NORTHWESTERN UNIVERSITY

Theoretical Considerations and Computational Analysis of the Complexity in Polyketide Synthesis Pathways

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Chemical Engineering

By

Joanna González-Lergier

EVANSTON, ILLINOIS

June 2007

ABSTRACT

Theoretical Considerations and Computational Analysis of the Complexity in Polyketide Synthesis Pathways

Joanna González-Lergier

The emergence of antimicrobial resistance and the growing concern to produce new drugs have influenced an increase in the amount of research directed towards the engineering of novel polyketides. The polyketide carbon backbone is synthesized by a set of enzymes known as polyketide synthases (PKSs). These catalyze the formation of a linear chain and its subsequent cyclization and thus control the variables in the synthesis, and a change in the organization of the PKS leads to the synthesis of a different polyketide structure. Although only 10,000 polyketide structures have been discovered to date, the theoretical analysis of the mechanism for the formation of the chain performed suggests that over a billion possible linear structures can be synthesized.

The complexity in the number of structures led to the implementation of this system in *Biochemical Network Integrated Computational Explorer* (BNICE), a computational framework that is being developed for the study of cellular reaction networks. This formulation allowed the analysis of the evolution of diversity in the synthesis mechanism and the construction of the pathway architecture of polyketide biosynthesis. However, the original framework utilized graph theory to represent reactions as a series of matrix operations, an approach that did not distinguish between stereochemical isomers. Since enzymes are known for their stereoselective catalysis, the framework was expanded to differentiate stereoisomers as distinct structures and

for the specification of the stereochemistry obtained through a reaction. Consequently, the framework can be used to identify all the possible polyketide structures that can theoretically be produced as well as the corresponding PKS organization required to synthesize each of the structures.

The feasibility of the implementation of polyketide synthesis in *Escherichia coli* as a heterologous host was assessed through the use of metabolic flux analysis in a genome-scale model of *E. coli*, showing that there is a complex interplay between cellular energetics, oxygen uptake rates and polyketide production yield. Consequently, this framework can be used to assess the cellular feasibility for the synthesis of novel polyketides, thereby guiding metabolic engineering actions to produce a potential therapeutic agent.

ACKNOWLEDGEMENTS

The completion of this project would not have been possible without the invaluable guidance and support of my advisors, Professors Linda J. Broadbelt and Vassily Hatzimanikatis, for which I am very grateful and appreciative. I would also like to thank all of the members of the Hatzimanikatis and Broadbelt labs throughout my five years at Northwestern for their collaboration, criticism and motivation.

The members of my committee, Professors Annelise Barron, Lonnie Shea and Thomas Meade, were also instrumental in my progress towards the completion of the requirements for a Ph.D. degree, and I would like to give them my gratitude for the generous donation of their time and their helpful evaluation of my work.

In addition, I would like to primarily thank my family and friends for their unwavering support throughout the years and for encouraging me to pursue a graduate degree. I would not be at Northwestern University without them and I therefore dedicate this work to them.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Introduction

Polyketides are cellular metabolites that are widely used in human and veterinary medicine, agriculture and animal nutrition and include a number of major pharmaceutical agents, such as the antibiotics erythromycin and the anti-tumor agent rapamycin [1, 2]. In fact, these metabolites are comprised of a large structural diversity, leading to a wide variety of pharmacological uses.

1.1 Motivation

The rise of antimicrobial resistance has led to a growing demand for the development of new drugs and vaccines, an interest that has motivated the study of polyketides. The past decade has revealed a striking increase in knowledge concerning the function and synthesis of polyketides, which is catalyzed by a set of enzymes that regulate the formation of the polyketide carbon backbone. These enzymes direct the length and structure of the chain through control of the number, identity and order of the monomer units and the sequence and amount of reduction used in its synthesis [3-7]. Experimental approaches have shown that mutations to the enzyme sequence leads to the synthesis of a different polyketide structure [3-7]. Combinatorial metabolic engineering has been applied to the generation of polyketide libraries, leading to the identification of an erythromycin library comprised of approximately 60 structures [8, 9]. This library, however, does not include all the possible mutations to the polyketide enzymes.

Therefore, a systematic way of identifying all the theoretically possible polyketide structures would prove useful in guiding future metabolic engineering approaches for the synthesis of polyketides with potential antibiotic properties.

1.2 Outline of Research

Polyketide synthesis is a complex biochemical process involving the formation of a linear chain through a series of elongation and reduction steps, as illustrated in Chapter 2. Consequently, a theoretical analysis, presented in Chapter 3, based on the synthesis mechanism for polyketide biosynthesis was performed to identify the variables in the synthesis of these metabolites, as well as the number of theoretically possible polyketide structures as a function of these variables. The large number of theoretical structures led to the implementation of the polyketide synthesis pathway in the *Biochemical Network Integrated Computational Explorer* (BNICE) framework, a computational framework that was developed to explore the chemistry in biological networks, leading to the study of the evolution of diversity in polyketide synthesis, as illustrated in Chapter 3. The structures identified by the framework were classified by the number of elongation steps required in their synthesis, allowing the identification of the pathway lengths required to produce structures of various lengths and degrees of reduction. Furthermore, the pathway architecture of the polyketide synthesis mechanism was constructed.

This analysis performed using the BNICE framework did not distinguish between stereochemical isomers, a necessary feature when studying biological networks. Therefore, the framework was modified, as shown in Chapter 4, in order to specify and manipulate the stereochemical information of a structure. This modified framework was then used to generate a complete library of polyketide structures and to identify their corresponding synthesis pathways,

as presented in Chapter 5. The synthesis reactions of these structures were then studied in order to evaluate their feasibility under cellular constraints, using metabolic flux analysis, as shown in Chapter 6.

Chapter 2

Background

The biosynthesis of polyketides is a complex biological process, first involving the formation of a linear chain followed by its cyclization. The synthesis of these metabolites is often contrasted with fatty acid biosynthesis. In both of these processes, the carbon backbone of the final structure is synthesized through successive Claisen condensations of acyl coenzyme A monomers, as shown in Figure 2.1 [10]. However, in contrast to fatty acid synthesis where the β-ketone group of the growing linear chain is fully reduced after each condensation reaction, the linear poly-β-ketone intermediate in polyketide biosynthesis may undergo no, partial or full reduction of the β-carbon during each elongation step. The presence of reduction during the elongation steps dictates the main group to which the polyketide belongs and, consequently, its final structural properties. As shown in Figure 2.1, the linear chain of aromatic polyketides does not undergo any reduction, whereas the linear chain of reduced polyketides experiences a series of reductions throughout its synthesis [1, 2, 11].

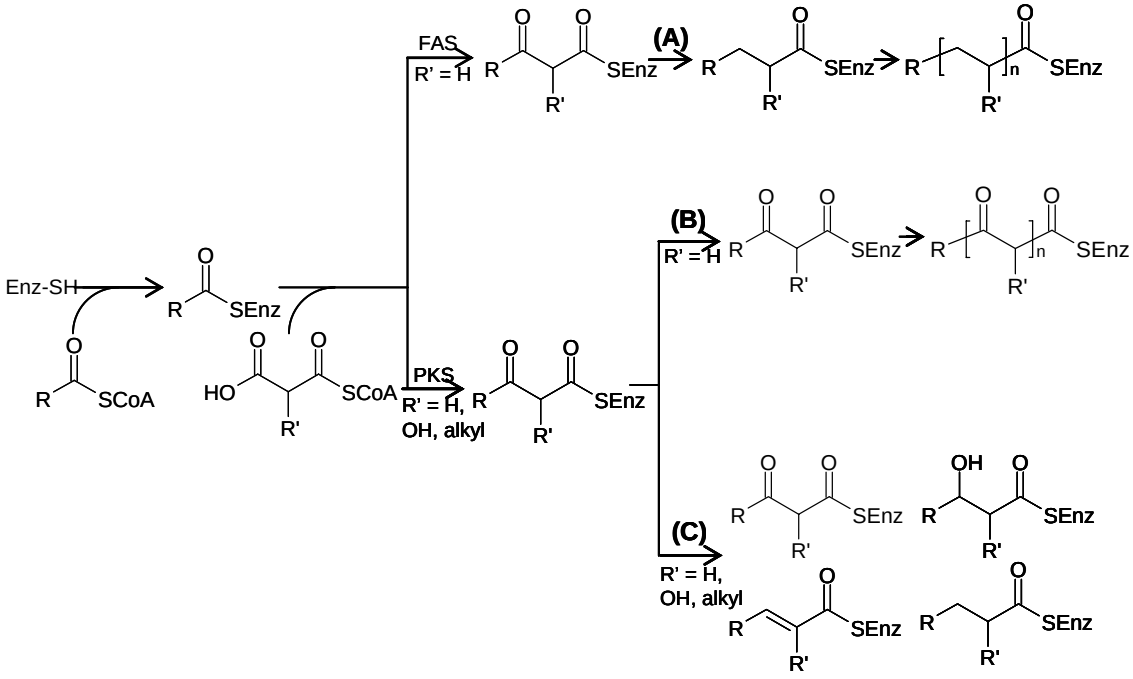


Figure 2.1. Pathways for polyketide and fatty acid biosynthesis. The general pathway separates into three distinct paths depending on the degree of reduction during each elongation step and on the structure of the extender molecules. The path labeled A leads to the complete reduction of the β-carbon during each elongation step, forming fatty acids. The paths labeled B and C, on the other hand, form polyketides; path B does not undergo any reduction during the elongation steps, forming aromatic polyketides, whereas path C allows no, partial or full reduction during the different elongation steps, leading to the formation of reduced polyketides.

Polyketide synthesis is catalyzed by proteins known as polyketide synthases, or PKSs, which are divided into two types depending on their structure and function. Type I PKSs are large, multifunctional proteins consisting of discrete active sites, each of which is responsible for one of the synthesis reactions. On the other hand, type II PKSs are complex structures involving the interaction of discrete monofunctional proteins that act sequentially throughout the synthesis, transferring the linear intermediate from active site to active site [1, 2, 10].

2.1 Aromatic Polyketides

Aromatic polyketides are distinguished by their polycyclic aromatic structures. Their synthesis is characterized by a lack of reduction of the β-carbon after each of the condensation reactions occurring between the growing linear intermediate and the extender units [2, 5].

2.1.1 Aromatic Polyketide Synthesis

In bacteria, aromatic polyketide synthesis is catalyzed by type II PKSs. Depending on the individual genetic sequence for this enzyme complex, the active sites that comprise the PKS may vary. However, a set of sites common among type II PKSs has been identified. The minimal PKS is believed to consist of a heterodimer formed by the interaction of a β-ketoacyl synthase/acyltranferase (KS/AT), or KS_α unit, and a chain length factor (CLF), referred to as the KS_β unit [12, 13]. The KS/AT enzyme complex catalyzes the condensation reactions that occur between the extender units and the growing linear chain and transfers the growing linear chain from the acyl carrier protein (ACP), which is responsible for bringing the extender units into close proximity with the heterodimer, to the KS_α unit [2].

The formation of the linear chain in aromatic polyketide synthesis consists of the consecutive addition of malonyl-CoA molecules, the extender units in the production of aromatic polyketides, to the growing linear intermediate, leading to the formation of a linear poly-β-ketone chain [2]. As shown in Figure 2.2, the synthesis is initiated with the priming of the enzyme complex through interaction between a malonyl-ACP complex and an unoccupied KS/AT-CLF heterodimer. The malonyl group then undergoes decarboxylation to form acetyl-ACP, which is subsequently transferred to the cysteine residue of the KS, at which point the ACP is released.

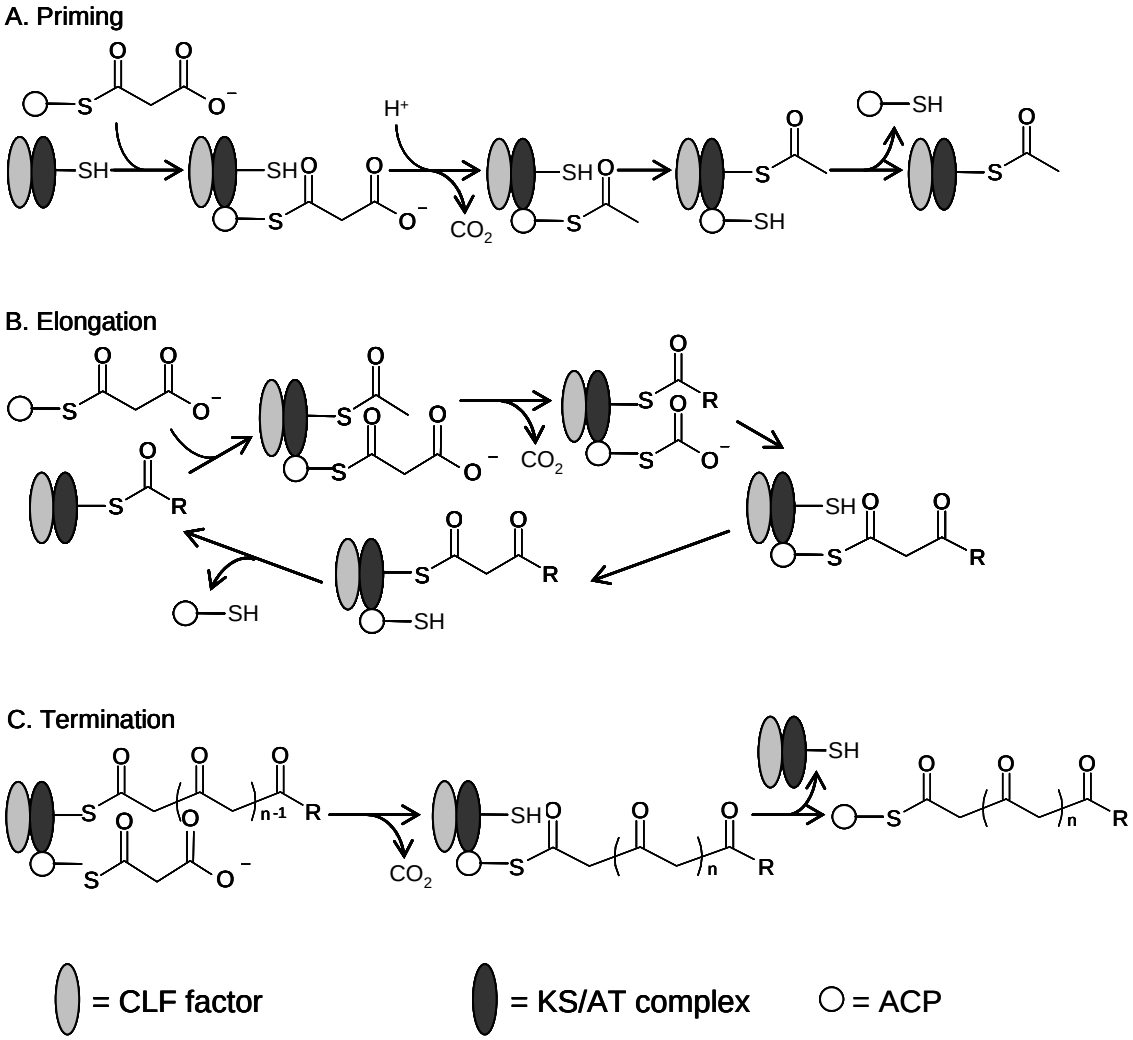


Figure 2.2. Pathway for the synthesis of the aromatic polyketide linear chain. The pathway is divided into three steps: priming of the enzyme complex (A), elongation of the chain (B) and termination of the synthesis (C).

The first chain elongation step involves a condensation reaction between the acetate starter unit formed during the priming step and a second malonyl-ACP extender unit attached to the heterodimer. As a result of the condensation reaction, the chain grows by two carbons and a molecule of carbon dioxide is released. The newly extended linear chain is then transferred from the ACP to the cysteine residue of the KS and the empty ACP is released. This series of steps,

involving a reaction between the growing linear chain and a malonyl-ACP extender unit, is repeated during each elongation step of the synthesis [13].

The length of the linear poly-β-ketone chain is controlled by the KS_β unit of the heterodimer; it is hypothesized that, since the KS/AT-CLF structure resembles a binding pocket, chain elongation terminates when the size limit of the pocket is reached by the growing linear chain [2, 6, 13]. After the last condensation reaction, termination involves the detachment of the full-length poly-β-ketone linear intermediate from the heterodimer. This final step of the synthesis is the rate-limiting step of the biosynthesis [13].

Based on the synthesis pathway, it is expected that the presence of malonyl-CoA:ACP malonyl transferase (MAT), an enzyme responsible for the activation of malonyl-CoA to malonyl-ACP, is necessary, since malonyl-ACP is required for chain elongation. However, this enzyme is not encoded in the minimal set of type II PKS genes. Its importance has therefore been studied, leading to the conclusion that MAT is required at limiting or equal concentrations of active ACP to KS/AT complex, while at high concentrations, active ACPs are capable of self-malonylation in the presence of malonyl-CoA [2, 13].

The linear poly-β-ketone chain synthesized during aromatic polyketide synthesis undergoes a number of cyclizations to achieve its characteristic polycyclic aromatic structure. These cyclization reactions, although some happen spontaneously, are often the result of additional PKS subunits, consisting of a series of ketoreductases, cyclases and aromatases [2]. While the PKS complex is responsible for the general structural variation among polyketides, these post-PKS enzymes produce an even greater molecular diversity [10, 12]. These tailoring enzyme actions, including oxidative changes and the addition of residues through esterification and

alkylation are responsible for the final structural and functional changes to the polyketide molecules, which often result in increased biological activity [1].

2.1.2 Engineering of Novel Aromatic Polyketides

A combinatorial substitution of naturally occurring genes to create unnatural combinations results in the production of type II PKSs capable of directing the synthesis of novel aromatic polyketides [12]. Based on the linear chain synthesis mechanism, there are a number of variables that can be modified in order to alter the final linear structure. Although most aromatic polyketides use acetate as a starter unit, cases have been observed where propionyl-CoA and malonamyl-CoA are chosen during the priming step. Furthermore, engineered aromatic polyketides have so far been synthesized with chains lengths in the range of ten to 24 carbons [2, 10].

However, the majority of the diversity in the structure of aromatic polyketide is obtained through the cyclization and post-PKS enzymes, which direct the regiospecific cyclization of the linear chain and the final modifications to the polycyclic aromatic structure; therefore, variety in engineered aromatic polyketides can be achieved through the deletion of these tailoring enzymes. An experimental approach introduced expression plasmids coding the genetic information for a naturally-occurring minimal PKS in the absence of its corresponding cyclases into an engineered host that had its original PKS genetic information deleted. This approach produced over thirty different polyketide structures, eight of which were fully studied and are shown in Figure 2.3 [12].

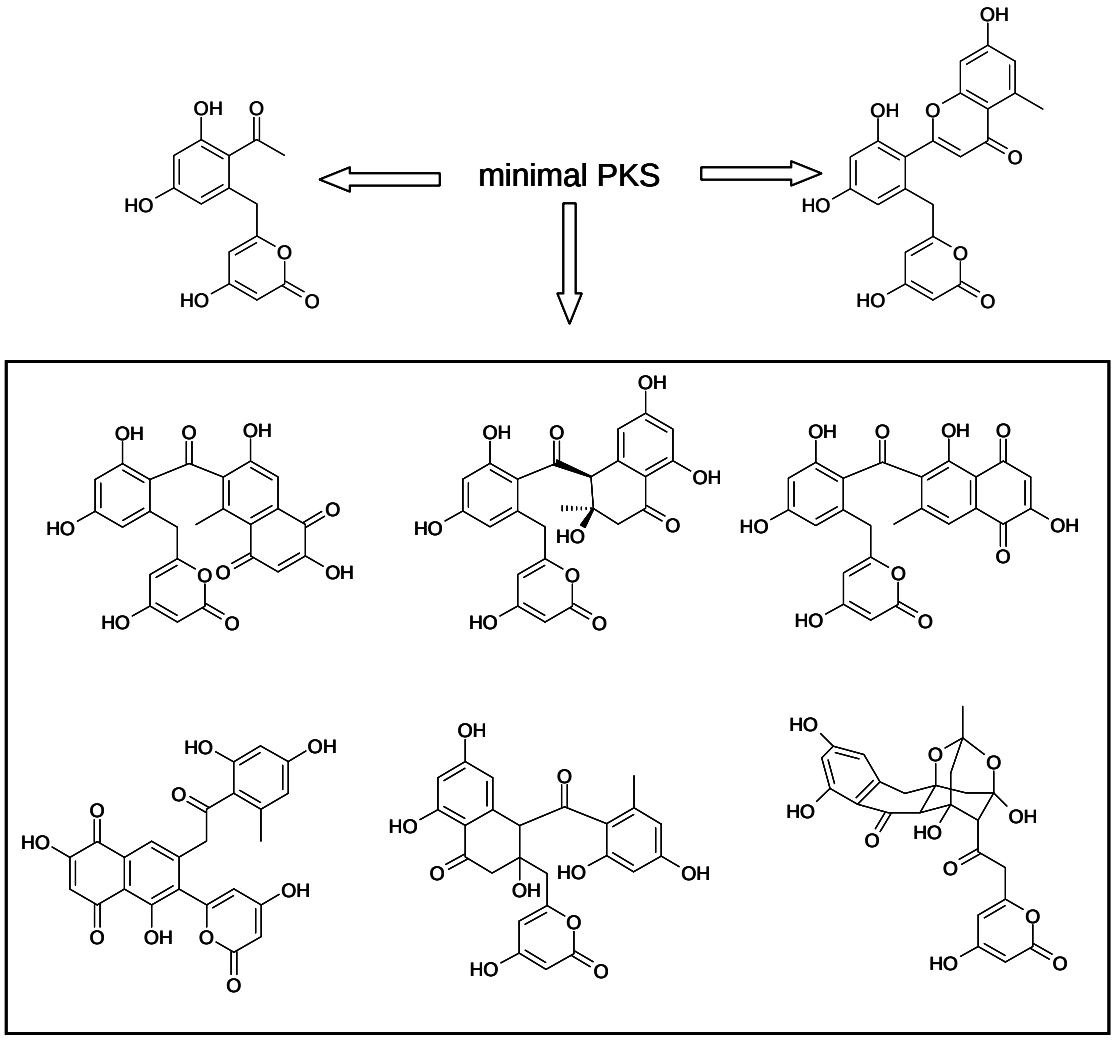


Figure 2.3. Novel polyketide structures obtained with a minimal PKS in the absence of the corresponding cyclase enzymes. The structure on the bottom right forms part of a novel structural polyketide class.

The structures produced in the experiment incorporated a number of chain lengths, ranging from heptaketides (14 carbons) to dodecaketides (24 carbons). Furthermore, the cyclization of these structures was spontaneous, leading to a large variety in ring structures, including a new structural class of polyketides which had not been observed previously in nature [12].

2.2 Reduced Polyketides

The synthesis of reduced polyketides is similar to aromatic polyketide synthesis. The main difference occurs after each condensation reaction in the elongation steps, at which point the linear molecule in reduced polyketide synthesis can undergo partial or full reduction of the β-carbon [2, 14]. Consequently, the β-carbon during each of the elongation steps is capable of having any of four different bonding arrangements: a carbonyl group, a hydroxyl group, an enoyl function or a methylene function [5]. This ability to undergo reduction leads to the formation of reduced, or complex, polyketides.

2.2.1 Reduced Polyketide Synthesis

In addition to aromatic polyketides, bacteria are capable of synthesizing reduced polyketides, which are catalyzed by large, multifunctional proteins classified as type I PKSs. As illustrated in Figure 2.4, type I PKSs are made up of a series of modules, each of which contains all the enzymatic active sites necessary for one elongation step [1, 10]. Reduced polyketides are synthesized through the transfer of the growing linear chain from one active site on the type I PKS to the next. These active sites are responsible for the enzymatic activities performed by the KS/AT complex in type II PKSs, as well as the reduction steps that can occur during synthesis, which are achieved through the presence of ketoreductase (KR), dehydratase (DH) and enoyl reductase (ER) domains [1, 5].

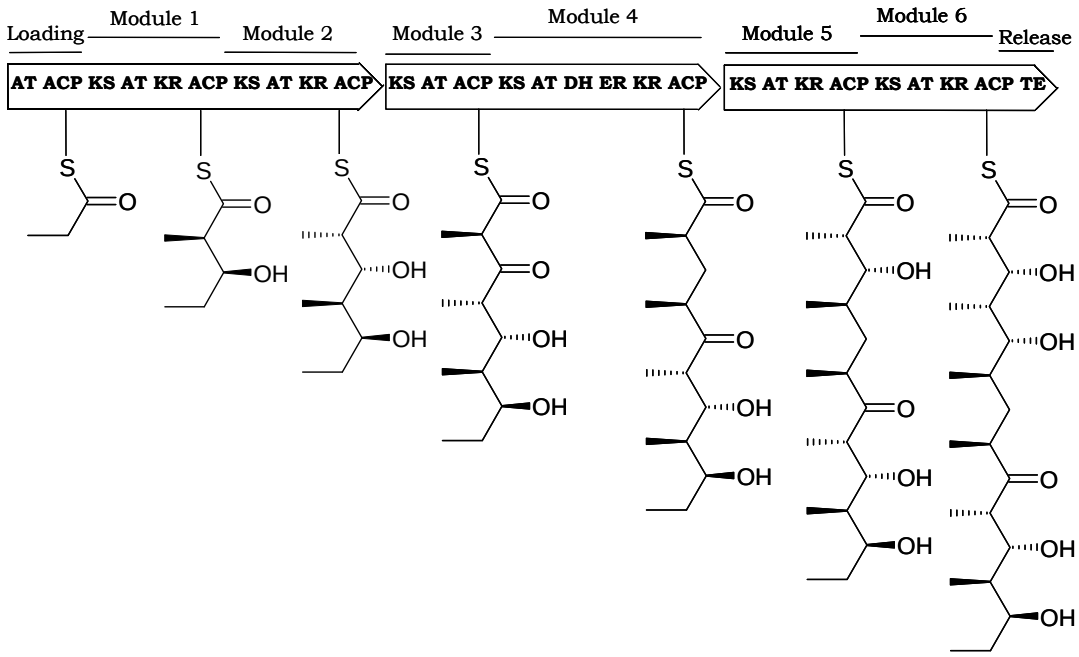


Figure 2.4. Pathway for the synthesis of the reduced polyketide linear chain. The extent of reduction achieved during each elongation step is the result of the presence of the reducing enzymatic active sites in the different modules.

The type I PKS shown in Figure 2.4 has one loading, six elongation and one release module. Synthesis starts by the transfer, catalyzed by the first AT domain, of the starter unit to the ACP site of the loading module. Then, the first elongation step, catalyzed by the KS and AT enzymes of module 1, takes place. The first condensation reaction, due to the presence of a KR active site in the first module as shown in the figure, is followed by a reduction of the β-carbon from a carbonyl to a hydroxyl group using $NADPH_2^{+1}$ as a hydrogen donor. The choice of malonyl-CoA as the extender unit during this elongation step is due to the specificity of the AT site in module 1. The second module contains the same enzymes as the first; consequently, the same reactions occur, extending the chain by two carbons and reducing the β-carbon of this elongation

[1]The formulation $NADPH_2^+$ is used for NADPH to account for the balance of charge and hydrogen atoms that are transferred during the reaction: $NADP^+ + 2H \leftrightarrow NADPH_2^+$. This formulation is used throughout the text and is extended to $NADH_2^+$ and $FADH_2^+$.

step to a hydroxyl group. Module 3 does not contain the KR domain; the chain does not undergo any reduction during the third elongation step. The fourth module, on the other hand, contains DH, ER and KR active sites, which are responsible for the β-carbon undergoing full reduction during the fourth elongation step of the synthesis. The dehydratase (DH) domain is responsible for release of the hydroxyl group formed during the first reduction step to form a double bond at the β-carbon. The enoyl reductase (ER) site catalyzes the second reduction that can occur during an elongation step, which results in the reduction of the enoyl functionality of the β-carbon to a methylene function using $NADPH_2^+$ as a hydrogen acceptor. The dehydration reaction cannot occur without a previous ketoreduction, and enoyl reduction cannot take place without a prior dehydration [10]. The fifth and sixth modules contain the same enzymes as the first and second modules; the β-carbons undergo partial reduction during these two steps before the linear chain is released from the PKS through a thioesterase (TE). Based on this synthesis mechanism, the oxidation level and stereochemistry of the β-carbon of the growing linear chain are determined during the corresponding chain elongation step, leading to the formation of a linear chain with a combination of fully-reduced, partially-reduced and unreduced carbons throughout its structure [1, 3, 4].

2.2.2 Engineering of Novel Reduced Polyketides

The structural diversity of reduced polyketides results from the controlled variation in the choice of starter unit, the number and type of extender units, the sequence of reduction and the stereochemistry obtained during each elongation cycle [10]. Variations in these variables can be achieved through manipulation of the genes that dictate the modular assembly of the type I PKS [3]. The AT sites of each of the modules is responsible for the choice of starter or extender unit

used during each step of the synthesis; if the extender molecule is chiral, these domains also determine the stereochemistry of the α-carbon during each elongation step. Type I PKSs have been observed to utilize a broad range of substrates as starter units, including acetyl-CoA, propionyl-CoA, isobutyryl-CoA, and isovaleyl-CoA. Similarly, malonyl-CoA, methyl-malonyl-CoA, ethylmalonyl-CoA and hydroxyl-malonyl-CoA, among others, have been used as extender units [3, 5]. In addition, the presence of the KR, DH and ER domains in each module dictates the amount of reduction that occurs during each elongation step. The length of the chain is dictated by the number of modules in the PKS before the release module.

One of the most investigated polyketides is erythromycin A, the linear chain of which is derived from one molecule of propionyl-CoA and six molecules of methylmalonyl-CoA, as illustrated in Figure 2.5. Isolated in 1952, erythromycin A is used against Gram-positive bacterial infections and as treatment for pulmonary diseases such as Legionnaire's disease [1]. Its importance in the medical profession has led to a number of directed genetic manipulations of its PKS genes in order to obtain a variety of different polyketide structures. Figure 2.5 shows a number of cases in which directed mutagenesis of the erythromycin A PKS led to the synthesis of novel structures.

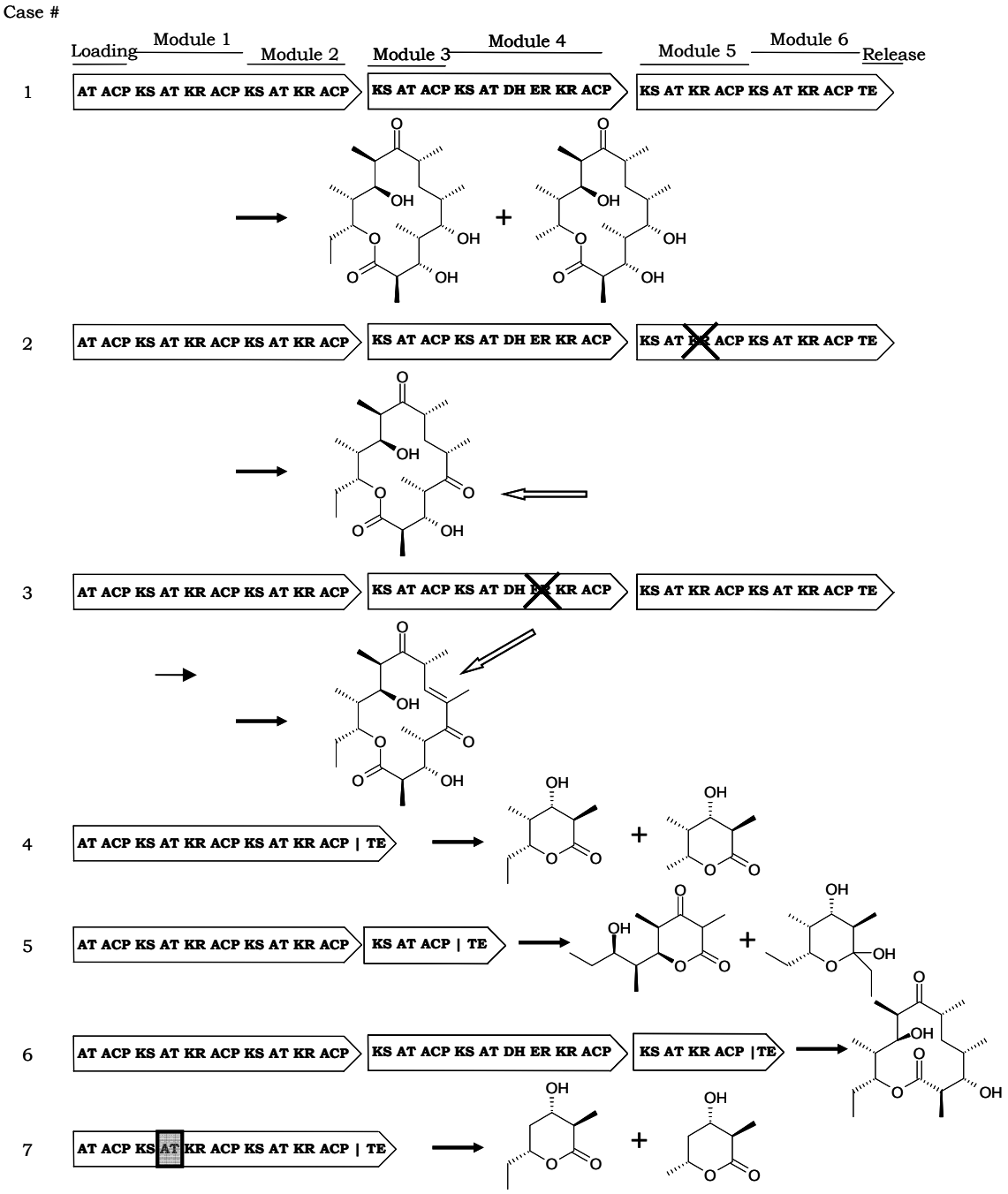


Figure 2.5. Mutagenesis of the erythromycin A PKS genes. Case 1 shows erythromycin A synthesis (a); cases 2 through 7 illustrate the synthesis of novel polyketides.

The first case in Figure 2.5 shows the sequence of reduction directed by the unaltered erythromycin PKS, leading to the synthesis of the antibiotic erythromycin A and an analogue

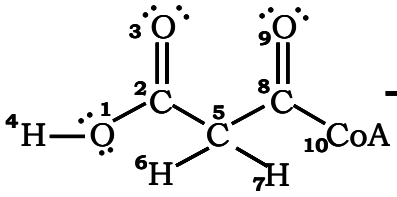
with an acetate starter. The second and third cases illustrate the deletion of an active site from the original erythromycin A PKS, which leads to decreased reduction during synthesis. Specifically, case 2 shows the deletion of the KR active site in module 5; therefore, the β-carbon during the fifth elongation step does not undergo any reduction and remains as a carbonyl group in the final polyketide structure obtained in the synthesis, as shown in the figure. Similarly, the third case involves the deletion of the ER domain in the fourth module, resulting in the β-carbon in the fourth elongation step not being fully reduced to a methylene functionality; thus, a double bond appears in the final polyketide structure in place of a single bond, as shown [3-5].

The fourth, fifth and sixth cases involve the deletion of modules from the original erythromycin A PKS, accomplished through the insertion of a TE active site at the end of a module to release the linear chain before it can reach the remaining modules. Case 4 involves the addition of TE at the end of the second module, resulting in only two chain extensions during synthesis and a final polyketide molecule with only nine carbons instead of the twenty carbons of erythromycin A. Similarly, in case 5, where the TE domain is added after module 3, a twelve-carbon polyketide structure is produced, as shown in the figure. The sixth case shows the addition of TE after the fifth module, forming a sixteen-carbon polyketide [3-5].

The seventh case, in addition to the deletion of modules 3 through 6, involves the replacement of the original AT domain in module 1 with an AT domain from a different PKS gene, a mutation that results in the use of a different extender molecule during the first chain elongation. In other words, the altered AT site chooses malonyl-CoA as an extender unit instead of methylmalonyl-CoA and, as shown in Figure 2.5, the lactone synthesized differs from the one produced in case 4 through the absence of a methyl group [3-5].

2.3 The BNICE Formalism

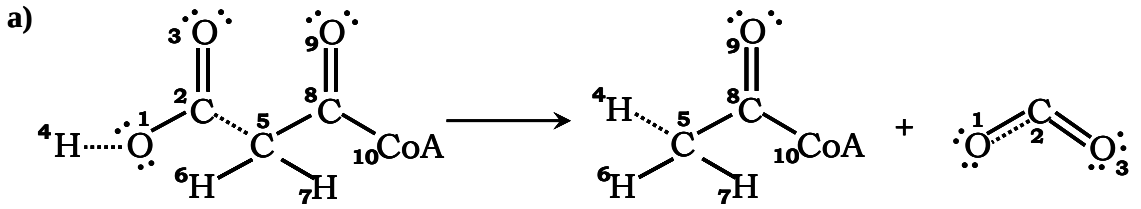
Automatic network generation has emerged as a tool to create and analyze complex reaction networks. A computational framework called NetGen, developed by Broadbelt and collaborators, has been applied to diverse problems including silicon nanoparticle production, tropospheric ozone formation and lubricant oxidation [15-18]. This framework utilizes graph theory to represent reactions as a series of matrix operations. Each substrate and product can be characterized by a unique matrix termed the bond-and-electron, or BE matrix, which denotes the atomic connectivity of the molecule. The BE-matrix is of size *n*-by-*n*, where *n* is the number of atoms or groups in the molecule [19]. As an illustrative example, the matrix representation of malonyl-CoA is shown in Figure 2.6. Malonyl-CoA is composed of nine atoms in addition to the CoA group; therefore, treating the CoA group as a single entity, the structure of malonyl-CoA can be represented by a 10-by-10 matrix. The {*i,j*} matrix elements represent the bond order of the bond between atoms *i* and *j*; similarly, the {*i,i*} elements correspond to the number of non-bonded electrons on atom *i*.



		1 O	2 C	3 O	4 H	5 C	6 H	7 H	8 C	9 O	10 CoA
1	O	4	1	0	1	0	0	0	0	0	0
2	C	1	0	2	0	1	0	0	0	0	0
3	O	0	2	4	0	0	0	0	0	0	0
4	H	1	0	0	0	0	0	0	0	0	0
5	C	0	1	0	0	0	1	1	1	0	0
6	H	0	0	0	0	1	0	0	0	0	0
7	H	0	0	0	0	1	0	0	0	0	0
8	C	0	0	0	0	1	0	0	0	2	1
9	O	0	0	0	0	0	0	0	2	4	0
10	CoA	0	0	0	0	0	0	0	1	0	0

Figure 2.6. Graph theory matrix representation of malonyl-CoA. The {*i,j*} and {*i,i*} matrix elements denote the bond arrangement and non-bonded valence electrons, respectively.

Reactions can also be represented using a reaction operator, the entries of which denote the changes in the connectivity and the change in the electronic configuration of the atoms of the reactant(s) to form the product(s). Addition of the reaction operator to the substrate matrix results in the product matrix for the reaction. Figure 2.10 shows the dehydrogenation of methanol using $NADP^+$ as a hydrogen acceptor, a reaction which is catalyzed by the enzyme alcohol dehydrogenase. The substrate matrix used in the matrix addition is not the full matrix shown in Figure 2.6; only those atoms which undergo connectivity changes or alterations in their electronic configuration during the reaction need to be included in the matrix addition operation. The product matrix formed after the addition operation is then inserted into the complete reactant matrix to determine the product(s) of the reaction.



		1 O	2 C	4 H	5 C
1	O	4	1	1	0
2	C	1	0	0	1
4	H	1	0	0	0
5	C	0	1	0	0

+

		1 O	2 C	4 H	5 C
1	O	0	1	-1	0
2	C	1	0	0	-1
4	H	-1	0	0	1
5	C	0	-1	1	0

=

		1 O	2 C	4 H	5 C
1	O	4	2	0	0
2	C	2	0	0	0
4	H	0	0	0	1
5	C	0	0	1	0

Figure 2.7. (a) Structural and (b) matrix representation of the decarboxylation of malonyl-CoA, catalyzed by the enzyme malonyl-CoA decarboxylase, E.C. 4.1.1.9.

The NetGen framework, originally intended for chemical systems, was recently adapted for biological systems. The new framework, referred to as *Biochemical Network Integrated Computational Explorer*, or BNICE, utilizes reaction operators to represent reactions that occur

in a cellular environment, most of which are catalyzed by enzymes [20]. Enzymes are classified by an Enzyme Commission number, or EC designation *i.j.k.l*, depending on the specific reaction that they catalyze [21]. The first number of this classification refers to the main class of the enzyme and identifies its primary action. The second characterizes the functional group the enzyme acts upon and the third identifies the cofactors or cosubstrates involved. The fourth number is specific to the substrate or set of substrates capable of interacting with the enzyme.

Taking advantage of this *i.j.k.l* designation, generalized enzyme functions can be formulated based on the first three levels of the enzyme classification system [22]. The generalized enzyme class would therefore have the same designation as the enzyme, with the exception of the fourth level classification; in other words, the generalized enzyme class would identify the main function, the target functional group and the cofactors involved in the reaction but would not specify the substrate. Each generalized enzyme function is then represented as a matrix operator; for example, the enzyme malonyl-CoA decarboxylase, EC number 4.1.1.9, forms part of the 4.1.1 generalized enzyme class, which has a unique 4.1.1 reaction operator. However, not all of the enzymes that belong to the same third level follow the same generalized rules; there are some exceptions to the generalized enzyme classes that are considered separately in order to accurately describe the varied chemistry found in the cell.

For a specific substrate, the BNICE framework identifies all its functional groups and applies all the generalized enzyme functions that can act on each of those functional groups, obtaining the set of all possible products. This methodology is then applied to this set of products, and the process is repeated until no new species are formed or a user-specified termination criterion is met [18].

Chapter 3

Theoretical Analysis of Polyketide Synthesis

The emergence of antimicrobial resistance has given rise to a growing concern for the development of new drugs, motivating the study of polyketides, cellular metabolites that are widely used in human and veterinary medicine, agriculture and animal nutrition [1]. Figure 3.1 shows a number of polyketides that, in addition to having large structural diversity, have a wide variety of pharmacological uses [23-25]. The synthesis of these metabolites is a complex biological process, involving the formation of a linear chain and its subsequent cyclization. The variables involved in the synthesis of the linear chain, such as the choice and sequence of monomer units and the degree and sequence of reduction, affect the final polyketide structure and are specifically controlled by the various enzymes that direct the synthesis [10]. Consequently, it would be useful to determine the number of possible structures that can be synthesized through manipulation of the different variables involved in the synthesis. A computational framework would allow a more thorough analysis of the complexity in the polyketide synthesis process. In addition, it would allow the identification of all the possible final polyketide structures, which would prove advantageous in identifying possible pharmacological targets; once targets are identified, their synthesis pathway can be determined,

thereby guiding the design of the metabolic actions required for production of the target molecule [4].

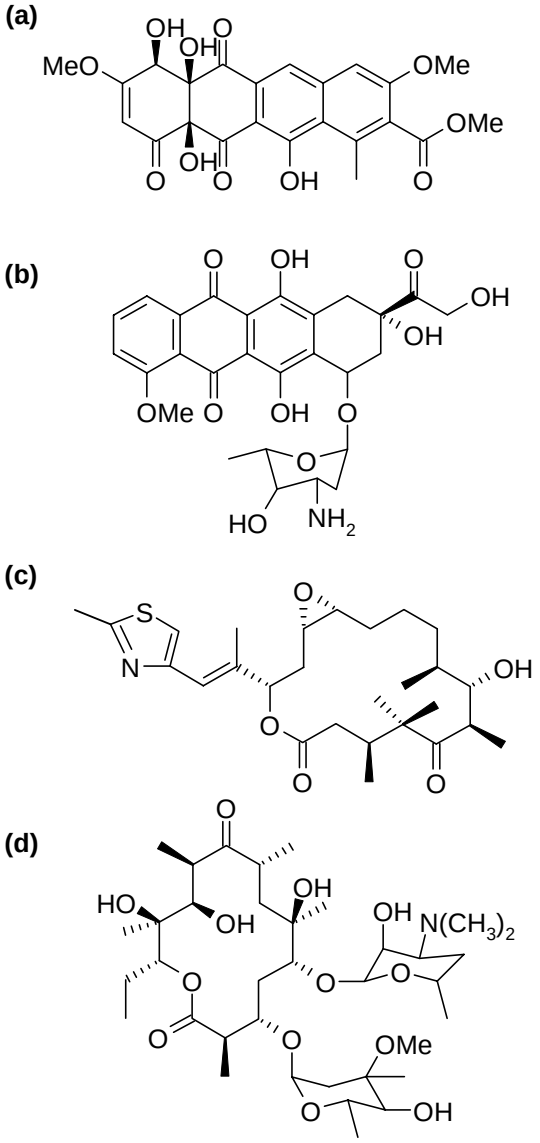


Figure 3.1 Polyketide structures. (a) Tetracenomycin C is an antibiotic; (b) doxorubibin is used as an anti-tumor agent; (c) epothilone A is an anti-cancer agent; and (d) erythromycin A is an antibiotic.

The carbon backbone of the polyketide structure is synthesized through successive Claisen condensations of acyl coenzyme A monomers, or chain elongation steps, a process catalyzed by

an enzyme complex termed polyketide synthase, or PKS. The PKS of reduced polyketides, specifically, consists of a series of modules each of which directs the synthesis of one elongation step. Each module is composed of a number of enzymes, including a ketosynthase, an acyl transferase, and an acyl carrier protein (ACP); additionally, it can include a ketoreductase, a dehydratase and an enoyl reductase enzyme for reduction [4, 8]. The synthesis originates through recruitment of the starter unit by the acyltransferase enzyme of the PKS loading module; the first module then catalyzes the first elongation step, and the chain is subsequently transferred through the remaining modules, which catalyze the remaining elongation steps [2, 4, 5]. After the full-length chain is synthesized, a number of cyclization and post-translational modifications synthesize the final polyketide molecule [2, 4, 5, 23, 26].

In an elongation step, as illustrated in Figure 3.2, the extender unit undergoes decarboxylation and is subsequently added to the growing linear chain, or the starter unit in the case of the first elongation step [13]. The resulting β-carbonyl can undergo reduction, using $NADPH_2^+$ as a hydrogen donor, to form a hydroxyl group. This reduction can be followed by a dehydration reaction, resulting in the β-carbon achieving enoyl functionality. A second reduction can then occur, reducing the β-carbon from enoyl to methylene functionality, through the use of $NADPH_2^+$. The first two reactions shown in Figure 3.2, which summarize the Claisen condensation, occur during every elongation step while the other three do not necessarily occur; however, if any of these three reactions do occur in an elongation step, they occur in the order described [1, 4, 5].

The structural diversity of polyketides is due in large part to the variation in their linear carbon backbones. The differences in the linear chain therefore result from the controlled

variation in the choice of starter unit, the number and type of extender units, the sequence and degree of reduction, and the stereochemistry obtained during each elongation cycle [3-5, 10, 27, 28], since changes to these variables lead to the synthesis of different polyketide structures.

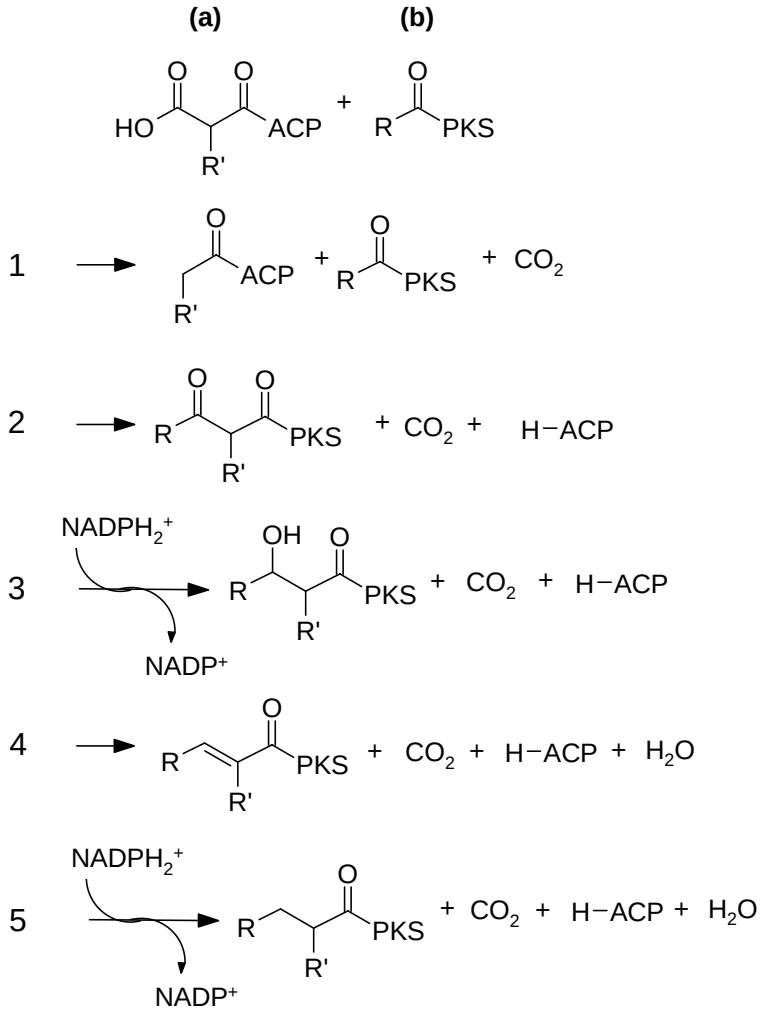


Figure 3.2. Elongation of polyketide linear chain. The substrates for the elongation reaction are the extender unit attached to an acyl carrier protein, or malonyl-ACP (a) as shown in this example, and the growing linear chain (b) attached to the PKS complex. The extender unit undergoes a decarboxylation, producing a molecule of carbon dioxide (1), and is then added to the growing linear chain releasing a free ACP (2), as shown in the first two reactions. The β-carbonyl can then undergo a reduction using $NADPH_2^+$ to form a hydroxyl group (3). This hydroxyl group can then be dehydrated, releasing a molecule of water, to achieve enoyl functionality (4). A second reduction can then occur, reducing the β-carbon to methylene functionality through the use of $NADPH_2^+$ (5).

Automatic network generation has emerged as a tool to create and analyze complex reaction networks. A computational framework called NetGen, developed by Broadbelt and

collaborators, has been applied to diverse problems including silicon nanoparticle production, tropospheric ozone formation and lubricant oxidation [15-18]. This framework utilizes graph theory to represent reactions as a series of matrix operations, where each substrate and product can be characterized by a unique matrix and each reaction by another matrix termed a reaction operator. The NetGen framework, originally intended for chemical systems, is currently being implemented for biological applications. In this new framework, referred to as *Biochemical Network Integrated Computational Explorer* (BNICE), the reaction operators are modified to correspond to reactions that occur in a cellular environment, which are generally catalyzed by enzymes [20]. The applications of this computational framework can be illustrated though the use of polyketide biosynthesis as a model system.

A theoretical analysis of polyketide synthesis, involving the use of stoichiometric balances and combinatorial theory, was employed to determine the effect of the different variables on the number of possible polyketides, resulting in over a billion possible structures. Implementation of the polyketide synthesis pathway in the BNICE framework allowed the identification of all of the possible structures. Since only 10,000 polyketides have been discovered experimentally to date, it is possible that the remaining predicted structures have not been discovered due to low yields, thus eliminating the possibility of detection, or, more importantly, they represent novel structures [1]. Additionally, the evolution of diversity in polyketide synthesis was studied. The structures identified from the application of the framework were classified by the number of elongation steps required in their synthesis, allowing the identification of the pathway lengths required to produce structures of various lengths and degrees of reduction. Furthermore, the pathway architecture of the polyketide synthesis mechanism was constructed. Based on this architecture,

the order and identity of the modules of the PKS responsible for the formation of any linear chain can be easily identified, thus guiding the implementation of the synthesis pathways for a potential antibiotic using metabolic engineering.

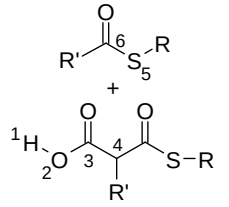
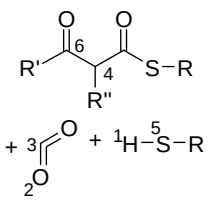
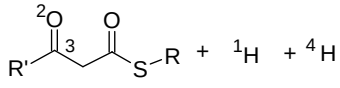
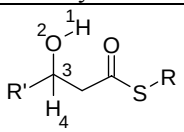
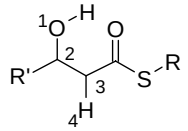
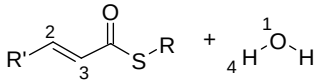
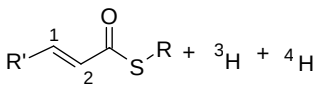
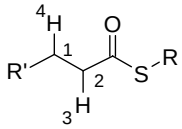
3.1 The BNICE Formalism

Enzymes are classified by an Enzyme Commission number, or EC designation *i.j.k.l*, depending on the specific reaction that they catalyze [21]. The first number of this classification refers to the main class of the enzyme and identifies its primary action. The second characterizes the functional group the enzyme acts upon and the third identifies the cofactors or cosubstrates involved. The fourth number is specific to the substrate or set of substrates capable of interacting with the enzyme. Taking advantage of this *i.j.k.l* designation, generalized enzyme functions can be formulated based on the first three levels of the enzyme classification system [22]. This generalized enzyme function representation makes it possible to study all the possible reactions that can occur in a cell with a variety of substrates. For a specific substrate, the BNICE framework identifies all its functional groups and applies all the generalized enzyme functions that can act on each of those functional groups, obtaining the set of all possible products. This methodology is then applied to this set of products, and the process is repeated until no new species are formed or a user-specified termination criterion is met [18].

The EC classification of the enzymes involved in polyketide biosynthesis is used to generate the list of generalized enzyme functions involved in polyketide biosynthesis, as listed in Table 3.1. As shown in the table, the ketosynthase/acyltransferase, which is responsible for the condensation reaction, can be divided into a series of two generalized enzyme actions; consequently, it is classified as belonging to the generalized enzyme class 4.1.1/2.3.1.

Additionally, the ketoreductase, the dehydratase and the enoyl reductase belong to the generalized enzyme classes 1.1.1, 4.2.1 and 1.3.1, respectively. Based on the reaction that a generalized enzyme class catalyzes, the functional group of the substrate(s) is identified; through the use of graph theory, a unique matrix representation of the functional group is constructed based on the bonding arrangement of the atoms in the group. An analogous representation of these atoms is created based on their bonding arrangement in the product molecule(s). The reaction operator is then determined by identifying the changes in the bond arrangement due to the reaction. The matrix representation of the functional group for the substrate and product, as well as the reaction operator, for the generalized enzyme functions involved in polyketide biosynthesis are illustrated in Table 3.1.

Table 3.1. Table of BNICE reaction operators for the generalized enzyme functions involved in the synthesis of the polyketide linear chain.

<table>
<tr><th>Reactant</th><th>Product</th><th>Reaction Operator</th></tr>
<tr><td colspan="3">Generalized Enzyme Function 4.1.1/2.3.1</td></tr>
<tr><td> 
<table>
<tr><th></th><th>1H</th><th>2O</th><th>3C</th><th>4C</th><th>5S</th><th>6C</th></tr>
<tr><td>1H</td><td>0</td><td>1</td><td>0</td><td>0</td><td>0</td><td>0</td></tr>
<tr><td>2O</td><td>1</td><td>4</td><td>1</td><td>0</td><td>0</td><td>0</td></tr>
<tr><td>3C</td><td>0</td><td>1</td><td>0</td><td>1</td><td>0</td><td>0</td></tr>
<tr><td>4C</td><td>0</td><td>0</td><td>1</td><td>0</td><td>0</td><td>0</td></tr>
<tr><td>5S</td><td>0</td><td>0</td><td>0</td><td>0</td><td>4</td><td>1</td></tr>
<tr><td>6C</td><td>0</td><td>0</td><td>0</td><td>0</td><td>1</td><td>0</td></tr>
</table></td><td> 
<table>
<tr><th></th><th>1H</th><th>2O</th><th>3C</th><th>4C</th><th>5S</th><th>6C</th></tr>
<tr><td>1H</td><td>0</td><td>0</td><td>0</td><td>0</td><td>1</td><td>0</td></tr>
<tr><td>2O</td><td>0</td><td>4</td><td>2</td><td>0</td><td>0</td><td>0</td></tr>
<tr><td>3C</td><td>0</td><td>2</td><td>0</td><td>0</td><td>0</td><td>0</td></tr>
<tr><td>4C</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>1</td></tr>
<tr><td>5S</td><td>1</td><td>0</td><td>0</td><td>0</td><td>4</td><td>0</td></tr>
<tr><td>6C</td><td>0</td><td>0</td><td>0</td><td>1</td><td>0</td><td>0</td></tr>
</table></td><td>
<table>
<tr><th></th><th>1H</th><th>2O</th><th>3C</th><th>4C</th><th>5S</th><th>6C</th></tr>
<tr><td>1H</td><td>0</td><td>-1</td><td>0</td><td>0</td><td>1</td><td>0</td></tr>
<tr><td>2O</td><td>-1</td><td>0</td><td>1</td><td>0</td><td>0</td><td>0</td></tr>
<tr><td>3C</td><td>0</td><td>1</td><td>0</td><td>-1</td><td>0</td><td>0</td></tr>
<tr><td>4C</td><td>0</td><td>0</td><td>-1</td><td>0</td><td>0</td><td>1</td></tr>
<tr><td>5S</td><td>1</td><td>0</td><td>0</td><td>0</td><td>0</td><td>-1</td></tr>
<tr><td>6C</td><td>0</td><td>0</td><td>0</td><td>1</td><td>-1</td><td>0</td></tr>
</table></td></tr>
<tr><td colspan="3">Generalized Enzyme Function 1.1.1</td></tr>
<tr><td> 
<table>
<tr><th></th><th>2O</th><th>3C</th><th>1H</th><th>4H</th></tr>
<tr><td>2O</td><td>0</td><td>2</td><td>0</td><td>0</td></tr>
<tr><td>3C</td><td>2</td><td>4</td><td>0</td><td>0</td></tr>
<tr><td>1H</td><td>0</td><td>0</td><td>1</td><td>0</td></tr>
<tr><td>4H</td><td>0</td><td>0</td><td>0</td><td>1</td></tr>
</table></td><td> 
<table>
<tr><th></th><th>2O</th><th>3C</th><th>1H</th><th>4H</th></tr>
<tr><td>2O</td><td>0</td><td>1</td><td>1</td><td>0</td></tr>
<tr><td>3C</td><td>1</td><td>4</td><td>0</td><td>1</td></tr>
<tr><td>1H</td><td>1</td><td>0</td><td>0</td><td>0</td></tr>
<tr><td>4H</td><td>0</td><td>1</td><td>0</td><td>0</td></tr>
</table></td><td>
<table>
<tr><th></th><th>2O</th><th>3C</th><th>1H</th><th>4H</th></tr>
<tr><td>2O</td><td>0</td><td>-1</td><td>1</td><td>0</td></tr>
<tr><td>3C</td><td>-1</td><td>0</td><td>0</td><td>1</td></tr>
<tr><td>1H</td><td>1</td><td>0</td><td>-1</td><td>0</td></tr>
<tr><td>4H</td><td>0</td><td>1</td><td>0</td><td>-1</td></tr>
</table></td></tr>
<tr><td colspan="3">Generalized Enzyme Function 4.1.2</td></tr>
<tr><td> 
<table>
<tr><th></th><th>1O</th><th>2C</th><th>3C</th><th>4H</th></tr>
<tr><td>1O</td><td>4</td><td>1</td><td>0</td><td>0</td></tr>
<tr><td>2C</td><td>1</td><td>0</td><td>1</td><td>0</td></tr>
<tr><td>3C</td><td>0</td><td>1</td><td>0</td><td>1</td></tr>
<tr><td>4H</td><td>0</td><td>0</td><td>1</td><td>0</td></tr>
</table></td><td> 
<table>
<tr><th></th><th>1O</th><th>2C</th><th>3C</th><th>4H</th></tr>
<tr><td>1O</td><td>4</td><td>0</td><td>0</td><td>1</td></tr>
<tr><td>2C</td><td>0</td><td>0</td><td>2</td><td>0</td></tr>
<tr><td>3C</td><td>0</td><td>2</td><td>0</td><td>0</td></tr>
<tr><td>4H</td><td>1</td><td>0</td><td>0</td><td>0</td></tr>
</table></td><td>
<table>
<tr><th></th><th>1O</th><th>2C</th><th>3C</th><th>4H</th></tr>
<tr><td>1O</td><td>0</td><td>-1</td><td>0</td><td>1</td></tr>
<tr><td>2C</td><td>-1</td><td>0</td><td>1</td><td>0</td></tr>
<tr><td>3C</td><td>0</td><td>1</td><td>0</td><td>-1</td></tr>
<tr><td>4H</td><td>1</td><td>0</td><td>-1</td><td>0</td></tr>
</table></td></tr>
<tr><td colspan="3">Generalized Enzyme Function 1.3.1</td></tr>
<tr><td> 
<table>
<tr><th></th><th>1C</th><th>2C</th><th>3H</th><th>4H</th></tr>
<tr><td>1C</td><td>0</td><td>2</td><td>0</td><td>0</td></tr>
<tr><td>2C</td><td>2</td><td>0</td><td>0</td><td>0</td></tr>
<tr><td>3H</td><td>0</td><td>0</td><td>1</td><td>0</td></tr>
<tr><td>4H</td><td>0</td><td>0</td><td>0</td><td>1</td></tr>
</table></td><td> 
<table>
<tr><th></th><th>1C</th><th>2C</th><th>3H</th><th>4H</th></tr>
<tr><td>1C</td><td>0</td><td>1</td><td>0</td><td>1</td></tr>
<tr><td>2C</td><td>1</td><td>0</td><td>1</td><td>0</td></tr>
<tr><td>3H</td><td>0</td><td>1</td><td>0</td><td>0</td></tr>
<tr><td>4H</td><td>1</td><td>0</td><td>0</td><td>0</td></tr>
</table></td><td>
<table>
<tr><th></th><th>1C</th><th>2C</th><th>3H</th><th>4H</th></tr>
<tr><td>1C</td><td>0</td><td>-1</td><td>0</td><td>1</td></tr>
<tr><td>2C</td><td>-1</td><td>0</td><td>1</td><td>0</td></tr>
<tr><td>3H</td><td>0</td><td>1</td><td>-1</td><td>0</td></tr>
<tr><td>4H</td><td>1</td><td>0</td><td>0</td><td>-1</td></tr>
</table></td></tr>
</table>

3.2 Theoretical Calculation of Number of Possible Polyketides

Since manipulation of the choice and sequence of monomer units and the degree and sequence of reduction lead to the synthesis of different structures, the total number of possible structures that can be produced was analyzed with respect to each of these variables. First, since the bonding arrangement of each of the β-carbons during each of the elongation steps is independent from the rest of the β-carbons, the total number of possible structures for the complete range of reduction is b^m, where b is the number of possible β-carbon configurations and m is the number of elongation steps. However, the amount and sequence of reduction is not the only variable involved in polyketide synthesis. The synthesis of reduced polyketides is capable of utilizing a number of different starter and extender units, depending on the acyltransferase enzyme that forms part of the PKS. Designating the number of different starter molecules available for polyketide synthesis by the variable s, the total number of possible linear structures that can be synthesized is $s \cdot b^m$, since the starter unit only affects the beginning of the chain. Unlike the starter unit, however, the choice of extender unit affects the bonding arrangement of the α-carbon in each elongation step. If the number of different extender molecules that can be used for polyketide synthesis is represented by the variable a, there are a total of a possible arrangements for the α-carbon; therefore, the total number of structures that can be produced is $s(ba)^m$. Additionally, some of these extender units, such as methylmalonyl-CoA and ethylmalonyl-CoA, produce chiral centers at the α-carbon thereby increasing the number of possible linear structures; this chirality is lost when the β-carbon has the double bond arrangement. Consequently, the total number of possible linear structures that can be synthesized can be calculated by the following formula:

$$N_{m,tot} = s\left[b(a - a_c) + (2b - 1)a_c\right]^m \quad (3.1)$$

where a_c is the number of different extender molecules that introduce stereochemistry to the linear chain.

Erythromycin, a polyketide naturally synthesized in *Saccharopolyspora erythrae*, is a commonly prescribed antibiotic. The cyclic precursor to erythromycin, 6-deoxyerythronolide (6dEB), is synthesized through six elongation steps using propionyl-CoA and methylmalonyl-CoA as the starter and extender units, respectively. For this synthesis, the corresponding values of b, s, a, a_c and m are 5, 1, 1, 1, and 6, respectively; consequently, over 100,000 possible structures can theoretically be produced in addition to 6dEB. Furthermore, a malonyl-CoA acyltransferase enzyme has been experimentally introduced into the erythromycin PKS [8, 9], increasing the number of possible extender units a that can be used in the synthesis to two; therefore, the theoretical number of possible structures increases by one order of magnitude to over three million possible structures. However, malonyl-CoA and methylmalonyl-CoA are not the only possible extender units; similarly, alternate values for the number of starter units and elongation steps are feasible. Based on naturally-occurring polyketides, realistic values of $s = 3$, $b = 5$, $a = 4$, $a_c = 3$ and $b = 6$ result in over 800 million possible linear polyketide structures, a number that continues to increase exponentially with respect to the number of elongation steps. Of these potential polyketide structures, only 10,000 have been identified, suggesting that a large number of new polyketide structures remain to be discovered [9].

3.3 Assessment of the Variation in Polyketide Structures

The previous theoretical analysis assumes an infinite source of carbon and reducing equivalents. However, the implementation of polyketide synthesis in a cell introduces

competition between polyketide synthesis and the native cellular metabolism for available carbon, energy and redox resources. Consequently, assuming that *m* extender units and *n* $NADPH_2^+$ molecules are available for synthesis, the resulting polyketide linear chain is formed through *m* elongation steps and *n* reduction reactions.

In general, the polyketide synthesis reaction is defined as

$$C_\varepsilon H_\kappa O\text{-PKS} + mC_\lambda H_\sigma O_3\text{-ACP} + n NADPH_2^+ \rightarrow C_\alpha H_\beta O_\gamma\text{-PKS} + cCO_2 + hH_2O + n NADP^+ + sH\text{-ACP}$$

where the starter unit consists of ε carbon, κ hydrogen and one oxygen atoms; *m* refers to the number of extender molecules, each of which is characterized by the presence of λ carbon, σ hydrogen and three oxygen atoms; *n* is the number of $NADPH_2^+$ molecules that act as hydrogen donors; α, β, and γ are the number of carbon, hydrogen and oxygen atoms in the resulting linear polyketide chain, respectively; *c* is the number of carbon dioxide molecules released during synthesis; *h* is the number of water molecules produced; and *s* is the number of unattached acyl carrier proteins.

A balance for each of the elements involved in the reaction gives:

$$\text{C:} \quad \lambda m + \varepsilon = \alpha + c \tag{3.2}$$

$$\text{H:} \quad 2n + \sigma m + \kappa = \beta + 2h + s \tag{3.3}$$

$$\text{O:} \quad 3m + 1 = \gamma + 2c + h \tag{3.4}$$

$$\text{ACP:} \quad m + 1 = s + 1 \tag{3.5}$$

The mechanism for the synthesis of the linear chain presented in Figure 3.2 led to the formulation of three additional constraints. During each elongation reaction, the extender molecule undergoes decarboxylation, releasing a molecule of carbon dioxide. Therefore, the number of malonyl-CoA molecules equals the number of carbon dioxide molecules produced:

$$m = c \tag{3.6}$$

Furthermore, since the maximum number of $NADPH_2^+$ molecules that can be used in each elongation step is two and the α-carbon in the linear chain after the last elongation step is not reduced, the number of $NADPH_2^+$ molecules is less than or equal to twice the number of malonyl-ACP molecules:

$$n \leq 2m \tag{3.7}$$

Similarly, only one water molecule is released during each elongation step and it can not be released without the first reduction step taking place; therefore, the number of water molecules is less than or equal to the minimum of the number of malonyl-ACP molecules or the number of $NADPH_2^+$ molecules:

$$h \leq \min(m, n) \tag{3.8}$$

In the case of erythromycin, for example, six methylmalonyl-ACP molecules are used as extender units and six molecules of $NADPH_2^+$ are used for reduction; however, theoretically, for six elongation steps, the number of molecules of $NADPH_2^+$ can range from zero to 12; therefore, the number of water molecules produced can range from zero to six, depending on the number of $NADPH_2^+$ molecules utilized in the synthesis. Consequently, a list of all the possible overall reactions that can occur can be generated.

However, since a set of values for *m*, *n* and *h* can represent more than one possible structure, the total number of reactions does not directly reflect the total number of linear structures that can be produced. For example, the synthesis of erythromycin uses six methylmalonyl-ACP molecules, six $NADPH_2^+$ molecules and produces one molecule of water; thus, the corresponding values of *m*, *n*, and *h* are 6, 6, and 1, respectively. These values describe the

presence of two carbonyl groups, four hydroxyl groups and one methylene group in the linear chain; thus, since one of the carbonyl groups is constrained to the final α-carbon, there are 30 different structures that can be produced, some of which are shown in Figure 3.3. Additionally, this set of values can also represent one carbonyl group, five hydroxyl groups and one double bond, adding six structures to the total number of structures that can be produced for this set of *m*, *n*, and *h* values.

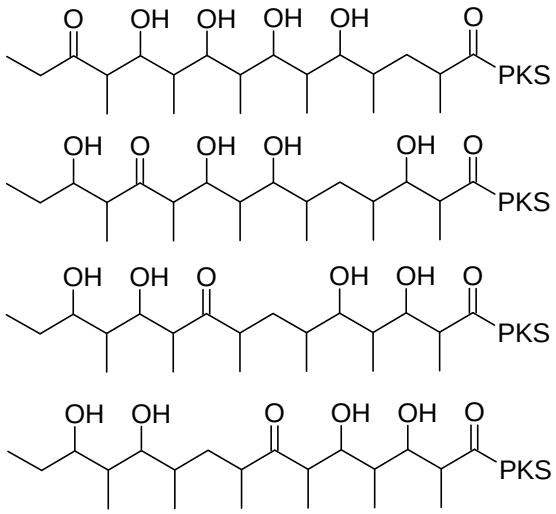


Figure 3.3. Four structures of the 30 possible polyketide linear chains with two carbonyl groups, four hydroxyl groups and one single bond, obtained with a synthesis requiring the use of one molecule of propionyl-ACP, six molecules of methylmalonyl-ACP and six molecules of $NADPH_2^+$ and producing one molecule of water.

From the formulated mechanism of linear chain synthesis shown in Figure 3.2, and without taking into account stereochemistry, the β-carbon during each elongation step can be transformed into any of four possible groups depending on the amount of reduction that the linear chain undergoes after each condensation reaction. The number of carbonyl groups (δ), hydroxyl groups (ζ), enoyl groups (θ) and methylene groups (ξ) on the linear structure, without taking into account the carbonyl group at the final α-carbon of the chain, all add up to *m*:

$$\delta + \zeta + \theta + \xi = m \tag{3.9}$$

Additionally, there can be more than one combination of β-carbon groups. Based on the values of *m*, *n*. and *h*, the number of carbonyl groups (δ), hydroxyl groups (ζ), double bonds (θ) and single bonds (ξ) are determined:

$$\begin{aligned} \delta &= m - n + i \\ \zeta &= n - h - i \\ \theta &= h - i \\ \xi &= i \end{aligned} \qquad n \le m, i \in [0, \min(n-h, h)] \tag{3.10}$$

and

$$\begin{aligned} \delta &= i \\ \zeta &= m - h - i \\ \theta &= h - (n-m) - i \\ \xi &= n - m + i \end{aligned} \qquad n > m, i \in [0, \min(h-n+m, m-h)] \tag{3.11}$$

where *i* is the number of combinations of β-carbon groups.

After the number of carbonyl groups (δ), hydroxyl groups (ζ), double bonds (θ) and single bonds (ξ) was determined, the total number of possible linear structures, *N*, for a combination of *m*, *n*, *h* and *i* values is calculated as

$$N_{m,n,h,i} = N(\delta_{m,n,h,i}) \cdot N(\zeta_{m,n,h,i}) \cdot N(\theta_{m,n,h,i}) \cdot N(\xi_{m,n,h,i}) \tag{3.12}$$

where the numbers of structures for each of the four different groups are calculated based on combinatorial theory:

$$N_{m,n,h,i}(\delta) = \frac{(m)!}{\delta!(m-\delta)!} \tag{3.13}$$

$$N_{m,n,h,i}(\zeta) = \frac{(m-\delta)!}{\zeta!(m-\delta-\zeta)!} \tag{3.14}$$

$$N_{m,n,h,i}(\theta) = \frac{(m-\delta-\zeta)!}{\theta!(m-\delta-\zeta-\theta)!} \tag{3.15}$$

$$N_{m,n,h,i}(\xi) = \frac{(m-\delta-\zeta-\theta)!}{\xi!(m-\delta-\zeta-\theta-\xi)!} = \frac{(m-\delta-\zeta-\theta)!}{\xi!} \tag{3.16}$$

Therefore, the equation for the total number of possible structures for a set number of elongation steps, reduction reactions and dehydration reactions is:

$$N_{m,n,h} = \sum_i \frac{m!}{\delta!\zeta!\theta!\xi!} \tag{3.17}$$

In order to assess the variation of polyketide structures, an analysis of a model system was performed. In this system, propionyl-ACP and methylmalonyl-ACP, the starter and extender units in the synthesis of erythromycin, were used. Figure 3.4 shows the distribution of possible structures produced with six molecules of methylmalonyl-CoA as a function of the number of $NAPDH_2^+$ molecules used, or number of reduction reactions in the synthesis. When zero molecules of $NADPH_2^+$ are used, no reduction occurs and only one possible structure can be obtained, which corresponds to the poly-β-ketone linear chain characteristic of aromatic polyketides [2]. However, when one molecule of $NADPH_2^+$ is used, the number of possible structures increases to 12, six of which are obtained without dehydration, leading to one hydroxyl group and six carbonyl groups in the linear chain, and six of which involve one dehydration reaction, corresponding to one carbon-carbon double bond and six carbonyl groups. The number of structures that are produced continues to increase until the molecules undergo six reduction reactions in their synthesis. Any additional increase in the amount of reduction involves the full reduction of some of the β-carbons generated during the synthesis, thus reducing the number of possible structures. This decrease in the number of possible structures

continues until the number of reduction reactions equals twice the number of elongation steps and only one structure is produced, corresponding to a fully reduced polyketide chain. Similar histograms can be produced for different number of elongation steps and different starter and extender units using Equations 3.1, 3.11 and 3.17.

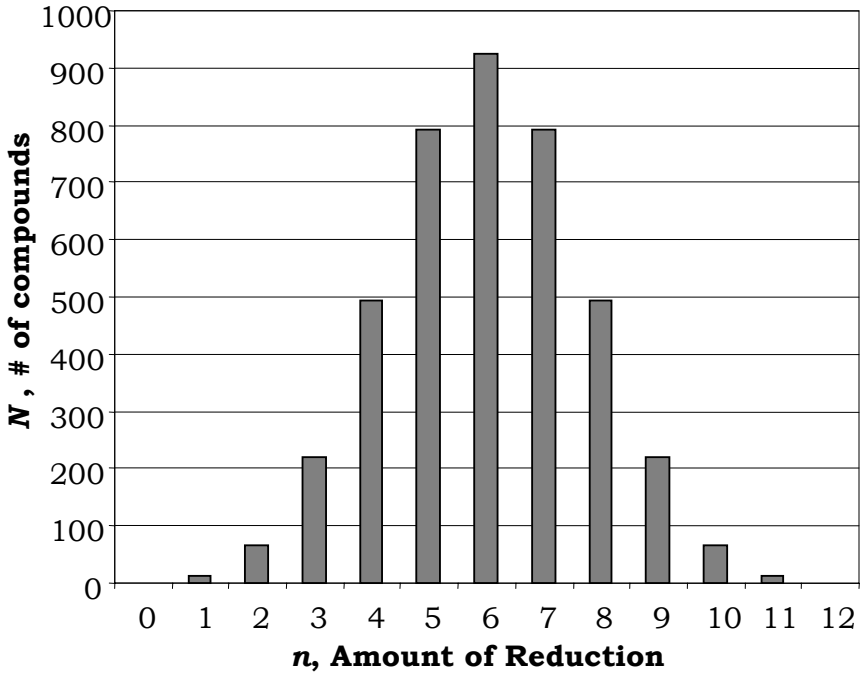


Figure 3.4. Distribution of possible polyketide structures synthesized with one propionyl-CoA and six methylmalonyl-ACP molecules as a function of the number of $NADPH_2^+$ molecules required for the synthesis, calculated using Equation 3.17.

3.4 Evolution of Complexity in Polyketide Synthesis

The BNICE framework is currently being developed to generate biological reaction networks. It involves the identification of a set of enzyme rules and their successive application to a set of substrates; therefore, it generates structures in successive iterations, corresponding to the number of reactions, or pathway length, that produce each structure from the starting

reactants. The complexity in polyketide biosynthesis leads to its use as a model system that can be analyzed more thoroughly through its implementation in the BNICE framework. Therefore, using methylmalonyl-ACP, propionyl-ACP and $NADPH_2^+$ as reactants, the first iteration of the framework results in the generation of one structure, corresponding to the product of one elongation step; this structure is designated as having been produced in generation 1. The second iteration leads to the formation of two generation 2 products, one the product of a second elongation and the other the result of a reduction. In essence, the pathway length from the input molecule to any of the output molecules is equal to g, where g is the generation in which the output molecule is generated. The ensuing emergence of identified structures from the framework is illustrated in Table 3.2.

Table 3.2. Number of structures identified by the BNICE framework in each generation, or iteration of the framework, using malonyl-ACP, acetyl-ACP and two hydrogen atoms as input molecules.

Number of	Generation, g												
Elongation Steps	0	1	2	3	4	5	6	7	8	9	10	11	Total
1	1	1	1	1	0	0	0	0	0	0	0	0	4
2	0	1	2	3	4	3	2	1	0	0	0	0	16
3	0	0	1	3	6	10	12	12	10	6	3	1	64
4	0	0	0	1	4	10	20	31	40	44	40	31	221
5	0	0	0	0	1	5	15	35	65	101	135	155	512
6	0	0	0	0	0	1	6	21	56	120	216	336	756
7	0	0	0	0	0	0	1	7	28	84	203	413	736
8	0	0	0	0	0	0	0	1	8	36	120	322	487
9	0	0	0	0	0	0	0	0	1	9	45	165	220
10	0	0	0	0	0	0	0	0	0	1	10	55	66
11	0	0	0	0	0	0	0	0	0	0	1	11	12
12	0	0	0	0	0	0	0	0	0	0	0	1	1
Total	1	2	4	8	15	29	56	108	208	401	773	1490	3095

From the analysis summarized by Equation 3.1, it is expected that using an input of propionyl-ACP (s = 1) and methylmalonyl-ACP (a = 1) as reactants and without taking into account stereochemistry (b = 4, a_c = 0) the number of structures generated should be 4^m, where m

is the number of elongation steps. This fact is supported by the output obtained from BNICE. For example, a linear chain that is the result of one elongation step, by definition, has an *m* value of one; consequently, four structures that are the product of one elongation reaction are identified by BNICE; these are all generated after four iterations, as shown in Table 3.2. Similarly, 16 and 64 structures, corresponding to two and three elongation steps, respectively, form part of the BNICE output, further verifying that the expected results are obtained from the framework.

As illustrated in Table 3.2, the number of molecules generated increases with generation number, or pathway length. Designating the generation number as g and the number of new molecules produced in generation g as $N(g)$, the following equation is derived to calculate the number of new molecules produced in a generation:

$$N(g) = 2N(g-1) - x_g \tag{3.18}$$

where $N(1)$ is one and the correction factor x_g, which accounts for the decrease in structures due to fully-reduced β-carbons, is calculated from the following:

$$x_g = \begin{cases} 0, g < 5 \\ 1, g = 5 \\ N(g-5), g > 5 \end{cases} \tag{3.19}$$

3.5 Analysis of the Diversity in Polyketide Synthesis Pathways

In addition to generating all the possible species, it is possible to track the growth of the molecules by studying the number of new molecules synthesized by a set number of elongation reactions produced in each generation. Molecules that are the result of *m* elongation steps are produced in a range of generations, as shown in Table 3.2. For example, molecules that are the result of two elongation reactions are produced in generations 1 through 7. Based on the

computational results, formulae are derived to predict the first generation, $g_{1,m}$, and the total number of different generations, $n_{g,m}$, where molecules resulting from m elongation steps are produced:

$$g_{1,m} = m \tag{3.20}$$

$$n_{g,m} = 2m + g_{1,m} + 1 \tag{3.21}$$

Implementation of polyketide biosynthesis in BNICE provides insight into the pathways required for the synthesis of these polyketides; each generation indicates a longer pathway. The aromatic polyketide linear chain is generated first, indicating that the syntheses of the linear chains of all the reduced polyketides involve longer pathways, or a larger number of enzyme actions, than the pathway used for aromatic polyketide synthesis.

The molecules identified by BNICE were also analyzed more closely in order to determine the pathway architecture of polyketide synthesis, depicted in Figure 3.5. The first structure in the figure corresponds to the product of the first elongation step, involving a Claisen condensation between the starter unit propionyl-ACP and the extender unit methylmalonyl-ACP. This product, identified as a generation 1 structure, leads to the production of two structures in generation 2 of the framework. One of these two products is the result of a condensation reaction between the product in generation 1 and the extender unit malonyl-ACP and the other is the product of a reduction of the product in generation 1. Each of these two products identified in generation 2 gives rise to two generation 3 products, producing a total of four structures in generation 3. Two of these structures are the result of an elongation of the two products in the previous generation and the remaining two reduced structures of the two products in the previous generation. Similarly, each of these four products, as shown in the figure, produces two

generation 4 products, one the result of an elongation reaction and the other the result of a reduction reaction. Thus, there are eight structures identified in generation 4. The β-carbon of one of the generation 4 structures, however, is fully reduced; consequently, this structure is not capable of undergoing another reduction reaction. Therefore, instead of producing 16 generation 5 products, the eight generation 4 structures produce 15 structures. This decrease in the expected number of structures generated gives rise to the correction factor in Equation 3.18 for the number of new products identified in a generation.

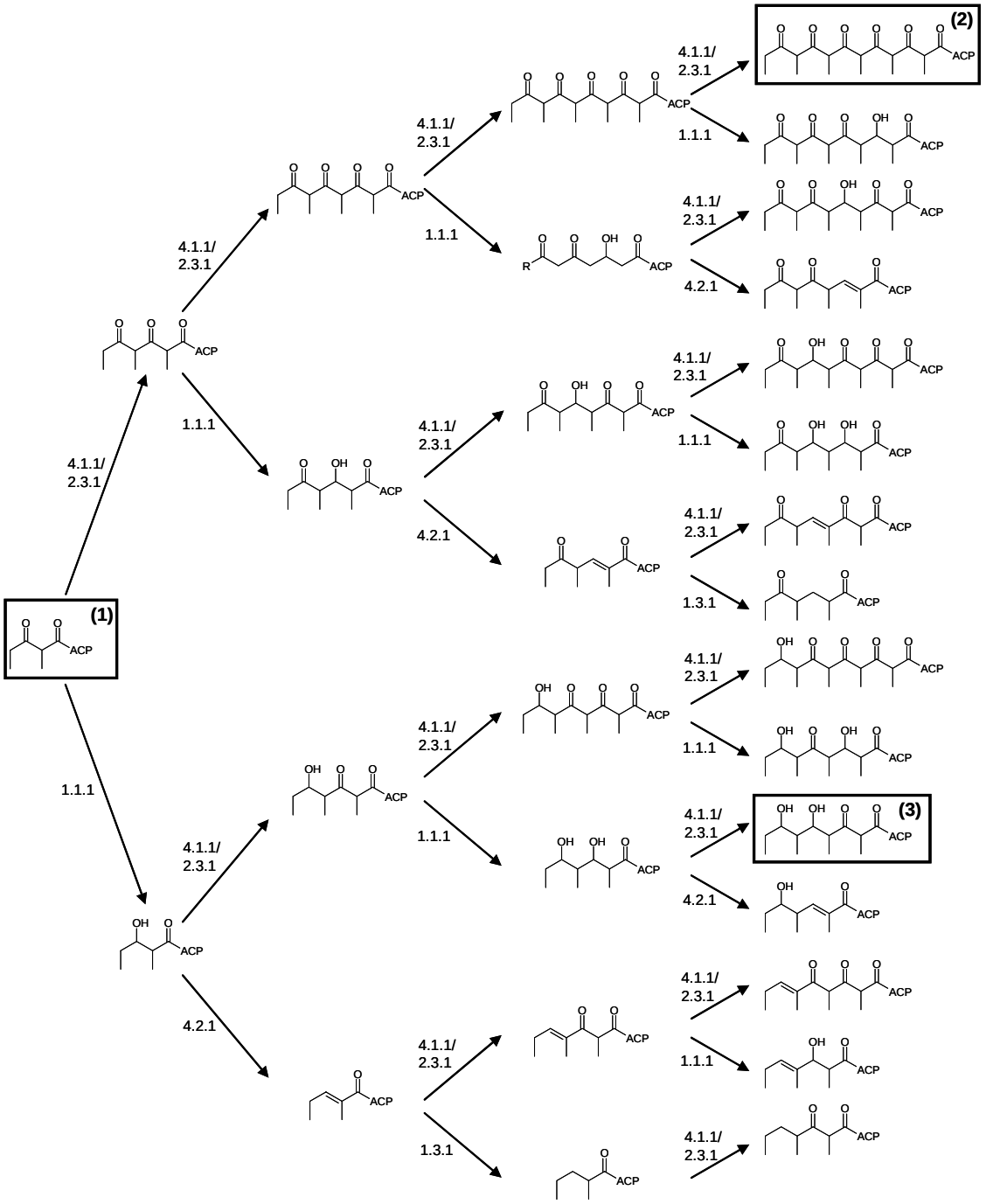


Figure 3.5. Metabolic pathway architecture of polyketide linear chain growth. The structure labeled as (1) is the starting molecule for the pathway; as shown, it has a carbonyl group at the β-carbon. The reactions are labeled according to the generalized enzyme function that catalyzes them. Molecule 2, an aromatic polyketide, and molecule 3 are examples of structures generated after four enzyme actions from the starting material.

The modular nature of the synthesis is also observed in the metabolic tree shown in Figure 3.5. Molecule 1 is the result of the first condensation reaction in polyketide synthesis. Every upward arrow, each of which corresponds to a condensation, or 4.1.1/2.3.1, reaction represents the start of a new module. For example, five modules are required for the synthesis of the linear chain designated as (2) in the figure; each of these modules catalyzes a condensation reaction leading to the production of an aromatic polyketide linear chain. On the other hand, three modules are responsible for the synthesis of molecule 3 even though this structure is the same number of enzyme actions from the starting structure as molecule 2, a difference that arises from the fact that reduction is present in the synthesis of molecule 3 and not in that of molecule 2. The first module in the synthesis of molecule 3 catalyzes the production of molecule 1 and a subsequent reduction reaction, the second consists of a condensation reaction followed by a reduction, and the third involves a condensation reaction.

The pathway architecture illustrated in Figure 3.5 is the minimal representation of the system, where the initial molecule 1 represents any linear polyketide structure with a carbonyl group attached to the β-carbon. In other words, the reaction pathway starting from the product of a Claisen condensation will be the same as the minimalist pathway tree shown. Based on these observations, it is easily shown that each polyketide linear chain is synthesized via a unique pathway. Therefore, the identity and sequence of the modules required for the synthesis of a target polyketide can be determined. Consequently, this pathway architecture clearly demonstrates the uniqueness of the sequence of enzyme actions required to produce a target polyketide, essentially illustrating the uniqueness of the identity and sequence of the modules that catalyze the target molecule. Therefore, the identity and sequence of the modules required

for the synthesis of a target polyketide can be determined. Implementation of the reverse reaction operators and the target structure as the reactant in BNICE results in the identification of the modules involved in the synthesis. For example, using molecule 3 as input results in an output of five reactions: 4.1.1/2.3.1 → 1.1.1 → 4.1.1/2.3.1 → 1.1.1 → 4.1.1/2.3.1; these reactions illustrate the reverse order of the synthesis, showing that the first module catalyzes a 4.1.1/2.3.1 and a reverse 1.1.1 reaction, the second module catalyzes a 4.1.1/2.3.1 and a reverse 1.1.1 reaction and the third catalyzes a 4.1.1/2.3.1 reaction. This allows the facile identification of the identity and sequence of the modules required to synthesize it.

3.6 Conclusions

Polyketides form an important class of biological molecules due to their wide array of biological properties and commercial applications. Based on the number of variables that can be altered during polyketide synthesis, there are over a billion possible linear structures that can be synthesized; assuming that the cyclical structures of each of these linear intermediates is different, there remain a large number of polyketides that can potentially be produced since only about 10,000 structures have been discovered so far. Therefore, it is reasonable to assume that novel polyketide structures can be engineered and that some of these might have properties that would prove useful to the medical community.

The mechanism for polyketide linear chain synthesis makes it simple to identify the effect of variable manipulation on the output linear structure. This feature of the system provides a basis for which the results from the BNICE framework were analyzed in order to understand the combinatorial nature of reaction networks in cellular organisms as well as the evolution of structural diversity in the synthesis of polyketides. As mentioned, reduced polyketides can be

synthesized with different starter and extender molecules; for example, the synthesis of erythromycin, a widely used antibiotic, utilizes propionyl-CoA and methylmalonyl-CoA as starter and extender units, respectively [8, 9]. The modular nature of the PKS of reduced polyketides allowed experimentalists to develop a combinatorial approach to the metabolic engineering of PKS genes, leading to the generation of polyketide libraries, such as the one that has been developed for erythromycin, which consists of over 100 structures [8, 9]. This library was generated using propionyl-CoA as the starter unit and methylmalonyl-CoA and malonyl-CoA as extender units. From the analysis presented, it is expected that over seven million deviations of the erythromycin linear chain can be produced for one starter unit (s = 1), two different extender units (a = 2), one of which is chiral (a_c = 1), five β-carbon arrangements (b = 5) if taking into account stereochemistry and six elongation steps (m = 6). Therefore, a large number of structures are missing from the erythromycin library. Consequently, BNICE can be used to identify all the possible structures that could be generated in order to determine which structures are missing from the erythromycin library in addition to generating other libraries of polyketides.

Furthermore, an analysis of the synthesis pathways was performed, through the use of the BNICE framework, in order to identify the pathway architecture in the biosynthesis of polyketides. The synthesis pathway architecture clearly demonstrates the uniqueness of the sequence of enzyme actions required to produce a target molecule, essentially illustrating the uniqueness in the identity and sequence of the modules that would catalyze the target polyketide. Therefore, the polyketide libraries can be supplemented by information about their synthesis

pathways, guiding researchers in determining the metabolic engineering actions that would lead to the production of a specific metabolite.

In order to identify the final polyketide structures, the cyclization reactions, as well as stereochemistry, will be implemented in the BNICE framework. Additionally, as mentioned, there are a large number of structures that to date have not been discovered experimentally. It is possible that these are produced in yields too low to allow detection or are unstable structures. Furthermore, it is possible that the synthesis of some of these undiscovered structures is not thermodynamically favorable. As described in Chapter 6, metabolic flux analysis will aid in determining the effect of cellular constraints on the synthesis yield of target polyketide structures.

Chapter 4

Effect of Stereochemistry on Polyketide Synthesis

Metabolism is a complex process involving a large number of biochemical reactions and transformations between the different cellular metabolites. In order to explore the complexity of this system, a computational framework called *Biochemical Network Integrated Computational Explorer* (BNICE) was developed to analyze metabolic networks. The framework utilizes graph theory to represent reactions as a series of matrix operations, where each substrate and product can be characterized by a bond and electron (BE) matrix and each reaction by another matrix termed a reaction operator. To date, it has been applied to the study of amino acid synthesis, the production of specialty chemicals and polyketide biosynthesis [20, 29, 30]. However, the framework used in the previous analyses did not distinguish between stereoisomers. Therefore, since enzymes are known for their efficient catalysis of enantioselective biochemical transformations, an expansion of the BNICE framework to incorporate stereochemical information is particularly important for the faithful reproduction of biological reaction systems.

Chemical isomerism, of which there are two types, arises from different atomic arrangements in compounds with the same chemical formula. The first is structural isomerism, which involves a different atomic connectivity, or bond structure, between the different isomeric compounds. Stereochemical isomers, on the other hand, are molecules with the same bond structure but with

a different geometrical positioning of the atoms in space. Chirality, or optical isomerism, arises from a lack of symmetry around an atom with four different substituents; in other words, chiral isomers are two different molecules with the same molecular formula and atomic bonding but a different symmetry around at least one atom. These four substituents are assigned a priority according to the Cahn Ingold Prelog (CIP) priority rules [31, 32]. A chiral atom is then labeled as R or S depending on the arrangement of the substituents; assuming that the bond to the substituent with priority 4 is pointed away from the viewer, if the arrangement of substituent 1 to substituent 2 to substituent 3 results in a clockwise rotation, the chiral center is labeled R versus if the rotation is counterclockwise, the chiral center is labeled S. Compounds with more than one stereogenic center might have an internal plane of symmetry, leading to an optically inactive structure, or meso compound. A second type of stereochemical isomers consists of geometric isomers in which the functional groups of the isomers are oriented differently. The substituents of the atoms from which geometric isomerism arises are labeled according to the CIP priority rules. If both of the substituents with the highest priority are on the same side, the isomer arrangement is labeled Z; on the other hand, if they are located on different sides, the arrangement is labeled E.

The traditional visual structural representation of compounds utilizes wedged and hatched bonds to denote the orientation of bonds, allowing the differentiation of stereoisomers. Consequently, the computational representation of the stereochemistry of a structure has been analyzed previously and a number of formulations have been developed including SMILES [33] and InChI [34]. These formulations identify the atomic connectivity of a structure and the relative stereochemistry without incorporating atomic coordinate information. For example, the

SMILES formalism, which characterizes a molecule as a series of characters where atoms are identified by their atomic symbols and single, double, triple and aromatic bonds are represented by the symbols "-", "=", "#" and ":", respectively, includes the option to identify the chirality of the molecule by distinguishing the chirality of an atom through the symbol @ and @@, following the symbol for the chiral center in the notation. The symbol is then followed by one of the substituents, which, although it can be any of the substituents, it is analogous to the substituent with priority 4 in the R/S notation; the following three substituents are arranged in a counterclockwise or clockwise rotation, denoted by @ and @@, respectively [33]. The SMILES formulation also allows the representation of geometric isomer using the symbols "/" and "\" to denote the location of the substituent groups of double bonds; for example, the strings "F/C=C/F" and "F/C=C\F" represent the trans, or E, isomer of difluoroethane (fluorine atoms on different sides of the double bond) and cis, or Z, isomer (fluorine atoms on the same side), respectively [33].

In addition, the manipulation and analysis of stereochemistry has also been studied previously. Algorithms have also been developed for the analysis of xyz structural coordinates in order to identify stereogenic centers [35]. These algorithms use the spatial information in order to label the stereochemistry of chiral atoms as R or S and geometric isomers as Z or E [36]. Additionally, algorithms have been developed for the automatic determination of possible stereoisomers as well as for the identification of stereogenic centers and the enumeration of all possible stereoisomers [37-39].

A number of frameworks have been previously developed for the analysis of reaction pathways. These include algorithms for the identification of one or more pathways that are

capable of transforming A into B using a given database of chemical transformations [40, 41]. To date, stereochemical isomerism has only been considered in the analysis of existing pathways as long as the stereochemistry is inherent to the reactions in the database through the use of different compound identifiers for the different isomers. These frameworks explore the existing chemistry and can not be used to analyze novel pathways with reactions not included in the database. On the other hand, a number of frameworks have been developed that utilize reaction rules to identify all the theoretically possible products of a given substrate [17, 42]. These frameworks have the capability of identifying novel reactions pathways. However, none of these algorithms have previously explored the effect of stereochemistry in pathway generation.

The present work fills in these gaps by including the capability to distinguish between stereoisomers and effect stereochemical reactions in automated pathway generation for the first time. In order to distinguish between stereoisomers in the BNICE framework, each chiral atom is labeled as R or S and those atoms that give rise to geometric isomerism as Z or E. In addition, the BE-matrix was expanded to a 3-D BE-matrix in order to denote the spatial arrangement of the bonds in the $\{i,j,2\}$ matrix elements. As illustrated with amino acid biosynthesis, the numbers of possible reactions and structures that are identified by the framework increase significantly due to the treatment of isomers as distinct structures. In addition, the geometry of the active site of enzymes allows their enantioselective catalysis; therefore, it is possible to more accurately represent biological reactions through the specification of the stereochemistry of the substrate and reaction operator as demonstrated with polyketide biosynthesis. Consequently, the introduction of stereochemical differentiation in automatic pathway analysis makes it possible to more completely and accurately explore a cellular reaction network.

4.1 The BNICE Formalism

4.1.1 Mathematical Representation of Stereochemical Isomers

The BNICE framework used in previous analyses represents substrate and product molecules through BE matrices that denote the atomic connectivity of the molecule. The BE-matrix is of size *n*-by-*n*, where *n* is the number of atoms or groups in the molecule [19]. The {*i*,*j*} and the {*i*,*i*} elements of the matrix illustrate the connectivity or bond order between atoms *i* and *j* and the number of non-bonded electrons in atom *i*, respectively. Since the BE-matrix illustrates the bond connectivity of a compound, this approach differentiates between structural isomers. However, stereoisomers, including both chiral and geometric isomers, have the same atomic connectivity and are therefore represented by the same BE-matrix.

Therefore, the stereochemistry of an atom was specified through labeling of a chiral atom as R or S and an atom that participates in a double bond that gives rise to geometric isomerism as Z or E. This representation was sufficient for the unique identification of most isomers.

However, geometric isomerism due to the presence of a ring required an additional representation of the stereochemistry. In a visual representation of a structure, wedged and hatched bonds are used to illustrate the relative orientation of atoms in the molecules. This same principle was applied to the mathematical representation of a chiral molecule through the expansion of the original *n*-by-*n* BE-matrix into an *n*-by-*n*-by-2 BE-matrix, where the {*i*,*j*,*1*} elements correspond to the original BE-matrix representation of a molecule and the {*i*,*j*,*2*} elements illustrate the stereochemistry of a molecule. These {*i*,*j*,*2*} elements all have a value of zero with the exception of those that correspond to wedged or hatched bonds in the

corresponding structural representation of the molecule; these {*i*,*j*,*2*} elements have a value of +1 and -1 to denote a wedged and hatched bond, respectively.

4.1.2 Mathematical Representation of a Stereochemical Reaction

In the framework, reactions are represented by the addition of the substrate BE-matrix to a reaction operator R-matrix to obtain the BE-matrix of the corresponding product. The previous BNICE framework did not constrain the stereochemistry of the substrate or product of a reaction. However, the framework was expanded such that the stereochemistry of the functional group of the substrate and/or the stereochemistry produced in the product can be specified.

4.1.3 Definition of the Stereochemical String Code

A string code is used by the framework in order to easily compare the different compounds to verify that a structure has not been generated previously by the framework [15, 43]. The string code is a unique string representation of a molecule; however, as with the BE-matrix representation, it does not differentiate between stereochemical isomers. Therefore, in order to distinguish stereochemical isomers, a stereochemical string code was created, where the strings "D" or "L" are added immediately after a chiral atom with a chirality of "R" or "S", respectively (the string "S" could not be used because it also represents a sulfur atom). Similarly, the strings "Z" or "E" were added immediately following each of the two atoms in a double bond that gives rise to Z or E geometric isomerism. As mentioned previously, additional information was necessary to identify all the isomers that correspond to atoms in a ring that give rise to geometric isomerism; therefore, the strings "+1" or "-1" are added immediately prior to an atom that is bonded to a one of these atoms; in other words, although bonds are not usually included in the

generation of the string code, if the {*i,j,2*} matrix element corresponding to a bond is +1 or -1, the string "+1" or "-1", respectively, is included in the stereochemical string code.

4.1.4 Overview of BNICE Framework

The computational framework utilizes two types of input structures: a specific isomer of a substrate or structural information without stereochemical information, in which case the framework uses all the isomers of the substrate. Therefore, each input substrate is first analyzed for isomerism such that if no or incomplete stereochemical information is provided, the framework creates and includes all possible isomers of the input substrate(s) specified in the initial species list. In addition to the substrates, the generalized enzyme rules are defined by the user; the stereochemistry of the functional groups of the substrate and/or any change in stereochemistry that can occur as a result of the reaction can also be defined, although it is not necessary. The framework then identifies all the available functional groups in each of these species in accordance with the specified generalized enzyme rules and the corresponding reactions are carried out via matrix addition including any changes in the stereochemistry of the substrate(s) to form the product(s). These products are then analyzed for stereochemistry and all possible isomers of the structures are created. The uniqueness of each product is verified using the stereochemical string code and the new products are then added to the species list. The process is then repeated until a user-defined termination criterion is met. (Refer to Appendix B for detailed algorithm.)

4.2 Effect of Stereochemistry in a Biological Reaction Network

An important function of the cell is the biosynthesis of the amino acids required for growth. The synthesis of the amino acid phenylalanine from chorismate, which has been previously analyzed using the BNICE framework, involves the use of five generalized enzyme functions: 2.6.1, 4.1.1, 4.2.1, 5.3.3 and 5.4.4 [20]. The previous study shows that these five generalized enzyme functions give rise to over a thousand reactions and 246 compounds from the input substrate chorismate and using glutamate as the amino donor; however, this investigation did not differentiate among chiral isomers. If all possible chiral isomers can be generated from chorismate using these five generalized enzyme reactions, the number of compounds identified by the framework increases by more than a factor of two, as shown in Figure 4.1, illustrating that a number of these compounds, including chorismate, have more than one chiral center resulting in three or more chiral isomers of the compounds being identified by the framework; the same is true for the number of reactions. As expected based on the previous analysis, no new compounds are identified after generation ten and no new reactions are possible after generation 11. The previous study performed by Hatzimanikatis et al. (2005) shows that there are seven native length pathways from chorismate to phenylalanine without taking into account stereochemistry. However, as illustrated in Figure 4.2, if stereochemical isomers are differentiated, it is possible to identify twelve native-length pathways from chorismate to phenylalanine.

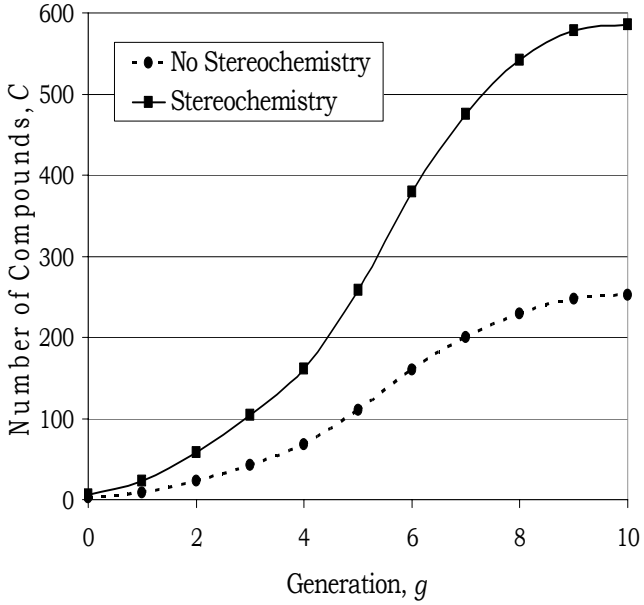


Figure 4.1. Total number of compounds identified by the framework as a function of generation and differentiating versus not differentiating between stereoisomers. These compounds are generated from chorismate using the generalized enzyme functions 2.5.1, 4.1.1, 4.2.1, 5.3.3 and 5.4.4.

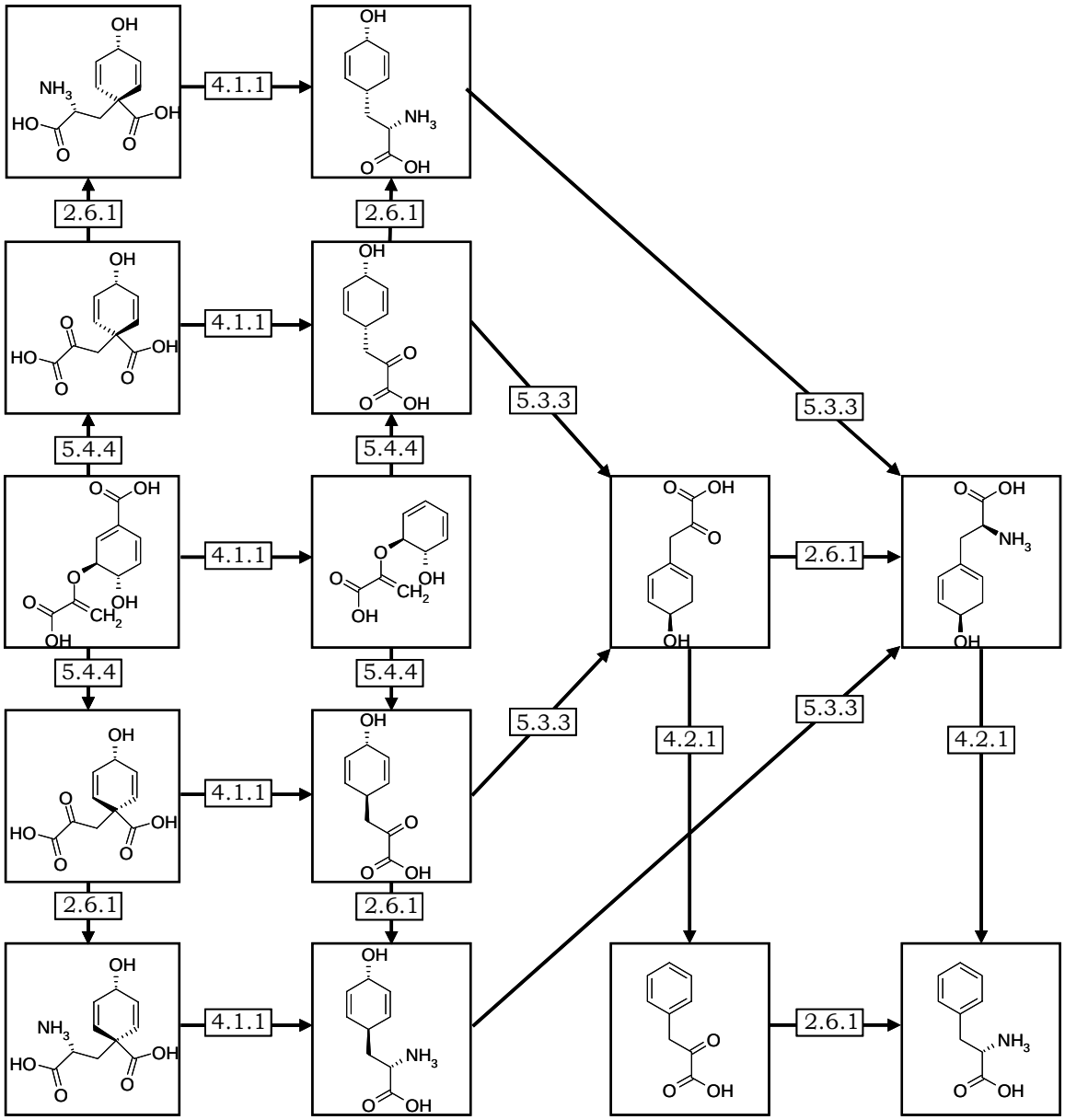


Figure 4.2. Illustration of native-length pathways from chorismate to phenylalanine. These were identified allowing the framework to distinguish between chiral isomers and include a higher number of alternate native-length pathways than when chirality was not accounted for in the framework.

An analysis of the reactions catalyzed by these five generalized functions shows that some reactions can destroy while others can create isomerism. For example, the dehydration reaction catalyzed by the generalized enzyme function 4.2.1 involves breaking a bond between a carbon and a hydroxyl group and a carbon-hydrogen bond, thus releasing water and forming a double carbon-carbon bond. Therefore, the generalized enzyme function 4.2.1 will always breaks a chiral center if the hydroxyl group is attached to a chiral carbon; in addition, since this reaction involves the formation of a double bond, it can never form a chiral center. The opposite is true for the generalized enzyme function 2.6.1; this reaction is capable of creating a chiral center but can never break one. The generalized enzyme functions 4.1.1 and 5.3.3 are capable of both functionalities: they can both make and destroy a chiral center. The isomerase generalized enzyme function 5.4.4 changes the location of a double bond in a molecule; it is therefore capable of creating and/or destroying a chiral center in the same reaction.

4.3 Analysis of Stereochemistry on a Biological Reaction Network

Polyketide synthesis has been previously implemented in the BNICE framework in order to identify all the possible polyketide structures, and the pathway architecture for the system was identified to involve a tree-like structure, which implies that each structure is produced via a unique pathway [29]. Allowing the framework to differentiate between stereoisomers, all the possible structures can be identified, resulting in an increase of almost double the number of possible structures. An analysis of the pathway architecture, depicted in Figure 4.3, shows that stereochemistry can significantly affect the properties of the reaction network. As in the amino acid synthesis system, the reaction pathway generated by the framework illustrates how some reactions create and others destroy chirality. For example, the reduction reaction catalyzed by

the KR domain of the PKS always creates a chiral center; however, the dehydration reaction always destroys the chiral center created by the previous KR reaction. A similar property is shown with geometrical isomerism; as illustrated in Figure 4.3, the reaction catalyzed by the DH domain creates geometrical isomerism, while the ER reaction destroys it. Therefore, whereas the previous analysis performed without stereochemistry shows that there is a unique pathway for the synthesis of a polyketide structure, the creation and destruction of isomerism in the polyketide linear chain leads to the identification of more than one synthesis pathway for a polyketide structure, as illustrated in Figure 4.3.

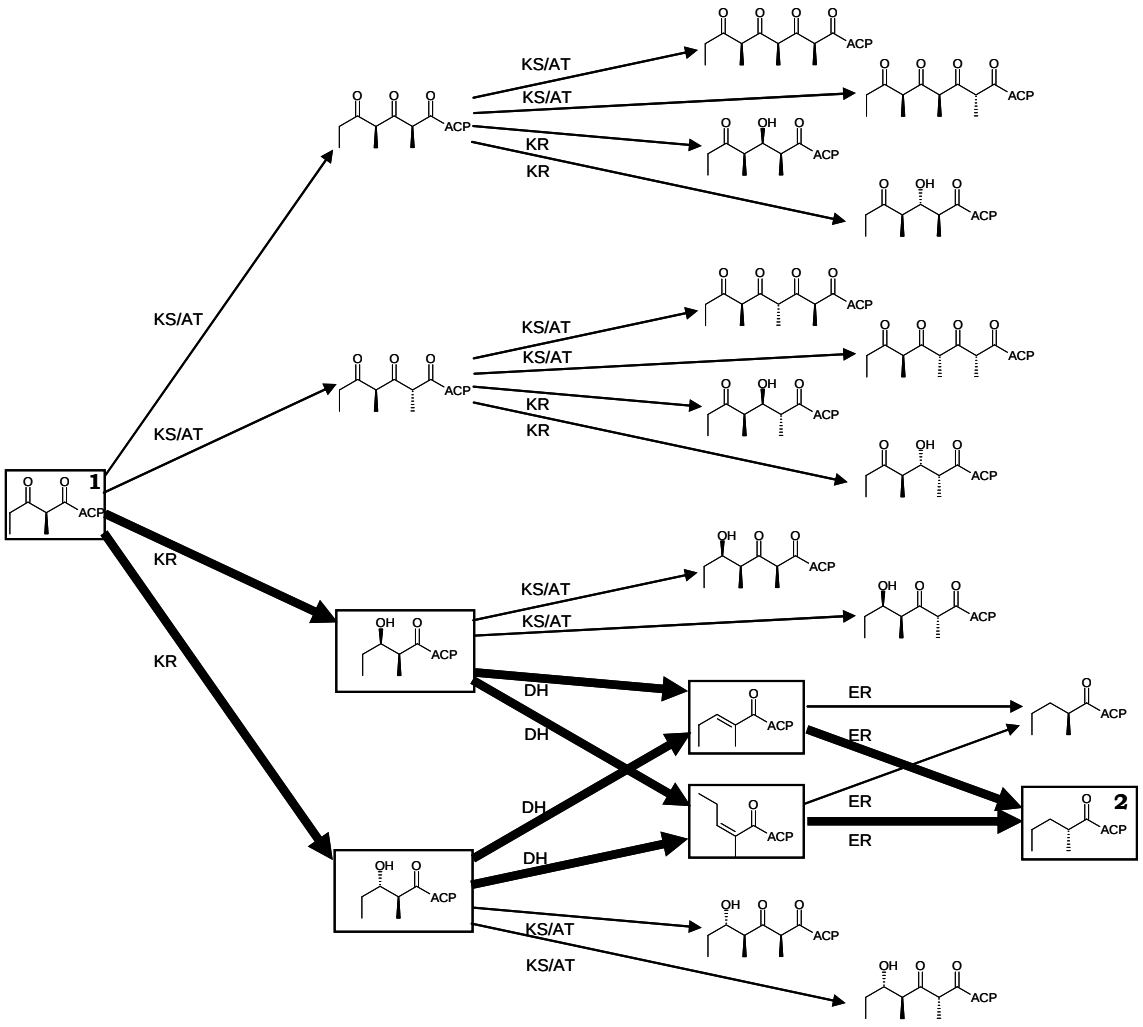


Figure 4.3. Pathway architecture of polyketide linear chain synthesis. The structure labeled 1 is the product of the first elongation reaction, assuming no change in stereochemistry. This structure can then undergo an elongation reaction (KS/AT), which creates a chiral center and therefore two new structures; similarly, it can also undergo a reduction catalyzed by the KR active site, which also leads to the production of two isomeric structures. The chiral center produced by the KR reaction can then be destroyed via the dehydration (DH) domain, which in turn creates a geometric isomerism in the structure that is then destroyed with the ER domain. Therefore, the creation and destruction of stereochemical isomerism leads to the identification of four distinct synthesis pathways for structure 2.

4.4 Accurate Reproduction of Biological Reactions

An analysis of the synthesis of the precursor to erythromycin, a currently prescribed antibiotic, shows that some modules result in the synthesis of different stereochemistry. For example, as depicted in Figure 4.4, the condensation reaction catalyzed by the KS/AT domains of module 1 results in a chiral center with a methyl group pointing upwards from the plane while the condensation reaction in module 2 creates a chiral center with an attached methyl group pointing downwards from the plane; a similar phenomenon is observed with the reduction reactions catalyzed by the KR domains of modules 1 and 2. These differences in the stereochemistry of the chiral centers suggest that the geometries of the active sites of the KS/AT and KR enzymes differ between modules 1 and 2.

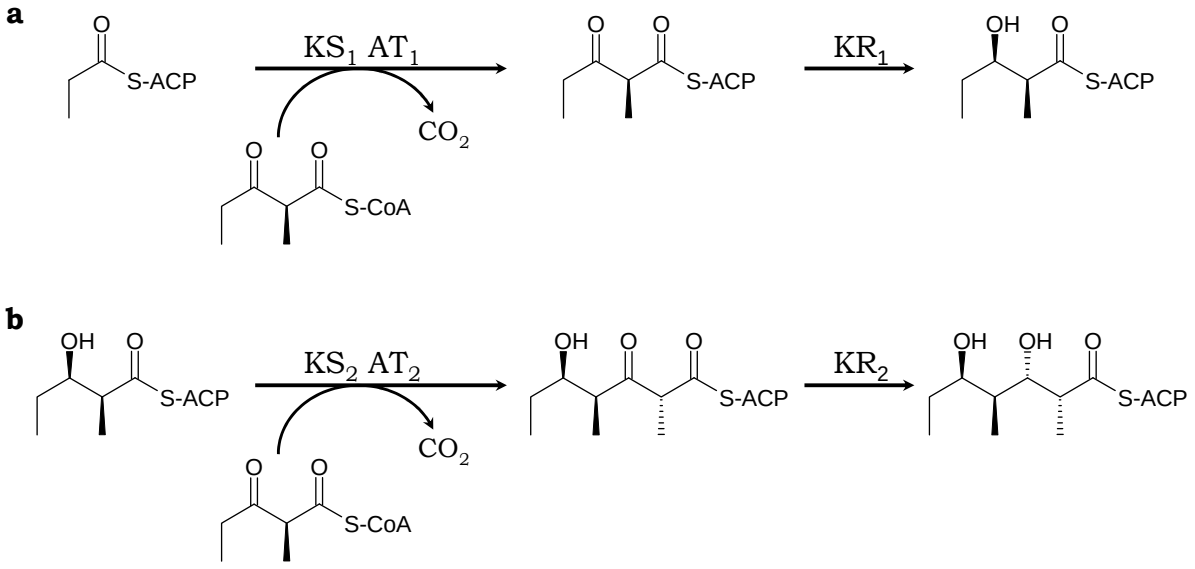


Figure 4.4. Schematic of elongation step catalyzed by module 1 (a) and module 2 (b) of the erythromycin PKS. Both of these modules catalyze a condensation reaction followed by a reduction reaction. However, the stereochemistry of the chiral center created by each of these reactions differs in modules 1 and 2: the carboxyl and methyl groups from module 1 are pointing upwards from the plane while those from module 2 are pointing downward. Therefore, the KS/AT and KR enzymes are different between these two modules.

Consequently, the 4.1.1/2.3.1 generalized reaction can be subdivided in order to account for the two possible stereochemical changes that can be produced as a result of the condensation reaction. Therefore, the 4.1.1/2.3.1S generalized reaction rule was created to represent the reaction catalyzed by the 4.1.1/2.3.1 domains of modules 1, 3 and 4 while the 4.1.1/2.3.1R reaction rule would catalyze the changes in the stereochemistry caused by the 4.1.1/2.3.1 domains in modules 2, 5 and 6; these two generalized reaction rules denote the same changes in the bond arrangement defined by the original KS/AT reaction rule but differ in the stereochemistry. Similar subdivision was performed for the other reactions involved in polyketide synthesis, leading to the identification of ten generalized reaction rules involved in the biosynthesis of polyketides, as opposed to the five operators used in the original work in Chapter 3. The ten generalized reaction rules for the synthesis of the polyketide linear chain are summarized in Table 4.1- Table 4.4. Utilizing these stereochemistry-specific generalized enzyme functions, the number of synthesis pathways illustrated in Figure 4.3 for the synthesis of molecule 2 is reduced from four to two, as shown in Figure 4.5.

Table 4.1. Table of reaction operators for the stereochemistry-specific 4.1.1/2.3.1 generalized enzyme function.

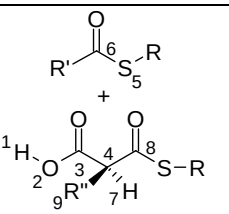
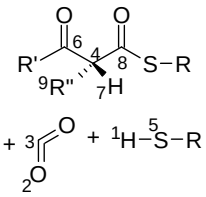
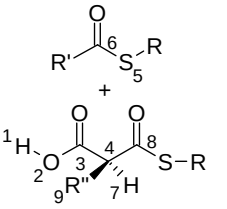
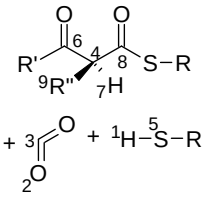
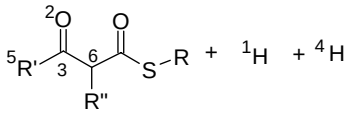
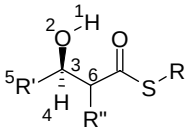
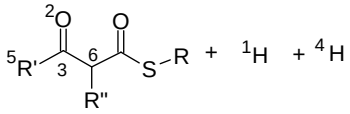
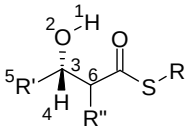
<table>
<tr><th>Reactant</th><th>Product</th><th>Reaction Operator</th></tr>
<tr><td colspan="3">Generalized Enzyme Function 4.1.1/2.3.1R</td></tr>
<tr>
<td> 
<table>
<tr><th></th><th>1H</th><th>2O</th><th>3C</th><th>4C</th><th>5S</th><th>6C</th></tr>
<tr><td>1H</td><td>0</td><td>1</td><td>0</td><td>0</td><td>0</td><td>0</td></tr>
<tr><td>2O</td><td>1</td><td>4</td><td>1</td><td>0</td><td>0</td><td>0</td></tr>
<tr><td>3C</td><td>0</td><td>1</td><td>0</td><td>1</td><td>0</td><td>0</td></tr>
<tr><td>4C</td><td>0</td><td>0</td><td>1</td><td>0</td><td>0</td><td>0</td></tr>
<tr><td>5S</td><td>0</td><td>0</td><td>0</td><td>0</td><td>4</td><td>1</td></tr>
<tr><td>6C</td><td>0</td><td>0</td><td>0</td><td>0</td><td>1</td><td>0</td></tr>
</table>
Chirality of 4C is S</td>
<td> 
<table>
<tr><th></th><th>1H</th><th>2O</th><th>3C</th><th>4C</th><th>5S</th><th>6C</th></tr>
<tr><td>1H</td><td>0</td><td>0</td><td>0</td><td>0</td><td>1</td><td>0</td></tr>
<tr><td>2O</td><td>0</td><td>4</td><td>2</td><td>0</td><td>0</td><td>0</td></tr>
<tr><td>3C</td><td>0</td><td>2</td><td>0</td><td>0</td><td>0</td><td>0</td></tr>
<tr><td>4C</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>1</td></tr>
<tr><td>5S</td><td>1</td><td>0</td><td>0</td><td>0</td><td>4</td><td>0</td></tr>
<tr><td>6C</td><td>0</td><td>0</td><td>0</td><td>1</td><td>0</td><td>0</td></tr>
</table></td>
<td><table>
<tr><th></th><th>1H</th><th>2O</th><th>3C</th><th>4C</th><th>5S</th><th>6C</th></tr>
<tr><td>1H</td><td>0</td><td>-1</td><td>0</td><td>0</td><td>1</td><td>0</td></tr>
<tr><td>2O</td><td>-1</td><td>0</td><td>1</td><td>0</td><td>0</td><td>0</td></tr>
<tr><td>3C</td><td>0</td><td>1</td><td>0</td><td>-1</td><td>0</td><td>0</td></tr>
<tr><td>4C</td><td>0</td><td>0</td><td>-1</td><td>0</td><td>0</td><td>1</td></tr>
<tr><td>5S</td><td>1</td><td>0</td><td>0</td><td>0</td><td>0</td><td>-1</td></tr>
<tr><td>6C</td><td>0</td><td>0</td><td>0</td><td>1</td><td>-1</td><td>0</td></tr>
</table>
4C achieves a chirality R if CIP Priorities:
8C = 1; 6C = 2;
9R” = 3; 7H = 4</td>
</tr>
<tr><td colspan="3">Generalized Enzyme Function 4.1.1/2.3.1S</td></tr>
<tr>
<td> 
<table>
<tr><th></th><th>1H</th><th>2O</th><th>3C</th><th>4C</th><th>5S</th><th>6C</th></tr>
<tr><td>1H</td><td>0</td><td>1</td><td>0</td><td>0</td><td>0</td><td>0</td></tr>
<tr><td>2O</td><td>1</td><td>4</td><td>1</td><td>0</td><td>0</td><td>0</td></tr>
<tr><td>3C</td><td>0</td><td>1</td><td>0</td><td>1</td><td>0</td><td>0</td></tr>
<tr><td>4C</td><td>0</td><td>0</td><td>1</td><td>0</td><td>0</td><td>0</td></tr>
<tr><td>5S</td><td>0</td><td>0</td><td>0</td><td>0</td><td>4</td><td>1</td></tr>
<tr><td>6C</td><td>0</td><td>0</td><td>0</td><td>0</td><td>1</td><td>0</td></tr>
</table>
Chirality of 4C is S</td>
<td> 
<table>
<tr><th></th><th>1H</th><th>2O</th><th>3C</th><th>4C</th><th>5S</th><th>6C</th></tr>
<tr><td>1H</td><td>0</td><td>0</td><td>0</td><td>0</td><td>1</td><td>0</td></tr>
<tr><td>2O</td><td>0</td><td>4</td><td>2</td><td>0</td><td>0</td><td>0</td></tr>
<tr><td>3C</td><td>0</td><td>2</td><td>0</td><td>0</td><td>0</td><td>0</td></tr>
<tr><td>4C</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>1</td></tr>
<tr><td>5S</td><td>1</td><td>0</td><td>0</td><td>0</td><td>4</td><td>0</td></tr>
<tr><td>6C</td><td>0</td><td>0</td><td>0</td><td>1</td><td>0</td><td>0</td></tr>
</table></td>
<td><table>
<tr><th></th><th>1H</th><th>2O</th><th>3C</th><th>4C</th><th>5S</th><th>6C</th></tr>
<tr><td>1H</td><td>0</td><td>-1</td><td>0</td><td>0</td><td>1</td><td>0</td></tr>
<tr><td>2O</td><td>-1</td><td>0</td><td>1</td><td>0</td><td>0</td><td>0</td></tr>
<tr><td>3C</td><td>0</td><td>1</td><td>0</td><td>-1</td><td>0</td><td>0</td></tr>
<tr><td>4C</td><td>0</td><td>0</td><td>-1</td><td>0</td><td>0</td><td>1</td></tr>
<tr><td>5S</td><td>1</td><td>0</td><td>0</td><td>0</td><td>0</td><td>-1</td></tr>
<tr><td>6C</td><td>0</td><td>0</td><td>0</td><td>1</td><td>-1</td><td>0</td></tr>
</table>
4C achieves a chirality S if CIP Priorities:
8C = 1; 6C = 2;
9R” = 3; 7H = 4</td>
</tr>
</table>

Table 4.2. Table of reaction operators for the stereochemistry-specific 1.1.1 generalized enzyme function.

Reactant	Product	Reaction Operator
Generalized Enzyme Function 1.1.1R		
		
		3C achieves a chirality of R if CIP Priorities: 2O = 1; 6C = 2; 5R' = 3; 4H = 4
Generalized Enzyme Function 1.1.1S		
		
		3C achieves a chirality of S if CIP Priorities: 2O = 1; 6C = 2; 5R' = 3; 4H = 4

	2O	3C	1H	4H
2O	0	2	0	0
3C	2	4	0	0
1H	0	0	1	0
4H	0	0	0	1

	2O	3C	1H	4H
2O	0	1	1	0
3C	1	4	0	1
1H	1	0	0	0
4H	0	1	0	0

	2O	3C	1H	4H
2O	0	-1	1	0
3C	-1	0	0	1
1H	1	0	-1	0
4H	0	1	0	-1

Table 4.3. Table of reaction operators for the stereochemistry-specific 4.1.2 generalized enzyme function.

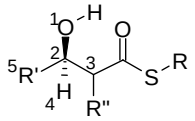
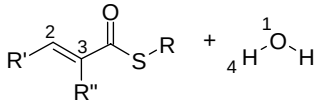
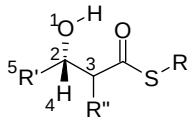
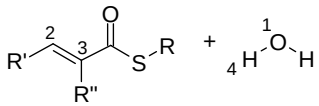
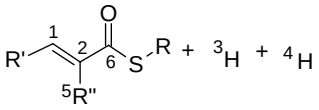
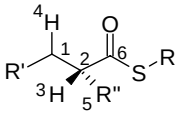
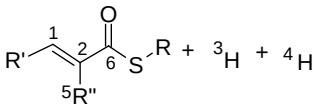
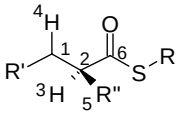
Reactant	Product	Reaction Operator
Generalized Enzyme Function 4.1.2R		
		
1O 2C 3C 4H 1O 4 1 0 0 2C 1 0 1 0 3C 0 1 0 1 4H 0 0 1 0 Chirality of 2C is R	1O 2C 3C 4H 1O 4 0 0 1 2C 0 0 2 0 3C 0 2 0 0 4H 1 0 0 0	1O 2C 3C 4H 1O 0 -1 0 1 2C -1 0 1 0 3C 0 1 0 -1 4H 1 0 -1 0 2C and 3C achieve Z chirality
Generalized Enzyme Function 4.1.2S		
		
1O 2C 3C 4H 1O 4 1 0 0 2C 1 0 1 0 3C 0 1 0 1 4H 0 0 1 0 Chirality of 2C is S	1O 2C 3C 4H 1O 4 0 0 1 2C 0 0 2 0 3C 0 2 0 0 4H 1 0 0 0	1O 2C 3C 4H 1O 0 -1 0 1 2C -1 0 1 0 3C 0 1 0 -1 4H 1 0 -1 0 2C and 3C achieve Z chirality

Table 4.4. Table of reaction operators for the stereochemistry-specific 1.3.1 generalized enzyme function.

Reactant	Product	Reaction Operator
Generalized Enzyme Function 1.3.1R		
 1C 2C 3H 4H 1C 0 2 0 0 2C 2 0 0 0 3H 0 0 1 0 4H 0 0 0 1 Chirality of 1C and 2C is Z	 1C 2C 3H 4H 1C 0 1 0 1 2C 1 0 1 0 3H 0 1 0 0 4H 1 0 0 0	1C 2C 3H 4H 1C 0 -1 0 1 2C -1 0 1 0 3H 0 1 -1 0 4H 1 0 0 -1 2C achieves a chirality of R if CIP Priorities: 6C = 1; 1C = 2; 5R"= 3; 3H = 4
Generalized Enzyme Function 1.3.1S		
 1C 2C 3H 4H 1C 0 2 0 0 2C 2 0 0 0 3H 0 0 1 0 4H 0 0 0 1 Chirality of 1C and 2C is Z	 1C 2C 3H 4H 1C 0 1 0 1 2C 1 0 1 0 3H 0 1 0 0 4H 1 0 0 0	1C 2C 3H 4H 1C 0 -1 0 1 2C -1 0 1 0 3H 0 1 -1 0 4H 1 0 0 -1 2C achieves a chirality of S if CIP Priorities: 6C = 1; 1C = 2; 5R" = 3; 3H = 4

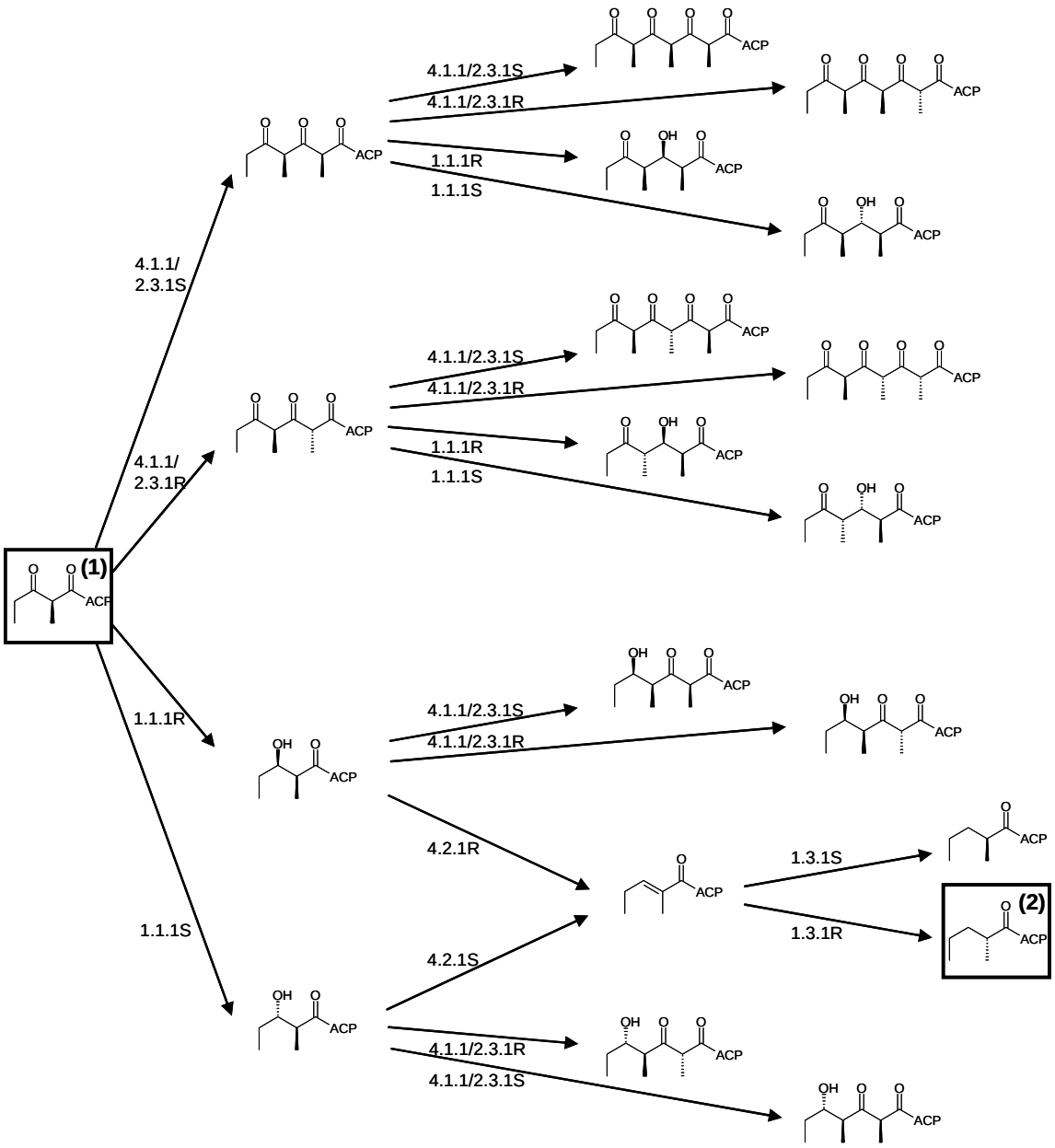


Figure 4.5. Pathway architecture of polyketide linear chain synthesis with stereochemistry-specific generalized enzyme functions. The structure labeled 1 can undergo either the 4.1.1/2.3.1R or 4.1.1/2.3.1S condensation reactions, resulting in the two different isomers; similarly, it can also be reduced by either of the 1.1.1 generalized enzyme functions, which also leads to the production of two isomeric structures. The chiral center produced by a 1.1.1 reaction can then be destroyed via a 4.2.1 reaction, which in turn creates a specific geometric isomer that is then destroyed with a 1.3.1 reaction. Therefore, the creation and destruction of stereochemical isomerism leads to the identification of two distinct synthesis pathways for structure 2.

4.5 Conclusions

Computational frameworks have been previously developed for the automatic generation of reaction networks. Previous work has shown that the BNICE framework can be used to analyze biological systems. However, the reactions in a metabolic network are catalyzed by enzymes, which are known for their stereospecific catalysis. Therefore, in order to more accurately represent the cellular metabolism, it was necessary to expand the framework such that the specific stereochemistry of an atom was represented, making it possible to differentiate stereochemical isomers. Although the framework does not address some stereoisomers such as axial and equatorial conformations, it is now capable of distinguishing chiral and geometric isomers. As illustrated by the analysis of the synthesis of phenylalanine from chorismate, the introduction of stereochemistry into the BNICE framework allows a more complete exploration of all the possible biochemical pathways in a biochemical network than was previously achieved. In addition, the stereochemistry generated by a generalized enzyme function can be specified, allowing a more accurate reproduction of a metabolic pathway, as shown in the study of polyketide biosynthesis. Therefore, the inclusion of stereochemical information in the representation of molecules and reactions makes it possible to accurately represent and explore metabolic networks.

Chapter 5

Generation of *in silico* Polyketide Libraries

The modularity of the erythromycin PKS led to the development of a combinatorial approach for the metabolic engineering of these genes in order to generate a library for the antibiotic erythromycin [8, 9]. This library, which is comprised of over 50 structures, was generated through the use of the AT, DH/KR and ER/DH/KR domains of the rapamycin PKS, as well as a linker between an AT domain and the KS of the following module; consequently, the full range of reduction can be achieved, as well as the incorporation of malonyl-ACP as an extender unit instead of methylmalonyl-ACP. Additionally, experimentalists have attached the thioesterase domain at the end of modules 2, 3 and 5, resulting in the synthesis of shorter polyketide structures [4, 5].

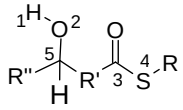
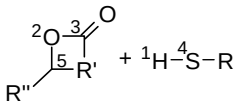
The theoretical analysis of the synthesis that was performed suggests that a large number of structures are missing from the experimentally generated erythromycin library [29]. Consequently, polyketide synthesis was implemented in BNICE for the high-throughput identification of all the erythromycin derivatives as well as for the generation of other polyketide libraries. In addition, reverse polyketide synthesis was also implemented in the framework and can also be used in order to determine the specific sequence of reactions, corresponding to the

order and identity of the modules in the PKS, required for the synthesis of every possible polyketide structure.

5.1 Implementation of Polyketide Synthesis in the BNICE Framework

The enzymes involved in polyketide biosynthesis were previously identified and the list of corresponding generalized enzyme functions was generated; these are listed in Table 4.1 - Table 4.4. As shown in Table 4.1, the condensation reaction, which is catalyzed by the 4.1.1/2.3.1 generalized enzyme functions, was subdivided into 4.1.1/2.3.1R and 4.1.1/2.3.1S in order to distinguish between the enzymes that synthesize the different chiral isomers. Additionally, the ketoreductase, the dehydratase and the enoyl reductase enzymes belong to the generalized enzyme functions 1.1.1, 4.2.1 and 1.3.1, respectively, which are subdivided into 1.1.1S, 1.1.1R, 4.2.1S, 4.2.1R, 1.3.1S, and 1.3.1R. The cyclization reaction, which is catalyzed by the thioesterase domain, does not change the chirality of the structure; therefore, only one generalized enzyme function is defined, as shown in Table 5.1. These 11 stereochemistry-specific generalized enzyme function operators were implemented in the BNICE framework in order to identify all possible polyketide structures.

Table 5.1. Table of reaction operator for the stereochemistry-specific cyclization generalized enzyme function.

Reactant	Product	Reaction Operator
Generalized Enzyme Function CyclizationR		
		

	1H	2O	3C	4S
1H	0	1	0	0
2O	1	4	0	0
3C	0	0	0	1
4S	0	0	1	4

Cyclization of 5C can be R or S

	1H	2O	3C	4S
1H	0	0	0	1
2O	0	4	1	0
3C	0	1	0	0
4S	1	0	0	4

	1H	2O	3C	4S
1H	0	-1	0	1
2O	-1	0	1	0
3C	0	1	0	-1
4S	1	0	-1	0

Chirality of 5C remains the same

Additionally, reverse polyketide biosynthesis was implemented in the framework. The generalized enzyme functions in the reverse synthesis are simply the reverse of those shown in Table 4.1 - Table 4.4 and Table 5.1: 4.1.1/2.3.1R_Rev, 4.1.1/2.3.1S_Rev, 1.1.1R_Rev, 1.1.1S_Rev, 4.2.1R_Rev, 4.2.1S_Rev, 1.3.1R_Rev, 1.3.1S_Rev, Cyclization_Rev. For example, in the 4.1.1/2.3.1R_Rev generalized enzyme function, the functional group of the reactant(s), and its corresponding matrix, is the functional group and matrix of the product(s) in the 4.1.1/2.3.1R reaction shown in Table 4.1; similarly, the functional group of the product(s) in the 4.1.1/2.3.1R_Rev reaction is the functional group of the substrate(s) in the 4.1.1/2.3.1R reaction. Consequently, the 4.1.1/2.3.1R_Rev matrix operator is the negative of the 4.1.1/2.3.1R matrix operator shown in Table 4.1; the same applies to the other generalized enzyme functions involved in reverse polyketide biosynthesis. Using a target polyketide structure as the input substrate, implementation of the reverse reaction operators in the framework led to the identification of the sequence of reactions required to synthesize the target structure. This

approach was illustrated using an experimentally generated library of 59 erythromycin derivatives [8] where each of the structures was used as the target structure in the framework and the corresponding order of reactions and thus the identity and order of the modules required for its synthesis was identified.

5.2 Generation of Polyketide Libraries

The experimental erythromycin library was generated using propionyl-CoA as the starter unit and both malonyl-CoA and methylmalonyl-CoA as extender units. Consequently, the total number of linear structures that can be produced is $s[b(a-a_c)+(2b-1)a_c]^m$, where s and a are the number of distinct starter and extender units that can be used, respectively; a_c is the number of chiral extender units; b is the number of β-carbon configurations; and m is the number of elongation steps [29]. However, since the cyclization of the linear chain involves a lactonization reaction between a hydroxyl group on the linear chain and the α-carbon after the final elongation step, the number of possible cyclic structures changes to $m(b-3)s[b(a-a_c)+(2b-1)a_c]^{m-1}$. Consequently, if propionyl-CoA is the starter unit (s = 1), malonyl-CoA and S-methylmalonyl-CoA are the extender units (a = 2, a_c = 1), the thioesterase module is placed after module 2 (m = 2) and chirality is studied (b = 5), the total number of cyclic structures that can theoretically be produced is 168. Therefore, as expected from this analysis, the implementation in the BNICE framework of the generalized enzyme functions for the reactions involved in polyketide synthesis results in the identification of 168 structures after two elongation steps, as illustrated in Figure 5.1. The framework can also be used with alternate choice and number of starter and

extender units as well as for a different number of elongation steps; consequently, the described methodology can be used to identify all the theoretically possible polyketide structures.

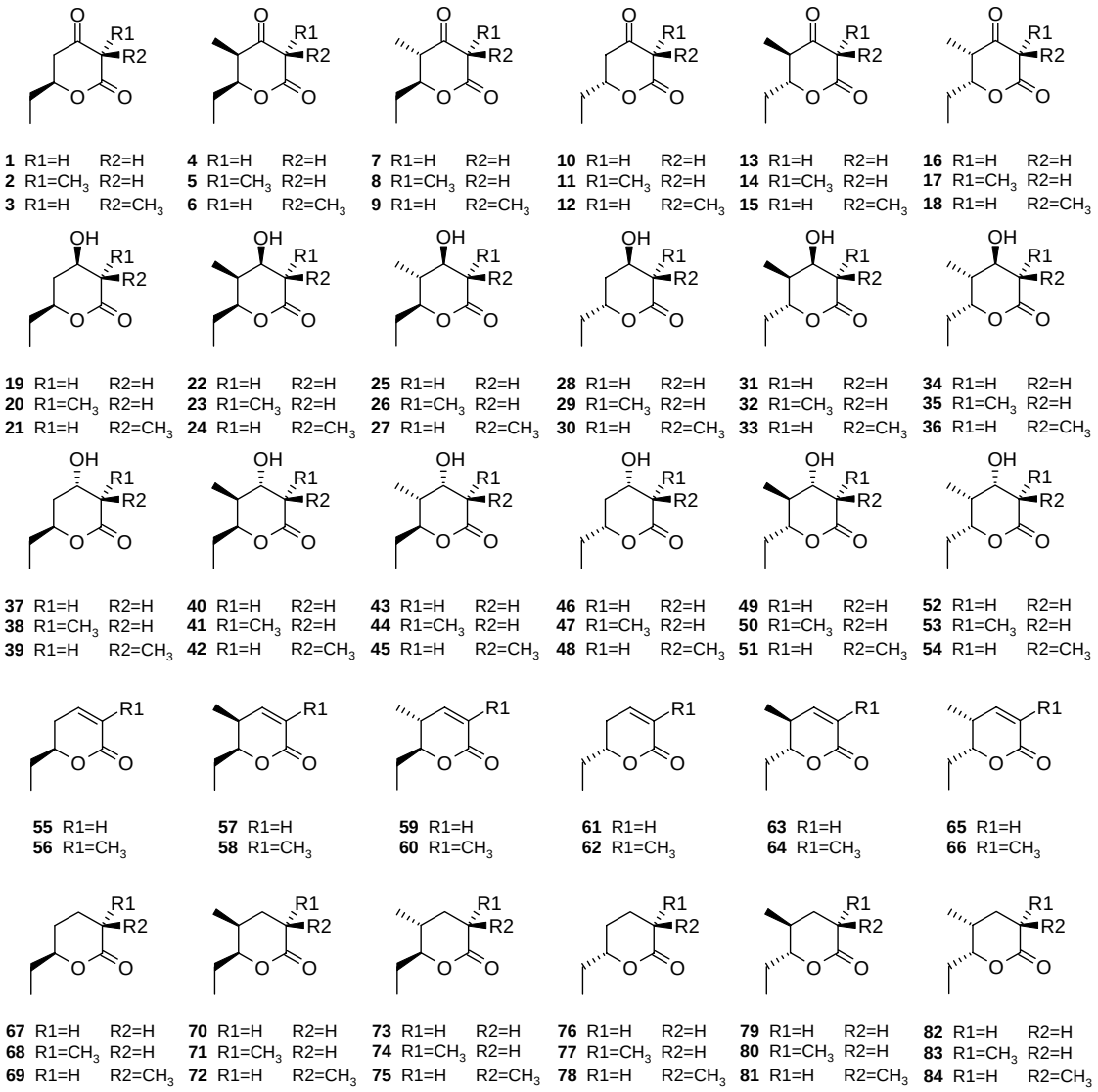


Figure 5.1. Erythromycin derivatives with six-member rings identified by the BNICE framework. These structures are synthesized after two elongation steps using propionyl-CoA as the starter unit and malonyl-CoA and methylmalonyl-CoA as the extender units. The same number of four-member ring structures is identified; these are omitted for figure clarity.

5.3 Identification of Genome Sequence

The BNICE framework utilizes a set of input structures to automatically identify all the possible products using the specified reaction rules. Therefore, the framework can theoretically be used in reverse to explore all the possible reactions that can give rise to one of those possible products. Since the synthesis of each erythromycin derivative depends directly on the choice and order of modules in the PKS, the framework can be used to identify the specific module sequence required to produce a specific polyketide structure. The reaction rules for reverse polyketide synthesis were therefore implemented in the framework which was used to identify the module sequence required for the synthesis of each of the 59 erythromycin derivatives that have been experimentally generated [8].

Of the 59 structures that were experimentally observed, 18 erythromycin derivatives were synthesized with mutations in only one of the modules of the erythromycin PKS. In every case, as illustrated in Table 5.2, a mutation in the genomic sequence results in a different polyketide structure. Therefore, it is possible to identify the specific mutations that were performed in the generation of the library. For example, the replacement of the AT site in module 1 by an AT domain specific for malonyl-CoA was the only experimental mutation in module 1; this change led to the synthesis of structure 35, as shown in Table 5.2. The same mutation was achieved in module 3, represented by structure 36. Additionally, as verified by the data in Table 5.2, the ER domain in module 4 was deleted, leading to the synthesis of structure 37. However, in the generation of the library, modules 2 and 5 were allowed to each have seven possible mutations, resulting from combinations of using AT domains specific for methylmalonyl-CoA or malonyl-CoA, as well as insertions/deletions of KR, DH and ER active sites, as illustrated in Figure 5.2a.

However, as shown in Table 5.2 only five structures from the seven possible derivatives were identified. It was determined that the missing mutations in module 2 involved the deletion of the KR domain. The same analysis was performed for the mutations in module 5, resulting in deficiencies in the simultaneous replacement of the AT domain and the introduction of the DH and DH/ER domains or deletion of the KR domain. These seven mutations were also implemented in module 6, in addition to the replacement of the KR domain to create a different chirality in the β-carbon, giving rise to a total of six additional mutation schemes although only two added possible structures, as illustrated in Figure 5.2; however, of these nine structures, only six were observed experimentally. An analysis of the missing structures shows that the addition of the DH/ER domains, as well as the simultaneous addition of the DH and DH/ER domains and change of the stereochemistry of the KR, did not yield products. Therefore, of the 26 possible structures that should be generated, only 18 were actually identified.

Table 5.2. Erythromycin derivatives synthesized with mutations in only one of the modules of the Ery PKS.

Structure # [a]	**Module #**	**Mutation**[b]
35	1	AT^{1}
2	2	AT^{1}
3	2	DH
4	2	DH, ER
5	2	AT^{1}, DH
6	2	AT^{1}, DH, ER
36	3	AT^{1}
37	4	Delete ER
7	5	AT^{1}
8	5	Delete KR
9	5	DH
10	5	DH, ER
11	6	AT^{1}
12	6	Delete KR
13	6	DH
14	6	AT^{1}, Delete KR
15	6	AT^{1}, KR^{S}
38	6	KR^{S}

[a]The structure number refers to the labeling used in the experimentally generated library of erythromycin derivatives.

[b]Notation: AT^{1} = replacement of AT domain for specificity for malonyl-CoA; DH = insertion of DH domain; ER = insertion of ER domain; Delete KR = deletion of KR domain; KR^{S} = replacement of KR domain for stereochemistry change in β-carbon

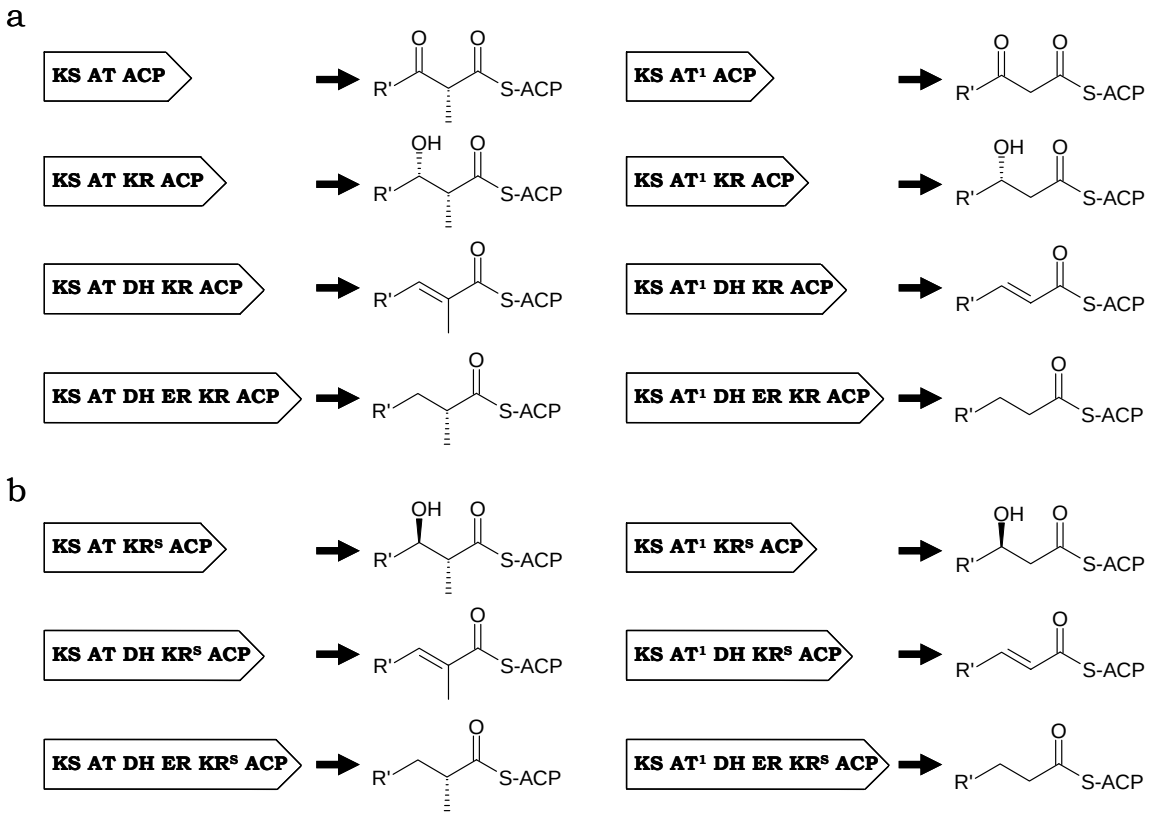


Figure 5.2. Schematic of the module sequences and corresponding structural features possible for modules 2, 5 and 6. The active sites in the illustrated proteins include ketosynthase (KS), acyltransferase specific for methylmalonyl-CoA (AT), acyltransferase specific for malonyl-CoA (AT^1), acyl carrier protein (ACP), ketoreductase (KR), ketoreductase with different stereochemical orientation (KR^S), dehydratase (DH), enoyl reductase (ER) and thioesterase (TE). The mutations performed in modules 2 and 5 are shown in a while the mutations for module 6 include those shown in (a) and (b). Some of the mutations shown in (a) and (b) synthesize the same structure.

The remaining 41 structures in the experimentally generated library were synthesized from a PKS with mutations in two modules. As shown in Table 5.3, of these 41 structures, 11 are double mutants involving mutations in modules 2 and 5. As expected, those mutations that did not result in a product in the single mutation study did not produce a polyketide structure as part of a double mutation; for example, the deletion of the KR domain in module 2 did not yield a structure and, similarly, this mutation did not result in the synthesis of a polyketide with an

additional mutation in module 5. An analysis of all the possible double mutations shows that, based on the genetic modifications undertaken in the library, 247 new structures should have been synthesized, not including those that would be produced from mutations in a single module. Therefore, since only 41 new structures were identified, there are 206 structures missing.

Table 5.3. Erythromycin derivatives synthesized with mutations in two of the modules of the Ery PKS.

Structure # [a]	Module #	Mutation[b]	Module #	Mutation[b]
16	2	AT^{I}	5	Delete KR
17	2	AT^{I}	5	DH
18	2	AT^{I}	5	DH, ER
19	2	DH	5	AT^{I}
20	2	DH	5	Delete KR
21	2	DH, ER	5	AT^{I}
22	2	DH, ER	5	Delete KR
23	2	DH, ER	5	DH
24	2	AT^{I}	5	AT^{I}
39	2	DH	5	AT^{I}, Delete KR
40	2	DH	5	AT^{I}, DH, ER
25	2	AT^{I}	6	Delete KR
26	2	AT^{I}	6	DH
27	2	DH	6	AT^{I}
28	2	DH	6	Delete KR
29	2	DH, ER	6	AT^{I}
30	2	DH, ER	6	Delete KR
31 (43[c])	2	AT^{I}	6	AT^{II}, Delete KR
32	2	DH	6	AT^{I}, Delete KR
33	2	DH, ER	6	AT^{I}, Delete KR
34	2	DH, ER	6	AT^{I}
41	2	DH	6	KR^{S}
42	2	DH, ER	6	KR^{S}
44	2	DH	6	AT^{I}, KR^{S}
45	2	DH, ER	6	AT^{I}, KR^{S}
46	2	AT^{I}, DH	6	AT^{I}
47	2	AT^{I}, DH	6	KR^{S}
48	2	AT^{I}, DH, ER	6	Delete KR
49	2	AT^{I}, DH, ER	6	AT^{I}
50	3	AT^{I}	6	AT^{I}
51	5	AT^{I}	6	AT^{I}
52	5	AT^{I}, Delete KR	6	AT^{I}
53	5	Delete KR	6	KR^{S}
54	5	DH	6	AT^{I}
55	5	DH	6	Delete KR
56	5	Delete KR	6	Delete KR
57	5	AT^{I}	6	AT^{I}, KR^{S}
58	5	AT^{I}	6	AT^{I}, Delete KR
59	5	AT^{I}, Delete KR	6	AT^{I}, KR^{S}
60	5	DH	6	AT^{I}, KR^{S}
61	5	DH	6	AT^{I}, Delete KR

[a]The structure number refers to the corresponding number in the experimentally generated library.
[b]Notation: AT^{I} = replacement of AT domain for specificity for malonyl-CoA; DH = insertion of DH domain; ER = insertion of ER domain; Delete KR = deletion of KR domain; KR^{S} = replacement of KR domain for stereochemistry change in β-carbon
[c]Structures 31 and 43 are equivalent.

5.4 Conclusions

The modular nature of reduced polyketides allows the manipulation of the variables involved in the synthesis of these metabolites. Therefore, libraries of polyketides have been experimentally generated, as in the case of the erythromycin library, which was constructed using propionyl-CoA as the starter unit and malonyl-CoA and methylmalonyl-CoA as extender units. However, a theoretical analysis of polyketide synthesis suggests that a large number of erythromycin derivatives were not identified in the generation of this library. Implementation of polyketide synthesis in the BNICE computational framework allows the identification of the complete erythromycin library. This analysis can also be extended for the identification of other polyketide libraries; more specifically, a library of all possible polyketide structures can be constructed by utilizing all the theoretically possible starter and extender units, number of elongation steps, and degree of reduction of the β-carbon in each elongation step.

In addition to library generation, the framework can also be used to identify the set of reactions that are required to produce a target structure. In other words, implementation of reverse polyketide synthesis in the BNICE framework allows the identification of the specific module arrangement required for the synthesis of each of the structures in the polyketide library. This reverse implementation of the framework uses the target structure as the input reactant and the reverse generalized enzyme functions as the reaction rules. Consequently, the module sequence for the synthesis of any structure missing from the library can be identified, and, therefore, the metabolic engineering actions required to produce each of the missing structures can be identified.

Chapter 6

Assessment of Synthetic Feasibility of Polyketide Derivatives

The rise of antibacterial resistance leads to the increased need for the development of new antibiotics; consequently, these have become a target for a number of metabolic engineering initiatives, which have focused on engineering an organism for the synthesis of either existing or altered antibiotic structures [24, 44-46]. Specifically, a number of the current antibiotics form part of the class of polyketides, cellular metabolites with a wide array of pharmacological properties [1, 25]. A large number of the polyketide structures are naturally produced in organisms that are not feasible for large-scale fermentation [44]. Therefore, in order to produce a target polyketide in the quantities necessary for commercial use, it would be beneficial to use a heterologous host. Recently, research efforts have been invested for the production of 6-deoxyerythronolide B, or 6dEB, the precursor to the antibiotic erythromycin, in *Escherichia coli* [24, 46].

Propionyl-CoA, the starter unit in the synthesis of 6dEB, is native to *E. coli*. However, the extender unit, (S)-methylmalonyl-CoA, is not naturally produced in this organism; therefore, the propionyl carboxylase genes from *S. coelicolor* were inserted into *E. coli*, allowing *E. coli* to synthesize (S)-methylmalonyl-CoA from propionyl-CoA . The erythromycin polyketide synthase (EryPKS), responsible for the synthesis of the erythromycin precursor from its starter

and extender units, was also introduced into *E. coli*. These cells were used in fed-batch fermentation experiments in which they were grown in glucose media at 37ºC, pH of 7.1 and a level of dissolved oxygen of 50% air saturation. Once the cells reached stationary phase, the glucose level was exhausted, the carbon source was then switched to propionate, and synthesis of 6dEB was induced. Although a high yield was expected, experimental results show that less than 10% of the propionate was actually converted into 6dEB . Therefore, although no other products beside 6dEB were observed, a 1.4% molar yield was obtained, illustrating that the carbon is being utilized in an alternate pathway in *E. coli*.

6.1 Cellular Feasibility of Synthesis of Erythromycin

6.1.1 Metabolic Flux Analysis of E. coli Metabolism

Metabolic flux analysis is used to determine the relative flux of all the reactions in a cell while satisfying all cellular constraints of mass, energy and redox. When the uptake rates are experimentally determined, metabolic flux analysis can be used to determine the maximum production rate, and subsequently yield, of the product of interest. Assuming quasi steady-state conditions for the intracellular metabolites, the metabolite mass balance equations can be written as:

$$\underline{\underline{N}} \cdot \underline{v} = \underline{0} \tag{6.1}$$

where $\underline{\underline{N}}$ is an *n*-by-*m* matrix of the stoichiometric coefficients of the mass balance of the *n* metabolites in each of the *m* reactions; $\underline{v}$ is a vector of the *m* metabolic fluxes; and $\underline{0}$ is the *n*-by-*1* zero vector. Since the *E. coli* metabolic system is underdetermined, with a larger number of reactions compared to metabolites, a set of feasible solutions is obtained. A particular solution

from this set can be found through the use of linear optimization [47, 48]. For example, in order to determine the maximum theoretical yield of a metabolite, the problem can be formulated as a linear optimization problem:

$$\begin{aligned} &\text{Maximize } v_{\text{product}} \\ &\text{Subject to} \\ &\qquad \underline{\underline{N}} \cdot \underline{v} = \underline{0} \\ &\qquad a_i \le v_i \le b_i \end{aligned} \tag{6.2}$$

where v_{product} is the production rate of the target molecule and a_i and b_i are thermodynamic and physiological constraints on the minimum and maximum values for each reaction flux. This technique has been successfully used to investigate the capability of *E. coli* to individually synthesize different precursors, as well as to study the ability of this network to meet the demands imposed during cellular growth [47-49]. The shadow prices, or dual variables of each of the constraints in the linear optimization problem, can also be determined. The shadow price of a constraint measures how much the objective would improve if the constraint were to be relaxed by a small amount; consequently, in metabolic flux analysis, they can be used to analyze the effect of each metabolite mass balance on the objective function of the problem.

The analysis of the effect of implementing the synthesis of 6dEB into the *E. coli* metabolism was performed using a model of *E. coli*, identified as *i*JR904, which was constructed from 904 genes [50]. This model includes 931 biochemical reactions and 625 metabolites, in addition to the transport reactions between the extracellular and intracellular environments. These reactions in the *i*JR904 model essentially describe the wild-type *E. coli* metabolism. Therefore, the reactions essential for polyketide biosynthesis were added to the model. Specifically, the reaction that synthesizes methylmalonyl-CoA from propionyl-CoA and the overall reaction for

the synthesis of 6dEB from its starter and extender units were added to the *i*JR904 model, resulting in a model with 933 reactions and 626 metabolites. The biomass reaction was not considered because 6dEB production is induced during the stationary phase .

6.1.2 Cellular Feasibility of 6dEB Production

6.1.2.1 Synthesis Pathway for 6dEB Synthesis

The reaction pathway for the production of 6dEB from propionate is illustrated in Figure 6.1; it consists of three reactions: (1) the synthesis of the starter unit propionyl-CoA from propionate; (2) the conversion of propionyl-CoA to the extender unit (S)-methylmalonyl-CoA and (3) the synthesis of 6dEB from propionyl-CoA and (S)-methylmalonyl-CoA, which is comprised of a total of six reduction reactions which utilize $NADPH_2^+$, and one dehydration reaction, releasing a molecule of water [1, 8]. Based on the stoichiometry of the reactions in this pathway, the overall synthesis equation for the production of 6dEB is:

$$7\ \text{Propionate} + 7\ \text{ATP} + 6\ NADPH_2^+ \rightarrow\ 6\text{dEB} + 7\ \text{ADP} + 7\ P_i + 6\ NADP^+ + H_2O$$

corresponding to a 14% molar yield for the production of 6dEB.

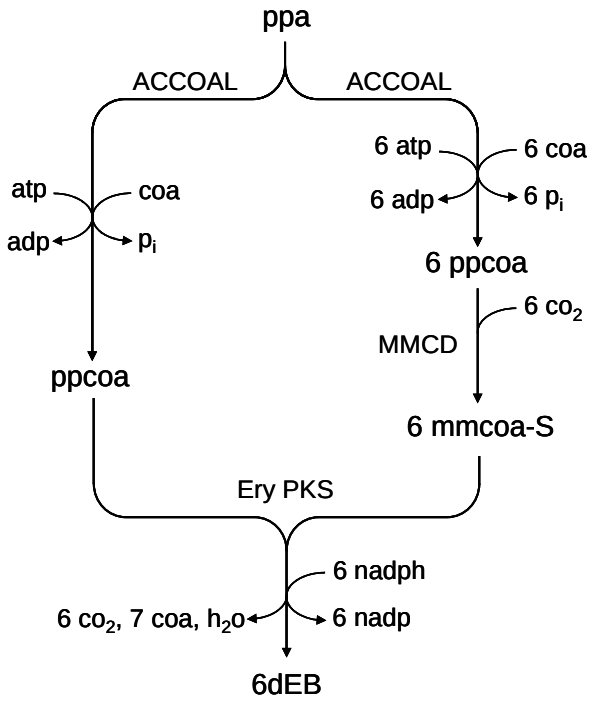


Figure 6.1. Production of 6dEB from propionate. The metabolites and reactions are detailed in Appendix A.

6.1.2.2 Metabolic Network Yield for 6dEB Production

The pathway analysis performed in the previous section assumes that excess NAD(P)H_2^+ and ATP are readily available in the cellular environment. However, the introduction of novel reactions into an organism generates competition with the native pathways for carbon, energy, and redox resources. Therefore, the introduction of 6dEB synthesis in *E. coli* requires an adjustment of the bacterial metabolism in order to produce the necessary amount of energy and redox required for the synthesis of the erythromycin precursor. Since no products besides 6dEB were observed, the resources that are influencing the decrease in the measured yield are unknown. Therefore, in order to quantitatively assess the effects of this competition, a metabolic flux analysis was performed using the set of all reactions in the *E. coli* metabolism.

The analysis suggests that a maximum 6dEB synthesis yield of 0.11 mmol 6dEB [mmol propionate]$^{-1}$, or an 11% molar yield, is feasible with no cellular growth and under zero maintenance energy requirements assuming that the transport of propionate into the cell and of 6dEB out of the cell does not require energy. The pathway shown in Figure 6.2 is the set of reactions necessary to obtain this maximum theoretical yield of 6dEB from propionate. The first two steps in the pathway involve the uptake of propionate by the cell and its subsequent conversion to propionyl-CoA. The propionyl-CoA is used for the synthesis of 6dEB, as well as for the production of $NADPH_2^+$ through the TCA cycle. However, as shown in Figure 6.2, two of the reactions in the TCA cycle involve the release of carbon dioxide, thereby reducing the carbon yield of 6dEB.

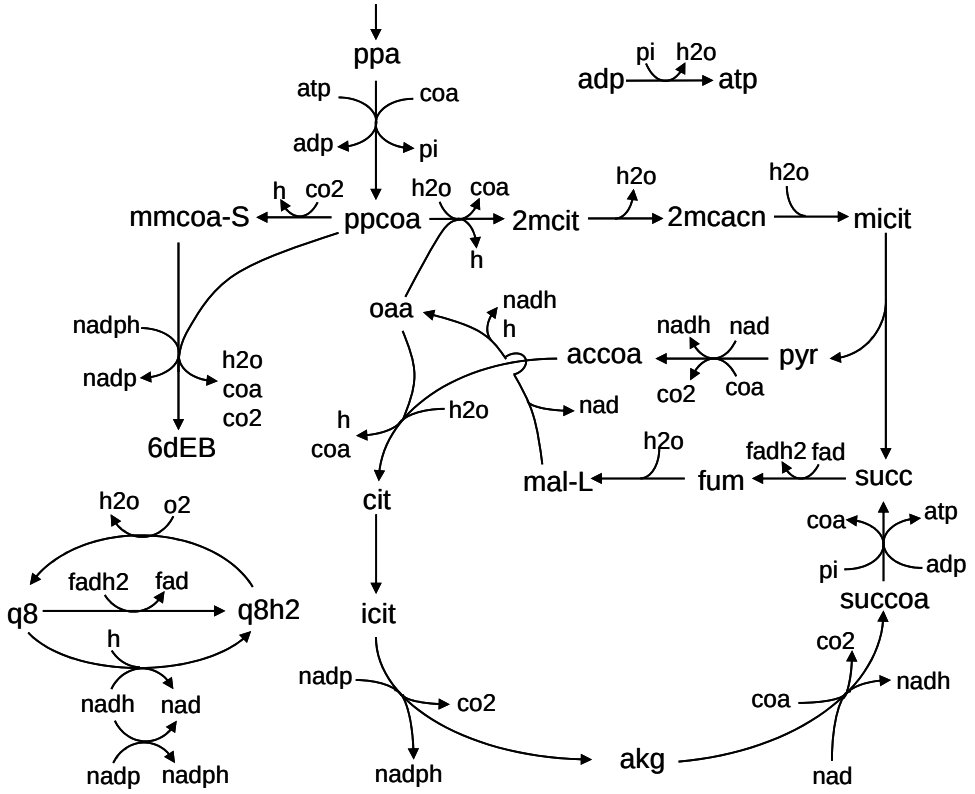


Figure 6.2. E. coli reactions for the synthesis of 6dEB from propionate (ppa). This reaction network represents one reaction set that maximizes the production of 6dEB assuming negligible

maintenance energy and aerobic conditions; more than one solution is possible. The full names of the metabolites are listed in Appendix A. The enzymes that catalyze the reactions, as well as the reaction fluxes, are not included for figure clarity.

The 11% molar yield obtained from this analysis is significantly higher than the 1.4% molar yield observed experimentally . The difference between the expected and the observed molar yields might be due to the fact that living cells require energy for maintenance. The previous analysis suggested that, in order to obtain the maximum theoretical yield, the maintenance energy must be zero; however, the experimentally measured physiological maintenance energy value for *E. coli* using glucose as the carbon source is 5.87 mmol ATP g[dry weight]$^{-1}$ hr^{-1} under non-growth conditions [51]. Metabolic flux analysis of the *E. coli* metabolism suggests that the organism is not capable of producing any 6dEB from the experimentally measured specific propionate uptake rate of 0.02 mmol g[dry weight]$^{-1}$ hr^{-1} (based on the reported 0.35 g L^{-1} hr^{-1} at an OD_{600} of 45 and assuming 0.00023 g[dry weight] OD_{600}^{-1}) and with an ATP maintenance requirement of 5.87 mmol ATP g[dry weight]$^{-1}$ hr^{-1}. In fact, the minimum specific propionate uptake rate that is able to support the experimentally measured maintenance energy of 5.87 mmol ATP g[dry weight]$^{-1}$ hr^{-1} is 0.65 mmol g[dry weight]$^{-1}$ hr^{-1}, and that does not result in any 6dEB production; an additional increase in the specific propionate uptake rate is required for the production of 6dEB, as shown in Figure 6.3.

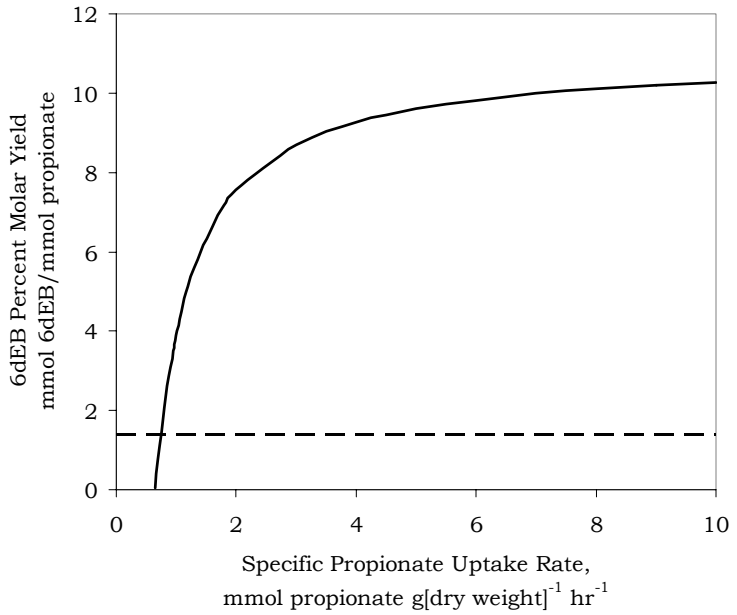


Figure 6.3. Percent molar yield of 6dEB as a function of the specific propionate uptake rate for the physiological ATP maintenance rate of 5.87 mmol ATP g[dry weight]$^{-1}$ hr^{-1} under aerobic conditions. The horizontal dashed line represents the experimentally observed yield of 6dEB.

Metabolic flux analysis shows that the maximum amount of ATP that the cells are capable of synthesizing with the experimentally observed specific propionate uptake rate is 0.18 mmol ATP g[dry weight]$^{-1}$ hr^{-1}, although under these conditions 6dEB is not produced. Therefore, it is possible that the ATP maintenance requirement is lower for these cells, which might result in minimized activity of the various cellular processes requiring energy. Consequently, cellular energetics can explain the experimentally observed 6dEB production rate of 0.00028 mmol g[dry weight]$^{-1}$ hr^{-1}. Since the previous analysis shows that energetics play a critical role in controlling the synthesis of 6dEB, a transport process requiring energy for the uptake of propionate or for the release of 6dEB would further reduce the maximum theoretical yield of 6dEB.

It was found that the experimentally observed rates are only feasible for a maximum maintenance energy of 0.158 mmol ATP g[dry weight]$^{-1}$ hr^{-1} and a specific oxygen uptake rate of 0.062 mmol O_2 g[dry weight]$^{-1}$ hr^{-1} (Figure 6.4) and the feasible range of specific oxygen uptake

rate depends on the ATP maintenance requirement. However, a microaerobic environment would limit the amount of oxygen the cells are able to uptake. Under zero maintenance energy requirements, the production of the observed 6dEB yield is feasible when the specific oxygen uptake rates range from 0.01 to 0.06 mmol g[dry weight]$^{-1}$ hr^{-1}. However, as the maintenance energy increases, the range of specific oxygen uptake rates that result in the observed yield decreases until, for a maintenance energy of 0.158 mmol g[dry weight]$^{-1}$ hr^{-1}, only the maximum oxygen uptake rate of 0.062 mmol g[dry weight]$^{-1}$ hr^{-1} allows the synthesis of the experimentally observed 6dEB yield. Beyond this limit, increased maintenance energy will require increased flow of carbon into the TCA cycle for the production of redox and the generation of energy through respiration. This redirection of carbon flux through the TCA cycle will result in a lower yield of 6dEB due to the increased loss of carbon into carbon dioxide.

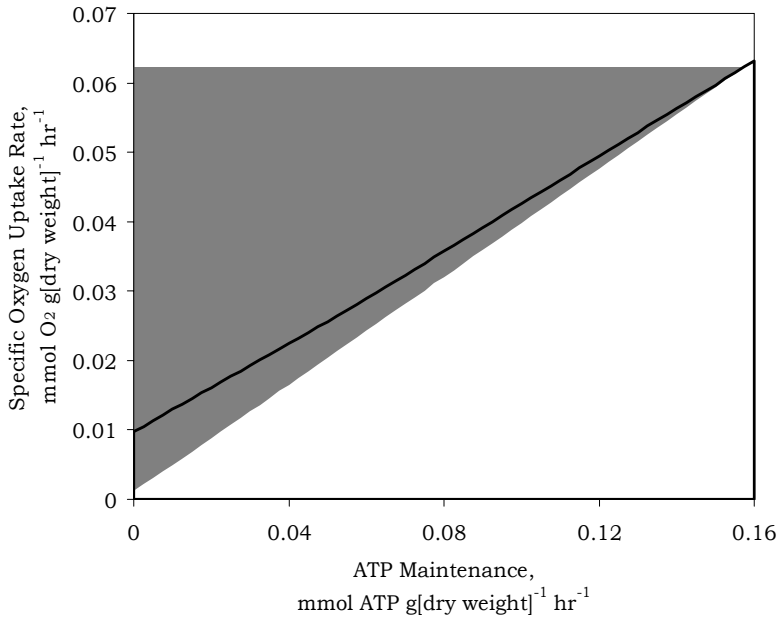


Figure 6.4. Feasible ATP maintenance requirements and specific oxygen uptake rates that result in the experimentally observed specific 6dEB production rate of 0.00028 mmol g[dry weight]$^{-1}$ hr^{-1} from the experimentally measured propionate uptake rate of 0.02 mmol g[dry weight]$^{-1}$ hr^{-1} (gray area). The solid line refers to the line of optimality for the synthesis of the maximum amount of 6dEB from propionate as a function of the ATP maintenance requirement and specific oxygen uptake rates for a specific propionate uptake rate of 0.02 mmol g[dry weight]$^{-1}$ hr^{-1}.

The maximum theoretical yield under increasing maintenance energy conditions is achieved with an increase in the specific oxygen uptake rate, as illustrated by the solid line in Figure 6.4. In order to determine the impact of the aeration culture conditions in the synthesis of 6dEB, a metabolic flux analysis was performed assuming a microaerobic environment, where the specific oxygen uptake rate is constrained. If the maximum specific oxygen uptake rate is 0.02 mmol O_2 g[dry weight]$^{-1}$ hr^{-1}, the maximum yield of 6dEB is the same as for an unconstrained specific oxygen uptake rate for low ATP maintenance requirements, as shown in Figure 6.5. However, for low ATP maintenance requirements, the maximum yield of 6dEB for a maximum specific oxygen uptake rate of 0.01 mmol O_2 g[dry weight]$^{-1}$ hr^{-1} is lower than the yield for an

unconstrained specific oxygen uptake rate. Additionally, as the ATP maintenance requirement increases above 0.03 mmol ATP g[dry weight]$^{-1}$ hr^{-1}, a specific oxygen uptake rate of 0.01 mmol O_2 g[dry weight]$^{-1}$ hr^{-1} cannot sustain any production of 6dEB and the amount of oxygen required for the synthesis of 6dEB increases. Therefore, as shown by Figure 6.5, a tight coupling exists between the specific oxygen uptake rate, the maintenance energy requirement, and the product yield for the experimentally observed specific propionate uptake rates.

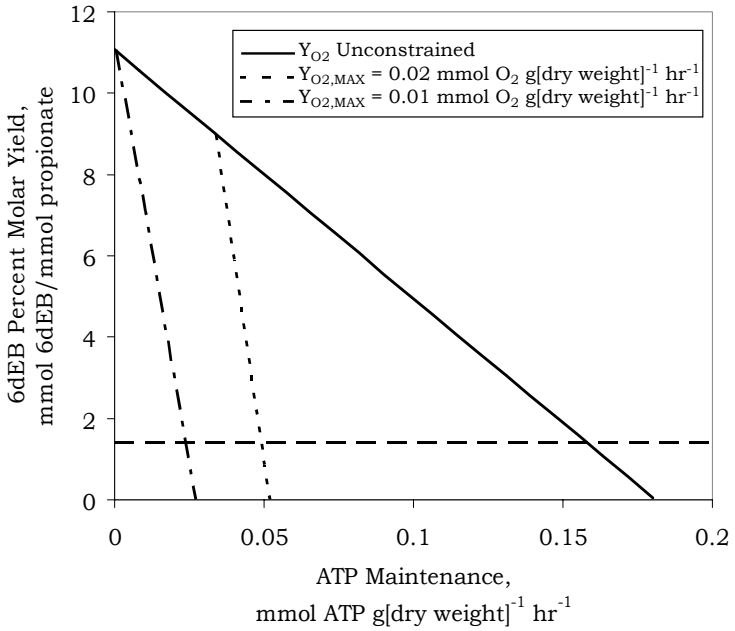


Figure 6.5. Percent molar yield of 6dEB from propionate as a function of the amount of ATP required for maintenance using a specific propionate uptake rate of 0.02 mmol g[dry weight]$^{-1}$ hr^{-1}. The solid line refers to an aerobic environment. The dotted lines refer to microaerobic environments in which the maximum specific oxygen uptake rates are 0.01 and 0.02 mmol O_2 g[dry weight]$^{-1}$ hr^{-1}. The horizontal dashed line represents the experimentally observed yield of 6dEB.

6.1.3 Effect of Glucose as the Carbon Source on 6dEB Production

Shadow price analysis suggests that, under energy and oxygen limitations, glucose could be a better carbon source for 6dEB production. The conversion of propionate to propionyl-CoA and methylmalonyl-CoA, the precursors in 6dEB synthesis, does not generate ATP or $NADPH_2^+$;

therefore, *E. coli* needs to utilize the TCA cycle in order to synthesize the necessary amounts of ATP and $NADPH_2^+$, resulting in a decrease in carbon yield due to the release of carbon dioxide. On the other hand, since the glycolysis pathway produces ATP and $NADH_2^+$ without the release of carbon dioxide, the use of glucose as the carbon source was analyzed in order to determine if there is an improvement in the yield of 6dEB.

Metabolic flux analysis, for a specific glucose uptake rate of 1 mmol g[dry weight]$^{-1}$ hr^{-1}, shows that the maximum theoretical molar yield for 6dEB is 21% with no cellular growth and under zero maintenance energy requirements, which is higher than the 11% molar yield obtained with propionate as the carbon source under the same conditions. However, glucose has six carbon atoms while propionate only contains three; thus, in order to compare the carbon yield for 6dEB between these two carbon sources, it is necessary to calculate the yield of 6dEB per C_3 carbon source. Utilizing this measure, the molar yield of 6dEB using glucose as the carbon source is 10.5% per C_3, which is lower than the yield obtained from propionate; the corresponding *E. coli* metabolism for the synthesis of 6dEB using glucose as the carbon source is shown in Figure 6.6. This reaction network includes the glycolysis and pentose phosphate pathways, in addition to the TCA cycle reactions utilized in the synthesis of 6dEB from propionate. Glucose is transformed into pyruvate through glycolysis, and pyruvate is then used in the TCA cycle for the production of succinyl-CoA, which is the precursor for the production of the starter and extender units in 6dEB synthesis from glucose.

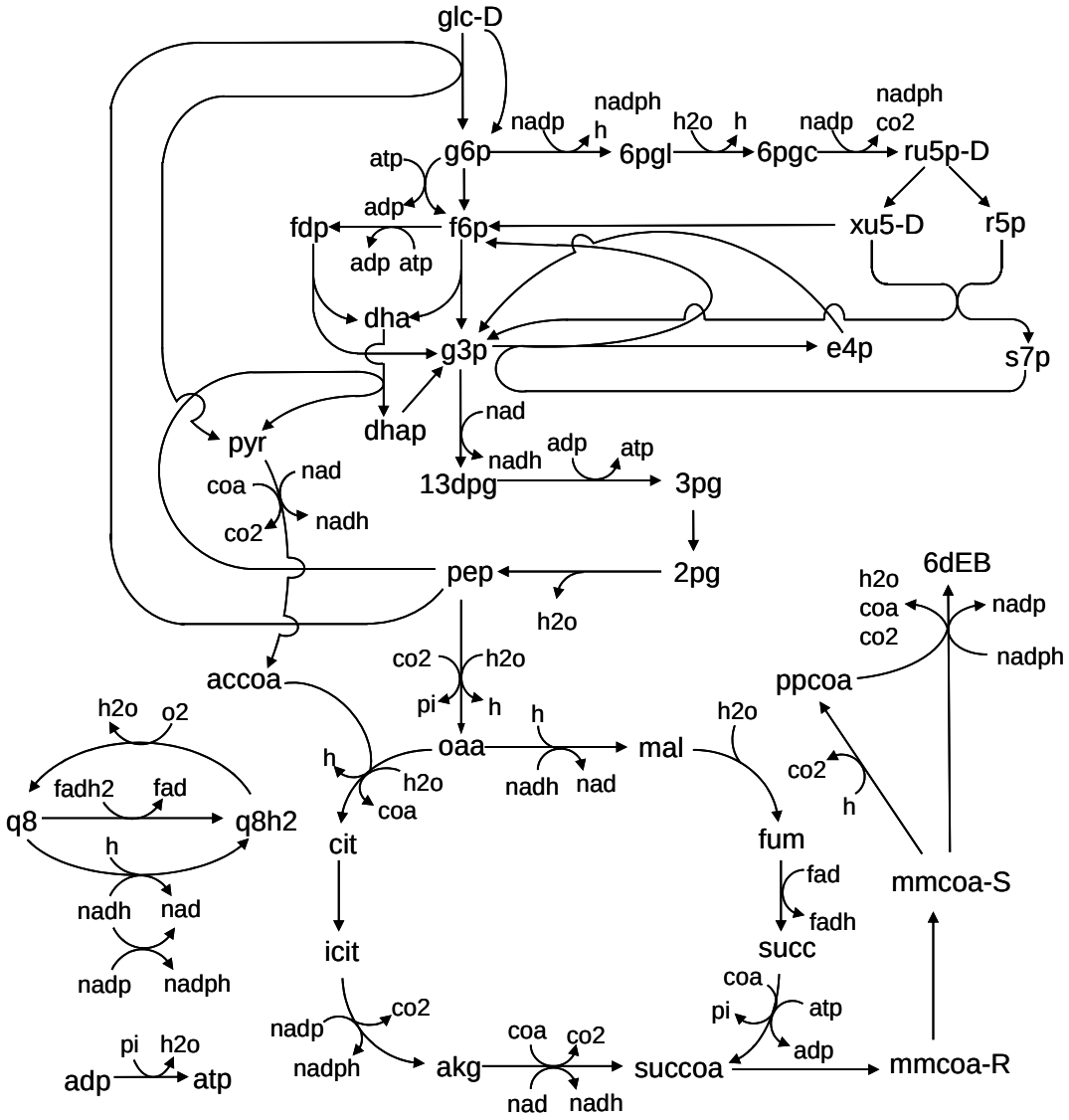


Figure 6.6. E. coli reactions for the synthesis of 6dEB from glucose (glc-D). This reaction network represents one reaction set that maximizes the production of 6dEB assuming negligible maintenance energy and aerobic conditions; more than one solution is possible. The full names of the metabolites are listed in Appendix A. The enzymes that catalyze the reactions, as well as the reaction fluxes, are not included for figure clarity.

As shown in Figure 6.7, the yield of 6dEB from glucose is lower than that determined from propionate for low ATP maintenance requirements. This lower yield from glucose might be due to the additional release of carbon dioxide from the required use of the TCA cycle since the 6dEB precursors are synthesized from succinyl-CoA, a metabolite in the TCA cycle. However,

as the maintenance requirement increases, the yield of 6dEB from glucose is higher than the yield from propionate. An analysis of the reaction fluxes in 6dEB synthesis from glucose shows utilization of the pentose phosphate pathway and the TCA cycle; the use of these two pathways involves the release of carbon dioxide, thereby decreasing the yield of 6dEB. As observed with propionate as the carbon source, the utilization of the pentose phosphate pathway and the TCA cycle increases for increasing ATP maintenance requirements; as shown in Figure 6.7, this increase in the maintenance requirement results in a decrease in the yield of 6dEB. However, the *E. coli* metabolic network is capable of sustaining higher ATP maintenance requirements when glucose is used instead of propionate as the carbon source, as illustrated in Figure 6.7.

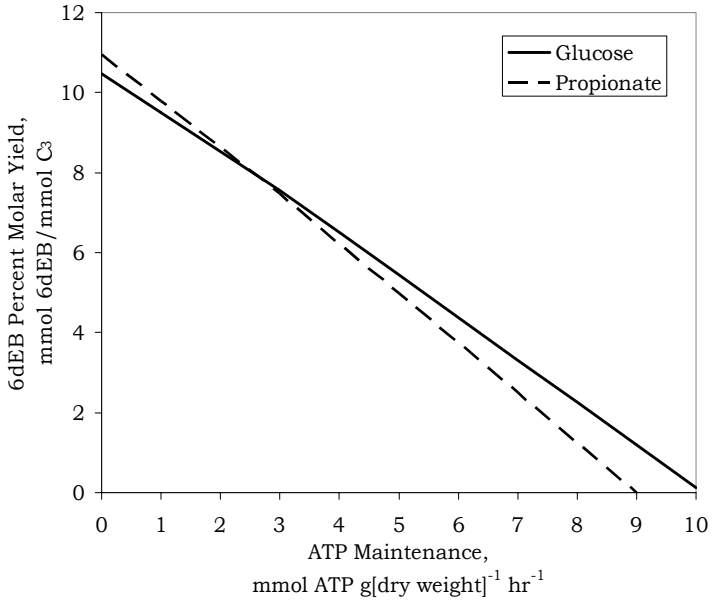


Figure 6.7. Comparison between glucose and propionate as carbon sources of the percent molar yield of 6dEB (per C_3) as a function of the amount of ATP required for maintenance, for a specific glucose uptake rate of 0.01 mmol g[dry weight]$^{-1}$ hr^{-1}and a specific propionate uptake rate of 0.02 mmol g[dry weight]$^{-1}$ hr^{-1}.

As in the propionate system, a specific glucose uptake rate of 0.01 mmol g[dry weight]$^{-1}$ hr^{-1}, which corresponds to the same carbon atom uptake rate as the experimentally measured specific propionate uptake rate of 0.02 mmol g[dry weight]$^{-1}$ hr^{-1}, is not able to sustain the physiological 5.87 mmol ATP g[dry weight]$^{-1}$ hr^{-1} maintenance energy requirement. The minimum specific glucose uptake rate that can sustain this maintenance energy requirement is 0.29 mmol g[dry weight]$^{-1}$ hr^{-1}, or 0.58 mmol C_3 g[dry weight]$^{-1}$ hr^{-1}, assuming a completely aerobic environment, which is lower than the corresponding minimum specific propionate uptake rate of 0.65 mmol g[dry weight]$^{-1}$ hr^{-1}.

The effect of microaerobic conditions was also analyzed for 6dEB synthesis from glucose. As shown in Figure 6.8, the same trend is observed for constrained specific oxygen uptake rates

using glucose as the carbon source as was shown for propionate in Figure 6.5. However, under the same microaerobic conditions examined with propionate as the carbon source, the *E. coli* metabolism for 6dEB synthesis from glucose is able to sustain higher ATP, as shown in Figure 6.9, and, as expected, the yield of 6dEB increases as the specific glucose uptake rate increases (Figure 6.10).

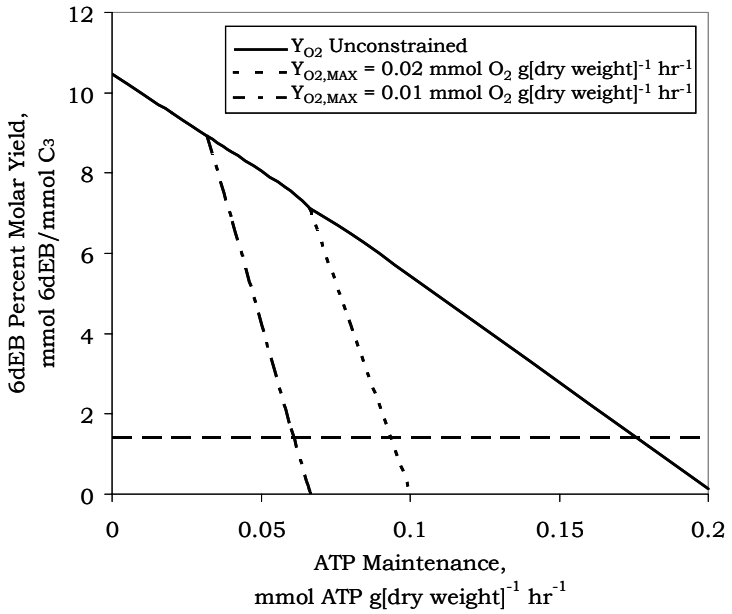


Figure 6.8. Percent molar yield of 6dEB from glucose as a function of the amount of ATP required for maintenance using a specific glucose uptake rate of 0.01 mmol g[dry weight]$^{-1}$ hr^{-1}. The solid line refers to an aerobic environment and the dotted lines refer to microaerobic environments in which the maximum specific oxygen uptake rates are 0.01 and 0.02 mmol O_2 g[dry weight]$^{-1}$ hr^{-1}. The horizontal dashed line represents the experimentally observed yield of 6dEB from propionate.

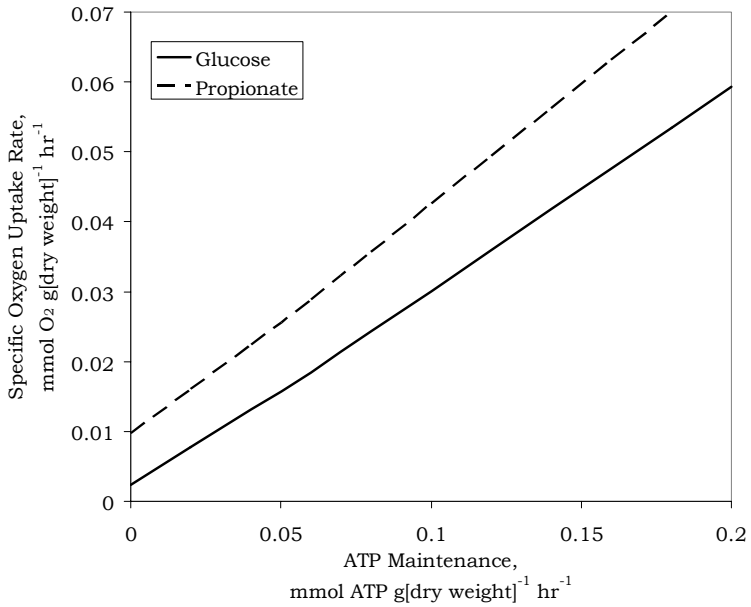


Figure 6.9. Specific oxygen uptake rates required for the synthesis of the maximum amount of 6dEB as a function of the ATP maintenance requirement, for specific glucose and propionate uptake rates of 0.01 mmol g[dry weight]$^{-1}$ hr^{-1} and 0.02 mmol g[dry weight]$^{-1}$ hr^{-1}, respectively.

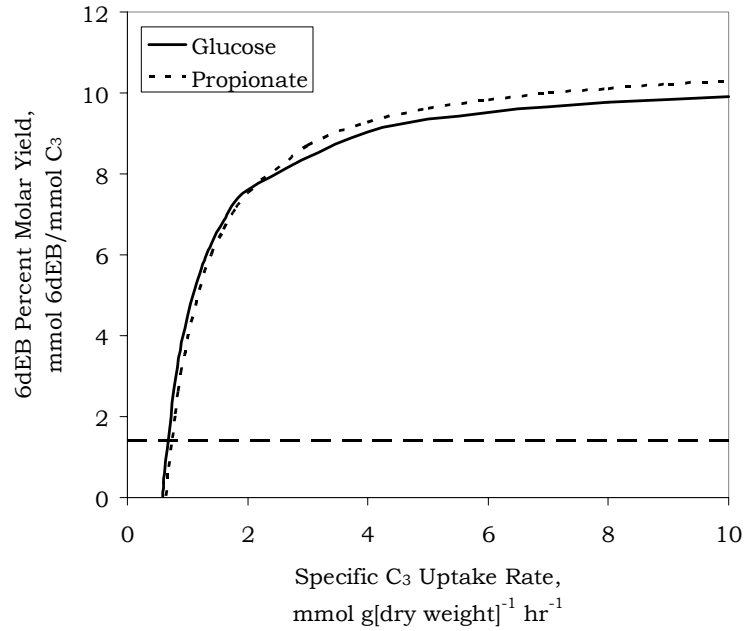


Figure 6.10. Comparison between glucose and propionate as carbon sources of the percent molar yield of 6dEB per C_3, as a function of the specific C_3 uptake rate for a physiological ATP maintenance rate of 5.87 mmol ATP g[dry weight]$^{-1}$ hr^{-1} under aerobic conditions. The horizontal dashed line represents the experimentally observed yield of 6dEB from propionate.

6.2 Cellular Feasibility for Synthesis of Polyketide Derivatives

The cellular feasibility of synthesizing polyketide derivatives in general was also investigated using *E. coli* as the heterologous host. As for erythromycin, the synthesis of polyketide derivatives from propionate consists of three main reactions: (1) the synthesis of the starter unit propionyl-CoA from propionate; (2) the conversion of propionyl-CoA to form extender unit (S)-methylmalonyl-CoA and (3) the synthesis of the derivative from propionyl-CoA and (S)-methylmalonyl-CoA, which also comprises of *m* elongation steps, *n* reduction reactions and *h* dehydration reactions. Therefore, the overall reaction for the biosynthesis of a polyketide derivative from propionate is:

$$(m+1)\text{Propionate} + (m+1)\text{ATP} + n\text{NADPH}_2^+ \rightarrow \text{Polyketide} + (m+1)\text{ADP} + (m+1)\text{P}_i + n\text{NADP}^+ + h\text{H}_2\text{O}$$

where *m* refers to the number of extender units of methylmalonyl-CoA or number of elongation steps; *n* is the number of $NADPH_2^+$ molecules that act as hydrogen donors; and *h* is the number of water molecules produced. A balance on carbon shows that the theoretical production yield for polyketide synthesis is:

$$Y_{polyketide} = \frac{1}{m+1} \tag{6.3}$$

which corresponds to a 14% molar yield for a polyketide synthesized with six elongation steps ($m = 6$), as shown for erythromycin in section 6.1.

However, the ATP and $NAD(P)H_2^+$ resources are constrained in a cellular environment, making it necessary to analyze the complete *E. coli* metabolism in order to analyze the theoretical yield for the synthesis of a polyketide derivative, which depends on the value of the number of elongation steps *m*, the degree of reduction *n* and the amount of dehydration *h*. In

order to analyze the effect of these variables on the theoretical yield, it was first necessary to formulate a series of constraints for these variables based on the polyketide synthesis reaction mechanism. For example, since the maximum number of $NADPH_2^+$ molecules than can be used in each elongation step is two, the α-carbon in the linear chain after the last elongation step is not reduced, and one of the β-carbons throughout the synthesis needs to be reduced only once in order to participate in the cyclization reaction. Therefore, the constraint $1 \leq n \leq 2m\text{-}1$ was derived. In addition, only one water molecule can be released during an elongation step, the water molecule can only be released if the first reduction step has occurred and one of the β-carbons cannot be dehydrated in order to participate in the cyclization reaction; therefore, the amount of dehydration is constrained by $max(n\text{-}m,0) \leq h \leq min(m\text{-}1,n)$.

Metabolic flux analysis was then used to determine the maximum theoretical yield of polyketide structures synthesized with six elongation steps but with different degrees of reduction n. As shown in Figure 6.11, an increase in the amount of $NADPH_2^+$ required leads to a decrease in the theoretical yield of the polyketide structure; an increase in the degree of reduction of the polyketide leads to an increased use of the TCA cycle, thereby releasing higher amounts of carbon dioxide and thus reducing the yield on carbon through the system. The same result is obtained for the range of feasible values of h.

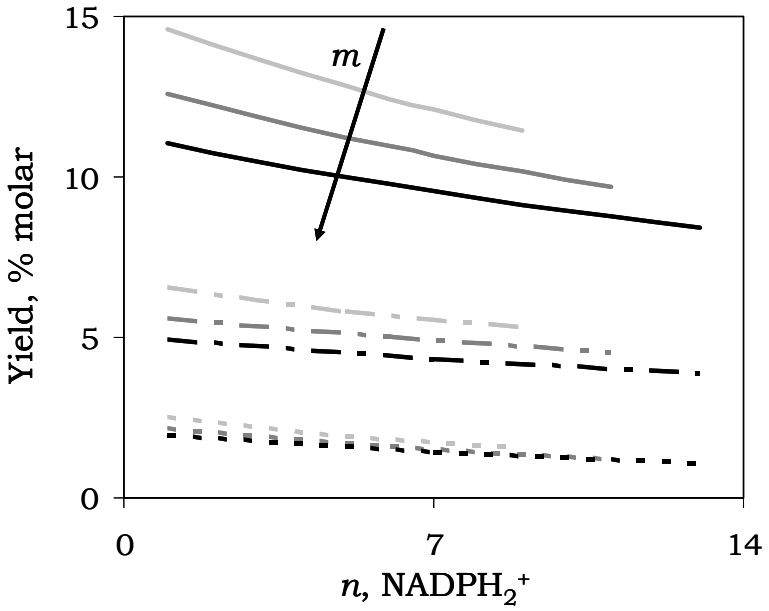


Figure 6.11. Percent molar yield for a polyketide derivative as a function of the degree of reduction or amount of $NADPH_2^+$, *n*, required for its synthesis. Data shown in the same hue corresponds to the same number of elongation steps *m*. Solid lines (—) were obtained with negligible ATP maintenance energy and an unconstrained oxygen uptake rate; dashed-dotted lines (– · –) were determined with an unconstrained oxygen uptake rate and an ATP maintenance energy of 5 mmol ATP g[dry weight]$^{-1}$ hr^{-1}; dotted lines (· · ·) were calculated with an ATP maintenance energy of 5 mmol ATP g[dry weight]$^{-1}$ hr^{-1} and a maximum oxygen uptake rate of 2 mmol O_2 g[dry weight]$^{-1}$ hr^{-1}.

The same analysis was performed to determine the effect of the number of elongation steps on the maximum theoretical yield of a polyketide. Therefore, as depicted in Figure 6.11, a structure synthesized with a higher number of elongation steps has a lower theoretical yield; this decrease is due to the fact that polyketide structures synthesized with more elongation steps are comprised of a higher amount of carbon atoms, thereby decreasing the yield of carbon. The amount of dehydration does not affect the maximum theoretical yield. However, as shown in the figure, the decreasing trend shown for different degrees of reduction is also observed for polyketides synthesized with different number of elongation steps *m*.

It was shown for erythromycin that the ATP maintenance energy and the aeration conditions of the fermentation affect the yield [52]. Therefore, metabolic flux analysis was used to analyze the effect of these cellular conditions on the maximum theoretical yield of other polyketide structures. As illustrated in Figure 6.11, a decrease of the maximum theoretical yield as a function of increasing degree of reduction *n* and increasing number of elongation steps *m* is also observed for an increase in the ATP maintenance energy requirement as well as for a decrease in the oxygen uptake rate.

6.3 Conclusions

Experimental approaches have obtained a maximum molar yield of 1.4% for the production of the erythromycin precursor, 6dEB, in *E. coli* using propionate as the carbon source, which corresponds to the use of less than 10% of the input propionate. Metabolic flux analysis was used with a mathematical model of the complete set of reactions, including the synthesis of 6dEB, in order to determine the maximum theoretical yield of 6dEB in *E. coli* and identify potential sources of carbon loss that result in the low yield observed experimentally. The analysis suggests that ATP maintenance requirements and specific oxygen uptake rates limit the production yield of 6dEB assuming that the uptake of propionate does not require energy. In order to meet the 6dEB synthesis requirements of ATP and $NADPH_2^+$, it is necessary to utilize the TCA cycle, thereby releasing carbon dioxide and thus reducing the yield of carbon in 6dEB production. These observations and shadow price analysis suggest that improvements in the yield of 6dEB from propionate could probably be achieved through the design of an efficient aerobic fermentation process.

The use of glucose as a carbon source might prove more effective in converting carbon to 6dEB under energy limiting conditions. The improvement of a process using glucose as a carbon source will require the consideration of the same genetic and process strategies as in the case of propionate: a decrease in the maintenance energy requirements, an increase in the specific glucose uptake yield, and an increase in the specific oxygen uptake rates would result in a higher yield of 6dEB. From an economic standpoint, although the yield of 6dEB from glucose is not significantly better than the yield from propionate, the use of a cheaper carbon source could improve the overall synthesis process of 6dEB in *E. coli*.

The analysis shows that there is a complex interplay between cellular energetics, oxygen uptake rates and production yield of 6dEB and polyketides in general. The studies suggest that further optimization and physiological experiments can more clearly elucidate this interplay, which could be quantified through additional metabolic flux analysis.

Chapter 7

Conclusions and Future Recommendations

Polyketides are cellular metabolites with a wide array of biological properties and commercial applications. However, although there is a large structural variation in the polyketide class, the carbon backbone of all the structures is synthesized using the same set of enzymatic biotransformations, the choice and order of which is dictated by the PKS. However, experimental approaches have shown that the PKS module organization can be altered in order to synthesize a different polyketide structure. A theoretical analysis of polyketide synthesis was therefore performed in order to identify the effect of the different variables in polyketide synthesis. This analysis suggests that over a billion possible structures can theoretically be synthesized. Consequently, a computational framework was developed for the identification of polyketide libraries and the assessment of their synthesis pathways.

The BNICE framework was created for the analysis of cellular reaction networks and it therefore provided a suitable tool for the study of polyketide synthesis. However, in order to more accurately represent enzymatic biotransformations, the framework was expanded to include stereochemical information and therefore distinguish between stereochemical isomers and effect stereochemical reactions. This expansion of the framework allowed the accurate implementation

of polyketide synthesis thereby leading to the identification of a complete *in silico* library of polyketide structures and their corresponding synthesis pathways.

7.1 Assessment of Synthetic Feasibility of Polyketide Derivatives

The introduction of a novel pathway in an organism generates competition within the existing cellular metabolism; therefore, it is necessary to assess the feasibility for the synthesis of each of the polyketide structures identified by the framework. Consequently, metabolic flux analysis was used on a genome-scale model of *Escherichia coli* to assess the effect of cellular constraints on polyketide synthesis, resulting in a complex interaction between cellular energetics, oxygen uptake rates and synthesis yield. Therefore, a lower production yield would be expected for those polyketide structures that achieve a higher degree of reduction during their synthesis suggesting that in the creation of the experimental erythromycin library, those mutations that involved the addition of the KR/DH/ER domains and did not lead to a derivative might actually have produced a structure at a level that was too low to detect.

However, cellular constraints are not sufficient in explaining all of the missing structures in the experimentally generated erythromycin library. Therefore, further analyses are necessary to better assess the synthetic feasibility of erythromycin derivatives. First, it is possible that the syntheses of some of these undiscovered structures are not thermodynamically favorable. Therefore, a thermodynamic analysis using group contribution theory would provide a better estimate of the thermodynamic feasibility for the synthesis of the derivatives. A more complete analysis using both thermodynamic and cellular constrains can also be performed in order to more accurately assess the synthesis pathways for each of the polyketide structures identified in the *in silico* library.

Second, in addition to cellular and thermodynamic constraints, mutations in a module of the PKS might affect the binding of the structure in subsequent modules. Therefore, an estimate of the binding between the structure produced after a mutation and the following domains in the PKS would prove useful in assessing whether a structure will be synthesized. However, in addition to computational time constraints, the structures of the domains in the elongation modules of the PKS are not known. On the other hand, the structure of the thioesterase (TE) domain has been elucidated (1kez, 1mo2) [53]. The geometry of the domain is very specific for the erythromycin linear chain with a number of amino acids interacting with the chain in order for the chain to achieve the correct conformation for the cyclization reaction. Consequently, the interaction between the amino acids in the substrate channel and the linear chain direct the cyclization, and it is therefore possible that a different linear chain would not interact correctly with the TE domain, potentially leading to missing structures in the experimentally generated erythromycin library.

Protein docking analyses between different linear chains and the erythromycin wild-type TE domain would prove useful in determining the effect of the TE domain in the cyclization of different polyketide structures. Additionally, for those structures that can not undergo cyclization with the wild-type TE domain, mutations to the amino acid sequence of the TE active site can be performed in order to determine what residues should be modified in order to engineer a TE domain that would allow the cyclization of these erythromycin derivatives.

Experimentalists have also inserted the TE domain after modules 2, 3 and 5 in the erythromycin PKS in order to synthesize shorter polyketides. The cyclization of the erythromycin linear chain involves a lactonization reaction between the α-carbon and the farthest

hydroxyl group in the linear chain. The same pattern is observed if the TE domain is inserted after modules 2 and 5 with the lactonization reaction involving the farthest hydroxyl group. However, placement of the TE domain after module 3 leads to a lactonization reaction bewteen the α-carbon and the second farthest hydroxyl group in the linear chain. Preliminary quantum chemical analyses have shown that the cyclizations observed with placement of the TE domain after modules 2 and 3 are spontaneous. Therefore, it can be hypothesized that these chains are too small to interact specifically with the TE substrate channel, leading to spontaneous cyclizations. This hypothesis should be tested through protein docking studies between the shorter linear chains and the erythromycin wild-type TE substrate channel. It would be interesting to determine if a TE domain can be engineered to obtain a different cyclization pattern either by modifications to the amino acids involved in the cyclization reaction or by changes to the geometry of the substrate channel to obtain a different conformation and thus a different cyclization.

7.2 Analysis of Polyketide Activity and Toxicity

A large amount of the antibacterial activity of polyketides is conferred through the tailoring steps that the cyclical structure undergoes. Therefore, in addition to the generalized enzyme functions for the synthesis of the polyketide carbon backbone, a set of reaction rules for the tailoring steps should be implemented in BNICE in order to identify all the final polyketide structures.

Based on previous analyses, it is expected that the framework will identify a large number of polyketide structures. However, not all of these structures will prove to possess antibiotic properties; consequently, docking approaches can be used to identify those structures that can

potentially become antibiotics. A number of different target proteins, as well as RNA and ribosomes, can be used for screening this large set of polyketides. In fact, a research initiative has predicted binding between erythromycin and the ribosome in approximately 30% of the trial runs [54]. Similar studies can be performed with structures identified by the BNICE framework in order to identify those structures that have a higher antibiotic potential.

In addition to estimating potential antibiotic structures, one of the most time-consuming issues related to the discovery of a new drug involves its safety evaluation. Therefore, once a list of potential polyketide structures has been identified, their toxicity can be evaluated. Previous research has shown that docking techniques can be used to identify undesirable protein targets with which a potential drug can interact; this has successfully been shown for drugs such as penicillin and aspirin, among others [55-57]. This approach can be used on the list of polyketides with potential antibiotic properties in order to eliminate those that can potentially have high-risk side effects.

References

1. Staunton, J. and B. Wilkinson. (1998) "The biosynthesis of aliphatic polyketides." Biosynthesis **195**: 49-92.

2. Shen, B. (2000) "Biosynthesis of aromatic polyketides." Biosynthesis: Aromatic Polyketides, Isoprenoids, Alkaloids **209**: 1-51.

3. Khosla, C. (1997) "Harnessing the biosynthetic potential of modular polyketide synthases." Chem Rev **97**(7): 2577-2590.

4. Hopwood, D.A. (1997) "Genetic contributions to understanding polyketide synthases." Chem Rev **97**(7): 2465-2498.

5. Katz, L. (1997) "Manipulation of modular polyketide synthases." Chem Rev **97**(7): 2557-2576.

6. McDaniel, R., S. Ebert-Khosla, D.A. Hopwood, and C. Khosla. (1993) "Engineered biosynthesis of novel polyketides." Science **262**(5139): 1546-50.

7. Rowe, C.J., I.U. Bohm, I.P. Thomas, B. Wilkinson, B.A.M. Rudd, G. Foster, A.P. Blackaby, P.J. Sidebottom, Y. Roddis, A.D. Buss, J. Staunton, and P.F. Leadlay. (2001) "Engineering a polyketide with a longer chain by insertion of an extra module into the erythromycin-producing polyketide synthase." Chem Biol **8**(5): 475-485.

8. McDaniel, R., A. Thamchaipenet, C. Gustafsson, H. Fu, M. Betlach, M. Betlach, and G. Ashley. (1999) "Multiple genetic modifications of the erythromycin polyketide synthase to produce a library of novel "unnatural" natural products." Proc Nat Acad Sci USA **96**(10): 5890-5890.

9. Xue, Q., G. Ashley, C.R. Hutchinson, and D.V. Santi. (1999) "A multiplasmid approach to preparing large libraries of polyketides." Proc Natl Acad Sci USA **96**(21): 11740-5.

10. Cane, D.E., C.T. Walsh, and C. Khosla. (1998) "Harnessing the biosynthetic code: combinations, permutations, and mutations." Science **282**(5386): 63-8.

11. Cane, D.E. (1997) "Introduction: polyketide and nonribosomal polypeptide biosynthesis. From collie to *coli*." Chem Rev **97**(7): 2463-2464.

12. Moore, B.S. and J. Piel. (2000) "Engineering biodiversity with type II polyketide synthase genes." Antonie Van Leeuwenhoek **78**(3-4): 391-8.

13. Dreier, J. and C. Khosla. (2000) "Mechanistic analysis of a type II polyketide synthase. Role of conserved residues in the beta-ketoacyl synthase-chain length factor heterodimer." Biochemistry **39**(8): 2088-95.

14. Carreras, C.W., R. Pieper, and C. Khosla. (1996) "Efficient synthesis of aromatic polyketides *in vitro* by the actinorhodin polyketide synthase." J Am Chem Soc **118**(21): 5158-5159.

15. Broadbelt, L.J., S.M. Stark, and M.T. Klein. (1996) "Computer generated reaction modelling: decomposition and encoding algorithms for determining species uniqueness." Comput Chem Eng **20**(2): 113-129.

16. Broadbelt, L.J., S.M. Stark, and M.T. Klein. (1994) "Computer generated reaction networks: on-the-fly calculation of species properties using computational quantum chemistry." Chem Eng Sci **49**: 4991-5010.

17. Broadbelt, L.J., S.M. Stark, and M.T. Klein. (1994) "Computer-generated pyrolysis modeling - on-the-fly generation of species, reactions, and rates." Ind Eng Chem Res **33**(4): 790-799.

18. Broadbelt, L.J., S.M. Stark, and M.T. Klein. (1995) "Termination of computer-generated reaction-mechanisms - species rank-based convergence criterion." Ind Eng Chem Res **34**(8): 2566-2573.

19. Ugi, I., J. Bauer, J. Brandt, J. Friedrich, J. Gasteiger, C. Jochum, and W. Schubert. (1979) "New applications of computers in chemistry." Angew Chem Int Ed **18**(2): 111-123.

20. Hatzimanikatis, V., C. Li, J.A. Ionita, C.S. Henry, M.D. Jankowski, and L.J. Broadbelt. (2005) "Exploring the diversity of complex metabolic networks." Bioinformatics **21**(8): 1603-9.

21. Tipton, K. and S. Boyce. (2000) "History of the enzyme nomenclature system." Bioinformatics **16**(1): 34-40.

22. Hatzimanikatis, V., C. Li, J.A. Ionita, and L.J. Broadbelt. (2004) "Metabolic networks: enzyme function and metabolite structure." Curr Opin Struct Biol **14**(3): 300-6.

23. Bentley, R. and J.W. Bennett. (1999) "Constructing polyketides: from collie to combinatorial biosynthesis." Annu Rev Microbiol **53**: 411-46.

24. Dayem, L.C., J.R. Carney, D.V. Santi, B.A. Pfeifer, C. Khosla, and J.T. Kealey. (2002) "Metabolic engineering of a methylmalonyl-CoA mutase-epimerase pathway for complex polyketide biosynthesis in *Escherichia coli*." Biochemistry **41**(16): 5193-201.

25. Staunton, J. and K.J. Weissman. (2001) "Polyketide biosynthesis: a millennium review." Nat Prod Rep **18**(4): 380-416.

26. Shen, B. (2003) "Polyketide biosynthesis beyond the type I, II and III polyketide synthase paradigms." Curr Opin Chem Biol **7**(2): 285-95.

27. Caffrey, P. (2003) "Conserved amino acid residues correlating with ketoreductase stereospecificity in modular polyketide synthases." ChemBioChem **4**(7): 654-7.

28. Khosla, C. and R.J.X. Zawada. (1996) "Generation of polyketide libraries via combinatorial biosynthesis." Trends Biotechnol **14**(9): 335-341.

29. González-Lergier, J., L.J. Broadbelt, and V. Hatzimanikatis. (2005) "Theoretical considerations and computational analysis of the complexity in polyketide synthesis pathways." J Am Chem Soc **127**(27): 9930-9938.

30. Li, C., J.A. Ionita, C.S. Henry, M.D. Jankowski, V. Hatzimanikatis, and L.J. Broadbelt. (2004) "Computational discovery of biochemical routes to specialty chemicals." Chem Eng Sci.

31. Cahn, R.S., C.K. Ingold, and V. Prelog. (1956) "The specification of asymmetric configuration in organic chemistry." Experientia **12**(3): 81-94.

32. Cahn, R.S., C. Ingold, and V. Prelog. (1966) "Specification of molecular chirality." Angew Chem Int Ed **5**(4): 385-&.

33. Weininger, D. (1988) "Smiles, a chemical language and information system.1. Introduction to methodology and encoding rules." J Chem Inf Comput Sci **28**(1): 31-36.

34. Adam, D. (2002) "Chemists synthesize a single naming system." Nature **417**(6887): 369-369.

35. McMillan, S.A., N.C. Haubein, R.Q. Snurr, and L.J. Broadbelt. (2003) "*Ab initio* stochastic optimization of conformational and many-body degrees of freedom." J Chem Inf Comput Sci **43**(6): 1820-1828.

36. Cieplak, T. and J.L. Wisniewski. (2001) "A new effective algorithm for the unambiguous identification of the stereochemical characteristics of compounds during their registration in databases." Molecules **6**(11): 915-926.

37. Akutsu, T. (1991) "A new method of computer representation of stereochemistry - transforming a stereochemical structure into a graph." J Chem Inf Comput Sci **31**(3): 414-417.

38. Contreras, M.L., R. Rozas, R. Valdivia, and R. Aguero. (1995) "Exhaustive generation of organic rsomers.4. Acyclic stereoisomers with one or more chiral carbon-atoms." J Chem Inf Comput Sc **35**(4): 752-758.

39. Contreras, M.L., G.M. Trevisiol, J. Alvarez, G. Arias, and R. Rozas. (1999) "Exhaustive generation of organic isomers.5. Unsaturated optical and geometrical stereoisomers and a new CIP subrule." J Chem Inf Comput Sci **39**(3): 475-482.

40. Mavrovouniotis, M.L., G. Stephanopoulos, and G. Stephanopoulos. (1990) "Computer-aided synthesis of biochemical pathways." Biotechnol Bioeng **36**(11): 1119-1132.

41. Seressiotis, A. and J.E. Bailey. (1986) "MPS - an algorithm and database for metabolic pathway synthesis." Biotechnol Lett **8**(12): 837-842.

42. Klopman, G., M. Dimayuga, and J. Talafous. (1994) "Meta.1. A program for the evaluation of metabolic transformation of chemicals." J Chem Inf Comp Sci **34**(6): 1320-1325.

43. Wong, H.W., X.G. Li, M.T. Swihart, and L.J. Broadbelt. (2004) "Detailed kinetic modeling of silicon nanoparticle formation chemistry via automated mechanism generation." J Phys Chem A **108**(46): 10122-10132.

44. Pfeifer, B.A. and C. Khosla. (2001) "Biosynthesis of polyketides in heterologous hosts." Microbiol Mol Biol Rev **65**(1): 106-18.

45. Cameron, D.C. and I.T. Tong. (1993) "Cellular and metabolic engineering - an overview." App Biochem Biotechnol **38**(1-2): 105-140.

46. Pfeifer, B.A., S.J. Admiraal, H. Gramajo, D.E. Cane, and C. Khosla. (2001) "Biosynthesis of complex polyketides in a metabolically engineered strain of *E. coli*." Science **291**(5509): 1790-2.

47. Varma, A. and B.O. Palsson. (1993) "Metabolic capabilities of *Escherichia coli*.1. Synthesis of biosynthetic precursors and cofactors." J Theor Biol **165**(4): 477-502.

48. Varma, A. and B.O. Palsson. (1993) "Metabolic capabilities of *Escherichia coli*.2. Optimal-growth patterns." J Theor Biol **165**(4): 503-522.

49. Price, N.D., J.A. Papin, C.H. Schilling, and B.O. Palsson. (2003) "Genome-scale microbial *in silico* models: the constraints-based approach." Trends Biotechnol **21**(4): 162-169.

50. Reed, J.L., T.D. Vo, C.H. Schilling, and B.O. Palsson. (2003) "An expanded genome-scale model of *Escherichia coli* K-12." Genome Biol **4**(9): R54.

51. Varma, A., B.W. Boesch, and B.O. Palsson. (1993) "Stoichiometric interpretation of *Escherichia coli* glucose catabolism under various oxygenation rates." Appl Environ Microbiol **59**(8): 2465-2473.

52. Gonzalez- Lergier, J., L.J. Broadbelt, and V. Hatzimanikatis. (2006) "Analysis of the maximum theoretical yield for the synthesis of erythromycin precursors in *Escherichia coli*." Biotechnol Bioeng **95**(4): 638-644.

53. Tsai, S.C., H.X. Lu, D.E. Cane, C. Khosla, and R.M. Stroud. (2002) "Insights into channel architecture and substrate specificity from crystal structures of two macrocycle-forming thioesterases of modular polyketide synthases." Biochemistry **41**(42): 12598-12606.

54. Detering, C. and G. Varani. (2004) "Validation of automated docking programs for docking and database screening against RNA drug targets." J Med Chem **47**(17): 4188-4201.

55. Chen, Y.Z. and C.Y. Ung. (2001) "Prediction of potential toxicity and side effect protein targets of a small molecule by a ligand-protein inverse docking approach." J Mol Graphics Modell **20**(3): 199-218.

56. Chen, X., C.Y. Ung, and Y.Z. Chen. (2003) "Can an *in silico* drug-target search method be used to probe potential mechanisms of medicinal plant ingredients?" Nat Prod Rep **20**(4): 432-444.

57. Chen, Y.Z. and C.Y. Ung. (2002) "Computer automated prediction of potential therapeutic and toxicity protein targets of bioactive compounds from Chinese medicinal plants." Am J Chinese Med **30**(1): 139-154.

Appendix A. Cellular Metabolites

The metabolites that are involved in the metabolism of *E. coli* used in 6dEB synthesis are summarized, with their abbreviations, in Table A.1.

Table A.1. Metabolites in the *E. coli* synthesis of the erythromycin precursor from propionate.

Abbreviation	Full Name
6dEB	6-deoxyerthronolide
adp	Adenosine Diphosphate
atp	Adenosine Triphosphate
co2	Carbon Dioxide
coa	Coenzyme A
h2o	Water
mmcoa-S	(S)-Methylmalonyl-CoA
nadp	Nicotinamide Adenine Dinucleotide Phosphate (Oxidized)
nadph	Nicotinamide Adenine Dinucleotide Phosphate (Reduced)
p_i	Phosphate
ppa	Propionate
ppcoa	Propionyl-CoA

The enzymes involved in the reactions for the synthesis of 6dEB from propionate are listed, with their abbreviations, in Table A.2

Table A.2. Enzymes in the *E. coli* synthesis of the erythromycin precursor from propionate.

Abbreviation	Full Name
ACCOAL	Acetate-CoA ligase (ADP-forming)
MMCD	Methylmalonyl-CoA Carboxylase
Ery PKS	Erythromycin Polyketide Synthase

The metabolites not included in the direct synthesis of 6dEB that are part of the reaction network required to synthesize the maximum theoretical yield of the 6dEB from propionate in *E. coli* are listed in Table A.3.

Table A.3. Metabolites in the *E. coli* reaction network for the synthesis of the erythromycin precursor from propionate.

Abbreviation	Full Name
2mcacn	Cis-2-Methylaconitate
2mcit	2-Methylcitrate
aacoa	Acetyl-CoA
akg	2-Oxoglutarate
cit	Citrate
fad	Flavin Adenine Dinuclotide (Oxidized);
fadh2	Flavin Adenine Dinuclotide (Reduced)
fum	Fumarate
icit	Isocitrate
mal-L	L-Malate
micit	Methylisocitrate
nad	Nicotinamide Adenine Dinucleotide (Oxidized)
nadh	Nicotinamide Adenine Dinucleotide (Reduced)
oaa	Oxaloacetate
o2	Oxygen
pyr	Pyruvate
q8:	Ubiquinone-8
q8h2:	Ubiquinol-8
oaa:	Oxaloacetate
succ	Succinate
succoa	Succinyl-CoA

The additional metabolites that are part of the reaction network required to synthesize the maximum theoretical yield of the 6dEB from glucose in *E. coli* are listed in Table A.4.

Table A.4. Metabolites in the *E. coli* reaction network for the synthesis of the erythromycin precursor from glucose.

Abbreviation	Full Name
2pg	D-glycerate 2-Phosphate
3pg	3-Phospho-D-Glycerate
6pgc	6-Phospho-D-Gluconate
6pgl	6-Phospho-D-Glucono-1,5-Lactone
13dpg	3-Phospho-D-Glyceroyl Phosphate
dha	Dihydroxyacetone
dhap	Dihydroxyacetone Phosphate
e4p	D-Erythrose 4-Phosphate
f6p	D-Fructose 6-Phosphate
fdp	D-Fructose 1,6-Bisphosphate
for	Formate
g3p	Glyceraldehyde 3-Phosphate
g6p	D-Glucose 6-Phosphate
glc-D	D-Glucose
pep	Phosphoenolpyruvate
r5p	Alpha-D-Ribose 5-Phosphate
ru5p-D	D-ribulose 5-Phosphate
s7p	Sedoheptulose 7-Phosphate

Appendix B. Detailed Description of the Algorithm

The BNICE framework was developed to explore all the possible reactions that can occur in a cellular organism. It utilizes the set of specified reaction rules to identify all the possible functional groups in a given set of input structures, and applies the specified reaction operators to obtain the set of all possible products that can be synthesized from the given reactants. This methodology is then applied to the previously identified set of products, and the process is repeated until no new species are formed or a user-specified termination criterion is met.

Analysis of Input Substrates

The structure(s) of the input substrate(s) for the framework are user-specified. To expand the BNICE framework to include stereochemical information, two types of input structures were accommodated: a specific isomer(s) of a substrate or structural information without stereochemical information. In the latter case, the framework was designed to use all isomers of the input substrate(s). Therefore, the BNICE framework was modified such that each input substrate is first analyzed for isomerism such that if no or incomplete stereochemical information is provided, the framework will utilize all possible isomers of the input substrate(s) specified.

In order to identify all the sources of isomerism of the input substrate(s), and therefore generate all possible isomers of the substrate(s), chiral centers as well as atoms that give rise to geometric isomerism are detected. First, all the atoms in the substrate, excluding those in a ring, that have four single bonds are marked as possible chiral centers. Chirality of an atom is defined as the atom having four different substituents. A hierarchical approach based on the CIP priority rules for determining the configuration of stereogenic carbons was implemented for the identification of chiral centers and the assignment of priorities to each of the four substituents of

a chiral atom. Using this approach, the atomic number of an immediately substituent atom determines its priority, where a higher atomic number results in a higher priority. Therefore, the substituent atom with the highest atomic number is assigned a 1, that with the second highest atomic number is assigned a 2, and so forth. If two or more of the immediately substituent atoms are equal, the atomic numbers of the atoms connected to these are used to determine the priority; for example, if two immediately substituent atoms to a chiral atom are carbon, an ethyl group would have a higher priority when compared to a methyl group due to a higher atomic mass. In addition, if the immediately substituent atom is connected to another atom through a double or triple bond, the bonds are treated as an equivalent set of single-bonded atoms; for instance, if one of the substituent groups is a carbonyl group, it is considered as a carbon with two single carbon-oxygen bonds instead of a double carbon-oxygen bond resulting in two oxygen atoms used in the evaluation of the substituent group instead of one. This approach is used iteratively until either a difference in the substituents is found and all of the substituents are assigned a priority or all the atoms in the substituent groups have been analyzed and no difference is found. If the atom is found to be chiral, and its chirality information is not user-specified, the framework labels the atom as chiral and generates the two corresponding stereoisomers. The first isomer is created by assigning the chirality label R to the chiral center and the second is assigned a chirality S. This process is continued until all possible chiral centers are analyzed and all 2^c possible chiral isomers of the input substrate(s) are identified, where c is the number of chiral atoms. Once the isomers are created, the original structure is erased from the set of input structures, and the isomers are added to the set. The presence of meso compounds in this set will be investigated later.

Once chiral isomerism has been analyzed, the possibility of geometric isomerism due to the presence of double bonds is studied. The two substituent groups attached to each of the atoms that form part of a double bond that is not part of a ring are analyzed; if each atom has two different substituents, then the geometric isomers of the structure are created: (1) two new structures are created; (2) the two substituents are assigned a priority according to the CIP priority rules and (3) the two atoms are labeled "Z" in the first isomer and "E" in the second.

Finally, geometric isomerism due to the presence of a ring is studied. All the non-chiral atoms that are part of a ring and have four single bonds are studied in order to identify those that are bonded to two different substituent groups, excluding the two bonds within the ring; if there is more than one ring, an analysis is done for each ring separately. In order for the structure to exhibit isomerism in the ring, the total number of atoms that contribute to geometric isomerism must be more than one. If more than one atom is found, these are marked and the priorities of the two substituents outside the ring for each of the atoms are determined using the CIP rules. The directionality of the two substituents outside the ring of the first marked atom are then determined by assigning values to the $\{i,j,2\}$ matrix elements; therefore, the $\{i,j,2\}$ element of the higher priority substituent of the first marked atom is assigned a value of +1 and the $\{i,j,2\}$ element of the lower priority substituent of the first marked atom is given a value of -1. The isomers are then created: (1) two new structures are created; (2) the $\{i,j,2\}$ elements of the substituents with priority 1 and 2 of the second marked atom in the first isomer are assigned values of +1 and -1, respectively; (3) the $\{i,j,2\}$ matrix elements of the substituents with priority 1 and 2 of the second marked atom in the second isomer are given values of -1 and +1, respectively. If there are additional marked atoms, then the $\{i,j,2\}$ elements of the substituents

with priority 1 and 2 of the remaining marked atoms are assigned values of +1 and -1, respectively. Additional isomers are then created if there is more than one additional marked atom: (1) two new structures are created equal to the second isomer from the previous iteration; (2) the third and fourth marked atoms in the first new structure are assigned values for the {*i*,*j*,*2*} matrix elements following the same pattern as the first two marked atoms; and (3) the process is repeated if there is more than one additional marked atom.

In order to identify all the possible isomers using the stereochemical string code, it is also necessary to assign {*i*,*j*,*2*} matrix elements to the chiral atoms in the ring: (1) the priorities are manipulated such that the substituents in the ring have priorities 1 and 4 where the substituent with priority 1 is the substituent in the ring with the higher priority; (2) the substituents outside the ring are similarly assigned priorities 2 and 3; (3) the values for the {*i*,*j*,*2*} matrix elements are then assigned in the same manner as was done for the atoms that give rise to geometric isomerism in the ring. Figure B.1 illustrates the generation of isomers when a structure has both chiral atoms in a ring and atoms that give rise to geometric isomerism in a ring. Note that meso compounds are not eliminated prior to identifying geometrical isomerism in a ring in order to be able to create all the possible geometric isomers.

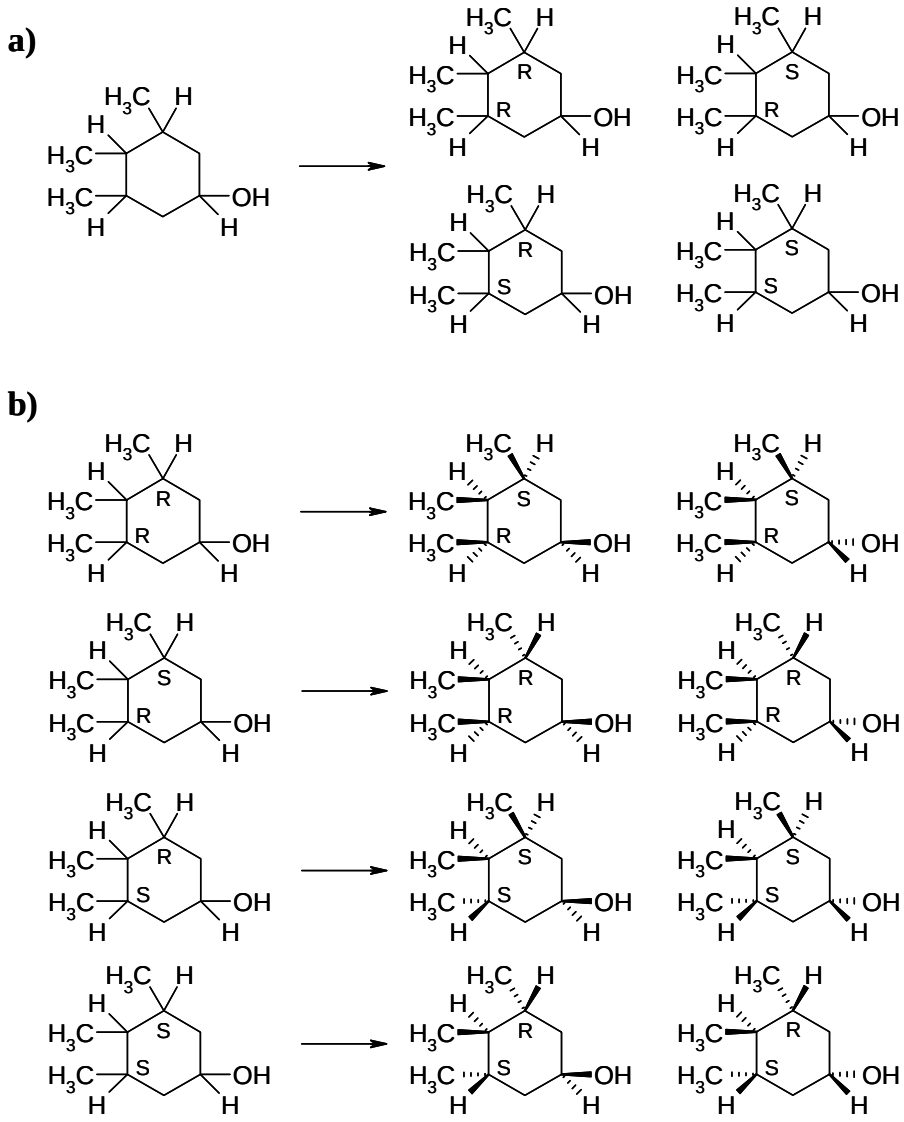


Figure B.1. Generation of isomers from input structure. (a) Identification of chiral atoms in a ring and generation of corresponding isomers for the chiral carbons. This includes the identification of a meso compounds. (b) Identification of atoms that give rise to geometric isomerism in a ring and generation of the corresponding isomers. Note that the meso compounds identified in (a) are both used in the identification of geometric isomers.

Verification of Substrate Uniqueness

If there is more than one input substrate, and due to the possibility of meso compounds when isomers have been generated, it is necessary to verify that all the structures are indeed unique.

This is accomplished by generating the stereochemical string code for each species, which is unique for each stereoisomer. The generation of the stereochemical string code is performed in the same way as the molecule's string code. Therefore, the root of the canonical decomposition tree is identified and its substituents are then added sequentially according to a lexicographical ordering of the string of each of the substituents. Therefore, for acyclic molecules, if the root of the canonical decomposition tree is an atom, the string code will be generated by comparing the strings of all of its descendents and ordering them alphabetically; similarly, if the root is a bond, the two substituents of the bond are ordered in the same way. Therefore, meso compounds will have the same string code and one of the isomers will then be eliminated from the species list. A similar approach is used in selecting the order of the atoms involved in a ring; consequently, it was necessary to add some orientation information to the bonds in order to differentiate between isomers that have both chiral cyclic atoms and atoms that exhibit geometrical isomerism in a ring.

Execution of Biochemical Reaction

The reaction rules used by the framework are defined by the user; the user therefore has the choice of specifying the stereochemistry of the functional groups of the substrate and/or any change in stereochemistry that can occur as a result of the reaction. If no stereochemistry is specified, the framework identifies all the possible functional groups in the set of substrates and executes the reaction via matrix addition. The products of the reaction are then analyzed in order to generate all possible isomers of the product found by the reaction using the same algorithm that was used in the analysis of the input substrates.

However, the active site of the enzyme often has a specific geometry that induces specific stereochemistry in the product(s) or operates only on a particular isomer of the substrate(s). Thus, the stereochemistry of the substrates and products that utilize a stereochemistry-specific enzyme is predetermined due to the geometry of the enzyme active site. For example, in a reduction reaction, as shown in Figure B.2, the chirality obtained by a product can be constrained such that the hydroxyl group is pointed forward and the hydrogen backwards from the plane of the molecule. The reduction reaction rule can therefore be defined to produce a product with a chirality R provided that the CIP priority of the groups follow: OH > x_2 > x_1 and the CIP priority of H is 4. This definition of the priorities would also synthesize a product with chirality R if the CIP priorities in the product of the reaction follow: x_2 > x_1 > OH *or* x_1 > OH > x_2. However, if the priorities of the structure produced in the reaction does not follow the defined priorities (i.e. the CIP priorities are OH > x_1> x_2 or x_1 > x_2 > OH or x_2 > OH > x_1), then the chirality of the product to defined as S. As shown in Figure B.2a, the substrate possessed two reactive sites that satisfy the reaction rule; therefore, it can react twice leading to both isomers, R and S, being produced by the generalized reaction rule. However, as shown in the figure, if the functional group is constrained more specifically (i.e. the carboxylic acid must be x_2), the framework is only able to find one reactive site for the substrate, leading to the production of only one isomer. Similarly, if the functional group of the substrate has a predetermined stereochemistry, the framework verifies that the reactive sites in the substrate correlates to the specified stereochemistry before carrying out the reaction.

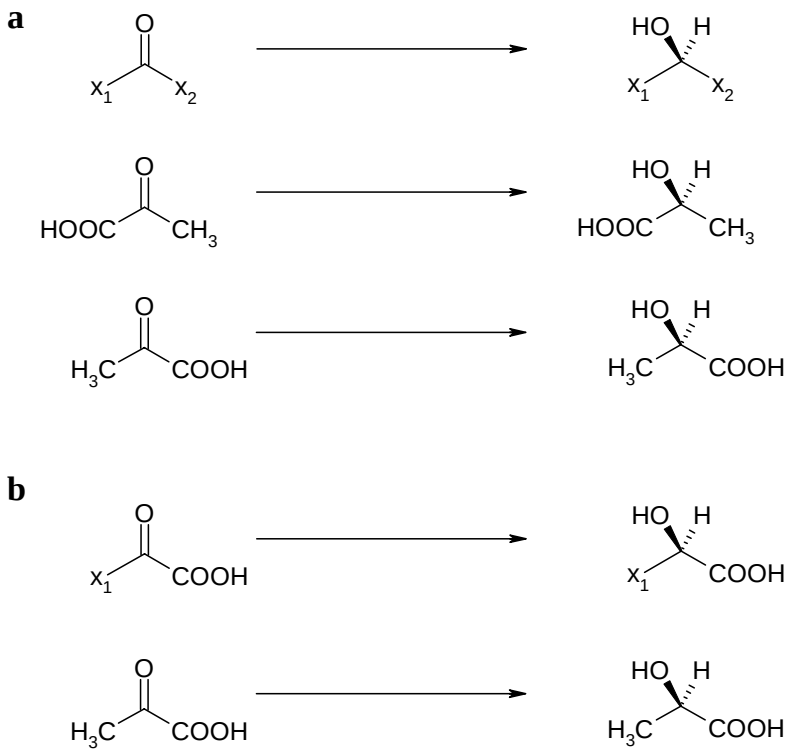


Figure B.2. Schematic of a stereochemical reaction in the BNICE framework. (a) The reduction reaction is capable of producing a chiral center which can be constrained to produce a final product as shown by the first reaction schematic. This produced a product with a chirality R provided that the CIP priority of OH > x_2 > x_1 and the CIP priority of H is 4. The structure shown satisfies the functional group in two ways since x_1 and x_2 can be either the methyl or the carboxylic acid group. However, if the priorities of the structure produced in the reaction are different than the priorities specified by the reaction rule, the chirality of the resulting structure changes. Therefore, both isomers, R and S, are produced by this reaction. (b) A more specific definition of the functional group leads to the substrate being able to participate in only one reaction, therefore producing only one isomer.

Identification of Product Uniqueness

Once a product 3-D BE-matrix is generated, it is necessary to determine if the molecule that it represents has been generated in a previous iteration of the framework. Therefore, the stereochemical string code is generated using the same method as was used in the analysis of the set of input substrates. This stereochemical string code is then compared to the other molecules identified by the framework in order to determine if the product is unique.