J Ind Microbiol Biotechnol (2014) 41:443–450
DOI 10.1007/s10295-013-1368-1

MINI-REVIEW

SIMB
Society for Industrial Microbiology and Biotechnology

Bioinformatics tools for genome mining of polyketide and non-ribosomal peptides

Christopher N. Boddy

Received: 16 September 2013 / Accepted: 14 October 2013 / Published online: 31 October 2013

Abstract Microbial natural products have played a key role in the development of clinical agents in nearly all therapeutic areas. Recent advances in genome sequencing have revealed that there is an incredible wealth of new polyketide and non-ribosomal peptide natural product diversity to be mined from genetic data. The diversity and complexity of polyketide and non-ribosomal peptide biosynthesis has required the development of unique bioinformatics tools to identify, annotate, and predict the structures of these natural products from their biosynthetic gene clusters. This review highlights and evaluates web-based bioinformatics tools currently available to the natural product community for genome mining to discover new polyketides and non-ribosomal peptides.

Keywords Genome mining · Polyketide · Non-ribosomal peptide · Biosynthesis · Bioinformatics · Natural product discovery

Introduction

Bacterial polyketides (PKs) and non-ribosomal peptides (NRPs) have been and continue to be essential sources of chemical diversity for drug discovery and development. These complex secondary metabolites have impacted all therapeutic areas leading to clinical agents including anti-infectives, anticancer agents, cholesterol lowering drugs, and immunosuppressants. Their continued discovery remains a very active and pressing scientific concern.

The explosion in microbial genome sequencing over the past 15 years has shown that many organisms encode a wealth of PK and NRP diversity. With the increase in speed and decrease in cost of genome sequencing as well as the rise of metagenomic sequencing projects, mining sequencing data sets is becoming increasingly important in PK and NRP discovery and characterization. Key goals for the natural products chemist with these data sets are to identify new biosynthetic pathways and predict their corresponding natural products (Fig. 1). This in silico genome mining step is essential to triage known pathways or gene clusters that likely produce compounds with limited new chemical diversity so that the challenging and time consuming step of isolating and characterizing new PKs and NRPs can be focused on the highest value pathways. This review summarizes and evaluates the web-based bioinformatic tools for genome mining available to the PK and NRP natural products community.

PK and NRP biosynthetic pathways are common and widely distributed in bacterial genomes

The growth in sequence data has revealed an enormous number of PK and NRP biosynthetic pathways. Analysis shows that PK and NRP pathways are broadly distributed across multiple and diverse bacterial genomes [13, 32, 51]. Data from the 223 complete bacterial genomes sequenced prior to 2005 shows that approximately 50 % of genomes contain at least one PK or NRP pathway and that there is a correlation between genome size above 5 Mb and total bp of polyketide synthase (PKS) and non-ribosomal peptide synthetase (NRPS) genes [13]. A more recent analysis

C. N. Boddy (✉)
Departments of Chemistry and Biology, Center for Advanced Research in Environmental Genomics, University of Ottawa, Ottawa, ON K1N 6N5, Canada
e-mail: cboddy@uottawa.ca

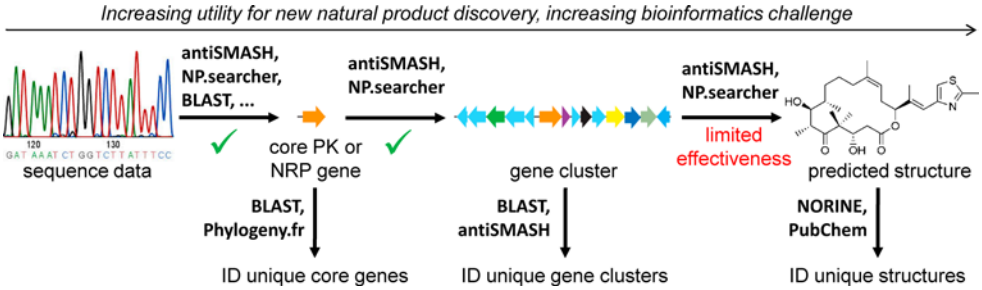

Fig. 1 Bioinformatics can play a key role in guiding natural products chemists in their genome mining projects, enabling them to focus on biosynthetic gene clusters likely encoding new natural product diversity

of 58 bacterial genomes containing representative examples of most of the sequenced bacterial phylogenic diversity, including major secondary metabolite producers, identified 180 PK and NRP biosynthetic pathways. This study also showed that greater than 50 % of genomes had one or more PK or NRP pathways [51]. The majority of pathways were from the Actinobacteria (≈9/genome) and the delta-Proteobacteria (≈7/genome), a class of bacteria not well represented in the 2005 data set. Finally, a study of 210 anaerobic bacteria genomes showed that 33 % possessed PK and NRP pathways [32].

While all three studies use different criteria for selecting the genomes to analyze as well as different methodologies for identifying PK and NRP biosynthetic pathways, a consistent trend emerges. Genomes of greater than 3 Mb are likely to have one or more PK and NRP gene clusters. Of the 2,500 complete bacterial genome sequences and the nearly 10,000 incomplete genomes, over 7,000 have genomes larger than 3 Mb (July 1 download of prokaryote.txt from NCBI) and are thus likely to have PK and NRP pathways. We can thus speculate that there are between 7,000 and 35,000 currently sequenced PK and NRP biosynthetic gene clusters, with this number projected to grow rapidly over the next few years. This is a rich data set to mine; however, its size and the diversity and complexity of PK and NRP biosynthesis have required the development of unique bioinformatics tools to identify and annotate gene clusters as well as to predict their encoded products.

Detection of PK and NRP biosynthetic pathways in genomes

There are a number of challenges that need to be met for the application of bioinformatics tools to PK and NRP genome mining. The first is to identify within a genome the location of a PK and NRP biosynthetic pathway. The most common approach is to query the translated genome with an ortholog of an expected protein from the pathway. Conserved catalytic domains such as ketosynthases (KS) from PKS pathways [22], and adenylation (A) or condensation (C) domains from NRPS pathways [23] are often used. A versatile query sequence is PksJ from *Bacillus subtilis* which contains both a type I PKS module and an NRPS module, enabling simultaneous identification of both types of pathways from a genome [61, 62]. Alternatively, highly specific target queries can be used. The recently characterized geranyl transferase from viridicatumtoxin biosynthesis was used as a query sequence for BLASTp analysis of the GenBank, JGI, and Broad-MIT fungal genome databases to identify nine gene clusters from nine different fungal genomes that likely generate geranylated aromatic polyketides [10].

A more robust tool for the detection of PK and NRP biosynthetic pathways is the use of hidden Markov models (HMMs) [15]. HMMs are statistical models generated from multiple sequences. As such they are superior to pairwise search methods like BLAST at detecting distantly related homologs. HMMs have been developed to signature proteins from type I, type II, and type III PK and NRP biosynthetic pathways [2, 18, 28, 31, 34, 44, 64]. These tools have successfully been incorporated into web-based search tools such as antiSMASH [4, 34], NP.searcher [33], NaPDoS [67], PKMiner [28] (which focuses exclusively on type II PKs), and SMURF [26] (which focuses exclusively on fungal genomes), downloadable tools like CLUSEAN [59], and commercial tools including ClustScan [50]. These tools all enable researchers to scan large DNA data sets for PK and NRP biosynthetic pathways.

Because biosynthetic pathways are typically composed of multiple enzymes, the second task is to identify all the genes involved in the biosynthesis of the metabolite. These include genes encoding PKSs and NRPSs as well as those encoding tailoring enzymes for oxidation, methylation, and glycosylation of the core structure. In addition, genes coding for biosynthesis of essential building blocks and metabolites, such as activated sugars or starter units, regulatory elements, and, if necessary, resistance mechanisms can also be present in a gene cluster. Typically these genes are all clustered tightly together on the chromosome, simplifying their identification. This task is relatively straightforward for a user manually investigating a single gene cluster. However it is challenging to automate reliably. In general

the solution has been to assume that these additional genes do not extend particularly far from the core signature genes. For example, antiSMASH defines clusters as groups of signature genes within 10 kb of each other and extends the cluster 20 kb on each side of the last signature gene to define the boundaries of PK and NRP biosynthetic gene clusters [34]. NP.searcher annotates pathways as extending 15 kb upstream and downstream from a PKS or NRPS gene. Any additional PKS or NRPS genes in this window are added to the cluster and the gene cluster is expanded a further 15 kb from the newly added gene [33]. SMURF annotates fungal pathways as the 20 genes upstream and downstream of an identified PKS or NRPS gene. This 20 gene window was established empirically by examination of 22 gene clusters from *Aspergillus fumigates* [26]. The use of large windows, 20 genes or 15-20 kb, ensures that the vast majority of pathway genes are included in the predicted gene cluster.

A concern with this automated analysis is that gene clusters located close together in the genome may be merged into superclusters, as occurs with the salinilactam biosynthetic gene cluster and the NRP siderophore gene cluster that follows less than 9 kb after it in *Salinispora tropica* [16]. While secondary metabolite biosynthesis gene clusters are often found in genomic islands [39], individual gene clusters are typically separated by greater than 30 genes or over 30 kb, minimizing supercluster formation.

For the analysis of bacterial whole genomes for gene cluster identification antiSMASH (http://antismash.secondarymetabolites.org/), NP.searcher (http://dna.sherman.lsi.umich.edu/), and NaPDoS (http://napdos.ucsd.edu/) are all viable options. Using the *Sorangium cellulosum* genome as the query sequence [48], antiSMASH detects 5 PK, 2 NRP, and 4 hybrid PK NRP gene clusters, NP.searcher identifies 2 PK, 1 NRP, and 2 hybrid PK NRP gene clusters, and NaPDoS locates 50 KS domains (from PK and hybrid gene clusters) and 20 C domains (from NRP and hybrid gene clusters). For comparison to high-quality manual annotation, the authors who annotated the *S. cellulosum* genome identified 3 PK, 2 NRP, and 4 hybrid PK NRP gene clusters [48]. The antiSMASH analysis thus compares favourably with detailed manual annotation for pathway identification. While the NaPDoS analysis provides excellent identification of the KS and C domains, it requires further analysis to identify the number of different gene clusters in the genome. Comparison of the antiSMASH data output with NP.searcher shows that antiSMASH provides a more detailed description of individual clusters identified, indicating for example if the PK gene clusters are type I, type II, type III or *trans*-AT and if a cluster also contains isoprenoid biosynthetic genes and enables the user to easily toggle between gene clusters in the browser interface. While NP.searcher also distinguishes between type I and *trans*-AT PK gene clusters, it is less flexible for toggling between gene clusters or visualizing the genetic organization of a cluster. Lastly, in terms of ease of use for whole genome analysis antiSMASH, NP.searcher, and NaPDoS are all easy to operate. While all enable users to upload DNA files, antiSMASH also enables users to enter NCBI accession numbers to perform analysis on sequenced, deposited genomes. NP.searcher and NaPDoS provide their results very rapidly, typically less than 10 min, where as antiSMASH results from a genome analysis are typically available in 4 h or longer depending on the server load.

Identifying unique gene clusters

Having identified PK or NRP biosynthetic gene clusters from genomic data sets, the key challenge is to prioritize which gene clusters to investigate experimentally. Product identification, genetic disruption, heterologous expression, and biochemical characterization of gene clusters are experimentally intensive and cannot be realistically performed for all pathways identified from a genome or metagenomic data set. Identifying the gene clusters that are likely to encode new molecules is thus a key priority for genome mining. This requires the ability to compare target gene clusters to known gene clusters to identify genetically distinct pathways, a fairly straightforward task, and to predict the structure, or at least the core, of the encoded PK or NRP, an extremely challenging bioinformatics problem.

BLAST analysis of pathways can be effective in identifying related biosynthetic pathways. For example BLASTn analysis of the oxytetracycline biosynthetic gene cluster from *Streptomyces rimosus* [65] identifies the chlorotetracycline (unpublished, NCBI accession number AB039379), SF2575 [40], and the dactylocycline [58] biosynthetic gene clusters (Fig. 2). However, BLASTn analysis of the pederin biosynthetic gene cluster [41] does not identify the onnamide biosynthetic gene cluster [42] even though these two pathways and compounds share substantial homology (Fig. 2). This is likely because the perderin biosynthetic pathway is separated into three genome regions. It is possible for nucleotide-based searches to not identify related gene clusters from organisms that differ greatly in GC content such as the actinomycetes (65–80 % GC) and firmicutes (≈35 % GC). Protein-based searches avoid this potential problem. BLASTp analysis of an individual core protein from a PK or NRP gene cluster generally provides a large number of good hits that must be manually curated to identify related pathways. For example BLASTp analysis of the last PKS protein in the biosynthesis of erythromycin, EryIIIA, from *Aeromicrobium erythreum* [5], provides over 100 hits that are statistically indistinguishable from

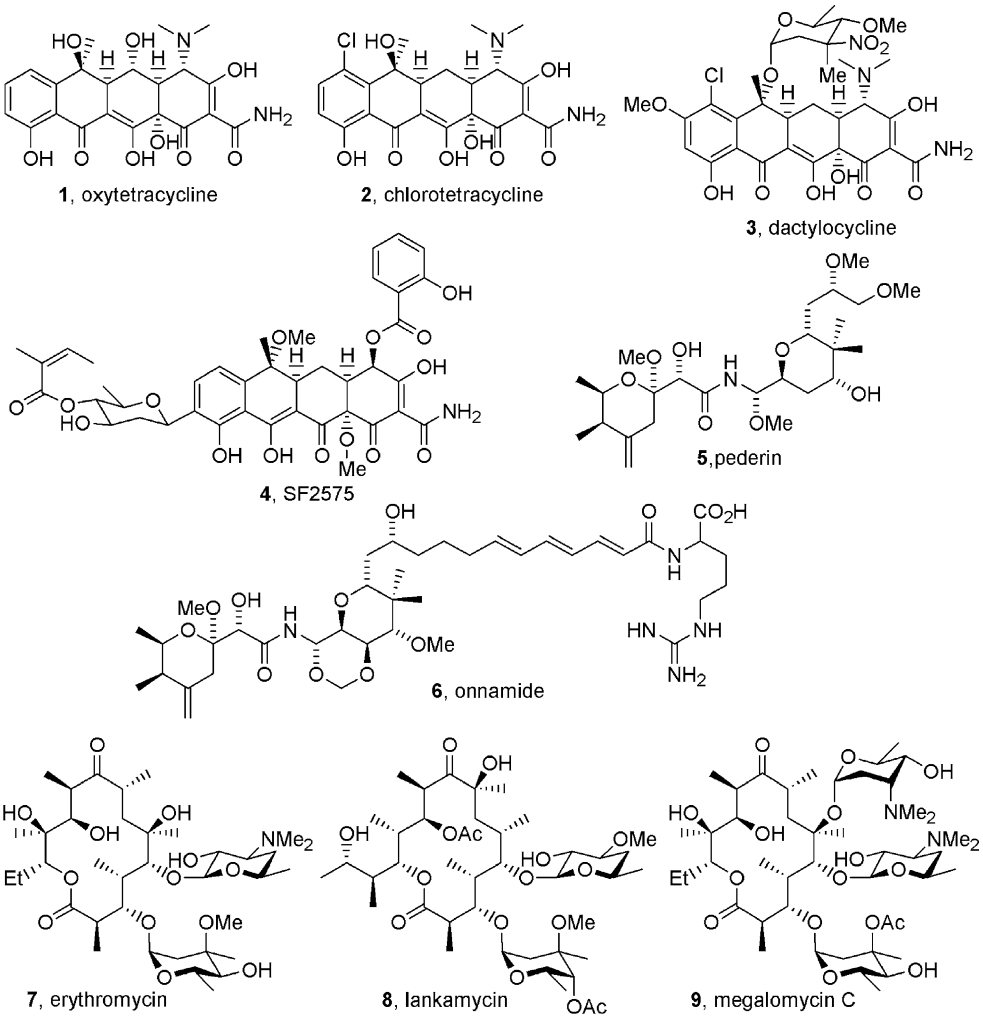

Fig. 2 The structure of polyketides from related type II (**1**–**4**), *trans*-AT (**5**,**6**), and type I (**7**–**9**) biosynthetic pathways

the query (*e*-values of 0) with greater than 90 % sequence coverage.

The ClusterBlast and Subcluster Blast tools from antiSMASH have provided a more automated approach to the identification of related gene clusters [4, 34]. The ClusterBlast algorithm sums the number of individual conserved genes between pathways with a bias to conserved core PKS and NRPS genes and sums the number of gene pairs with synteny between clusters again with a bias for core PKS and NRPS genes to quantify similarity between clusters. This tool enables rapid comparison of a new gene cluster to gene clusters deposited in the NCBI database. For example, analysis of the erythromycin biosynthetic gene cluster form *Aeromicrobium erythreum* immediately shows it to be highly related to the erythromycin gene cluster from *Saccharopolyspora erythraea* [38], the lankamycin biosynthetic gene cluster from *Streptomyces rochei* [53], and the megalomycin biosynthetic gene cluster from *Micromonospora megalomicea* [57] (Fig. 2). Subcluster Blast searches translated query sequences against the protein sequences for 126 known subclusters that encode the production of starter units like 6-methylsalicylic acid, extender units like ethylmalonyl-CoA, and sugars such as activated deoxysugars [4]. This enables rapid identification of the capacity for a particular gene cluster to produce specific chemical moieties. The automation of ClusterBlast and its graphical display of the gene cluster architecture for the query and hits make it very easy and rapid for a user to identify if a new gene cluster has a close homolog in the current NCBI database.

Phylogentic analysis of PKS and NRPS proteins is also a very powerful tool for comparison of PK and NRP biosynthetic pathways. For example, phylogenic analyses of type II PKS KS domains [35] and cyclase (CYC) domains [19, 28] show that they cluster based on the overall structure of the aromatic PK product such as angucycline, tetracycline, or anthracycline type aromatic PKs. This analysis has proven useful in the discovery of new aromatic polyketides from bacterial genomes [52] as well as metagenomic DNA [17]. Of particular interest for new natural product discovery in these phylogenic analyses are sequences that cluster away from the known groups, which suggest possible new

structural types. A large number of tools are currently available for phylogentic analysis. A particularly approachable, web-based, user friendly interface is the Phylogeny.fr platform (http://www.phylogeny.fr) [12], which provides non-experts a 'One Click' mode that links multiple sequence alignment, tree building, and tree rendering together to rapidly and with high accuracy, construct a phylogenic tree.

Predicting structure from gene cluster sequence data

De novo structure prediction from gene cluster DNA sequence data is extremely challenging. Outstanding tools have been developed to predict the function of individual catalytic domains within PK and NRP biosynthetic pathways. Using these tools, it is possible to identify reliably a PK or NRP pathway from a genome data set that is responsible for the production a known metabolite. Detailed manual prediction of structure from gene cluster data can often provide a reasonable estimate for the core structure of a PK or NRP but automated structure prediction is unreliable in most cases.

As many of the catalytic domains in PK and NRP biosynthesis have been well-characterized, it has become possible to predict their function reliably. In NRP pathways, the substrate selectivity of the A domains, which select the amino acid building blocks, can be predicted with a good degree of accuracy. Based on high-resolution structural characterization of A domains, key residues in the binding cavity responsible for substrate specificity were identified and used to build predictive models based on sequence alignment for substrate specificity [9, 49]. These models have been refined into HMMs [27, 43] and Support Vector Machines (SVMs) [47], which can be used to predict specificity with a high degree of reliability for 40–50 different substrates. While there are fewer substrates in PK pathways, prediction has been somewhat more challenging. The most common substrates for the acyltransferase (AT) domains, which are responsible for selection of the correct building block for PK biosynthesis [14], are malonyl-CoA (MCoA) and methylmalonyl-CoA (MMCoA). MCoA-selective ATs cluster apart from MMCoA-selective ATs in phylogenetic analyses with a few exceptions [25, 46, 63]. ATs with selectivity toward less common substrates such as ethylmalonyl-CoA and methoxymalonyl-CoA appear to have evolved convergently from MCoA and MMCoA ancestors leading to their clustering in both the MCoA and MMCoA selective clades [46]. Models based on key residues identified through structural studies of the substrate binding pocket [1, 33, 63] and HMM based on multiple sequence alignment [27, 36] have been successful at predicting MCoA and MMCoA specificity and moderately successful at predicting other substrates. Many of these predictive tools are easily accessible in the comprehensive bioinformatics platforms antiSMASH and NP.searcher as well as more specific tools such as NRPSpredictor2 (nrps.informatik.uni-tuebingen.de) [47], NRPSsp (http://www.nrpssp.com) [43], SBSPKS (http://www.nii.ac.in/sbspks.html) [1] and NRPS-PKS-substrate-predictor (http://www.cmbi.ru.nl/NRPS-PKS-substrate-predictor) [27].

Predicting the order of connectivity of the building blocks used by PK and NRP pathways has generally relied on collinearity. This term refers to the observation that in the vast majority of cases the order of the genes in the gene cluster corresponds to the order in which the proteins construct the PK or NRP product [8]. Therefore, the predicted connectivity of individual building blocks selected by the A and AT domains is defined by the order of the A and AT coding regions in the gene cluster. This approach is used by antiSMASH and NP.searcher to connect building blocks and assemble putative structures. However, not all PK and NRP biosynthetic pathways follow collinearity. To predict the order of PKS protein–protein interactions, which define the order of substrate addition, a bioinformatics tool has been developed to examine the N- and C-termini of these proteins to identify complementary intermolecular contacts [64] and this has been incorporated into the SBSPKS tool [1]. Core PK structure prediction from type I PKSs currently lacks good bioinformatics tools to evaluate module skipping as seen in 10-deoxymethynolide biosynthesis [3] and iterative module use as seen in stigmatellin biosynthesis [20].

A key aspect of many PKs and NRPs is the presence of stereogenic elements. Prediction of stereochemistry in NRPs is robust. A domain specificity can predict the configuration of the amino acid building blocks. The presence of an epimerization domain indicates that the stereochemistry of the amino acid building block in the peptidyl donor-intermediate will be inverted. C domains have been shown to be selective for the α-carbon stereochemistry of the acceptor amino acid and the donor peptidyl groups [11]. Phylogenetic analysis of C domains shows that they cluster based on the stereochemistry of the donor groups enabling further prediction of the configuration of the final peptide product [44]. In PKs, stereogenic alcohols are common and frequently introduced via reduction of a β-ketothioester intermediate by a ketoreductase (KR) domain. Multiple sequence alignment has identified type A and type B KRs from type I PKSs, which differ based on conserved residues in the binding cavity and produce the 3D and 3L configured alcohols, respectively [7, 45]. There is a proposed correlation between the stereogenic configuration of the beta-hydroxy thioesters and the configuration of olefins generated by dehydratase domains (DH) in type I PKSs, with the 3L alcohols generating the *E* olefins and 3D alcohols forming the much less common *Z* olefins [21, 45, 55, 60]. However, it is clear from a number of pathways that

this correlation has many exceptions [56]. There are also emerging models to predict the stereochemical outcome of enoyl reductases (ER), which generate the fully saturated acyl-thioester intermediate in type I PKS [30]. However, this model likely needs further refinement to have strong predictive value [29]. Stereogenic elements in iterative fungal and type II PKSs have proven to be much harder to predict [24, 66] and no reliable models currently exist for their prediction.

Some of the predictive models for stereochemistry have been incorporated into bioinformatics tools. For NRPs, antiSMASH, NP.searcher, and NaPDoS all provide clear predictions for amino acid α-carbon stereochemistry. AntiSMASH explicitly provides the stereochemistry of amino acid building blocks in its predictive core structure image as does NP.searcher in its SMILES output of predicted structures. NaPDoS identifies E domains and the stereoselectivity of C domains, enabling the user to predict stereochemistry. Most of the predictive models for PK stereochemistry are insufficiently reliable for incorporation into bioinformatics tools. However, Clustscan has a robust prediction tool for the stereochemistry of KR reductions in type I PKS pathways. In addition the legacy version of antiSMASH (1.0) assigned KRs to type A or B, enabling the user to predict stereochemistry. This feature, however, is absent in the most recent version of antiSMASH (2.0).

Clear challenges exist in automating bioinformatics-based structure prediction of PK and NRP structures. However, current state-of-the-art in NRP structure prediction is sufficient to provide natural products researchers with insight into genome mining projects. Using the predicted amino acid composition, it is possible to search databases such as PubChem or NORINE (bioinfo.lifl.fr/norine/), which contains the structures of over 1,100 NRPs [6]. This search is highly complementary to the sequence-based searches as it can identify NRPs analogous to the search query whose biosynthetic gene clusters have yet to be sequenced. While antiSMASH and NP.searcher have robust automated prediction tools for amino acid composition and connectivity, backbone heterocyclization and tailoring chemistries, two common events in NRP biosynthesis, are not well accounted for. To address this issue, NP.searcher has user controlled features enabling dimerization, heterocyclization, and glycosylation to be included in structure prediction, generating a diverse set of predicted products for each pathway.

Automated structure prediction for PKs is not currently effective. Detailed manual prediction can, however, be effective in predicting core structures as was seen for in the discovery of thailandamide A [37] and elansolide D [54]. Among PK pathways, structure prediction is the most advanced for the type I pathways. Prediction of products from *trans*-AT pathways is improving and is based on phylogenic analysis of the KS domains, which cluster based on the structure at the α and β carbons of the upstream growing acyl-chain [37, 54]. Prediction of type II PK structures is limited to the type of core structure based on the KS [35] and CYC [19, 28] domains; however, the length of type II PKs product and its final tailoring are very hard to predict. Lastly, little progress has been made on the prediction of fungal PK structures from their corresponding gene clusters. With advances in PK structure prediction, it may be possible in the future to access high quality structure predictions through automated methods. For now, however, development of these predictive tools is an active and exciting area of research.

Conclusions

With the wealth of minable sequence data now available to natural products chemists, a number of powerful bioinformatics tools have been developed to identify PK and NRP pathways and determine if they are likely to encode unique compounds. The most comprehensive tool currently available is antiSMASH. It possesses an array of features, including the ability to search genomes and metagenomes for core PK and NRP genes using HMMs, the ability to define the boundaries of a particular gene cluster, a search tool to identify related gene clusters in the NCBI database, and structural prediction tools to provide an rough estimate of the core structure of the NRP or PK product. A number of more specialized tools also exist for natural products researchers focusing on, for example, fungal pathways (SMURF) or aromatic polyketide pathways (PKMiner).

Because structure prediction is generally poor, the most reliable method for genome mining to find new natural products is currently a comparative genetic approach using tools such as ClusterBlast in antiSMASH. This type of tool can rapidly identify unique gene clusters from sequence data sets. However, identified unique gene clusters may not necessarily encode new chemical diversity, rather they represent PK and NRP pathways that have not yet been sequenced. A detailed manual structure prediction, guided by a number of bioinformatics tools, can be used to further test if a pathway encodes new chemical diversity. This refined unique subset represents high value pathways that likely encode new natural products. Experimental characterization of these pathways will hopefully lead to the discovery of new natural product structures.

References

1. Anand S, Prasad MVR, Yadav G et al (2010) SBSPKS: structure based sequence analysis of polyketide synthases. Nucleic Acids Res 38:W487–W496. doi:10.1093/nar/gkq340

2. Ansari MZ, Sharma J, Gokhale RS, Mohanty D (2008) In silico analysis of methyltransferase domains involved in biosynthesis of secondary metabolites. BMC Bioinformatics 9:454. doi:10.1186/1471-2105-9-454
3. Beck BJ, Yoon YJ, Reynolds KA, Sherman DH (2002) The hidden steps of domain skipping: macrolactone ring size determination in the pikromycin modular polyketide synthase. Chem Biol 9:575–583
4. Blin K, Medema MH, Kazempour D et al (2013) antiSMASH 2.0—a versatile platform for genome mining of secondary metabolite producers. Nucleic Acids Res 41:W204–W212. doi:10.1093/nar/gkt449
5. Brikun IA, Reeves AR, Cernota WH et al (2004) The erythromycin biosynthetic gene cluster of *Aeromicrobium erythreum*. J Ind Microbiol Biotechnol 31:335–344. doi:10.1007/s10295-004-0154-5
6. Caboche S, Pupin M, Leclère V et al (2008) NORINE: a database of nonribosomal peptides. Nucleic Acids Res 36:D326–D331. doi:10.1093/nar/gkm792
7. Caffrey P (2003) Conserved amino acid residues correlating with ketoreductase stereospecificity in modular polyketide synthases. ChemBioChem 4:654–657. doi:10.1002/cbic.200300581
8. Callahan B, Thattai M, Shraiman BI (2009) Emergent gene order in a model of modular polyketide synthases. Proc Natl Acad Sci USA 106:19410–19415. doi:10.1073/pnas.0902364106
9. Challis GL, Ravel J, Townsend CA (2000) Predictive, structure-based model of amino acid recognition by nonribosomal peptide synthetase adenylation domains. Chem Biol 7:211–224
10. Chooi Y-H, Wang P, Fang J et al (2012) Discovery and characterization of a group of fungal polycyclic polyketide prenyltransferases. J Am Chem Soc 134:9428–9437. doi:10.1021/ja3028636
11. Clugston SL, Sieber SA, Marahiel MA, Walsh CT (2003) Chirality of peptide bond-forming condensation domains in nonribosomal peptide synthetases: the C5 domain of tyrocidine synthetase is a (D)C(L) catalyst. Biochemistry 42:12095–12104. doi:10.1021/bi035090+
12. Dereeper A, Guignon V, Blanc G et al (2008) Phylogeny.fr: robust phylogenetic analysis for the non-specialist. Nucleic Acids Res 36:W465–W469. doi:10.1093/nar/gkn180
13. Donadio S, Monciardini P, Sosio M (2007) Polyketide synthases and nonribosomal peptide synthetases: the emerging view from bacterial genomics. Nat Prod Rep 24:1073–1109. doi:10.1039/b514050c
14. Dunn BJ, Khosla C (2013) Engineering the acyltransferase substrate specificity of assembly line polyketide synthases. J R Soc Interface 10:20130297. doi:10.1098/rsif.2013.0297
15. Eddy SR (2004) What is a hidden Markov model? Nat Biotechnol 22:1315–1316. doi:10.1038/nbt1004-1315
16. Eustáquio AS, McGlinchey RP, Liu Y et al (2009) Biosynthesis of the salinosporamide A polyketide synthase substrate chloroethylmalonyl-coenzyme A from *S*-adenosyl-L-methionine. Proc Natl Acad Sci USA 106:12295–12300. doi:10.1073/pnas.0901237106
17. Feng Z, Kallifidas D, Brady SF (2011) Functional analysis of environmental DNA-derived type II polyketide synthases reveals structurally diverse secondary metabolites. Proc Natl Acad Sci USA 108:12629–12634. doi:10.1073/pnas.1103921108
18. Finn RD, Mistry J, Tate J et al (2010) The Pfam protein families database. Nucleic Acids Res 38:D211–D222. doi:10.1093/nar/gkp985
19. Fritzsche K, Ishida K, Hertweck C (2008) Orchestration of discoid polyketide cyclization in the resistomycin pathway. J Am Chem Soc 130:8307–8316. doi:10.1021/ja800251m
20. Gaitatzis N, Silakowski B, Kunze B et al (2002) The biosynthesis of the aromatic myxobacterial electron transport inhibitor stigmatellin is directed by a novel type of modular polyketide synthase. J Biol Chem 277:13082–13090. doi:10.1074/jbc.M111738200
21. Guo X, Liu T, Valenzano CR et al (2010) Mechanism and stereospecificity of a fully saturating polyketide synthase module: nanchangmycin synthase module 2 and its dehydratase domain. J Am Chem Soc 132:14694–14696. doi:10.1021/ja1073432
22. Hertweck C (2009) The biosynthetic logic of polyketide diversity. Angew Chem Int Ed Eng 48:4688–4716. doi:10.1002/anie.200806121
23. Hur GH, Vickery CR, Burkart MD (2012) Explorations of catalytic domains in non-ribosomal peptide synthetase enzymology. Nat Prod Rep 29:1074–1098. doi:10.1039/c2np20025b
24. Javidpour P, Das A, Khosla C, Tsai S-C (2011) Structural and biochemical studies of the hedamycin type II polyketide ketoreductase (HedKR): molecular basis of stereo- and regiospecificities. Biochemistry 50:7426–7439. doi:10.1021/bi2006866
25. Jenke-Kodama H, Sandmann A, Müller R, Dittmann E (2005) Evolutionary implications of bacterial polyketide synthases. Mol Biol Evol 22:2027–2039. doi:10.1093/molbev/msi193
26. Khaldi N, Seifuddin FT, Turner G et al (2010) SMURF: Genomic mapping of fungal secondary metabolite clusters. Fungal Genet Biol 47:736–741. doi:10.1016/j.fgb.2010.06.003
27. Khayatt BI, Overmars L, Siezen RJ, Francke C (2013) Classification of the adenylation and acyl-transferase activity of NRPS and PKS systems using ensembles of substrate specific hidden Markov models. PLoS ONE 8:e62136. doi:10.1371/journal.pone.0062136
28. Kim J, Yi G-S (2012) PKMiner: a database for exploring type II polyketide synthases. BMC Microbiol 12:169. doi:10.1186/1471-2180-12-169
29. Kwan DH, Leadlay PF (2010) Mutagenesis of a modular polyketide synthase enoylreductase domain reveals insights into catalysis and stereospecificity. ACS Chem Biol 5:829–838. doi:10.1021/cb100175a
30. Kwan DH, Sun Y, Schulz F et al (2008) Prediction and manipulation of the stereochemistry of enoylreduction in modular polyketide synthases. Chem Biol 15:1231–1240. doi:10.1016/j.chembiol.2008.09.012
31. Letunic I, Doerks T, Bork P (2009) SMART 6: recent updates and new developments. Nucleic Acids Res 37:D229–D232. doi:10.1093/nar/gkn808
32. Letzel A-C, Pidot SJ, Hertweck C (2013) A genomic approach to the cryptic secondary metabolome of the anaerobic world. Nat Prod Rep 30:392–428. doi:10.1039/c2np20103h
33. Li MHT, Ung PMU, Zajkowski J et al (2009) Automated genome mining for natural products. BMC Bioinformatics 10:185. doi:10.1186/1471-2105-10-185
34. Medema MH, Blin K, Cimermancic P et al (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. doi:10.1093/nar/gkr466
35. Metsä-Ketelä M, Halo L, Munukka E et al (2002) Molecular evolution of aromatic polyketides and comparative sequence analysis of polyketide ketosynthase and 16S ribosomal DNA genes from various streptomyces species. Appl Environ Microbiol 68:4472–4479
36. Minowa Y, Araki M, Kanehisa M (2007) Comprehensive analysis of distinctive polyketide and nonribosomal peptide structural motifs encoded in microbial genomes. J Mol Biol 368:1500–1517. doi:10.1016/j.jmb.2007.02.099
37. Nguyen T, Ishida K, Jenke-Kodama H et al (2008) Exploiting the mosaic structure of trans-acyltransferase polyketide synthases for natural product discovery and pathway dissection. Nat Biotechnol 26:225–233. doi:10.1038/nbt1379
38. Oliynyk M, Samborskyy M, Lester JB et al (2007) Complete genome sequence of the erythromycin-producing bacterium *Saccharopolyspora erythraea* NRRL23338. Nat Biotechnol 25:447–453. doi:10.1038/nbt1297

39. Penn K, Jenkins C, Nett M et al (2009) Genomic islands link secondary metabolism to functional adaptation in marine Actinobacteria. ISME J 3:1193–1203. doi:10.1038/ismej.2009.58
40. Pickens LB, Kim W, Wang P et al (2009) Biochemical analysis of the biosynthetic pathway of an anticancer tetracycline SF2575. J Am Chem Soc 131:17677–17689. doi:10.1021/ja907852c
41. Piel J (2002) A polyketide synthase-peptide synthetase gene cluster from an uncultured bacterial symbiont of Paederus beetles. Proc Natl Acad Sci USA 99:14002–14007. doi:10.1073/pnas.222481399
42. Piel J, Hui D, Wen G et al (2004) Antitumor polyketide biosynthesis by an uncultivated bacterial symbiont of the marine sponge *Theonella swinhoei*. Proc Natl Acad Sci USA 101:16222–16227. doi:10.1073/pnas.0405976101
43. Prieto C, García-Estrada C, Lorenzana D, Martín JF (2012) NRPSsp: non-ribosomal peptide synthase substrate predictor. Bioinformatics 28:426–427. doi:10.1093/bioinformatics/btr659
44. Rausch C, Hoof I, Weber T et al (2007) Phylogenetic analysis of condensation domains in NRPS sheds light on their functional evolution. BMC Evol Biol 7:78. doi:10.1186/1471-2148-7-78
45. Reid R, Piagentini M, Rodriguez E et al (2003) A model of structure and catalysis for ketoreductase domains in modular polyketide synthases. Biochemistry 42:72–79. doi:10.1021/bi0268706
46. Ridley CP, Lee HY, Khosla C (2008) Evolution of polyketide synthases in bacteria. Proc Natl Acad Sci USA 105:4595–4600. doi:10.1073/pnas.0710107105
47. Röttig M, Medema MH, Blin K et al (2011) NRPSpredictor2—a web server for predicting NRPS adenylation domain specificity. Nucleic Acids Res 39:W362–W367. doi:10.1093/nar/gkr323
48. Schneiker S, Perlova O, Kaiser O et al (2007) Complete genome sequence of the myxobacterium *Sorangium cellulosum*. Nat Biotechnol 25:1281–1289. doi:10.1038/nbt1354
49. Stachelhaus T, Mootz HD, Marahiel MA (1999) The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. Chem Biol 6:493–505. doi:10.1016/S1074-5521(99)80082-9
50. Starcevic A, Zucko J, Simunkovic J et al (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. doi:10.1093/nar/gkn685
51. Stevens DC, Conway KR, Pearce N et al (2013) Alternative sigma factor over-expression enables heterologous expression of a Type II polyketide biosynthetic pathway in *Escherichia coli*. PLoS ONE 8:e64858. doi:10.1371/journal.pone.0064858
52. Sun W, Peng C, Zhao Y, Li Z (2012) Functional gene-guided discovery of type II polyketides from culturable actinomycetes associated with soft coral *Scleronephthya* sp. PLoS ONE 7:e42847. doi:10.1371/journal.pone.0042847
53. Suwa M, Sugino H, Sasaoka A et al (2000) Identification of two polyketide synthase gene clusters on the linear plasmid pSLA2-L in *Streptomyces rochei*. Gene 246:123–131
54. Teta R, Gurgui M, Helfrich EJN et al (2010) Genome mining reveals trans-AT polyketide synthase directed antibiotic biosynthesis in the bacterial phylum bacteroidetes. ChemBioChem 11:2506–2512. doi:10.1002/cbic.201000542
55. Valenzano CR, You Y-O, Garg A et al (2010) Stereospecificity of the dehydratase domain of the erythromycin polyketide synthase. J Am Chem Soc 132:14697–14699. doi:10.1021/ja107344h
56. Vergnolle O, Hahn F, Baerga-Ortiz A et al (2011) Stereoselectivity of isolated dehydratase domains of the borrelidin polyketide synthase: implications for cis double bond formation. ChemBioChem 12:1011–1014. doi:10.1002/cbic.201100011
57. Volchegursky Y, Hu Z, Katz L, McDaniel R (2000) Biosynthesis of the anti-parasitic agent megalomicin: transformation of erythromycin to megalomicin in *Saccharopolyspora erythraea*. Mol Microbiol 37:752–762
58. Wang P, Kim W, Pickens LB et al (2012) Heterologous expression and manipulation of three tetracycline biosynthetic pathways. Angew Chem Int Ed Eng 51:11136–11140. doi:10.1002/anie.201205426
59. Weber T, Rausch C, Lopez P et al (2009) CLUSEAN: a computer-based framework for the automated analysis of bacterial secondary metabolite biosynthetic gene clusters. J Biotechnol 140:13–17
60. Wu J, Zaleski TJ, Valenzano C et al (2005) Polyketide double bond biosynthesis. Mechanistic analysis of the dehydratase-containing module 2 of the picromycin/methymycin polyketide synthase. J Am Chem Soc 127:17393–17404. doi:10.1021/ja055672+
61. Wyatt MA, Ahilan Y, Argyropoulos P et al (2013) Biosynthesis of ebelactone A: isotopic tracer, advanced precursor and genetic studies reveal a thioesterase-independent cyclization to give a polyketide β-lactone. J Antibiot. doi:10.1038/ja.2013.48
62. Wyatt MA, Lee J, Ahilan Y, Magarvey NA (2013) Bioinformatic evaluation of the secondary metabolism of antistaphylococcal environmental bacterial isolates. Can J Microbiol 59:465–471. doi:10.1139/cjm-2013-0016
63. Yadav G, Gokhale RS, Mohanty D (2003) Computational approach for prediction of domain organization and substrate specificity of modular polyketide synthases. J Mol Biol 328:335–363
64. Yadav G, Gokhale RS, Mohanty D (2009) Towards prediction of metabolic products of polyketide synthases: an in silico analysis. PLoS Comput Biol 5:e1000351. doi:10.1371/journal.pcbi.1000351
65. Zhang W, Ames BD, Tsai S-C, Tang Y (2006) Engineered biosynthesis of a novel amidated polyketide, using the malonamyl-specific initiation module from the oxytetracycline polyketide synthase. Appl Environ Microbiol 72:2573–2580. doi:10.1128/AEM.72.4.2573- 2580.2006
66. Zhou H, Gao Z, Qiao K et al (2012) A fungal ketoreductase domain that displays substrate-dependent stereospecificity. Nat Chem Biol 8:331–333. doi:10.1038/nchembio.912
67. Ziemert N, Podell S, Penn K et al (2012) The natural product domain seeker NaPDoS: a phylogeny based bioinformatic tool to classify secondary metabolite gene diversity. PLoS ONE 7:e34064. doi:10.1371/journal.pone.0034064