iScience

CellPress
REVIEWS

Review

Computational Tools for Discovering and Engineering Natural Product Biosynthetic Pathways

Hengqian Ren,[1] Chengyou Shi,[1] and Huimin Zhao[1,2,3,*]

Natural products (NPs), also known as secondary metabolites, are produced in bacteria, fungi, and plants. NPs represent a rich source of antibacterial, antifungal, and anticancer agents. Recent advances in DNA sequencing technologies and bioinformatics unveiled nature's great potential for synthesizing numerous NPs that may confer unprecedented structural and biological features. However, discovering novel bioactive NPs by genome mining remains a challenge. Moreover, even with interesting bioactivity, the low productivity of many NPs significantly limits their practical applications. Here we discuss the progress in developing bioinformatics tools for efficient discovery of bioactive NPs. In addition, we highlight computational methods for optimizing the productivity of NPs of pharmaceutical importance.

INTRODUCTION

Natural products (NPs) are molecules of great medical importance that usually originate from bacteria, fungi, and plants. Since the early 20th century when penicillin was first discovered and clinically used for controlling infection, tens of thousands of NPs were identified in all three domains of life and many were approved as drugs. Based on the statistics from 1981 to 2014, around half of the US Food and Drug Administration-approved drugs have at least a NP origin (Newman and Cragg, 2016). To find drug leads for currently untreatable diseases such as Alzheimer disease and deal with the rise of antimicrobial resistance, there has always been a pressing need for discovering new NPs (Dey et al., 2017; Hernando-Amado et al., 2019). However, after the golden age of NP discovery in the 1960s to the 1970s, the rate of NP discovery drastically decreased (Shen, 2015). Moreover, another related challenge is that the manufacturing cost for NP-based drugs is relatively high, which to a large extent is due to the low productivity of NPs in native hosts (Pham et al., 2019).

The biosynthesis of NPs usually involves multi-step enzymatic reactions, which are usually referred to as *biosynthetic pathways*. Traditional methods for discovering NPs mostly relied on bioactivity assays and product isolation and purification technologies, whereas advances in genomics and bioinformatics led to an alternative strategy called *genome mining* in which biosynthetic pathways are computationally predicted and prioritized for downstream experiments including NP isolation and characterization (Smanski et al., 2016). More importantly, coupled with synthetic biology tools, this genome mining strategy can overcome some inherent limitations of the traditional strategies, such as lack of appropriate cultivation methods for target organisms and lack of ways to bypass negative regulation of target pathways in native hosts. As many biosynthetic pathways have undetectable expression levels in native hosts unless specific signals exist, one may manipulate the pathway of interest and remove elements that are potentially related to repression (Rutledge and Challis, 2015). However, identification of biosynthetic pathways from the genome context is not easy, which necessitates the development of computational tools for genome mining.

On the other hand, design and optimization of biosynthetic pathways for overproducing NPs with demonstrated medical importance is also extremely valuable. Long time and high resource cost for cultivating NP-producing organisms often result in prohibitive prices of the corresponding drugs (Pham et al., 2019). For example, taxol, one of the most widely used anticancer drug, is naturally synthesized by *Taxus brevifolia* forest tree. The most common strategy to produce this drug is through extraction of the tree bark, which needs 3,000 trees to obtain a kilogram of taxol (Nazhand et al., 2019). Although producing NPs in a cell factory through reconstitution of the corresponding biosynthetic pathways can be a very promising way to address this challenge, many issues exist when expressing a pathway in heterologous hosts, such as

[1] Department of Chemical and Biomolecular Engineering, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

[2] Departments of Chemistry, Biochemistry, and Bioengineering, Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

[3] Center for Advanced Bioenergy and Bioproducts Innovation, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

*Correspondence: zhao5@illinois.edu

https://doi.org/10.1016/j.isci.2019.100795

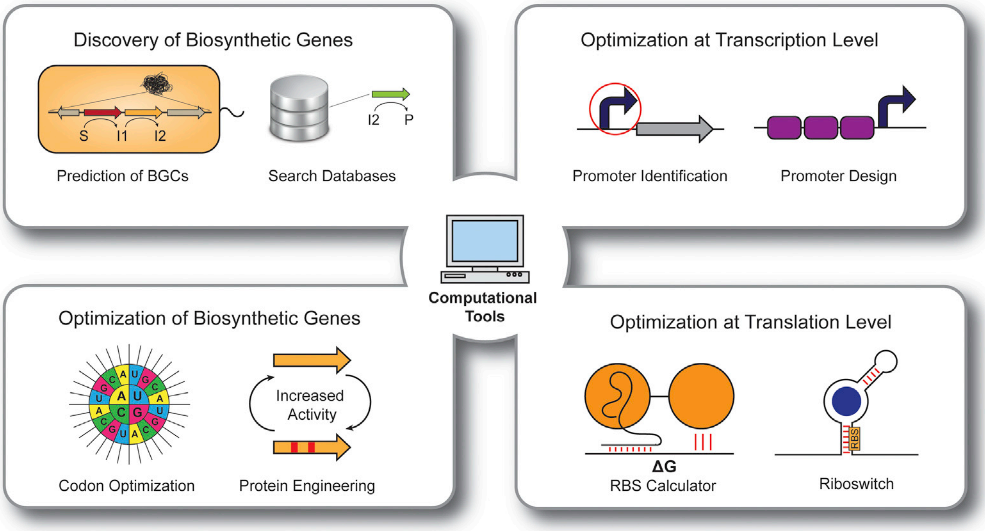

Figure 1. Summary of Computational Strategies for Natural Product Biosynthetic Pathway Prediction and Optimization Described in This Review

loss of activity or malfunction of biosynthetic enzymes as well as metabolic burden on the host. Therefore, computational approaches that can optimize the target pathway at different levels for NP overproduction are highly desirable.

In this review, we will highlight some computational methods that have been experimentally demonstrated or can be potentially used for the identification and optimization of NP biosynthetic pathways (Figure 1). For readers who may have a more general interest in discovery and metabolic engineering of NPs, we refer them to a recently published themed collection named Engineering of Cell Factories for the Production of Natural Products (Weber, 2019).

IDENTIFICATION AND ANALYSIS OF NATURAL PRODUCT BIOSYNTHETIC PATHWAYS

Bioinformatics Tools for Identifying Natural Product Biosynthetic Gene Clusters

With rapid advances in DNA sequencing technologies, many bioinformatics tools have been developed for mining sequenced microbial genomes (see Table 1 for a summary of biosynthetic gene clusters (BGCs) prediction tools mentioned in this review). Most of them are signature-based genome mining tools, which utilize profile hidden Markov models (HMMs) or Basic Local Alignment Search Tool (BLAST) searches to identify signature genes responsible for specific NP biosynthesis. Examples include ClustScan (Starcevic et al., 2008), NP.searcher (Li et al., 2009), SMURF (Khaldi et al., 2010), and antiSMASH (Blin et al., 2013, 2017, 2019; Medema et al., 2011; Weber et al., 2015). Notably, databases for known BGCs such as MIBiG (Minimum Information about a Biosynthetic Gene Cluster) 2.0 and antiSMASH (antibiotics & Secondary Metabolite Analysis Shell) database can also be leveraged for genome mining (Kautsar et al., 2019; Blin et al., 2019).

ClustScan (Cluster Scanner) (Starcevic et al., 2008) was developed to specifically analyze modular clusters, including polyketide synthase (PKS), non-ribosomal peptide synthetase (NRPS), and hybrid PKS/NRPS BGCs in bacterial genomes. It employs a top-down approach based on profile HMMs to annotate BGCs encoding modular biosynthetic enzymes, allowing the semi-automatic structural prediction of the products. The predicted chemical structures of products can be exported in a SMILES/SMARTS format for further analysis by standard chemistry programs.

NP.searcher (Natural Product searcher) (Li et al., 2009) web server is one of the first open-source genome mining tools specialized in predicting the possible chemical structures resulting from PKS, NRPS, and PKS/NRPS BGCs in bacterial genomes. Its algorism uses BLAST to align sequences with seed NRPS and PKS sequences and recognize clusters of catalytic domains and auxiliary domains that function during the

Computational Tools	Target Organism	BGC Prediction	Input	Chemical Structure Prediction	Key Features	URL	Reference
AntiSMASH	Bacteria and fungi	Unrestricted	DNA sequences	Yes	Integrate multiple BGC prediction tools/algorithms: ClusterFinder, NaPDoS, RODEO	https://antismash.secondarymetabolites.org	(Blin et al., 2013, 2017, 2019; Medema et al., 2011; Weber et al., 2015)
PlantiSMASH	Plant	Unrestricted	DNA sequences	No	Plant-adapted pHMMs and cluster detection rules and support for co-expression analysis	http://plantismash.secondarymetabolites.org	(Kautsar et al., 2017)
NP.searcher	Bacteria	NRPS, PKS, NRPS/PKS	DNA sequences	Yes	Predict 2D and 3D structure of NRPS/PKS	http://dna.sherman.lsi.umich.edu	(Li et al., 2009)
SMURF	Fungi	NRPS, PKS, NRPS/PKS, DMATS	Protein sequences and chromosomal coordinates of genes	No	Use gene coordinates as well as protein sequences as input	www.jcvi.org/smurf	(Khaldi et al., 2010)
ClustScan	Bacteria	NRPS, PKS	DNA sequences	Yes	First employ pHMMs of signature genes for BGC prediction	Obtain by request at novalis@novalis.hr	(Starcevic et al., 2008)
eSNaPD	Bacteria	Unrestricted	Metagenomic DNA	No	Uncover biosynthetic diversity from metagenomic data	http://esnapd2.rockefeller.edu	(Reddy et al., 2014)
ClusterFinder	Bacteria	Unrestricted	DNA sequences	No	Prediction is based on Pfam domain frequencies	https://github.com/petercim/ClusterFinder	(Cimermancic et al., 2014)
EvoMining	Bacteria	Unrestricted	DNA sequences	No	Genome mining based on evolutionary principles	https://github.com/nselem/evomining	(Selem-Mojica et al., 2019)
NRPS-PKS/SBPKS	Bacteria	NRPS, PKS	Protein sequences	Yes	Model 3D structures of individual PKS catalytic domains	http://www.nii.ac.in/sbspks.html	(Anand et al., 2010)
NaPDoS	Metagenomic sample	NRPS, PKS	Protein or DNA sequences	No	Phylogenic approach for domain analysis, various query types including genome contigs	http://npdomainseeker.ucsd.edu	(Ziemert et al., 2012)

Table 1. Summary of Computational Tools for Pathway Prediction Highlighted in This Review

(Continued on next page)

Table 1. Continued

Computational Tools	Target Organism	BGC Prediction	Input	Chemical Structure Prediction	Key Features	URL	Reference
BAGEL	Bacteria	Bacteriocin, RIPP	DNA sequences	No	Single-input whole-genome analysis for bacteriocin and RIPP BGC detection	http://bagel.molgenrug.nl	(van Heel et al., 2013)
RODEO	Bacteria	RIPP	Protein accession number	Yes	Combine hidden Markov model-based analysis, heuristic scoring and machine learning	http://ripp.rodeo	(Tietz et al., 2017)

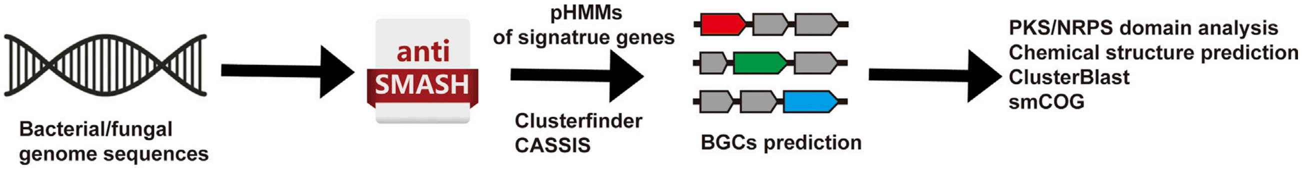

Figure 2. General Workflow of antiSMASH

Primarily, antiSMASH identifies signature biosynthetic genes (hit on their pHMMs) that encode enzymes responsible for generating a class-specific scaffold and locates a cluster based on a set of manually curated BGC cluster rules. Alternatively, it uses the ClusterFinder algorithm for BGC prediction. For fungal BGCs, Cluster Assignment by Islands of Sites (CASSIS) is integrated for better prediction of gene cluster boundaries. In addition, several downstream analyses can be performed: NRPS/PKS domain analysis and annotation, prediction of the core chemical structure of PKSs and NRPSs, substrate specificity prediction for NRPS 'A' domains (via NRPSpredictor2) (Röttig et al., 2011), ClusterBlast gene cluster comparative analysis, and smCOG secondary metabolism protein family analysis.

assembly of the core products. BLAST is also used to identify additional enzymes within clusters, which serve to further tailor the products following core molecule assembly. Unlike ClustScan, which only predicts basic linear and cyclical conformations of a polyketide compound, NP.searcher outputs putative candidate non-ribosomal peptide and polyketide specialized metabolites in SMILES format, which enables 2D and 3D structure representations in assorted chemical software.

The SMURF (Secondary Metabolite Unknown Regions Finder) tool (Khaldi et al., 2010) was specifically designed for mining secondary metabolite biosynthetic gene clusters in fungi, including PKS, NRPS, PKS/NRPS, and dimethylallyl tryptophan synthase (DMATS). It relies on profile HMMs much in the same way as CLUSEAN (Weber et al., 2009) to annotate signature genes, and then the algorithm scans a window of ±20 genes neighboring the signature gene to look for other tailoring enzymes depending on the presence of protein domains from a training set of 22 *Aspergillus fumigatus* BGCs. Unlike ClustScan and NP.searcher, which require DNA sequences as an input, SMURF requires gene coordinates as well as protein sequences as input and it lacks the prediction of the chemical structure of the final product.

AntiSMASH (Medema et al., 2011) is a comprehensive pipeline for identifying a broad range of secondary metabolite BGCs in bacterial and fungal genomes, including those producing polyketides, non-ribosomal peptides, terpenes, aminoglycosides, aminocoumarins, indolocarbazoles, lantibiotics, bacteriocins, nucleosides, beta-lactams, butyrolactones, siderophores, melanins, and others. On top of using signature gene profile HMMs for gene cluster identification, antiSMASH uses a "greedy" algorithmic approach to extend the gene cluster by 5, 10, or 20 kb on both sides; therefore closely spaced clusters can be merged into "superclusters." In the latest antiSMASH pipeline (v.5.0) (Blin et al., 2019), the so-called superclusters is renamed as "regions," which contain multiple mutually exclusive candidate clusters for better interpretation of hybrid BGCs. Furthermore, antiSMASH provides additional options of NRPS/PKS domain analysis and annotation, prediction of the core chemical structure of polyketides and non-ribosomal peptides, ClusterBlast gene cluster comparative analysis, and secondary metabolism protein family analysis (smCOG). Finally, the output can be visualized in one user-friendly interactive XHTML page (see Figure 2 for a general workflow of antiSMASH). It is noteworthy that plantiSMASH, an antiSMASH derivative for plant genome mining, includes plant-specific cluster detection rules, co-expression analysis to identify pathways within and between clusters, and comparative genomic analysis to study the evolutionary conservation of each cluster (Kautsar et al., 2017).

Other novel genome mining tools have also been developed for the prediction of new classes of NP BGCs. ClusterFinder (Cimermancic et al., 2014), an open-source pHMM-based probabilistic algorithm, predicts BGCs by analyzing the probability that locally encoded protein domains belong to a BGC using a training set of 677 experimentally characterized gene clusters as well as 100 randomly selected non-BGC regions. It is worth mentioning that the updated antiSMASH pipeline (v.3.0) integrates ClusterFinder algorithm for the detection of gene clusters of unknown types (Weber et al., 2015). Another genome mining tool, EvoMining (evolutionarily driven genome mining) (Cruz-Morales et al., 2016; Selem-Mojica et al., 2019), is based on the identification of functionally diverged paralogs of primary metabolic enzymes that have acquired functions in specialized metabolism during evolution.

In addition to the above-mentioned bioinformatics tools developed for BGC prediction from individual genomes, a web-based bioinformatics tool eSNaPD (environmental Surveyor of Natural Product Diversity)

(Reddy et al., 2014) was developed to survey NP BGC diversity in metagenomic DNA sequences. In this approach, PCR-generated sequence tags from metagenomic DNA are compared with a reference database of characterized gene clusters to estimate the biosynthetic diversity hidden within a pool of metagenomic DNA and discover BGCs encoding novel NPs.

Many genome mining tools are designed to predict a particular type of secondary metabolite BGC of interest. Two bioinformatics tools devised for NRPS and PKS prediction are NRPS-PKS/SBSPKS and NaPDoS (Anand et al., 2010; Ziemert et al., 2012). SBSPKS (Structure Based Sequence analysis of PolyKetide Synthases) is a web-based software tool that can be used to model 3D structures of bacterial type I PKS and allows a sequence-based comparison of query PKS with a database of 167 experimentally characterized NRPS/PKS gene clusters consisting of roughly 4,400 catalytic domains. The NaPDoS (Natural Product Domain Seeker) web portal employs a phylogeny-based classification system to identify candidate KS and C domains from amino acid or DNA sequences. It features the identification of new domain lineages from diversified query types, including the incomplete genome assemblies obtained by next-generation sequencing technologies.

Ribosomally synthesized and post-translationally modified peptide (RiPPs) BGCs are notoriously difficult to detect computationally (Arnison et al., 2013) due to the absence of a shared signature biosynthetic gene across all pathways. BAGEL (van Heel et al., 2013) and antiSMASH both use single-input whole-genome analysis for detecting several known classes of RiPPs. In 2017, RODEO (Rapid ORF Description and Evaluation Online) was reported for rapid RiPP BGC prediction, which combines pHMMs, heuristic scoring, and machine learning to identify BGCs and predict RiPP precursor peptides (Tietz et al., 2017). AntiSMASH 4.0 integrates RODEO algorithm for more accurate RiPP BGC prediction (Blin et al., 2017).

Once the BGCs have been identified computationally, there is often a need to find out whether the NPs encoded by the target BGCs are biologically active or not. Nature has evolved multiple strategies to protect antibiotic producers from being killed by the antibiotics themselves. One of them is that the producer encodes a second copy of the antibiotic target gene in the corresponding BGC, which is called *resistance gene*. Such resistance gene was successfully used as a signature for genome mining of antibiotics and providing insights for its mode of action (Thaker et al., 2014; Tang et al., 2015). To automate the discovery and identification of resistance genes, Ziemert and coworkers developed the antibiotic resistant target seeker (ARTS) tool, which can detect resistance genes in actinobacterial genomes based on three criteria: duplication, localization within a BGC, and evidence of horizontal gene transfer (Alanjary et al., 2017). In addition, some BGC prediction algorithms were also recently updated to incorporate functionality for resistance gene identification, such as PRediction Informatics for Secondary Metabolomes (PRISM) 3 (Skinnider et al., 2017).

Computational Methods for Identifying Unclustered Natural Product Biosynthetic Pathways

Computational approaches have also been developed to identify NP biosynthetic pathways even when the involved genes are not clustered, such as in plants. Co-expression and evolutionary genomics are two promising strategies to elucidate non-clustered specialized metabolic pathways.

In a typical co-expression analysis, candidate genes encoding biosynthetic enzymes for the biosynthesis of a compound of interest can be identified by using the genes for characterized enzymes as bait and ranking all other genes by correlation coefficient to the bait. In cases where the final structure of compound is unknown, the co-expression analysis can be complemented by untargeted metabolomics. For example, by using the gene encoding characterized cytochrome P450 as a bait to detect other genes that were strongly co-expressed, a previously unknown 4-hydroxyindole-3-carbonyl nitrile pathway in *A. thaliana* was uncovered (Rajniak et al., 2015). Using a similar strategy, the podophyllotoxin pathway from mayapple was also identified by the same laboratory (Lau and Sattely, 2015). In addition to co-expression analysis using bait genes, WGCNA (weighted correlation network analysis) can be used for measuring gene co-expression patterns. Candidate pathways can be identified by extracting modules of highly correlated genes from the correlation network (Langfelder and Horvath, 2008). Co-expression networks can also be constructed across multiple species, as exemplified by CoExpNetViz (Tzfadia et al., 2016). The rationale behind cross-species co-expression networks is that genes involved in the same pathway remain co-expressed during evolution and identifying conserved co-expression of orthologous groups of genes will aid in pathway identification.

Evolutionary genomic analysis is another promising approach of predicting functional connections between biosynthetic genes by using phylogenetic profiling to find co-occurrence across genomes. For example, CLIME (clustering by inferred models of evolution) is a tree-structured algorithm to cluster gene sets based on evolutionary history and has the potential to predict new members of a pathway based on shared inferred ancestry (Li et al., 2014).

RECONSTITUTION AND OPTIMIZATION OF NATURAL PRODUCT BIOSYNTHETIC PATHWAYS IN MODEL ORGANISMS

Reconstitution of Natural Product Biosynthesis in Model Organisms

Many NPs of pharmaceutical interest originate from organisms, such as plants, whose cultivation requires excessive time and resource, which usually ends up in expensive prices for corresponding drugs in the market and restricts their availability for patients. From a synthetic biology perspective, such obstacles can be potentially addressed through biosynthetic pathway reconstitution in microorganisms that are amenable to modern industrial fermentation. Model organisms such as *E. coli* for prokaryotes and *Saccharomyces cerevisiae* for eukaryotes are ideal chassis for heterologous expression of NP biosynthetic pathways with various origins. However, NP biosynthetic pathways are usually uncharacterized to some extent and the identification of missing enzymes can be very challenging. While missing enzymes can be identified by analyzing native producers as aforementioned, nature provides a vast number of enzymes for catalyzing numerous reactions. Therefore, mining enzymes with desired properties from organisms rather than the native host to construct artificial pathways for synthesizing NPs represents an attractive alternative strategy.

Similar to the strategy in organic synthesis, retrosynthesis can also be used for artificial pathway design, which is sometimes referred as *retrobiosynthesis* (Hadadi and Hatzimanikatis, 2015). An NP of interest can be deconstructed one step reaction at a time into simpler precursors, and the process can be performed iteratively until the precursors are readily available (e.g., from primary metabolism). Afterward, biosynthetic enzymes that are able to catalyze each proposed reaction step need to be discovered from databases and used for reconstructing the whole biosynthetic route. To date, 10 retrobiosynthesis-based pathway design tools are available, and only RetroPath has been experimentally tested for designing biosynthetic routes to pinocembrin, which has been reviewed elsewhere (Table 2) (Wang et al., 2017; Cravens et al., 2019; Delepine et al., 2018). Possible reasons for that can be to some extent the universal complicated chemical structures of NPs whose synthesis demands enzymes with properly balanced substrate specificity and tolerance, making the selection of appropriate enzymes a difficult task. Therefore, retrobiosynthetic tools that can rank the candidate enzymes by properties such as substrate similarity and/or promiscuity would be more suitable for designing artificial NP pathways (Cravens et al., 2019).

Although retrobiosynthesis is a generally applicable method for artificially designing pathways for synthesizing any compound even when the natural biosynthetic mechanism is uncharacterized, one may be interested in synthesizing NPs with demonstrated medical values. Owing to their pharmaceutical importance, such NPs have usually been studied for years and the corresponding pathways are characterized to some extent. However, the biosynthesis reconstitution can still be very challenging because the pathway may be incomplete and enzyme malfunctions can be observed in the selected host. Therefore, computational methods that can search enzymes with proper function from databases and prioritize for experimental test are particularly desirable in such scenarios (Figure 1). Currently, genomic databases such as Kyoto Encyclopedia of Genes and Genomes (KEGG) and medicinal plant genome project, as well as transcriptomic databases such as the 1000 plants (1KP) project, are frequently used for biosynthetic enzyme discovery (Table 2) (Kanehisa and Goto, 2000; Chen et al., 2011; Matasci et al., 2014). One recently reported application is the successful reconstitution of the cannabinoid biosynthetic pathway in *S. cerevisiae* (Luo et al., 2019). Cannabinoids constitute a series of compounds that act on cannabinoid receptors with emerging clinical applications (Aizpurua-Olaizola et al., 2016). The key precursor in cannabinoid biosynthesis, cannabigerolic acid, is produced from olivetolic acid and geranyl pyrophosphate by a geranylpyrophosphate:olivetolate geranyltransferase (GOT). Although the activity of GOT was detected in *Cannabis* extracts two decades ago, no GOT activity was ever reconstituted in yeast. By mining the transcriptomics of *Cannabis* as well as its close relative *Humulus lupulus* by BLAST using characterized GOTs as queries, Keasling and coworkers identified eight candidates and one exhibited activity in yeast. Another exemplary research is the identification of DRS-DRR, which converts *S*-reticuline to *R*-reticuline, a key step in opioid biosynthesis (Galanie et al., 2016). Previous isotope feeding studies indicate that the conversion from *S*-reticuline to *R*-reticuline proceeds via oxidation to Schiff base intermediate and stereospecific reduction. In addition,

Name	Category	Target	Function	Key Features	URL	Reference
RetroPath 2.0	Computational tool	Biosynthetic genes	Artificial pathway design	Automated open source workflow for retrosynthesis based on generalized reaction rules	https://www.myexperiment.org/workflows/4987.html	(Delepine et al., 2018)
Kyoto Encyclopedia of Genes and Genomes (KEGG)	Database	Biosynthetic Genes		Database for systematic analysis of gene functions, linking genomic information with higher order functional information	https://www.genome.jp/kegg/	(Kanehisa and Goto, 2000)
1000 Plants (1KP) Project	Database	Biosynthetic genes		Transcriptome data from over 1,000 plant species	http://www.bioinfodata.org/app/Blast4OneKP/home/	(Matasci et al., 2014)
CoStar	Computational tool	Biosynthetic genes	Codon optimization	Optimization of gene sequences by avoiding hairpins, GC content variation and repeat	http://life.sysu.edu.cn/COStar/COStar.html	(Liu et al., 2014)
Presyncodon	Computational tool	Biosynthetic genes	Codon optimization	Design of gene sequences for optimized heterologous expression by machine learning	http://www.mobioinfor.cn/presyncodon_www/index.html	(Tian et al., 2018)
BacPP	Computational tool	Promoter	Promoter prediction	Identification of promoters from genome sequences by a machine learning algorithm	www.bacpp.bioinfoucs.com/home	(Silva et al., 2011, 2016)
WebGeSter DB	Database	Terminator		Database containing a million terminators identified in 1,060 bacterial genome sequences and 798 plasmids	http://pallab.serc.iisc.ernet.in/gester	(Mitra et al., 2011)
RBS Calculator	Computational tool	RBS	RBS design	Control of expression level by designing RBSs with various strength	http://www.salis.psu.edu/software	(Salis et al., 2009; Salis, 2011; Borujeni et al., 2014; Farasat et al., 2014)
ClusterCAD	Computational tool	Biosynthetic genes	Protein engineering	Design of chimeric PKSs	https://clustercad.jbei.org/	(Eng et al., 2018)

Table 2. Summary of Computational Tools and Databases for Pathway Optimization Highlighted in This Review

plant gene silencing studies indicate that codeinone reductase (COR) knockdown results in reticuline accumulation and in one case, specifically *S*-reticuline accumulation. Therefore, the reported virus-induced gene silencing sequence of the COR enzyme was used as a BLAST query against *Papaver* species in the 1KP project and PhvtoMetaSyn transcriptome database. Four enzymes with both a cytochrome P450 oxidase (CYP) 82Y1 domain and a COR-like domain fusion protein were identified, which led to the identification of Pbr.89405. The activity of Pbr.89405 was then confirmed by heterologous expression in yeast.

Optimization of Enzymes in a Natural Product Biosynthetic Pathway

Codon Optimization of Biosynthetic Genes

Owing to the variation of biological context for gene expression from one organism to another, codon optimization is a widely used strategy to increase the success rate of heterologous gene expression, and development of computational tools that can automate the design process is both desirable and necessary, particularly in cases such as NP pathway reconstitution when multiple genes need to be optimized at the same time (Figure 1). Early codon optimization tools, including UpGene, JCat, and OPTIMIZER, rely on codon usage data for different organisms and optimize the target gene by the abundancy of synonymous codons for each amino acid residue, which is usually referred as "CAI = 1" theory (Gao et al., 2004; Grote et al., 2005; Puigbo et al., 2007). However, synonymous codons are not randomly used in nature and these methods overlooked many criteria for synonymous codon selection. For example, local GC content can change the stability and form secondary structures in mRNAs, which is harmful to translation (Kudla et al., 2006). In addition, varying ratios of the low-frequency-usage and high-frequency-usage codons in a gene would control the translation speed of the protein, which is critical for the folding of the synthesized protein fragment and the assembly of the structural elements of the protein (Li et al., 2014). Therefore, many codon optimization tools using more sophisticated algorithms were developed.

Global properties of the resulting DNA, such as GC content and repetitive sequences, after codon optimization have a significant impact on protein expression, which was considered by many codon optimization tools as designing criteria. Using the sliding window approach, Fath and coworkers developed a multiparameter RNA and codon optimization tool to assess and enhance autologous mammalian gene expression, in which nine sequence-based parameters, including increasing GC content, removing destabilizing RNA elements, and avoiding RNA secondary structures, were taken into account for sequence design (Fath et al., 2011). Following RNA and codon optimization, 50 candidate genes representing five classes of human proteins—transcription factors, ribosomal and polymerase subunits, protein kinases, membrane proteins, and immunomodulators—all showed reliable and elevated expression. Later on, Liu and coworkers combined a sliding window approach with a dynamic search using the D-star lite algorithm and developed CoStar algorithm (Liu et al., 2014). Compared with other codon optimization tools such as GeneOptimizer and EuGene, COStar gave lower scores of hairpins, lower variance of GC content, and fewer repeats (Table 2).

Besides individual codon bias, usage of sequential codons is also not random and unique to each species, which is known as codon pair usage or codon context bias (Quax et al., 2015). Although the evolutionary basis for such phenomena is still not completely clear, codon optimization tools considering codon context bias have been developed and proved to be successful. To compare the importance of individual codon usage (ICU) and codon context (CC) in sequence optimization, Lee and coworkers developed novel computational procedures to formulate appropriate mathematical expressions to quantify the ICU and CC fitness of a coding sequence and applied optimization procedures to maximize ICU and/or CC fitness (Chung and Lee, 2012). The resultant *in silico* optimized DNA sequence suggested that CC is a more relevant design criterion than the commonly considered ICU in four normally used organisms for heterologous expression, including *E. coli*, *Lactococcus lactis*, *Pichia pastoris,* and *S. cerevisiae*. More recently, Alexaki and coworkers constructed a database to include genomic codon-pair and dinucleotide statistics of all organisms with sequenced genomes available in the GenBank, which provides an invaluable resource for recombinant gene design (Alexaki et al., 2019).

In recent years, algorithms based on machine learning were also developed for codon optimization. Tian and coworkers developed Presyncodon, which learned codon usage patterns of the residue in the context of specific fragments and predicted synonymous codon selection in *E. coli* (Table 2) (Tian et al., 2017). By using Presyncodon, the authors designed eGFP and mApple and expressed in *E. coli*, which gave 2.3- and 1.7-fold higher fluorescence, respectively, than their counterparts that were optimized by only using high-frequency-usage codons. Later on, the same research group further extended this method to *Bacillus subtilis* and *S. cerevisiae* (Tian et al., 2018).

Although different codon optimization methods taking design criteria from various perspectives were mentioned above, one may have specific considerations for optimizing the gene of interest as these methods may not be generally applicable. For example, Claassens and coworkers evaluated the expression of six wild-type membrane-integrated proteins in *E. coli* and compared them with their codon-optimized and codon-harmonized counterparts and found that not a single algorithm performed consistently best for the protein production (Claassens et al., 2017). Therefore, packages of different codon optimization algorithms that provide sufficient designer flexibility were also developed (Jayaraj et al., 2005; Jung and McDonald, 2011; Gaspar et al., 2012; Chin et al., 2014; Sequeira et al., 2017; Yu et al., 2017; Rehbein et al., 2019).

Protein Engineering of Biosynthetic Enzymes

Codon optimization tools can be used to increase the expression level of biosynthetic genes in the heterologous host, whereas computation-aided protein engineering strategies can also be used for engineering biosynthetic enzymes with desired properties for NP synthesis (Figure 1). Many computational tools for template-based or *de novo* protein structure prediction and sequence design have been developed, but their applications in NP synthesis are rarely reported (Kuhlman and Bradley, 2019). However, recently, Keasling and coworkers developed ClusterCAD, a computational tool for rational design of chimeric type I modular PKS to synthesize polyketides, a major family of NPs, with designer structures (Eng et al., 2018). A type I modular PKS consists of multiple catalytic domains arranged in a specific order and works together as an assembly line to produce the polyketide in a stepwise manner, and such a biosynthetic machinery enables researchers to produce designer polyketides by rearranging catalytic domains in PKSs accordingly (Fischbach and Walsh, 2006; Alanjary et al., 2019). However, engineering PKSs to produce designer polyketides was achieved with varying success rates that can be largely affected by the way of module selection (Alanjary et al., 2019). Through analysis of sequence similarity between PKS modules and structure similarity of their cognate polyketide intermediates, ClusterCAD helps its users to identify the best starting PKS that can generate the polyketide with the highest similarity to the designed structure and select donor modules for catalytic domain exchange. By using ClusterCAD, the authors successfully designed a chimeric PKS to produce adipic acid.

Through iterative rounds of genetic diversity generation and screening/selection for improved properties such as activity, substrate selectivity, regio- or steroselectivity, and stability, directed evolution has been proved to be a highly effective approach for protein engineering. However, traditional directed evolution methods have their inherent limitations. For example, protein variant libraries constructed through traditional methods such as random mutagenesis can hardly achieve a comprehensive sequence space, whereas such a library may still be too large for high-throughput screening because developing an appropriate high-throughput screening method for a particular reaction of interest is not trivial. Therefore, computational methods are also developed, which could provide effective ways to identify mutational hotspots and therefore make a ''smart library,'' which has much a smaller library size with increased possibility for positive hits. Such computational tools, including ones that identify mutational hotspots through sequence or structure comparison, as well as active sites through molecular docking, have been developed, which are readily accessible with a user-friendly interface, and were recently reviewed (Ebert and Pelletier, 2017). Assisted by machine learning, further development of algorithms for designing smart libraries is also under progress (Yang et al., 2019). Recently, Maranas and coworkers provided a comprehensive review of enzyme engineering milestones and summarized 50 successful cases of computational enzyme design (Chowdhury and Maranas, 2019). Although very few of these cases are related to NP synthesis, we envision that computational tools will be increasingly applied for designing biosynthetic enzymes with improved or altered activity for NP biosynthesis.

Optimization of Regulation in a Natural Product Biosynthetic Pathway

Identification of Regulatory Elements

To reconstruct a biosynthetic pathway for heterologous expression in the selected host, the availability of *cis*-elements, such as promoters and terminators, that regulate the expression of each biosynthetic gene are of particular importance. Although methods such as direct cloning are also applicable for heterologous expression of gene clusters, the phylogenetic distance between the organism from which the BGC originates and the heterologous host does matter for the success of expression (Smanski et al., 2016; Cobb and Zhao, 2012). One underlying reason could be the failure of recognition of the original promoter sequences by the transcription machinery in the heterologous host if too much phylogenetic difference exists. Therefore, replacement of the original regulatory elements would be sometimes critical. In addition, pathway optimization strategies, such as fine-tuning the expression level of each biosynthetic enzyme to balance the metabolic flux or applying dynamic control to reduce metabolic burden, also require promoters with varying strength or regulatory systems that may respond to particular compounds or signal molecules (Hammer et al., 2006). Thus, identification of regulatory elements with various properties is of great importance for the successful reconstitution and optimization of NP biosynthetic pathways (Figure 1).

One effective way for identifying regulatory elements for particular application is searching existing databases. In 2018, Szymczyk and coworkers summarized 40 databases that centralize experimental and theoretical knowledge regarding the organization of promoters, interacting transcription factors, and microRNAs in many eukaryotic and prokaryotic species, including some model organisms such as *E. coli*, *B. subtilis*, *S. cerevisiae,* and *A. thaliana* (Majewska et al., 2018). These databases can provide information on not only promoter sequences but also annotations such as transcription start sites that are sufficient for applications such as pathway reconstruction and even provide insight for design of artificial promoters. Although relatively less reported, other regulatory elements in addition to promoter sequences are also indispensable elements for pathway reconstruction. For example, terminator databases, such as WebGeSter DB, are also of particular interest, especially for pathway reconstruction in eukaryotes in which terminators have to be involved for successful transcription (Table 2) (Mitra et al., 2011).

Based on the knowledge about signatures in typical regulatory sequences, computational tools developed in the early days can predict *cis*-elements such as promoters and riboswitches mainly by searching existing databases and score the sequence in query by pattern matching (Münch et al., 2005; Gautheret and Lambert, 2001; Abreu-Goodger and Merino, 2005; de Jong et al., 2012; Dreos et al., 2015). Although proven to be successful, accuracy of prediction is still compromised in these tools by relatively simplified algorithms, and regulatory elements with uncharacterized features can be overlooked. In recent years, machine learning has found many applications in life sciences and bioengineering. Besides assisting development of codon optimization tools as aforementioned, machine learning has also been used for promoter prediction from complex genome sequence contexts. For example, Gerhardt and coworkers developed BacPP for promoter prediction using neural network (NN) approach, which is able to identify degenerated, imprecise, and incomplete patterns merged within those sequences and can achieve high performance when processing extended genome sequences (Table 2) (Silva et al., 2011). BacPP was based on rules derived from NN learning process of six sigma factor-dependent promoter sequences and can predict promoter sequences from *E. coli* genome with high accuracy (at least above 80%), and BacPP implemented within a user-friendly platform is also available online (Silva et al., 2016). More recently, Umarov and coworkers successfully predicted promoters from five distant organisms including human, mouse, plant (*Arabidopsis*). and two bacteria (*E. coli* and *B. subtilis*) by using convolutional NNs to build predictive models. Machine learning approaches have also been applied for predicting other regulatory elements, such as histone marks, which will not be described in detail in this review (Zhou and Troyanskaya, 2015).

Artificial Design of Regulatory Elements

Besides the identification of naturally occurring regulatory elements, efforts have also been made in developing computational tools for artificially designing regulatory elements with preferred properties (Figure 1). Such strategies have been proved to be very successful in designing regulatory elements at the translational level. One of the most widely known computational tool for automated design of ribosomal-binding sites on mRNA sequences to control corresponding gene expression level is the RBS calculator (Figure 3 and Table 2) (Salis et al., 2009; Salis, 2011; Borujeni et al., 2014). The RBS calculator relies on the sophisticated biophysical modeling of translation initiation process and proved to be able to predict the translation initiation rate within a high dynamic range on a proportional scale. As in most cases the initiation is the rate-limiting step in bacterial translation process, RBS calculator is therefore a valuable tool for controlling protein expression level by designing 5′ UTR sequences in mRNAs. The quantitative prediction of expression level with high accuracy also makes RBS calculator an ideal tool for controlling expression levels of multiple genes simultaneously, while avoiding combinatorial explosion when using library-based approaches (Figure 3) (Farasat et al., 2014; Jeschek et al., 2016). In addition to using RBSs to constitutively tune gene

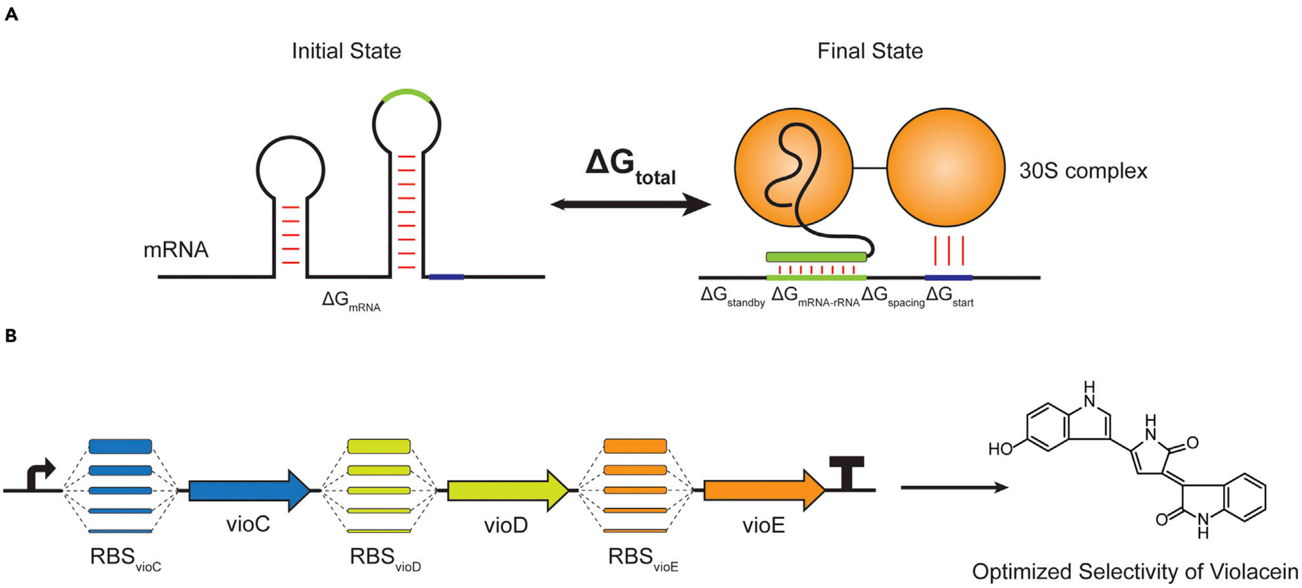

Figure 3. Overview of the RBS Calculator

(A) Description of the thermodynamic model of bacterial translation initiation used in RBS calculator.

(B) Application of RBS calculator in rational design of RBS library and optimization of violacein biosynthesis (Jeschek et al., 2016).

expression level, riboswitch aptamers can also be used for controlling gene expression in response to specific signals. Riboswitches are able to sense a broad range of small molecules including vitamins, amino acids, and nucleotides and control the transcription or translation of the host mRNA. Computational tools have been developed for riboswitch identification and structure prediction, which has been comprehensively reviewed elsewhere (Antunes et al., 2018).

Although less reported, algorithms and computation-assisted methods have also been used for designing promoter sequences (Figure 1). One pioneering work by Mogno and coworkers provided a thermodynamic model of combinatorial regulation and explained 75% variance in gene expression in synthetic promoter libraries with different strength TATA boxes (Mogno et al., 2010). Possibly due to the more complicated regulatory mechanism of promoters, artificial design of promoters usually requires multiple design-build-test cycles, which cannot only be achieved *in silico*. A recent exemplary work on synthetic promoter design combined next-generation sequencing approach with machine learning, in which promoters with enhanced cell-state specificity (SPECS) were identified that exhibit distinct spatiotemporal activity during the programmed differentiation of induced pluripotent stem cells, as well as for breast cancer and glioblastoma stem-like cells (Wu et al., 2019). Another challenging task for artificial regulatory element design is *de novo* construction of regulatory proteins. To address this challenge, Baker and coworkers recently developed a latching orthogonal cage-key proteins (LOCKR) system and used it for controlling yeast mating pathways, which shed light on designing protein-based regulatory elements (Langan et al., 2019; Ng et al., 2019).

CONCLUSION AND FUTURE PERSPECTIVES

NP biosynthetic pathways are essentially part of their hosts' metabolic network, which are complicated systems involving both internal delicate and sophisticated catalysis and regulation and external interactions with the metabolism context. Therefore, identification and optimization of NP biosynthetic pathways usually involve very intricate data interpretation and design tasks that cannot be finished without assistance of computational tools. In this review, we highlighted some important computational tools and their application for discovery and optimization of NP pathways.

Discovery of NPs through identification of BGCs using bioinformatics analysis is very promising and has been experimentally demonstrated many times. However, for pathways whose biosynthetic genes are not clustered, identification of even an individual gene can be very challenging. Comparison of omics data sometimes generates too many hits to be tested experimentally, and ranking algorithms can have inconsistent performance. Although retrobiosynthesis can be an alternative strategy, the underlying design rules, such as to quantitatively evaluate enzyme candidates for specific reaction by both promiscuity and specificity, are still vague. Besides incorporating new scientific findings and updating algorithms for natural pathway identification or artificial pathway construction, automation coupled with synthetic biology and machine learning can accelerate the design-build-test cycle and help develop models with better accuracy. Such concept was recently demonstrated in the optimization of a lycopene biosynthetic pathway (HamediRad et al., 2019).

Optimization of biosynthetic pathways for increasing productivity can happen in both catalysis and regulation aspects. Successful expression and acceptable performance of biosynthetic enzymes are critical for heterologous biosynthesis, whereas appropriate regulation can balance the metabolic flux that makes the biosynthesis more efficiently. Assisted by computational tools, factors that may affect the productivity of target molecules can be optimized independently. Besides developing better models and algorithms to make the process more accurately and efficiently, scoring methods to prioritize the rate-limiting factor for optimization in a particular pathway design would also be very useful.

ACKNOWLEDGMENTS

This work was supported by awards GM077596 and AI144967 (to H.Z.) from the U.S. National Institutes of Health (NIH) and awards DE-SC0018260 and DE-SC0018420 (to H.Z.) from the U.S. Department of Energy.

AUTHOR CONTRIBUTIONS

H.R. and H.Z outlined the manuscript. All authors wrote the manuscript and prepared the figures. All authors edited and provided feedback on the manuscript.

REFERENCES

Abreu-Goodger, C., and Merino, E. (2005). RibEx: a web server for locating riboswitches and other conserved bacterial regulatory elements. Nucleic Acids Res. *33*, W690–W692.

Aizpurua-Olaizola, O., Soydaner, U., Öztürk, E., Schibano, D., Simsir, Y., Navarro, P., Etxebarria, N., and Usobiaga, A. (2016). Evolution of the cannabinoid and terpene content during the growth of Cannabis sativa plants from different chemotypes. J. Nat. Prod. *79*, 324–331.

Alanjary, M., Kronmiller, B., Adamek, M., Blin, K., Weber, T., Huson, D., Philmus, B., and Ziemert, N. (2017). The Antibiotic Resistant Target Seeker (ARTS), an exploration engine for antibiotic cluster prioritization and novel drug target discovery. Nucleic Acids Res. *45*, W42–W48.

Alanjary, M., Cano-Prieto, C., Gross, H., and Medema, M.H. (2019). Computer-aided re-engineering of nonribosomal peptide and polyketide biosynthetic assembly lines. Nat. Prod. Rep. *36*, 1249–1261.

Alexaki, A., Kames, J., Holcomb, D.D., Athey, J., Santana-Quintero, L.V., Lam, P.V.N., Hamasaki-Katagiri, N., Osipova, E., Simonyan, V., Bar, H., et al. (2019). Codon and codon-pair usage tables (CoCoPUTs): facilitating genetic variation analyses and recombinant gene design. J. Mol. Biol. *431*, 2434–2441.

Anand, S., Prasad, M.V., Yadav, G., Kumar, N., Shehara, J., Ansari, M.Z., and Mohanty, D. (2010). SBSPKS: structure based sequence analysis of polyketide synthases. Nucleic Acids Res. *38*, W487–W496.

Antunes, D., Jorge, N.A.N., Caffarena, E.R., and Passetti, F. (2018). Using RNA sequence and structure for the prediction of riboswitch aptamer: a comprehensive review of available software and tools. Front. Genet. *8*, 231.

Arnison, P.G., Bibb, M.J., Bierbaum, G., Bowers, A.A., Bugni, T.S., Bulaj, G., Camarero, J.A., Campopiano, D.J., Challis, G.L., Clardy, J., et al. (2013). Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature. Nat. Prod. Rep. *30*, 1568.

Blin, K., Medema, M.H., Kazempour, D., Fischbach, M.A., Breitling, R., Takano, E., and Weber, T. (2013). antiSMASH 2.0-a versatile platform for genome mining of secondary metabolite producers. Nucleic Acids Res. *41*, W204–W212.

Blin, K., Wolf, T., Chevrette, M.G., Lu, X., Schwalen, C.J., Kautsar, S.A., Suarez Duran, H.G., de Los Santos, E.L.C., Kim, H.U., Nave, M., et al. (2017). antiSMASH 4.0-improvements in chemistry prediction and gene cluster boundary identification. Nucleic Acids Res. *45*, W36–W41.

Blin, K., Shaw, S., Steinke, K., Villebro, R., Ziemert, N., Lee, S.Y., Medema, M.H., and Weber, T. (2019). antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. Nucleic Acids Res. *47*, W81–W87.

Blin, K., Pascal Andreu, V., de Los Santos, E.L.C., Del Carratore, F., Lee, S.Y., Medema, M.H., and Weber, T. (2019). The antiSMASH database version 2: a comprehensive resource on secondary metabolite biosynthetic gene clusters. Nucleic Acids Res. *47*, D625–D630.

Borujeni, A.E., Channarasappa, A.S., and Salis, H.M. (2014). Translation rate is controlled by coupled trade-offs between site accessibility, selective RNA unfolding and sliding at upstream standby sites. Nucleic Acids Res. *42*, 2646–2659.

Chen, S., Xiang, L., Guo, X., and Li, Q. (2011). An introduction to the medicinal plant genome project. Front. Med. *5*, 178–184.

Chin, J.X., Chung, B.K.S., and Lee, D.Y. (2014). Codon Optimization OnLine (COOL): a web-based multi-objective optimization platform for synthetic gene design. Bioinformatics *30*, 2210–2212.

Chowdhury, R., and Maranas, C.D. (2019). From directed evolution to computational enzyme engineering-A review. Aiche J. https://doi.org/10.1002/aic.16847.

Chung, B.K.S., and Lee, D.Y. (2012). Computational codon optimization of synthetic gene for protein expression. BMC Syst. Biol. *6*, 134.

Cimermancic, P., Medema, M.H., Claesen, J., Kurita, K., Wieland Brown, L.C., Mavrommatis, K., Pati, A., Godfrey, P.A., Koehrsen, M., Clardy, J., et al. (2014). Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. Cell *158*, 412–421.

Claassens, N.J., Siliakus, M.F., Spaans, S.K., Creutzburg, S.C.A., Nijsse, B., Schaap, P.J., Quax, T.E.F., and van der Oost, J. (2017). Improving heterologous membrane protein production in *Escherichia coli* by combining transcriptional tuning and codon usage algorithms. PLoS One *12*, e0184355.

Cobb, R.E., and Zhao, H.M. (2012). Direct cloning of large genomic sequences. Nat. Biotechnol. *30*, 405–406.

Cravens, A., Payne, J., and Smolke, C.D. (2019). Synthetic biology strategies for microbial biosynthesis of plant natural products. Nat. Commun. *10*, 1–12.

Cruz-Morales, P., Kopp, J.F., Martínez-Guerrero, C., Yáñez-Guerra, L.A., Selem-Mojica, N., Ramos-Aboites, H., Feldmann, J., and Barona-Gómez, F. (2016). Phylogenomic analysis of natural products biosynthetic gene clusters allows discovery of arseno-organic metabolites in model streptomycetes. Genome Biol. Evol. *8*, 1906–1916.

Delepine, B., Duigou, T., Carbonell, P., and Faulon, J.L. (2018). RetroPath2.0: a retrosynthesis workflow for metabolic engineers. Metab. Eng. *45*, 158–170.

Dey, A., Bhattacharya, R., Mukherjee, A., and Pandey, D.K. (2017). Natural products against Alzheimer's disease: Pharmaco-therapeutics and biotechnological interventions. Biotechnol. Adv. *35*, 178–216.

Dreos, R., Ambrosini, G., Perier, R.C., and Bucher, P. (2015). The eukaryotic promoter database: expansion of EPDnew and new promoter analysis tools. Nucleic Acids Res. *43*, D92–D96.

Ebert, M.C.C.J.C., and Pelletier, J.N. (2017). Computational tools for enzyme improvement: why everyone can - and should - use them. Curr. Opin. Chem. Biol. *37*, 89–96.

Eng, C.H., Backman, T.W.H., Bailey, C.B., Magnan, C., García Martín, H., Katz, L., Baldi, P., and Keasling, J.D. (2018). ClusterCAD: a computational platform for type I modular polyketide synthase design. Nucleic Acids Res. *46*, D509–D515.

Farasat, I., Kushwaha, M., Collens, J., Easterbrook, M., Guido, M., and Salis, H.M. (2014). Efficient search, mapping, and optimization of multi-protein genetic systems in diverse bacteria. Mol. Syst. Biol. *10*, 731.

Fath, S., Bauer, A.P., Liss, M., Spriestersbach, A., Maertens, B., Hahn, P., Ludwig, C., Schäfer, F., Graf, M., and Wagner, R. (2011). Multiparameter RNA and codon optimization: a standardized tool to assess and enhance autologous mammalian gene expression. PLoS One *6*, e17596.

Fischbach, M.A., and Walsh, C.T. (2006). Assembly-line enzymology for polyketide and nonribosomal peptide antibiotics: Logic, machinery, and mechanisms. Chem. Rev. *106*, 3468–3496.

Galanie, S., Thodey, K., Trenchard, I., Interrante, M.F., and Smolke, C. (2016). Complete biosynthesis of opioids in yeast. Science *349*, 1095–1100.

Gao, W.T., Rzewski, A., Sun, H.J., Robbins, P.D., and Gambotto, A. (2004). UpGene: application of a web-based DNA codon optimization algorithm. Biotechnol. Progr. *20*, 443–448.

Gaspar, P., Oliveira, J.L., Frommlet, J., Santos, M.A.S., and Moura, G. (2012). EuGene: maximizing synthetic gene design for heterologous expression. Bioinformatics *28*, 2683–2684.

Gautheret, D., and Lambert, A. (2001). Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles. J. Mol. Biol. *313*, 1003–1011.

Grote, A., Hiller, K., Scheer, M., Münch, R., Nörtemann, B., Hempel, D.C., and Jahn, D. (2005). JCat: a novel tool to adapt codon usage of a target gene to its potential expression host. Nucleic Acids Res. *33*, W526–W531.

Hadadi, N., and Hatzimanikatis, V. (2015). Design of computational retrobiosynthesis tools for the design of de novo synthetic pathways. Curr. Opin. Chem. Biol. *28*, 99–104.

HamediRad, M., Chao, R., Weisberg, S., Lian, J., Sinha, S., and Zhao, H. (2019). Towards a fully automated algorithm driven platform for biosystems design. Nat. Commun. *10*, 5150.

Hammer, K., Mijakovic, I., and Jensen, P.R. (2006). Synthetic promoter libraries - tuning of gene expression. Trends Biotechnol. *24*, 53–55.

van Heel, A.J., de Jong, A., Montalban-Lopez, M., Kok, J., and Kuipers, O.P. (2013). BAGEL3: automated identification of genes encoding bacteriocins and (non-)bactericidal posttranslationally modified peptides. Nucleic Acids Res. *41*, W448–W453.

Hernando-Amado, S., Coquet, T.M., Baquero, F., and Martinez, J.L. (2019). Defining and combating antibiotic resistance from one Health and global Health perspectives. Nat. Microbiol. *4*, 1432–1442.

Jayaraj, S., Reid, R., and Santi, D.V. (2005). GeMS: an advanced software package for designing synthetic genes. Nucleic Acids Res. *33*, 3011–3016.

Jeschek, M., Gerngross, D., and Panke, S. (2016). Rationally reduced libraries for combinatorial pathway optimization minimizing experimental effort. Nat. Commun. *7*, 11163.

de Jong, A., Pietersma, H., Cordes, M., Kuipers, O.P., and Kok, J. (2012). PePPER: a webserver for prediction of prokaryote promoter elements and regulons. BMC Genomics *13*, 299.

Jung, S.K., and McDonald, K. (2011). Visual gene developer: a fully programmable bioinformatics software for synthetic gene optimization. BMC Bioinformatics *12*, 340.

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. *28*, 27–30.

Kautsar, S.A., Duran, H.G.S., Blin, K., Osbourn, A., and Medema, M.H. (2017). plantiSMASH: automated identification, annotation and expression analysis of plant biosynthetic gene clusters. Nucleic Acids Res. *45*, W55–W63.

Kautsar, S.A., Blin, K., Shaw, S., Navarro-Muñoz, J.C., Terlouw, B.R., van der Hooft, J.J.J., van Santen, J.A., Tracanna, V., Suarez Duran, H.G., Pascal Andreu, V., et al. (2019). MIBiG 2.0: a repository for biosynthetic gene clusters of known function. Nucleic Acids Res. https://doi.org/10.1093/nar/gkz882.

Khaldi, N., Seifuddin, F.T., Turner, G., Haft, D., Nierman, W.C., Wolfe, K.H., and Fedorova, N.D. (2010). SMURF: genomic mapping of fungal secondary metabolite clusters. Fungal Genet. Biol. *47*, 736–741.

Kudla, G., Lipinski, L., Caffin, F., Helwak, A., and Zylicz, M. (2006). High guanine and cytosine content increases mRNA levels in mammalian cells. PLoS Biol. *4*, 933–942.

Kuhlman, B., and Bradley, P. (2019). Advances in protein structure prediction and design. Nat. Rev. Mol. Cell Biol. *20*, 681–697.

Langan, R.A., Boyken, S.E., Ng, A.H., Samson, J.A., Dods, G., Westbrook, A.M., Nguyen, T.H., Lajoie, M.J., Chen, Z., Berger, S., et al. (2019). *De novo* design of bioactive protein switches. Nature *572*, 205–210.

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics *9*, 559.

Lau, W., and Sattely, E.S. (2015). Six enzymes from mayapple that complete the biosynthetic pathway to the etoposide aglycone. Science *349*, 1224–1228.

Li, M.H.T., Ung, P.M.U., Zajkowski, J., Garneau-Tsodikova, S., and Sherman, D.H. (2009). Automated genome mining for natural products. BMC Bioinformatics *10*, 185.

Li, Y., Calvo, S.E., Gutman, R., Liu, J.S., and Mootha, V.K. (2014). Expansion of biological pathways based on evolutionary inference. Cell *158*, 213–225.

Li, G.W., Burkhardt, D., Gross, C., and Weissman, J.S. (2014). Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. Cell *157*, 624–635.

Liu, X.W., Deng, R.Q., Wang, J.W., and Wang, X.Z. (2014). COStar: a D-star Lite-based dynamic search algorithm for codon optimization. J. Theor. Biol. *344*, 19–30.

Luo, X., Reiter, M.A., d'Espaux, L., Wong, J., Denby, C.M., Lechner, A., Zhang, Y., Grzybowski, A.T., Harth, S., and Lin, W. (2019). Complete biosynthesis of cannabinoids and their unnatural analogues in yeast. Nature *567*, 123–126.

Majewska, M., Wysokinska, H., Kuzma, L., and Szymczyk, P. (2018). Eukaryotic and prokaryotic promoter databases as valuable tools in exploring the regulation of gene transcription: a comprehensive overview. Gene *644*, 38–48.

Matasci, N., Hung, L.H., Yan, Z., Carpenter, E.J., Wickett, N.J., Mirarab, S., Nguyen, N., Warnow, T., Ayyampalayam, S., Barker, M., et al. (2014). Data access for the 1,000 Plants (1KP) project. Gigascience *3*, 17.

Medema, M.H., Blin, K., Cimermancic, P., de Jager, V., Zakrzewski, P., Fischbach, M.A., Weber, T., Takano, E., and Breitling, R. (2011). antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. Nucleic Acids Res. *39*, W339–W346.

Mitra, A., Kesarwani, A.K., Pal, D., and Nagaraja, V. (2011). WebGeSTer DB-a transcription terminator database. Nucleic Acids Res. *39*, D129–D135.

Mogno, I., Vallania, F., Mitra, R.D., and Cohen, B.A. (2010). TATA is a modular component of synthetic promoters. Genome Res. *20*, 1391–1397.

Münch, R., Hiller, K., Grote, A., Scheer, M., Klein, J., Schobert, M., and Jahn, D. (2005). Virtual footprint and PRODORIC: an integrative framework for regulon prediction in prokaryotes. Bioinformatics *21*, 4187–4189.

Nazhand, A., Durazzo, A., Lucarini, M., Mobilia, M.A., Omri, B., and Santini, A. (2019). Rewiring cellular metabolism for heterologous biosynthesis of Taxol. Nat. Prod. Res. *34*, 1–12.

Newman, D.J., and Cragg, G.M. (2016). Natural products as sources of new drugs from 1981 to 2014. J. Nat. Prod. *79*, 629–661.

Ng, A.H., Nguyen, T.H., Gómez-Schiavon, M., Dods, G., Langan, R.A., Boyken, S.E., Samson, J.A., Waldburger, L.M., Dueber, J.E., Baker, D., and El-Samad, H. (2019). Modular and tunable biological feedback control using a *de novo* protein switch. Nature *572*, 265–269.

Pham, J.V., Yilma, M.A., Feliz, A., Majid, M.T., Maffetone, N., Walker, J.R., Kim, E., Cho, H.J., Reynolds, J.M., Song, M.C., et al. (2019). A review of the microbial production of bioactive natural products and biologics. Front. Microbiol. *10*, 1404.

Puigbo, P., Guzman, E., Romeu, A., and Garcia-Vallve, S. (2007). OPTIMIZER: a web server for optimizing the codon usage of DNA sequences. Nucleic Acids Res. *35*, W126–W131.

Quax, T.E.F., Claassens, N.J., Soll, D., and van der Oost, J. (2015). Codon bias as a means to fine-tune gene expression. Mol. Cell *59*, 149–161.

Rajniak, J., Barco, B., Clay, N.K., and Sattely, E.S. (2015). A new cyanogenic metabolite in Arabidopsis required for inducible pathogen defence. Nature *525*, 376–379.

Reddy, B.V.B., Milshteyn, A., Charlop-Powers, Z., and Brady, S.F. (2014). eSNaPD: a versatile, web-based bioinformatics platform for surveying and mining natural product biosynthetic diversity from metagenomes. Chem. Biol. *21*, 1023–1033.

Rehbein, P., Berz, J., Kreisel, P., and Schwalbe, H. (2019). "CodonWizard" - an intuitive software tool with graphical user interface for customizable codon optimization in protein expression efforts. Protein Expr. Purif. *160*, 84–93.

Röttig, M., Medema, M.H., Blin, K., Weber, T., Rausch, C., and Kohlbacher, O. (2011). NRPSpredictor2-a web server for predicting NRPS adenylation domain specificity. Nucleic Acids Res. *39*, W362–W367.

Rutledge, P.J., and Challis, G.L. (2015). Discovery of microbial natural products by activation of silent biosynthetic gene clusters. Nat. Rev. Microbiol. *13*, 509–523.

Salis, H.M. (2011). The ribosome binding site calculator. Methods Enzymol. *498*, 19–42.

Salis, H.M., Mirsky, E.A., and Voigt, C.A. (2009). Automated design of synthetic ribosome binding sites to control protein expression. Nat. Biotechnol. *27*, 946–950.

Selem-Mojica, N., Aguilar, C., Gutierrez-Garcia, K., Martinez-Guerrero, C.E., and Barona-Gomez, F. (2019). EvoMining reveals the origin and fate of natural product biosynthetic enzymes. Microb. Genom. https://doi.org/10.1099/mgen.0.000260.

Sequeira, A.F., Brás, J.L.A., Fernandes, V.O., Guerreiro, C.I.P.D., Vincentelli, R., and Fontes, C.M.G.A. (2017). A novel platform for high-throughput gene synthesis to maximize recombinant expression in Escherichia coli. Methods Mol. Biol. *1620*, 113–128.

Shen, B. (2015). A new golden age of natural products drug discovery. Cell *163*, 1297–1300.

Silva, S.D.E., Echeverrigaray, S., and Gerhardt, G.J.L. (2011). BacPP: bacterial promoter prediction-A tool for accurate sigma-factor specific assignment in enterobacteria. J. Theor. Biol. *287*, 92–99.

Silva, S.D.E., Notari, D.L., Neis, F.A., Ribeiro, H.G., and Echeverrigaray, S. (2016). BacPP: a web-based tool for Gram-negative bacterial promoter prediction. Genet. Mol. Res. *15*, gmr7973.

Skinnider, M.A., Merwin, N.J., Johnston, C.W., and Magarvey, N.A. (2017). PRISM 3: expanded prediction of natural product chemical structures from microbial genomes. Nucleic Acids Res. *45*, W49–W54.

Smanski, M.J., Zhou, H., Claesen, J., Shen, B., Fischbach, M.A., and Voigt, C.A. (2016). Synthetic biology to access and expand nature's chemical diversity. Nat. Rev. Microbiol. *14*, 135–149.

Starcevic, A., Zucko, J., Simunkovic, J., Long, P.F., Cullum, J., and Hranueli, D. (2008). ClustScan: an integrated program package for the semi-automatic annotation of modular biosynthetic gene clusters and in silico prediction of novel chemical structures. Nucleic Acids Res. *36*, 6882–6892.

Tang, X., Li, J., Millán-Aguiñaga, N., Zhang, J.J., O'Neill, E.C., Ugalde, J.A., Jensen, P.R., Mantovani, S.M., and Moore, B.S. (2015). Identification of thiotetronic acid antibiotic biosynthetic pathways by target-directed genome mining. ACS Chem. Biol. *10*, 2841–2849.

Thaker, M.N., Waglechner, N., and Wright, G.D. (2014). Antibiotic resistance-mediated isolation of scaffold-specific natural product producers. Nat. Protoc. *9*, 1469–1479.

Tian, J., Yan, Y., Yue, Q., Liu, X., Chu, X., Wu, N., and Fan, Y. (2017). Predicting synonymous codon usage and optimizing the heterologous gene for expression in E. coli. Sci. Rep. *7*, 9926.

Tian, J., Li, Q.B., Chu, X.Y., and Wu, N.F. (2018). Presyncodon, a web server for gene design with the evolutionary information of the expression hosts. Int. J. Mol. Sci. *19*, 3872.

Tietz, J.I., Schwalen, C.J., Patel, P.S., Maxson, T., Blair, P.M., Tai, H.C., Zakai, U.I., and Mitchell, D.A. (2017). A new genome-mining tool redefines the lasso peptide biosynthetic landscape. Nat. Chem. Biol. *13*, 470–478.

Tzfadia, O., Diels, T., De Meyer, S., Vandepoele, K., Aharoni, A., and Van de Peer, Y. (2016). CoExpNetViz: comparative co-expression networks construction and visualization tool. Front. Plant Sci. *6*, 1994.

Wang, L., Dash, S., Ng, C.Y., and Maranas, C.D. (2017). A review of computational tools for design and reconstruction of metabolic pathways. Synth. Syst. Biotechnol. *2*, 243–252.

Weber, T. (2019). Engineering of cell factories for the production of natural products. Nat. Prod. Rep. *36*, 1231–1232.

Weber, T., Rausch, C., Lopez, P., Hoof, I., Gaykova, V., Huson, D.H., and Wohlleben, W. (2009). CLUSEAN: a computer-based framework for the automated analysis of bacterial secondary metabolite biosynthetic gene clusters. J. Biotechnol. *140*, 13–17.

Weber, T., Blin, K., Duddela, S., Krug, D., Kim, H.U., Bruccoleri, R., Lee, S.Y., Fischbach, M.A., Müller, R., Wohlleben, W., et al. (2015). antiSMASH 3.0-a comprehensive resource for the genome mining of biosynthetic gene clusters. Nucleic Acids Res. *43*, W237–W243.

Wu, M.R., Nissim, L., Stupp, D., Pery, E., Binder-Nissim, A., Weisinger, K., Enghuus, C., Palacios, S.R., Humphrey, M., Zhang, Z., et al. (2019). A high-throughput screening and computation platform for identifying synthetic promoters with enhanced cell-state specificity (SPECS). Nat. Commun. *10*, 2880.

Yang, K.K., Wu, Z., and Arnold, F.H. (2019). Machine-learning-guided directed evolution for protein engineering. Nat. Methods *16*, 687–694.

Yu, K., Ang, K.S., and Lee, D.Y. (2017). Synthetic gene design using codon optimization on-line (COOL). Methods Mol. Biol. *1472*, 13–34.

Zhou, J., and Troyanskaya, O.G. (2015). Predicting effects of noncoding variants with deep learning-based sequence model. Nat. Methods *12*, 931–934.

Ziemert, N., Podell, S., Penn, K., Badger, J.H., Allen, E., and Jensen, P.R. (2012). The natural product domain seeker NaPDoS: a phylogeny based bioinformatic tool to classify secondary metabolite gene diversity. PLoS One *7*, e34064.