REVIEW

Evolution and taxonomic distribution of nonribosomal peptide and polyketide synthases

Grigoris D Amoutzias, Yves Van de Peer & Dimitris Mossialos†

†Author for correspondence
Department of Biochemistry & Biotechnology, University of Thessaly, Ploutonos & Aiolou 26, GR-41221 Larissa, Greece
Tel.: +30 241 056 5283; Fax: +30 241 056 5290; mosial@bio.uth.gr

The majority of nonribosomal peptide synthases and type I polyketide synthases are multimodular megasynthases of oligopeptide and polyketide secondary metabolites, respectively. Owing to their multimodular architecture, they synthesize their metabolites in assembly line logic. The ongoing genomic revolution together with the application of computational tools has provided the opportunity to mine the various genomes for these enzymes and identify those organisms that produce many oligopeptide and polyketide metabolites. In addition, scientists have started to comprehend the molecular mechanisms of megasynthase evolution, by duplication, recombination, point mutation and module skipping. This knowledge and computational analyses have been implemented towards predicting the specificity of these megasynthases and the structure of their end products. It is an exciting field, both for gaining deeper insight into their basic molecular mechanisms and exploiting them biotechnologically.

A large number of antibiotics, antitumor agents, immunosuppressants, toxins and siderophores are produced in microbes by two classes of related multidomain modular enzymes (megasynthases), the nonribosomal peptide synthases (NRPSs) and polyketide synthases (PKSs) type-I (PKS-I) [1–4]. Owing to their modular architecture, these enzymes have the potential to synthesize a vast number of polyketides and oligopeptides, which are low molecular weight secondary metabolites. Up to now, the structures of approximately 10,000 polyketides have been identified, but theoretical considerations raise the number of possible structures to over 1 billion [5]. The importance of NRPSs and PKSs in the fields of biomedicine, biotechnology and food technology is already significant and is expected to grow even more in the following years. Examples of widely used products of NRPSs and PKSs include the immunosuppressant molecule ciclosporin as well as the antibiotic erythromycin [1]. Ongoing research is unraveling the molecular mechanisms that underlie the biosynthesis of NRPS and PKS products. In this way, proper manipulation of existing enzymes could modify their product, or product activity [6,7].

In the last decade, a plethora of microbial genomes, together with experimental data on individual enzymes, have provided the opportunity to apply various bioinformatics tools and computational analyses. These tools and analyses focus on mining the new genomes for NRPSs and PKSs, understanding the evolution and diversity of these synthases and predicting their products.

Keywords: antibiotics, bioinformatics, distribution, evolution, nonribosomal peptide synthase, NRPS, PKS, polyketide synthase, prediction, siderophores

future medicine part of fsg

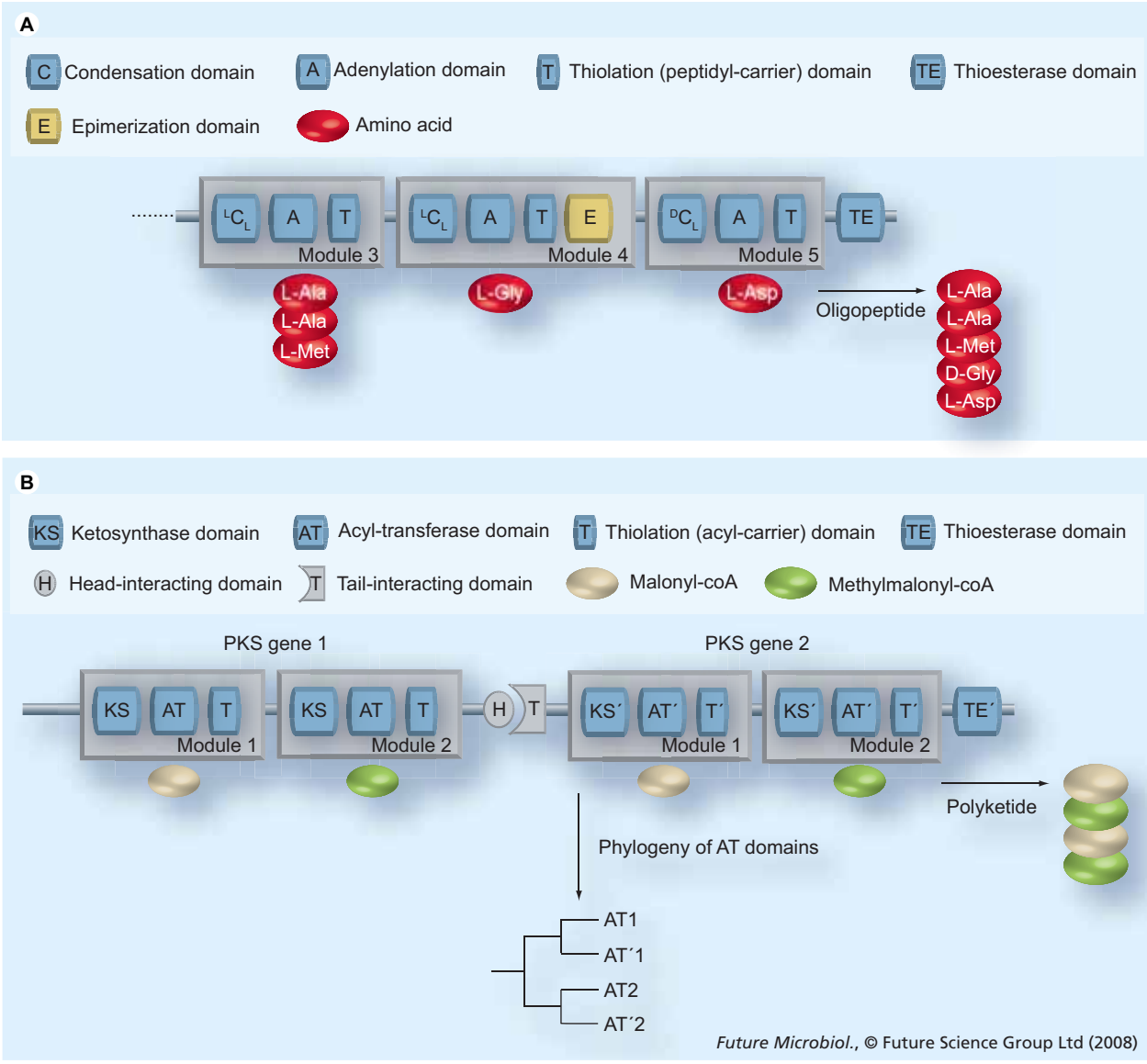
Enzyme organization

PKSs have been categorized into three distinct groups (type I, II and III), but in this review we will focus on type I PKSs, which are mostly multimodular and have a biosynthetic machinery similar in organization and function to NRPSs [1,8]. PKS-I and NRPSs both constitute megasynthases that polymerize acyl-coenzyme A or organic acid (amino acid and hydroxy acid) monomers into more complex products. A chain of polymerized monomers is produced, which is elongated in steps. This chain can undergo further modifications, such as cyclization, epimerization, reduction, methylation, and so on [1].

The structure and function of these multimodular enzymes is reviewed in detail in [1]. A module of at least three different domains is responsible for every cycle of the elongation process (Figure 1). The order of the modules actually defines the order of the polymerized monomers in the produced polyketide or oligopeptide chain. This constitutes the colinearity rule. Based on their relative position in the pathway they are designated as initiation, elongation or termination modules [9]. Every module is responsible for the incorporation of a certain type of monomer. In each module, two domains are catalytic, whereas the third is a carrier.

Specifically, in NRPSs, the adenylation (A) domain is responsible for recognition and activation of its cognate amino acid or hydroxy acid and transfer to the peptidyl-carrier (PCP) domain of the same module [1]. The A domain identifies and incorporates the 20 proteinogenic amino acids as well as 300 nonproteinogenic

Figure 1. Domain organization of nonribosomal peptide synthases and type I polyketide synthases.



(A) Domain organization of NRPS proteins. The core domains of each module are C, A and T. Nevertheless, other accessory domains may be present as well, such as the E domain. In this example, an L-amino acid is incorporated from module 4, but the E domain changes it to a D-amino acid. The C domain of the downstream module needs to form a C–N bond between the D-amino acid of the elongated chain and the downstream L-amino acid of module 5. The C domain of module 5 is designated as $^{D}C_{L}$, whereas the other C domains are designated as $^{L}C_{L}$. $^{D}C_{L}$ domains can be distinguished from $^{L}C_{L}$ domains in a phylogenetic analysis. **(B)** Domain organization of PKS type I genes. KS, AT and T domains are organized in modules. Every module is responsible for the incorporation of one monomer in the elongated polyketide. PKS type I proteins may interact in a head to tail fashion, thus forming an assembly line megasynthase. AT domains that recognise malonyl-CoA form a phylogenetic group that is distinct from AT domains that recognise methylmalonyl-CoA.
PKS: Polyketide synthase.

residues [10]. The condensation (C) domain is responsible for formation of a C–N bond between the elongated amino acid chain (bound at the preceding PCP domain) and the activated amino acid (that is bound at the PCP domain of this module) [11]. Several types of C domains exist [11]. The $^{L}C_{L}$ subtype forms the bond between two L-amino acids, whereas the $^{D}C_{L}$ subtype forms the bond between an L-amino acid and an oligopeptide that ends in a D-amino acid (Figure 1A). Sometimes, a special type of C domain, the cyclization domain, is responsible for both condensation

and cyclization of the elongated chain. Starter C domains acylate the first amino acid of the elongated chain, whereas the homologous epimerization domain changes the chirality from an L- to a D-amino acid (Figure 1A). Dual epimerization (E/C) domains are responsible for both epimerization and condensation. The last NRPS module often contains a thioesterase domain that releases the peptide product from the enzyme by cyclization or hydrolysis [10–12].

In PKSs, the acyl-transferase (AT), ketosynthase (KS) and acyl-carrier (ACP) domains are equivalent in function to the A, C and PCP domains of NRPS, respectively (Figure 1B) [8]. Specifically, the AT domain identifies and incorporates malonyl or methyl-malonyl CoA, among other substrates. The KS domain is responsible for the formation of a C–C bond, whereas in most cases a thioesterase (TE) domain is identified in the last module. The ACP and PCP domains of PKSs and NRPSs are also called thiolation (T) domains.

In addition to the three core domains found in each module of PKSs and NRPSs, other accessory domains can be found as well. These accessory domains in NRPSs can catalyze N-methylation and epimerization, whereas in PKSs they perform keto-reduction (KR), enoyl-reduction and dehydration (DH).

One or more PKS or NRPS modules may constitute an open reading frame (ORF). An ORF may contain either PKS or NRPS modules, or both of them, thus forming hybrids. Also, more than one ORF can be organized in an operon or gene cluster to form a certain product (Figure 2); the order of the ORFs in the genome does not necessarily reflect the order of the polymerized monomers in the end product [8]. PKS and NRPS proteins form protein complexes by interacting in a head to tail fashion, thus forming an assembly line (Figure 2) [9,13–19]. The order in which they interact will affect the structure of the metabolite, owing to the colinearity rule. In PKS-I, a 19 amino acid C-terminal docking domain (head region) and a 27 amino acid N-terminal docking domain (tail region) are responsible for the interaction between two PKS proteins [13,16,19]. In a similar fashion, NRPSs have short communication mediating domains at the C- and N-terminal region of two consecutive proteins of the assembly line [9,15].

Genomic mining & phylogenetic distribution

In the last decade, a plethora of sequenced genomes (~700) [20], together with the application of homology-search bioinformatics tools (see Box 1), such as pairwise and profile search, have contributed significantly to the identification and annotation of new NRPS and PKS genes from diverse organisms. So far, PKS-I genes have been identified in bacteria, fungi, chromalveolates and chlorophytes [21], whereas a search in the Pfam database shows that NRPS genes are present in bacteria and fungi, as well as in a few metazoa [22].

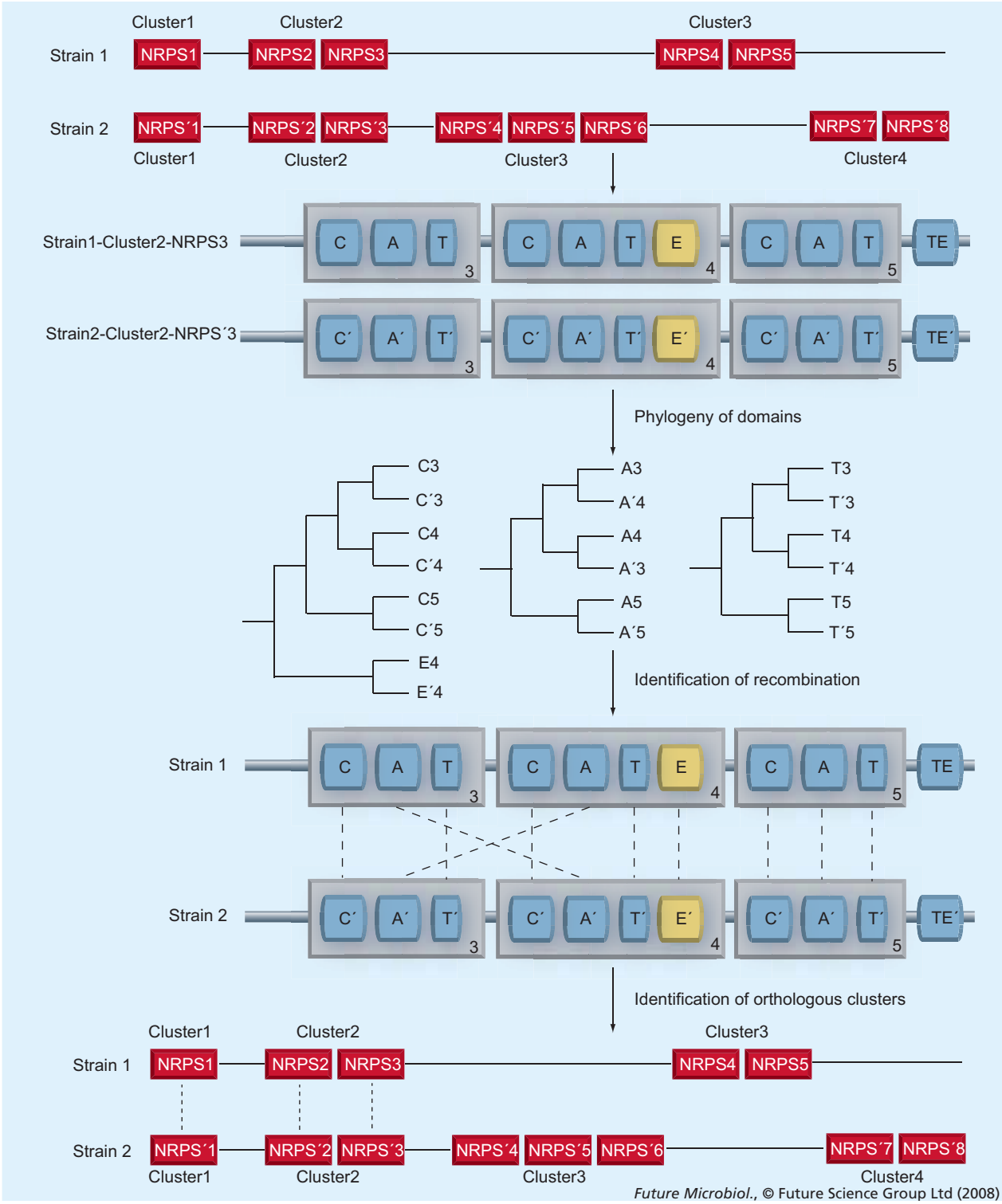
The Basic Local Alignment Search Tool (BLAST) algorithm tries to find general similarity between a (region of a) query and a (region of a) target sequence, whereas the profile hidden Markov model (HMM) suite of tools, HMMER, tries to identify a specific domain or set of domains in the query sequence.

In practical terms, it means that BLAST may retrieve, as significant hits, proteins that are functionally unrelated with PKSs or NRPSs, but share some short region of similarity with them, for example, an accessory domain. On the contrary, HMMER will identify only proteins that have a particular domain. The C domain is characteristic of NRPSs, whereas the KS domain is also found in fatty acid synthases. Therefore, the C domain is a good marker of how many nonribosomal peptide molecules a given organism produces; however the same does not apply to polyketides.

The plethora of genomic data is very well documented in the database GOLD: Genomes online [20]. Since January 2008, approximately 700 genomic projects have been completed and 2800 are ongoing. A total of 590 of the completed genome projects involve bacteria, 50 Archaea, 36 fungi and 11 plants. A search for the C domain in Pfam shows that from all the lineages, bacteria and fungi are especially enriched in NRPS genes, whereas animals and archaea possess almost no NRPS genes. The rare presence of NRPS genes in the metazoan *Caenorhabditis elegans* is probably the result of horizontal gene transfer (HGT) from bacteria [23].

Currently sequenced bacterial genomes do not represent the full spectrum of bacterial lineages, but instead are biased in favor of animal and plant pathogens, and also human pathogens [24]. GOLD statistics show that almost half of the bacterial genomic projects are on proteobacteria, with funding coming mostly from the biomedical field. Therefore, the genomic projects are influenced in favor of bacterial organisms that are expected to harbor many PKS or NRPS genes. On the contrary, archaeal genomic projects are influenced in favor of extremophiles. Owing to their extreme environment, these

Figure 2. Identification of orthology and evolutionary events among domains, genes and clusters of two organisms.



Phylogenetic analyses of the various domains will reveal which ones are orthologous and which ones duplicated or shuffled by recombination events, as shown for the A domains of modules 3 and 4 of NRPS3 of gene cluster 2. Once orthology among domains has been established, we can deduce the orthologous relationships between genes and clusters of two organisms and identify events of gene loss or horizontal gene transfer.

A: Adenylation; C: Condensation; E: Epimerization; NRPS: Nonribosomal peptide synthase; T: Thiolation; TE: Thioesterase.

Box 1. Bioinformatics tools for homology search, clustering and classification.

- The most widely used algorithm for rapidly identifying homologous sequences is BLAST [54]. This tool compares a query sequence against a database of known sequences and tries to identify homologous regions between the query sequence and a target sequence in the database. In essence, BLAST compares the query sequence against each one of the database sequences and tries to find the common fragments. The longer and more similar the fragments, the higher the score. A statistical expectation value (e-value) is also calculated, which shows the probability of obtaining the same score against a database (of the same size) of random sequences. The calculation of the score takes into account a general substitution matrix of amino acids or nucleotides. Compared with other pairwise similarity tools, BLAST does not guarantee to find the optimal alignment between two sequences, especially if they are very distantly related, but it is a very fast algorithm and provides satisfactory solutions that are comparable to the other, more computationally intensive algorithms.
- When we want to identify homology between distantly related sequences, a more sensitive and effective tool is the HMM, which is based on profile searches [22]. As a first step, a protein multiple alignment is generated from many diverse sequences of a certain domain. In this way, the algorithm can identify which positions of the domain are highly conserved, which are more variable, and what the amino acid or gap frequencies are. Then, a HMM is trained with this alignment to identify distantly related domains in a query sequence. The HMMER suite of tools is publicly available for training of custom-made HMMs [102]. In addition, a collection of various HMMs for many domains is available in the Pfam database [55]. Therefore, a specific query sequence or all of the ORFs of a new genome can be scanned against either a custom made HMM or a certain collection of HMM models, or even against the whole Pfam database.
- Phylogenetic analysis clusters homologous proteins or domains based on sequence similarity. First, the domains of the same category are extracted and aligned. Next, various algorithms that take into consideration the mutation frequency from one nucleotide to the other, or from one amino acid to the other, are used to cluster them, most commonly forming bifurcation trees. The most common methods are neighbour joining, maximum likelihood and maximum parsimony [56]. By mapping additional information, such as ORFs and species origin of each domain or functional information, we can understand how the functions and the multimodular proteins evolved and whether this methodology could help us predict the function of new unknown proteins **(Figure 1B)**.
- Support vector machines are linear classifiers based on supervised learning [57]. They map vectors in high-dimensional space, and based on the training data they calculate hyperplanes that seperate the various classes of data.

BLAST: Basic local search alignment tool; HMM: Hidden Markov model; ORF: Open reading frame.

archaea generally do not confront as fierce competition from other microbes as much as the sequenced pathogenic bacteria do. This is reflected on metagenomic projects, which demonstrate the limited biodiversity (mainly archaeal) in extreme environments, compared with the biodiversity found in soil or water samples [25]. Therefore, it is conceivable that the currently observed absence of NRPS genes in archaea could be an artefact, stemming from the genomic sampling bias, although this is yet to be confirmed. Regarding the animal lineage, many representatives from various phylogenetic groups have been sequenced and therefore their general lack of NRPS genes seems to be true.

An extensive literature review and genomic scanning of PKS and NRPS genes with BLAST on 223 bacterial strains showed that most genes are found in γ-proteobacteria, actinobacteria and β-proteobacteria [24]. In addition, there were no functional data for most of these genes. Their annotation relied on homology against a small number of genes with experimentally determined metabolite structures. Our search [Amoutzias GD, Mossialos D, Unpublished Data] in Pfam revealed that the top molecular factories of NRPSs in bacteria are predicted to be *Pseudomonas syringae pv. syringae* (*strain B728a*; 17 NRPS genes, 65 C domains), *Myxococcus xanthus* (*strain DK 1622*; 26 NRPS genes, 79 C domains) and *Rhodococcus sp.* (*strain RHA1*; 23 NRPS genes, 118 C domains). The same Pfam search in fungi revealed that the top molecular factories are predicted to belong to the species of the genus *Aspergillus* (e.g., *Aspergillus terreus* NIH2624; 25 NRPS genes, 62 C domains) and several other species, such as *Phaeosphaeria nodorum* SN15 (15 NRPS genes, 49 C domains), *Chaetomium globosum* CBS 148.51 (15 NRPS genes, 54 C domains) and *Gibberella zeae* (*Fusarium graminearum*; 12 NRPS genes, 64 C domains). A very interesting finding is that in bacteria one strain may have a considerably different number of NRPS genes and C domains than other strains of the same species. The three *Pseudomonas syringae* strains are a very good example [101] where the NRPS gene content and C domains may vary from 10 to 17 genes and 20–65 C domains. Such a variation may occur either by extensive gene loss in any of the strains, by HGT, or both. This process has a significant impact on the way genes are annotated using a BLAST search against a well studied

bacterial strain. The best BLAST hit against a well annotated strain does not necessarily mean that the query sequence is actually the true orthologue (Figure 2). Therefore, regarding these multidomain synthases, caution is needed when performing automatic annotation of a new genome. The analysis system for modular polyketide synthases (ASMPKS) and NRPS/PKS bioinformatics tools are freely available to help identify and annotate NRPS and PKS genes in a sequenced genome [26,27].

NRPS and PKS proteins are a huge energetic burden on an organism [1]. In the case of NRPS, a module of three domains that spans 1000 amino acids is responsible for the incorporation of one monomer in the elongated chain. The most striking example is the ciclosporin synthase, a protein of 15,000 amino acids, which produces an undecapeptide [28]. Such a great energetic burden must be counterbalanced by the protective or adaptive effect of the synthesized secondary metabolite in a given environment. Once the environment changes and there is no more need for these secondary metabolites, the NRPS and PKS megasynthases become an immense burden, compromising the fitness of the microorganism. Given the strong link between effective population size and the efficiency of selection in a certain species, which the nearly neutral theory of molecular evolution suggests [29], it is expected that microorganisms will jettison unnecessary gene clusters very rapidly. It is probably the ecology of the microorganisms that affects their PKS and NRPS gene content, rather than their evolutionary history. Indeed, a study in cyanobacteria shows that the phylogenetic distribution of these molecules is very patchy and is the result of extensive gene loss in many derived cyanobacterial lineages [30]. Such a patchy distribution is also observed for the PKS-I in unicellular eukaryotes [21].

Given the bias in the selection of sequenced genomes, the great variation in the number of PKS and NRPS genes (even among strains), the documented HGT among bacteria–bacteria, bacteria–fungi and fungi–fungi, and the extensive PKS and NRPS gene loss in derived lineages, we believe that it is actually impossible so far (from the currently limited genomic sampling) to conclude confidently that any certain microbial taxonomic lineage lacks these genes.

Evolution & diversity

The modular PKS-I and NRPS megasynthases are an energetically expensive solution for producing small secondary metabolites. There are probably two reasons why evolution has favored the emergence and fixation of such energetically expensive megasynthases: the incorporation of nonproteinogenic substrates (for NRPS) in the elongated chain and the rapid evolution and diversity of products, which results from the underlying modular structure of these synthases.

Multidomain proteins generally evolve by gene duplication, recombination, gene fusion/fission, domain deletions/substitutions and circular permutations, where the sequential order between two genes is inverted [31–34]. Indeed, in NRPS and PKS, circular permutations have been detected and are attributed to a duplication/deletion mechanism [33]. In addition, phylogenetic analyses, especially of the KS and C domains, show that intragenic as well as intergenic duplications are mainly responsible for the evolution of a given pathway [35].

Gene and domain duplications are responsible for creating longer products, but recombination and point mutations are responsible for increasing the diversity of the new product. The phylogeny of domains that are responsible for variation in the product substrate, such as the A, AT, KR and DH domains, clearly shows that they are moving by recombination events from one gene cluster to the other, or within a cluster, with a mechanism that resembles 'copy/cut-paste' (Figure 2) [35]. A striking case of intragenic swapping of adenylation domains only, and not of the whole module, was observed when comparing the almost identical iturin A and mycosubtilin NRPS proteins from two strains of *Bacillus subtilis* (RB14 and ATCC6633, respectively) [36]. Recombination events may also result in loss of domain function [35]. Interdomain regions flanking AT domains as well as regions within domains likely function as recombination points [35].

An evolutionary study of PKS-I genes in *Streptomyces* highlights the importance of recombination in the evolution of pathways and generation of metabolite diversity [35]. Via recombination, modules and accessory domains are shuffled and thus diversity in the metabolite product is generated. However, point mutations do not seem to significantly affect the evolution and diversity of the polyketide product. This could be due to the fact that the various isoforms of each domain diverged a long time ago and it takes more than a couple of mutations to change the specificity of the domains found in PKS-I [35]. Also, as yet there are no signs of positive selection in PKS-I genes, rather purifying

selection. NRPSs diversify by duplication and recombination, as well as by point mutations, which change the specificity of the A domain [10,36–42]. This is due to the fact that ten amino acids of the A domain catalytic pocket appear to be mostly responsible for the recognition and activation of an amino acid substrate. Many A domains have side specificities for noncognate amino acids that have similar biochemical properties with the cognate substrate [40]. It seems that in some cases one or two point mutations are enough to change the specificity in favor of the noncognate amino acid [40,41].

Apart from the aforementioned mechanisms, extensive gene loss as well as HGT also occur among bacteria and fungi, as well as between bacteria and fungi [30,35,43–46]. There are several ways to infer whether the presence of a gene is due to HGT. A good indication is the positioning of genes in highly mobile genomic regions, such as pathogenicity islands and plasmids. Also, genomic anomalies or phylogenetic context are used, for example, GC content, codon usage and noncanonical phylogenetic distribution [44]. Occasionally, it is difficult to deduce whether noncanonical phylogenetic distribution is due to HGT, or due to extensive gene loss in many branches of a phylogenetic group. In α-, β- and γ-proteobacteria, most PKS-I are found in pathogenicity islands or plasmids [43]. PKS-I genes, found in fungi, mostly evolved by gene duplication, divergence and gene loss, and HGT that occurred between fungi–fungi and fungi–bacteria [45,46]. Several PKS and NRPS genes were found in genomic islands of the marine bacterium *Hahella chejuensis* [44]. It seems that HGT is occurring frequently within phyla, between phyla and even between kingdoms, for both NRPSs and PKSs. These new genes may further shuffle with the cognate genes, via recombination, thus increasing the diversity of metabolites.

A mechanism that may disrupt colinearity and generate modified metabolites is module skipping. A 'loss of function' mutation in the core motif of a PCP domain in NRPS results in a whole module skipping. [42]. In addition, module skipping has also been observed for PKS-I, but via a different mechanism [47].

Code specificity

Evolutionary studies show that extensive shuffling occurs between modules and individual domains, which is reminiscent of a copy/cut-paste function [35,36]. This shuffling (in addition to point mutations) is responsible for generating a vast diversity of nonribosomal peptide and polyketide metabolites. Nevertheless, the same studies show that certain combinations of domains and modules are preferred. Understanding the rules of domain architecture, how the various mutations can change the specificity of a megasynthase and how the polyketide or oligopeptide product can be predicted from the primary sequence of unknown megasynthases is a focal point of research.

A well studied case of specific domain architecture is that of $^{L}C_{L}$ and $^{D}C_{L}$ domains in NRPSs (Figure 1). If one module ends with an epimerization domain, then the following module starts with a $^{D}C_{L}$ domain, in order to form the C–N bond between an L- and a D-amino acid. Phylogenetic analysis can classify the various types of C domains in the functional groups that we mentioned earlier [11,48]. From these groups, profile HMMs have been generated to automatically detect the type of C domain in a new sequence [11].

In addition, other phylogenetic studies coupled with functional data and bioinformatics tools allow us to predict, to some extent, the product of a PKS, purely from its primary sequence. Specifically, the AT domains that recognize malonyl-CoA can be distinguished (in a phylogenetic analysis) from the AT domains that recognize methylmalonyl-CoA (Figure 1B) [43]. In the same way, KR domains form two distinct phylogenetic groups, where each group creates a certain stereoisomer [35,49]. Software that has been developed for the prediction of polyketide products, based on the domain architecture of PKS sequences, include NRPS/PKS [26], ASMPKS [27], as well as the method developed by Minowa *et al* [50].

Regarding the prediction of NRPS A domain specificity, initial attempts clustered (in a phylogenetic analysis) new domains of unknown specificity with A domains, whose specificity was experimentally determined. For this clustering, a stretch of 200 amino acids was used [41,51]. Nevertheless, when information about the structure of gramicidin synthetase (GrsA) A domain (which recognizes Phe) was integrated, ten amino acids were detected in the binding pocket that are mainly responsible for the specificity of the A domain [38,41]. In this way, a specificity conferring code was formulated [10,37,38,41]. The clustering together of new unknown domains with experimentally determined A domains based on these ten amino acids and not on the whole 200 amino acid

stretch, increased the accuracy of specificity prediction from 43 to 86% [41]. Based on this seminal work, several bioinformatics tools were developed to predict the specificity of A domains. Challis *et al*. used eight out of the ten amino acids originally proposed to cluster domains and predict their specificity [37]. Ansari *et al*. used a combination of BLAST and the ten key residues to predict the specificity of an A domain [26]. Rausch *et al*. used another approach, where they extracted the 34 amino acids of the A domain, which are positioned at a distance of 8 Å around the substrate (based on the crystal structure of GrsA) [40]. In this way, they included not only the amino acids of the binding pocket, but also amino acids that affect the structure of the pocket. They also extracted the corresponding amino acid positions from A domains of known specificity. The various biochemical properties of these pocket-proximity amino acids, as well as the specificities of their corresponding A domains were used to train a support vector machine (SVM). This SVM could then predict the A domain specificity from primary sequence alone, with an improved performance [40].

Another promising direction of research regarding the prediction of the metabolic product is on the prediction of protein–protein interactions among PKS-I or NRPSs of the same pathway [9,13,15,16]. Frequently, two or more proteins from different genes interact head to tail to form a chain of proteins (Figure 1B). The order in which they interact will affect the structure of the metabolite, owing to the colinearity rule. In PKS-I, a 19 amino acid C-terminal head region and a 27 amino acid N-terminal tail region are responsible for the interaction between two PKS proteins [13]. Based on sequence alignment and functional data on true interactions, Thattai *et al*. [19] and Burger *et al*. [52] tried to predict which PKS-I may interact in a head to tail fashion.

Future perspective

As the cost of DNA sequencing decreases, even more genomic and metagenomic projects will identify new sources of these secondary metabolites [53]. The integration of current genomic and functional data has allowed us to understand how these enzymes evolve, how their specificity is encoded on the primary sequence of the megasynthases, and how to transform this evolutionary information into biotechnological applications. Although there is an abundance of sequence data, the analysis of Rausch *et al*. [40] highlights the need for more functional data, especially for the experimental characterization of A domains, in order to develop even more reliable specificity prediction algorithms.

While the problem of A domain specificity is approaching a solution, manipulation of protein–protein interactions is already emerging as a focal point of research. The NRPSs and PKSs are working as protein complexes, and the order of the linear protein complexes defines the product.

Once we can accurately predict the structures of the metabolites, the challenge will be to actually find them experimentally and assign biological functions to them. There is already interesting work in this field and we expect it to increase in the following years [50].

Executive summary

- The majority of nonribosomal peptide synthases (NRPSs) and type I polyketide synthases (PKS-I) are multimodular megasynthases of oligopeptide and polyketide secondary metabolites, respectively. Owing to their multimodular architecture, they can synthesize metabolites in assembly line logic.
- PKS-I are mostly found in bacteria and fungi, but also in chromalveolates and chlorophytes. NRPSs are also found mostly in bacteria and fungi, whereas a few cases of their presence in metazoa is attributed to horizontal transfer from bacteria. Owing to biases in the current selection of sequenced genomes, extensive gene loss in many lineages and horizontal gene transfer, it is too early to characterize any taxonomic lineages as NRPS- or PKS-free. We can be sure of their presence, but not of their absence in a large phylogenetic group.
- The modular structure of PKS-I and NRPSs allows them (in some cases) to evolve rapidly by recombination. In addition, duplications occur at the level of domains, modules and entire genes. The organization of pathways in operons or gene clusters renders it relatively easy to copy them from one organism to another, via horizontal gene transfer. A few point mutations in the catalytic pocket of substrate-selecting domains allow NRPSs to change the specificity for a certain substrate and evolve slightly modified metabolite structures. Other reasons for rapid evolution and diversification of the end products are module skipping as well as protein–protein interactions between the upstream and downstream synthase of the pathway.
- Phylogenetic analyses of domains, coupled with functional data about substrate and protein–protein interaction specificity, have led to the development of several bioinformatics tools that can predict, at least to some extent, the structure of the synthesized metabolite.

Finally, systems biology approaches will attempt to model the pathways that control the synthesis of these metabolites and investigate how their yield can be increased in a controlled laboratory environment. Then, the ultimate goal will be to produce a specific metabolite from an engineered organism of desirable properties with a high yield.

Acknowledgements

We would like to thank two anonymous reviewers for constructive comments on the manuscript.

Financial & competing interests disclosure

G Amoutzias is supported by an EMBO long-term fellowship (ALTF-930–2007) in the laboratory of Y Van de Peer. D Mossialos is financially supported by the Research Committee, University of Thessaly (Project No 3551). The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

Bibliography

Papers of special note have been highlighted as either of interest (•) or of considerable interest (••) to readers.

1. Fischbach MA, Walsh CT: Assembly-line enzymology for polyketide and nonribosomal peptide antibiotics: logic, machinery, and mechanisms. *Chem. Rev.* 106(8), 3468–3496 (2006).
•• Extensive review on the domain organization and biochemical mechanisms of polyketide and nonribosomal peptide synthetase (NRPS) function.
2. Grunewald J, Marahiel MA: Chemoenzymatic and template-directed synthesis of bioactive macrocyclic peptides. *Microbiol. Mol. Biol. Rev.* 70(1), 121–146 (2006).
3. Mossialos D, Amoutzias GD: Siderophores in fluorescent pseudomonads: new tricks from an old dog. *Future Microbiol.* 2, 387–395 (2007).
4. Oide S, Moeder W, Krasnoff S *et al.*: NPS6, encoding a nonribosomal peptide synthetase involved in siderophore-mediated iron metabolism, is a conserved virulence determinant of plant pathogenic ascomycetes. *Plant Cell* 18(10), 2836–2853 (2006).
5. Gonzalez-Lergier J, Broadbelt LJ, Hatzimanikatis V: Theoretical considerations and computational analysis of the complexity in polyketide synthesis pathways. *J. Am. Chem. Soc.* 127(27), 9930–9938 (2005).
6. Kopp F, Marahiel MA: Where chemistry meets biology: the chemoenzymatic synthesis of nonribosomal peptides and polyketides. *Curr. Opin. Biotechnol.* 18(6), 513–520 (2007).
7. Nguyen KT, Ritz D, Gu JQ *et al.*: Combinatorial biosynthesis of novel antibiotics related to daptomycin. *Proc. Natl Acad. Sci. USA* 103(46), 17462–17467 (2006).
8. Cane DE, Walsh CT: The parallel and convergent universes of polyketide synthases and nonribosomal peptide synthetases. *Chem. Biol.* 6(12), R319–325 (1999).
9. Hahn M, Stachelhaus T: Selective interaction between nonribosomal peptide synthetases is facilitated by short communication-mediating domains. *Proc. Natl Acad. Sci. USA* 101(44), 15585–15590 (2004).
•• Identification of key residues that affect the specificity of protein–protein interactions among NRPSs. By harnessing this information, the authors developed a method where various NRPSs may form all possible pairs of interactions and synthesize many different metabolites.
10. Von Dohren H, Dieckmann R, Pavela-Vrancic M: The nonribosomal code. *Chem. Biol.* 6(10), R273–279 (1999).
11. Rausch C, Hoof I, Weber T, Wohlleben W, Huson DH: Phylogenetic analysis of condensation domains in NRPS sheds light on their functional evolution. *BMC Evol. Biol.* 7, 78 (2007).
12. Von Dohren H: Biochemistry and general genetics of nonribosomal peptide synthetases in fungi. *Adv. Biochem. Eng. Biotechnol.* 88, 217–264 (2004).
13. Broadhurst RW, Nietlispach D, Wheatcroft MP, Leadlay PF, Weissman KJ: The structure of docking domains in modular polyketide synthases. *Chem. Biol.* 10(8), 723–731 (2003).
14. Gokhale RS, Tsuji SY, Cane DE, Khosla C: Dissecting and exploiting intermodular communication in polyketide synthases. *Science* 284(5413), 482–485 (1999).
15. Hahn M, Stachelhaus T: Harnessing the potential of communication-mediating domains for the biocombinatorial synthesis of nonribosomal peptides. *Proc. Natl Acad. Sci. USA* 103(2), 275–280 (2006).
16. Lai JR, Fischbach MA, Liu DR, Walsh CT: A protein interaction surface in nonribosomal peptide synthesis mapped by combinatorial mutagenesis and selection. *Proc. Natl Acad. Sci. USA* 103(14), 5314–5319 (2006).
17. O'Connor SE, Walsh CT, Liu F: Biosynthesis of epothilone intermediates with alternate starter units: engineering polyketide-nonribosomal interfaces. *Angew. Chem. Intl Ed. Engl.* 42(33), 3917–3921 (2003).
18. Richter CD, Nietlispach D, Broadhurst RW, Weissman KJ: Multienzyme docking in hybrid megasynthetases. *Nat. Chem. Biol.* 4(1), 75–81 (2008).
19. Thattai M, Burak Y, Shraiman BI: The origins of specificity in polyketide synthase protein interactions. *PLoS Comput. Biol.* 3(9), 1827–1835 (2007).
20. Liolios K, Mavromatis K, Tavernarakis N, Kyrpides NC: The Genomes On Line Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res.* 36, D475–D479 (2008).
21. John U, Beszteri B, Derelle E *et al.*: Novel insights into evolution of protistan polyketide pynthases through phylogenomic analysis. *Protist* 159(1), 21–30 (2008).
22. Sonnhammer EL, Eddy SR, Birney E, Bateman A, Durbin R: Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res.* 26(1), 320–322 (1998).
23. Di Vincenzo L, Grgurina I, Pascarella S: *In silico* analysis of the adenylation domains of the freestanding enzymes belonging to the eukaryotic nonribosomal peptide synthetase-like family. *FEBS J.* 272(4), 929–941 (2005).
24. Donadio S, Monciardini P, Sosio M: Polyketide synthases and nonribosomal peptide synthetases: the emerging view from bacterial genomics. *Nat. Prod. Rep.* 24(5), 1073–1109 (2007).
25. Foerstner KU, Von Mering C, Bork P: Comparative analysis of environmental sequences: potential and challenges. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 361(1467), 519–523 (2006).
26. Ansari MZ, Yadav G, Gokhale RS, Mohanty D: NRPS–PKS: a knowledge-based resource for analysis of NRPS/PKS megasynthases. *Nucleic Acids Res.* 32, W405–W413 (2004).
27. Tae H, Kong EB, Park K: ASMPKS: an analysis system for modular polyketide synthases. *BMC Bioinformatics* 8, 327 (2007).

28. Lawen A, Traber R: Substrate specificities of ciclosporin synthetase and peptolide SDZ 214–103 synthetase: comparison of the substrate specificities of the related multifunctional polypeptides. *J. Biol. Chem.* 268(27), 20452–20465 (1993).
29. Ohta T: Near-neutrality in evolution of genes and gene regulation. *Proc. Natl Acad. Sci. USA* 99(25), 16134–16137 (2002).
30. Rantala A, Fewer DP, Hisbergues M *et al.*: Phylogenetic evidence for the early evolution of microcystin synthesis. *Proc. Natl Acad. Sci. USA* 101(2), 568–573 (2004).
•• Demonstrates that the patchy distribution of microcystins in cyanobacteria is not the result of horizontal gene transfer, but of extensive gene loss. This finding is consistent with the nearly neutral theory of molecular evolution and energetic burden of NRPSs and polyketide synthases (PKSs).
31. Kummerfeld SK, Teichmann SA: Relative rates of gene fusion and fission in multi-domain proteins. *Trends Genet.* 21(1), 25–30 (2005).
32. Vogel C, Teichmann SA, Pereira-Leal J: The relationship between domain duplication and recombination. *J. Mol. Biol.* 346(1), 355–365 (2005).
33. Weiner J 3rd, Bornberg-Bauer E: Evolution of circular permutations in multidomain proteins. *Mol. Biol. Evol.* 23(4), 734–743 (2006).
34. Weiner J 3rd, Beaussart F, Bornberg-Bauer E: Domain deletions and substitutions in the modular protein evolution. *FEBS J.* 273(9), 2037–2047 (2006).
35. Jenke-Kodama H, Borner T, Dittmann E: Natural biocombinatorics in the polyketide synthase genes of the actinobacterium *Streptomyces avermitilis*. *PLoS Comput. Biol.* 2(10), e132 (2006).
•• Detailed analysis of the PKS evolution in *Streptomyces*, which highlights the key role of recombination.
36. Tsuge K, Akiyama T, Shoda M: Cloning, sequencing, and characterization of the iturin A operon. *J. Bacteriol.* 183(21), 6265–6273 (2001).
37. Challis GL, Ravel J, Townsend CA: Predictive, structure-based model of amino acid recognition by nonribosomal peptide synthetase adenylation domains. *Chem. Biol.* 7(3), 211–224 (2000).
38. Conti E, Stachelhaus T, Marahiel MA, Brick P: Structural basis for the activation of phenylalanine in the non-ribosomal biosynthesis of gramicidin S. *EMBO J.* 16(14), 4174–4183 (1997).
•• This paper identifies the ten amino acids of the A domain catalytic pocket, which are mainly responsible for the specificity of the domain towards substrate amino acids.
39. Fewer DP, Rouhiainen L, Jokela J *et al.*: Recurrent adenylation domain replacement in the microcystin synthetase gene cluster. *BMC Evol. Biol.* 7, 183 (2007).
40. Rausch C, Weber T, Kohlbacher O, Wohlleben W, Huson DH: Specificity prediction of adenylation domains in nonribosomal peptide synthetases (NRPS) using transductive support vector machines (TSVMs). *Nucleic Acids Res.* 33(18), 5799–5808 (2005).
41. Stachelhaus T, Mootz HD, Marahiel MA: The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. *Chem. Biol.* 6(8), 493–505 (1999).
42. Wenzel SC, Meiser P, Binz TM, Mahmud T, Muller R: Nonribosomal peptide biosynthesis: point mutations and module skipping lead to chemical diversity. *Angew. Chem. Intl Ed. Engl.* 45(14), 2296–2301 (2006).
43. Jenke-Kodama H, Sandmann A, Muller R, Dittmann E: Evolutionary implications of bacterial polyketide synthases. *Mol. Biol. Evol.* 22(10), 2027–2039 (2005).
44. Jeong H, Yim JH, Lee C *et al.*: Genomic blueprint of *Hahella chejuensis*, a marine microbe producing an algicidal agent. *Nucleic Acids Res.* 33(22), 7066–7073 (2005).
45. Khaldi N, Collemare J, Lebrun MH, Wolfe KH: Evidence for horizontal transfer of a secondary metabolite gene cluster between fungi. *Genome Biol.* 9(1), R18 (2008).
46. Kroken S, Glass NL, Taylor JW, Yoder OC, Turgeon BG: Phylogenomic analysis of type I polyketide synthase genes in pathogenic and saprobic ascomycetes. *Proc. Natl Acad. Sci. USA* 100(26), 15670–15675 (2003).
47. Thomas I, Martin CJ, Wilkinson CJ, Staunton J, Leadlay PF: Skipping in a hybrid polyketide synthase. Evidence for ACP-to-ACP chain transfer. *Chem. Biol.* 9(7), 781–787 (2002).
48. Roongsawang N, Lim SP, Washio K *et al.*: Phylogenetic analysis of condensation domains in the nonribosomal peptide synthetases. *FEMS Microbiol. Lett.* 252(1), 143–151 (2005).
49. Caffrey P: Conserved amino acid residues correlating with ketoreductase stereospecificity in modular polyketide synthases. *Chembiochem.* 4(7), 654–657 (2003).
50. Minowa Y, Araki M, Kanehisa M: Comprehensive analysis of distinctive polyketide and nonribosomal peptide structural motifs encoded in microbial genomes. *J. Mol. Biol.* 368(5), 1500–1517 (2007).
51. Turgay K, Krause M, Marahiel MA: Four homologous domains in the primary structure of GrsB are related to domains in a superfamily of adenylate-forming enzymes. *Mol. Microbiol.* 6(18), 2743–2744 (1992).
52. Burger L, Van Nimwegen E: Accurate prediction of protein–protein interactions from sequence alignments using a Bayesian method. *Mol. Syst. Biol.* 4, 165 (2008).
53. Zhao J, Yang N, Zeng R: Phylogenetic analysis of type I polyketide synthase and nonribosomal peptide synthetase genes in Antarctic sediment. *Extremophiles* 12(1), 97–105 (2008).
54. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: Basic local alignment search tool. *J. Mol. Biol.* 215(3), 403–410 (1990).
55. Finn RD, Tate J, Mistry J *et al.*: The Pfam protein families database. *Nucleic Acids Res.* 36, D281–D288 (2008).
56. Felsenstein J: *Inferring phylogenies*. Sinauer Associates, MA, USA (2003).
57. Byvatov E, Schneider G: Support vector machine applications in bioinformatics. *Appl. Bioinformatics* 2(2), 67–77 (2003).

Websites

101. HHMI Janelia Farm Research Campus. http://pfam.janelia.org/family?acc=PF00668
102. HMMER. http://hmmer.janelia.org/

Affiliations

- *Grigoris D Amoutzias*
  Department of Plant Systems Biology, VIB & Department of Molecular Genetics, Ghent University, Technologiepark 927, B-9052 Ghent, Belgium
  Tel.: +320 9331 3758;
  Fax: +320 9331 3809;
  gramo@psb.ugent.be
- *Yves Van de Peer*
  Department of Plant Systems Biology, VIB & Department of Molecular Genetics, Ghent University, Technologiepark 927, B-9052 Ghent, Belgium
  Tel.: +320 9331 3807;
  Fax: +320 9331 3809;
  yves.vandepeer@psb.ugent.be
- *Dimitris Mossialos*
  Department of Biochemistry & Biotechnology, University of Thessaly, Ploutonos & Aiolou 26, GR-41221 Larissa, Greece
  Tel.: +30 241 056 5283;
  Fax: +30 241 056 5290;
  mosial@bio.uth.gr