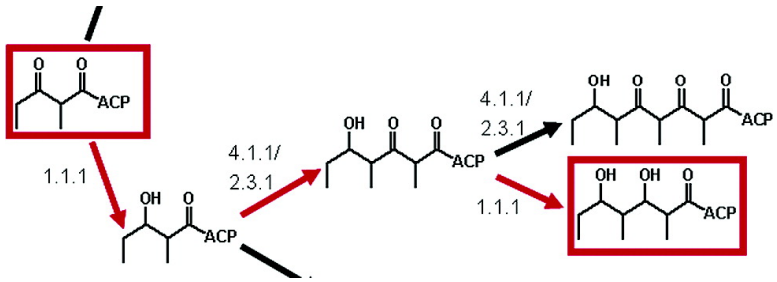
Article

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

Joanna Gonzlez-Lergier, Linda J. Broadbelt, and Vassily Hatzimanikatis

J. Am. Chem. Soc., **2005**, 127 (27), 9930-9938• DOI: 10.1021/ja051586y • Publication Date (Web): 16 June 2005



More About This Article

Additional resources and features associated with this article are available within the HTML version:

- Supporting Information
- Links to the 2 articles that cite this article, as of the time of this article download
- Access to high resolution figures
- Links to articles and content related to this article
- Copyright permission to reproduce figures and/or text from this article

View the Full Text HTML

J|A|C|S
ARTICLES
Published on Web 06/16/2005

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

Joanna González-Lergier, Linda J. Broadbelt,* and Vassily Hatzimanikatis*

Contribution from the Department of Chemical and Biological Engineering, Northwestern University, Evanston, Illinois 60208

Received March 11, 2005; E-mail: vassily@northwestern.edu

Abstract: The emergence of antimicrobial resistance has led to an increase in research directed toward the engineering of novel polyketides. To date, less than 10 000 polyketide structures have been discovered experimentally; however, the theoretical analysis of polyketide biosynthesis performed suggests that over a billion possible structures can be synthesized. Polyketide synthesis, which involves the formation of a linear chain and its subsequent cyclization, is catalyzed by an enzyme complex called polyketide synthase (PKS). There are a number of variables in the linear chain synthesis controlled by the PKS: the number, identity, stereochemistry and sequence of the monomer units used in the elongation steps, and the degree of reduction that occurs after each of the condensation reactions. The theoretical analysis performed demonstrates that changes in these variables lead to the formation of different polyketide linear chains and, consequently, a high diversity of polyketide structures. The complexity in the number of possible structures led to the implementation of this system in BNICE, a computational framework that generates all possible biochemical pathways using a given set of enzyme reaction rules. This formulation allowed the analysis of the evolution of diversity in the synthesis mechanism and the construction of the pathway architecture of polyketide biosynthesis. It is expected that, after future implementation of the cyclization reactions, this framework can be used to identify all possible polyketides and their corresponding synthesis pathways. Consequently, this formulation would prove useful in guiding experimental approaches to engineer novel polyketides, a number of which will likely have medicinal properties.

Introduction

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 1 shows a number of polyketides that, in addition to having large structural diversity, have a wide variety of pharmacological uses.[2−4] 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.[5] 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.[6]

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 direct 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.[6,7] 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.[6,8,9] After the full-length chain is synthesized, a number of cycliza-

(1) Staunton, J.; Wilkinson, B. *Top. Curr. Chem.* **1998**, *195*, 49−92.
(2) Bentley, R.; Bennett, J. W. *Annu. Rev. Microbiol.* **1999**, *53*, 411−46.
(3) Dayem, L. C.; Carney, J. R.; Santi, D. V.; Pfeifer, B. A.; Khosla, C.; Kealey, J. T. *Biochemistry* **2002**, *41* (16), 5193−201.
(4) Staunton, J.; Weissman, K. J. *Nat. Prod. Rep.* **2001**, *18* (4), 380−416.
(5) Cane, D. E.; Walsh, C. T.; Khosla, C. *Science* **1998**, *282* (5386), 63−8.
(6) Hopwood, D. A. *Chem. Rev.* **1997**, *97* (7), 2465−2498.
(7) McDaniel, R.; Thamchaipenet, A.; Gustafsson, C.; Fu, H.; Betlach, M.; Betlach, M.; Ashley, G. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96* (10), 5890−5890.
(8) Shen, B. *Biosynthesis: Aromatic Polyketides, Isoprenoids, Alkaloids* **2000**, *209*, 1−51.
(9) Katz, L. *Chem. Rev.* **1997**, *97* (7), 2557−2576.

10.1021/ja051586y CCC: $30.25

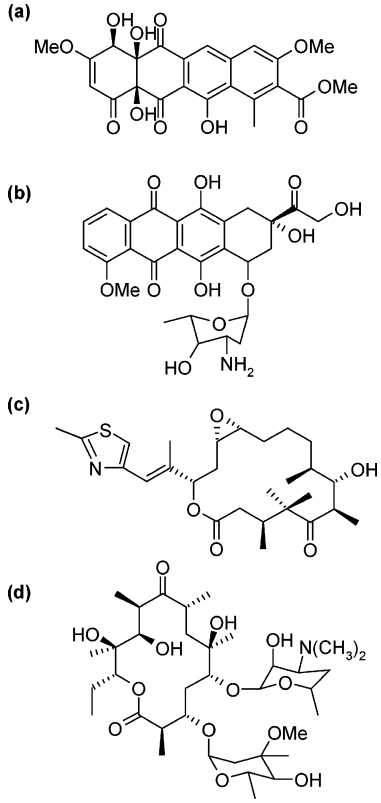


Figure 1. Polyketide structures. (a) Tetracenomycin C is an antibiotic; (b) doxorubibin is used as an antitumor agent; (c) epothilone A is an anticancer agent; and (d) erythromycin A is an antibiotic.

tion and posttranslational modifications synthesize the final polyketide molecule.[2,6,8−10]

In an elongation step, as illustrated in Figure 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.[11] 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 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 does occur in a chain extension step, it occurs in the order described.[1,6,9]

The structural diversity of polyketides is due in part to the variation in their linear carbon backbones. These differences in the linear chain 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.[5,6,9,12−14] Modification of these variables leads to the synthesis of different polyketide structures. Based on the complexity of this process, which arises primarily from the large number of structures that can be produced, a computational framework would prove useful in the automatic generation and analysis of polyketide synthesis.

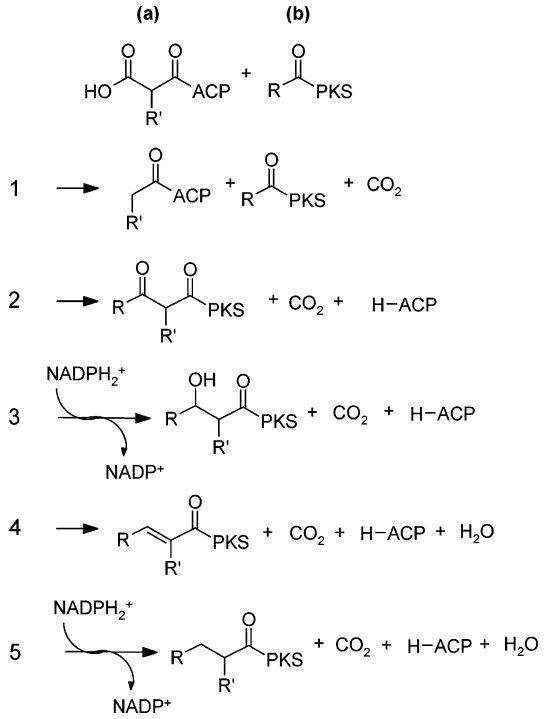


Figure 2. Schematic of an elongation step. The substrates for the elongation reaction are the extender unit attached to an acyl carrier protein, or malonyl-ACP (a) as shown in this example, and the growing linear chain (b) attached to the PKS complex. The extender unit undergoes a decarboxylation, producing 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, and application of this novel computational framework allows the de novo synthesis of metabolic pathways.[19]

A theoretical analysis of polyketide synthesis, involving the use of stoichiometric balances and combinatorial theory, was

(10) Shen, B. *Curr. Opin. Chem. Biol.* **2003**, *7* (2), 285−95.
(11) Dreier, J.; Khosla, C. *Biochemistry* **2000**, *39* (8), 2088−95.
(12) Khosla, C. *Chem. Rev.* **1997**, *97* (7), 2577−2590.
(13) Caffrey, P. *ChemBioChem* **2003**, *4* (7), 654−7.
(14) Khosla, C.; Zawada, R. J. X. *Trends Biotechnol.* **1996**, *14* (9), 335−341.
(15) Broadbelt, L. J.; Stark, S. M.; Klein, M. T. *Comput. Chem. Eng.* **1996**, *20* (2), 113−129.
(16) Broadbelt, L. J.; Stark, S. M.; Klein, M. T. *Chem. Eng. Sci.* **1994**, *49*, 4991−5010.
(17) Broadbelt, L. J.; Stark, S. M.; Klein, M. T. *Ind. Eng. Chem. Res.* **1994**, *33* (4), 790−799.
(18) Broadbelt, L. J.; Stark, S. M.; Klein, M. T. *Ind. Eng. Chem. Res.* **1995**, *34* (8), 2566−2573.
(19) Hatzimanikatis, V.; Li, C.; Ionita, J. A.; Henry, C. S.; Jankowski, M. D.; Broadbelt, L. J. *Bioinformatics* **2004**, *21* (8), 1603−1609.

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.

Methods

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.[20] 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.[21] 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.[19,22]

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 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 subtracting the substrate matrix from the product matrix. 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 1. These reaction operators are then implemented in BNICE, and the framework was used to identify all the possible linear polyketide chains.

Results and Discussion

Theoretical Calculation of Number of Possible Polyketides. Manipulation of the variables that control polyketide biosynthesis leads to the synthesis of different structures. Therefore, the total number of possible structures that can be produced was analyzed with respect to the different variables involved in polyketide synthesis. 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 are not the only variables 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

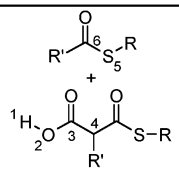
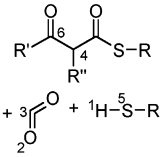
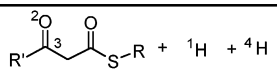
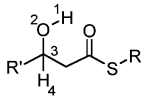
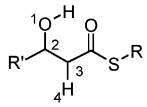
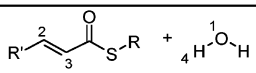
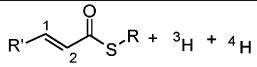
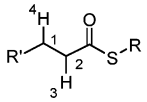
$$N_{m,tot} = s[b(a - a_c) + (2b - 1)a_c]^m \qquad (1)$$

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 the result of a cyclization of the linear chain, which 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,[7,23] increasing the number of possible extender units that can be used in the synthesis, a, to two; therefore, the theoretical number of possible structures increases by 1 order of magnitude to over three million possible structures. However, malonyl-CoA and methylmalonyl-CoA are not the

(20) Tipton, K.; Boyce, S. *Bioinformatics* **2000**, *16* (1), 34−40.
(21) Hatzimanikatis, V.; Li, C.; Ionita, J. A.; Broadbelt, L. J. *Curr. Opin. Struct. Biol.* **2004**, *14* (3), 300−6.
(22) Li, C.; Ionita, J. A.; Henry, C. S.; Jankowski, M. D.; Hatzimanikatis, V.; Broadbelt, L. J. *Chem. Eng. Sci.* **2004**, *21* (8), 1603−1609.
(23) Xue, Q.; Ashley, G.; Hutchinson, C. R.; Santi, D. V. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96* (21), 11740−5.

Table 1. Table of BNICE Reaction Operators for the Four Reactions in the Synthesis of the Polyketide Linear Chain

Generalized Enzyme Function	Reactant	Product	Reaction Operator
4.1.1/ 2.3.1			
	1H 2O 3C 4C 5S 6C 1H 0 1 0 0 0 0 2O 1 4 1 0 0 0 3C 0 1 0 1 0 0 4C 0 0 1 0 0 0 5S 0 0 0 0 4 1 6C 0 0 0 0 1 0	1H 2O 3C 4C 5S 6C 1H 0 0 0 0 1 0 2O 0 4 2 0 0 0 3C 0 2 0 0 0 0 4C 0 0 0 0 0 1 5S 1 0 0 0 4 0 6C 0 0 0 1 0 0	1H 2O 3C 4C 5S 6C 1H 0 -1 0 0 1 0 2O -1 0 1 0 0 0 3C 0 1 0 -1 0 0 4C 0 0 -1 0 0 1 5S 1 0 0 0 0 -1 6C 0 0 0 1 -1 0
1.1.1			
	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
4.2.1			
	1O 2C 3C 4H 1O 4 1 0 0 2C 1 0 1 0 3C 0 1 0 1 4H 0 0 1 0	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
1.3.1			
	1C 2C 3H 4H 1C 0 2 0 0 2C 2 0 0 0 3H 0 0 1 0 4H 0 0 0 1	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

only 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.[23]

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

Scheme 1. Overall Polyketide Synthesis Reaction

$$\mathrm{CHO\text{-}PKS} + m\mathrm{CHO_3\text{-}ACP} + n\mathrm{NADPH_2^+} \rightarrow$$

$$\mathrm{CHO\text{-}PKS} + c\mathrm{CO_2} + h\mathrm{H_2O} + n\mathrm{NADP^+} + s\mathrm{H\text{-}ACP}$$

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 in Scheme 1where 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

$$\text{C: } \lambda m + \epsilon = \alpha + c \quad (2)$$

$$\text{H: } 2n + \sigma m + \kappa = \beta + 2h + s \quad (3)$$

$$\text{O: } 3m + 1 = \gamma + 2c + h \quad (4)$$

$$\text{ACP: } m + 1 = s + 1 \quad (5)$$

The mechanism for the synthesis of the linear chain presented in Figure 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:

$$m = c \quad (6)$$

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:

$$n \leq 2m \quad (7)$$

Similarly, only one water molecule is released during each elongation step, and it cannot 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:

$$h \leq \min(m, n) \quad (8)$$

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 0 to 12; therefore, the number of water molecules produced can range from 0 to 6, 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 (Supporting Information).

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. 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.

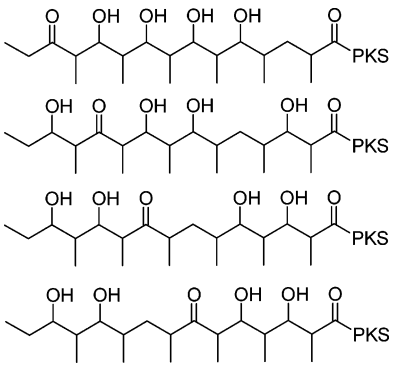


Figure 3. Four structures of the 30 possible polyketide linear chains with two carbonyl groups, four hydroxyl groups, and one single bond, obtained with a synthesis requiring the use of one molecule of propionyl-ACP, six molecules of methylmalonyl-ACP, and six molecules of $NADPH_2^+$ and producing one molecule of water.

From the formulated mechanism of linear chain synthesis shown in Figure 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:

$$\delta + \zeta + \theta + \xi = m \quad (9)$$

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) were determined:

$$\begin{bmatrix} \delta = m - n + i \\ \zeta = n - h - i \\ \theta = h - i, \qquad n \leq m, i \in [0, \min(n - h, h)] \\ \xi = i \end{bmatrix} \quad (10)$$

and

$$\begin{bmatrix} \delta = i \\ \zeta = m - h - i \\ \theta = h - (n - m) - i, \quad n > m, i \in [0, \min(h - n + m, m - h)] \\ \xi = n - m + i \end{bmatrix} \quad (11)$$

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 was calculated as

$$N_{m,n,h,i} = N(\delta_{m,n,h,i}) \cdot N(\zeta_{m,n,h,i}) \cdot N(\theta_{m,n,h,i}) \cdot N(\xi_{m,n,h,i}) \quad (12)$$

where the number of structures for each of the four different groups were calculated based on combinatorial theory:

$$N_{m,n,h,i}(\delta) = \frac{(m)!}{\delta!(m-\delta)!} \quad (13)$$

$$N_{m,n,h,i}(\zeta) = \frac{(m-\delta)!}{\zeta!(m-\delta-\zeta)!} \quad (14)$$

$$N_{m,n,h,i}(\theta) = \frac{(m-\delta-\zeta)!}{\theta!(m-\delta-\zeta-\theta)!} \quad (15)$$

$$N_{m,n,h,i}(\xi) = \frac{(m-\delta-\zeta-\theta)!}{\xi!(m-\delta-\zeta-\theta-\xi)!} = \frac{(m-\delta-\zeta-\theta)!}{\xi!} \quad (16)$$

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

$$N_{m,n,h} = \sum_i \frac{m!}{\delta!\zeta!\theta!\xi!} \quad (17)$$

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 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.[8] 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 numbers of elongation steps and different starter and extender units using eqs 10, 11, and 17.

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 led 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 2.

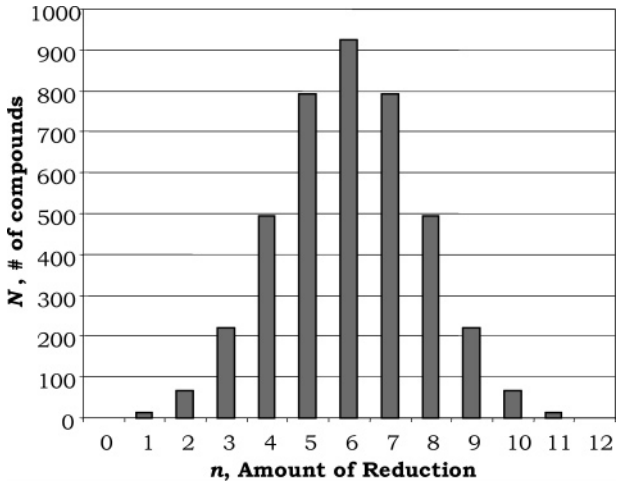


Figure 4. Distribution of possible polyketide structures synthesized with one propionyl-CoA and six methylmalonyl-ACP molecules as a function of the number of $NADPH_2^+$ molecules required for the synthesis, calculated using eq 17.

Table 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 elongation steps	generation, g												
	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 eq 1, it is expected that using an input of propionyl-ACP and methylmalonyl-ACP as reactants and without taking into account stereochemistry 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 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 2, the number of molecules generated increases with generation number or pathway length. Designat-

ing the generation number as g and the number of new molecules produced in generation g as $N(g)$, the following equation was derived to calculate the number of new molecules produced in a generation:

$$N(g) = 2N(g-1) - x_g \tag{18}$$

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:

$$x_g = \begin{cases} 0, g < 5 \\ 1, g = 5 \\ N(g-5), g > 5 \end{cases} \tag{19}$$

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 2. For example, molecules that are the result of two elongation reactions are produced in generations 1 through 7. Based on the computational results, formulas were 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:

$$g_{1,m} = m \tag{20}$$

$$n_{g,m} = 2m + g_{1,m} + 1 \tag{21}$$

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 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 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 eq 18 for the number of new products identified in a generation.

The modular nature of the synthesis is also observed in the metabolic tree shown in Figure 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 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. 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.

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. We presented here a theoretical analysis

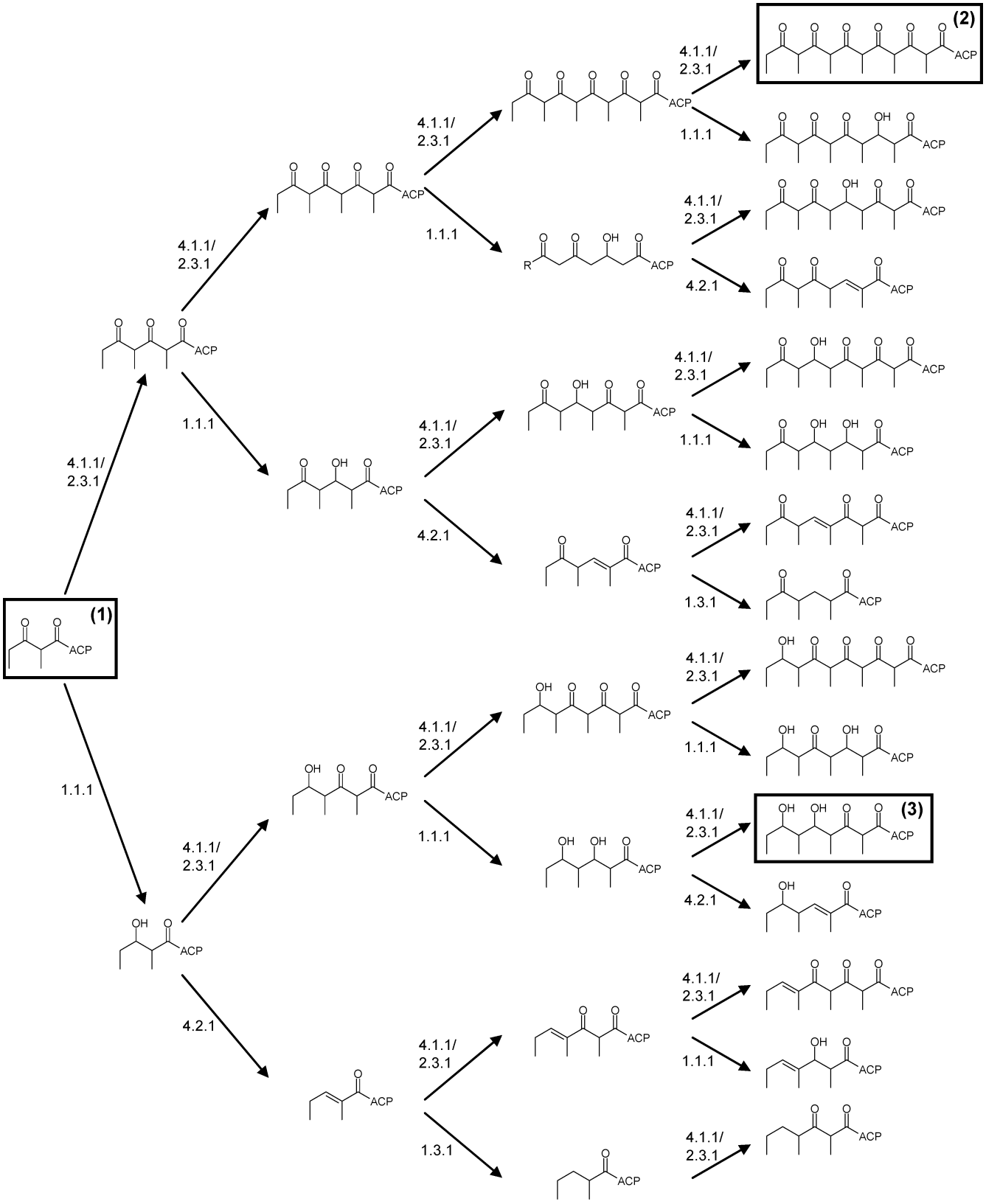


Figure 5. Metabolic pathway architecture of polyketide linear chain growth. The structure labeled as (1) is the starting molecule for the pathway; as shown, it has a carbonyl group at the β-carbon. This molecule is capable of undergoing a condensation, or 4.1.1/2.3.1, and a reduction, or reverse 1.1.1, reaction, giving rise to two structures. These two structures in turn give rise to four structures through two condensation and two reduction reactions. The four structures then produce eight structures through four condensation and four reduction reactions. These eight structures are capable of generating 15 structures via eight condensation reactions and seven reductions. Molecule 2, an aromatic polyketide, and Molecule 3 are examples of structures generated after four enzyme actions from the starting material.

that allows the quantification and characterization of the combinatorial complexity of the polyketide pathways. In addition, the reaction rules of the enzymes involved in polyketide biosynthesis were implemented in BNICE, a novel computa-

tional framework that allows the de novo synthesis of metabolic pathways based on a set of enzyme reaction rules.

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.[7,23] 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.[7,23] 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, 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.

To identify the final polyketide structures, the cyclization reactions and tailoring steps, 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. Additionally, it is possible that the synthesis of some of these undiscovered structures is not thermodynamically favorable. Therefore, the BNICE framework will be expanded to include a thermodynamic analysis, involving quantum chemical calculations, of each of the reactions involved in a synthesis in order to estimate their feasibility in a cellular environment. Additionally, metabolic flux analysis will aid in determining the effect of cellular constraints on the synthesis yield of target polyketide structures.

Acknowledgment. This work was partially supported by the US Department of Energy, Genomes to Life Program. Additionally, J.G.L. received fellowships from the NIH Biotechnology Training Program, NIH Grant 2 T32 GM008449-11, and from the GEM Foundation. V.H. received support from DuPont through a DuPont Young Professor Award.

Supporting Information Available: List of theoretically possible reactions for the synthesis of polyketides formed after six elongation steps and variable amount of reduction and dehydration. This material is available free of charge via the Internet at http://pubs.acs.org.

JA051586Y