Diversity of monomers in nonribosomal peptides: towards the prediction of origin and biological activity.

Ségolène Caboche, Valérie Leclère, Maude Pupin, Gregory Kucherov, Philippe Jacques

JOURNAL OF BACTERIOLOGY, Oct. 2010, p. 5143–5150 Vol. 192, No. 19
0021-9193/10/$12.00 doi:10.1128/JB.00315-10

Diversity of Monomers in Nonribosomal Peptides: towards the Prediction of Origin and Biological Activity▽†

Ségolène Caboche,[1,2]* Valérie Leclère,[1] Maude Pupin,[2] Gregory Kucherov,[2] and Philippe Jacques[1]

ProBioGEM (UPRES EA 1026), Université Lille Nord de France, USTL, Polytech-Lille/IUTA, F59655 Villeneuve d'Ascq, France,[1] *and LIFL, UMR USTL/CNRS 8022, INRIA Lille-Nord Europe, F59655 Villeneuve d'Ascq, France*[2]

Received 22 March 2010/Accepted 26 July 2010

Nonribosomal peptides (NRPs) are molecules produced by microorganisms that have a broad spectrum of biological activities and pharmaceutical applications (e.g., antibiotic, immunomodulating, and antitumor activities). One particularity of the NRPs is the biodiversity of their monomers, extending far beyond the 20 proteogenic amino acid residues. Norine, a comprehensive database of NRPs, allowed us to review for the first time the main characteristics of the NRPs and especially their monomer biodiversity. Our analysis highlighted a significant similarity relationship between NRPs synthesized by bacteria and those isolated from metazoa, especially from sponges, supporting the hypothesis that some NRPs isolated from sponges are actually synthesized by symbiotic bacteria rather than by the sponges themselves. A comparison of peptide monomeric compositions as a function of biological activity showed that some monomers are specific to a class of activities. An analysis of the monomer compositions of peptide products predicted from genomic information (metagenomics and high-throughput genome sequencing) or of new peptides detected by mass spectrometry analysis applied to a culture supernatant can provide indications of the origin of a peptide and/or its biological activity.

Nonribosomal peptides (NRPs) are molecules produced by microorganisms and synthesized by huge multienzymatic complexes (38, 41), called *n*on*r*ibosomal *p*eptide *s*ynthetases (NRPSs). These megaenzymes are organized into modules, one for each amino acid to be built into the peptide product. This is accomplished by division of each catalytic step into specialized semiautonomous domains. The basic set of domains (adenylation, thiolation, and condensation) within a module can be extended by substrate-modifying domains, including domains for substrate epimerization, β hydroxylation, N methylation, and heterocyclic ring formation. The peptide release is catalyzed by a thioesterase domain which can also, in many cases, be involved in an intramolecular reaction leading to a cyclic or partially cyclic peptide or, in fewer cases, in the oligomerization of peptide units (iterative biosynthesis). NRPs show a broad spectrum of biological activities and pharmaceutical applications. They can harbor antimicrobial, immunomodulator, or antitumor activities. Cyclosporine (5), an immunosuppressant drug widely used in organ transplantation, daptomycin (60) (marketed in the United States under the trade name Cubicin), used in the treatment of certain infections caused by Gram-positive bacteria, aminoadipyl-cysteinyl-valine (ACV)-tripeptide, which is the precursor of cephalosporin and penicillin (29), the most famous antibiotic, and also bleomycin (57), used in the treatment of several cancers, are some common examples of NRPs of high therapeutic importance. Two main structural traits distinguish these peptides from ribosomally synthesized peptides: first, their primary structure is more frequently cyclic (partially or totally) branched or polycyclic rather than linear and, second, the biodiversity of monomers incorporated in NRPs goes far beyond the 20 proteogenic amino acids residues. NRP monomers include modified versions of the proteogenic amino acids (e.g., methylated, hydroxylated, and D-forms) but also other monomers, such as, for example, 2-aminoisobutyric acid (Aib), hydroxyphenylglycine (Hpg), and 2,3-dihydroxybenzoic acid (diOH-Bz). However, essential characteristics of this diversity and its relationship with biological functions and producing organisms have been poorly understood until now.

The development of the Norine database, the first resource entirely dedicated to NRPs (8, 9), filled this gap. Based on Norine data, we performed the first large-scale analysis of about a thousand peptides which represent a total coverage of more than 10,000 monomer occurrences, revealing the presence of as many as 500 different monomer types. A data-mining analysis of the monomeric compositions of NRPs allowed us to reveal a strong relationship between certain monomeric characteristics of NRPs and their biological function and producing organism. In addition to providing a comprehensive overview of monomeric biodiversity in NRPs, this work demonstrated (i) a dissimilarity of structural properties between bacterial and fungal NRPs; (ii) a significant relationship between NRPs synthesized by bacteria and those isolated from metazoa, especially from sponges, supporting the hypothesis that the peptides isolated from sponges are in reality synthesized by symbiotic bacteria rather than by the sponges themselves; and (iii) a certain monomer specificity to a class of biological activities. Those observations are supported by successful statistical predictions of biological activities of NRPs based on their monomeric compositions.

* Corresponding author. Mailing address: ProBioGEM (UPRES EA 1026), Université Lille Nord de France, USTL, F59655 Villeneuve d'Ascq, France. Phone: 33(0) 328 76 7440. Fax: 33(0) 328 76 7356. E-mail: segolene.caboche@lifl.fr.

† Supplemental material for this article may be found at http://jb.asm.org/.

▽ Published ahead of print on 6 August 2010.

TABLE 1. Repartition of NRPs in groups showing great diversity

Group	All NRPs ($n = 9$)	Only curated NRPs ($n = 4$)
Dolastatines	4	4
Kahalalides	16	16
Pyoverdins	57	57
Serrawettins	2	2
Guineamides	6	
Hymenamides	10	
Kapakahines	5	
Phakellistatins	14	
Stylopeptides	2	

MATERIALS AND METHODS

NRP set. Here, we define the data sets we used in our analyses. We started with the whole set of the first 1,071 peptides stored in the Norine database (http://bioinfo.lifl.fr/norine). Based on the annotations of the Norine database, we selected several training sets as described below.

First, we distinguished between curated and putative NRPs as annotated in the Norine database. For curated peptides, either corresponding synthetase genes have been identified or their nonribosomal origin has been universally accepted by the scientific community, as is the case for polytheonamide, for example (32). For putative peptides, there is no experimental evidence of their synthesis pathway, and other characteristics, such as their nonlinear structure and/or found nonproteogenic amino acids, suggest their nonribosomal origin. Examples are oscillarin (27) and aurilide (62). Of the 1,071 Norine peptides, 790 (that is, nearly three-quarters) are curated.

Second, some Norine peptides are considered variants belonging to the same group. Until now, no universal definition of the NRP variants has been proposed. Most of the time, peptides are called variants if they show similar compositions. For example, surfactin (3) and [Val7]surfactin (44) differ by only one monomer at position 7. More rarely, variants may have different structures and/or sizes, such as cyclic gramicidin S (11), composed of 10 monomers, and linear gramicidin A (31), composed of 16 monomers. However, some peptides can also be considered variants when they share a specific function or have features in common. For example, pyoverdins (67) form a large group of siderophores (7) produced by species of *Pseudomonas* or a related genus; they have a large diversity in structure and monomeric composition but have features in common, such as the presence of a fluorescent chromophore.

To reduce a bias in learning structural or compositional properties of NRPs belonging to different groups, we eliminated close variants and kept only one variant per group. For example, in some groups, only few monomers vary, and therefore invariable monomers of these groups may be given an overestimated significance. In order to identify groups for which we need to keep only one representative variant, we computed an average distance between all variants of the group (see the supplemental material). If the peptides of a given group were similar, i.e., the average phylogenetic distance in the group was low, we kept only one random member of this group; otherwise, we kept all the peptides of the group. The 1,071 NRPs (790 curated) were divided into 183 groups (100 curated), of which 62 groups (24 curated) contain only one peptide. Among the 121 groups (76 curated) containing at least two variants, 9 (4 curated) showed a high diversity (Table 1): dolastatines (47), guineamides (64), hymenamides (58), kahalalides (26), kapakahines (68), phakellistatins (42), pyoverdins (67), serrawettins (35), and stylopeptides (6). For those groups, we kept all the variants in all analyses. We want to mention that 130 peptaibols (16) are stored in the Norine database, divided into 20 groups of variants. All the peptaibols are curated NRPs. This means that there are 20 representative peptaibols among 290 (175 curated) representative NRPs. When we consider a data set containing only one variant per group (except for the 9 groups mentioned above), we use the term "excluding variants." The NRP set excluding variants contains 290 peptides, and the curated NRP set excluding variants contains 175 peptides.

Correlation coefficient. The correlation coefficient (CC) is computed in order to evaluate the relationship between two data sets. It is comprised between -1 and 1; the closer the CC is to these extreme values, the more the data are correlated. In two series, X ($x1, \ldots, xn$) and Y ($y1, \ldots, yn$), the CC is computed as follow:

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

In our case, X and Y will be either two monomeric distributions, i.e., the number of monomer occurrences in two sets of NRPs, or two distributions of peptide sizes.

RESULTS

The arrival of the Norine database provided us with the possibility of obtaining a general overview of NRPs. Below we present three analyses that we performed on Norine data. First, we studied some general statistical characteristics of NRPs, and then we focused on properties related to producing organisms; finally, we analyzed the monomeric distribution depending on biological activities. The observations that we drew from these analyses should be used for predicting the activities and origins of newly identified products.

General statistics. To avoid a bias in our analysis, we chose to restrict the data set for general statistics to the curated peptides excluding variants, which refers to 175 peptides (see Materials and Methods). However, the results obtained with the total set of NRPs lead to the same conclusions (data not shown).

(i) How variable are nonribosomal peptide structures? It can be deduced from Norine data that nonlinear structures represent nearly three-quarters of the NRP structures (Fig. 1). The majority (64%) of NRPs contain at least one cycle, but only a few peptides (1%) possess only branchings, and up to 8% present complex structures with overlapping cycles and branching. Those complex structures are found mainly in glycopeptides (15), a large group of antibiotics, the most famous of which is vancomycin (Fig. 2), used in treatment of infections caused by Gram-positive bacteria (28).

(ii) What is the size of a nonribosomal peptide? We have defined the size of an NRP to be the total number of monomers, with fatty acids in lipopeptides being considered individual monomers, as they are most frequently added as a single block by a condensation domain (as in arthrofactin synthesis [51]), or by using a single module in PKS-NRPS hybrid synthetases (as in the mycosubtilin synthetase [17]). As the synthetases are generally constituted of as many modules as monomers incorporated into the peptide (except in the iterative mode of synthesis), the sizes of NRPs are limited (Fig. 3). The most frequent sizes are 7 to 9 monomers, sizes shared by about one-third of the set. The sizes vary between 2 and 23 monomers, except for polytheonamide B (not shown in Fig. 3), which has 49 monomers. Polytheonamide B has been extracted

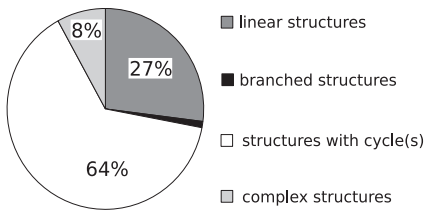

FIG. 1. Distribution of primary structures in curated peptides excluding variants (175 peptides).

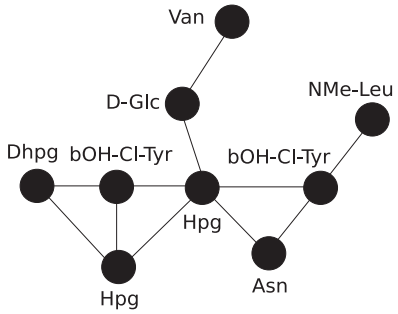

FIG. 2. Monomeric structure of vancomycin. Hpg, hydroxyphenylglycine; NMe-Leu, *N*-methylleucine; bOH-Cl-Tyr, beta-hydroxychloro-tyrosine; Asn, asparagine; Dhpg, 3,5-dihydroxyphenylglycine; D-Glc, D-glucose; Van, vancosamine.

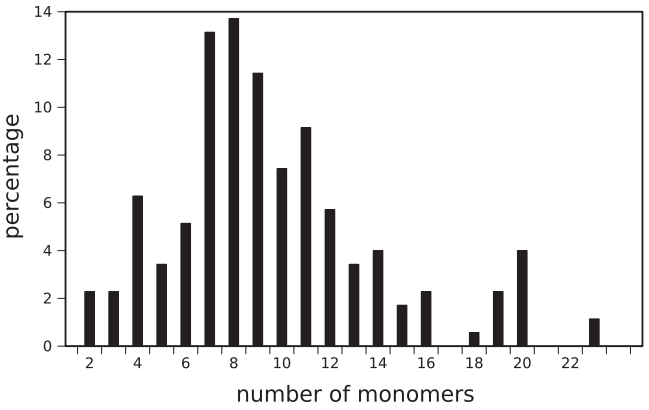

FIG. 3. Size distribution of curated peptides excluding variants (175 peptides). Polytheonamide B, composed of 49 monomers, does not appear in the figure, as it is the only peptide known with more than 23 monomers.

from the marine sponge *Theonella swinhoei* (25), but its synthesis process is not known. However, it contains several nonproteogenic and D-amino acids that suggest that its synthesis is nonribosomal (32).

(iii) What are the most frequently found monomers in nonribosomal peptides? In addition to having specific primary structures, NRPs contain nonproteogenic amino acids and other monomers. Among 1,725 monomers contained in 175 curated peptides excluding variants, 1,614 are incorporated and/or modified directly by nonribosomal peptide synthetases, 44 are lipids, 11 are carbohydrates, 5 are polyketides, and 51 are of other or unknown origins. Proteogenic L-amino acids represent 40% of monomers found in NRPs. The most frequent monomer in curated peptides excluding variants is 2-aminoisobutyric acid (Aib) (Fig. 4). The monomer Aib is characteristic of peptaibols (2), which are linear antibiotics produced only by fungi. The Norine database contains 130 peptaibols, forming 20 groups of variants (see Materials and Methods), all of which contain at least one Aib residue. Aib can occur several times in the same peptaibol, on average 6 times per peptide, which explains why Aib is the most frequent monomer in Norine NRPs. Note that serine (Ser), threonine (Thr), and their derivatives are very frequent in NRPs. These amino acids present a hydroxyl function that allows the formation of an additional chemical bond in order to obtain nonlinear primary structures. For example, in syringomycins (56), the hydroxyl group of serine is responsible for the branching, and the hydroxyl group of two threonines is used to form two cycles in actinomycin D (45). Many amino acids appear in their D-form in NRPs. D-Monomers are epimerized mainly by a specific domain of the synthetase or, in some cases, can be directly incorporated into their D-form, as was shown for arthrofactin synthesis (51). The monomer isomery can play an important role in a peptide's structure and properties, also by limiting protease degradation. Nonproteogenic amino acids such as 2,4-diaminobutyric acid (Dab) or ornithine (Orn) and derivates are often found in NRPs, for example, in marinobactins (39) or pyoverdins (67), respectively.

Finally, it is interesting to note that the two proteogenic amino acids containing the thiol function are underrepresented in NRPs. Methionine (Met) is present in only one curated peptide, oscillamide B (55), synthesized by bacteria, and in four putative peptides extracted from sponges (haligramides A and B [48], hymenamide F [33], and phakellistatin 5 [43]). Cysteine (Cys) occurs in the famous ACV (the penicillin precursor) synthesized by both bacteria and fungi but also occurs in 9 bacitracins (40) and in different siderophores (curated or not), such as pyochelin and related compounds (wataseymycin, thiazostatin, aeruginoic acid, desferriferrithiocin, micacocidin, yersiniabactin, and anguibactin) (7), synthesized by bacteria where Cys forms a cycle with another monomer. The high reactivity of the sulfhydryl group may explain the low representativeness of free cysteine in NRPs.

The Norine database provides an interface to query the monomers composing nonribosomal peptides (see http://bioinfo.lifl.fr/norine/search_amino.jsp). The entire list of monomers can be browsed, and information on each monomer can be consulted.

Study of producing organisms. Most of the peptides stored in the Norine database are isolated from inoculated media or natural environments, but only a few are inferred from a synthetase protein sequence (154 database peptides are linked to synthetases). Until now, synthetase genes have been identified only in bacteria and fungi; none have been identified in archaea or nonfungal eukaryotes. The major part of curated Norine peptides (61%) are synthesized by bacteria, and 34% of them are synthesized by fungi (Fig. 5). However, some peptides are extracted from other organisms (5% of data), such as sponges (like cyclotheonamides [20] and polytheonamides [25, 32] from *Theonella*), tunicates (like didemnins [66] isolated from the *Didemnidae*), gastropoda (such as antitumor peptide dolastatines [7] extracted from species of *Dolabella*), or even plants (e.g., a putative cyclolinopeptide [46] from linseed oil). In these cases, the NRPS genes have not been identified, but based on the presence of unusual monomers and/or structural features in these peptides, a nonribosomal origin can be hypothesized. For example, didemnins form a group of cyclic depsipeptides isolated from two groups of tunicates, *Didemnidae* and *Polyclinidae*, and show antibiotic, antitumor, and immunomodulating activities. Didemnins contain several unusual monomers, such as isostatine (Ist) and hydroxyisovalerylpropionyl (Hip), suggesting their nonribosomal origin (24). However, in the case of organisms other than bacteria and fungi,

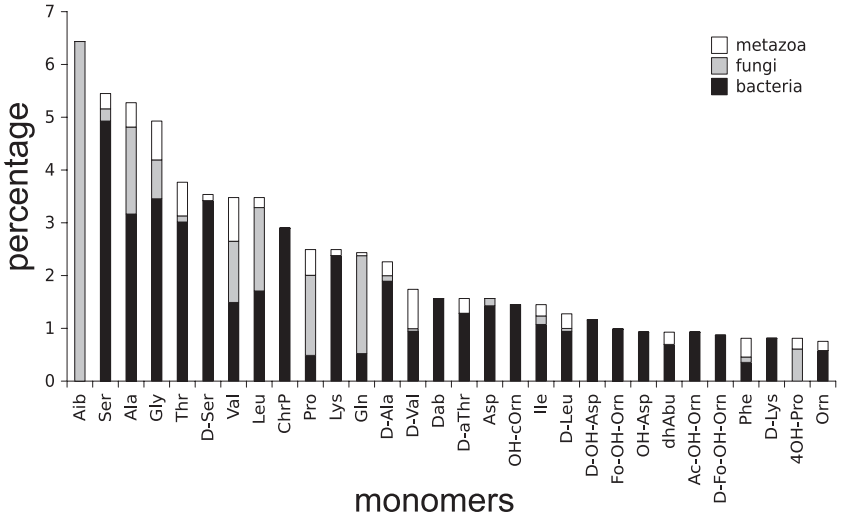

FIG. 4. Thirty of the most frequently found monomers among curated peptides excluding variants (175 peptides). The monomers' proportions occurring in bacteria, fungi, and metazoa are represented. dhAbu, 2,3-dehydro-2-aminobutyric acid.

the hypothesis that NRPs are actually synthesized by symbiotic bacteria cannot be excluded. Indeed, several recent studies have shown that NRPs extracted from sponges are in fact synthesized by symbiotic bacteria (for recent studies, see references 30, 36, 53, and 70).

We compared properties of NRPs isolated from bacteria, fungi, and metazoa, the last being mainly sponges. In this study, we considered curated NRPs for bacteria and fungi and all NRPs for metazoa. We did not use sets excluding variants or curated metazoan NRPs in order to keep a significant number of peptides in each set. To begin, we studied the size distribution of NRPs in the three groups of producing organisms (Fig. 6). We observed that the NRP sizes are different depending on the group (bacteria, fungi, or metazoa). For example, both bacteria and metazoa display a peak for sizes 7 and 8, while those sizes are nearly absent in fungal peptides. Numerous fungal peptides have a size between 14 and 20, while few bacterial and no metazoan peptides have those sizes. A peak for size 4 is shared by fungi and metazoa. However, the high proportion of size 4 metazoan peptides comes from geodiamolides (10), a family of 19 variants extracted only from sponges. We have computed the correlation coefficients (CCs) (Materials and Methods) between size distributions of NRPs synthesized by bacteria or fungi or extracted from metazoa (Table 2). The closer the CC is to 0, the less the monomeric distributions are related. This experiment confirms the previous observations that fungal NRPs are unrelated to both bacterial and metazoan NRPs (CC close to 0), while metazoan NRPs can be related to bacterial NRPs (CC close to 1).

Furthermore, we have computed the CC between peptide monomeric distributions depending on the producing organism. The results are shown in Table 3.

The CC between the monomeric distribution of NRPs synthesized by bacteria and those synthesized by fungi is the lowest observed, pointing out differences between the monomers used by both (super)kingdoms. On the other hand, the CC between the distributions of bacteria and metazoa is the highest, highlighting their similarity. Analyzing the monomers contained in peptides of different (super)kingdoms confirms the correlation coefficient tendency. For example, Aib (2-aminoisobutyric acid), the characteristic monomer of peptaibols, occurs in 131 fungal peptides, from which only one putative cyclic antitumor peptide is not a peptaibol: chlamydocin (4). Only one bacterial peptide of the Norine database, microcystin L-Aib (21), contains an L-Aib and no metazoan peptide (see Fig. 4). Other monomers are specific to fungal NRPs and, more

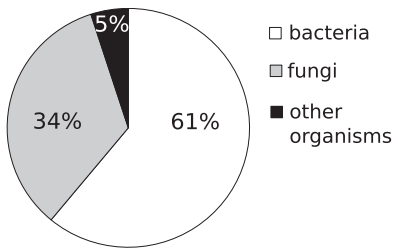

FIG. 5. Distribution of producing organisms for curated peptides (790 peptides).

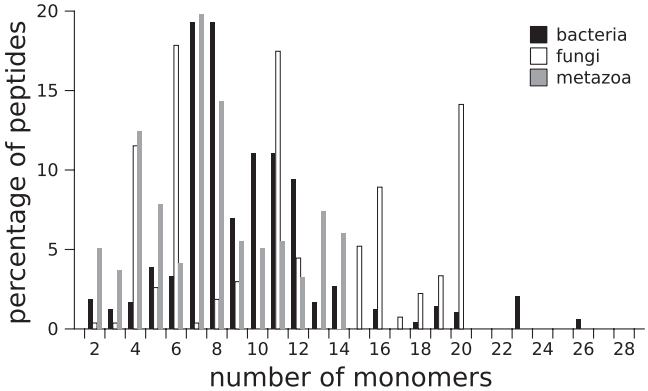

FIG. 6. Size distribution of curated peptides synthesized by bacteria (n = 488), curated peptides synthesized by fungi (n = 269), and peptides extracted from metazoa (n = 218).

TABLE 2. Correlation coefficients between size distributions of NRPs synthesized by bacteria or fungi or extracted from metazoa

Organism 1	Organism 2	CC
Bacteria (488 peptides)	Fungi (269 peptides)	0.236
Bacteria (488 peptides)	Metazoa (218 peptides)	0.817
Fungi (269 peptides)	Metazoa (218 peptides)	0.254

precisely, to peptaibols, such as isovaline (Ival), occurring in 23 peptaibols, phenylalaninol (Pheol), occurring in 68 peptaibols, Leucinol (Leuol), occurring in 43 peptaibols, and valinol (Valol), occurring in 29 peptaibols. The C-terminal amino alcohol in the peptaibols plays an important role in liposome permeabilization and ion channel formation (18).

Other monomers seem to be found only in bacterial NRPs (Fig. 4), which have specific properties. The monomer hydroxyphenylglycine (Hpg) is present in 56 NRPs, all of them produced by bacteria. This monomer is the only amino acid able to form 5 bonds with other monomers by oxidative ring closure. Hpg is usually found in peptides with complex primary structures, forming overlapping cycles and branching, such as vancomycin (Fig. 2) (28), balhimycin (49), decaplanin (54), eremomycin (22), and galacardin (63). Other monomers specific to bacteria are chromophores (Chr), which occur mainly in siderophores, except for actinomycins.

Study of biological activity. (i) Statistical results. In this section, we present a statistical analysis of some peptide characteristics depending on the peptide's biological activity. Here we consider all the variants because, in spite of their close compositions and structures, two variants can exhibit different activities. For example, actinomycin D is known to have antibiotic and antitumoral activities, but for the majority of other actinomycin variants, an antitumoral activity has not been reported. Figure 7 shows the distribution of six main activities found in curated peptides from the Norine database. Note that some peptides can show more than one activity. For example, there are 9 curated peptides in the Norine database that present both antibiotic and immunomodulator activities (such as edeines [12]) and 14 didemnins (66) that present three activities, antibiotic, antitumor, and immunomodulator.

We analyzed the monomeric distribution of NRPs depending on their activities. The distribution of the 30 most frequent monomers found in the Norine database-curated siderophores are presented in Fig. 8.

At neutral and alkaline pHs, ferric ions form insoluble polymeric hydroxide complexes that cannot be assimilated by microorganisms. To tackle this low iron bioavailability, many bacteria and fungi biosynthesize and excrete high-affinity iron

TABLE 3. Correlation coefficients between monomeric distributions of NRPs synthesized by bacteria or fungi or extracted from metazoa

Organism 1	Organism 2	CC
Bacteria (488 peptides)	Fungi (269 peptides)	0.252
Bacteria (488 peptides)	Metazoa (218 peptides)	0.534
Fungi (269 peptides)	Metazoa (218 peptides)	0.359

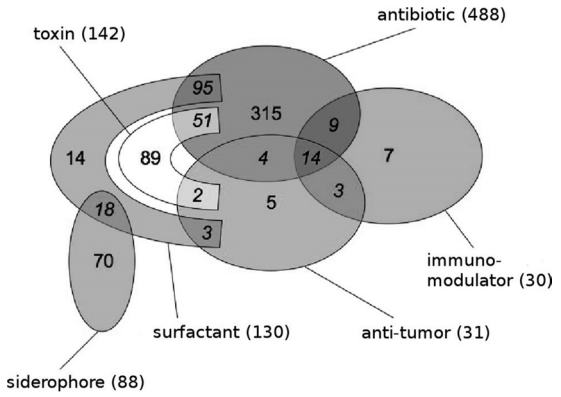

FIG. 7. Repartition of six main biological activities displayed by curated peptides in the Norine database (790 NRPs).

chelators known as siderophores. A common structural trait of most of these molecules is the presence of three bidentate groups which can ensure the 6-fold coordination of the ferric iron. Catecholate, hydroxamate, and hydroxycarboxylate groups are the three found most frequently that are able to play this role in the coordination of iron. Different monomers of NRPS products harbor such groups. For example, 2,3-dihydroxybenzoate (diOH-Bz, often denoted Dhb, which may result in confusion with another monomer, dehydrobutyrin, present in ribosomal bacteriocins like nisin or subtilin [19]), found in enterobactin or bacillibactin (1), or the chromophore of pyoverdins (ChrP, a dihydroxyquinoline group which results from the condensation and subsequent modification of diaminobutyric acid, tyrosine, and a dicarboxylic acid or its monoamide [7]) contains a catecholate group. *N*-Formyl-*N*-OH-ornithine (Fo-OH-Orn), found in ornibactins (61), *N*-acetyl-*N*-OH-ornithine (Ac-OH-Orn), found in aquachelins (69), *N*-OH-cyclo-ornithine (OH-cOrn), found in pseudobactins (65), OH-histidine (OH-His), found in corrugatin (50), and OH-lysine (OH-Lys), found in mycobactins (59) all contain an hydroxamate group. OH-aspartate (OH-Asp) of the azotobactin D (13) contains a hydroxycarboxylate group. Note that these monomers appear in the 30 most frequent siderophore monomers (Fig. 8). We also analyzed the monomeric distribution of NRPs in the other five classes of main activities (data not shown). The results suggest that the monomeric composition of a peptide can be used as a determiner of its biological activity. From this observation, we developed a method helping to predict the biological activity of a peptide from its monomeric composition.

(ii) Toward the prediction of biological activities. For each monomer of the database, we precomputed its frequency in each of the six activity classes. Then, for a given peptide, we computed the average frequency of its monomers to appear within each activity class and then deduced *P* values reflecting the probability of the peptide belonging to each class. The lower the *P* value for a given activity class, the more likely the peptide is to present this given activity.

We tested this method on several NRPs which have not yet been annotated in the Norine database. Table 4 gives some examples of resulting predictions.

Orfamide is a surfactant showing antibiotic activity (23). A

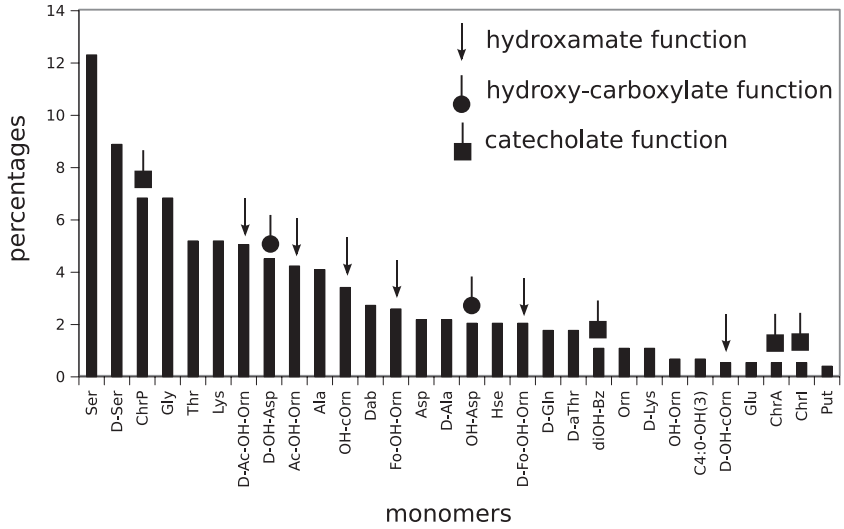

FIG. 8. Repartition of 30 of the most frequently found monomers in Norine database-curated siderophores (88 peptides). Hse, homoserine; ChrA, azotobactin chromophore; ChrI, isopyoverdin chromophore; Put, putrescine.

significant *P* value of 3.4E−4 is obtained for this peptide for the surfactant class, which suggests its surfactant properties. Fuscachelin is a nonribosomal siderophore (14), for which our method provided a strong indication. Aibellin is an antibiotic NRP (34). The *P* value obtained for the antibiotic class is close to 0, accurately predicting the biological activity of aibellin. These results suggest that the monomeric compositions of peptides can be of great help in predicting their biological activities.

DISCUSSION

General statistics. NRPs have numerous particularities. First, more than 500 different monomers are found in those peptides. Fatty acids, carbohydrates, and nonproteogenic amino acids are often incorporated in NRPs. We showed that in NRPs, a large part of monomers appear in their D-form. The D-residues play an important role in the structural conformation of the peptide, crucial for its biological activity and resistance against degradation process. NRPs can have a linear primary structure (less than 30% of peptides in the Norine database) but often present a more complex primary structure, including branchings and cycles. These nonlinear structures are essential for NRPs to have important biological activities, such as antibiotic and antitumor activities, preserving molecules from being degraded by proteases, which is essential for their biological functions.

Study of producing organisms. NRPS genes are experimentally identified only for bacteria and fungi. In this paper, we showed that NRPs produced by bacteria and those produced by fungi present different characteristics. The monomers used by bacteria are very different from those used by fungi. In addition, the NRP sizes are also very different for bacterial and fungal NRPs. These results showed that monomeric and other features of NRPs provide strong indications of their bacterial or fungal origin.

In addition to NRPs from fungi and bacteria, various peptides extracted from other organisms, such as metazoa, are believed to have a nonribosomal origin with regard to the presence of nonproteogenic amino acids (as in apramides [37], containing *N*-methyl-glycine-thiazole, or the cyclic barangamides [52], containing D-alloisoleucine) or their nonlinear structure. The current hypothesis is that the peptides extracted from metazoa are in fact produced by symbiotic bacteria rather than the metazoa themselves. Many recent papers tend to confirm this hypothesis by experimental validation of special cases of symbiosis (30, 36, 53, 70). In this paper, we have shown that monomers and structural features of NRPs produced by bacteria and those isolated from metazoa are very similar, which provides a supplementary argument supporting this hypothesis or the idea of a potential transfer of bacterial genes to metazoan genomes.

Study of biological activities. NRPs harbor a large spectrum of important biological activities. In this paper, we demonstrated that each activity class is associated with some specific monomers. By using frequencies of monomers in each activity class, we showed that the monomeric composition of a peptide

TABLE 4. Examples of biological activity predictions for some NRPs not annotated in the Norine database

Tested NRP	Known activity(ies)	*P* value obtained for the indicated class of activity					
		Antibiotic	Antitumor	Immunomodulator	Siderophore	Surfactant	Toxin
Orfamide	Surfactant, antibiotic	0.46	0.52	0.86	0.68	3.40E−04	0.29
Fuscachelin	Siderophore	0.65	0.82	0.96	4.90E−05	0.99	0.99
Aibellin	Antibiotic	<1.0E−30	0.98	0.97	0.97	0.97	0.93

can be an indicator of its biological activities. This can be of great interest for the study of new NRPs and can provide guidance for experimental studies.

Conclusion. For the last decade, the interest in NRPs and their biosynthetic enzymes has been considerably increased, as witnessed by an exponentially growing number of publications in this field. These peptides, indeed, are or can be used in many existing or potential biotechnological and pharmaceutical applications. A major part of published reviews focus on synthesis enzymes or on a genomic analysis, and much less work has been devoted to the global diversity of NRPs themselves. This situation led us to develop the Norine database, which is the only public resource devoted to NRPs and currently contains more than a thousand peptides. This tool allowed us to review, for the first time, the remarkable monomer diversity of these compounds. To our knowledge, no extensive study of monomers incorporated into NRPs has previously been done. In this paper, we presented the first large-scale analysis of monomers incorporated in NRPs. In addition to providing an overview of this biodiversity, this work demonstrated (i) a dissimilarity of structural properties between bacterial and fungal NRPs, (ii) a significant relationship between NRPs synthesized by bacteria and those isolated from metazoa, especially from sponges, supporting the hypothesis that the peptides isolated from sponges are in reality synthesized by symbiotic bacteria rather than by the sponges themselves, and (iii) a certain monomer specificity to classes of biological activities. An analysis of the monomer compositions of peptide products predicted from genomic information (metagenomics and high-throughput genome sequencing) can provide an indication of the origin of a peptide and/or its biological activity. Furthermore, new peptides detected by mass spectrometry analysis applied to a culture supernatant or carried out directly on colonies could be studied in the same way, leading to predictions concerning their origin or activities. Finally, those observations can be of great interest for developing combinatorial biosynthesis of NRPS and for the design of more-active antibiotic, immunomodulator, or anticancer drugs.

ACKNOWLEDGMENTS

This work was supported by the PPF bioinformatique program of Lille 1 University. S.C. was supported by an INRIA/Région Nord-Pas-de-Calais fellowship. The ProBioGEM lab is supported by the Région Nord-Pas-de-Calais, the Ministère de l'Enseignement Supérieur et de la Recherche, the French ANR Agency, and European Funds for Regional Development.

REFERENCES

1. **Abergel, R., A. Zawadzka, T. Hoette, and K. Raymond.** 2009. Enzymatic hydrolysis of trilactone siderophores: where chiral recognition occurs in enterobactin and bacillibactin iron transport. J. Am. Chem. Soc. **131:**12682–12692.
2. **Aravinda, S., N. Shamala, and P. Balaram.** 2008. Aib residues in peptaibiotics and synthetic sequences: analysis of non helical conformations. Chem. Biodivers. **5:**1238–1262.
3. **Arima, K., A. Kakinuma, and G. Tamura.** 1968. Surfactin, a crystalline peptidelipid surfactant produced by *Bacillus subtilis*: isolation, characterization and its inhibition of fibrin clot formation. Biochem. Biophys. Res. Commun. **31:**488–494.
4. **Bernardi, E., J. Fauchere, G. Atassi, P. Viallefont, and R. Lazaro.** 1993. Antitumoral cyclic peptide analogues of chlamydocin. Peptides **14:**1091–1093.
5. **Borel, J.** 2002. History of the discovery of cyclosporine and of its early pharmacological development. Wien. Klin. Wochenschr. **114:**433–437.
6. **Brennan, M., C. Costello, S. Maleknia, G. Pettit, and K. Erickson.** 2008. Stylopeptide 2, a proline-rich cyclodecapeptide from the sponge *Stylotella* sp. J. Nat. Prod. **71:**453–456.
7. **Budzikiewicz, H.** 2004. Siderophores of the *Pseudomonadaceae* sensu strict (fluorescent and non-fluorescent *Pseudomonas* spp.). Fortschr. Chem. Org. Naturst. **87:**81–237.
8. **Caboche, S., M. Pupin, V. Leclère, A. Fontaine, P. Jacques, and G. Kucherov.** 2008. Norine: a database of nonribosomal peptides. Nucleic Acids Res. **36:**D326–D331.
9. **Caboche, S., M. Pupin, V. Leclère, P. Jacques, and G. Kucherov.** 2009. Structural pattern matching of nonribosomal peptides. BMC Struct. Biol. **18:**9–15.
10. **Coleman, J., R. V. Soest, and R. Andersen.** 1999. New geodiamolides from the sponge *Cymbastela* sp. collected in Papua New Guinea. J. Nat. Prod. **62:**1137–1141.
11. **Consden, R., A. Gordon, and A. Martin.** 1947. Gramicidin S; the sequence of the amino-acid residues. Biochem. J. **41:**596–602.
12. **Czajgucki, Z., R. Andruszkiewicz, and W. Kamysz.** 2006. Structure activity relationship studies on the antimicrobial activity of novel edeine A and D analogues. J. Pept. Sci. **12:**653–662.
13. **Demange, P., A. Bateman, A. Dell, and M. Abdallah.** 1988. Structure of azotobactin D, a siderophore of *Azotobacter vinelandii* strain D (ccm289). Biochemistry **27:**2745.
14. **Dimise, E., P. Widboom, and S. Bruner.** 2008. Structure elucidation and biosynthesis of fuscachelins, peptide siderophores from the moderate thermophile *Thermobifida fusca*. Proc. Natl. Acad. Sci. U. S. A. **105:**15311–15316.
15. **Donadio, S., and M. Sosio.** 2008. Biosynthesis of glycopeptides: prospects for improved antibacterials. Curr. Top. Med. Chem. **8:**654–666.
16. **Duclohier, H.** 2007. Peptaibiotics and peptaibols: an alternative to classical antibiotics? Chem. Biodivers. **4:**1023–1026.
17. **Duitman, E., L. Hamoen, M. Rembold, G. Venema, H. Seitz, W. Saenger, F. Bernhard, R. Reinhardt, M. Schmidt, C. Ullrich, T. Stein, F. Leenders, and J. Vater.** 1999. The mycosubtilin synthetase of *Bacillus subtilis* atcc6633: a multifunctional hybrid between a peptide synthetase, an aminotransferase, and a fatty acid synthase. Proc. Natl. Acad. Sci. U. S. A. **96:**13294–13299.
18. **Duval, D., P. Cosette, S. Rebuffat, H. Duclohier, B. Bodo, and G. Molle.** 1998. Alamethicin-like behaviour of new 18-residue peptaibols, trichorzins PA. Role of the C-terminal amino-alcohol in the ion channel forming activity. Biochim. Biophys. Acta **1369:**309–319.
19. **Entian, K., and W. de Vos.** 1996. Genetics of subtilin and nisin biosyntheses: biosynthesis of lantibiotics. Antonie Van Leeuwenhoek **69:**109–117.
20. **Fusetani, N., and S. Matsunaga.** 1990. Cyclotheonamides, potent thrombin inhibitors, from a marine sponge *Theonella* sp. J. Am. Chem. Soc. **112:**7053–7054.
21. **Gathercole, P., and P. Thiel.** 1987. Liquid chromatographic determination of the cyanoginosins, toxins produced by the cyanobacterium *Microcystis aeruginosa*. J. Chromatogr. **408:**435–440.
22. **Gause, G., M. Brazhnikova, N. Lomakina, T. Berdnikova, G. Fedorova, N. Tokareva, V. Borisova, and G. Batta.** 1989. Eremomycin: new glycopeptide antibiotic: chemical properties and structure. J. Antibiot. (Tokyo) **42:**1790–1797.
23. **Gross, H., V. Stockwell, M. Henkels, B. Nowak-Thompson, J. Loper, and W. Gerwick.** 2007. The genomisotopic approach: a systematic method to isolate products of orphan biosynthetic gene clusters. Chem. Biol. **14:**53–63.
24. **Grubb, D., E. Wolvetang, and A. Lawen.** 1995. Didemnin B induces cell death by apoptosis: the fastest induction of apoptosis ever described. Biochem. Biophys. Res. Commun. **215:**1130–1136.
25. **Hamada, T., S. Matsunaga, G. Yano, and N. Fusetani.** 2005. Polytheonamides A and B, highly cytotoxic, linear polypeptides with unprecedented structural features, from the marine sponge, *Theonella swinhoei*. J. Am. Chem. Soc. **127:**110–118.
26. **Hamann, M., C. Otto, P. Scheuer, and D. Dunbar.** 1996. Kahalalides: bioactive peptides from a marine mollusk *Elysia rufescens* and its algal diet *Bryopsis* sp. J. Org. Chem. **61:**6594–6600.
27. **Hanessian, S., M. Tremblay, and J. Petersen.** 2004. The N-acyloxyiminium ion aza-prins route to octahydroindoles: total synthesis and structural confirmation of the antithrombotic marine natural product oscillarin. J. Am. Chem. Soc. **126:**6064–6071.
28. **Hubbard, B., and C. Walsh.** 2003. Vancomycin assembly: nature's way. Angew. Chem. Int. Ed. Engl. **42:**730–765.
29. **Kallow, W., T. Neuhof, B. Arezi, P. Jungblut, and H. von Döhren.** 1997. Penicillin biosynthesis: intermediates of biosynthesis of delta-alpha-aminoadipyl-cysteinyl-valine formed by ACV synthetase from *Acremonium chrysogenum*. FEBS Lett. **414:**74–78.
30. **Kennedy, J., P. Baker, C. Piper, P. Cotter, M. Walsh, M. Mooij, M. Bourke, M. Rea, P. O'Connor, R. Ross, C. Hill, F. O'Gara, J. Marchesi, and A. Dobson.** 2009. Isolation and analysis of bacteria with antimicrobial activities from the marine sponge *Haliclona simulans* collected from Irish waters. Mar. Biotechnol. **11:**384–396.
31. **Kessler, N., H. Schuhmann, S. Morneweg, U. Linne, and M. Marahiel.** 2004. The linear pentadecapeptide gramicidin is assembled by four multimodular nonribosomal peptide synthetases that comprise 16 modules with 56 catalytic domains. J. Biol. Chem. **279:**7413–7419.

32. **Kleinkauf, H., and H. V. Döhren.** 1996. Polytheonamide, the longest peptide reported of presumably enzymatic origin. J. Biochem. **236:**335–351.
33. **Kobayashi, J., T. Nakamura, and M. Tsuda.** 1996. Hymenamide F, new cyclic heptapeptide from marine sponge *Hymeniacidon* sp. Tetrahedron **52:** 6355–6360.
34. **Kumazawa, S., M. Kanda, H. Aoyama, M. Utagawa, J. Kondo, S. Sakamoto, H. Ohtani, T. Mikawa, I. Chiga, and T. Hayase.** 1994. Structural elucidation of aibellin, a new peptide antibiotic with efficiency enhancing activity on rumen fermentation. J. Antibiot. (Tokyo) **47:**1136–1144.
35. **Li, H., T. Tanikawa, Y. Sato, Y. Nakagawa, and T. Matsuyama.** 2005. Serratia marcescens gene required for surfactant serrawettin W1 production encodes putative aminolipid synthetase belonging to nonribosomal peptide synthetase family. Microbiol. Immunol. **49:**303–310.
36. **Luesch, H., G. Harrigan, G. Goetz, and F. Horgen.** 2002. The cyanobacterial origin of potent anticancer agents originally isolated from sea hares. Curr. Med. Chem. **9:**1791–1806.
37. **Luesch, H., W. Yoshida, R. Moore, and V. Paul.** 2000. Apramides A–G, novel lipopeptides from the marine cyanobacterium *Lyngbya majuscula*. J. Nat. Prod. **63:**1106–1112.
38. **Marahiel, M., and L. Essen.** 2009. Nonribosomal peptide synthetases mechanistic and structural aspects of essential domains. Methods Enzymol. **458:** 337–351.
39. **Martinez, J., and A. Butler.** 2007. Marine amphiphilic siderophores: marinobactin structure, uptake, and microbial partitioning. J. Inorg. Biochem. **101:**1692–1698.
40. **Ming, L., and J. Epperson.** 2002. Metal binding and structure-activity relationship of the metalloantibiotic peptide bacitracin. J. Inorg. Biochem. **91:** 46–58.
41. **Mootz, H., D. Schwarzer, and M. Marahiel.** 2002. Ways of assembling complex natural products on modular nonribosomal peptide synthetases. Chembiochem **3:**490–504.
42. **Pettit, G., Z. Cichacz, J. Barkoczy, A. Dorsaz, D. Herald, M. Williams, D. Doubek, J. Schmidt, L. Tackett, and D. Brune.** 1993. Isolation and structure of the marine sponge cell growth inhibitory cyclicpeptide phakellistatin 1. J. Nat. Prod. **56:**260–267.
43. **Pettit, G., B. Toki, J. Xu, and D. Brune.** 2000. Synthesis of the marine sponge cycloheptapeptide phakellistatin 5. J. Nat. Prod. **63:**22–28.
44. **Peypoux, F., J. Bonmatin, H. Labbé, B. Das, M. Ptak, and G. Michel.** 1991. Isolation and characterization of a new variant of surfactin, the [Val7]surfactin. Eur. J. Biochem. **202:**101–106.
45. **Pfennig, F., F. Schauwecker, and U. Keller.** 1999. Molecular characterization of the genes of actinomycin synthetase I and of a 4-methyl-3-hydroxyanthranilic acid carrier protein involved in the assembly of the acylpeptide chain of actinomycin in *Streptomyces*. J. Biol. Chem. **274:**12508–12516.
46. **Picur, B., M. Cebrat, J. Zabrocki, and I. Siemion.** 2006. Cyclopeptides of *Linum usitatissimum*. J. Pept. Sci. **12:**569–574.
47. **Poncet, J.** 1999. The dolastatins, a family of promising antineoplastic agents. Curr. Pharm. Des. **5:**139–162.
48. **Rashid, M., K. Gustafson, J. Boswell, and M. Boyd.** 2000. Haligramides A and B, two new cytotoxic hexapeptides from the marine sponge *Haliclona nigra*. J. Nat. Prod. **63:**956–959.
49. **Recktenwald, J., R. Shawky, O. Puk, F. Pfennig, U. Keller, W. Wohlleben, and S. Pelzer.** 2002. Nonribosomal biosynthesis of vancomycin-type antibiotics: a heptapeptide backbone and eight peptide synthetase modules. Microbiology **148:**1105–1118.
50. **Risse, D., H. Beiderbeck, K. Taraz, H. Budzikiewicz, and D. Gustine.** 1998. Corrugatin, a lipopeptide siderophore from *Pseudomonas corrugata*. Z. Naturforsch. C **53:**295–304.
51. **Roongsawang, N., K. Hase, M. Haruki, T. Imanaka, M. Morikawa, and S. Kanaya.** 2003. Cloning and characterization of the gene cluster encoding arthrofactin synthetase from *Pseudomonas* sp. MIS38. Chem. Biol. **10:**869–880.
52. **Roy, M., I. Ohtani, J. Tanaka, T. Higa, and R. Satari.** 1999. Barangamide A, a new cyclic peptide from the Indonesian sponge *Theonella swinhoei*. Tetrahedron Lett. **40:**5373–5376.
53. **Salomon, C., N. Magarvey, and D. Sherman.** 2004. Merging the potential of microbial genetics with biological and chemical diversity: an even brighter future for marine natural product drug discovery. Nat. Prod. Rep. **21:**105–121.
54. **Sanchez, M., R. Wenzel, and R. Jones.** 1992. In vitro activity of decaplanin (M86-1410), a new glycopeptide antibiotic. Antimicrob. Agents Chemother. **36:**873–875.
55. **Sano, T., T. Usui, K. Ueda, H. Osada, and K. Kaya.** 2001. Isolation of new protein phosphatase inhibitors from two cyanobacteria species, *Planktothrix* spp. J. Nat. Prod. **64:**1052–1055.
56. **Segre, A., R. Bachmann, A. Ballio, F. Bossa, I. Grgurina, N. Iacobellis, G. Marino, P. Pucci, M. Simmaco, and J. Takemoto.** 1989. The structure of syringomycins A1, E and G. FEBS Lett. **255:**27–31.
57. **Shen, B., L. Du, C. Sanchez, D. Edwards, M. Chen, and J. Murrell.** 2001. The biosynthetic gene cluster for the anticancer drug bleomycin from *Streptomyces verticillus* ATCC 15003 as a model for hybrid peptide-polyketide natural product biosynthesis. J. Ind. Microbiol. Biotechnol. **27:**378–385.
58. **Shiki, Y., M. Onai, D. Sugiyama, S. Osada, I. Fujita, and H. Kodama.** 2009. Synthesis and biological activities of cyclic peptide, hymenamide analogs. Adv. Exp. Med. Biol. **611:**323–324.
59. **Snow, G.** 1970. Mycobactins: iron-chelating growth factors from mycobacteria. Bacteriol. Rev. **34:**99–125.
60. **Steenbergen, J., J. Alder, G. Thorne, and F. Tally.** 2005. Daptomycin: a lipopeptide antibiotic for the treatment of serious gram-positive infections. J. Antimicrob. Chemother. **55:**283–288.
61. **Stephan, H., S. Freund, W. Beck, G. Jung, J. Meyer, and G. Winkelmann.** 1993. Ornibactins—a new family of siderophores from *Pseudomonas*. Biometals **6:**93–100.
62. **Suenaga, K., S. Kajiwara, S. Kuribayashi, T. Handa, and H. Kigoshi.** 2008. Synthesis and cytotoxicity of aurilide analogs. Bioorg. Med. Chem. Lett. **18:**3902–3905.
63. **Takeuchi, M., S. Takahashi, R. Enokita, Y. Sakaida, H. Haruyama, T. Nakamura, T. Katayama, and M. Inukai.** 1992. Galacardins A and B, new glycopeptide antibiotics. J. Antibiot. (Tokyo) **45:**297–305.
64. **Tan, L., N. Sitachitta, and W. Gerwick.** 2003. The guineamides, novel cyclic depsipeptides from a Papua New Guinea collection of the marine cyanobacterium *Lyngbya majuscula*. J. Nat. Prod. **66:**764–771.
65. **Teintze, M., and J. Leong.** 1981. Structure of pseudobactin A, a second siderophore from plant growth promoting *Pseudomonas* B10. Biochemistry **20:**6457–6462.
66. **Vera, M., and M. Joullié.** 2002. Natural products as probes of cell biology: 20 years of didemnin research. Med. Res. Rev. **22:**102–145.
67. **Visca, P., F. Imperi, and I. Lamont.** 2007. Pyoverdine siderophores: from biogenesis to biosignificance. Trends Microbiol. **15:**22–30.
68. **Yeung, B., Y. Nakao, R. Kinnel, J. Carney, W. Yoshida, P. Scheuer, and M. Kelly-Borges.** 1996. The kapakahines, cyclic peptides from the marine sponge *Cribrochalina olemda*. J. Org. Chem. **61:**7168–7173.
69. **Zhang, G., S. Amin, F. Küpper, P. Holt, C. Carrano, and A. Butler.** 2009. Ferric stability constants of representative marine siderophores: marinobactins, aquachelins, and petrobactin. Inorg. Chem. **48:**11466–11473.
70. **Zhang, W., Z. Li, X. Miao, and F. Zhang.** 2009. The screening of antimicrobial bacteria with diverse novel nonribosomal peptide synthetase (NRPS) genes from South China Sea sponges. Mar. Biotechnol. **11:**346–355.