DOI: 10.1002/cbic.200800389

Strategies for the Discovery of New Natural Products by Genome Mining

Malek Zerikly and Gregory L. Challis*[a]

Natural products have a very broad spectrum of applications. Many natural products are used clinically as antibacterial, antifungal, antiparasitic, anticancer and immunosuppressive agents and are therefore of utmost importance for our society. When in the 1940s the golden age of antibiotics was ushered in, a "gold rush fever" of natural product discovery in the pharmaceutical industry ensued for many decades. However, the traditional process of discovering new bioactive natural products is generally long and laborious, and known natural products are frequently rediscovered. A mass-withdrawal of pharmaceutical companies from new natural product discovery and natural products research has thus occurred in recent years. In this article, the concept of genome mining for novel natural product discovery, which promises to provide a myriad of new bioactive natural compounds, is summarized and discussed. Genome mining for new natural product discovery exploits the huge and constantly increasing quantity of DNA sequence data from a wide variety of organisms that is accumulating in publicly accessible databases. Genes encoding enzymes likely to be involved in natural product biosynthesis can be readily located in sequenced genomes by use of computational sequence comparison tools. This information can be exploited in a variety of ways in the search for new bioactive natural products.

1. Introduction

Although our scientific knowledge of natural products has only developed over the last 100 years or so, the practical application of antibiotics existed for a long time before the term was even known. Ancient cultures already knew that the use of certain crude plant extracts or moldy curd of cheese and soybeans had certain therapeutic benefits. Evidence for exposure to antibiotics in an archaeological population was presented by Basset and his colleagues when they showed traces of tetracyclines at therapeutic levels found in bones from the ancient Sudanese–Nubian civilization (350 AD).[1] In mediaeval Europe, different plant extracts, as well as "magic" unguents and many different blood-curdling elixirs, were used to fight infections and diseases, and some of them actually helped without there being a real understanding of what was happening.[2] Although many cultures used antibiotics, the general principles of antibiotic action were not understood until the twentieth century.

The development of modern antibiotics was led by a few key individuals who showed that material derived from microorganisms could be used to cure or protect against infectious diseases. An effective vaccine that protected against anthrax, for example, was developed by Toussaint, Greenfield, and Pasteur in 1880 and 1881.[3] This whole-cell vaccine (although not an antibiotic) demonstrated the medicinal potential of bacteria. When Sir Alexander Fleming noticed a halo of inhibition of bacterial growth around a contaminant mold on a *Staphylococcus* plate culture in 1928, he ushered in a new era. With the further development of penicillin into a medicine by Florey and Chain in the 1940s and its following mass production and success, the golden age of antibiotics had begun.[4,5] During the following 60 years, the pharmaceutical industry was deeply engaged in the search for new natural products; this search almost exclusively relied on the detection of bioactivity in extracts from natural sources and purification of the corresponding compounds. Many natural products with antibacterial, antifungal, antiparasitic, anticancer, and immunosuppressive properties were discovered.[2]

In the late 1990s, the one strain/many compounds (OSMAC) approach began to reveal nature's chemical diversity and showed that, depending on the environment and cultivation conditions, single strains of fungi or bacteria often have the potential to produce a variety of different natural products.[6] With the development and establishment of new genetic engineering techniques in recent years, the molecular basis for this phenomenon has begun to be understood.[7]

About a decade ago, various factors contributed to a decrease in interest in natural products as part of the drug discovery process. On the one hand there was very frequent rediscovery of already known natural products, technical challenges associated with their purification and structure elucidation, as well as the financial cost of new natural product discovery. On the other hand the advent of combinatorial chemistry promised to provide an overabundance of new compounds to screen for biological activity. As a result, the large pharmaceutical companies started to realign their strategies, and a mass-withdrawal from new natural product discovery and natural products research ensued.

[a] *M. Zerikly, Prof. G. L. Challis*
Department of Chemistry, University of Warwick
Coventry CV4 7AL (UK)
Fax: (+44) 24-7652-4112
E-mail: g.l.challis@warwick.ac.uk

Genomics has resulted in the deposition of a huge quantity of DNA sequence data from a wide variety of organisms in public databases. At present, there are more than 1150 complete microbial genome sequences (http://www.ncbi.nlm.nih.gov/sutils/genom_table.cgi). This large quantity of publicly accessible DNA data has paved the way for exploitation of new bioinformatics tools for the discovery of novel natural products. This endeavor has been underpinned by substantial and continuous increases in the understanding of the genetics and enzymology of natural product biosynthesis made throughout the 1990s,[8] which has greatly facilitated the identification and analysis of gene clusters likely to encode natural product biosynthetic pathways in sequenced genomes. These advances allow the OSMAC approach to be applied to the detection of metabolites—produced only under certain conditions—that are postulated as the final products of novel secondary metabolite gene clusters discovered by genomics. However, the recent advances in genomics have also stimulated the development of a variety of new approaches for the identification of novel natural products, including several elegant methods not available in the pre-genomic era.[9–13]

Many microbial natural products such as polyketides and nonribosomal peptides are biosynthesized by modular megasynthases and synthetases.[8] It has been shown that in many cases the number of modules in such an assembly line corresponds to the number of metabolic building blocks incorporated into the final product, although several exceptions to this rule have been observed recently.[14] The presence or absence of domains with tailoring functions in modules allows, in many cases, the way in which a building block gets modified during the process of its incorporation into a natural product to be predicted.[8] Furthermore, substrate specificity can be predicted by sequence analysis of particular domains within each module.[15–18] The acyl transferase (AT) domains of polyketides synthases (PKSs) and, analogously, the adenylation (A) domains of nonribosomal peptide synthetases (NRPSs) are responsible for recognition of the correct substrates from among the many molecules in the cellular pool.[8] The AT and A domains catalyze tethering of their substrates as thioesters to the adjacent acyl carrier protein (ACP) domains or peptidyl carrier protein (PCP) domains in PKSs and NRPSs, respectively.[8]

Several groups have applied sequence analysis tools to the prediction of novel metabolic products of NRPSs and PKSs over the last eight years. Actinobacteria, in particular those belonging to the *Streptomyces* genus, are known to be prolific antibiotic producers. They produce the majority of clinically used antibiotics of natural origin, such as streptomycin, tetracycline, neomycin, or chloramphenicol. One particular strain, which has been studied extensively and is often regarded as a model streptomycete, is *Streptomyces coelicolor* A3(2).[19] In 2001, the *S. coelicolor* genome sequencing project was completed.[20] It was one of the first sequenced microbes that was shown to have more gene clusters encoding biosynthetic pathways for natural products than there were known natural products of the organism. These gene clusters encoding biosynthetic systems not associated with production of a known metabolite are often referred to as "cryptic." Previously unknown natural products of *S. coelicolor*, including several novel compounds, have been discovered as products of such cryptic biosynthetic gene clusters.[21–25] New natural products have also been discovered in several other sequenced organisms by genome mining; one example is the novel triterpenoid thalianol in *Arabidopsis thaliana*, a model plant.[26] These findings established genome mining as a powerful new tool for novel natural product discovery. They also demonstrated that the reservoir of novel natural products has not yet been used up and that the traditional screening-based strategies for bioactive natural product discovery can be usefully augmented and extended by the genome mining concept. Contrary to the apparent perception of many large pharmaceutical companies, numerous novel natural compounds with structural complexities and bioactivities that are unlikely to be achieved by combinatorial chemistry—on which pharmaceutical companies nowadays mostly rely for discovery of new lead compounds—remain to be discovered.

This concepts article discusses recently developed strategies for the discovery of new natural products by genome mining and illustrates them with important recent examples. Genome mining has resulted in the discovery of many novel natural products: the reader is directed to other sources for a comprehensive list of these.[27]

The first approach presented can be applied when an insight into the structural features of the metabolic products of a biosynthetic assembly line encoded by a cryptic gene cluster can be derived from bioinformatics analyses. These insights can lead to the prediction of putative physicochemical properties, which can then in turn be used to target only metabolites with the predicted properties; this facilitates isolation and characterization.

For cryptic gene clusters in which the substrates of the encoded biosynthetic enzymes can be predicted, two other approaches are discussed. The first one, the "genomisotopic approach," combines initial genomic sequence analysis with an isotope-incorporation-guided fractionation strategy. The second approach relies on in vitro analysis of the reaction catalyzed by purified enzymes when incubated with the predicted precursors.

If the substrate specificities of a cryptic biosynthetic system cannot be predicted with sufficient accuracy and confidence, more generic approaches are required. Two related strategies that have already proven successful in this area are the combination of biosynthetic gene knockouts with comparative metabolic profiling and heterologous biosynthetic gene expression with comparative metabolic profiling. The former involves inactivation of a biosynthetic gene presumed to be essential within the cryptic biosynthetic gene cluster, followed by chromatographic comparison of the metabolites extracted from the wild-type organism and the non-producing mutant. The metabolites present in the wild-type but missing in the mutant are likely to be products of the cryptic gene cluster and can be isolated and structurally characterized. The latter approach involves cloning and expressing the entire biosynthetic gene cluster in a heterologous host. Again, with the appropriate analytical techniques the metabolic profiles in culture superna-

tants or extracts are compared. Metabolites lacking in the wild-type host but present in the heterologous host containing the cloned cryptic biosynthetic gene cluster are likely to be products of the cryptic biosynthetic pathway.

All the above approaches combine an elaborate and refined strategy to discover, isolate, and characterize novel natural products. All of them (with the exception of the in vitro reconstitution approach), however, would fail to yield new compounds if the cryptic biosynthetic gene cluster were not expressed (under typical growth conditions) in the organism in question. Thus, in addition to empirical manipulation of culture conditions, three rational approaches that exploit knowledge in the field of regulation of biosynthetic gene expression to solve the above problem are also presented.

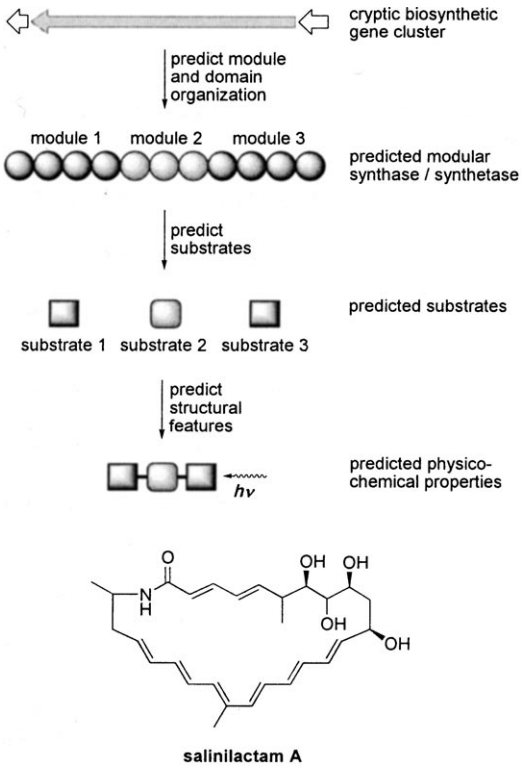
2. Identification of the Products of Cryptic Biosynthetic Gene Clusters by Prediction of Physicochemical Properties

As mentioned above, many microbial natural products such as polyketides and nonribosomal peptides are biosynthesized by modular mega-synthases and synthetases. The principles underlying the relationship between the organization of the multienzymes and the structures of their metabolic products are now quite well understood,[8] and models to predict the substrate specificities of domains responsible for building block selection in each module of nonribosomal peptide synthetases (NRPSs) and polyketide synthases (PKSs)[15–18] have been developed.

Applying such bioinformatics tools to cryptic biosynthetic systems can give important insight into the structural features of the metabolic product(s). Such structural information can lead to the prediction of putative physicochemical properties, which allows only metabolites with the predicted properties to be targeted. This can greatly simplify the analytical challenge of identifying the metabolic products of cryptic biosynthetic gene clusters (Scheme 1).

Several novel metabolic products of cryptic biosynthetic gene clusters have been discovered by these techniques. One recent example is salinilactam (Scheme 1), which was discovered by mining the recently sequenced genome of the marine actinomycete *Salinispora tropica*.[28] Sequence analysis of a novel modular PKS system encoded within the genome suggested it was responsible for assembly of a lysine-primed polyene macrolactam. Polyenes absorb UV light at characteristic wavelengths. This predicted physicochemical property was used to guide the isolation of several related polyene macrolactams, one of which was salinilactam A (Scheme 1). The structural proposal for salinilactam A was developed by a combination of detailed predictive sequence analyses of the PKS with NMR and MS analyses.

Whereas the fact that genetic manipulation is not needed in order to identify the putative metabolic product(s) of cryptic biosynthetic gene clusters in this approach may seem advantageous, such manipulations will ultimately be required to confirm that the identified metabolite(s) is/are indeed the product(s) of the gene cluster. Furthermore, the prediction of a compound's structural features must be done with a high degree of confidence, because the corresponding physicochemical properties hugely depend on it. It is important also to bear in mind that the predicted physicochemical properties of the putative metabolic product of a cryptic biosynthetic gene cluster might be similar to those of other common natural products, thus raising the possibility of being misled by this approach.

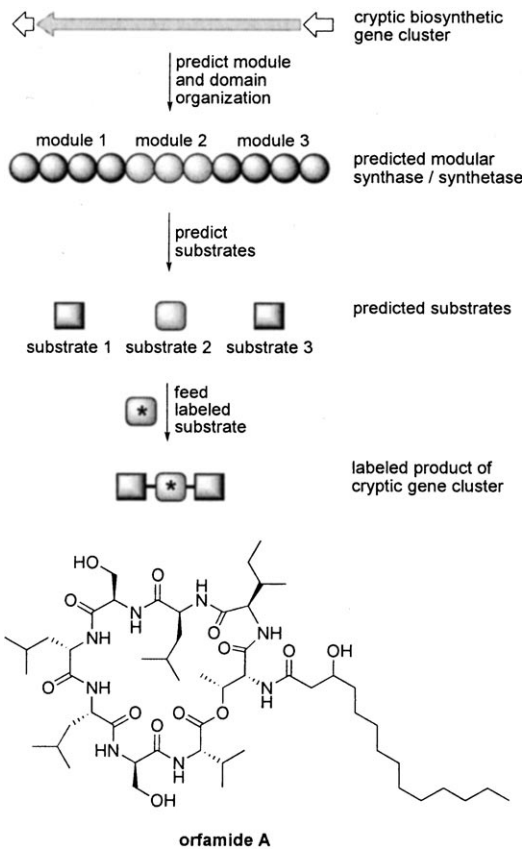


Scheme 1. Principles underlying the prediction of the physicochemical properties of a putative metabolic product (or products) of a cryptic biosynthetic gene cluster and the structure of salinilactam A, a novel natural product identified by application of this approach.

Despite these caveats, this method is a powerful tool. The particular strength of this approach lies in its rapidity and in its ability to significantly focus the search for metabolic products of cryptic biosynthetic gene clusters; this allows large numbers of culture extracts to be analyzed and many different growth conditions, as well as extraction methods and analysis techniques, to be tested with relatively little effort, and potentially allows the problems connected with low gene expression and silent gene clusters to be bypassed in an empirical manner.

3. The Genomisotopic Approach

Another approach to the discovery of natural products by genome mining has been dubbed the "genomisotopic approach" by its authors.[29] This tool, reported by William Gerwick and co-workers in 2007, works by combining predictions of substrate specificity for cryptic biosynthetic pathways discovered from genomic data with the exploitation of stable-isotope-labeled precursor incorporation (Scheme 2). A gene clus-



Scheme 2. The "genomisotopic" approach for identifying the putative metabolic product(s) of a cryptic biosynthetic gene cluster and the structure of orfamide A, a new natural product identified by this approach.

ter encoding an NRPS system was identified in the *Pseudomonas fluorescens* Pf-5 genome, but its "encoded" natural product was unknown. The gene cluster was analyzed with bioinformatics tools, and the substrate specificities of the domains of each module of the NRPS were predicted. This led to the hypothesis that the product of the cryptic gene cluster should be a novel lipopeptide containing four leucine residues. The strategy used to identify and purify this putative new compound from *P. fluorescens* involved the feeding of cultures with ^{15}N-labeled leucine, which as a predicted precursor should be incorporated into the lipopeptide. ^{1}H–^{15}N HMBC NMR experiments were used to identify metabolites containing the labeled leucine in culture extracts and fractions from chromatographic separations of the extract. This led to the isolation and structure elucidation of the novel metabolite orfamide A as the major product of this cryptic NRPS system (Scheme 2).

The example of the discovery of orfamide A indicates that the genomisotopic approach may be a valuable tool for new natural product discovery. In the broader context, the genomisotopic approach has the potential to be a quite general tool allowing the isolation not only of novel nonribosomal peptides but also of natural products belonging to other structural classes for which reliable predictions of biosynthetic enzyme substrate specificity can be made. Its strength lies in its potential for rapid isolation of new compounds without the need for any genetic manipulation of the producing organism, which in some cases can be very difficult.

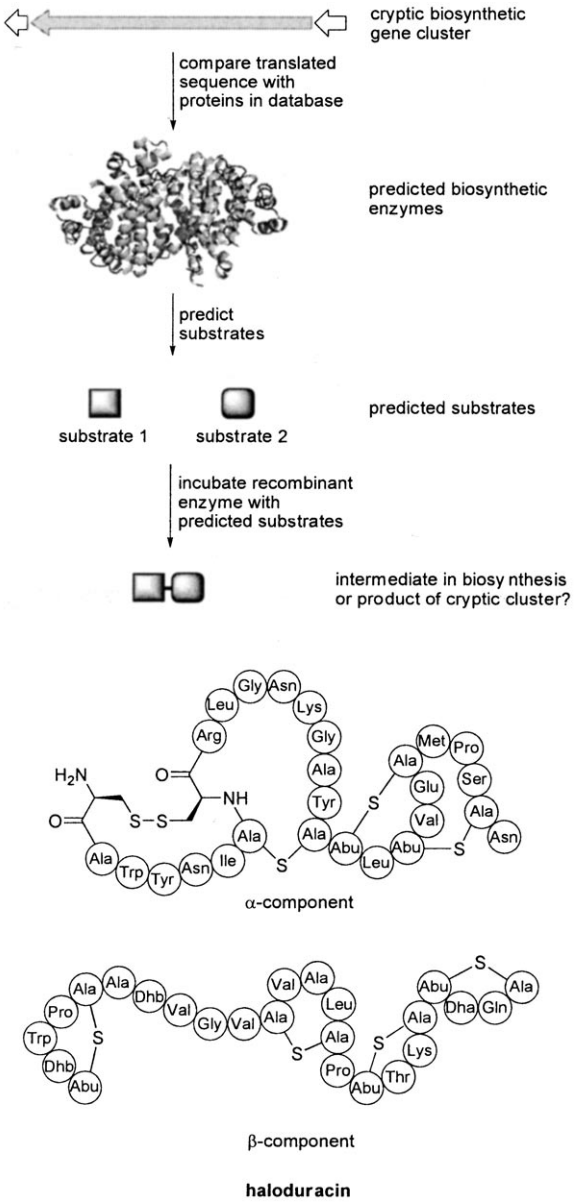
When the substrate specificity of the cryptic biosynthetic system is harder to predict, it becomes more difficult to identify appropriate labeled precursors that can be fed and used for NMR-guided isolation experiments. Besides the accuracy with which the substrate prediction from DNA sequencing data can be made, another limiting factor can be the availability of appropriate labeled precursors. Furthermore, for best results the fed substrate concentration and time of feeding have to be empirically optimized and the incorporation must occur to a reasonably high level in a direct and specific way. Incorporation of labeled precursors into other metabolites unrelated to the cryptic biosynthetic gene cluster of interest could cause the researcher to chase a red herring.

The potential disadvantages of the genomisotopic approach, as discussed above, makes other genome mining tools such as the heterologous gene expression or gene knock-out approach combined with comparative metabolic profiling potentially more general and powerful. As with the prediction of physicochemical properties approach, the linkage between novel compounds isolated by the genomisotopic approach and the cryptic gene cluster proposed to direct their biosynthesis needs to be established. This can be done, for example, by a gene knock-out experiment to inactivate a biosynthetic gene and abrogate metabolic production, or by transferring the gene cluster to a heterologous host and demonstrating that the metabolite is produced. Thus, genetic manipulations are ultimately needed.

Continuing developments in the field of bioinformatics and NMR spectroscopy should help evolution of the potential of the genomisotopic tool, which is currently limited in some respects, as discussed above, but nevertheless represents a valuable and rapid genome mining approach for the identification and isolation of novel natural products.

4. The In Vitro Reconstitution Approach

As its name implies, in the in vitro reconstitution approach one or more biosynthetic enzymes encoded by the cryptic gene cluster of interest is/are produced in pure recombinant form and incubated with the predicted substrate(s) to investigate what products are formed (Scheme 3). This approach was applied to the elucidation of the metabolic products of a cryptic putative antibiotic biosynthetic system discovered by analysis of the *Bacillus halodurans* C-125 genome sequence.[30] The sequences encoding the putative biosynthetic enzymes and their ribosomally synthesized substrates were overexpressed in *E. coli* and the resulting recombinant proteins were purified, mixed, and incubated with the appropriate co-factors; this led to the formation of the two-component lantibiotic haloduracin (Scheme 3).[30] The α- and β-components of this antibiotic complex were also identified as metabolic products of *B. halodurans* C-125. The structures of these compounds were proposed on the basis of a combination of bioinformatics, chemical modification, and MS analyses.



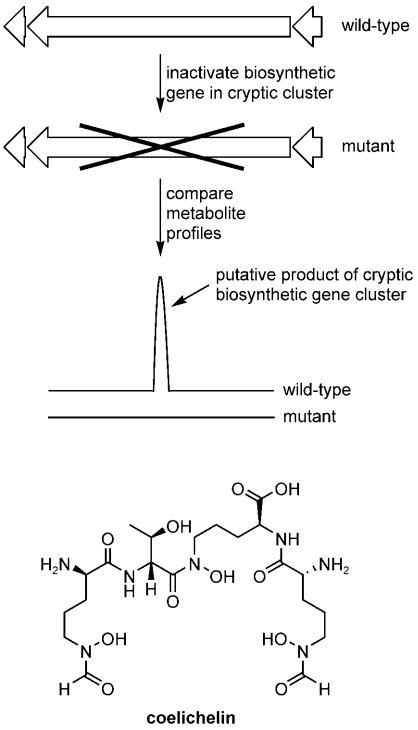
Scheme 3. Outline of the in vitro reconstitution approach for identifying the metabolic products of cryptic biosynthetic enzymes, together with the structure of haloduracin, a novel two-component antibiotic identified as a metabolic product of *B. halodurans* C-125 by this approach.

A problem commonly encountered in the other methods described in this article—that a cryptic biosynthetic gene cluster is not expressed in the wild-type organism (or in a heterologous host)—does not affect the in vitro reconstitution approach, which removes the biosynthetic genes from their natural regulatory context and expresses them under the control of a heterologous (usually inducible) promoter. On the other hand, using this approach to identify the product of a complex biosynthetic pathway would usually require separate overexpression of each biosynthetic gene and purification of the resulting overproduced protein, so the discovery of a fully elaborated metabolic product by this approach is likely to be very laborious.

The in vitro reconstitution approach represents a powerful approach for defining the functions of genes and potentially for discovering novel natural products, but may only find widespread application for relatively simple cryptic biosynthetic pathways, because other approaches require less effort and offer the potential for more rapid results.

5. The Biosynthetic Gene Inactivation/ Comparative Metabolic Profiling Approach

Comparison of metabolite profiles in wild-type and mutant strains has been widely used to identify genes that direct the production of known natural products. For new natural product discovery by genome mining this approach can be turned on its head. Genes within a cryptic biosynthetic gene cluster can be inactivated, and the profiles of metabolites produced by the wild-type and the mutant strains can be compared by an appropriate analytical technique such as HPLC or LC-MS to identify the product of the cryptic gene cluster (Scheme 4). The discovery of the novel *S. coelicolor* nonribosomal peptide coelichelin was one of the first examples in which this approach was applied successfully (Scheme 4).[21,23]



Scheme 4. Principles of the gene knockout/comparative metabolic profiling strategy, together with the structure of coelichelin, a new nonribosomal peptide discovered by this strategy in combination with the prediction of physicochemical properties approach.

In 2000, the discovery of the *cch* gene cluster in the partially completed genomic DNA sequence of *S. coelicolor* A3(2) was reported, and it was predicted that this gene cluster encoded a cryptic nonribosomal peptide synthetase (NRPS).[21] The substrates of the putative NRPS were predicted by sequence anal-

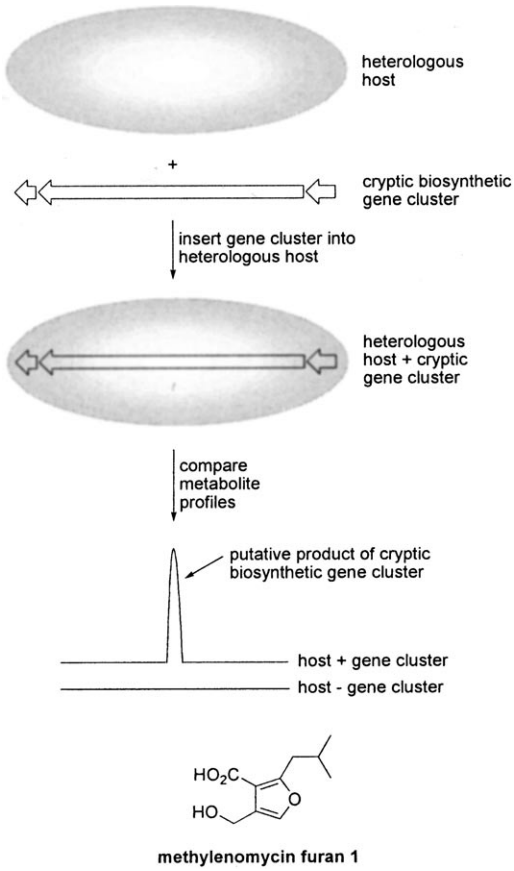
ysis, and it was proposed that this NRPS system catalyzes the assembly of a new ferric-iron-chelating tripeptide, which was called coelichelin.[21]

The predicted hydroxamic acid functional groups of coelichelin suggested that it functions in *S. coelicolor* as an iron-chelating siderophore. This provided information of vital importance for its identification: it indicated that for the *cch* cluster to be expressed, an iron-deficient medium would be required, and it also suggested that coelichelin should be easily identifiable by its conversion into the corresponding ferric-hydroxamate complex, which should absorb light at predictable wavelengths because of ligand-to-metal charge-transfer transitions. The predicted physicochemical properties of coelichelin and growth conditions to enhance expression of the *cch* gene cluster were exploited in comparative metabolic profiling experiments of wild-type *S. coelicolor* and a mutant in which the NRPS-encoding *cchH* gene had been inactivated. This strategy led to the identification of a tris-hydroxamate tetrapeptide as the product of the *cch* gene cluster, providing the first example of a quite unexpected biosynthetic phenomenon: the assembly of a tetrapeptide by an NRPS containing only three modules.[23]

In general, the gene knock-out/comparative metabolic profiling approach provides a powerful tool for discovery of novel natural products by genome mining. Structural predictions for the metabolic products of cryptic biosynthetic gene clusters are not required for this method to be effective. However, if such predictions are available, the approach can be combined with predictions of physicochemical properties or the genomisotopic approach to increase its power and utility further. Nevertheless, its success depends on various factors. One potential obstacle is the construction of metabolite non-producing mutants. For some organisms, limited or no genetic tools are available. This makes the construction of knockout mutants difficult or impossible. Additionally, the change in metabolite profile between the wild-type strain and the mutant has to be large enough to allow it to be easily identified.

6. The Heterologous Gene Expression/ Comparative Metabolic Profiling Approach

An approach similar to that described in the previous section involves heterologous expression of the cryptic biosynthetic gene cluster in combination with comparative metabolic profiling. Like the gene knock-out/comparative metabolic profiling approach, this strategy is especially useful when the substrate specificity of a cryptic biosynthetic system cannot be predicted. One big advantage of this approach in relation to the in vitro reconstitution approach is that purified biosynthetic enzymes encoded by the cryptic gene cluster are not required; this saves a lot of time and effort. The metabolite profiles in the culture supernatants or extracts obtained from the wild-type host and from the host containing the cloned cryptic biosynthetic gene cluster are compared by an appropriate analytical technique such as LC-MS. Metabolites absent in the wild-type host but present in the host containing the cloned cryptic biosynthetic gene cluster are likely to be products of the cryptic biosynthetic pathway (Scheme 5). Compounds identified by this analysis can be purified and structurally characterized. An interesting recent example of a family of novel compounds discovered by this approach is the methylenomycin furans, which are the products of a cryptic three-gene cluster imbedded within the *S. coelicolor* methylenomycin biosynthetic gene cluster (Scheme 5). These compounds were shown to act as inducers of methylenomycin production and are likely to be the first in a new family of autoinducers used by *Streptomyces* species to regulate antibiotic production.[25]

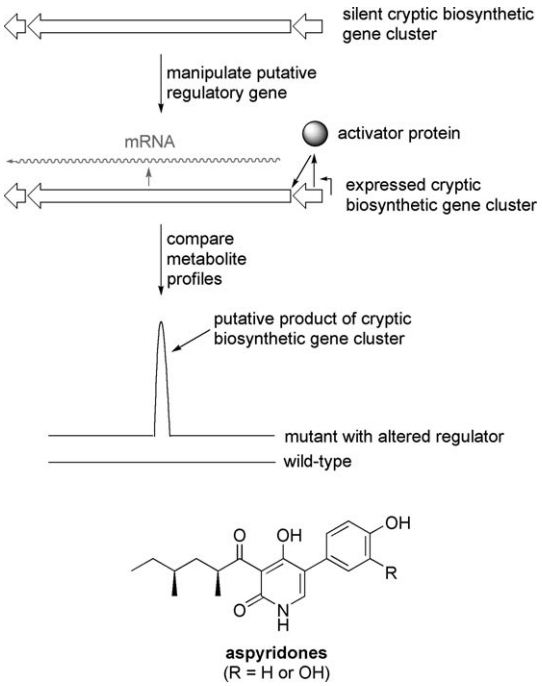


Scheme 5. The heterologous expression approach, together with the structure of methylenomycin furan 1, which is representative of a novel family of autoinducers of antibiotic biosynthesis discovered in *S. coelicolor* by this approach.

This method can be used to investigate cryptic biosynthetic gene clusters from a wide range of organisms that direct the production of a variety of natural product classes. One potential drawback of this approach is the possibility that the natural product biosynthetic gene clusters are too large (typically >40 kb) to be cloned easily into a single vector (for example, a cosmid vector). A homologous recombination approach to overcome this generic problem was pioneered by Hopwood and co-workers in the 1980s.[31] They used DNA derived from two clones containing different but overlapping fragments of the *red* cluster, which directs undecylprodiginine and streptorubin B biosynthesis in *S. coelicolor*, to "retrieve" a single clone containing the entire gene cluster by homologous recombination with the *S. coelicolor* chromosome. The resulting plasmid

was introduced into the heterologous host *Streptomyces parvulus*; this resulted in production of red-pigmented prodiginine antibiotics and indicated that the entire gene cluster had been cloned. Recently, this approach has been extended by Müller and co-workers to the assembly of clones containing an entire biosynthetic gene cluster from two cosmid clones containing different partial fragments of the gene cluster.[32] However, the largest gene cluster thus far assembled by such technology is ~60 kb. Many natural product biosynthetic gene clusters (especially NRPSs and modular PKSs) are significantly bigger than this (for example, the 128 kb daptomycin biosynthetic gene cluster). The ability of homologous recombination approaches to assemble entire large biosynthetic gene clusters from cosmids or clones with larger inserts (such as bacterial artificial chromosomes, BACs) containing gene cluster fragments remains to be demonstrated.

Expressing a cryptic biosynthetic gene cluster in a well-studied host with a well-characterized metabolic profile can be a significant advantage. Growth conditions to maximize expression of the cryptic gene cluster can be easier to optimize and comparison of metabolic profiles can be less difficult, especially if the heterologous host produces fewer metabolites than the original organism in which the cryptic biosynthetic gene cluster was identified.



Scheme 6. Steps involved in the identification of metabolic products of silent cryptic biosynthetic gene clusters by the manipulation of a pathway-specific activator. Aspyridones are novel products of a hybrid PKS-NRPS discovered by this strategy.

7. Activation of Silent Cryptic Biosynthetic Gene Clusters by Manipulation of Regulatory Genes and Chromatin Remodeling

None of the above strategies for identifying novel metabolic products of cryptic biosynthetic gene clusters is likely to be effective if the gene cluster of interest is poorly expressed or not expressed at all. As a result, the organism (or heterologous host) containing the cluster of interest does not produce the corresponding metabolite in sufficient quantities for detection, isolation, and characterization.

Whether a cryptic biosynthetic gene cluster is expressed in the first place can be rapidly established by using the reverse transcription polymerase chain reaction (RT-PCR). This method is highly sensitive, allowing a very low number of RNA molecules to be detected. The abundance of specific RNA molecules within cells can be qualitatively determined and used as an indication of gene expression. Real-time PCR or northern blot analysis can be used to compare levels of gene expression in a more quantitative manner.

Environmental factors—that is, different growth conditions—can lead to silent gene clusters being expressed in, for example, the coelichelin example discussed above. Empirical variation of growth conditions is thus one method for activation of silent cryptic biosynthetic gene clusters.

An alternative approach involves the manipulation of regulatory genes to activate the expression of silent gene clusters (Scheme 6). Two examples have been reported. The first involved manipulation of pleiotropic regulators of secondary metabolism, whereas the second involved overexpression of pathway-specific activator genes.[33,34] Both of these approaches were investigated in *Aspergillus nidulans*.

LaeA, a nuclear protein, has been shown to function as a global regulator of secondary metabolism in the *Aspergillus* genus. Deletion of *laeA* blocks the expression of some secondary metabolic gene clusters, including a variety of well known gene clusters such as the one that directs production of penicillin.[33] In contrast, overexpression of *laeA* triggers increased expression of certain secondary metabolic gene clusters, resulting in increased transcription of, for example, the penicillin gene cluster and consequently increased product formation.[33] This method has strong potential as a strategy for activating silent cryptic biosynthetic gene clusters in *Aspergillus* species but it has yet to be demonstrated by the discovery of a novel *Aspergillus* natural product through awakening of the expression of a silent biosynthetic gene cluster. As with some of the approaches described above, this approach also involves comparative metabolite profiling. The wild-type metabolite profile can be compared to those of mutants with the pleiotropic regulator of secondary metabolism—that is, *laeA*—either deleted or overexpressed. The differences in the metabolic profiles should allow potential products of gene clusters that have had their expression activated to be identified.

In the second approach, putative pathway-specific activator genes from within silent cryptic biosynthetic gene clusters were expressed under the control of an inducible promoter. Upon addition of the inducer a gene cluster that would normally be silent was expressed. On use of this method in *Aspergillus nidulans,* aspyridones—novel products of a cryptic hybrid PKS-NRPS system—were discovered (Scheme 6). These metabolic products were identified by comparative metabolic profiling of the wild-type and mutant strains.[34]

The example of the discovery of aspyridones by manipulation of pathway-specific regulatory genes and the example of manipulation of the *laeA* pleiotropic regulator of secondary metabolism show the potential of this approach for activation of silent cryptic biosynthetic gene clusters and discovery of novel natural products. This approach offers a potentially much faster and more rational way to activate or optimize metabolite production than the empirical OSMAC approach. However, it is not without potential drawbacks. It can only be effective if appropriate candidate regulators of cryptic biosynthetic gene clusters can be identified and if the host of the cryptic cluster can be genetically manipulated. Although this approach promises to be a powerful tool for identifying novel metabolic products of silent cryptic biosynthetic gene clusters, its potential for new natural product discovery in the broader context (that is, beyond *Aspergilli*) has yet to be demonstrated.

In a related recent approach, DNA methyl transferase and histone deacetylase inhibitors were used to examine the effect on secondary metabolite production in fungi, in the expectation that chromatin remodeling induced by these compounds might result in the activation of silent biosynthetic gene clusters. Comparative metabolic profiling of a *Diatrype* sp. treated and not treated with the DNA methyltransferase inhibitor 5-azacytidine led to the discovery of a novel family of glycosylated polyketides.[35]

Promising approaches for the activation of silent biosynthetic gene clusters in *Streptomyces* species have also recently been developed. One of these involves addition of *N*-acetylglucosamine to minimal media or deletion of the *dasR* gene that encodes an *N*-acetylglucosamine-responsive regulatory protein, both of which appear to induce antibiotic production.[36] Another involves selection for drug resistance mutations, which result in mutated ribosomes with altered protein and ppGpp biosynthesis activities.[37] Such mutants also appear to overproduce antibiotics. However, none of these approaches has thus far been applied to the discovery of novel metabolic products through activation of silent cryptic biosynthetic gene clusters.

8. Conclusion

As we have demonstrated with the few selected examples discussed in this article, genome mining is a powerful approach for new natural product discovery. It offers several advantages over traditional approaches. For example, the ability to predict structural features of cryptic biosynthetic gene clusters greatly facilitates "dereplication" to avoid rediscovery of already known compounds or compound classes and significantly accelerates the process of spectroscopic structure elucidation of novel metabolites. Additionally, the coupling of genome sequencing with transcriptional analyses and genetic manipulation of regulatory genes allows semi-rational and rational approaches for maximizing the diversity of bioactive metabolites produced by a given strain to be used instead of the traditionally employed empirical approach. Furthermore, genome mining links the discovery of novel natural products directly with the genes that direct their biosynthesis; this thus provides all the information required to produce analogues of the natural product by combinatorial biosynthesis[38] for structure–activity relationship (SAR) and absorption-distribution-metabolism-excretion (ADME) investigations.

The central prerequisite for genome mining as an approach for new natural product discovery is obviously the availability of a complete genome sequence or genomic sequence data. While the acquisition of such data used to be expensive and time-consuming, the continuing, rapid advances in DNA sequencing technology mean that this is no longer the case. It is conceivable that, in five years time, an instrument capable of delivering high-coverage sequence data for an average microbial genome within 24 h will be standard in many research laboratories. Literally thousands of cryptic natural product biosynthetic gene clusters should thus become available, making the discovery of the metabolic products of these clusters the bottleneck.

The various strategies for identifying the metabolic products of cryptic biosynthetic gene clusters have different strengths and weaknesses, depending on the accuracy with which the structures of the products can be predicted from bioinformatics analyses, the physicochemical properties of the natural product, the size of the gene cluster, and whether it is expressed in laboratory cultures. When choosing the best approach, these various factors need to be carefully considered.

Genome mining for new natural product discovery is far from being a mature science. Doubtless, several further strategies for discovering the metabolic products of cryptic biosynthetic gene clusters will be added to the already impressive and versatile pool of approaches available. Nevertheless, the potential of this approach has clearly been demonstrated. With further exploration and development, it will likely become an integral part of natural product discovery in the majority of industrial and academic laboratories.

Acknowledgements

Research on genome mining in the authors' laboratory is supported by grants from the BBSRC (Ref. Nos. BB/E008003/1, EGH16081, and B16610) and the EU (Contract Nos. MEIF-CT-2003-501686 and 005224).

Keywords: biosynthesis · genomics · natural products · peptides · polyketides

[1] E. J. Bassett, M. S. Keith, G. J. Akmelagos, D. L. Martin, A. R. Villanueva, *Science* **1980**, *209*, 1532–1534.
[2] D. J. Newman, G. M. Cragg, K. M. Snader, *Nat. Prod. Rep.* **2000**, *17*, 215–234.
[3] A. Scorpio, T. E. Blank, W. A. Day, D. J. Chabot, *Cell. Mol. Life Sci.* **2006**, *63*, 2237–2248.
[4] A. Fleming, *Br. J. Exp. Pathol.* **1929**, *10*, 226–236.
[5] E. Chain, H. W. Florey, A. D. Gardner, N. G. Heatley, M. A. Jennings, J. Orr-Ewing, A. G. Sanders, *Lancet* **1940**, *236*, 226–228.
[6] H. B. Bode, B. Bethe, R. Hofs, A. Zeeck, *ChemBioChem* **2002**, *3*, 619–627.
[7] S. J. Rokem, A. E. Lantz, J. Nielsen, *Nat. Prod. Rep.* **2007**, *24*, 1262–1287.
[8] M. A. Fischbach, C. T. Walsh, *Chem. Rev.* **2006**, *106*, 3468–3496.
[9] C. Corre, G. L. Challis, *Chem. Biol.* **2007**, *14*, 7–9.
[10] H. B. Bode, R. Müller, *Angew. Chem.* **2005**, *117*, 6988–7007; *Angew. Chem. Int. Ed.* **2005**, *44*, 6828–6846.

[11] B. Wilkinson, J. Micklefield, *Nat. Chem. Biol.* **2007**, *3*, 379–386.
[12] G. L. Challis, *Microbiology* **2008**, *154*, 1555–1569.
[13] H. Gross, *Appl. Microbiol. Biotechnol.* **2007**, *75*, 267–277.
[14] S. W. Haynes, G. L. Challis, *Curr. Opin. Drug Discovery Dev.* **2007**, *10*, 203–218.
[15] S. F. Haydock, J. F. Aparicio, I. Molnar, T. Schwecke, L. E. Khaw, A. Konig, A. F. Marsden, I. S. Galloway, J. Staunton, P. F. Leadlay, *FEBS Lett.* **1995**, *374*, 246–248.
[16] T. Stachelhaus, H. D. Mootz, M. A. Marahiel, *Chem. Biol.* **1999**, *6*, 493–505.
[17] G. L. Challis, J. Ravel, C. A. Townsend, *Chem. Biol.* **2000**, *7*, 211–224.
[18] T. Nguyen, K. Ishida, H. Jenke-Kodama, E. Dittmann, C. Gurgui, T. Hochmuth, S. Taudien, M. Platzer, C. Hertweck, J. Piel, *Nat. Biotechnol.* **2008**, *26*, 225–233.
[19] D. A. Hopwood, *Microbiology* **1999**, *145*, 2183–2202.
[20] S. D. Bentley, K. F. Chater, A.-M. Cerdeno-Tarraga, G. L. Challis, N. R. Thomson, K. D. James, D. E. Harris, M. A. Quail, H. Kieser, D. Harper, A. Bateman, S. Brown, G. Chandra, C. W. Chen, M. Collins, A. Cronin, A. Fraser, A. Goble, J. Hidalgo, T. Hornsby, S. Howarth, C.-H. Huang, T. Kieser, L. Larke, L. Murphy, K. Oliver, S. O'Neil, E. Rabbinowitsch, M.-A. Rajandream, K. Rutherford, S. Rutter, K. Seeger, D. Saunders, S. Sharp, R. Squares, S. Squares, K. Taylor, T. Warren, A. Wietzorrek, J. Woodward, B. G. Barrell, J. Parkhill, D. A. Hopwood, *Nature* **2002**, *417*, 141–147.
[21] G. L. Challis, J. Ravel, *FEMS Microbiol. Lett.* **2000**, *187*, 111–114.
[22] B. Gust, G. L. Challis, K. Fowler, T. Kieser, K. F. Chater, *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 1541–1546.
[23] S. Lautru, R. J. Deeth, L. Bailey, G. L. Challis, *Nat. Chem. Biol.* **2005**, *1*, 265–269.
[24] L. Song, F. Barona-Gomez, C. Corre, L. Xiang, D. W. Udwary, M. B. Austin, J. P. Noel, B. S. Moore, G. L. Challis, *J. Am. Chem. Soc.* **2006**, *128*, 14754–14755.
[25] C. Corre, L. Song, S. O'Rourke, K. F. Chater, G. L. Challis, *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 17510–17515.
[26] G. C. Fazio, R. Xu, S. P. T. Matsuda, *J. Am. Chem. Soc.* **2004**, *126*, 5678–5679.
[27] G. L. Challis, *J. Med. Chem.* **2008**, *51*, 2618–2628.
[28] D. W. Udwary, L. Zeigler, R. N. Asolkar, V. Singan, A. Lapidus, W. Fenical, P. R. Jensen, B. S. Moore, *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 10376–10381.
[29] H. Gross, V. O. Stockwell, M. D. Henkels, B. Nowak-Thompson, J. E. Loper, W. H. Gerwick, *Chem. Biol.* **2007**, *14*, 53–63.
[30] A. L. McClerren, L. E. Cooper, C. Quan, P. M. Thomas, N. L. Kelleher, W. A. van der Donk, *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 17243–17248.
[31] F. Malpartida, J. Niemi, R. Navarrete, D. A. Hopwood, *Gene* **1990**, *93*, 91–99.
[32] S. C. Wenzel, F. Gross, Y. Zhang, J. Fu, A. F. Stewart, R. Müller, *Chem. Biol.* **2005**, *12*, 349–356.
[33] J. W. Bok, D. Hoffmeister, L. A. Maggio-Hall, R. Murillo, J. D. Glasner, N. P. Keller, *Chem. Biol.* **2006**, *13*, 31–37.
[34] S. Bergmann, J. Schuemann, K. Scherlach, C. Lange, A. A. Brakhage, C. Hertweck, *Nat. Chem. Biol.* **2007**, *3*, 213–217.
[35] R. B. Williams, J. C. Henrikson, A. R. Hoover, A. E. Lee, R. H. Cichewicz, *Org. Biomol. Chem.* **2008**, *6*, 1895–1897.
[36] S. Rigali, F. Titgemeyer, S. Barends, S. Mulder, A. W. Thomae, D. A. Hopwood, G. P. van Wezel, *EMBO Rep.* **2008**, *9*, 670–675.
[37] G. Wang, K. Hoska, K. Ochi, *Appl. Environ. Microbiol.* **2008**, *74*, 2834–2840.
[38] W. Zhang, Y. Tang, *J. Med. Chem.* **2008**, *51*, 2629–2633.

Received: June 9, 2008
Published online on January 22, 2009

