Published in final edited form as:

Methods Enzymol. 2009 ; 458: 575–596. doi:10.1016/S0076-6879(09)04823-X.

Cyanobactin ribosomally synthesized peptides — a case of deep metagenome mining

Eric W. Schmidt and Mohamed S. Donia

Department of Medicinal Chemistry, University of Utah, Salt Lake City, UT 84112 USA

Abstract

Deep metagenome mining is a new method for engineering natural product pathways, focusing on examining symbiotic organisms. The method has been applied to a family of compounds known as cyanobactins, which are ribosomally synthesized peptides produced by cyanobacteria. Often, these cyanobacteria live symbiotically with marine animals, leading to production of natural products in whole animal samples. Here, we focus on methods to identify, clone, and study cyanobactin natural product genes from axenic organisms and metagenomic environments. The application to deep metagenome mining is described, along with other potential targets of this methodology.

1. Introduction

Genetic engineering of natural product pathways has led to the synthesis of many new small molecules and to elegant solutions to biochemical problems (Wilkinson and Mickelfield 2007). However, significant hurdles still exist preventing rapid engineering in many cases. One solution to this problem is to look deeply at very closely related pathways to identify natural evolutionary solutions to structural diversity (Schmidt 2005, Schmidt 2008, Schmidt et al. 2000). For example, one could examine a family of closely related natural products to determine what genetic changes underlie structural differences, assuming only a few discrete mutations lead to a new structure. A good source of variability for these studies is found in certain metagenomic environments, particularly bacteria living in symbiotic relationships with animals. By observing this natural variability, it is possible to determine engineering rules for individual pathways and genes. The resulting methodology, dubbed “deep metagenome mining”, was first applied to small ribosomally synthesized peptide natural products, the cyanobactins, but in principle it can be applied to any natural product type.

Deep metagenome mining does not refer to the depth of the ocean, but rather to the depth of scanning individual gene types. That is, instead of a shallow approach to genetics in which a large amount of data about different genes is sought, the deep mining approach seeks to examine numerous nearly identical genes, especially from environmental (metagenomic) samples. Discrete mutations leading to new compounds can be observed by this approach, as highlighted by the cyanobactins example described below. This stands in contrast to pathways with a greater number of mutations, in which it is not possible to determine which mutations are important to the resulting chemical structures without further experiment.

Cyanobactins are produced by many cyanobacterial strains (Figure 1) (Donia et al. 2008, Ireland and Scheuer 1980). Over 100 cyanobactins have been isolated, making them one of the biggest compound classes in cyanobacteria. While the compounds bear many similarities to the larger microcin group (Duquesne et al. 2007), key features that separate the cyanobactins from other compound classes include N-C terminal cyclization and the presence of either heterocycles or isoprenoid derivatives or both. Cyanobactins are found in many different cyanobacteria but were first discovered in marine animals known as ascidians (Figure 2). In fact, the initial

isolation of ulithiacyclamide from an ascidian by Ireland and Scheuer represented the first chemical characterization of any representative of the now well-known microcin class (Ireland and Scheuer 1980). The ascidian animals containing cyanobactins harbored specific cyanobacterial symbionts, *Prochloron* spp., leading Ireland and Scheuer to propose that these cyanobacteria could be responsible for producing the compounds found in whole animals. This prediction was validated 25 years later by genetic methods (Long et al. 2005, Schmidt et al. 2005). Cyanobactins were isolated mostly on the basis of their anticancer activities, which are sometimes in the low nanomolar range against human cell lines (Ireland and Scheuer 1980, Carroll et al. 1996). Not all cyanobactins exhibit this activity: the structures and the resulting activities are quite diverse.

The chemical relationships defining the cyanobactin class can be clearly discerned by comparing structures and amino acid sequence alignments of known compounds, but they are most apparent in the gene clusters for cyanobactins (Figure 3). These clusters always encode two paralogous serine proteases, related to PatA and PatG from the patellamide gene cluster, *pat* (Donia et al. 2006, Donia et al. 2008, Schmidt et al. 2005, Sudek et al. 2006, Ziemert et al. 2008). These proteases catalyze the steps required for N-C cyclization. In addition, all known gene clusters also contain homologs of *patD*, *patF*, or both, which are almost certainly responsible for heterocyclization and / or prenylation. Two further genes that are not required for synthesis when transferred to *E. coli* but are conserved in cyanobacteria are *patB* and *patC*. Finally, an oxidase domain is present in PatG homologs and sometimes as a free-standing protein that probably oxidizes heterocycles, e.g. from thiazoline to thiazole.

The cyanobactin amino acid sequences are directly encoded on a precursor peptide, PatE and its relatives (Figure 4). Usually, more than one cyanobactin structure is encoded on a single precursor peptide. The leader sequence of about 40 amino acids, which is removed by the PatA family of proteases, is relatively well conserved among cyanobactins. In addition, flanking each cyanobactin sequence are two recognition sequences, one of which is recognized by the PatA protease and the other by PatG protease. PatG catalyzes the cyclization event (Lee et al., manuscript submitted).

Our initial question in examining the cyanobactins was, “what genetic modifications are required to go from one structure to another?”. This question was about both the changes to the encoded peptide sequence and post-translational changes in functionality, such as prenylation, heterocyclization, and oxidation.

Symbiotic *Prochloron* spp. provided the ideal model to study pathway evolution. Prior to our work, there were about 60 known cyanobactins from ascidians, representing about 6% of all known ascidian natural products (MarinLit). These cyanobactins formed overlapping families of natural products that had lengths of 6-8 amino acids, variable side chains, and variable modification. First, we examined pathways that did not lead to prenylated molecules. We found that pathway gene clusters from ascidians collected across the Western Pacific were >99% identical at the DNA sequence level (Donia et al. 2006). *patE* genes were also ~99% identical to each other, except in the exact region encoding the peptide products (Table 1). These hypervariable regions encoded very different peptides, yet the modifying genes and intergenic regions were essentially identical. In total, we identified mutants leading to ~30 different products, in which each position was substituted at least once, with identical modifying enzymes. This is basically a natural alanine-scanning mutagenesis, since every position is substituted naturally, but the modifying enzymes were not mutated. This made it possible to identify functional mutations by simple PCR. Inspired by this, we made a completely different peptide, eptidemnamide, by genetic engineering, demonstrating conclusively that the pathway is extremely substrate-tolerant.

We next examined changes leading to prenylation. To our surprise, there were no new enzyme classes in the prenylating pathways such as that for trunkamide (Figure 3) (Donia et al. 2008). Prenylating and non-prenylating pathways were virtually identical (>98% DNA sequence identity) except in a region encoding the C-terminus of PatD, PatF, and the oxidase region of PatG. A PatG-like oxidase was missing from the prenylated pathways; this makes sense since the prenylated molecules were not oxidized. In place of a single *patF* gene, there were two encoded PatF proteins, which were about 40% identical to each other and to the non-prenylating PatF. Thus, simply by sequence gazing, the difference between prenylating and non-prenylating pathways was easily discerned. This would not have been possible without deep metagenome mining because there are no characterized homologs of PatF. Thus, if all proteins had multiple modifications, the basis of prenylation (now explained by the change in PatD C-terminal and / or PatF) would not be clear. Deep metagenome mining allows rapid identification of functional features for genetic engineering.

Several other cyanobactin pathways from free-living cyanobacteria have been examined (Donia et al. 2008, Sudek et al. 2006, Ziemert et al. 2008). These pathways are structurally more diverse than those found in *Prochloron* spp., including peptides 6-12 amino acids in length, with relatives containing nearly every proteinogenic amino acid. The structures, biological activities, and potential roles of diverse cyanobactins are summarized elsewhere (Donia and Schmidt 2009).

Deep metagenome mining revealed extremely closely related pathways with clear engineering rules. The cyanobactins from free-living organisms have more divergent biosynthetic enzymes, but nonetheless useful rules can be deduced by careful comparison of sequences. In particular, the presence of certain types of amino acids at certain positions can be readily discerned (Figure 5). In the natural cyanobactins known to date, the C-terminal amino acid (position 1) is always a heterocycle derived from Cys, Ser, or Thr, or the usual proteinogenic heterocycle, Pro. There is only one exception found by deep metagenome mining, although the resulting compound has not yet been isolated. Thus, position 1 might require heterocycles. Similarly, although cyanobactins have an overall high Cys content, Cys is never found in position 2. In fact, this residue is almost always hydrophobic, with a few exceptions. Thus, some selectivity exists at position 2 for aliphatic and hydrophobic groups, although the presence of Ser / Thr in some compounds suggests the substrate selectivity may be broader at position 2 than position 1. The remaining residues are substituted by nearly every other amino acid and are thus not selective, although they contain a high proportion of aliphatic amino acids and Ser / Thr / Cys. Although these natural experiments provided a great deal of information about substrate selectivity, we were still not sure how much Trp or the more hydrophilic amino acids would be tolerated by the patellamide modifying enzymes. We selected a simple engineering approach to address these issues. The new compound, eptidemnamide, was synthesized by PCR mutagenesis and producing in *E. coli* (Donia et al. 2006). In eptidemnamide, Pro was kept in position 1, Trp (which is almost never found in cyanobactins) was placed in position 2, and the remainder of the amino acids, such as Arg, Gly, Asp, and Gln, were extremely rare in cyanobactins. The engineered compound thus explored much of the sequence selectivity space not encountered by the natural molecules. Together, these studies illustrate the power of traditional engineering methods coupled with deep metagenome mining to rapidly assess design rules for natural products.

2. Some Remaining Questions

The work described above represents just the start of the application of deep metagenome mining to natural products. Numerous questions remain. Among the more pressing questions are: will similar observations be made with other pathway types, such as those to polyketides and nonribosomal peptides, or will different rules be discerned? What drives the changes

described, which are still otherwise unprecedented? Finally, it should be noted that so far cyanobactin genes and compounds have been found only in cyanobacteria, where they are ubiquitous and second only to the nonribosomal peptide / polyketide group. Given the wealth of genome sequencing, the ancient origin of cyanobacteria, and the prevalence of lateral gene transfer, this is bizarre to say the least. One possible example of deep metagenome mining in the polyketide world is found in the area of polyketides derived from mycobacteria (Pidot et al., 2008). These pathogenic (i.e., symbiotic) bacteria were shown to encode different polyketide products from nearly identical genes that exhibited clear deletions or insertions. With this first example, there may be many more such systems to be found in closely related bacteria. The symbiotic pool provides an excellent source of such symbiotic bacteria (Schmidt 2008).

The methods described below are sufficient to enable researchers to work on cyanobactins, either in ascidians or in free-living bacteria. In addition, methods for deep metagenome mining are described.

3. Obtaining *Prochloron* cells and DNA

3.1. Identification and processing of *Prochloron*-containing ascidians

Many of these methods are derived from protocols described in a key book about *Prochloron* and references therein (Lewin and Cheng 1989). *Prochloron* bacteria have not yet been stably cultivated. They have mostly been found as obligate symbionts of ascidians from the family *Didemnidae*. In a handful of cases, they have also been reported from other invertebrates and from stromatolites in Australia (Burns et al. 2005). Over 30 different ascidian species harbor *Prochloron* (Kott 2001), although this number is likely to be an underestimate (Hirose et al. 2008), and these ascidians are found in tropical waters throughout the Indo-Pacific and Red Sea regions. Closely related symbionts are found in Caribbean ascidians (Lafargue and Duclaux 1979). It should be noted that *Prochloron* does not always contain the genes for cyanobactin synthesis; however, they are a relatively reliable source (Donia et al. 2006).

Prochloron-containing ascidians are extremely diverse in morphology and can be readily observed from intertidal depths to the edges of the photic zone (Figure 2) (Lewin and Cheng 1989). Patches of *Prochloron* are spinach-green due to the relatively unusual presence of chlorophyll *b* in the bacteria. Sometimes, this green color has a slightly bluish tinge. Other ascidians sometimes appear to be the same color due to the presence of eukaryotic algae; however, these are readily distinguished upon close inspection because of the punctate appearance of the eukaryotes, whereas *Prochloron* is usually evenly dispersed within a patch. Conveniently, in most ascidian species the *Prochloron* cells are readily displaced from the ascidian by squeezing the animal. Cells obtained in this way clump within a few minutes, providing good additional evidence of the presence of *Prochloron*. In addition, *Prochloron* can be readily observed even with relatively low-power field microscopes, since they are often >10 μ in size. In some cases, *Prochloron* cannot be dissociated from ascidian tissues. Most notably, they are seemingly irretrievably embedded in a mucous matrix in the extremely common ascidians, *Didemnum molle* and relatives. Fortunately, these ascidians are easy to recognize based upon gross morphology (Colin and Arneson 1995).

3.2. DNA purification from enriched *Prochloron* cells

If *Prochloron* can be dissociated from a sample, the following protocol is recommended (Schmidt et al. 2004):

1. Collect and maintain the ascidians in leak-proof bags, maintaining temperature and keeping them wet. Transfer ascidians to larger containers containing fresh seawater as soon as feasible after collection. If possible, process immediately. However,

reasonable *Prochloron* samples have been obtained from specimens kept up to ~24 h if the seawater is changed and kept at a reasonable temperature.

2. Rinse the outside of the ascidians with 0.2 μ -filtered artificial seawater (ASW). ASW can be obtained from pet stores or aquarium supply stores and brought to the collection site as a dry, pre-weighed powder. It is conveniently dissolved in bottled water or tap water and filtered on-site.
3. Press the ascidian against the inner surface of a sterile Petri dish. Resuspend the obtained cells using a pipet equipped with an aerosol tip, pipetting up and down to disperse clumped cells. Transfer to 1.5 mL microcentrifuge tubes.
4. Pellet the cells using a lightweight microcentrifuge.
5. Further squeeze the animal sample and add more cells to the microcentrifuge tubes. Centrifuge as in step 4 and repeat as necessary to maximize cell yield.
6. Rinse cells in 1 mL ASW, resuspend, and repeat step 4. Up to 500 μ L of *Prochloron* cells can be obtained from an animal that fits in a standard Petri dish. These preparations are not pure, in the sense that often spicules and other bacteria are present. However, the resulting DNA preparations originate almost entirely from *Prochloron*.
7. In contrast to other reports, we have not found *Prochloron* to be particularly delicate. To be cautious, it is worth processing samples immediately after step 6.
8. To extract DNA in the field, numerous methods are possible. We usually use the Qiagen Genomic Tip kit, following the manufacturer's instructions. Lysozyme is not required to lyse *Prochloron*, making it unnecessary to transport frozen or refrigerated samples to the site of collection. The resulting DNA is stable at room temperature for travel and is used successfully in the synthesis of fosmid libraries after up to several years of storage at 4 °C. Whole ascidian samples can also be stored in a commercial nucleic acid preservation reagent, such as RNALater, at -20 °C for DNA extraction in the laboratory.

Lightweight, field-appropriate equipment can be purchased for all of these experiments and packed into a single container for travel. Of note, these types of procedures can also be used for many other symbiotic organisms, although the details of microbial isolation differ. Also of note, no refrigeration or freezing is required for these techniques, eliminating many of the problems associated with transporting goods and reagents for field research. 70% ethanol and isopropanol purchased at pharmacies are perfectly acceptable alternatives to the pure chemicals suggested in protocols.

3.3 DNA purification from whole animals

If *Prochloron* cannot be dissociated from samples, as in the case of *D. molle*, mixed samples can still be stored as above. Often, DNA can still be obtained using standard methods. However, this is problematic with highly mucous-laden animals such as *D. molle*, for which the following protocol is suggested.

1. Ascidians stored in RNALater or similar preservative are processed using this method, inspired by the large scale DNA isolation methods optimized for actinomycetes (Kieser et al. 2000).
2. Resuspend dissected pieces of the animal (or preferably freeze-dried) in 5 ml of TE buffer, add 10 mg lysozyme (final conc 2 mg/ml) and swirl to dissolve. Incubate at 30 °C for 30-60 min, monitoring every 15 min. On first sign of increased viscosity or after 1 h, proceed to step 3.

3. Add 1.2 ml 0.5 M EDTA (final conc 0.1 M), mix gently and add 0.13 ml of Proteinase K (final conc 0.2 mg/ml), mix gently and incubate at 30 °C for 5 min.
4. Add 6.5 ml 10% SDS (final conc 1%), tilt immediately and incubate at 37 °C for up to 2 h, checking for lysis. (This could be extended to overnight, depending on the speed of lysis.)
5. Add 1.2 ml of 5 M NaCl, mix, and add 1.0 ml of CTAB/NaCl solution (10% CTAB in 0.7 M NaCl; dissolve 4.1 g NaCl in 80 ml water, slowly add 10 g CTAB [hexadecyltrimethyl ammonium bromide] while heating and stirring. Adjust final volume to 100 ml.)
6. Incubate at 65 °C for 20 min (could be extended to 1 h).
7. Perform a standard phenol / chloroform extraction, followed by an isopropanol precipitation.

The scale of the extraction can be adjusted depending on the amount of sample to be processed.

Note: It is critical to follow all relevant travel and customs regulations when transporting scientific equipment and reagents. In addition, proper local and national permits must be obtained prior to collection.

4. Chemical analysis

Over 100 different cyanobactins have been structurally characterized, with an additional >30 identified from metagenomics experiments. Therefore, exact protocols depend upon the structural classes under consideration. The following applies to isolation of known cyanobactins from ascidians. These compounds tend to be of relatively low polarity in comparison to many from other sources. Samples stored in ethanol or isopropanol (70-100%) can be used to obtain these compounds, but it is generally better to extract fresh or frozen specimens if possible. In the procedure below, it is desirable to start with at least 50 g wet weight of ascidian if pure material is required, but the LC-MS method will work with a few mg of tissue (Donia et al. 2006).

1. Using frozen or lyophilized ascidians, cut tissue into small (~1-8 cm³) segments and place in at least a 10-fold excess of methanol. Leave at room temperature for several hours, decant, and extract at least twice more with fresh methanol. The last extraction step is usually left overnight. After proper extraction, the green color is almost entirely bleached from the sample.
2. Dry by rotary evaporation to remove methanol. Partition remaining water against ethyl acetate, or add water to dried samples and extract. Dry the ethyl acetate fraction and partition it between hexane and methanol. The methanol fraction will be greatly enriched for cyanobactins.
3. The dried methanol fraction can be further purified by silica or C₁₈ chromatography. The exact method depends upon the compounds in question. In order to rapidly determine which cyanobactins are present in a sample, the following represents a universal method.
4. Prepare a small fraction of the methanol extract for LC-MS by pushing it through a plug of end-capped C₁₈ resin in methanol. Dry this sample, which is pure enough for mass spectrometry.
5. In a typical LC-MS run, samples are dissolved in 1:1 water : acetonitrile and injected on a Waters Alliance 2795 HPLC, using an analytical C₁₈ column (Phenomenex) and an acetonitrile / water gradient at a flow rate of 0.2 ml / min. Gradient conditions are

initially 50% water, 50% acetonitrile to 1% water, 99% acetonitrile over 15 or 25 minutes and held at that level for another 5 minutes. 1% formic acid is used in the water and / or acetonitrile to enhance ionization. For high resolution HPLC mass spectrometry (HPLCHRESIMS), a Micromass Q-TOF mass spectrometer is used with a LockSpray running concurrently to ensure higher mass accuracy. This method is relatively universal for detection of cyanobactins (which are usually the major compounds in the methanol extract) from ascidians and for moderately non-polar cyanobactins from cyanobacteria. In all cases, we use authentic standards of peptides that have been purified to homogeneity from whole ascidians and analyzed by NMR and MS to confirm the identity of cyanobactins.

It is difficult to detect more polar cyanobactins by this method because of their near co-elution with salts. Other methods have been described for these compounds. In our hands, it is not useful to use the host taxonomy in deconvoluting the masses obtained as cyanobactins are usually produced in diverse hosts. For example, trunkamide was only reported from *Lissoclinum patella* prior to our work; however, we purified it from *Didemnum molle* (Donia et al. 2008). Finally, the presence of a two-mass-unit difference in the mass spectrometric trace is usually a sign of different oxidation states (Donia et al. 2006).

5. Cyanobactin gene cloning and identification

Cyanobactin pathway clusters can be obtained either from whole DNA samples or from large-insert libraries (Donia et al. 2006, Donia et al. 2008, Long et al. 2005, Schmidt et al. 2005). In addition, cyanobactin clusters are readily recognized in genome sequences, leading to useful structure prediction from genomes (Donia et al. 2008, Sudek et al. 2006, Ziemert et al. 2008). Cloned pathways are useful in providing new natural products from miniscule environmental samples by heterologous expression (Donia et al. 2008).

5.1 Genome mining and structure prediction

To identify cyanobactin clusters in genome sequences, it is sufficient to perform a BLAST search for homologs of PatA, PatD, PatF, or PatG. Once the contig containing this homolog is identified, areas surrounding that gene should be carefully annotated by translation in both directions and all frames. All cyanobactin genes identified to date are clustered in one locus of the genome. However, some clusters appeared to be truncated or interrupted due to insertion events, which probably inactivate the pathway.

Genome sequences can be used for accurate structure prediction, as previously demonstrated. One of the key features allowing accurate prediction is the substrate specificity of PatA and PatG homologs, which defines the N-C sequences of the actual cyanobactin products. Alignment of existing PatE sequences identifies G (hydrophobic) X A/P S, where X is usually D or E, as the recognition site for PatA homologs, while A/S F/Y D is the substrate for PatG homologs. The canonical PatG protease recognizes AYD and SYD with apparently equal affinity, while PatA protease is known to recognize GLEAS, GVEPS, and GVDAS. All known PatA and PatG relatives recognize similar sequences, even if the proteins are only ~ 60% identical. Thus, a peptide with a leader sequence terminating in ...GVEPSITVCISVCAYDGE can be reliably predicted to synthesize the underlined peptide. The presence of PatD and / or PatF and application of a phylogenetic analysis can be used to determine whether a heterocycle is formed. The presence or absence of an oxidase can be used to determine whether that heterocycle is oxidized. Structural hypotheses can be readily differentiated with extremely small samples by applying MS/MS techniques using FT-ICR (Sudek et al. 2006). Because the amino acid sequences are directly genome-encoded, and in many cases whole genome sequences are available, high-resolution MS/MS peptide sequencing experiments are absolutely definitive as to structure. *patE* homologs are often not predicted to encode a protein

by auto-annotation tools, so it pays in some cases to repeat the analysis of sequenced regions. For example, this was necessary in the annotation of *Lyngbya aestuarii* and *Trichodesmium erythraeum*.

A caution is that individual cyanobactin gene types are often found outside of cyanobactin clusters and are probably not associated with cyanobactin synthesis. For example, a PatD homolog is found in the microcin-like pathway leading to goadsporin biosynthesis in actinomycetes (Onaka et al. 2005). PatA- and PatG-like proteases are found outside of cyanobactin pathways. The oxidase domain is also found in other pathway types, including nonribosomal peptide synthetases. Thus far, we have not found PatF homologs outside of a cyanobactin gene cluster, although there is at least one cyanobactin cluster lacking PatF. Since N-C cyclization requires PatA and PatG, a cyanobactin cluster should at minimum contain homologs of both of these proteases and a PatE-like precursor peptide. This is the minimum gene set required to make cyanobactin N-C cycles, although we have never found a pathway containing just this set.

Another issue, as described above, is how to determine whether or not a pathway is prenylating. We have derived a phylogenetic model (Donia and Schmidt, manuscript in preparation) that is useful in determining whether a pathway is prenylating or heterocyclizing. In this model, modifying enzymes as well as the precursor peptide sequences are correlated to the type of post-translational modification observed.

5.2 Cyanobactin cloning from *Prochloron* spp.

To clone pathways from DNA samples, several strategies can be attempted. Because cyanobactin pathways from *Prochloron* are so similar to each other, it is possible to clone entire, intact pathways out of ascidian samples using primers that exactly match the 5'-end of PatA and the 3'-end of PatG (Donia et al. 2008). These pathways can be directly transferred to *E. coli* for expression. A limitation to this strategy is that there is usually more than one cyanobactin pathway in a single animal. Therefore, sequencing over the PatE homolog is required to confirm that the desired pathway has been obtained. If DNA is degraded or is problematic, fragments of the desired gene sequence can be directly cloned. For example, a prenylated cyanobactin pathway can be cloned as follows:

1. Exactly matching primers internal to PatD and PatF homologs are designed and applied to the sample;
2. The resulting product is sequenced. If it encodes the desired peptide, homologous recombination can be used to make hybrids with an intact prenylating pathway for expression in *E. coli*. Methods for performing this cross are described in section 9 below. This method has been used successfully to express the anticancer clinical candidate, trunkamide, in *E. coli* culture.

5.3. Cyanobactin cloning from non-*Prochloron* cyanobacteria

When cloning from organisms other than *Prochloron*, DNA sequences are not similar enough to use an exact-match strategy. In these cases, a degenerate priming strategy is required. In order to obtain whole pathways, they must then be cloned out of DNA libraries, such as cosmid or bac libraries. We have previously reported a degenerate strategy to obtain the tenuencyclamides pathway from a free-living cyanobacterium (Donia et al. 2008).

Genomic DNA from *Nostoc* sp. was cloned into the pCC1FOS fosmid vector (Epicentre) following the manufacturer's protocol. A library of about 800 colonies was constructed. Colonies were then picked into 96-well plates and stored in 20% glycerol. Using a Clustal X alignment of different members, conserved regions in cyanobactin proteases were identified.

Degenerate primers were then designed using CODEHOP to specifically amplify cyanobactin proteases and not any other serine proteases. The primers used were Protease-F (TGGGCACCCTCGGCTAYGAYTTYGG) and Protease-R (GCGTTCACGTTCCAGCCRTANADNCC). This pair of primers was used initially to amplify a ~400 bps cyanobactin protease region. Exact match primers were then designed based on the obtained sequence and used to screen the fosmid library. Each 96 well plate was replicated in one tube by the use of a 96-well replicator immersed in the plate then in sterile water. This mixed preparation of fosmids proved to be an actual representation of every colony in the plate and was screened directly by PCR. After identifying one plate containing the desired sequence, further screening was performed to identify the appropriate single clone. This time, one mixed sample was prepared from each row and each column by mixing 10-20 μ l from each well. A plasmid from the cyanobactin-containing clone was prepared and used for PCR to further prove the screening success. This clone containing the intact *ten* cluster was then sequenced to 6x coverage.

6. Heterologous expression in *E. coli*

A key advantage of many cyanobactin pathways is that they can be expressed in *E. coli*. In our hands, removal of PatA, PatD, PatE, PatF, or PatG homologs from a pathway results in no detectable product *in vivo*, while PatB and PatC homologs can be safely left out of expression constructs if desired (Donia et al. 2006). The best yields are obtained when genes are left intact in apparent operons, although placing individual promoters in front of every gene also leads to cyanobactin production in *E. coli*. There is a report that cyanobactins are extremely robustly expressed from large plasmids such as bacmids (Long et al. 2005), but we have been unable to replicate this result despite extensive investigation. It may be that certain pathways from some samples are more readily expressed in the completely native context, while most are much better expressed after some minor engineering.

The most robust expression we have obtained so far was by cloning the entire patellin biosynthetic pathway (Donia et al. 2008). PCR with exact-match primers provided the entire ~11 kb pathway, which was cloned into the pCR2.1 vector. Fortuitously, PatA was cloned in-frame with the *lacZ* gene N-terminal fragment found in the vector. Patellins were expressed from ~100 μ g-1 mg per L, and expression was shown to be responsive to lactose and glucose concentrations. Expression is performed as follows:

1. Due to rearrangement problems, all experiments are performed at 30 °C.
2. A freshly picked colony of Top10' cells containing the above described vector was seeded overnight into 50 mL LB with appropriate antibiotic selection. The resulting seed culture is transferred to a 10 L fermenter in the same medium and allowed to grow for 72 h. Aeration is performed with 7.5 cc/min airflow and 250 rpm mixing.
3. After 3 days, the culture is extracted by addition of ~10 g HP20 resin. The resin is washed with water (1.5 L) and eluted with methanol (1L). Following rotary evaporation of the methanol fraction, the residue is dissolved in water (250 ml) and extracted three times with equal amounts of ethyl acetate.
4. The resulting ethyl acetate fraction is dried and fractionated through a C₁₈ column. Fractions are then dried, dissolved in 1:1 acetonitrile: water and analyzed by HR-HPLCMS.

A problem that is sometimes encountered is that individual starting *E. coli* colonies produce different levels of compounds, leading to variability in production.

7. Deep metagenome mining

This process refers to the rapid comparison of many closely related pathways to determine functional differences that are useful in engineering. The process will be described with ascidians containing *Prochloron* as an example. In principle, this procedure could be applied to any pathway type involving symbiotic interactions.

First, obtain sample ascidians as described above. DNA can be extracted, or PCR can be performed directly on enriched *Prochloron* cells. The protocol we developed for *Prochloron* PCR works mainly for small PCR products, especially *patE* variants. Bigger PCR products require better quality DNA isolated by any of the methods described above.

7.1 Whole-cell PCR with *Prochloron*

1. Make a 10x betaine solution: add 16 μ l of betaine (6.5 M) to 0.5 μ l DMSO.
2. Add up to 10 μ l *Prochloron* cells to 200 μ l DMSO and grind the mixture with a tissue grinder.
3. Make serial dilution to 1/10 x and 1/100 x in sterile water.
4. Perform PCR using each dilution (2.5 μ l) and the following components, diluted to 20 μ l per reaction: 10x betaine solution (2 μ l); recommended amounts of buffer and Mg, dNTP mix (0.4 μ l of 10 mM solution); HiFi Taq polymerase (Invitrogen, 0.1 μ l). In addition, each primer is added to give a final concentration of 1 mM.

7.2 Example: deep metagenome mining in *Prochloron*

1. To determine regions of interest, use PCR to obtain fragments of the whole pathway using multiple overlapping exact match primers. At this step, products must be cloned prior to sequencing. Sequence analysis of overlapping segments reveals which portions are likely to be variable and which are conserved. In the case of cyanobactins, we noticed that only the hypervariable region of PatE contained consistent mutations.
2. PCR over candidate variable regions “deeply,” i.e. perform PCR using multiple, closely related samples, preferably of mixed (metagenomic) origin. In the cyanobactin case, we performed PCR from a region inside PatF to a region inside PatD. In this way, we could monitor for mutations within proteins and intergenic regions easily. In addition, PatE itself was analyzed by direct PCR from samples. Although it is preferable to examine as many samples as possible, in fact numerous variable products originated from single animal samples, revealing a diverse mixed colony within a single animal. In principle, therefore, a single animal could be sufficient for some deep metagenome projects.
3. Clone products and sequence multiple isolates in 96-well plates. Use at least two PCR reactions per sample to minimize the probability of identifying PCR mutations as opposed to natural mutations. In the case of cyanobactins, we sequenced 12 96-well plates, identifying 30 cyanobactin pathway relatives. Because flanking sequences were available, it was possible to determine rapidly that other enzymatic changes did not contribute to product diversity.
4. For a few representative samples, sequence pathway clusters completely to ensure that the observed mutations are focused in the identified variable regions and not elsewhere.

Using this methodology, it was possible to examine numerous mutations in a short time frame. Finally, as mentioned above, comparison of prenylating and non-prenylating pathways also followed a variation of the deep metagenome mining method.

8. Enzymatic analysis of cyanobactin biosynthesis

Thus far, we have reported the activities of PatA and PatG proteins (Lee et al., manuscript submitted), but not yet PatD or PatF. A critical advance allowing us to perform this analysis was as follows. We experimented with many different PatE peptides, but they were very difficult to handle, often precipitating from solution and not reacting well in enzyme assays. In addition, PatE relatives were often not detected on SDS-PAGE with various stains, despite a huge abundance of the peptides as shown by tryptic digest-mass spectrometry. The advance came by using N-terminally His-tagged PatEdm, an engineered peptide containing the wholly artificial “eptidemnamide” sequence in cassette 2. This peptide was better behaved because it has only two Cys residues, rather than the four to six commonly found in these small peptides, and moreover it was readily detectable by SDS-PAGE. In addition, a key step in the successful purification was to purify under denaturing conditions in urea, followed by exhaustive dialysis against water over several days. This protocol provided the best yield of precursor.

Preparation of PatA and PatG was relatively more standard and will not be described here. As an outcome of these experiments, we were able to show that PatA cuts the N-terminal recognition sequence, while PatG cuts the C-terminus and cyclizes the peptides. This represented the first natural ribosomal peptide cyclase to be characterized *in vitro* as a pure protein.

9. Applying deep metagenome mining: pathway engineering

As described above, it is possible to modify just the PatE hypervariable region to generate new products. Larger regions must be modified in order to switch from non-prenylating to prenylating pathways. Much remains unknown about the size limitations of the pathways. Several methods have been used to generate new molecules *in vivo*. Standard PCR mutagenesis was used to synthesize PatEdm from the natural PatE2 gene. Whole pathways have been expressed either as an operon under control of the *lac* promoter or in individual plasmids under control of T7 promoters. In order to manipulate whole operons, we have used recombination in yeast and by lambdaRed (Gust et al. 2003). The latter protocol uses the standard method as described. The only exception is that we use a multi-copy plasmid, so the correct mutant must be obtained by various purification methods, such as re-streaking followed by colony PCR or Southern hybridization. This is necessary because the lambdaRed method generates individual colonies containing mixtures of plasmids, and at least in our hands this problem is greatly exacerbated with multi-copy plasmids. However, it is often technically easier to re-purify these mixtures than to use single-copy vectors.

We also use the yeast recombination method for the rapid manipulation of pathways. The example described below enabled us to produce the rare compound, trunkamide, which is a potential anticancer agent occasionally found in ascidian samples. The trunkamide pathway is very closely related to that of patellin biosynthesis. Previously, we had cloned the patellin biosynthetic cluster. Thus, we rescued the trunkamide genes for heterologous expression in *E. coli* using yeast recombination. The following is a modification of a previously reported protocol (Kunes et al. 1987, Sikorski and Hieter 1989).

9.1 Yeast recombination for pathway manipulation

1. The yeast-*E. coli* shuttle vector harboring the original *truI* cluster is linearized near the site of modification by restriction digestion.
2. The region harboring a desired modification is amplified by PCR or synthesized. In this example, we amplified the *truE2* gene with sufficient flanking regions to allow homologous recombination.

3. The two linear pieces from steps 1 and 2 are mixed and used to transform yeast following the standard high-lithium yeast transformation protocol:
 - a. Inoculate 0.5 ml of stationary phase cells into 10 ml of fresh YPD medium, which should give an initial O.D.600 of 1.0. Shake the culture in a 30 °C incubator for about 4 h, until the O.D.600 is about 4.0.
 - b. Spin cells at 3000 rpm for 5 min. Remove supernatant and resuspend cells in 5 ml of lithium solution (10 mM Tris PH 7.5 , 1 mM EDTA pH 8.0, 0.1 M lithium acetate).
 - c. Spin again, pour off supernatant. Resuspend cells in residual lithium solution remaining in the tube. Transfer the mixture to a microcentrifuge tube.
 - d. Boil salmon sperm DNA (10 mg/ml) for 10 min. Place on ice before adding to yeast cells.
 - e. Add DNA to a microcentrifuge tube. Add 50 µl of yeast cells and 7 µl of salmon sperm DNA to the DNA solution. Mix well. Add 0.45 ml of 40 % PEG solution (10mM Tris pH 7.5, 1mM EDTA pH 8.0, 0.1 M lithium acetate, 40% PEG 4000 or 3350) to each tube. Mix well.
 - f. Incubate at 30 °C for 20 min.
 - g. Heat-shock the cells at 42 °C for 15 min. Leave cells at room temperature thereafter.
 - h. Spin at 4000 rpm, resuspend in buffer TE (400 µl), and plate 100-150 µl of the mixture onto selective medium plates. Incubate at 30 °C for 2-3 days.
4. Combine at least 15-20 yeast transformant colonies and transfer to 10 ml YPD. Grow overnight until a high density is achieved, then rescue plasmids using standard yeast miniprep procedures. Lyticase incubations prior to mini-preps are sometimes helpful to facilitate recovery of larger plasmids. We usually use uracil auxotrophy as a convenient yeast selection marker.

Acknowledgments

This work was funded by NIH GM071425. We are grateful for the participation of our colleagues led by Margo Haygood (OHSU) and Jacques Ravel (U. Maryland) in these endeavors.

References

- Burns BP, Seifert A, Goh F, Pomati F, Jungblut AD, Serhat A, Neilan BA. Genetic potential for secondary metabolite production in stromatolite communities. *FEMS Microbiol. Lett* 2005;243:293–301. [PubMed: 15668032]
- Carroll AR, Coll JC, Bourne DJ, MacLeod JK, Zabriskie TM, Ireland CM, Bowden BF. Patellins 1-6 and trunkamide A: novel cyclic hexa-, hepta- and octa-peptides from colonial ascidians, *Lissoclinum* sp. *Aust. J. Chem* 1996;49:659–667.
- Colin, PL.; Arneson, C. *Tropical Pacific Invertebrates*. Coral Reef Press; Beverly Hills: 1995.
- Donia MS, Hathaway BJ, Sudek S, Haygood MG, Rosovitz MJ, Ravel J, Schmidt EW. Natural combinatorial peptide libraries in cyanobacterial symbionts of marine ascidians. *Nat. Chem. Biol* 2006;2:729–735. [PubMed: 17086177]
- Donia MS, Ravel J, Schmidt EW. A global assembly line for cyanobactins. *Nat. Chem. Biol* 2008;4:341–343. [PubMed: 18425112]
- Donia, MS.; Schmidt, EW. *Comprehensive Natural Products II*. 2009. Cyanobactins, ubiquitous cyanobacterial ribosomal peptide metabolites. in press

- Duquesne S, Destoumieux-Garzon D, Peduzzi J, Rebuffat S. Microcins, gene-encoded antibacterial peptides from enterobacteria. *Nat. Prod. Rep* 2007;24:708–34. [PubMed: 17653356]
- Gust B, Challis GL, Fowler K, Kieser T, Chater KF. PCR-targeted *Streptomyces* gene replacement identifies a protein domain needed for biosynthesis of the sesquiterpene soil odor geosmin. *Proc. Natl. Acad. Sci. U S A* 2003;100:1541–1546. [PubMed: 12563033]
- Hirose M, Yokobori S, Hirose E. Potential speciation of morphotypes in the photosymbiotic ascidian *Didemnum molle* in the Ryukyu Archipelago, Japan. *Coral Reefs*. 2008 epub.
- Ireland C, Scheuer PJ. Ulicyclamide and ulithiacyclamide, two new small peptides from a marine tunicate. *J. Am. Chem. Soc* 1980;102:5688–91.
- Kieser, T.; Bibb, MJ.; Buttner, MJ.; Chater, KF.; Hopwood, DA. *Practical Streptomyces Genetics*. John Innes Foundation; Norwich: 2000.
- Kott P. The Australian Ascidiacea part 4, Aplousobranchia (3), Didemnidae. *Mem. Qld. Mus* 2001;4:1–408.
- Kunes S, Ma H, Overbye K, Fox MS, Botstein D. Fine structure recombination analysis of cloned genes using yeast transformation. *Genetics* 1987;115:73–81. [PubMed: 3549444]
- Lafargue F, Duclaux D. Premier exemple, en Atlantique tropical, d'une association symbiotique entre une ascidie didemnidae et une cyanophycée chroococcale: *Trididemnum cyanophorum* nov. sp. et *Synechocystis trididemni* nov. sp. *Ann. Inst. Océanogr* 1979;55:163–184.
- Lewin, RA.; Cheng, L., editors. *Prochloron: A Microbial Enigma*. Chapman and Hall; New York: 1989.
- Long PF, Dunlap WC, Battershill CN, Jaspars M. Shotgun cloning and heterologous expression of the patellamide gene cluster as a strategy to achieving sustained metabolite production. *ChemBioChem* 2005;6:1760–1765. [PubMed: 15988766]
- Onaka H, Nakaho M, Hayashi K, Igarashi Y, Furumai T. Cloning and characterization of the goadsporin biosynthetic gene cluster from *Streptomyces* sp. TP-A0584. *Microbiology* 2005;151:3923–33. [PubMed: 16339937]
- Pidot SJ, Hong H, Seeman T, Porter JL, Yip MJ, Men A, Johnson M, Wilson P, Davies JK, Leadley PF, Stinear TP. Deciphering the genetic basis for polyketide variation among mycobacteria producing mycolactones. *BMC Genomics* 9:462. [PubMed: 18840298]
- Schmidt EW. From chemical structure to environmental biosynthetic pathways: navigating marine invertebrate-bacteria associations. *Trends Biotechnol* 2005;23:437–440. [PubMed: 16038996]
- Schmidt, EW. *Nat. Chem. Biol.* 2008. Trading molecules and tracking targets in symbiotic interactions. in press
- Schmidt EW, Nelson JT, Rasko DA, Sudek S, Eisen JA, Haygood MG, Ravel J. Patellamide A and C biosynthesis by a microcin-like pathway in *Prochloron didemni*, the cyanobacterial symbiont of *Lissoclinum patella*. *Proc. Natl. Acad. Sci. U. S. A* 2005;102:7315–7320. [PubMed: 15883371]
- Schmidt EW, Obraztsova AY, Davidson SK, Faulkner DJ, Haygood MG. Identification of the antifungal peptide-containing symbiont of the marine sponge *Theonella swinhoei* as a novel delta-proteobacterium, “*Candidatus Entotheonella palauensis*”. *Mar. Biol* 2000;136:969–977.
- Schmidt EW, Sudek S, Haygood MG. Genetic evidence supports secondary metabolic diversity in *Prochloron* spp., the cyanobacterial symbiont of a tropical ascidian. *J. Nat. Prod* 2004;67:1341–1345. [PubMed: 15332852]
- Sikorski RS, Hieter P. A system of shuttle vectors and yeast host strains designed for efficient manipulation of DNA in *Saccharomyces cerevisiae*. *Genetics* 1989;122:19–27. [PubMed: 2659436]
- Sudek S, Haygood MG, Youssef DTA, Schmidt EW. Structure of trichamide, a cyclic peptide from the bloom-forming cyanobacterium *Trichodesmium erythraeum*, predicted from the genome sequence. *Appl. Environ. Microbiol* 2006;72:4382–4387. [PubMed: 16751554]
- Wilkinson B, Mickelfield J. Mining and engineering natural-product biosynthetic pathways. *Nat. Chem. Biol* 2007;3:379–386. [PubMed: 17576425]
- Ziemert N, Ishida K, Quillardet P, Bouchier C, Hertweck C, Tandeau de Marsac N, Dittmann E. Microcyclamide biosynthesis in two strains of *Microcystis aeruginosa*: from structure to genes and vice versa. *Appl. Environ. Microbiol* 2008;74:1791–1797. [PubMed: 18245249]

Figure 1.

Representative cyanobactin structures. Both trunkamide and ulithiacyclamide are from symbiotic *Prochloron* strains living with ascidians. Trichamide is from free-living cyanobacteria. Trunkamide exemplifies a prenylated cyanobactin, while the others are modified only by N-C cyclization and heterocyclization.

Figure 2.

Prochloron-containing didemnid ascidians from a tropical reef. The green color in these animals is due to symbiotic cyanobacteria, *Prochloron* spp. Three species of didemnid ascidians are shown, including a bluish encrusting form, a plate-like green form, and the large animal in the center of the frame, *Lissoclinum patella*. This picture was taken in the Solomon Islands by Chris Ireland (U. Utah).

Figure 3.

Cyanobactin gene clusters. At top, the *pat* gene cluster from *Prochloron* is shown. This leads to heterocyclic small molecules such as ulithiacyclamide. Second from top is the *tru* cluster from *Prochloron*, which leads to prenylated cyanobactins such as trunkamide. These clusters are ~99% identical to each other (red regions) except in the region that is important for prenylation (blue region). Note that pathways from non-symbiotic cyanobacteria have a lower level of identity to other known pathways, highlighting information that can be obtained by deep metagenome mining. Dashed lines indicate changes in gene order.

87654321

>PatE1 MNKKNILPQQGQPVIRTAGQLSSQLAELSEEALGDAGLEASVTACITFCAYDGVEPSITVCISVCAYDGE
 >PatE2 MNKKNILPQQGQPVIRTAGQLSSQLAELSEEALGDAGLEASVTACITFCAYDGVEPSCTLCCTLCAYDGE
 >PatE3 MNKKNILPQQGQPVIRTAGQLSSQLAELSEEALGDAGLEAS:ACFPPTICAYDGVEPS:FCFPPTVCAYDGE

Figure 4.

Deep metagenome mining with *pat* pathway. The three PatE peptide variants shown encode the three cyanobactin molecules in their second cassettes (numbered 1-8 above the sequences). The underlined sequences encode cyanobactin products, while sequences in bold are sequences recognized by the PatA and PatG proteases that define the boundaries of the final products. Note that the peptides are exactly identical except where they encode these different products. They are also identical at the DNA level. In addition, all three *patE* variants are flanked by identical genes encoding identical enzymes.

Figure 5.

Comparison of amino acid use in cyanobactins by (a) *Prochloron*; (b) all cyanobacteria; (c) the engineered peptide, eptidemnamide. Note that, in natural cyanobactins, either Pro or other heterocycles derived from Cys / Ser are always present in position 1. This was replicated with Pro in position 1 of eptidemnamide. In position 2, the natural molecules never contain Cys and virtually never contain Ser / Thr. In eptidemnamide, position 2 is taken by Trp, which is almost never found in natural cyanobactins. The other positions in natural cyanobactins are more varied, although they contain a high percentage of Cys, Pro, Ser/Thr, and Hy (hydrophobic amino acids Ala, Ile, Leu and Val). Very few Gly or other more hydrophilic amino acids are found, and these amino acids are virtually never found in *Prochloron*-derived cyanobactins. In eptidemnamide, these amino acids constitute the last six amino acids. This figure illustrates the degree to which deep metagenome mining coupled with chemical analysis illuminates engineering principles. Questions that could not be answered by observing natural diversity were addressed by the construction of the artificial molecule, eptidemnamide. Finally, this

figure does not include compounds predicted from sequence, but which have not yet been purified.

Table 1

Mutations in the PatE sequence identified by deep metagenome mining. Position number is from the C-N terminus to be consistent with Figures [3] and 4. The numbers in the rows refer to the number of times an individual amino acid is found in that position, out of 34 mutants identified. In all of these cases, the pathway enzymes were almost 100% identical, and mutations were only identified within the 24-nt product-coding region of the patE-homologous genes. Note that every site is mutated at least once, revealing broad-substrate selectivity for the pathway enzymes

Position	1	2	3	4	5	6	7	8
Ala	0	1	4	1	0	9	2	0
Arg	1	0	0	0	0	0	0	0
Cys	33	0	0	7	24	6	0	8
Ile	0	1	0	13	1	0	1	3
Leu	0	9	0	0	1	7	1	5
Phe	0	11	0	1	4	0	5	0
Pro	0	1	9	0	0	2	0	0
Ser	0	1	2	0	0	2	1	1
Thr	0	0	27	1	0	3	23	0
Tyr	0	0	0	1	0	0	0	0
Val	0	11	0	1	4	5	0	7
None	0	0	0	0	0	0	1	8