, though ideally with less manual supervision, can be basically taken for the automatic tentative identification of predicted biosynthetic gene clusters, by comparing each predicted pathway in a newly sequenced genome to the reference library of well-characterized pathways. Finally, a “compare all-to-all” approach can be taken to establish a similarity matrix based on gene cluster architectures, in order to reveal groups of similar clusters and outliers.

Taking care of all these requirements, a “conceptual genome mining” module was developed and integrated into the Mxbase application. The newly added functionality comprises of three main components: a windows-based application interface, additions to the Mxbase back-end (MySQL data

structures on the server, being the foundation of Myxobase) and a worker thread, implemented as a remote procedure call (RPC) framework that communicates between functions located on the server and the graphical user interface. The design and functions of components added to Mxbase in the course of this project have been depicted in Figure 3.1 and earlier in Figure 1.12, and their use is described in the following section.

The interface is equipped with numerous options and enables the user to make a query against the Myxobase to generate results covering all pathways from the BiosynML pathway repository. The pathways which make up the search space are then passed to the algorithms where major computational tasks are involved. Myxobase is the database (the knowledge repository for secondary metabolite annotation) which contains information about the known gene clusters along with their meta-data, as well as genome sequences from different strains of myxobacteria.

After the retrieval of pathways from the database the worker thread starts computing the 2nd pass calculations such as functional deviations, substrate deviations, domain status deviations and scores them, to finally return compiled results to the user.

3.2.2.1 Genome mining using the BiosynML engine

Predicted pathways from genomes enter the BiosynML repository via a route briefly consisting of these steps: The raw data from sequencing is passed through the genome assembly pipeline. The assembled data is passed through antiSMASH to extract biosynthetic meaningful information from complex datasets and then inventorize the full complement of putative gene clusters in the form of representative domains and their properties (the pathway models). The dataset obtained after processing through antiSMASH is deposited in the BiosynML repository inside Myxobase and consists of domains found and their function and meta-information attached. The “Conceptual genome mining” tool has the function to identify the known models from a genome using any of the algorithms developed such as LocalBestGreedy (LBG), LocalBestCoalesce (LBC) or GlobalBestSubset (GBS). Thereby, known biosynthetic clusters are highlighted and the obtained report can in principle provide the information regarding similar pathways that are present in various strains. Thereby, users can also take advantage of extended and up-to-date information from the BiosynML collection which is supplied by other users. Most importantly, the BiosynML workflow decouples the pathway analysis from the need to access the raw sequences or GenBank files. This application is used to answer common questions like,

- which of the predicted pathways in a new genome is similar to any of the known models

- is there a pathway which matches the hypothetical biosynthesis for a newly elucidated compound

The user's expectation for the algorithm is to reliably report similar pathways ("primary hits"), when using a biosynthetic model as input, without sequence-based comparison ("conceptual") and in presence of typical sequencing- or prediction-derived errors and uncertainties. Furthermore, the identification of putatively related pathways must work robustly also in a background of many pathways per genome, containing a multitude of modules when some of them would just by chance have a domain composition similar to a small portion of the query model (distinguishing them as "secondary hits").

The Pathway query module is implemented with input categories, the available search settings provided are:

1. Input of a pathway(s) (one specific pathway or a complete set of pathways from a genome, or model for a hypothetical pathway designed by researcher using BiosynML Editor, see 3.1.5)
2. Choice of available search and match algorithms (LBG, LBC, GBS)
3. Search space (entire database or restricted to selected genus or species)
4. Search parameters (window size, substrate specificity, additional domain penalty, collinearity and pathway completeness)

Input

The input for a pathway query is information relating to biosynthetic pathway(s). Based on the types of available information and specific interests of the user, three kinds of input are implemented and available, which might be biosynthetic pathway(s) from repository linked to a strain or compound

- pathway from external source
- pathway model designed through the BiosynML editor

For the pathways which are given as input, the interface is able to handle a single pathway or a set of pathways that are obtained from a genome linked to strain. Characterized pathways typically originate from the Geneious plugin; hypothetical pathways are created by users in the BiosynML editor.

Search and match algorithms

See section 3.2.1 for detailed explanation of algorithms implemented for matching and scoring of biosynthetic pathways.

Search space

Myxobase contains an organized collection of biosynthetic clusters that are linked to strain and compound information. The researcher can limit the search space to a particular genus or species prior to searching for pathways similar to the query pathway. While the researcher aims to find the best similar pathway among the all pathways across all strains, each pathway in search space represent one feasible solution marked by its value or fitness for the query.

Search parameters

The scoring reflects not only the overall similarity, but allows the investigator also to evaluate the differences regarding domain composition of modules, their distribution across genes, pathway organization, and match of substrate specificity (“meta information”) for domains in corresponding positions. The scores also accounts the differences between the predicted and established pathways like the missing modules, extra modules, ambiguity, positions of the genes and domains and their additional properties (meta information like specificity, active or inactive, etc.). The possibility to parametrize the search was developed which is exemplified by adding options for intuitive parameters:

Window size: the algorithms GBG and GBC requires user to define window size to select various aspects of limitations for the search in order to yield more confident results.

Substrate specificity: it defines the importance of the substrate that has to be matched during the search influencing the overall score of the hits. The range of the substrate specificity is [0, 1] where

0 – no importance

1 – very important

Pathway completeness: it defines the importance of finding complete set of domains irrespective of additional domains present. This will boost the overall score of the hits which has all domains with respect to the query.

Collinearity: this is the parameter is to search efficiently for a subsequence of domains or generally a pattern in large sequences of domains arranged in the same order in various pathways preserving collinearity. This will boost the score of the hits which have highly similar arrangements of the domains spanning an extended range. The importance can range between 0 and 1

0 – no importance

1 – very important

Additional domain penalty: it defines how strongly extra domains present or missing domains in the genes that are not predicted by the tool are penalized. Values ranges between 0 and 1.

0 – no importance

1 – very important

The screenshot displays the BiosynML interface. The 'Query input' section on the left allows users to select a query from Mxbase (Strain: Cmc5, Key: 4264) or from a local BiosynML editor/file. The 'Settings' section on the right includes algorithm selection (LocalBestGreedy (LBG), LocalBestCoalesce (LBC), GlobalBestSubset (GBS)), search space restrictions (Genus, Species), and adjustable parameters like Window size (3), Substrate specificity (0.3), Collinearity (0), and Extra domains penalty (0.1). A 'Run' button is located at the bottom right.

Below the settings, a pathway diagram shows the Chondramid biosynthesis pathway with domains m1 through m7 and genes cmdA through cmdD. The 'Results (Most recent first)' table is shown below:

JobID	Query identifier	Algorithm	Substrate penalty	Domain penalty	Collinearity	Date
bb1af48d562544...	4264	GBS	0.3	0.1	0	11/16/2015 1:19...
47772d0f0ab141...						11/16/2015 1:16...
707aa969f18444...	4264	GBS	0.3	0.1	0	11/16/2015 1:16...
7bc698079c1144...	7693	GBS	0.3	0.1	0	11/16/2015 12:2...

The 'Status' window on the right shows: Pathways searched: 1867, No of sources: 123, Calculated maximum score: 268.929.

Figure 3.23: BiosynML interface for submitting query pathway to the search and match class with parameters, integrated in the Mxbase Explorer application.

These parameters influence the scoring function of the algorithm used for search and compare, reflecting the positioning of pathways in the result list.

To reduce the burden of the processing on the user's machine, the search and match functions are located on Mxbase server which performs the operations based on user input and stores the results in the database which are subsequently retrieved by the client. Access to BiosynML functions on Mxbase server is through a dedicated interface integrated into the Mxbase Explorer application (Figure 3.23). This interface covers all the aspects with default values of the parameters. Typical use can be exemplified using the Chondramid gene cluster from strain Cmc5 (*Chondromyces crocatus*) where the query pathway (known and characterized for the production of Chondramides (126)) is submitted to the search and match class, using GBS algorithm and setting substrate specificity to 0.3 and additional domain penalty to 0.1. The results can be accessed through the results window designed with a datagridview, giving an overview of the matching results (Figure 3.25), a bit score plot and also a ball scheme representation for visualization (Figure 3.24 and 3.26).

The bit score plot generated reveals a confidence interval where the hits above the confidence interval (the “jump” in the plot) have domain architecture highly similar to that of the query pathway (Figure 3.24).

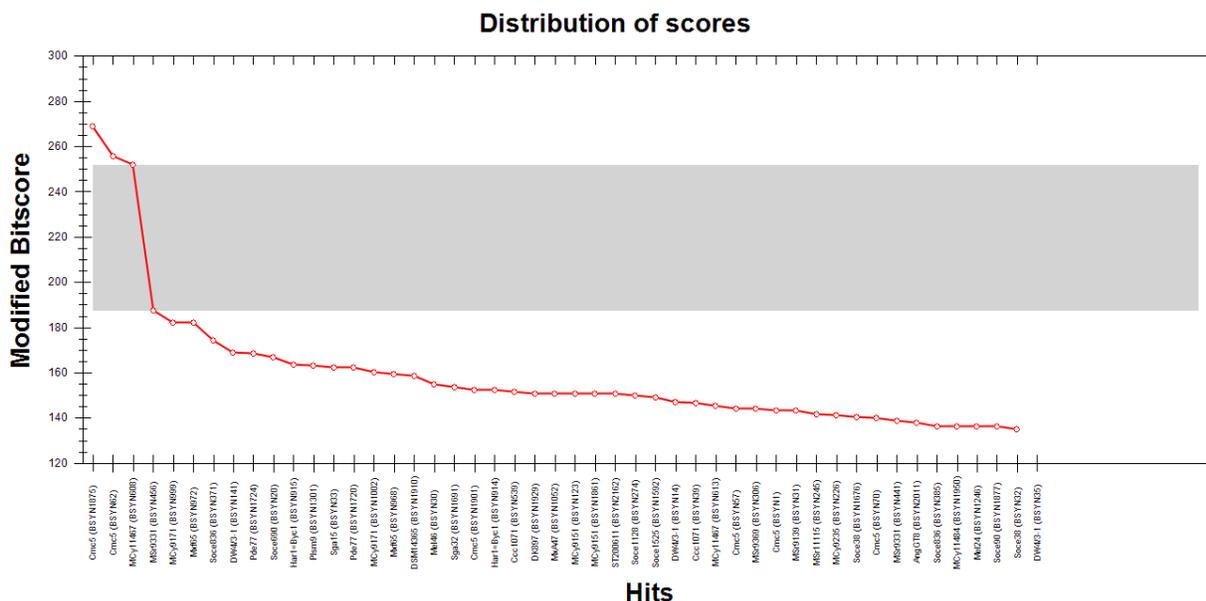


Figure 3.24: Bit score plot generated for Chondramid pathway from *C. crocatus* Cmc5 used as query. The grey box represents the confidence “interval” (in the sense of a gap separating high-confidence and low-confidence hits). The pathways above the grey box are highly similar to the query pathway. The high affinity hits above the confidence interval are from auto-annotated genomes of strains Cmc5 (*Chondromyces crocatus*) and MSr9030 (*Chondromyces catenulatus*) genomes. Note, the first hit is always the “self-hit” which sets the theoretical maximum score. Due to differences resulting from manual curation, the auto-annotated version of the same pathway from a genome might not reach the same score.

We here emphasize that the first hit is always a “self-hit” which sets the theoretical maximum score based on which the confidence interval is generated. The second hit in this example is from the auto-annotated Cmc5 genome and the third hit is from MSr9030 (*Chondromyces catenulatus*) (ball scheme representation of query and target pathway are displayed in Figure 3.25). The latter pathway is highly similar to the Chondramid (query) pathway; however it lacks the genes for a tailoring halogenase and the β Tyr precursor-generating aminomutase. This finding might be explained by missing prediction during auto-annotation, or the genes are located elsewhere in the genome (in theory the pathway might also produce only non-halogenated derivatives; however the strain has been shown before to produce the chlorinated chondramides, too (127)). Distribution of PKS domains across genes is also slightly different, a possible consequence of operon re-arrangement during evolution, or as an artefact during sequencing since the split occurs in an unusual way between genes that together harbours the complement of domains from cmdA. Despite these deviations, the detection of a high-scoring candidate *cmd* gene

cluster in *C. catenulatus* by the BiosynML approach in this example is plausible and anticipated as the strain is known as a chondramide producer (128).

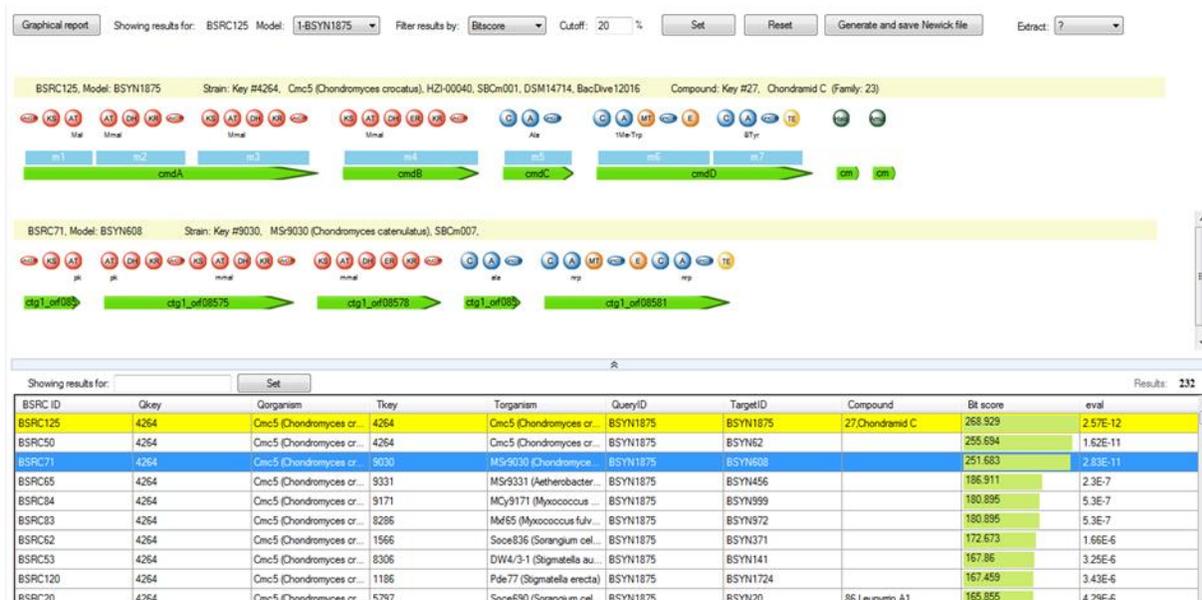


Figure 3.25: Results window with summary table and ball scheme representation of the query and hit.

Furthermore, analysing the first hit below the confidence interval, a NRPS/PKS hybrid pathway from the strain MSr9331 (*Aetherobacter rufus*) is found (Figure 3.26).

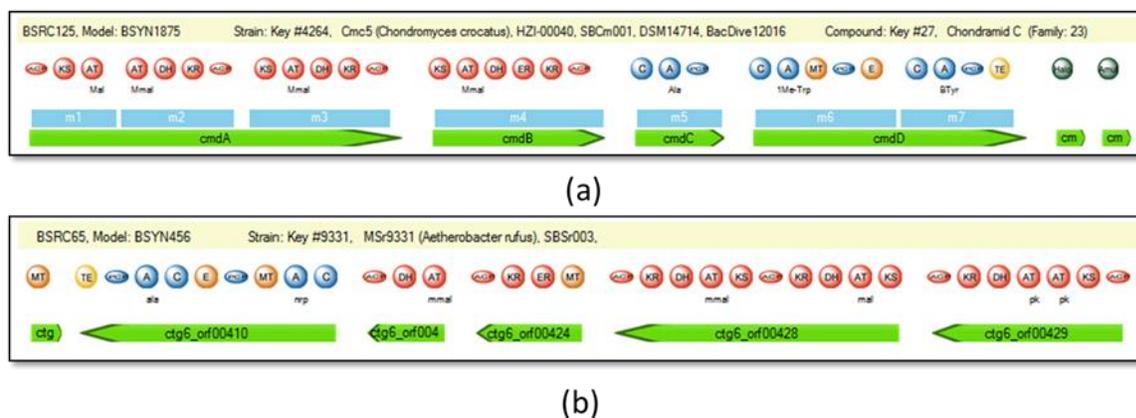


Figure 3.26: Example of the first hit (MSr9331 (*Aetherobacter rufus*)) below the confidence interval from where the similarity deteriorates

The target gene (*ctg6_orf00410*) has NRPS domain modules of same functionality but different substrate specificity as that of the query gene (*cmdD*), and overall number of NRPS modules is 2 versus 3 in chondramide biosynthesis. In addition, there are differences regarding domain composition and operon organization compared to that of the query, such as additional MT domains (internal as in

ctg6_orf00410 and single-standing) Although still rudimentary similar to the cmd pathway, it appears justified that in this search the *A. rufus* candidate pathway is significantly down-ranked compared to the original pathway from *C. crocatus* and the model detected from *C. catenulatus*.

By choosing different algorithms and varying setting of parameters the result set may change depending on the complexity of the matching, as exemplained earlier in section 3.2.1. Thus, in the following sections the limitations of the algorithms developed for search and match approach and their response to varying parameters will be exemplarily demonstrated.

3.2.2.2 Comparison of BiosynML methods

The architectural comparison of biosynthetic pathways is a crucial part for several steps of the workflows established here as part of the biosynthetic pathway mining tools. The characteristics of functions tried out in the course of this project are described and discussed in the following.

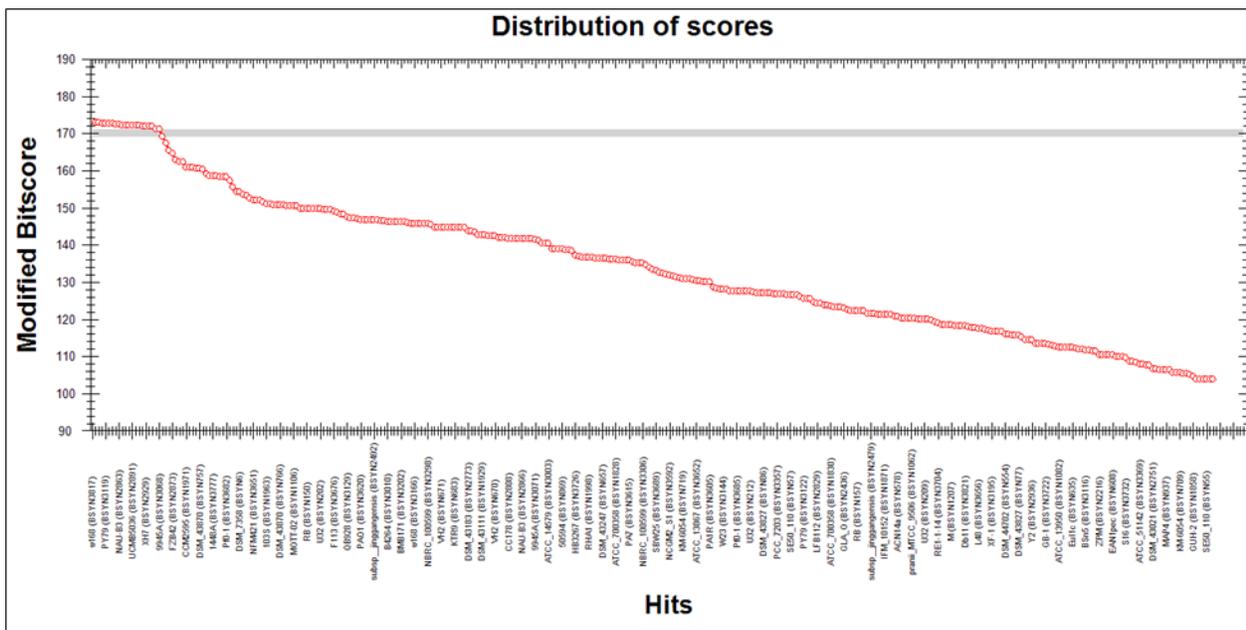


Figure 3.27: Bitscore plot generated for Surfactin pathway from *Bacillus subtilis* strain w168by using LBG algorithm.

As an example, applying all the three basic algorithms for comparing the Surfactin pathway against predicted models from biosynthetic pathways of published genomes (downloaded from NCBI and biosynthetic pathways predicted through antiSMASH), the results of LBG (Figure 3.27), LBC (Figure 3.28) and GBS (Figure 3.29) are plotted taking bitscore on y-axis and strain identifiers on x-axis. The query is executed with substrate weighting value of 0.3 and additional domain penalty of 0.3.

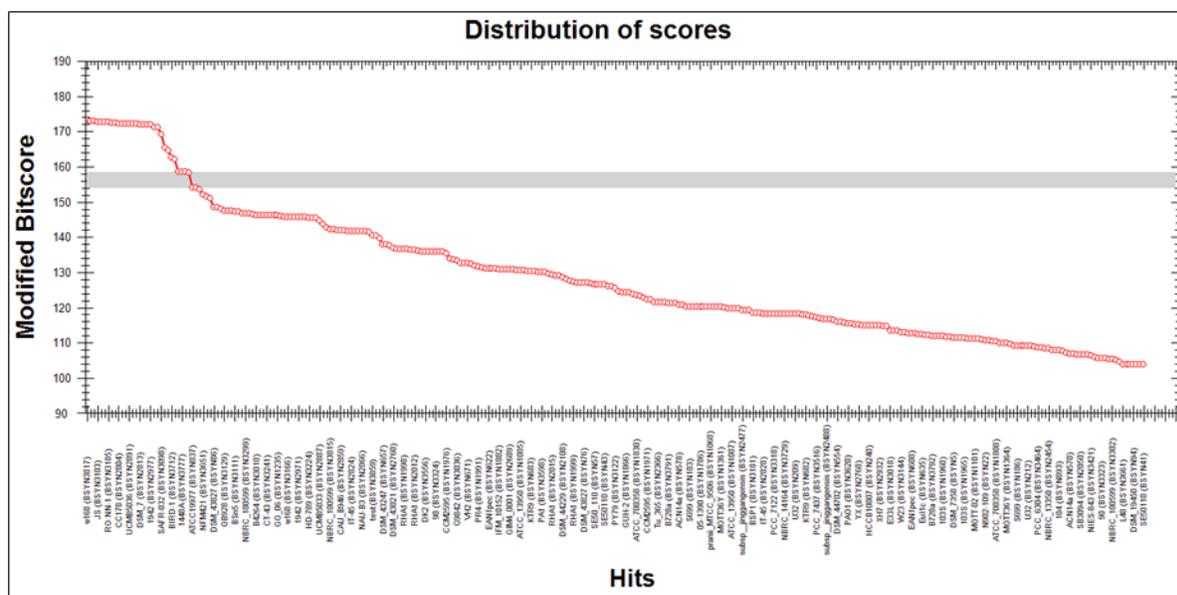


Figure 3.28: Bitscore plot generated for Surfactin pathway from *Bacillus subtilis* strain w168by using LBC algorithm

Since the scoring methods are applied on the datasets, it will also be more interesting to see the efficiency of scoring function for pathways which has overall similar domain composition but doesn't have the same domain architecture. A careful consideration of the hits in the output reveals differences in the results generated by the three algorithms.

The initial hits (Table 3.1) in the result set, identified by all three algorithms shows similar set of pathways from organisms since these pathways exhibit a highly similar domain architecture to that of the query (Figure 3.30). However, the interesting aspect is to observe the fate of the pathways which has similar domain composition and dissimilar domain architecture. The LBG and LBC failed to identify biosynthetic pathways which deviate from the query domain architecture (in the sense of placing these hits above the confidence interval) and as the algorithms assign lower scores to such pathways, they are misplaced in the output result set, which might be considered as unspecific result by a chemist. These

two algorithms (LBG and LBC) thus generate results with different levels of false negative rates, reflecting the lost opportunities of numerous interesting results that a natural product researcher would like to know or expect

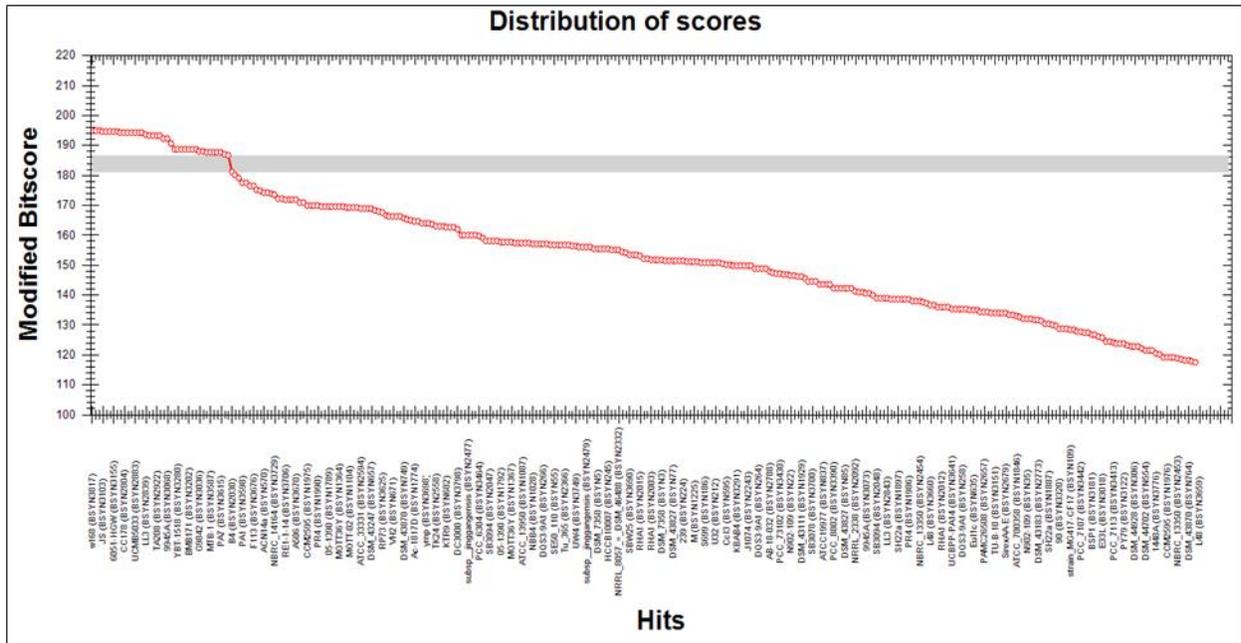


Figure 3.29: Bitscore plot generated for Surfactin pathway from *Bacillus subtilis* strain w168 by using GBS algorithm

However, GBS tends to reduce the false negative results by matching the domains composition at a global level. Due to the concept of subset intersection, GBS has identified the pathways (Table 3.2) which does not follow exactly the query's domain architecture (Figure 3.31) but having similar composition and are scored accordingly. Thus, GBS generates a result set which is highly relevant and is devoid of false negative results. The bitscore plot for LBG and LBC may be deceiving as some pathways are misplaced because of the limitations in both the algorithms, but the plot generated by GBS shows a set of results that is more accurate from the point of view of the natural products chemist. In fact, the score plot hints at the presence of two groups of similar pathways (two plateaus in Figure 3.29), one representing the Surfactin pathway family with relatively little deviations, and the other group being "Surfactin-like", showing overall similar domain composition but some degree of deviation in architecture and substrate specificity. These two groups are well discriminated against the backdrop of many NRPS pathways in the test dataset using the GBS algorithm, while the second group was not revealed by the LBG and LBC algorithms.

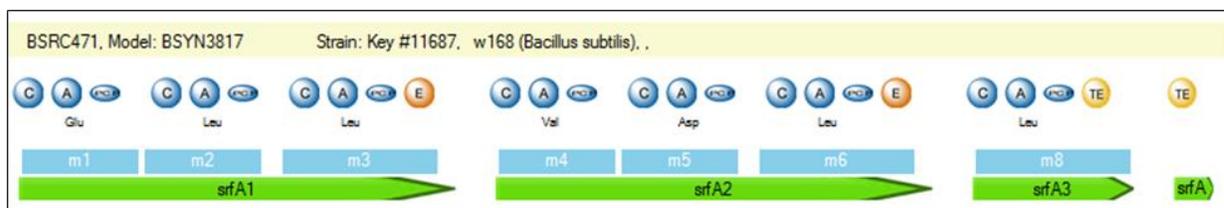
It has to be considered that the values of the parameters used for the processing the results affect the shape and confidence levels of the bitscore plot generated as these parameters influence on the scoring function used to measure the similarity between pathways.

BSRC ID	BSYN ID	Strain Key	Organism
BSRC471	BSYN3817	11687	w168 (Bacillus subtilis)
BSRC329	BSYN3126	11682	QB928 (Bacillus subtilis)
BSRC334	BSYN3163	11687	w168 (Bacillus subtilis)
BSRC326	BSYN3103	11679	JS (Bacillus sp)
BSRC328	BSYN3119	11681	PY79 (Bacillus subtilis)
BSRC331	BSYN3141	11684	W23 (Bacillus subtilis)
BSRC333	BSYN3155	11686	6051-HGW (Bacillus subtilis)
BSRC337	BSYN3185	11690	RO-NN-1 (Bacillus subtilis)
BSRC282	BSYN2855	11635	CAU_B946 (Bacillus amyloliquefaciens)
BSRC276	BSYN2804	11629	CC178 (Bacillus amyloliquefaciens)
BSRC281	BSYN2846	11634	AS43_3 (Bacillus amyloliquefaciens)
BSRC284	BSYN2873	11637	FZB42 (Bacillus amyloliquefaciens)
BSRC285	BSYN2883	11638	UCMB5033 (Bacillus amyloliquefaciens)
BSRC286	BSYN2891	11639	UCMB5036 (Bacillus amyloliquefaciens)
BSRC287	BSYN2900	11640	UCMB5113 (Bacillus amyloliquefaciens)
BSRC280	BSYN2839	11633	LL3 (Bacillus amyloliquefaciens)
BSRC277	BSYN2813	11630	DSM_7 (Bacillus amyloliquefaciens)
BSRC290	BSYN2929	11643	XH7 (Bacillus amyloliquefaciens)
BSRC289	BSYN2922	11642	TA208 (Bacillus amyloliquefaciens)
BSRC297	BSYN2977	11650	1942 (Bacillus atrophaeus)
BSRC319	BSYN3077	11672	DSM13 (Bacillus licheniformis)
BSRC318	BSYN3068	11671	9945A (Bacillus licheniformis)
BSRC324	BSYN3098	11677	SAFR-032 (Bacillus pumilus)

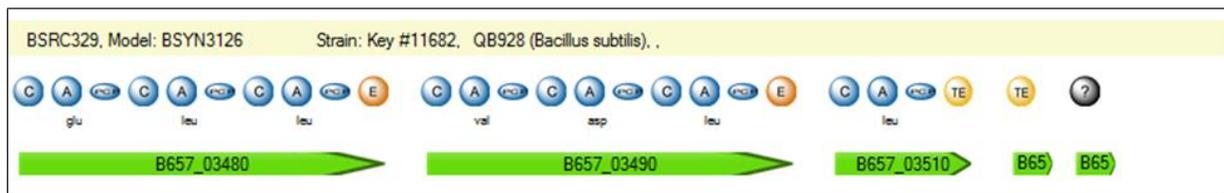
Table 3.1: Hits identified by algorithms (LBG, LBC and GBS) with Surfactin as a query

BSRC ID	BSYN ID	Strain Key	Organism
BSRC304	BSYN3010	11657	B4264 (<i>Bacillus cereus</i>)
BSRC350	BSYN3280	11703	YBT-1518 (<i>Bacillus thuringiensis</i>)
BSRC348	BSYN3269	11701	IS5056 (<i>Bacillus thuringiensis</i>)
BSRC344	BSYN3241	11697	CT-43 (<i>Bacillus thuringiensis</i>)
BSRC339	BSYN3202	11692	BMB171 (<i>Bacillus thuringiensis</i>)
BSRC340	BSYN3209	11693	Bt407 (<i>Bacillus thuringiensis</i>)
BSRC347	BSYN3260	11700	HD73 (<i>Bacillus thuringiensis</i>)
BSRC309	BSYN3036	11662	G9842 (<i>Bacillus cereus</i>)
BSRC342	BSYN3224	11695	HD-789 (<i>Bacillus thuringiensis</i>)
BSRC423	BSYN3592	11776	NCGM2_S1 (<i>Pseudomonas aeruginosa</i>)
BSRC422	BSYN3587	11775	MTB-1 (<i>Pseudomonas aeruginosa</i>)
BSRC421	BSYN3580	11774	M18 (<i>Pseudomonas aeruginosa</i>)
BSRC418	BSYN3556	11771	DK2 (<i>Pseudomonas aeruginosa</i>)
BSRC426	BSYN3615	11779	PA7 (<i>Pseudomonas aeruginosa</i>)
BSRC231	BSYN1999	11584	RHA1 (<i>Rhodococcus jostii</i>)
BSRC431	BSYN3651	11784	NFM421 (<i>Pseudomonas brassicacearum</i>)
BSRC232	BSYN2030	11585	B4 (<i>Rhodococcus opacus</i>)
BSRC228	BSYN1960	11581	103S (<i>Rhodococcus equi</i>)
BSRC230	BSYN1991	11583	PR4 (<i>Rhodococcus erythropolis</i>)
BSRC424	BSYN3598	11777	PA1 (<i>Pseudomonas aeruginosa</i>)
BSRC425	BSYN3605	11778	PA1R (<i>Pseudomonas aeruginosa</i>)
BSRC436	BSYN3682	11789	Pf0-1 (<i>Pseudomonas fluorescens</i>)
BSRC435	BSYN3676	11788	F113 (<i>Pseudomonas fluorescens</i>)
BSRC437	BSYN3689	11790	SBW25 (<i>Pseudomonas fluorescens</i>)
BSRC443	BSYN3705	11796	RE1-1-14 (<i>Pseudomonas poae</i>)
BSRC116	BSYN570	11469	ACN14a (<i>Frankia alni</i>)
BSRC116	BSYN570	11469	ACN14a (<i>Frankia alni</i>)
BSRC447	BSYN3722	11800	GB-1 (<i>Pseudomonas putida</i>)

Table 3.2: Additional results displayed by GBS for Surfactin pathways which are not available from LBG and LBC



(a)



(b)

Figure 3.30: Target pathways from strain QB928 (*Bacillus cereus*) (b) having similar domain composition, substrate specificities and domain architecture to that of the query pathway from strain w168 (*Bacillus subtilis*)(a)

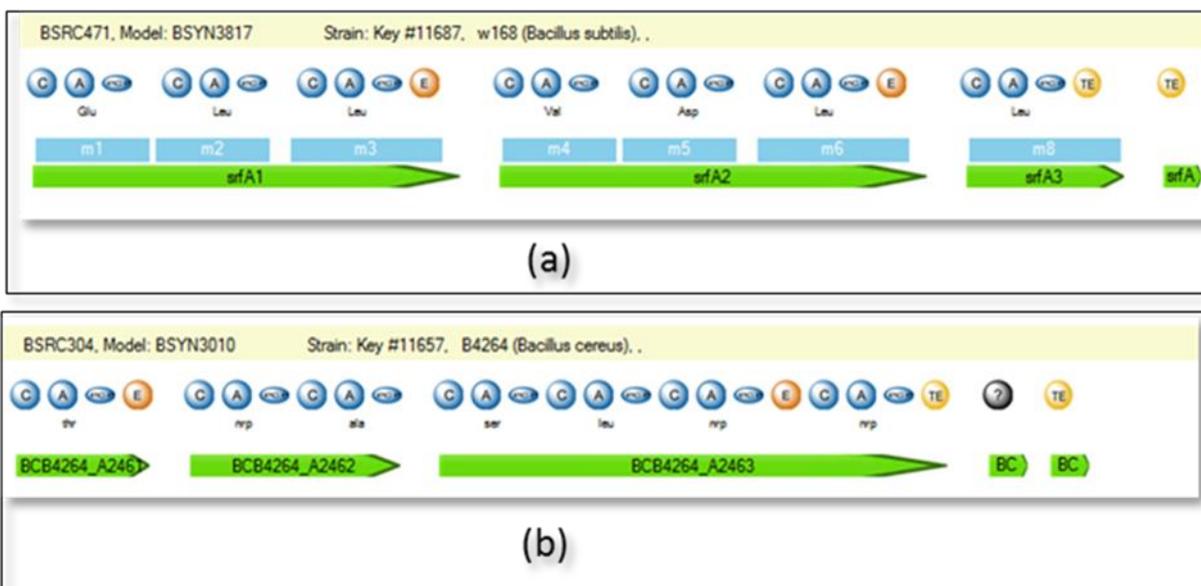


Figure 3.31: Target pathway from strain B4264 (*Bacillus subtilis*) (b) which has similar domain composition but with varied domain architecture to that of the query pathway from strain w168 (*Bacillus subtilis*) (a), and also different predicted substrate specificities.

From the experiments mentioned in this section choosing a right algorithm is important to obtain comprehensive results, where throughout this project the most relevant results were stringently obtained by using GBS. But it is also important to evaluate the influence of the parameter settings on the results generated by search and match engine.

3.2.3 Influence of parameter settings on the outcome of the pathway comparison

A matching and scoring algorithm ideally should have parameters whose effect is intuitively clear and predictable. In the course of this study, we found that parameters such as substrate specificity match, extra domains penalty, collinearity and pathway completeness are important to feed the BiosynML matching algorithms in order to downgrade the appearance of irrelevant data from the possible outcome. These parameters influence to a great extent in the accuracy and performance of the algorithm.

Influence of substrate specificity

Intuitively, the substrate specificity parameter defines the influences of missing or wrongly assigned substrate specificity on the scoring function. This is currently relevant mostly for monomer-incorporating domains such as A- and AT domains (although it should be noted that there is no inherent limitation to certain domain types). The value ranges from 0 to 1 where low values meaning ‘liberal scoring of mismatched substrates’ and high values meaning ‘strict comparison and scoring of substrates’. To demonstrate the influence of this parameter, a query was executed in the BiosynML repository using

GBS algorithm with a substrate specificity weighting of 0.8 and 0.3, all other parameters are kept constant. As a test case, Myxoprincomide pathway from strain DK1622 (*Myxococcus xanthus*) (Figure 3.32) was used to investigate the influence of different substrate specificity parameters on the outcome.

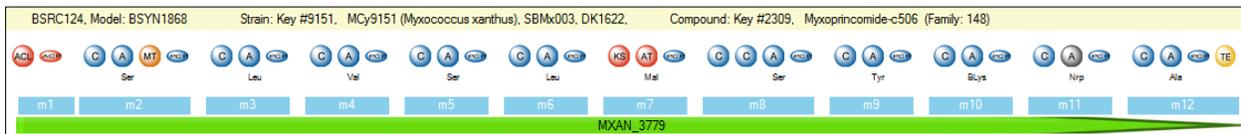


Figure 3.32: Myxoprincomide pathway from strain DK1622 (*Myxococcus xanthus*) used to evaluate the influence of substrate specificity parameters on the results set. The grey-colored A-domain in module 11 was marked as “inactive” during manual curation (1), and substrate specificities were inferred from the elucidated structure myxoprincomide-C506.

From Figure 3.33 (query using weighting factor 0.8), it can be observed that only one pathway is found which shows above the confidence interval. This pathway is itself a curated pathway where substrate specificities have been manually corrected by the researcher. All the pathways which have the same domain functionality but different substrate specificities, possibly due to the prediction tool uncertainty are found below the confidence interval. By decreasing the influence of substrate specificity to 0.3, the results changed, as it can be observed that there two more pathways moved above the confidence interval as the specificity scoring is relaxed (Figure 3.34). This means, giving a high value to

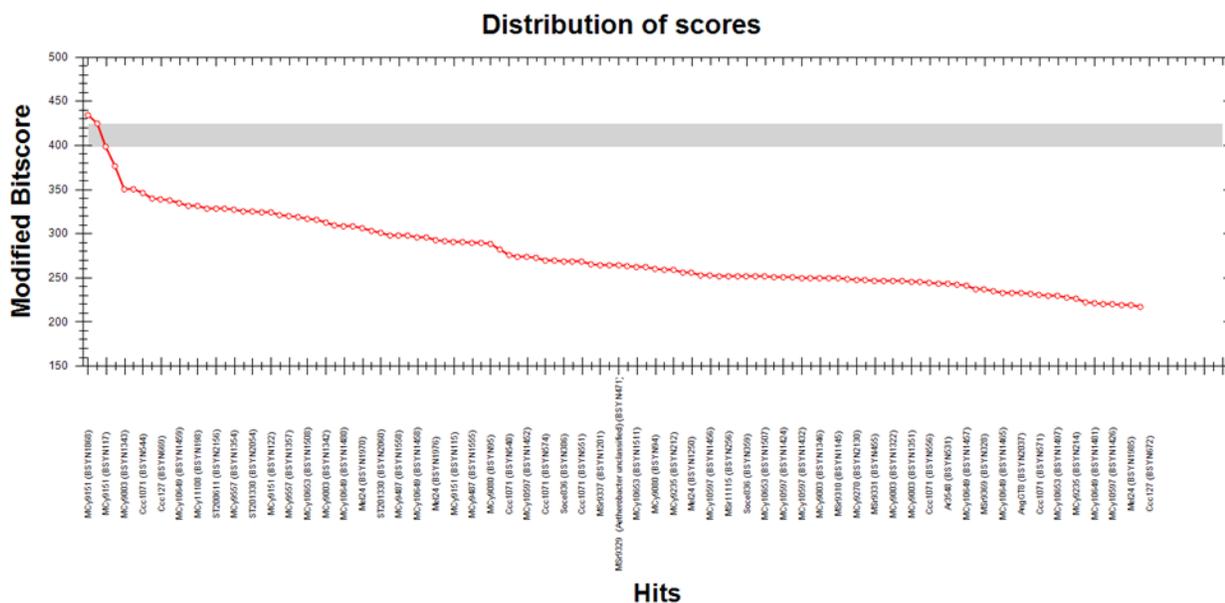


Figure 3.33: Distribution of scores for Myxoprincomide pathway using GBS algorithm with substrate specificity emphasized by a weighting of 0.8

this parameter will decrease the score of the hit pathways whose substrate specificities deviate with that of the query.

It has to be noted that there is no change in the order of the hits resulted by the algorithm (some of these shown in Figure 3.35) with the change in substrate specificity but this parameter has a great influence in deciding the range of confidence interval. It is a parameter of interest for the researcher during automated mining of the pathways from a newly annotated genome, to decide the extent of having the same postulated monomeric substrate for query and hit pathways. Note that the utility of a stringent parameter setting also depends critically on the overall state of the pathway repository: queries using a high substrate specificity weighting can make sense when searching in a library of well-curated pathways, whereas a collection of auto-annotated genomes might make it necessary to accept a higher degree of mis-assigned substrates due to imperfection of predictive tools.

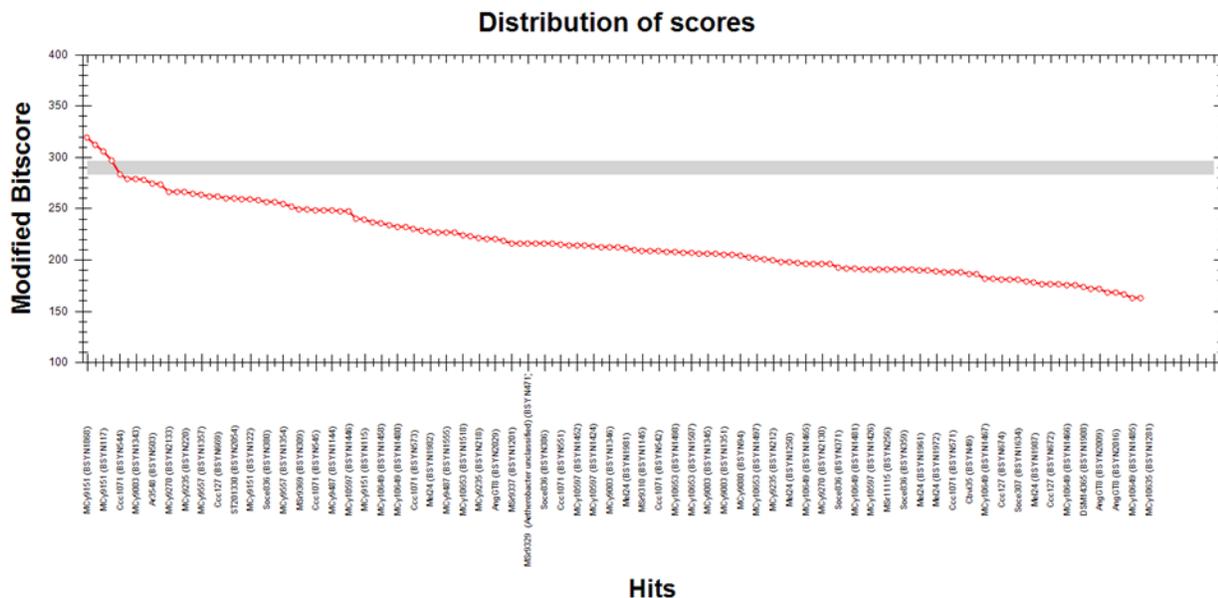


Figure 3.34: Distribution of scores for Myxoprincomide pathway using GBS algorithm with a more relaxed substrate specificity weighting of 0.3.

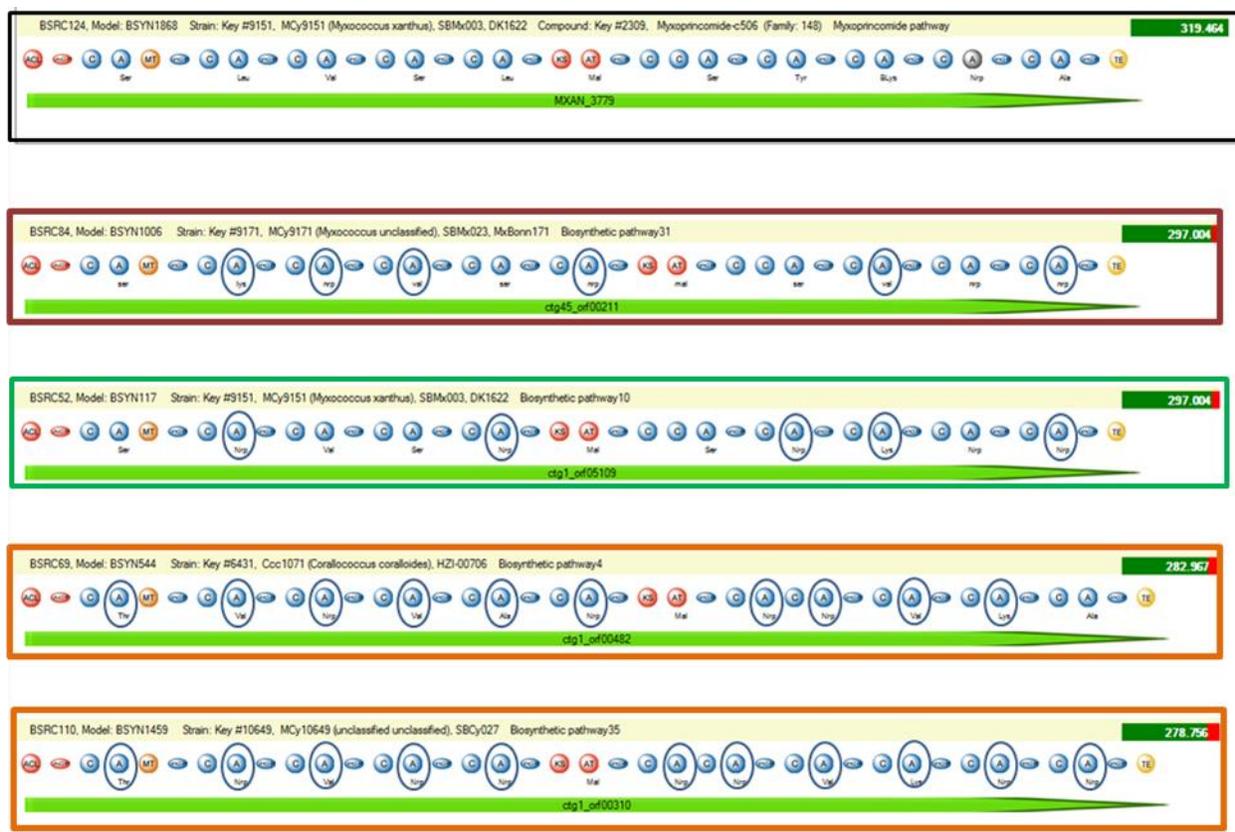


Figure 3.35: Hits reported by GBS algorithm for Myxoprincomide gene cluster from strain DK1622 (*Myxococcus xanthus*). Query is represented in black box. Blue circle represents the domains which deviate in substrate specificity. Red box represents the pathways with translocated modules and orange box represents the pathways with an extra (C-A-PCP) module. The pathway in green box has identical domain architecture with deviations in substrate specificities.

Influence of penalty imposed on additional domains

Another important factor of the comparison process between two pathways is the additional domain penalties i.e., how strongly are additional domains not present in the query pathway penalized. The value ranges from 0 to 1 where low values meaning 'weak penalty' and high values meaning 'strong penalty'. This parameter does not take into account the domains which are missing in the middle of multimodular genes but penalize domains from additional or missing genes from the query and target pathways. As an example Myxochelin from strains Sga15 (*Stigmatella aurantiaca*) was used to evaluate the influence of penalty on additional domains. The query is executed with additional domain penalty of 0.1 and 0.8, keeping the rest of the parameters constant.

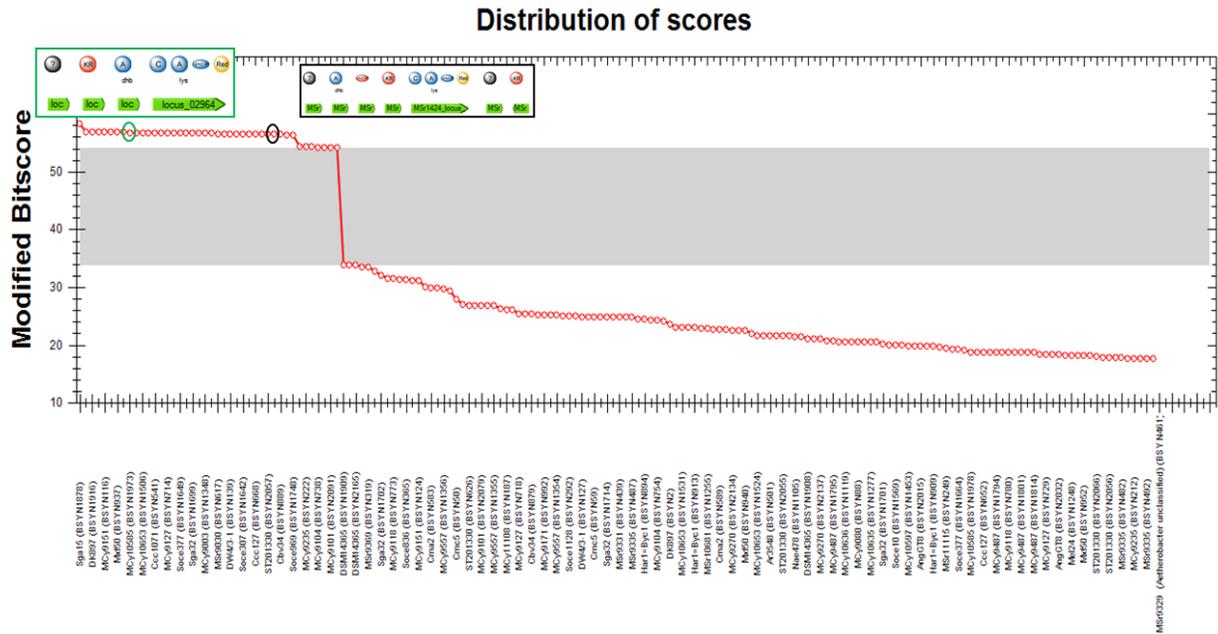


Figure 3.36: Distribution of scores for Myxochelin pathway using GBS algorithm with low additional domain penalty of 0.1. The pathway represented in green and black outline has similar domain architecture to myxochelin pathway. Though the black outlined pathway has additional domains there is no significant distinction of scores.

Form the Figure 3.36, it can be observed that there is no distinction of hit pathways which have only the “myxochelin” domain set and pathways exhibiting additional domains. This is due to the low value (0.1) used for additional domain penalty.

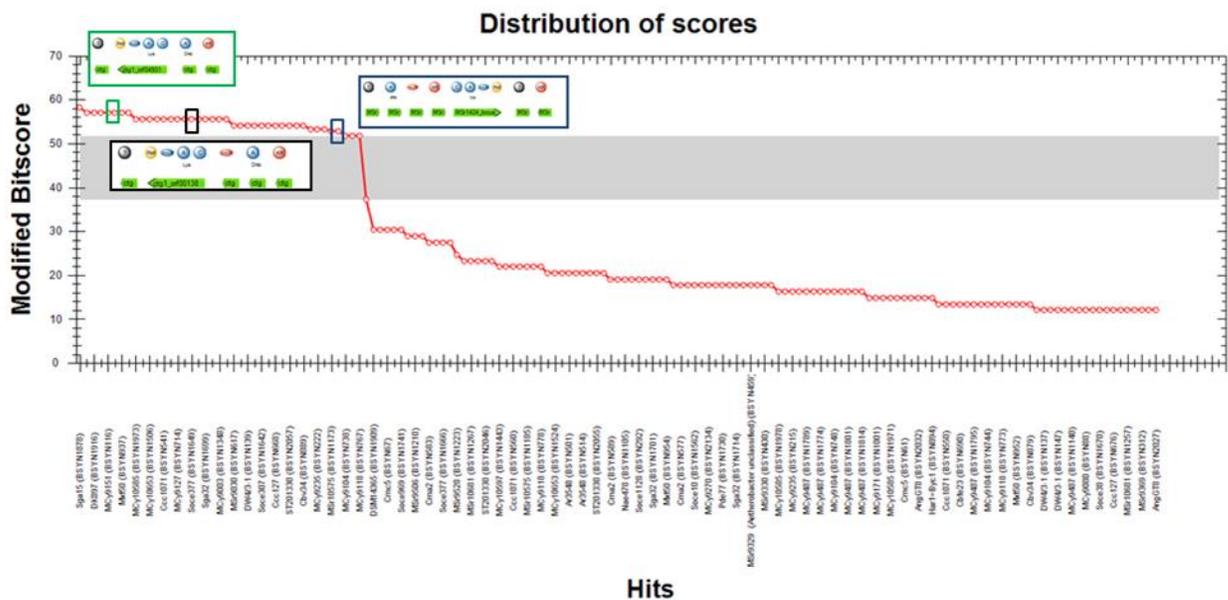


Figure 3.37: Distribution of scores for Myxochelin pathway using GBS algorithm with increased additional domain penalty of 0.8. There is now a noticeable distinction of scores for the pathways with additional domains as exemplified by the pathways outlined in green, black and blue colors.

Therefore, the result set contains unsorted list of pathways which has exact matching domains with respect to query as well as pathways with (varying numbers of) additional domains. With increase in the penalty of the additional domains to 0.8, the result set is sorted and pathways with additional domains are well distinguished (Figure 3.37). This parameter is useful to prioritize the hits showing sets of matching domains which a chemist might be interested in.

Impact of Collinearity weighting on the outcome

Collinearity weighting refers to the relationship where there is a high correlation on both composition and order of domains between two pathways. As an example, a surfactin like pathway from strain CT-43 (*Bacillus thuringiensis*) (Figure 3.40) was used to evaluate the impact of collinearity parameter. The tests were done with collinearity values for 0.1 and 0.8, all the other parameters are set to zero.

With a value of 0.8, the algorithm reported hits which are above the level of confidence interval. These hits have very high domain composition similarities as well as the order of the domains is preserved (Figure 3.38). On reducing the value of collinearity to 0.1 there is a significant change in the range of confidence interval (Figure 3.39) where the hits with well-matching domain composition but has lower similarity in the sequence order of the domains are also considered.

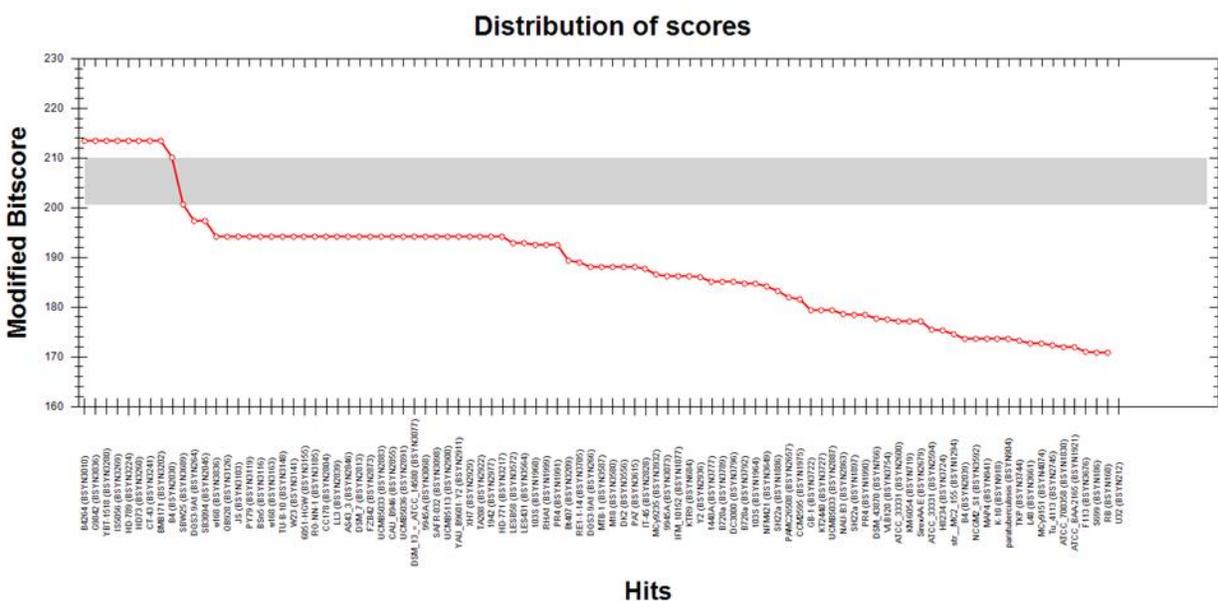


Figure 3.38: Results displayed using collinearity weighting value of 0.8. Only eight hits are reported above the confidence interval

This parameter is very useful when a search claims to have hits with highly similar domain composition and preserving also the order of domains. A manual observation also confirmed that, with a value of 0.8, the resulted top hits have similar domain arrangements (Figure 3.40). With a decrease in the collinearity value to 0.1, there is a change in the range of confidence interval, thus the pathway with high similar domain composition but with dissimilar arrangement (Figure 3.41) were also seen above the confidence interval.

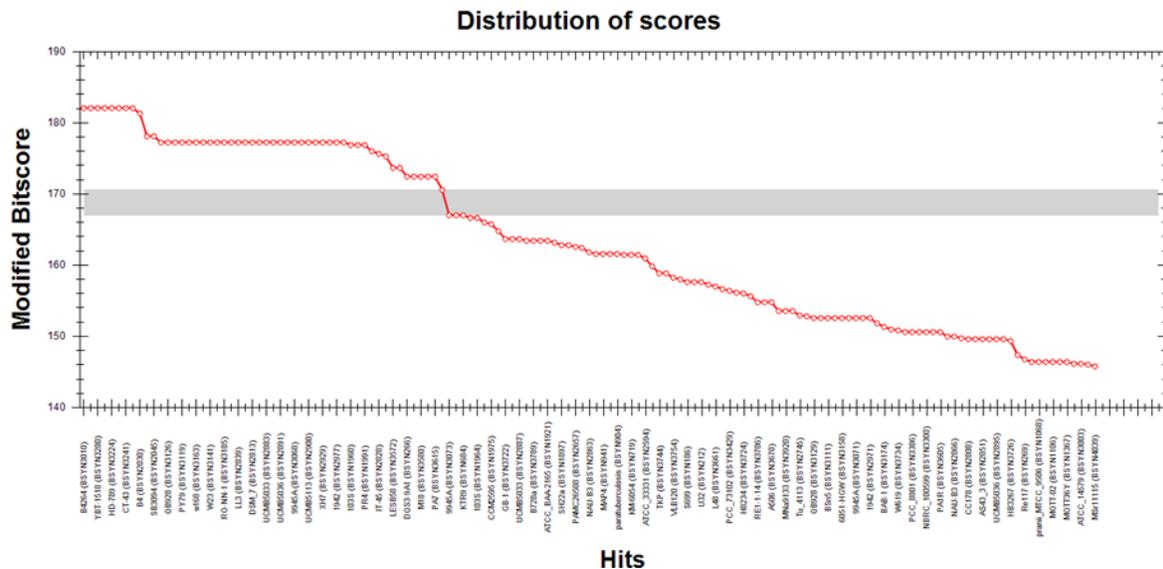


Figure 3.39: Results displayed using low collinearity weighting value of 0.1. Several additional hits are reported above the confidence interval

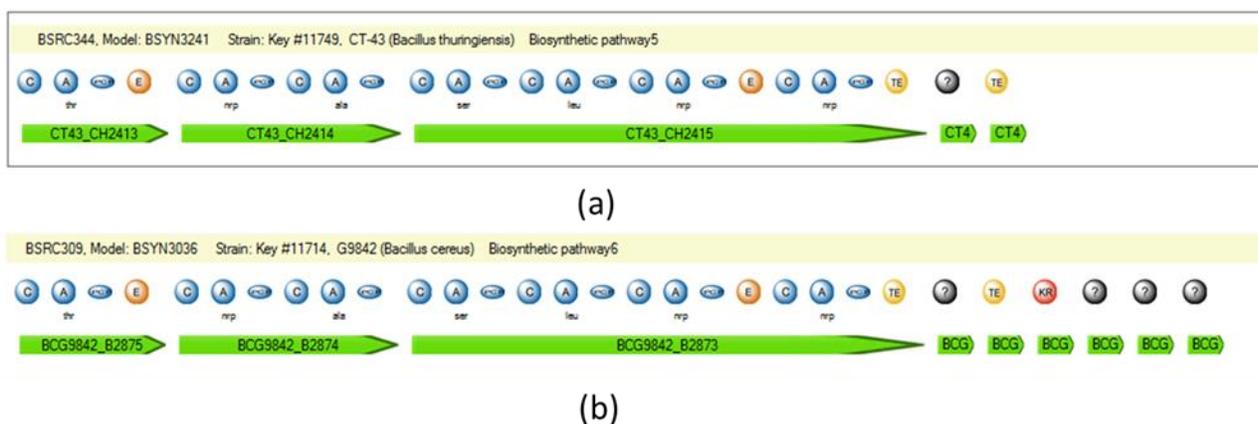


Figure 3.40: High scoring hit pathways with similar domain arrangement. (a) Query pathway from strain CT-43 (*Bacillus thuringiensis*) (b) hit pathway from G9842 (*Bacillus cereus*).

There are domains which are missing in the hit pathways compared to the query pathway (blue box in Figure 3.43), due to which the hits are below the range of confidence interval upon activating the Pathway completeness parameter (Figure 3.42). Deactivating the parameter modified the confidence interval range, showing the hits which have very similar domain composition with that of the query (Figure 3.44) although preceding domain detection was apparently incomplete for some of the auto-annotated pathways.

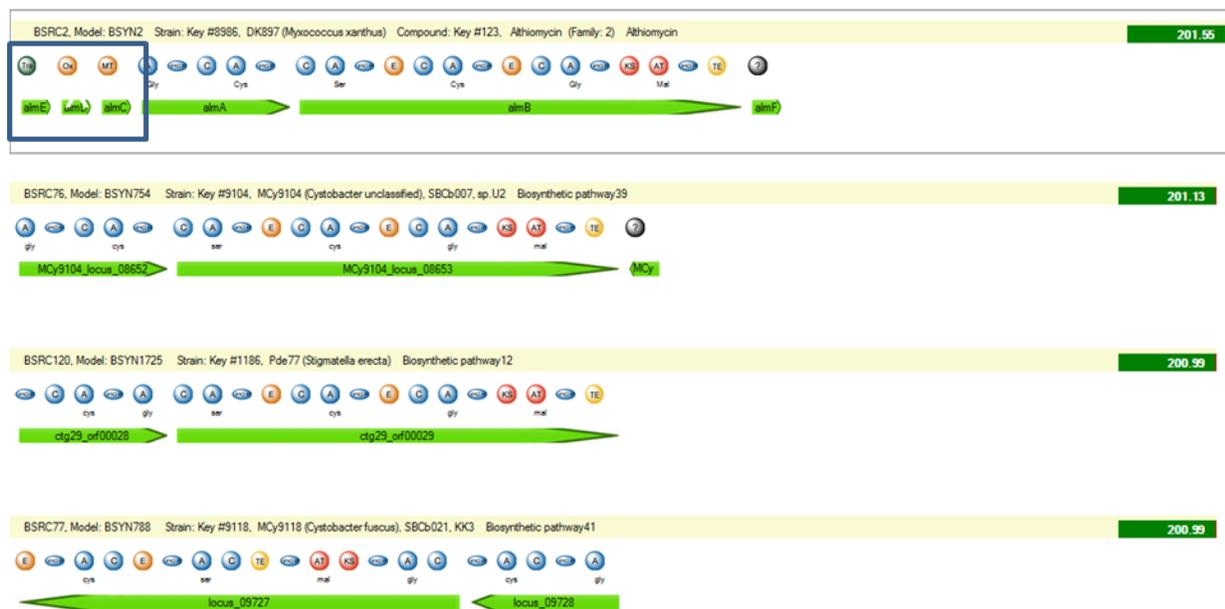


Figure 3.43: Ball scheme representation of query pathway Althiomycin (pathway shown with thin border) from strain DK897 (*Myxococcus xanthus*) and hits pathways from various producers. The blue box represents the domains that are missing in the hits

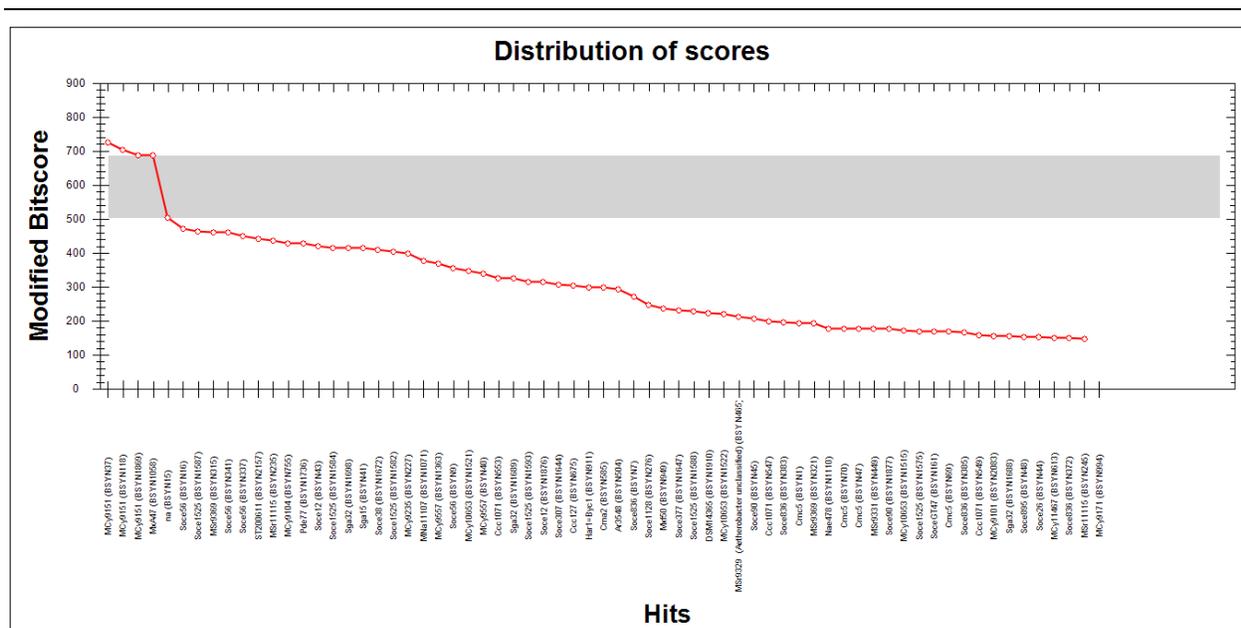


Figure 3.44: Distribution of scores for Althiomycin with Pathway completeness parameter deactivated

3.2.4 Pathway query based on signature domains

An additional mode of using the BiosynML database system should be briefly mentioned here, although it is not a strictly “concepts-based” approach for genome-mining.

The BiosynML record representing a gene cluster contains meta data of domains that are involved in the biosynthesis of a product. Meta data contains description of properties such as building blocks, functions, status, position of the domains in the genes and many more that are linked to domains. Thus, availability of these data enables researchers to simply search and identify the pathways that possess specific information across the database. To search by metadata, researcher needs to specify information about a domain or a group of domains. The respective research scenario would be to retrieve pathways (or connected information, e.g. domain sequences) stored in the database based on the query provided by the user. An example would be of query of the database to find pathways from any strain which use a tyrosine-derived building block: It retrieves all the pathways with tyrosine-derived building block and outputs the detailed information of the pathway. Similarly, one might want to extract the protein sequences of all TE domains from hybrid NRPS/PKS pathways: it extracts the protein sequence from all domains of annotated type “TE” from all the pathways (Figure 3.45).

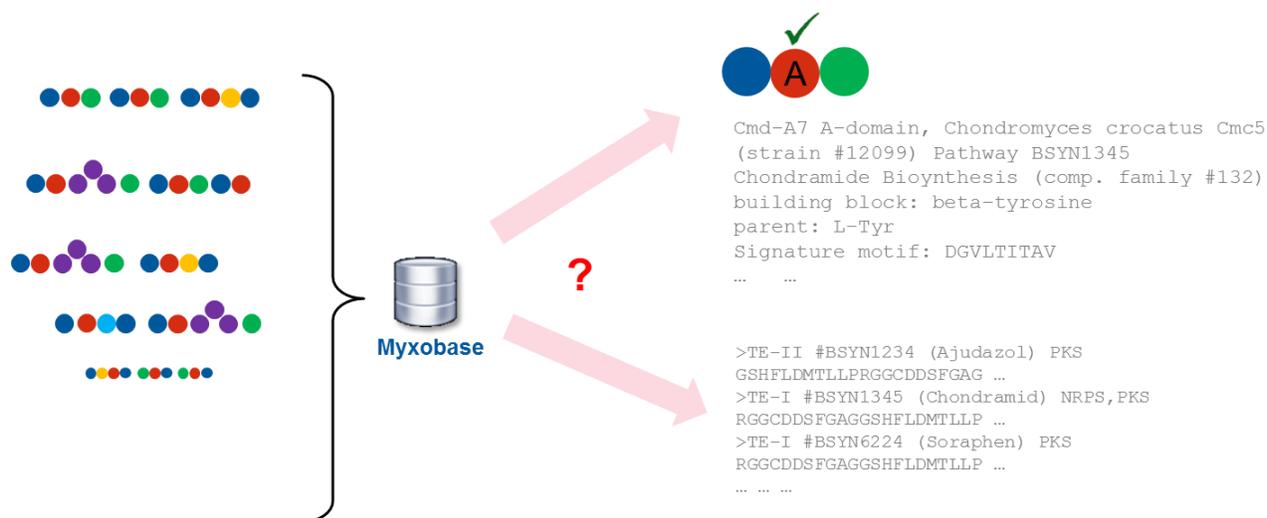


Figure 3.45: Querying database for pathways with specific information such as domains and its meta information

A query can contain multiple domains and their properties linked by connector to control the output of the pathways from the biosynthetic pathway repository. Each operand in the connectors is considered as a condition that contains either true or false value. This determines the overall results of the query that is provided by the user.

Connectors

Connectors are words used to connect two or more domains and their metadata in a valid way to retrieve pathways of researcher's interest from the database.

AND (logical and): the query returns all the pathways which contain all of the specified information provided by the user.

OR (logical or): the query returns all the pathways which contains contain at least one of the specified information provided by the user.

NOT (logical not): Finds content items that contain the information that precedes the operator (if any), and ignores content items that contain the information that follows it.

An interface has been implemented for the researchers to retrieve pathways from the database based on a given query. This interface provides multiple options for the researcher to search pathways containing a specific set of domains and their metadata information based on the keywords provided by the user (Figure 3.46).

For this type of query, the search engine routinely performs a basic SQL search based on domains types and their properties that are considered appropriate. In the second step, the interface (Figure 3.46) starts with initiating a search against the database where the function performs the domain

matching tasks and outputs the results. The results consist of the information about hits which are estimated to have domains and their properties to that of the query given. As mentioned earlier in the framework section, the parameters for the search are *domain type, properties that the domains possess (building block, parent of the building block, subtype and status) and conditions which fits according to the researcher needs (AND, OR and NOT)*. The output is delivered by the result window depends on the pathways that are considered as hits, and generally comprises the BSRC key, BSYN key, Strain identifier key, compound identifier as well as compound name and strain name.

A part from the simple display of the results, a graphical output of the result pathways is generated, highlighting the hit domains in the pathways. In addition, to the search and visualization function of the interface also allows researcher to extract the protein sequences of the domains such as AT, A or KS which would be further helpful to perform sequence related analysis of the pathways.

The screenshot displays a web-based interface for querying pathways. It is divided into three main sections:

- Table (Blue box):** A table with columns: BSRC key, BSYN key, MXID, Strain, Compound key, and Compound. The first row is highlighted in blue.

BSRC key	BSYN key	MXID	Strain	Compound key	Compound
BSRC1	BSYN14	6431	Coc1071		
BSRC3	BSYN150	11108	MCy11108		
BSRC10	BSYN158	9557	MCy9557		
BSRC19	BSYN159	9557	MCy9557		
BSRC1	BSYN2	6431	Coc1071		
BSRC1	BSYN21	6431	Coc1071		
BSRC17	BSYN323	7455	Ang518		
BSRC18	BSYN332	8384	AJ3548		
BSRC2	BSYN37	4264	Cnc5		
BSRC1	BSYN4	6431	Coc1071		
BSRC21	BSYN405	8986	DK397		
BSRC24	BSYN467	10984	MCy10984		
BSRC3	BSYN52	9151	MCy9151	2309	Myxopiticos...
BSRC3	BSYN58	9151	MCy9151		
BSRC5	BSYN71	9235	MCy9235		
BSRC62	BSYN809	11324	CHN365		
BSRC5	BSYN83	9235	MCy9235		
BSRC69	BSYN835	9101	MCy9101	40	Argem A
- Query Builder (Red box):** A form with fields for Condition, Domain, Property, and Keywords. The condition is set to "and", Domain to "A", Property to "Building block", and Keywords to "ser". There are "Filter" and "Reset" buttons at the bottom.
- Extract protein sequences (Green box):** A section with a radio button selected for "Protein sequence of:" and a dropdown menu set to "A". An "Extract" button is located below.

Below the query builder, the following information is displayed:

- Time take for query: 00:00:00.1318971
- # of rows filtered: 18

Figure 3.46: Interface for querying pathways with specific domain set. Red box shows the query builder, blue box indicates the basic result set generated by the query and green box shows the options for extracting protein sequences for the selected domains from the results set.

As a simple test case, a “domain query” was conducted to reveal candidate pathways which are indicative for presence of two activating domains “A” whose building blocks information contains “Serine” and “Alanine”. The domain “A” is selected from the dropdown, the building block is selected from the property list, the information of the building block is given in the input filed keyword and an appropriate condition has to be chosen which in our test case is “AND”. The aim of this query is to verify the functions using domains and their properties, aiming to retrieve all candidate pathways exhibiting similar information in the database. After the execution of the query, the results can be seen in

datagridview and the graphical output can be generated in the graphical output tab as shown in Figure 3.47.

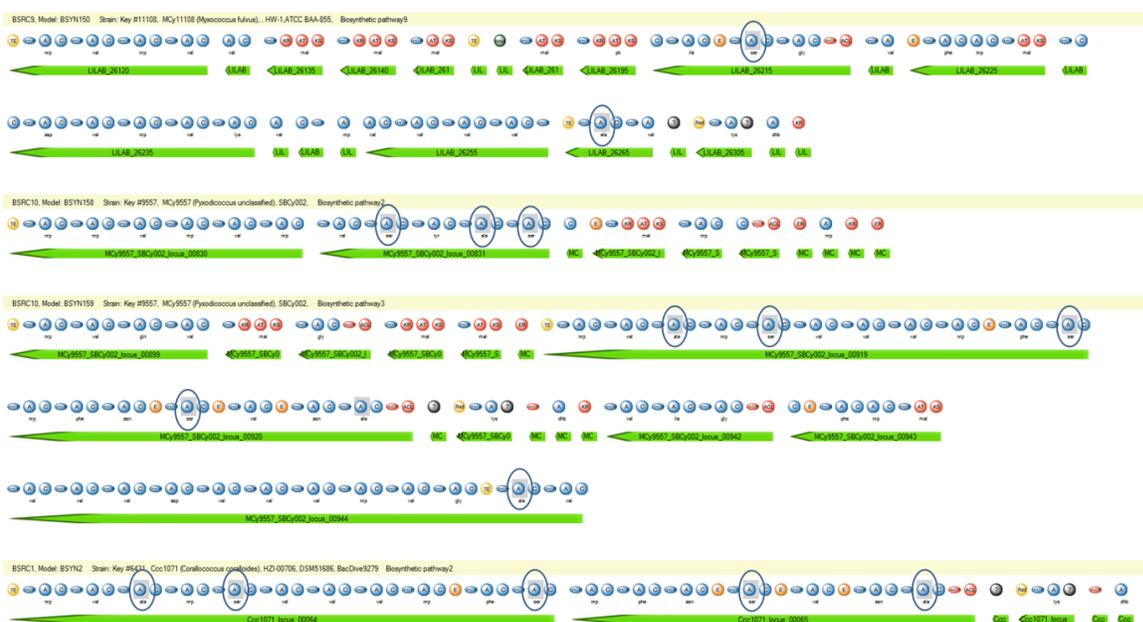


Figure 3.47: Graphical interface displaying ball scheme representation of pathways with highlighted hit domains in blue circle

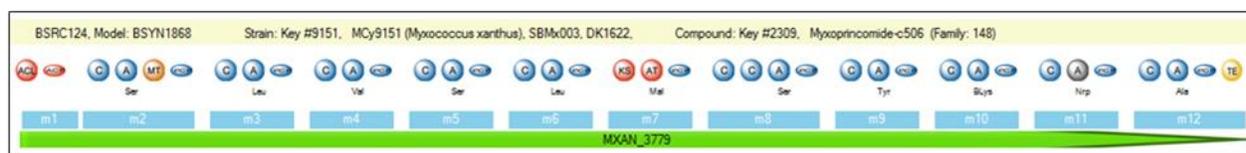
From the Figure 3.47, it can be seen that the best possible hits among all the pathways available in the database. The output in the datagridview shows minimal information such as pathway key, strain information and compound information (if the information is available for the hits). Along with this basic information the graphical interface provides the ball scheme representation of the hit pathways where the hit domains are highlighted (Figure 3.47). In addition, the interface also shows the details associated with the pathways such as internal Ids, strain details and compound details. This search provides useful information for the researchers to find similar pathway with signature domains and their properties of interest.

In addition to the resulting pathway information, researchers can extract protein or DNA sequences of the domains from the pathways set for further use. The interface also has a function which can generate a graphical output of ball scheme representation of the resulted pathways which can explicitly show clusters in terms of individually identified domains and their properties highlighted.

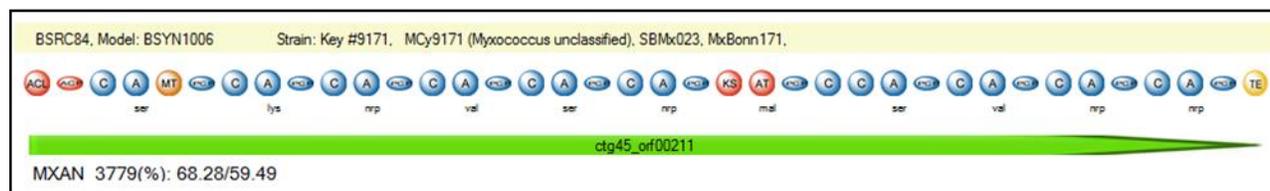
Addition of sequence-based analysis functions

Traditionally sequence analysis tools play a major role in identifying regions of similarity in DNA, RNA and protein through which functional, structural, or evolutionary relationships between the sequences are constructed. For several reasons it may be desirable to conduct sequence alignments between query and hit pathways following their retrieval from the pathway repository as the result of a conceptual query. One motivation could be to increase the confidence of the high affinity hits resulted from conceptual matching. In order to amend the BiosynML analysis engine with the capacity to perform analysis also at sequence level, MUSCLE, a program for multiple protein sequence alignment was fully integrated into the BiosynML framework. The researcher can chose to submit the desired hit to the sequence analysis tool through BiosynML interface in Mxbase Explorer to find the similarity of pathways at sequence level. As an example the gene clusters of Myxoprincomide from strain DK1622 (*Myxococcus xanthus*) and Althiomycin from strain DK897 (*Myxococcus xacnthus*) were used to evaluate the similarity of the top hits from conceptual matching at the protein sequence level (Figure 3.48 and Figure 3.49).

From Figure 3.48 and Figure 3.49, it is clear that the Myxoprincomide pathway has hits from the strains MxBonn171 (*Myxococcus unclassified*) and Ccc1071 (*Corallocooccus coralloides*). These hits have convincing architectural similarity. However, upon investigating the hit pathways at sequence level the hit from strain MxBonn171 (*Myxococcus unclassified*) showed overall only 68.28% similarity and the hit from Ccc1071 (*Corallocooccus coralloides*) showed 66.36% similarity. The obvious reason is the deviating arrangement of NRPS modules up- and downstream of the small PKS part in the center, possibly as a common result of pathway evolution. This intuitive example essentially underlines the fact that sequence alignments (including methods like blastP on that matter) are not an adequate means for genome mining (especially unsupervised) with long multimodular pathways.



(a)



(b)

Figure 3.48: Sequence similarity (Similarity/Identity) of (b) hit from Strain MxBonn171 (*Myxococcus unclassified*) and (a) Myxoprincomide pathway from DK1622 (*Myxococcus xanthus*) used as a query.

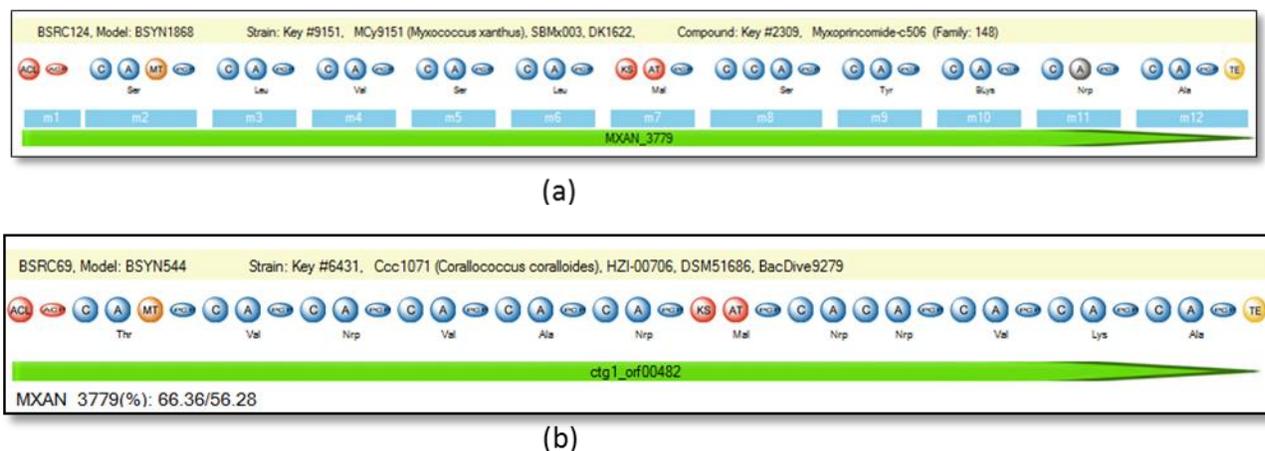


Figure 3.49: Sequence similarity (Similarity/Identity) of (b) hit from Strain Ccc1071 (*Coralloccoccus coralloides*) and (a) Myxoprincomide pathway from DK1622 (*Myxococcus xanthus*) used as a query.

To introduce a different example: Using Althiomycin gene cluster as the query pathway against a database filled with public genomes, the GBS algorithm resulted in a single hit from strain Db11 (*Serratia marcescens* (129)). Investigating the protein sequence similarity, the query gene almA has 60.62% similarity with target gene SMDB11_RS11325, query gene almB has 58.57% similarity with target gene SMDB11_RS11320 and query gene almC has 54% similarity with target gene SMDB11_RS11330 (single standing MT domain) (Figure 3.51). Simultaneously on the contrary, comparing protein sequence of Althiomycin from DK897 (*Myxococcus xanthus*) with a top hit which is from myxobacterial strain Pde77 (*Stigmatella erecta*) showed, the target gene ctg29_orf00028 has a similarity of 81.48% with query gene almA and target gene ctg29_orf00029 has a similarity of 85.38% with query gene almB (Figure 3.50). Thus, the althiomycin example shows that a conclusion about the suitable sequence similarity threshold - especially when thinking of any automated and largely unsupervised analysis workflow covering a wide taxonomic range - is not easily made.

This results suggest that the current sequence based tool that are widely used to identify known biosynthetic pathways using protein sequence similarity, have anticipated difficulties in matching protein sequences of genes with translocated modules as in the myxoprincomide example (Figure 3.48 b and Figure 3.49b). Such pathways results in lower sequence similarity score because of the gaps generated by translocated module (Figure 3.52), although the hit pathways have similar domain compositions and plausibly similar architecture. In case of larger genes, partial sequences are used to identify the pathways which produce similar compounds; hence success critically depends on first identifying the correct part of the gene cluster that is to be used to find the matching pathways. Moreover, a definite confidence level of similarity which could determine that the query and target gene clusters encode pathways for

structurally similar products is not easily defined. Similar domain architecture on the contrary are likely to produce similar compounds even if sequence similarity is weak, commonly observed as a result of taxonomic distance. This could lead to unspecific results where the researcher has to invest time in manually filtering the result set to identify the appropriate result. This limitation highlights the necessity of having an automated approach which can efficiently match the biosynthetic pathways based on architectural similarity (with or without presence of the sequence information) and outputs the results in a way which is understandable to the researcher.

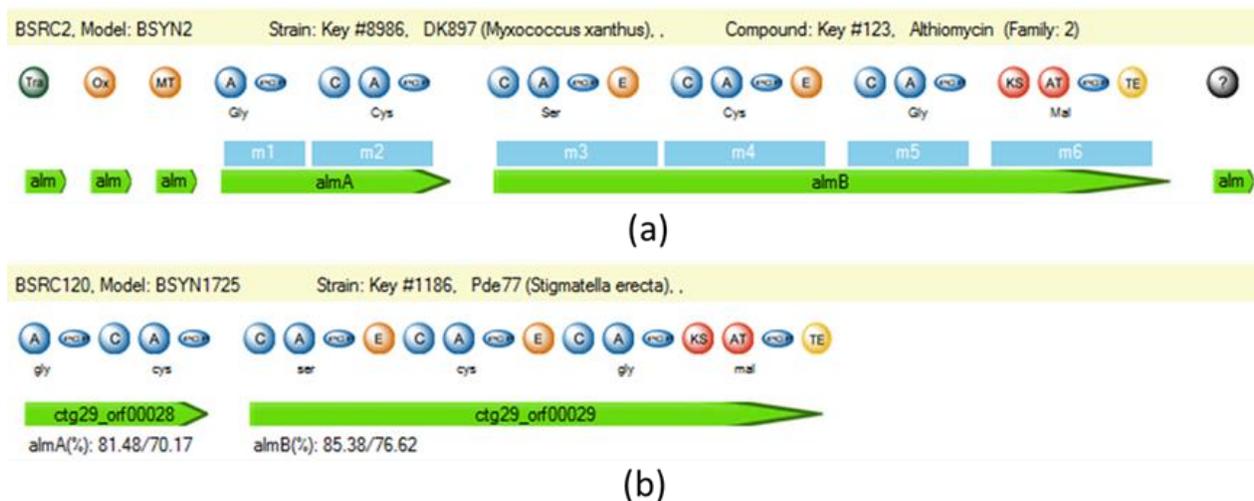


Figure 3.50: High sequence similarity of Althiomycin pathway between (a) query strain and DK897 (*Myxococcus xanthus*) and (b) Pde77 (*Stigmatella erecta*)

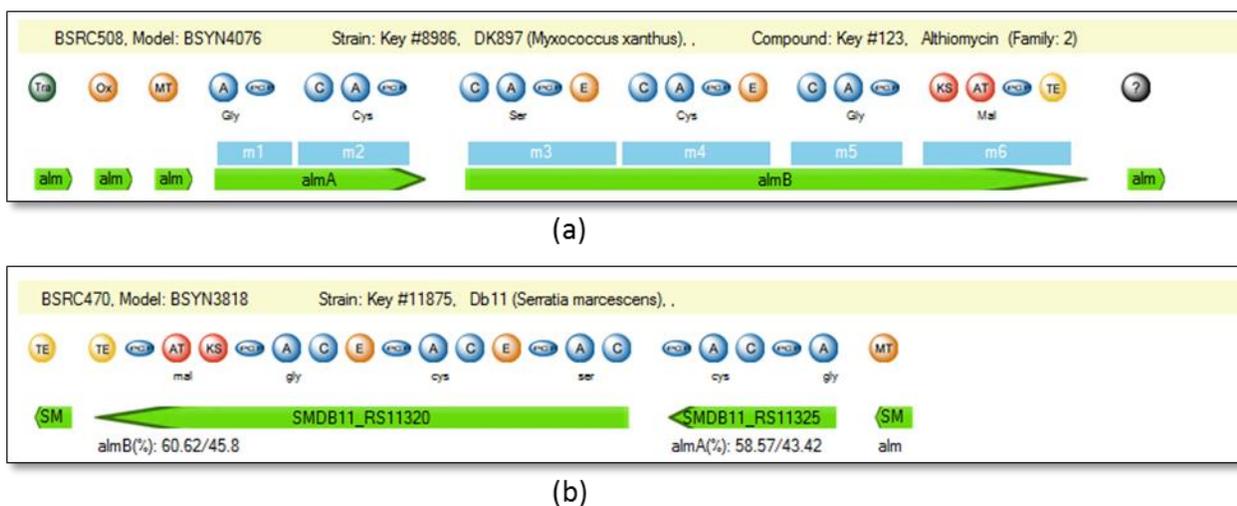


Figure 3.51: Lower sequence similarity of Althiomycin pathway between (a) query strain and DK897 (*Myxococcus xanthus*) and (b) Db11 (*Serratia marcescens*)

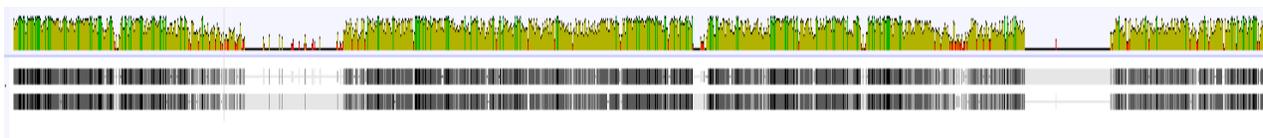


Figure 3.52: Result of protein alignment of the single-gene Myxoprincomide pathway from DK1622 (*Myxococcus xanthus*) and hit pathway from MxBonn171 (*Myxococcus unclassified*) with an overall similarity score of 68.2%

3.3 Conceptual genome mining with natural products sources

In the previous sections, the establishment of the framework and the concepts of the BioynML analysis engine were discussed along with the preliminary results. The following part mainly focuses on applying the new tools on real world data. The key considerations are 1) to conduct targeted queries across a BiosynML repository to facilitate the identification of alternative producers of a biosynthetic gene cluster of interest, 2) evaluating the automated process of identifying the known pathways from a genome based on the architectural similarity and finally, 3) creating an overview of biosynthetic model diversity within a potentially extensive database of (predicted and characterized) pathways.

3.3.1 Overview of datasets used in this study

Two different datasets were used in this study to evaluate the performance of the BiosynML conceptual genome mining toolbox. The first one is the myxobacteria set, comprising 42 characterized biosynthetic gene clusters (see Table 2.4) as well as 71 myxobacterial genomes from various sequencing projects (few of them already published, many available only in-house).

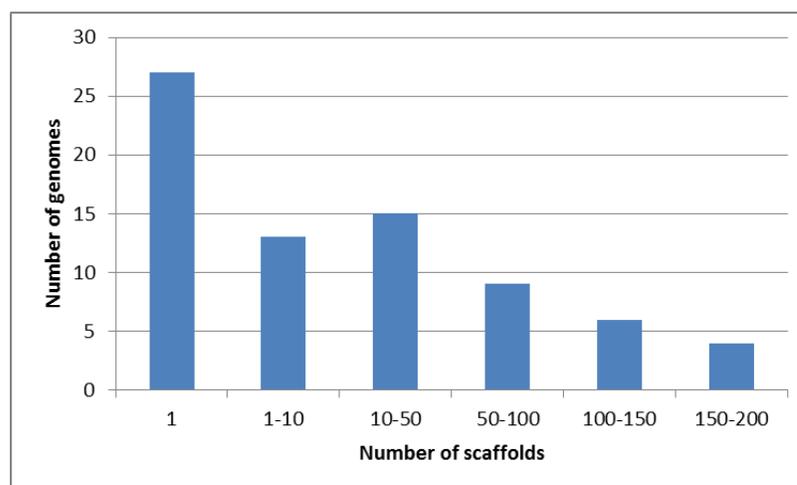


Figure 3.53: Distribution of scaffolds across genomes used in this study

These include complete (27) and draft genomes with up to 200 scaffolds (44) as shown in Figure 3.53. Including the draft genomes is important as the antiSMASH predictions of gene clusters from draft genomes necessarily produce annotated pathways which are partial or have some missing domains or even missing genes. Their inclusion in the test dataset thus could be of use to check the dependency of the conceptual genome mining method in matching and identifying biosynthetic pathways with incomplete information.

In addition, genomes of Actinomycetes (277), Bacilli (77), Cynobacteria (63), Proteobacteria (47) and Serratia (3) were downloaded from NCBI and used as second dataset for the analysis. Focus was on covering a taxonomically diverse set of known and potential natural product sources.

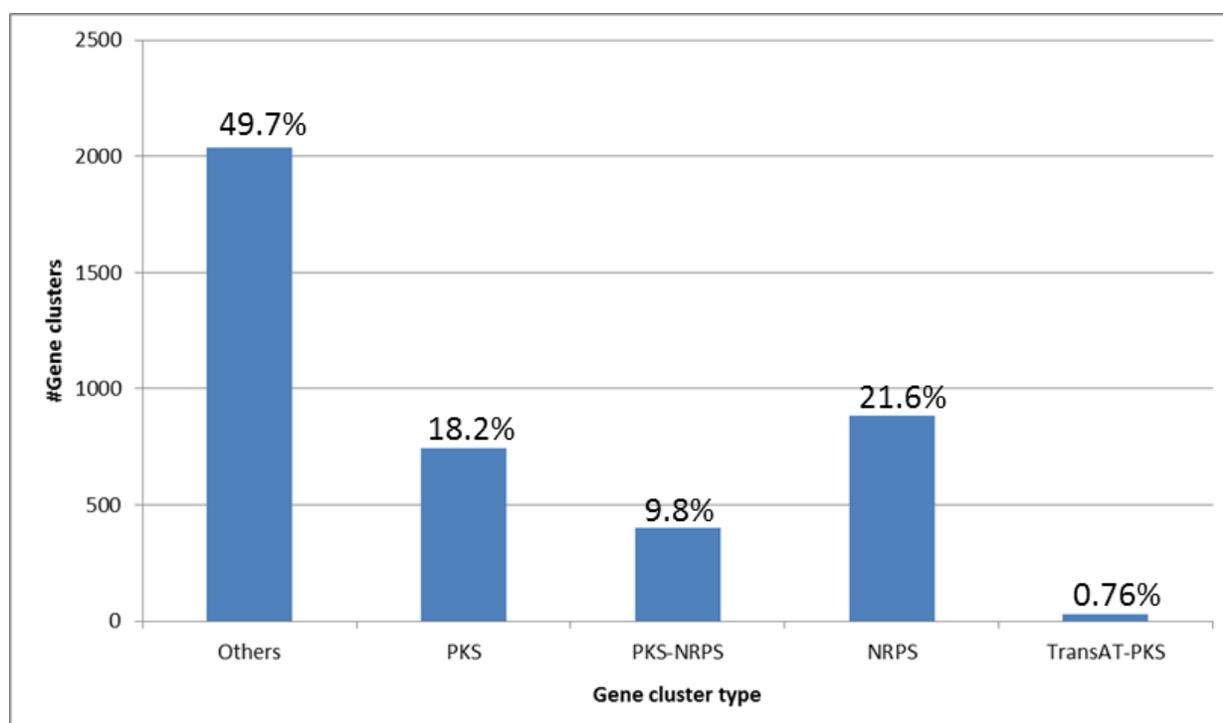


Figure 3.54: Distribution of gene clusters from public genome dataset according to their types.

Further, the gene clusters were classified into PKS, NRPS, PKS-NRPS, TansAT-PKS and others based on the presence of core domains. For the myxobacteria dataset the majority of the gene clusters belonged to hybrid PKS-NRPS (490), NRPS (373), PKS (176), TransAT-PKS (52) and rest of the gene cluster types predicted from antiSMASH are counted as others (254) (Figure 3.55). Similarly, analysing the distribution of gene clusters from the public collection resulted in more PKS (746), NRPS (884), hybrid PKS-NRPS (402), TransAT-PKS (31) and others (2036) (Figure 3.54). This observation suggests that the myxobacterial genomes produce a higher number of hybrid gene clusters (PKS-NRPS) compared to autonomous PKS and NRPS gene clusters. Since hybrid gene clusters contain both PKS and NRPS

modules, they support the production of even more structurally diverse products as opposed to pure PKS and NRPS compounds, which might be seen as a hint at myxobacteria following an evolutionary strategy to produce a large variety of natural products using multimodular NRPS, PKS and hybrid pathways. Notably, the genomes from the public collection have significantly more of the “other” cluster types than the core PKS and NRPS clusters compared to myxobacteria.

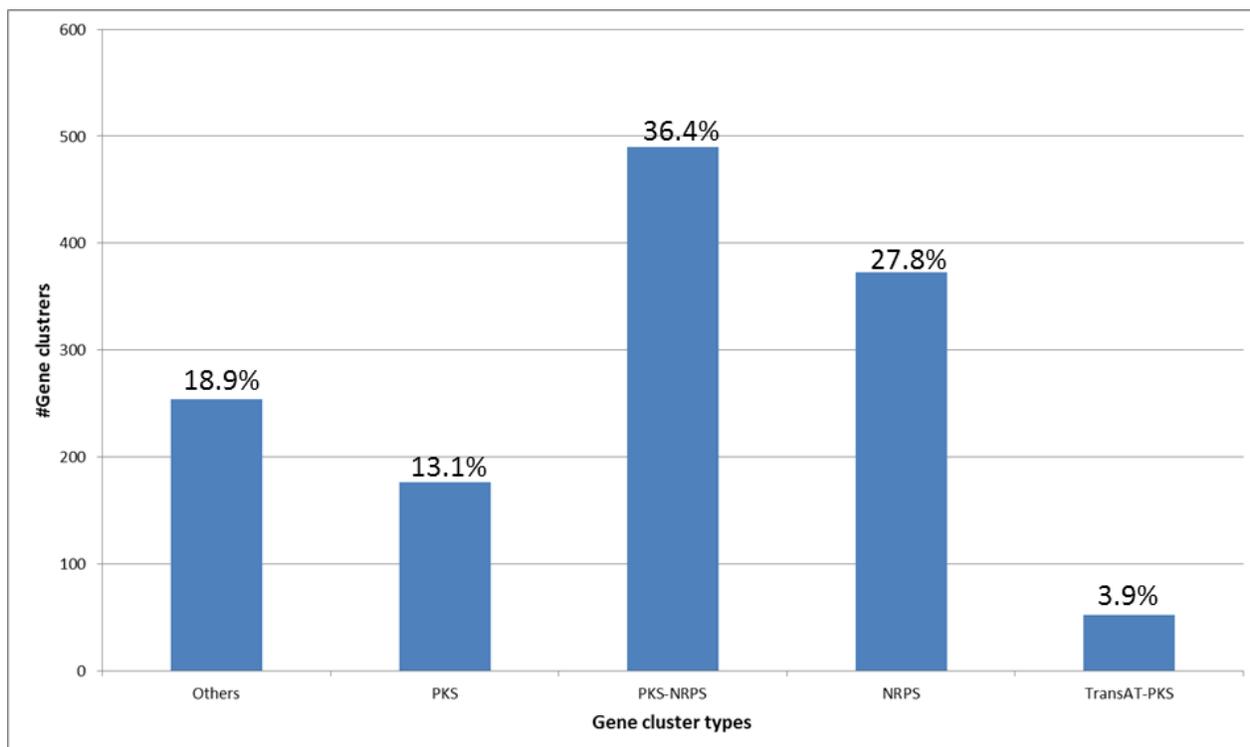


Figure 3.55: Distribution of gene clusters from myxobacterial genome dataset according to their types.

3.3.2 Targeted genome mining: identification of similar gene clusters

In the context of conceptual genome mining with secondary metabolite gene clusters, a “targeted query” refers to the procedure of identifying all gene clusters inside Myxobase which contain pathways giving rise to matching functional, substrate and status properties. The functionality, substrate and status patterns of candidate pathways are compared to the query pathway.

When the user creates a new “targeted query” job, the interface starts with initiating a search against complete database or restricted by genus and species which have been selected by the user and then the working thread performs the pathway matching tasks and outputs the results. The results consist of the information about hits which are scored to be similar or identical to that of the target query. As mentioned earlier in the methods section, the parameters for hit evaluation are overall combined into bitscore and e-value. The additional information delivered by the result interface depends

on the database part which is the source of a hit, and generally comprises biosynthetic pathway dataset key/strain collection key/compound collection key as well as the name of the organism and compounds of both the query input and the hits, if previously assigned.

As a simple test case, a “targeted query” was conducted to reveal biosynthetic pathways which are indicative for presence of the secondary metabolite cluster for Surfactin biosynthetic pathway (Figure 3.56) from the *Bacillus subtilis* strain W168, which already served as a basic example in section 3.2.2.2.

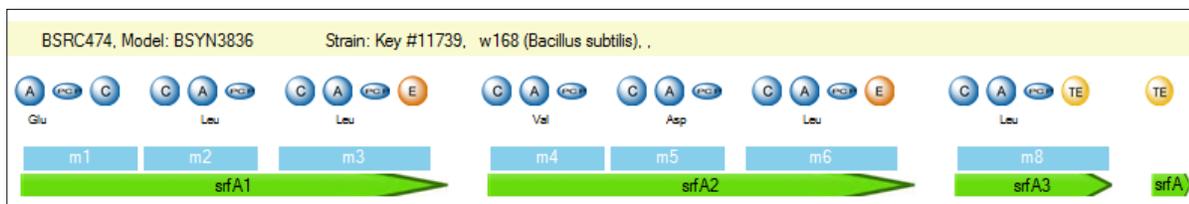


Figure 3.56: Ball scheme representation of surfactin pathway from *Bacillus subtilis* strain W168, used in the following as an example for targeted genome mining with the BiosynML toolbox

First this analysis the search parameters are chosen, setting substrate specificity weighting to 0.3, and using GBS as the method for pathway evaluation and additional domain penalty weighting of 0.3. The collinearity restriction is set to 0 (“off”) in order to avoid the chance of losing pathways which might be variants of the query pathway (similar domain composition but varied domain arrangements with different substrate specificity).

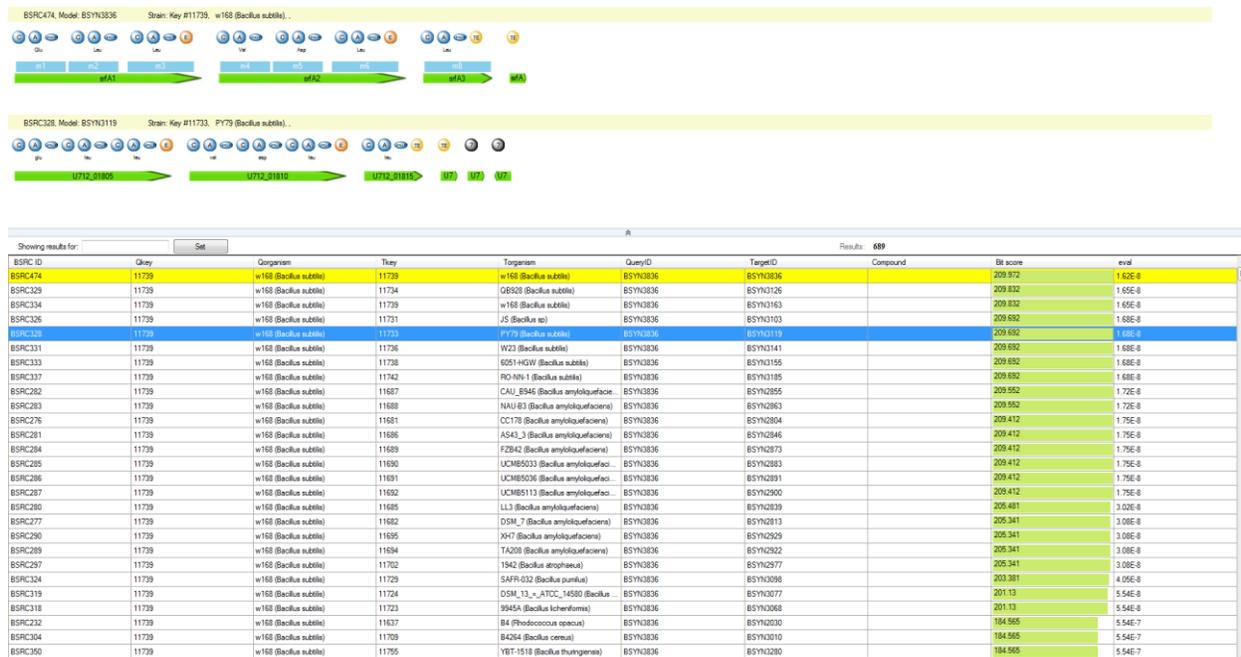


Figure 3.57: Results for surfactin biosynthetic pathway from *Bacillus subtilis* strain W168 queried against public genome dataset. The result highlighted in yellow is the “selfhit”.

The aim of this query is to verify the functions using a known and curated “textbook” pathway, aiming to retrieve all candidate pathways exhibiting similar architecture in the database. After the execution of the query, the results can be seen in hit summary tab as shown in Figure 3.57.



Figure 3.58: Biosynthetic pathway with near-identical domain composition and architecture (besides additional “?” domains). The box around the pathway represents query surfactin pathway from *Bacillus subtilis* strain W168

According to Figure 3.58, several known pathways are identified which are expected as a result because of prior knowledge on the genome dataset from the biosynthetic repository. More specifically, the genome annotation in this case reveals the surfactin pathway leads to identical pathways used by various strains of *Bacillus subtilis* (Table 3.3) and different species of *Bacillus* (Table 3.4). The change in score is because of additional predicted domains that are present in the pathway and minor changes in the substrate specificity predicted by antiSMASH. There are also hits from *Bacillus* genus (Table 3.3) showing identical domain composition but with varied substrate specificity, which reduced the similarity score of the hits as we chose to give high priority to the substrate specificity. We emphasize here, that the pathways whose absolute domain arrangement and operon organization differs (Figure 3.59) from that of the query pathway (Table 3.5) but has similar composition and relative domain arrangement (Figure 3.59) are identified and have a decent similarity score, which are noticeable to a chemist for

investigation. These pathways and could produce compound derivatives, some of which in practice may exhibit reduced toxicity and improved pharmaceutical properties.

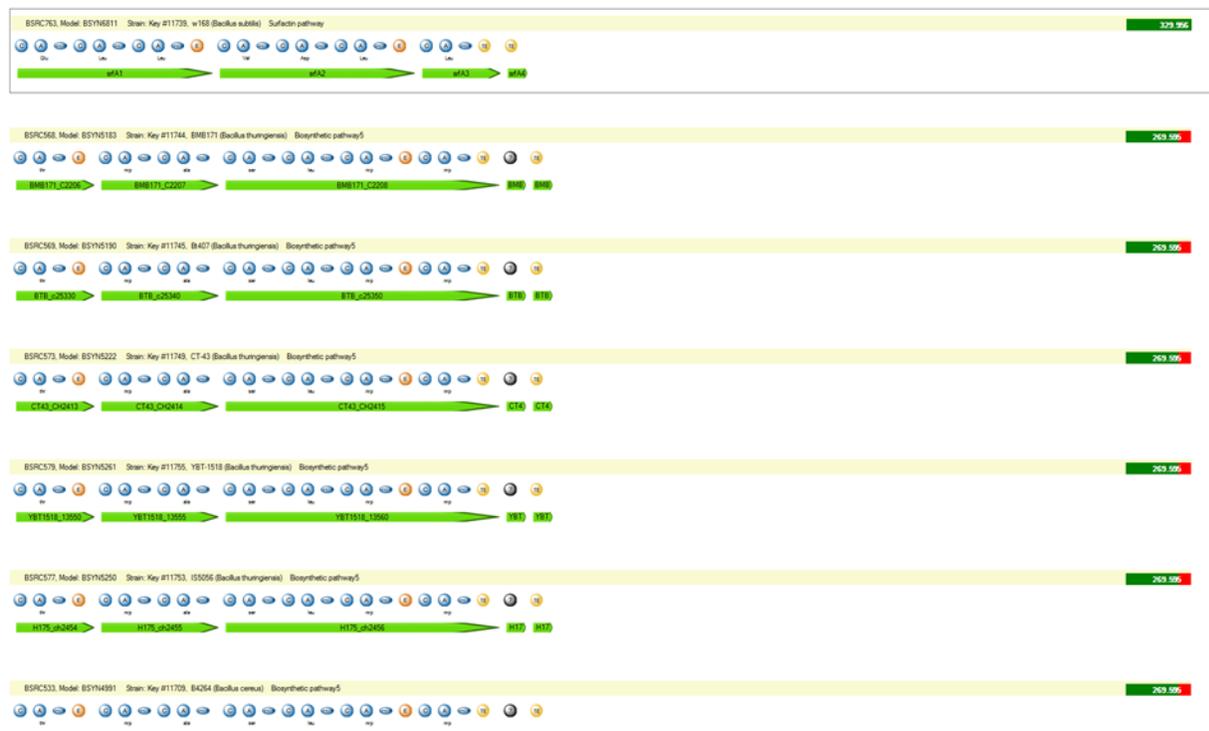


Figure 3.59: Hit pathway with near-identical domain composition but apparently different operon organization and substrate specificities compared to the query surfactin pathway from *Bacillus subtilis* strain W168 (represented in box)

BSYN ID	Strain key	Organism	Score
BSYN5107	11739	W168 (<i>Bacillus subtilis</i>)	329.956
BSYN5107	11734	QB928 (<i>Bacillus subtilis</i>)	329.756
BSYN5144	11739	w168 (<i>Bacillus subtilis</i>)	329.756
BSYN5084	11731	JS (<i>Bacillus</i> sp)	329.755
BSYN5100	11733	PY79 (<i>Bacillus subtilis</i>)	329.755
BSYN5166	11742	RO-NN-1 (<i>Bacillus subtilis</i>)	329.755
BSYN5136	11738	6051-HGW (<i>Bacillus subtilis</i>)	329.755
BSYN5122	11736	W23 (<i>Bacillus subtilis</i>)	329.755

Table 3.3: Identical hits for Surfactin pathway from various strains

BSYN ID	Strain key	Organism	Score
BSYN4836	11687	CAU_B946 (<i>Bacillus amyloliquefaciens</i>)	329.355
BSYN4864	11690	UCMB5033 (<i>Bacillus amyloliquefaciens</i>)	329.154
BSYN4854	11689	FZB42 (<i>Bacillus amyloliquefaciens</i>)	329.154
BSYN4881	11692	UCMB5113 (<i>Bacillus amyloliquefaciens</i>)	329.154
BSYN4872	11691	UCMB5036 (<i>Bacillus amyloliquefaciens</i>)	329.154
BSYN4785	11681	CC178 (<i>Bacillus amyloliquefaciens</i>)	329.154
BSYN4827	11686	AS43_3 (<i>Bacillus amyloliquefaciens</i>)	329.154
BSYN4820	11685	LL3 (<i>Bacillus amyloliquefaciens</i>)	319.529

BSYN4958	11702	1942 (Bacillus atrophaeus)	319.328
BSYN4903	11694	TA208 (Bacillus amyloliquefaciens)	319.328
BSYN4910	11695	XH7 (Bacillus amyloliquefaciens)	319.328
BSYN4794	11682	DSM_7 (Bacillus amyloliquefaciens)	319.328
BSYN5079	11729	SAFR-032 (Bacillus pumilus)	316.328
BSYN5058	11724	DSM_13 = ATCC_14580 (Bacillus licheniformis)	309.301
BSYN5049	11723	9945A (Bacillus licheniformis)	309.301

Table 3.4: Near-identical hits for Surfactin pathway from various species

BSYN ID	Strain key	Organism	Score
BSYN5183	11744	BMB171 (Bacillus thuringiensis)	269.595
BSYN5190	11745	Bt407 (Bacillus thuringiensis)	269.595
BSYN5222	11749	CT-43 (Bacillus thuringiensis)	269.595
BSYN5261	11755	YBT-1518 (Bacillus thuringiensis)	269.595
BSYN5250	11753	IS5056 (Bacillus thuringiensis)	269.595
BSYN4991	11709	B4264 (Bacillus cereus)	269.595
BSYN5241	11752	HD73 (Bacillus thuringiensis)	269.395
BSYN5017	11714	G9842 (Bacillus cereus)	268.793
BSYN5205	11747	HD-789 (Bacillus thuringiensis)	268.793

Table 3.5: Hits for Surfactin pathway with similar domain composition but varied substrate specificity

In fact this output compares well to the results typically obtained from sequence analysis, a finding which will be elaborated in more detail during performance evaluation. In contrast to the sequence analysis procedure, however, no access to the sequence data was necessary, and the user did not have to supply the protein sequence knowledge of representative domains as the starting point for analysis. Annotation is based instead on the pathway knowledge by applying it to the domain feature-predicted clusters deposited in the BiosynML biosynthetic pathway repository. For that, the raw data needs to be processed only once through antiSMASH, and the pathway content is then available inside the BiosynML repository (here: Myxobase) for multiple analyses independently done by researchers. The interface is equipped with a function to visualize the ball scheme representation of the biosynthetic pathways where the researcher can analyse the disturbances in the similarity scores (Figure 3.58).

Additionally, the interface can also plot distribution of similarity scores against all pathways from the database (Figure 3.60). The plot shows a sudden decline in the scores, which indicates the ability of the algorithm to distinguish pathways which have high confidence. A low confidence is assigned to pathways that disagree with modularity of domain composition and high confidence to those that comply with it. This provides the researcher an opportunity to identify the pathways that need more attention. However, the selection of the parameter values also affects the outcome of the result. So, researcher should follow-up analysing the results until the second jump in the plot, as some pathways are scored less because of various factors such as sequencing errors where the antiSMASH pipeline missed prediction of possible domains and errors in the prediction of substrate leads to lower score of

the pathways even though the pathway is highly similar to that of the query pathway used for matching. A detailed result from the targeted results for *Surfactin* is presented in Appendix 6.1-Tab

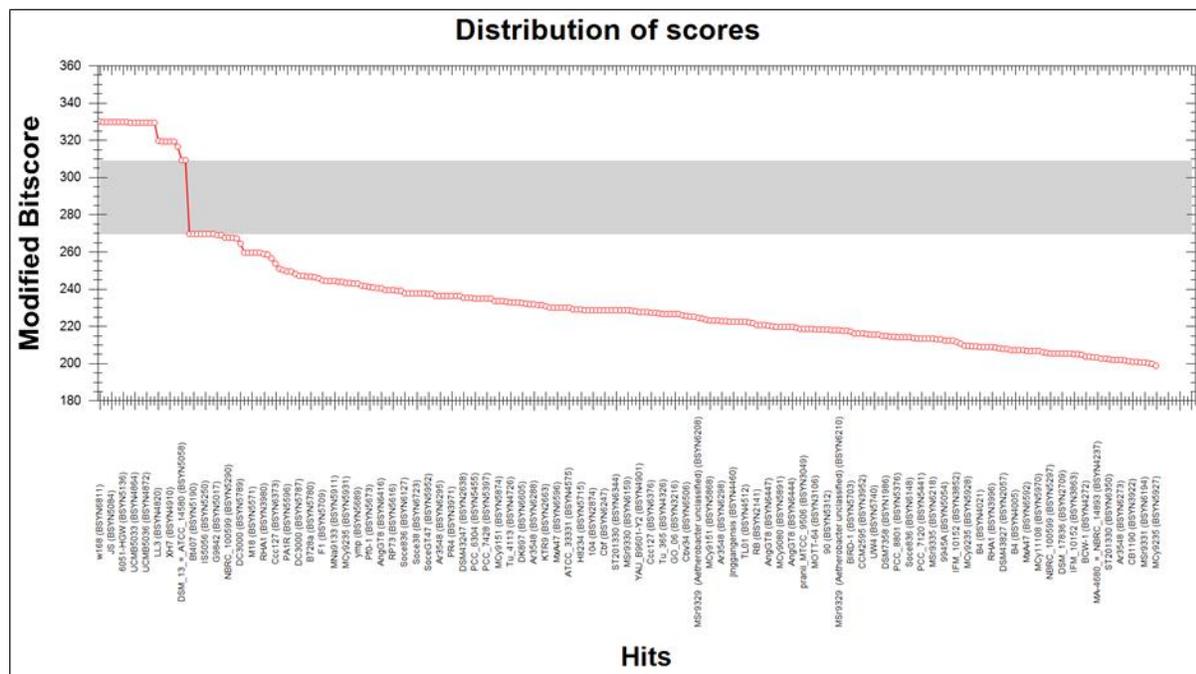


Figure 3.60: Distribution of bit scores for *Surfactin* pathway across all the pathways in the database. The grey box denotes confidence interval of the scores where the domain composition of the pathways starts varying significantly.

3.3.3 Architectural matching vs. sequence-based comparison

Using *Surfactin* pathway as a query, the architectural matching succeeded to reveal pathways which are highly similar to that of the query pathway. It is of interest to check the similarity of the hits at sequence level, too – because this has potentially severe implications for the ability or inability of genome-mining approaches to reveal pathway similarity at high or low levels of taxonomic relatedness. To study this aspect further, the protein sequences of the high scoring gene pairs (query and target) are extracted from the BiosynML files that are stored in the BiosynML repository of Myxobase. These protein sequences from the query and hit are passed through MUSCLE aligner for multiple sequence alignment.

Here, the query is *Surfactin* cluster obtained from W168 strain of *Bacillus subtilis*, the top hit reported by architectural matching was from the strain CAUB946 belonging to *Bacillus amyloliquefaciens*

was chosen. The best hits that have better scores than pathway from CAUB946 are from different strains of *Bacillus subtilis*, so, we chose a pathway from a different species but of the same genus *Bacillus*.

The query Surfactin cluster has 4 genes that are matched against 7 genes of CAUB946, out of which 4 are reported as hits. The protein sequences of the high scoring pairs are aligned and the results are as shown in Table 3.6.

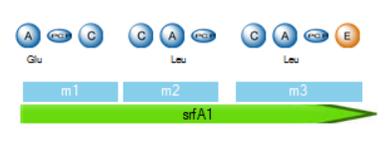
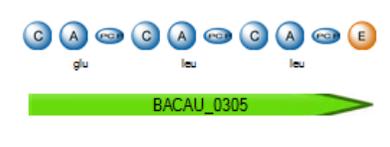
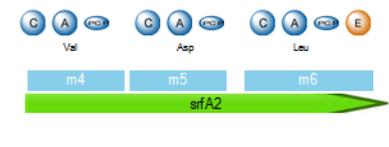
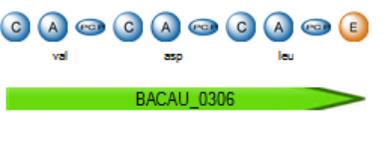
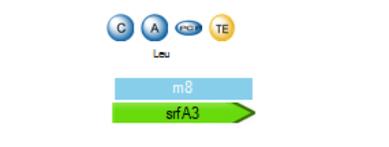
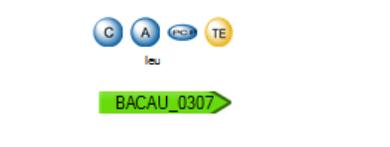
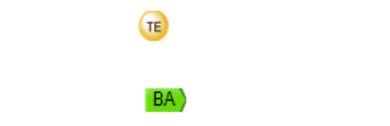
Query W168 (<i>Bacillus subtilis</i>)	Hit CAUB946 (<i>Bacillus amyloliquefaciens</i>)	Similarity (%)	Identity (%)
		83.54	72.09
		83.41	73.84
		88.89	83.72
sfrA4 	BACAU_0308 	83.95	75.31

Table 3.6: Protein sequence similarity of genes between query pathways from W168 (*Bacillus subtilis*) and hit pathway from CAUB946 (*Bacillus amyloliquefaciens*)

As the pathway from CAUB946 shows similar domain compositions with exact substrate specificities, the sequence analysis also resulted in high similarity.

Query W168 (<i>Bacillus subtilis</i>)	Hit B4264 (<i>Bacillus cereus</i>)	Similarity (%)	Identity (%)
		54.7	35.4
		55.5	34.2
		52.3	32.5
		55.9	34.9
		55.3	30.5

Table 3.7: Protein sequence similarity of genes between query pathway from W168 (*Bacillus subtilis*) and hit pathway from B4264 (*Bacillus cereus*). The red boxes represent the modules that are distributed differently across genes, still the full complement of domains could in principle bring about a Surfactin-like molecule.

Another hit which can be a potential variant of Surfactin was chosen to check the sequence similarity between genes. A hit from strain B4264 from *Bacillus cereus* was chosen to show that the specificities of the pathway vary with that of the query (Table 3.7). The sequence similarity between the query and hit shows very low score while the overall domain composition of the pathway looks to be similar, albeit with altered distribution of modules across genes and varying substrate specificity.

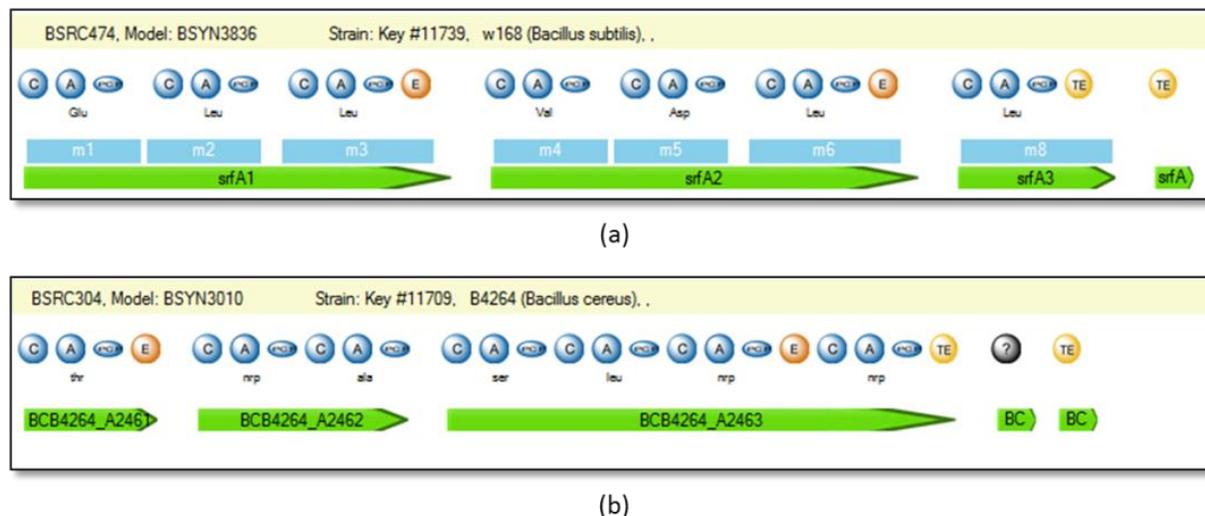


Figure 3.61: Ball scheme representation of (a) Surfactin pathway from W168 (*Bacillus subtilis*) (b) hit pathway from B4264 (*Bacillus cereus*). The target gene BCB4264_A2463 has hits with both query genes *srfA1* and *srfA3*, query gene *srfA2* has two hit target genes BCB4264_A2461 and BCB4264_A2462 and query gene *srfA4* has a hit with target gene BCB4264_A2461

Hits of such type are very easy to be missed by chemist using sequence based analysis as the sequence similarity results in very low scores for genes in query pathways. The query gene *srfA1* has a similarity of 54.7% with BCB4264_A2463. The target gene BCB4264_A2463 partially is also a hit for query gene *srfA3* has a similarity of 55.5%, *srfA2* has two hit genes BCB4264_A2461 and BCB4264_A2462 having a similarity of 52.3% and 55.9% respectively and query gene *srfA4* has a hit to gene BCB4264_A2461 of similarity 55.3% (Figure 3.61).

We furthermore constructed a cladogram for the hits based on the similarity scores obtained through architectural matching. Each leaf of the cladogram displays the label that provides a unique biosynthetic pathway identification key (BSYN ID) as well as the strain, genus and species. Considering the pathways which are reported as hits compared to the best possible score (score of self-hit), yields five distinct groups of “Surfactin-like” candidate gene clusters.

A clade is a set of closely related pathways which is represented as a subtree as shown in Figure 3.62. The hit pathways which resembles to Surfactin pathway in terms of both domain functionality and specificity are grouped together as a clade (Figure 3.62 a), the hit pathways which have high similarity in domain functionality but changes in predicted substrate specificity are grouped as another clade (Figure 3.62 b) where e.g. the leucine-activating A-domain was replaced with isoleucine-specific A-domain. This variant of surfactin with isoleucine substitution was in fact already characterized and is found in Norine database (130). The pathways which have high similarity in the functionality but have more different substrate specificities are also roughly grouped together (Figure 3.62 c, Figure 3.62 d). These clades form

subclades to the difference in the scores obtained because of presence of additional domains in the pathways which are penalised. Clade from Figure 3.62 (e) is an example for heterogeneous clade as the domain composition disintegrates with respect to the query domain composition.

Finally, to analyse the behaviour of the conceptual method with small gene clusters, the Myxochelin pathway from strain Sga15 (*Stigmatella aurantiaca*) was used (131). This pathway represents a challenging test case since it has merely a handful of distinct domains, as opposed to the large set of domains comprised in the pathway models used in previous examples. In that respect, myxochelin represents the “borderline” case for using the conceptual genome mining approach in this study.

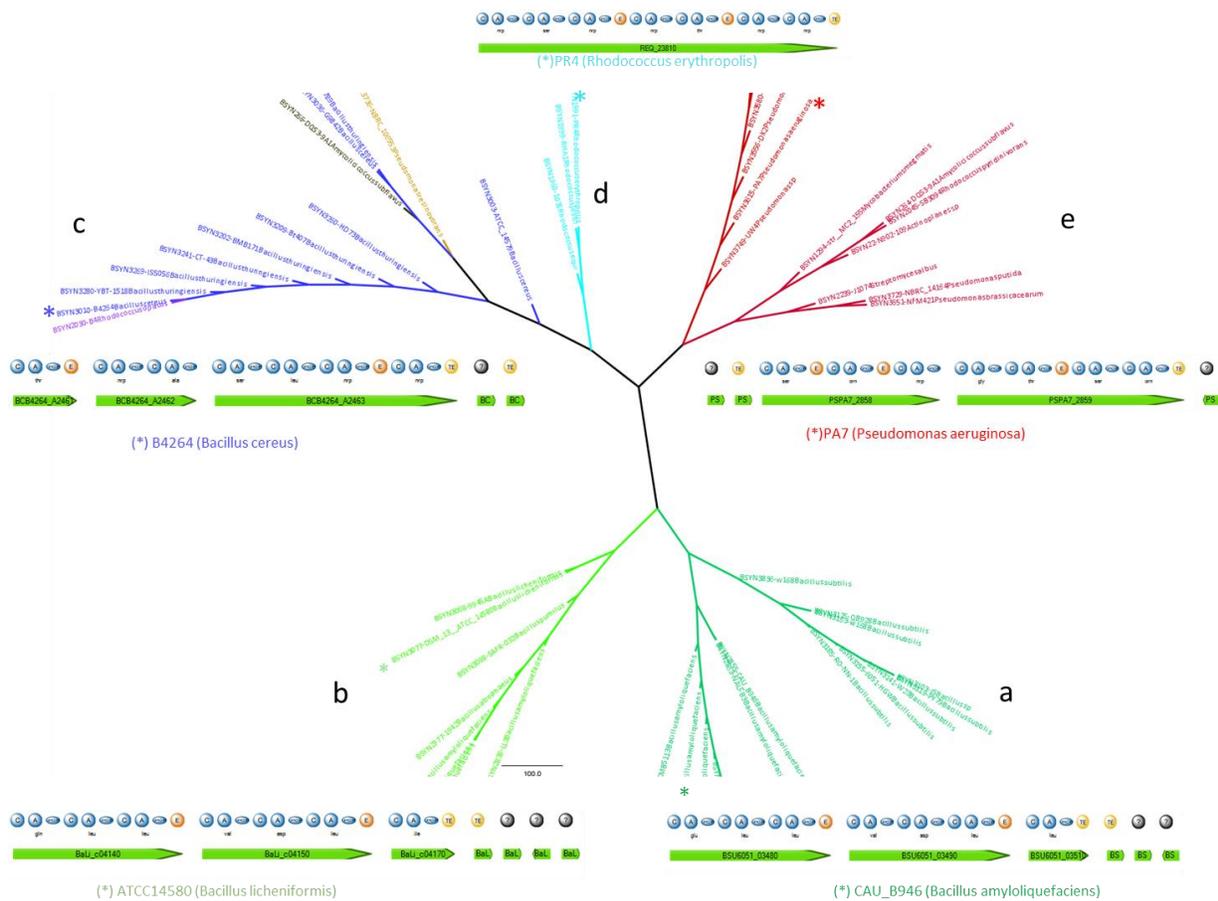


Figure 3.62: A cladogram of the pathway established based on the scores obtained from architectural matching of Surfactin pathway. a) Pathways with similar domain functionality and substrate specificity. b) Pathways with similar domain functionality and minor deviation in the substrate specificities. c, d) Pathways with similar domain functionality with deviations in the substrate specificities. e) Pathways with disintegrating similarity between domain functionality as well as substrate specificities.

The method reported nine hit pathways from the public genome dataset which resembles myxochelin pathway. Among these, there are five hits which have identical domain composition as well as substrate specificity (Figure 3.63 a) and four pathways with deviating domain composition and

deviations in the substrate specificities (Figure 3.63 b) along with missing KR domain (which is supposedly inactive and has no assigned role in the myxochelin biosynthesis). The pathways which are below the confidence interval have overall similar domain composition (small NRPS cluster with terminal Red domain) but deviates with substrate specificities (Figure 3.63 c) along with missing and additional domains.

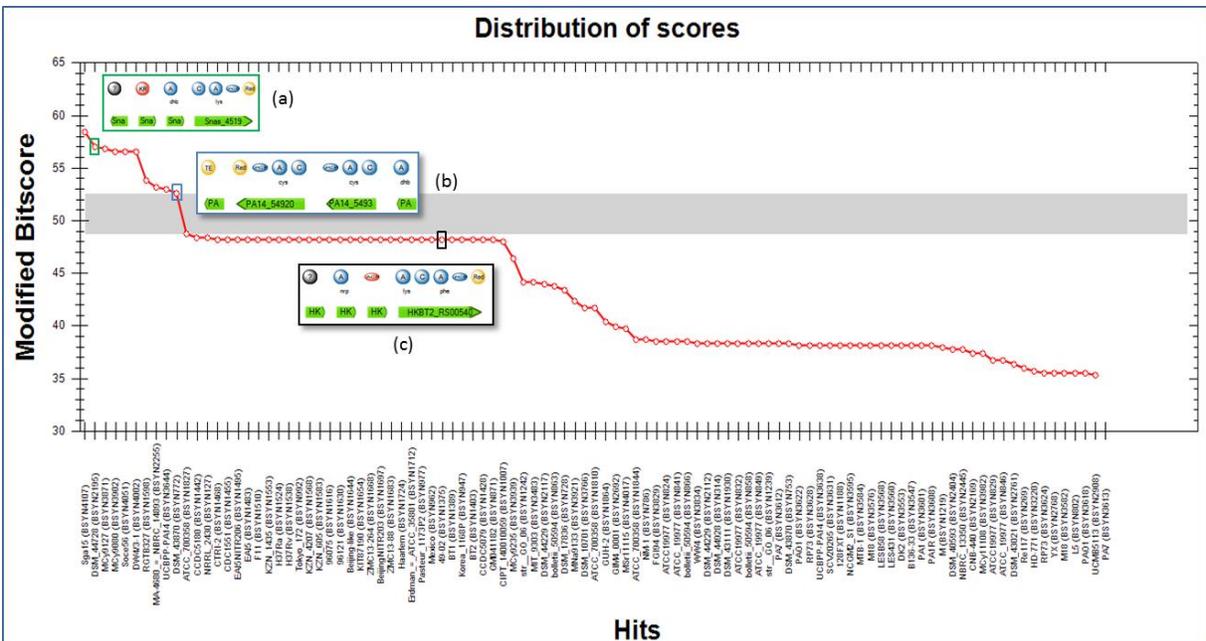


Figure 3.63: Bitscore plot for Myxochelin genecluster. (a) A plausible myxochelin pathway from strain DSM44728 (*Stackebrandtia nassauensis*) which is a near identical hit. (b) A pathway from strain UCBPP-PA14 (*Pseudomonas aeruginosa*) showing deviations in the substrate specificity, but retaining the “C-A-PCP-Red” arrangement as in the myxochelin pathway and also harbouring a *dhb*-activating A-domain. (c) A representative pathway from strain EAI5 (*Mycobacterium tuberculosis*) that has deviations in the substrate specificities, additional and missing domains and not preserving the myxochelin-typical four domain NRPS gene.

On analysing the protein sequence similarity of best hit from strain DSM44728 (*Stackebrandtia nassauensis*) for Myxochelin pathway gave notable results (Table 3.8). The query gene *mxcG* has only a similarity of 47.4 % with the target gene *Snas_4519*, the query gene *mxcE* has similarity of 47.4 % with the target gene *Snas_4516* and the query gene *mxcC* has 64.1% similarity with target gene *Snas_4514*.

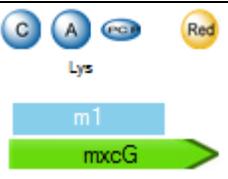
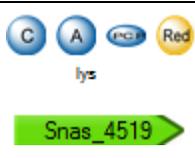
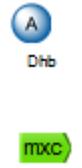
Query Sga15 (<i>Stigmatella aurantiaca</i>)	Hit DSM44728 (<i>Stackebrandtia nassauensis</i>)	Similarity (%)	Identity (%)
		47.4	36.1
		67.8	56.4
		64.1	51.6

Table 3.8: Protein sequence similarity of genes between query pathways Sga15 (*Stigmatella aurantiaca*) and hit pathway from DSM44728 (*Stackebrandtia nassauensis*)

Considering another top hit from strain Soce56 (*Sorangium cellulosum*) for analysing protein sequence similarity showed convincing agreement with the query myxochelin pathway (Table 3.9). The query gene mxcG has similarity of 91.6% with the target gene ctg1_orf09037, the query gene mxcE has similarity of 78.9 % with the target gene ctg1_orf09020 and the query gene mxcC has 79.8% similarity with target gene ctg1_orf09024.

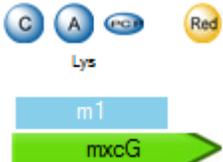
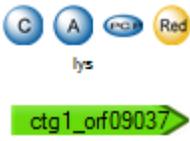
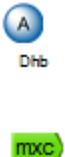
Query Sga15 (<i>Stigmatella aurantiaca</i>)	Hit Soce56 (<i>Sorangium cellulosum</i>)	Similarity (%)	identity (%)
		91.6	87.1
		78.9	67.9
		79.8	68.2

Table 3.9: Protein sequence similarity of genes between query pathways Sga15 (*Stigmatella aurantiaca*) and hit pathway from Soce56 (*Sorangium cellulosum*)

Analysing the similarity at protein level with a hit from strain UCBPP-PA14 (*Pseudomonas aeruginosa*) which has partially similar domain composition but deviations in predicted substrate specificities (Table 3.10), the query gene *mxcG* has 41.4% similarity with target gene *PA14_54920*. But the query gene *mxcE* which has a single standing, *dhb*-activating “A” domain showing 68.6% similarity with hit gene *PA14_54940*. The *mxcC* gene coding for (inactive/unnecessary) KR domain is not found in the hit pathway.

Query Sga15 (<i>Stigmatella aurantiaca</i>)	Hit CBPP-PA14 (<i>Pseudomonas aeruginosa</i>)	Similarity (%)	Identity (%)
		40.91	25.3
		68.6	52.2

Table 3.10: Protein sequence similarity of genes between query pathways Sga15 (*Stigmatella aurantiaca*) and hit pathway from CBPP-PA14 (*Pseudomonas aeruginosa*)

On the basis of examples provided in this section, the targeted query module within the BiosynML conceptual genome mining framework is able to annotate identical, near-identical and similar to remote-similar pathways from the database of candidate pathways on the basis of architectural similarity in terms of domain compositions, domain arrangement and associated meta data. Some of the reported hits might have been achieved also on the basis of sequence similarity (e.g. using blast), however with the inherent difficulty of choosing an appropriate sequence stretch and determining the similarity threshold, a problematic decision especially in light of similar gene clusters appearing in phylogenetically distant taxa. This difficulty is illustrated by the above examples of althiomycin, surfactin and myxochelin.

In the case of myxochelin, a top hit pathway from strain DSM44728 (*Stackebrandtia nassauensis*) was detected which has discouraging low protein similarity compared to the myxobacterial pathway, but has the identical domain composition as well as substrate specificities. It is not known whether this species is able to produce myxochelin, but the conceptual genome mining result strongly suggest it could be a producer of a structurally very similar molecule. On the contrary, observing the protein similarity of the top hit pathway from myxobacterial strain Soce56 (*Sorangium cellulosum*) showed high similarity with the genes in the *Stigmatella* myochelin pathway and thus would have been easily identified by sequence analysis, too. Since targeted query performs search on architectural level, these myxochelin candidates from different taxa were detected and reported; decreasing the allowed “substrate specificity” and “additional domain penalty” tolerance could possibly also report more hits above the confidence interval generated, but at the cost of expecting an increased false discovery rate in terms of deviations in substrate specificities and functionality. These incidents were investigated in-detail by manual inspection of ball scheme represented clusters generated in the descending order of bit-score, and it depends on the specific application whether it is desired to include less specific hits in the result set.

Using a biosynthetic pathway as an input, the targeted genome mining module reports on the occurrence of high-scoring pathways with highly similar domain compositions and arrangements, regardless whether these belong to a known compound or to an unidentified feature. Output is a set of pathways that have highly similar domains architecture ordered in the decreasing order of bit scores obtained and thus can be used to rapidly identify alternative producers for the similar pathway from a large collection of genomic datasets. The special importance of this analysis is due to the fact that the genomics-guided biosynthetic pathway discovery strategies may reveal some novel pathways for compounds which are initially only known from the respective discovery strain. When an instant overview of potential alternative sources, based on the presence of the underlying pathway, is easily available using the targeted query, then chances to purify and structurally elucidate the novel compound using an alternative producer are much increased. In a different scenario, the targeted query result may highlight a genome-sequenced strain as a producer of the novel compound, using a hypothetical assembly line as query input; in that case the result paves the way to the new assignment of a previously “orphan” biosynthesis gene cluster to the metabolite it produces.

Overall, the comparison of biosynthetic pathways across the BiosynML pathways repository, facilitated here by the targeted conceptual query tool in Myxobase, should provide a powerful method for the natural product researcher for the discovery and identification of biosynthetic gene clusters.

3.3.4 Genome annotation and dereplication analysis of biosynthetic gene clusters

Genome annotation refers to the following procedure: using domain properties in a genome dataset stored in Myxobase, for each of the biosynthetic pathway in a given genome a query is executed against all the pathways in the database. A matching is performed of the candidates and hits are evaluated using function, substrate specificity and status deviation as criteria.

The framework significantly enhances the search of conceptual information stored in the database by integrating meta information such as function, specificity and status of the domains in the searches. A genome sequence submitted to the procedure is pathway-wise evaluated for high scoring hits in the known pathways library and gene clusters are auto-assigned to pathways showing plausible architectural similarity. This meta information is used by search and match algorithms which outputs the result both in the form of a table showing scores and hits and bitscore plot showing the graphical representation of distribution of hits with reference to each pathway from genome queried. The bitscore plot generates a confidence interval based on the self-hit score, the best possible score for the query pathway; the hits above the confidence interval can be recognised as the clusters which are identical or

highly similar to that of the query cluster. When these hit clusters above the confidence interval are linked to the compound information of the characterized pathway in the library, any new genome-encoded pathway used as a query could be tentatively assigned to that metabolite, too.

As a test case for the genome annotation module, we used datasets from the myxobacterial strains DK1622, Cmc5 and Soce56. A function was added to the Myxobase interface, which imports the data from the .xml file delivered by the antiSMASH annotation pipeline into the database. All the annotations from the dataset are stored in the Myxobase biosynthetic pathway repository with a unique pathway key. For dataset from DK1622 some of the pathways namely Myxochromid, Myxoprincomide and Myxovirescin are manually curated using BiosynML plugin for Geneious such that the reference library contained both, curated and non-curated pathways.

The conceptual genome auto-annotation for the following example is triggered with these query settings: selection of genome set to be annotated, substrate specificity penalty (0.3), additional domain penalty (0.3), and search against all the pathways in the database. The information content is in principle similar to that of the targeted query results.

Observing the results of conceptual genome annotation for strain DK 1622 (Table 3.11), six known pathways which are Myxovirescin, Myxochelin, Myxochromid, DKxanthen, Myxalamid and Myxoprincomide are identified which are expected as a result because of prior knowledge on the pathways from the *M. xanthus* genome (85, 132), which was used to create the dataset. These known pathways from the DK 1622 genome were reliably assigned by the algorithm.

Query ID	Hits above confidence interval	Bitscore (best hit score/ self-hit score)	Assignment to pathway (number of domains)	Hits in suborders (frequency)
BSYN109	0			
BSYN110	23	49.9/49.9 (100%)	Uncharacterized Short NRPS-PKS (6)	Cystobacterineae (20), Sorangiineae (2) and Nannocystineae (2)
BSYN111	2	176.2/176.2 (100%)	Uncharacterized NRPS (21)	Cystobacterineae (3)
BSYN112	10	66.8/66.8 (100%)	Uncharacterized Short NRPS (8)	Cystobacterineae (8) and Sorangiineae (3)
BSYN114	4	201.2/201.6 (99.8%)	Uncharacterized NRPS-PKS (24)	Cystobacterineae (5)
BSYN115	1	541.4/572.1 (94.6%)	Uncharacterized long NRPS-PKS (68)	Cystobacterineae (2)
BSYN116	42	56.9/58.4 (97.4%)	Myxochelin (7)	Cystobacterineae (35) and Sorangiineae (7)
BSYN117	2	313.8/319.5 (98.2%)	Myxoprincomide (38)	Cystobacterineae (2)
BSYN118	2	622.8/639.5 (97.4%)	Myxovirescin (76)	Cystobacterineae (2)
BSYN119	1	239.1/243.7 (98.1%)	Uncharacterized NRPS-PKS (29)	Cystobacterineae (2)
BSYN120	10	226.8/226.8 (100%)	Myxochromid (27)	Cystobacterineae (10)
BSYN121	10	285.8/285.8 (100%)	DKxanthene (34)	Cystobacterineae (10)
BSYN122	1	403.7/403.7 (100%)	Uncharacterized long NRPS-PKS (48)	Cystobacterineae (2)
BSYN123	9	386.8/386.8 (100%)	Myxalamid (46)	Cystobacterineae (9)
BSYN124	1	252.1/252.1 (100%)	Uncharacterized NRPS (30)	Cystobacterineae (3)
BSYN125	15	49.7/49.9 (99.5%)	Uncharacterized short PKS (6)	Cystobacterineae (15)

Table 3.11: Results of genome annotation for DK 1622 using architectural matching of domains functionality, substrate specificity and status.

According to the outcome of this analysis, strain *Myxococcus xanthus* DK1622 has no unique secondary metabolite pathways, as each biosynthetic gene cluster is found in at least two other myxobacterial strains. This picture may however at least in part be because of a fair number of additional *Myxococcus* genomes in the test database, implicitly increasing the chance/risk to encounter evolutionary “close” genomic content.

Genome wide detection and identification of known pathways

The availability of the BiosynML matching method made it possible to construct a dereplicated library for biosynthetic pathways, allowing to highlight known pathways from newly sequenced genomes, thereby revealing the unknown pathways which might be of interest to the chemist. The BiosynML framework efficiently compares the predicted pathways from a genome and assigns compound information to the highly similar characterized pathways identified as a hit which are in general found indicated above the confidence interval. This automated analysis requires very little human interaction to identify the known pathways from the newly annotated genome dataset. As an example DK1622 (*Myxococcus xanthus*), Cm c5 (*Chondromyces crocatus*) and So ce56 (*Sorangium cellulosum*) genomes were used to test the functionality and ability of the algorithm to identify the known pathways by comparing them to the characterized pathways stored in the database. The queries are executed using GBS algorithm, Substrate specificity of 0.3 and domain penalty of 0.3.

The algorithm identified the known clusters of Myxovirescin, Myxochelin, Myxochromid, DKxanthen-534 and Myxalamid clusters from DK1622 genome, Crocacin, Chondramid, Thuggacin and Ajudazol, Chondrochloren clusters from Cmc5 genome and Etnangien, Chivosazol and Myxochelin clusters from Soce56 genome as expected (Table 3.12). For example, querying myxochelin pathway from strain DK1622 identified the similar pathway from Soce56 with high confidence (Figure 3.64). Bitscore plots for the respective pathways from the three genomes can be seen in Appendix 6.2.

Genome	Cluster acknowledged
DK1622	Myxovirescin Myxochelin Myxochromid DKxanthen-534 Myxalamid Myxoprincomide
Cmc5	Crocacin Chondramid Thuggacin Ajudazol Chondrochloren
Soce56	Etnangien Chivosazol Myxochelin

Table 3.12: Overview of known pathways recognized by BiosynML framework algorithm from genomes DK1622, Cmc5 and Soce56. Here, true negatives are the known pathways which are not identified by the algorithm and false positives are the hits reported by the algorithm as known pathways which in reality it is not. No such mis-assignments occurred.

This auto recognition process facilitates the rapid identification of strains that has great potential to produce novel secondary metabolites as well as the strains that produce known compounds which has the potential for improving the process of natural product pathway characterization to a great extent. For example, querying myxochelin pathway from strain DK1622 identified the similar pathway from Soce56 with high confidence, together with a high number of additional myxochelin pathways from other strains (Figure 3.64).

With development in the sequencing technologies, the secondary metabolite gene cluster information is increasing rapidly, and there is a great demand for a medium which can store the derived information efficiently, such that this can be accessible and used by researchers in a collaborative manner. A database system is the logical medium for this purpose, but few such specialized systems exist for research in the field of natural products. In that respect the Mxbase system represents a novelty, by integrating both chemical compound and biological data management and additionally providing the tools for analytical workflows in one database-driven environment for natural products research.

It should also be highlighted that the research questions which the tools developed here are able to answer, are in fact quite significant for the novel biosynthetic gene clusters discovery.

The genome wide annotation is used to classify all known and unknown clusters from a genome which are stored in the database. In this case, the input is the full complement of pathways in a genome dataset. Each pathway features contained therein is compared to clusters stored in the database using the annotation properties, and matching is done using the chosen algorithm in order to confirm a plausible hit. Thus, this function can highlight pathways, which are to date unidentified and are actually likely to make the containing strain a producer of novel compounds.

The results obtained using the newly developed tools are promising and routine operation of the analysis framework is feasible. For reasons of computational performance matching and scoring algorithms are placed on the Mxbase server that performs the calculations through a job submission system which leaves the interface useable for the user.

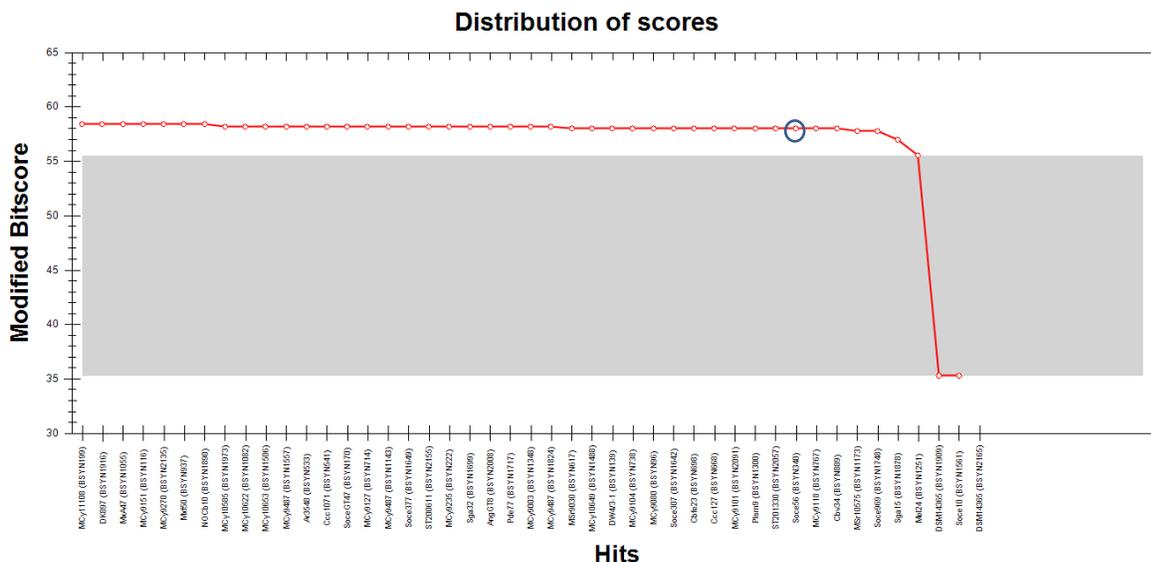


Figure 3.64: Myxochelin pathway from strain DK1622 query resulted in the identification of a hit pathway from Soce56 (blue circle) with high confidence. The plot also reveals that the myxochelin pathway is ubiquitous in genomes of the suborders Cystobacterinaea and Sorangineae but not in the Nannocystineae

Although the tools for matching are server-side functions, with the increase in the size of biosynthetic pathway repository, the tools require relatively high computational power for performing the matching functions which make the users wait for long time to obtain results. In order to reduce the waiting time a future version could perform parallelized processing of the pathway comparison. Implementing the parallelized version of mining tools on a server-side application would surely increase the effective usage of the tools. The generalized workflow of the BiosynML-enhanced genome-mining framework is illustrated in the Figure 3.65.

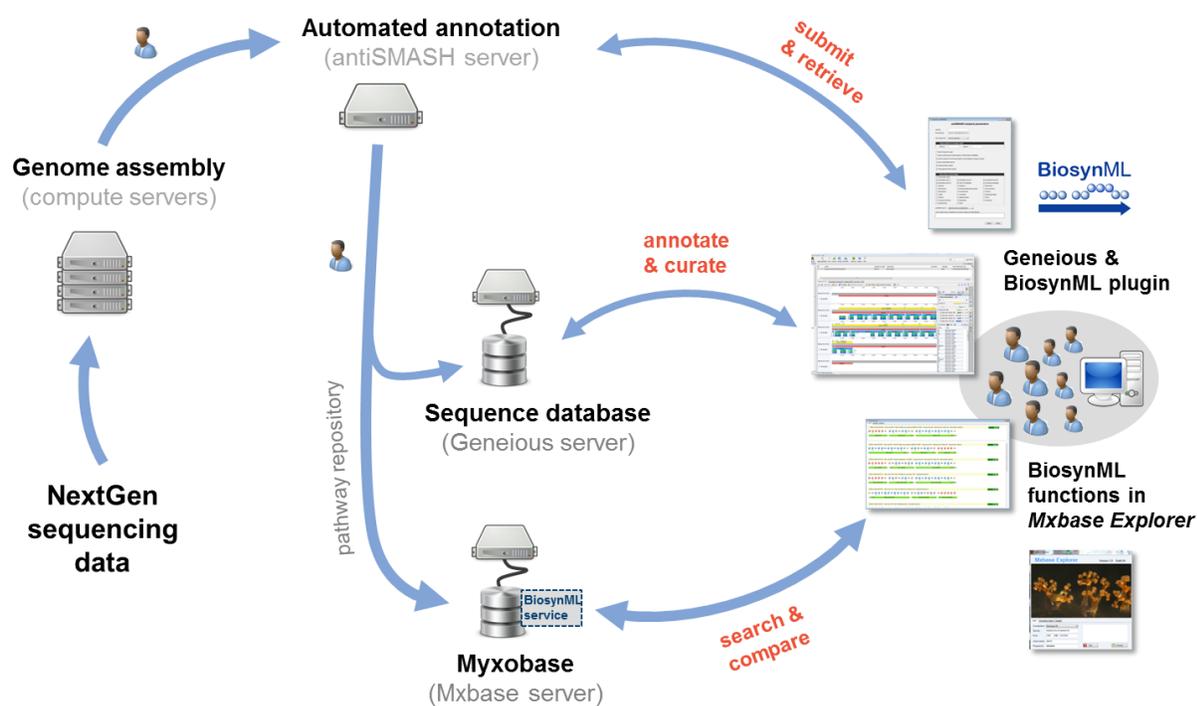


Figure 3.65: An overview of BiosynML enhanced genome-mining framework

3.3.5 Exposing the diversity of secondary metabolite pathways in myxobacterial genomes

The present understanding of the diversity of secondary metabolic pathways in myxobacterial strains are mostly based on the studies concentrated on the biosynthesis of compounds linked to PKS and NRPS pathways. In this study, we performed a large-scale analysis of myxobacterial biosynthetic pathways by combining the conceptual information of pathways obtained from the biosynthesis of compounds screened from myxobacterial strains as well as the genomic data of myxobacterial strains available in the Myxobase biosynthetic pathway repository. This suborder-level investigation allowed identification of biosynthetic pathways which are restricted to suborder(s) and the pathways which are distributed across all the suborder genomes.

The gene clusters from the genomes along with the characterized pathways were analysed using GBS methods and results later sorted based on the originating suborder. Due to the differences in the quality of the genomes used in this study, analysis in the following is carried out with relaxed parameters. This analysis should be repeated at a later timepoint using more stringent parameters with genomes of good quality.

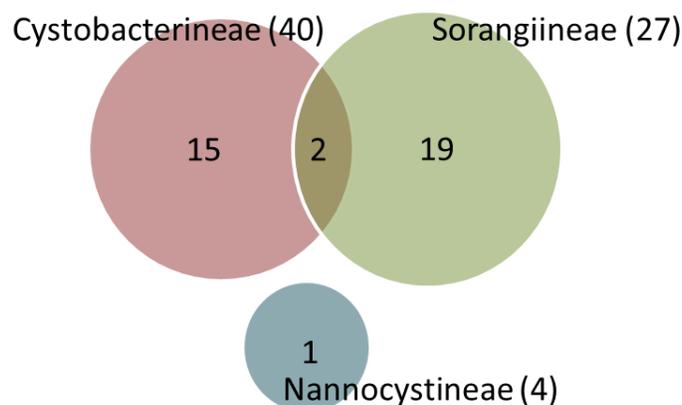


Figure 3.66: A Venn diagram showing the distribution of previously characterized gene clusters with known products, from suborders Cystobacterineae (red), Sorangiineae (green and Nannocystineae (blue)). The number of genomes in each suborder was shown in brackets.

Analysing the distribution of characterized clusters (Figure 3.66) among the suborders, 34 pathways are observed to be suborder specific among which 15 pathways exclusively belong to Cystobacterineae (40 genomes) encoding gene clusters for NRPS (2), PKS (4), NRPS-PKS (8) and trans-AT PKS (1), 19 pathways belonging exclusively to Sorangiineae (27 genomes) encoding gene clusters for NRPS (1), PKS (3), NRPS-PKS (10) and trans-AT PKS (5) and a single pathway (NRPS-PKS) was found to be suborder specific in Nannocystineae (4 genomes). The latter is the biosynthetic pathway for Nannochelin, the gene cluster characterized from a member of Nannocystineae. There are two pathways shared between Cystobacterineae and Sorangiineae suborders, these are the pathways for myxochelin and tubulysin. Myxochelin is known as an iron chelator produced by all species of Cystobacterineae and Sorangiineae analysed to date, albeit mass spectrometry data indicate that production titers generally depend largely on iron supply in the respective cultivation. Tubulysin production, however, has so far only been attributed to strains from Cystobacterineae suborder (133), and Myxobase contains no mass spectrometric evidence that strain So ce836 - which is suggested by the analysis here (see Figure 3.67) to contain a tubulyin gene cluster – actually produces the compound. Close inspection of the candidate

cluster from that strain, however, reveals an unusual domain arrangement where an A-domain seems to have been duplicated (Figure 3.67 c). Thus, it could be conceived that the gene cluster might have become non-functional as consequence of this mutation, which might explain non-production of tubulysins despite the fact that the candidate gene cluster must be clearly classified as a tubulysin pathway based on overwhelming architectural similarity. Another explanation may be the production of a tubulysin analogue which is not detected in our target screening for MS data. In fact the tubulysin pathway has been shown before to produce a large diversity of tubulysin congeners (134).

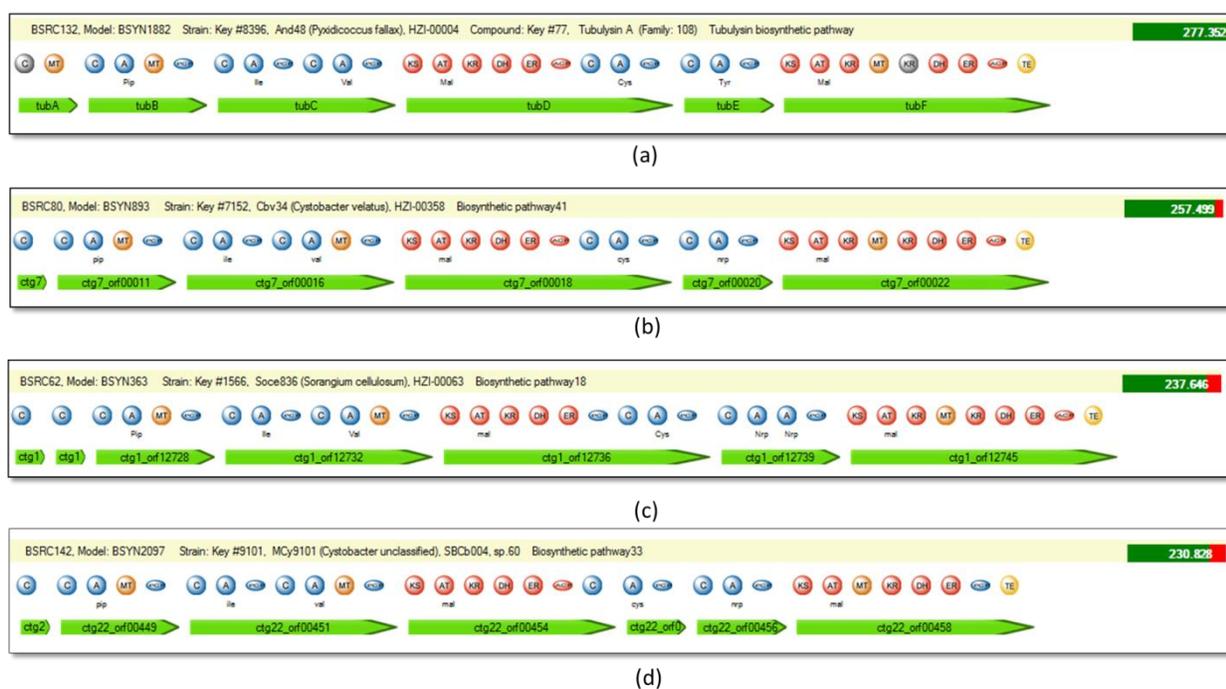


Figure 3.67: Ball scheme representation of tubulysin pathway from And48 (*Pyxidicoccus fallax*) and hits reported. (a) Tubulysin (query) pathway from And48 (*Pyxidicoccus fallax*) (b) hit pathway from Cbv34 (*Cystobacter velatus*) (c) hit pathway from Soce836 (*Sorangium cellulosum*) (d) hit pathway from MCy9101 (*Cystobacter unclassified*)

An overview of the pathways found in the available myxobacterial genomes which are plausibly similar to characterized myxobacterial pathways can be seen in Tables 3.13 and 3.14.

	Genomes	Total pathways	Average pathways per genome	#Characterized pathways	NRPS	PKS	Hybrid NRPS PKS	Trans AT-PKS
Myxobacteria	71	1347	18.9	42	6	7	18	6
Cystobacterineae	40	889 (66%)	22.2	15 (1.7%)	2	4	8	1
Sorangineae	27	419 (31.1%)	15.5	19 (4.5%)	1	3	10	5
Nannocystineae	4	39 (2.9%)	9.75	1 (4%)	1			
Shared between Cystobacterineae and Sorangineae	67			2	2			

Table 3.13: Overview table for characterized pathways types distribution

Among the characterized pathways, suborder specific secondary metabolite pathways occurred in 40 cystobacterineae at following frequency such as Myxochromide (in 10 genomes) (Figure 3.68), DKxanthene (in 10 genomes), Myxalamid (in 9 genomes), Myxothiazol (in 7 genomes), Althiomycin (in 4 genomes), Aurafuron (in 3 genomes), Rhizopodin (in 3 genomes), Cystobactamide (in 3 genomes), Myxovirescin (in 2 genomes), Myxoprincomide (in 2 genomes), Pyxidienon (in 1 genome), Phenalamid (in 1 genome), Stigmatellin (in 1 genome), Melithiazol (in 1 genome) and Myxoalargin (in 1 genome). Other characterized pathways, suborder specific secondary metabolites occurred in 27 sorangineae at following frequency such as Lipothiazole (in 3 genomes) (Figure 3.70), Epothilon (in 2 genomes), Microsclerodermin (in 1 genome), Disorazol (in 1 genome), Etnangien (in 1 genome), Chondrochloren (in 1 genome), Carolacton (in 1 genome), Chivosazol (in 2 genomes), Chondramide (in 2 genome), Ajudazol (in 1 genome), Ripostatin (in 1 genome), Ambruticin (in 1 genome), Thuggacin (in 1 genomes), Leupyrrin (in 1 genome), Pellasoren (in 1 genome), Crocacin (in 1 genome), Crocapeptin (in 1 genome) and Sorangicin (in 1 genome) are observed in suborder Sorangineae. Nannochelin (in 1 genome) (Figure 3.69) was only observed in Nannocystineae suborder.

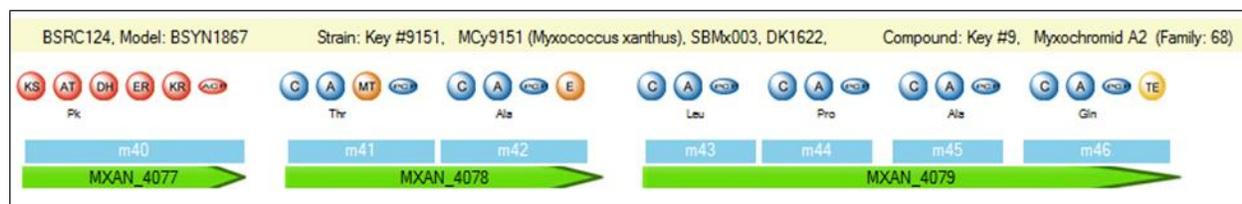


Figure 3.68: Ball scheme representation of Myxochromide, a suborder specific gene cluster found in 10 Cystobacterineae genomes

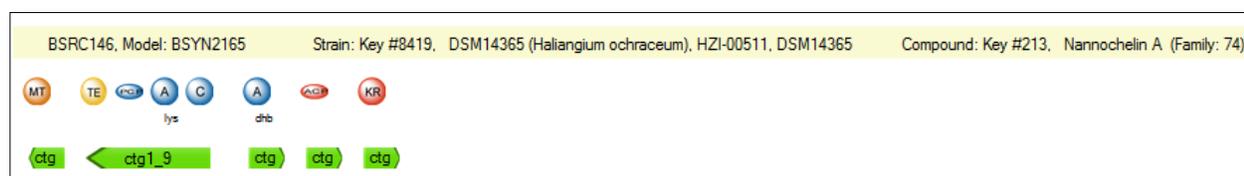


Figure 3.69: Ball scheme representation of Nannochelin, a suborder specific gene cluster found exclusively in Nannocystineae genome

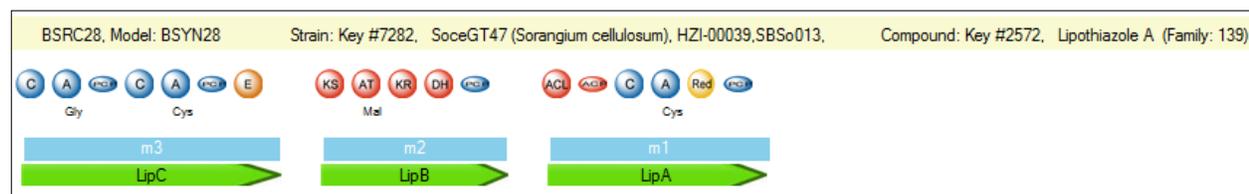


Figure 3.70: Ball scheme representation of Lipothiazole A, a suborder specific gene cluster found in 3 Sorangiineae genomes.

The pathways appearing in multiple suborders Cystobacterineae and Sorangiineae are Myxochelin (42) (Figure 3.71) and Tubulyisin (3) (Figure 3.67); regarding the latter see comment above

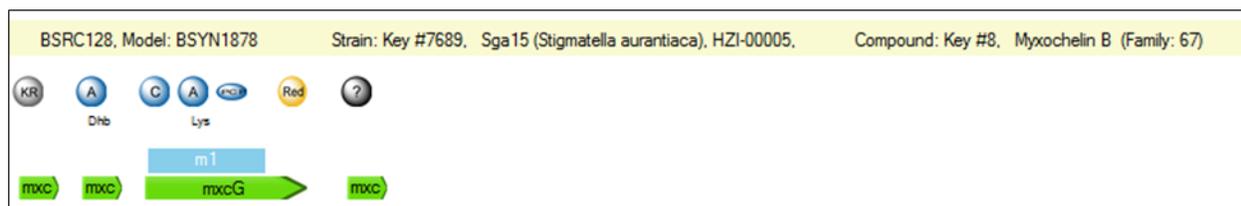
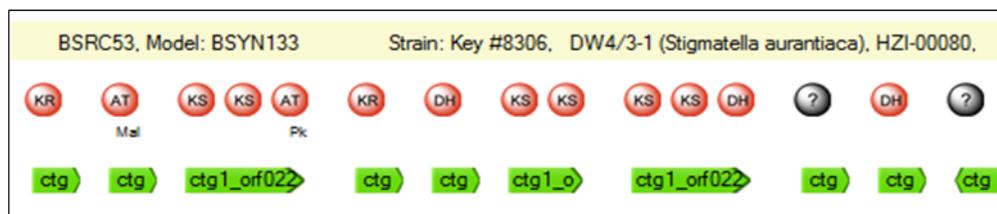


Figure 3.71: Ball scheme representation of Myxochelin, the most common characterized pathway appeared in 48 42 myxo genomes, both in Cystobacterineae (35) and Sorangiineae(7) but not on Nannocystineae.

Due to the bad prediction of pathway from DW4/3 genome, the hit pathway expected for Dawenol was not found in the results set (Figure 3.72). Note that the characterized pathways (marked with * in Table 3.14) that are not found by the algorithm are missing simply because of the current unavailability of genome sequences with sufficient quality to be included in the analysis.



(a)



(b)

Figure 3.72: Dawenol pathways from DW4/3 genome. (a) Curated pathway (b) predicted pathway from antiSMASH

As a special case, the Soraphen gene cluster produced three more unspecific hits along with the specific hit. This is due to high similarity in domain composition by combining domains from genes (Figure 3.73). These pathways actually lack precise domain architecture compared to the query Soraphen pathway. Since, the algorithm performs matching based on the set intersection, the unspecific hits were scored high with the parameter settings used (substrate specificity (0.3), additional domain penalty (0.1) and collinearity (0)). The latter was chosen because of the presence of many draft genomes in the test dataset where the information is scattered across up to 200 scaffolds. However, with the increase in the

Etnangien	1	Sorangineae	trans-AT PKS
Leupyrrin	1	Sorangineae	nrps-t1pks
Pellasoren	1	Sorangineae	nrps-t1pks
Ripostatin	1	Sorangineae	t1pks-t2pks
Sorangicin	1	Sorangineae	transatpks
Thuggacin	1	Sorangineae	t1pks-nrps
Microsclerdermin	1	Sorangineae	t1pks-nrps
Soraphen A(***)	1	Sorangineae	t1pks
Myxoprincomide	2	Cystobacterineae	NRPS-PKS
Myxovirescin	2	Cystobacterineae	PKS-NRPS
Chivosazol	2	Sorangineae	nrps-t1pks
Chondramide	2	Sorangineae	t1pks-nrps
Epothilon	2	Sorangineae	t2pks-transatpks
Tubulysin	3	Cystobacterineae and Sorangineae	nrps-t1pks
Cystobactamide	3	Cystobacterineae	nrps
Rhizopodin	3	Cystobacterineae	transatpks-nrps
Aurafuron	3	Cystobacterineae	t1pks
Lipothiazole	3	Sorangineae	nrps-t1pks
Althiomycin	4	Cystobacterineae	t1pks-nrps
Myxothiazol	7	Cystobacterineae	t1pks
Myxalamid	9	Cystobacterineae	nrps-t1pks
Myxochromide	10	Cystobacterineae	NRPS-PKS
DKxanthene	10	Cystobacterineae	PKS-NRPS
Myxochelin	42	Cystobacterineae and Sorangineae	nrps

Table 3.14: Distribution of characterized pathways between suborders. (*) these pathways belong to metabolites produced by strains for which genome data was not present in the database, (**) the hit pathway expected for Dawenol has missing several domains and mispredicted domains due to which the pathway could not be found. (***) increase in parameter values removed unspecific hits from Soraphen which appeared because of their high similarity in domain composition but lack similarity in domain architecture.

For most of the genome-sequenced myxobacterial strains the Myxobase contains mass spectrometric evidence from which production of the metabolites connected to the gene clusters in its genome can be confirmed (or disproved). As an example, strains revealed to contain a Rhizopodin gene cluster based on the present analysis (Table 3.14) could be shown subsequently to produce this metabolite based on high-resolution LC-MS measurements. Similarly, existing LC-MS data (not shown here; information from Myxobase and personal communication, Daniel Krug) underpin the production of suborder-specific compounds by at least one representative of the respective suborder, so that we can regard the above analysis of non-overlapping pathways as a realistic picture.

However, specific gene clusters were also found in strains which have not been known as producers of the respective candidate compound to date. In such cases the absence of compounds in LC-MS results could be because of the strain not expressed the genes for the pathway under laboratory conditions or because of degradation in the cultivating media. The cluster may also be defective, such as in the case of tubulysin (see above), the only other overlap seen between Sorangiineae and Cystobacterineae in this study besides myxochelin. This low overlap regarding known characterized pathways should be seen in light of only 70 currently available myxobacterial genomes.

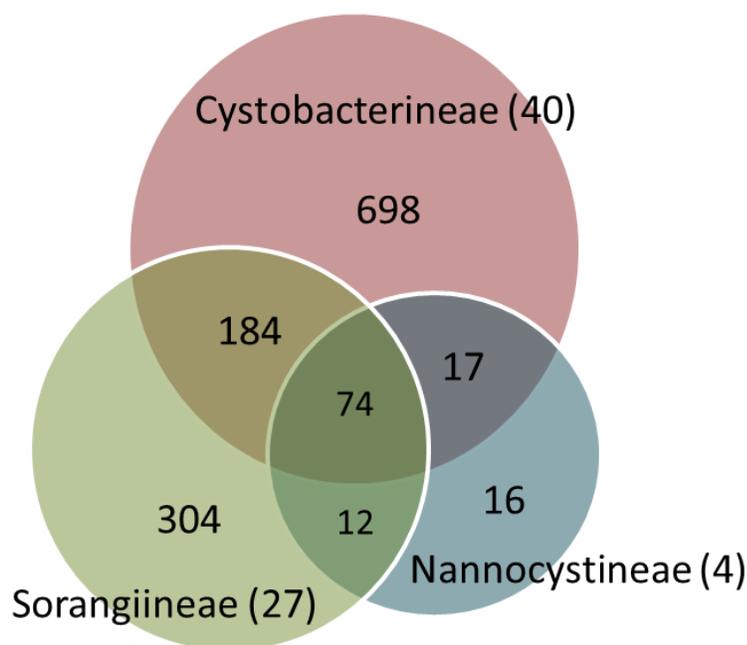


Figure 3.74: A Venn diagram showing the distribution of uncharacterized gene clusters from suborder Cystobacterineae (red), Sorangiineae (green) and Nannocystineae (blue).

On analysing the distribution of uncharacterized biosynthetic gene clusters from all the available genomes (Figure 3.74), resulted in suborder specific gene clusters belonging to Cystobacterineae (698) encoding gene clusters for NRPS (217), PKS (63), NRPS-PKS (295), Trans-AT-PKS (32), Trans-AT-PKS- NRPS (7) and Others (95) pathways. Similarly, Sorangiineae (304) contains gene clusters encoding for NRPS (80), PKS (59), NRPS-PKS (96), Trans-AT-PKS (8), NRPS-Trans-AT-PKS (8) and others (53) pathways. Nannocystineae (16) contains very few clusters encoding gene clusters for NRPS (4), PKS (3), NRPS-PKS (8) and Trans-AT-PKS (1) pathways (Table 3.15).

Several of these uncharacterized gene clusters are shared between suborders. There are 184 clusters belonging to both Cystobacterineae and Sorangiineae encoding gene clusters for NRPS (60), PKS (37), NRPS-PKS (34), Trans-AT-PKS (1) and others (52). There are fourteen gene clusters that belong to Cystobacterineae and Nannocystineae encoding gene clusters for NRPS-PKS (13), PKS (9) and others (1). There are four gene clusters belonged to Nannocystineae and Sorangiineae, all encoding gene clusters for NRPS-PKS (5), PKS (1) and others (6) pathways. Among all, there are only seventy four pathways observed in all the three suborders which potentially comprise the gene clusters for the “myxobacterial core secondary metabolome”, encoding gene clusters for NRPS (4), PKS (5), NRPS-PKS (19), and Others (46).

	Genomes	Total pathways	Average pathway per genome	Number of Uncharacterized pathways	Average uncharacterized clusters per genome	NRPS	PKS	Hybrid NRPS PKS	Trans AT-PKS
Myxobacteria	71	1347	18.9	1305	18.3				
Cystobacterineae	40	889 (66%)	22.2	698	17.45	217	63	295	28
Sorangiiineae	27	419 (31.1%)	15.5	304	11.25	80	59	96	8
Nannocystineae	4	39 (2.9%)	9.75	16	4	4	3	8	3
Shared between Cystobacterineae and Sorangiiineae	67	1317		184	2.7	60	37	34	1
Cystobacterineae and Nannocystineae	44	912		17	0.38	3		13	
Nannocystineae and Sorangiiineae	31	449		12	0.39		5	6	
Cystobacterineae, Sorangiiineae and Nannocystineae	71	1347		74	1	4	19	48	

Table 3.15: Overview table for uncharacterized pathways types distribution

An overview distribution of combined characterized and uncharacterized pathways across all genomes is summarized in a venn diagram shown in Figure 3.75. The picture is essentially similar to the diagram drawn for uncharacterized pathways due to the overwhelming prevalence of the latter.

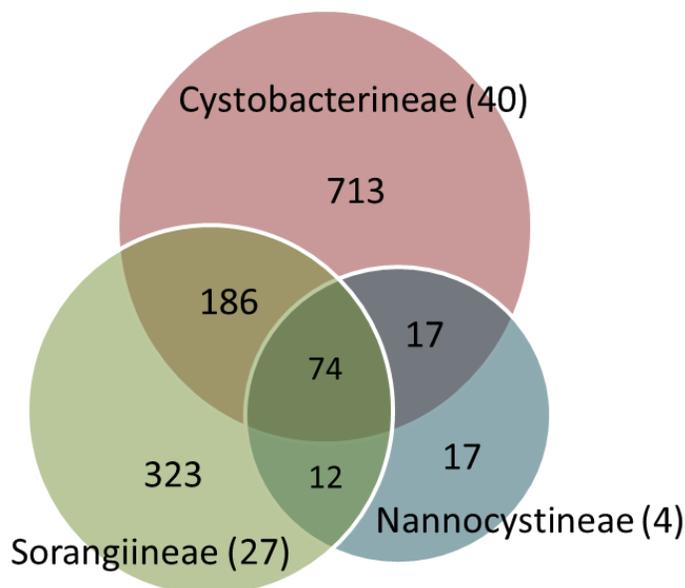


Figure 3.75: A Venn diagram showing the distribution of characterized and uncharacterized gene clusters from suborder Cystobacterineae (red), Sorangiineae (green) and Nannocystineae (blue).

Myxobacterial strains are well known for producing compounds with structural complexity and with diverse biological activities. However, there are plenty of natural products which are unexplored, which might be because of genes that are not expressed under laboratory conditions or simply because the metabolites have evaded their analytical detection to date. More attention should be put on previously unrecognised secondary metabolite pathways in the genomes and in extracting the end products that the unknown metabolic gene clusters could offer. In this analysis, using conceptual genome mining tool we were able to get a preliminary results on the pathways which are not yet characterized. Moreover, this approach allows to classify the uncharacterized pathway models which are rare (unique), common (2 to 5 occurrences) and rather frequent (>5 occurrences), as listed in (Table 3.16).

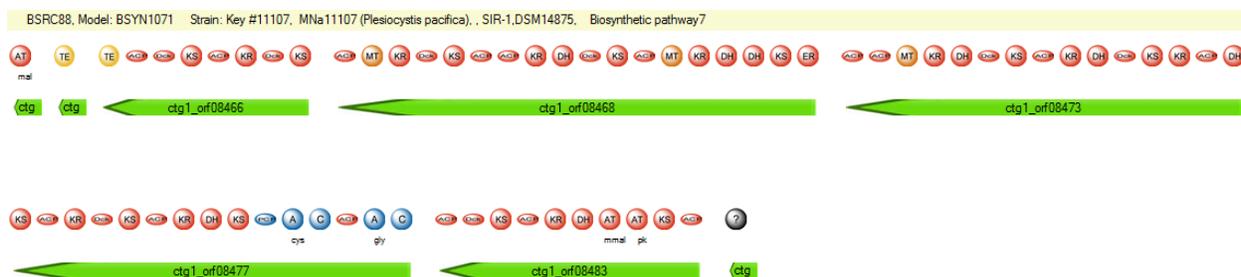


Figure 3.78: An example of the rare pathways among Nannocystinea was from MNa11107 (*Plesiocystis pacifica*) with 69 domains.

An example of common occurring pathway from strain MMy9151 (*Myxococcus xanthus*) with 24 domains belonged to suborder Cystobacterineae (Figure 3.79) appeared for five times.



Figure 3.79: An example of the commonly occurring pathway among Cystobacterineae was from MMy9151 (*Myxococcus xanthus*) with 24 domains was found 5 genomes

A pathway from strain MSr9337 (*Aetherobacter fasciculatus*) with 65 domains belonged to Sorangiineae (Figure 3.80) appeared for four times. Given the current knowledge these pathways reported were uncharacterized.

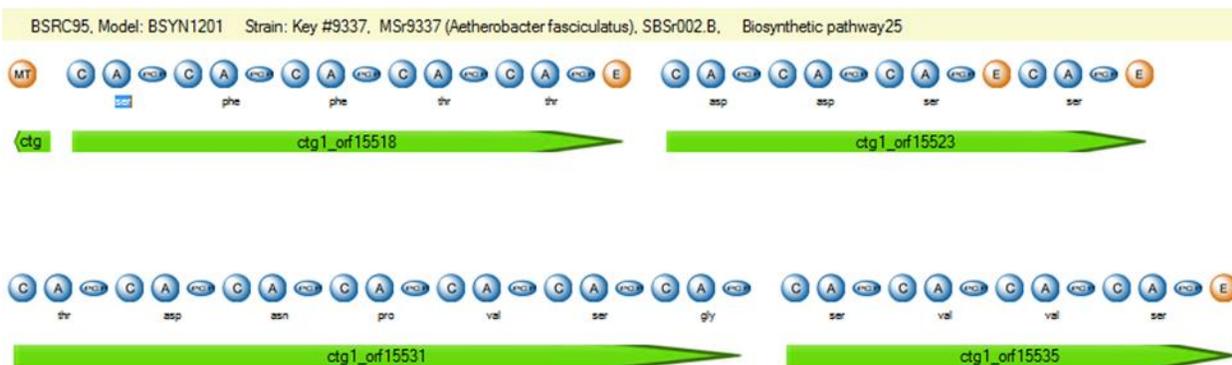


Figure 3.80: An example of the common pathways among Sorangiineae was from MSr9337 (*Aetherobacter fasciculatus*) with 65 domains was found 4 genomes

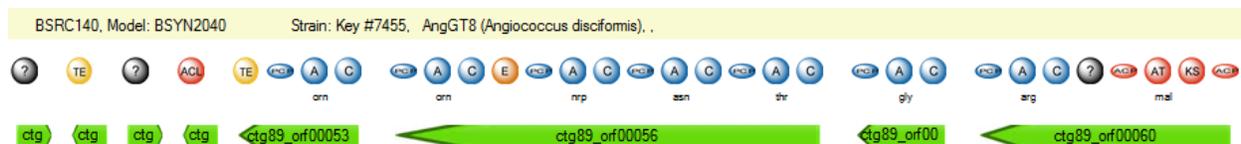


Figure 3.81: An example of the frequently occurring pathway among *Cystobacterineae* was from AngGT8 (*Angiococcus disciformis*) with 32 domains, it was found in 16 genomes.

The most frequent occurring pathways reported among the suborders among *Cystobacterineae* was from strain AngGT8 (*Angiococcus disciformis*) with 32 domains (Figure 3.82), among *Sorangiiineae* was from strain SoceGT47 (*Sorangium cellulosum*) with 32 domains (Figure 3.83). Due to the limited abundance of genomes, there was no frequent pathway which is exclusively specific to *Nannocystineae* reported.



Figure 3.82: An example of the frequent occurring most common pathway among *Sorangiiineae* was from SoceGT47 (*Sorangium cellulosum*) with 32 domains, it was found in 9 genomes.

This analysis suggests that there are unexplored pathways abundant in myxobacterial strains. Prospects to identify the expected novel metabolites and to assign them to their corresponding biosynthetic pathways are best for those pathways where monomer cooperation can be reliably predicted, i.e. especially for NRPS systems with presumed incorporation of specific amino acids. To name only some approaches, feeding studies could be conducted to identify the new compounds as stably isotope-labelled products using mass spectrometry (135), and recently developed strategies like peptidogenomics could be applied (136). Where genetic manipulation of strains is feasible, targeted gene inactivation (“knockout”) studies with subsequent statistical evaluation of production profiles can be performed (85) in order to identify novel natural products. Obstacles for identification may arise – among other possibilities - from the production of these metabolites being dependent on certain cultivation conditions and media ingredients or being subject to regulation, making their detection under laboratory conditions a challenge. In that case heterologous expression in alternative host strains or synthetic biology approaches may offer a solution, although this comes with its own specific challenges including possibly non-recognized promoters, complicated operon structures, precursor supply bottlenecks or problems with codon adaptation (137).

Towards the estimation of myxobacterial biosynthetic pathway richness

Estimating the pathways richness is a persistent challenge which is interesting to the natural product research field in general. In the following, an approach to pathway richness estimation for myxobacteria is evaluated with the results obtained from conceptual genome mining. For estimating the richness of distinct pathways, 71 genomes of Cystobacterineae (40) and Sorangiineae (27) and Nannocystinae (4) were used in this study, containing a total of 1347 pathways of which 783 pathways are regarded as distinct models (complete distribution matrix shown in fold-out Appendix NN). Among the 1347 pathways, 419 pathways belonged to Sorangiineae of which 264 distinct pathways are reported by this analysis. Similarly, there are 889 pathways belonging to strains of suborder Cystobacterineae of which 492 distinct pathways are reported (Table 3.16). The software EstimateS was used to analyze these data with an subsequent extrapolation of rarefaction curves; classic formula of Chao1 and Chao2 was chosen to compute the data (122). The rarefaction curves were visualized using R based on the results obtained from EstimateS tool (Figure 3.83). The pathway Richness estimation (y-axis) is plotted versus the number of genome samples (x-axis). This allows to compare rarefaction curves from suborders Cystobacterineae, Sorangiineae and combination of suborders Cystobacterineae and Sorangiineae (C+S). The rarefaction curve for the distinct pathways by suborders can be seen in Figure 3.83 and separate rarefaction curves for the (currently prevalent) genera *Sorangium* and *Myxococcus* are shown in Figure 3.84. Each of the rarefaction curves represents the rate of discovering new pathways as sampling size increases. The steep slope of rarefaction curve for Sorangiineae, Cystobacterineae and C+S indicates that the census is far from complete, when considering the currently sequenced genomes. Extrapolating each of the suborder populations individually, and calculating in addition also the combined suborders, reveals that slopes remain significantly positive when sampling was projected forward to the sequencing of 142 genomes, twice the amount currently available (Figure 3.83). The extrapolated rarefaction curve shows that, a total of 1050 distinct pathways could be estimated from those 142 myxobacterial genomes. Considering that only around 150 structurally distinct classes of secondary metabolites have been characterized from myxobacteria so far (according to Myxobase; internal statistics from Institute), these numbers support the notion of myxobacteria as a still underexploited resource for the discovery of novel natural products.

Here, it should be emphasized that the genome samples for Sorangiineae are currently majorly contributed from genus *Sorangium* (13 out of 27) with a total of 151 distinct pathways. Similarly, the suborder Cystobacterineae is dominated by *Myxococcus* species (16 out of 40) with a total of 309 distinct pathways (Figure 3.84). Several available genomes of suborder Cystobacterineae were not yet classified at genus level but belong to family Myxococcaceae. Since *Myxococcus* is a well-represented member of

family Myxococcaceae, we assumed in the following that the genomes which were unclassified under this family could be hypothetically *Myxococcus species*, too. Therefore, rarefaction curves were calculated separately for genera *Sorangium* and *Myxococcus* (including the alleged *Myxococcaceae members*), and it can be contended from the flattening-out of plots in Figure 3.84 that the discovery rate of new pathways in the two genera *Sorangium* and *Myxococcus* is limited. Importantly, addition of genomes from other genera is apparently able to increase the slope of the rarefaction curve significantly (compare Figure 3.83). Future isolation and sequencing efforts should be soon able to make sufficient genomes available to assay this anticipated effect quantitatively. Altogether, despite the currently still restricted sample size of myxobacterial genomes, this analysis aided to obtain a richness estimation of the pathways within and across suborders. It should be noted that the present analysis is more likely to under- than to over-estimate pathway richness, because the heterogeneous quality of genomes involved in the study required to compromise during concepts-based pathway comparison. Search parameters such as pathway collinearity and matching substrate specificity were set to rather relaxed values to compensate for possible prediction gaps in pathways retrieved from multi-scaffold genomes. Thus, in this analysis pathways may have been grouped together to form one distinct model although they might actually represent individual models. Since the overall sample size (71 genomes) is not extraordinarily large, such effects could have noticeable impact. Nevertheless, the approach taken here can be further developed into a blueprint workflow for estimating multimodular pathway richness, to be executed again as fresh and improved whole-genome sequence information becomes available from diverse myxobacterial taxa.

	Genomes	Total count of pathways	Distinct pathway models	Avg. number of distinct models per genome	Uncharacterized distinct models (may appear once or several times)	Unique models (singletons, appear only once)
Myxobacteria	71	1347	783 (58%)	11		
Cystobacterineae	40	889 (66%)	492 (55.3%)	12.3	475 (53.4 %)	261 (29.3%)
Sorangineae	27	419 (31.1%)	264 (63.3%)	9.8	245 (58.4%)	162 (38.6%)
Nannocystineae	4	39 (2.9%)	27 (69%)	6.7	26 (66%)	17 (43.5%)

Table 3.16: An overview of pathway analysis among the suborders of Cystobacterineae, Sorangineae and Nannocystineae reported using BiosynML algorithm

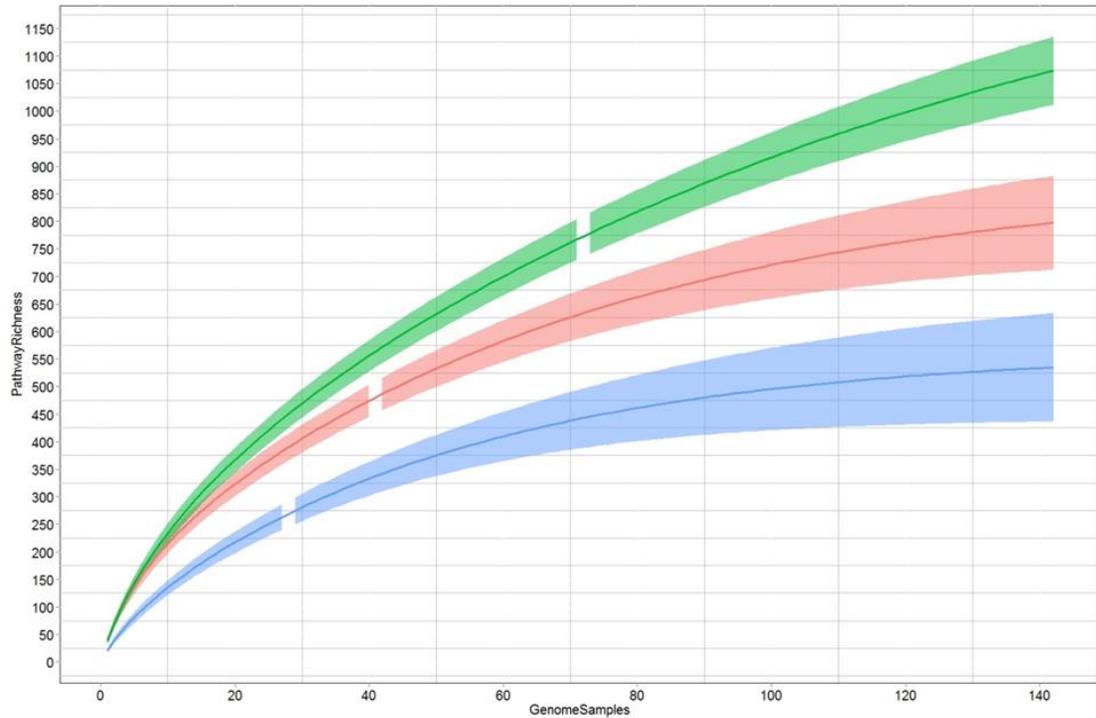


Figure 3.83: Estimation of distinct pathway richness among suborders *Cystobacterineae* (red), *Sorangiineae* (blue) and combination of both (C+S) (green). The gap in each rarefaction curve is separation between experimental (left) and extrapolated curves (to the right)

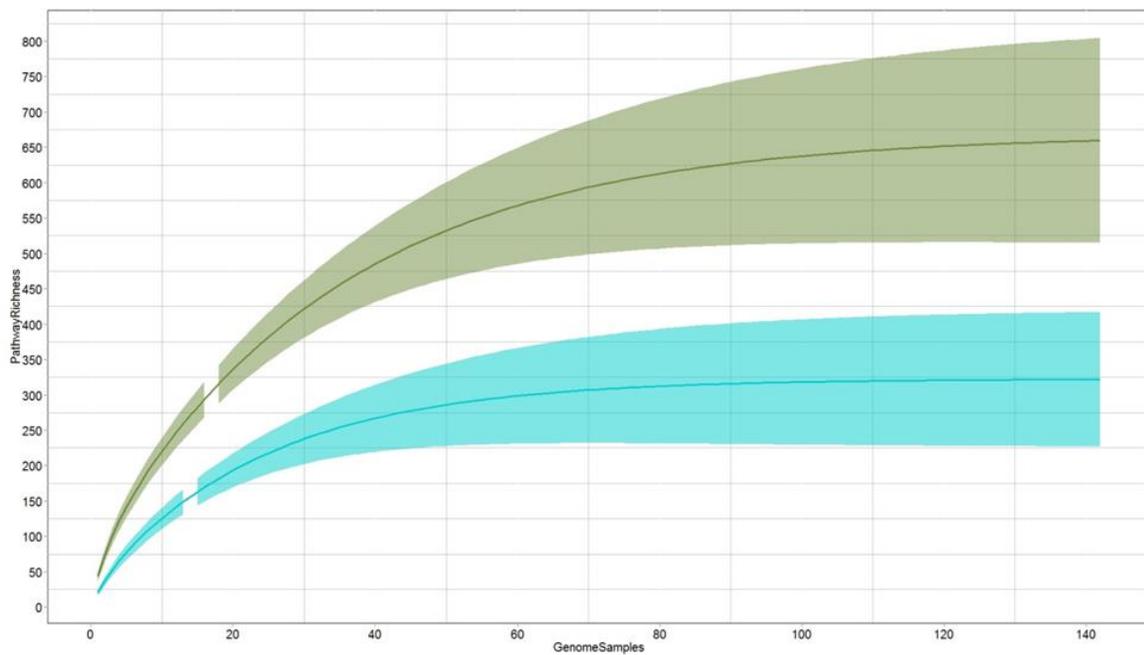


Figure 3.84: Estimation of distinct pathway richness among suborders *Myxococcus* (olivegreen) and *Sorangium* (cyan). The gap in each rarefaction curve is separation between experimental (left) and extrapolated curves (to the right)

4 Conclusion and outlook

In this work a bioinformatic analysis framework summarized as “BiosynML toolkit” was developed, where the most important component is a searching and matching engine for the identification and conceptual comparison of multimodular biosynthetic pathways in microbial genomes. Contrary to existing genome-mining approaches, the methods devised here rely on previously annotated sequences and pathways are matched on the basis of architectural similarity, thereby taking into account overall domain composition of pathways, the available functional annotations for domains, their relative arrangement, alignment between subsets of domains and operon organization. The underlying rationale is that these “enzymatic collectives” together constitute the crucial determinants for the small-molecule structure produced by a given pathway, and thus it was anticipated that it should be possible to specifically recognize and compare pathways by their evolutionarily established domain pattern (or “concept”). Different algorithms were tested and the “global best subset” (GBS) approach was empirically validated to provide best analytical power and robustness for this purpose.

On the one hand the conceptual mining approach can achieve similar results like conventional sequence-base genome mining, a congruence that is understandable because the traditional procedures working with sequence data usually employ aligning biosynthetic signature domains, as the protein sequences of those domains are conserved. On the other hand, the BiosynML approach is less dependent on which part of a pathway is used as input for comparison, and can even accept hypothetical assembly lines (“sequence-free”, i.e. manually constructed by natural product chemists on the basis of retro-biosynthetic considerations) as input. The possibility to set several parameters for the scoring of results allows fine-tuning the BiosynML comparison tool for a range of applications. Examples in this study show that selectivity can be achieved even when using very short biosynthetic assembly lines (such as the myxochelin pathway) with a small number of domains as input, but also with larger pathways showing little diversified domain composition (such as the myxoprincomide pathway, a giant NRPS assembly line). Suitable parameters for running the analysis are in practice readily found for varying research interests, i.e. depending on whether only high-ranking near-exact matches are desired or remotely similar gene cluster architectures should also be reported. As a consequence of increased parameter tolerance hits will be included showing significant deviations to the query pathway which could be meaningful in a biosynthetic sense, but at the cost of expecting an increased false discovery rate. It is especially encouraging that the false discovery rate using the conceptual genome mining

approach was very low in this study when it was applied to the recognition of known biosynthetic pathways in myxobacterial whole-genome sequences.

It is clear that biosynthetic gene cluster prediction, a prerequisite for using the BiosynML approach, sets a limit for sensitivity, because certain domains or even entire genes could be missed by the upstream gene-finding and annotation pipeline. Moreover, the whole-genome assembly process may not always yield closed genomes, whereas every gap comes with the risk of partially losing information for a biosynthetic gene cluster. However, in both fields – whole-genome sequencing and assembly as well as bioinformatics prediction of pathways – advancements have been recently made (and are continuously being made) which underpin the expectation that genome sequence quality and pathway prediction quality will be less and less of an obstacle for using the BiosynML genome mining approach in the future.

The sequencing and characterization of novel myxobacterial strains and biosynthetic pathways is an on-going process. Updates on pathway identification and their assignment to compounds need to be carried out as soon as new knowledge becomes available and will contribute in turn to make the picture for newly annotated genomes more complete. Therefore, the BiosynML module developed in this work has been tightly integrated into the in-house database system Myxobase to support everyday efforts of scientists working on the discovery of new myxobacterial natural products. Usage of the tool is facilitated by the implementation of BiosynML import and annotation functions within the Geneious plugin and through the addition of BiosynML export functionality to the antiSMASH system. The stringent client-server design of the BiosynML core engine (implemented as remote procedure framework) allows for its re-use in other projects, whereas the perhaps most foreseeable application could be its interfacing with the MIBiG pathway repository which is currently being built through a community initiative. Future work on the BiosynML engine could incorporate more options for the matching function which could help to further fine-tune the search, e.g. by introducing a fundamental differentiation between scaffold—generating and tailoring domains as it is done by the MIBiG standard. Also the algorithm GBS is regarded as computationally expensive; implementing it with multi thread capability should speed up the running time.

The BiosynML methods were ultimately used in this study to generate for the first time an overview of myxobacterial biosynthetic pathway diversity, covering NRPS, PKS and hybrid pathways from all myxobacterial genomes currently available at the institute (with sufficient quality in terms of contig/scaffold numbers). In light of 71 genomes used for the study - with an uneven distribution across the three myxobacterial suborders and vastly unbalanced coverage of the 23 known genera (many genomes belonged to *Myxococcus* and *Sorangium*) - results must be regarded as preliminary. Altogether

1347 biosynthetic gene clusters from all genomes were involved in the analysis of which 783 were classified by the BiosynML comparison to represent distinct pathway architectures. Rarefaction curves were calculated and project significant potential for the discovery of novel pathways in the upcoming genomes to be sequenced. However, a more complete appreciation of myxobacterial pathway richness will have to await the availability of more genome sequences from more diverse species. It should be particularly interesting to see how isolates from hitherto underexploited habitats, belonging to novel genera and families and thus likely to constitute sources of additional (bio) chemical diversity, are able to contribute to myxobacterial pathway richness.

5 Literature Cited

1. Butler MS. The role of natural product chemistry in drug discovery. *Journal of natural products* 2004; 67(12):2141–53.
2. Robert Cruickshank. Sir Alexander Fleming. *Journal of Clinical Pathology* 1955; 8(4):355–6.
3. Jesse W.-H. Li and John C. Vederas. Drug discovery and natural products: end of an era or an endless frontier? *SCIENCE* 2009; 325:161–5.
4. Mishra BB, Tiwari VK. Natural products: an evolving role in future drug discovery. *European journal of medicinal chemistry* 2011; 46(10):4769–807.
5. Menche D, Arikian F, Perlova O, Horstmann N, Ahlbrecht W, Wenzel SC et al. Stereochemical determination and complex biosynthetic assembly of etnangien, a highly potent RNA polymerase inhibitor from the myxobacterium *Sorangium cellulosum*. *Journal of the American Chemical Society* 2008; 130(43):14234–43.
6. Gaitatzis N, Silakowski B, Kunze B, Nordsiek G, Blöcker H, Höfle G et al. The biosynthesis of the aromatic myxobacterial electron transport inhibitor stigmatellin is directed by a novel type of modular polyketide synthase. *The Journal of biological chemistry* 2002; 277(15):13082–90.
7. Erol O, Schäberle TF, Schmitz A, Rachid S, Gurgui C, El Omari M et al. Biosynthesis of the myxobacterial antibiotic coralopyronin A. *Chembiochem : a European journal of chemical biology* 2010; 11(9):1253–65.
8. James Staunton. Biosynthesis of Erythromycin and Rapamycin. *Chemical reviews* 1997; 97:2611–29.
9. Mulzer J, Altmann K, Höfle G, Müller R, Prantz K. Epothilones – A fascinating family of microtubule stabilizing antitumor agents. *Comptes Rendus Chimie* 2008; 11(11-12):1336–68.
10. Thomas ES, Gomez HL, Li RK, Chung H, Fein LE, Chan VF et al. Ixabepilone plus capecitabine for metastatic breast cancer progressing after anthracycline and taxane treatment. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 2007; 25(33):5210–7.
11. Tacar O, Sriamornsak P, Dass CR. Doxorubicin: an update on anticancer molecular action, toxicity and novel drug delivery systems. *The Journal of pharmacy and pharmacology* 2013; 65(2):157–70.
12. Parker C, Waters R, Leighton C, Hancock J, Sutton R, Moorman AV et al. Effect of mitoxantrone on outcome of children with first relapse of acute lymphoblastic leukaemia (ALL R3): an open-label randomised trial. *The Lancet* 2010; 376(9757):2009–17.

-
13. Xue Q, Ashley G, Hutchinson CR, Santi DV. A multiplasmid approach to preparing large libraries of polyketides. *P. Natl. Acad. Sci. USA (Proceedings of the National Academy of Sciences of the United States of America)* 1999; 96(21):11740–5.
 14. CHARLES FLETCHER. First clinical use of penicillin. *BRITISH MEDICAL JOURNAL* 1984; 289:22–9.
 15. János Bérdy. Bioactive Microbial Metabolites. *JOURNAL OF ANTIBIOTICS* 2004; 58:1–26.
 16. Walsh CT, Wencewicz TA. Prospects for new antibiotics: a molecule-centered perspective. *The Journal of antibiotics* 2014; 67(1):7–22.
 17. Walsh. C. Where will new antibiotics come. *NATURE REVIEWS* 2003:65–70.
 18. Auerbach D, Thaminy S, Hottiger MO, Stagljar I. The post-genomic era of interactive proteomics: Facts and perspectives. *Proteomics* 2002; 2:611–23.
 19. Boucher HW, Talbot GH, Bradley JS, Edwards JE, Gilbert D, Rice LB et al. Bad bugs, no drugs: no ESKAPE! An update from the Infectious Diseases Society of America. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* 2009; 48(1):1–12.
 20. Donadio S, Monciardini P, Sosio M. Polyketide synthases and nonribosomal peptide synthetases: the emerging view from bacterial genomics. *Natural product reports* 2007; 24(5):1073–109.
 21. Staunton J, Weissman KJ. Polyketide biosynthesis: a millennium review. *Natural product reports* 2001; 18(4):380–416.
 22. Marahiel MA. Working outside the protein-synthesis rules: insights into non-ribosomal peptide synthesis. *Journal of peptide science : an official publication of the European Peptide Society* 2009; 15(12):799–807.
 23. Chaitan Khosla, Rajesh S. Gokhale, John R. Jacobsen, David E. Cane. TOLERANCE AND SPECIFICITY OF POLYKETIDE SYNTHASES. *Annual Review of Biochemistry* 1999; 68:219–53.
 24. Jenke-Kodama H, Sandmann A, Müller R, Dittmann E. Evolutionary implications of bacterial polyketide synthases. *Molecular biology and evolution* 2005; 22(10):2027–39.
 25. Shen B. Polyketide biosynthesis beyond the type I, II and III polyketide synthase paradigms. *Current Opinion in Chemical Biology* 2003; 7(2):285–95.
 26. Hertweck C, Luzhetskyy A, Rebets Y, Bechthold A. Type II polyketide synthases: gaining a deeper insight into enzymatic teamwork. *Natural product reports* 2007; 24(1):162–90.
 27. Wawrik B, Kerkhof L, Zylstra GJ, Kukor JJ. Identification of Unique Type II Polyketide Synthase Genes in Soil. *Applied and Environmental Microbiology* 2005; 71(5):2232–8.
 28. Shen B. Biosynthesis of aromatic polyketides. *Top. Curr. Chem. (Topics in Current Chemistry)* 2000; 209:1–51.

-
29. Yu D, Xu F, Zeng J, Zhan J. Type III polyketide synthases in natural product biosynthesis. *IUBMB life* 2012; 64(4):285–95.
30. Moore BS, Hertweck C, Hopke JN, Izumikawa M, Kalaitzis JA, Nilsen G et al. Plant-like biosynthetic pathways in bacteria: from benzoic acid to chalcone. *Journal of natural products* 2002; 65(12):1956–62.
31. Meier JL, Burkart MD. The chemical biology of modular biosynthetic enzymes. *Chemical Society reviews* 2009; 38(7):2012–45.
32. Moore BS, Höpke JN. Discovery of a New Bacterial Polyketide Biosynthetic Pathway. *ChemBioChem (ChemBioChem)* 2000; 2(1):35–8.
33. Crosby J, Crump MP. The structural role of the carrier protein--active controller or passive carrier. *Natural product reports* 2012; 29(10):1111–37.
34. Cheng Y, Coughlin JM, Lim S, Shen B. Chapter 8 Type I Polyketide Synthases That Require Discrete Acyltransferases. In: Hopwood DA, editor. *Complex Enzymes in Microbial Natural Product Biosynthesis, Part B: Polyketides, Aminocoumarins and Carbohydrates*: Elsevier; 2009. p. 165–86 (Methods in Enzymology).
35. C. Yanofsky, B. C. Carlton, J. R. Guest, D. R. Helinski, and U. Henning. ON THE COLINEARITY OF GENE STRUCTURE AND PROTEIN STRUCTURE. *Proceedings of the National Academy of Sciences of the United States of America* 1964.
36. Jungmann K, Jansen R, Gerth K, Huch V, Krug D, Fenical W et al. Two of a Kind-The Biosynthetic Pathways of Chlorotoniol and Anthracimycin. *ACS chemical biology* 2015.
37. Piel J. Biosynthesis of polyketides by trans-AT polyketide synthases. *Natural product reports* 2010; 27(7):996–1047.
38. GOKHALE R, HUNZIKER D, CANE D, KHOSLA C. Mechanism and specificity of the terminal thioesterase domain from the erythromycin polyketide synthase. *Chemistry & Biology* 1999; 6(2):117–25.
39. Yolande A. Chan, Angela M. Podevels, Brian M. Kevany, and Michael G. Thomas. *Biosynthesis of Polyketide Synthase Extender Units* 2009.
40. Gyles C, Boerlin P. Horizontally transferred genetic elements and their role in pathogenesis of bacterial disease. *Veterinary pathology* 2014; 51(2):328–40.
41. Eugene V. Koonin, Kira S. Makarova, L. Aravind. *HORIZONTAL GENE TRANSFER IN PROKARYOTES: Quantification and Classification* 1.
42. Nguyen T, Ishida K, Jenke-Kodama H, Dittmann E, Gurgui C, Hochmuth T et al. Exploiting the mosaic structure of trans-acyltransferase polyketide synthases for natural product discovery and pathway dissection. *Nature biotechnology* 2008; 26(2):225–33.
43. Henning D, Mootz, Dirk Schwarzer, and Mohamed A. Marahiel. *Ways of Assembling Complex Natural Products on Modular Nonribosomal Peptide Synthetases* 2002.

-
44. Schwarzer D, Finking R, Marahiel MA. Nonribosomal peptides: from genes to products. *Natural product reports* 2003; 20(3):275.
45. Felnagle EA, Jackson EE, Chan YA, Podevels AM, Berti AD, McMahon MD et al. Nonribosomal peptide synthetases involved in the production of medically relevant natural products. *Molecular pharmaceutics* 2008; 5(2):191–211.
46. HENNING D. MOOTZ AND MOHAMED A. MARAHIEL. The tyrocidine biosynthesis operon of *Bacillus brevis* complete nucleotide sequence and biochemical characterization of functional internal adenylation domains 1997.
47. Torsten Stachelhaus, Henning D. Mootz, Veit Bergendahl, and Mohamed A. Marahiel. *Peptide Bond Formation in Nonribosomal Peptide Biosynthesis* 1998.
48. Stachelhaus T1, Hüser A, Marahiel MA. Biochemical characterization of peptidyl carrier protein (PCP), the thiolation domain of multifunctional peptide synthetases 1996.
49. Luis E. N. Quadri, Paul H. Weinreb, Ming Lei, Michiko M. Nakano, Peter Zuber, and Christopher T. Walsh. Characterization of Sfp, a *Bacillus subtilis* Phosphopantetheinyl Transferase for Peptidyl Carrier Protein Domains in Peptide Synthetases 1997.
50. LAMBALOT R, GEHRING A, FLUGEL R, ZUBER P, LACELLE M, MARAHIEL M et al. A new enzyme superfamily ? the phosphopantetheinyl transferases. *Chemistry & Biology* 1996; 3(11):923–36.
51. Walsh CT, Gehring AM, Weinreb PH, Quadri LEN, Flugel RS. Post-translational modification of polyketide and nonribosomal peptide synthetases. *Current Opinion in Chemical Biology* 1997; 1(3):309–15.
52. John W. Trauger*, Rahul M. Kohli*, Henning D. Mootz². Peptide cyclization catalysed by the thioesterase domain of tyrocidine synthetase 2000.
53. Kohli RM, Trauger JW, Schwarzer D, Marahiel MA, Walsh CT. Generality of Peptide Cyclization Catalyzed by Isolated Thioesterase Domains of Nonribosomal Peptide Synthetases †. *Biochemistry* 2001; 40(24):7099–108.
54. Christian P. Ridley, Ho Young Lee, and Chaitan Khosla. Evolution of polyketide synthetases in bacteria. *PNAS* 2008; 105(12):4595–600.
55. A. Laupacis, P. A. Keown, R. A. Ulan, N. McKenzie, and C. R. Stiller. Cyclosporin A: a powerful immunosuppressant. *Canadian Medical Association Journal* 1982; 126(9):1041–6.
56. Henken S, Bohling J, Martens-Lobenhoffer J, Paton JC, Ogunniyi AD, Briles DE et al. Efficacy profiles of daptomycin for treatment of invasive and noninvasive pulmonary infections with *Streptococcus pneumoniae*. *Antimicrobial agents and chemotherapy* 2010; 54(2):707–17.
57. P M Small and H F Chambers. Vancomycin for *Staphylococcus aureus* endocarditis in intravenous drug users. *1990; 34 (6):1227–31.*

-
58. Walsh CT. Insights into the chemical logic and enzymatic machinery of NRPS assembly lines. *Natural product reports* 2015.
59. Caboche S, Pupin M, Leclère V, Fontaine A, Jacques P, Kucherov G. NORINE: a database of nonribosomal peptides. *Nucleic acids research* 2008; 36(Database issue):D326-31.
60. Lautru S, Deeth RJ, Bailey LM, Challis GL. Discovery of a new peptide natural product by *Streptomyces coelicolor* genome mining. *Nature chemical biology* 2005; 1(5):265–9.
61. Park SR, Yoo YJ, Ban Y, Yoon YJ. Biosynthesis of rapamycin and its regulation: past achievements and recent progress. *The Journal of antibiotics* 2010; 63(8):434–41.
62. Liu F, Garneau S, Walsh CT. Hybrid nonribosomal peptide-polyketide interfaces in epothilone biosynthesis: minimal requirements at N and C termini of EpoB for elongation. *Chemistry & Biology* 2004; 11(11):1533–42.
63. Du L, Sánchez C, Chen M, Edwards DJ, Shen B. The biosynthetic gene cluster for the antitumor drug bleomycin from *Streptomyces verticillus* ATCC15003 supporting functional interactions between nonribosomal peptide synthetases and a polyketide synthase. *Chemistry & Biology* 2000; 7(8):623–42.
64. Du, L., et al. Review: Hybrid Peptide–Polyketide Natural Products: Biosynthesis and Prospects toward Engineering Novel Molecules. *Metabolic Engineering* 2001:78–95.
65. Meiser P, Bode HB, Müller R. The unique DKxanthene secondary metabolite family from the myxobacterium *Myxococcus xanthus* is required for developmental sporulation. *PNAS* 2006; 103:19128–33.
66. Omura S, Ikeda H, Ishikawa J, Hanamoto A, Takahashi C, Shinose M, Takahashi Y, Horikawa H, Nakazawa H, Osonoe T, Kikuchi H, Shiba T, Sakaki Y, Hattori M. 10, Genome sequence of an industrial microorganism 2001; 98(21).
67. Bentley SD, Chater KF, Cerdeño-Tárraga AM, Challis GL, Thomson NR, James KD, Harris DE, Quail MA, Kieser H, Harper D, Bateman A, Brown S, Chandra G, Chen CW, Collins M, Cronin A, Fraser A, Goble A, Hidalgo J, Hornsby T, Howarth S, Huang CH, Kieser T, Larke L, Murphy L, Oliver K, O'Neil S, Rabbinowitsch E, Rajandream MA, Rutherford K, Rutter S, Seeger K, Saunders D, Sharp S, Squares R, Squares S, Taylor K, Warren T, Wietzorrek A, Woodward J, Barrell BG, Parkhill J, Hopwood DA. Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). *Nature* 2002; 417:141–7.
68. Chris M. Farnet and Emmanuel Zazopoulos. *Improving Drug Discovery From Microorganisms*; 2005.
69. Matt Ridley. *The autobiography of a species in 23 chapters*.
70. R.Staden. A strategy of DNA sequencing employing computer programs 1979; 6(7).
71. F. SANGER, S. NICKLEN, AND A. R. COULSON. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA* 1977; 74(12):5463–7.

-
72. F. SANGER, G. M. AIR, B. G. BARRELL, N. L. BROWN, A. R. COULSON, J. C. FIDDES, C. A. HUTCHISON III, P. M. SLOCOMBE & M. SMITH. Nucleotide sequence of bacteriophage ϕ X174 DNA. *Nature* 1977; 265:687–95.
73. M. Susan Lindee. The origins of bioinformatics. *NATURE REVIEWS* 2000; 1:231–6.
74. Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, Connell CR, Heiner C, Kent SB, Hood LE. Fluorescence detection in automated DNA sequence analysis. 1986.
75. J. Craig Venter^{1,*}, Mark D. Adams¹, Eugene W. Myers¹, Peter W. Li¹, Richard J. Mural¹, Granger G. Sutton¹, Hamilton O. Smith¹, Mark Yandell¹, Cheryl A. Evans¹, Robert A. Holt¹, Jeannine D. Gocayne¹, Peter Amanatides¹, Richard M. Ballew¹, Daniel H. Huson¹, Jennifer Russo Wortman¹, Qing Zhang¹, Chinnappa D. Kodira¹, Xiangqun H. Zheng¹, Lin Chen¹, Marian Skupski¹, Gangadharan Subramanian¹, Paul D. Thomas¹, Jinghui Zhang¹, George L. Gabor Miklos², Catherine Nelson³, Samuel Broder¹, Andrew G. Clark⁴, Joe Nadeau⁵, Victor A. McKusick⁶, Norton Zinder⁷, Arnold J. Levine⁷, Richard J. Roberts⁸, Mel Simon⁹, Carolyn Slayman¹⁰, Michael Hunkapiller¹¹, Randall Bolanos¹, Arthur Delcher¹, Ian Dew¹, Daniel Fasulo¹, Michael Flanigan¹, Liliana Florea¹, Aaron Halpern¹, Sridhar Hannenhalli¹, Saul Kravitz¹, Samuel Levy¹, Clark Mobarry¹, Knut Reinert¹, Karin Remington¹, Jane Abu-Threideh¹, Ellen Beasley¹, Kendra Biddick¹, Vivien Bonazzi¹, Rhonda Brandon¹, Michele Cargill¹, Ishwar Chandramouliswaran¹, Rosane Charlab¹, Kabir Chaturvedi¹, Zuoming Deng¹, Valentina Di Francesco¹, Patrick Dunn¹, Karen Eilbeck¹, Carlos Evangelista¹, Andrei E. Gabrielian¹, Weiniu Gan¹, Wangmao Ge¹, Fangcheng Gong¹, Zhiping Gu¹, Ping Guan¹, Thomas J. Heiman¹, Maureen E. Higgins¹, Rui-Ru Ji¹, Zhaoxi Ke¹, Karen A. Ketchum¹, Zhongwu Lai¹, Yiding Lei¹, Zhenya Li¹, Jiayin Li¹, Yong Liang¹, Xiaoying Lin¹, Fu Lu¹, Gennady V. Merkulov¹, Natalia Milshina¹, Helen M. Moore¹, Ashwinikumar K Naik¹, Vaibhav A. Narayan¹, Beena Neelam¹, Deborah Nusskern¹, Douglas B. Rusch¹, Steven Salzberg¹², Wei Shao¹, Bixiong Shue¹, Jingtao Sun¹, Zhen Yuan Wang¹, Aihui Wang¹, Xin Wang¹, Jian Wang¹, Ming-Hui Wei¹, Ron Wides¹³, Chunlin Xiao¹, Chunhua Yan¹, Alison Yao¹, Jane Ye¹, Ming Zhan¹, Weiqing Zhang¹, Hongyu Zhang¹, Qi Zhao¹, Liansheng Zheng¹, Fei Zhong¹, Wenyan Zhong¹, Shiaoping C. Zhu¹, Shaying Zhao¹², Dennis Gilbert¹, Suzanna Baumhueter¹, Gene Spier¹, Christine Carter¹, Anibal Cravchik¹, Trevor Woodage¹, Feroze Ali¹, Huijin An¹, Aderonke Awe¹, Danita Baldwin¹, Holly Baden¹, Mary Barnstead¹, Ian Barrow¹, Karen Beeson¹, Dana Busam¹, Amy Carver¹, Angela Center¹, Ming Lai Cheng¹, Liz Curry¹, Steve Danaher¹, Lionel Davenport¹, Raymond Desilets¹, Susanne Dietz¹, Kristina Dodson¹, Lisa Doup¹, Steven Ferriera¹, Neha Garg¹, Andres Gluecksmann¹, Brit Hart¹, Jason Haynes¹, Charles Haynes¹, Cheryl Heiner¹, Suzanne Hladun¹, Damon Hostin¹, Jarrett Houck¹, Timothy Howland¹, Chinyere Ibegwam¹, Jeffery Johnson¹, Francis Kalush¹, Lesley Kline¹, Shashi Koduru¹, Amy Love¹, Felecia Mann¹, David May¹, Steven McCawley¹, Tina McIntosh¹, Ivy McMullen¹, Mee Moy¹, Linda Moy¹, Brian Murphy¹, Keith Nelson¹, Cynthia Pfannkoch¹, Eric Pratts¹, Vinita Puri¹, Hina Qureshi¹, Matthew Reardon¹, Robert Rodriguez¹, Yu-Hui Rogers¹, Deanna Romblad¹, Bob Ruhfel¹, Richard Scott¹, Cynthia Sitter¹, Michelle Smallwood¹, Erin Stewart¹, Renee Strong¹, Ellen Suh¹, Reginald Thomas¹, Ni Ni Tint¹, Sukyee Tse¹, Claire Vech¹, Gary Wang¹, Jeremy Wetter¹, Sherita Williams¹, Monica Williams¹, Sandra Windsor¹, Emily Winn-Deen¹, Keriellen Wolfe¹, Jayshree Zaveri¹, Karena Zaveri¹, Josep F. Abril¹⁴, Roderic Guigó¹⁴, Michael J. Campbell¹, Kimmen V. Sjolander¹, Brian Karlak¹, Anish Kejariwal¹, Huaiyu Mi¹, Betty

Lazareva1, Thomas Hatton1, Apurva Narechania1, Karen Diemer1, Anushya Muruganujan1, Nan Guo1, Shinji Sato1, Vineet Bafna1, Sorin Istrail1, Ross Lippert1, Russell Schwartz1, Brian Walenz1, Shibu Yooseph1, David Allen1, Anand Basu1, James Baxendale1, Louis Blick1, Marcelo Caminha1, John Carnes-Stine1, Parris Caulk1, Yen-Hui Chiang1, My Coyne1, Carl Dahlke1, Anne Deslattes Mays1, Maria Dombroski1, Michael Donnelly1, Dale Ely1, Shiva Esparham1, Carl Fosler1, Harold Gire1, Stephen Glanowski1, Kenneth Glasser1, Anna Glodek1, Mark Gorokhov1, Ken Graham1, Barry Gropman1, Michael Harris1, Jeremy Heil1, Scott Henderson1, Jeffrey Hoover1, Donald Jennings1, Catherine Jordan1, James Jordan1, John Kasha1, Leonid Kagan1, Cheryl Kraft1, Alexander Levitsky1, Mark Lewis1, Xiangjun Liu1, John Lopez1, Daniel Ma1, William Majoros1, Joe McDaniel1, Sean Murphy1, Matthew Newman1, Trung Nguyen1, Ngoc Nguyen1, Marc Nodell1, Sue Pan1, Jim Peck1, Marshall Peterson1, William Rowe1, Robert Sanders1, John Scott1, Michael Simpson1, Thomas Smith1, Arlan Sprague1, Timothy Stockwell1, Russell Turner1, Eli Venter1, Mei Wang1, Meiyuan Wen1, David Wu1, Mitchell Wu1, Ashley Xia1, Ali Zandieh1, Xiaohong Zhu1. The Sequence of the Human Genome 2001.

76. Fleischmann RD1, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM, et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. 1995; 269.

77. Weber, J. et al. Human Whole-Genome Shotgun Sequencing.

78. Eugene W. Myers1,*, Granger G. Sutton1, Art L. Delcher1, Ian M. Dew1, Dan P. Fasulo1, Michael J. Flanigan1, Saul A. Kravitz1, Clark M. Mobarry1, Knut H. J. Reinert1, Karin A. Remington1, Eric L. Anson1, Randall A. Bolanos1, Hui-Hsien Chou1, Catherine M. Jordan1, Aaron L. Halpern1, Stefano Lonardi1, Ellen M. Beasley1, Rhonda C. Brandon1, Lin Chen1, Patrick J. Dunn1, Zhongwu Lai1, Yong Liang1, Deborah R. Nusskern1, Ming Zhan1, Qing Zhang1, Xiangqun Zheng1, Gerald M. Rubin2, Mark D. Adams1, J. Craig Venter1. A Whole-Genome Assembly of *Drosophila* 2000.

79. Shendure J, Ji H. Next-generation DNA sequencing. *Nature biotechnology* 2008; 26(10):1135–45.

80. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 2008; 456(7218):53–9.

81. Rusk N. Torrents of sequence. *Nat Meth* 2010; 8(1):44.

82. M. J. Levene, J. Korlach1, S. W. Turner, M. Foquet, H. G. Craighead, W. W. Webb. Zero-Mode Waveguides for Single-Molecule Analysis at High Concentrations 2003; 299:682–6.

83. Fischbach MA, Walsh CT. Assembly-line enzymology for polyketide and nonribosomal Peptide antibiotics: logic, machinery, and mechanisms. *Chemical reviews* 2006; 106(8):3468–96.

84. Ikeda H, Ishikawa J, Hanamoto A, Shinose M, Kikuchi H, Shiba T et al. Complete genome sequence and comparative analysis of the industrial microorganism *Streptomyces avermitilis*. *Nature biotechnology* 2003; 21(5):526–31.

-
85. Cortina NS, Krug D, Plaza A, Revermann O, Müller R. Myxoprincomide: a natural product from *Myxococcus xanthus* discovered by comprehensive analysis of the secondary metabolome. *Angewandte Chemie (International ed. in English)* 2012; 51(3):811–6.
86. Silakowski B, Kunze B, Müller R. Multiple hybrid polyketide synthase/non-ribosomal peptide synthetase gene clusters in the myxobacterium *Stigmatella aurantiaca*. *Gene* 2001; 275(2):233–40.
87. Brigitte Kunze, Hans Reichenbach, Rolf Müller, Gerhard Höfle. Aurafuron A and B, New Bioactive Polyketides from *Stigmatella aurantiaca* and *Archangium gephyra* (Myxobacteria). *JOURNAL OF ANTIBIOTICS* 2005; 58:244–51.
88. Walsh CT, Fischbach MA. Natural products version 2.0: connecting genes to molecules. *Journal of the American Chemical Society* 2010; 132(8):2469–93.
89. Sattely ES, Fischbach MA, Walsh CT. Total biosynthesis: in vitro reconstitution of polyketide and nonribosomal peptide pathways. *Natural product reports* 2008; 25(4):757–93.
90. Jones AC, Monroe EA, Eisman EB, Gerwick L, Sherman DH, Gerwick WH. The unique mechanistic transformations involved in the biosynthesis of modular natural products from marine cyanobacteria. *Natural product reports* 2010; 27(7):1048–65.
91. Koglin A, Walsh CT. Structural insights into nonribosomal peptide enzymatic assembly lines. *Natural product reports* 2009; 26(8):987–1000.
92. Starcevic A, Zucko J, Simunkovic J, Long PF, Cullum J, Hranueli D. ClustScan: an integrated program package for the semi-automatic annotation of modular biosynthetic gene clusters and in silico prediction of novel chemical structures. *Nucleic acids research* 2008; 36(21):6882–92.
93. Anand S, Prasad, M V R, Yadav G, Kumar N, Shehara J, Ansari MZ et al. SBSPKS: structure based sequence analysis of polyketide synthases. *Nucleic acids research* 2010; 38(Web Server issue):W487-96.
94. Weber T, Rausch C, Lopez P, Hoof I, Gaykova V, Huson DH et al. CLUSEAN: A computer-based framework for the automated analysis of bacterial secondary metabolite biosynthetic gene clusters. *Journal of Biotechnology* 2009; 140(1-2):13–7.
95. Tilmann Weber, Kai Blin, Srikanth Duddela, Daniel Krug, Hyun Uk Kim, Robert Brucoleri, Sang Yup Lee, Michael A. Fischbach, Rolf Müller, Wolfgang Wohlleben, Rainer Breitling, Eriko Takano and Marnix H. Medema. antiSMASH 3.0—a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic acids research*.
96. Stephen F. Altschul Warren Gish Webb Miller Eugene W. Myers David J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology* 1990; 215:403–10.
97. Erol O, Schaberle TF, Schmitz A, Rachid S, Gurgui C, El OM, et al. Biosynthesis of the myxobacterial antibiotic coralopyronin A 2010.

-
98. Demain AL. Antibiotics: natural products essential to human health. *Medicinal research reviews* 2009; 29(6):821–42.
99. Schneiker S, Perlova O, Kaiser O, Gerth K, Alici A, Altmeyer MO et al. Complete genome sequence of the myxobacterium *Sorangium cellulosum*. *Nature biotechnology* 2007; 25(11):1281–9.
100. Reichenbach H. Myxobacteria, producers of novel bioactive substances 2001.
101. Bode HB, Müller R. Analysis of myxobacterial secondary metabolism goes molecular. *Journal of industrial microbiology & biotechnology* 2006; 33(7):577–88.
102. Lee, Francis Y F, Borzilleri R, Fairchild CR, Kamath A, Smykla R, Kramer R et al. Preclinical discovery of ixabepilone, a highly active antineoplastic agent. *Cancer chemotherapy and pharmacology* 2008; 63(1):157–66.
103. Schäberle TF, Goralski E, Neu E, Erol O, Hölzl G, Dörmann P et al. Marine myxobacteria as a source of antibiotics--comparison of physiology, polyketide-type genes and antibiotic production of three new isolates of *Enhygromyxa salina*. *Marine drugs* 2010; 8(9):2466–79.
104. FLORENZ SASSE*, HEINRICH STEINMETZa, THOMAS SCHUPPb, FRANK PETERSENb, KLAUS MEMMERTb, HANS HOFMANNb, CHRISTOPH HEUSSERb, VOLKER BRINKMANNb, PETER VON MATT and GERHARD HOFLEa and HANS REICHENBAC. Argyrins, Immunosuppressive Cyclic Peptides from Myxobacteria. *The Journal of antibiotics* 2002.
105. Baumann S, Herrmann J, Raju R, Steinmetz H, Mohr KI, Hüttel S et al. Cystobactamids: myxobacterial topoisomerase inhibitors exhibiting potent antibacterial activity. *Angewandte Chemie (International ed. in English)* 2014; 53(52):14605–9.
106. Berod L, Friedrich C, Nandan A, Freitag J, Hagemann S, Harmrolfs K et al. De novo fatty acid synthesis controls the fate between regulatory T and T helper 17 cells. *Nature medicine* 2014; 20(11):1327–33.
107. Brigitte Kunze, Norbert BEDORF and Hofleá" and Hans Reichenbach. MYOCHELIN A, A NEW IRON-CHELATING COMPOUND FROM ANGIOCOCCUS DISCIFORMIS (MYXOBACTERALES). *The Journal of antibiotics* 1989; 42(1):14–7.
108. itte Kunze, Wolfram Trowitzsch-Kienast1, Gerhard Hofle1 and ns Reichenbach. NANNOCHELINS A, B AND C, NEW IRON-CHELATING COMPOUNDS FROM Nannocystis exedens (MYXOBACTERIA). *The Journal of antibiotics* 1991.
109. Takashi Iizuka, Ryosuke Fudou, Yasuko Jojima, Sumie Ogawa, Shigeru Yamanaka, Asutaka Inukai, Makoto Ojika. Miuraenamides A and B, Novel Antimicrobial Cyclic Depsipeptides from a New Slightly Halophilic Myxobacterium: Taxonomy, Production, and Biological Properties. *International Academic Printing Co. Ltd.* 2006; 59:385–91.
110. Wenzel SC, Müller R. Myxobacteria--'microbial factories' for the production of bioactive secondary metabolites. *Molecular bioSystems* 2009; 5(6):567–74.
111. Velicer GJ, Vos M. Sociobiology of the myxobacteria. *Annual review of microbiology* 2009; 63:599–623.
112. Bode HB, Müller R. The impact of bacterial genomics on natural product research. *Angewandte Chemie (International ed. in English)* 2005; 44(42):6828–46.

-
113. Medema MH, Kottmann R, Yilmaz P, Cummings M, Biggins JB, Blin K et al. Minimum Information about a Biosynthetic Gene cluster. *Nature chemical biology* 2015; 11(9):625–31.
 114. Roberts RJ, Carneiro MO, Schatz MC. The advantages of SMRT sequencing. *Genome biology* 2013; 14(7):405.
 115. Koehn FE, Carter GT. The evolving role of natural products in drug discovery. *Nature reviews. Drug discovery* 2005; 4(3):206–20.
 116. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S et al. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics (Oxford, England)* 2012; 28(12):1647–9.
 117. Mark Slee, Aditya Agarwal and Marc Kwiatkowski. Thrift: Scalable Cross-Language Services Implementation. Technical report, Facebook 2007.
 118. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* 2004; 32(5):1792–7.
 119. Singhal A. Modern Information Retrieval: A Brief Overview. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering* 2001; 24(4):35–43.
 120. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology* 1970; 48(3):443–53.
 121. Smith, T. F. and Waterman, M. S. Identification of common molecular subsequences. *Journal of Molecular Biology* 1981; 147(1):195–7.
 122. Colwell RK, Elsensohn JE. EstimateS turns 20: statistical estimation of species richness and shared species from samples, with non-parametric extrapolation. *Ecography* 2014; 37(6):609–13.
 123. Rausch C, Weber T, Kohlbacher O, Wohlleben W, Huson DH. Specificity prediction of adenylation domains in nonribosomal peptide synthetases (NRPS) using transductive support vector machines (TSVMs). *Nucleic acids research* 2005; 33(18):5799–808.
 124. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* 1990:403–10.
 125. Megginson RE. An introduction to Banach space theory. New York: Springer-Verlag; 1998.
 126. Rachid S, Krug D, Kunze B, Kochems I, Scharfe M, Zabriskie TM et al. Molecular and biochemical studies of chondramide formation-highly cytotoxic natural products from *Chondromyces crocatus* Cm c5. *Chemistry & Biology* 2006; 13(6):667–81.
 127. Brigitte Kunze, Rolf JANSEN, Florenz Sasse, HOFLE and Hans Reichenbach. Chondramides A - D, New Antifungal and Cytostatic Depsipeptides from *Chondromyces crocatus* (Myxobacteria) Production, Physico-chemical and Biological Properties. *JOURNAL OF ANTIBIOTICS* 1995:1262–6.

-
128. Jennifer Herrmann, Stephan Hüttel and Rolf Müller. Discovery and Biological Activity of New Chondramides from *Chondromyces* sp. *ChemBioChem* (ChemBioChem) 2013; 14:1573–80.
129. Gerc AJ, Song L, Challis GL, Stanley-Wall NR, Coulthurst SJ. The insect pathogen *Serratia marcescens* Db10 uses a hybrid non-ribosomal peptide synthetase-polyketide synthase to produce the antibiotic althiomycin. *PLoS one* 2012; 7(9):e44673.
130. F. Baumgart, B. Kluge, C. Ullrich, J. Vater, and D. Ziessow. IDENTIFICATION OF AMINO ACID SUBSTITUTIONS IN THE LIPOPEPTIDE SURFACTIN USING 2D NMR SPECTROSCOPY. *BIOCHEMICAL AND BIOPHYSICAL RESEARCH COMMUNICATIONS*:998–1005.
131. Barbara Silakowski, Brigitte Kunze, Gabriele Nordsiek, Helmut Bloecker, Gerhard Hoefle and Rolf Mueller. The myxochelin iron transport regulon of the myxobacterium *Stigmatella aurantiaca* Sg a15. *European journal of biochemistry* 2000; 267:6476±6485.
132. Silke C. Wenzel and Rolf Müller. *Myxobacteria – Unique Microbial Secondary Metabolite Factories*. Elsevier 2010; 2:189–222.
133. Chai Y, Pistorius D, Ullrich A, Weissman KJ, Kazmaier U, Müller R. Discovery of 23 natural tubulysins from *Angiococcus disciformis* An d48 and *Cystobacter* SBCb004. *Chemistry & Biology* 2010; 17(3):296–309.
134. Chai Y, Pistorius D, Ullrich A, Weissman KJ, Kazmaier U, Müller R. Discovery of 23 natural tubulysins from *Angiococcus disciformis* An d48 and *Cystobacter* SBCb004. *Chemistry & Biology* 2010; 17(3):296–309.
135. Gross H, Stockwell VO, Henkels MD, Nowak-Thompson B, Loper JE, Gerwick WH. The genomisotopic approach: a systematic method to isolate products of orphan biosynthetic gene clusters. *Chemistry & Biology* 2007; 14(1):53–63.
136. Medema MH, Paalvast Y, Nguyen DD, Melnik A, Dorrestein PC, Takano E et al. Pep2Path: automated mass spectrometry-guided genome mining of peptidic natural products. *PLoS computational biology* 2014; 10(9):e1003822.
137. Wenzel SC, Gross F, Zhang Y, Fu J, Stewart AF, Müller R. Heterologous expression of a myxobacterial natural products assembly line in pseudomonads via red/ET recombineering. *Chemistry & Biology* 2005; 12(3):349–56.