Rational Engineering of Expression Level of Multi-Gene Systems encoding Natural Product Biosynthesis in *Streptomyces*

A Dissertation
SUBMITTED TO THE FACULTY OF
UNIVERSITY OF MINNESOTA
BY

SZU-YI (SUZIE) HSU

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Advisor: Dr. Michael J. Smanski

August 2020

Acknowledgements

I want to thank my PhD advisor, Dr. Michael J. Smanski, for taking a chance on me and accepted me to join his new laboratory in 2014. It has been quite a journey. I would not be at the place where I am now if not for Dr. Smanski's mentorship and rigorous training in synthetic biology for natural product discovery. His tirelessly guides me to hone my research and science communication skills. If I had to choose one takeaway from working with Dr. Smanski, it would be his enthusiasm towards science. It was his enthusiasm that drew me to join his laboratory at the first place and his enthusiasm continues to inspire me even during the toughest phase of my doctoral journey. I want to thank my PhD committee members for providing valuable feedbacks to my doctoral research and professional development: my committee chair Dr. Claudia Schmidt-Dannert, Dr. Christine Salomon, Dr. Burckhard Seelig and Dr. Larry Wackett. I want to thank BioTechnology Institute for supporting my doctoral research through Biocatalysis Grant and the Johnson family's generosity of funding support through the Johnson Arnold Fellowship.

I also want to thank my lab mates in the Smanski group. I am honored to have had the chance working with these creative, hardworking individuals. They not only provide technical guidance on various topics, but also are the sounding boards for fostering research ideas. I will miss the lunch break when we were immersed in conversations and debates about the sci-fi imaginations, terraforming, drug discovery, and DIY biology. I also want to thank my fellow scientists in the BioTechnology Institute, a great community to talk science. I want to especially thank Dr. LeeAnn Higgins and Dr. Stephen Harvey who provide expertise in troubleshooting and maintenance of liquid-chromatography mass-spectrometer. I also want to thank the undergraduate students I worked with: Tom Hougard on chemical synthesis of serofendic acid, Mariela Rivera De Jesus on DNA assembly, Jihaeng Lee on the colony morphology documentation, Alesksandra Bajer and Dakota Snustad on transcriptional regulation of biosynthetic gene cluster in *Streptomyces*, and Jessica Halverson on liver microsome screening. They made the projects fun to work with as well as taught me a lot about effective mentoring. Finally, I would like to thank Audrey Chia-Yu Kuei for helping me to polish my dissertation defense presentation.

Dedication

To my family in Taiwan and Minnesota, especially my parents and grandmother who taught me the value of humor, hard work, and fearlessness.

To my partner, Chris, who endured many of my sleepless nights, become my personal statistical consultant, and support me on many levels.

Abstract

It has been shown that genomes of bacteria, especially members of the *Streptomyces* genus, harbor unprecedented numbers of biosynthetic gene clusters (BGCs) potentially encoding novel compounds. However, cloning and controlled expression of these large BGCs in a heterologous host require tedious optimization on a case-by-case basis. This dissertation presents a synthetic biology platform to rapidly reconstitute BGCs by refactoring and physically piecing DNA fragments together in a hierarchical manner. The two core technologies are (i) a high-throughput DNA assembly pipeline for high GC organisms and (ii) synthetic genetic elements to control gene expression in *Streptomyces*. As a proof of concept, a small library of synthetic gene clusters was constructed to encode *ent*-atisanoic acid, a late-stage intermediate of the neuroprotectant serofendic acid. We successfully controlled the relative expression level of individual genes, identified the tailoring enzyme required for the oxidation as well as demonstrated the utility of this DNA assembly pipeline. Next, we rationally optimize isoprenoid biosynthesis by perturbing relative expression of eight enzymes in the methylerythritol phosphate (MEP) pathway. One of the Design of Experiment (DoE) methods called Plackett-Burman design was used to guide the optimization effort for this eight-gene system. A five-level Plackett-Burman design was used to guide the design of 125 unique synthetic gene clusters encoding the MEP pathway, which was required to fully screen the effects of expression of each gene on the output measured by isoprenoid titer. In the eight-gene pathway, each gene has one of five expression levels. Total screening of the entire pathway variants has revealed a surprising degree of robustness in actinobacterial secondary metabolism. In sum, the DNA assembly pipeline will become a powerful tool to fuel future rational optimization efforts of multi-gene systems, including large BGCs.

Table of Contents

List of Tables

List of Figures

Chapter 1 Introduction

1.1 Exploring the Biosynthetic Potential of *Streptomyces* through Combinatorial Design of Uncharacterized Biosynthetic Gene Clusters

1.1.1 General Introduction

With the advent of DNA sequencing [1] and *de novo* synthesis [2], the molecular toolkits available to engineer biology have quickly expanded. One of the growing applications of these techniques is in the genome mining field [3], especially in efforts to access previously untapped biosynthetic potential in *Streptomyces* genomes [4]–[7]. *Streptomyces* biosynthetic genes have a particular organization as "biosynthetic gene clusters (BGCs)", which are genes grouped together in a genome that concertedly encode specialized molecules. More than 60% of clinically used antibiotics are derived from *Streptomyces*-produced natural products [8], [9]. There have been a substantial number of biosynthetic gene clusters identified through bioinformatics approaches [10], estimated to encode hundreds of thousands of future drug leads [11]. However there has been little success at expressing the BGCs under laboratory conditions [12], therefore their products and properties remain unknown.

A conventional technique for characterizing these silent BGC is called heterologous expression [13]. During heterologous expression, a target BGC is transferred to a non-native host for its functional characterization. However, several caveats prevent heterologous expression from becoming a universal application. First, the traditional approach is to select expression hosts that are genetically close to the native host, but subtle genetic and metabolic differences between the native host and the expression hosts can cause unexpected expression levels [14]. Natural BGCs often have complex internal regulatory elements embedded in the sequences that interact with non-native host's cellular machinery via complex mechanisms [15], [16]. Several examples of expressing identical multi-gene constructs in several heterologous hosts of the same genus leads to drastic changes in the behaviors [17], [18]. Second, the productivity of the multi-gene system in the non-native host is usually different from wild type activity [19], [20]. Transferring BGCs to a heterologous host without modification leads to unpredictable changes in gene expression. In addition, our current understanding of the regulation of BGC gene expression in *Streptomyces* is not detailed enough to predict functions of the

core regulatory elements. Many heterologous expression studies require substantial optimization on a case-by-case basis, which is laborious and time-consuming. Properly tuned expression levels in a multi-gene pathway is important for maximizing its performance as well as identifying the encoded molecules and understanding the biosynthetic pathway. Thus, it is imperative to develop a systematic approach for tuning gene expression levels in BGCs to access novel natural products.

There are many innovative strategies to overcome the challenges of investigating BGCs via heterologous expression [21]. One of the emerging approaches is re-designing the genetics of uncharacterized BGCs. This approach is based on using abstraction hierarchy to simplify these complex systems. Abstraction hierarchy [22] refers to viewing complex systems as composites of simpler functional devices, which are in turn composed of even simpler parts. In biological systems, including BGCs, functional devices include individual operons or monocistronic expression units and are composed of genetic parts (promoters, ribosome, coding DNA sequences (CDSs), and terminators). A technique called refactoring aims to systematically rewrite the underlying genetics without changing the functionality by (1) deconstructing native complex regulatory elements, (2) introducing well characterized genetic parts, and (3) reconstituting the functionality of BGCs [23]–[25]. The “parts approach” in synthetic biology exploits interchangeability of genetic parts to build large libraries of BGC variants for functional characterization. The goals of such libraries can be to explore combinatorial genetic design space [26], [27], to optimize expression for BGC characterization[28], and even to produce analogs of natural products via combinatorial biosynthesis [29], [30]. In this chapter, I will focus on large-scale reengineering BGCs and outline the considerations of codon composition, cis-acting regulatory elements, combinatorial design and algorithmic DNA assembly that are important for re-designing multi-part systems in *Streptomyces*.

1.1.2 Design of Codon Usage

The degeneracy of the genetic code offers ample opportunities to recode CDSs without changing the primary structure of the encoded protein. However, such changes can impact important aspects of heterologous expression including gene expression levels, construct stability, and ease of DNA assembly. Codon usage in *Streptomyces* is highly GC-biased (65 – 70 % on average) [31]. Traditionally, codon optimization substitutes native codons of the genes of interest with the most preferred codons in the heterologous

host [32]–[34]. Alternatively, codon composition is adjusted to mirror that of similar genes in the host organisms according to codon adaptation index (CAI) [32]. Recent work demonstrated an alternative strategy that uses rare codons to increase protein expression [35]. Goodman et al. measured more than 14,000 combinations of promoter, RBSs and N-terminal codon usage to express a synthetic reporter gene in *Escherichia coli*. Surprisingly, the N-terminal peptide sequence encoding most rare codon variants show increased protein expression by 15-fold on average. Although changing codon composition using synonymous codon pairs maintained the amino acid sequence, the change in mRNA transcript secondary structure has substantial impact on protein expression. The analysis of the mRNA transcripts showed reduced mRNA structure at the +10 region is most correlated with expression increase.

Codon randomization, which is based on the random assignment of codons to genes of interest according to a codon table, uses a weighted frequency distribution of the codons in the host genome or a subset of highly expressed genes [36]. Codon randomization aims to create variant CDSs whose sequences are as divergent from wild type as possible. The benefits of codon randomization include more flexible codon selection and elimination of internal regulatory elements, repetitive sequences and internal restriction sites.

Another interesting study demonstrated unique regulation of codon availability in *Streptomyces.* A tRNA gene *bldA* is needed to decode a rare leucine codon, TTA, during translation [37]. *bldA* is only expressed during the stationary phase in *Streptomyces,* and many genes required for sporulation contain this rare TTA codon. Overall, this is a mechanism of post-transcriptional regulation, where *bldA* availability limits the translation of developmental genes that are transcribed too early. This developmental regulation should be accounted for when designing the codon composition of genes of interest for heterologous expression in *Streptomyces* [38]–[41].

Besides the considerations for CDS composition, other parameters need to be considered in genetic design for efficient translation, including local GC content [42], local context of a given codon and presence of mRNA sequence motifs [43]–[45]. Overall, design of codon composition of BGCs for heterologous expression can be used to control translational efficiency and temporal expression.

1.1.3 Design of Cis-acting Regulatory Elements

Non-coding DNA elements including intergenic regions, promoters, ribosomal binding sites (RBSs) and terminators have substantial effects on gene expression. In natural BGCs, these elements often overlap each other or CDSs, making it difficult to replace them without unintended side-effects [46]–[49]. To make BGC engineering more feasible, applying a “parts approach” to these genetic regulatory elements allows (1) decoupling intrinsic complex pathway regulation embedded in the natural sequence and (2) introducing a modular organization, in which characterized regulatory elements can be easily swapped in or out.

Promoter swapping, or replacing the native promoters with strong, characterized constitutive promoters is often used to increase expression level [50]. Several novel molecules were isolated by promoter swapping to transcriptionally activate uncharacterized silent BGCs in *Streptomyces albus* J1074 (which was reclassified as *Streptomyces albidoflavus* J1074) [51]. Additionally, in the context of multi-gene systems, balancing relative expression levels of biosynthetic enzymes is more important than maximal overexpression of the entire gene cluster [52]. Expressing every gene in a BGC as strongly as possible can lead to sub-optimal stoichiometries of relative gene expression that negatively impact the flux of metabolites through a pathway. It is possible that the flux imbalance leads to accumulation of shunt metabolites or accumulation of intermediates to toxic levels, both of which affect productivity and/or host fitness [53], [54]. Thus, it is highly desirable to have a platform where the expression of separate genes can be individually tuned to optimize flux.

There is ongoing effort to build and characterize large libraries of regulatory elements for use is synthetic biology. These include synthetic promoters [55] and terminators [56], as well as models for designing and predicting the strength of RBSs *de novo* [26]. Part libraries often provide control of gene expression that spans several orders of magnitude, providing varying strengths to increase the range of tuning relative expression levels in a multi-gene system [56], [57]. In addition, there are more sequence-based tools to tune the translational efficiency including synthetic *Streptomyces* riboswitches [58] and RBSs [59]. Recent sequence-based biophysical models of gene regulation and expression, such as RBS calculator developed by the Salis *et al*., are available as web applications [60].

1.1.4 Redesigning Gene Clusters to Aid in Construction and Permutation

Although we have numerous tools to design underlying genetics and predict the multi-gene expression dynamics *in silico*, the technical capabilities remain limited to physically fabricate large-scale libraries of complex, large multi-part variants. Furthermore, the size of natural BGCs ranges from 10 to 100 kilobases [7], [11], [61]. As a result, the refactored BGC variants encompassing multiple synthetic regulatory elements and coding sequences can be quite large. There are two alternative strategies for producing a large synthetic genetic construct: de novo synthesis and ligation-based assembly. Either are technically possible for constructs the size of BGCs. Indeed, entire bacterial genomes have been synthesized *de novo*. However, *de novo* synthesis is cost-prohibitive for the production of large libraries of constructs. Additionally, the high GC bias of the coding sequences in *Streptomyces* also complicates preparation of DNA fragments by PCR or *de novo* synthesis. For these reasons, ligation-based DNA assembly is preferred for the construction of large combinatorial libraries.

Strategies for DNA assembly have evolved in the past decade with the growth of Synthetic Biology. The conventional approach is to tailor a DNA assembly plan according to the sequence characteristics of a particular BGC of interest, which is time-consuming, low-throughput, and requires optimization on a case-by-case basis. Recently, many techniques have been developed for large multi-part assembly for biosynthetic pathways from small DNA fragments including MoClo[62], Gibson assembly [63], and Golden Gate assembly. There is an emerging trend to establish systematic pipelines that incorporate genetic design rules, assembly standards, and hierarchical assembly methods that can be extended to any BGC re-engineering project. This “one-pipeline-fits-all” approach can be called “algorithmic DNA assembly.” An essential factor for fast and reliable engineering of complex systems is standardization and modularity of its basic parts. This accelerates the design-build-test cycle by allowing diverse BGC designs to be built from a relatively small collection of genetic parts. In the context of re-engineering BGCs, standardization allows the reuse of previously characterized genetic elements to generate large combinatorial libraries of coding sequences as well as variants of their regulatory element sequences. Also, standardization facilitates exchange and testing composed parts to rapidly optimize multi-part constructs (**Figure 1.1**).

Combinatorial DNA assembly can accelerate hypothesis testing by allowing diverse genetic designs to be built and screened in parallel. The permutations of regulatory parts, gene order, gene orientations, and operon occupancy can be radically different from

those seen in natural systems [64]. For example, a BGC composed of polycistronic operons can be redesigned as monocistronic transcription units, with a unique promoter, RBS, and terminator controlling the expression of each gene to modulate individual expression dynamics. In addition, varying expression of the genes in the constructs can also be modulated by gene order. The alternative approach is designing BGC variants as artificial operons (in which all the genes are transcriptionally regulated by a single promoter), knowing that transcription distance has a linear relationship with the expression rate [65]. Permuting gene positions is an effective method to identify the variant with higher performance [66]. This technique is especially useful when there are limited genetic parts to control transcription and translation.

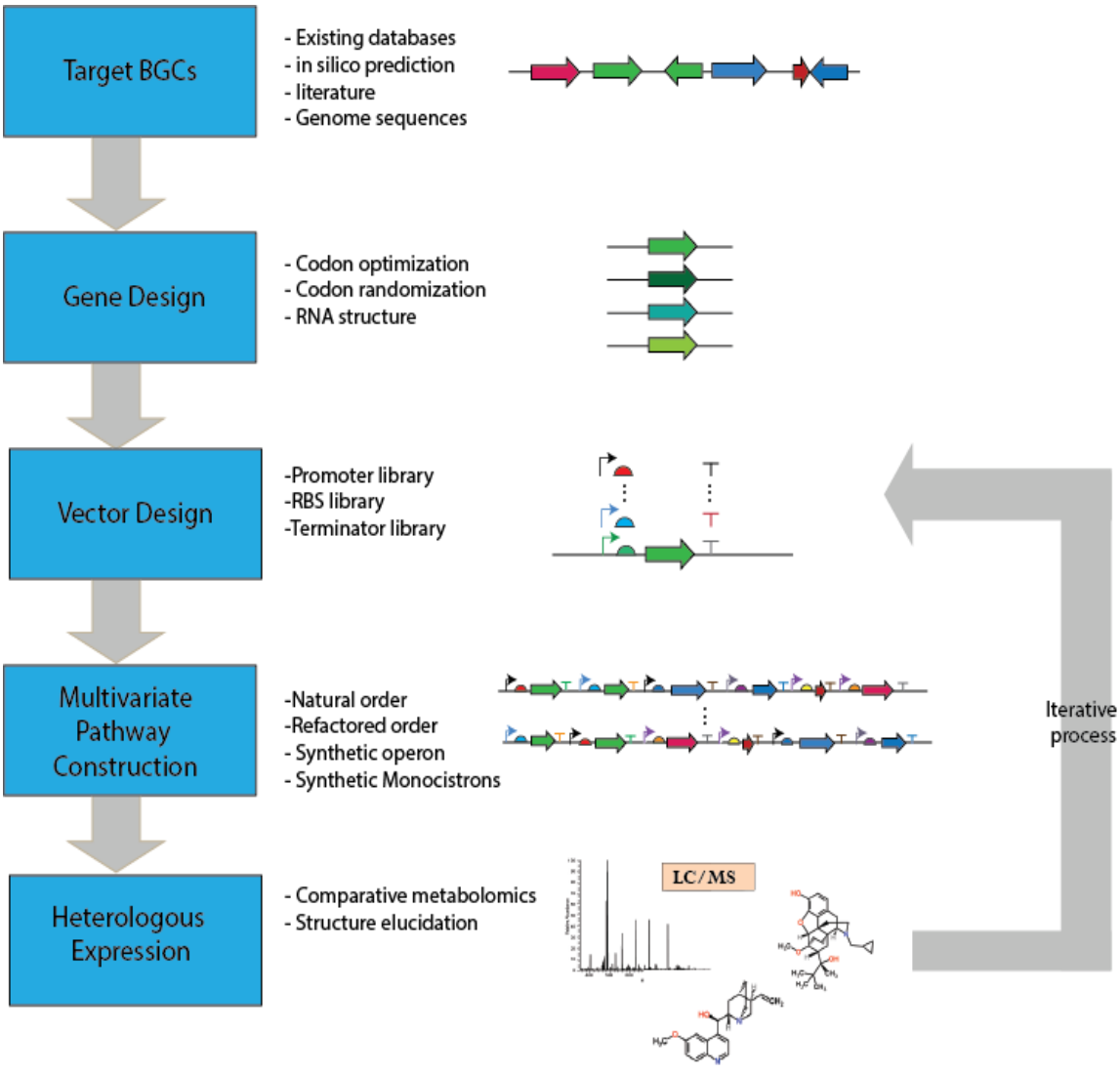


Figure 1.1. Schematic representation multi-part construct design considerations in *Streptomyces*

There are several sequence considerations to building large libraries of multi-gene constructs. First, homologous sequences among fragments of the construct should be avoided to prevent recombination or deletion when introduced into the expression host. In *Streptomyces*, the BGCs often contain repetitive regions, such as BGCs encoding multi-domain polyketide synthases that have high sequence homology and repetitive motifs [67]–[69]. Codon randomization can be used to eliminate highly similar or identical nucleotide sequences. Second, artifacts of repetitive sequences or internal regulatory elements in the DNA fragments should be screened before assembly. Repetitive sequences can render unexpected secondary structures difficult for cloning, DNA synthesis, or translation. This can be prevented by using unique regulatory elements to construct biosynthetic pathway variants. Third, internal restriction enzyme recognition sites should be eliminated to assist assembly of larger constructs [70].

1.1.5 Conclusion and Outlook

Redesigning BGCs at the DNA sequence level has tremendous potential to accelerate natural product discovery. Mining the genomes of *Streptomyces* poses an interesting challenge including selection of appropriate cis-acting regulatory elements, tuning optimal expression dynamics for multi-part constructs encoding large BGCs, and physical preparation and assembly of high-GC DNA fragments for larger constructs. The need for innovative strategies to reliably assemble large DNA constructs remains. The ability to iterate through ‘design-build-test-learn cycles’ quickly will accelerate efforts to understand and engineer natural product biosynthesis. This potential in the natural product field is just beginning to be realized.

1.2 Thesis Overview

The aim of this dissertation is to apply algorithmic DNA assembly methods to engineer and understand natural product biosynthesis in *Streptomyces.* The scope of the dissertation is focused on terpenoid biosynthesis in *Streptomyces*, a genus of Gram-positive bacteria known for prolific production of antibiotics and specialized metabolites. This dissertation presents a synthetic biology platform which integrates a high-throughput algorithmic DNA assembly pipeline for high GC DNA sequences and a library of synthetic

genetic elements to control gene expression. A refactored gene cluster encoding a novel diterpene acid, *ent*-atiserenoic acid (eAA), is used as a model multi-gene system to demonstrate utility of the *Streptomyces* synthetic biology platform developed during my PhD. This dissertation also explored the application of Design of Experiments (DoE) to rationally reduce the number of genetic constructs required to identify optimal genetic design encoding desired function (in this case, maximal eAA production titer).

The motivation of this doctoral dissertation is to bridge the gap between the promise and the reality of *Streptomyces*-derived natural products as sources for new drug leads. On average, *Streptomyces* bacteria harbor 40+ biosynthetic gene clusters (BGCs) per genome based on the analysis of 1,100 publicly available *Streptomyces* genomes.[7] Hypothetically, there are hundreds, if not thousands, of novel metabolites encoded in the *Streptomyces* genomes waiting to be discovered. However, it is a non-trivial task to experimentally validate the products predicted to be associated with these BGCs; the number of new natural products empirically characterized through genome mining of *Streptomyces* is far fewer than the number of BGCs identified computationally.[71] One of the most prominent challenges in genome mining is to identify conditions for proper expression of the BGCs and detect encoded products. Most BGCs encoded in *Streptomyces* strains are not expressed at high-enough levels to detect the corresponding metabolites when the strains are cultivated in laboratory conditions. Transferring unmodified or partially modified BGCs to heterologous hosts is time consuming and does not guarantee proper expression of the associated gene products. With these challenges of the genome mining of *Streptomyces* in mind, I aimed to streamline the workflow of BGCs characterization from high-throughput DNA assembly to combinatorial refactoring to identify genetic designs for the optimal expression of cryptic or synthetic BGCs in *Streptomyces*.

While working with Dr. Smanski, I developed two technologies which have laid the foundation for combinatorial optimization of multi-gene systems in *Streptomyces*. The first technology is the high-throughput algorithmic DNA assembly pipeline designed for cloning high-GC nucleotide sequences (chapters 2 and 3). Standardization of genetic parts, including cloning vectors, CDSs, and synthetic regulatory elements, leads to facile assembly of large DNA constructs in parallel. The modularity enables permutation of parts to be assembled with ease. We have developed a hierarchical assembly pipeline in which each step is a one-pot reaction with >95% efficiency. A special feature of this algorithmic DNA assembly pipeline is the use of two Type IIS restriction enzymes, *Aar*I and *Sap*I,

instead of the *Bbs*I and *Bsa*I used in conventional Modular Cloning.[62] The high GC content of *Streptomyces* genomes leads to an increased frequency of *Bbs*I and *Bsa*I restriction sites compared to of *Aar*I and *Sap*I; on average, restriction sites of *Bbs*I and *Bsa*I appear 81 times whereas the restriction sites of *Aar*I and *Sap*I appear 28 times per 50 kB of *Streptomyces* genomic DNA. By building the algorithmic DNA assembly pipeline for high GC genes with *Sap*I and *Aar*I, the domestication of CDSs is more efficient and economical, allowing us to target more gene clusters.

The second technology indispensable to the success of my project is the parts library of 19 synthetic promoter-RBSs and 9 transcriptional terminators (Chapters 3 and 4). The parts library allows controlling and tuning expression levels ranging over three orders of magnitude. These synthetic expression control elements with varying strengths extend the tuning range of relative expression levels. In addition, using synthetic promoter sequences derived from the housekeeping sigma factor-, σhrdB-, dependent promoters[72] allow production titer to be decoupled from media components and tightly linked with biomass accumulation; the expression level of each promoter-RBS is constitutive and consistent in two different media. Furthermore, diverse sequences of the DNA parts prevent undesired homologous recombination during physical assembly of genetic parts.

As a proof of concept, a small library of refactored gene clusters encoding eAA was constructed using the synthetic biology platform in *Streptomyces* described in this dissertation (Chapter 3). Tuning relative expression levels of nine genes gave several insights to the relationship with genotype and chemotype. Following that, I aimed to increase the supply of isoprenoid precursor to increase eAA production titer (Chapter 4). I applied a DoE methodology called Plackett-Burman design to rationally design the expression of eight genes encoding the methylerythritol phosphate (MEP) pathway, an isoprenoid precursor pathway that most strongly impact eAA production levels. Specifically, a five-level Plackett-Burman design generated 125 unique MEP pathway variants which have different relative expression level for each of the eight genes.

A chapter-by-chapter breakdown of the contents of this thesis follows:

Chapter 1. This chapter reviews the background of the current approaches to study uncharacterized BGCs, particularly in *Streptomyces*. This chapter also highlights untapped opportunities in synthetic biology tool development for *Streptomyces*, as most of the novel molecular and synthetic biology tools are developed for *E. coli*, *Saccharomyces cerevisiae* and other model organisms.

Chapter 2. The goal of this chapter is to delineate the considerations to design and implement an algorithmic DNA assembly pipeline to increase the throughput of prototyping multi-gene designs. This chapter covers three new molecular biology techniques: PCR ligation, isothermal assembly, and Golden Gate assembly to assemble DNA of different sizes. Finally, I present an algorithmic DNA assembly pipeline developed during the first year of my PhD in the Smanski group. Specifically, I incorporated all three DNA assembly techniques mentioned in this chapter to build large gene clusters from small, functional genetic parts. This chapter is adapted from a published book chapter "Designing and implementing algorithmic DNA assembly pipelines for multi-gene systems" [73] (https://doi.org/10.1007/978-1-4939-7295-1_9)

Chapter 3. This chapter presents the application of the *Streptomyces* synthetic biology platform for the combinatorial optimization of natural products biosynthesis in *Streptomyces*. As a proof of concept, I constructed a small library of gene clusters to afford a synthetic biosynthetic pathway of *ent*-atiserenoic acid (*e*AA), a late stage intermediate in the total chemical synthesis of serofendic acid. This project demonstrated that the high-throughput combinatorial genetic refactoring effectively introduced synthetic control over individual expression levels, chemical diversity, and the production of desired molecules. This chapter is adapted from published work: "Semisynthesis of the Neuroprotective Metabolite, Serofendic Acid" [74] (https://doi.org/10.1021/acssynbio.9b00261).

Chapter 4. This chapter presents the exploration of multivariate statistical approaches to identify optimal relative expression level for a synthetic eight-gene cluster encoding the methylerythritol phosphate (MEP) pathway, which produces isoprenoid precursors. The main hypothesis is that mathematical optimization paradigms, which have been long used in industrial and chemical engineering, can be extended to well-defined genetic systems. The library design of the MEP pathway was rationally guided by a five-level Plackett-Burman design, a multivariate design principle which allows screening of large number of variables while minimizing the number of tests. The five-level Plackett-Burman design yielded 125 unique genetic designs which were constructed and screened. The analysis identifies *dxs* as the most important gene affecting isoprenoid titer, which corroborated the same conclusions in the MEP pathway literature.

Chapter 5. This chapter summarizes the major results of my thesis work and discusses logical future directions for my specific projects as well as the field of microbial genome mining for drug discovery in general.

Appendices: Additional information is appended and relates to: (i) eAA biosynthetic pathway design, and (ii) rational refactoring of MEP pathway:

Appendix 1. This appendix includes additional information of the refactored eAA gene cluster described in chapter 3. Specifically, it documents supplementary methods, materials, figures, data, and NMR spectra of novel metabolites and structural analogs of serofendic acid. A separate excel file called "**Supplementary Data File A1**" contains supplementary materials of chapter 3.

Appendix 2. This appendix includes additional information of rational genetic design of a synthetic MEP pathway described in chapter 4. Specifically, it documents supplementary methods, materials, figures, data and legend of a separate excel file called "**Supplementary Data File A2**" which contains supplementary materials of chapter 4.

Chapter 2 Designing and Implementing Algorithmic DNA Assembly Pipelines for Multi-gene Systems

This chapter is a reformatted reprint of the book chapter Hsu SY and Smanski, M.J. (2018) “Designing and Implementing Algorithmic DNA Assembly Pipelines for Multi-Gene Systems” In: Jensen M.K., Keasling J.D. (eds) Synthetic Metabolic Pathways. Methods in Molecular Biology, vol 1671. Humana Press, New York, NY.

Book chapter hyperlink:

https://doi.org/10.1007/978-1-4939-7295-1_9

SYH wrote the original draft. SYH and MJS reviewed and edited the manuscript.

Summary

Advances in DNA synthesis and assembly technology allow for the high-throughput fabrication of hundreds to thousands of multi-part genetic constructs in a short time. This allows for rapid hypothesis-testing and genetic optimization in multi-gene biological systems. Here we discuss key considerations to design and implement an algorithmic DNA assembly pipeline that provides the freedom to change nearly any design variable in a multi-gene system. In addition to considerations for pipeline design, we describe protocols for three useful molecular biology techniques in plasmid construction.

2.1 Introduction

In the past 50 years, recombinant DNA (rDNA) technology has been applied to produce life-saving medicines, herbicide-resistant or nutrient-enriched food crops, and has enabled to numerous discoveries in the life sciences. The vast majority of rDNA-enabled applications rely on rather simple systems comprising one or a few genes. Recent advances in DNA synthesis, DNA assembly, and synthetic biology have allowed for the engineering of more complex biological capabilities that require the coordinated expression of a dozen or more genes. These include multigene systems that control cellular computation [75], [76], biosynthesis of structurally complicated drugs [52], [77] and enhanced biomaterials [78] and energy systems [79]. The engineering of these massively multipart genetic systems is facilitated by integrated pipelines of automated genetic design, plasmid construction, and experimental characterization. Engineering biology at this scale

is facilitated by robust and algorithmic DNA assembly pipelines that allow hundreds of variant multigene constructs to be built and tested in parallel.

DNA synthesis and DNA assembly are two related yet distinct technology areas that both have seen tremendous advances in the past decade. We define DNA synthesis as the process of polymerizing nucleotide monomers using chemical processes. This is most commonly done with phosphoramadite chemistry, but the platforms vary from solid phase synthesis on resin to photochemistry on glass slides. DNA synthesis technology has been reviewed recently [2] and is not the focus of this chapter. DNA assembly refers to the suite of biochemical methods available to combine individual fragments of single- or double-stranded DNA into larger composite constructs. DNA assembly has been possible for the past half century using restriction enzyme-mediated cloning reactions; however, a number of new techniques now allow DNA assembly of many parts with high levels of efficiency and complete control over genetic design. For genetic engineering projects that require construction and testing of combinatorial libraries of constructs composed of the same DNA sequences (e.g. genes in a metabolic pathway), DNA assembly is more economical than DNA synthesis alone. A DNA assembly 'pipeline' is a set of protocols that can be used in succession to go all the way from the individual fragments of synthetic DNA to a large multigene construct.

Refactored systems, in which the genetics have been systematically rewritten to enable engineering efforts, can be designed to be compatible with any DNA assembly pipeline. The main advantages to working with refactored genetic systems are that (i) every functional genetic element is characterized and its role in expression of the system is understood, (ii) regulation can be decoupled from host chromosome, for example by using orthogonal RNA polymerases, and (iii) they contain a modular genetic architecture. This allows the "parts approach" in synthetic biology, where genetic elements like promoters and ribosome binding sites are swapped in a combinatorial fashion to control gene expression. Many complex systems have been refactored in recent years, including bacteriophage [80], nitrogen fixation [64], and natural product biosynthesis pathways [42], [81]. Even a 272 kb yeast chromosome has been redesigned and assembled from synthetic oligonucleotides [82].

An important facet of refactored systems is that they can be designed to accommodate the DNA synthesis and assembly protocols used for fabrication. This contrasts with modification of plasmids encoding wild-type sequences, which is highly constrained by the presence or absence of restriction recognition sequences and by the

overlapping nature of encoded genetic elements. This means that unique DNA assembly routes had to be created for each recombinant plasmid design on a case-by-case basis. With refactored systems, it is possible to pre-define a set of robust DNA assembly protocols that will be used repeatedly to build any construct via an algorithmic, efficient pipeline. In other words, the DNA sequence is adjusted to accommodate the DNA assembly, not vice versa. Diversity in the genetic designs is controlled by varying the substrate DNA fragments that are included in each assembly reaction.

Recently, many novel assembly techniques have been developed for large multi-part assembly for multi-gene systems from small DNA fragments, including isothermal assembly [83], Golden Gate assembly [84], and yeast recombination [67]. These cloning methods can be coupled with standardized vector designs to provide hierarchical DNA assembly systems, such as Modular Cloning [62], GoldenBraid [85], BioBrick/Bglbrick [86], and TNT-cloning [87]. However, understanding the strengths and weaknesses of each cloning method allows the genetic engineering to customize a DNA assembly system to meet the design constraints of a specific multigene engineering project.

This review covers major design considerations of an algorithmic DNA assembly pipeline and three useful DNA assembly techniques that can be incorporated in the assembly pipeline that fit individual assembly projects. We describe protocols to design primers and perform PCR-ligation, isothermal assembly, and Golden Gate assembly, and provides example of an algorithmic DNA assembly pipeline that integrates all three techniques. Lastly, we will discuss the advantages and limitations of each assembly technique so that the users are aware how to design, implement, test, and troubleshoot an algorithmic pipeline for given DNA assembly applications (*see* **Note 2.4.3**).

2.2 Materials

Prepare all the enzymatic reactions in PCR tubes and on ice. Prevent unnecessary thawing and freezing of the buffers and enzymes, for which are generally heat labile and therefore should always be kept at -15 to -25°C. When preparing reactions, the enzymes can be kept on ice or small freezer boxes for a brief amount of time. Always prepare reactions using ultrapure water (which is often purified by Milli-Q® water purification system) that has resistivity of 18.2 MΩ·cm. All the buffers and enzymatic reactions should be mixed well before use.

2.2.1 Computational and Instrumental Requirements

1. Plasmid editing software facilitates primer design and visualization of intermediate and final constructs. Free software that the authors recommend includes ApE (A Plasmid Editor) for computers running the Windows operating system. The software can be downloaded at http://biologylabs.utah.edu/jorgensen/wayned/ape/. For Mac users we recommend SnapGene (http://www.snapgene.com/).
2. Nanodrop or similar spectrophotometer
3. Thermocycler
4. DNA gel electrophoresis equipment
5. Gel imaging system

2.2.2 PCR-ligation

1. NEB T4 DNA ligase (400,000 units/ml; Catalog No. M0202S)
2. NEB T4 DNA polynucleotide kinase (10,000 units/ml; Catalog No. M0201S)
3. NEB 10x T4 ligase buffer
4. NEB *Dpn*I (20,000 units/ml; Catalog No. R0176S)
5. Q5® High-Fidelity PCR Kit (200 reactions; Catalog No. E0555L)
6. Deoxynucleotide (dNTP) Solution Mix
7. Plasmid template(s)
8. Primers
9. PCR tubes
10. Zymoclean™ Gel DNA Recovery Kit (200 preps; Catalog No. D4002)
11. Agarose gel for DNA gel electrophoresis
12. 6X DNA loading Dye
13. 1 Kb DNA plus ladder (250 µg; Catalog No. 10787018)

2.2.3 Isothermal Assembly

1. NEB *Dpn*I (20,000 units/ml; Catalog No. R0176S)
2. Q5® High-Fidelity PCR Kit (200 reactions; Catalog No. E0555L)
3. Deoxynucleotide (dNTP) Solution Mix
4. NEB Gibson Assembly 2X Master mix (50 reactions; Catalog No. E2611L)
5. Zymoclean™ Gel DNA Recovery Kit (200 preps; Catalog No. D4002)

6. Up to eight DNA fragments with 20-40 bp overlapping sequences
7. PCR tubes
8. Plasmid template(s)
9. Primers
10. PCR tubes

2.2.4 Golden Gate Assembly

1. Type IIS restriction endonuclease
2. Promega T4 DNA ligase (HC) (500 U; Catalog No. M1794)
3. 10X T4 DNA ligase buffer
4. Plasmid DNA with Type IIS restriction recognition sites flaking the desired sequences and compatible 5′ overhangs
5. Primers
6. PCR tubes

2.3 Methods

2.3.1 Key Considerations for Designing an Algorithmic DNA Assembly Pipeline

Developing a customized DNA assembly pipeline can be organized into three stages. The first stage is definition of design parameters for genetic constructs needed for the research project (**Figure 2.1**). For example, what types of genetic elements (promoters, ribosomal binding sites (RBSs), insulator sequences, etc) will be varied during library design? Where are scar sequences (vestigial DNA sequences that remain as a result of the cloning strategy, for example, restriction recognition sequences) between genetic elements to be tolerated, and how large can these scars be? Will higher-level architectural variables including operon occupancy, gene order, or gene orientation be permuted? Libraries that contain more design constraints can often be constructed with more streamlined assembly pipelines compared with libraries that encode diverse designs. In the second stage, the DNA assembly methods most suited to the assembly needs are selected. In the Notes section below, we discuss the advantages, limitations, failure modes and troubleshooting methods for three DNA assembly techniques: PCR-ligation, isothermal assembly, and Golden Gate assembly. Other assembly techniques have been reviewed recently [70]. Included in the second stage is specifying the specific molecular

components that will be used for DNA assembly (*see* **Note 2.4.1**). This is particularly important for Golden Gate assembly, where there are many type IIS restriction enzymes that could be used in a pipeline, the selection of which will impact genetic part domestication. Aside from restriction enzyme choice, other components to determine include molecular features of cloning and expression vectors, antibiotic resistance cassette as selectable markers, and reporter genes for rapid and efficient screening of correct construct in each step of the pipeline. With these variables set, it is possible to construct all required vector plasmids needed for the assembly system. The third stage is to plan how intermediate and final constructs will be validated. Plasmid verification can take even more time and resources than DNA assembly, and the unique failure modes of each technique dictate different verification methods (*see* **Note 2.4.2**). The following subsections cover three useful DNA assembly methods that can be integrated into an algorithmic pipeline.

2.3.2 PCR-ligation

PCR-ligation is useful for adding or modifying short DNA sequences (1-200 bp) in existing plasmids. The entire plasmid is amplified in a single PCR reaction, with new sequences added via the 5′-ends of the forward and/or reverse primers (**Figure 2.2A**). Intramolecular blunt-end ligation of the linear PCR product re-circularizes the plasmid and readies it for transformation. PCR-ligation is particularly useful to build libraries of small functional DNA sequences such as promoters, RBS, and terminators with flanking sequences that will allow them to enter later stages of an assembly pipeline.

2.3.2.1 Primer Design

1. The 3’-ends of both the forward and reverse primer should have sufficient complementarity to the template to provide a Tm of 55°C-70°C (usually 18- 35 bp that will hybridize to the template; this can be checked using New England Biolabs’ Tm calculator, http://tmcalculator.neb.com). Note that the primers should anneal at the regions of the template plasmid that allow the user to amplify molecular features that the user wants to include in the final construct.
2. New DNA sequence is added to the 5′ end of either forward primer, reserve primer or both and will be incorporated at the ligation junction. It is highly suggested to

use plasmid editing software to construct the desired final sequences *in silico* before designing primers.

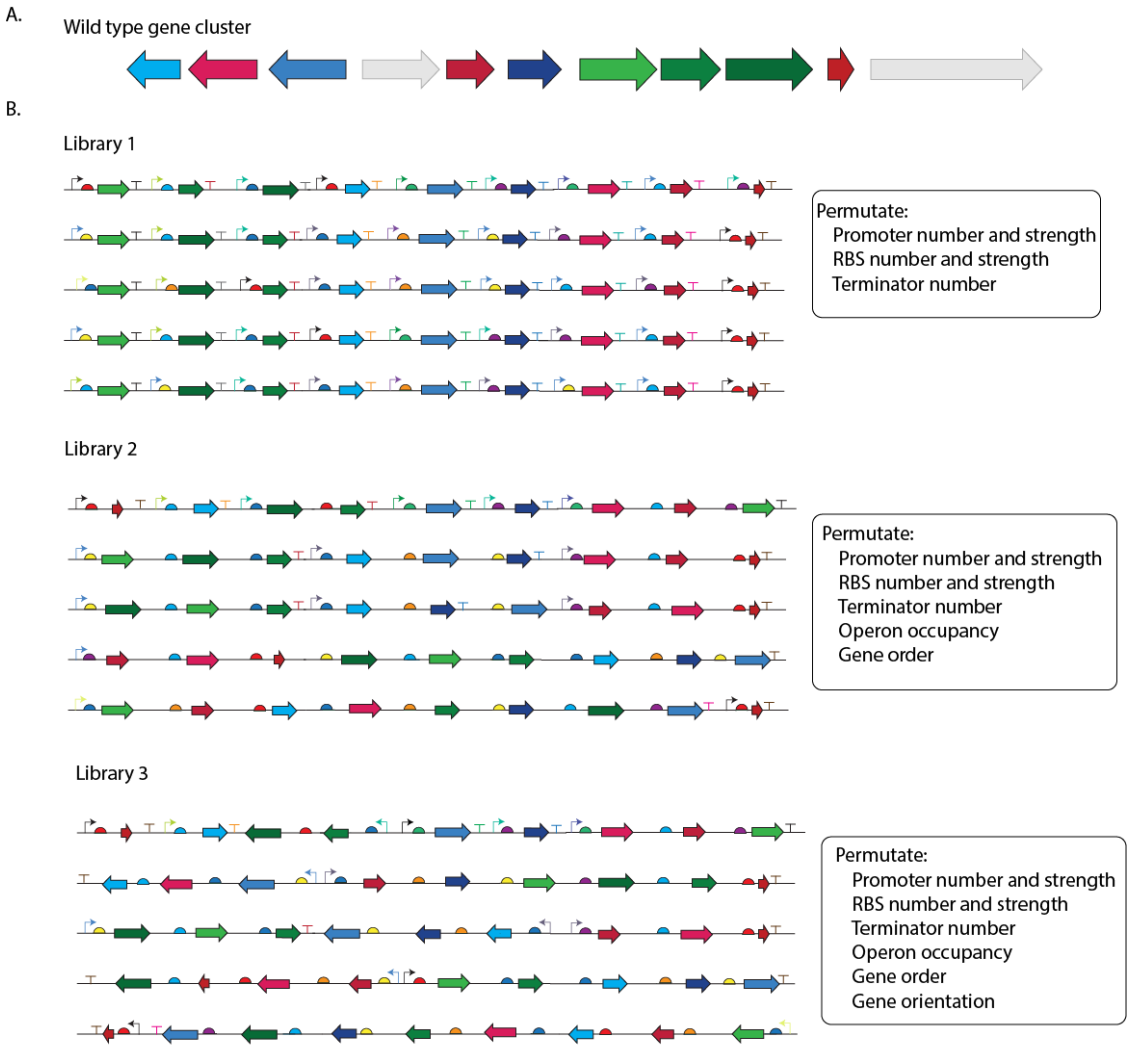


Figure 2.1. Three library designs with different design parameters.

(A) A wild type gene cluster. Arrows represent genes. Grey arrows are the genes not essential to gene functions. (B) Example of three libraries with different design parameters.

2.3.2.2 Protocol

1. Thaw dNTPs, Q5 PCR buffer, primers and template DNA on ice.
2. Prepare the PCR mix on ice according to the standard NEB Q5® polymerase PCR protocol (**Table 2.1**). Add Q5 polymerase to the mix the last and avoid letting polymerase warm above -20°C.
3. Calculate the appropriate annealing temperature for specific primer pairs using NEB Tm calculator (http://tmcalculator.neb.com/#!/). Make sure to select “Q5”

under product group drop down menu, “Q5 High-Fidelity DNA Polymerase” under polymerase/Kit drop down menu. Type “500” in Primer concentration (nM) box. Only input DNA sequences that anneal to the template. Otherwise the calculated anneal temperature will be overestimated if input is the entire primer sequence.

4. Calculate the extension time for your construct. Q5 DNA polymerase extends at 20-30 seconds per kilobase.
5. Start thermal cycler reaction with the following parameters given in **Table 2.2**.

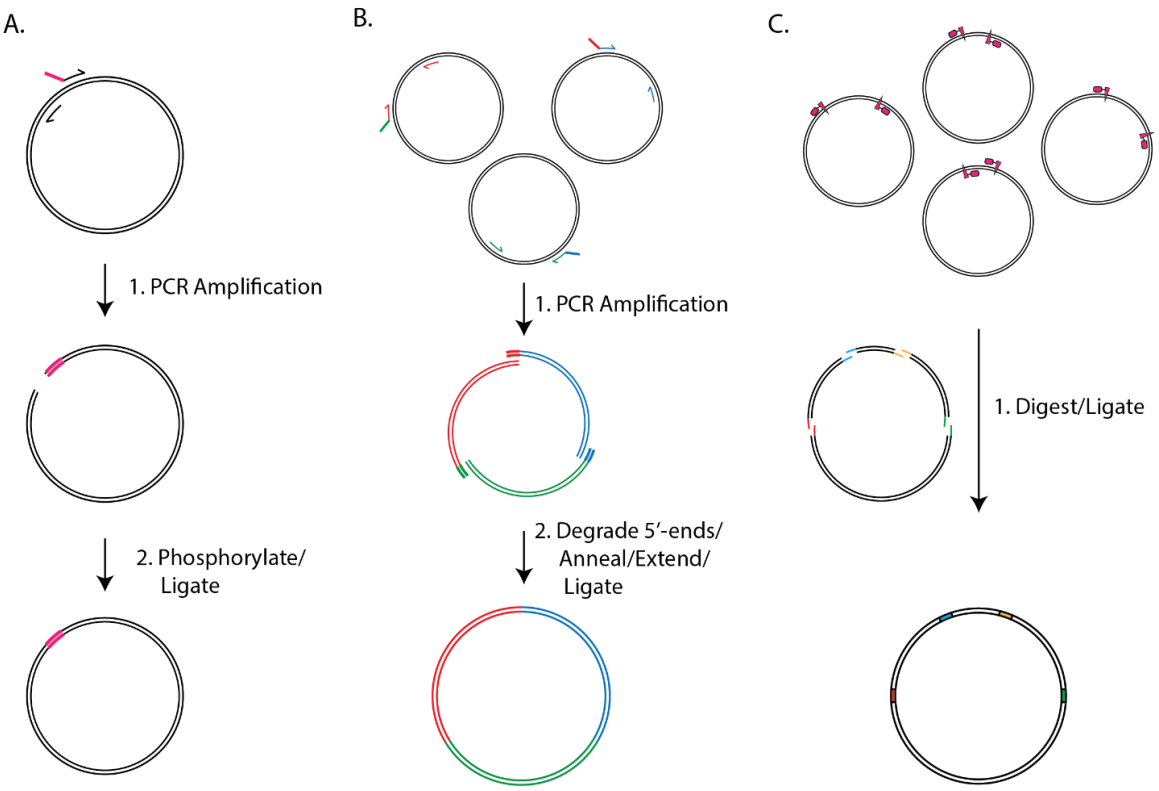


Figure 2.2. Overall schematics of PCR-ligation, isothermal assembly, and Golden Gate assembly.

(A) Key steps for PCR-ligation. Half arrows represent oligonucleotide primers. Red color denotes new DNA sequence being added to the final construct (bottom). (B) Isothermal assembly. The double stranded lines with different colors denote PCR products with unique set of oligonucleotide primers. (C) Golden Gate assembly. Red rectangle represents arbitrary type IIS restriction recognition sequences. Red triangle represents arbitrary cleavage site. The single line highlighted in different colors represents arbitrary 5′ overhang generated by restriction digest.

Table 2.1. Composition of Q5 PCR reaction mixture.

Component	Volume	Final Concentration
5X Q5 Reaction Buffer	5 µl	1X
10 mM dNTPs	0.5 µl	200 µM
10 µM Forward Primer	1.25 µl	0.5 µM
10 µM Reverse Primer	1.25 µl	0.5 µM
Template DNA	variable	< 1,000 ng
Q5 High-Fidelity DNA Polymerase	0.25 µl	0.02 U/µl
5X Q5 High GC Enhancer (optional)	(5 µl)	(1X)
Nuclease-Free Water	to 25 µl	

Table 2.2. Thermocycling program for Q5 reaction.

Step	Temperature	Time
Initial Denaturation	98°C	30 seconds
25–35 Cycles	98°C	5-10 seconds
	*50–72°C	10-30 seconds
	72°C	20-30 seconds/kb
Final Extension	72°C	2 minutes
Hold	4–10°C	∞

6. Once the PCR is finished, freeze the mixture in the -20°C or proceed to the step 7.
7. Analyze the presence of PCR product by DNA gel electrophoresis: mix 4 µL PCR and with 0.6 µL 6x DNA loading dye. Load the PCR mixture and 1 Kb DNA ladder on a 1% agarose TAE gel to check for expected product size.
8. *Dpn*I digest to remove template plasmid: add 1 µL *Dpn*I to the remaining PCR mixture and incubate at 37°C for 30 minutes to one hour. Heat inactivate *Dpn*I at 80°C for 20 minutes.
9. Purify the digested mixture using Zymoclean Gel Extraction Kit: Add 5 volume of DNA binding buffer to your PCR and load to the column. Spin down for 30 seconds. Discard flow-through. Wash each column using 200 µL column wash buffer two times. Add minimal amount of water (≥ 6.5 µL) to elute the purified PCR product so it is as concentrated as possible.
10. Perform T4 DNA ligation reaction: For each ligation, add the entire purified PCR product, 1 µL T4 ligase, 5 units T4 polynucleotide kinase (PNK), 1 µL T4 ligase buffer, and add water to bring up the final reaction volume to 10 µL. T4 ligase and T4 PNK should be added to the reaction last. Incubate the ligation mixture at room temperature for 1 hour. Transform 1-5 µL of the ligation mixture into 20-50 µL chemically competent cells.

2.3.3 Isothermal Assembly

Isothermal assembly [83], also known as Gibson assembly, is a homology-based, restriction enzyme-independent method for stitching together multiple pieces of linear DNA (**Figure 2.2B**). The reaction utilizes a master mix of a 5′ exonuclease to 'chew back' one strand of the double stranded DNA, a DNA polymerase that fills in the gaps that are created and eventually overtake the exonuclease, and a DNA ligase that covalently joins independent pieces by repairing nicks. This method is fast, efficient, and reliable for multipart DNA assembly reactions, and has been used in the complete chemical synthesis of a bacterial genome [88].

2.3.3.1 Primer Design

1. Primers for adjoining DNA fragments must encode 20-40 bases of overlapping sequence for annealing and ligation. We highly recommend designing the final construct sequences *in silico* prior to primer design. For each DNA fragment to be assembled, first design annealing 3′-ends of both the forward and reverse primers. They should have sufficient complementarity to the template to provide a Tm of 55°C-70°C (usually 18-35 bp that will anneal to the template).
2. Overlap sequences of 20-40 nucleotides from neighboring DNA fragments are added into 5′ end of both forward and reverse primers to allow for annealing of overhangs.
3. For sections of DNA that must be completely synthesized *de novo*, single stranded oligos can be used directly by 'tiling' them with successive 20 bp overlaps. The key to this single-stranded Gibson assembly is that the terminal oligo on each end must create a 3′ overhang.
4. Alternatively, the primers can be designed by NEBuilder, a web-based Gibson assembly visualization tool provided by NEB. This can be found on http://nebuilder.neb.com.

2.3.3.2 Protocol

1. Perform PCR with primers designed to incorporate overlapping ends as described in PCR protocol in **Method 2.3.2.2**, including *Dpn*I digestion and purification.
2. Measure the concentration of purified DNA fragments using Nanodrop.

3. Thaw 10 µL Gibson assembly master mix on ice.
4. Add 100 ng of the vector and equimolar amounts of other DNA and add water to bring the final volume up to 20 µL.
5. Incubate the isothermal reaction mixture at 50°C for 1 hour.
6. Transform 1-5 µL of the isothermal reaction into 20-50 µL chemically competent cells.

2.3.4 Golden Gate Assembly

Golden Gate assembly [84] utilize type IIS restriction enzymes to generate 3-4 bp sticky ends outside their recognition sequences that can be subsequently joined by T4 ligase (**Figure 2.2C**) in a one-pot reaction. That type IIS restriction enzymes cleave outside their recognition site provides several advantages. First, the overhangs generated upon cleavage can be customized because they are independent of the restriction recognition sequences, allowing for scarless assemblies or combinatorial assembly between user-defined 3-4 base junctions. Also, because recognition sites are not present in the final assembled product, Golden Gate assembly can proceed in a one-pot digestion/ligation reaction with substantially higher efficiency than traditional cloning. Golden Gate reactions can routinely be used to incorporate several (>5) fragments into large (>25 kb) plasmid designs. An advanced version of Golden Gate assembly utilizes two type IIS restriction enzymes alternating between assembly stages to build an “infinite cloning loop” (**Figure 2.3**).

2.3.4.1 Primer Design

As with Isothermal assembly reactions, we suggest creating a plasmid sequence file of the final construct before designing primers. Here we discuss two aspects of Golden Gate assembly that require extra attention: the orientation of restriction recognition sequence and the position and design of scar site. Note that unlike isothermal reactions, it is possible and even preferable to start with circular plasmid substrates instead of linear fragments. Whether using circular or linear substrates, it is essential to design the orientation of restriction recognition sites and overhangs correctly.

1. For each DNA fragment, design annealing 3′ ends of both the forward and reverse primers. They should have sufficient complementarity to the template to provide a Tm of 55°C-70°C (usually 18-35 bp that will anneal to the template).

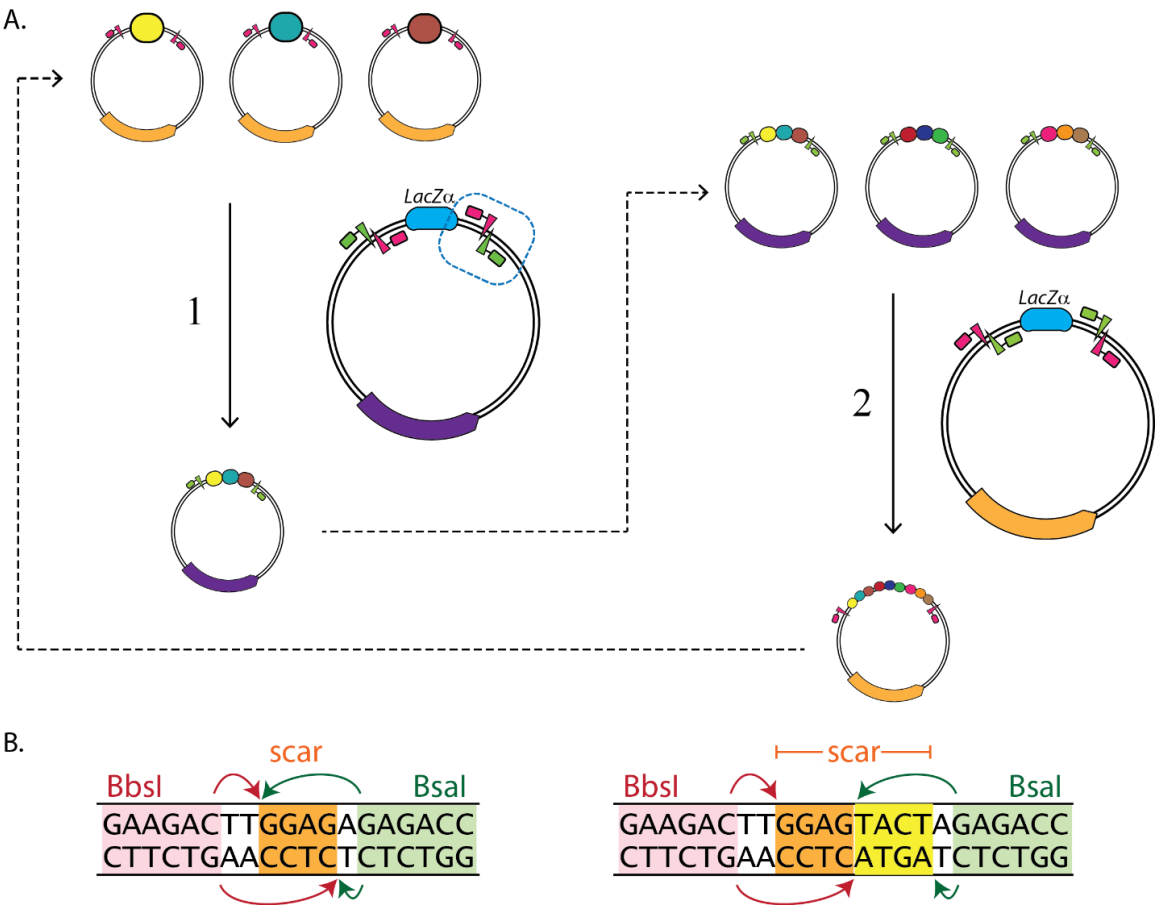


Figure 2.3. Schematic of an iterative Golden Gate assembly.

(A) Alternating between two arbitrary type IIS restriction enzyme allows infinite cloning loop. The oval in different colors denotes DNA part. The arbitrary antibiotic resistance cassette on the donor and destination vector are highlighted in orange and purple. Type IIS restriction recognition sequence is represented as rectangle. The cleavage site is represented as triangle. (B) Two design strategies of customizable scar sites. The left diagram shows that the directionality of *Bbs*I and *Bsa*I recognition sequence embedded in the plasmid share the same cut site. The right diagram shows different scar sites generated by *Bsa*I or *Bbs*I site. The cohesive ends can be customized based on how the recognition sites are positioned relative to each other and the Type IIS recognition sequence. Pink box: *Bbs*I regognition sequence. Green box: *Bsa*I recognition sequence. Orange and yellow box: scar sequence generated by either *Bbs*I or *Bsa*I cleavage.

2. Orientation of the restriction recognition site: Type IIS restriction recognition sites are not palindromic, and DNA cleavage occurs on one side of the site (represented by carrot symbols in **Figure 2.2C**). It is important to orient each restriction recognition site so that cleavage occurs between the recognition site and the fragment to be assembled. Cleavage with the Type IIS restriction enzymes will produce an insert that lacks the recognition site.

3. Design of annealing cohesive ends: The spacing between restriction recognition sequence and cleavage site depends on the choice of type IIS enzyme. Similar cohesive ends should be avoided in a single Golden Gate reaction, and the likelihood of two cohesive ends to join during the digestion/ligation reaction can be predicted using thermodynamic models [89].

2.3.4.2 Protocol

1. Measure the DNA concentration of each fragment to be assembled on a NanoDrop spectrophotometer.
2. Calculate the mass of each DNA fragment equivalent to 20 fmol: dilute the DNA stock so each DNA fragment is approximately 20 fmol / µL
3. Prepare a 10 µL reaction mixture by adding the DNA fragments, water and 1 µL 10X T4 ligase buffer. Finally, add type IIS restriction enzyme (10U) and T4 ligase (10U). If performing many reactions in parallel, the water, buffer, and enzymes can be combined to form a master mix immediately before adding the DNA fragments.
4. Mix the reaction by pipetting up and down 3-4 times and centrifuge briefly
5. Incubate the reaction mixture on thermal cycler using the cycling conditions described in Table 2.3.
6. Transform 1-5 µL of each Golden Gate assembly reaction into 20-50 µL chemically competent cells.

Table 2.3. Golden Gate assembly cycling conditions.

Step	Temperature	Time
Initial Digest	37°C	5 minutes
Initial Ligation	16°C	5 minutes
10-30 cycles	37°C	1 minute
	16°C	1 minute
Ligase Heat inactivation	50°C	5 minutes
Restriction enzyme Heat inactivation	65°C*	10-20 minutes*
Hold	4-10°C	∞

*Restriction enzymes have different heat inactivation temperatures. Refer to vendor information to set the proper inactivation time and temperature.

2.4 Notes

2.4.1 Molecular Specifications for Algorithmic DNA Assembly

Choice of type IIS restriction endonuclease. Many type IIS restriction endonucleases can be used in an algorithmic DNA assembly pipeline, including *Bsa*I, *Bbs*I, *Aar*I, and *Sap*I. Selection of suitable restriction endonuclease is influenced by (i) differences in cohesive end size (*i.e.* 3-base overhangs generated by enzymes like *Ear*I vs 4-base overhangs generated by enzymes like *Bsm*BI) and (ii) frequency of recognition site in vectors or parts. Smaller cohesive ends generate smaller vestigial scars, but also decrease the number of fragments that can be reasonably assembled in a single reaction. More commonly, selection of restriction endonuclease is based on the frequency of the recognition sites in the substrate genetic parts or assembly vectors. These sites will have to be removed via synonymous mutations to 'domesticate' genetic parts that will enter the assembly pipeline. For example, *Aar*I recognition sites are less frequent than *Bbs*I

recognition sites in *Streptomyces* genes, so using *Aar*I would be preferred for building multigene pathways from *Streptomyces* genes.

Choice of selectable marker in vector backbones. The vector backbone for initial, intermediate, and final constructs should be tailored for the specific assembly project. One failure mode for every DNA assembly method is the growth of colonies resulting from undigested substrate plasmids. This is easily avoided by changing selection markers in successive 'levels' in a DNA assembly pipeline. For example, 'monocistron' plasmids (**Figure 2.4**) contain a kanamycin resistance marker while 'partial cluster' plasmids contain an ampicillin marker. Using two unique selectable markers is sufficient for creating an infinite cloning loop (**Figure 2.3**). Undigested destination vectors can be easily identified if a lacZα reporter cassette is included between the Golden Gate cloning scars (**Figure 2.3** and **2.4**).

Choice of origin of replication in vector backbones. High-copy origins of replication provide the best plasmid yields following purification, and thus are suggested for all vectors holding intermediate assembly constructs (*i.e.* partial clusters). However, the use of low- or medium-copy origins of replication might mitigate potential toxicity of constructs. The final destination vector should be customized for the expression host. Origin of replication copy number as well as the choice between replicative or chromosome-integrating vectors will depend on the specific project requirements.

2.4.2 Failure Mode, Screening Techniques and Troubleshooting

When screening plasmids to verify the fidelity of DNA assembly reactions, it is good to consider the common failure modes of individual methods. These are summarized in Table 2.4, where we consider Golden Gate cloning from previously sequence-verified plasmid substrates, not PCR products.

2.4.3 Strengths and Weaknesses of Assembly Techniques

PCR-Ligation. The advantage of PCR-ligation is its utility to generate large combinatorial library with relatively small amount of primers. It is particularly suited to building out large libraries of promoter-RBS pairs in a combinatorial fashion. Primer requirement scales with the sum of parts, not the product of parts, so assembling 10 promoter sequences and 10 RBS sequences into 100 combinations only requires purchasing 20 oligonucleotides. Also, the use of blunt-end ligations yields scarless part-

junctions, which can be useful between regulatory sequences in the 5′-UTR that are sensitive to relative spacing. Lastly, this is a restriction enzyme-independent technique which limits sequence constraints. The largest weakness of PCR-ligation is that it is limited to small genetic parts that can be fully encoded in an oligonucleotide primer. Another weakness is the most common failure mode of small deletions at the ligation junction, which require sequence verification and add to the validation costs.

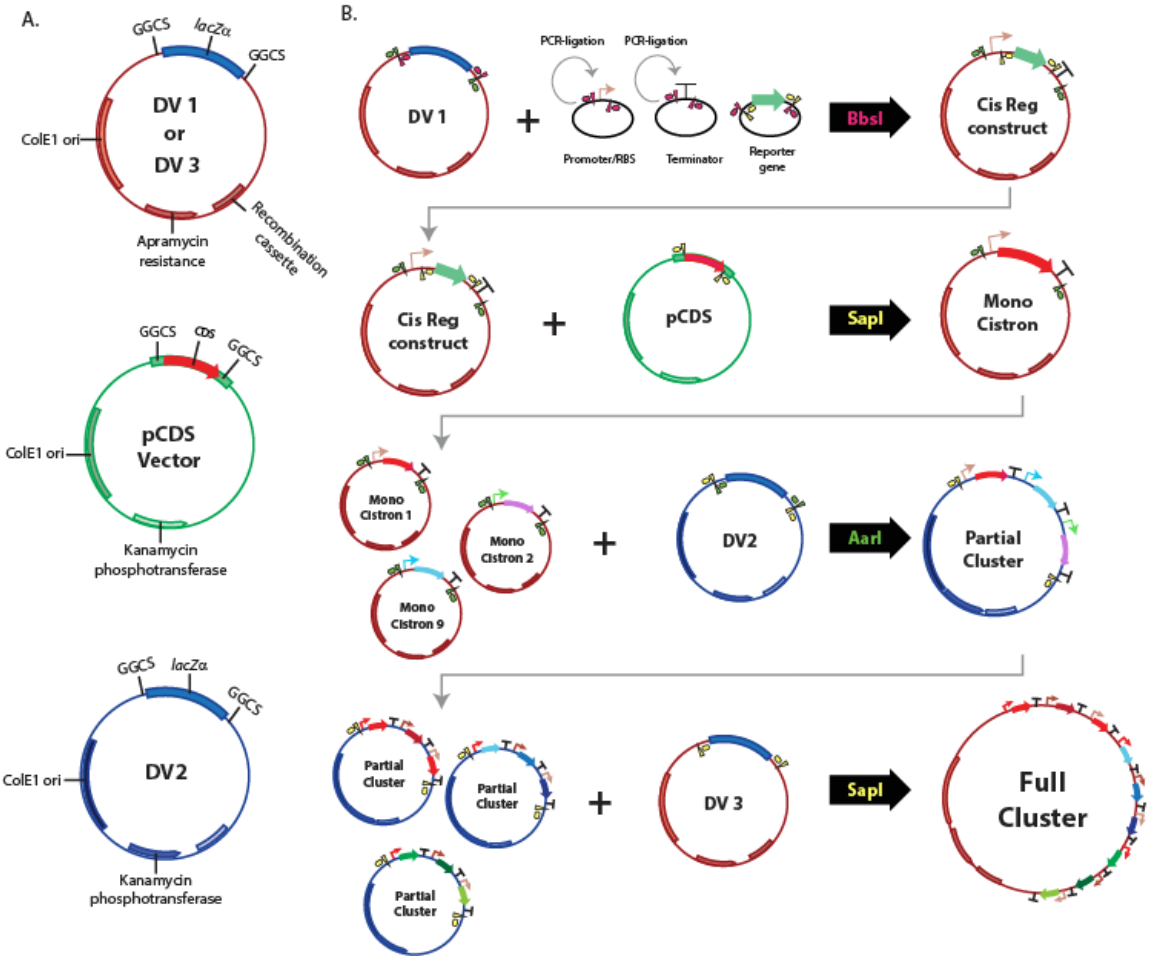


Figure 2.4. Example of an algorithmic DNA assembly pipeline incorporating isothermal assembly, PCR-ligation, and Golden Gate assembly.

(A)Plasmid designs of destination vectors for different DNA assembly levels. The resistance cassette on the destination vectors alternates between apramycin and kanamycin to facilitate selection of desired constructs. The destination vectors were constructed using Isothermal assembly to introduce various combinations of Golden Gate assembly scar site. (B) PCR-ligation is used to build promoter/RBS and terminator part plasmids. Level 1 assembly starts with a DV1, part plasmids of promoter/RBS, terminator and reporter gene, and CisReg construct is assembled by *Bbs*I Golden Gate reaction. Level 2 assembly starts with CisReg plasmid and a target CDS stored in pCDS plasmid, and a monocistron is assembled by a *Sap*I Golden Gate reaction. Level 3 assembly starts with multiple monocistrons and a partial cluster is assembled by *Aar*I Golden Gate

reaction. Level 4 assembly uses multiple partial clusters as substrates and a final cluster is assembled by *Sap*I Golden Gate reaction.Green arrow represents reporter gene. GGCS, Golden Gate Cassette Site. DV, destination vector. The black arrow contains the name of Type IIS restriction enzyme used in individual level of Golden Gate assembly.

Table 2.4. Failure modes and screening methods for assembly methods described here.

Assembly Techniques	Failure Modes	Best Screening Method
PCR-ligation	Small deletions at site of ligation Point mutations in construct introduced through PCR	DNA sequencing
Isothermal Assembly	Point mutations in the construct introduced through PCR or junction gap-filling Low efficiency due to too many fragments, high GC fragments, or extremely long fragments	DNA sequencing
Golden Gate Assembly	Missing fragments	Colony PCR or diagnostic restriction digest

Isothermal Assembly. Isothermal assembly is an incredibly useful method with diverse applications. Primer design is straightforward, and there are few sequence-constraints since it is restriction enzyme-independent. The ease of building scarless constructs allows the genetic engineer to specify every base in a plasmid design. Further, many fragments can be assembled in a single reaction with high efficiency. Drawbacks to isothermal assembly are that it does not lend itself to combinatorial assembly. Joining parts need ~20 bases of homologous sequence for the reaction to proceed meaning that for combinatorial libraries, either (i) a 20 base scar needs to be designed between neighboring genetic parts, or (ii) primers need to be designed to accommodate every possible unique part junction to avoid scars. This latter option would make oligonucleotide requirements scale by the product of parts (*i.e.* a 10 promoters x 10 RBSs library would require ~100 oligonucleotides). Isothermal reactions aimed to produce constructs greater than ~12 kb require many-fold more substrate DNA, so this technique does not lend itself to large construct assembly. Lastly secondary structure in the substrate fragments, for example that caused by high-GC DNA, can reduce the efficiency of this technique.

Golden Gate Assembly. Golden Gate assembly is known for its scalability and high efficiency. First, the ability to define annealing cohesive ends allows either scarless assembly or combinatorial assembly with small scar sequences. This method can readily be used to assemble >10 fragments in a single reaction, although reaction efficiencies decrease with more parts. Of the methods described here, Golden Gate assembly is the best choice for assembling large (>10 kb) constructs. Weaknesses include additional complexity in primer design compared with the other methods that requires substantial practice to become experienced. Also, as a restriction enzyme-dependent technique, substrate parts and vectors need to be domesticated prior to entry into the pipeline. However, this can be leveraged as an advantage, where substrate fragments are domesticated and cloned into a plasmid before entry into the pipeline. Sequencing of these substrate plasmids decreases future validation costs and starting with plasmid substrates increases the efficiency of the reaction.

2.4.4 Sample DNA Assembly Pipeline with Integrated Expression Analysis

Figure 2.4 shows an example DNA assembly pipeline designed to accommodate differences in promoter strength, RBS strength, gene order, and operon occupancy. Prior to the assembly, the destination vectors (DVs) are constructed using isothermal assembly. Libraries of promoter-RBS combinations (represented in a thin arrow) and terminators (represented as Ts) are built using PCR-ligation. The coding sequences are domesticated in the pCDS vector via Golden Gate assembly (details steps shown in the pipeline.) The first step of the pipeline is the *Bbs*I Golden Gate assembly of promoter-RBS, terminator and reporter gene, resulting in a transcription unit containing cis-regulation elements (CisReg construct). The reporter gene in the CisReg construct has two functions. First, it serves as a place holder that will be swapped out in the next assembly stage. Second, it allows the users to characterize expression strength of promoter::RBS combination in the pipeline. Third, the reporter gene is swapped out by the desired coding sequence by Golden Gate assembly using *Sap*I, resulting a monocistronic construct (Cistron). Third, multiple monoscitronic constructs are pieced together by *Aar*I Golden Gate assembly to yield a partial cluster. Third, partial clusters are pieced together by *Sap*I assembly to yield a final cluster. Theoretically, by alternating the last two type IIS restriction endonucleases, *Sap*I and *Aar*I, can create an infinite cloning loop as described in **Figure 2.3**.

Chapter 3 Semi-synthesis of the Neuroprotective Metabolite, Serofendic Acid

The following is a reprint of the article Hsu SY, Perusse D., Hougard T., and Smanski M.J. (2019) “Semi-synthesis of the Neuroprotective Metabolite, Serofendic Acid” *ACS Synth. Biol.*, Sep 5; 8(10), 2397–2403.

Article hyperlink:
https://doi.org/10.1021/acssynbio.9b00261

MJS conceived the project. SYH, DP, TH, and MJS developed methodology, performed investigation and analyzed data. SYH wrote the original draft. SYH, DP, and MJS reviewed and edited the manuscript. Dr. Stephen Harvey from the UMN Mass Spectroscopy Center, Todd Rappe from the UMN NMR Center, and Dr. Victor Young Jr. from the UMN X-ray Crystallography Laboratory for their contributions in acquiring analytical data to support this work.

Summary

Serofendic acid is a natural neuroprotective molecule found in fetal calf serum. It is able to protect neurons against mechanisms of cell death associated with neurodegenerative disease. Because only trace quantities are present in fetal calf serum and complete chemical syntheses are long and inefficient, its development as a therapeutic agent has been slow. We engineered a heterologous metabolic pathway in *Streptomyces* to produce a late-stage synthetic intermediate, *ent*-atiserenoic acid, at high titers. We completed the total synthesis of serofendic acid from this intermediate in four steps.

3.1 Introduction

Acute and chronic neurodegenerative diseases are responsible for 13% of deaths in the USA each year and cause a combined annual economic burden in excess of $300B [90], [91]. A general characteristic of neurodegenerative diseases is tissue damage

caused by programmed cell death. This tissue damage is irreversible, as cells in the central nervous system have limited regenerative capabilities. Neurodegenerative diseases are drug-poor, with only a small number of clinically approved therapeutics available for treatment. Recently, serofendic acid (**1**) (**Figure 3.1a**) has been identified as a diterpenoid natural product [92] that has potent (nanomolar) bioactivity in decreasing tissue damage caused by stroke [93], [94], neurodegenerative disease [95], and other ischemic injuries [96] and improving recovery outcomes in animal models [94].

Despite its promise in pre-clinical investigations, the lack of a scalable source for the compound has limited its development. It is only present in trace quantities in fetal calf serum (3.1 mg isolated from 250 L of fetal calf serum) [92], and total chemical synthesis is challenging due to the nine chiral centers. Several total syntheses have been reported [97]–[100], but the many steps and low overall yields (around 1%) make these impractical to support pre-clinical development or clinical trials.

Heterologous production in a microorganism is an attractive solution to this problem. In recent years, several groups have engineered microbes to produce clinical drugs or drug-precursors, including opioid painkillers [54], anti-malarial agents [101], and anti-cancer agents [102]. Each of these projects benefited from the existence of a known producing organism from which biosynthetic genes could be identified and used in engineering. In contrast, the producer of **1** is unknown. The *ent*-atisane diterpene scaffold is unprecedented as a mammalian metabolite and the required biosynthetic genes are not present in the cow's genome. Without a known producing organism, we had to design and build a metabolic pathway using genes present in public sequence databases.

Here we describe our approach to produce **1** at scales needed for pre-clinical development. We make a late stage diterpene intermediate, *ent*-atiserenoic acid (**2**), via a heterologous metabolic pathway engineered in an industrial host organism. Compound **2** is converted to **1** and analogs via a short chemical synthesis. With optimized strains able to produce **2** at high titers, we have established a reliable and economical source of the *ent*-atisane scaffold of **1** to support mechanistic and structure-activity-relationship (*i.e.* SAR) studies.

3.2 Results

We first set out to identify enzymes that could be used for the biosynthesis of **1**. Atisane-derived scaffolds are found in other bioactive natural products including *ent*-19-hydroxy-atis-16-ene-3,14-dione (**3**), platencin (**4**), and spiramines (**5**) (**Figure 3.1a**). The lack of a known organism producing **1** precluded genome mining as an approach to obtain biosynthetic genes. Instead, we performed a retro-(bio)synthetic analysis using reactions either described in the literature or proposed based on chemical logic (**Figure 3.1b**). The methylsulfinyl group at C17 and the hydroxyl group at C15 are predicted to be late-stage tailoring modifications of **2**. This compound is predicted to derive from a six-electron oxidation of the C19 methyl of *ent*-atiserene (**6**), analogous to the first tailoring step in gibberellin biosynthesis [103]. As a labdane diterpene, **6** originates from a type-I cyclization of an *ent*-copalyl diphosphate intermediate (**7**). Finally, **7** would come from a type-II cyclization of geranylgeranyl diphosphate (**8**), the common intermediate in all diterpene metabolism. Our preliminary goal was to reconstitute a biosynthetic pathway to produce **2**, which can be converted to **1** via a four-step chemical synthesis.

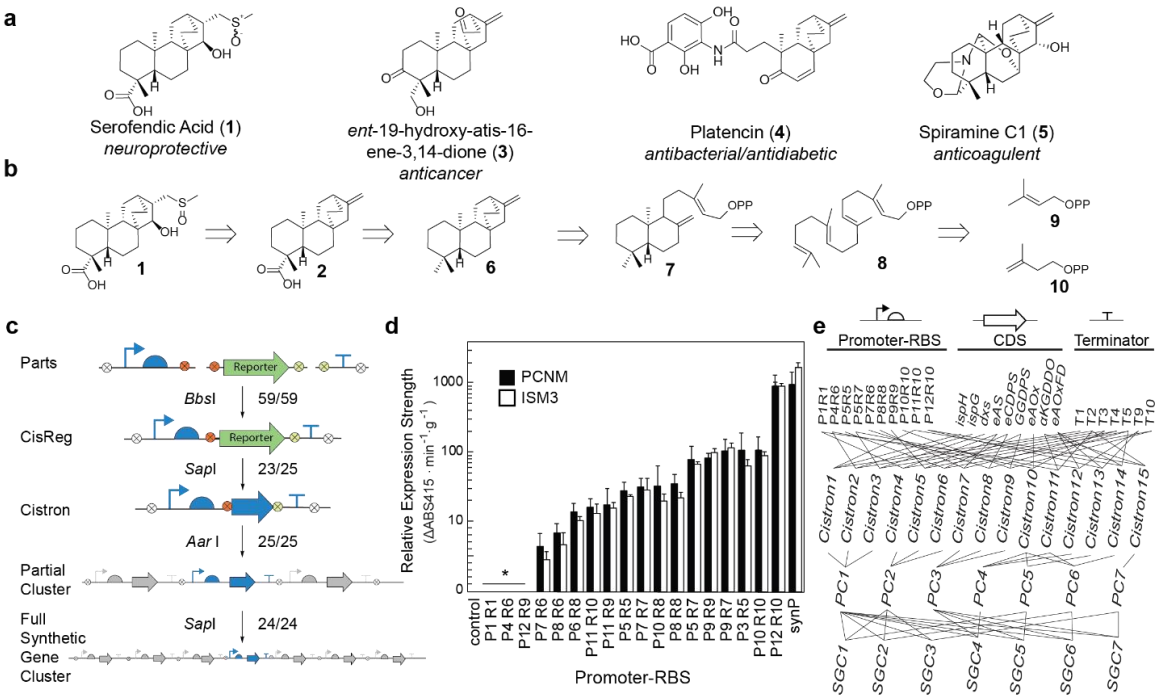


Figure 3.1. Design and construction of biosynthetic pathways.

(a) Examples of bioactive natural products containing atisane scaffold [93], [104]–[106]. (b) Retro-(bio)synthetic analysis of **1** from primary metabolites required for isoprenoid biosynthesis. (c) Schematic representation of DNA assembly pipeline from genetic parts (top) to synthetic gene clusters (SGCs, bottom). Key type IIS restriction enzymes used in assembly (left of arrows) and representative assembly efficiencies (right of arrows; reported as a fraction of correct constructs over total colonies tested) are shown on the schematic. Circle with x in it, 4-bp MoClo assembly scar site; orange circle, start codon; yellow circle, stop codon. (d) Characterization of library of promoter-RBS combinations in 60-hour fermentation in two growth media. Black bar, PCNM medium; white bar, ISM3 medium. Error bars denote standard deviation from three replicates over multiple days. *expression detected after 96-hour fermentation (Figure S4). (e) DNA assembly graph for initial seven SGCs built for this study. P denotes promoter; R denotes RBS; T denotes terminator; *e*AS, *ent*-atiserene synthase; *e*CDPS, *ent*-copalyl diphosphate synthase; GGDPS, geranylgeranyl diphosphase synthase; αKG-DDO, α-ketoglutarate dependent dioxygenase; *e*AOx, *ent*-atiserene oxidase;; *e*AOxFD, *ent*-atiserene oxidase ferredoxin.

We surveyed the literature and existing genome sequence databases to select genes encoding the desired enzymes. We focused our search in *Streptomyces* genomes, as these bacteria are known to host a large diversity of metabolic enzymes. Further, there are sufficient tools for genetic manipulation of model *Streptomyces* that allow us to avoid many hurdles to heterologous expression in distant taxa. Several geranylgeranyl diphosphate synthases (GGDPS) have been described in diterpenoid-producing *Streptomyces* that can produce compound **8** from primary metabolites. Further, putative *ent*-copalyl diphosphate synthases (*e*CDPS) were identified in genomes on known labdane-producing *Streptomyces* by phylogenetic analysis of previously characterized *ent*-copalyl diphosphate synthase, PtmT2 (**Supplementary Figure A1.1**) [107], [108]. We identified several putative *ent*-atiserene synthases (*e*AS) in the sequence databases based on their homology to the platencin biosynthetic gene *ptmT1* (**Supplementary Figure A1.2**). The required six-electron oxidation of a methyl group on **6** has never been reported, although a P450 monooxygenase CYP117 was reported to oxidize C19 of the methyl group of *ent*-kaurene, a similar labdane-diterpene scaffold to **6** [109]. We selected several putative oxidoreductases and P450 monooxygenases from gene clusters containing *ptmT1* homologs as likely candidates to perform this reaction. Since our initial experiment, Shen et al. have reported the activity of one of these enzymes, the α-ketoglutarate dependent dioxygenase, PtmO6, as a specific C7 hydroxylase [110].

Genes were recoded for entry into a multi-gene DNA assembly pipeline designed for engineering bacteria of the genus *Streptomyces* (**Figure 3.1c** and **Supplementary**

Figure A1.3). To control and vary individual gene expression of the synthetic gene clusters, we used previously published cis-acting regulatory elements including synthetic promoters [111] and bacterial terminators [56]. A set of synthetic ribosomal binding site sequences of varying translational rates were generated by RBS calculator [112] (**Supplementary Data File A1, Worksheet 1**). Promoter and RBS were concatenated as a single genetic part in our genetic design and the collection of promoter-RBS combinations were characterized by β-glucuronidase (GUS) assay. To minimize the confounding effects of medium composition on small molecule titers/yield, we only included used cis-regulatory elements that gave consistent expression of a reporter gene in two growth media, ISM3 and PCNM (**Figure 3.1d** and **Supplementary Figure A1.4**). Combining these natural and synthetic genetic elements using one-pot type IIS restriction digestion/ligation reactions [73], an initial set of seven synthetic gene clusters (SGCs) were produced (**Figure 3.1e**). We opted for an iterative DNA assembly pipeline comprising modest (4-5 part) assembly reactions to keep the efficiency high at each step. After verifying a minimum of 25 reactions of each step, our efficiency (defined as percentage of colonies on a plate containing the desired plasmid construct) ranged from 92% (23/25) to 100% (59/59) (**Figure 3.1c**). SGC1-7 vary according to the relative expression levels of genes encoding the terpene scaffold and according to gene composition for the oxidative tailoring reactions. These SGCs were introduced to a *Streptomyces albidoflavus* J1074 (formerly *S. albus* J1074) [113] heterologous host and culture broth was analyzed by LC-MS.

Each recombinant strain produced a unique chemical profile with two, SGC4/5, producing a predominant peak with a mass-to-charge ratio of 301 amu (**Figure 3.2**). This mass is consistent with the $[M-H]^-$ ion of our desired product, **2**. The structure was confirmed following purification from a 50 mL liquid culture and 1D and 2D NMR spectroscopy (**Supplementary Figure A1.5-A1.8**) [114]. Other strains produced prominent peaks with masses in the range expected for diterpenoid metabolites that were isolated and their structures determined by spectroscopic methods. These compounds, (4*E*,8*E*,12*E*)-14-hydroxy-4,8,12-trimethyltetradeca-4,8,12-trienoic acid (**11**) and *ent*-isocupressic acid (**12**) **(Figure 3.2** and **Supplementary Figure A1.9** - **A1.15**), are

presumably shunt metabolites derived from biosynthetic intermediates **7** and **8** (**Supplementary Figure A1.16** - **A1.17**).

Comparing the genotypes to the chemical profiles from this initial library revealed several insights important for the production of compound **2**. First, we identified *WP_030985276*, an ortholog of *ptmO2*, as a cytochrome P450 capable of catalyzing the complete six-electron oxidation from compound **6** to **2**. Only the subset of SGCs containing this gene and its cognate ferredoxin reductase, *WP_030990007*, accumulate **2** or the corresponding oxidized copalyl derivative, **12** (**Figure 3.2**). Based on primary sequence similarity to characterized cytochrome P450's, *WP_030985276* is assigned as the first member of a new CYP family, CYP1408A1 (D.R. Nelson, personal communication) and is given the shorthand name *ent*-atiserene oxidase (*e*AOx). 2D NMR analysis of isolated metabolites confirms the regio- and stereo-specificity of the oxidation catalyzed by *e*AOx (**Supplementary Figure A1.8**).

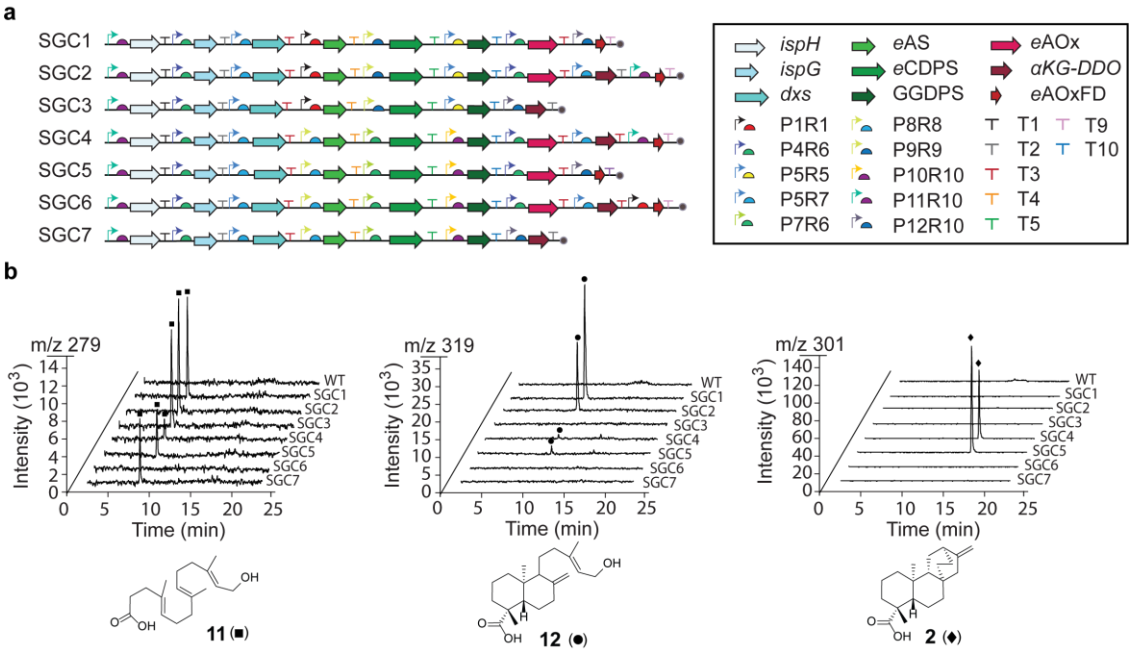


Figure 3.2. LC-MS analyses of key metabolites produced by seven SGCs.

(a) Schematic of seven SGC designs and keys of genetic parts. (b) The selected ion chromatograms correspond to m/z values for the oxidized shunt metabolite (**11**) of **8** (left), the oxidized shunt metabolite (**12**) of **7** (center) and desired final product **2** (right).

The importance of properly balancing gene expression levels is evident from comparing SGCs with identical gene content. SGC1 and SGC5 contain the same set of genes, but the genes required for terpene cyclization (GGDPS, *e*CDPS, and *e*AS) are controlled by promoters and RBSs of different strength (**Supplementary Data File A1, Worksheet 2**). In SGC1, *e*AS is controlled by promoter-RBS P1R1 with basal activity (2.7 ± 1.4 relative expression unit (REU)), and the shunt metabolite **12** is the major product. In contrast, the SGC5 design expresses *e*AS from a medium-strength promoter P8R8 (21.9 ± 4.6 REU), and the desired compound **2** is produced at 50 mg/L. The shunt metabolite **12** was only observed in strains containing the P450 monooxygenase, *e*AOx, suggesting this enzyme is promiscuous and can oxidize **7** or **6**. Thus, both the *e*AS and the *e*AOx compete for the common substrate, **7**. The expression ratio of these two enzymes determines whether **12** or **2** preferentially accumulates during fermentation.

Lastly, we see evidence that perturbing the expression of on-pathway genes can influence shunt metabolite production by endogenous host enzymes. Compound **11** is a 17-carbon metabolite that we predict arises from (i) the β-oxidation and degradation of **8** to remove the terminal three carbons, and (ii) hydrolysis of the diphosphate tail (**Supplementary Figure A1.16**). Because **11** accumulates regardless of the complement of oxidases encoded by the synthetic gene cluster, it is most likely a shunt product made by endogenous host enzymes and not by oxidases encoded within the SGC. Despite this, the titer of **11** is affected by the relative ratio of the geranylgeranyl diphosphate (GGDP) synthase and the *e*CDPS, as can be seen by comparing production of **11** in isogenic SGCs 1-5 and 3-7. In both cases, SGCs in which GGDPS is expressed at higher levels (22.9 ± 1.0 REU in SGC1 and SGC3 versus 88.0 ± 14.1 REU in SGC5 and SGC7) accumulate **11** at higher concentrations at the expense of desired downstream metabolites. These results reinforce the observation made by others [101], [102] that designing biosynthetic pathways such that every gene is maximally expressed is rarely optimal. Additionally, these results show metabolic shunting by endogenous host enzymes can be ameliorated by fine-tuning expression levels within the biosynthetic gene cluster.

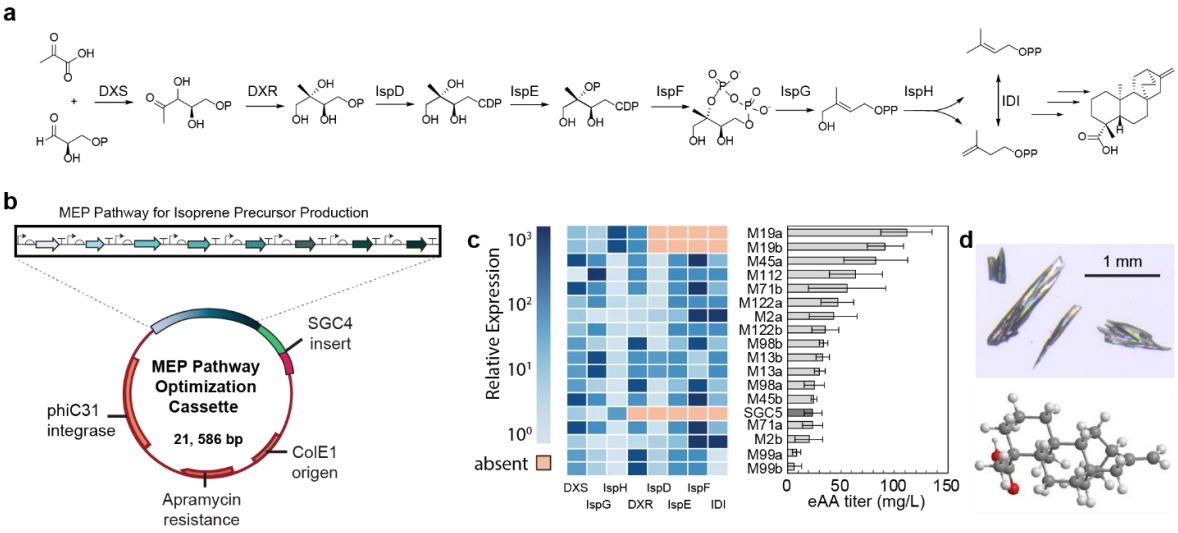


Figure 3.3. Titer improvement through precursor pool pathways.

(a) The MEP pathway for isoprenoid precursor supply. (b) Plasmid components of MEP pathway optimization SGCs. (c) Relative expression level of the MEP pathway genes in SGCs and corresponding titer of 2. Error bars denote standard deviation from three replicates measured on multiple days. (d) Photographs of crystals (top) obtained from large-scale production of 2 and crystal structure obtained by X-ray diffraction (bottom).

To further increase titers of compound **2**, we bolstered the supply of precursor metabolites by adding genes for the entire methylerythritol phosphate (MEP) pathway plus the isopentenyl pyrophosphate-dimethylallyl pyrophosphate (IPP-DMAPP) isomerase, *idi*, to the SGC4 construct. Each of the eight genes were cloned from *Streptomyces* sp. NRRL S-1813 into our coding DNA sequence (CDS) domestication plasmid, pMJS_CDS (**Supplementary Figure A1.3** and **Supplementary Table A1.1**). A library of nine SGCs was built in which the relative expression levels of each of the eight new genes was perturbed (**Supplementary Data File A1**). Following conjugation into our *S. albidoflavus* J1074 host, two unique recombinant strains of each genotype (designated 'a' and 'b') were fermented in 0.5 mL ISM3 medium for five days. The titer of **2** was determined by LC-MS. Eight out of nine constructs produced more **2** than strains harboring SGC5 (**Figure 3.3c**). In 500 mL shake-flask cultures, strain M19a produced 547.9 ± 204.8 mg/L (**Supplementary Figure A1.18**). At this titer, **2** spontaneously crystalizes from an ethyl acetate crude extract, giving gram-scale quantities of >99.9% pure natural product in

laboratory-scale fermentations without the need for chromatography. Additionally, our use of cis-regulatory elements with consistent behavior across growth media allowed us to see prodigious production in every medium tested, with titers correlating to cell growth/doubling (**Supplementary Figure A1.19**).

Using this high-titer bacterial strain to access compound **2**, we next used chemical synthesis to produce **1**. The methyl ester of **2** is a late stage intermediate in several total syntheses of **1** [100]. Esterification in acidic conditions led only to methanolysis of the C16 double bond (compounds **14** and **15**, **Supplementary Figure A1.20-A1.27**). In basic conditions with iodomethane as a methylating agent, we attained the methyl ester, **16,** without any nucleophilic attack on the double bond (**Supplementary Figure A1.28-A1.32**). This constitutes a formal synthesis of **1**, as **16** has been converted to **1** with an overall yield of 17% in 5 steps [100].

We hypothesized that protecting the carboxylic acid as a methyl ester was unnecessary and designed a more direct total synthesis of **1** using compound **2** as our starting material (**Figure 3.4**). Allylic oxidation of **2** using benzeneseleninic anhydride produced enone **17** with same yield previously reported (75%) [100]. The thiomethylation step as described in literature [100], monitored by LC-MS, was slow and inefficient using **17** as substrate. Four successive additions of a methanethiolate in aqueous solution (15% m/v, 2eq) every 15 min led to a quantitative conversion of enone to a 1:1 mixture of α- and β-thiomethyl ketones **18** and **19**. Despite several attempts, this mixture of diastereomers was inseparable, similar to what was reported for their methyl ester derivatives [100]. Borane proved to be unsuitable as a reducing agent, likely due to attack of the carboxylic acid moiety in **18** and **19**. However, sodium borohydride led to the exclusive reduction of the ketone at C15 to form diastereomers **20** (10% yield from **17**), **21** (30% yield from **17**), **22** (39% yield from **17**) and **23** (11% yield from **17**) which were separable by flash column chromatography or semi-preparative HPLC. Finally, use of Davis oxaziridine [99], [100], [115] on the isomer **20** led to a mixture of serofendic acids A and B (**1A** and **1B**) in similar ratio reported in literature (**Supplementary Figure A1.33-A1.58**). Davis oxidation of diastereomers **21-23** produced serofendic acids isomers **24A** and **24B** (**Supplementary Figure A1.59-S64**) and putative isomers **25** (15R,16R) and **26** (15S,16R) which were not fully characterized.

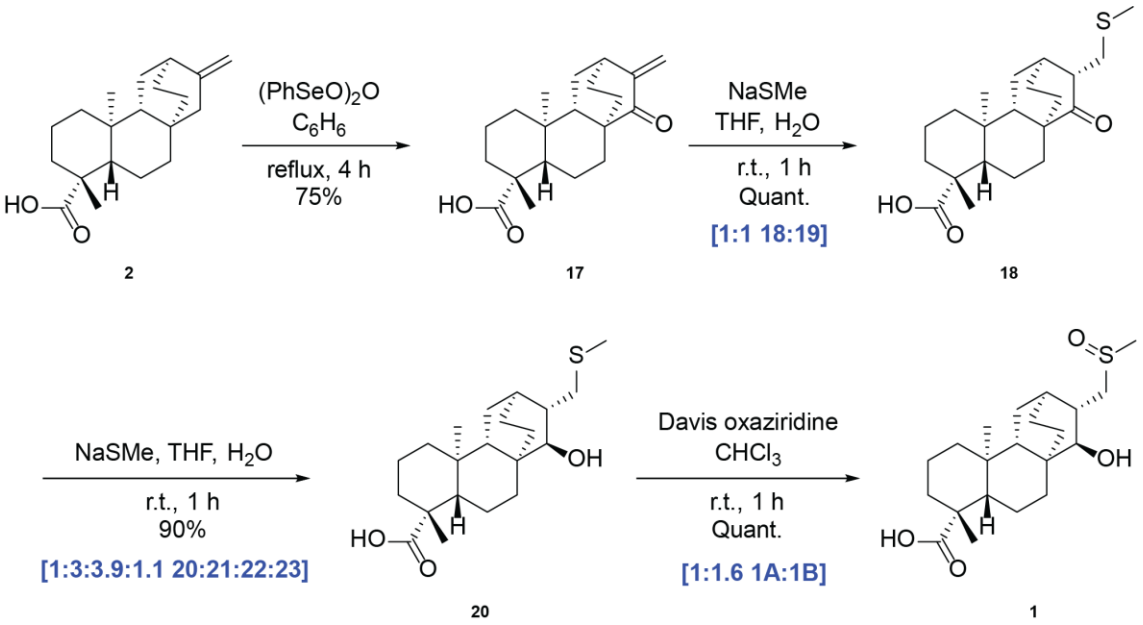


Figure 3.4. Total synthesis of 1 from 2.

Reaction conditions are provided above arrows and isolated yields are provided below. The blue bracketed text describes ratios of related diastereomers.

3.3 Discussion

We have created a pipeline for constructing entire multi-gene SGCs from genetic parts and expressing them in model *Streptomyces* hosts [116]–[120]. DNA synthesis and assembly technology has improved dramatically in the past decade to enable genetic engineering projects at a much larger scale than previously possible [23]. We designed our pipeline to be compatible with *Streptomyces* hosts to take advantage of their diverse metabolic potential and importance in drug development. With a moderate set of synthetic cis-regulatory elements, we are able to perturb relative expression levels of pathway genes across three orders of magnitude. Further, these regulatory elements behave consistently in different growth media, allowing us to focus efforts on genetic optimization instead of media optimization. We have made several modifications of the DNA assembly pipeline we reported previously [64], including: (i) using Type IIS restriction endonucleases that are rare in actinomycete genes to reduce the requirement for *de novo* DNA synthesis,

(ii) creating an intermediate cloning step that allows all unique combinations of cis-regulatory elements to have their strength quantified prior to inserting the biosynthetic CDS, and (iii) creating destination vectors that allow direct integration into the chromosome of *Streptomyces* spp.

A defining feature of **1** compared to recent heterologous production targets is the fact that the producing organism is unknown. The *ent*-atisane carbon scaffold is unprecedented in mammalian systems, and has only been seen before in a handful of plant and microbial metabolites. As such, we did not have a natural host genome from which to mine biosynthetic genes [54], [101], [102], [121] and instead had to turn to genes present in public databases. We used genes from the publicly available *Streptomyces* sp. NRRL S-1813 that are orthologous to those from the platencin biosynthetic gene cluster [17], [110], [122]. Our engineered production of **2** provides access to **1** as well as other bioactive *ent*-atisanes (**Figure 3.1a**).

Completing the biosynthesis of **1** from **2** will require the discovery and incorporation of enzymes to install the C15 hydroxyl group and the C17 methylsulfinyl group. Several plant *ent*-atisanes with the needed C15 modification have been described, including the Eriocatisin A [123] and *ent*-(1S,6S,15S)-1,5,6,15-tetrahydroxy-14,19-diacetyloxylatiserene [124] and their producers are candidates for future genome-mining to identify a suitable enzyme (**Supplementary Figure A1.65**). Methylthiolation is often seen in drug metabolism studies and there are several theories on the molecular origin of the methylthiol group [125]. Since **1** is isolated from fetal calf serum as an epimeric mixture at the sulfoxide of diastereomers, we predict that the final oxidation of the methylthiol to a methylsulfinyl group occurs non-enzymatically. An intriguing possibility is that production of **1** in nature involves a biosynthetic pathway split between different hosts, with the carbon scaffold made by a plant or microbe and the final tailoring steps occurring in the cow. While efforts to complete the biosynthesis of serofendic acid are ongoing in our lab, **2** is an attractive intermediate for medicinal chemistry efforts, as its olefin moiety provides a reactive handle to facilitate chemical derivatization for SAR studies.

In conclusion, we report the design and engineering of a synthetic gene clusters to enable the high-titer semi-synthesis of a potent neuroprotectant, **1**. We have completed a formal and total synthesis of serofendic acid from our microbially-derived intermediate,

and demonstrated the semi-synthesis of several potent structural analogs. This represents the first sustainable route towards developing new neuroprotectants based on the serofendic acid scaffold.

3.4 Materials and Methods

3.4.1 Strains and Chemicals

E. coli DH5α and NEB® Stable were used for routine cloning. *E. coli* ET12567/pUZ8002 was used for intergenic conjugation. *S. albidoflavus* J1074 was used as a heterologous host for screening SGCs. LB medium (10 g/L tryptone, 5 g/L yeast extract, 10 g/L NaCl; MP Biomedicals # 113002022) supplemented with proper antibiotics was used for strain maintenance and plasmid construction in *E. coli* strains. Liquid media for *Streptomyces* maintenance and fermentation are R2YE medium (103 g/L sucrose, 10 g/L glucose, 0.25 g/L K_2SO_4, 5 g/L yeast extract, 0.1 g/L Difco casamino acid, 100 ml TES buffer (5.73%, pH 7.2), 10 ml KH2PO4 (0.5%), 80 ml $CaCl_2.2H_2O$ (3.68%), 15 ml L-proline (20%), 2 ml trace element solution and 5 ml NaOH (1N)), ISM3 medium (15 g/L yeast extract, 10 g/L malt extract, 0.5 g/L $MgSO_4$, 0.3 g/L $FeCl_3 \cdot 6H_2O$, 20 g/L glucose, adjusted to pH 7.0 with NaOH), PCNM medium (6 g/L yeast extract, 15 g/L malt extract, 6 g/L glucose, 20 g/L MOPS, adjust pH to 7.4, then add trace element aseptically), and TSB (30 g/L tryptic soy broth). Solid medium for *Streptomyces* spore isolation is IWL-4 (37 g/L Difco ISP medium 4, 0.5 g/L yeast extract, 1 g/L tryptone). Fermentations at 500- and 50-mL scale were supplemented with 3% w/v hydrophobic resin (Amberlite - XAD16; Alfa Aesar L19565-36). Unless stated otherwise, antibiotics apramycin, nalidixic acid, kanamycin, and chloramphenicol were supplemented to liquid or solid media at final concentration of 50 µg /mL, 25 µg /mL, 50 µg /mL and 25 µg /mL, respectively. Media components and all other chemicals were purchased from standard commercial sources.

3.4.2 Plasmid Construction

Promoter and RBS sequences were chemically synthesized as ultramer® oligonucleotide primers (Integrated DNA Technologies, Coraville, IA). Sequence of

synthetic promoter, s*ynP,* was synthesized as gblocks (Integrated DNA Technologies, Coraville, IA). Combinations of promoters and RBSs were organized so that there were no additional sequences in between promoter and RBS sequences. All promoter-RBS parts were flanked by one of the MoClo scar sequences (4 bp; upstream) and AATG (downstream, start codon underlined), CDS parts were flanked by AATG (upstream, start codon underlined) and **TGA**T (downstream, stop codon bolded), and terminators parts were flanked by **TGA**T (upstream, stop codon bolded) and one of the MoClo scar sequences (4 bp; downstream). These overhang sequences correspond to 5' overhanging single-stranded cohesive ends when digested with type IIs restriction enzymes. Promoter-RBS part library and terminator part library were constructed by PCR-ligation using pMJS2AF as the template using published protocol [73]. Standard PCR was set up using Q5® High-Fidelity DNA polymerase (New England Biolabs Inc.; Cat. No. M0491S) and manufacturer's recommended protocols unless stated otherwise.

Each CDS was PCR-amplified and domesticated by *Aar*I Golden Gate assembly. Genomic DNA of *Streptomyces* sp. NRRL S-1813 was extracted with standard protocol [126] and used as template for PCR amplification of target CDS. Each CDS was PCR-amplified with gene specific primers flanked by *Aar*I restriction sites and a four-nucleotide *Aar*I docking sequence. If the native CDSs had internal *Sap*I and/or *Aar*I restriction sites, they were removed by introducing synonymous mutations by additional PCR amplifications. The mutagenic primers were designed so that each contained homologous sequence to at least 18 nucleotides 5' of the internal *Aar*I /*Sap*I sites, a few nucleotides that interrupt *Aar*I /*Sap*I site with a single base substitution, an *Aar*I restriction site and a four-nucleotide *Aar*I docking sequence. The amplified fragments each had unique 5' overhang sequences to allow orderly and seamless assembly. The PCR products were analyzed by gel electrophoresis and extracted with standard gel purification protocol (Zymoclean Gel DNA Recovery Kit, Zymo Research Cat. No. D4007). Domestication of PCR-amplified CDS were done by *Aar*I golden gate reaction with 30 ng of pCDS, three equimolar amounts of PCR-amplified CDS fragments and 2 µL of 2X *Aar*I master mix (Thermo Fisher Scientific, GeneArt™ Type IIs Assembly Kit, *Aar*I; Cat. No. A15916). The incubation parameters are described below: 37^{o}C initial digest for 15 min, initial ligation at 16^{o}C for 5 min, then cycling between 37^{o}C for 1 min and 16^{o}C for 1 min for at least 20

cycles. Half of the Golden Gate assembly reaction was introduced to *E. coli* via chemical transformation. The clones were verified by Sanger sequencing.

Three type of destination vectors are constructed: level 1 destination vector for cistron storage, level 2 destination vector for partial cluster storage, and level 3 destination vector for full cluster storage. Level 1 destination vector library was constructed by isothermal assembly with two PCR products, *lacZ* fragment and a linearized vector with apramycin resistance cassette. Level 1 destination vector library has unique combinations of upstream and downstream MoClo scar flanked by *Bbs*I and *Aar*I recognition sites. PCR was performed with Q5® High-Fidelity DNA Polymerase. First, the linearized vector was amplified from pMJS1000 using primers that contain vector-specific sequences and *Aar*I recognition site. A *lacZ* fragment was amplified from pMJS2AF using oligonucleotide primers that contain *lacZ*-specific sequences, a *Bbs*I recognition site, MoClo scar sequence, *Aar*I recognition site, and 24 nucleotides homologous to the linearized vector. Total of 30 unique, non-redundant combinations of upstream and downstream MoClo sites flanking *lacZ* was analyzed by gel electrophoresis. The PCR fragments were gel purified. Isothermal assembly was used to piece together the linearized vector and *lacZ* fragment. Specifically, 100 ng of PCR-linearized vector and three equimolar amounts of *lacZ* fragment were used in individual isothermal assembly (15 µL isothermal assembly pre-mix and bring the final volume to 20 µL). The isothermal assembly were incubated at 50^{o}C for 1 h. The isothermal assembly mixtures were introduced into *E. coli* DH5α by chemical transformation. The clones were verified via Sanger sequencing.

Level 2 (partial cluster) destination vector library was constructed by *BbsI* golden gate assembly of PCR-amplified *lacZ* fragment and pMJS1AC, a vector of kanamycin resistance. *LacZ* is amplified from pMJS1AC using oligonucleotide primers that contain *lacZ*-specific sequences, an *Aar*I recognition site, MoClo scar sequence, a *Sap*I recognition site, and a *Bbs*I recognition site. Total of 30 unique, non-redundant combinations of upstream and downstream MoClo sites flanking *lacZ* was analyzed by gel electrophoresis. The PCR fragments were extracted by gel purification. Forty femtomoles of pMJS1AC and three equimolar amounts of purified *lacZ* fragment were pieced together by *Bbs*I Golden Gate assembly. In each reaction, 5 U *Bbs*I (New England BioLabs; Cat. No. R0539S), 5 U DNA T4 ligase (Promega; Cat. No. M1794), 1 µL 10x T4 ligase buffer,

DNA substrates and ddH_2O were mixed together to make a 10-µL reaction. The incubation parameters are the same as the *Aar*I Golden Gate assembly. Half of the Golden Gate assembly reaction was introduced to *E. coli* via chemical transformation. The clones were verified by Sanger sequencing.

Level 3 (full synthetic gene cluster) destination vector library was constructed by isothermal assembly with two PCR products, *lacZ* fragment and a linearized vector with apramycin resistance cassette. Level 3 destination vector library has unique combinations of upstream and downstream MoClo scar sequence flanked by *Sap*I and *Aar*I recognition sequences. PCR was performed with Q5® High-Fidelity DNA Polymerase. First, a *lacZ* fragment was amplified from pMJS2AF using oligonucleotide primers that contain *lacZ*-specific sequences, a *Sap*I recognition site, MoClo scar sequence, an *Aar*I recognition site, and 24 nucleotides homologous to the vector. The linearized vector was amplified from pMJS1000 using primers that contain vector specific sequences and an *Aar*I recognition site. Total of 30 unique, non-redundant combinations of upstream and downstream MoClo sites flanking *lacZ* was analyzed by gel electrophoresis. The PCR fragments were purified by gel extraction. Isothermal assembly was used to piece together the linearized vector and *lacZ* fragment as described in the level 1 destination vector assembly. The isothermal assembly mixtures were introduced into *E. coli* by chemical transformation. The clones were verified via Sanger sequencing.

3.4.3 Synthetic RBS Design

Ten of novel RBS sequences were designed using RBS calculator (https://www.denovodna.com/software/). The following parameters were used: no pre-sequence was used; *Streptomyces*-codon optimized eGFP was used as protein coding sequence; free energy model version 2.x was used; 16S rRNA of *S. albidoflavus* J1074 was used. RBSs were generated with a range of predicted strengths, but actual strengths were determined by reporter assay in **3.4.4**.

3.4.4 Promoter-RBS Strength Characterization by GUS Assay

The GUS assay protocol was modified from a previous study [111]. Briefly, the synchronization culture of each *S. albidoflavus* J1074 GUS strains was used to inoculate 2 mL PCNM and ISM3 in 24-well MTPs. The fermentations were incubated in 28°C shaker with 250 rpm for 60 h. The cell pellets were harvested in pre-weighted 1.7 mL Eppendorf microcentrifuge tubes at 3,000 rpm for 2 min. The supernatant was removed, and the cell pellets were washed with 4°C sterile ddH_2O for at least five times or until the supernatant was clear to make sure the residual fermentation media were removed as much as possible (centrifuged at 3,000 rpm for 1 min per wash). The cells were resuspended in 0.5 mL sterile ddH_2O and frozen at -80°C for 24 h. The frozen samples were then freeze-dried for 24 h or until the moisture was removed completely. The freeze-dried samples were weighted on precise scales and the dry cell weight (DCW) was calculated.

All samples and buffers were incubated on ice before measurement. During the lysis step of GUS assay, the dried cell samples of *S. albidoflavus* J1074 GUS strains that predicted with low activity were incubated with 450 µL lysis buffer (negative vector control, P1R1, P1R3, P4R6, p6R8, P3R5, P5R7) for 20 min at 37°C. The other dried cell samples were incubated in 900 µL lysis buffer (50 mM phosphate buffer [pH 7.0], 5 mM dithiothreitol (DTT], 0.1% Triton X-100, 1 mg/mL lysozyme) for 20 min at 37°C, and these lysates were diluted two-fold by mixing 200 µL lysate with 200 µL base buffer (50 mM phosphate buffer [pH 7.0], 5 mM DTT, 0.1% Triton X-100). The samples were centrifuged at 21,300 rcf for 10 min at 4°C. Samples of 100 µL supernatant from each lysate were transferred to 96-well plate (Sarstedt microtest plate 96 well F; order no. 82.1581) and keep on ice until 100 µL GUS buffer (50mM phosphate buffer [pH 7.0], 5mM DTT, 0.1% Triton X-100 supplemented with 2 mM *p*-nitrophenyl-β-D-glucuronide) was added to each lysate. The absorbance was immediately measured at 415 nm at 25°C for 40 min by plate reader (SpectraMax® Plus 384 Microplate Reader). As a reference, 100 µL of supernatant was mixed with 100 µL base buffer. The slope of the linear range of the absorption curve was used to calculate the enzymatic activity. The relative strength of the promoter was calculated as ΔABS_{415} per minute per gram of DCW.

3.4.5 Bottom-up Hierarchical DNA Assembly of SGCs

*Bbs*I golden gate assembly was used to construct CisReg plasmid. Level 1 destination vector and part plasmids of promoter-RBS, transcriptional terminator, and reporter gene (*Streptomyces* codon-optimized *egfp* or *uidA*) were the substrates. Forty femtomoles of level 1 destination vector, three equimolar amounts of part plasmids, 5 U *Bbs*I, 5 U T4 DNA ligase, 1 µL 10x T4 DNA ligase buffer and water were added to bring the final reaction volume to 10 µL. The incubation parameter is the same as the previous *Aar*I Golden Gate reaction. Half of the Golden Gate assembly reaction was introduced to *E. coli* via chemical transformation. The clones were verified by Sanger sequencing.

*Sap*I Golden Gate assembly was used to construct monocistronic plasmids with CisReg plasmids and pCDS plasmids. Forty femtomoles of CisReg plasmid, three equimolar amounts of pCDS plasmid, 5 U *Sap*I, 5 U T4 DNA ligase, 1 µL 10x T4 DNA ligase buffer and water were added to bring the final reaction volume to 10 µL. The incubation parameter is the same as the previous *AarI* Golden Gate reaction. Half of the Golden Gate assembly reaction was introduced to *E. coli* via chemical transformation. The monocistrons were verified by either colony PCR using OneTaq® PCR (New England Biolabs Inc.; Cat. No. M0480S) or Sanger sequencing.

*Aar*I Golden Gate assembly was used to construct partial clusters with monocistronic plasmid and level 2 destination vector. Thirty nanogram of level 2 destination vector, three equimolar amounts of monocistronic plasmids, and equivolume amount of 2x *Aar*I master mix were used to set up *Aar*I golden gate reaction. The incubation parameter is the same as the previous *Aar*I Golden Gate reaction. Half of the Golden Gate assembly reaction was introduced to *E. coli* via chemical transformation. The monocistrons were verified by colony PCR.

*Sap*I Golden Gate assembly was used to construct full cluster plasmids with partial plasmids and level 3 destination vectors. Forty femtomoles of level 3 destination vector, three equimolar amounts of partial cluster plasmids, 5 U *Sap*I (New England Biolabs Inc; Cat. No. R0569S), 5 U T4 DNA ligase, 1 µL 10x T4 DNA ligase buffer and water were added to bring the final reaction volume to 10 µL. The incubation parameter is the same as the previous *Aar*I Golden Gate reaction. Half of the Golden Gate assembly reaction was introduced to *E. coli* via chemical transformation. The monocistronic part were verified

by colony PCR. The colony PCR primers are the same as the CDS cloning primers. One gene from each cassette was targeted for colony PCR.

3.4.6 Intergenic Conjugation

SGCs were introduced into conjugation donor *E. coli* ET12567/pUZ8002 via electroporation (BioRad electrocuvette, 1 mm gap; 1.8 kV; BioRad Gene Pulser Xcell electroporation system). The electroporation mixtures were recovered with 1 mL S.O.C. media for 1 h in 37°C shaker. Upon recovery, the electroporation mixture was plated on LB agar supplemented with kanamycin, chloramphenicol and half concentration of apramycin. The transformations were incubated at 37°C overnight. The transformants were verified with colony PCR. Two primer pairs (SYH241, SYH242 and SYH 233, SYH234) were used as colony PCR primers to amplify the *ispH-ispG* and *ispD-ispE* junctions in the final MEP SGCs. Each verified transformant was inoculated in 2 mL liquid LB supplemented with apramycin, kanamycin, and chloramphenicol and incubated in 37°C shaker overnight. On the next day, 2 mL of LB supplemented with apramycin, kanamycin, chloramphenicol and $CaCl_2$ were added to the overnight culture (20 mM $CaCl_2$ final concentration) and incubated for four more hours in 37°C shaker.

Preparation and germination of the spore of the wild type *S. albidoflavus* J1074 for conjugation was performed in the following steps. First, the frozen spore stocks of *S. albidoflavus* J1074 were thawed on ice for 10 min. The thawed spore stocks were centrifuged at max speed for 1 min, and the supernatant was removed. Every 50 µL of packed spores was resuspended in 500 µL TSB supplemented with 10% (wt/vol) sucrose and glycine. Germination of spores was initiated in 28°C shaker for at least 2 h.

Once conjugation donors were overgrown for 4 h, the cells were pelleted by centrifugation at 3,000 rpm for 30 s. The supernatant was removed, and the cell pellet was washed gently three times, each time with 1 mL LB at room temperature to remove antibiotics. The final cell pellet was resuspended in 250 µL of germinated spore solution. The *E. coli* – *Streptomyces* conjugation mixtures were centrifuged 3,000 rpm for 1 min and the excess supernatant was removed so the final resuspension volume was 50 µL. Each conjugation mixture was plated on IWL-4 solid media supplemented with 20 mM

$CaCl_2$ (final concentration) and incubated overnight at 28ºC. After 24 h, the entire conjugation mixture was scrapped and re-plated on IWL-4 solid media supplemented with nalidixic acid and half concentration of apramycin to select for successful *Streptomyces* exconjugants. The final *Streptomyces* exconjugants were transferred to an IWL-4 agar plate supplemented with full concentration of apramycin for maintenance and downstream experiments.

3.4.7 Fermentation of *S. albidoflavus* J1074

The fermentation protocol of *S. albidoflavus* J1074 was modified from Deutz and colleagues [127]. The *Streptomyces* fermentation was divided in three stages and they were (i) fresh sporulation of strains on solid media, (ii) germination and synchronization of cultures and (iii) inoculation of the fermentation media. All 50 mL and 500 mL shake flask cultivations were done in 250 mL and 2 L standard baffled Erlenmeyer flasks, respectively. High-throughput 2 mL and 0.5 mL microcultivation were done in low-evaporation, 24- or 96-square-deep-well microtiter plates (MTPs), respectively (Enzyscreen CR1424a and CR1496a). At the sporulation stage, the frozen spore stocks of recombinant *S. albidoflavus* J1074 strains were streaked on IWL-4 solid media supplemented with apramycin and incubated in 28°C for 2 days to obtain fresh spores. For the spore germination stage, fresh spores from each recombinant strain were grown in 2 mL R2YE supplemented with apramycin in 15 mL culture tube. The R2YE liquid cultures were incubated in 28°C shaker with 250 rpm for 48 h. Each R2YE preculture was used to inoculate (1/100 (v/v) inoculum) 2 mL ISM3 supplemented with apramycin in 15 mL culture tube to synchronize growth of various strains. The synchronization culture for 500 mL fermentation was scaled up to 50 mL ISM3 supplemented with apramycin and was also inoculated with 1/100 (v/v) R2YE preculture. The ISM3 cultures were incubated in 28°C shaker with 250 rpm for 48 h. Each synchronization culture was standardized so that the cell pellet inoculated across samples was the same. Briefly, 1 mL of the synchronization culture was gently pelleted in sterile 1.7 mL Eppendorf microcentrifuge tubes with 3,000 rpm for 30 s. Excess supernatant was removed so the volume of cell pellet and the supernatant were 1 to 1 ratio. For larger synchronization cultures, the standardization was

done in 50 mL sterile falcon tubes. The fermentation media were inoculated with 1% of the standardized synchronization cultures in 500 mL, 50 mL and 2 mL fermentations. Each microfermentation (0.5 mL) was inoculated with 10% of standardized synchronization cultures. The fermentation cultures were grown in 28°C shaker with 250 rpm for 4-7 days.

3.4.8 Extraction of Metabolites from *Streptomyces* Fermentation

The 50 mL fermentations were transferred to 50 mL conical tubes and centrifuged at 4,500 rpm for 15 min. The supernatant was discarded, and the cell-resin pellets were washed with de-ionized water (ddH_2O) and centrifuged for 30 min at 4600 rpm for three times. The samples were frozen at -80°C for 24 h. The frozen samples were then freeze-dried for 24 h or until the moisture was completely removed (0.01 mbar, -84°C; Labconco FreeZone Plus 2.5 Liter Cascade Benchtop Freeze Dry System, catalog number 7670020). Each dried sample was extracted in methanol (3 x 5 mL).

The 500 mL fermentations were centrifuged at 4600 rpm for 30 min. The supernatant was discarded, and the cell-resin pellets were washed with ddH_2O and centrifuged for 30 min at 4600 rpm for three times. The pellets collected were put together and the all was extracted with methanol (250 mL and stirring for 15 min) 6 times. The solutions collected were put together, evaporated under vacuum pressure and the residue was extracted with ethyl acetate (3 x 50 mL). The organic solution was then washed with brine (2 x 50 mL), dried over anhydrous magnesium sulfate, filtrated, and evaporated under vacuum pressure. The powder obtained was washed with *n*-hexane and was collected over a frit to afford **2** as white powder (356.1 mg, 24%). For ^{1}H and ^{13}C NMR data, see **Supplementary Figure A1.5-A1.8**. The NMR attributions are in agreement with the ones previously reported in literature [114]. HRESIMS: *m/z* calculated for $C_{20}H_{29}O_2^-$ $[M\text{-}H^+]^-$ 301.2173, found 301.2173.

The 96-well MTP containing 0.5 mL fermentations was frozen at -80°C for 24 h. The frozen samples were then freeze-dried for 24 h or until the moisture was completely removed. Each sample were extracted with 1.25 mL methanol.

3.4.9 Liquid Chromatography–mass Spectrometry (LC-MS) Analysis

LC-MS analyses were performed on Thermo Scientific MSQ plus mass detector equipped with Ultimate 3000 Rapid Separation (RS) system. The extracts were diluted in 100% MeOH, and 10 µL of each extract was loaded onto a Thermo Scientific analytical BetaSil C18 column (2.1 x 150 mm, 3 µm). Chromatography was performed at flow rate of 500 µL/min using 10 mM ammonium formate buffer (solvent A, pH 8.3) and 100% acetonitrile (solvent B) with the following gradient: 0 – 0.5 min, 95 % solvent A; 0.5 – 14.5 min, 95% solvent B; 14.5 – 16 min, 95% solvent B; 16 – 18 min, 95% solvent A; 18 – 23 min, 95% solvent A. The initial parameters of MSQ plus mass detector for analyzing **2** and shunt metabolites were described below: ionization mode, electrospray; cone ionization, 75 V; capillary probe temperature, 350°C. Full negative MS spectra were acquired for the mass range m/z 150 to 500 for all samples. The optimized parameters for the detection of **2** were described below: ionization mode, electrospray; cone ionization, 110 V; capillary probe temperature, 450°C. Full negative MS spectra were acquired for the mass range m/z 150 to 500 for all samples.

3.4.10 Phylogenetic Analysis of Biosynthetic Genes for Production of 2

Amino acid sequence of *ent*-atiserene synthase PtmT1 (accession ACO31274.1) was used to perform BlastP search of publicly available genome sequences. A total 100 amino acid sequences were reported in the initial result. Redundant or partial sequences were removed from the list. Total of 92 amino acid sequences were used to construct phylogenetic tree of PtmT1 (**Supplementary Data File A1, Worksheet 5**). The sequence alignment was performed by ClustalW multiple alignment application (1000 bootstraps) to calculate protein distance matrix. The phylogenetic analysis was performed in Fitch-Margoliash method in PHYLIP, and the phylogenetic tree was plotted in FigTree.

Phylogenetic tree of *ent*-copalyl diphosphate synthase PtmT2 (accession ACO31274.1) was constructed likewise with 69 amino acid sequences (**Supplementary Data File A1, Worksheet 6**).

3.4.11 General Materials and Methods for Chemical Synthesis

All chemicals and reagents were purchased from commercial sources and used directly without further purification. All non-aqueous reactions were performed under an atmosphere of argon in flame-dried glassware. Reaction progress was monitored by thin-layer chromatography (TLC) using silica gel plates (silica gel 60 F254). Eluted TLC plates were visualized with UV light (254 nm) and dyed using either ethanol acidified with sulfuric acid (5% v/v) or iodine vapors. Compounds were purified either by flash silica gel (230-400 mesh) column chromatography or by Büchi C-700 chromatograph apparatus with 4 g Büchi FlashPure EcoFlex cartridges. Diastereoisomers were purified by HPLC (Ultimate 3000 Rapid Separation (RS) system) using a semi-preparative column Betasil C18, 250 x 10 mm, 5µm particle size (Thermo Scientific) and the products were observed at 210 nm. Chromatography was performed at flow rate of 1.5 mL/min using 10 mM ammonium formate buffer (solvent A, pH 8.3) and 100% acetonitrile (solvent B) with the following gradients: for **20** and **21**: 0 – 1 min, 95 % solvent A; 1 – 3 min, 95% solvent B; 3 – 7 min, 95% solvent B; 7 – 9 min, 95% solvent A; 9 – 18 min, 95% solvent A. For **1** and **24**: 0 – 0.5 min, 95 % solvent A; 0.5 – 1 min, 75% solvent A; 1 – 10 min, 75% solvent A; 10 – 11 min, 50% solvent A; 11 – 18 min, 50% solvent A; 18 – 19 min, 95% solvent B; 19 – 21 min, 95% solvent B; 21 – 22 min, 95% solvent A; 22– 25 min, 95% solvent A. The fractions collected were first concentrated by removing acetonitrile and second the aqueous fractions were extracted with ethyl acetate. NMR experiments were performed on Bruker Advance III 500 MHz (Broadband Observe SmartProbe) or Bruker Advance 700 MHz (5-mm triple resonance cryoprobe) spectrometers. Chemical shifts were reported as ppm relative to either chloroform-d (7.26 ppm for ^{1}H, 77.16 ppm for ^{13}C) or methanol-d_4 (3.31 ppm for ^{1}H and 49.00 ppm for ^{13}C). ^{1}H constant coupling (*J*) are expressed in hertz (Hz), and multiplicity is described as follows: s = singlet, d = doublet, t = triplet, br = broad, m = multiplet. Compounds were analyzed by ESI-MS (Thermo Scientific MSQ plus mass detector) and HR-ESI-MS (ThermoScientific, Q Exactive, Quadrupole, Orbitrap, Heated-Electrospray Ionization probe source (HESI-II)).

3.4.12 Synthesis of *ent*-16-methoxyatisan-19-oic Acids (14-15)

To a solution of **2** (17.9 mg, 0.06 mmol) in methanol (4.0 mL) cooled down to 0°C was added dropwise thionyl chloride (17.6 mg, 0.15 mmol). Once the addition done, the mixture was heated under reflux for 2 h. After completion, the solution was concentrated under vacuum pressure and the residue was purified by column chromatography with a *n*-hexane to *n*-hexane/ethyl acetate (9:1 v/v) gradient to afford the two diastereoisomers of **14** (6.4 mg, 32%) and **15** (6.6 mg, 34%) as white solids. For ^{1}H and ^{13}C NMR data, see **Supplementary Figure A1.21-A1.24** (**14**) and **Supplementary Figure A1.25-A1.27** (**15**). HRESIMS, **14**: *m/z* calculated for $C_{21}H_{33}O_3^-$ $[M\text{-}H^+]^-$ 333.2435, found 333.2435, **15**: *m/z* calculated for $C_{21}H_{33}O_3^-$ $[M\text{-}H]^-$ 333.2435, found 333.2435.

3.4.13 Synthesis of methyl *ent*-atis-16-en-19-oate (16)

To a solution of **2** (10.7 mg, 0.04 mmol) in anhydrous methanol (500 µL) under argon atmosphere was added diisopropylethylamine (22.9 mg, 0.18 mmol). The resulting mixture was vigorously stirred and after 5 min methyl iodide was added (25.1 mg, 0.18 mmol). After reaction completion (96 h), the mixture was acidified with acetic acid and diluted in water (5 mL) and extracted with ethyl acetate (2 x 5 mL). The combined organic extracts were washed with brine (5 mL), dried over anhydrous magnesium sulfate, filtered and the solvent was evaporated with a rotary evaporator. The residue was purified by column chromatography to afford **16** as a white solid (10.2 mg, 91 %). For ^{1}H and ^{13}C NMR data, see **Supplementary Figure A1.29-A1.32** The NMR attributions are in adequation with the ones previously reported in literature [100], [114]. Single-quad ESIMS: *m/z* calculated for $C_{21}H_{33}O_2$ $[M+H]^+$ 317.2, found 317.8.

3.4.14 Synthesis of *ent*-15-oxoatis-16-en-19-oic Acid (17)

Under argon atmosphere, a solution of **2** (56.5 mg, 0.19 mmol) and benzeneseleninic anhydride (70.6 mg, 0.196 mmol) in benzene (36 mL) was refluxed for 4 h. The solution was then concentrated under vacuum pressure and the residue was purified by silica gel column chromatography using a *n*-hexane to *n*-hexane / ethyl acetate (4:1 v/v) gradient to afford **17** (44.6 mg, 75%) as a white solid. For ^{1}H and ^{13}C NMR data,

see **Supplementary Figure A1.34**-**A1.36**. The NMR attributions are in adequation with the ones previously reported in literature.[100] HRESIMS: *m/z* calculated for $C_{20}H_{27}O_3^-$ [M-H]$^-$ 315.1966, found 315.1966.

3.4.15 Synthesis of *ent*-17-methylsulfenyl-15-oxoatisan-19-oic Acids (18-19)

To a solution of **17** (167.3mg, 0.53 mmol) in THF (30 mL) was added every 15 min an aqueous solution of sodium thiomethoxide (15% m/v, 512 µL, 1.06 mmol). After the fourth addition the medium was stirred for 15 min and brine (50 mL) was added to the solution. The mixture was extracted with ethyl acetate (2 x 50 mL) and the organic phase was washed with hydrochloric acid 1 M (2 x 20 mL). The organic solvent was then dried over anhydrous magnesium sulfate, evaporated under vacuum pressure. The solid collected was dried thoroughly under vacuum pressure overnight to afford a mixture of inseparable diastereoisomers of **18** and **19** (186.1 mg, 97%) as a white solid. HRESIMS: **18:** *m/z* calculated for $C_{21}H_{27}O_3S^-$ [M-H]$^-$ 363.1999 (100%), [M+2-H]$^-$ 365.1968 (4%), found 363.2013 (100%), 365.1966 (4%), **19:** *m/z* calculated for $C_{21}H_{27}O_3S^-$ [M-H]$^-$ 363.1999 (100%), [M+2-H]$^-$ 365.1968 (4%), found 363.2013 (100%), 365.1967 (4%).

3.4.16 Synthesis of *ent*-15-hydroxy-17-methylsulfenylatisan-19-oic Acids (20-23)

To a solution of sodium borohydride (185.2 mg, 4.90 mmol) in ethanol (200 proof, 15 mL) cooled down to -5° - 0° C was added dropwise a mixture of **18** and **19** (178.5 mg, 0.49 mmol) in ethanol (200 proof, 5.0 mL). After 25 min of stirring at -5° - 0° C hydrochloric acid 1M was added to mixture until the bubbling stopped. The solution was then evaporated, and the residue was dissolved in ethyl acetate (20 mL), washed with brine (10 mL), dried over anhydrous magnesium sulfate and concentrated. Column chromatography using an *n*-hexane to *n*-hexane/ethyl acetate (7:3 v/v) gradient allowed the isolation of a mixture composed of **20** and **21**, **22** (69.8 mg, white powder, 39%) and **23** (19.5 mg, white powder, 11%). Semi preparative HPLC purification on the latter mixture allowed the separation of **20** (17.8 mg, white powder, 10%) from **21** (53.0 mg, white powder, 30%). For ^{1}H and ^{13}C NMR data, see **Supplementary Figure A1.40-A1.42** (**20**), **Supplementary Figure A1.43-A1.45** (**21**), **Supplementary Figure A1.46-A1.48** (**22**) and

Supplementary Figure A1.49-A1.51(**23**). The NMR attributions for **20** and **21** are in adequation with the ones previously reported in literature [100]. HRESIMS: **20**: *m/z* calculated for $C_{21}H_{29}O_3S^-$ $[M-H]^-$ 365.2156 (100%), $[M+2-H]^-$ 367.2114 (4%) found 365.2155 (100%), 367.2113 (4%), **21:** *m/z* calculated for $C_{21}H_{29}O_3S^-$ $[M-H]^-$365.2156 (100%), $[M+2-H]^-$ 367.2114 (4%) found 365.2155 (100%), 367.2113 (4%), **22**: *m/z* calculated for $C_{21}H_{29}O_3S^-$ $[M-H]^-$ 365.2156 (100%), $[M+2-H]^-$ 367.2114 (4%) found 365.2155 (100%), 367.2113 (4%), **23:** *m/z* calculated for $C_{21}H_{29}O_3S^-$ $[M-H]^-$365.2156 (100%), $[M+2-H]^-$ 367.2114 (4%) found 365.2155 (100%), 367.2113 (4%).

3.4.17 Synthesis of Serofendic Acids and Diastereoisomers (1, 24)

To a solution of sulfides **20** and **21** (9.7 mg, 26.5 µmol) in chloroform (500 µL) was added Davis' oxaziridine (2-benzenesulfonyl-3-phenyloxaziridine) (7.3 mg, 27.8 µmol). After completion (1 h) the mixture was purified on silica gel column chromatography using a dichloromethane to dichloromethane/methanol (10:1 v/v) gradient to yield **24A** (3.2 mg, 33%) and a mixture composed of **1A**, **1B** and **24B** (6.1 mg). Semi preparative HPLC purification on the latter mixture allowed the separation of **24B** (4.2 mg, 43%) from **1A** (0.43 mg, in mixture with **1B** (0.91:0.09, **1A**:**1B**), 4%) and **1B** (0.85 mg, in mixture with **1A** (0.81:0.19, **1B**:**1A**), 9%). For ^1H and ^{13}C NMR data, see **Supplementary Figure A1.52**-**A1.56** (**1A**), **Supplementary Figure A1.55**-**A1.58** (**1B), Supplementary Figure A1.59-A1.61** (**24A**) and **Supplementary Figure A1.62**-**A1.64** (**24B**). The NMR attributions for **1A** and **1B** are in agreement with the ones previously reported in literature [97]. HRESIMS: *m/z* calculated for $C_{21}H_{29}O_4S^-$ $[M-H]^-$ 381.2105 (100%), $[M+2-H]^-$ 383.2063 (4%), found **1A**: 381.2109 (100%), 383.2068 (4%), **1B**: 381.2110 (100%), 383.2068 (4%), **24A**: 381.2109 (100%), 383.2068 (4%), **24B**: 381.2110 (100%), 383.2068 (4%).

Additional methods and supplementary figures can be found in the **Appendix 1** and **Supplementary Data File A1** (separate excel file).

Chapter 4 Rational Search of Genetic Design Space to Improve Terpene Production in *Streptomyces*

The contents of this chapter are currently in preparation for publication. MJS designed the 5-level Plackett-Burman matrix. SYH planned and performed all the experiments. Colony morphology photography was assisted by Jihaeng Lee. SYH wrote the original draft. SYH and MJS reviewed and edited the manuscript.

Summary

Modern tools in DNA synthesis and assembly give genetic engineers complete control over the nucleotide-level design of complex, multi-gene systems. Systematic approaches to explore genetic design space and optimize the performance of genetic constructs are lacking. Here we demonstrate the application of a five-level Plackett-Burman fractional factorial design to improve the titer of a heterologous terpene biosynthetic pathway. Combinatorial perturbation of a refactored methylerythritol phosphate (MEP) pathway expression affects the titer of the target diterpenoid, *ent*-atiserenoic acid (eAA) over two orders of magnitude. A library of 125 prototypes of synthetic gene clusters encoding a natural MEP pathway was constructed and introduced into *Streptomyces albidoflavus* J1047 for heterologous expression. The eAA production titer of all strains was measured, which showed surprisingly robust isoprenoid metabolism; the changes in average eAA titer did not change significantly when the expression of seven of eight genes was changed. Analysis of Plackett- Burman design identifies that the expression of *dxs*, the first and the flux-controlling enzyme, is the most dominant gene on eAA titer with negative correlation. This study showed Plackett-Burman design is as an effective screening approach as the analysis of the experiments replicate the same finding, which is that DXS is the bottleneck of MEP pathway in the literature.

4.1 Introduction

The recent advances in DNA synthesis and assembly allow scientists to engineer more complex genetic systems at various scales. For example, biosynthesis of pharmaceuticals [54], [77], [128] and functionalized materials [129]–[131], synthetic

genetic circuits [132], [133] and organisms with synthetic chromosomes [88], [134]–[136] have all been built from scratch. A common feature of these systems is the requirement to coordinate gene expression of 10s to 1000s of genes for the system to work.

Optimizing gene expression to improve performance in such systems becomes a combinatorial problem. Common design variables that impact relative gene expression include promoters [137], ribosomal binding sites [26], intergenic region sequences [138], [139], codon usage[140], [141], gene order [65], [66], [142], and gene orientation [143]–[145]. The exponential scaling of the multi-dimensional design space makes an exhaustive search impossible or impractical given limits on time and labor resources. For instance, production of hydrocodone required overexpression of 26 genes [146]. Testing five design variants of each expression cassette in a full-factorial design would require 5^{26} (~1.49 x 10^{18}) unique synthetic constructs to be analyzed. A new challenge that emerges in designing multi-gene systems is that of accelerating genetic optimization and reducing cost by maximizing the information content of each iterative design cycle [147].

Statistical design of experiments (DoE) is a rational approach that can guide the design of synthetic genetic constructs. It has been used effectively in optimizing media compositions [148], [149], recombinant protein production [150] and industrial processes [151]. DoE employs statistical methodologies to systematically design and analyze experiments. It can be used to specify the minimal number of experiments needed to rigorously identify the relative impacts of each variable on system performance. Various DoE approaches have been used to guide library design of multigene systems and have been applied to improve nitrogen fixation rates [64] and production titers of 6-aminocaproic acid [152], itaconic acid [153] and violacein [154], [155]. All of these studies aimed to measure the effects of the input (gene expression strength or media composition) on system performance and explore the genotype-phenotype landscape of the multigene systems.

Plackett-Burman design is a type of fractional factorial experimental design that allows the variables in a system to be prioritized based on their overall importance (main effect) to system performance [156]. Plackett-Burman design is especially advantageous when the experimenters need to study large number of variables of a system which they do not have complete *a priori* knowledge. In the literature, a two-level Plackett-Burman

design is commonly used in which each variable is set at a high level and a low level to screen the variable's effect on the output. A two-level Plackett-Burman only requires a small set of experiments to identify the most important variables. For instance, an 8-experiment design can screen up to 7 variables of which each variable is tested at low and high level. In practice, experimenters design Plackett-Burman matrices that can accommodate more variables than they will interrogate in their system [148], [157], [158]. The extra 'dummy variables' function as negative controls during data analysis and provide the required context to calculate experimental error and statistical significance [159]–[161]. Another method to increase the accuracy in measuring main effects is to use a Plackett-Burman design with more than two levels per variable. This is particularly true for genetic systems, where setting the gene expression variables to target levels is imprecise. Including more levels for each variable decreases the likelihood that a primary effect will be erroneous due to noise in setting the level.

In this chapter, I present the investigation of the effectiveness of a five-level Plackett-Burman design to screen important genes encoding the methyl-erythritol phosphate (MEP) pathway for the production of a diterpene acid, *ent*-atiserenoic acid (eAA) [74], in *Streptomyces albidoflavus J1074*. eAA is a labdane diterpene whose carbon scaffold is present in a number of bioactive natural products [74], [124], [162] We quantify the effects of the expression level of seven genes of the MEP pathway (*dxs*, *dxr*, *ispD*, *ispE*, *ispF*, *ispG*, and *ispH*) plus the gene encoding isopentenyl-diphosphate delta-isomerase (*idi*). We chose the MEP pathway as a model multigene system for several reasons. First, the MEP pathway produces C_5 building blocks to terpenoids, a highly industrially and medicinally relevant group of molecules in the biomanufacturing field [163]–[165]. Second, it is an excellent testbed for new method development, as there is a substantial body of literature for MEP pathway engineering to which experimental results can be compared. Lastly, there are seemingly contradictory reports on MEP pathway optimization, with many different enzymes in the pathway identified as the rate limiting step in different systems [166], [167].

4.2 Results

4.2.1 Design of a Five-level Plackett-Burman Experiment via Cyclic Permutation

A five-level Plackett-Burman design was used to determine which gene's expression has the largest effect on eAA titer. Specifically, the independent variables being screened are the expressions of seven genes of the MEP pathway plus *idi* and their effect on the response variable, the eAA titer. A five-level Plackett-Burman design comprising a 125 × 31 matrix was designed as described previously [156]. The matrix is balanced in that each column (i.e. variable) contains an equal number of rows in each of the five levels (i.e. 1-5). When rows are grouped into five sets of 25 based on the values in a single column, each of the other columns will contain a balanced number of variable states 1-5 in each set. In other words, assuming that the variables are completely independent, the effects of all other system variables cancel out when grouping the experiments in a way to examine the effect of a variable of interest. (**Supplementary Data File A1**). In this Plackett-Burman matrix, each column represents an independent variable, and each row represents the overall genetic design of a full cluster which is constructed and tested. Thus, the 125 x 31 matrix allowed for screening of up to 31 independent variables. When less than 31 independent variables are tested, the remaining columns represent 'dummy variables' that do not correspond to any physical changes in the genetic system and can serve as negative controls during data analysis to determine experimental error.

The seven genes of the MEP pathway and *idi* were selected as the independent variables, and they are *ispH* (X_1), *ispG* (X_2), *dxs* (X_3), *dxr* (X_4), *ispD* (X_5), *ispE* (X_6), *ispF* (X_7), and *idi* (X_8). Each gene was represented at five gene expression levels: 1 for low level, 2 for medium-low level, 3 for medium level, 4 for medium-high level, 5 for high level (**Figure 4.1b**). The production titer of *ent*-atiserenoic acid (*e*AA) is measured as the dependent variable (Y).

Two of the most important properties of a Plackett-Burman design is that it is balanced and orthogonal. Each variable is being tested the same number of times for all levels in a balanced experimental design. In the five-level Plackett-Burman genetic design, each gene is expressed at each of the 5 expression levels in exactly 25 Synthetic Gene Clusters (SGCs). The orthogonality of Plackett-Burman design means that none of the

independent variables co-vary across the library as a whole. In other words, for the 25 genetic designs in which *dxr* is expressed at a low level (*i.e.* Level 1), each of the other genes are present in all 5 levels an equal number of times. This is important, as it allows the experimenter to measure the main effect of one variable by plotting the average performance of sub-groups of gene clusters wherein a gene is expressed at levels 1-5. The impact of other genes on those sub-group averages cancels out due to the orthogonal design. There is an underlying assumption in Plackett-Burman designs that the variables are completely independent, or uncorrelated. While this is a prior not expected to be the case for genetic systems, PB designs are still effective so long as the correlation factor is not high. This is evident from the successful application of PB designs to optimize complex systems such as fermentation media [168].

Eight columns were chosen to minimize the number of intermediary DNA assemblies and maximize re-use of DNA parts for the construction of the Plackett-Burman Full Cluster (PBFC) library. The resulting 125 × 8 matrix was used as the blueprint of the genetic design of the synthetic gene cluster library encoding natural MEP pathway (**Supplementary Figure A2.1**). The rest of columns are treated as dummy variables which do not correspond to any variables in our physical system.

4.2.2 Hierarchical Assembly of 125 Synthetic Gene Clusters Encoding eAA Biosynthesis with Variant Expression of MEP Pathway Genes

To control proper gene expression level in *Streptomyces*, we used a previously characterized *Streptomyces* promoter-RBS library [74]. The promoter-RBS parts were ranked by the $\log_{10}$-transformed expression strength. Then, a set of five unique promoter-RBS parts were selected for each gene to tune the expression of the five levels represented in the Plackett-Burman design.

Total of 125 unique PBFC designs were generated based on the selected eight columns from the Plackett-Burman matrix. We used an iterative, hierarchical assembly strategy to construct PBFC designs in parallel from small functional genetic parts and intermediary partial clusters, to obtain full clusters encoding MEP pathway plus *idi* (**Supplementary Figure A2.1**). The assembly started with each gene pieced together with selected promoter-RBS and terminator parts to form monocistronic units (**Figure 4.1b**).

Four monocistronic units were grouped as partial clusters: the upper partial clusters (*ispH*/*ispG*/*dxs*/*dxr*) and lower partial clusters (*ispD*/*ispE*/*ispF*/*idi*). Next, one upper and one lower partial cluster were assembled to form each full cluster. All full clusters contain three invariant modules: (i) genes for terpene cyclization and eAA oxidation, (ii) apramycin resistance gene for selection, and (iii) ΦC integrase gene to integrate full cluster into *S. albidoflavus* J1074 genome [113] for stable expression. Every stage of the assembly was done by plasmid-based Golden Gate assembly.

In total, the library design requires 40 unique monocistronic parts, 25 unique upper and lower partial clusters, and 125 unique full clusters (FCs) (**Figure 4.1c**). All but one FC (#81) were successfully assembled and verified in *E. coli* via colony PCR (**Supplementary Figure A2.2**). FC #81 assembly failed despite multiple efforts, suggesting an unexpected toxic effect from the combination of genetic parts [57]. FCs were conjugated into *S. albidoflavus* J1074 and verified colonies were obtained for all but FC #34, and #108. Again, the failure of these constructs despite multiple attempts suggests they are toxic in the heterologous host. The efficiency of the assembly steps ranges from 92% to 100% [74].

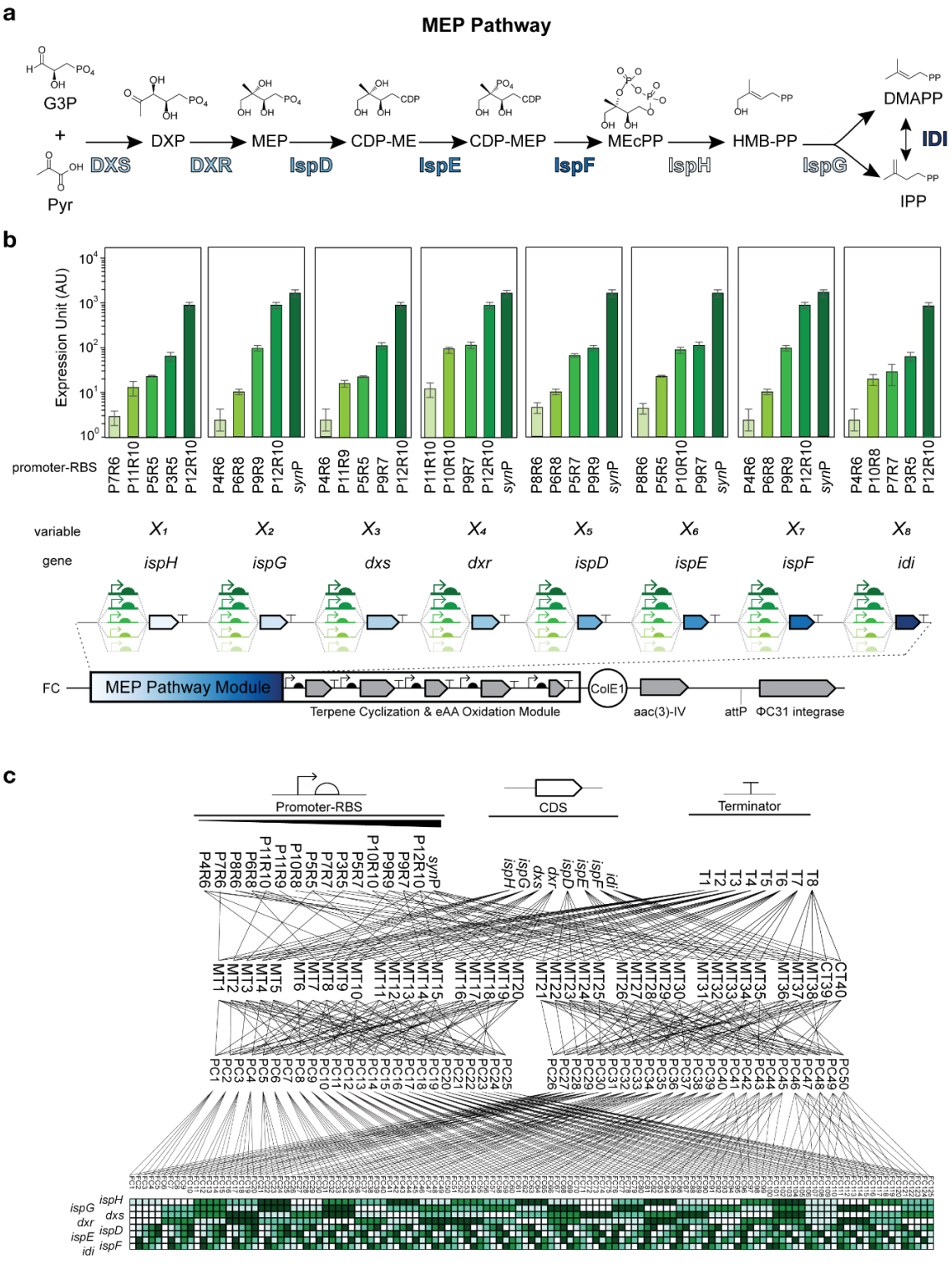


Figure 4.1. The design and assembly of PBFC library.

(a) The MEP pathway and (b) the promoter-RBS combinations used to control the expression of eight genes are shown. Each gene is designed to express at five different levels; level 1, 2, 3, 4, and 5 correspond to low, medium-low, medium, medium-high, and high expression level. The shades of greens indicate the expression strength, with darker meaning stronger. The expression strength was measured previously as ΔABS415nm per minute per gram dried cell weight by beta-glucuronidase reporter assay [74]. The architecture of the full cluster reflects the gene order in the constructs. (c) The bottom-up assembly plan of PBFC library is shown. The top section lists the individual DNA parts. The strength of promoter-RBS parts is ranked from the weakest to the strongest (left to right). The second section lists monocistronic units. The third section lists partial clusters. PC1-25 and PC26-50 are upper and lower partial clusters, respectively. The bottom section lists the full clusters. Abbreviations are: P, promoter; R, RBS; ispH, 4-hydroxy-3-methylbut-2-enyl diphosphate reductase; ispG, 4-hydroxy-3-methylbut-2-en-1-yl diphosphate synthase; dxs, 1-deoxy-D-xylulose-5-phosphate synthase; dxr, 1-deoxy-D-xylulose 5-phosphate reductoisomerase; ispD, 2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase; ispE, 4-diphosphocytidyl-2-C-methyl-D-erythritol kinase; ispF, 2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase; idi, isopentenyl-diphosphate delta-isomerase; MT, monocistronic unit; PC, partial cluster; FC, full cluster.

4.2.3 Performance of PBFC library in *S. albidoflavus* J1074

The PBFC strains were fermented in duplicate in 0.5 mL of ISM3 medium for five days in high-throughput 96 deep-well plates. The cultures were freeze-dried, extracted with methanol, and analyzed by LC-MS to quantify the eAA production titer. PBFC 34, 81 and 108 did not yield any *S. albidoflavus* J1074 exconjugants, so their eAA production titers were considered as 0.0 mg/mL in data analysis. The eAA production titer spans from 0.0 mg/L (non-detectable) to 152.2 ± 4.8 mg/L with FC#44 as the highest producer (**Figure 4.2a**). The distribution of eAA titer of the PBFC library is lognormal, with a median titer of 10.8 mg/L and an interquartile range of 20.7 mg/L.

Interestingly, many PBFC strains displayed different colony morphologies from the *S. albidoflavus J1074*-empty vector control when grown on solid media. We used six basic colony morphology features, shape, margin, color, opacity, elevation, and texture, to categorize the PBFC strains into ten morphology classes (**Figure 4.2b**). Most notably, the PBFC strains in morphology classes 7, 8 and 9 have translucent colonies which partially or completely lost the sporulation phenotype even after extensive incubation at 28°C. Strains in morphology class 2 exhibited smooth concentric ring morphology that was not previously observed in wild-type *S. albidoflavus J1074.* Class 1, 5, 7, and 10 have more elevated and wrinkled surface than wild-type *S. albidoflavus J1074.*

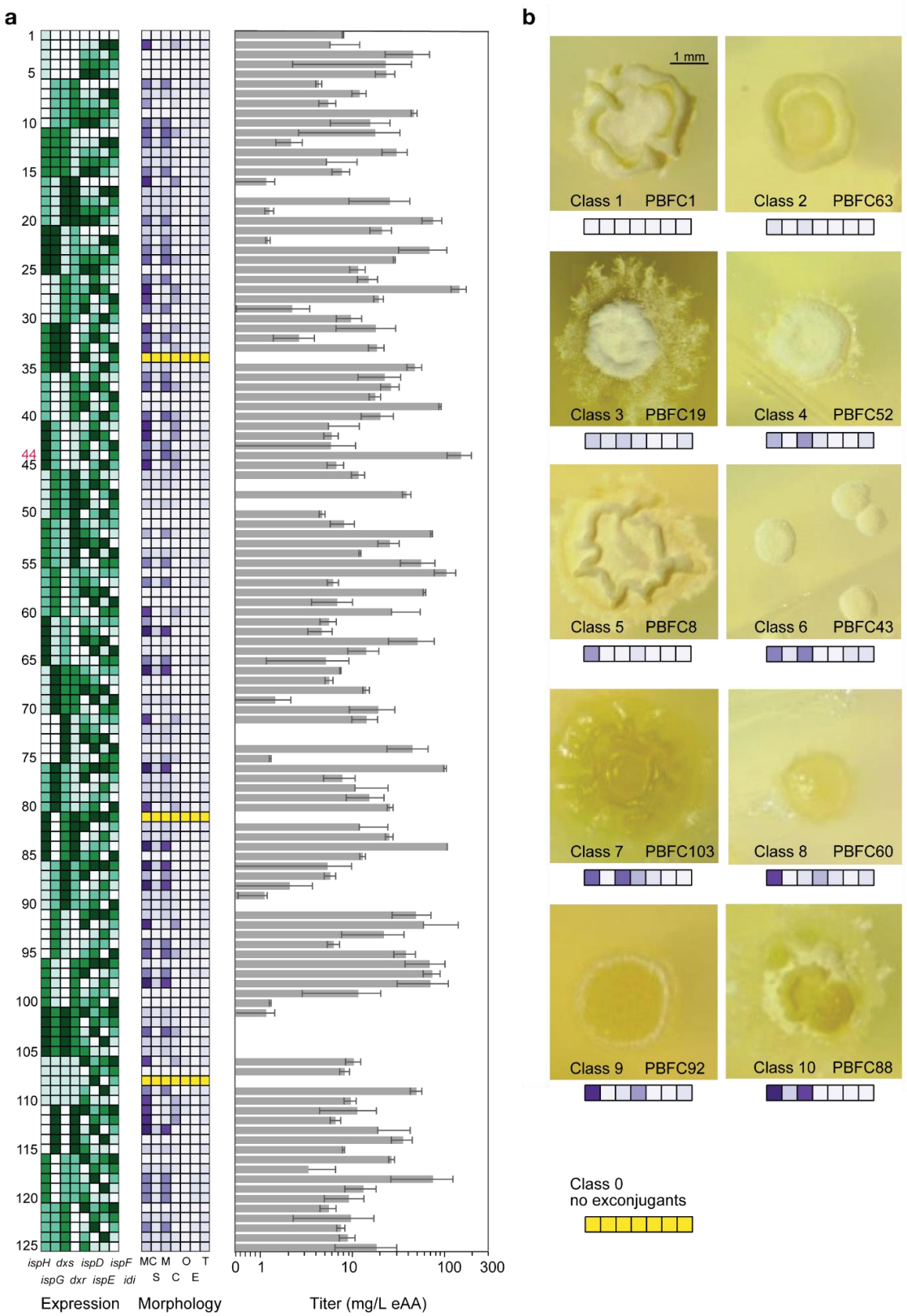

a
b
1 mm
Class 1 PBFC1
Class 2 PBFC63
Class 3 PBFC19
Class 4 PBFC52
Class 5 PBFC8
Class 6 PBFC43
Class 7 PBFC103
Class 8 PBFC60
Class 9 PBFC92
Class 10 PBFC88
Class 0
no exconjugants
ispH dxs ispD ispF
ispG dxr ispE idi
MC M O T
S C E
0 1 10 100 300
Expression
Morphology
Titer (mg/L eAA)

Figure 4.2. Expression level, eAA titer and morphological features of strains of PBFC library.

(a) The shades of green indicate the expression level of individual genes, which is the same as in **Figure 4.1**. The shades of purple indicate different morphology classes and morphological features, with lightest purple representing Class 1, and darkest purple representing Class 10. Yellow means no exconjugants were obtained for the specific genetic designs. The average eAA production titer (grey bar) of each PBFC strain plotted on a semi-log scale (n = 2, standard deviation plotted). PBFC#44 is the highest eAA producer in this library (red text). (b) Morphological feature representation of 10 morphology classes in the PBFC library. MC abbreviates for morphology class; S, shape; M; margin; C, color, O, opacity; E, elevation; T, texture. The PBFC strains were fermented on two different days.

4.2.4 Main effect of overexpression of MEP pathway genes on eAA production titer

We analyzed the output of this pathway, eAA production, by one-way analysis of variance (ANOVA) to screen the genes that have the largest effect on the eAA titer. Since the distribution of PBFC library eAA titer is lognormal, the titer data were log10-transformed prior to any statistical analysis. The 'log eAA titer' from now on represents 'log10-transformed eAA titer' to prevent any confusion. For each gene expression group, the arithmetic mean of log eAA titer was calculated to quantify the effect of each gene on the eAA production. **Figure 4.3a** shows the overall changes in the average log eAA titer caused by changing expression levels. One-way ANOVA showed that only *dxs* expression is statistically significant on the eAA titer.

Next, we analyzed the correlation between a gene's expression level and the average eAA titer. An ordinary least square regression analysis was used to fit a linear model of the effects of each gene on the mean log eAA titer. The 'main effect' can be approximated by the slope of the linear regression. The statistical significance of the slope of each linear regression was evaluated by t-test. The most dominant variable over eAA titer is *dxs*, with lower expression correlating with higher titer (Figure **4.3b** and Figure **4.3c**). The gene expression of *ispF* has weak positive correlation with eAA titer. The gene expressions of *ispH*, *ispG, dxr*, and *idi* have weak negative correlations with the eAA titer. Gene expression of *ispE* has little or no association with titer. Although the linear fit of the genes other than *dxs* is not statistically significant to the eAA titer, the analysis reveals a few insights about an individual gene's optimal expression level in our system.

Interestingly, the optimal expression level of *ispG*, *dxr*, *ispE*, is at the lowest (level 1); *dxs* and *idi*, at medium low (level 2); *ispH*, medium (level 3); *ispD*, highest level (level 5).

The main effects of 21 dummy variables (unassigned columns) were also calculated to estimate experimental error used in statistical interpretation. If the components of a system are independent from each other and their two- or higher-order interactions are negligible, the level changes of dummy variables should not influence the system performance. In a perfectly independent system, changes in the levels of dummy variables should not correlate with eAA titer at all, and the main effect/slope of the linear regression line would be zero. The fact that a handful of the dummy variables have effect (non-zero) on eAA titers (**Figure 4.3c**), indicates either the presence of interaction between the genes or a non-trivial amount of stochastic experimental noise/variance.

The dummy variables serve as negative controls during data analysis. In **Figure 4. 3c** we plotted the distribution of eAA titer variance (between the five sets of 25 grouped rows) and calculated main effect for each of the 23 dummy variables. The distribution of the main effects of the dummy variables tightly clustered around zero, indicating the overall statistical variance caused by experimental error or higher-order interactions are quite small; the distribution of the variation of average titer clustered around 0.1. The experimental error was not to a statistically significant as the indicated by T-test.

Importantly, plotting the experimental variables on the same axes as the dummy variables in **Figure 4.3c** shows that five of the eight variables fall outside of 2 standard deviations of the negative controls. This is not expected to result from random chance, and supports that conclusion that our PB library design returned actionable information about the system that stands out from the stochastic noise that is inherent in *Streptomyces* fermentations.

Across the range of expression values that we tested in this experiment, *ispD* is the only gene with a meaningful positive main effect (i.e. increase in titer corresponding to increase in *ispD* expression). Three genes, *dxs*, *dxr*, *idi*, and *ispH*, have a meaningful negative main effect, with lower expression levels preferred (**Figure 4.3c**), however it is noteworthy that this negative main effect is in some cases driven primarily by low titers only at the highest expression level (e.g. *ispH* and *idi* in **Figure 4.3b**). For these genes, medium-level expression of the genes is optimal.

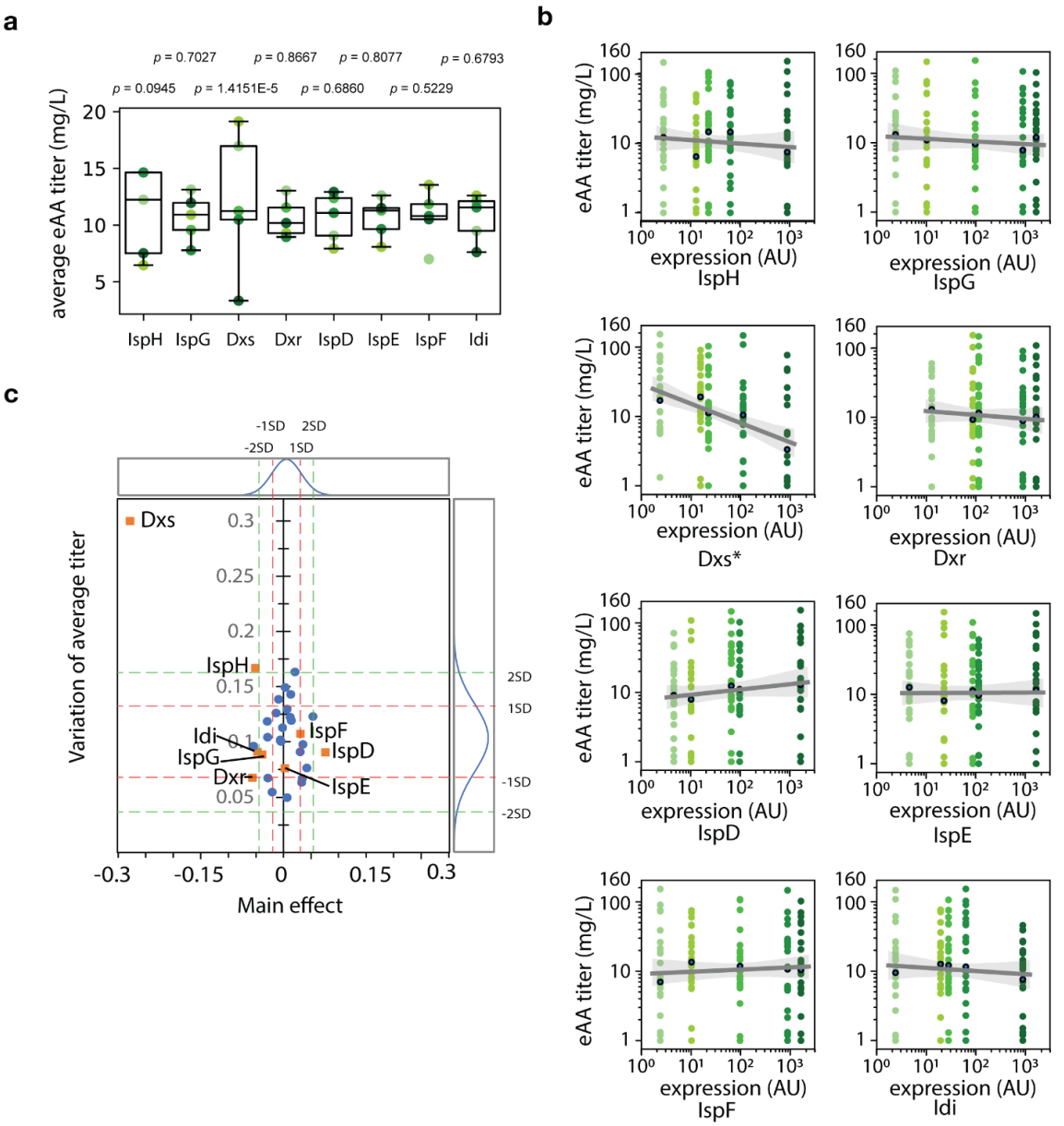


Figure 4.3. Main effects of MEP pathway genes plus idi on eAA titer.

(a) overall magnitude of average eAA titer changes caused by gene expression. (b) Expression level of each genes (in log10 scale) is plotted against eAA titer of strains containing that specific expression level. The geometric mean of the eAA titer of each gene expression group is plotted as a black dot. A linear regression is represented as the grey line. Grey area denotes 95% confidence interval. α= 0.05. (c) Main effects and variation of average log eAA titer changes caused by genes (orange squares) or dummy variables (blue circles). Main effect is the same as the

slope of the linear regression fit. Variation of average log eAA titer is calculated as the standard deviation of the geometric average of eAA titer of five expression levels. The five shades of green correspond to the expression level in Figure 4.1. * represents the *p*-value <0.05.

4.2.5 Contribution of discrete genetic motifs to overall eAA titer

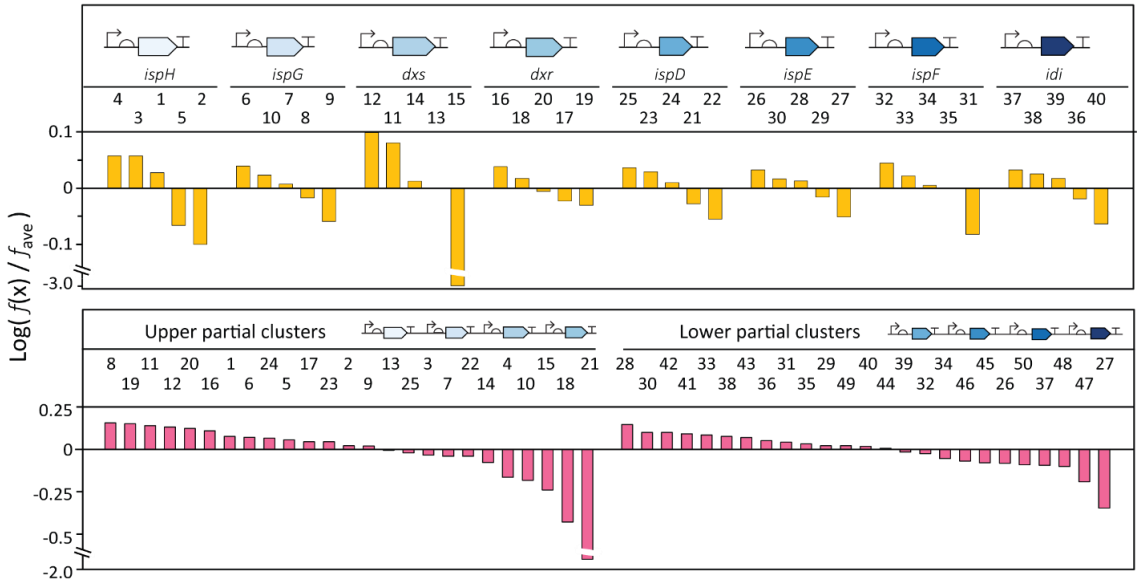
We decomposed the activity of the final eAA gene cluster into the activities of its constituent parts, namely the monocistronic unit and partial clusters. Here we attempt to analyze the extent of how the system behavior can be predicted by the combination of its parts using additive, multiplicative and dominant negative effect models which are used routinely in genetic interaction studies [169], [170].

First, the contribution of each intermediate genetic motif to full cluster performance was evaluated by calculating the arithmetic mean of log eAA titer of the strains containing the specific part. We define this value as a 'part fitness.' **Figure 4.4a** showed the $\log_{10}$ ratio of each part fitness value is to the average full cluster fitness, f_{ave} , which is defined as the arithmetic mean of the entire log eAA titer data (10.5 mg/L eAA). If a part has higher fitness than the average full cluster fitness, then the bar is above zero; worse, than below zero.

Based on the part fitness values, a retroactive prediction of full cluster performance is calculated using three models. First, the additive model predicts the full cluster fitness to be the average of the corresponding single part fitness values. Second, the multiplicative mode predicts the full cluster fitness to be the product of the corresponding single part fitness values. Third, the dominant negative effect model assumes that a full cluster is only as good as its worse part and would be true if genetic motifs with low part fitness values have a dominant effect over motifs with high part fitness values. (**Fig 4.4b**). The analysis showed that the multiplicative model has the highest R^2 value amongst the three models, indicating better predictability of the full cluster fitness. Overall, multiplicative model using partial cluster fitness is a better predictor of full cluster fitness than MT fitness in all models, as 39.23% of all the variance in the fitness can be accounted for by the partial cluster multiplicative model. This result is likely due to smaller genetic context changes in partial cluster assembly. The multiplicative model suggests that the MEP

pathway components interacts on a non-linear scale rather than the linearly, which is also observed in anthocyanin biosynthesis [171], [172].

a



b

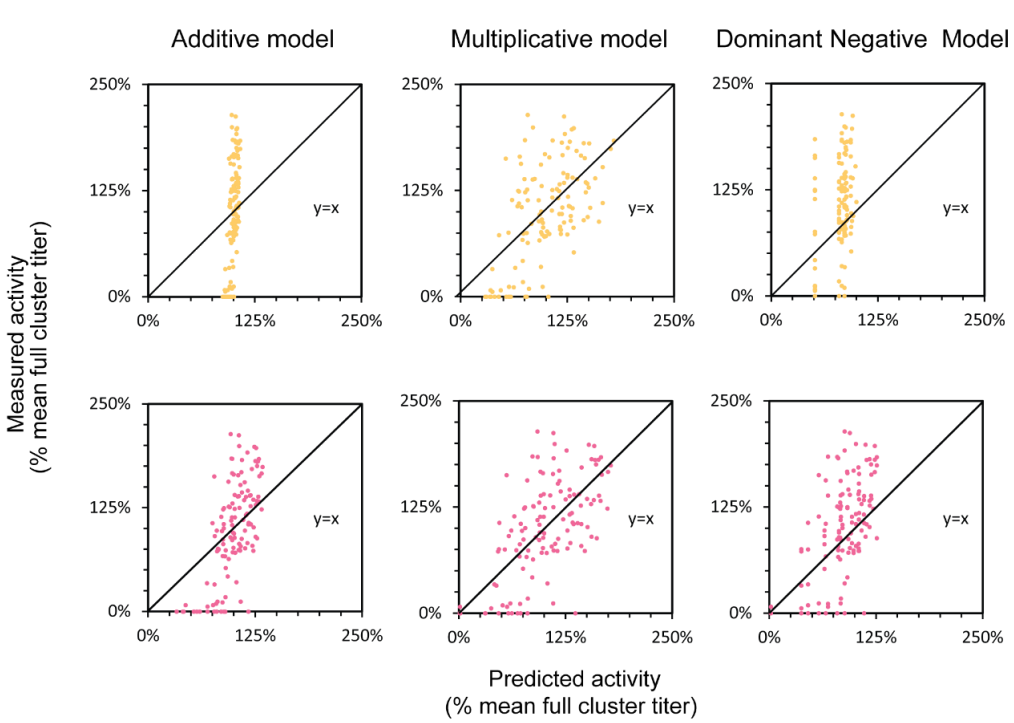


Figure 4.4. Behavior of parts in full clusters.

(a) fitness values of monocistronic unit (yellow bars) and partial clusters (magenta bars), presented as a log ratio normalized to mean fitness value of PBFCs, are shown. Part names corresponding to the assembly graph in Figure 1c are labeled above bars. f_{ave} is the geometric mean eAA titer of the PBFC library. (b) Predicted full cluster activity is plotted against measured activity. Colored points represent the 125 full cluster constructs. Level of correlation to the y = x line improves at more advanced levels of assembly from R^2 = 0.0875, 0.3313, and 0.0361 for monocistronic units (yellow) and R^2 = 0.3355, 0.3923, and 0.2740 for partial clusters (magenta) in additive, multiplicative, and dominant negative models, respectively.

4.3 Discussion

The expression level of two genes, *dxs* and *ispD*, were shown to have the most noticeable effect on eAA production. Previous research has shown that DXS is the major flux-controlling enzyme [173], and its overexpression leads to increase isoprenoid production titer [26], [167], [173], [174]. In our system, overexpression of *dxs* leads to lower eAA titers, and the best expression level for *dxs* is medium-low (level 2). Others have reported a negative effect for *dxs* overexpression on cell growth and productivity [175]–[177]. One plausible explanation for the negative correlation between *dxs* expression and eAA production is that overexpression of *dxs* which causes carbon and energy burden in the host cell. Dxs catalyzes the condensation of pyruvate and glyceral-aldephyde-3-phsophate, two key glycolytic intermediates for generating amino acid building block, reducing equivalents, and ATP. While glycolysis displays a certain tolerance to flux perturbations [178], overexpression of *dxs* diverts too much carbon flux towards MEP pathway and robs the cellular resources from central metabolism to function properly, eventually sabotages the eAA production level. Overall, the MEP pathway is quite robust, as the gene expression changes in six out of eight genes do not leads to significant changes in the eAA titer. Our result may explain why there are so many different results as to which enzyme is the rate-limiting step in the MEP pathway literature.

Previously, a few publications highlighted the nonlinearity of the fitness/production landscape of the MEP pathway [102], [173], [179]. Similarly, we observed multiplicative (non-linear) interaction of partial clusters and monocistronic units can better predict the full cluster activity encoding MEP pathway rather than the additivity (linear) interaction.

We have two hypotheses to explain the origin of the nonlinearity of the eAA landscape. The first hypothesis is the observed nonlinearity is an artifact of the gene expression noise [180], rather than the real nonlinearity caused by higher-order interaction of the MEP pathway components. The promoter-RBS parts used in the full clusters were characterized in isolation, and their actual behaviors in a full cluster may vary from expected due to genetic context effect [14], [57], [181]. To mitigate the potential effect of genetic expression noise, we chose the five-level Plackett-Burman design, instead of the usual two-level seen in the literature, to build in redundancy of testing multiple level of gene expression.

Our second hypothesis is that the observed nonlinearity represents the higher-order interactions of MEP pathway components. In this work, the only modification to the *S. albidoflavus* J1074 genome is the integration of the refactored MEP pathway. The endogenous MEP pathway and regulatory processes still operate in the background, which may confound the result of Plackett-Burman library. The flux of the refactored MEP pathway may be regulated at multiple layers just like how endogenous MEP pathway is regulated [182]. One of the regulations of MEP pathway is at the protein level; the activity of DXR and IspD is modulated by phosphorylation of key active site residues. Another regulation mechanism is through metabolite control including feedback inhibition of DXS by IPP/DMAPP [183], and feed-forward activation of *ispF* by MEP [184]. MEP pathway also interacts with other cellular processes. For example, the MEP pathway intermediate, MEcPP, modulates the stress response at the transcription level. DXP, another pathway intermediate, is a precursor to pyridoxal and cofactor biosynthesis. In addition, the first step of MEP pathway draws from two primary metabolites of glycolysis (glyceraldehyde-3-phosphate and pyruvate). The complex interplay of the MEP pathway and other metabolic processes is also reflected in the diverse morphological changes in PBFC strains. However, it is unclear how the MEP pathway modulates morphology development in *Streptomyces*.

A weakness of implementation of Plackett-Burman design is that the analysis assumes the effect of independent variables has a linear relationship of with the overall system performance [185]. The linear assumption of the main effect makes the interpretation of the result intuitive and straightforward, but it may not capture the quadratic

effect of variables. For example, the width of the confidence interval of the linear regression of expression of *ispH*, *ispF*, and *idi* are quite large, indicating a weak correlation with eAA production titer (**Figure 4.3b**). However, the distributions of the grouped average log eAA titer in *ispH, ispF* and *idi* suggests a non-linear correlation with the gene expression, and therefore a linear model with polynomial terms or a quadratic model may be more appropriate to fit the curvature effects. Another weakness of Plackett-Burman design is its aliasing structure, meaning that the main effects can be partially confounded by second- or higher-order interactions. Therefore, Plackett-Burman design is assumed to be effective in estimating true main effects when the effect of higher-order interactions is negligible. When a main effect is large, it is less likely to be masked by higher-order interactions. Plackett-Burman design is not capable of tease apart interaction effect and main effect when the main effect is small, and will require either advance statistical analyses [186], [187] or more experiments to de-confound interaction effects and main effects.

While it has become increasingly easy to explore the genetic design space, we have demonstrated that a subset of carefully chosen designs from a full-factorial design is sufficient to the systematically explore and understand the landscape of isoprenoid biosynthesis in *Streptomyces*. Other may argues to increase coverage of the design space and gain maximal information about the landscape. This has been demonstrated by generation of diverse genetic variants of multi-gene systems with targeted directed evolution [167], [188], *in vivo* mutagenesis[189], and combinatorial assemblies [64], [102]. A limitation of these diversification-driven methods is that they are difficult to apply to the cases which multi-gene systems do not have phenotypes compatible with current high-throughput screening methods. In contrast, fractional factorial designs like Plackett-Burman design is effective to identify the variable that has large main effect on the overall system performance with minimal number of experiments.

DoE can be considered as a complementary framework to systematically reduce library size and design-build-test cycle time. In this work, we were able to identify *dxs* as the important genes on eAA titer with a small library of 125 unique genetic designs, which covers 0.032% of the entire design space (125 designs of all 5^8 possible designs). By applying a Plackett-Burman design to a well-studied system like the MEP pathway allowed

us to compare our DoE-based screening results to the large body of literature and rank order the effects of these eight genes. The result of this study showed that Plackett-Burman design was able to replicate the findings that *dxs* is the rate-limiting step and provided additional insights to MEP pathway with a medium library size. Plackett-Burman design also allows systematic access to unexplored sequence-function space landscape. In sum, Plackett-Burman design is an effective method in screening and prioritizing variables in order to optimize a model multi-gene system, which gives us confidence that Plackett-Burman design can be extended to study less characterized genetic systems.

4.4 Material and Methods

4.4.1 Strains and Chemicals

E. coli DH5α and NEB® Stable were used for routine cloning. *E. coli* ET12567/pUZ8002 was used for intergenic conjugation. *E. coli* NEB ® Stable/pUZ8002 was used as an alternative intergenic conjugation strain for the PBFC21 which was not able to be conjugated by using *E. coli* ET12567/pUZ8002. *S. albidoflavus* J1074 was used as a heterologous expression host for screening PBFCs and FFCs. LB medium (10 g/L tryptone, 5 g/L yeast extract, 10 g/L NaCl; MP Biomedicals # 113002022) supplemented with proper antibiotics was used for strain maintenance and plasmid construction in *E. coli* strains. *Streptomyces* liquid media for maintenance and fermentation include R2YE medium (103 g/L sucrose, 10 g/L glucose, 0.25 g/L K_2SO_4, 5 g/L yeast extract, 0.1 g/L Difco casamino acid, 100 ml TES buffer (5.73%, pH 7.2), 10 ml KH2PO4 (0.5%), 80 ml $CaCl_2.2H_2O$ (3.68%), 15 ml L-proline (20%), 2 ml trace element solution and 5 ml NaOH (1N)), ISM3 medium (15 g/L yeast extract, 10 g/L malt extract, 0.5 g/L $MgSO_4$, 0.3 g/L $FeCl_3 \cdot 6H_2O$, 20 g/L glucose, adjusted to pH 7.0 with NaOH and TSB (30 g/L tryptic soy broth). Solid medium for *Streptomyces* spore isolation is ISP4 (37 g/L Difco ISP medium 4). Fermentations at 500- and 50-mL scale were supplemented with 3% w/v hydrophobic resin (Amberlite – XAD16; Alfa Aesar L19565-36). Unless stated otherwise, antibiotics apramycin, nalidixic acid, kanamycin, and chloramphenicol were supplemented to liquid or solid media at final concentration of 50 µg /mL, 25 µg /mL, 50 µg /mL and 25 µg /mL, respectively. Media components and all other chemicals were purchased from standard

commercial sources. T4 DNA ligase (Promega; Cat. No. M1794) and its 10X buffer used in this study were purchased from Promega unless stated otherwise.

4.4.2 Generation of *E. coli* NEB ® Stable/pUZ8002

Plasmid pUZ8002 (www.lifescience-market.com; Cat. No. PVT12448) was electroporated into electrocompetent *E. coli* NEB ® Stable cells (BioRad electrocuvette, 1 mm gap; 1.8 kV; BioRad Gene Pulser Xcell electroporation system). The electroporation mixture was plated on LB solid media supplemented with kanamycin to select for transformants. The plate was incubated at 37°C overnight. The colonies appeared on selective solid media were transferred to liquid LB medium supplemented with kanamycin.

4.4.3 Statistical Analysis

Microsoft Excel and Python were used for statistical analysis in this study.

4.4.4 Generation of Five-level Plackett-Burman Design

A five-level Plackett-Burman design is generated by cyclic permutation as described previously [156]. Specifically, the design cycles through five numbers, which are 1, 2, 3, 4 and 5. The orthogonality is verified in JMP's design diagnostics. The resulted Plackett-Burman matrix is found in **Supplementary Data File A2, Worksheet 1**. The columns selected to guide genetic design of synthetic MEP pathway are highlighted in grey in the **Supplementary Data File A2, Worksheet 1**.

4.4.5 DNA Parts Construction

The DNA assembly strategy, DNA parts, and plasmids used to construct synthetic MEP pathway prototypes in this study is the same as previously described in chapter 3 and [74]. The promoter-RBS parts, and terminator parts were chemically synthesized as Ultramer® DNA Oligonucleotides (Integrated DNA Technologies, Coraville, IA). Subsequently, the DNA parts were assembled into part storage vector. The CDSs of *ispH*, *ispG*, *dxs*, *dxr*, *ispD*, *ispE*, *ispF*, and *idi* were previously cloned and domesticated by PCR

amplification with gene specific primers flanked by *Aar*I restriction sites and a four-nucleotide *Aar*I docking sequence. The domesticated CDS are free of any internal *Sap*I and/or *Aar*I restriction sites. Eight genes were cloned from genomic DNA of *Streptomyces* sp. NRRL S-1813. The PCR-amplified CDS fragments were domesticated by *Aar*I Golden Gate assembly into CDS domestication vector pCDS. Each PCR-amplified CDS fragment was analyzed by gel electrophoresis and extracted with standard gel purification protocol (Zymoclean Gel DNA Recovery Kit, Zymo Research Cat. No. D4007). Domestication of PCR-amplified CDS were done by minimized *Aar*I Golden Gate reaction with minor modifications to the manufacture's protocol: 30 ng of pCDS, three equimolar amounts of PCR-amplified CDS fragments and 2 µL of 2X *Aar*I master mix (Thermo Fisher Scientific, GeneArt™ Type IIs Assembly Kit, *Aar*I; Cat. No. A15916), and water to final volume of 4 µL. The reaction mixtures were incubated as per manufacturer's *Aar*I Golden Gate incubation protocol, which is described below: 37°C initial digest for 15 min, initial ligation at 16°C for 5 min, then cycling between 37°C for 1 min and 16°C for 1min for at least 20 cycles. Half of the Golden Gate assembly reaction was introduced to *E. coli* via chemical transformation. The clones were verified by Sanger sequencing.

4.4.6 Bottom-up Hierarchical DNA Assembly of Synthetic Gene Clusters

The method of hierarchical DNA assembly is the same as described in chapter 3. The workflow of DNA assembly pipeline is summarized in **Supplementary Figure A2.1** Three type of destination vectors were used: level 1 destination vector for cisReg/cistron storage, level 2 destination vector for partial cluster storage, and level 3 destination vector for full cluster storage. Level 1 (CisReg/Cistron) destination vector library has unique combinations of upstream and downstream MoClo scars flanked by *Bbs*I and *Aar*I recognition sites and apramycin resistance cassette. Each of the level 2 (partial cluster) destination vector has unique combinations of upstream and downstream MoClo scars flanked by *Aar*I and *Sap*I recognition sites and kanamycin resistance cassette. Level 3 (full synthetic gene cluster) destination vector has of upstream and downstream MoClo scar sequences A and E flanked by *Sap*I and *Aar*I recognition sequences and apramycin resistance cassette.

*Bbs*I golden gate assembly was used to construct CisReg plasmid. Level 1 destination vector and part plasmids of promoter-RBS, transcriptional terminator, and reporter gene (*Streptomyces* codon-optimized *egfp*) were the substrates. Forty femtomoles of level 1 destination vector, three equimolar amounts of each part plasmids, 5 U *Bbs*I, 5 U T4 DNA ligase, 1 µL 10x T4 DNA ligase buffer and water were added to bring the final reaction volume to 10 µL. The incubation parameter is the same as the previous *Aar*I Golden Gate reaction. Half of the Golden Gate assembly reaction was introduced to *E. coli* via chemical transformation. The clones were verified by Sanger sequencing.

*Sap*I Golden Gate assembly was used to construct monocistronic plasmids with CisReg plasmids and pCDS plasmids. Forty femtomoles of CisReg plasmid, three equimolar amounts of pCDS plasmid, 5 U *Sap*I (New England Biolabs Inc; Cat. No. R0569S), 5 U T4 DNA ligase, 1 µL 10x T4 DNA ligase buffer and water were added to bring the final reaction volume to 10 µL. The incubation parameter is the same as the previous *AarI* Golden Gate reaction. Half of the Golden Gate assembly reaction was introduced to *E. coli* via chemical transformation. The monocistrons were verified by either colony PCR using OneTaq® PCR (New England Biolabs Inc.; Cat. No. M0480S) or Sanger sequencing.

*Aar*I Golden Gate assembly was used to construct partial clusters with monocistronic plasmid and level 2 destination vector. Thirty nanogram of level 2 destination vector, three equimolar amounts of monocistronic plasmids, 2 µL 2X *Aar*I master mix, and water to final volume of 4 µL. The incubation parameter is the same as the previous *Aar*I Golden Gate reaction. Half of the Golden Gate assembly reaction was introduced to *E. coli* via chemical transformation. The monocistrons were verified by colony PCR.

*Sap*I Golden Gate assembly was used to construct full cluster plasmids with partial plasmids and level 3 destination vectors. Forty femtomoles of level 3 destination vector, three equimolar amounts of partial cluster plasmids, 5 U *Sap*I, 5 U T4 DNA ligase, 1 µL 10x T4 DNA ligase buffer and water were added to bring the final reaction volume to 10 µL. The incubation parameter is the same as the previous *Aar*I Golden Gate reaction. Half of the Golden Gate assembly reaction was introduced to *E. coli* via chemical

transformation. The monocistronic part were verified by colony PCR. The colony PCR primers are the same as the CDS cloning primers. One gene from each cassette was targeted for colony PCR.

4.4.7 Intergenic Conjugation

Final synthetic gene clusters were introduced into conjugation donor *E. coli* ET12567/pUZ8002 or *E. coli* NEB® stable/pUZ8002 via electroporation (BioRad electrocuvette, 1 mm gap; 1.8 kV; BioRad Gene Pulser Xcell electroporation system). The electroporation mixtures were recovered with 1 mL S.O.C. media for 1 h in 37°C shaker. Upon recovery, the electroporation mixture was plated on LB agar supplemented with kanamycin, chloramphenicol, and half concentration of apramycin. The transformations were incubated at 37°C overnight. The transformants were verified with colony PCR. Two primer pairs (SYH241-SYH242 and SYH 233-SYH234) were used as colony PCR primers to amplify the *ispH-ispG* and *ispD-ispE* junctions in the full clusters (**Supplementary Figure A2.2)**. Alternative primer pairs (SYH237-SYH238 and SYH 243-SYH244) were used as colony PCR primers to amplify *ispG-dxr* and *ispE-ispF* junction. Each verified transformant was inoculated in 2 mL liquid LB supplemented with apramycin, kanamycin, and chloramphenicol and incubated in 37°C shaker overnight. On the next day, 2 mL of LB supplemented with apramycin, kanamycin, chloramphenicol and $CaCl_2$ were added to the overnight culture (20 mM $CaCl_2$ final concentration) and incubated for four more hours in 37°C shaker.

Preparation and germination of the spore of the wild type *S. albidoflavus* J1074 for conjugation was performed in the following steps. First, the frozen spore stocks of *S. albidoflavus* J1074 were thawed on ice for 10 min. The thawed spore stocks were centrifuged at max speed for 1 min, and the supernatant was removed. Every 50 µL of packed spores was resuspended in 500 µL TSB supplemented with 10% (wt/vol) sucrose and glycine. Germination of spores was initiated in 28°C shaker for at least 2 h.

Once conjugation donors were overgrown for 4 h, the cells were pelleted by centrifugation at 3,000 rpm for 30 s. The supernatant was removed, and the cell pellet was washed gently three times, each time with 1 mL LB at room temperature to remove antibiotics. The final cell pellet was resuspended in 250 µL of germinated spore solution.

The *E. coli* – *Streptomyces* conjugation mixtures were centrifuged 3,000 rpm for 1 min and the excess supernatant was removed so the final resuspension volume was 50 µL. Each conjugation mixture was plated on ISP4 solid media supplemented with 20 mM $CaCl_2$ (final concentration) and incubated overnight at 28°C. After 24 h, the entire conjugation mixture was scrapped and re-plated on ISP4 solid media supplemented with nalidixic acid and half concentration of apramycin to select for successful *Streptomyces* exconjugants. The final *Streptomyces* exconjugants were transferred to an ISP4 solid medium supplemented with full concentration of apramycin for maintenance and downstream experiments.

The PBFC plasmid were verified via colony PCR. All PBFC constructs except PBFC81 were successfully electroporated into *E. coli* ET12567/pUZ8002 or *E. coli* NEB® stable/pUZ8002 as the conjugation donor. With four failed attempts, we confirmed that was not able to be introduced into *E. coli* ET12567/ pUZ8002 for unknown reason. Instead, PBFC21 was introduced into *E. coli* NEB stable®/ pUZ8002 and confirmed with colony PCR. Intergenic conjugation of PBFC plasmids into *S. albidoflavus* J1074 were successful except for PBFC 34 and 108. PBFC 34, 81 and 108 could not yield any *S. albidoflavus* J1074 exconjugants so we considered their eAA production titers as 0 mg/L in data analysis.

4.4.8 Fermentation of *S. albidoflavus* J1074

The fermentation protocol of S. *albidoflavus* J1074 was the same as previously described in chapter 3 and [74]. The *Streptomyces* fermentation was divided in three stages and they were (i) fresh sporulation of strains on solid media, (ii) germination and synchronization of cultures and (iii) inoculation of the fermentation media. All 50 mL and 500 mL shake flask cultivations were done in 250 mL and 2 L standard baffled Erlenmeyer flasks, respectively. High-throughput 2 mL and 0.5 mL microcultivation were done in low-evaporation, 24- or 96-square-deep-well microtiter plates (MTPs), respectively (Enzyscreen CR1424a and CR1496a).

At the sporulation stage, the frozen spore stocks of recombinant S. *albidoflavus* J1074 strains were streaked on ISP4 solid media supplemented with apramycin and incubated in 28°C for 2 days to obtain fresh spores. For the PBFC strains that do not

sporulate fully, their vegetative mycelia were used for inoculation. At the spore germination stage, fresh spores from each recombinant strain were grown in 2 mL R2YE supplemented with apramycin in 15 mL culture tube. The R2YE liquid cultures were incubated in 28°C shaker with 250 rpm for 48 h. Each R2YE preculture was used to inoculate (1/100 (v/v) inoculum) 2 mL ISM3 supplemented with apramycin in 15 mL culture tube to synchronize growth of various strains. The synchronization culture for 500 mL fermentation was scaled up to 50 mL ISM3 supplemented with apramycin and was also inoculated with 1/100 (v/v) R2YE preculture. The ISM3 cultures were incubated in 28°C shaker with 250 rpm for 48 h. Each synchronization culture was standardized so that the cell pellet inoculated across samples was the same. Briefly, 1 mL of the synchronization culture was gently pelleted in sterile 1.7 mL Eppendorf microcentrifuge tubes with 3,000 rpm for 30 s. Excess supernatant was removed so the volume of cell pellet and the supernatant were 1 to 1 ratio. Each microfermentation (0.5 mL) was inoculated with 10% of standardized synchronization cultures. The fermentation cultures were grown in 28°C shaker with 250 rpm for 4-7 days.

4.4.9 Extraction of Metabolites from *Streptomyces* Fermentation

The 50 mL fermentations were transferred to 50 mL conical tubes and centrifuged at 4,500 rpm for 15 min. The supernatant was discarded, and the cell-resin pellets were washed with de-ionized water (ddH_2O) and centrifuged for 30 min at 4600 rpm for three times. The samples were frozen at -80°C for 24 h. The frozen samples were then freeze-dried for 24 h or until the moisture was completely removed (0.01 mbar, -84°C; Labconco FreeZone Plus 2.5 Liter Cascade Benchtop Freeze Dry System, catalog number 7670020). Each dried sample was extracted in methanol (3 x 5 mL).

The 96-well MTP containing 0.5 mL fermentations was frozen at -80°C for 24 h to 48 h. The frozen samples were then freeze-dried for 24 h or until the moisture was completely removed. Each sample were extracted with 1.25 mL methanol.

4.4.10 Liquid Chromatography–mass Spectrometry (LC-MS) Analysis

LC-MS analyses were performed on Thermo Scientific MSQ plus mass detector equipped with Ultimate 3000 Rapid Separation (RS) system. The extracts were diluted in 100% MeOH, and 10 µL of each extract was loaded onto a Thermo Scientific analytical BetaSil C18 column (2.1 x 150 mm, 3 µm). Chromatography was performed at flow rate of 500 µL/min using 10 mM ammonium formate buffer (solvent A, pH 8.3) and 100% acetonitrile (solvent B) with the following gradient: 0 – 0.5 min, 95 % solvent A; 0.5 – 14.5 min, 95% solvent B; 14.5 – 16 min, 95% solvent B; 16 – 18 min, 95% solvent A; 18 – 23 min, 95% solvent A. The parameters of MSQ plus mass detector for analyzing eAA were described below: ionization mode, electrospray; cone ionization, 110 V; capillary probe temperature, 450°C.

4.4.11 One-way ANOVA

One-way ANOVA was used to calculate the significance of the expression of each gene to *e*AA titer. To conduct ANOVA properly, the eAA titer data, which follows a lognormal distribution, is required to be transformed to normal data. To make zero values in the dataset countable, 1.0 mg/L was added to each of data point. Then the $\log_{10}$-transformed eAA titer data y were used in one-way ANOVA. Microsoft Excel Data Analysis ToolPak was used to conduct one-way ANOVA.

4.4.12 Main Effects Calculation

First, the effect E of each expression level L of individual genes is calculated by taking the arithmetic mean of $\log_{10}$-transformed eAA titer y of the number n of PBFC strains containing that specific expression level L.

$$\boldsymbol{E_L = \left[\sum \mathbf{y}(PBFC_i, PBFC_j, \ldots PBFC_k)\right]/n} \text{ (**Equation 4.1**)}$$

In this case, $n = 25$ for each L.

Then the effect E is plotted against the actual $\log_{10}$-transformed expression strength x corresponding to the expression level L. An ordinary least squares model (OLS) function in the Python statsmodel package was used to conduct linear regression and calculate the slope β. The main effects of gene expression on production titer of eAA was

defined as the slope β of the simple linear fit model. The linear regression line has an equation of the form:

$$y = a + \beta x \text{ (**Equation 4.2**)}$$

The statistical significance of the linear fit was calculated by t-test.

4.4.13 Contribution of Discrete Genetic Motifs to Overall eAA Titer

The fitness f of a genetic motif x, whether for a monocistronic unit MT or a partial cluster PC in the assembly, is defined as the arithmetic mean of the $\log_{10}$-transformed eAA titer of full clusters containing that part.

Genetic motif fitness is shown for monocistronic unit and partial clusters in **Figure 4.4** as the logarithm of the ratio of individual part fitness to average part fitness. Better-than-average parts fall above the x-axis and worse-than-average parts fall below the x-axis.

4.4.14 Using Genetic Motif Fitness to Predict Full Cluster Activity

Additivity, multiplicity, and dominant negative models were used to retroactively calculate the predicted fitness of a full cluster f_{PBFCi} from the part fitness values of monocistronic units f_{MT} or of partial clusters f_{PC}.

The additivity model predicts the full cluster fitness to be the average fitness of its constituent monocistronic units or partial clusters:

$$f_{PBFCi} = \frac{1}{8}(f_{MTi} + f_{MTii} + f_{MTiii} + f_{MTiv} + f_{MTv} + f_{MTvi} + f_{MTvii} + f_{MTviii})$$

or

$$f_{PBFCi} = \frac{1}{2}(f_{upper\_PC} + f_{lower\_PC})$$

The multiplicativity model predicts the full cluster fitness to be the product fitness of its constituent monocistronic units or partial clusters:

$$f_{PBFCi} = f_{MTi} \cdot f_{MTii} \cdot f_{MTiii} \cdot f_{MTiv} \cdot f_{MTv} \cdot f_{MTvi} \cdot f_{MTvii} \cdot f_{MTviii}$$

or

$$f_{PBFCi} = f_{upper\_PC} \cdot f_{lower\_PC}$$

The multiplicativity model predicts the full cluster fitness to be the product fitness of its constituent monocistronic units or partial clusters:

$$f_{PBFCi} = \min\left(f_{MTi}, f_{MTii}, f_{MTiii}, f_{MTiv}, f_{MTv}, f_{MTvi}, f_{MTvii}, f_{MTviii}\right)$$

or

$$f_{PBFCi} = \min\left(f_{upper\_PC}, f_{lower\_PC}\right)$$

4.4.15 *Streptomyces* Morphology Photography

Plates of ISP4 solid medium were made using automatic plate pourer two days before streaking PBFC strains. Each plate had 22 mL of ISP4 medium (100 x 15 mm petri dishes, VWR, cat. No 25384-342). On the day of streaking, ISP4 plates were pre-warmed in 28°C incubator for 20 minutes. Frozen glycerol stocks of PBFC strains were placed on ice to prevent thawing. Each PBFC strain was streaked onto ISP4 plate supplemented with 50 ug/mL Apramycin. These plates were incubated at 28°C for seven days.

Nikon stereo zoom microscope (SMZ-2B), Meiji Techno RZBD/LED focus stand, AmScope microscope digital camera (MU1000-HS-CK) were used for photography of PBFC colonies. The photography setup is shown in **Supplementary Figure A2.3**. The left arm of the AmScope LED Dual Gooseneck Illuminator (item #: T9FM99286) was used as an additional light source. The light source from the Amscpe illuminator was 5 inch above the plate. The angle of the lighting is approximately 45°. The brightness of lighting was set at 8.00.

Images of PBFC colony morphology were captured using Amscope digital camera and its Amscope software for MU series (https://www.amscope.com/software-download#toup1; version v3.7.13522). Before the photography, the white balance was adjusted a plate that had more than one colony in the edge with white paper as the reference in the background. The color temperature was 6243 and the tint was 1000. The parameters of exposure and resolution were set at follows: exposure target, 120; exposure time, 92.536 ms; analog gain, 1.00; The parameters of capture and resolution were set as follow: live: 3584 x 2748; snap: 3584 x 2748; format: RGB24.

Additional methods and supplementary figures can be found in the **Appendix 2** and **Supplementary Data File A2** (separate excel file).

Chapter 5 Conclusion

Summary

Dr. Richard Feynman's well-known quote "What I cannot create, I do not understand" echoes the sentiment of efforts to determine the underlying principles of biosynthetic multi-gene systems by reconstructing them at the DNA level. While many questions remain to be answered and new ones have surfaced, I have presented a platform to rapidly construct metabolic pathway and efficiently balance the expression of genes encoding the biosynthesis of natural products in *Streptomyces*. In the quest to identify relative expression levels that lead to optimal performance of a multi-gene system, algorithmic DNA assembly enables the rapid, robust generation of complex, diverse libraries.

In this dissertation, I have described multiple contributions I made during my doctoral work to advance field of *Streptomyces* natural product discovery with the focus of engineering expression level of multi-gene systems. These contributions centered on learning how to efficiently balance gene expression in a refactored metabolic pathway. The two technologies I developed, a high-throughput algorithmic assembly of synthetic BGCs and a library of synthetic promoter-RBS parts to combinatorial control gene expression, resulted in a platform for rapid metabolic pathway prototyping in *Streptomyces* bacteria. In addition, my doctoral work also explores novel strategies for the optimization of multi-gene systems with a focus on isoprenoid metabolism. The main hypothesis is that mathematical optimization paradigms, which have long been used in industrial and chemical engineering, can be extended to well-defined genetic systems. As a proof of concept, a multivariate DoE method called Plackett-Burman design is used to guide the design of a library of 125 synthetic gene clusters encoding a natural MEP pathway. Each MEP pathway synthetic gene cluster is composed of monocistronic unit of eight gene and the relative expression level of these eight genes are designed tuned to one of five possible expression level. The library of 125 synthetic gene clusters only covers 0.032% of the entire design space (125 designs of all 5^8 possible designs). Although it seems that the tradeoff for smaller library size is the resolution of the MEP pathway landscape, we successfully showed that a medium size library guided by Plackett-Burman design

identified DXS as the same bottleneck enzyme of the MEP pathway as previously reported. The approach provided confidence that Plackett-Burman design can be used reliably to identify important genetic features for the optimization of less well-characterized multi-gene systems.

During my PhD, I overcame a few technical challenges in designing prototypes of multi-gene systems to address the main questions of re-engineering BGCs. First, the traditional way of restriction-enzyme based cloning is labor-intensive and usually requires case-by-case optimization. Generating different genetic prototypes is low-throughput and thus creates a major bottleneck prior to the hypothesis testing step. To overcome the challenge, I established an algorithmic high-throughput DNA assembly pipeline. In other words, the process of this DNA assembly is an algorithm: a defined series of Golden Gate assembly-based steps a user always performs, one after another. While it is possible to assemble the final multi-gene construct from small, standardized DNA parts in a single cloning reaction, the assembly efficiency drops drastically as the number of DNA fragments increases. To minimize the number of transformants that needed to be screened to validate assembly reactions, we limited the complexity of each DNA assembly step. The first step combines CDSs with cis-regulatory elements like promoters, CDSs and terminators, which form a 'CisReg' plasmid, which is a storage plasmid that contains cis-regulatory elements. CisReg plasmid is also used to quantitatively measure the promoter strength. The second step is a scarless replacement of the reporter gene with the desired CDS, forming a monocistronic unit. The third step is bringing multiple monocistronic units together to form a partial cluster; and finally, multiple partial clusters to form a full cluster.

Banking all intermediate plasmids during the hierarchical assembly plan has a few technical benefits. First, plasmids can be amplified in cloning strains if the subsequent assemblies require more of a particular plasmid as substrate. Second, the intermediary and final assemblies in a plasmid format can be stored long term, reducing the need to repeat assembly reactions. Third, Golden Gate assembly achieves higher efficiency with plasmids as substrates when compared to linear DNA fragments (unpublished data). Fourth, this facilitates combinatorial re-use of parts of which their reliability is known. The

DNA parts stored in standardized format ensure design compliance to the algorithmic DNA assembly pipeline, which reduces overall engineering and design time.

In sum, the features of this algorithmic DNA assembly pipeline are carefully incorporated to optimize assembly efficiency, accuracy, and throughput. This algorithmic DNA assembly pipeline serves as a universal scheme that can be applied to any genetic system construction project and shortens the time between prototyping and testing. This innovation to the traditional cloning is analogous to the transition to mass production of standardized goods from craft production during the Industrial Revolution.

Second, a central challenge to constructing a functional multi-gene system is the proper control of its gene expression. When I started my dissertation project in 2015, the number of characterized synthetic cis-regulatory elements to control gene expression in *Streptomyces* was limited. To address the need of controlling expression for multiple genes, I constructed a library of synthetic, constitutive promoter-RBS (combined promoter and RBS parts) to tune gene expression at the levels of transcription and translation. For my project, the ideal properties of a synthetic expression control element are: (i) the ability to activate gene expression (ii) the ability to bypass the host regulatory machinery and (iii) consistent activity at different environmental conditions. I also attempted to use a native, strong constitutive *Streptomyces* promoter called ermE* as a benchmark in the expression strength measurement. However, domesticating ermE* to make it compatible with the algorithmic DNA assembly pipeline, which introduced a single point mutation to remove an internal *Aar*I restriction site in the ermE* sequence, completely abolished its activity (unpublished data). So, my work has therefore been fully focused on building synthetic regulatory elements in the first year of my PhD work. A handful of synthetic promoter sequences with varying expression strengths were selected from the synthetic promoter library constructed by Siegl et [111]. However, these synthetic promoters were concatenated with the native RBS region of the reporter gene *gusA* for expression strength measurement in the original work. To add control at the level of translation, 10 *de novo* RBS sequences of varying translational initiation efficiency were generated by the RBS calculator [26], [112]. The synthetic promoter and RBS sequences were chemically synthesized to form diverse sequences, yielding 21 unique promoter-RBS parts. This library of synthetic promoter-RBS provides expression control over three orders of

magnitude. Another benefit of synthetic promoter-RBS parts is that media components do not interfere with the expression strength of promoter-RBS parts, as their expression levels are consistent in two different media. As most BGCs reach maximal expression during stationary phase, replacing their native promoters and RBSs with synthetic expression elements decouples their expression from the temporal control of the host regulation; the synthetic, promoter-RBSs are constitutive and tightly coupled with growth, allowing accumulation of the encoded product at earlier growth phase. This is observed in the time-dependent accumulation of eAA in a 2L flask fermentation in chapter 3.

The technology developments of the algorithmic DNA assembly pipeline and the library of synthetic promoter-RBS enables the scale up of agile genetic prototyping. As a result, I was able to assemble a small library of 14 prototypes to constitute eAA biosynthesis in *Streptomyces* and another library of 125 prototypes to investigate the robustness of isoprenoid precursor metabolism. The main goal of constructing the initial library of seven eAA synthetic gene clusters was to study the relationship between the genotype and chemotype as well as identify the gene encoding the cognate atisane oxidation enzyme. The genes were selected based on their predicted biochemical functions. The genes are first paired with individual promoter-RBS and terminator parts to create monocistronic units, subsequently they were assembled together to form the final eAA synthetic gene clusters. The genetic design permutated the gene content in the oxidation cassette, which contained genes encoding an alpha-ketoglutarate dependent dioxygenase (PtmO6), a cytochrome P450 (PtmO2), or a ferredoxin (PtmO9) or combinations of these genes. The second design consideration was to perturb the gene expression level of the predicted *ent*-atiserene synthase PtmT1 to verify its function. Based on these design considerations, the eAA gene cluster prototypes were constructed and revealed a few insights about *e*AA biosynthesis. First, tuning the relative expression level of genes encoding the eAA pathway created chemical diversity at the metabolite level. Each variant produced different ratios of shunt metabolites and the target compound, eAA. The shunt metabolite species identified in this work are novel chemical entities, which are derived from two pathway intermediates, geranylgeranyl diphosphate and *ent*-copalyl diphosphate. We also identified a novel cytochrome P450 enzyme, PtmO2, which catalyzes stereospecific oxidation of the C19 methyl group of atiserene scaffold to afford

final product eAA. In addition, we found that the relative expression of PtmO2 and the *ent*-atiserene synthase, PtmT1, is crucial to the production titer of eAA. PtmO2 was found to oxidize the intermediate *ent*-copalyl diphosphate, the substrate of PtmT1, and form a shunt metabolite if PtmT1 expression is too low. Based on the eAA synthetic gene cluster as the model system, the second iteration of synthetic gene cluster (SGC) library design aimed to perturb the relative expression level of eight genes encoding the MEP pathway. We increased the eAA titer from 50 mg/L to 500+ mg/L by adding genes for the entire MEP pathway plus the isopentenyl pyrophosphate-dimethylallyl pyrophosphate (IPP-DMAPP) isomerase, *idi*, to one of the synthetic eAA gene clusters. Surprisingly, the genetic designs of the best producers from the eAA library only have the ferredoxin gene expressed at the highest level and the rest of the gene expressed at medium level. This is contrary to the conventional wisdom in metabolic engineering, which is to use strong promoters to drive strong expression of individual genes. The lesson learned from constructing a synthetic eAA biosynthetic pathway support the emerging concept that production titer is not linear with the expression level [102].

Once the utility of the synthetic biology platform I developed was validated by the eAA synthetic gene cluster model system, I moved on to build a more elaborate library of 125 prototypes of the synthetic gene clusters of the natural MEP pathway. The main goal was to investigate the relationship between the expression levels for the MEP pathway genes and the performance of the isoprenoid precursor metabolism, ultimately to optimize the isoprenoid production titer. The design approach for this specific library is based on unbiased, multivariate statistical DoE method rather than a heuristic approach like the previous library. The conventional approach to characterize a pathway and its optimization is an iterative process in which one variable is changed at a time while keeping other variables fixed to identify the optimal genetic design. A downside to this approach is that it might miss the global optimal solution, as the quality of the search is determined by the starting point. The basis of variable selection in DoE either requires *a priori* knowledge or uses random starting points. Another disadvantage is that it is often impractical to test all possible genetic designs. To address these issues of identifying optimal genetic design, I used Plackett-Burman design, which simultaneously changes the variables to maximize information while minimizing the number of tests, to screen the most important gene for

the isoprenoid titer. Based on the synthetic biology platform and a five-level Plackett-Burman design, I built 125 prototypes of an eight-gene system encoding a synthetic MEP pathway. In this library design, the gene order and gene content remain constant; the variable is the expression level of individual gene, which is set at one of the five expression strengths. A full factorial experiment requires 5^8 (= 32,768) experiments, while a 5-level Plackett-Burman design only needs 125 combinations to fully gather the effect of individual gene expression. The screening of the entire 125 variants has revealed a surprising degree of robustness in actinobacterial secondary metabolism. The MEP pathway metabolism is quite robust, as perturbation of the expression of 6 out of 8 genes does not substantially change the average production titer. In addition, the analysis of the Plackett-Burman result replicated findings reported in the MEP pathway literature, which identify *dxs*, the gene encoding the first and flux-controlling enzyme of the MEP pathway, as the most dominant gene to influence the isoprenoid titer. This gives us confidence that Plackett-Burman design is an effective DoE method to identify the most important variable in a extensively studied multi-gene system like the MEP pathway, and can be extended to screen important genes in other multi-gene system with little or no working knowledge.

Future directions

For the immediate future, I propose a few potential projects as the expansion of research presented in this dissertation. The first project will be to understand the mechanism of action of serofendic acid, a neuroprotective drug candidate. I tested *e*AA, serofendic acid and the structural analogs synthesized by Dr. Dimitri Perusse in the yeast chemical-genetics platform developed by Dr. Chad Myers' group (University of Minnesota, Department of Computer Science and Engineering). The results, which were quantitative chemical-genetic interaction profiles, contained functional information useful to predict bioprocess targets of serofendic acid. Sarah Jo DeVore (a Bioinformatics and Computational Biology master student), is currently tasked with performing the bioinformatic analysis to predict the molecular target of serofendic acid.

The second potential project will be to complete the biosynthesis of serofendic acid, which has an additional C15 hydroxyl group and the C17 methylsulfinyl group compared

to the structure of eAA. I propose two approaches to discover the enzymes performing these specific reactions. First, creating a cDNA library of cytochrome P450 genes from *Isodon eriocalyx* [123] and *Lepidolaena clavigera* [124], the two plant species reported to produce *ent*-atisanoid with a specific C15 hydroxyl group. The second approach is using the panels of previously published cytochromes P450 of fungi known for biotransformation of xenobiotics. As for installing a sulfur-containing functional group on eAA, liver microsomes of different organisms could be an interesting tool as they catalyze unspecific methylthiolation [125], [190].

The third project will be the follow-up project of chapter 4, the rational search of genetic designs for optimal isoprenoid production. The insights provided by the analyses of Plackett-Burman design and the part fitness will be used to design a second iteration of the synthetic MEP pathway library. The screening result of this second library can be used to evaluate the predictive power of the rules identified in the Plackett-Burman design and part fitness analysis. We plan to validate the findings resulted from the analysis of Plackett-Burman design. Specifically, A synthetic gene cluster will be constructed with eight genes with the optimal expression level identified in **Figure 4.3**. The further optimization improvement will be focused on constructing additional FCs from PCs with the highest fitness values (**Figure 4.4a**). Three upper PCs were selected, and they are PC8, PC19, and PC11. Three lower PCs were selected, and they are PC28, PC30, and PC42. A library of full factorial combination (3 × 3) of partial clusters will be constructed. RNA-seq analysis can be performed on the best producer to measure the transcriptional profile of the synthetic MEP pathway. It would be interesting to see how the overexpression of the synthetic MEP pathway alters the global transcription landscape of the host cell. Furthermore, it would be powerful to integrate kinetics data of MEP pathway enzymes, transcriptomics data, and flux-based analysis to establish a MEP pathway model in *Streptomyces*.

In addition to enabling bioproduction of value-added isoprenoids [101], [102], [191]–[194], we are hoping to apply our findings about MEP pathway for new terpene scaffold discovery. The recent whole genome data reveals that the *Streptomyces* genomes harbor many novel terpene synthases which potentially encode new terpenoids[195]. Several *Streptomyces* diterpenoids have been discovered since 2015:

tsukubadiene [196], fusicomycins [197], labdanmycins [198], and tiancilactone [199], just to name a few. The bottleneck of discovering novel terpene synthases is that (i) either activity loss in heterologous expression host of a more distance genetic relationship (i.e. *E. coli*), or (ii) tightly regulated pools of isoprenoid building blocks for specialized metabolism in the native host, which leads to little or non-detectable level of encoded products. By designing and introducing a synthetic MEP pathway gene cluster into the *Streptomyces* strains harboring new terpene synthase genes, their isoprenoid production would be potentiated for production of novel terpene scaffold. Alternatively, the *Stretpomyces* terpene synthase genes can be expressed in a model *Streptomyces* isoprenoid overproducing strain. The high-throughput DNA assembly methodology I developed during my doctoral training can increase the scale of refactoring *Streptomyces* cryptic gene clusters for novel natural product discovery. The first goal will be to increase the yield of the encoded product, which is useful for its detection, isolation, and characterization. There is also an opportunity for combinatorial biosynthesis, as the algorithmic DNA assembly pipeline is readily amenable for generating combinatorial libraries from smaller modular expression cassettes. For example, a plug-and-play approach allows swapping novel genes encoding terpene synthases and tailoring enzymes for scaffold and functionalization discovery.

I believe an automated platform which incorporates robotics, DNA design and assembly, and high-throughput screening can substantially increase the productivity in natural product discovery. The two technologies I developed during my PhD have the potential to be integrated with existing automated workflows, which will further streamline the ‘design’ and ‘build’ aspect of the “design-build-test-learn” cycle in synthetic biology. There are already software packages dedicated to automating the generation of multivariate statistical DoE designs (e.g. JMP, Minitab, and R). The methodical experimentation approach can be used to optimize an objective function of a multi-gene system, as scientists are becoming more capable of precisely tuning expression of multiple genes. In a refactored system, the expression levels may be tuned as variables analogous to independent variables in an industrial process (time or quantity) The implementation of algorithmic DNA assembly is repeatable, straightforward, and the steps can be easily translated to a sequence of actions in a software program. The Densmore

group identified major staff time savings in automated DNA assembly for a large number of assemblies; assemblies of 42 constructs required 51 minutes of manual execute time and 15 minutes of automated time [200]. During the last five years, I have manually assembled and validated at least 550 plasmids for my doctoral project, which are approximately equivalent to 8.2 megabases of DNA. I speculate that an automated DNA assembly workflow can potentially assemble the same number of plasmids only half of time. The automation of the DNA assembly will also encourage a formalized protocol for better sharing of reliable results and reproducibility.

In the last two decades, the developments in *de novo* synthesis and combinatorial DNA assembly have allowed us to perform more sophisticated experiments. Specifically, the ability to synthesize user-defined oligonucleotides and long DNA sequences improves access to the unexplored nucleotide sequence space. Novel combinatorial assembly methods increase the throughput of generating complex, diverse multi-gene constructs from libraries of standardized, functional DNA parts. These technological capabilities provide unprecedented opportunities to radically engineer the regulation of multi-gene systems like natural product BGCs at the DNA level. My dissertation provides a platform of rapid prototyping of metabolic pathway and the application of multivariate statistical approach to optimize designed multi-gene system in *Streptomyces*, in the hopes to augment the contribution of more data-driven reasoning to design, construct, and test multi-gene system encoding more complex functions.

Reference

[1] M. L. Metzker, “Sequencing technologies the next generation,” *Nature Reviews Genetics*, vol. 11, no. 1. pp. 31–46, Jan-2010.

[2] S. Kosuri and G. M. Church, “Large-scale de novo DNA synthesis: technologies and applications.,” *Nat. Methods*, vol. 11, no. 5, pp. 499–507, 2014.

[3] N. Ziemert, M. Alanjary, and T. Weber, “The evolution of genome mining in microbes-a review,” *Nat. Prod. Rep.*, vol. 33, no. 8, pp. 988–1005, 2016.

[4] S. D. Bentley *et al.*, “Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2),” *Nature*, vol. 417, no. 6885, pp. 141–147, May 2002.

[5] S. Omura *et al.*, “Genome sequence of an industrial microorganism *Streptomyces avermitilis*: Deducing the ability of producing secondary metabolites,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 98, no. 21, pp. 12215–12220, Oct. 2001.

[6] D. A. van Bergeijk, B. R. Terlouw, M. H. Medema, and G. P. van Wezel, “Ecology and genomics of Actinobacteria: new concepts for natural product discovery,” *Nature Reviews Microbiology*. Nature Publishing Group, pp. 1–13, 01-Jun-2020.

[7] K. C. Belknap, C. J. Park, B. M. Barth, and C. P. Andam, “Genome mining of biosynthetic and chemotherapeutic gene clusters in *Streptomyces* bacteria,” *Sci. Rep.*, vol. 10, no. 1, pp. 1–9, 2020.

[8] M. Watve, R. Tickoo, M. Jog, and B. Bhole, “How many antibiotics are produced by the genus *Streptomyces*?,” *Arch. Microbiol.*, vol. 176, no. 5, pp. 386–390, Nov. 2001.

[9] J. Bérdy, “Bioactive microbial metabolites,” *J. Antibiot.*, vol. 58, pp. 1–26, 2005.

[10] M. H. Medema and M. A. Fischbach, “Computational approaches to natural product discovery,” *Nature Chemical Biology*, vol. 11, no. 9. pp. 639–648, 18-Aug-2015.

[11] J. R. Doroghazi *et al.*, “A roadmap for natural product discovery based on large-scale genomics and metabolomics,” *Nat. Chem. Biol.*, vol. 10, no. 11, pp. 963–968, Nov. 2014.

[12] S. S. G. W. H Zhu, “Triggers and cues that activate antibiotic production by actinomycetes,” *J. Ind. Microbiol. Biotechnol.*, vol. 41, pp. 371–386, 2014.

[13] L. Huo, J. J. Hug, C. Fu, X. Bian, Y. Zhang, and R. Müller, “Heterologous expression of bacterial natural product biosynthetic pathways,” *Natural Product Reports*, vol. 36, no. 10. Royal Society of Chemistry, pp. 1412–1436, 01-Oct-2019.

[14] S. Cardinale and A. P. Arkin, “Contextualizing context for synthetic biology - identifying causes of failure of synthetic biological systems,” *Biotechnol. J.*, vol. 7, no. 7, pp. 856–866, 2012.

[15] J. Huang, “Cross-regulation among disparate antibiotic biosynthetic pathways of *Streptomyces coelicolor*,” *Mol. Microbiol.*, vol. 58, pp. 1276–1287, 2005.

[16] K. C. G. C. G. N. H. T. G Liu, “Molecular regulation of antibiotic biosynthesis in *Streptomyces*,” *Microbiol. Mol. Biol. Rev.*, vol. 77, pp. 112–143, 2013.

[17] M. J. Smanski, J. Casper, R. M. Peterson, Z. Yu, S. R. Rajski, and B. Shen, “Expression of the platencin biosynthetic gene cluster in heterologous hosts yielding new platencin congeners,” *J. Nat. Prod.*, vol. 75, no. 12, 2012.

[18] R. G. Egbert and E. Klavins, “Fine-tuning gene networks using simple sequence repeats,” *Proc. Natl. Acad. Sci.*, vol. 109, no. 42, pp. 16817–16822, 2012.
[19] H.-J. Nah, H.-R. Pyeon, S.-H. Kang, S.-S. Choi, and E.-S. Kim, “Cloning and Heterologous Expression of a Large-sized Natural Product Biosynthetic Gene Cluster in *Streptomyces* Species,” *Front. Microbiol.*, vol. 8, no. March, 2017.
[20] G. Wang *et al.*, “CRAGE enables rapid activation of biosynthetic gene clusters in undomesticated bacteria,” *Nat. Microbiol.*, vol. 4, no. 12, pp. 2498–2510, Dec. 2019.
[21] W. Xu, E. Klumbys, E. L. Ang, and H. Zhao, “Emerging molecular biology tools and strategies for engineering natural product biosynthesis,” *Metabolic Engineering Communications*, vol. 10. Elsevier B.V., p. e00108, 01-Jun-2020.
[22] E. Andrianantoandro, S. Basu, D. K. Karig, and R. Weiss, “Synthetic biology: new engineering rules for an emerging discipline,” *Mol. Syst. Biol.*, vol. 2, no. 1, Jan. 2006.
[23] M. J. Smanski, H. Zhou, J. Claesen, B. Shen, M. A. Fischbach, and C. A. Voigt, “Synthetic biology to access and expand nature’s chemical diversity,” *Nat. Rev. Microbiol.*, vol. 14, no. 3, 2016.
[24] H. J. Frasch, M. H. Medema, E. Takano, and R. Breitling, “Design-based re-engineering of biosynthetic gene clusters: Plug-and-play in practice,” *Current Opinion in Biotechnology*, vol. 24, no. 6. Elsevier Current Trends, pp. 1144–1150, 01-Dec-2013.
[25] M. Fischbach and C. A. Voigt, “Prokaryotic gene clusters: A rich toolbox for synthetic biology,” *Biotechnol. J.*, vol. 5, no. 12, pp. 1277–1296, Dec. 2010.
[26] I. Farasat, M. Kushwaha, J. Collens, M. Easterbrook, M. Guido, and H. M. Salis, “Efficient search, mapping, and optimization of multi-protein genetic systems in diverse bacteria,” *Mol. Syst. Biol.*, vol. 10, no. 6, p. 731, Jun. 2014.
[27] M. Jeschek, D. Gerngross, and S. Panke, “Combinatorial pathway optimization for streamlined metabolic engineering,” *Curr. Opin. Biotechnol.*, vol. 47, pp. 142–151, Oct. 2017.
[28] G. Naseri, J. Behrend, L. Rieper, and B. Mueller-Roeber, “COMPASS for rapid combinatorial optimization of biochemical pathways based on artificial transcription factors,” *Nat. Commun.*, vol. 10, no. 1, pp. 1–18, 2019.
[29] E. Kim, B. S. Moore, and Y. J. Yoon, “Reinvigorating natural product combinatorial biosynthesis with synthetic biology,” *Nature Chemical Biology*, vol. 11, no. 9. Nature Publishing Group, pp. 649–659, 18-Aug-2015.
[30] F. Yan *et al.*, “Synthetic biology approaches and combinatorial biosynthesis towards heterologous lipopeptide production,” *Chem. Sci.*, vol. 9, no. 38, pp. 7510–7519, Oct. 2018.
[31] Z. Zhou, J. Gu, Y.-L. Du, Y.-Q. Li, and Y. Wang, “The -omics Era- Toward a Systems-Level Understanding of *Streptomyces*,” *Curr. Genomics*, vol. 12, no. 6, pp. 404–416, Sep. 2011.
[32] G. Wu, D. E. Culley, and W. Zhang, “Predicted highly expressed genes in the genomes of *Streptomyces coelicolor* and *Streptomyces avermitilis* and the implications for their metabolism,” *Microbiology*, vol. 151, no. 7, pp. 2175–2187, 2005.
[33] T. E. F. Quax, N. J. Claassens, D. Söll, and J. van der Oost, “Codon Bias as a Means to Fine-Tune Gene Expression,” *Molecular Cell*, vol. 59, no. 2. Cell Press,

pp. 149–161, 16-Jul-2015.
[34] Y. Saito *et al.*, “Developing a codon optimization method for improved expression of recombinant proteins in actinobacteria,” *Sci. Rep.*, vol. 9, no. 1, pp. 1–10, Dec. 2019.
[35] D. B. Goodman, G. M. Church, and S. Kosuri, “Causes and Effects of N-Terminal Codon Bias in Bacterial Genes,” *Science (80-.).*, no. October, pp. 475–480, 2013.
[36] M. Welch *et al.*, “Design Parameters to Control Synthetic Gene Expression in *Escherichia coli*,” *PLoS One*, vol. 4, no. 9, p. e7002, Sep. 2009.
[37] Y. Ueda, S. Taguchi, K. Nishiyama, I. Kumagai, and K. Miura, “Effect of a rare leucine codon, TTA, on expression of a foreign gene in *Streptomyces lividans*,” vol. 1172, pp. 262–266, 1993.
[38] E. P. Guthrie, C. S. Flaxman, J. White, D. A. Hodgson, M. J. Bibb’, and K. F. Chaterl, “A response-regulator-like activator of antibiotic synthesis from *Streptomyces coelicolor* A3(2) with an amino-terminal domain that lacks a phosphorylation pocket,” 1998.
[39] S. Hackl and A. Bechthold, “The Gene *bldA* , a Regulator of Morphological Differentiation and Antibiotic Production in *Streptomyces*,” *Arch. Pharm. (Weinheim).*, vol. 348, no. 7, pp. 455–462, Jul. 2015.
[40] A. M. Cerdeño, M. J. Bibb, and G. L. Challis, “Analysis of the prodiginine biosynthesis gene cluster of *Streptomyces coelicolor* A3(2): new mechanisms for chain initiation and termination in modular multienzymes,” *Chem. Biol.*, vol. 8, no. 8, pp. 817–829, Jan. 2001.
[41] L. Kalan *et al.*, “A cryptic polyene biosynthetic gene cluster in *Streptomyces calvus* is expressed upon complementation with a functional bldA gene,” *Chem. Biol.*, vol. 20, no. 10, pp. 1214–1224, Oct. 2013.
[42] C. Oßwald *et al.*, “Modular construction of a functional artificial epothilone polyketide pathway,” *ACS Synth. Biol.*, p. 121025132953007, 2012.
[43] Y. Lee *et al.*, “The Transcription Unit Architecture of *Streptomyces lividans* TK24,” *Front. Microbiol.*, vol. 10, p. 2074, Sep. 2019.
[44] D. J. Studholme, S. D. Bentley, and J. Kormanec, “Bioinformatic identification of novel regulatory DNA sequence motifs in *Streptomyces coelicolor*,” *BMC Microbiol.*, vol. 4, no. 1, pp. 1–25, Apr. 2004.
[45] K. Yanai, T. Murakami, and M. Bibb, “Amplification of the entire kanamycin biosynthetic gene cluster during empirical strain improvement of *Streptomyces kanamyceticus*,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 103, no. 25, pp. 9661–9666, Jun. 2006.
[46] E. J. Gehrke *et al.*, “Silencing cryptic specialized metabolism in *Streptomyces* by the nucleoid-associated protein Lsr2,” *Elife*, vol. 8, Jun. 2019.
[47] Y. Jeong *et al.*, “The dynamic transcriptional and translational landscape of the model antibiotic producer *Streptomyces coelicolor* A3(2),” *Nat. Commun.*, vol. 7, no. 1, pp. 1–11, Jun. 2016.
[48] A. Chatterjee, L. Drews, S. Mehra, E. Takano, Y. N. Kaznessis, and W.-S. Hu, “Convergent Transcription in the Butyrolactone Regulon in *Streptomyces coelicolor* Confers a Bistable Genetic Switch for Antibiotic Biosynthesis,” *PLoS One*, vol. 6, no. 7, p. e21974, Jul. 2011.
[49] F. Santos-Beneit, A. Rodríguez-García, and J. F. Martín, “Overlapping binding of PhoP and AfsR to the promoter region of glnR in *Streptomyces coelicolor*,”

Microbiol. Res., vol. 167, no. 9, pp. 532–535, Oct. 2012.
[50] Q. Zhou, S. Ning, and Y. Luo, “Coordinated regulation for nature products discovery and overproduction in *Streptomyces*,” *Synth. Syst. Biotechnol.*, vol. 5, no. 2, pp. 49–58, Jun. 2020.
[51] C. Olano *et al.*, “Activation and identification of five clusters for secondary metabolites in *Streptomyces albus* J1074.,” *Microb. Biotechnol.*, vol. 7, no. 3, pp. 242–56, May 2014.
[52] P. K. Ajikumar *et al.*, “Isoprenoid pathway optimization for Taxol precursor overproduction in *Escherichia coli*.,” *Science*, vol. 330, no. 6000, pp. 70–74, 2010.
[53] Z. Shao, G. Rao, C. Li, Z. Abil, Y. Luo, and H. Zhao, “Refactoring the silent spectinabilin gene cluster using a plug-and-play scaffold,” *ACS Synth. Biol.*, vol. 2, no. 11, pp. 662–669, Nov. 2013.
[54] S. Galanie, K. Thodey, I. J. Trenchard, M. F. Interrante, and C. D. Smolke, “Complete synthesis of opiods in yeast,” *Science (80-.).*, vol. 349, no. 6252, pp. 1095–1100, 2015.
[55] M. Myronovskyi and A. Luzhetskyy, “Native and engineered promoters in natural product discovery,” *Natural Product Reports*, vol. 33, no. 8. Royal Society of Chemistry, pp. 1006–1019, 01-Aug-2016.
[56] Y.-J. Chen *et al.*, “Characterization of 582 natural and synthetic terminators and quantification of their design constraints.,” *Nat. Methods*, vol. 10, no. 7, pp. 659–64, 2013.
[57] S. Kosuri *et al.*, “Composability of regulatory sequences controlling transcription and translation in *Escherichia coli*.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 110, no. 34, pp. 14024–9, 2013.
[58] M. M. Rudolph, M. P. Vockenhuber, and B. Suess, “Synthetic riboswitches for the conditional control of gene expression in *Streptomyces coelicolor*,” *Microbiol. (United Kingdom)*, vol. 159, no. PART7, pp. 1416–1422, Jul. 2013.
[59] L. Horbal, T. Siegl, and A. Luzhetskyy, “A set of synthetic versatile genetic control elements for the efficient expression of genes in Actinobacteria,” *Sci. Rep.*, vol. 8, no. 1, p. 491, Dec. 2018.
[60] A. E. Borujeni, D. Cetnar, I. Farasat, A. Smith, N. Lundgren, and H. M. Salis, “Precise quantification of translation inhibition by mRNA structures that overlap with the ribosomal footprint in N-terminal coding sequences,” *Nucleic Acids Res.*, vol. 45, no. 9, p. 5448, 2017.
[61] W. M. JR Doroghazi, “Comparative genomics of actinomycetes with a focus on natural product biosynthetic genes,” *BMC Genomics*, vol. 14, 2013.
[62] S. Werner, C. Engler, E. Weber, R. Gruetzner, and S. Marillonnet, “A Modular Cloning System for Standardized Assembly of Multigene Constructs,” *Bioeng. Bugs*, vol. 3, no. 1, pp. 38–43, Jan. 2012.
[63] D. G. Gibson, H. O. Smith, C. a Hutchison, J. C. Venter, and C. Merryman, “Chemical synthesis of the mouse mitochondrial genome.,” *Nat. Methods*, vol. 7, no. 11, pp. 901–3, Nov. 2010.
[64] M. J. Smanski *et al.*, “Functional optimization of gene clusters by combinatorial design and assembly,” *Nat. Biotechnol.*, vol. 32, no. 12, pp. 1241–1249, Dec. 2014.
[65] H. N. Lim, Y. Lee, and R. Hussein, “Fundamental relationship between operon organization and gene expression.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 108, no.

26, pp. 10626–10631, 2011.
[66] A. Hiroe, K. Tsuge, C. T. Nomura, M. Itaya, and T. Tsuge, “Rearrangement of gene order in the phaCAB operon leads to effective production of ultrahigh-molecular-weight poly[(R)-3-hydroxybutyrate] in genetically engineered *Escherichia coli*,” *Appl. Environ. Microbiol.*, vol. 78, no. 9, pp. 3177–3184, 2012.
[67] Z. Shao, H. Zhao, and H. Zhao, “DNA assembler, an in vivo genetic method for rapid construction of biochemical pathways,” *Nucleic Acids Res.*, vol. 37, no. 2, pp. 1–10, 2009.
[68] L. Maschio *et al.*, “Cloning, expression, and purification of intact polyketide synthase modules,” in *Methods in Enzymology*, vol. 617, Academic Press Inc., 2019, pp. 63–82.
[69] H. Jenke-Kodama, T. Börner, and E. Dittmann, “Natural Biocombinatorics in the Polyketide Synthase Genes of the Actinobacterium *Streptomyces avermitilis*,” *PLoS Comput. Biol.*, vol. 2, no. 10, p. e132, Oct. 2006.
[70] A. Casini, M. Storch, G. S. Baldwin, and T. Ellis, “Bricks and blueprints: methods and standards for DNA assembly,” *Nat. Rev. Mol. Cell Biol.*, vol. 9, no. June, pp. 1–9, 2015.
[71] O. Genilloud, “Actinomycetes: Still a source of novel antibiotics,” *Nat. Prod. Rep.*, vol. 34, no. 10, pp. 1203–1232, 2017.
[72] T. Siegl, B. Tokovenko, M. Myronovskyi, and A. Luzhetskyy, “Design, construction and characterisation of a synthetic promoter library for fine-tuned gene expression in actinomycetes,” *Metab. Eng.*, vol. 19, pp. 98–106, Sep. 2013.
[73] S. Y. Hsu and M. J. Smanski, “Designing and implementing algorithmic DNA assembly pipelines for multi-gene systems,” in *Methods in Molecular Biology*, vol. 1671, M. K. Jensen and J. D. Keasling, Eds. Humana Press Inc., 2018, pp. 131–147.
[74] S. Y. Hsu, D. Perusse, T. Hougard, and M. J. Smanski, “Semisynthesis of the Neuroprotective Metabolite, Serofendic Acid,” *ACS Synth. Biol.*, vol. 8, no. 10, pp. 2397–2403, 2019.
[75] A. Regev and E. Shapiro, “Cells as computation,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 2602, no. September, pp. 1–3, 2003.
[76] J. a N. Brophy and C. a Voigt, “Principles of genetic circuit design.,” *Nat. Methods*, vol. 11, no. 5, pp. 508–20, 2014.
[77] D. Ro *et al.*, “Production of the antimalarial drug precursor artemisinic acid in engineered yeast,” *Nature*, vol. 440, no. April, pp. 3–6, 2006.
[78] T. Scheibel, “Spider silks: recombinant synthesis, assembly, spinning, and engineering of synthetic proteins.,” *Microb. Cell Fact.*, vol. 3, no. 1, p. 14, 2004.
[79] M. Miyao, “Molecular evolution and genetic engineering of C4 photosynthetic enzymes,” *J. Exp. Bot.*, vol. 54, no. 381, pp. 179–189, 2003.
[80] L. Y. Chan, S. Kosuri, and D. Endy, “Refactoring bacteriophage T7.,” *Mol. Syst. Biol.*, vol. 1, p. 2005.0018, 2005.
[81] Z. Shao, G. Rao, C. Li, Z. Abil, Y. Luo, and H. Zhao, “Refactoring the silent spectinabilin gene cluster using a plug-and-play scaffold,” *ACS Synth. Biol.*, vol. 2, no. 11, pp. 662–669, 2013.
[82] N. Annaluru *et al.*, “Total Synthesis of a Functional Designer Eukaryotic Chromosome,” *Science*, vol. 344, no. April, pp. 55–59, 2014.

[83] D. G. Gibson *et al.*, “Enzymatic assembly of DNA molecules up to several hundred kilobases.,” *Nat. Methods*, vol. 6, no. 5, pp. 343–5, 2009.
[84] C. Engler, R. Kandzia, and S. Marillonnet, “A one pot, one step, precision cloning method with high throughput capability,” *PLoS One*, vol. 3, no. 11, 2008.
[85] A. Sarrion-Perdigones *et al.*, “GoldenBraid 2.0: a comprehensive DNA assembly framework for plant synthetic biology.,” *Plant Physiol.*, vol. 162, no. 3, pp. 1618–31, Jul. 2013.
[86] J. C. Anderson *et al.*, “BglBricks: A flexible standard for biological part assembly.,” *J. Biol. Eng.*, vol. 4, no. 1, p. 1, 2010.
[87] H. C. De Paoli, G. A. Tuskan, and X. Yang, “An innovative platform for quick and flexible joining of assorted DNA fragments.,” *Sci. Rep.*, vol. 6, no. August 2015, p. 19278, 2016.
[88] C. A. Hutchison *et al.*, “Design and synthesis of a minimal bacterial genome,” *Science (80-.).*, vol. 351, no. 6280, pp. aad6253–aad6253, Mar. 2016.
[89] D. T. W. Ng and C. A. Sarkar, “NP-Sticky: A web server for optimizing DNA ligation with non-palindromic sticky ends,” *J. Mol. Biol.*, vol. 426, no. 8, pp. 1861–1869, 2014.
[90] A. B. Nager and R. D. Atkinson, “A Trillion-Dollar Opportunity: How Brain Research Can Drive Health and Prosperity,” *Inf. Technol. Innov. Found.*, no. July, pp. 1–31, 2016.
[91] A. Di Carlo, “Human and economic burden of stroke,” *Age Ageing*, vol. 38, no. 1, pp. 4–5, 2009.
[92] T. Kume *et al.*, “Isolation of a diterpenoid substance with potent neuroprotective activity from fetal calf serum.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 99, no. 5, pp. 3288–3293, 2002.
[93] T. Ioroi, K. Taguchi, Y. Izumi, Y. Takada-Takatori, A. Akaike, and T. Kume, “Protective effect of serofendic acid, administered intravenously, on cerebral ischemia-reperfusion injury in rats,” *Brain Res.*, vol. 1532, pp. 99–105, 2013.
[94] T. Nakamura, T. Kume, H. Katsuki, T. Niidome, H. Sugimoto, and A. Akaike, “Protective effect of serofendic acid on ischemic injury induced by occlusion of the middle cerebral artery in rats,” *Eur. J. Pharmacol.*, vol. 586, no. 1–3, pp. 151–155, 2008.
[95] M. Inden *et al.*, “Serofendic acid prevents 6-hydroxydopamine-induced nigral neurodegeneration and drug-induced rotational asymmetry in hemi-parkinsonian rats,” *J. Neurochem.*, vol. 95, no. 4, pp. 950–961, 2005.
[96] O. Kitamura, K. Uemura, H. Kitamura, H. Sugimoto, A. Akaike, and T. Ono, “Serofendic acid protects from iodinated contrast medium and high glucose probably against superoxide production in LLC-PK1 cells,” *Clin. Exp. Nephrol.*, vol. 13, no. 1, pp. 15–24, 2009.
[97] T. Terauchi *et al.*, “Synthesis and pharmacological profile of serofendic acids A and B,” *Bioorganic Med. Chem.*, vol. 15, no. 22, pp. 7098–7107, 2007.
[98] T. Terauchi *et al.*, “Synthesis and neuroprotective effects of serofendic acid analogues,” *Bioorg. Med. Chem. Lett.*, vol. 16, no. 19, pp. 5080–5083, 2006.
[99] T. Terauchi, N. Asai, M. Yonaga, T. Kume, A. Akaike, and H. Sugimoto, “Synthesis and absolute configuration of serofendic acids,” vol. 0, pp. 1807–1809, 2002.
[100] M. Toyota, T. Asano, and M. Ihara, “Total synthesis of serofendic acids A and B

employing tin-free homoallyl-homoallyl radical rearrangement," *Org. Lett.*, vol. 7, no. 18, pp. 3929–3932, 2005.
[101] C. J. Paddon *et al.*, "High-level semi-synthetic production of the potent antimalarial artemisinin.," *Nature*, vol. 496, no. 7446, pp. 528–32, Apr. 2013.
[102] P. K. Ajikumar *et al.*, "Isoprenoid pathway optimization for Taxol precursor overproduction in *Escherichia coli*.," *Science*, vol. 330, pp. 70–74, 2010.
[103] R. J. Peters, "Two rings in them all: the labdane-related diterpenoids.," *Nat. Prod. Rep.*, vol. 27, no. 11, pp. 1521–30, Nov. 2010.
[104] X. Kuang *et al.*, "ent-Atisane diterpenoids from Euphorbia fischeriana inhibit mammosphere formation in MCF-7 cells," *J. Nat. Med.*, vol. 70, no. 1, pp. 120–126, Jan. 2016.
[105] J. Wang *et al.*, "Discovery of platencin, a dual FabF and FabH inhibitor with in vivo antibiotic properties.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 104, no. 18, pp. 7612–7616, 2007.
[106] Q. U. Ain, H. Khan, M. S. Mubarak, and A. Pervaiz, "Plant alkaloids as antiplatelet agent: Drugs of the future in the light of recent developments," *Frontiers in Pharmacology*. 2016.
[107] T. Kawasaki *et al.*, "Presence of copalyl diphosphate synthase gene in an actinomycete possessing the mevalonate pathway," *J. Antibiot. (Tokyo).*, vol. 57, no. 11, pp. 739–747, 2004.
[108] J. D. Rudolf *et al.*, "Structure of the ent-Copalyl Diphosphate Synthase PtmT2 from *Streptomyces platensis* CB00739, a Bacterial Type II Diterpene Synthase," *J. Am. Chem. Soc.*, vol. 138, no. 34, pp. 10905–10915, 2016.
[109] R. S. Nett *et al.*, "Elucidation of gibberellin biosynthesis in symbiotic rhizobia reveals convergent evolution," *Nat. Chem. Biol.*, vol. 13, no. November, 2016.
[110] L.-B. Dong *et al.*, "Cryptic and Stereospecific Hydroxylation, Oxidation, and Reduction in Platensimycin and Platencin Biosynthesis," *J. Am. Chem. Soc.*, vol. 141, no. 9, pp. 4043–4050, Mar. 2019.
[111] T. Siegl, B. Tokovenko, M. Myronovskyi, and A. Luzhetskyy, "Design, construction and characterisation of a synthetic promoter library for fine-tuned gene expression in actinomycetes," *Metab. Eng.*, vol. 19, pp. 98–106, 2013.
[112] H. M. Salis, E. A. Mirsky, and C. A. Voigt, "Automated design of synthetic ribosome binding sites to control protein expression," *Nat. Biotechnol.*, vol. 27, no. 10, pp. 946–950, Oct. 2009.
[113] D. P. Labeda, J. R. Doroghazi, K. S. Ju, and W. W. Metcalf, "Taxonomic evaluation of *Streptomyces albus* and related species using multilocus sequence analysis and proposals to emend the description of *Streptomyces albus* and describe *Streptomyces pathocidini* sp. nov.," *Int. J. Syst. Evol. Microbiol.*, vol. 64, no. PART 3, pp. 894–900, 2014.
[114] E. C. Cherney, J. M. Lopchuk, J. C. Green, and P. S. Baran, "A Unified Approach to ent-Atisane Diterpenes and Related Alkaloids: Synthesis of (−)-Methyl Atisenoate, (−)-Isoatisine, and the Hetidine Skeleton," *J. Am. Chem. Soc.*, vol. 136, pp. 12592–12595, 2014.
[115] F. A. Davis and O. D. Stringer, "Chemistry of Oxaziridines. 2. Improved Synthesis of 2-Sulfonyloxaziridines," *J. Org. Chem.*, vol. 47, no. 9, pp. 1774–1775, 1982.
[116] M. M. Zhang *et al.*, "CRISPR-Cas9 strategy for activation of silent *Streptomyces* biosynthetic gene clusters," *Nat. Chem. Biol.*, vol. 13, no. 6, pp. 607–609, 2017.

[117] M. H. Medema, R. Breitling, and E. Takano, “Synthetic Biology in *Streptomyces* Bacteria,” in *Methods in enzymology*, 1st ed., vol. 497, Elsevier Inc., 2011, pp. 485–502.
[118] Y. Luo, L. Zhang, K. W. Barton, and H. Zhao, “Systematic identification of a panel of strong constitutive promoters from *Streptomyces albus*,” *ACS Synth. Biol.*, vol. 4, no. 9, pp. 1001–1010, 2015.
[119] N. Zaburannyi, M. Rabyk, B. Ostash, V. Fedorenko, and A. Luzhetskyy, “Insights into naturally minimised *Streptomyces albus* J1074 genome.,” *BMC Genomics*, vol. 15, no. 1, p. 97, Jan. 2014.
[120] Y. Ahmed, Y. Rebets, B. Tokovenko, E. Brötz, and A. Luzhetskyy, “Identification of butenolide regulatory system controlling secondary metabolism in *Streptomyces albus* J1074,” *Sci. Rep.*, vol. 7, no. 1, 2017.
[121] S. Zhao *et al.*, “Improvement of catechin production in *Escherichia coli* through combinatorial metabolic engineering,” *Metab. Eng.*, vol. 28, pp. 43–53, 2015.
[122] M. J. Smanski *et al.*, “Dedicated ent-kaurene and ent-atiserene synthases for platensimycin and platencin biosynthesis,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 108, no. 33, 2011.
[123] X. N. Li *et al.*, “Structure and cytotoxicity of diterpenoids from Isodon eriocalyx,” *J. Nat.*
Prod., vol. 73, no. 11, pp. 1803–1809, 2010.
[124] N. B. Perry, E. J. Burgess, S. H. Baek, and R. T. Weavers, “The first atisane diterpenoids from a liverwort: Polyols from Lepidolaena clavigera,” *Org. Lett.*, vol. 3, no. 26, pp. 4243–4245, 2001.
[125] W. G. Stillwell, “Methylthiolation: a new pathway of drug metabolism,” *Trends Pharmacol. Sci.*, vol. 2, no. C, pp. 250–252, 1981.
[126] T. Kieser, M. J. Bibb, M. J. Buttner, K. F. Chater, and D. A. Hopwood, “Practical *Streptomyces* Genetics,” *John Innes Cent. Ltd.*, p. 529, 2000.
[127] W. Minas, J. E. Bailey, and W. Duetz, “Streptomycetes in micro-cultures: Growth, production of secondary metabolites, and storage and retrieval in the 96-well format,” *Antonie van Leeuwenhoek, Int. J. Gen. Mol. Microbiol.*, vol. 78, no. 3–4, pp. 297–305, 2000.
[128] X. Luo *et al.*, “Complete biosynthesis of cannabinoids and their unnatural analogues in yeast,” *Nature*, vol. 567, no. 7746, pp. 123–126, Mar. 2019.
[129] H. Pyles, S. Zhang, J. J. De Yoreo, and D. Baker, “Controlling protein assembly on inorganic crystals through designed protein interfaces,” *Nature*, vol. 571, no. 7764, pp. 251–256, Jul. 2019.
[130] C. Gilbert and T. Ellis, “Biological Engineered Living Materials: Growing Functional Materials with Genetically Programmable Properties,” vol. 15, p. 54, 2019.
[131] G. Zhang, T. Johnston, M. B. Quin, and C. Schmidt-Dannert, “Developing a Protein Scaffolding System for Rapid Enzyme Immobilization and Optimization of Enzyme Functions for Biocatalysis,” *ACS Synth. Biol.*, vol. 8, no. 8, pp. 1867–1876, Aug. 2019.
[132] A. A. K. Nielsen *et al.*, “Genetic circuit design automation,” *Science (80-.).*, vol. 352, no. 6281, 2016.
[133] A. Khakhar, A. R. Leydon, A. C. Lemmex, E. Klavins, and J. L. Nemhauser, “Synthetic hormone-responsive transcription factors can monitor and re-program

plant development," *Elife*, vol. 7, May 2018.
[134] L. A. Mitchell *et al.*, "Synthesis, debugging, and effects of synthetic chromosome consolidation: synVI and beyond.," *Science*, vol. 355, no. 6329, p. eaaf4831, Mar. 2017.
[135] S. M. Richardson *et al.*, "Design of a synthetic yeast genome," *Science (80-.).*, vol. 355, no. 6329, pp. 1040–1044, Mar. 2017.
[136] Y. Wu *et al.*, "Bug mapping and fitness testing of chemically synthesized chromosome X.," *Science*, vol. 355, no. 6329, p. eaaf4706, Mar. 2017.
[137] R. S. Cox, M. G. Surette, M. B. Elowitz, and M. B. Elowitz, "Programming gene expression with combinatorial promoters.," *Mol. Syst. Biol.*, vol. 3, p. 145, 2007.
[138] C. Lou, B. Stanton, Y. J. Chen, B. Munsky, and C. A. Voigt, "Ribozyme-based insulator parts buffer synthetic circuits from genetic context," *Nat. Biotechnol.*, vol. 30, no. 11, pp. 1137–1142, Nov. 2012.
[139] B. F. Pfleger, D. J. Pitera, C. D. Smolke, and J. D. Keasling, "Combinatorial engineering of intergenic regions in operons tunes expression of multiple genes," *Nat. Biotechnol.*, vol. 24, no. 8, pp. 1027–1032, Aug. 2006.
[140] G. Kudla, A. W. Murray, D. Tollervey, and J. B. Plotkin, "Coding-sequence determinants of expression in *Escherichia coli*," *Science*, vol. 324, no. 5924, pp. 255–258, Apr. 2009.
[141] Z. Zhou *et al.*, "Codon usage is an important determinant of gene expression levels largely through its effects on transcription.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 113, no. 41, pp. E6117–E6125, Oct. 2016.
[142] J. N. Wells, L. T. Bergendahl, J. A. M. Correspondence, and J. A. Marsh, "Operon Gene Order Is Optimized for Ordered Protein Complex Assembly," *CellReports*, vol. 14, pp. 679–685, 2016.
[143] Y.-Q. Feng, M. C. Lorincz, S. Fiering, J. M. Greally, and E. E. Bouhassira, "Position Effects Are Influenced by the Orientation of a Transgene with Respect to Flanking Chromatin," *Mol. Cell. Biol.*, vol. 21, no. 1, pp. 298–309, 2001.
[144] G.-Z. Wang, M. J. Lercher, and L. D. Hurst, "Transcriptional Coupling of Neighboring Genes and Gene Expression Noise: Evidence that Gene Orientation and Noncoding Transcripts Are Modulators of Noise," *Genome Biol. Evol.*, vol. 3, pp. 320–331, Jan. 2011.
[145] C. Sauer *et al.*, "Effect of Genome Position on Heterologous Gene Expression in *Bacillus subtilis*: An Unbiased Analysis," *ACS Synth. Biol.*, vol. 5, no. 9, pp. 942–947, 2016.
[146] S. Galanie, K. Thodey, I. J. Trenchard, M. Filsinger Interrante, and C. D. Smolke, "Complete biosynthesis of opioids in yeast.," *Science*, vol. 349, no. 6252, pp. 1095–100, Sep. 2015.
[147] J. Beal, "Bridging the Gap: A Roadmap to Breaking the Biological Design Barrier," *Front. Bioeng. Biotechnol.*, vol. 2, no. January, pp. 1–16, 2015.
[148] A. Zhao *et al.*, "Use of real-time cellular analysis and Plackett-Burman design to develop the serum-free media for PC-3 prostate cancer cells," *PLoS One*, vol. 12, no. 9, p. e0185470, Sep. 2017.
[149] C. Singleton, J. Gilman, J. Rollit, K. Zhang, D. A. Parker, and J. Love, "A design of experiments approach for the rapid formulation of a chemically defined medium for metabolic profiling of industrially important microbes," *PLoS One*, vol. 14, no. 6, p. e0218208, Jun. 2019.

[150] A. Uhoraningoga *et al.*, “The Goldilocks Approach: A Review of Employing Design of Experiments in Prokaryotic Recombinant Protein Production,” *Bioengineering*, vol. 5, no. 4, p. 89, Oct. 2018.
[151] C.-F. Mandenius and A. Brundin, “Bioprocess optimization using design-of-experiments methodology,” *Biotechnol. Prog.*, vol. 24, no. 6, pp. 1191–1203, Nov. 2008.
[152] H. Zhou, B. Vonk, J. A. Roubos, R. A. L. Bovenberg, and C. A. Voigt, “Algorithmic co-optimization of genetic constructs and growth conditions: application to 6-ACA, a potential nylon-6 precursor,” *Nucleic Acids Res.*, vol. 43, no. 21, p. gkv1071, Oct. 2015.
[153] E. M. Young *et al.*, “Iterative algorithm-guided design of massive strain libraries, applied to itaconic acid production in yeast,” *Metab. Eng.*, vol. 48, no. May, pp. 33–43, 2018.
[154] P. Xu, E. A. Rizzoni, S. Y. Sul, and G. Stephanopoulos, “Improving metabolic pathway efficiency by statistical model-based multivariate regulatory metabolic engineering,” *ACS Synth. Biol.*, vol. 6, no. 1, pp. 148–158, 2017.
[155] M. E. Lee, A. Aswani, A. S. Han, C. J. Tomlin, and J. E. Dueber, “Expression-level optimization of a multi-enzyme pathway in the absence of a high-throughput assay,” *Nucleic Acids Res.*, vol. 41, no. 22, pp. 10668–10678, 2013.
[156] R. L. Plackett and J. P. Burman, “The Design of Optimum Multifactorial Experiments,” *Biometrika*, vol. 33, no. 4, pp. 305–325, 1946.
[157] A. P. Karlapudi, S. Krupanidhi, R. R. E., I. M., N. B. Md., and V. T.C., “Plackett-Burman design for screening of process components and their effects on production of lactase by newly isolated Bacillus sp. VUVD101 strain from Dairy effluent,” *Beni-Suef Univ. J. Basic Appl. Sci.*, vol. 7, no. 4, pp. 543–546, 2018.
[158] A. K. Das and S. Dewanjee, “Optimization of Extraction Using Mathematical Models and Computation,” in *Computational Phytochemistry*, S. D. Sarker and L. Nahar, Eds. 2018, pp. 75–106.
[159] H. Ebrahimi-Najafabadi, R. Leardi, and M. Jalali-Heravi, “Experimental design in analytical chemistry -Part I: Theory,” *Journal of AOAC International*, vol. 97, no. 1. Oxford Academic, pp. 3–11, 01-Jan-2014.
[160] A. M. C. T. Briefs, “Experimental design and optimisation (4): Plackett-Burman designs,” *Anal. Methods*, vol. 5, no. 8, pp. 1901–1903, 2013.
[161] Y. Vander Heyden, C. Hartmann, D. L. Massart, L. Michel, P. Kiechle, and F. Erni, “Ruggedness tests for a high-performance liquid Chromatographic assay: Comparison of an evaluation at two and three levels by using two-level Plackett-Burman designs,” *Anal. Chim. Acta*, vol. 316, no. 1, pp. 15–26, 1995.
[162] O. L. Rakotonandrasana *et al.*, “Cytotoxic 3,4- seco-atisane diterpenoids from Croton barorum and Croton goudotii,” *J. Nat. Prod.*, vol. 73, no. 10, pp. 1730–1733, 2010.
[163] R. Jaeger and E. Cuny, “Terpenoids with special pharmacological significance: A review,” *Nat. Prod. Commun.*, vol. 11, no. 9, pp. 1373–1390, 2016.
[164] J. Kirby *et al.*, “Enhancing Terpene Yield from Sugars via Novel Routes to 1-Deoxy-d-Xylulose 5-Phosphate,” *Appl. Environ. Microbiol.*, vol. 81, no. 1, pp. 130–138, Jan. 2015.
[165] R. Mewalal *et al.*, “Plant-Derived Terpenes: A Feedstock for Specialty Biofuels,” *Trends Biotechnol.*, vol. 35, no. 3, pp. 227–240, 2017.

[166] G. Daletos, C. Katsimpouras, and G. Stephanopoulos, “Novel Strategies and Platforms for Industrial Isoprenoid Engineering,” *Trends Biotechnol.*, pp. 1–12, 2020.
[167] H. H. Wang *et al.*, “Programming cells by multiplex genome engineering and accelerated evolution,” *Nature*, vol. 460, no. 7257, pp. 894–898, Aug. 2009.
[168] N. Zhang *et al.*, “Improved production of the tallysomycin H-1 in Streptoalloteichus hindustanus SB8005 strain by fermentation optimization,” *Appl. Microbiol. Biotechnol.*, vol. 86, no. 5, pp. 1345–1353, May 2010.
[169] R. Mani, R. P. St. Onge, J. L. Hartman IV, G. Giaever, and F. P. Roth, “Defining genetic interaction,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 105, no. 9, pp. 3461–3466, 2008.
[170] B. Boucher and S. Jenna, “Genetic interaction networks: Better understand to better predict,” *Front. Genet.*, vol. 4, no. DEC, pp. 1–16, 2013.
[171] S. Jana and W. Seyffert, “Simulation of quantitative characters by genes with biochemically definable action,” *Theor. Appl. Genet.*, vol. 42, no. 1, pp. 16–24, 1972.
[172] C. Dillmaim and J. L. Foulley, “Another look at multiplicative models in quantitative genetics,” *Genet. Sel. Evol.*, vol. 30, no. 6, pp. 543–564, Nov. 1998.
[173] D. C. Volke, J. Rohwer, R. Fischer, and S. Jennewein, “Investigation of the methylerythritol 4-phosphate pathway for microbial terpenoid production through metabolic control analysis,” *Microb. Cell Fact.*, vol. 18, no. 1, p. 192, Nov. 2019.
[174] L. P. Wright *et al.*, “Deoxyxylulose 5-phosphate synthase controls flux through the methylerythritol 4-phosphate pathway in arabidopsis,” *Plant Physiol.*, vol. 165, no. 4, pp. 1488–1504, 2014.
[175] H. Alper, C. Fischer, E. Nevoigt, and G. Stephanopoulos, “Tuning genetic control through promoter engineering,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 102, no. 36, pp. 12678–12683, 2005.
[176] K. L. Jones, S. W. Kim, and J. D. Keasling, “Low-copy plasmids can perform as well as or better than high-copy plasmids for metabolic engineering of bacteria,” *Metab. Eng.*, vol. 2, no. 4, pp. 328–338, 2000.
[177] A. Rodríguez-Villalón, J. Pérez-Gil, and M. Rodríguez-Concepción, “Carotenoid accumulation in bacteria with enhanced supply of isoprenoid precursors by upregulation of exogenous or endogenous pathways,” *J. Biotechnol.*, vol. 135, no. 1, pp. 78–84, May 2008.
[178] J. Stelling, U. Sauer, Z. Szallasi, F. J. Doyle, and J. Doyle, “Robustness of cellular functions.,” *Cell*, vol. 118, no. 6, pp. 675–85, Sep. 2004.
[179] H. Alper, K. Miyaoku, and G. Stephanopoulos, “Construction of lycopene-overproducing *E. coli* strains by combining systematic and combinatorial gene knockout targets,” *Nat. Biotechnol.*, vol. 23, no. 5, pp. 612–616, Apr. 2005.
[180] J. M. Raser and E. O’Shea, “Noise in Gene Expression : Origins, Consequences, and Control,” vol. 309, no. September 2005, pp. 2010–2014, 2005.
[181] C. Bai *et al.*, “Exploiting a precise design of universal synthetic modular regulatory elements to unlock the microbial natural products in *Streptomyces*,” *Proc. Natl. Acad. Sci.*, vol. 112, no. 39, pp. 12181–12186, 2015.
[182] A. Banerjee and T. D. Sharkey, “Methylerythritol 4-phosphate (MEP) pathway metabolic regulation,” *Nat. Prod. Rep.*, vol. 31, no. 8, pp. 1043–1055, Jul. 2014.
[183] A. Banerjee, Y. Wu, R. Banerjee, Y. Li, H. Yan, and T. D. Sharkey, “Feedback

inhibition of deoxy-D-xylulose-5-phosphate synthase regulates the methylerythritol 4-phosphate pathway," *J. Biol. Chem.*, vol. 288, no. 23, pp. 16926–16936, 2013.
[184] Z. Li and T. D. Sharkey, "Metabolic profiling of the methylerythritol phosphate pathway reveals the source of post-illumination isoprene burst from leaves," *Plant, Cell Environ.*, vol. 36, no. 2, pp. 429–437, 2013.
[185] B. Jones and C. J. Nachtsheim, "A class of three-level designs for definitive screening in the presence of second-order effects," *J. Qual. Technol.*, vol. 43, no. 1, pp. 1–15, 2011.
[186] J. C. Wang and C. F. J. Wu, "A HIDDEN PROJECTION PROPERTY OF PLACKETT-BURMAN AND RELATED DESIGNS," *Stat. Sin.*, vol. 5, pp. 235–250, 1995.
[187] J. F. Magallanes and A. C. Olivieri, "The effect of factor interactions in Plackett-Burman experimental designs. Comparison of Bayesian-Gibbs analysis and genetic algorithms," *Chemom. Intell. Lab. Syst.*, vol. 102, no. 1, pp. 8–14, 2010.
[188] M. Jeschek, D. Gerngross, and S. Panke, "Rationally reduced libraries for combinatorial pathway optimization minimizing experimental effort," *Nat. Commun.*, vol. 7, no. 1, p. 11163, Sep. 2016.
[189] N. Crook, J. Abatemarco, J. Sun, J. M. Wagner, A. Schmitz, and H. S. Alper, "In vivo continuous evolution of genes and pathways in yeast," *Nat. Commun.*, vol. 7, no. 1, p. 13051, Dec. 2016.
[190] R. S. Obach, G. S. Walker, R. Sharma, S. Jenkinson, T. P. Tran, and A. F. Stepan, "Lead Diversification at the Nanomole Scale Using Liver Microsomes and Quantitative Nuclear Magnetic Resonance Spectroscopy: Application to Phosphodiesterase 2 Inhibitors," *J. Med. Chem.*, vol. 61, no. 8, pp. 3626–3640, Apr. 2018.
[191] M. D. Leavell, D. J. McPhee, and C. J. Paddon, "Developing fermentative terpenoid production for commercial usage," *Current Opinion in Biotechnology*, vol. 37. Elsevier Ltd, pp. 114–119, 01-Feb-2016.
[192] L. Caputi and E. Aprea, "Use of Terpenoids as Natural Flavouring Compounds in Food Industry," *Recent Patents Food, Nutr. Agric.*, vol. 3, no. 1, pp. 9–16, 2012.
[193] C. E. Vickers and S. Sabri, *Biotechnology of Isoprene*, vol. 148. 2015.
[194] L. C. Mata-Gómez, J. C. Montañez, A. Méndez-Zavala, and C. N. Aguilar, "Biotechnological production of carotenoids by yeasts: An overview," *Microbial Cell Factories*, vol. 13, no. 1. BioMed Central, p. 12, 21-Jan-2014.
[195] Y. Yamada *et al.*, "Terpene synthases are widely distributed in bacteria.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 112, no. 3, pp. 857–62, 2015.
[196] Y. Yamada *et al.*, "Novel terpenes generated by heterologous expression of bacterial terpene synthase genes in an engineered *Streptomyces* host," *J. Antibiot. (Tokyo).*, vol. 68, no. 6, pp. 385–394, Jun. 2015.
[197] D. Zheng *et al.*, "Cytotoxic Fusicoccane-Type Diterpenoids from *Streptomyces violascens* Isolated from *Ailuropoda melanoleuca* Feces," *J. Nat. Prod.*, vol. 80, no. 4, pp. 837–844, Apr. 2017.
[198] Z.-J. Xiong *et al.*, "Isolation and biosynthesis of labdanmycins: four new labdane diterpenes from endophytic *Streptomyces*," *Org. Chem. Front.*, vol. 5, no. 8, pp. 1272–1279, Apr. 2018.
[199] L. Bin Dong, J. D. Rudolf, M. R. Deng, X. Yan, and B. Shen, "Discovery of the Tiancilactone Antibiotics by Genome Mining of Atypical Bacterial Type II

Diterpene Synthases," *ChemBioChem*, vol. 19, no. 16, pp. 1727–1733, 2018.
[200] D. I. Walsh *et al.*, "Standardizing Automated DNA Assembly: Best Practices, Metrics, and Protocols Using Robots," *SLAS Technol.*, vol. 24, no. 3, pp. 282–290, Jun. 2019.
[201] Y. Miura, "The biological significance of ω-oxidation of fatty acids.," *Proc. Jpn. Acad. Ser. B. Phys. Biol. Sci.*, vol. 89, no. 8, pp. 370–82, 2013.
[202] W. B. Rizzo, "Fatty aldehyde and fatty alcohol metabolism: review and importance for epidermal structure and function.," *Biochim. Biophys. Acta*, vol. 1841, no. 3, pp. 377–89, Mar. 2014.
[203] A. E. DeBarber, L. A. Bleyle, J.-B. O. Roullet, and D. R. Koop, "ω-Hydroxylation of farnesol by mammalian cytochromes P450," *Biochim. Biophys. Acta - Mol. Cell Biol. Lipids*, vol. 1682, no. 1–3, pp. 18–27, Jun. 2004.

Appendix 1

Supplementary Materials for

Chapter 3 Semi-synthesis of the Neuroprotective Metabolite, Serofendic Acid

This appendix includes:

Figures. A1.1 to A1.65
Tables A.1 to A.2
Captions for Data file A1

Other Supplementary Materials for this chapter include the following:

Data File A1

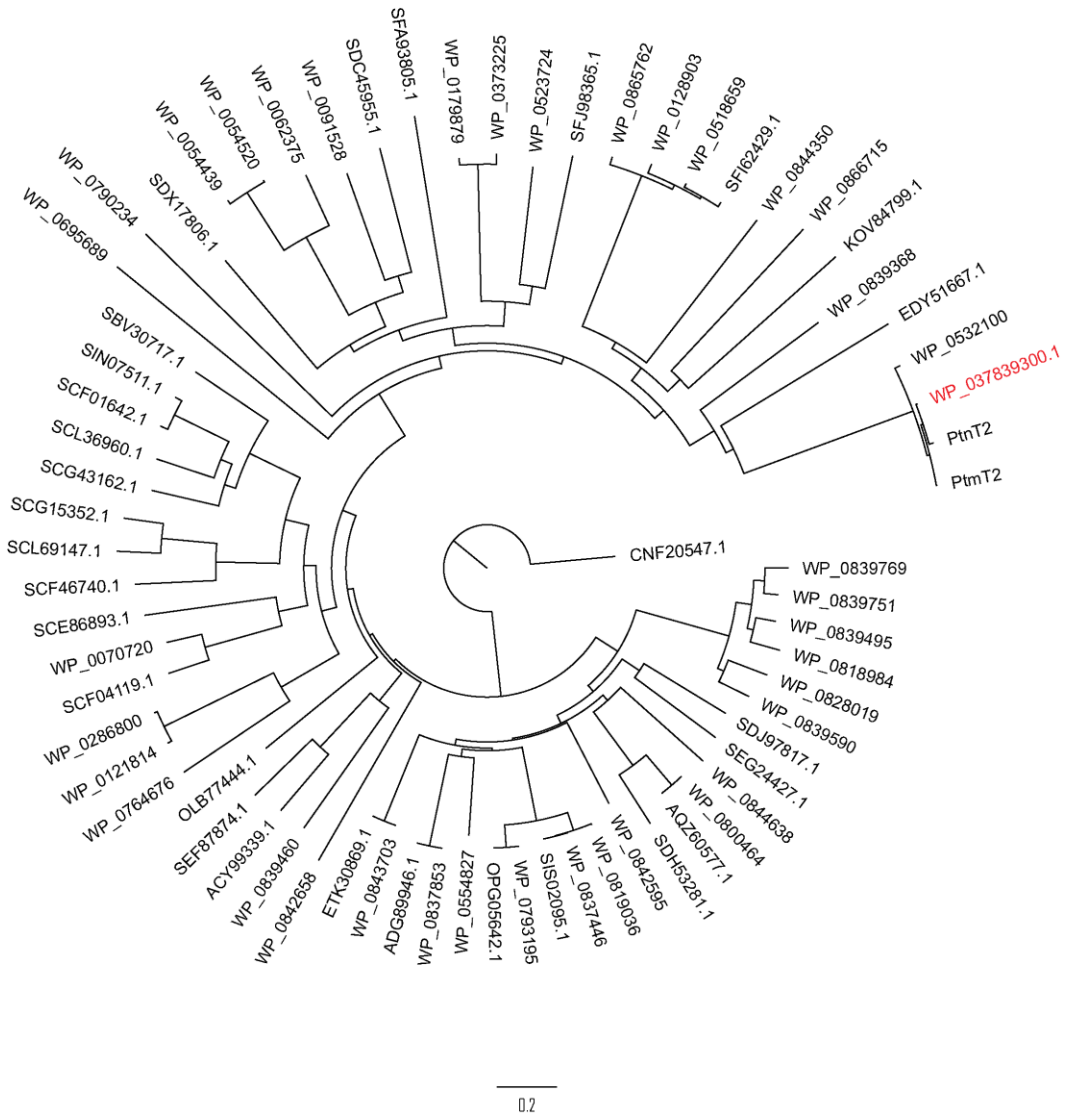


Figure A1. 1. Phylogenetic analysis of 68 primary amino acid sequences similar to *ent*-copalyl diphosphate synthase PtmT2 (accession ACO31276.1) from *Streptomyces platensis*.

WP_037839300.1 (highlighted in red) is a hypothetical protein and later confirmed as *ent*-copalyl diphosphate synthase from *Streptomyces sp.* NRRL S-1813 in this study.

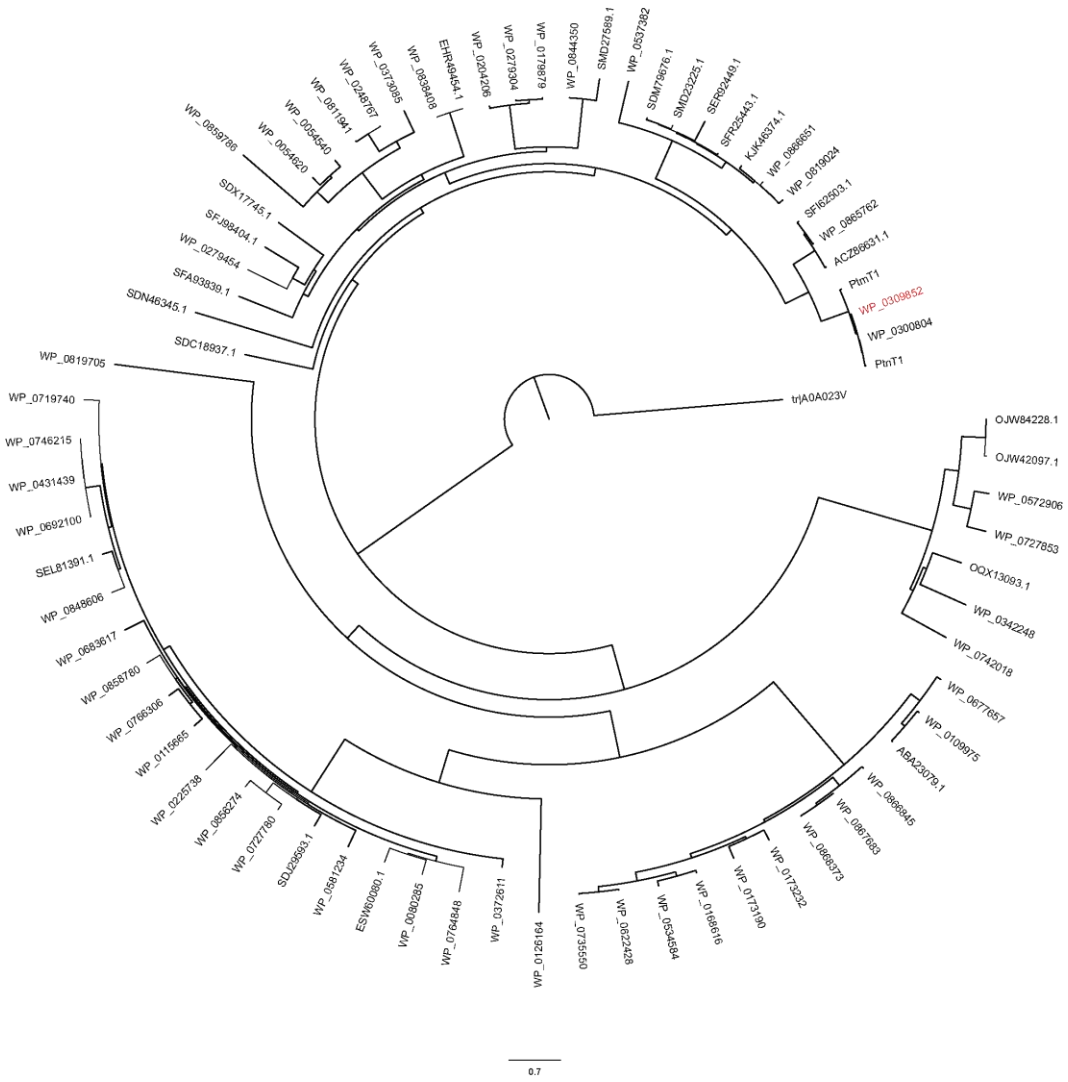


Figure A1.2. Phylogenetic analysis of 91 primary amino acid sequences similar to *ent*-atiserene synthase PtmT1 (accession ACO31274.1) from *Streptomyces platensis*.

WP_0309852 (highlighted in red) is a hypothetical protein, and later confirmed as an *ent*-atiserene synthase from *Streptomyces sp.* NRRL S-1813 in this study.

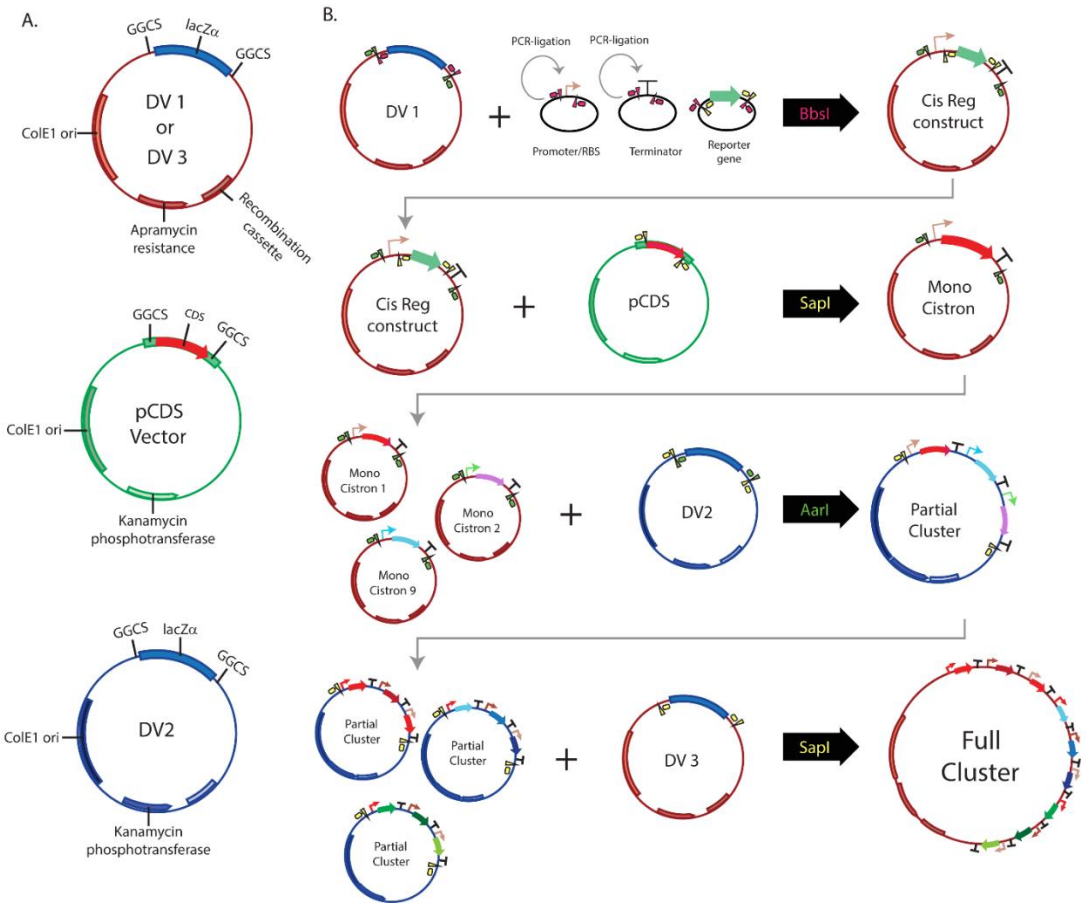


Figure A1.3. Plasmid maps and overall scheme for DNA assembly pipeline.

(A) Plasmid maps for backbone vectors used in the assembly pipeline. Ori: origin of replication; GGCS: golden gate cloning site. (B) Plasmids and enzymes used during DNA assembly. Top row shows construction of CisReg (cis-regulatory) plasmids containing a reporter gene CDS for part characterization. Second row shows replacement of reporter CDS with CDS from biosynthetic gene. Third row shows assembly of three monocistronic expression cassettes to a partial cluster. Bottom row shows assembly of multiple partial clusters to a full cluster plasmid. Black arrows are labeled with Type IIS restriction endonuclease used at each step.

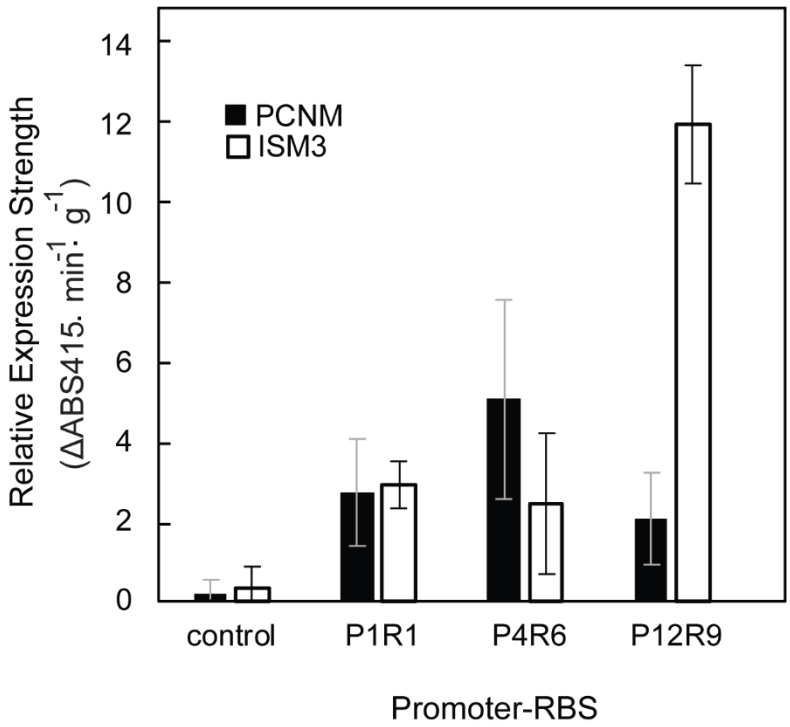
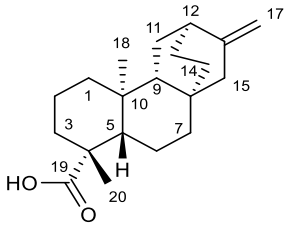


Figure A1.4. Characterization of promoter-RBS strength in 96-hour fermentation in 2-mL PCNM (black bar) and ISM3 (white bar) media in 24-well MTP.

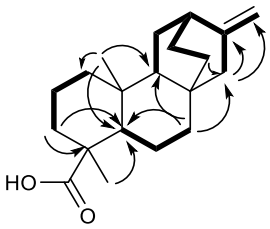
Bars represent average of three biological replicates over multiple days; error bars denote standard deviation.



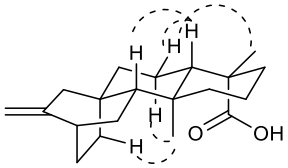
***ent*-atis-16-en-19-oic acid (2)**

Annotated NMR data for 2 in chloroform-*d*.

No.	δ_H (#H, mult., *J*(Hz))	δ_C
1	0.85 (1H, td, 3.3, 13.2)	39.6
	1.59 (1H, m)	
2	1.40 (1H, m)	18.8
	1.87 (1H, m)	
3	1.01 (1H, m)	38.3
	2.15 (1H, d, 14.0)	
4	-	43.9
5	1.05 (1H, dd, 1.5, 12.1)	57.2
6	1.75 (1H, m)	20.2
	1.82 (1H, m)	
7	1.11 (1H, m)	39.6
	1.47 (1H, td, 2.9, 13.2)	
8	-	33.6
9	1.13 (1H, m)	52.2
10	-	38.5
11	1.40 (1H, m)	28.7
	1.59 (1H, m)	
12	2.22 (1H, m)	36.6
13	1.54 (1H, m)	27.3
	1.60 (1H, m)	
14	1.00 (1H, m)	28.3
	1.97 (1H, m)	
15	1.88 (1H, d, 16.6)	48.3
	2.04 (1H, d, 16.6)	
16	-	152.6
17	4.56 (1H, d, 1.9)	104.7
	4.73 (1H, d, 1.9)	
18	0.90 (3H, s)	12.1
19	-	185.0
20	1.24 (3H, s)	29.0



Key COSY (bold) and HMBC (arrows) correlations



Key NOESY correlations

Figure A1.5. Summary of key 1D and 2D NMR spectroscopy data for compound 2.

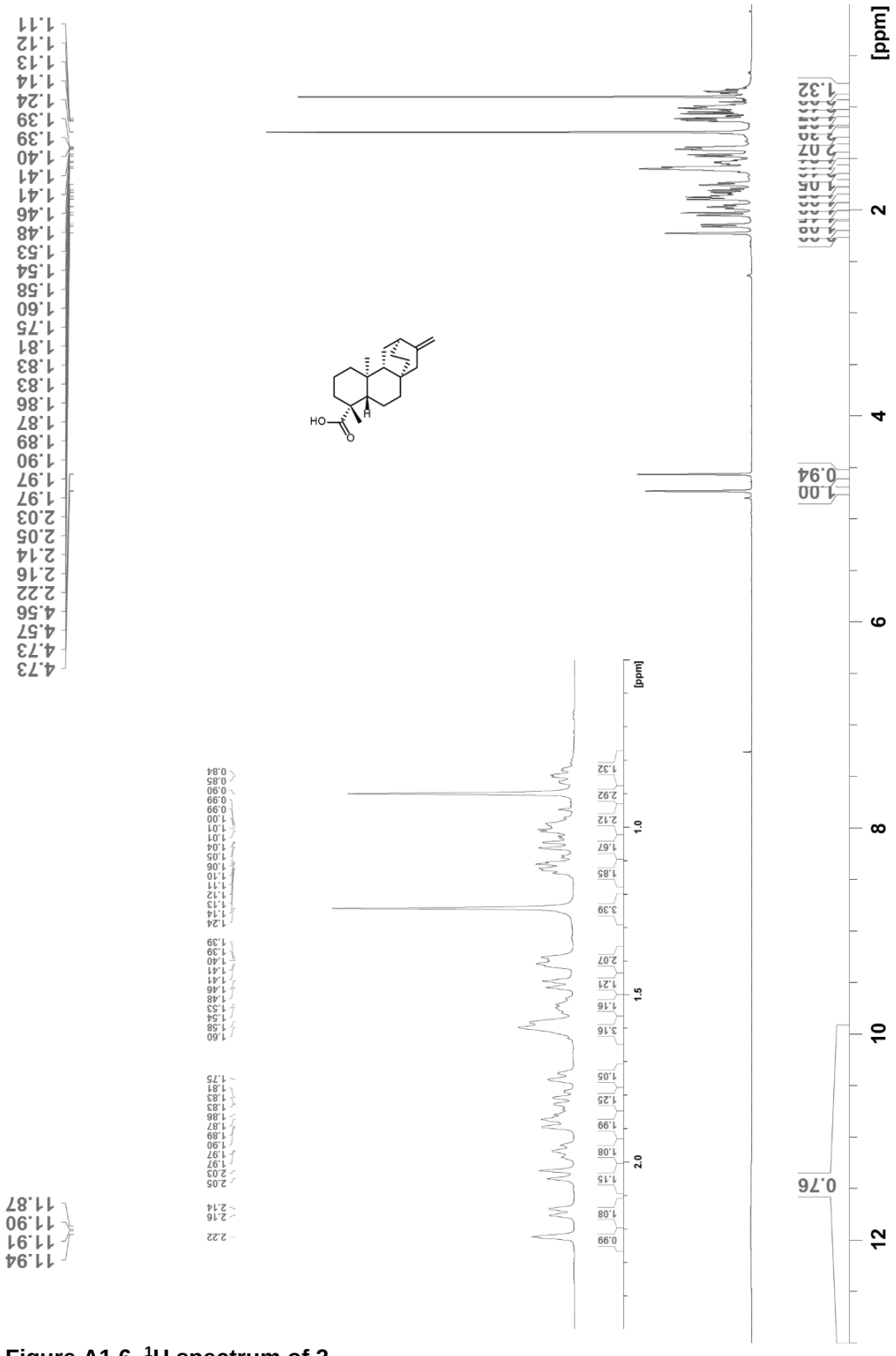


Figure A1.6. ^{1}H spectrum of 2.

Figure A1.7. ^{13}C spectrum of 2.

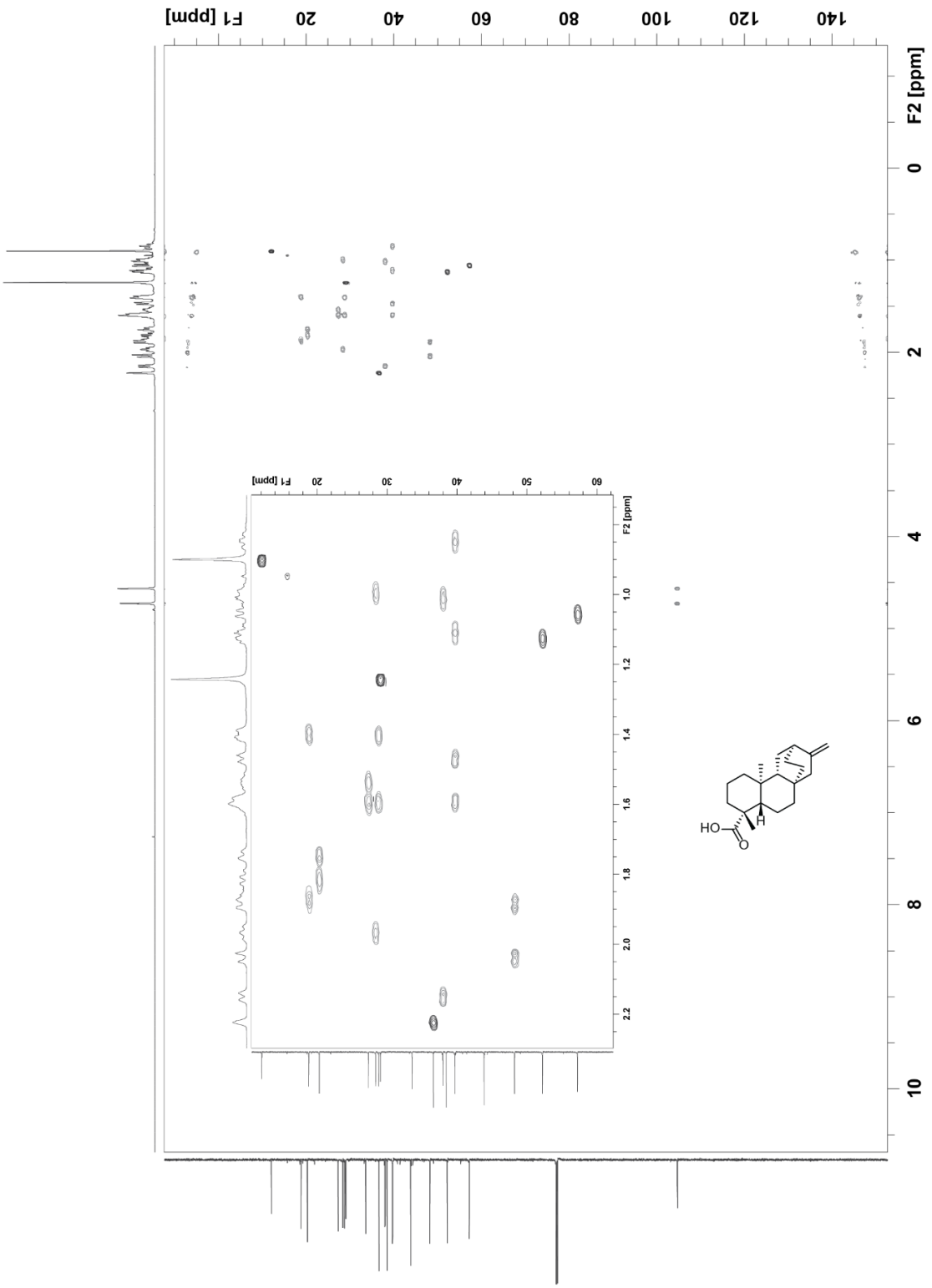
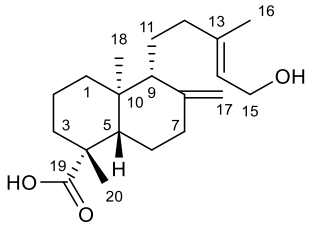


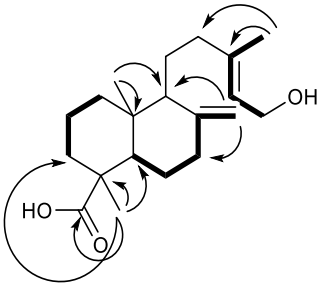
Figure A1. 8. ^{13}C-^{1}H HSQC spectrum of 2.



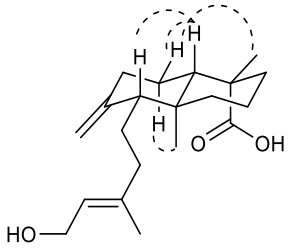
***ent*-isocupressic acid (12)**

Annotated NMR data for 12 in chloroform-*d*.

No.	δ_H (#H, mult., *J*(Hz))	δ_C
1	1.04 (1H, m)	38.5
	2.15 (1H, m)	
2	1.51 (1H, m)	20.0
	1.85 (1H, m)	
3	1.81 (1H, m)	38.1
	2.16 (1H, m)	
4	-	44.3
5	1.32 (1H, dd, 2.8, 11.9)	56.4
6	1.87 (1H, m)	26.4
	1.97 (1H, m)	
7	1.88 (1H, m)	38.8
	2.40 (1H, m)	
8	-	148.1
9	1.56 (1H, d, 11.0)	55.6
10	-	40.6
11	1.43 (1H, m)	22.2
	1.62 (1H, m)	
12	1.05 (1H, m)	39.2
	1.85 (1H, m)	
13	-	140.7
14	5.38 (1H, t, 6.8)	123.1
15	4.14 (2H, m)	59.5
16	1.66 (3H, s)	16.5
17	4.52 (1H, s)	106.6
	4.85 (1H, s)	
18	0.59 (3H, s)	12.9
19	-	183.7
20	1.23 (3H, s)	29.1



Key COSY (bold) and HMBC (arrows) correlations



Key NOESY correlations
ESIMS *m/z* $[M\text{-}H^{+}]^{-}$ 319.23, found 319.0

Figure A1.9. Summary of key 1D and 2D NMR data for compound 12.

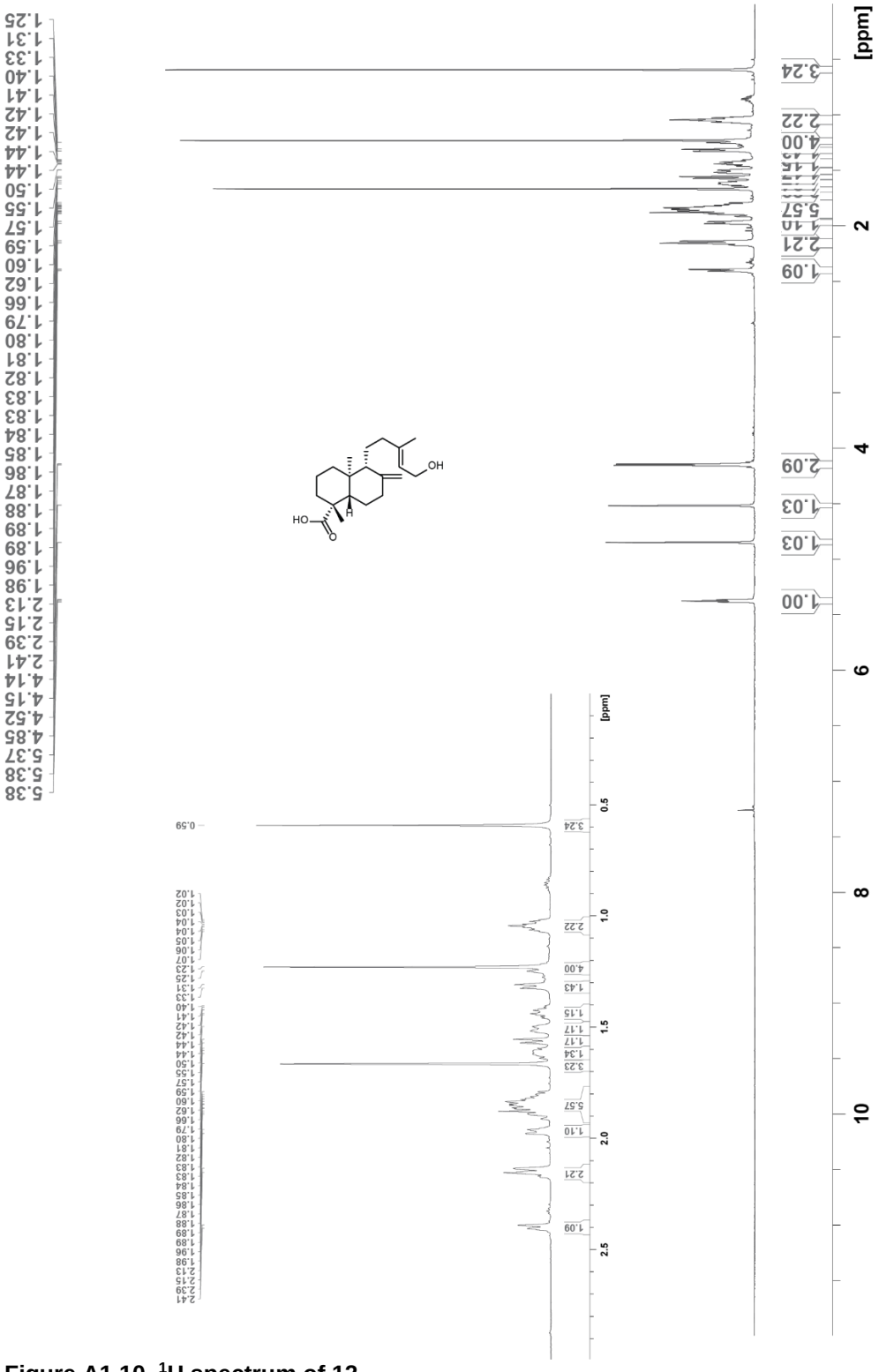


Figure A1.10. ^{1}H spectrum of 12.

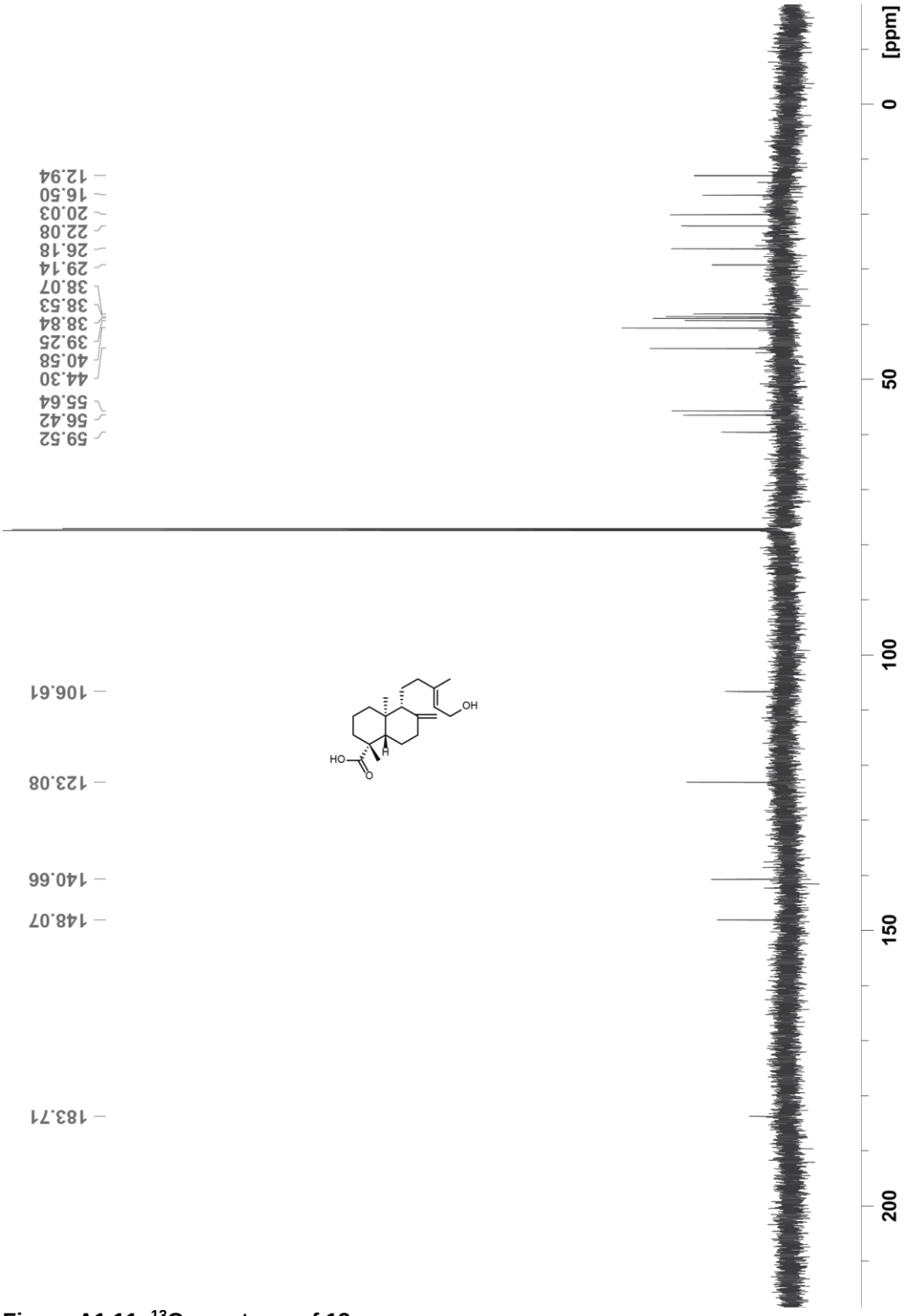
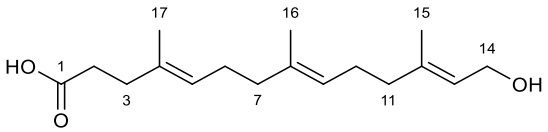


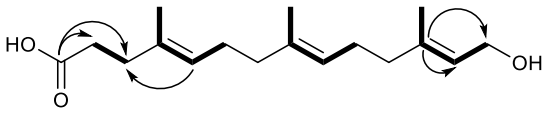
Figure A1.11. ^{13}C spectrum of 12.



(4*E*,8*E*,12*E*)-14-hydroxy-4,8,12-trimethyltetradeca-4,8,12-trienoic acid (11)

Annotated NMR data for 11 in chloroform-*d*.

No.	δ_H (#H, mult., *J*(Hz))	δ_C
1	-	177.1
2	2.44 (2H, t, 7.3)	32.7
3	2.30 (2H, t, 7.3)	34.5
4	-	132.9
5	5.15 (1H, bt, 7.0)	125.0
6	2.08 (2H, m)	26.5
7	1.98 (2H, m)	39.5
8	-	134.8
9	5.10 (1H, bt, 7.0)	124.2
10	2.12 (2H, m)	26.2
11	2.04 (2H, m)	39.6
12	-	139.9
13	5.42 (1H, bt, 7.1)	123.5
14	4.17 (2H, d, 7.0)	59.6
15	1.68 (3H, s)	16.5
16	1.59 (3H, s)	16.1
17	1.61 (3H, s)	16.1



Key COSY (bold) and HMBC (arrows) correlations

ESIMS *m/z* $[M\text{-}H^+]^-$ 279.20, found 279.00

Figure A1.12. Summary of key 1D and 2D NMR data for compound 11.

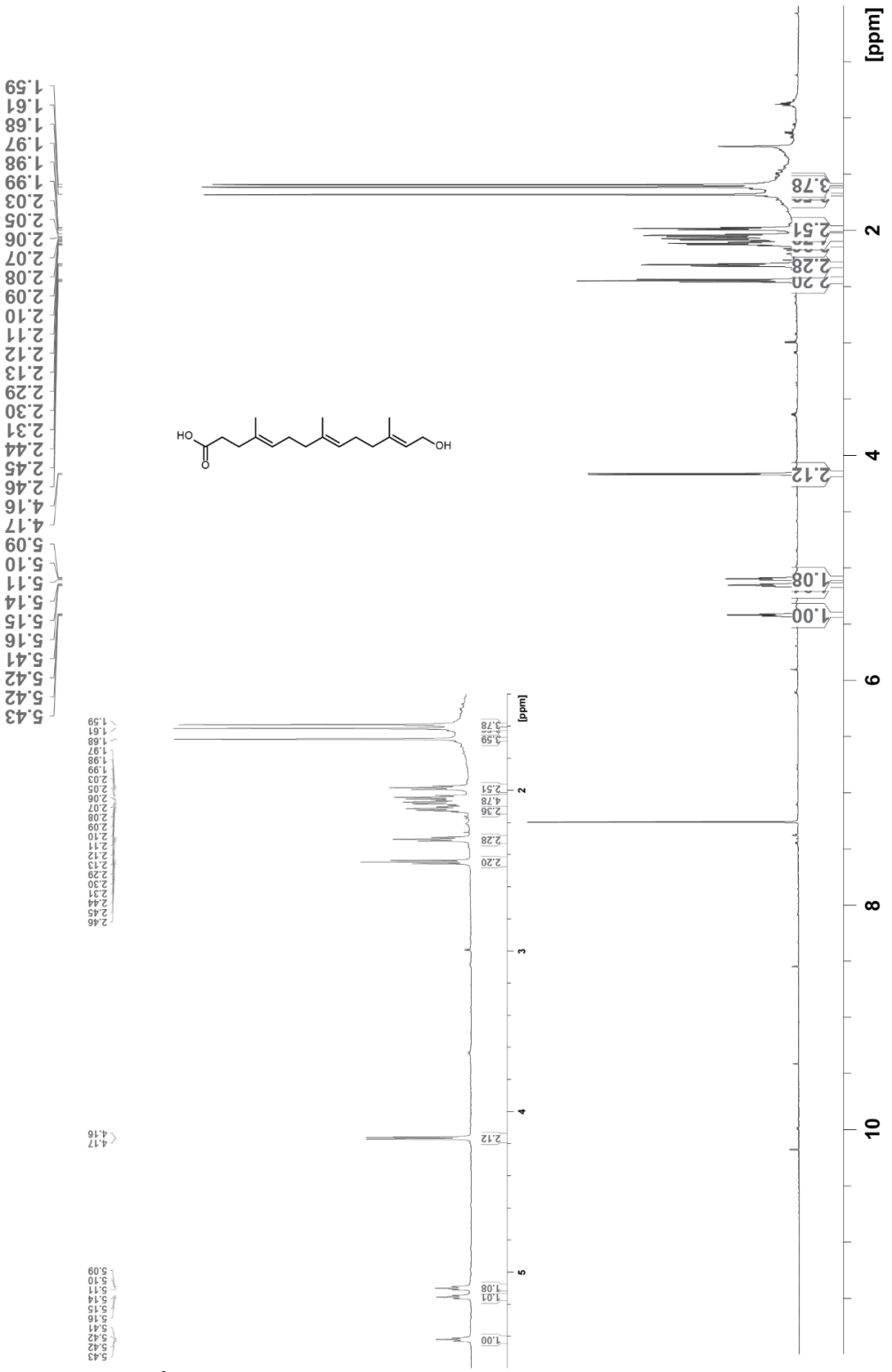


Figure A1.13. ^{1}H spectrum of 11.

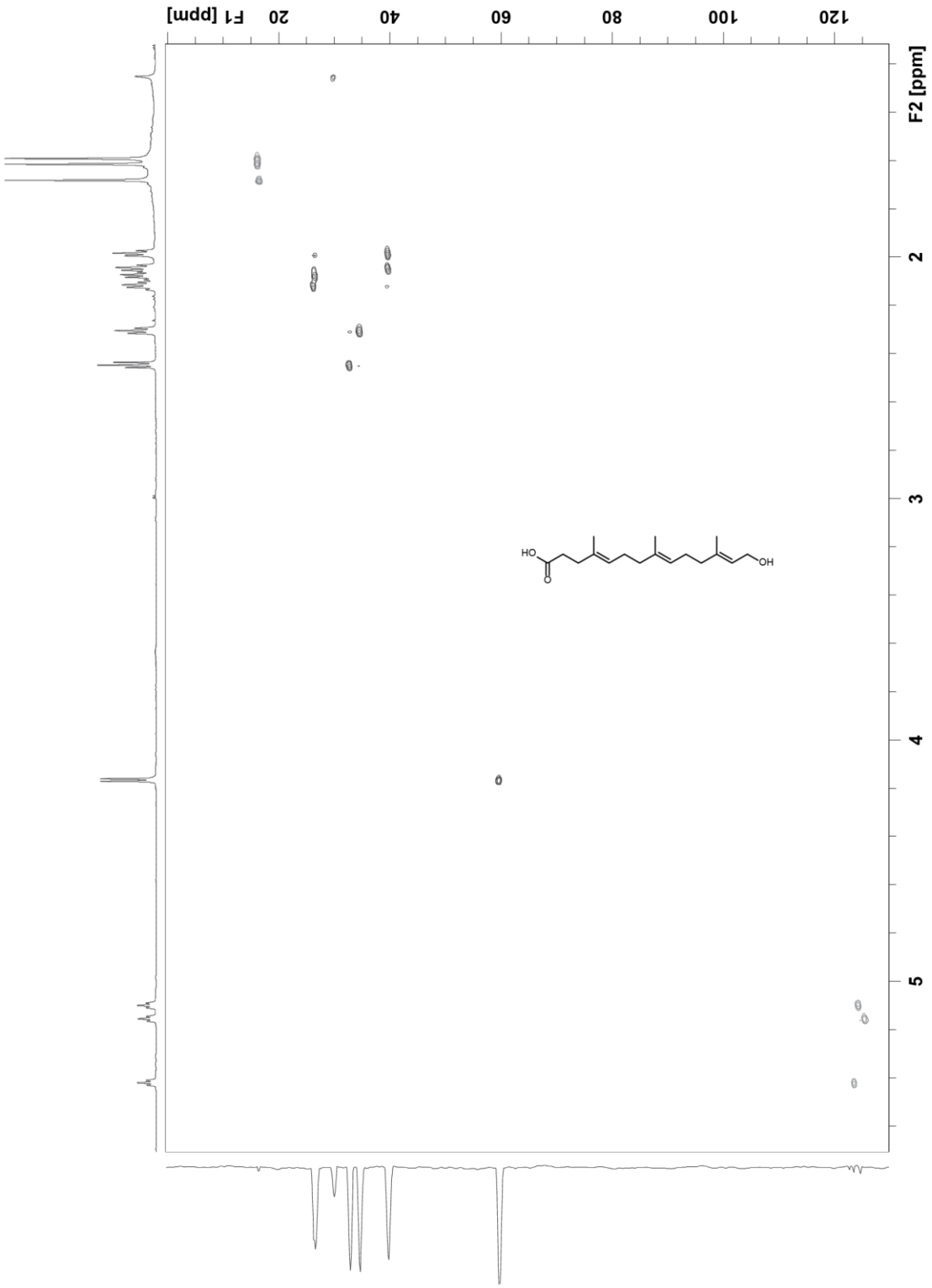


Figure A1.14. ^{1}H-^{13}C HSQC spectrum of 11.

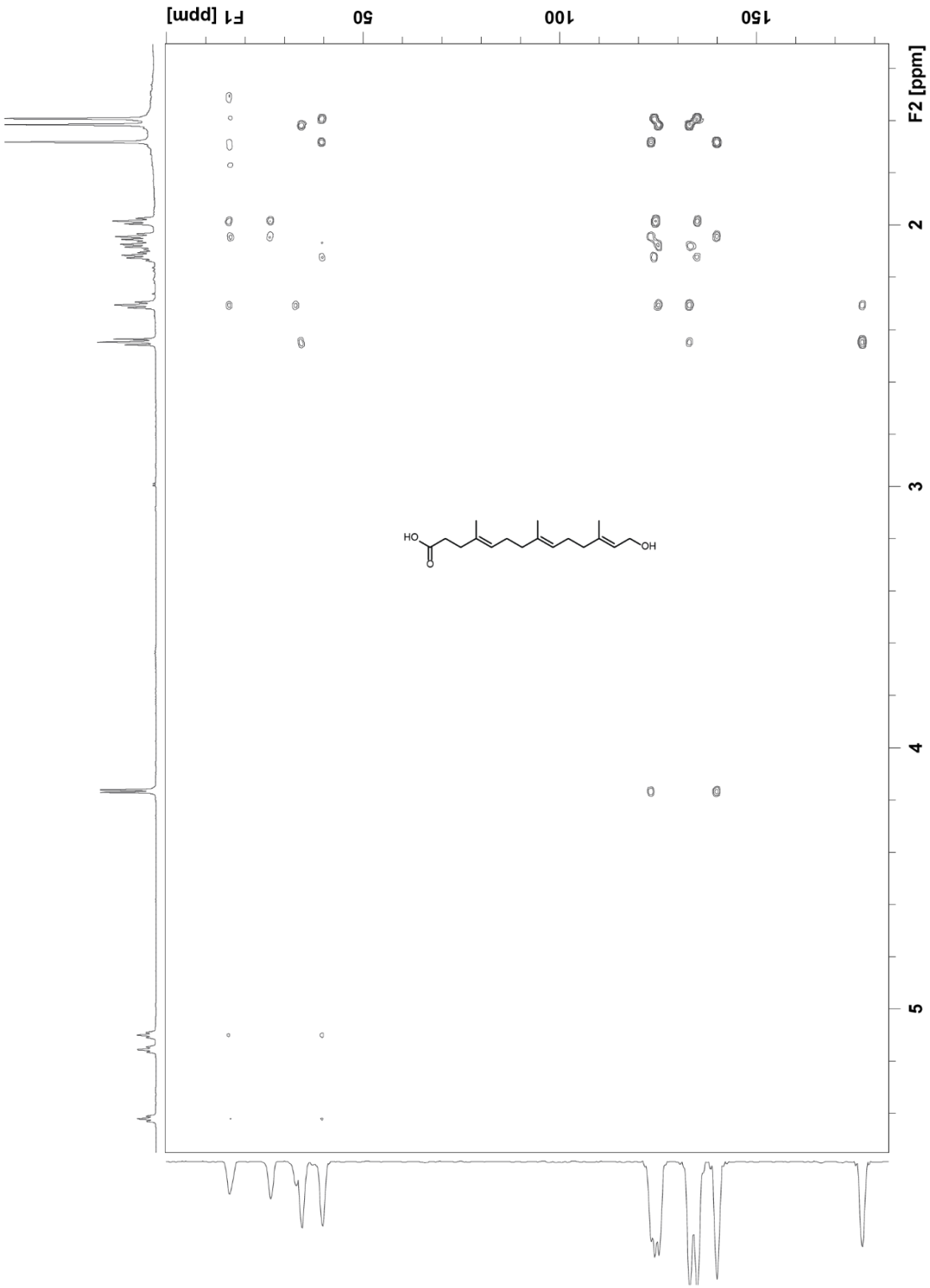


Figure A1.15. ^{1}H-^{13}C HMBC spectrum of 11.

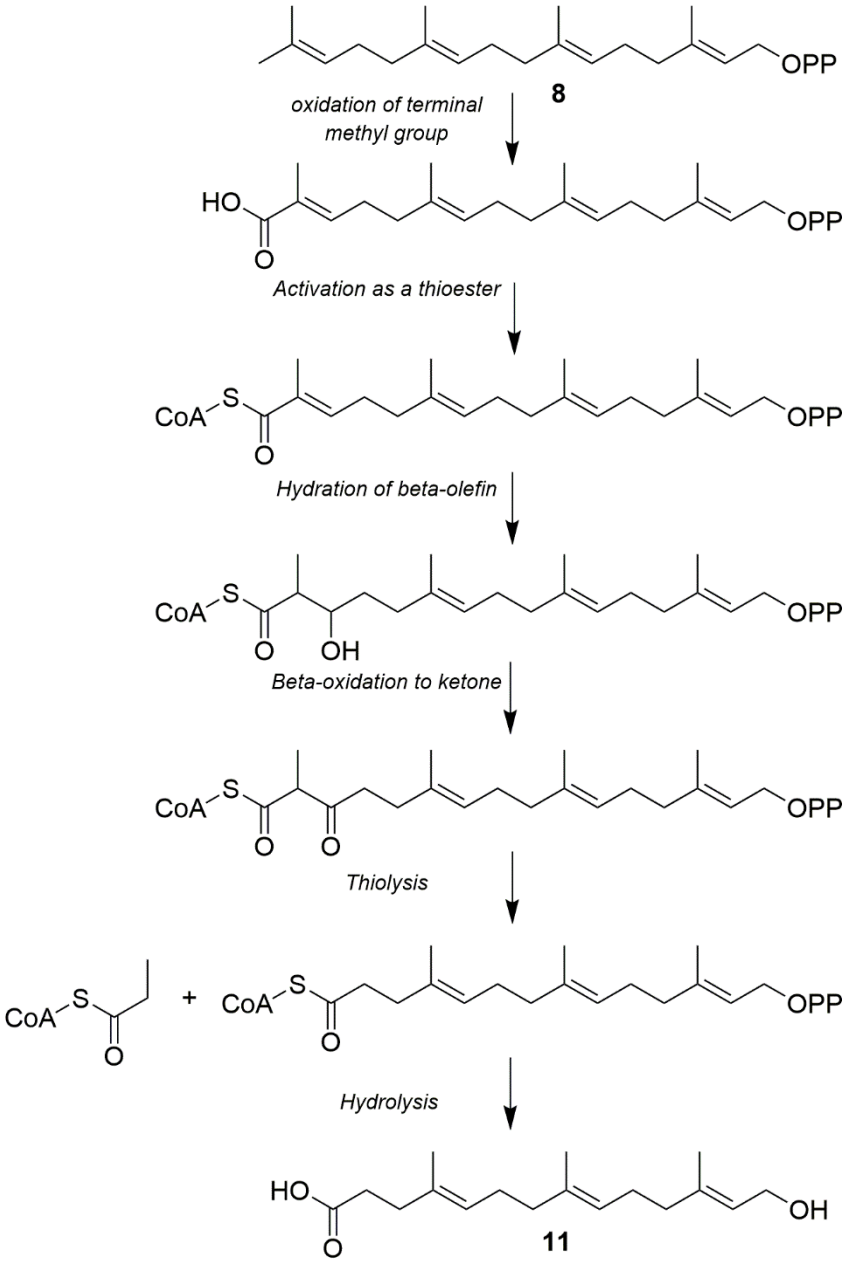


Figure A1.16. Putative biosynthesis of shunt metabolite 11 from intermediate 8.

The proposed mechanism is based on references [201]–[203].

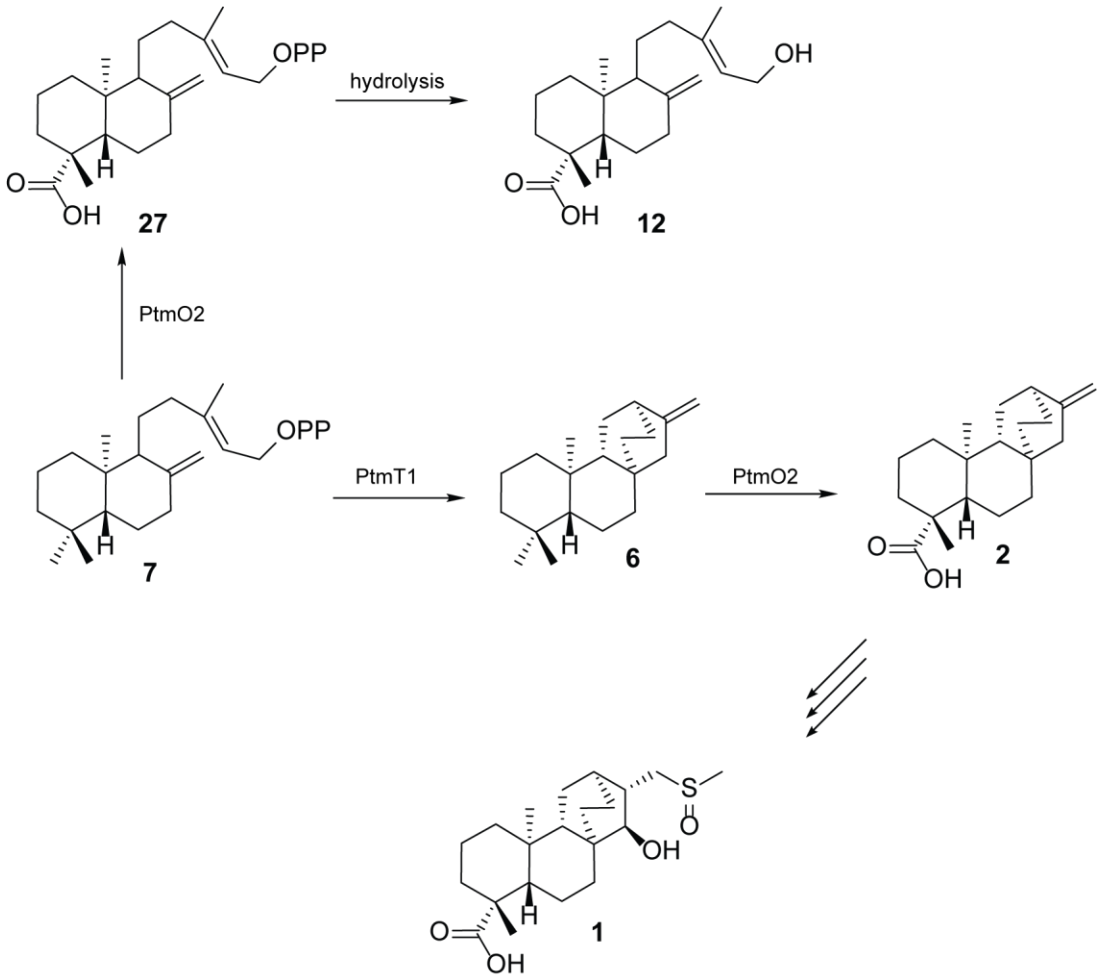


Figure A1.17. Putative biosynthesis of shunt metabolite 12 from intermediate 7.

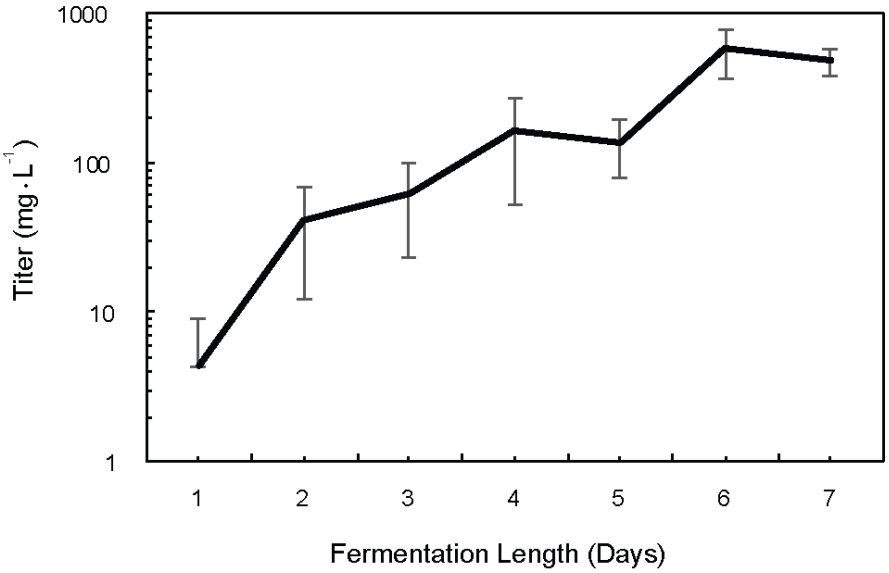


Figure A1.18. Production titer of 2 in 500 mL ISM3 fermentation medium over seven days.

Points plot the mean of six biological replicates with error bars corresponding to standard deviation.

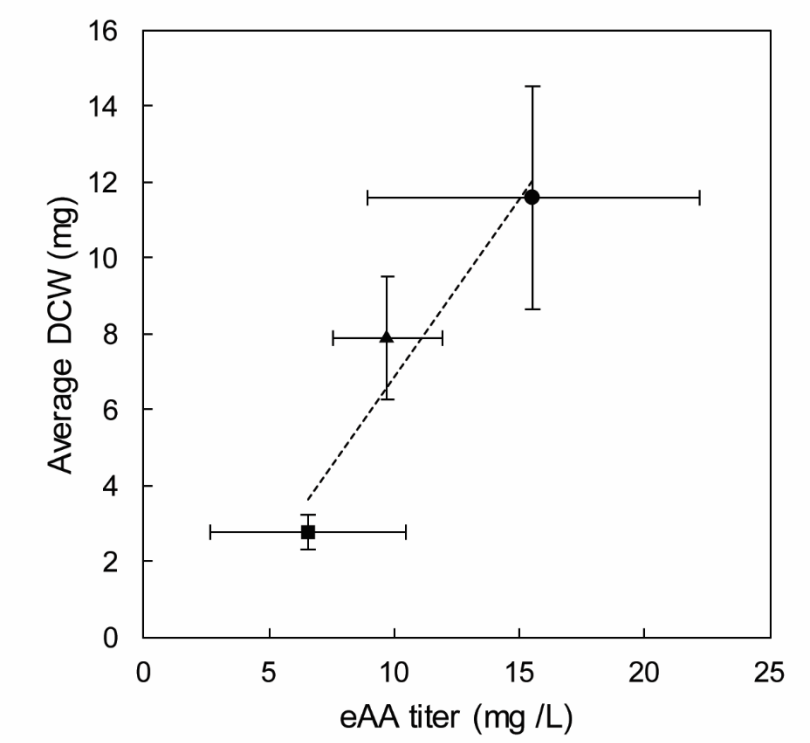


Figure A1.19. Dried cell weight versus titer of 2 for *S. albidoflavus* J1074 SGC5 cultured in 2 mL of PCNM (●), ISM3 (▲), or TSB (■) medium in 24-well MTP.

Data points represent mean of at least three experimental replicates over multiple days with error bars showing standard deviation.

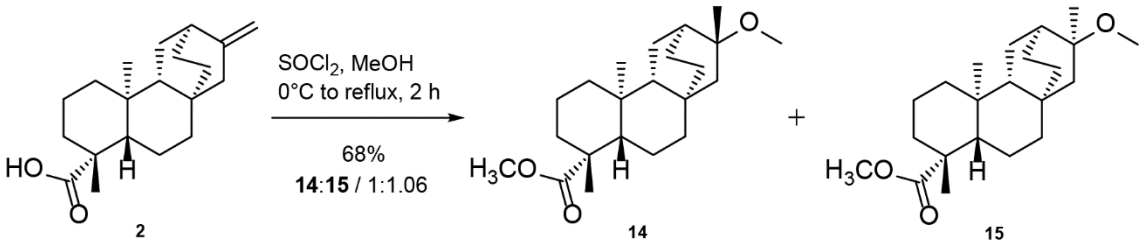
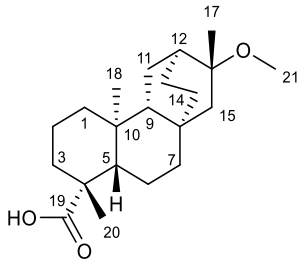


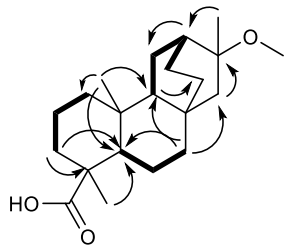
Figure A1.20. Methanol addition on the double bond of 2 in acidic conditions.



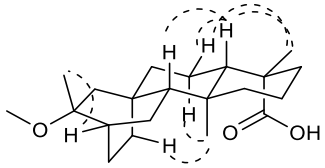
***ent*-(16S)-16-methoxyatisan-19-oic acid (14)**

Annotated. NMR data for 14 in chloroform-*d*.

No.	δ_H (#H, mult., *J*(Hz))	δ_C
1	0.84 (1H, td, 3.9, 13.3)	39.9
	1.59 (1H, m)	
2	1.40 (1H, m)	18.9
	1.86 (1H, m)	
3	1.01 (1H, m)	38.4
	2.15 (1H, d, 13.3)	
4	-	44.0
5	1.02 (1H, m)	57.5
6	1.72 (1H, m)	20.3
	1.78 (1H, m)	
7	1.02 (1H, m)	40.0
	1.43 (1H, m)	
8	-	33.4
9	0.95 (1H, dd, 7.3, 11.0)	51.6
10	-	38.4
11	1.30 (1H, td, 2.6, 6.9)	25.2
	1.56 (1H, m)	
12	1.74 (1H, m)	33.5
13	1.20 (1H, m)	22.1
	1.86 (1H, m)	
14	1.01 (1H, m)	27.0
	1.83 (1H, m)	
15	1.02 (1H, m)	55.8
	1.37 (1H, d,13.7)	
16	-	76.6
17	1.21 (3H, s)	24.6
18	0.89 (3H, s))	12.5
19	-	183.2
20	1.23 (3H, s)	29.1
21	3.16 (3H, s)	49.1



Key COSY (bold) and HMBC (arrows) correlations



Key NOESY correlations

Figure A1.21. Summary of key 1D and 2D NMR data for compound 14.

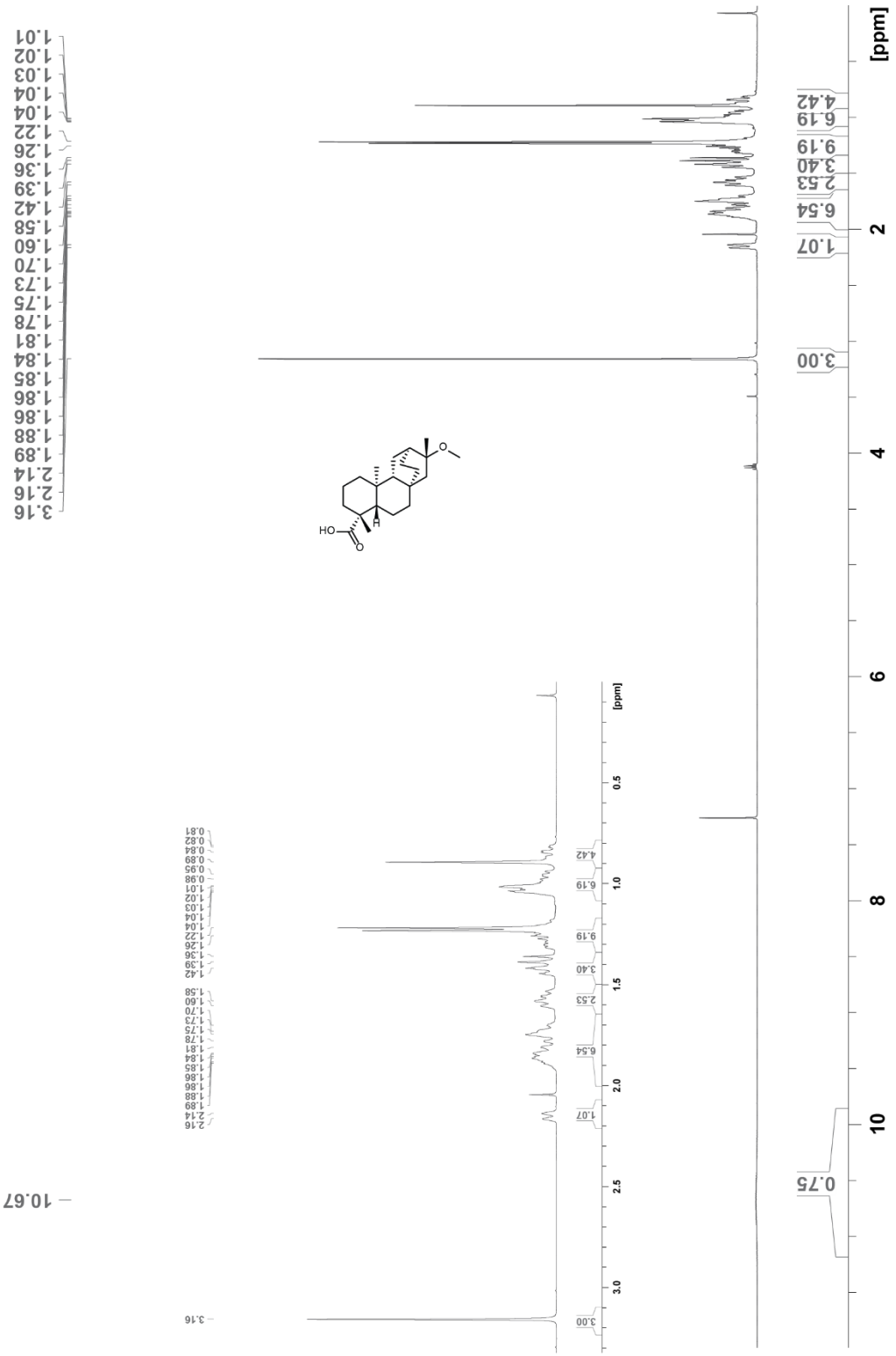


Figure A1.22. ^{1}H spectrum of 14.

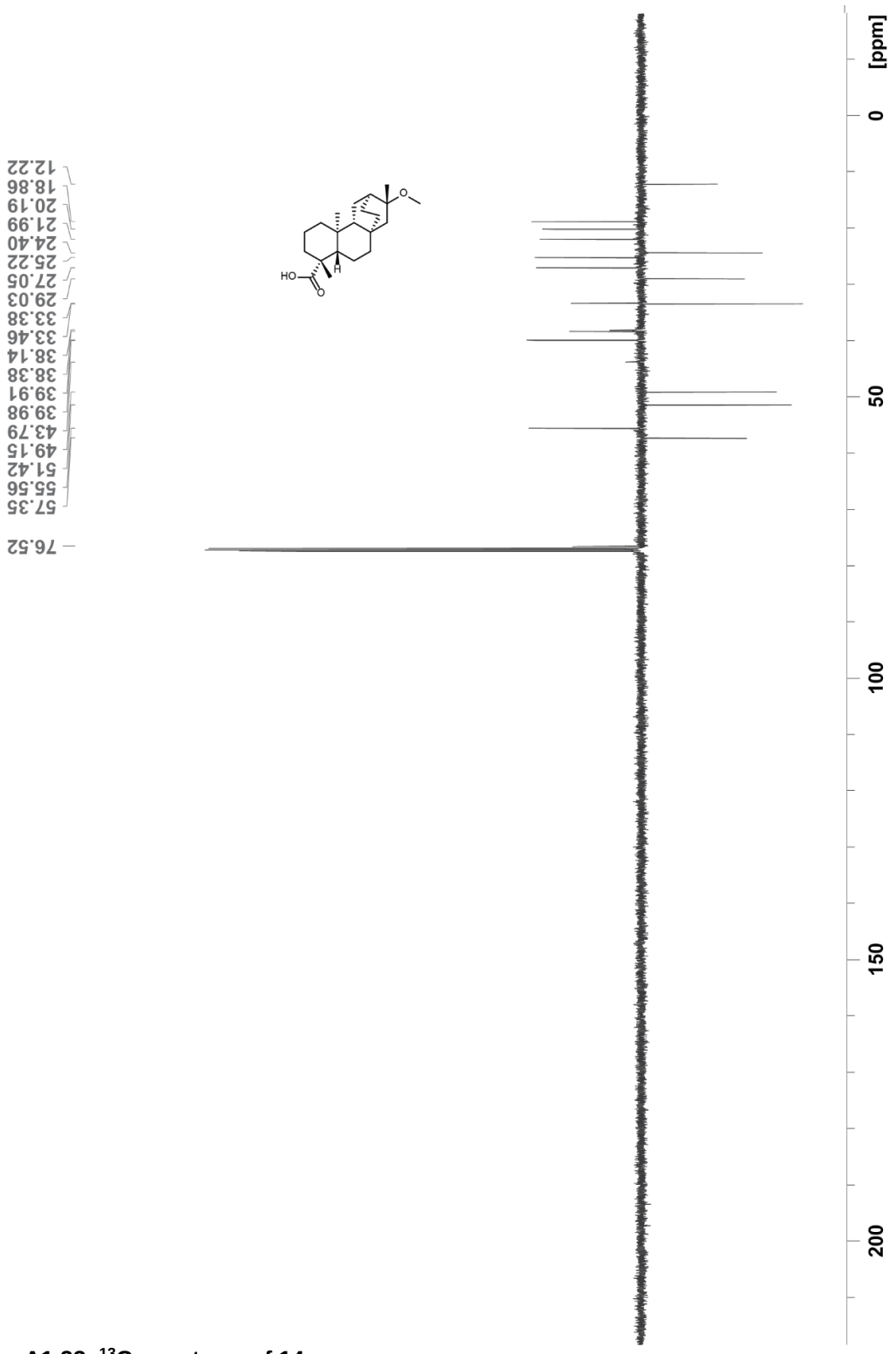


Figure A1.23. ^{13}C spectrum of 14.

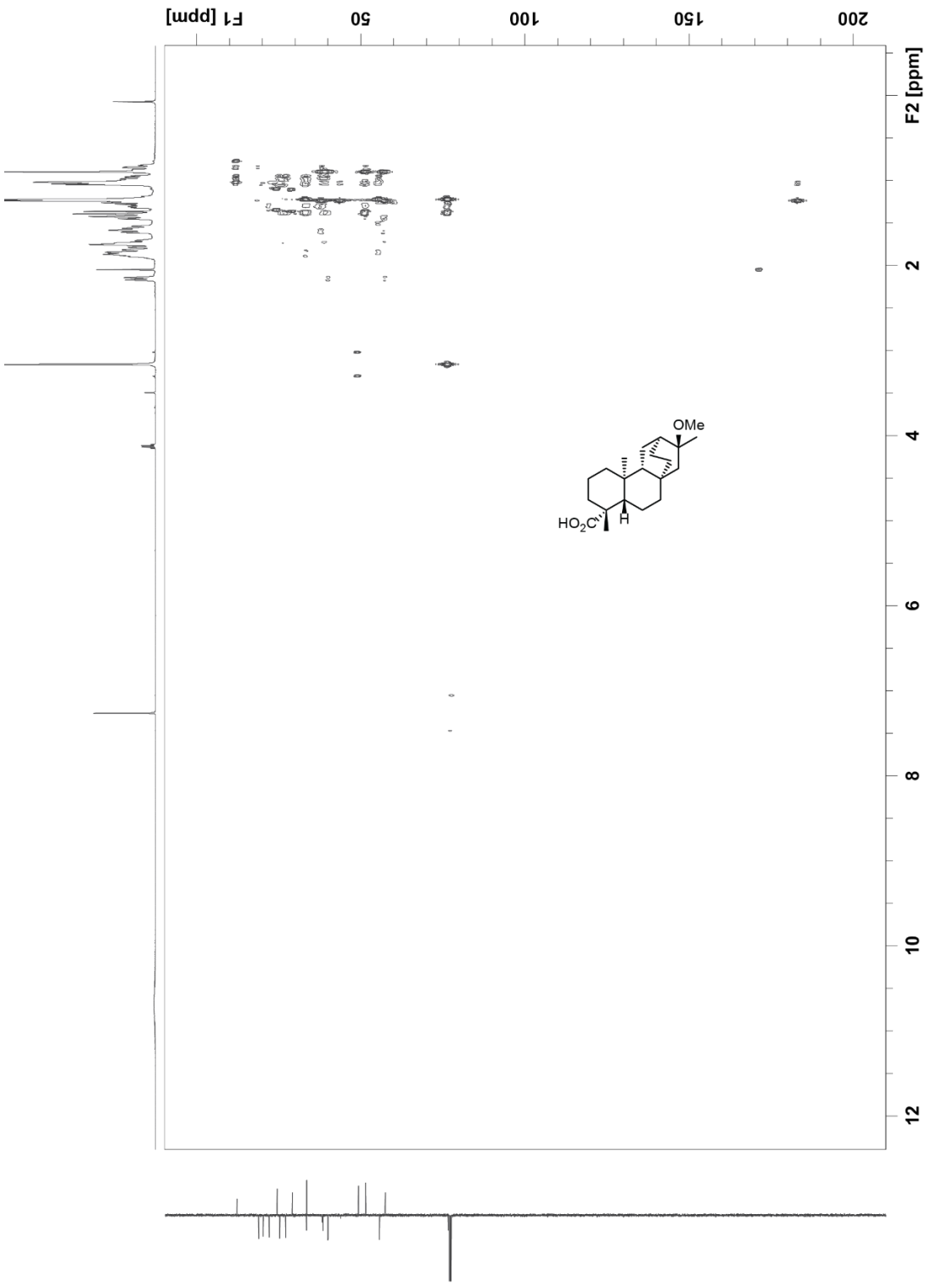
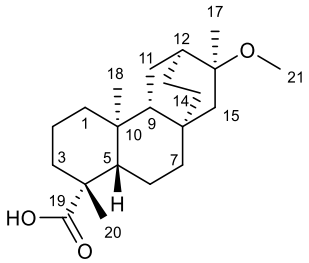


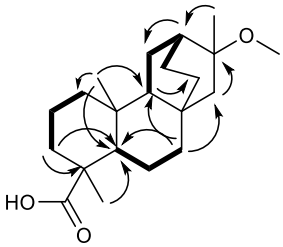
Figure A1.24. ^{1}H-^{13}C HMBC spectrum of 14.



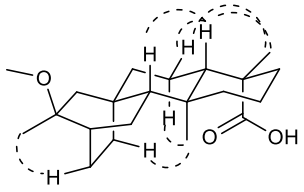
***ent*-(16R)-16-methoxyatisan-19-oic acid (15)**

Annotated NMR data for 15 in chloroform-*d*.

No.	δ_H (#H, mult., *J*(Hz))	δ_C
1	0.90 (1H, m)	39.6
	1.60 (1H, d, 13.0)	
2	1.41 (1H, m)	18.9
	1.86 (1H, m)	
3	1.01 (1H, td, 4.3, 13.1)	38.1
	2.14 (1H, d, 13.1)	
4	-	43.9
5	1.06 (1H, m)	57.3
+6	1.74 (1H, m)	20.3
	1.79 (1H, m)	
7	1.07 (1H, m)	39.9
	1.39 (1H, m)	
8	-	33.5
9	1.17 (1H, m)	50.6
10	-	38.4
11	1.07 (1H, m)	23.4
	1.84 (1H, m)	
12	1.76 (1H, m)	33.4
13	1.44 (1H, m)	23.7
	1.57 (1H, m)	
14	0.83 (1H, m)	27.4
	1.88 (1H, m)	
15	1.22 (2H, m)	55.8
16	-	76.4
17	1.22 (3H, s)	24.1
18	0.88 (3H, s)	12.2
19	-	183.9
20	1.23 (3H, s)	29.0
21	3.16 (3H. s)	49.1



Key COSY (bold) and HMBC (arrows) correlations



Key NOESY correlations

Figure A1.25. Summary of key 1D and 2D NMR data for compound 15.

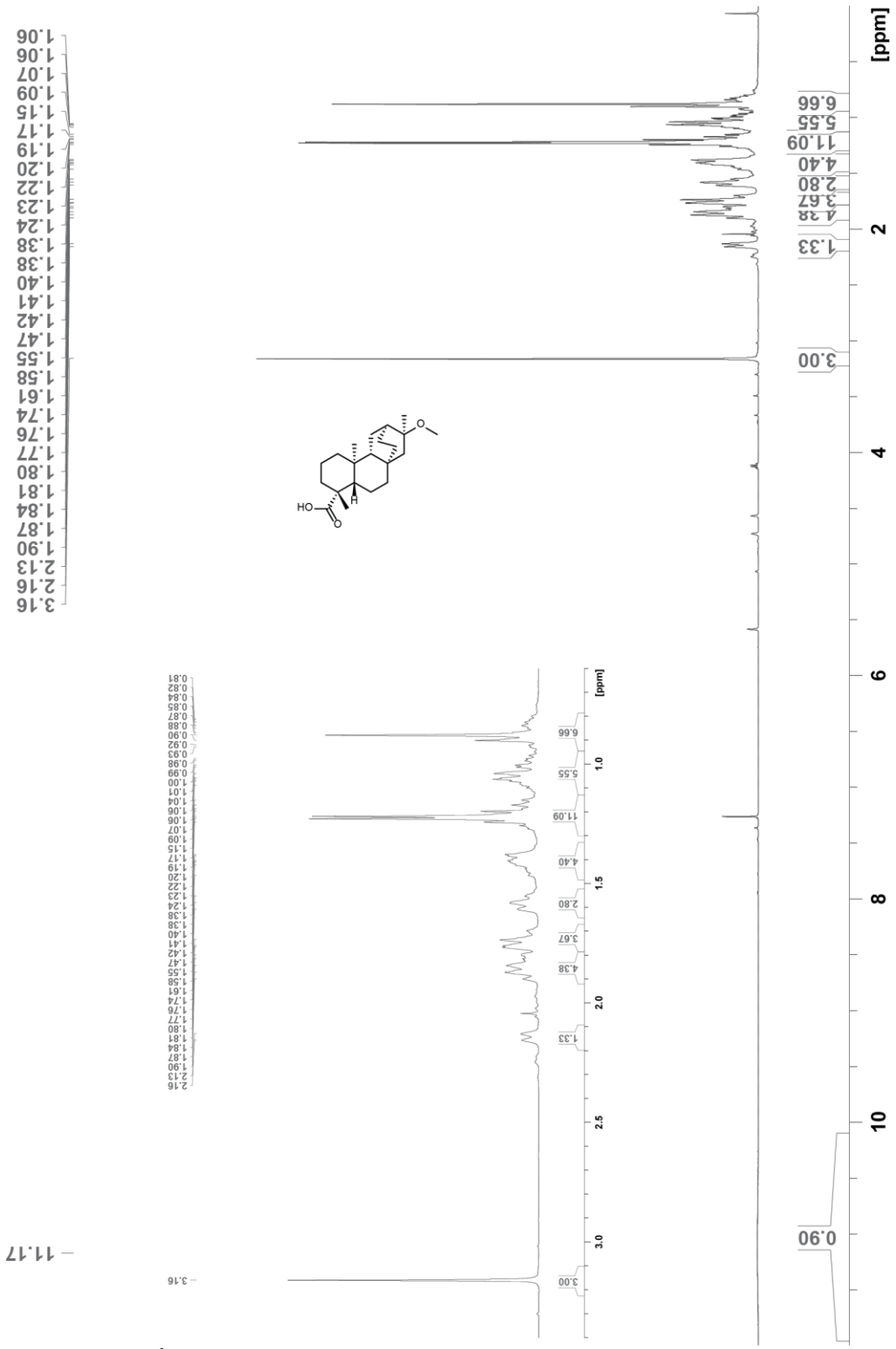


Figure A1.26. ^{1}H spectrum of 15.

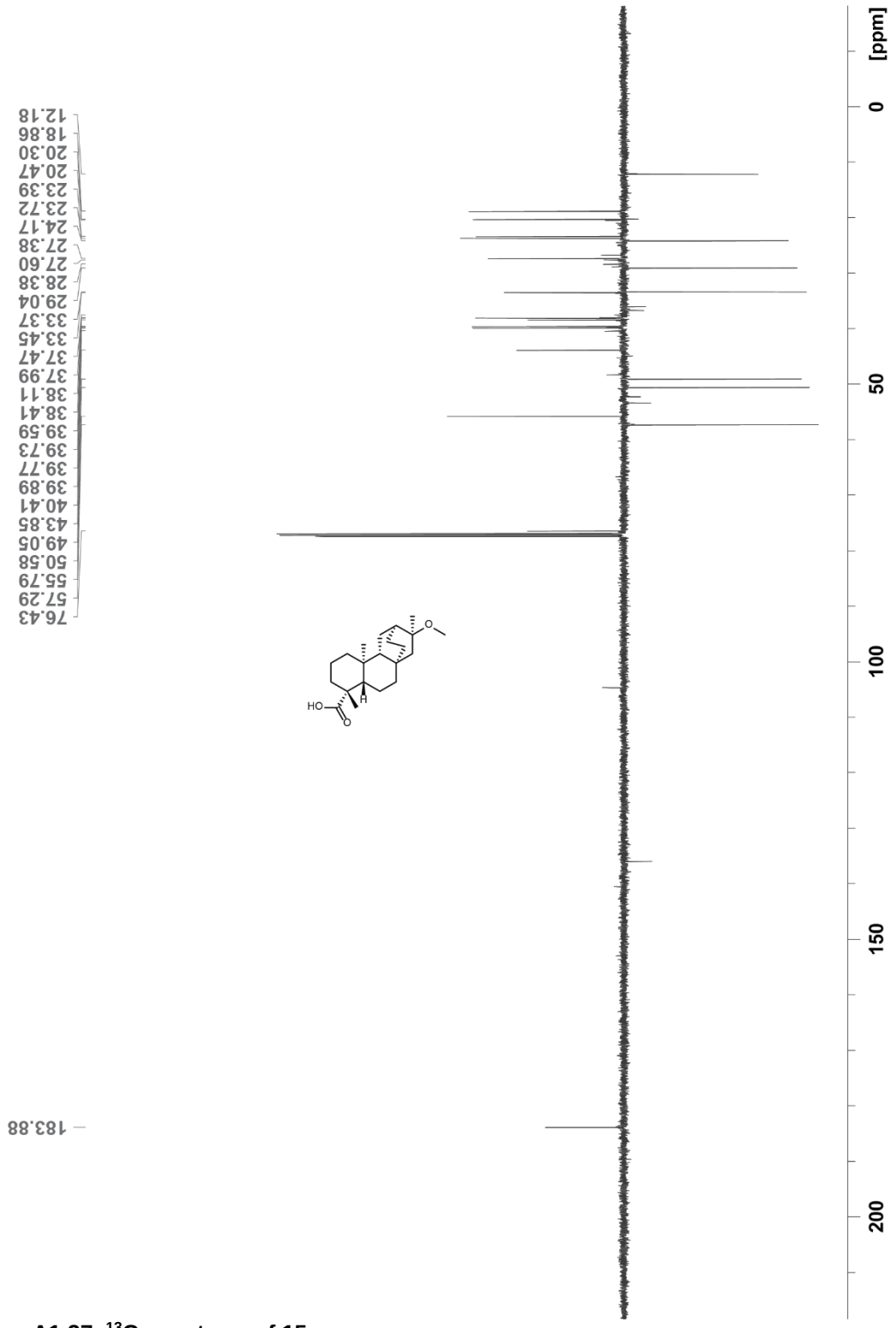


Figure A1.27. ^{13}C spectrum of 15.

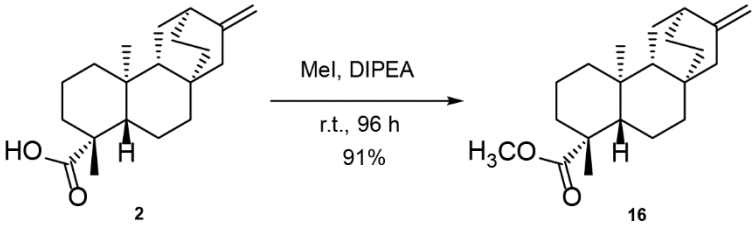
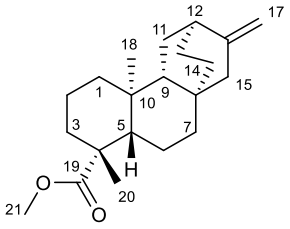


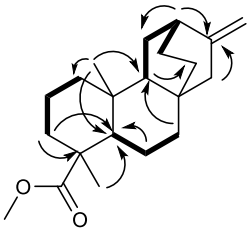
Figure A1.28. Esterification of 2 to 16 in basic conditions.



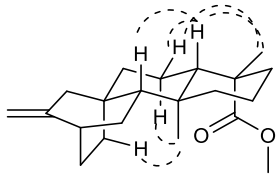
methyl *ent*-atis-16-en-19-oate (16)

Annotated NMR data for 16 in chloroform-*d*.

No.	δ_H (#H, mult., *J*(Hz))	δ_C
1	0.84 (1H, m)	39.8
	1.58 (1H, bd, 13.6)	
2	1.40 (1H, m)	18.9
	1.83 (1H, m)	
3	1.01 (1H, m)	38.3
	2.17 (1H, bd, 13.6)	
4	-	43.6
5	1.03 (1H, m)	57.3
6	1.74 (2H, m)	20.4
7	1.11 (1H, m)	39.8
	1.47 (1H, bd, 13.6)	
8	-	33.6
9	1.13 (1H, m)	52.2
10	-	38.5
11	1.40 (1H, m)	28.8
	1.59 (1H, m)	
12	2.22 (1H, br s)	36.7
13	1.59 (2H, m)	27.4
14	0.99 (1H, m)	28.4
	1.95 (1H, m)	
15	1.88 (1H, br d, 17.0)	48.3
	2.02 (1H, br d, 17.0)	
16	-	152.5
17	4.56 (1H, br s)	104.6
	4.73 (1H, br s)	
18	0.78 (3H, s)	12.0
19	-	178.1
20	1.18 (3H, s)	28.9
21	3.64 (3H, s)	51.3



Key COSY (bold) and HMBC (arrows) correlations



Key NOESY correlations

Figure A1.29. Summary of key 1D and 2D NMR data for compound 16.

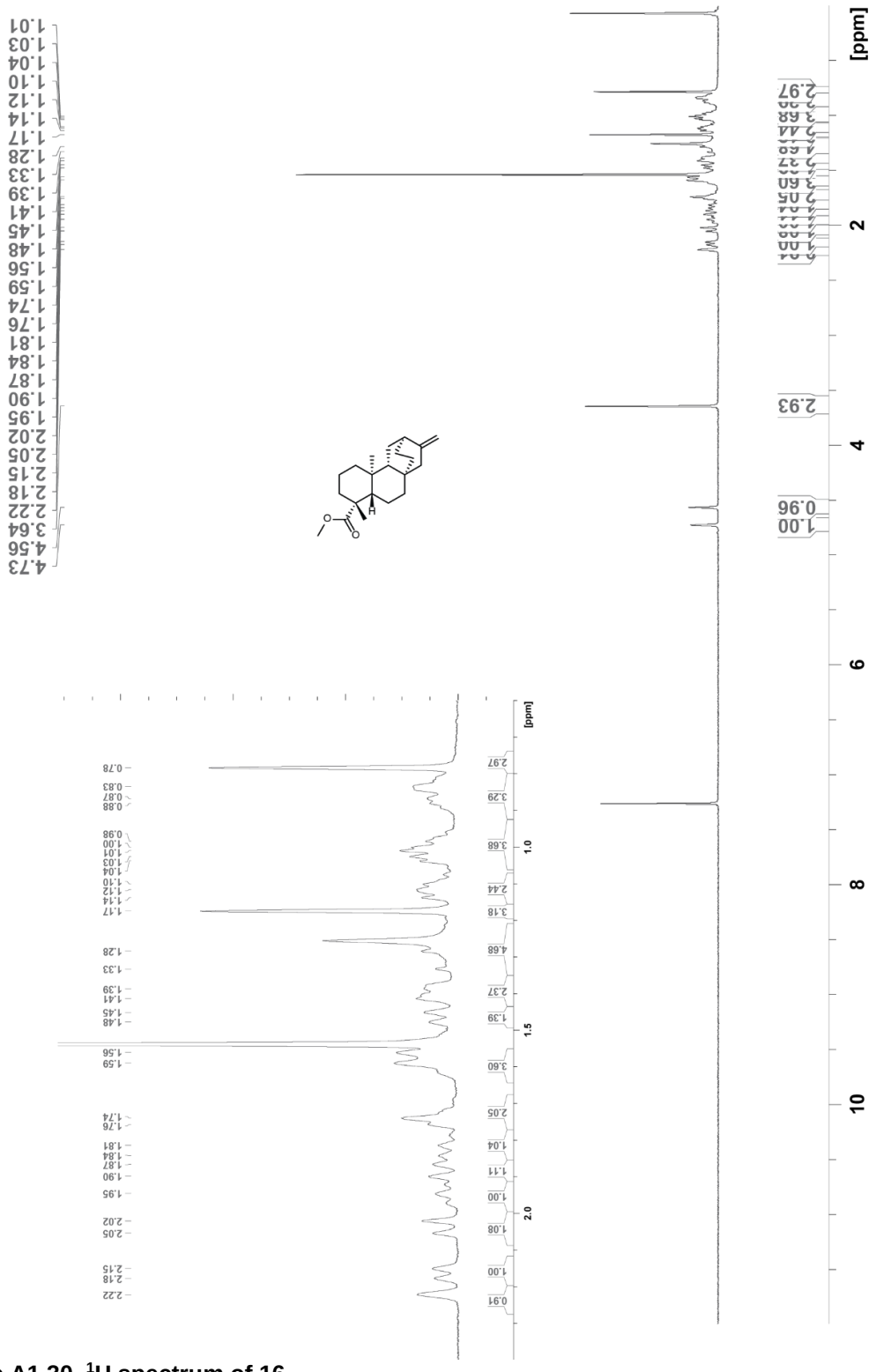


Figure A1.30. ^{1}H spectrum of 16.

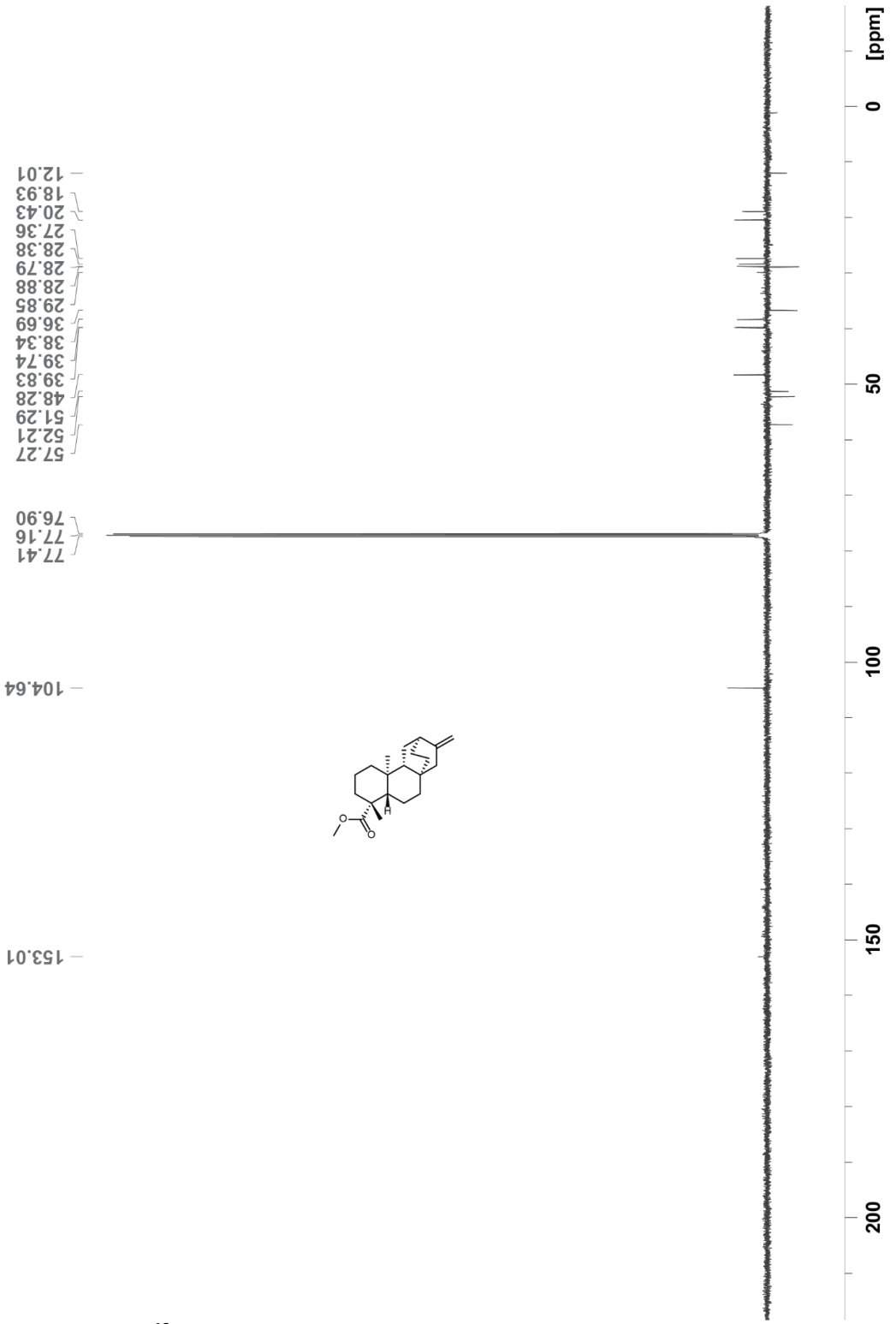


Figure A1.31. ^{13}C spectrum of 16.

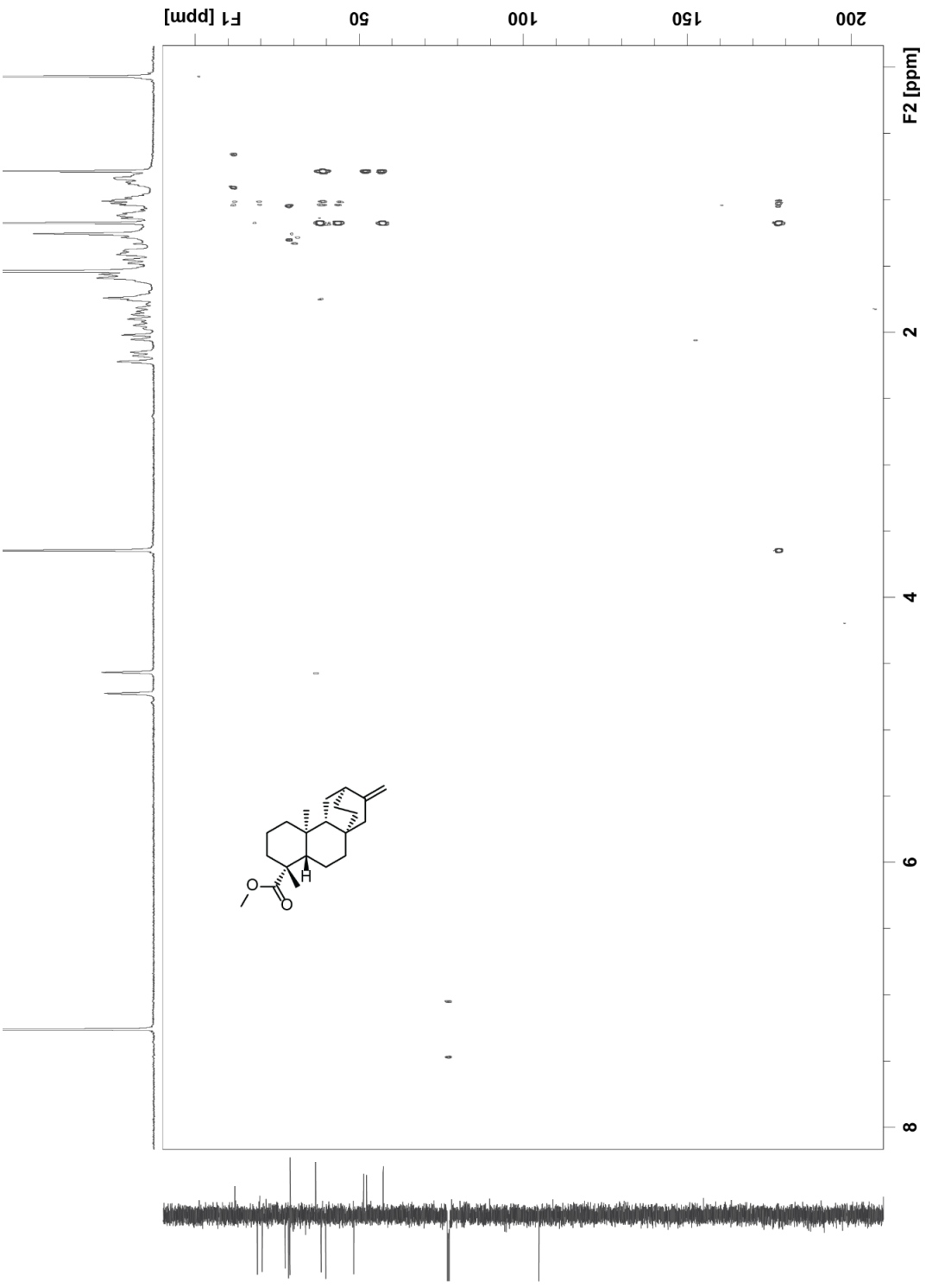


Figure A1.32. ^{1}H-^{13}C HMBC spectrum of 16.

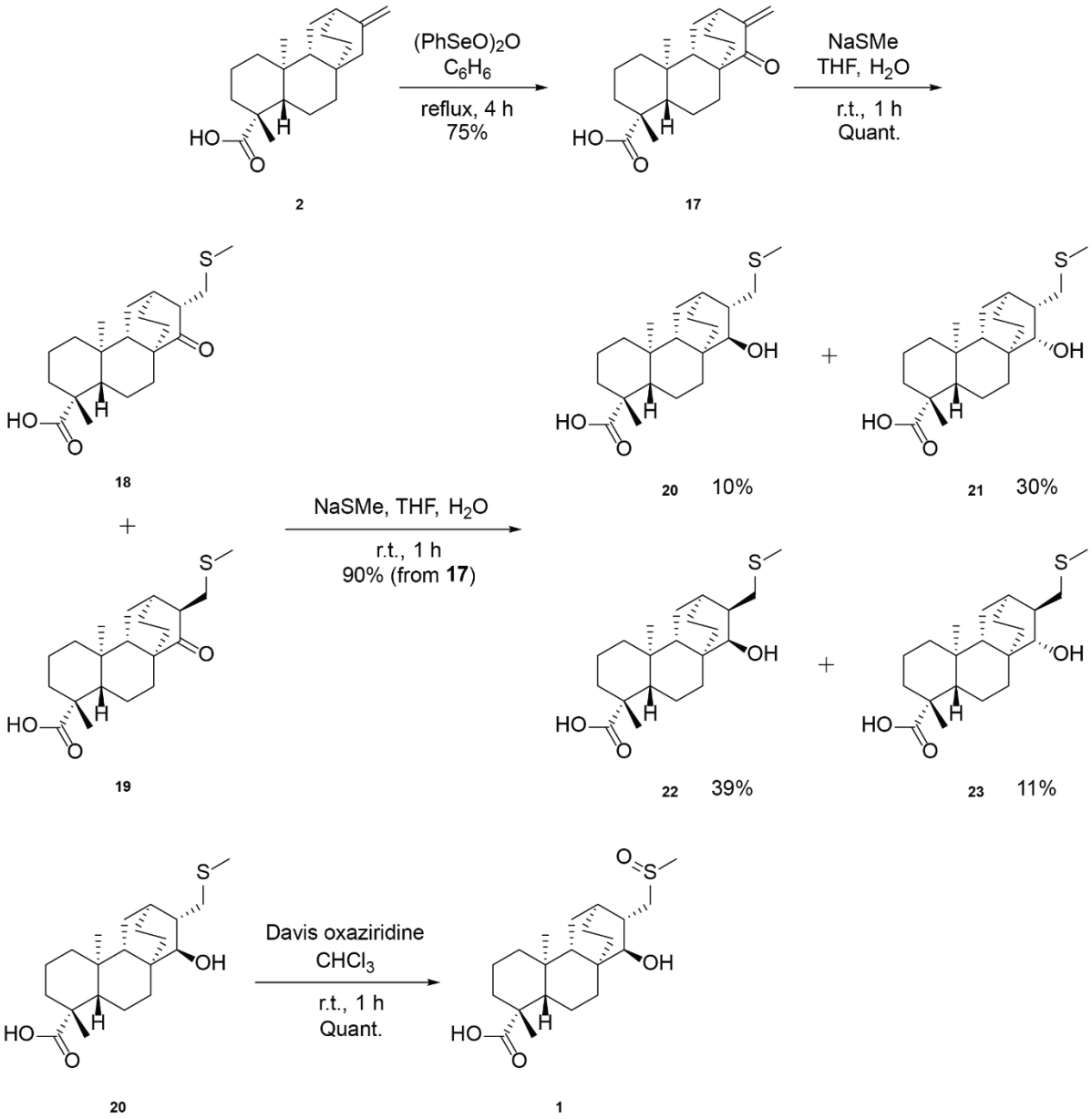
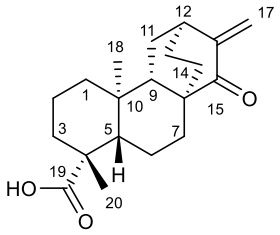


Figure A1.33. Total synthesis of 1 from 2.

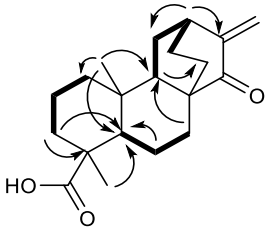
Reaction conditions are provided above arrows and isolated yields are provided below.



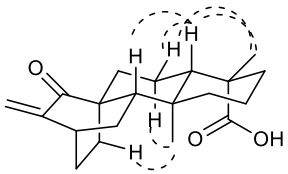
***ent*-15-oxoatis-16-en-19-oic acid (17)**

Annotated NMR data for 17 in chloroform-*d*

No.	δ_H (#H, mult., *J*(Hz))	δ_C
1	0.85 (1H, m)	39.8
	1.62 (1H, m)	
2	1.42 (1H, m)	18.7
	1.86 (1H, m)	
3	1.02 (1H, m)	37.9
	2.17 (1H, d, 13.5)	
4	-	43.9
5	1.05 (1H, m)	56.3
6	1.84(1H, m)	19.6
	1.92 (1H, m)	
7	1.50 (1H, m)	30.6
	1.83 (1H, m)	
8	-	45.3
9	1.32 (1H, dd, 6.8, 11.4)	46.7
10	-	39.3
11	1.57 (1H, m)	28.4
	1.71 (1H, m)	
12	2.76 (1H, br s)	36.2
13	1.72 (2H, m)	26.2
14	1.25 (1H, m)	24.8
	2.23 (1H, ddd, 3.2, 11.2, 14.3)	
15	-	204.0
16	-	147.3
17	5.20 (1H, d, 1.3)	117.0
	5.94 (1H, d, 1.3)	
18	1.01 (3H, s)	12.8
19	-	182.6
20	1.27 (3H, s)	29.0



Key COSY (bold) and HMBC (arrows) correlations



Key NOESY correlations

Figure A1.34. Summary of key 1D and 2D NMR data for compound 17.

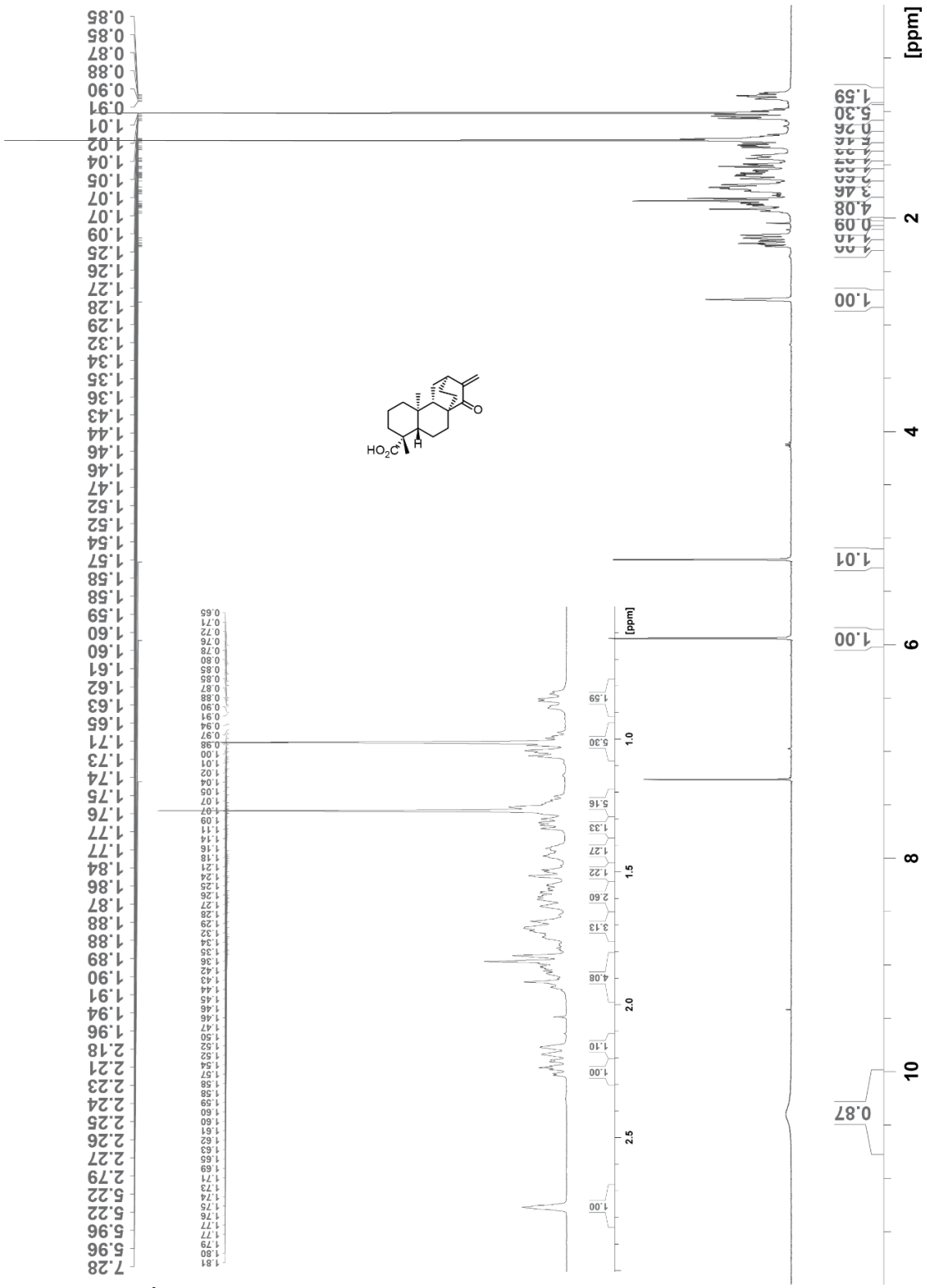


Figure A1.35. ^{1}H spectrum of 17.

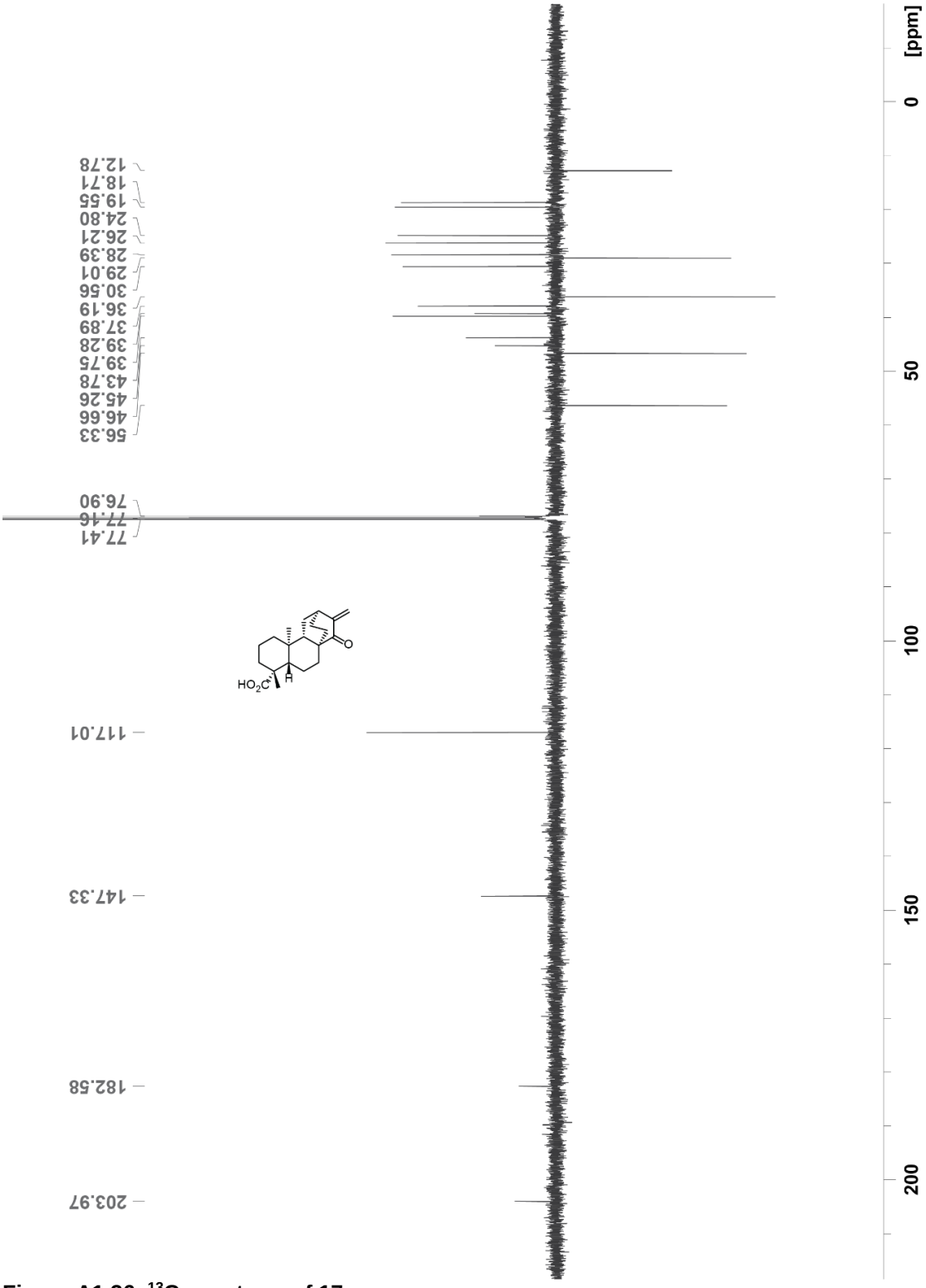
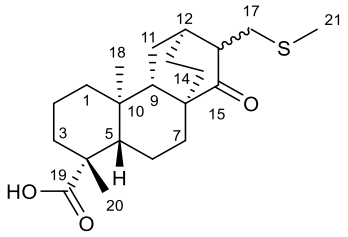


Figure A1.36. ^{13}C spectrum of 17.



***ent*-17-methylsulfenyl-15-oxoatisan-19-oic acid (18 and 19)**

Figure A1.37. Structure of compounds 18 and 19

Compounds purified as a racemic mixture. Chemical shifts not assigned

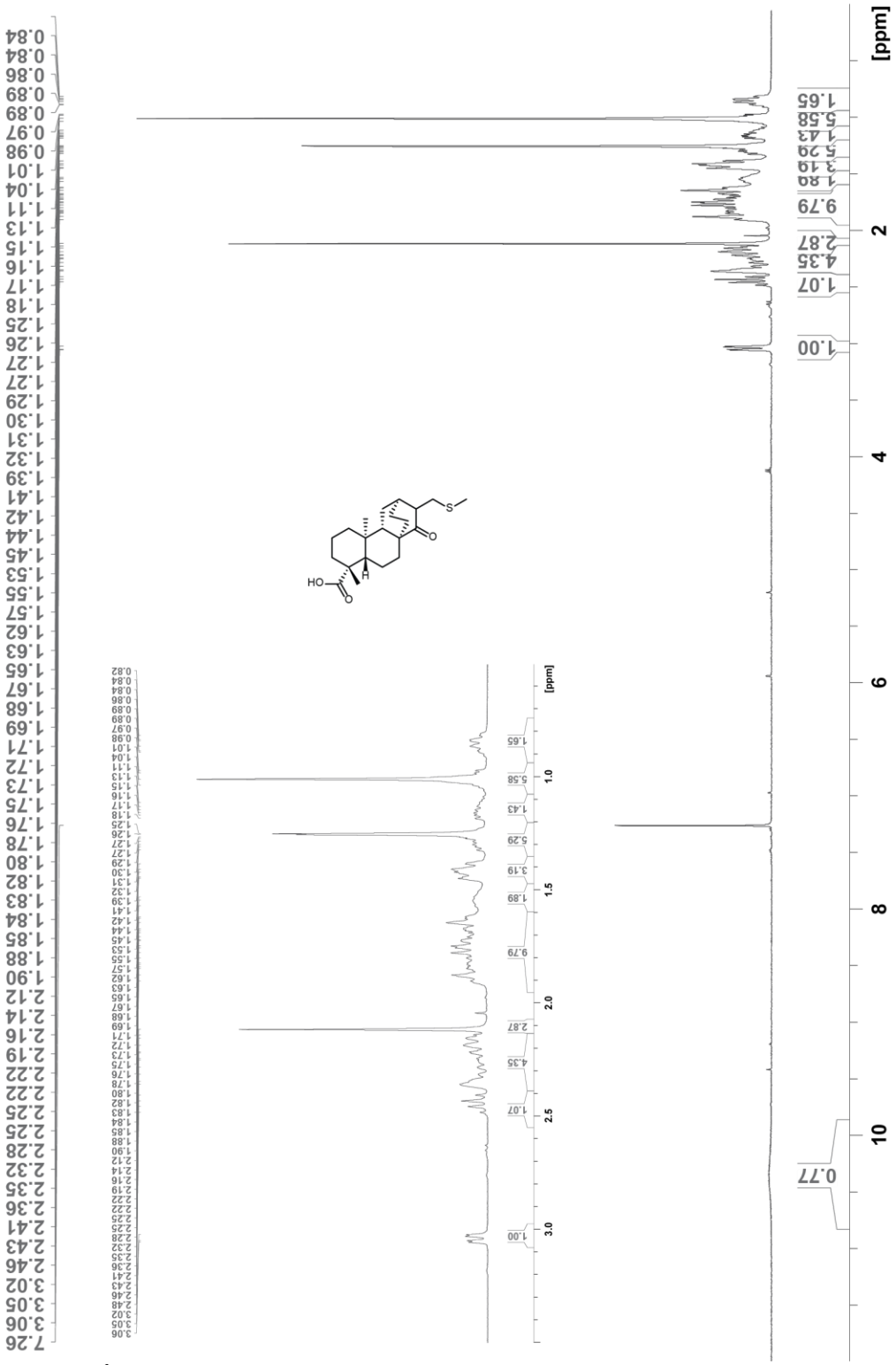


Figure A1. 38. ^{1}H spectrum of 18 and 19.

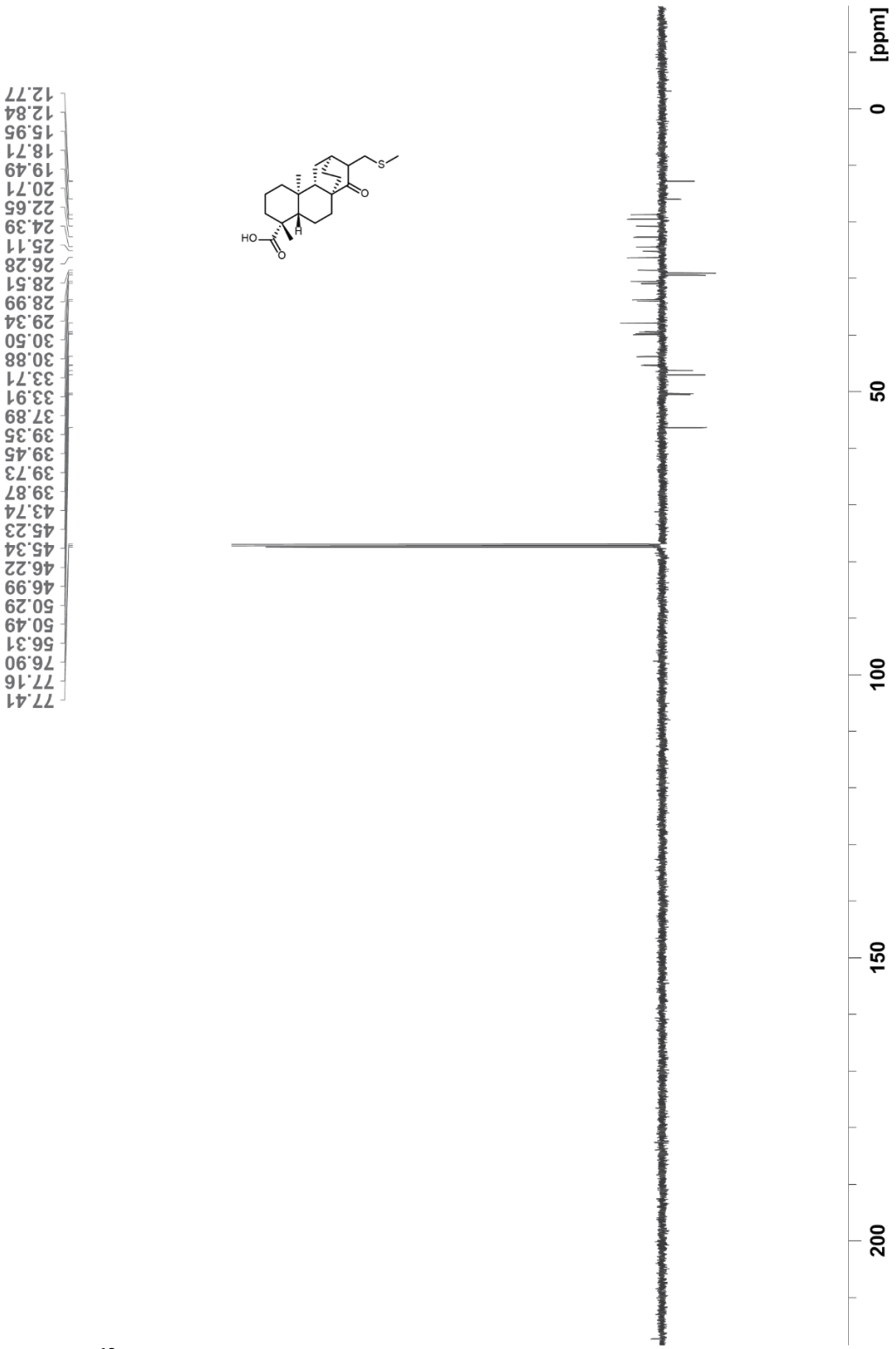
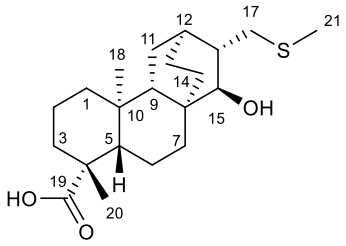


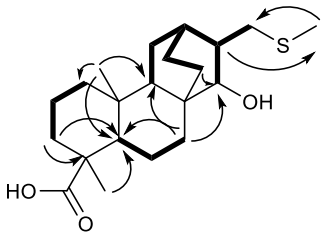
Figure A1.39. ^{13}C spectrum of 18 and 19.



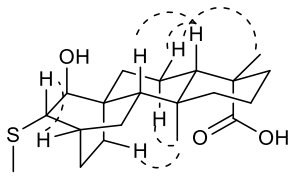
***ent*-(15S,16R)-15-hydroxy-17-methylsulfenylatisan-19-oic acid (20)**

Annotated NMR data for 20 in chloroform-*d.*

No.	δ_H (#H, mult., *J*(Hz))	δ_C
1	0.88 (1H, m)	39.9
	1.63 (1H, m)	
2	1.42 (1H, m)	18.9
	1.85 (1H, m)	
3	1.03 (1H, m)	37.6
	2.15 (1H, m)	
4	-	43.8
5	1.08 (1H, m)	56.6
6	1.80 (2H, m)	19.7
7	1.12 (1H, dt, 2.7, 13.1)	33.2
	1.69 (1H, m)	
8	-	36.5
9	1.61 (1H, m)	47.4
10	-	38.1
11	1.48 (1H, m)	29.3
	1.53 (1H, m)	
12	1.66 (1H, m)	30.0
13	1.30 (1H, m)	21.3
	1.53 (1H, m)	
14	0.74 (1H, td, 6.5, 13.1)	27.5
	1.97 (1H, m)	
15	2.95 (1H, d, 3.7)	81.1
16	1.54 (1H, m)	42.1
17	2.53 (2H, d, 7.8)	39.3
18	0.91 (3H, s)	13.0
19	-	182.5
20	1.25 (3H, s)	29.0
21	2.14 (3H, s)	15.7



Key COSY (bold) and HMBC (arrows) correlations



Key NOESY correlations

Figure A1.40. Summary of key 1D and 2D NMR data for compound 20.

Figure A1.41. ^{1}H spectrum of 20.

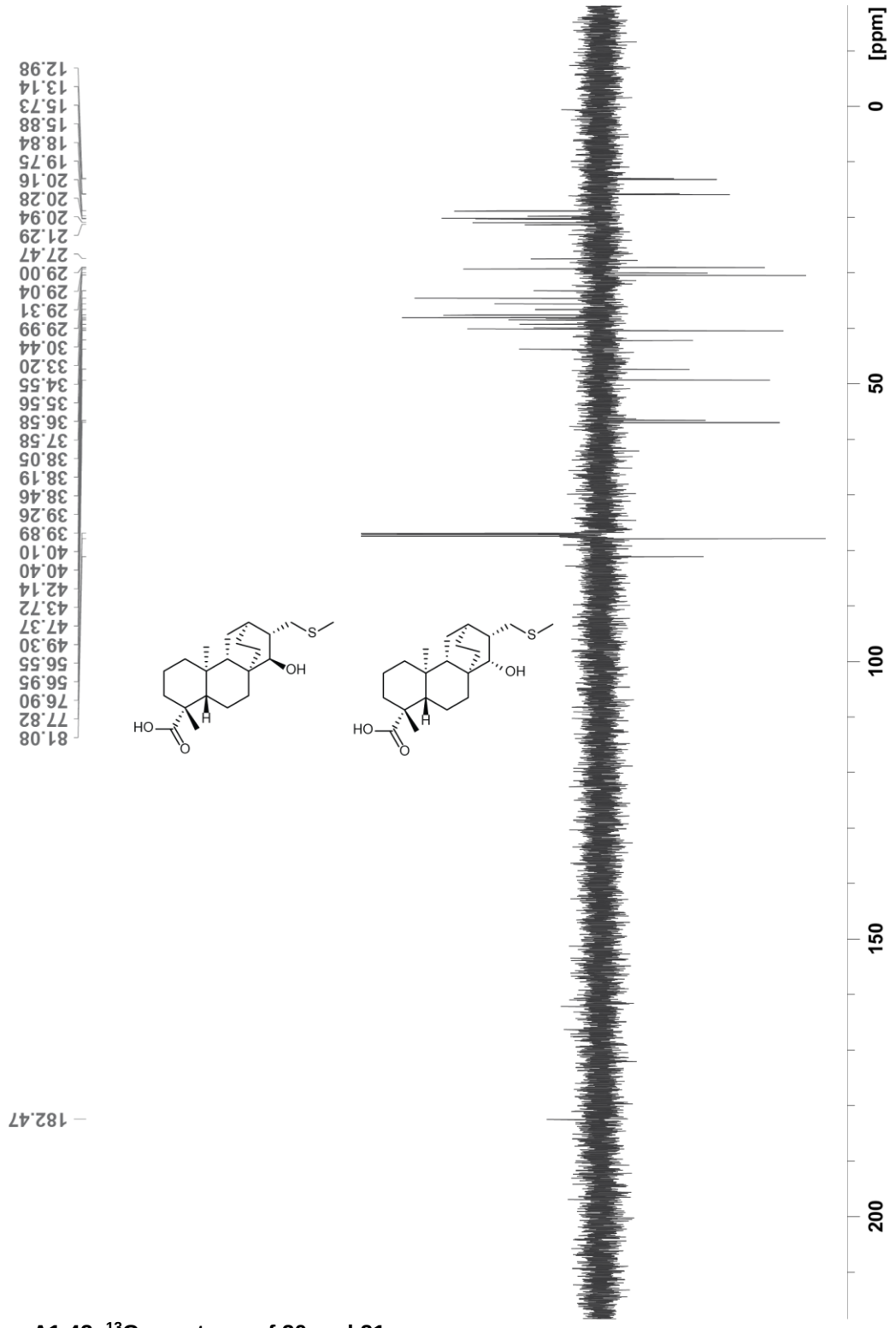
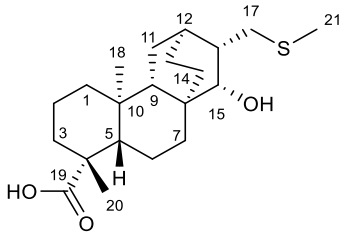


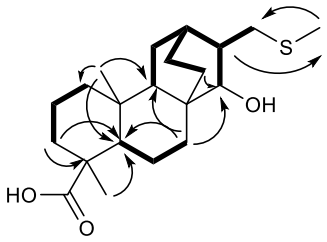
Figure A1.42. ^{13}C spectrum of 20 and 21.



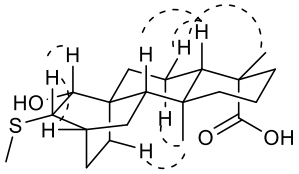
ent-(15S,16S)-15-hydroxy-17-methylsulfenylatisan-19-oic acid (21)

Annotated NMR data for 21 in chloroform-_d_.

No.	δ_H (#H, mult., J(Hz))	δ_C
1	0.88 (1H, m)	40.1
	1.63 (1H, m)	
2	1.41 (1H, m)	18.8
	1.87 (1H, m)	
3	1.02 (1H, m)	38.0
	2.15 (1H, m)	
4	-	43.8
5	1.02 (1H, m)	57.0
6	1.72 (1H, m)	20.1
	1.78 (1H, m)	
7	1.06 (1H, m)	37.6
	1.80	
8	-	35.6
9	0.99 (1H, m)	49.3
10	-	38.5
11	1.37 (1H, m)	29.3
	1.48 (1H, m)	
12	1.54 (1H, m)	30.4
13	1.27 (1H, m)	20.9
	1.64 (1H, m)	
14	1.47 (1H, m)	20.3
	1.64 (1H, m)	
15	3.45 (1H, d, 8.8)	77.8
16	1.97 (1H, m)	40.4
17	2.51 (1H, dd, 5.3, 12.2)	34.4
	2.89 (1H, m)	
18	0.94 (3H, s)	13.1
19	-	182.5
20	1.25 (3H, s)	29.0
21	2.14 (3H, s)	15.9



Key COSY (bold) and HMBC (arrows) correlations



Key NOESY correlations

Figure A1.43. Summary of key 1D and 2D NMR data for compound 21.

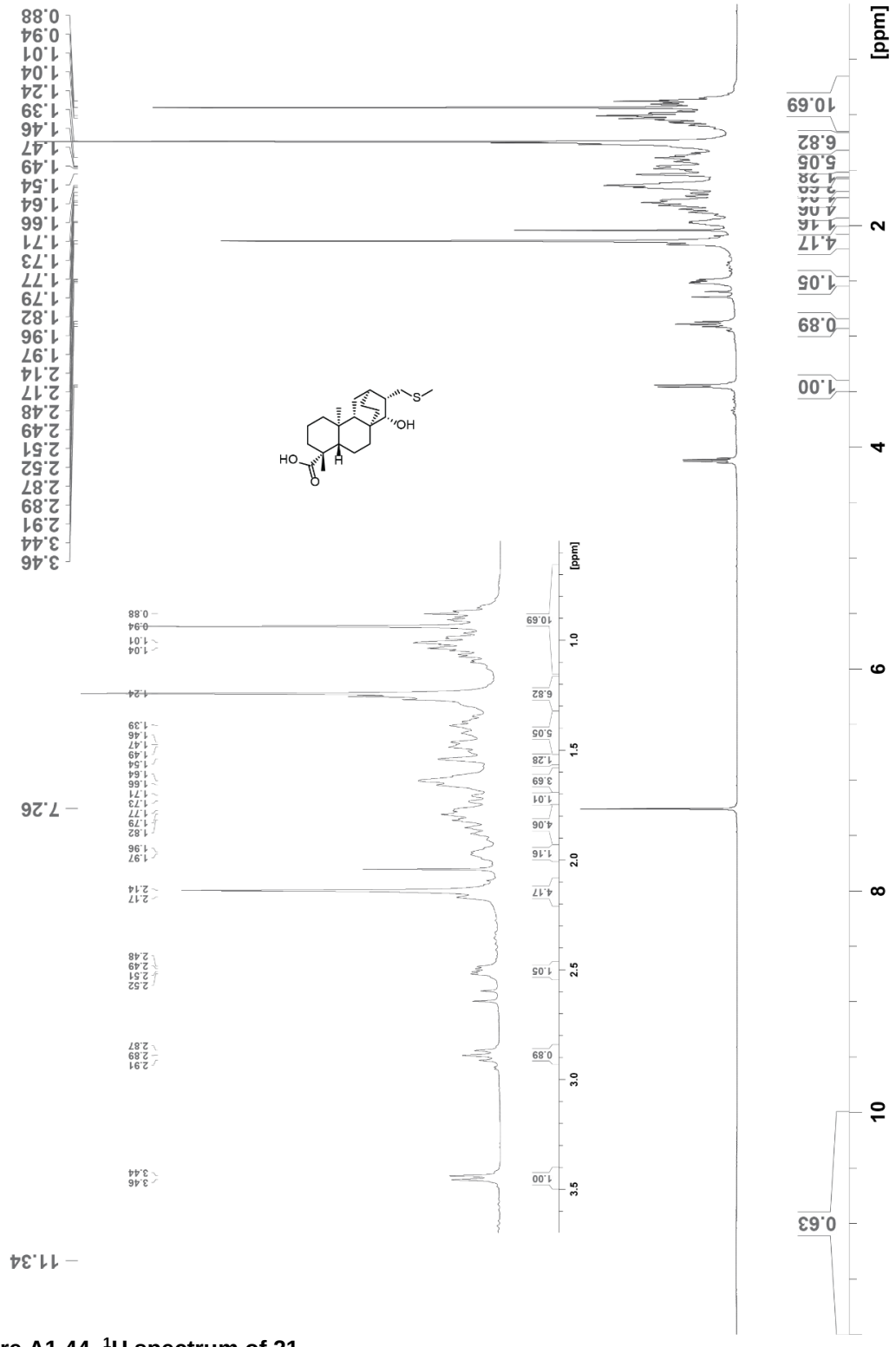


Figure A1.44. ^{1}H spectrum of 21.

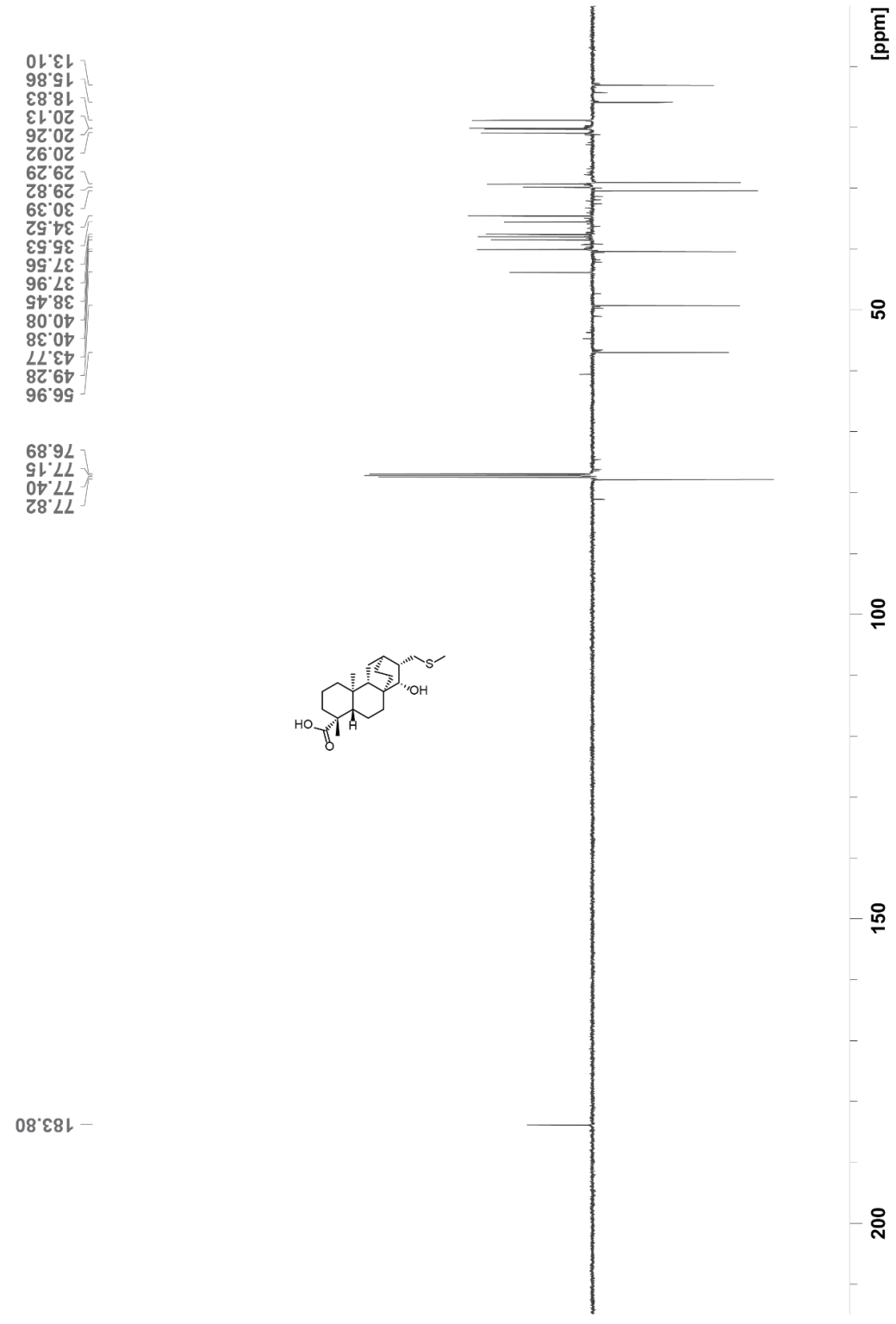
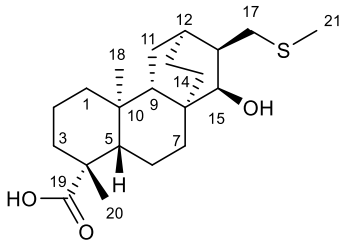


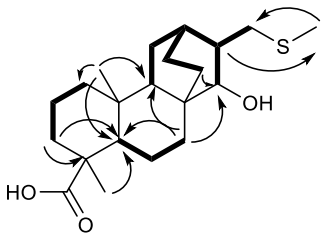
Figure A1.45.^{13}C spectrum 21.



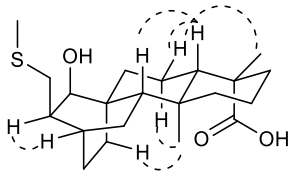
***ent*-(15S,16S)-15-hydroxy-17-methylsulfenylatisan-19-oic acid (22)**

Annotated NMR data for 22 in chloroform-*d.*

No.	δ_H (#H, mult., *J*(Hz))	δ_C
1	0.92 (1H, m)	40.1
	1.58 (1H, m)	
2	1.40 (1H, m)	18.7
	1.80 (1H, m)	
3	1.01 (td, 4.2, 13.5, 17.6)	38.0
	2.14 (1H, m)	
4	-	43.9
5	1.10 (1H, m)	56.7
6	1.70 (1H, m)	20.0
	1.78 (1H, m)	
7	1.11 (1H, m)	34.0
	1.70 (1H, m)	
8	-	35.8
9	1.63 (1H, m)	41.7
10	-	38.2
11	1.1 (1H, m)	22.5
	1.63 (1H, m)	
12	1. 51 (1H, m)	31.3
13	1.47 (2H, m)	27.8
14	0.89 (1H, m)	26.9
	1.95 (1H, m)	
15	3.46 (1H, d, 8.8)	74.6
16	1.97 (1H, m)	40.1
17	2.51 (1H, dd, 4.8, 12.1)	34.9
	2.87 (t, 11.8)	
18	0.90 (3H, s)	12.7
19	-	184.4
20	1.23 (3H, s)	29.0
21	2.15 (3H, s)	15.8



Key COSY (bold) and HMBC (arrows) correlations



Key NOESY correlations

Figure A1.46. Summary of key 1D and 2D NMR data for compound 22.

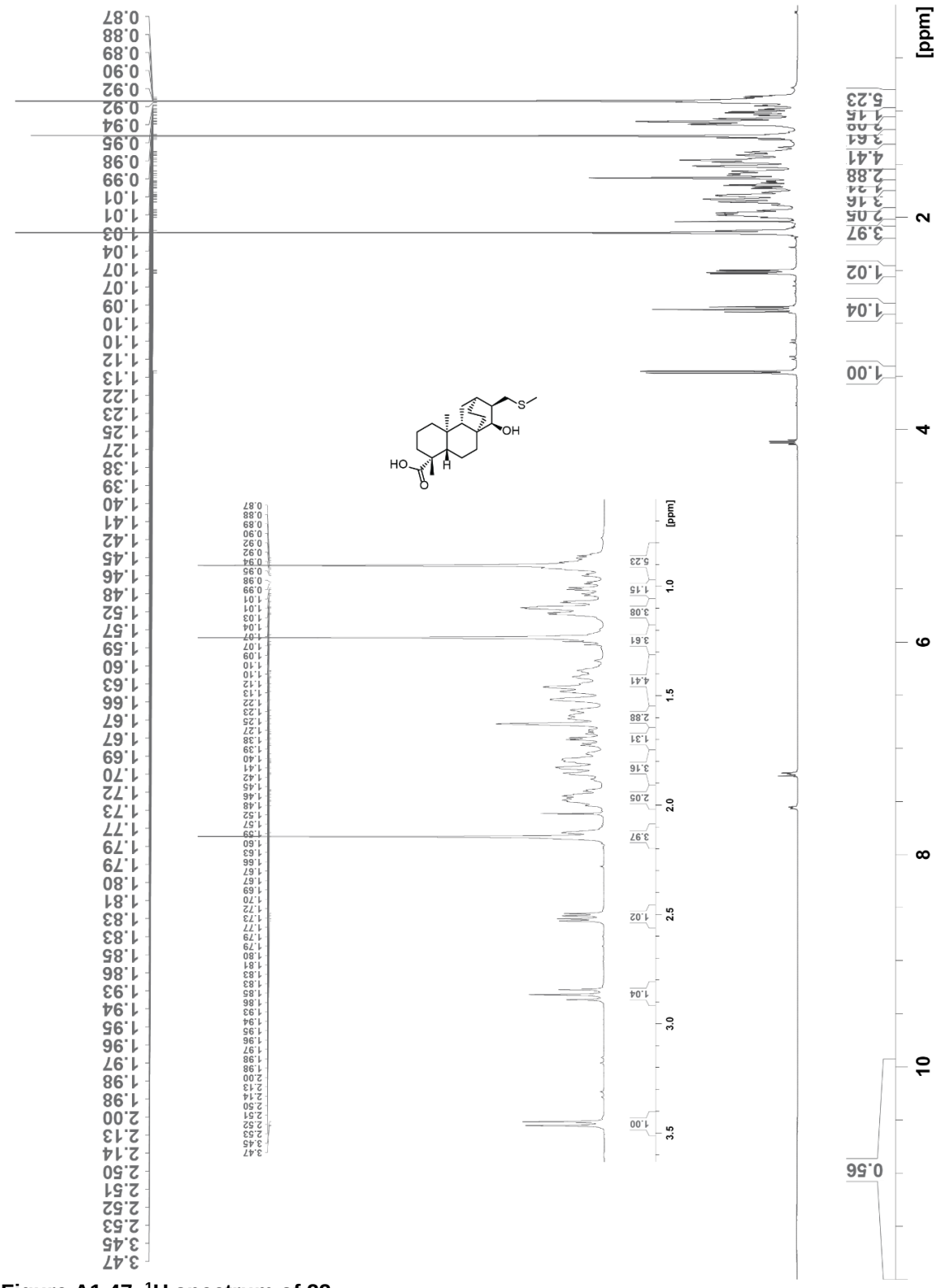


Figure A1.47. ^{1}H spectrum of 22.

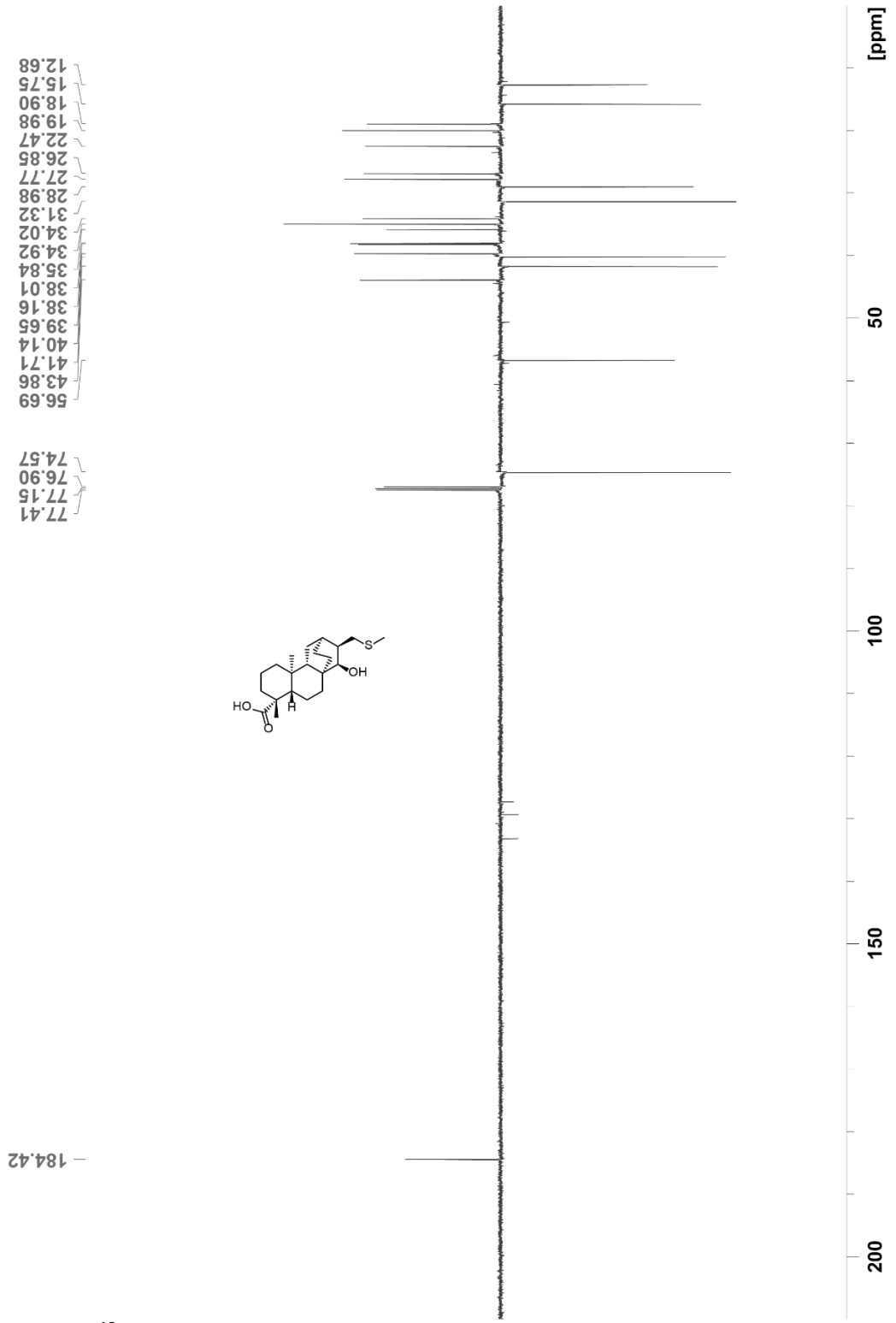
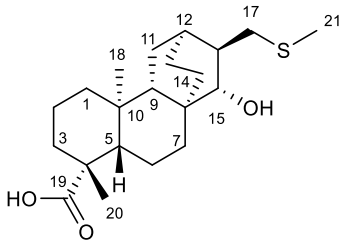


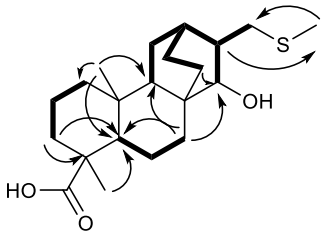
Figure A1.48.^{13}C spectrum of 22.



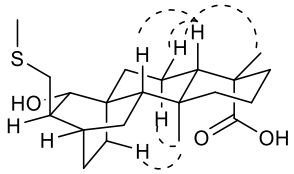
***ent*-(15R,16S)-15-hydroxy-17-methylsulfenylatisan-19-oic acid (23)**

Annotated NMR data for 23 in chloroform-*d.*

No.	δ_H (#H, mult., *J*(Hz))	δ_C
1	0.84 (1H, m)	39.9
	1.62 (1H, m)	
2	1.41 (1H, m)	18.8
	1.87 (1H, m)	
3	1.01 (1H, m)	38.0
	2.16 (1H, m)	
4	-	43.8
5	1.02 (1H, m)	57.1
6	1.77 (2H, m)	20.0
7	1.03 (1H, m)	35.5
	1.70 (1H, m)	
8	-	36.2
9	0.86 (1H, m)	50.8
10	-	38.5
11	1.16 (1H, dd, 6.4,13.9)	22.8
	1.51 (1H, m)	
12	1.65 (1H, m)	29.3
13	1.47 (1H, m)	20.2
	1.66 (1H, m)	
14	1.47 (1H, m)	27.6
	1.58 (1H, m)	
15	2.92 (1H, dd, 1.7, 4.4)	84.4
16	1.61 (1H, m)	46.7
17	2.53 (2H, m)	39.1
18	0.92 (3H, s)	12.5
19	-	183.9
20	1.24 (3H, s)	29.0
21	2.13 (3H, s)	15.7



Key COSY (bold) and HMBC (arrows) correlations



Key NOESY correlations

Figure A1.49. Summary of key 1D and 2D NMR data for compound 22.

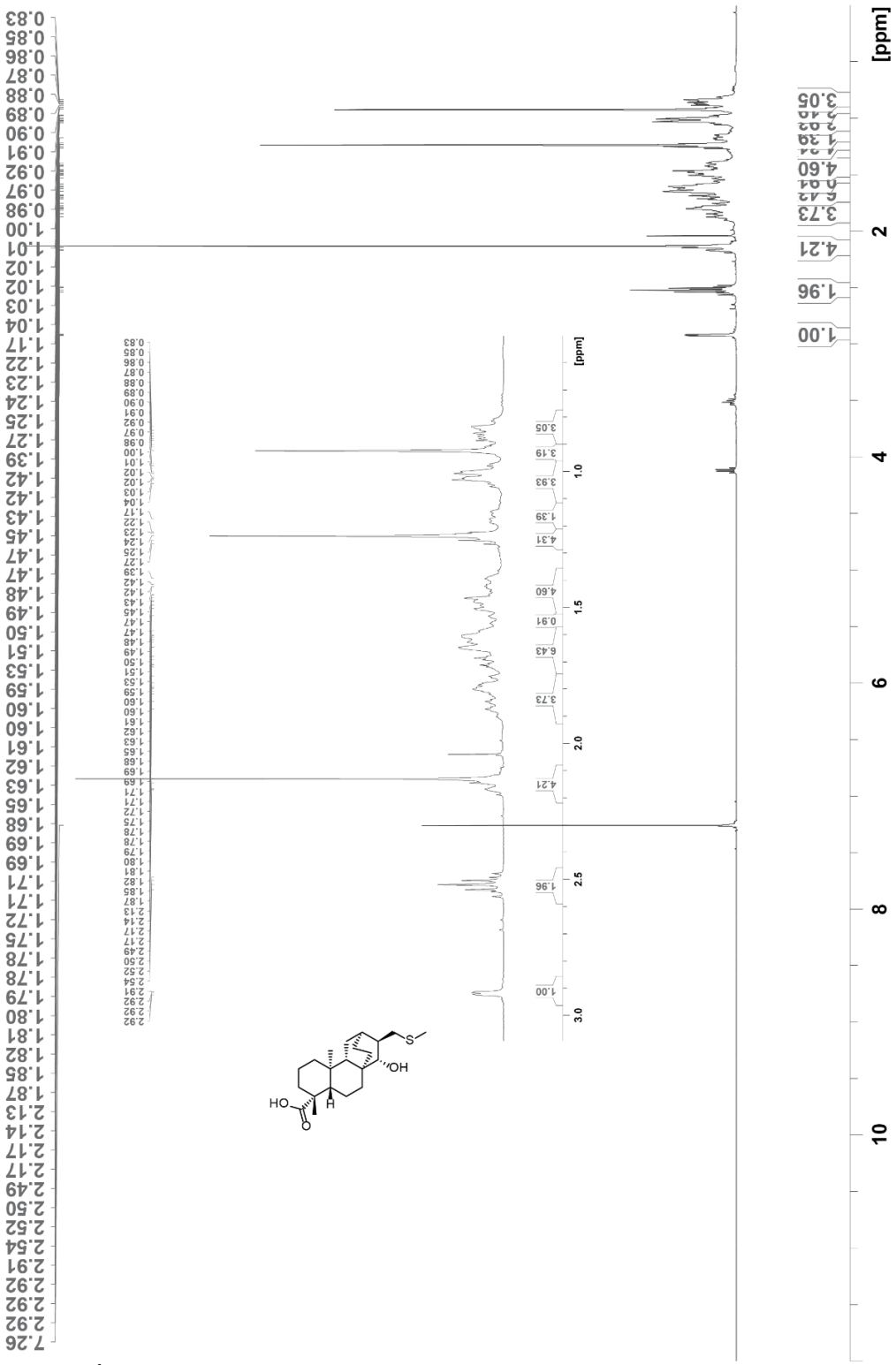


Figure A1.50. ^{1}H spectrum of 23.

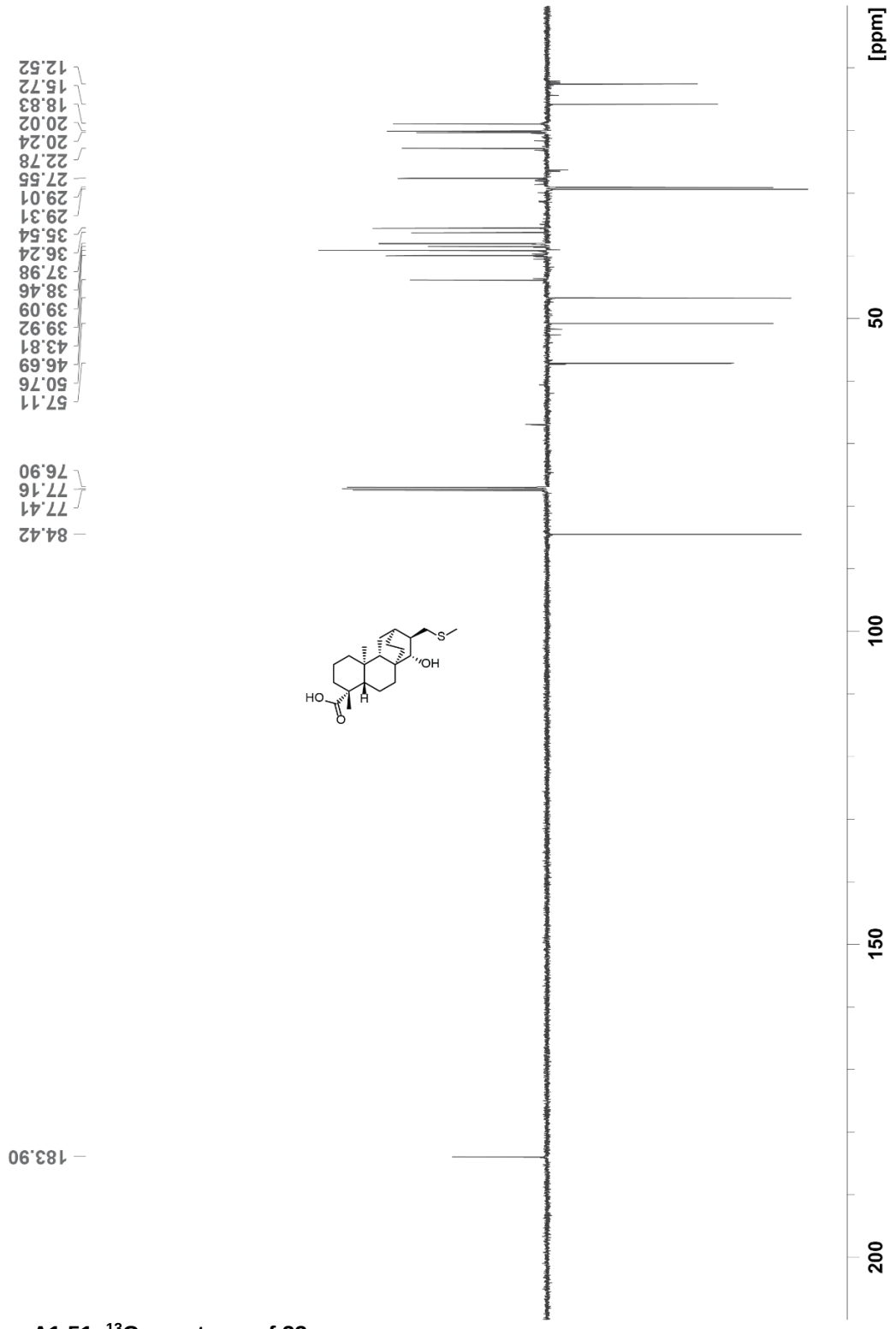
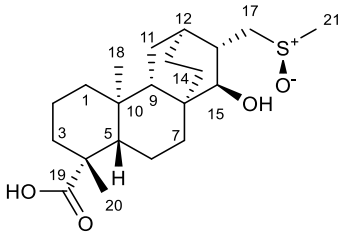


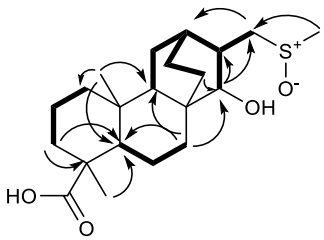
Figure A1.51. ^{13}C spectrum of 23.



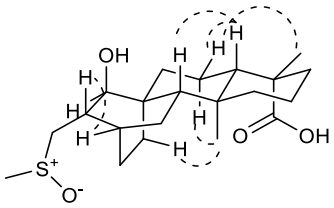
Serofendic acid A (1A)

Annotated. NMR data for 1A in methanol-*d*$_4$.

No.	δ_H (#H, mult., *J*(Hz))	δ_C
1	0.88 (1H, m)	41.2
	1.62 (1H, m)	
2	1.38 (1H, m)	20.0
	1.89 (1H, m)	
3	1.03 (1H, m)	39.2
	2.13 (1H, d, 13.0)	
4	-	44.7
5	1.04 (1H, m)	57.9
6	1.79 (1H, m)	20.9
	1.86 (1H, m)	
7	1.07 (1H, m)	34.3
	1.71 (1H, m)	
8	-	37.7
9	1.57 (1H, m)	42.8
10	-	39.2
11	1.41 (1H, m)	30.0
	1.59 (1H, m)	
12	1.77 (1H, m)	32.2
13	1.38 (1H, m)	22.2
	1.64 (1H, m)	
14	0.83 (1H, m)	28.2
	2.02 (1H, ddd, 2.8, 11.8, 14.1)	
15	2.93 (1H, d, 4.4)	81.3
16	1.92 (1H, m)	44.4
17	2.83 (1H, dd, 6.1, 13.1)	60.9
	2.95 (1H, dd, 9.3, 13.1)	
18	0.95 (3H, s)	13.4
19	-	181.7
20	1.20 (3H, s)	29.6
21	2.68 (3H, s)	38.7



Key COSY (bold) and HMBC (arrows) correlations



Key NOESY correlations

Figure A1.52. Summary of key 1D and 2D NMR data for compound 1A.

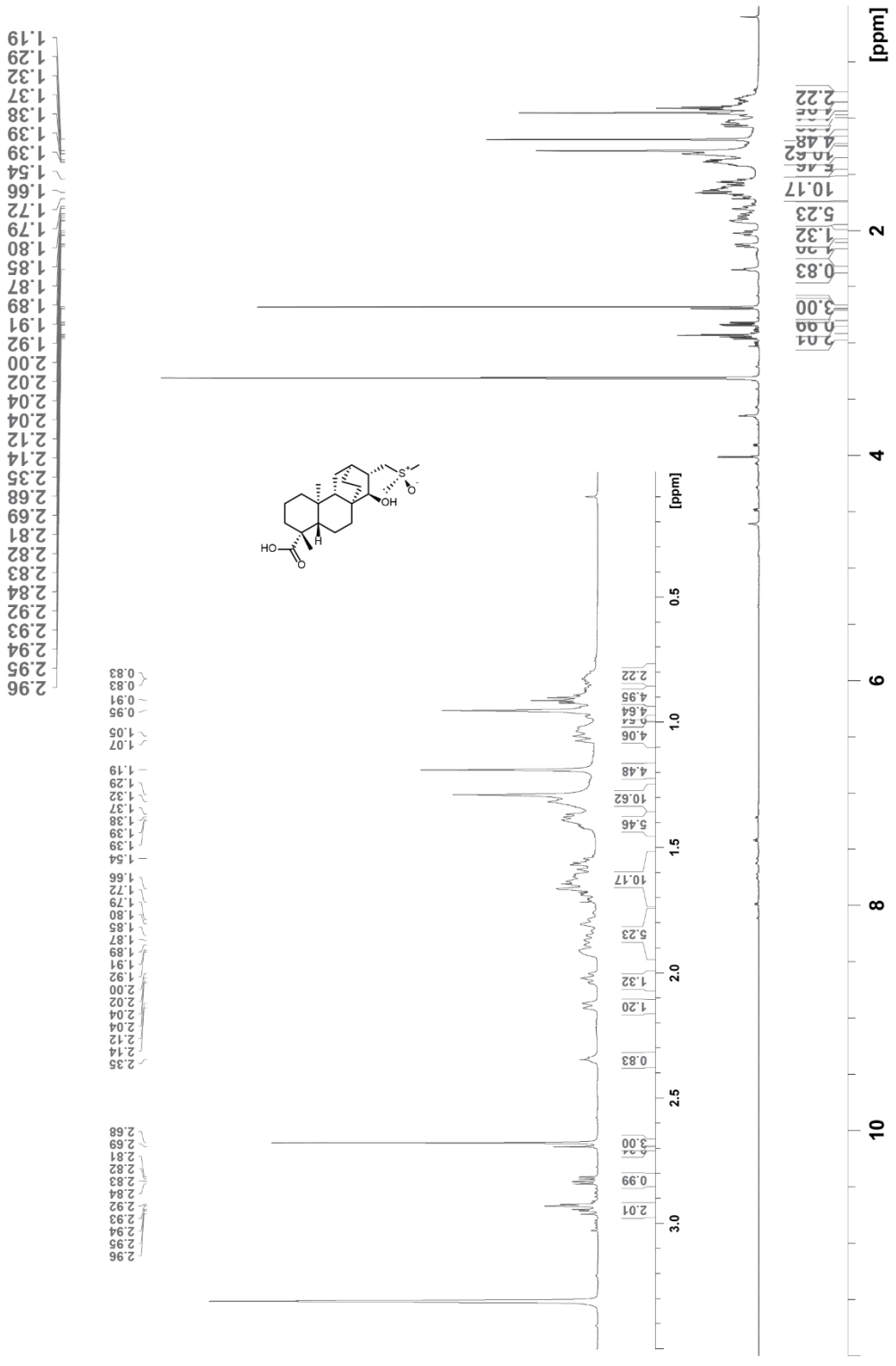


Figure A1.53. ^{1}H spectrum of 1A.

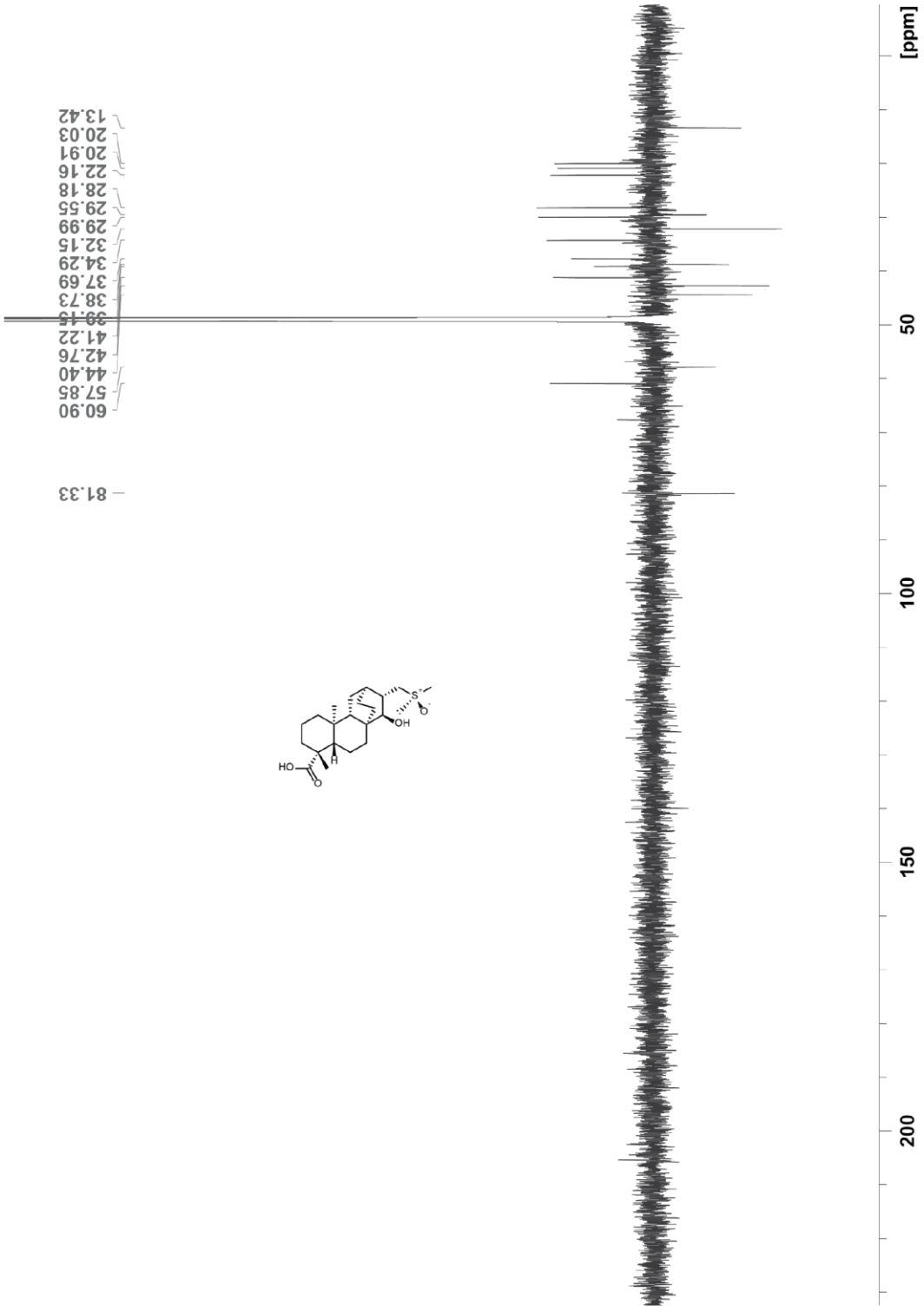


Figure A1.54. ^{13}C spectrum of 1A.

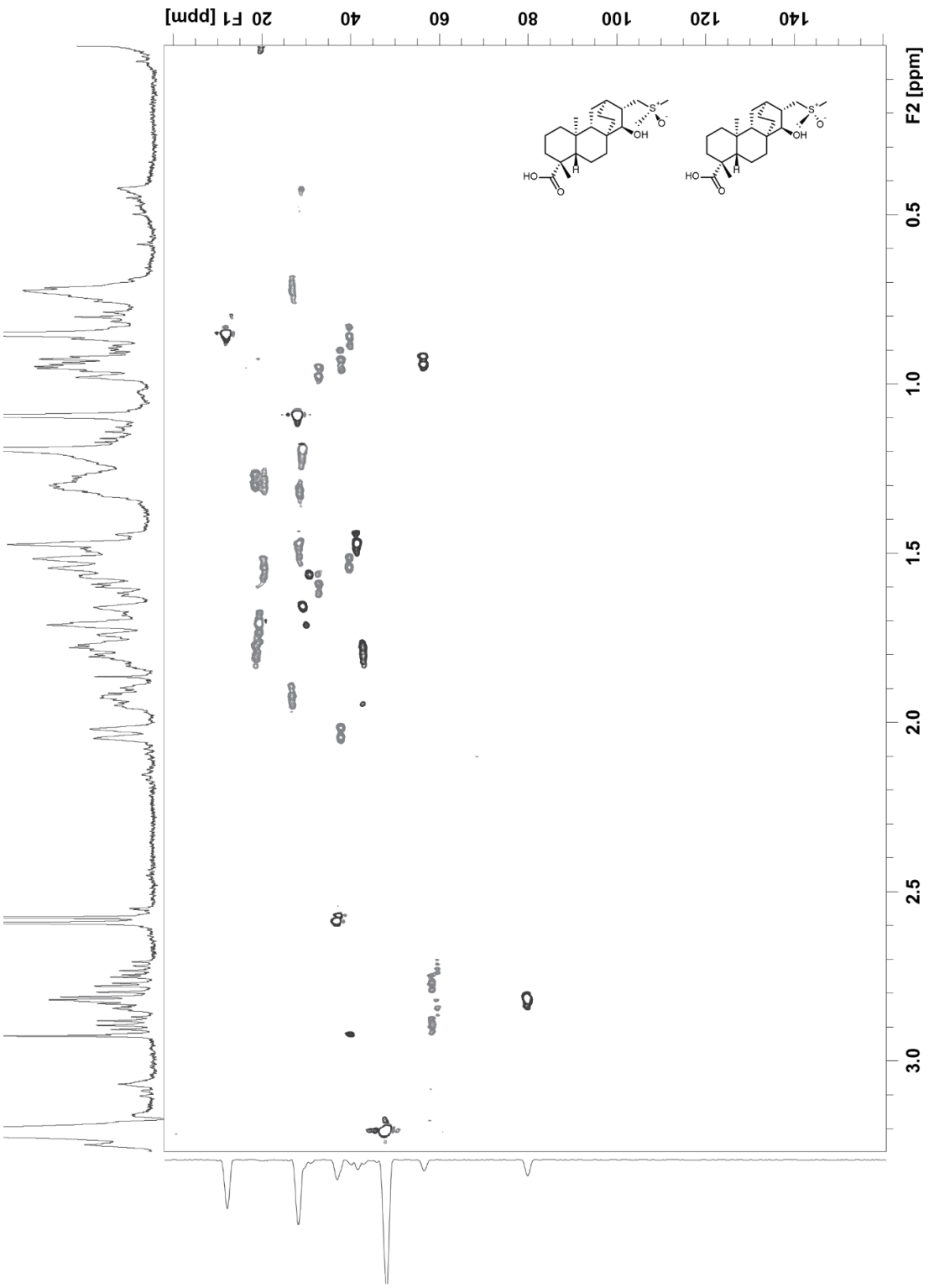


Figure A1.55. HSQC ^{1}H-^{13}C spectrum of 1A and 1B.

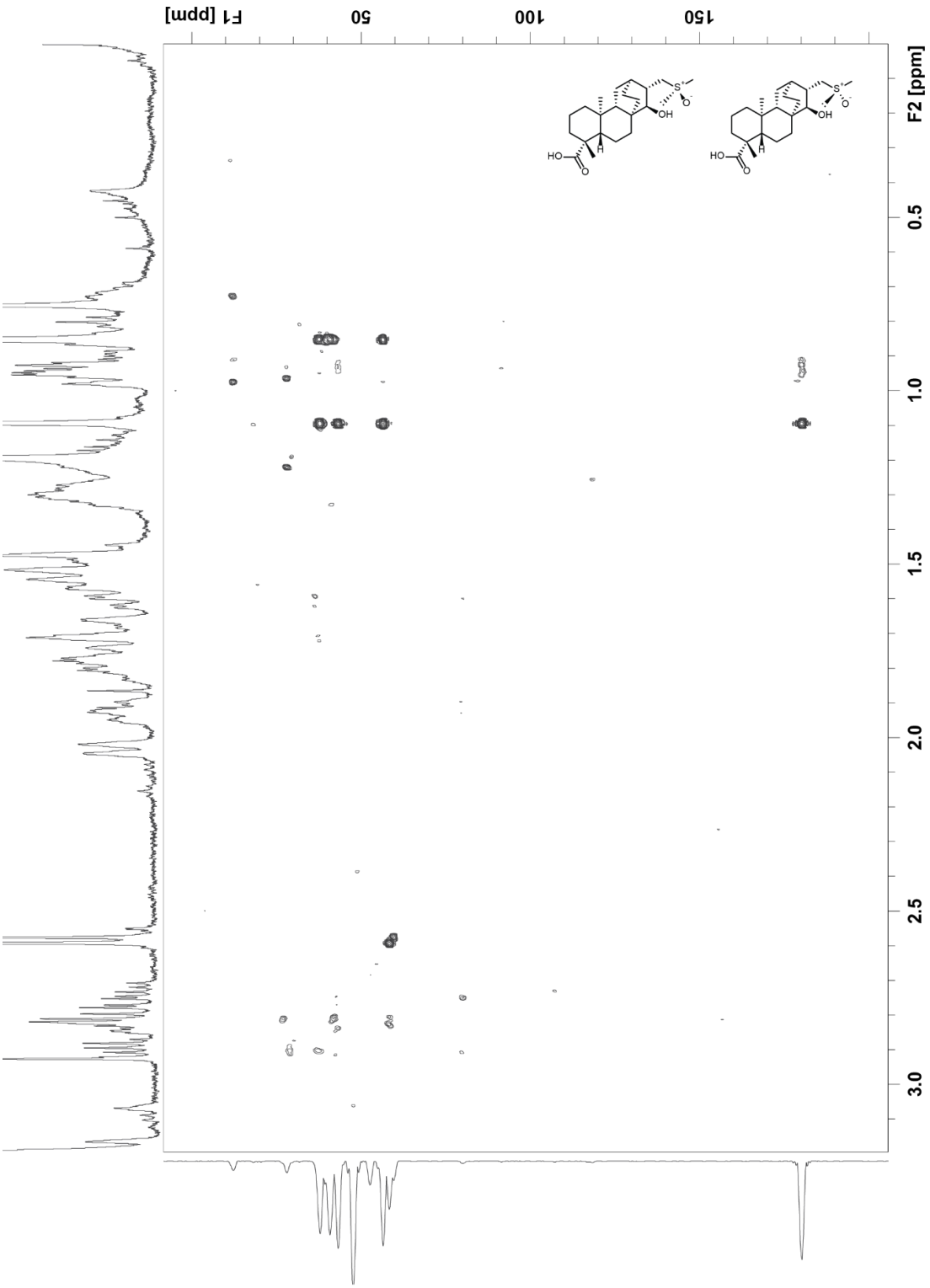
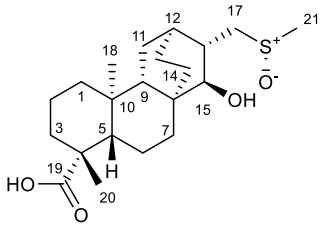


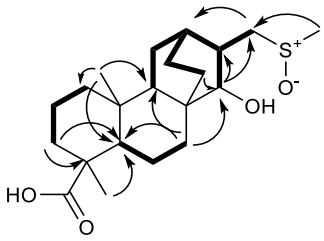
Figure A1.56. HMBC ^{1}H-^{13}C spectrum of 1A and 1B.



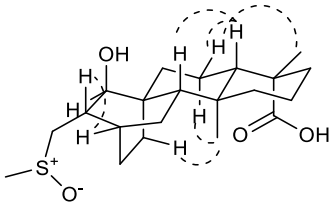
Serofendic acid B (1B)

Annotated NMR data for 1B in methanol-*d*$_4$.

No.	δ_H (#H, mult., *J*(Hz))	δ_C
1	0.96 (1H, m)	41.3
	1.62 (1H, m)	
2	1.38 (1H, m)	20.1
	1.92 (1H, m)	
3	1.04 (1H, m)	39.4
	2.13 (1H, br. d, 13.2)	
4	-	44.3
5	1.04 (1H, m)	57.9
6	1.79 (1H, m)	21.0
	1.86 (1H, m)	
7	1.06 (1H, m)	34.3
	1.69 (1H, m)	
8	-	37.7
9	1.58 (1H, m)	42.9
10	-	39.2
11	1.42 (1H, m)	29.9
	1.59 (1H, m)	
12	1.75 (1H, m)	30.8
13	1.40 (1H, m)	21.9
	1.62 (1H, m)	
14	0.82 (1H, ddd, 6.2, 12.2, 14.0)	28.4
	2.04 (1H, ddd, 3.0, 11.7, 14.0)	
15	2.91 (1H, d, 4.4)	81.2
16	1.88 (1H, m)	44.3
17	2.88 (1H, dd, 9.3, 13.1)	59.8
	3.00 (1H, dd, 6.8,13.1)	
18	0.95 (3H, s)	13.5
19	-	181.0
20	1.19 (3H, s)	29.6
21	2.69 (3H, s)	38.3

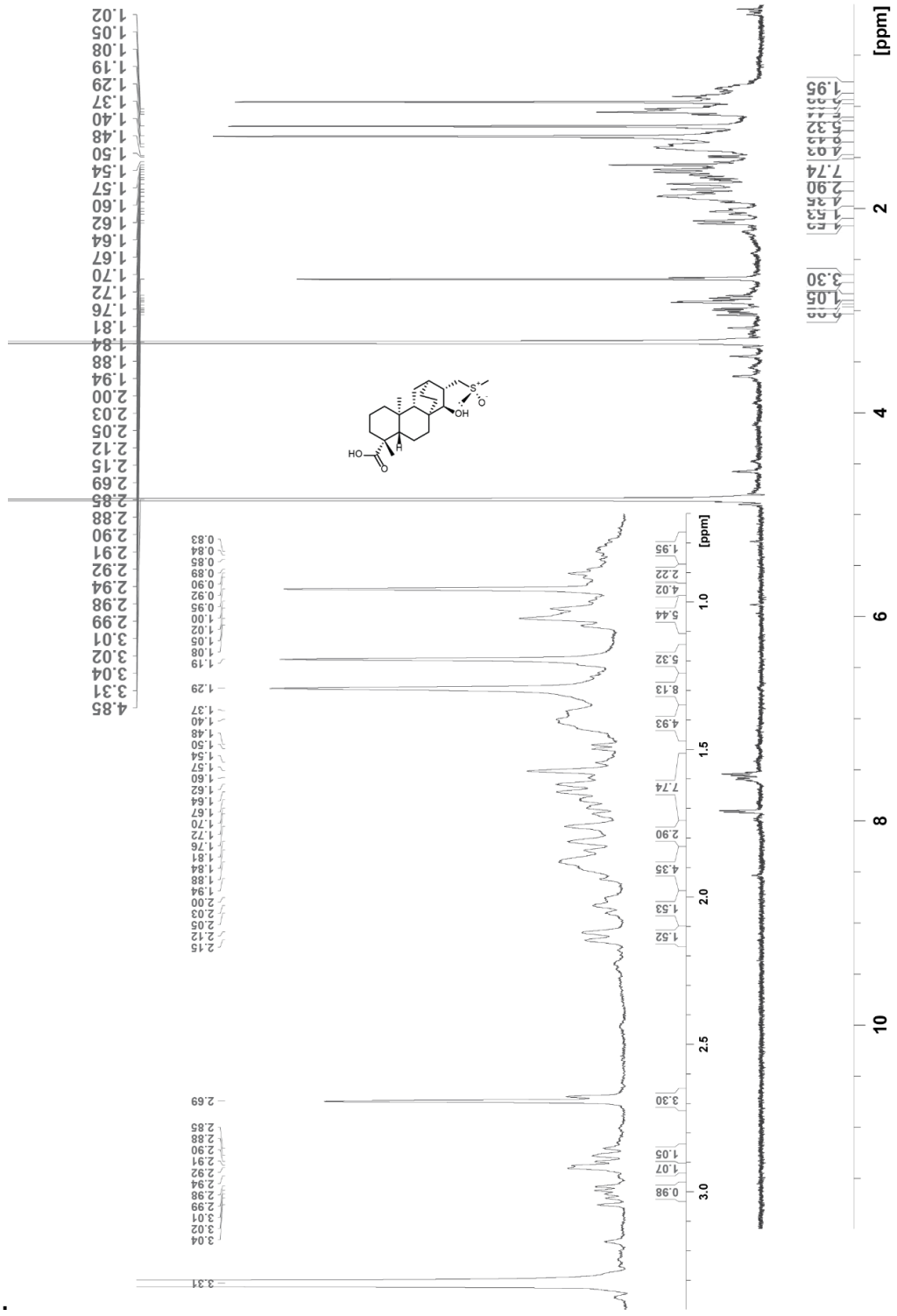


Key COSY (bold) and HMBC (arrows) correlations



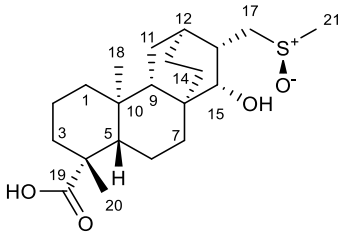
Key NOESY correlations

Figure A1.57. Summary of key 1D and 2D NMR data for compound 1B



.

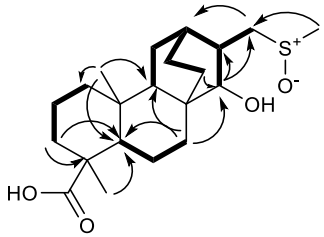
Figure A1.58. ^{1}H spectrum of 1B.



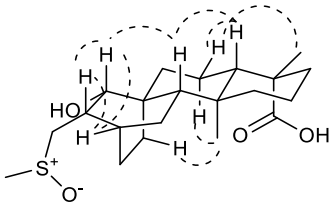
(15S,16S)-*ent*-15-hydroxy-17-methanesulfinylatisan-19-oic acid (24A)

Annotated NMR data for 24A in chloroform-*d*.

No.	δ_H (#H, mult., *J*(Hz))	δ_C
1	0.95 (1H, m)	41.3
	1.67 (1H, m)	
2	1.39 (1H, m)	20.0
	1.91 (1H, m)	
3	1.05 (1H, m)	39.1
	2.13 (1H, d, 13.1)	
4	-	44.6
5	1.05 (1H, m)	58.0
6	1.74-1.78 (2H, m)	21.3
7	1.08 (1H, m)	38.5
	1.75 (1H, m)	
8	-	36.7
9	1.08 (1H, m)	50.7
10	-	39.5
11	1.46 (1H, m)	29.9
	1.57 (1H, m)	
12	1.69 (1H, m)	31.1
13	1.34 (1H, m)	21.6
	1.68 (1H, m)	
14	1.49 (1H, m)	21.0
	1.66 (1H, m)	
15	3.36 (1H, d, 9.2)	78.1
16	2.33 (1H, m)	38.9
17	2.78 (1H, dd, 7.6, 13.2)	55.3
	3.29 (1H, dd, 7.6, 13.2)	
18	0.98 (3H, s)	13.6
19	-	181.6
20	1.20 (3H, s)	29.4
21	2.67 (3H, s)	38.5



Key COSY (bold) and HMBC (arrows) correlations



Key NOESY correlations

Figure A1.59. Summary of key 1D and 2D NMR data for compound 24A.

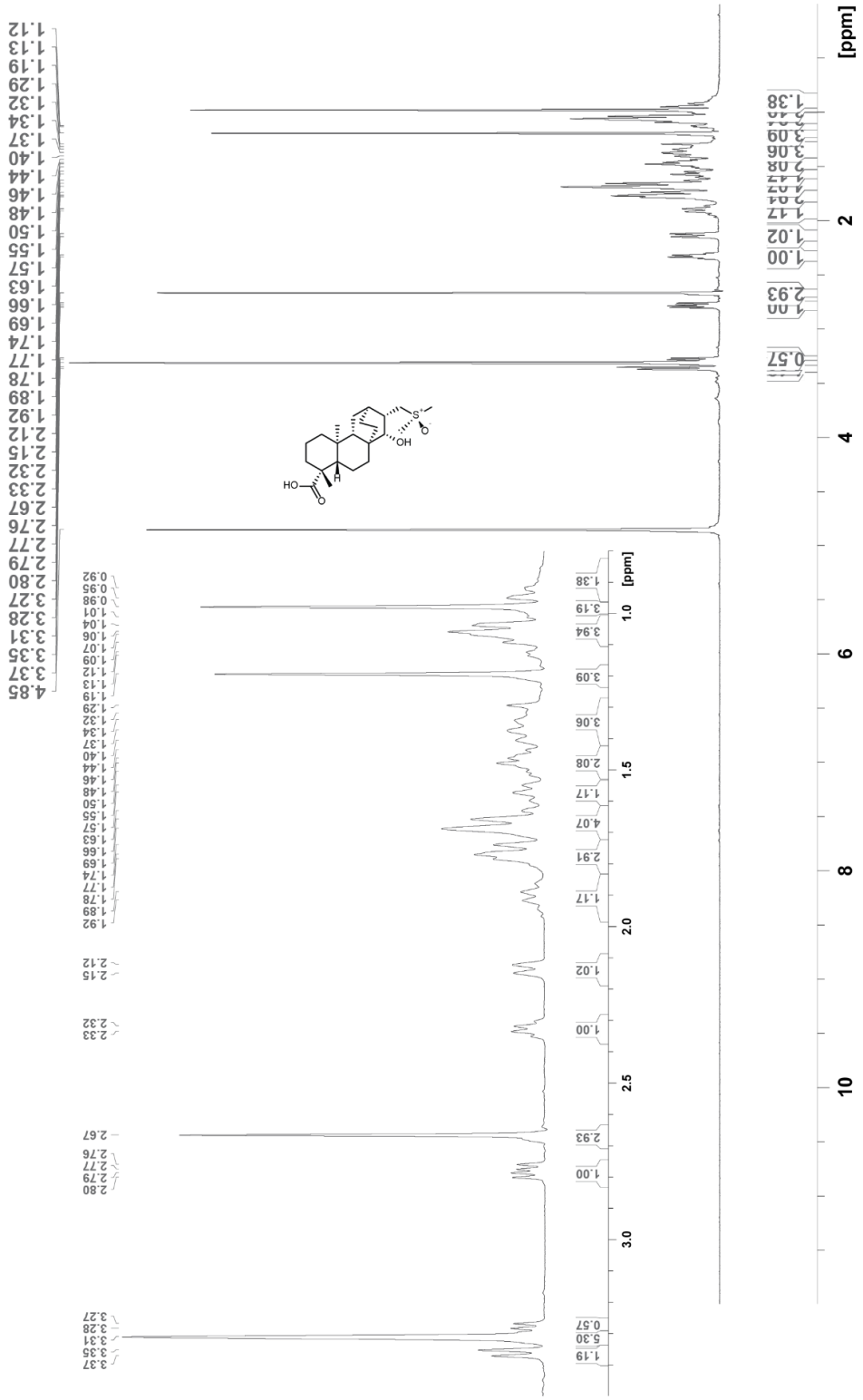


Figure A1.60. ^{1}H spectrum of 24A.

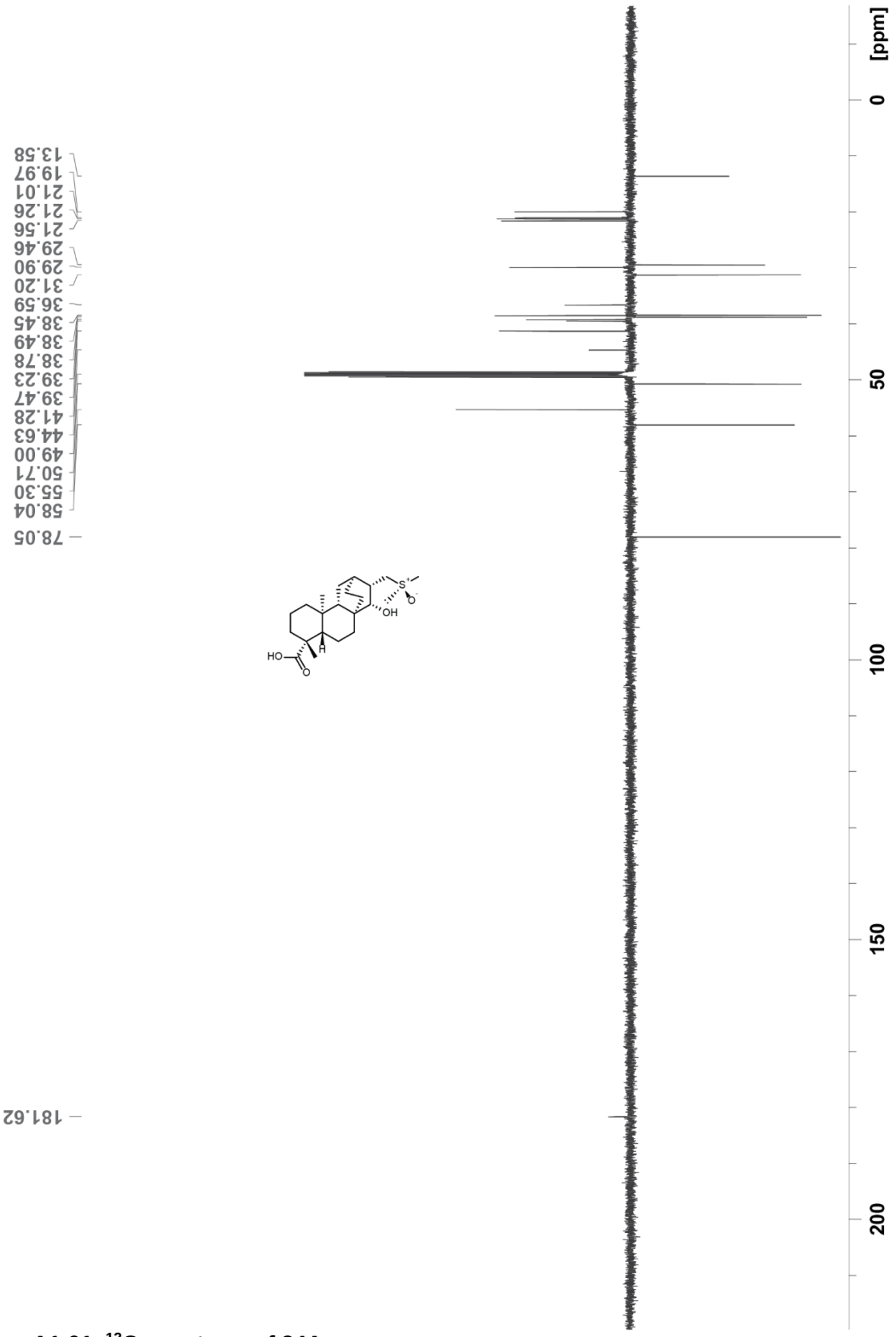
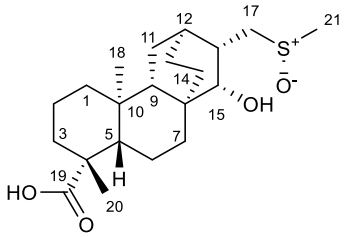


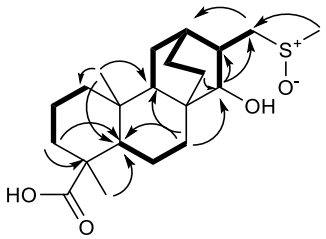
Figure A1.61. ^{13}C spectrum of 24A.



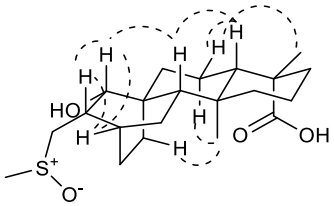
(15S,16S)-*ent*-15-hydroxy-17-methanesulfinylatisan-19-oic acid (24B)

Annotated. NMR data for 24B in methanol-*d*$_4$

No.	δ_H (#H, mult., *J*(Hz))	δ_C
1	0.93 (1H, m)	41.5
	1.67 (1H, m)	
2	1.38 (1H, m)	20.2
	1.94 (1H, m)	
3	1.01 (1H, m)	39.2
	2.14 (1H, br. d, 13.2)	
4	-	45.0
5	1.00 (1H, m)	58.3
6	1.80 (2H, m)	21.4
7	1.03 (1H, m)	38.5
	1.71 (1H, m)	
8	-	36.8
9	1.06 (1H, m)	50.7
10	-	39.5
11	1.45 (1H, m)	30.1
	1.58 (1H, m)	
12	1.56 (1H, m)	32.7
13	1.34 (1H, m)	21.9
	1.67 (1H, m)	
14	1.48 (1H, m)	21.0
	1.67 (1H, m)	
15	3.39 (1H, m)	77.4
16	2.42 (1H, m)	37.8
17	2.54 (1H, dd, 4.0, 12.7)	57.1
	3.40 (1H, m)	
18	0.99 (3H, s)	13.7
19	-	182.5
20	1.17 (3H, s)	29.7
21	2.65 (3H, s)	38.7



Key COSY (bold) and HMBC (arrows) correlations



Key NOESY correlations

Figure A1.62. Summary of key 1D and 2D NMR data for compound 24B.

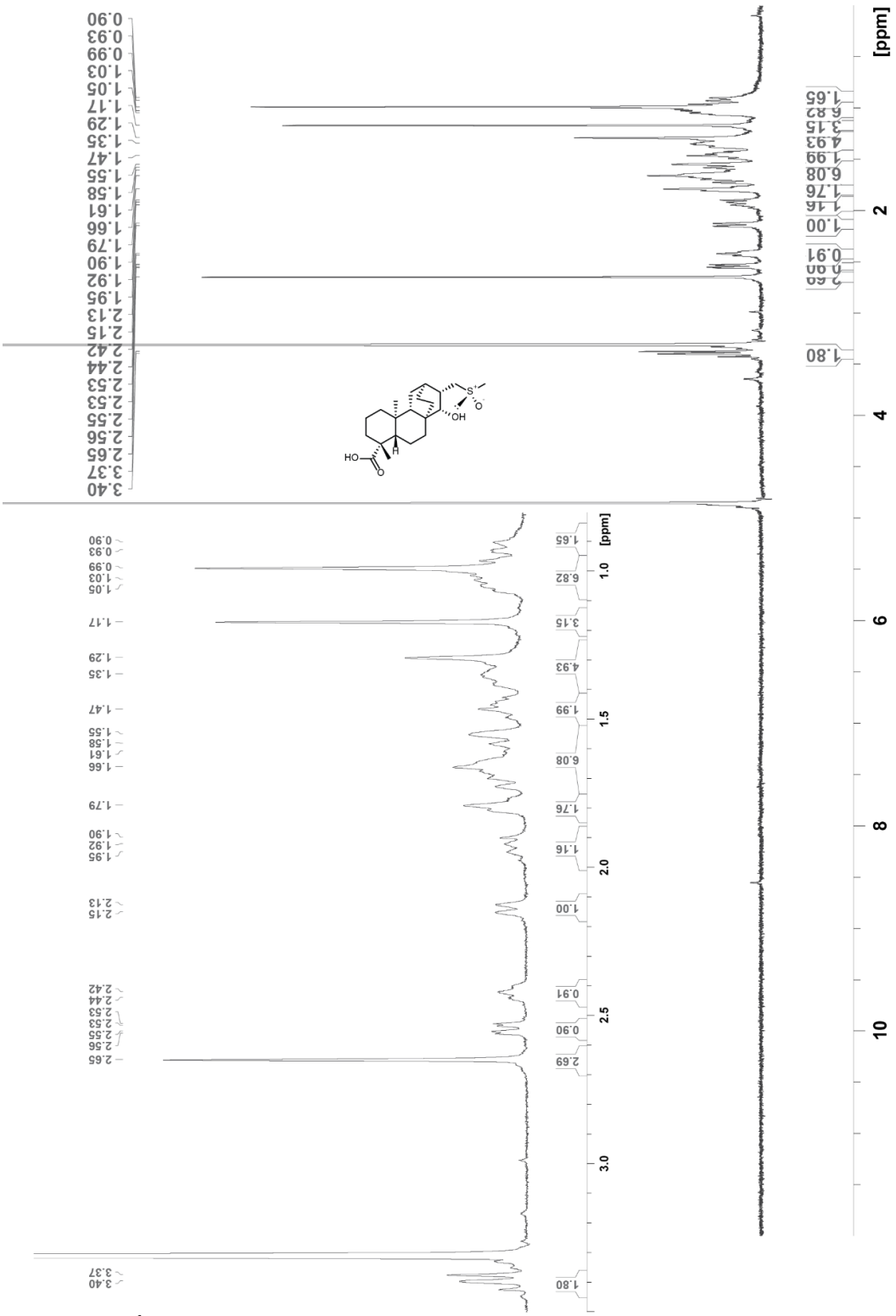


Figure A1.63. ^{1}H spectrum of 24B.

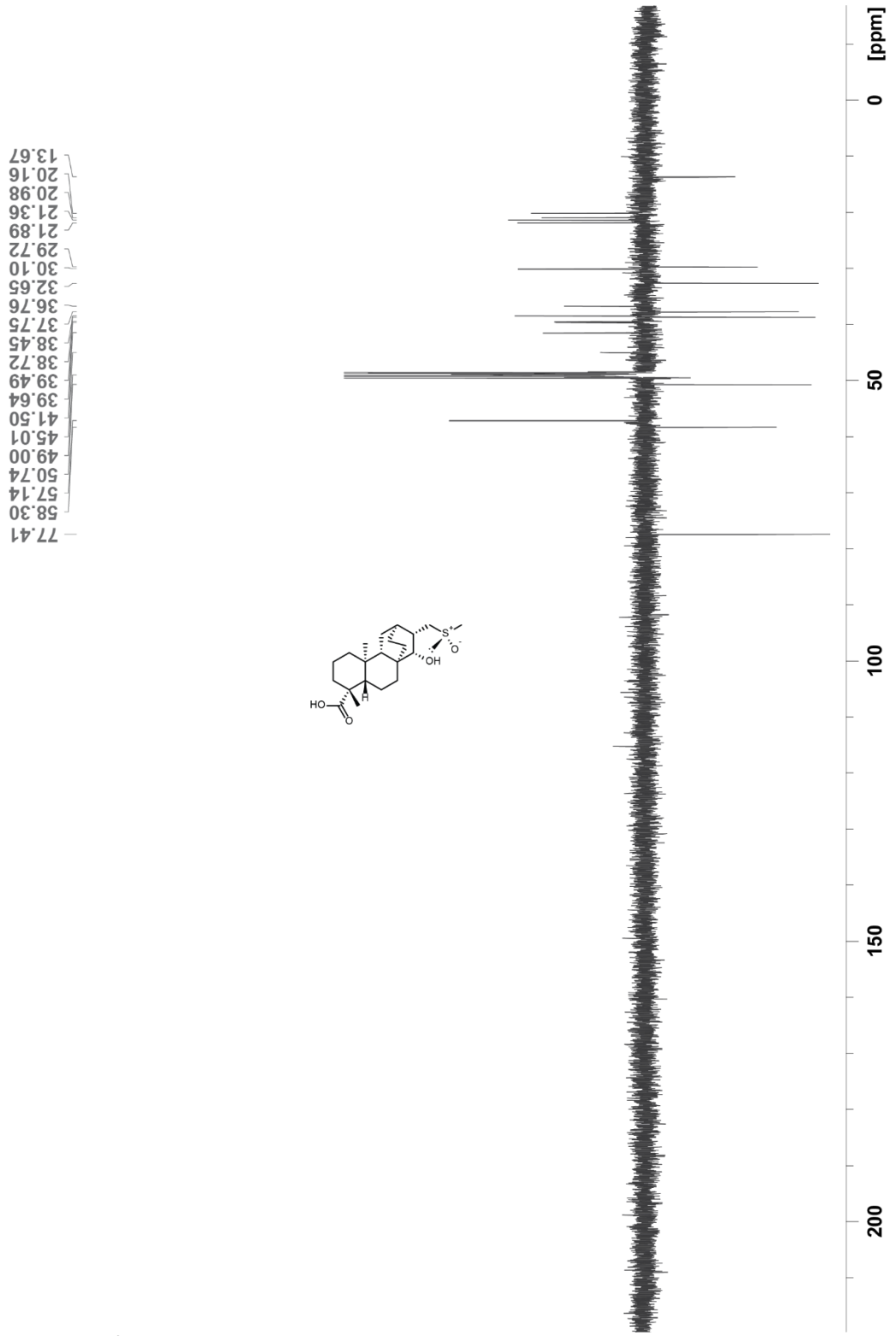


Figure A1.64. ^{13}C spectrum of 24B.

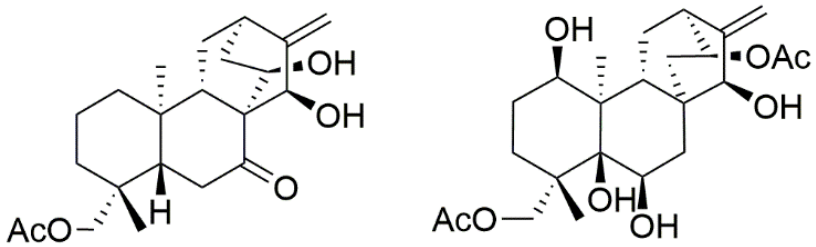


Figure A1.65. Chemical structures of C15-hydroxylated atisane compounds.

They are Eriocatisin A (left) from *Isodon eriocalyx*, and *ent*-(1S,6S,15S)-1,5,6,15-tetrahydroxy-14,19-diacetyloxylatis-16-ene (right) from New Zealand liverwort.

Table A1.1. Name, nucleotide length, and predicted function of the genes used in building SGCs.

Gene	***NTs***	***Predicted Function***	***Accession Number***
ispH	1347	4-hydroxy-3-methylbut-2-enyl diphosphate reductase (MEP pathway)	WP_078889659
ispG	1101	4-hydroxy-3-methylbut-2-en-1-yl diphosphate synthase (MEP pathway)	WP_030990036
dxs	1764	1-deoxy-D-xylulose-5-phosphate synthase (MEP pathway)	WP_037839917
eAS	927	*ent*-atiserene synthase	WP_030985278
eCPDPS	1602	*ent*-copalyl diphosphate synthase	WP_037839300
GGDPS	1047	geranylgeranyl diphosphate synthase	WP_078889660
eAOx	1326	cytochrome P450	WP_030985276
αKG-DDO	843	α-ketoglutarate dependent dioxygenase	WP_078889904
eAOxFD	336	ferredoxin	WP_030990007
dxr	1275	1-deoxy-D-xylulose-5-phosphate reductoisomerase (MEP pathway)	WP_030991996
ispD	749	2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase (MEP pathway)	WP_030992901
ispE	907	4-(cytidine 5'-diphospho)-2-C-methyl-D-erythritol kinase (MEP pathway)	WP_030985411
ispF	528	2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase (MEP pathway)	WP_078889846
idi	612	isopentenyl-diphosphate delta-isomerase (MEP pathway)	WP_030078955

Table A1.2. Crystal data and structure refinement of 2.

Identification code	17104c	
Empirical formula	$C_{20}H_{30}O_2$	
Formula weight	302.44	
Temperature	123(2) K	
Wavelength	0.71073 Å	
Crystal system	Orthorhombic	
Space group	$P2_12_12_1$	
Unit cell dimensions	*a* = 11.6388(11) Å	α = 90°
	b = 13.4757(16) Å	β = 90°
	c = 21.856(2) Å	γ = 90°
Volume	3427.9(6) $Å^3$	
Z	8	
Density (calculated)	1.172 Mg/m^3	
Absorption coefficient	0.073 mm^{-1}	
F(000)	1328	
Crystal color, morphology	Colourless, Plate	
Crystal size	0.230 x 0.120 x 0.030 mm^3	
Theta range for data collection	2.312 to 28.275°	
Index ranges	$-15 \le h \le 10$, $-17 \le k \le 13$, $-29 \le l \le 28$	
Reflections collected	23054	
Independent reflections	8437 [*R*(int) = 0.0436]	
Observed reflections	6753	
Completeness to theta = 25.242°	99.8%	
Absorption correction	Multi-scan	
Max. and min. transmission	0.7457 and 0.6872	
Refinement method	Full-matrix least-squares on F^2	
Data / restraints / parameters	8437 / 0 / 407	
Goodness-of-fit on F^2	1.029	
Final *R* indices [*I*>2sigma(*I*)]	*R*1 = 0.0468, *wR*2 = 0.1037	

R indices (all data)	*R*1 = 0.0679, *wR*2 = 0.1177
Absolute structure parameter	-0.9(6)
Largest diff. peak and hole	0.373 and -0.230 e.Å^{-3}

Captions of Supplementary Data File A1. (separate file)

The file titled ‘Supplementary Data A1.xlsx’ contains six worksheets.

Worksheet 1 (Genetic Parts):

Column A: Name of genetic parts
Column B: Sequence of genetic parts
Column C: Sources of genetic parts
Column D: Reference of genetic parts
Column 25-45F: Average value of relative expression level of promoter-RBS parts in 60-hour fermentation in ISM3 medium. Relative expression level is measured as changes in absorbance at 415 nm per minute per grams of dried cell weight. The value is calculated from three biological replicates.
Column 25-45G: Standard deviation of the value of relative expression level of promoter-RBS parts in 60-hr fermentation in ISM3 medium. Relative expression level is measured as changes in absorbance at 415 nm per minute per grams of dried cell weight. The value is calculated from three biological replicates.
Column 25-45H: Average value of relative expression level of promoter-RBS parts in 60-hour fermentation in PCNM medium. Relative expression level is measured as changes in absorbance at 415 nm per minute per grams of dried cell weight. The value is calculated from three biological replicates.
Column 25-45I: Standard deviation of the value of relative expression level of promoter-RBS parts in 60-hour fermentation in PCNM medium. Relative expression level is measured as changes in absorbance at 415 nm per minute per grams of dried cell weight. The value is calculated from three biological replicates.

Worksheet 2 (SGC Compositions):

Column A: Construct name, with ‘SGC’ denoting a synthetic gene cluster variant, ‘M’ a MEP pathway variant, ‘A’ indicating an independent exconjugant A from the same conjugation, and ‘B’ an independent exconjugant B from the same conjugation.
Column B-BH: Parts composition, with ‘P’ denoting a promoter part, ‘R’ an RBS part, ‘T’ a transcriptional terminator part, and ‘scar’ a 4-bp MoClo scar sequence.

Worksheet 3 (SGC Sequences):

Column A: Construct name, with ‘SGC’ denoting a synthetic gene cluster variant, ‘M’ a MEP pathway variant, ‘A’ indicating an independent exconjugant A from the same conjugation, and ‘B’ an independent exconjugant B from the same conjugation.
Column B: Complete nucleotide sequence of a construct

Worksheet 4 (SGC Characterization):

Column A: Construct name, with ‘SGC’ denoting a synthetic gene cluster variant, ‘M’ a MEP pathway variant, ‘A’ indicating an independent exconjugant A from the

same conjugation, and ‘B’ an independent exconjugant B from the same conjugation.
Column B: Media type
Column C: Fermentation volume (mL)
Column D-I: biological replicates of production titer of *ent*-atiserenoic acid (*e*AA). Unit: mg/mL
Column J: Average production titer of of *ent*-atiserenoic acid (*e*AA). Unit: mg/mL
Column J: Standard deviation of production titer of of *ent*-atiserenoic acid (*e*AA). Unit: mg/mL

Worksheet 5 (*e*As Phylogenetic Analysis):

Column A: Accession number of protein sequence used to construct phylogenetic analysis of *ent*-atiserene synthase (*e*As)

Worksheet 6 (*e*CDPS Phylogenetic Analysis):

Column A: Accession number of protein sequence used to construct phylogenetic analysis of *ent*-copalyl diphosphate synthase (*e*CPDS)

Appendix 2

Supplementary Materials for

Chapter 4 Rational search of genetic design space to improve terpene production in *Streptomyces*

This appendix includes:

Figure A2.1 to A2.3

Table A2.1

Captions for Data File A2

Other Supplementary Materials for this manuscript include the following:

Data File A2

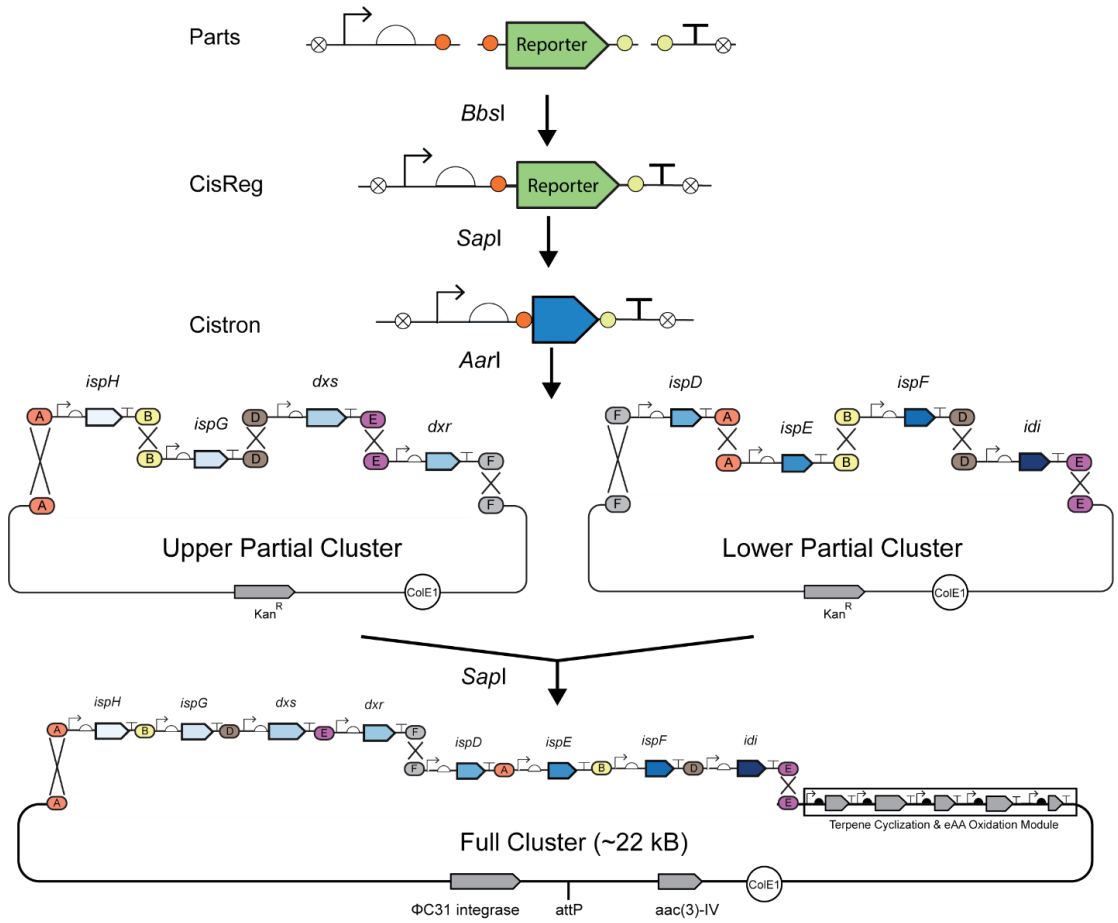


Figure A2.1. Schematic diagram of hierarchical DNA assembly used in this study.

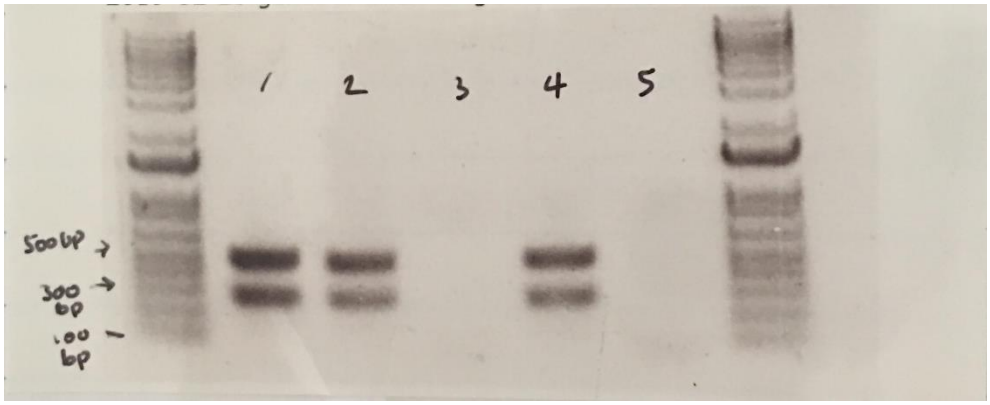


Figure A2.2. DNA gel electrophoresis of PCR products from colony PCR experiment.

Template DNA used in each lane are: purified plasmid of PBFC81 (lane 1; positive control), purified plasmid of PBFC21 (lane 2; positive control), purified plasmid of CisReg-DE-P12R10-GUS-T7 (lane 3; negative control), genomic DNA from a single *E. coli* NEBstable transformant of PBFC21 (lane 4) and water only control (lane 5). The DNA ladder is the 1kb+ DNA ladder (New England BioLabs).

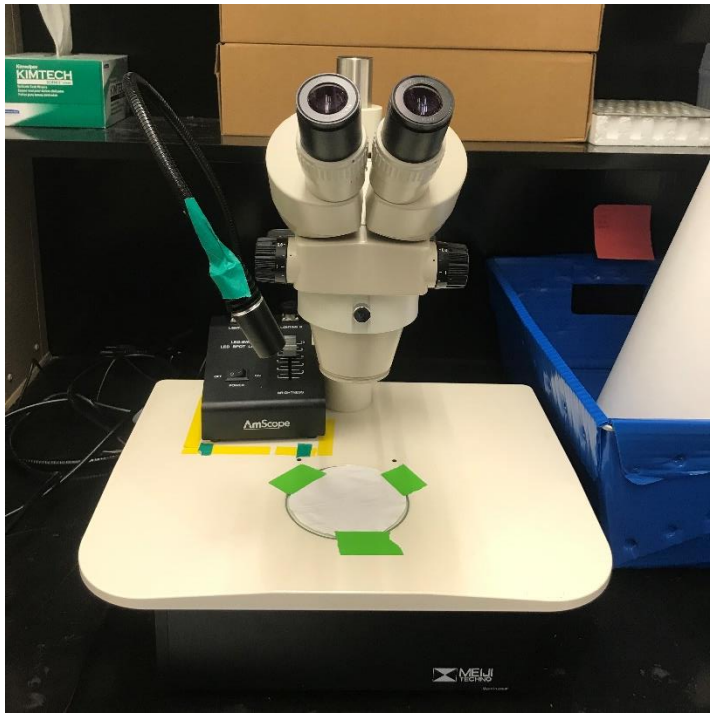
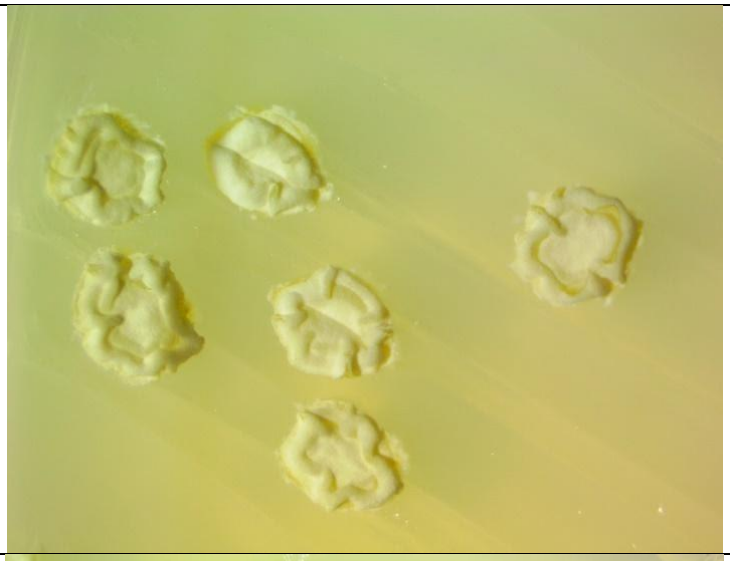


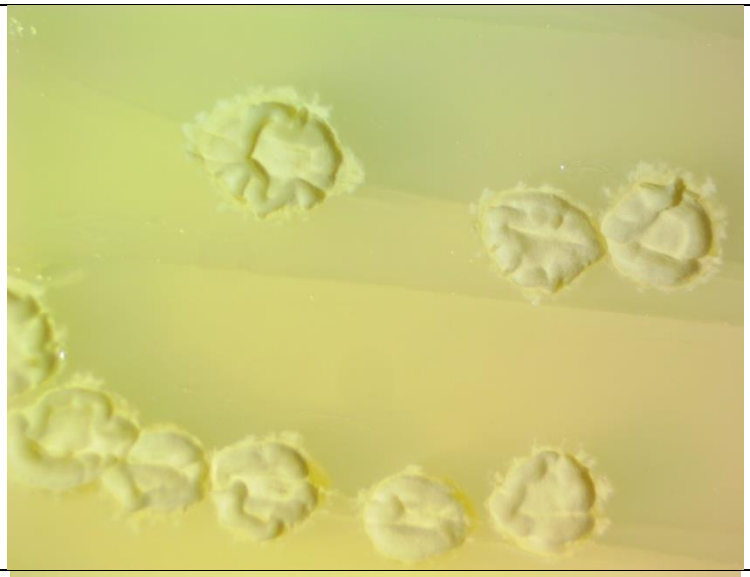
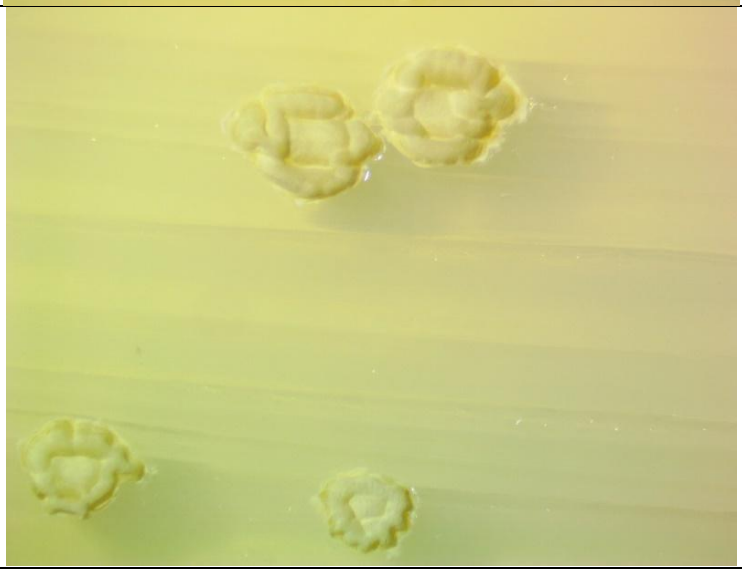
Figure A2.3. Microscope set up for colony morphology photography.

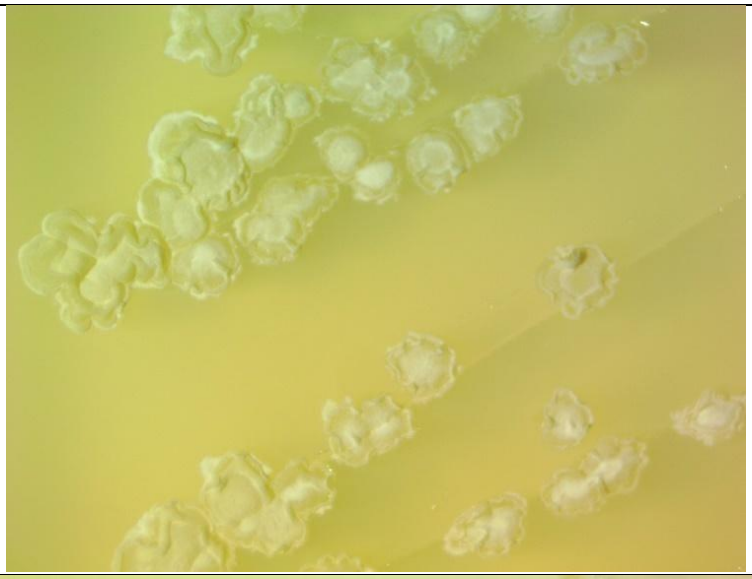
Note the Amscope digital camera was not shown in this picture, and the LED light was not turn on. During photography, the Amscope digital camera replaced the eye piece and is connected to a computer for digital image capture.

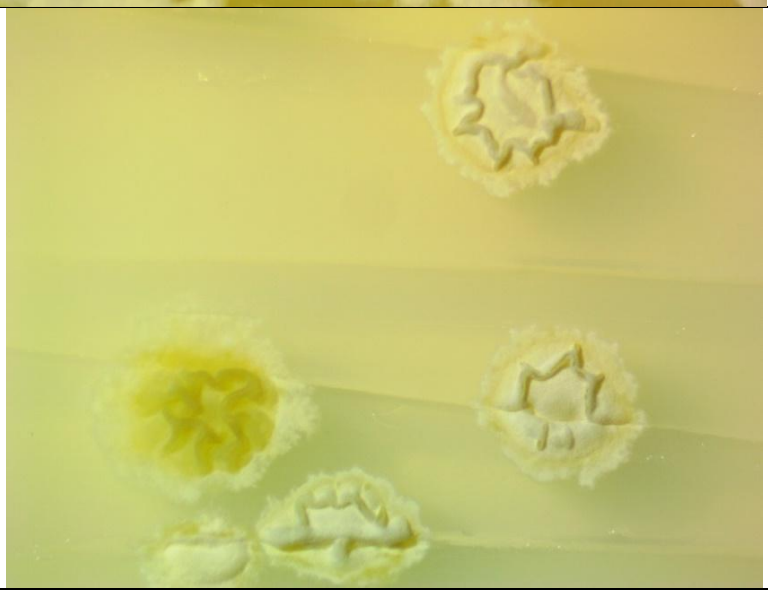
Table A2.1. Images of colony morphology of PBFC strains.

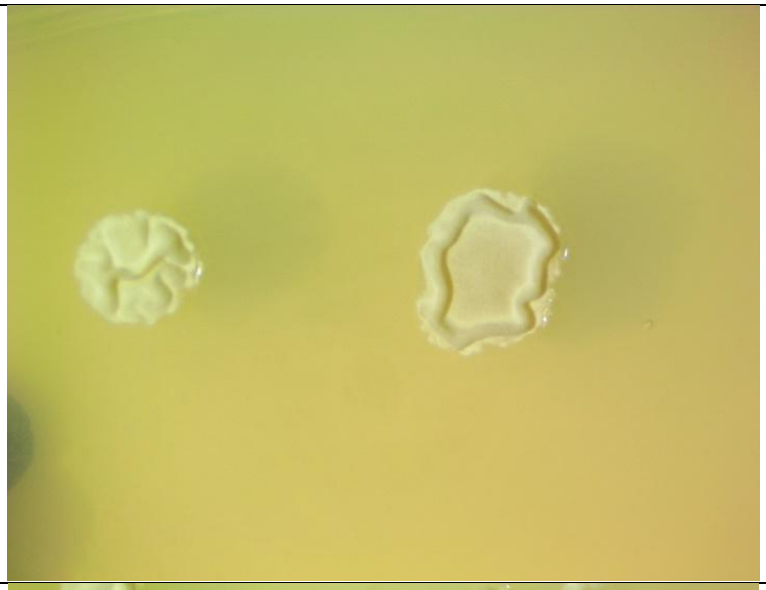
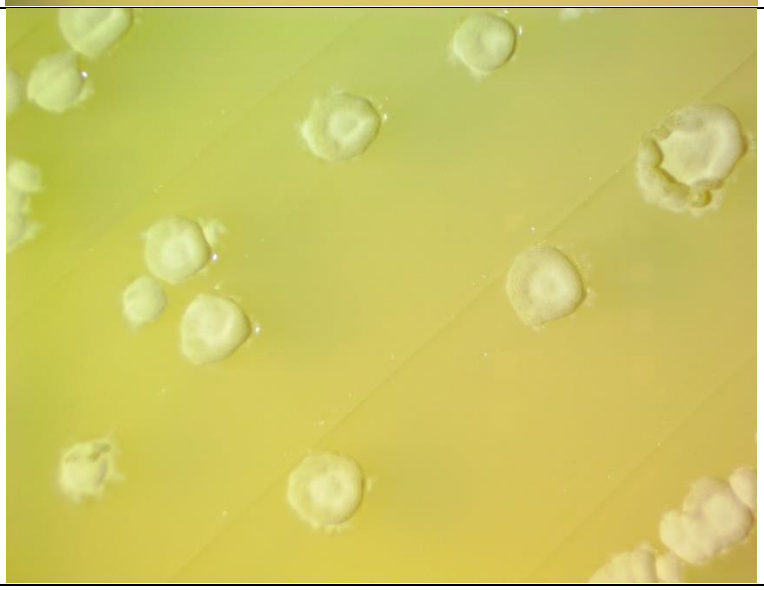
These strains are categorized into 10 different morphology classes.

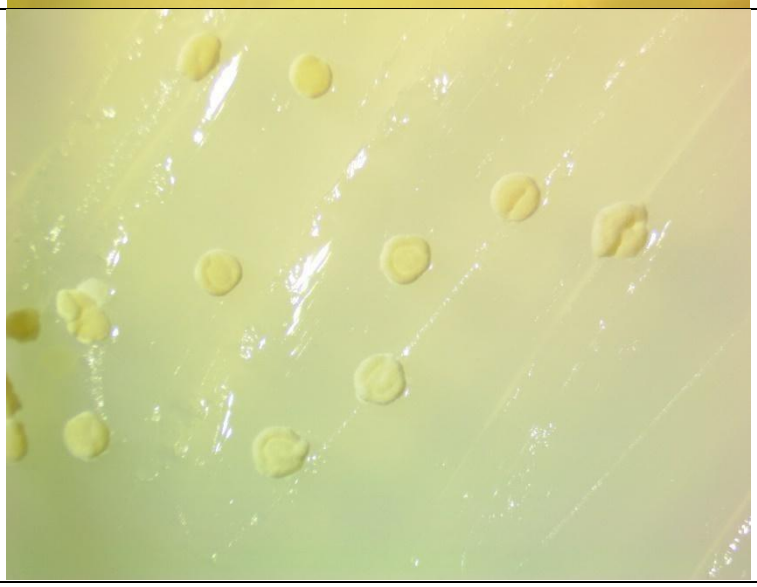
PBFC strain	Type	Image
1	1	
2	8	

3	1	
4	1	

5	1	
6	6	

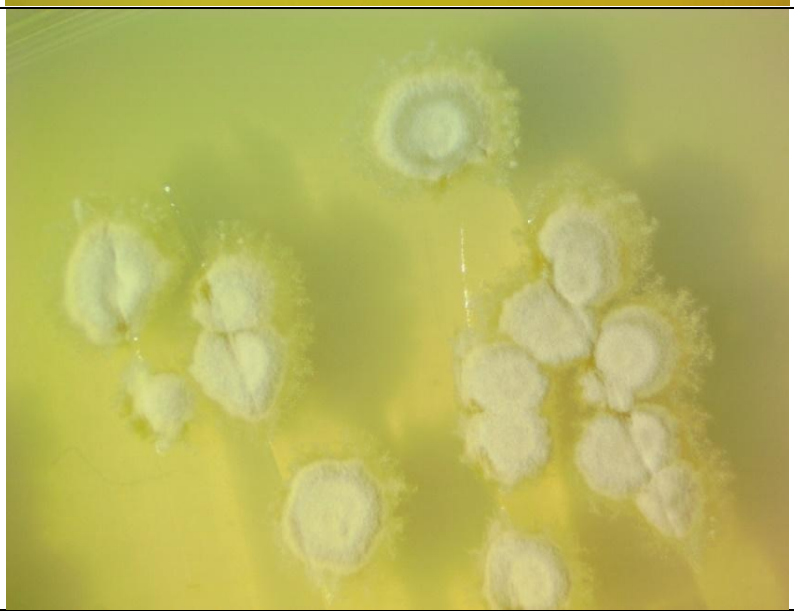
7	3	
8	5	

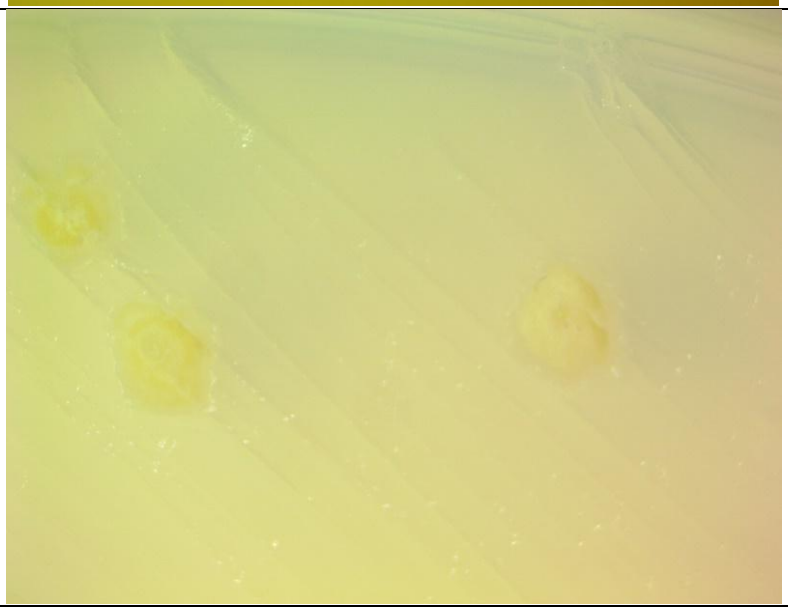
9	1	
10	6	

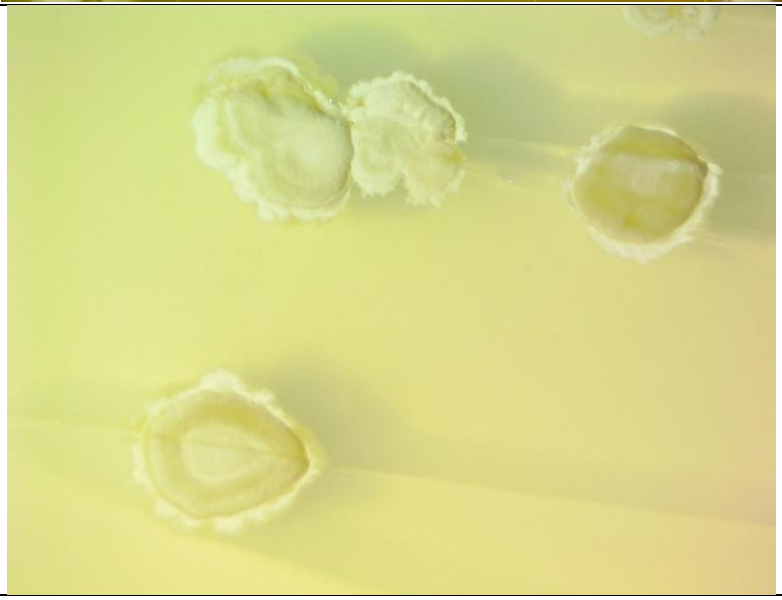
11	7	
12	6	

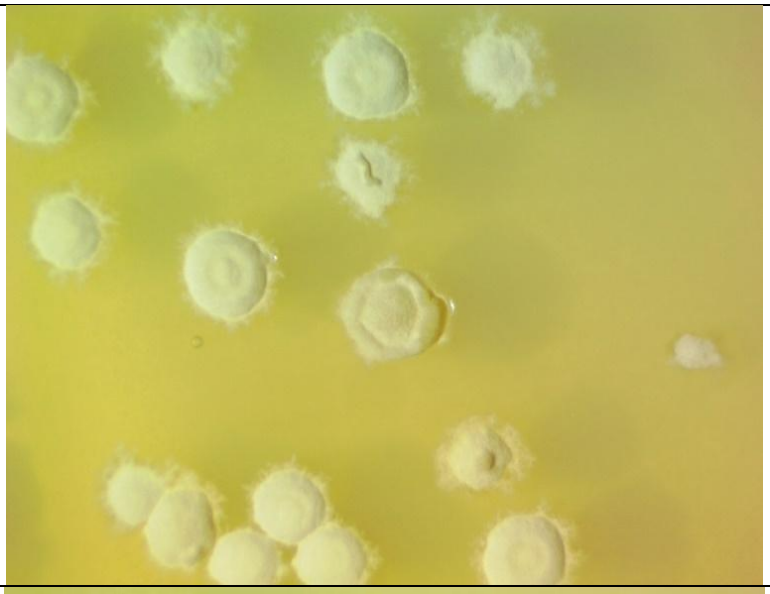
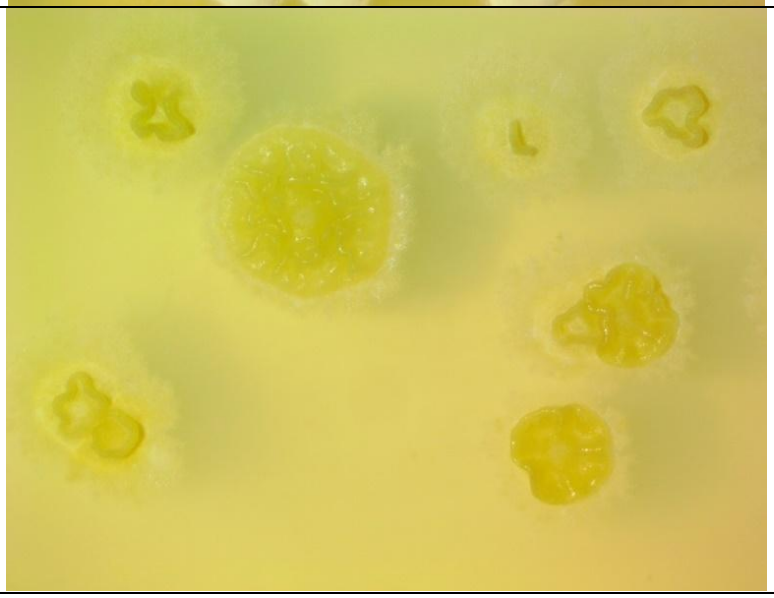
13	3	
14	3	

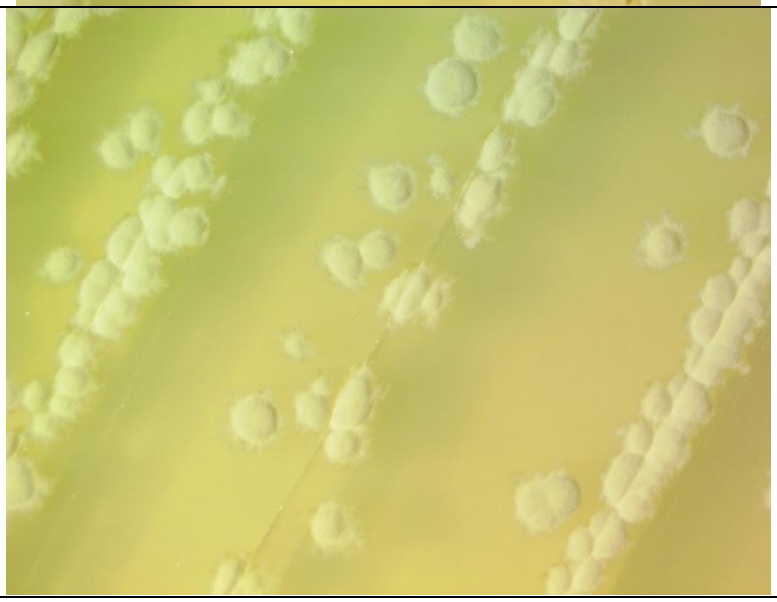
15	6	
16	9	

17	3	
18	3	

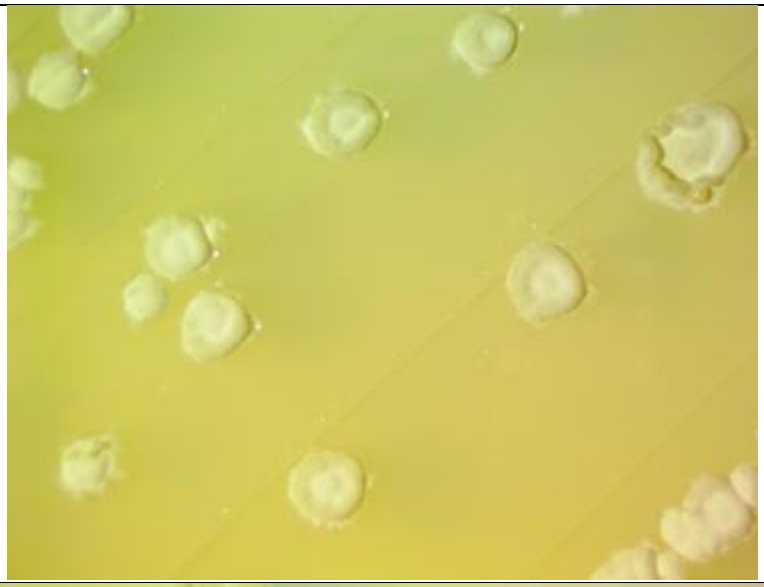
19	3	
20	6*	

21	3	
22	4	

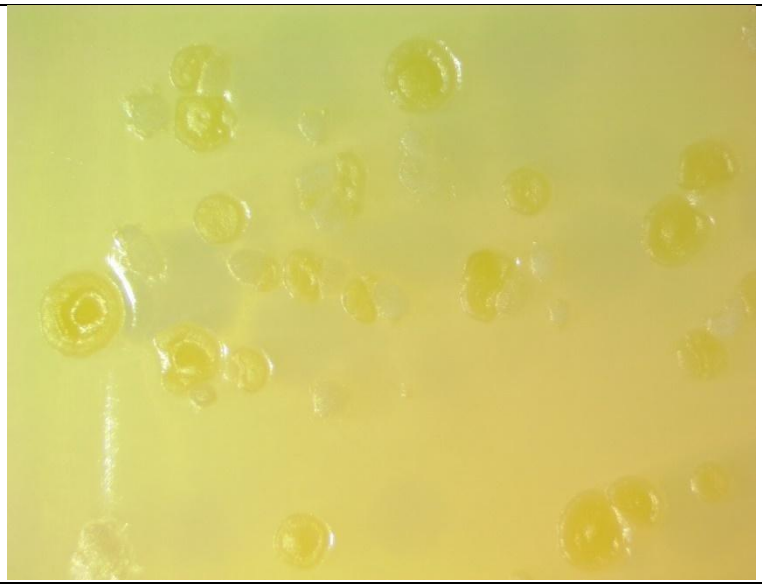
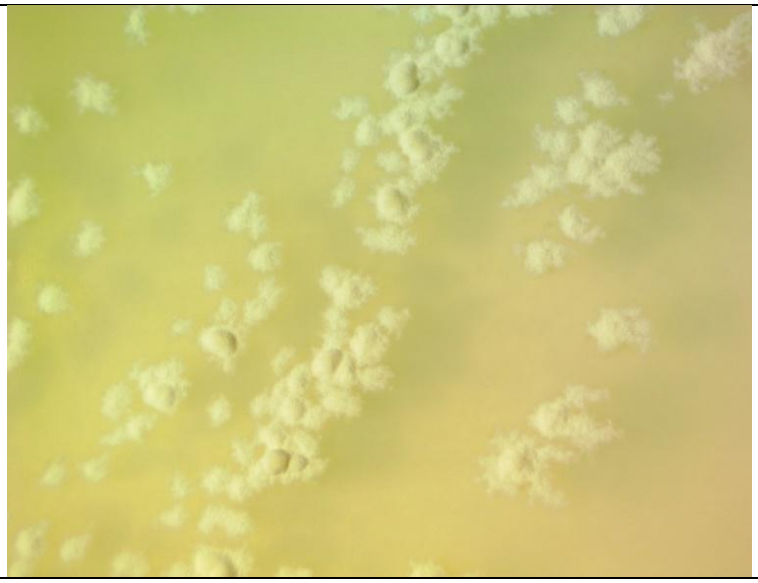
23	6	
24	7	

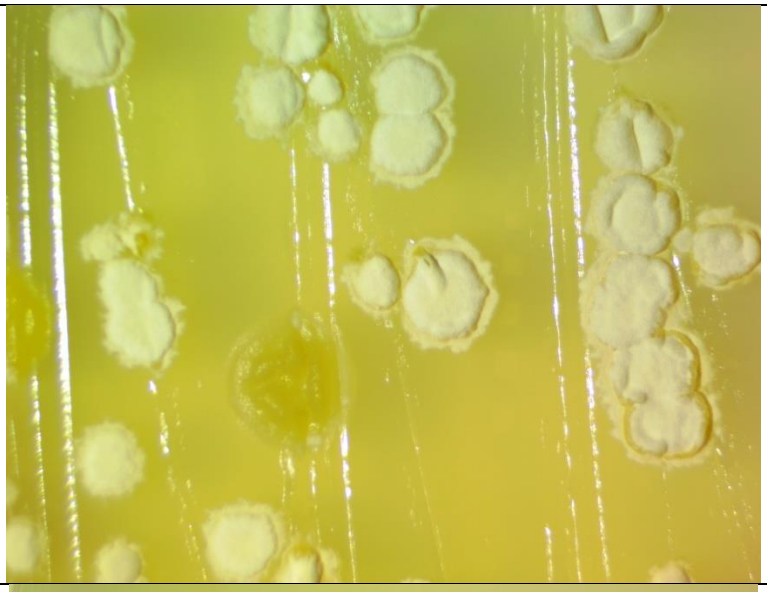
25	1	
26	6	

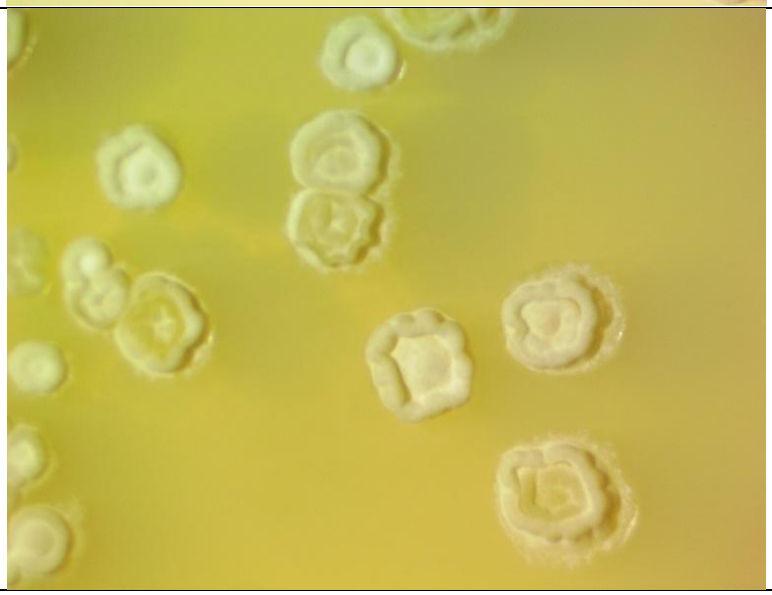
27	9	
28	8	

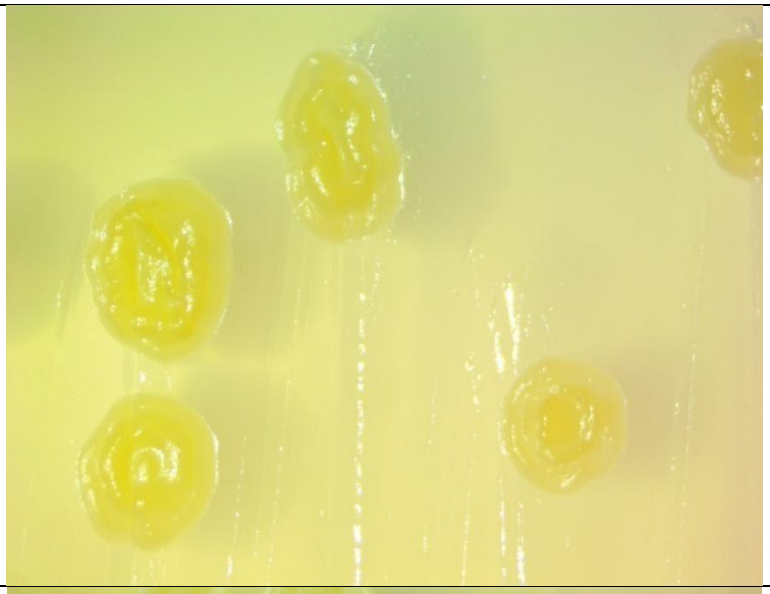
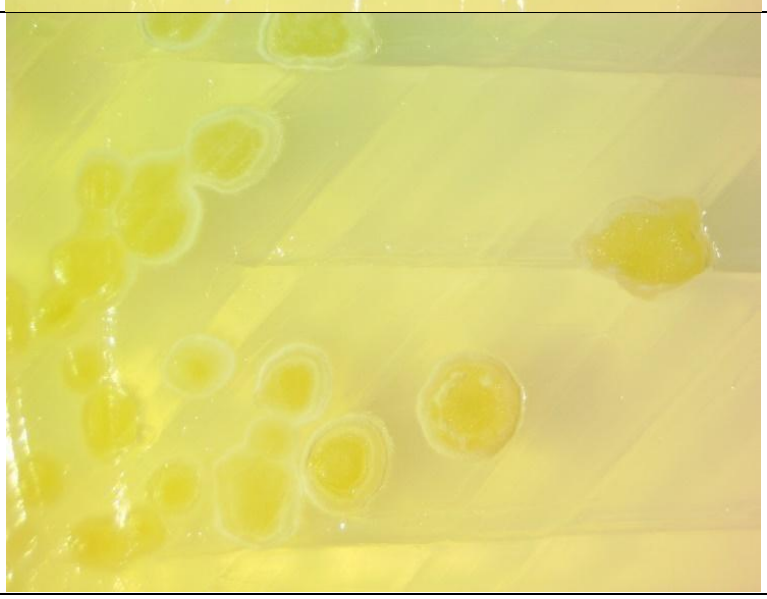
29	6	
30	1	

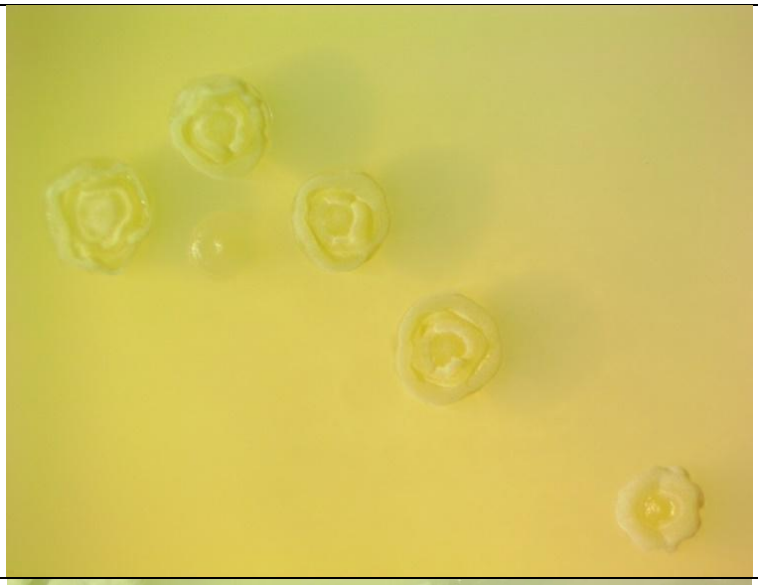
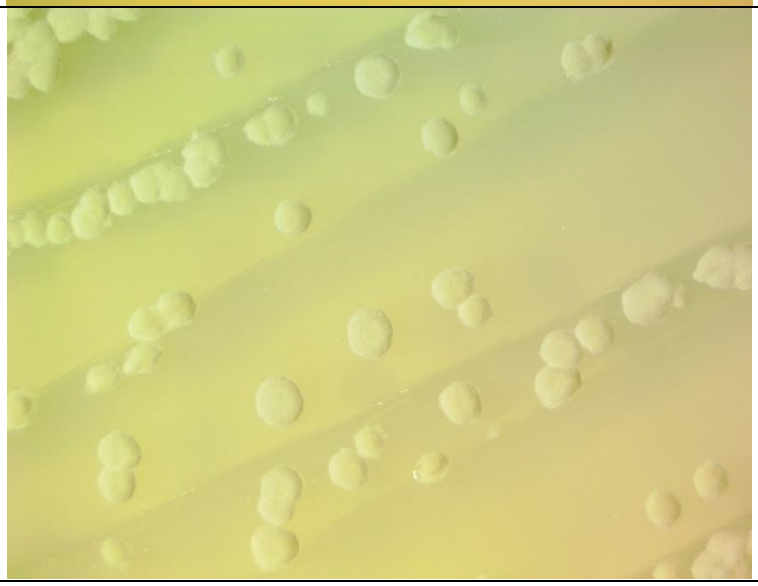
31	8	
32	6	

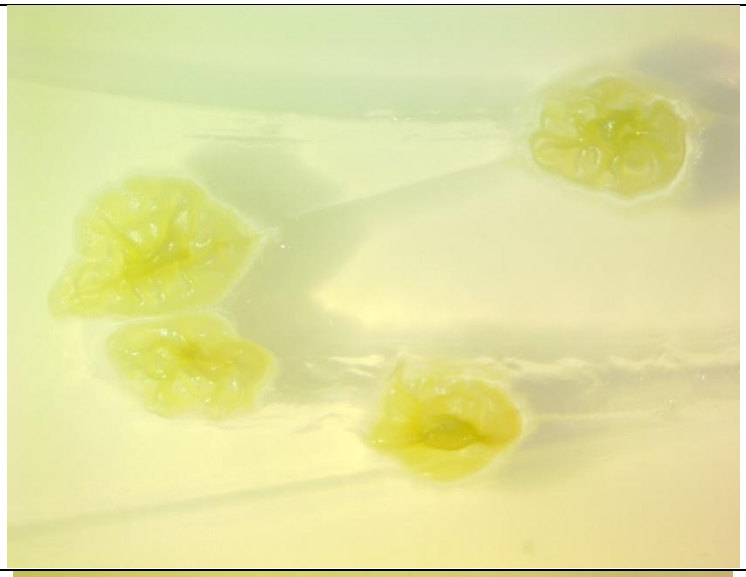
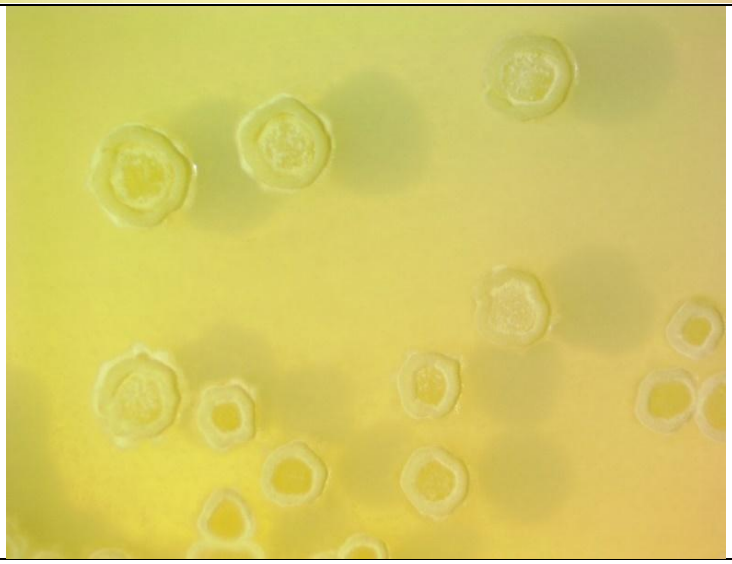
33	8	
34	0	No colonies
35	3	

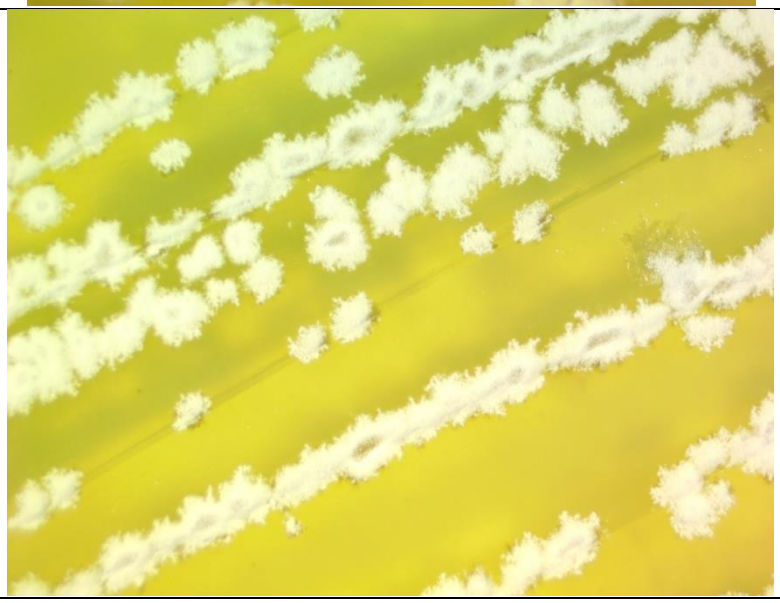
36	6	
37	7	

38	1	
39	2	

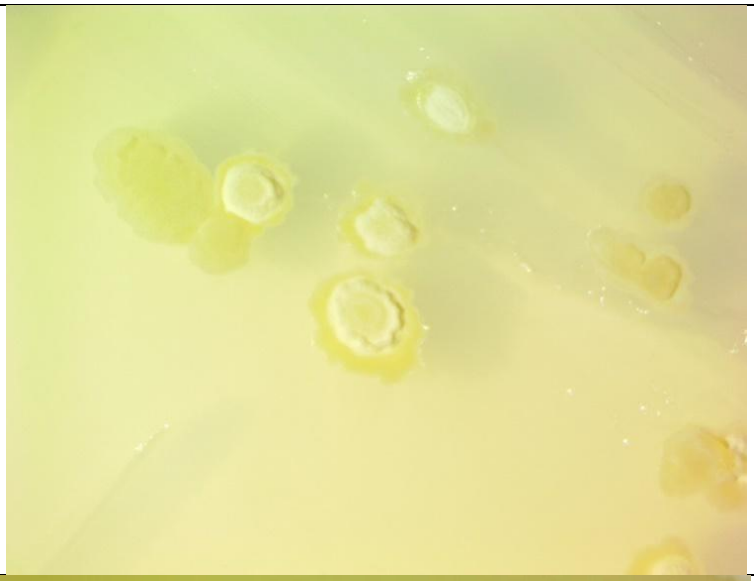
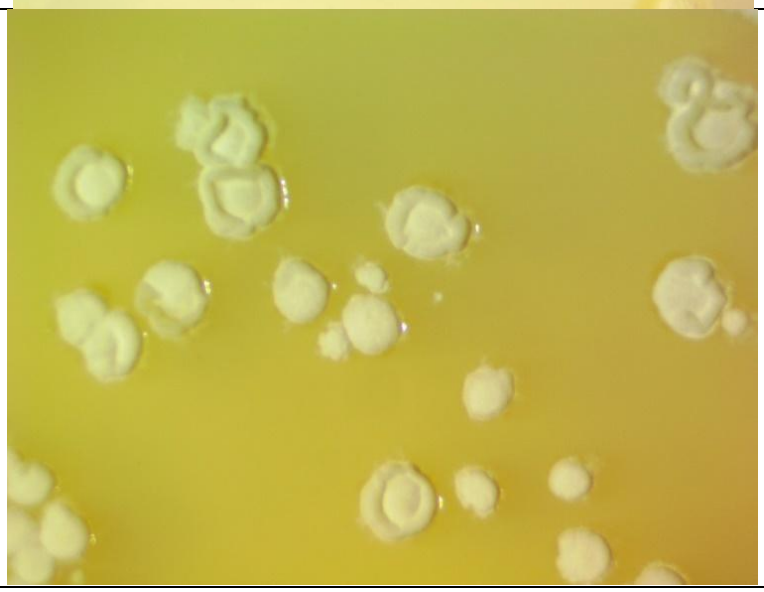
40	7	
41	9	

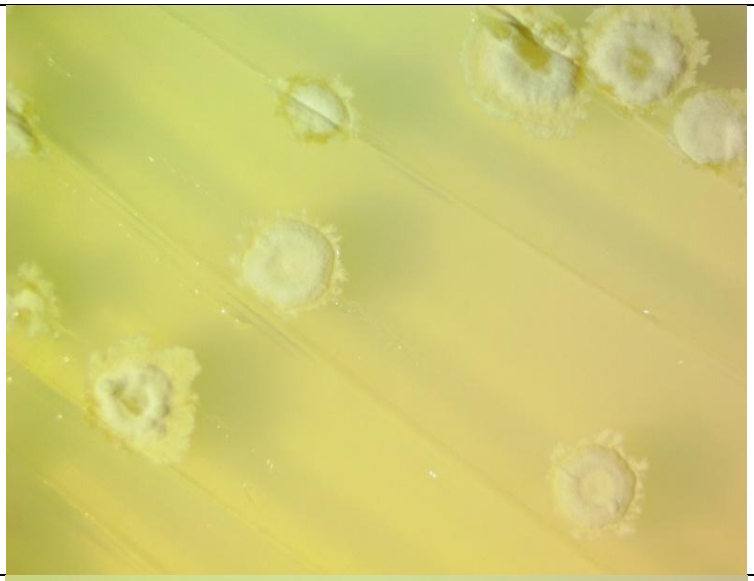
42	9	
43	6	

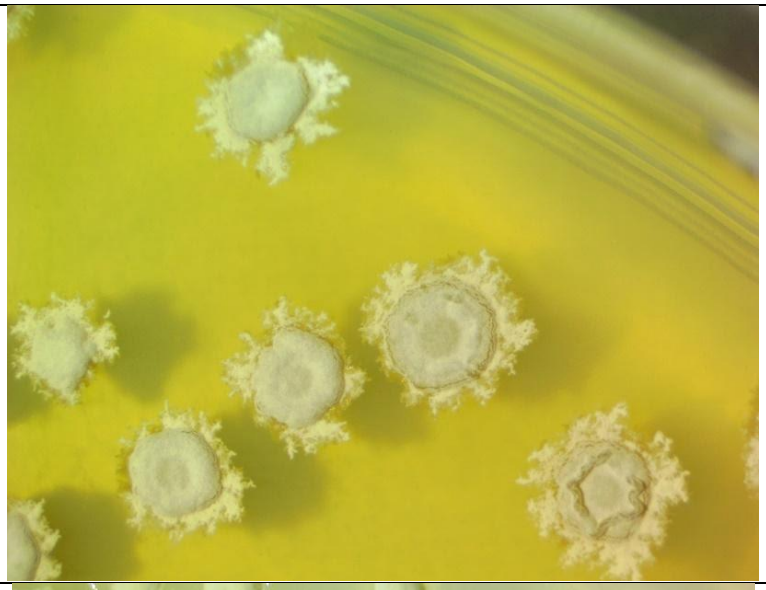
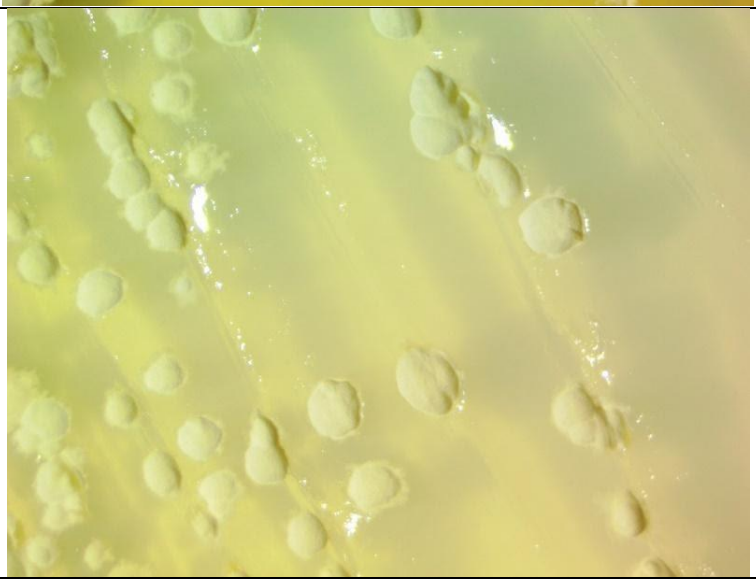
44	7	
45	9	

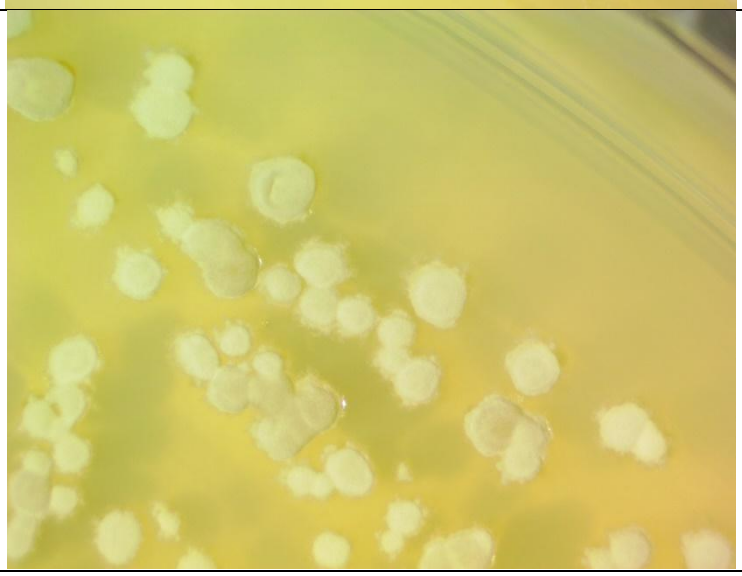
46	3	
47	3	

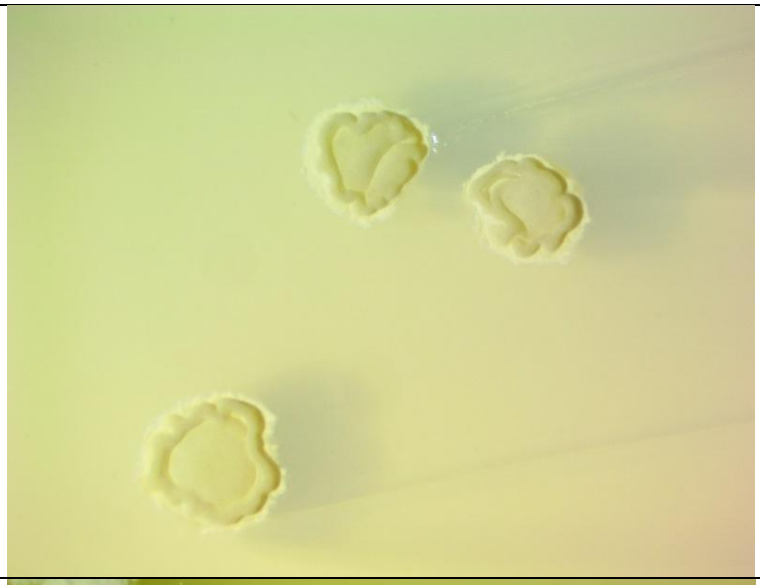
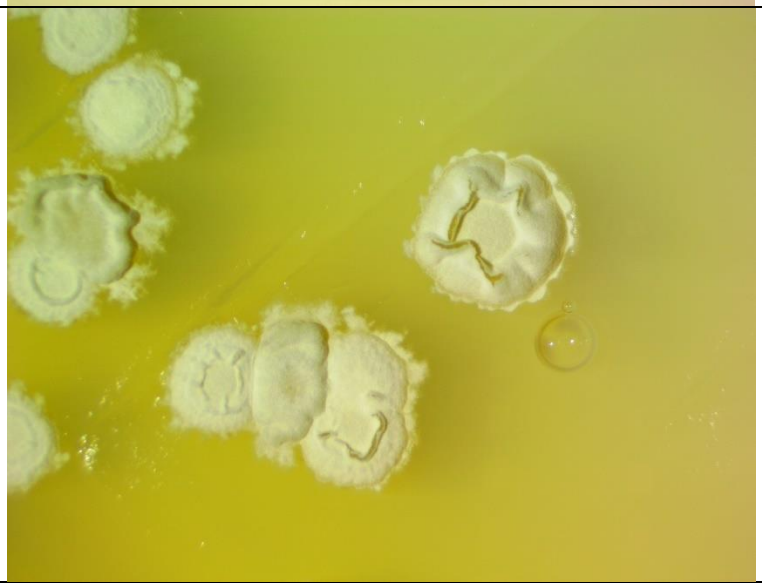
48	1	
49	3	

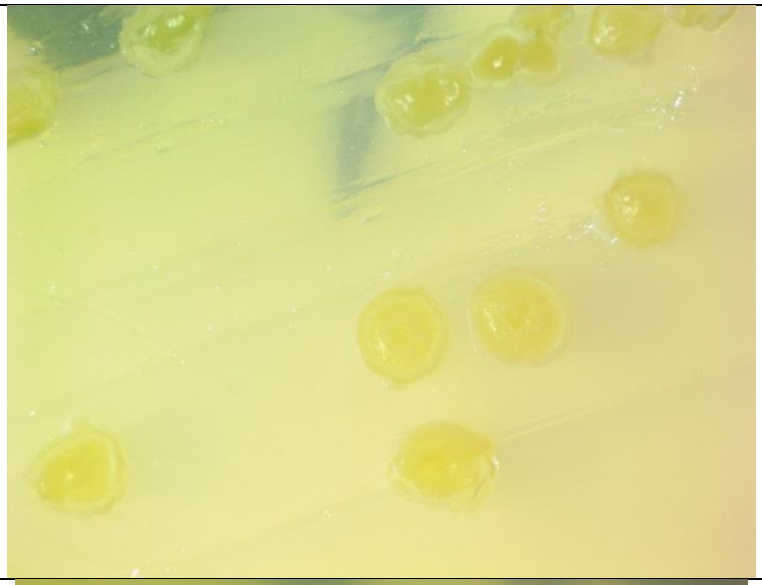
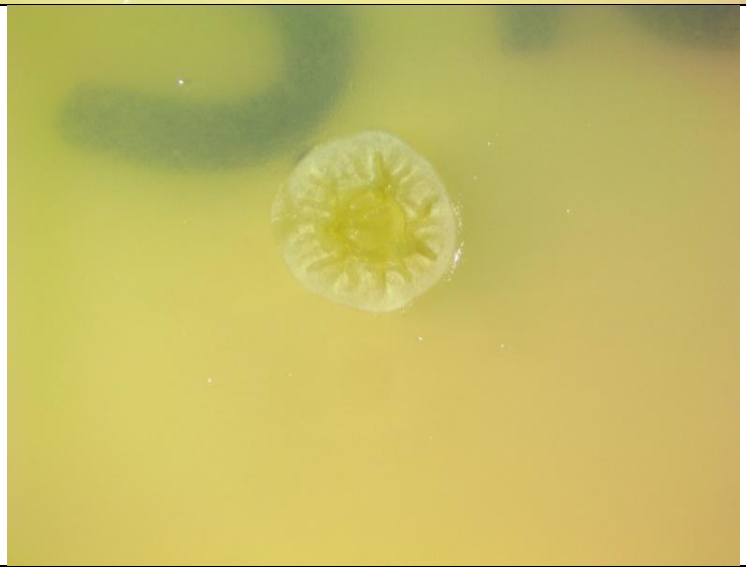
50	1	
51	1	

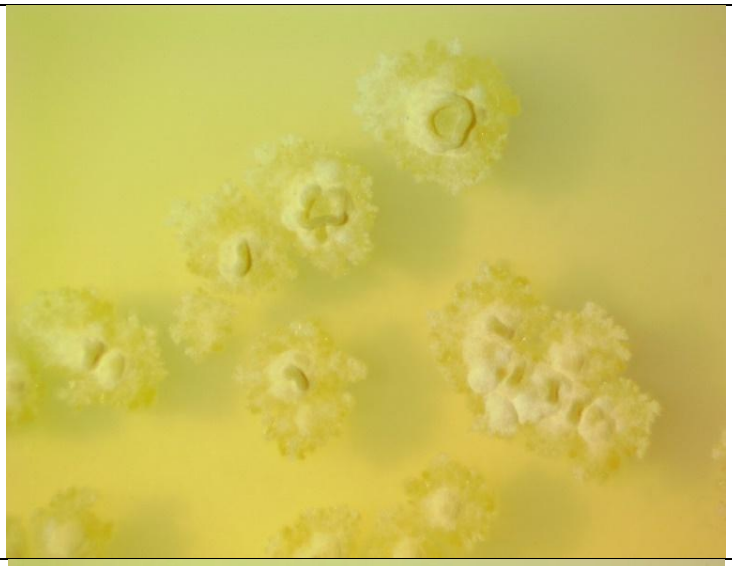
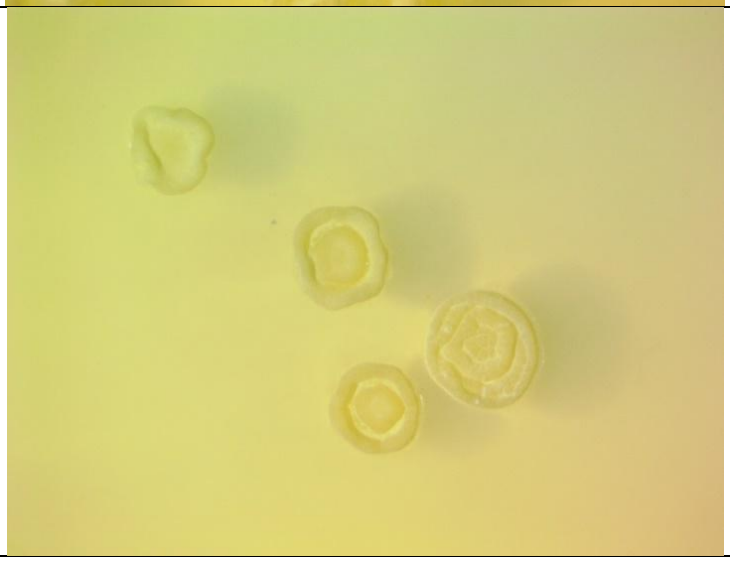
52	4	
53	1	

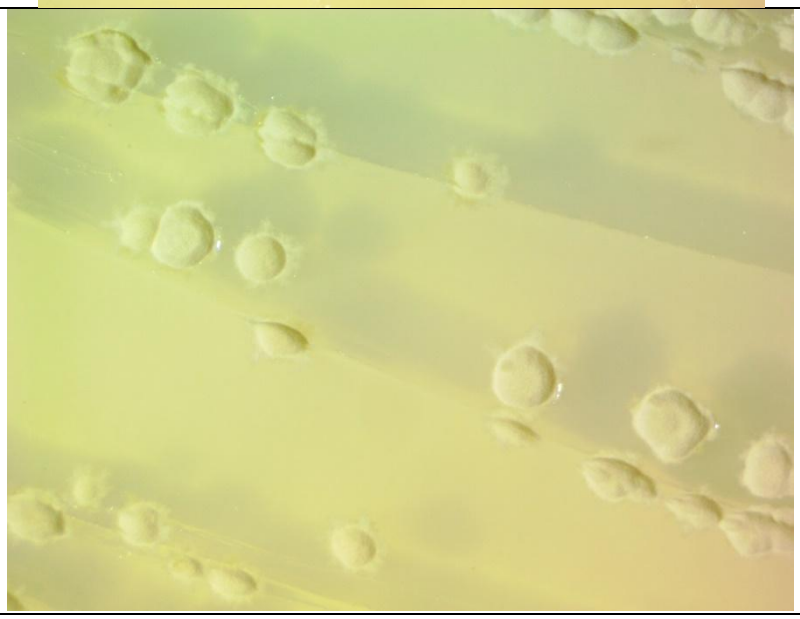
54	3	
55	6	

56	1		
57	6		

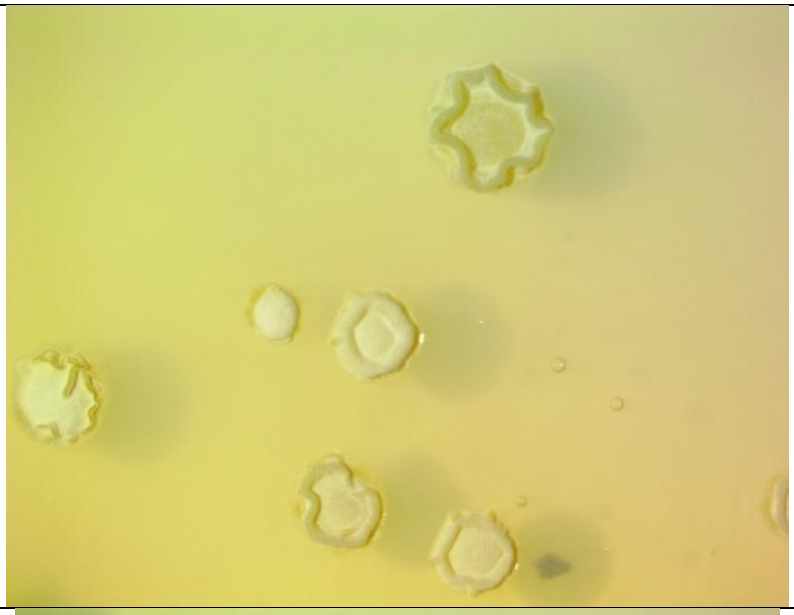
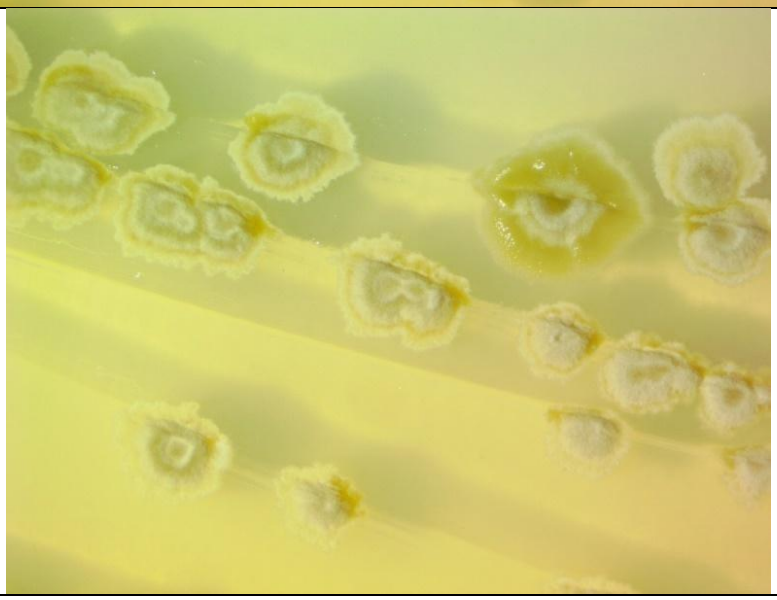
58	1	
59	1	

60	8	
61	5	

62	10		
63	2		

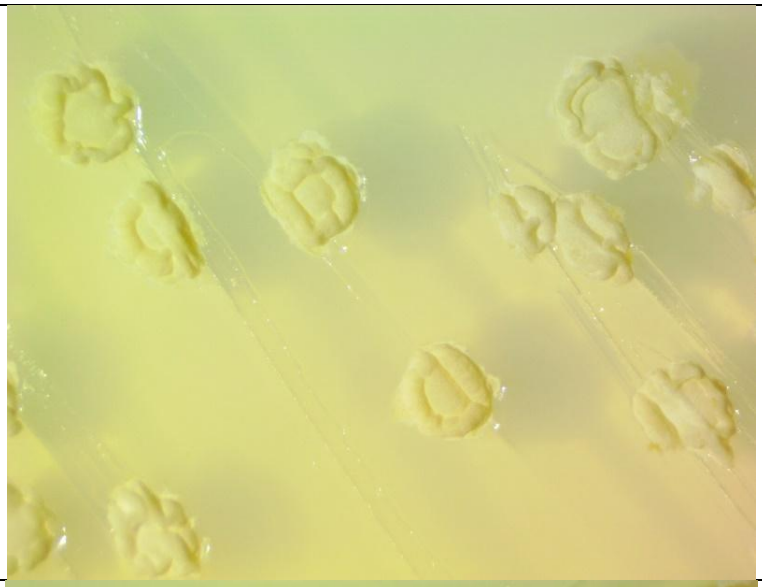
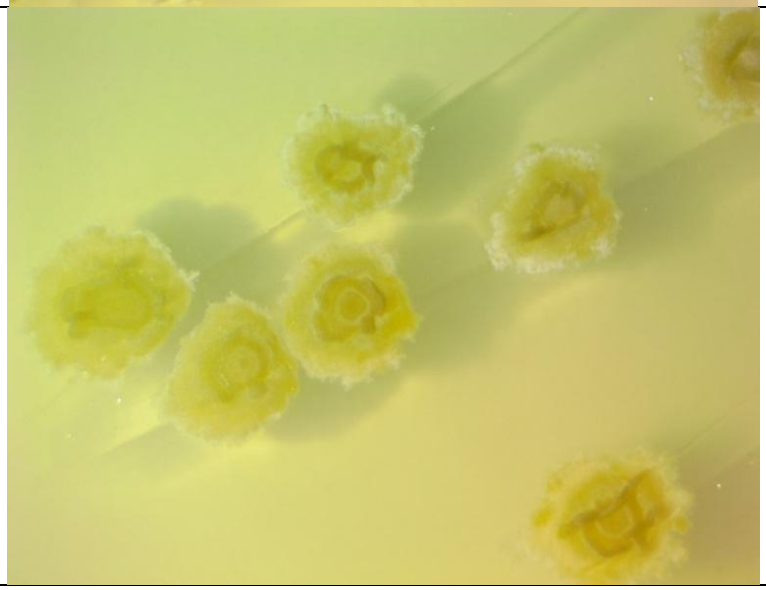
64	2	
65	6	

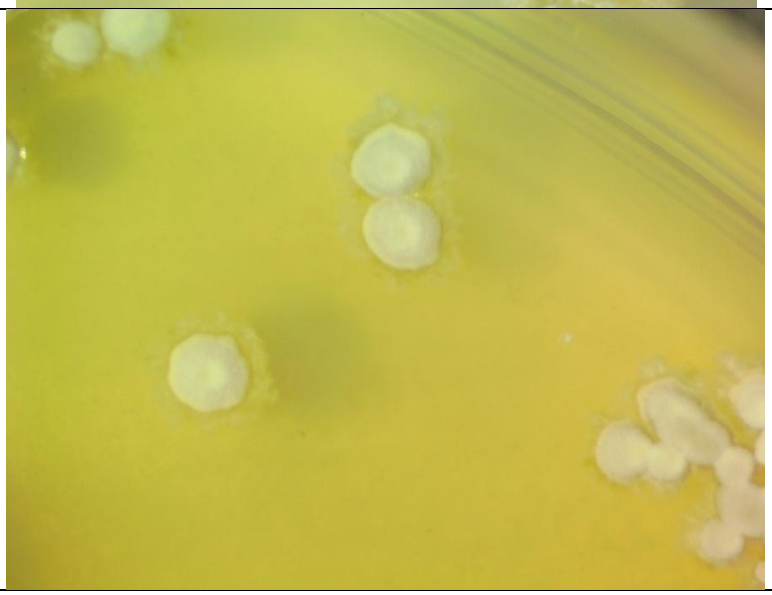
66	10	
67	2	

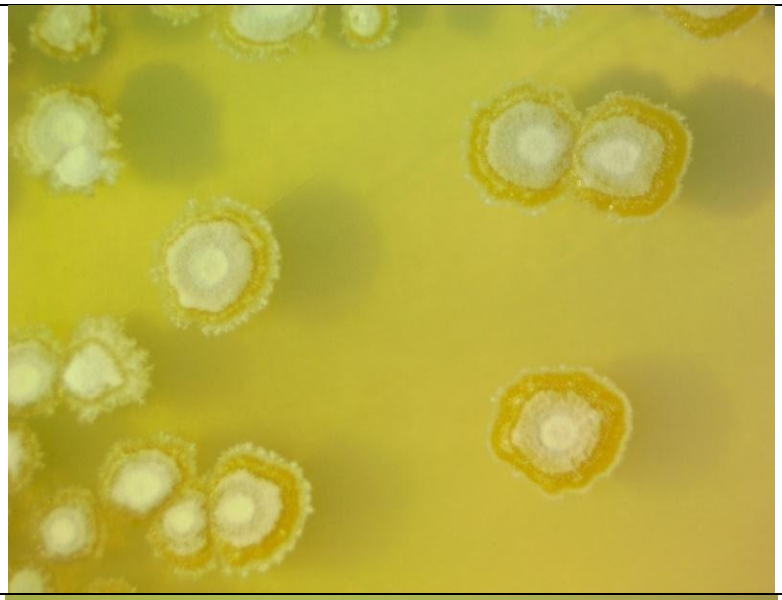
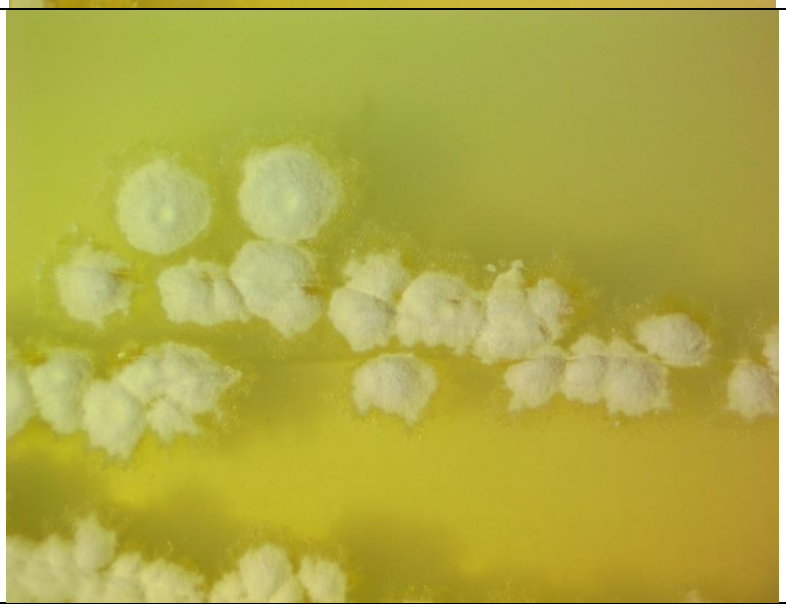
68	1	
69	3	

70	2	
71	8	

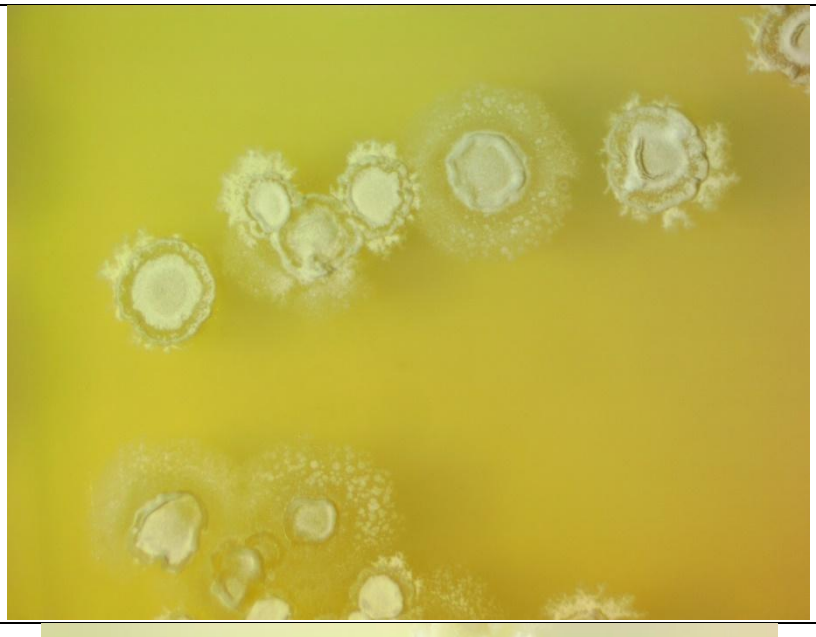
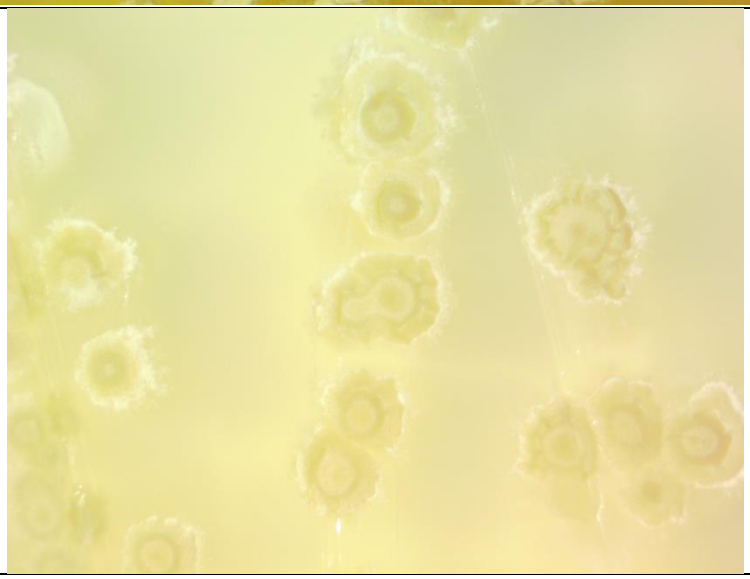
72	3	
73	3	

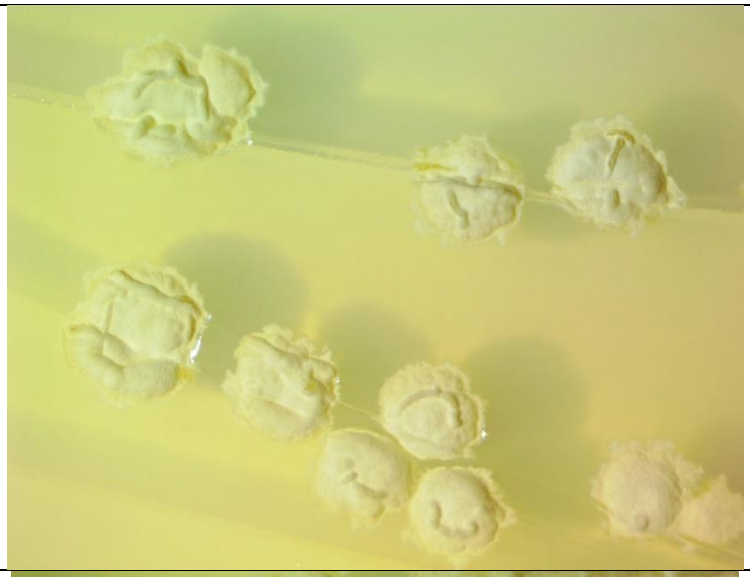
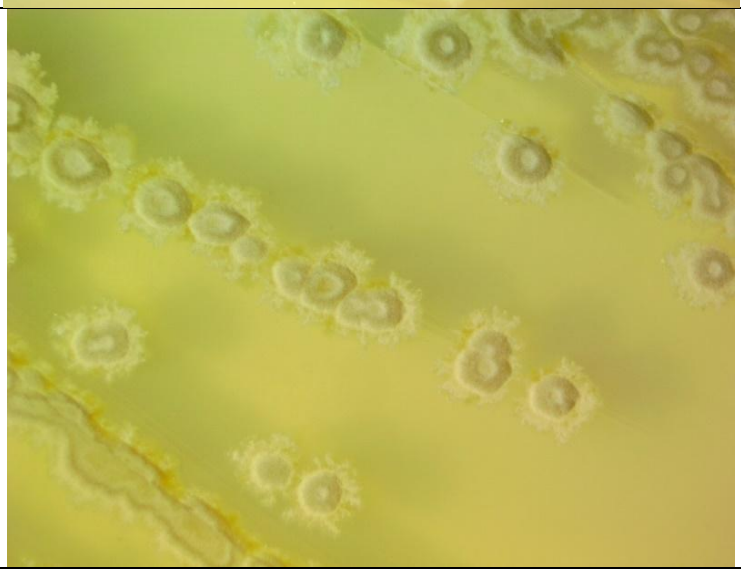
74	1		
75	3*		

76	10	
77	3	

78	3	
79	3	

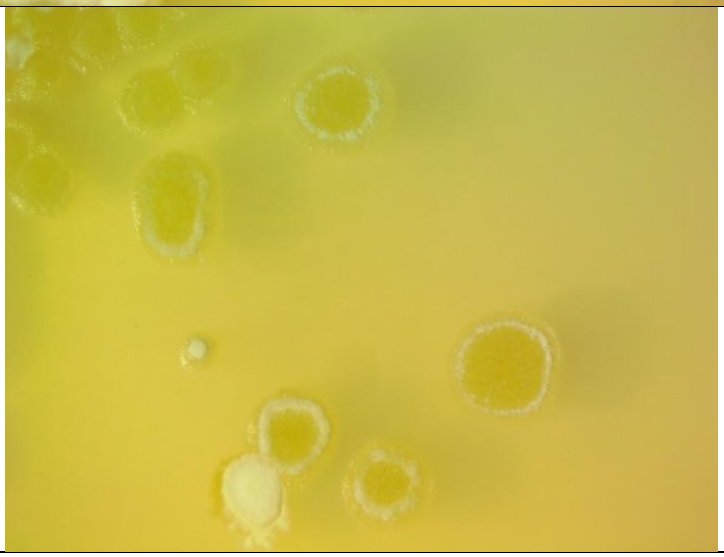
80	8	
81	0	No colonies
82	3	

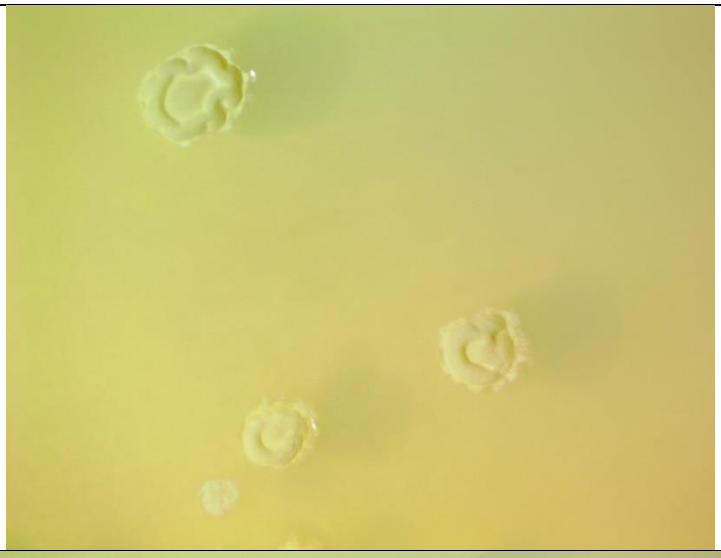
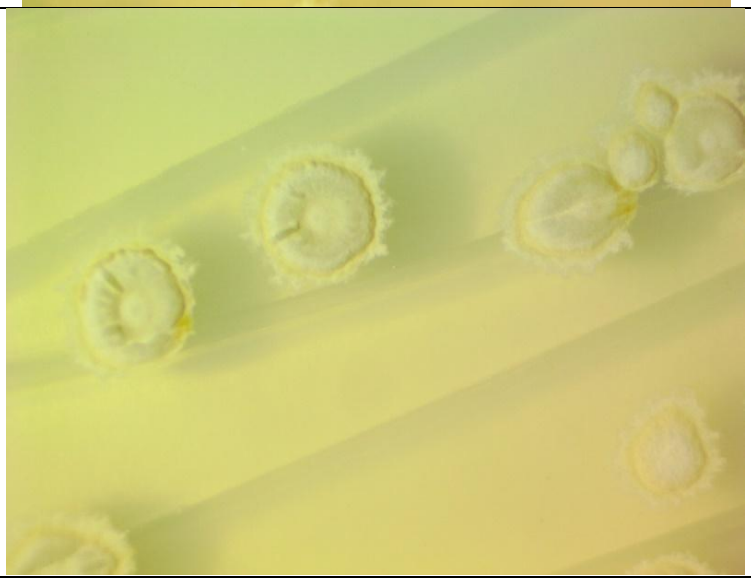
83	3	
84	10	

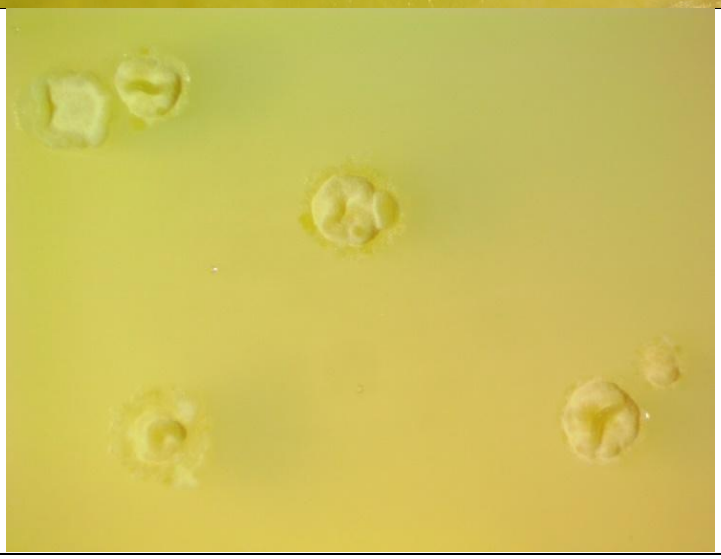
85	1		
86	10		

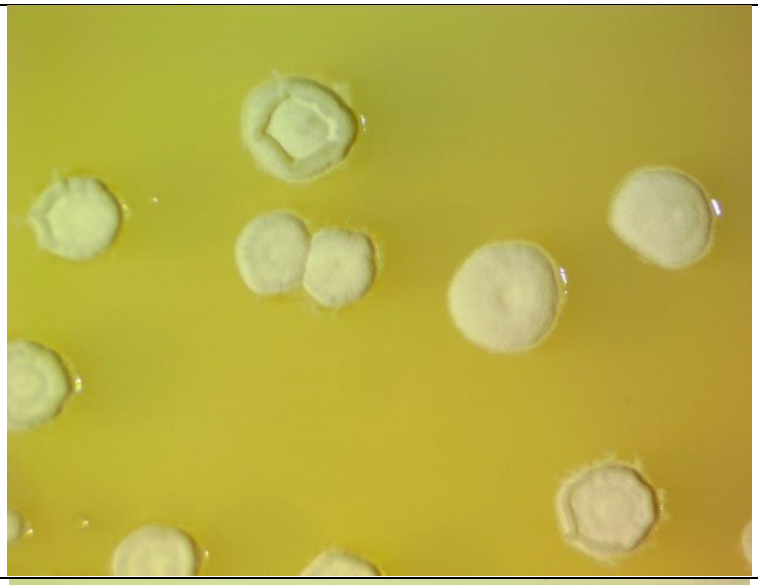
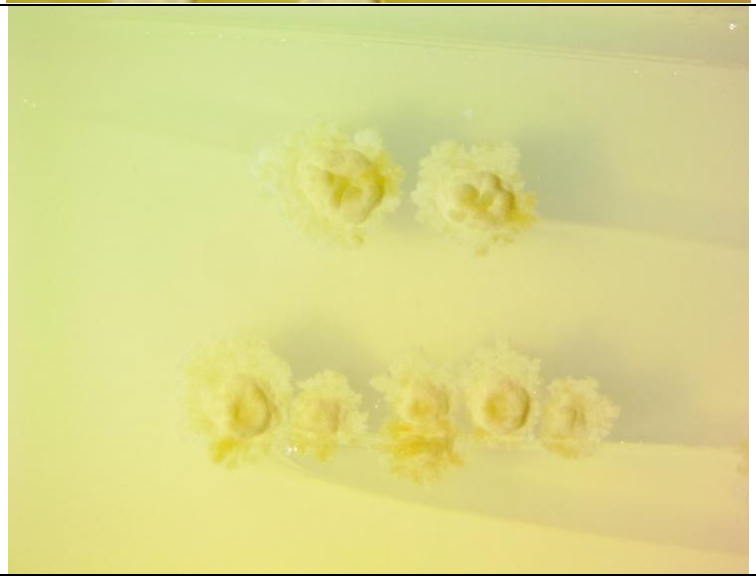
87	4		
88	10		

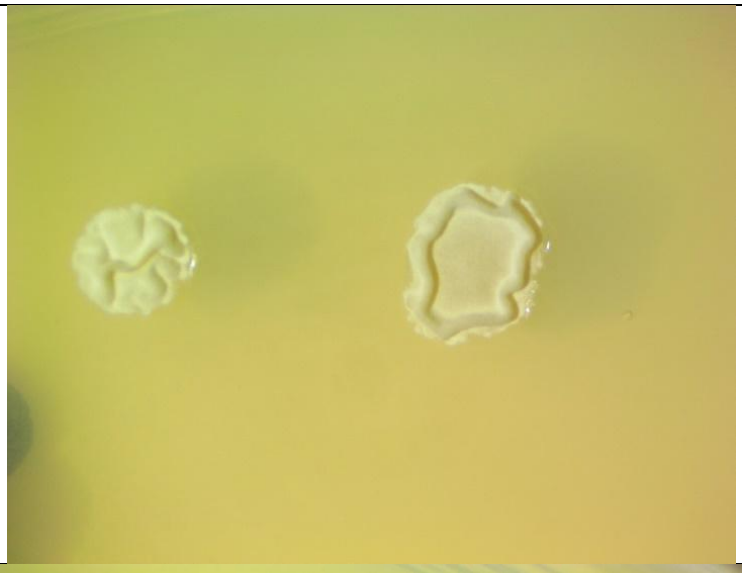
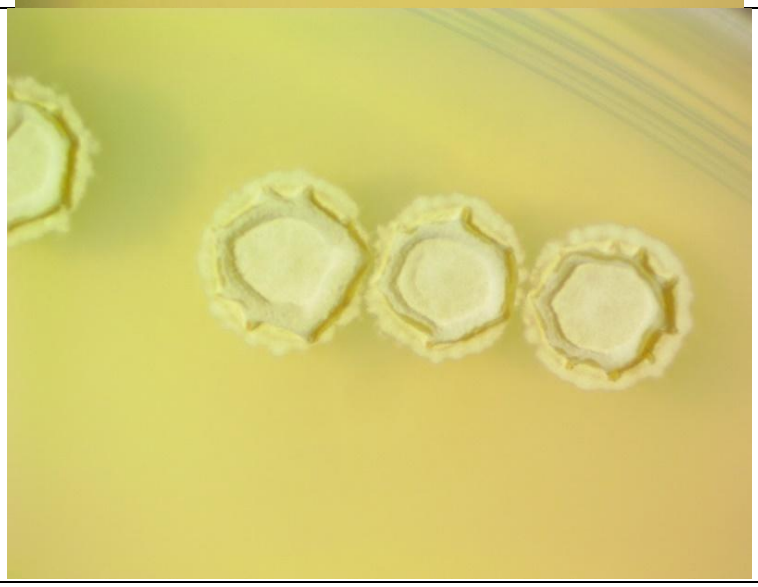
89	3	
90	3	

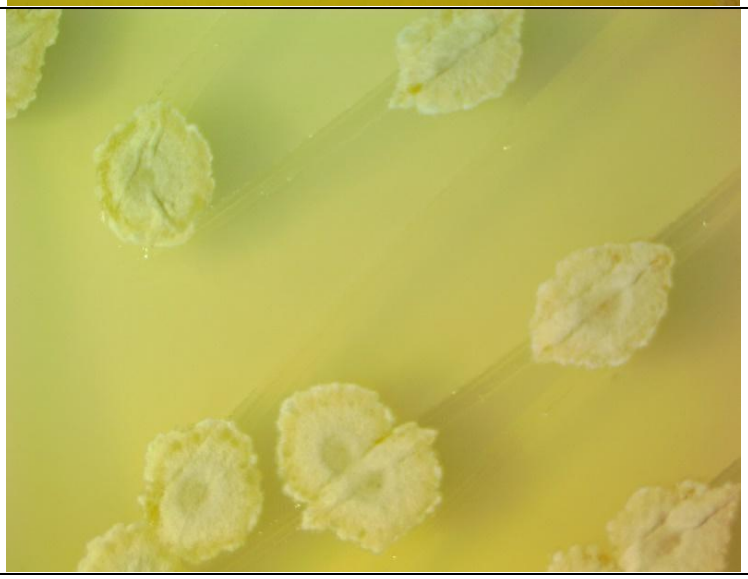
91	3	
92	9	

93	1	
94	6	

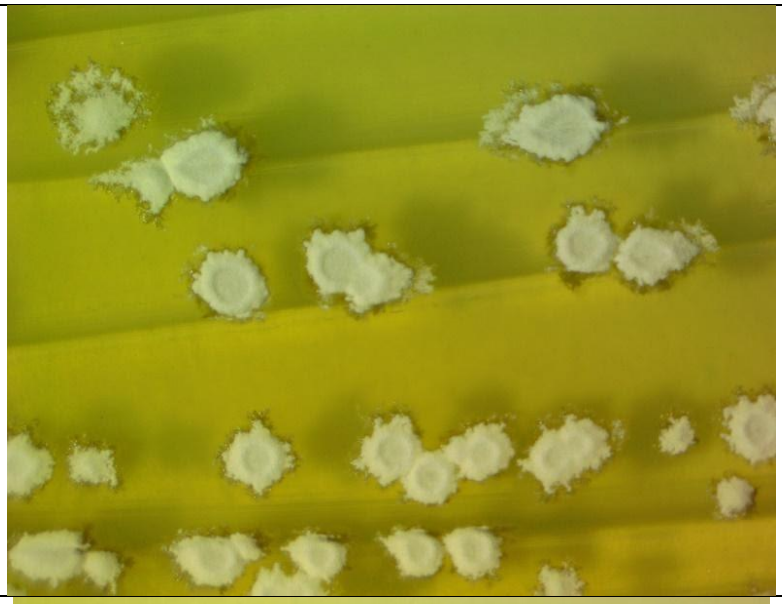
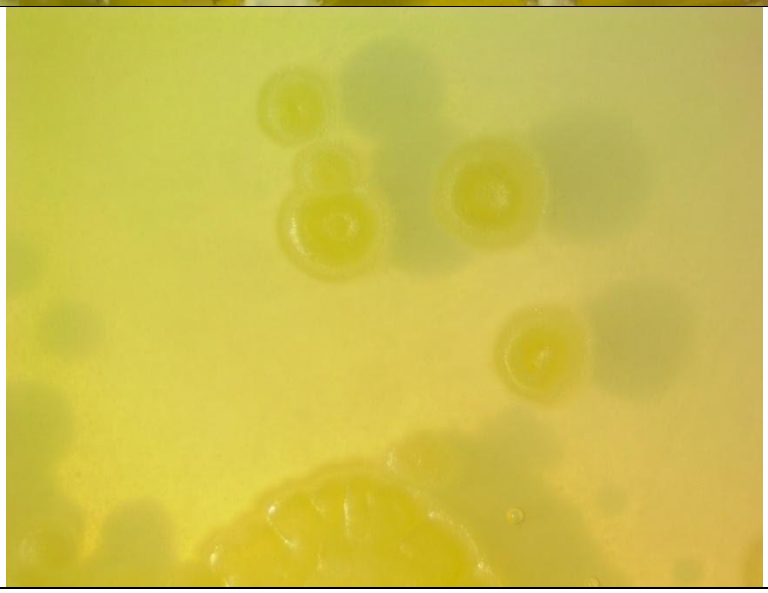
95	7	
96	2	

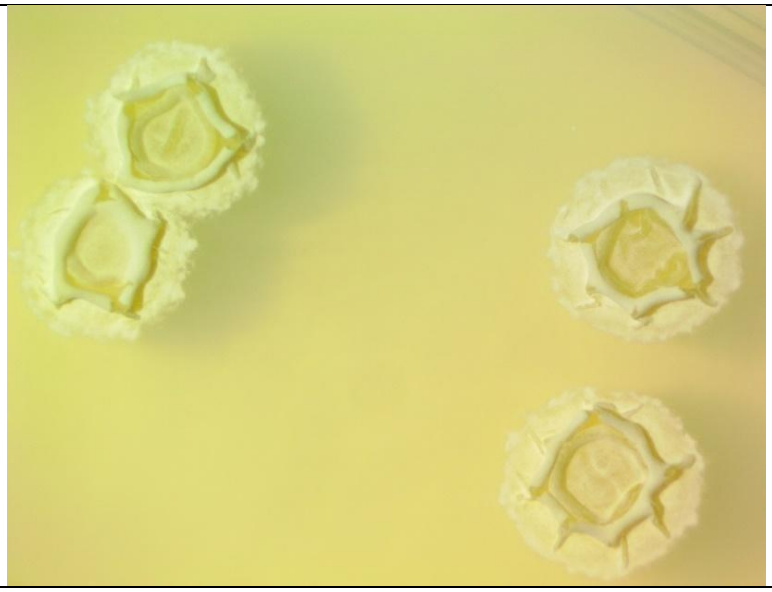
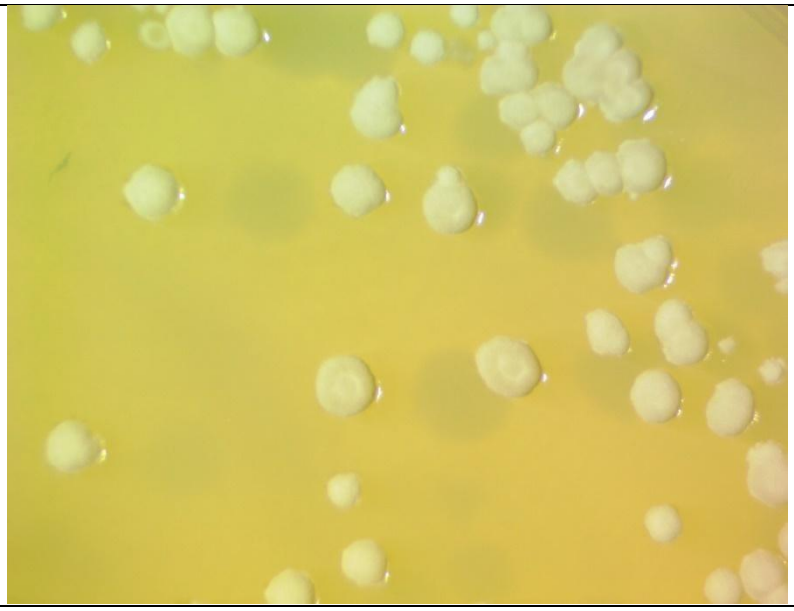
97	6	
98	10	

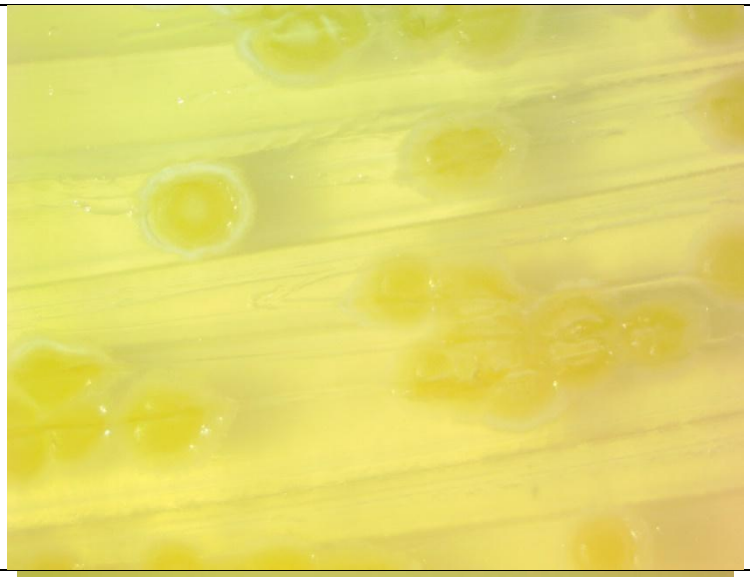
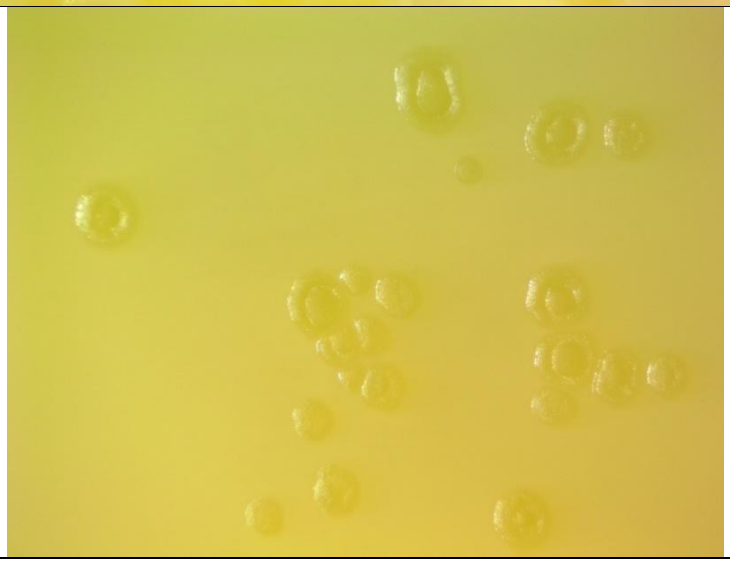
99	1	
100	1	

101	3	
102	3	

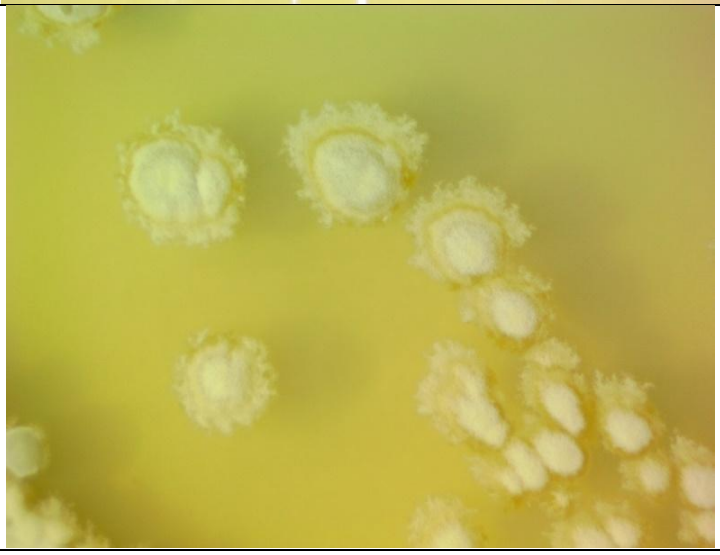
103	7	
104	3	

105	3	
106	8	

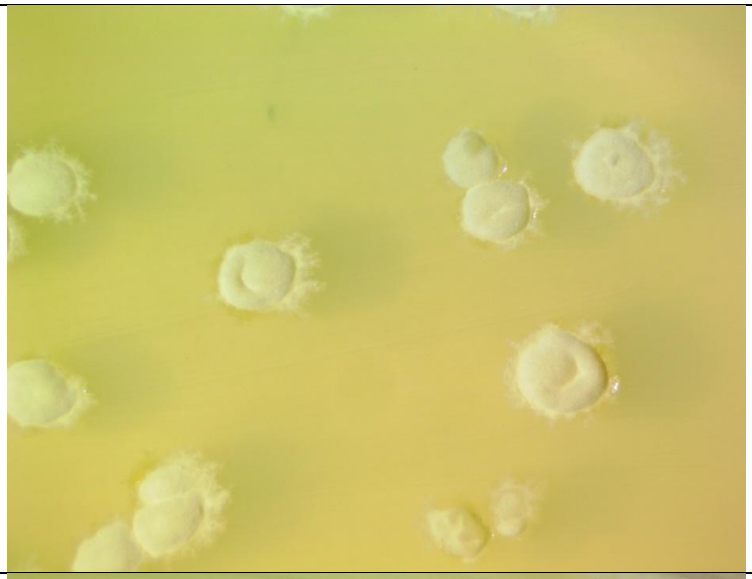
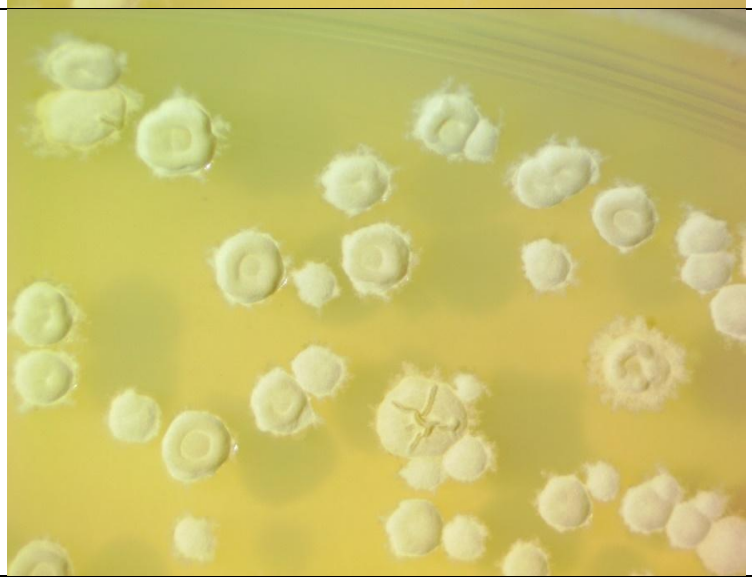
107	1	
108	0	No colonies
109	6	

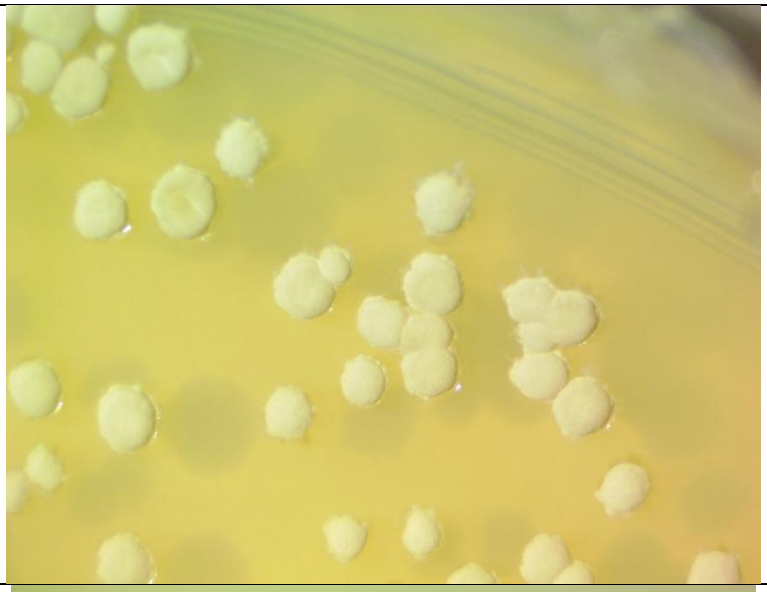
110	8	
111	8	

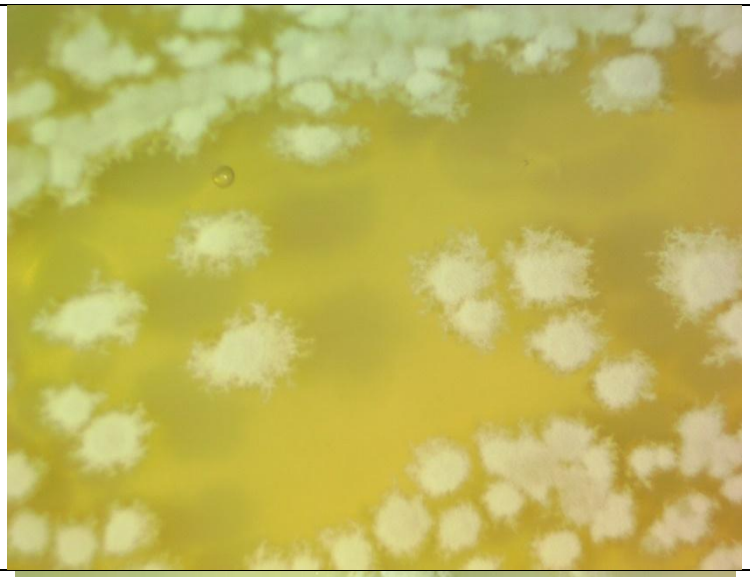
112	9	
113	10	

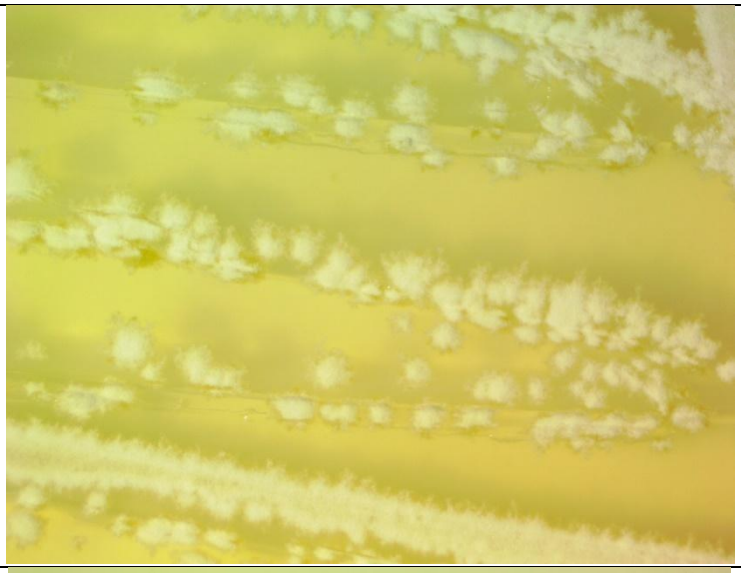
114	1	
115	3	

116	1		
117	3		

118	6		
119	6		

120	6	
121	1	

122	3	
123	6	

124	3	
125	3	

Captions of Supplementary Data File A2. (separate file)

The file titled ‘Supplementary Data File A2.xslx’ contains six worksheets.

Worksheet 1 (1. PB matrix):
This worksheet contains the 125 x 31 Plackett-Burman matrix. Column A through column H, which are highlighted in grey, are used to guide the expression level design. The unhighlighted columns (column I - AE) are not use in the genetic design but are used as dummy variables (dv).

Worksheet 2 (2. Genetic Design):
This worksheet describes the architecture of the PBFC design at monocistronic level and partial cluster level.

- Column A – G: The genetic design details of each monocistronic unit.
  - Column A: expression level corresponding to PB matrix in worksheet 1
  - Column B: The name of monocistronic units. This corresponds to the ‘MT’ name in **Figure 4.1c**.
  - Column C, G: MoClo scars flanking the monocistronic unit
  - Column D,F: Promoter-RBS and terminator part controlling the expression of the gene of interest.
  - Column E: gene of interest in each monocistronic unit
- Column I – P: The genetic design details of each partial cluster.
  - Column I: The name of upper and lower partial clusters. The name corresponds to the ‘PC’ names in **Figure 4.1c**.
  - Column J: This indicates the type of partial cluster, either upper or lower.
  - Column K, P: MoClo scars flanking the partial cluster
  - Column L – P: The detailed makeup of each partial cluster in terms of monocistronic units. The composition order reflects the gene order in the genetic design.
- Column R – V: The genetic design details of each Plackett-Burman full cluster in terms of partial clusters.
  - Column R: The name of PBFC. The name corresponds to the ‘FC’ names in **Figure 4.1c**.
  - Column S, V: MoClo scars flanking the full cluster
  - Column T, U: The detailed makeup of each full cluster in terms of partial cluster. The composition order reflects the cluster order in the genetic design.
- Column X – AH: The genetic design details of each Plackett-Burman full cluster in terms of monocistronic units.
  - Column X: The name of PBFC. The name corresponds to the ‘FC’ names in **Figure 4.1c**.
  - Column Y, AH: MoClo scars flanking the full cluster
  - Column Z – AG: The detailed makeup of each full cluster in terms of monocistronic unit. The composition order reflects the gene order in the genetic design.

Worksheet 3 (3. Exp,Titer, Morph):

This worksheet describes the expression strength of individual genes and dummy variables, eAA production titer data, and colony morphology features of each PBFC.

Column A: Plackett-Burman full cluster name

Column B – H: expression level of individual genes corresponding to the promoter-RBS parts controlling the expression. The genetic design is reflected in Worksheet 2 Column Z – AG.

Column J – AF: The levels of dummy variables dv1 through dv23. These columns are used as negative control in the experimental error calculation.

Column AG, AH: The composition of PBFC in terms of partial clusters

Column AI – AL: eAA production titer measurement of each PBFC strain.

Column AI, AJ: biological duplicate measurement of eAA titer

Column AK, AL: average value of eAA titer at linear scale and $\log_{10}$ scale

Column AM: FC relative fitness. This value is calculated by taking average $\log_{10}$ eAA titer of each PBFC strain and divide by the average $\log_{10}$ eAA titer of the overall PBFC library.

Column AN – AT: Morphological class and features of PBFC strains.

Column AN: Morphological classes

Column AO – AT: Morphological features defining each morphological class. The features are colony shape, margin, color, opacity, elevation, and texture.

Worksheet 4 (4. DNA Parts and Vectors):

This worksheet describes promoter-RBS, terminator, and CDS parts and vectors used to construct PBFCs.

Worksheet 5 (5. Fitness Calculation):

This worksheet contains fitness value of monocistronic units and partial clusters.

Column A – C: The column structure is identical to worksheet 2 Column R- V, excluding the scar information. The partial cluster name is replaced with its calculated fitness value.

Column D - F: Predicted fitness value of full clusters based on partial cluster fitness values using additive, multiplicative, or dominant negative model.

Column H – P: The column structure is identical to worksheet 2 Column X- AH, excluding the scar information. The monocistronic unit name is replaced with its calculated fitness value.

Column Q – S: Predicted fitness value of full clusters based on the fitness value of monocistronic units using additive, multiplicative, or dominant negative model.

Worksheet 6 (6. Morphology Description):

Column A- G: Text-based description of morphology features of each morphology class.

Column I – O: Discretized, numerical representation of morphological features of each morphology class.