# 5

# Improving Drug Discovery From Microorganisms

**Chris M. Farnet and Emmanuel Zazopoulos**

**Summary**

Microorganisms remain unrivalled in their ability to produce bioactive small molecules for drug development. However, the core technologies used to discover microbial natural products have not evolved significantly over the past several decades, resulting in a shortage of new drug leads. Advances in DNA-sequencing and bioinformatics technologies now make it possible to rapidly identify the clusters of genes that encode bioactive compounds and to make computer predictions of chemical structure based on gene sequence information. These structure predictions can be used to identify new chemical entities and provide important physicochemical "handles" that guide compound purification and structure confirmation. Industrialization of this process provides a model for improving the efficiency of natural-product discovery. The application of advanced genomics and bioinformatics technologies is now poised to revolutionize natural-product discovery and lead a renaissance of interest in microorganisms as a source of bioactive compounds for drug development.

**Key Words:** Natural products; genomics; drug discovery; bioinformatics; actinomycetes; dereplication; fermentation; structure elucidation.

## 1. Introduction

Microorganisms produce some of the most important medicines ever developed. They are the source of lifesaving treatments for bacterial and fungal infections (e.g., penicillin, erythromycin, streptomycin, tetracycline, vancomycin, amphotericin), cancer (e.g., daunorubicin, doxorubicin, mitomycin, bleomycin), transplant rejection (e.g., cyclosporin, FK-506, rapamycin), and high cholesterol (e.g., statins such as lovastatin and mevastatin) (**Fig. 1**). Microbial natural products are notable not only for their potent therapeutic activities, but also for the fact that they frequently possess the desirable pharmacokinetic properties required for clinical development. The drugs shown in **Fig. 1** are just a few of the many microbial natural products that reached the market without any chemical modifications required, a testimony to the remarkable ability of microorganisms to produce drug-like small molecules. Indeed, the potential to hit a "home run" with a single discovery distinguishes natural products from all other sources of chemical diversity and fuels the ongoing efforts to discover new compounds.

Traditionally, the search for new natural products has started by growing microorganisms in the laboratory and testing the fermentation broths for bioactivity. However, we now know that microorganisms have many natural-product gene clusters that they do not readily express when grown in the laboratory *(1–3)*. So despite decades of fermentation-broth screening, it is likely that a vast supply of bioactive microbial com-

From: *Natural Products: Drug Discovery and Therapeutic Medicine*
Edited by: L. Zhang and A. L. Demain © Humana Press Inc., Totowa, NJ

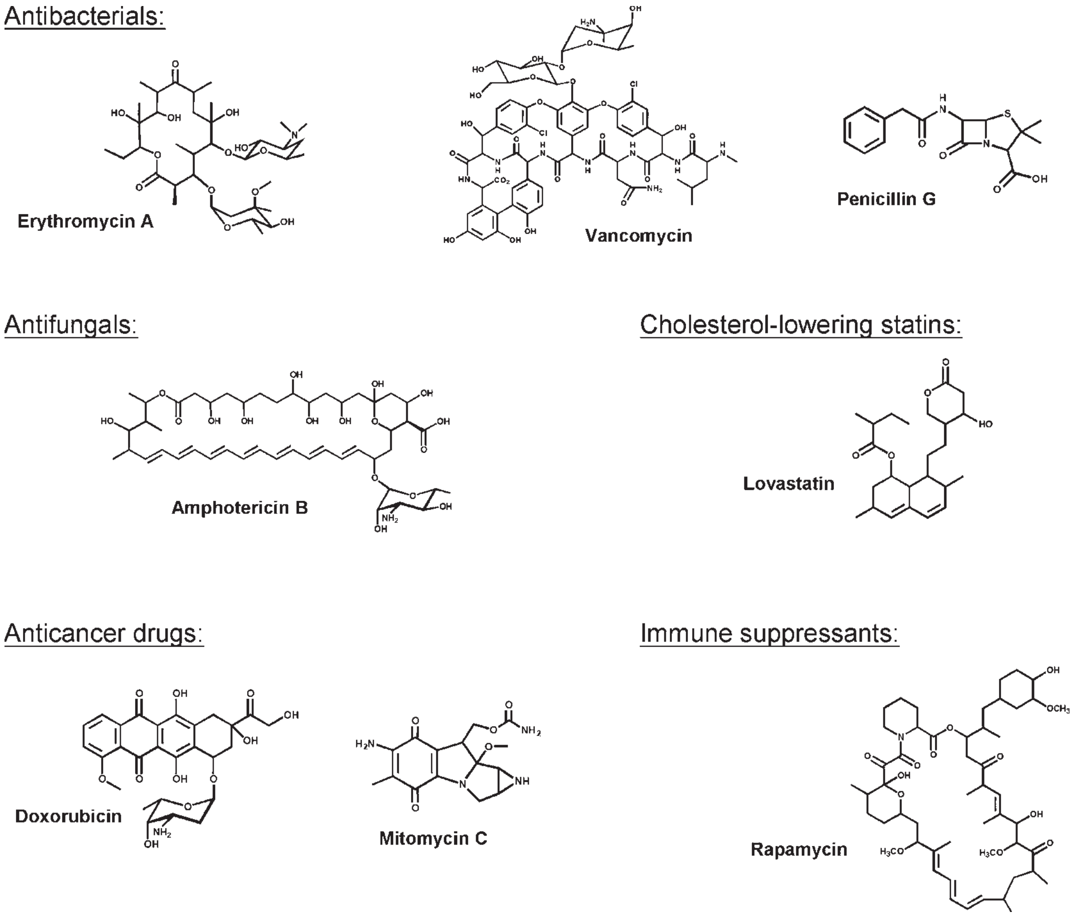


**Fig. 1.** A few of the landmark medicines produced by microorganisms.

pounds remains to be discovered. This untapped potential provided the impetus for us to develop an automated genomics platform that predicts the chemical structures of natural products by reading the sequences of the gene clusters that direct their synthesis. By surveying the genome, we can identify all of the natural products that a microorganism can make before fermentation studies begin, and specifically tailor the downstream production and purification strategies to isolate likely new chemical entities (NCEs) and avoid the re-isolation of known compounds. The integration of new genomics technologies greatly increases the efficiency of discovery and makes it possible to build a robust pipeline of NCEs from a small collection of microorganisms, providing a new paradigm for natural-product discovery. Here, we describe the core technologies behind the discovery platform developed at Ecopia BioSciences Inc. and present the discovery of a new antifungal agent as a case study to illustrate the power of the genomics-guided approach.

## 2. Genomics-Guided Natural-Product Discovery

### 2.1. Genome Scanning Technology

The development of a discovery platform that can predict chemical structures from gene sequences required the development of several new technologies and resources, including a means to efficiently isolate and sequence natural-product gene clusters from microbial genomes; a large reference database of gene clusters linked to the structures of the compounds they encode; and specialized computer applications that can detect correlations between gene sequence and chemical substructure elements. We developed a high-throughput genome-scanning method to sequence natural-product gene clusters without sequencing entire genomes *(3)*. Our strategy was to scan the genomes of selected microorganisms that were reported to produce known, structurally diverse natural products and to build a database of gene clusters covering the full range of natural-product chemical diversity (the Ecopia Decipher® database). The approach proved to be successful, as in all cases the gene clusters corresponding to known natural products were identified by deductive analysis, providing an important training set for chemical structure predictions. The most striking finding, however, was the large number of unexpected gene clusters found in these previously well-studied microorganisms. Genome scanning of approximately 60 actinomycete strains revealed some 700 natural product gene clusters, or an average of a dozen gene clusters per organism. This number exceeds by at least a factor of ten the number of natural products that would have been detected from these organisms by traditional screening approaches (**Fig. 2**). It is now clear that many gene clusters are expressed only under certain growth conditions. Furthermore, even when they are expressed, some gene clusters produce compounds only at very low levels, below the limit of detection of conventional screening methods. This may explain why so many natural products have eluded detection in the past, as it was common practice to screen only one to three growth conditions for each strain.

### 2.2. Genomics-Guided Discovery Platform

To capitalize on the wealth of gene clusters revealed by genome scanning, we developed a genomics-guided discovery platform designed to rapidly identify clusters encoding likely NCEs and target the compounds for purification (**Fig. 3**). A suite of specialized software and computer applications predicts the structures of compounds encoded by new gene clusters via automated comparisons with known clusters in the database. These structure predictions identify possible NCEs and provide important physicochemical "handles" (including molecular weight, ultraviolet [UV] absorbance, lipophilicity, and other properties) that are then used to detect the desired compound in fermentation broths. To fully exploit their potential, each microorganism is grown in as many as 50 different fermentation media in order to maximize the probability that each of its gene clusters will be expressed. A number of custom-made analytical tools are used to simultaneously display and analyze mass spectroscopy, UV, and bioactivity data generated from extracts prepared from all the growth conditions used. These tools make it possible to detect compounds that may be produced only rarely in fermentation broths, or at very low levels, and to identify those compounds whose properties match the gene-cluster predictions. The structure information provides practical handles to

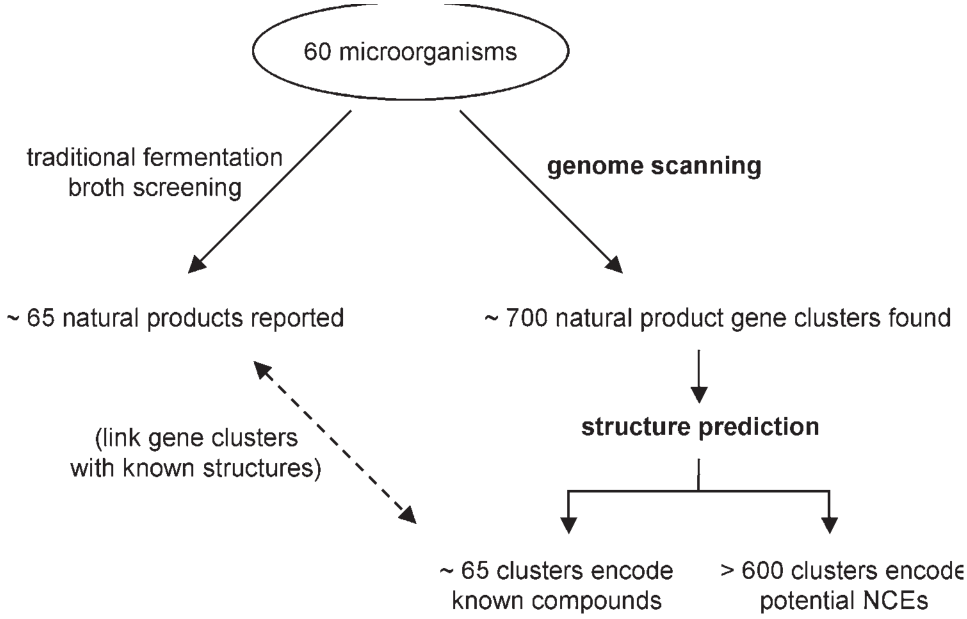

60 microorganisms
traditional fermentation broth screening
genome scanning
~ 65 natural products reported
~ 700 natural product gene clusters found
(link gene clusters with known structures)
structure prediction
~ 65 clusters encode known compounds
> 600 clusters encode potential NCEs

**Fig. 2.** Genome scanning reveals a vast supply of undiscovered natural products.

guide the purification of targeted compounds and greatly facilitates the final structure elucidation by spectroscopic methods.

The tremendous power of genomics-guided discovery is driven by a database and computational platform that "learns" from previous results and improves with each new compound discovered. In the final step of the discovery cycle, the confirmed structure of a new molecule is linked in the Decipher® database to the gene cluster that encodes it, thus enhancing the ability of the system to make future correlations between genes and chemical structures. Even the "rediscovery" of known compounds adds valuable new information to the database, as the genetic blueprint for each structure identifies new genes-to-molecules correlations. In addition, all of the chemical and biological data generated during the fermentation, extraction, purification, and bioactivity screening stages are fed back into the database and integrated with the genomics information. Sophisticated bioinformatics applications are then able to identify relationships between the diverse data sets that can be used to guide the production and purification of targeted metabolites—for example, by defining the fermentation media that are likely to support the expression of a gene cluster and by identifying purification schemes that have proved successful with isolations from the specific medium and for similar structure types.

## 3. Genomics-Guided Discovery: A Case Study

### *3.1. Genome Scanning of* Streptomyces aizunensis

The untapped potential of microorganisms to produce bioactive NCEs is illustrated by our experience with the actinomycete *Streptomyces aizunensis*, a producer of the antibiotic bicyclomycin. The bicyclomycin gene cluster was targeted for isolation because the compound contains some unusual functional groups and the genes required

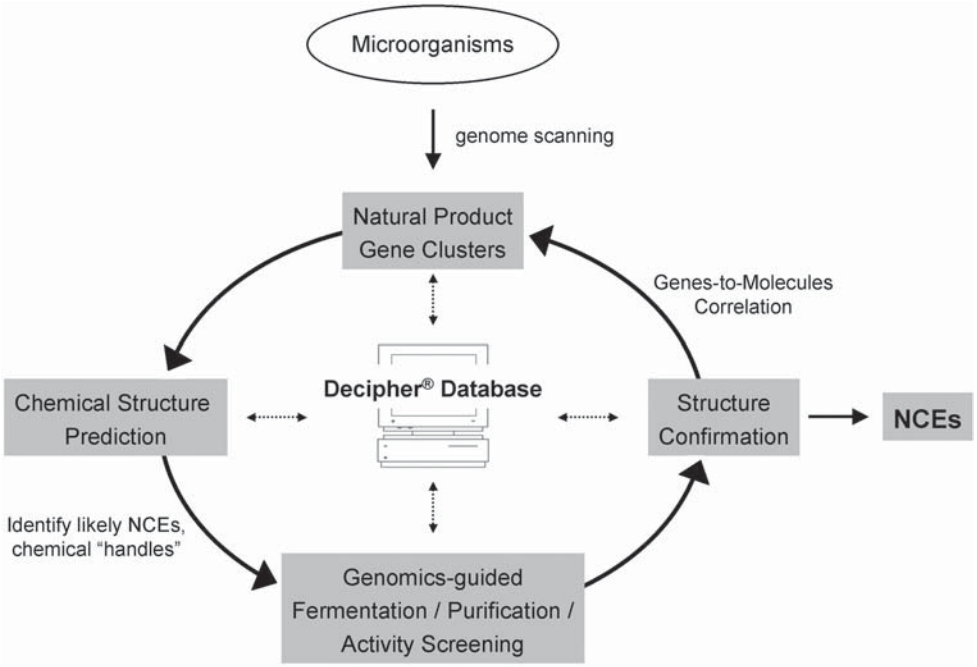

Microorganisms
genome scanning
Natural Product Gene Clusters
Genes-to-Molecules Correlation
Decipher® Database
Chemical Structure Prediction
Structure Confirmation
NCEs
Identify likely NCEs, chemical "handles"
Genomics-guided Fermentation / Purification / Activity Screening

**Fig. 3.** The genomics-guided discovery platform developed at Ecopia, a new paradigm for natural product discovery. The platform identifies all of the natural-product gene clusters in a genome, predicts the chemical structures of the compounds they encode, and targets compounds that are likely to be NCEs for production and purification.

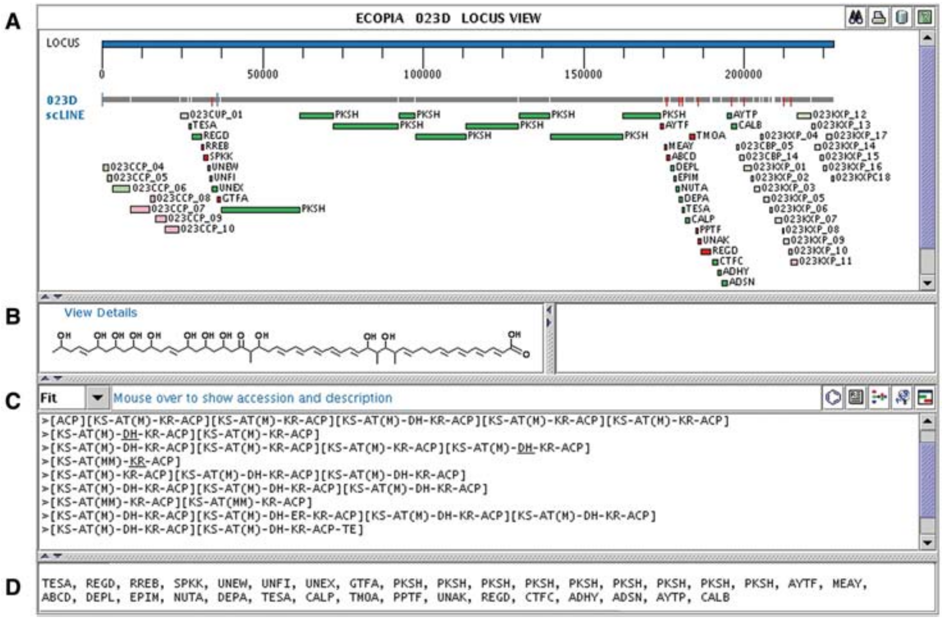
to make these kinds of structures were not known. Surprisingly, genome scanning of the *S. aizunensis* genome identified 11 natural-product gene clusters in addition to the bicyclomycin cluster, even though this organism produced only bicyclomycin in published fermentation screening studies *(4,5)*. One of the additional gene clusters was predicted to encode a compound similar to the known antibiotic streptothricin, based on computer-based comparisons to other clusters in the database. This information proved to be valuable, as a streptothricin-like compound was indeed detected during subsequent fermentation experiments. Knowing in advance that this compound would be produced made it very easy to identify and circumvent, while purifying other compounds from the fermentation broths. More importantly, the structure predictions generated for the remaining 10 gene clusters did not match any compounds present in databases of known natural products, indicating that they are likely to encode NCEs. Each of these clusters presented an exciting opportunity for new compound discovery. In the following sections, we demonstrate how automated gene sequence analysis was used to predict the structure of an NCE encoded by one of these gene clusters and how genomics information was used to guide the purification of the compound.

### 3.2. Automated Analysis of Gene Clusters and Chemical Structure Prediction

All gene clusters identified by genome scanning enter a fully automated analysis cascade of specialized software applications that identify open reading frames (ORFs), assign functions to each gene in the cluster, and predict chemical substructure elements. The output of the automated analysis is displayed in an interactive graphical user interface designed to allow scientists to quickly assess the accuracy of the computer predictions and to determine whether a cluster is likely to encode a known compound or an NCE. The automated analysis of one of the *S. aizunensis* gene clusters, designated 023D, is shown in **Fig. 4**. In this cluster, the computer assigned 35 ORFs to protein families based on homology comparisons to proteins in the Decipher database. The disposition of these ORFs in the cluster is shown in window A of **Fig. 4**, where each ORF carries a four-letter code indicating the protein family to which it was assigned. Nine of the ORFs in the 023D cluster were designated as polyketide synthases (PKSs). PKSs and other multimodular protein families (such as nonribosomal peptide synthetases) are further processed by an automated software application that parses the proteins into individual enzymatic domains. Each domain sequence is then compared to a series of protein models of active domains to identify domains that are likely to be nonfunctional. Additional computer scripts are also invoked when particular domains are encountered. For example, the substrate specificity of each acyltransferase (AT) domain is readily assigned by a phylogenetic comparison to AT domains of known specificity, while a similar analysis of thioesterase (TE) domains very effectively distinguishes domains that generate linear polyketide products from those that catalyze the formation of cyclic products. The result is an automated "domain string" (displayed in **Fig. 4**, window C) that captures the structure of the polyketide backbone in a line notation that can then be translated into a chemical structure prediction (**Fig. 4**, window B). The automated analysis of the 023D PKS system predicted a long, linear polyketide chain bearing polyene chromophores. While many cyclic (macrolide) polyene natural products are known, linear polyenes remain relatively rare. Chemical substructure searches using the predicted polyketide backbone identified no similar structures in natural-products databases, providing the first indication that the 023D gene cluster encoded a NCE.

### 3.3. Correlating Genes With Chemical Substructures

The "family string" generated by the analysis cascade (shown in window D of **Fig. 4**) provides a representation of the cluster that is used in an automated search of the database to identify gene clusters with similar families. The structures of the compounds linked to these clusters are then compared to identify common structural elements. For example, three gene clusters in the Decipher database contain the ADSN, AYTP, and CALB families found in the 023D cluster. Structure analysis of the corresponding compounds identified a single common structural element, a 2-amino-3-hydroxycyclopentenone ($C_5N$) group in amide linkage to a polyketide carboxylate (**Fig. 5**, upper). Inspection of the computer-predicted function of each family suggested a plausible pathway for the biosynthesis of a $C_5N$ group from glycine and 5-aminolevulinic acid. Thus, the presence of these three genes in a cluster provides a marker for the presence of this functional group in a natural product. Similarly, computer analysis correlated four families in the 023D cluster with the presence of a four-carbon, amine-containing ($C_4N$) polyketide starter unit (**Fig. 5**, middle) and five families with a 6-deoxyhexose sugar moiety (**Fig. 5**, lower). In both cases the predicted functions of the families suggested likely biosynthetic pathways and strongly supported

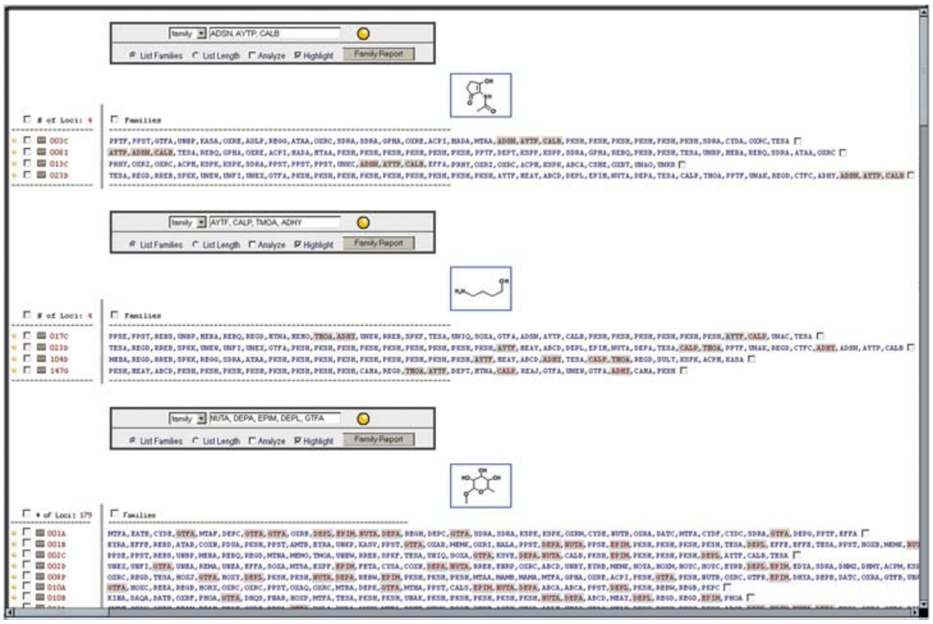


Fit ▾ Mouse over to show accession and description

```
>[ACP][KS-AT(M)-KR-ACP][KS-AT(M)-KR-ACP][KS-AT(M)-DH-KR-ACP][KS-AT(M)-KR-ACP][KS-AT(M)-KR-ACP]
>[KS-AT(M)-DH-KR-ACP][KS-AT(M)-KR-ACP]
>[KS-AT(M)-DH-KR-ACP][KS-AT(M)-KR-ACP][KS-AT(M)-KR-ACP][KS-AT(M)-DH-KR-ACP]
>[KS-AT(MM)-KR-ACP]
>[KS-AT(M)-KR-ACP][KS-AT(M)-DH-KR-ACP][KS-AT(M)-DH-KR-ACP]
>[KS-AT(M)-DH-KR-ACP][KS-AT(M)-DH-KR-ACP][KS-AT(M)-DH-KR-ACP]
>[KS-AT(MM)-KR-ACP][KS-AT(MM)-KR-ACP]
>[KS-AT(M)-DH-KR-ACP][KS-AT(M)-DH-ER-KR-ACP][KS-AT(M)-DH-KR-ACP][KS-AT(M)-DH-KR-ACP]
>[KS-AT(M)-DH-KR-ACP][KS-AT(M)-DH-KR-ACP-TE]
```

```
TESA, REGD, RREB, SPKK, UNEW, UNFI, UNEX, GTFA, PKSH, PKSH, PKSH, PKSH, PKSH, PKSH, PKSH, PKSH, PKSH, AYTF, MEAY,
ABCD, DEPL, EPIM, NUTA, DEPA, TESA, CALP, TMOA, PPTF, UNAK, REGD, CTFC, ADHY, ADSN, AYTP, CALB
```

**Fig. 4.** Automated analysis of the 023D gene cluster: (**A**) overview of the automated gene finding and family calling; (**B**) predicted structure of the polyketide backbone; (**C**) automated domain string; (**D**) automated family string.



family ▾ ADSN, AYTP, CALB
● List Families ○ List Length □ Analyze ☑ Highlight Family Report

□ # of Loci: 4 □ Families

□ 003C PPTF, PPST, GTFA, UNBP, KASA, OXRE, ADLP, REGG, ATAA, OXRC, SDRA, SDRA, GPHA, OXRE, ACPI, MADA, MTAA, ADSN, AYTP, CALB, PKSH, PKSH, PKSH, PKSH, PKSH, PKSH, SDRA, CYDA, OXRC, TESA □
□ 008I AYTP, ADSN, CALB, TESA, REBQ, GPHA, OXRE, ACPI, MADA, MTAA, PKSH, PKSH, PKSH, PKSH, PKSH, PKSH, PPTF, DEPT, KSPP, KSPP, SDRA, GPHA, REBQ, PKSB, PKSH, TESA, UNBP, MEBA, REBQ, SDRA, ATAA, OXRC □
□ 013C PRNY, OXRI, OXRC, ACPH, KSPK, KSPK, SDRA, PPST, PPST, PPST, UNKC, ADSN, AYTP, CALB, EFFA, PRNY, OXRI, OXRC, ACPH, KSPK, ABCA, CSHE, OXBT, UNAO, UNKR □
□ 023D TESA, REGD, RREB, SPKK, UNEW, UNFI, UNEX, GTFA, PKSH, PKSH, PKSH, PKSH, PKSH, PKSH, PKSH, PKSH, PKSH, AYTF, MEAY, ABCD, DEPL, EPIM, NUTA, DEPA, TESA, CALP, TMOA, PPTF, UNAK, REGD, CTFC, ADHY, ADSN, AYTP, CALB □

family ▾ AYTF, CALP, TMOA, ADHY
● List Families ○ List Length □ Analyze ☑ Highlight Family Report

□ # of Loci: 4 □ Families

□ 017C PPSE, PPST, REBS, UNBP, MEBA, REBQ, REGD, MTNA, MEMO, TMOA, ADHY, UNEW, RREB, SPKF, TESA, UNIQ, SOXA, GTFA, ADSN, AYTP, CALB, PKSH, PKSH, PKSH, PKSH, PKSH, PKSH, AYTF, CALP, UNAC, TESA □
□ 023D TESA, REGD, RREB, SPKK, UNEW, UNFI, UNEX, GTFA, PKSH, PKSH, PKSH, PKSH, PKSH, PKSH, PKSH, PKSH, PKSH, AYTF, MEAY, ABCD, DEPL, EPIM, NUTA, DEPA, TESA, CALP, TMOA, PPTF, UNAK, REGD, CTFC, ADHY, ADSN, AYTP, CALB □
□ 104D MEBA, REGD, RREB, SPKK, REGG, SDRA, ATAA, PKSH, PKSH, PKSH, PKSH, PKSH, PKSH, PKSH, PKSH, PKSH, AYTF, MEAY, ABCD, ADHY, TESA, CALP, TMOA, REGD, SULT, KSPK, ACPH, KASA □
□ 147G PKSH, MEAY, ABCD, PKSH, PKSH, PKSH, PKSH, PKSH, PKSH, PKSH, PKSH, CANA, REGD, TMOA, AYTF, DEPT, MTNA, CALP, REAJ, GTFA, UNEW, GTFA, ADHY, CANA, PKSH □

family ▾ NUTA, DEPA, EPIM, DEPL, GTFA
● List Families ○ List Length □ Analyze ☑ Highlight Family Report

□ # of Loci: 179 □ Families

□ 001A MTFA, EATB, CYDE, GTFA, MTAF, DEPC, GTFA, GTFA, OXRB, DEPL, EPIM, NUTA, DEPA, REGH, DEPC, GTFA, SDRA, SDRA, KSPK, KSPK, OXRM, CYDE, NUTB, OXRA, DATC, MTFA, CYDF, CYDC, SDRA, GTFA, DEPG, PPTF, EFFA □
□ 001B EYRA, EFFE, REBD, ATAB, COXN, PDUA, PKSH, PPST, AMTB, EYRA, UNKP, KASV, PPST, GTFA, OXAB, MEMK, OXRI, HALA, PPST, DEPA, NUTA, PPSE, EPIM, PKSH, PKSH, PKSH, PKSH, TESA, DEPL, EFFE, EFFE, TESA, PPST, HOXB, MEMK, NU
□ 002C PPSE, PPST, REBS, UNBP, MEBA, REBQ, REGD, MTNA, MEMO, TMOA, UNEW, RREB, SPKF, TESA, UNIQ, SOXA, GTFA, KSVE, DEPA, NUTA, CALB, PKSH, EPIM, PKSH, PKSH, PKSH, DEPL, AYTF, CALB, TESA □
□ 002D UNEX, UNFI, GTFA, UNEA, REMA, UNEA, EFFA, SOXA, MTSA, EXPF, EPIM, PETA, CYSA, COXN, DEPA, NUTA, RREB, ENRP, OXRC, ABCD, UNBY, EYRB, MEME, HOYA, HOXM, HOYC, HOYC, EYRB, DEPL, EPIM, EDYA, SDRA, DHMZ, DHMY, ACPM, KS
□ 008F OXRC, REGD, TESA, HOXJ, GTFA, HOXY, DEPL, PKSH, PKSH, NUTA, DEPA, REBW, EPIM, PKSH, PKSH, PKSH, MTAA, MAMB, MAMA, MTFA, GPHA, OXRE, ACPI, PKSH, GTFA, PKSH, NUTB, OXRC, GTFB, EPIM, DHYA, DEPB, DATC, OXRA, GTFB, UN
□ 010A GTFA, HOXC, REEA, REGB, HORX, OXRC, OXRC, PPST, OXAQ, OXRC, MTBA, DEPK, GTFA, MTNA, PPST, CALS, EPIM, NUTA, DEPA, ABCA, ABCA, PPST, DEPL, PKSH, REBW, REGB, PKPC □
□ 010B KIRA, DAQA, DATB, OXBF, PHOA, GTFA, DHQS, PHAR, HOXP, MTFA, TESA, PKSH, PKSH, UNAK, PKSH, PKSH, PKSH, PKSH, PKSH, NUTA, DEPA, ABCD, MEAT, DEPL, REGD, REGD, EPIM, PHOA □

**Fig. 5.** Correlating genes and structures. An automated application identifies gene clusters having similar gene families (highlighted). The compounds produced by these gene clusters are then compared to identify common structural elements (boxed). This analysis identified three genes that correlate with the $C_5N$ structure element (upper), four genes that correlate with the $C_4N$ polyketide starter group (middle), and five genes that correlate with a 6-deoxyhexose group (lower).

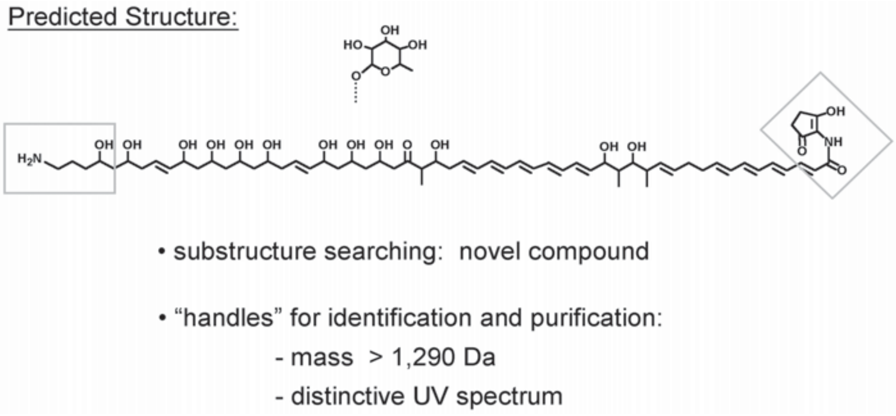
the gene-structure correlations. The family- and domain-string notations demonstrate the utility of reducing gene-sequence information to a series of computable properties, or "genetic descriptors," that a computer can learn to associate with chemical-structure descriptors. As the number of sequenced gene clusters in the Decipher database climbs, the ability to predict chemical structure directly from gene sequence is increasingly refined.

The results of the automated analysis cascade provided a very precise prediction of the structure of the compound encoded by the 023D gene cluster, as shown in **Fig. 6**. Tools for substructure searching are also integrated into the discovery platform and allow a scientist to quickly assess whether a predicted structure has already been reported, providing an early opportunity for *in silico* dereplication. For example, substructure searches of the AntiBase database (Wiley Publishers, 2003) of over 30,000 microbial natural products revealed only 64 products that contain the $C_5N$ group. This example illustrates how even a small amount of structure information can greatly limit the number of structures that need to be considered as candidate products. More importantly, the addition of a second structure element to the query returned no hits from the database, providing further evidence for the novelty of the compound encoded by the 023D gene cluster (**Fig. 6**).

### 3.4. Finding the Needle in a Haystack: Genomics-Guided Purification

Having strong evidence for an NCE, the compound encoded by the 023D gene cluster was targeted for purification. The structure prediction immediately identified physicochemical properties or "handles" that could be used to guide the purification of the compound. For example, the compound was predicted to have a molecular mass in excess of 1290 Daltons (Da) and a distinctive UV spectrum imparted by the pentaene chromophore. To ensure that the gene cluster was expressed, *S. aizunensis* was grown in more than 50 different fermentation media in 25-mL shake-flask cultures. Methanol extracts of each culture were subjected to high-performance liquid chromatography (HPLC)-UV-mass spectrometry (MS) analyses, and metabolites were monitored using a specially designed system that makes it possible to analyze HPLC fractions simultaneously across all the different media conditions (**Fig. 7**, upper). An overview of the MS traces showed that the profile of metabolites varied considerably from medium to medium. The interface shown in **Fig. 7** is fully interactive, so that the underlying chemical data can be rapidly searched using queries that incorporate multiple physicochemical parameters. For example, a mass filtering function allows the user to search all fractions for masses within a particular range. A search for metabolites having a mass greater that 1290 Da identified a single peak that appeared only in some fermentation media but not in others (**Fig. 7**, middle). Clicking on any fraction pops up a new window that displays the full spectral data set for that fraction. When this was done for one of the peak fractions from the mass filtered search, the data revealed molecular ions consistent with a major isotope of mass 1296.7 Da and a UV absorption spectrum characteristic of a pentaene (**Fig. 7**, lower inset), fully consistent with the structure predictions generated by computer analysis. Thus, the chemical data provided strong evidence that the 1296.7-Da metabolite corresponded to the compound encoded by the 023D gene cluster. In addition to the chemical data, aliquots of each HPLC fraction are routinely tested for antimicrobial activity against a panel of bacterial and fungal patho-

Predicted Structure:



- substructure searching: novel compound
- “handles” for identification and purification:
  - mass > 1,290 Da
  - distinctive UV spectrum

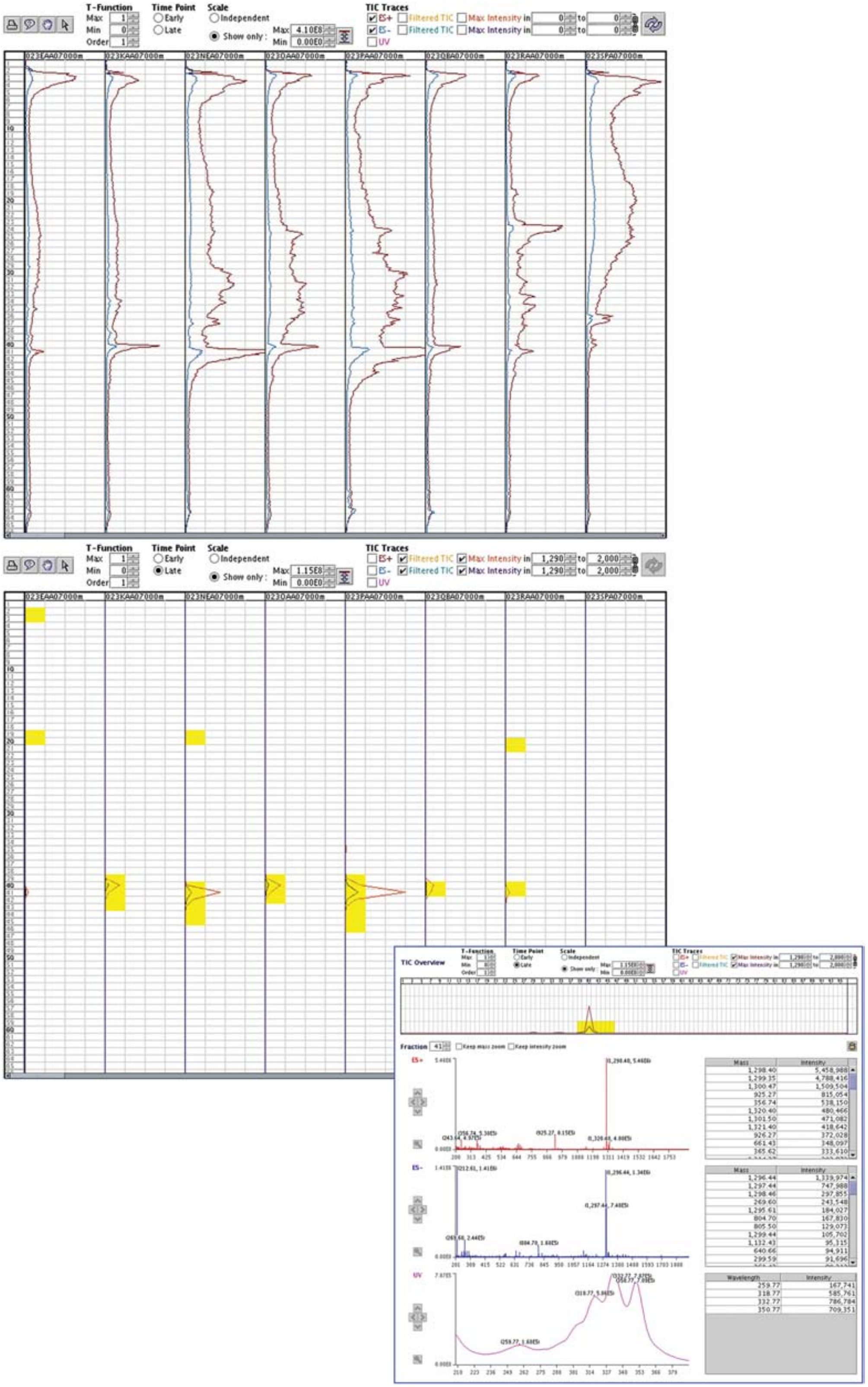
**Fig. 6.** Computer-generated prediction of the structure of the compound encoded by the 023D gene cluster. The boxed portions indicate examples of substructure elements that can be used to search against a database of known natural products.

gens, and the bioassay results can also be displayed along with the chemical data in the graphical interface. In **Fig. 7** (middle), the screening data for *Candida albicans* is shown, as the fractions containing the 1296.7-Da compound exhibited a potent antifungal activity in the bioassay screens.

Normally, the purification of a natural product from a complex fermentation mixture presents a formidable task that may take months to achieve. In this case, however, having the mass, UV, and bioactivity data in hand made the subsequent purification of the 1296.7-Da compound straightforward, and this was accomplished in a matter of days. Similarly, the elucidation of a complex natural-product structure using the standard analytical techniques can be exceedingly difficult, but in this case it was greatly facilitated by the *in silico* structure prediction. The final structure was confirmed by multidimensional NMR spectrometry, and proved to be entirely consistent with the structure prediction generated by gene-sequence analysis (**Fig. 8**). The new compound, named ECO-02301, displayed potent in vitro activity against numerous fungal pathogens, including *Aspergillus fumigatus* and azole-resistant strains of *Candida albicans*, and was shown to be efficacious in a mouse model of disseminated candidiasis, where treatment of infected mice resulted in a statistically significant increase of the median survival as compared to nontreated animals. ECO-02301 thus represents an exciting new chemical class of natural product and a promising agent for development as a treatment for serious fungal infections.

## 4. Summary

The discovery of ECO-02301 is one of the early successes of the genomics-guided platform developed at Ecopia and the new paradigm for natural-product discovery. The discovery of this compound from *S. aizunensis* provides direct evidence that the capacity for microorganisms to produce natural products has been greatly underestimated, and that exciting new natural products can be discovered from microorganisms that were already screened using traditional approaches.



T-Function Max 1 Min 0 Order 1
Time Point ○ Early ○ Late
Scale ○ Independent ● Show only : Max 4.10E8 Min 0.00E0
TIC Traces ☑ ES+ □ Filtered TIC □ Max Intensity in 0 to 0
☑ ES- □ Filtered TIC □ Max Intensity in 0 to 0
□ UV

T-Function Max 1 Min 0 Order 1
Time Point ○ Early ● Late
Scale ○ Independent ● Show only : Max 1.15E8 Min 0.00E0
TIC Traces □ ES+ ☑ Filtered TIC ☑ Max Intensity in 1,290 to 2,000
□ ES- ☑ Filtered TIC ☑ Max Intensity in 1,290 to 2,000
□ UV

TIC Overview
T-Function Max 1 Min 0 Order 1
Time Point ○ Early ● Late
Scale ○ Independent ● Show only : Max 1.15E8 Min 0.00E0
TIC Traces □ ES+ □ Filtered TIC ☑ Max Intensity in 1,290 to 2,000
□ ES- □ Filtered TIC ☑ Max Intensity in 1,290 to 2,000
□ UV

Fraction 41 □ Keep mass zoom □ Keep intensity zoom

| Mass | Intensity |
|---|---|
| 1,298.40 | 5,458,988 |
| 1,299.35 | 4,788,416 |
| 1,300.47 | 1,509,504 |
| 925.27 | 815,054 |
| 356.74 | 538,150 |
| 1,320.40 | 480,466 |
| 1,301.50 | 471,082 |
| 1,321.40 | 418,642 |
| 926.27 | 372,028 |
| 661.43 | 348,097 |
| 365.62 | 333,610 |

| Mass | Intensity |
|---|---|
| 1,296.44 | 1,339,974 |
| 1,297.44 | 747,988 |
| 1,298.46 | 297,855 |
| 269.60 | 243,548 |
| 1,295.61 | 184,027 |
| 804.70 | 167,830 |
| 805.50 | 129,073 |
| 1,299.44 | 105,702 |
| 1,132.43 | 95,315 |
| 640.66 | 94,911 |
| 299.59 | 91,696 |

| Wavelength | Intensity |
|---|---|
| 259.77 | 167,741 |
| 318.77 | 585,761 |
| 332.77 | 786,784 |
| 350.77 | 709,351 |

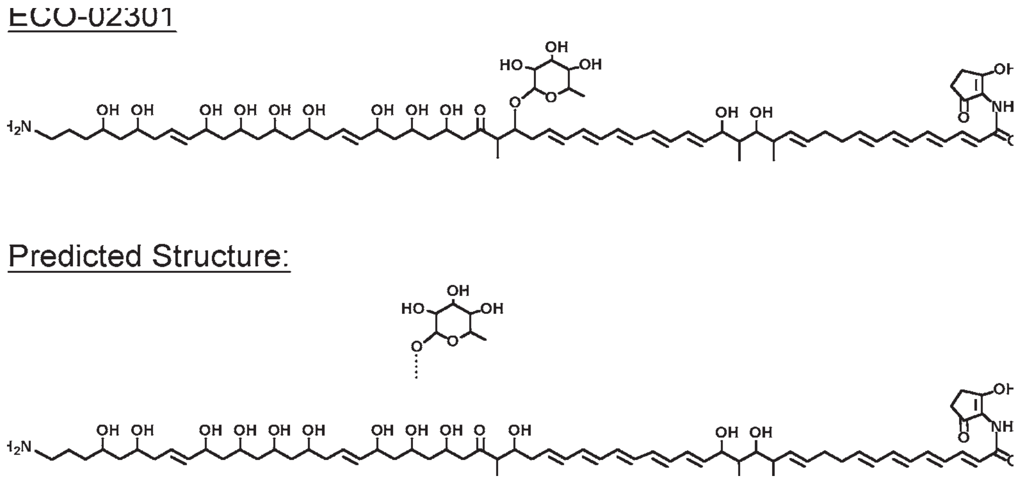


**Fig. 8.** The structure of ECO-02301, a new class of antifungal agent. The confirmed structure matches the computer-generated prediction.

Critical to the success of genomics-guided discovery are the database and data-mining tools that enable comparative analysis of gene clusters and the prediction of chemical structures from gene-sequence information. With each new gene cluster added to the Decipher database, the structure-prediction capabilities of the platform become increasingly accurate. The integration of specialized applications to search for metabolites simultaneously across multiple growth conditions makes it possible to correlate the metabolome of an organism with the gene clusters discovered by genome scanning and to identify metabolites that are likely to be NCEs.

Given that only a tiny fraction of the microbial world has been explored, the potential for discovering new natural products is virtually unlimited. Increasingly, new compounds will be discovered at the computer, and bioinformatics technologies will be used to tailor strategies for their production and purification, thus overcoming many of the technical hurdles previously associated with natural-product discovery. With the introduction of powerful new genomics technologies, the remarkable fifty-year track record of microorganisms in producing landmark medicines is now poised to extend well into the new millennium.

**Fig. 7.** (*opposite page*) Finding the predicted compound in fermentation broths. Upper, parallel high-performance liquid chromatography (HPLC)-mass spectrometry analyses of extracts of *S. aizunensis* grown in multiple fermentation media. Each column represents a different fermentation medium, and each row in a column represents a single HPLC fraction. All fifty media can be analyzed simultaneously by scrolling across the window. Middle, the same interface, filtered for metabolites having a mass greater than 1290 Da, showing a single peak found in only some fermentation media; the *Candida albicans* screening data are also shown in this view (yellow highlights) as an antifungal activity correlated with the peak fractions. Clicking on a peak fraction pulls up a new window (lower inset) showing detailed spectroscopic data consistent with the predicted compound.

## Acknowledgments

The authors would like to acknowledge Drs. Jim McAlpine and Brian Bachmann for their contributions to the discovery of ECO-02301.

## References

1. Bentley SD, Chater KF, Cerdeno-Tarraga AM, et al. Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). Nature 2002;417:141–147.
2. Omura S, Ikeda H, Ishikawa J, et al. Genome sequence of an industrial microorganism *Streptomyces avermitilis*: deducing the ability of producing secondary metabolites. Proc Natl Acad Sci USA 2001;98:12,215–12,220.
3. Zazopoulos E, Huang K, Staffa A, et al. A genomics-guided approach for discovering and expressing cryptic metabolic pathways. Nat Biotechnol 2003;21:187–190.
4. Miyamura S, Ogasawara N, Otsuka H, et al. Antibiotic No. 5879, a new water-soluble antibiotic against Gram-negative bacteria. J Antibiotics 1972;25:610–612.
5. Miyamura S, Ogasawara N, Otsuka H, et al. Antibiotic 5879 produced by *Streptomyces aizunensis*, identical with bicyclomycin. J Antibiotics 1973;26:479–484.