nature chemical biology

ARTICLE

PUBLISHED ONLINE: 3 AUGUST 2015 | DOI: 10.1038/NCHEMBIO.1868

Analysis of 51 cyclodipeptide synthases reveals the basis for substrate specificity

Isabelle B Jacques[1,3], Mireille Moutiez[1,3,4], Jerzy Witwinowski[2,4], Emmanuelle Darbon[2], Cécile Martel[2], Jérôme Seguin[1,3], Emmanuel Favry[1,3], Robert Thai[1], Alain Lecoq[1], Steven Dubois[1], Jean-Luc Pernodet[2], Muriel Gondry[1,3]* & Pascal Belin[1,3]*

Cyclodipeptide synthases (CDPSs) constitute a family of peptide bond–forming enzymes that use aminoacyl-tRNAs for the synthesis of cyclodipeptides. Here, we describe the activity of 41 new CDPSs. We also show that CDPSs can be classified into two main phylogenetically distinct subfamilies characterized by specific functional subsequence signatures, named NYH and XYP. All 11 previously characterized CDPSs belong to the NYH subfamily, suggesting that further special features may be yet to be discovered in the other subfamily. CDPSs synthesize a large diversity of cyclodipeptides made up of 17 proteinogenic amino acids. The identification of several CDPSs having the same specificity led us to determine specificity sequence motifs that, in combination with the phylogenetic distribution of CDPSs, provide a first step toward being able to predict the cyclodipeptides synthesized by newly discovered CDPSs. The determination of the activity of ten more CDPSs with predicted functions constitutes a first experimental validation of this predictive approach.

CDPSs constitute a family of enzymes that divert aminoacyl-tRNAs (aa-tRNAs) from the ribosomal machinery to catalyze the formation of various cyclodipeptides, the precursors of many secondary metabolites with interesting biological activities[1]. Crystallographic structures of three CDPSs are now available, and indicate that they share a common architecture reminiscent of the catalytic domain of class Ic aminoacyl-tRNA synthetases (aaRSs)[2–4]. They use a ping-pong catalytic mechanism initiated by the binding of the first aa-tRNA. The aminoacyl moiety of this substrate is transferred to a conserved active site serine to generate an aminoacyl-enzyme intermediate[2–5]. This intermediate reacts with the aminoacyl moiety of the second aa-tRNA to form a dipeptidyl-enzyme. Finally, this second intermediate undergoes an intramolecular cyclization through the involvement of a conserved tyrosine, leading to the cyclodipeptide product[6]. Two recent studies on the CDPS AlbC give clues as to how CDPSs and their two substrates interact[6,7]. The two aa-tRNAs bind at different sites: the aminoacyl moiety of the first aa-tRNA is accommodated in pocket P1—structurally similar to the amino acid–binding pocket in class Ic aaRSs—and its tRNA moiety interacts with a patch of basic residues on the amphipathic helix α4; the aminoacyl moiety of the second aa-tRNA is accommodated in a wider pocket, P2, with its tRNA moiety interacting with residues that belong to the flexible loop α6-α7. The specificity of recognition of the first substrate is determined by its aminoacyl moiety, whereas that of the second substrate depends on both its aminoacyl moiety and its tRNA sequence, especially the N^1-N^{72} base pair of the acceptor stem. Only ten bacterial CDPSs and one eukaryotic CDPS have been biochemically characterized[1,5,8,9]. Most show a degree of substrate promiscuity but incorporate essentially five hydrophobic amino acids—F, L, Y, M and W—into cyclodipeptides. Iterative PSI-BLAST searches have identified numerous putative CDPSs[10–12], and many of them diverge from the known conserved features of the family. To provide a better description of the CDPS family, we first report the identification and characterization of 41 new active members and propose a classification for the CDPS family. Then, we describe how our data support a predictive tool for CDPS specificity. Finally, we demonstrate the reliability of prediction for few selected groups by characterizing ten additional CDPSs of unknown function.

RESULTS

Search for active CDPSs identified 41 new members

We updated the list of putative CDPSs by searching the National Center for Biotechnology Information (NCBI) protein database. We retrieved about 80 new sequences corresponding to putative CDPSs (May 2013) and used these to construct a phylogenetic tree (**Fig. 1**). We selected 49 sequences representative of the whole sequence set (designated CDPS 1–49; **Supplementary Results**, **Supplementary Data Set 1**, **Fig. 1**) and investigated whether the selected putative CDPSs can synthesize cyclodipeptides. We cloned their genes in a homemade expression vector and produced the proteins in *Escherichia coli* strains, as previous work showed that cyclodipeptides can be recovered in culture supernatants upon CDPS expression in this host[1]. We implemented a medium-throughput method for CDPS expression in *E. coli*, recovery of culture supernatants for cyclodipeptide detection by LC/MS/MS, and bacterial fractionation for protein content analysis by SDS-PAGE (**Supplementary Fig. 1**). We found that 41 of the 49 recombinant CDPS candidates tested had cyclodipeptide-synthesizing activity (**Fig. 1**, **Supplementary Data Set 2**, **Table 1**).

We found that eight of the putative CDPSs were inactive, although they were effectively produced (**Supplementary Fig. 2**). Sequences analysis revealed several possible explanations for this absence of activity. CDPSs 12 and 15 lack the catalytic serine residue (replaced by A and G, respectively). In CDPS 45 this residue is replaced by a cysteine; this is, however, not sufficient to rule it out from being an active CDPS as the variant AlbC S37C is able to synthesize cyclodipeptides[3]. CPDS 45 is also eukaryotic, and therefore *E. coli* may lack an appropriate set of tRNA substrates for it to act on. We found no obvious explanations for the lack of activity of the other candidates.

[1]Commissariat à l'énergie atomique et aux énergies alternatives (CEA), Institut de Biologie et de Technologies de Saclay (iBiTec-S), Service d'Ingénierie Moléculaire des Protéines, Gif-sur-Yvette, France. [2]Université Paris-Sud, Centre National de la Recherche Scientifique (CNRS), Unité Mixte de Recherche (UMR) 8621, Institut de Génétique et Microbiologie, Orsay, France. [3]Present address: CEA, iBiTec-S, Service de Biologie Intégrative et Génétique Moléculaire, UMR 9198, Gif-sur-Yvette, France. [4]These authors contributed equally to this work. *e-mail: pascal.belin@cea.fr or muriel.gondry@cea.fr

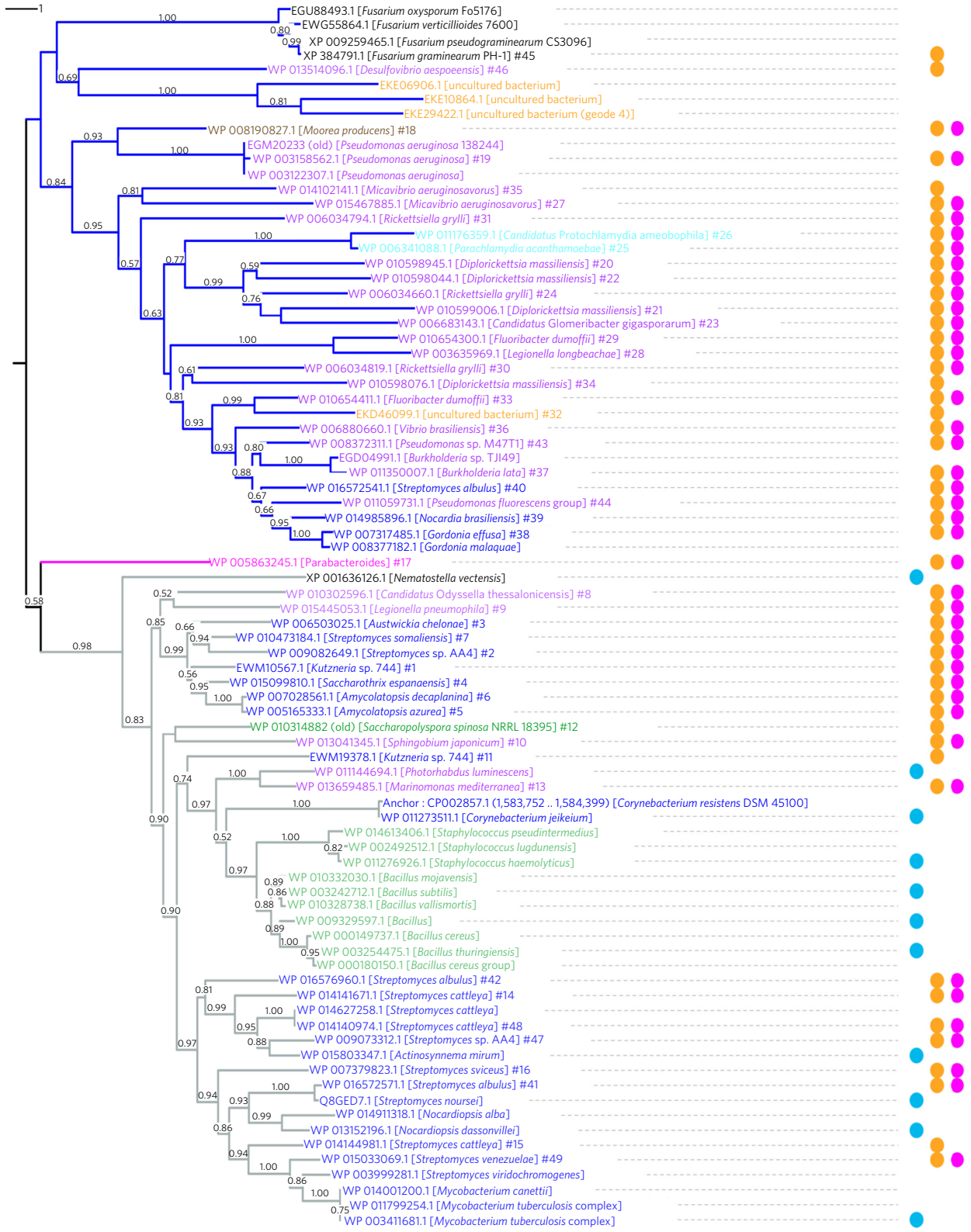

Figure 1 | Phylogenetic tree of known CDPSs and putative CDPS retrieved from bioinformatics searches in databases (May 2013). The set of sequences was curated: seven partial sequences were removed and the start of eight sequences manually corrected (see **Supplementary Figs. 6** and **7**). CDPSs are named by their protein accession numbers in NCBI and their host organisms; the numbers given to the CDPSs studied herein are added at the end of the name. CDPS names are colored according to their taxonomic origin: Proteobacteria, purple; Actinobacteria, dark blue; Chlamydiae, cyan; Cyanobacteria, brown; Bacteroidetes, fuchsia; Firmicutes, green; Eukaryotes, black (including 4 fungi and 1 metazoa (XP_001636126.1 [*Nematostella vectensis*])). CDPSs from metagenomic approaches are colored orange. The CDPS members can be classified into three subfamilies that clearly separate into three main branches on the tree; the branches classified as NYH, XYP and SYQ are in gray, blue and fuchsia, respectively. The 11 known CDPSs, the 49 putative CDPSs studied herein, and the 41 new active CDPSs are labeled with cyan, orange, and magenta dots, respectively.

Table 1 | The major cyclodipeptides produced by the newly identified CDPSs

CDPS	Species	CDPS subfamily	*In vivo* activity[a]
1	*Kutzneria* sp. 744	NYH	cCC (100%)
2	*Streptomyces* sp. AA4	NYH	cCC (100%)
3	*Austwickia chelonae*	NYH	cCC (63%), cCA (25%), cCV (12%)
4	*Saccharothrix espanaensis*	NYH	cCC (100%)
5	*Amycolatopsis azurea*	NYH	cCC (100%)
6	*Amycolatopsis decaplanina*	NYH	cCC (100%)
7	*Streptomyces somaliensis*	NYH	cCC (99%)
8	*Candidatus* Odyssella thessalonicensis	NYH	cGV (99%)
9	*Legionella pneumophila*	NYH	cCC (100%)
10	*Sphingobium japonicum*	NYH	cPM (95%)
11	*Kutzneria* sp. 744	NYH	–
12	*Saccharopolyspora spinosa*	NYH	–
13	*Marinomonas mediterranea*	NYH	cLL (98%)
14	*Streptomyces cattleya*	NYH	cWW (100%)
15	*Streptomyces cattleya*	NYH	–
16	*Streptomyces sviceus*	NYH	cLV (90%)
17	*Parabacteroides* sp. 20_3	*SYQ*	cHF (100%)
18	*Moorea producens*	XYP	cAP (100%)
19	*Pseudomonas aeruginosa*	XYP	cIL (93%)
20	*Diplorickettsia massiliensis*	XYP	cFN (90%)
21	*Diplorickettsia massiliensis*	XYP	cLL (80%)
22	*Diplorickettsia massiliensis*	XYP	cYV (30%), cYP (23%), cYT (11%), cYA (10%), cFP (10%), cFV (7%)
23	*Candidatus* Glomeribacter gigasporarum	XYP	cFF (93%) cFL (7%)
24	*Rickettsiella grylli*	XYP	cFF (53%) cFL (47%)
25	*Parachlamydia acanthamoebae*	XYP	cPP (75%)
26	*Candidatus* Protochlamydia amoebophila	XYP	cPP (100%)
27	*Micavibrio aeruginosavorus*	XYP	cGN (100%)
28	*Legionella longbeachae*	XYP	cAA (95%)
29	*Fluoribacter dumoffii*	XYP	cAA (97%)
30	*Rickettsiella grylli*	XYP	cMA (23%) cMG (27%)
31	*Rickettsiella grylli*	XYP	cAG (86%) cAA (14%)
32	Uncultured bacterium ACD_69C00020	XYP	–
33	*Fluoribacter dumoffii*	XYP	cGG (100%)
34	*Diplorickettsia massiliensis*	XYP	–
35	*Micavibrio aeruginosavorus*	XYP	–
36	*Vibrio brasiliensis*	XYP	cLE (58%), cLA (20%), cLP (14%)
37	*Burkholderia lata*	XYP	cEA (100%)
38	*Gordonia effusa*	XYP	cEA (97%)
39	*Nocardia brasiliensis*	XYP	cEA (96%)
40	*Streptomyces albulus*	XYP	cEA (97%)
41	*Streptomyces albulus*	NYH	cFL (48%), cFF (23%), cFY (19%), cFM (6%)
42	*Streptomyces albulus*	NYH	cWW (100%)
43	*Pseudomonas* sp. M47T1	XYP	cEA (100%)
44	*Pseudomonas protegens*	XYP	cLE (99%)
45	*Gibberella zeae*	XYP	–
46	*Desulfovibrio aespoeensis*	XYP	–
47	*Streptomyces* sp. AA4	NYH	cWW (100%)
48	*Streptomyces cattleya*	NYH	cWW (100%)
49	*Streptomyces venezuelae*	NYH	cYY (92%)

[a]The major cyclodipeptides produced by recombinant strain M15 pREP4 expressing the corresponding CDPS are given; cyclodipeptides are ranked according to the peak area on chromatograms recorded at 214 nm (see **Supplementary Data Set 2**); percentages in brackets indicate the proportion of each cyclodipeptide and were determined from peak areas on chromatograms recorded at 214 nm; cyclodipeptides that made up <5% of the total are not indicated; cCC refers to the detection of cyclocystine in bacterial culture supernatants: when a CDPS synthesizes cCC, it is recovered in its oxidized form, i.e., cyclocystine, in the supernatant; –, no cyclodipeptide detected.

Different catalytic residues for the two CDPS subfamilies

Thus, the CDPS family now includes 52 proteins (the 11 previously characterized CDPSs plus the 41 active CDPSs characterized in this study), which allows re-examination of the features of this family. Most known CDPSs are found in three bacterial phyla, Actinobacteria, Firmicutes and Proteobacteria, and a few are found in Bacteroidetes (1), Chlamydiae (2) and Cyanobacteria (1). Only one active CDPS has been found in a eukaryotic phylum (Cnidaria)[5], and no CDPSs have been found in Archaea (**Fig. 1**). CDPSs are typically 200–300 residues long. Most of the newly confirmed CDPSs share less than 30% amino acid sequence identity with any of the 11 previously characterized CDPSs, and 22 share only between 6% and 20% sequence identity with them (**Supplementary Data Set 3**). However, HHPRED searches with each of the 41 new CDPS sequences retrieved systematically as best hits the three CDPSs of known structure (**Supplementary Data Set 4**). We used the secondary-structure predictions and sequence similarities identified with HHPRED to adjust the multiple sequence alignment of the 52 CDPSs (**Supplementary Data Set 5**). The six residues previously identified as catalytic residues are mostly conserved (N40, Y178, E182, H203) or even strictly conserved (S37, Y202) (AlbC numbering, used throughout the paper)[1,3,6], allowing a functional sequence signature for the CDPS enzymes to be defined (**Fig. 2a**). We previously showed that the strictly conserved catalytic residue Y202 participates in a hydrogen bond network that also involves residues N40 and H203, which are essential for the accurate positioning of the two loops bearing the catalytic residues in AlbC[6]. We observed that about half of the CDPSs do indeed have the 'N40, H203' pair, whereas others have an 'X40, P203' pair (X being a nonconserved residue) and a very small number have other pairs (see Discussion section). Examination of the phylogenetic tree of CDPSs shows an early divergence into two different branches (**Fig. 1**). Comparative analysis between CDPS distribution in the tree and their sequences shows a strong correlation between the distribution and the presence of one or the other pair in the sequences. CDPSs clearly divide into two subfamilies or classes that we named 'NYH' and 'XYP' according to the 'X40, Y202, X203' sequence. CDPS 17 has a 'S40, Q203' pair and is also clearly separate from the other CDPSs in the phylogenetic tree, suggesting that it belongs to a probable third subfamily that we named 'SYQ'. Each subfamily can be associated with a

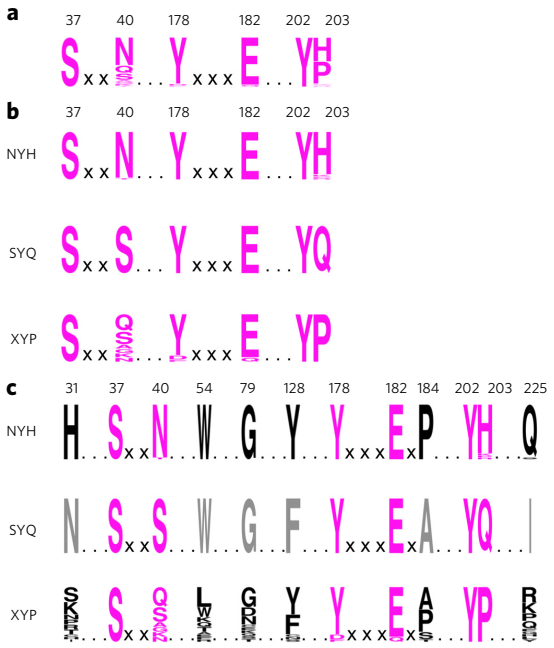

Figure 2 | CDPS subfamilies. (**a**) Functional sequence signature for CDPSs shown with sequence logos corresponding to the frequency plot of the catalytic amino acids at positions 37, 40, 178, 182, 202 and 203 (AlbC numbering). (**b**) Functional subsequence signatures for the CDPS subfamilies. Subsequence signatures are indicated for the NYH and XYP subfamilies and for CDPS 17, which is the lone member of a probable third subfamily, SYQ. (**c**) Subsequence signatures for the CDPS subfamilies with catalytic and conserved residues taken into account. Conserved residues, indicated in black for the NYH and XYP subfamilies and in gray for the probable SYQ subfamily (containing only one known member), are shown at positions 31, 54, 79, 128, 184 and 225 (AlbC numbering). The sequence of the eukaryotic NYH member, XP_001636126.1 [*Nematostella vectensis*], was removed from the set used to define the signatures because it contains modifications probably corresponding to phylum.

functional sub-sequence signature (**Fig. 2b**)—whose differences are likely to reflect different ways of positioning catalytic residues—and possesses further distinguishing features. Notably, the 11 previously characterized CDPSs all belong to the NYH subfamily.

The NYH subfamily has six conserved residues in addition to the six catalytic residues (**Supplementary Data Set 5**; **Supplementary Fig. 3**), including three residues identified in AlbC as being essential for enzyme activity (H31 and Y128) or structural integrity (W54)[3]. The XYP subfamily has a more degenerate sub-sequence, with only five catalytic residues (and no noncatalytic residues) conserved. The single SYQ sequence has half of the noncatalytic conserved residues of the NYH subfamily (**Fig. 2c**). The basic patch involved in the binding of the tRNA moiety of the first substrate[7] is found in all members of the NYH subfamily and contains 5–11 basic residues. This patch is much smaller in the XYP class, with only 2–6 basic residues. The XYP subfamily is generally far more diverse than the NYH subfamily, with variants of even the highly conserved catalytic sequence YxxxE found: CDPSs 25 and 26 have the sequence DxxxQ and not YxxxE (**Supplementary Data Set 5**). The role of the E residue, which acts as a catalytic base essential for generating the dipeptidyl-AlbC intermediate, and that of the Y residue in positioning the aminoacyl moiety of the first substrate in AlbC[3,6] could be fulfilled by D and Q, respectively. This provides support for the idea that the positioning of the catalytic residues is probably not the same in the different subfamilies. The XYP subfamily also contains members containing sequence insertions up to 30 residues long between predicted secondary structures, and deletions of a few residues, mostly in predicted α helices (CDPSs 21, 23, 25, 26, 28 and 29) (**Supplementary Fig. 4**). In addition, CDPSs 28 and 29 contain an additional ~230-residue C-terminal domain of unknown function, suggesting the possibility of a CDPS domain in a multifunctional protein (**Supplementary Fig. 4**).

An expanded diversity of cyclodipeptides

The 19 cyclodipeptides synthesized by the 11 NYH enzymes known before this study are mostly hydrophobic, with aromatic and aliphatic side chains[1,5,8,9]. Here, we identify 35 cyclodipeptides that have not previously been described as CDPS products. Many of these contain amino acids not previously observed in products of CDPS activity, including G, amino acids with hydrophobic side chains (V, P, I) and polar and charged amino acids (E, H, N, Q, S, T, C) (**Fig. 3a**). Consequently, CDPSs now incorporate 17 of 20 proteinogenic amino acids into cyclodipeptides.

We looked at how cyclodipeptide diversity is distributed among CDPS subfamilies. NYH members incorporate ten different amino acids into cyclodipeptides, with C being the only additional amino acid used by these new NYH subfamily CDPSs. XYP members use a larger set of amino acids—16 of the 20 proteinogenic amino acids—including those used by NYH members and also the polar and charged residues Q, T, N, S and E (**Fig. 3b**). Consistent with this diversity of amino acid utilization, the number of cyclodipeptides synthesized differs between the families, with 26 and 39 cyclodipeptides found for NYH and XYP members, respectively. This greater product diversity may be related to the larger diversity of sequences within the XYP subfamily (**Supplementary Fig. 3**).

The previously characterized NYH members show a degree of substrate promiscuity, with Amir_4628 being the only one synthesizing specifically one cyclodipeptide: cWW[8]. About a third of the newly characterized CDPSs show high substrate specificity (**Table 1**; **Supplementary Fig. 5**). These high specificities are not related to their membership in one or other subfamily; they are probably a consequence of the nature of the substrates and reflect interaction with either a particular tRNA sequence—the N^1-N^{72} base pair is important for recognition of the second substrate and, although well conserved in the prokaryotic world, it displays some differences (**Supplementary Data Set 6**)—or an aminoacyl moiety with specific physical properties or chemical properties[7]. The other newly characterized CDPSs demonstrated promiscuity toward different amino acids in our experimental conditions (**Table 1**; **Supplementary Data Set 2**; **Supplementary Fig. 5**), but most synthesized primarily one cyclodipeptide and other cyclodipeptides in smaller amounts.

Toward prediction of CDPS specificity

Substrate prediction requires biochemical characterization of a sufficient number of enzymes and the knowledge of the substrate-binding pocket determinants; this was illustrated by the pioneering works on the substrate specificity of the adenylation domains of nonribosomal peptide synthetases[13,14]. The large number of active CDPSs described herein and the identification of the cyclodipeptides they produce can be exploited in a similar way to help understand CDPS specificity.

The NYH subfamily member AlbC has two binding pockets, P1 and P2, accommodating the aminoacyl moieties of the first and second aa-tRNA; there are eight residues lining P1 and seven others lining P2 (ref. 6) (**Fig. 4a**). We used the multiple sequence alignment of CDPSs manually adjusted with HHPRED (**Supplementary Data Set 5**) to identify the residues of the active CDPSs corresponding to those lining P1 and P2 of AlbC (**Supplementary Data Set 7**). We predicted, for all active CDPSs, secondary structures similar to those observed in the crystal structures of AlbC, YvmC and Rv2275, and we assumed that the positions of residues lining P1 and P2 were conserved (**Supplementary Data Set 4**). CDPSs possessing similar activities, such as cLL-synthesizing enzymes, are present in both

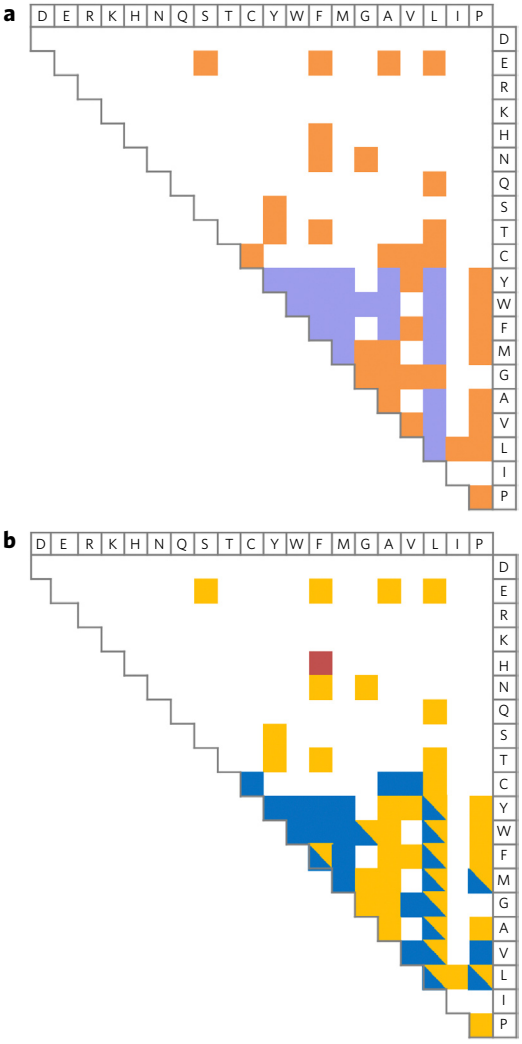

Figure 3 | Diversity of cyclodipeptides produced by the CDPS subfamilies. (**a**) Cyclodipeptides produced by the 11 previously characterized CDPSs (in violet) and the newly identified CDPSs (in orange). The two amino acids constituting the cyclodipeptides are indicated in the one-letter code at the top and right. The presence of a colored square at the intersection of a row and a column indicates that the corresponding cyclodipeptide was detected in the culture supernatant of bacteria expressing a CDPS. (**b**) Cyclodipeptides produced by the NYH (in blue), XYP (in yellow) and SYQ (in dark red) members. The presence and nature of cyclodipeptides are indicated as described in **a**.

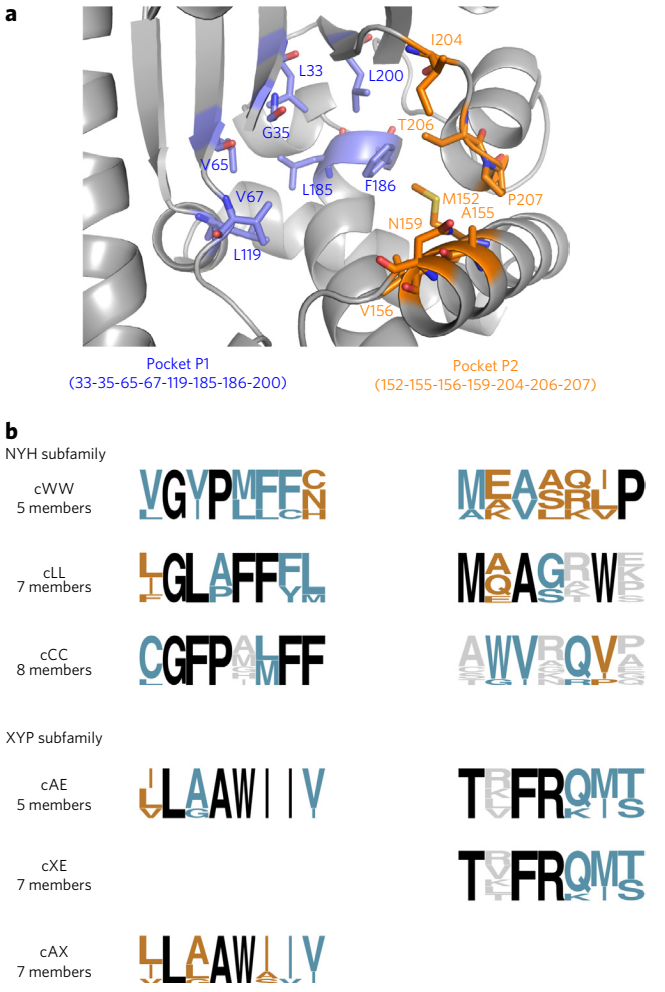

Figure 4 | Sequence logos of the amino acids constituting the two binding pockets, P1 and P2, for groups of CDPSs synthesizing the same cyclodipeptides. (**a**) Three-dimensional view of the binding pockets P1 and P2 accommodating the aminoacyl moieties of the first and second aa-tRNA in AlbC[6]. The eight residues constituting P1 are in violet and the seven constituting P2 are in orange. (**b**) Residues suspected to delineate P1 and P2 were determined for each CDPS (**Supplementary Data Set 7**). Logo sequences corresponding to the frequency plot of amino acids at positions in P1 and P2 were generated for groups of CDPSs belonging to the same subfamily and having similar cyclodipeptide-synthesizing activity: cWW (CDPSs 14, 42, 47, 48 and Amir_4627: 35–55% sequence identity); cLL (CDPSs 13, jk0923, Plu0297, pSHaeC06, YvmC_BSU, YvmC_BLIC and YvmC_BTHU: 26–70% sequence identity); cCC (CDPSs 1, 2, 3, 4, 5, 6, 7 and 9: 28–93% sequence identity; cAE (CDPSs 37, 38, 39, 40 and 43: 40–58% sequence identity); cXE (CDPSs 36, 37, 38, 39, 40, 43 and 44: 38–58% sequence identity); cAX (CDPSs 28, 31, 37, 38, 39, 40 and 43: 16–58% sequence identity). At each position, the presence of a single, two, three or more amino acids is indicated in black, teal blue, tan or gray, respectively.

NYH and XYP subfamilies. For CDPSs with the same activities, the residues constituting P1 and P2 appear well conserved within a subfamily, but show significant differences between subfamilies, indicating that different combinations of amino acids are used to synthesize the same product, in agreement with separate evolution. Thus, CDPSs of the same subfamily and having the same activity appear to constitute a specific group (**Supplementary Data Set 7**).

For groups that were sufficiently populated, that is had at least five members, we created sequence logos to represent sequence motifs of the P1 and P2 pockets. This was performed for four groups: those with cLL-, cWW-, cCC- and cAE-synthesizing members (**Fig. 4b**). As P1 and P2 pockets are independent[6], we considered combining the groups accommodating the same residue in one of the two pockets, regardless the residue accommodated in the other pocket. However, the identification of the site involved in amino acid binding is a key concern for CDPSs. We can use our knowledge of AlbC to infer the roles of P1 and P2 in the other CDPSs. Most CDPSs are promiscuous like AlbC and produce several cyclodipeptides of general formulae c(AA_1-X), where AA_1 is the preferred amino acyl and X is any of the incorporated amino acids. Assuming a similar mechanism for all CDPSs—sequential ping-pong mechanism with acyl enzyme intermediate formation—we can suggest amino acids accommodated by P1 and P2 for each CDPS, with the preferred amino acyl moiety binding to P1. Thus, combination was possible for cAA- and cAE- synthesizing groups, for which P1 accommodates A (its hydrophobicity excludes the accommodation of E). The resulting cAX-synthesizing group was sufficiently populated (seven members) to create a sequence logo for P1. Similarly, we combined

Table 2 | Prediction and validation of the cyclodipeptide-synthesizing activities of CDPSs

CDPS[a]	Species	Pocket residues[b] P1	Pocket residues[b] P2	Prediction	Sequence identity[c]	Cyclodipeptides synthesized[d]
50	*Streptomyces* sp. F12	CGFPWLFF	AWVGQPDV	cCC	34–52%	cCC (100%)
51	*Nocardiopsis gilva*	CGYPSLFF	AWVGRPEF	cCC	34–52%	–
52	*Kutzneria albida* DSM 43870	CGFPSLFF	AWVRQVWF	cCC	32–57%	cCC (100%)
53	*Staphylococcus pseudintermedius*	IGLAFFFM	MQASAWE	cLL	26–76%	–
54	*Bacillus* sp. 171095_106	LGLAFFFL	MAAGKWR	cLL	30–63%	cLL (56%), cLF (23%), cLM (20%)
55	*Bacillus* sp. NSP9.1	LGLAFFFL	MAAGRWK	cLL	29–82%	cLL (52%), cLF (28%), cLM (20%)
56	*Streptomyces albus*	VGIPMFFC	ATVARLP	cWW	35–56%	cWW (100%)
57	*Streptomyces* sp. CNH287	VGVPMFFC	MDVAQLP	cWW	39–60%	cWW (100%)
58	*Nocardiopsis halophila*	VGVPMFFN	MEVARLP	cWW	39–61%	cWW (100%)
59	*Xenorhabdus doucetiae*	VLAAWIII	TVFRKMT	cAE	37–66%	cAE (92%), cSE (7%)
60	*Photorhabdus luminescens* BA1	ILAAWIIV	TIFRQMT	cAE	42–54%	cAE (100%)
61	*Mycobacterium neoaurum*	ILAAWIIV	TRFRQIS	cAE	41–69%	cAE (80%), cAP (18%)

[a]CDPSs are grouped according to the prediction of activity: 50–52, cCC; 53–55, cLL; 56–58, cWW; 59–61, cAE. [b]The predicted residues of the P1 and P2 pockets are indicated for each CDPS. [c]Clustal Omega at the EMBL website was used to determine sequence identities between each CDPS and members of the predicted group (as defined in **Fig. 4b**). [d]–, no cyclodipeptide detected.

the cAE- and cLE-synthesizing groups to give the cXE-synthesizing group (seven members) and create a sequence logo of P2 (**Fig. 4b**). The production of any particular cyclodipeptide correlates with the conservation at specific positions in sequence motifs. However, with the knowledge currently available, sequence motif composition is insufficient to provide a complete understanding of why a particular aminoacyl is incorporated, although some pieces of information emerge. For example, for the cLE- and cAE-synthesizing groups (**Supplementary Data Set 7**), the discrimination between A and L may be based on the second residue of P1 (position 35), for which the presence of an alanine or a leucine correlates with the synthesis of cLE or cAE, respectively.

Sequence motifs allow discrimination between CDPSs that are clustered on the tree but have different cyclodipeptide-synthesizing activities. For example, in the same NYH subfamily, CDPS 8 is clustered on the phylogenetic tree with cCC-synthesizing CDPSs even though it produces cGV (**Fig. 1**; **Table 1**). However, its P1 and P2 residues differ substantially from those in the sequence motif of the cCC-synthesizing CDPSs (**Supplementary Data Set 7**).

We investigated whether sequence motifs could be used to predict the specificity of putative CDPSs. We searched databases for new putative CDPSs at the time of writing (September 2014) and retrieved about 200 further sequences. The set of sequences was curated to remove partial sequences, and about 30 sequences with N-terminal truncations were completed as for the experiment set. We then constructed a new phylogenic tree representing the extended putative CDPS family (**Supplementary Data Set 8**). All new putative CDPSs were distributed into the three subfamilies and fulfilled the functional subsequence signatures described above. We next compared the residues constituting P1 and P2 of all the new putative CDPSs to the sequence motifs described above (**Fig. 4b**). This led us to the identification of CDPSs likely to be members of the cLL-, cWW-, cCC- and cAE-synthesizing groups. We selected three proteins for each group, sharing as little sequence identity as possible with the characterized CDPSs of the group, and determined their cyclodipeptide-synthesizing activities. We found ten CDPSs to be active, and in all cases the main cyclodipeptides produced were as predicted (**Table 2**; **Supplementary Data Set 2**). Using the link between sequence motif and CDPS specificity, we were able to predict the activity of putative CDPSs, in spite of the only moderate sequence identity between these CDPSs and members of the predicted group (**Table 2**). Among the 200 newly identified putative CDPSs, about 30% have P1- and P2-constituting residues very similar to those of poorly populated groups (those with fewer than five members). These putative CDPSs are listed with the appropriate groups in **Supplementary Data Set 9**. Accumulating sequence and biochemical data should lead to the identification of additional sequence motifs corresponding to as yet undescribed cyclodipeptide-synthesizing groups.

DISCUSSION

Our study reveals that all CDPS members—whether or not they are biochemically characterized—can be classified into two main, phylogenetically distinct subfamilies. The NYH and XYP subfamilies contain almost all CDPS members, with the NYH subfamily being the most numerous. We identified only CDPS 17 and one additional member, both belonging to the same Bacteroidetes phylum (**Supplementary Data Set 8**), as belonging to a probable third subfamily named SYQ. This distribution between subfamilies may not accurately reflect that in nature, as the choice of genomes for sequencing is heavily biased toward organisms pathogenic to humans, animals and plants.

Only very few active CDPS members do not comply with all criteria for subfamily membership. In particular, several NYH members have a noncanonical pair X40-X203. Availability of new CDPS sequences is likely to reveal whether the atypical CDPSs form relevant subgroups in the different subfamilies and to what extent their study may reveal previously unknown features of CDPSs.

Current understanding of CDPSs has mostly come from the crystallographic structures of three NYH members determined so far[2–4,6] and from related biochemical studies[7]. The identification of XYP and SYQ members shows that the two residues forming the pair 'N40, H203' and previously demonstrated to be essential for the catalysis in NYH members are not conserved. Although the different catalytic steps of the reaction are probably the same throughout the whole CDPS family, this observation means that alternative solutions have been found to fulfill the roles of these two residues. Structural characterization of XYP members will be especially informative as the P203 residue may induce structural variations around the catalytic loops as compared to NYH members, and affect both the positioning of catalytic residues and the composition of amino acids delineating the binding pocket P2 accommodating the aminoacyl moiety of the second aa-tRNA substrate[6]. Structural characterization associated with biochemical studies will probably also improve our ability to predict the cyclodipeptides synthesized by new CDPSs.

The CDPSs characterized to date synthesize 54 different cyclodipeptides, containing 17 of the 20 proteinogenic amino acids. About

55% of the 200 new putative CDPSs are predicted to synthesize cCC, cLL, cWW or cAE (**Supplementary Data Set 8**), and about 30% may be members of already identified but poorly populated cyclodipeptide-synthesizing groups. By implication, therefore, the remaining putative CDPSs (about 15%) may produce different cyclodipeptides. These observations clearly show that CDPSs are able to synthesize a large variety of cyclodipeptides.

The physiological relevance of the substrate specificities of CDPSs determined upon overexpression in *E. coli* is an issue. We previously identified the specificity determinants for the two substrates of the model CDPS AlbC: the aminoacyl moiety is essential for the recognition of the first substrate, and both the aminoacyl moiety and the N^1-N^{72} base pair of the tRNA moiety participate in the recognition of the second substrate[6,7]. As (i) the sequence of the tRNA moiety does not matter for the binding of the first substrate, (ii) the N^1-N^{72} base pair is strongly conserved in prokaryotic sequences (**Supplementary Data Set 6**) and (iii) aminoacyl-tRNAs are present in significant quantities in all living cells, it is likely that the main product(s) of CDPS activity observed in our study are similar to (or the same as) those produced in the native organisms. Indeed, the activities of a few CDPSs in their native organisms have been characterized, and they correspond to the main products observed upon overexpression in *E. coli*: cLL for YvmC from *Bacillus* species[1,15], cFL for AlbC from *Streptomyces noursei*[16,17] and cFY for Ndas_1148 from *Nocardiopsis dassonvillei*[9]. The main products observed upon CDPS overexpression in *E. coli* also correspond to the CDPS activity detected *in vitro*, with the aminoacyl-tRNAs being in excess with respect to the CDPS[1,7–9]. However, the experimental conditions used in our study probably do not reflect the conditions, and particularly the level of expression of the CDPS, in the native organism. Although these conditions do not appear to influence the nature of the main cyclodipeptides produced, they may cause the production of minor products that are not synthesized *in vivo* in the native organism.

CDPSs are generally genetically associated with cyclodipeptide-tailoring enzymes in biosynthetic pathways dedicated to the synthesis of diketopiperazines[11,18]. Only five CDPS-dependent pathways have been fully characterized so far, and the incorporation of hydrophobic amino acids into cyclodipeptides and their modification by oxidation and methylation reactions have been described[8,9,17,19–21]. The identification of the cyclodipeptide-synthesizing activities of many CDPSs provides valuable data for unraveling a large number of CDPS-dependent pathways and for elucidating and describing the chemical diversity encoded by these pathways.

In conclusion, our work improves knowledge of the CDPS family and gives clues for deciphering CDPS-dependent biosynthetic pathways and their engineering with the aim of producing novel natural products.

Received 20 October 2014; accepted 5 June 2015;
published online 3 August 2015

METHODS

Methods and any associated references are available in the online version of the paper.

References

1. Gondry, M. *et al.* Cyclodipeptide synthases are a family of tRNA-dependent peptide bond-forming enzymes. *Nat. Chem. Biol.* **5**, 414–420 (2009).
2. Vetting, M.W., Hegde, S.S. & Blanchard, J.S. The structure and mechanism of the *Mycobacterium tuberculosis* cyclodityrosine synthetase. *Nat. Chem. Biol.* **6**, 797–799 (2010).
3. Sauguet, L. *et al.* Cyclodipeptide synthases, a family of class-I aminoacyl-tRNA synthetase-like enzymes involved in non-ribosomal peptide synthesis. *Nucleic Acids Res.* **39**, 4475–4489 (2011).
4. Bonnefond, L. *et al.* Structural basis for nonribosomal peptide synthesis by an aminoacyl-tRNA synthetase paralog. *Proc. Natl. Acad. Sci. USA* **108**, 3912–3917 (2011).
5. Seguin, J. *et al.* Nonribosomal peptide synthesis in animals: the cyclodipeptide synthase of *Nematostella*. *Chem. Biol.* **18**, 1362–1368 (2011).
6. Moutiez, M. *et al.* Unravelling the mechanism of non-ribosomal peptide synthesis by cyclodipeptide synthases. *Nat. Commun.* **5**, 5141 (2014).
7. Moutiez, M. *et al.* Specificity determinants for the two tRNA substrates of the cyclodipeptide synthase AlbC from *Streptomyces noursei*. *Nucleic Acids Res.* **42**, 7247–7258 (2014).
8. Giessen, T.W., von Tesmar, A.M. & Marahiel, M.A. A tRNA-dependent two-enzyme pathway for the generation of singly and doubly methylated ditryptophan 2,5-diketopiperazines. *Biochemistry* **52**, 4274–4283 (2013).
9. Giessen, T.W., von Tesmar, A.M. & Marahiel, M.A. Insights into the generation of structural diversity in a tRNA-dependent pathway for highly modified bioactive cyclic dipeptides. *Chem. Biol.* **20**, 828–838 (2013).
10. Aravind, L., de Souza, R.F. & Iyer, L.M. Predicted class-I aminoacyl tRNA synthetase-like proteins in non-ribosomal peptide synthesis. *Biol. Direct* **5**, 48 (2010).
11. Belin, P. *et al.* The nonribosomal synthesis of diketopiperazines in tRNA-dependent cyclodipeptide synthase pathways. *Nat. Prod. Rep.* **29**, 961–979 (2012).
12. Giessen, T.W. & Marahiel, M.A. The tRNA-dependent biosynthesis of modified cyclic dipeptides. *Int. J. Mol. Sci.* **15**, 14610–14631 (2014).
13. Stachelhaus, T., Mootz, H.D. & Marahiel, M.A. The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. *Chem. Biol.* **6**, 493–505 (1999).
14. Challis, G.L., Ravel, J. & Townsend, C.A. Predictive, structure-based model of amino acid recognition by nonribosomal peptide synthetase adenylation domains. *Chem. Biol.* **7**, 211–224 (2000).
15. Tang, M.R., Sternberg, D., Behr, R.K., Sloma, A. & Berka, R.M. Use of transcriptional profiling & bioinformatics to solve production problems. *Ind. Biotechnol. (New Rochelle N.Y.)* **2**, 66–74 (2006).
16. Fukushima, K., Yazawa, K. & Arai, T. Biological activities of albonoursin. *J. Antibiot. (Tokyo)* **26**, 175–176 (1973).
17. Lautru, S., Gondry, M., Genet, R. & Pernodet, J.L. The albonoursin gene cluster of *S. noursei:* biosynthesis of diketopiperazine metabolites independent of nonribosomal peptide synthetases. *Chem. Biol.* **9**, 1355–1364 (2002).
18. Gu, B., He, S., Yan, X. & Zhang, L. Tentative biosynthetic pathways of some microbial diketopiperazines. *Appl. Microbiol. Biotechnol.* **97**, 8439–8453 (2013).
19. Gondry, M. *et al.* Cyclic dipeptide oxidase from *Streptomyces noursei*. Isolation, purification and partial characterization of a novel amino acyl alpha,beta-dehydrogenase. *Eur. J. Biochem.* **268**, 1712–1721 (2001).
20. Belin, P. *et al.* Identification and structural basis of the reaction catalyzed by CYP121, an essential cytochrome P450 in *Mycobacterium tuberculosis*. *Proc. Natl. Acad. Sci. USA* **106**, 7426–7431 (2009).
21. Cryle, M.J., Bell, S.G. & Schlichting, I. Structural and biochemical characterization of the cytochrome P450 CypX (CYP134A1) from *Bacillus subtilis*: a cyclo-L-leucyl-L-leucyl dipeptide oxidase. *Biochemistry* **49**, 7282–7296 (2010).

Acknowledgments

This work was supported by the CEA, the CNRS, the Paris-Sud University and a grant from the French National Research Agency (ANR 2010/Blan 1501 01) to M.G. and J.-L.P. I.B.J. was supported by a doctoral fellowship from the CEA. The Service d'Ingénierie Moléculaire des Protéines is member of the Laboratory of Excellence LERMIT. We warmly thank V. Dive, head of the Service d'Ingénierie Moléculaire des Protéines, for his continuous support and encouragement throughout this work. We thank O. Lespinet for advice about the building of phylogenetic trees, D. Vallenet, M. Stam and M. Sorokina for helpful discussion on bioinformatics, and A. Ponties for technical assistance in cloning experiments. We are indebted to L. Beuvier for skillful assistance in mass data analysis. We thank F. Fenaille for kindly performing the experiments using the Orbitrap mass spectrophotometer. We thank P. Kessler and O. Lequin for skillful assistance in NMR experiments.

Author contributions

M.G. obtained funding. M.M., J.-L.P., M.G. and P.B. developed the hypothesis and designed the study. I.B.J., M.M. and P.B. performed bioinformatic analyses. I.B.J., J.W., E.D. and C.M. performed cloning experiments and prepared culture supernatants. I.B.J. performed LC/MS/MS analysis. I.B.J. and M.M. analyzed MS/MS data. I.B.J., J.S., E.F. and P.B. purified cyclodipeptides from culture supernatants. S.D. performed amino acid composition analyses. A.L. and S.D. chemically synthesized cyclodipeptides. R.T. performed high-resolution mass spectrometry. I.B.J., M.M., J.W., E.D., J.S., J.L.P., M.G. and P.B. analyzed and discussed the results. I.B.J., M.M., M.G. and P.B. prepared the draft manuscript. All of the authors participated in the production of the final version of the manuscript.

Competing financial interests

The authors declare no competing financial interests.

Additional information

Supplementary information is available in the online version of the paper. Reprints and permissions information is available online at http://www.nature.com/reprints/index.html. Correspondence should be addressed to P.B. or M.G.

ONLINE METHODS

Chemicals, bacterial strains and plasmids. All chemicals were from Sigma-Aldrich unless otherwise stated. We used *Escherichia coli* DH5α obtained from Invitrogen for cloning experiments. *E. coli* BL21AI and M15 were obtained from Invitrogen and Qiagen, respectively.

Plasmid pQE60 (Qiagen: ColE1 origin of replication, ampicillin resistance) is an expression vector that carries an IPTG-inducible promoter for protein expression. Plasmid pREP4 (Qiagen: p15A origin of replication, kanamycin resistance) is a repressor plasmid that allows the constitutive expression at high levels of the *lac* repressor for tight control of IPTG-inducible promoters. For CDPS expression, we constructed pIJ196, a pQE60/pREP4 hybrid vector that carries the IPTG-inducible promoter and allows strong expression of the LacI repressor. Briefly, a 1.5 kb *Bst*Z17I-*Sma*I fragment carrying the *lacI* gene and its promoter was excised from pREP4 and ligated into the *Bst*Z17I restriction site of pQE60 such that replication from ColE1 and transcription from *lacI* promoter are in the same direction on the plasmid (**Supplementary Fig. 1**). The cloning sites in the resulting plasmid, pIJ196, were checked by sequencing (MWG-Eurofins). Restriction enzymes were from Thermo Scientific. Standard methods were used for DNA manipulations[22].

Bioinformatics. *Identification of the new putative CDPSs.* Sequence databases at NCBI and Broad Institute were screened for genes encoding putative CDPSs: PSI-Blast searches[23,24] were conducted with an *E*-value threshold of 0.005 and using sequences of several biochemically characterized CDPSs as query. After each PSI-Blast run, hits irrelevant to a CDPS sequence were omitted for the next run. Five or six iterative runs were performed and we obtained ~90 different hits. Sequences encoded by database entries C791_7200, SsomD4_010100015357, SJA_C1-32620, Dmas2_010100010809, K530_11977 and pc1814 showed first residues aligning with positions 56, 32, 24, 66, 50 and 26 of AlbC, respectively (**Supplementary Fig. 6**), suggesting possible start codon misannotation. Examination of the 5′ surrounding DNA sequence of each corresponding coding sequence led us to identify an alternative start codon located upstream, leading to longer amino acid sequences that correctly matched the N-terminal part of AlbC. Thus, these extended versions were selected for CDPS activity screening. The PSI-Blast hit SCAT_0901 encodes an amino acid sequence with an extra N-terminal part that has no homology with AlbC and is rich in proline and alanine residues (**Supplementary Fig. 7**). Examination of the coding sequence indicated the presence of an ATG triplet coding Met52. As Met52 encoded by SCAT_0901 aligned with Ala8 of AlbC, we decided to select the shortened version starting at Met52 for CDPS activity screening. Seven other partial sequences were removed from the set.

Tree calculation. Tree of sequences available in May 2013 (**Fig. 1**): After PSI-Blast searches and a set of relevant sequences had been obtained as described above, we generated a multiple sequence alignment using Muscle[25] integrated into Seaview[26]. This alignment was not manually adjusted, but the accurate alignment of catalytic and conserved residues was checked. The tree was calculated using the PhyML program (v 3.1)[26] based on the maximum-likelihood method. The LG substitution model and the NNI tree searching operation were selected. Reliability of internal branching was assessed using the aLRT test (SH-Like). The iTOL software was used for graphical representation and edition of the phylogenetic tree[27,28]. Tree of sequences available at the time of writing (September 2014) (**Supplementary Data Set 8**): After PSI-Blast searches performed as described above, we obtained about 300 hits. We corrected sequences truncated at the N terminus when possible and we removed sequences truncated at the C terminus, redundant sequences and sequences lacking the catalytic serine (especially putative CDPS 12 and 15). The sequences of CDPSs 26 and 25 were added to the set. Multiple sequence alignment was performed with Muscle[25] integrated into Seaview[26] and was manually adjusted to align the conserved and catalytic residues. Insertions or domains in C terminus which did not align were removed from the multiple alignment. The tree was constructed as described above.

Selection of a set of 49 sequences. The putative CDPSs that shared more than 75% sequence identity with an already characterized CDPS or with another selected CDPS were excluded from the selection (19 sequences), with the exception of CDPS 41 (88% identity with AlbC), and CDPSs 5 and 6 (93% identity with each other) (**Supplementary Data Set 3**). Three sequences originating from metagenomes and likely to contain sequence errors were also excluded. The final set of 49 proteins selected for CDPS activity screening is given in **Supplementary Data Set 1**.

Selection of a set of 12 additional sequences. To assess the predictive value of sequence logos of P1 and P2 pockets determined for the four groups of CDPSs synthesizing cWW, cLL, cCC and cAE, respectively (**Fig. 4b**), we choose 12 CDPS sequences predicted to belong to one of the four groups (three per group) as listed in **Supplementary Data Set 8**. The selected sequences clustered with members of the group on the phylogenetic tree (**Supplementary Data Set 8**) and displayed sequence motifs for P1 and P2 that are consistent with the sequence logos for the corresponding group. Furthermore, the CDPSs were selected to display as little sequence identity as possible with the biochemically characterized CDPSs of the corresponding group. The final set of the 12 additional proteins is given in **Supplementary Data Set 1**.

HHPred Prediction. HHPred at the Max Planck Institute website was used to predict secondary structures for each putative CDPS by searching the RCSB Protein Data Bank (PDB) using default parameters (multiple sequence alignment generation method, HHblits; *E*-value threshold for multiple sequence alignment generation, 10^{-3}; alignment mode, local)[29].

Topology diagrams. Topology diagrams of the three structurally characterized CDPSs were obtained with PDBSum at the EMBL-EBI website, and a topology diagram was deduced by conserving secondary structures found in all CDPS crystal structures.

Sequence logos. The sequence logos of CDPSs consist of stacks of symbols representing amino acids in one letter code[30]. They were obtained at the WebLogo website[31]. Each stack presents a single position in the sequence. The height of the symbols within the stack indicates the relative frequency of each amino acid in the sequence.

CDPSs genes. Synthetic genes encoding CDPSs 1–46 and optimized for expression in *E. coli* were obtained from GeneArt. They were designed to have an *Nco*I restriction site containing the ATG start codon and a *Bgl*II restriction site located downstream from the last codon of the coding sequence. If required for further DNA manipulations, an alanine-encoding codon was introduced after the start codon. The synthetic genes were provided in GeneArt-specific cloning vectors. Their sequences are given in **Supplementary Data Set 1**. Genes encoding CDPSs 47–49 were obtained by PCR amplification of chromosomal DNA from *Streptomyces* sp. AA4 (now referred to as *Amycolatopsis* sp. AA4), *Streptomyces cattleya* NRRL 8057 and *Streptomyces venezuelae* ATCC 10712, respectively. Chromosomal DNA was prepared as follows. Mycelium was disrupted using glass beads and a FastPrep-24 instrument (MP Biomedicals) in the presence of 400 µl of a phenol/chloroform/isoamyl alcohol mixture (25:24:1 v/v/v) and 400 µl of water. The aqueous phase was collected after centrifugation (16,000*g*, 10 min) and precipitated by incubation for 30–60 min at −20 °C with 800 µl of propan-2-ol and 120 µl of 3 M sodium acetate. The pellet was centrifuged (16,000*g*, 15 min), washed with 70% ethanol and resuspended in TE buffer containing RNAse at 50 µg/ml. Phusion High Fidelity DNA Polymerase (ThermoScientific) and the GC protocol provided in the manufacturer's instructions were used for PCR amplification; betaine was added to the mix to the final concentration of 1 M. Oligonucleotides used for PCR amplification were obtained from Integrated DNA Technologies and are described in **Supplementary Table 1**. The PCR cycle was as follows: 98 °C for 30 s; 30 cycles of 98 °C for 10 s, 72 °C for 60 s; 72 °C for 10 min. After incubation with Taq DNA polymerase to add 3′ adenine overhangs to each end (Qiagen; according to manufacturer instructions), PCR products were ligated into pGEM-T Easy (Promega). Plasmid DNA was prepared from positive transformants and verified by DNA sequencing (Beckman Coulter Genomics). Sequences of the amplicons corresponding to CDPS 47–49 are given in **Supplementary Data Set 1**.

Synthetic genes encoding the 12 additional CDPSs (numbered 50–61) and optimized for expression in *E. coli* were obtained from GeneArt. They were designed as described above. Their sequences are given in **Supplementary Data Set 1**.

CDPS coding sequences were then inserted between the *Nco*I and *Bgl*II restriction sites of pIJ196 for protein expression in *E. coli*. All recombinant plasmids were prepared using DH5α bacteria and verified by DNA sequencing of the promoter and CDPS-encoding regions (Eurofins-MWG).

Expression of CDPSs in medium-throughput format. Initially, we worked with strain BL21AI in which protein production can be triggered using controlled culture medium appropriate for autoinduction. We did not obtain recombinant clones following transformation with pIJ196-derived plasmids

doi:10.1038/nchembio.1868

encoding CDPS 8, 21, 22, 25, 26, 27, 30, 31 and 36. We reasoned that this may have been due to toxicity associated with the CDPS-encoding plasmids and that better control of the repression of CDPS expression was necessary. Indeed, transformation of BL21AI harboring the repressor plasmid pREP4 with the nine plasmids led to appropriate transformants in all cases except for plasmids encoding CDPS 22 and 31. All CDPS-encoding plasmids were derived from pQE60, whose supplier recommends the use of strain M15 pREP4 to alleviate toxicity problems. Accordingly, we transformed M15 pREP4 cells with each of the CDPS-encoding plasmids, and successfully obtained transformants. Finally, we used BL21AI and BL21AI pREP4 and M15 pREP4 to express the recombinant putative CDPSs.

Each recombinant putative CDPS was expressed in *E. coli* M15 pREP4 and BL21AI (harboring or not harboring pREP4) from the corresponding pIJ196-derived plasmid. Bacteria were cultured in 10 ml 24-well plates with round-bottomed wells (Dutscher Scientific) containing 2 ml of the appropriate growth medium, covered with a hydrophobic porous film (VWR), and shaken at 200 rpm. Starter cultures were M9-derived minimum medium supplemented with trace elements and vitamins[1], 200 µg/ml ampicillin, 25 µg/ml kanamycin and 0.5% glucose. They were inoculated with several colonies from competent bacteria freshly transformed with plasmids encoding CDPSs. After an overnight incubation at 37 °C, the starter culture was used to inoculate (1/50) the same M9-derived minimum medium except that glucose was replaced by 0.5% glycerol for M15 bacteria or by a combination of 0.5% glycerol, 0.05% glucose and 0.02% lactose for BL21AI bacteria[32]. M15 bacteria were grown at 37 °C until the OD_{600} reached 0.6, and expression of the putative CDPS was induced by the addition of isopropyl-β-D-thiogalactopyranoside (IPTG, 2 mM final concentration). Cultivation was continued for 24 h at 20 °C. BL21AI bacteria were grown in an autoinduced medium[32] and thus did not need the addition of IPTG for CDPS expression. After inoculation of the expression cultures, BL21AI bacteria were grown at 37 °C for 3.5 h, and transferred to 20 °C for 20.5 h. At the end of cultivation, cells were pelleted by centrifugation of the plates. The supernatants were collected, acidified (2% TFA final concentration) and frozen at −20 °C. Bacterial pellets were stored at −80 °C until fractionation and protein content analysis.

Cell pellets were thawed on ice, and suspended in 400 µl of ice-cold lysis buffer (150 mM NaCl, 0.1% Triton X-100, 0.5 mg/ml lysozyme, 10 µM phosphoramidon,1 mM phenylmethylsulfonylfluoride (PMSF), 100 mM Tris HCl, pH 8.0)[33]. After 1.5 h on ice, $MgCl_2$ (10 mM final concentration) and 5 units of benzonase (Sigma-Aldrich) were added and the samples incubated for 1 h in an ice-cold bath. The soluble protein fraction was separated from the insoluble fraction by centrifugation (45 min, 3,000*g*, 4 °C), and both fractions were analyzed by SDS-PAGE using the Mini-Protean 3 Dodeca Cell system from Bio-Rad.

Cyclodipeptide identification. Cyclodipeptides were detected by LC/MS/MS analyses on an Agilent 1100 HPLC coupled via a split system to an Esquire HCT ion trap mass spectrometer (Bruker Daltonik GmbH) set in positive mode. Samples were loaded onto an Altantis dC18 column (4.6 mm × 150 mm, 3 µm, 100 Å, Waters) or a Hypercarb column (4.6 mm × 150 mm, 5 µm, 250 Å, ThermoScientific) developed over 50 min with the linear gradient 1 for CDPSs 1–47 (0% to 50% (v/v)) or the linear gradient 2 for CDPSs 48–49 (15% to 65% (v/v)) (solvent A: 0.1% (v/v) formic acid in H_2O, solvent B: 0.1% (v/v) formic acid in acetonitrile/H_2O (90/10), flow rate, 0.6 ml/min). Previous work showed that the Atlantis dC18 3 µm 4.6 × 150 mm column effectively separates cyclodipeptides containing at least one residue with a hydrophobic side chain[1]. However, we observed that these HPLC conditions were unsatisfactory for the retention of less hydrophobic cyclodipeptides (**Supplementary Table 2**). Retention on the column is important in our screening assay involving HPLC coupled to mass spectrometry because of a lag phase (around 7 min) before mass data acquisition. We therefore tested the Hypercarb column (Thermo Scientific), described to be effective for the separation of polar compounds[34]. We indeed observed substantial retention of the cyclodipeptides tested on the Hypercarb column (**Supplementary Table 2**). However, the analysis on Hypercarb of highly hydrophobic compounds such as cFF resulted in peaks with long tails, probably due to strong retention on the column. Therefore, we analyzed the bacterial culture supernatants with both the Atlantis dC18 column and the Hypercarb column. Positive electrospray ionization and mass analysis were optimized for the detection of compounds in the range of natural cyclodipeptides as previously described[1,20].

Cyclodipeptides were detected by both their *m/z* value and their daughter-ion spectra[35–41]. Their identities were confirmed in various ways. For most, the nature of the detected cyclodipeptides was unambiguously established either by comparison with data from literature (cPV and cPP[35]; cWW[8]; cLL, cLM, cFL, cMM, cFM, cYM, cFF, cYF, cYA, cYL and cYY[1]) or by comparison with authentic standards purchased from Bachem (cGG, cGA, cAA, cLG, cVV, cEA, cPL, cPM, cFP, cYS, cYP, cHF) or from Sigma (cLA, cLV) or chemically synthesized in our laboratory (cFA, cFV, cFT, cFN, and cFE; see below). The cyclodipeptides cYV, cGN, cGV, cAP, cLE and cyclo(L-cystine) were purified from culture supernatants of recombinant *E. coli* expressing CDPS 22, 27, 8, 18, 36 and 9, respectively. cYV was purified using a Purospher STAR RP-18 endcapped LiChroCART 250-10 column (10 mm × 250 mm, 5 µm, 120 Å, Merck); cGN, cGV, cAP, cLE and cyclocystine were purified with a Hypercarb column (10 mm × 150 mm, 5 µm, 250 Å, Thermo Scientific). The HPLC purifications were carried out at 4.75 ml/min using buffer A consisting of 0.1% TFA in water, buffer B of 90% acetonitrile in water with 0.09% TFA, and the gradients depicted in the **Supplementary Table 3**. Purified cyclodipeptides were obtained at purity >90% as estimated by UV chromatograms of analytical HPLC recorded at 220 nm (cYV: Atlantis dC18, 4.6 × 150 mm, 3 µm, 100 Å (Waters), linear gradient of acetonitrile in water from 0% to 50% (1% per min) in 0.1% TFA at a flow rate of 0.6 ml/min; cGN, cGV, cAP, cLE and cyclocystine: Hypercarb column 4.6 mm × 150 mm, 5 µm, 250 Å (Thermo Scientific), the same gradients as those used for purification, but at a flow rate of 0.6 ml/min). They were analyzed by determination of amino acid composition as below. No clear data could be obtained for cyclocystine. Purified cyclodipeptides were also analyzed by high-resolution mass spectrometry using either MALDI-TOF/TOF as previously described[1] (cYV, cGV, cAP, cLE) or a high-resolution/high-mass-accuracy LTQ-Orbitrap instrument (Thermo Scientific) with a resolution set at 30,000 at *m/z* 400 (cyclocystine and cGN). HRMS (*m/z*): cYV, [MH$^+$] calculated for $C_{14}H_{19}N_2O_3$ 263.1396, found 263.1393; cGN, [MH$^+$] calculated for $C_6H_{10}N_3O_3$ 172.0722, found 172.0713; cGV, [MH$^+$] calculated for $C_7H_{13}N_2O_2$ 157.0977, found 157.0974; cAP, [MH$^+$] calculated for $C_8H_{13}N_2O_2$ 169.0977, found 169.0966; cLE, [MH$^+$] calculated for $C_{11}H_{19}N_2O_4$ 243.1345, found 243.1324; cyclocystine, [MH$^+$] calculated for $C_6H_9N_2O_2S_2$ 205.0100, found 205.0096. Finally, the identity of ten cyclodipeptides produced in only small amounts (cCA, cMG, cCV, cMA, cLT, cLC, cSE, cCE, cLQ and cPW) and one produced in larger amounts (cYT) was not investigated beyond MS/MS analyses.

Chemical synthesis of cFA, cFV, cFT, cFN and cFE. Published procedures were used for the chemical synthesis of cFA, cFV, cFT, cFN, and cFE[42]. They were purified by semi-preparative RP-HPLC (Purospher STAR RP-18 endcapped LiChroCART, 10 mm × 250 mm, 5 µm, 120 Å, Merck) using a linear gradient of acetonitrile in water from 0% to 50% (1% per min) in 0.1% trifluoroacetic acid at a flow rate of 4.75 ml/min. After lyophilization of the samples, a white powder was observed for all compounds. Purified cyclodipeptides were obtained at purity >95% as estimated by UV chromatograms of analytical RP-HPLC recorded at 220 nm (Atlantis dC18, 4.6 × 150 mm, 3 µm, 100 Å, Waters; linear gradient of acetonitrile in water from 0% to 50% (1% per min) in 0.1% formic acid at a flow rate of 0.6 ml/min). The purified cyclodipeptides were characterized by determination of amino acid composition under standard conditions: samples (100 µM solution in 1% DMSO) were vacuum dried and sealed in glass tubes using the PicoTag system (Waters) and hydrolyzed under the vapor phase of 6N HCl with a crystal of phenol for 17h at 110 °C. The hydrolyzed sample was dissolved in 20–50 µl of Milli-Q water, and 5–20 µl of the resulting HCl hydrolysate (containing a minimum of 200 pmol of each amino acid) was analyzed and quantified by ninhydrin derivatization on an aminoTac JLC-500/V amino-acid analyzer (JEOL). Standard amino acid solutions were used for calibration at the beginning of each analysis series. Purified, chemically synthesized cyclodipeptides were also characterized by high-resolution mass spectrometry using MALDI-TOF/TOF as previously described[1] and NMR spectroscopy. The NMR experiments were recorded on a Bruker Avance 250 spectrometer. The lyophilized samples were dissolved in DMSO-d_6 (Eurisotop) and spectra were recorded at 25 °C. The data from NMR experiments were processed and analyzed with Bruker TOPSPIN 2.0 program.

cFA. ^{1}H NMR (250 MHz, DMSO): δ 8.13 (s, 1H, H^N F or A), 8.02 (s,1H, H^N F or A), 7.32–7.13 (m, 5H, $H^{\delta\text{-}\varepsilon\text{-}\zeta}$ F), 4.17 (m, 1H, H^α F), 3.61 (qt, J = 7, 1.5 Hz, 1H, H^α A), 3.13 (dd, J = 13.5, 3.75 Hz, 1H, H^β F), 2.85 (dd, J = 13.5, 5 Hz, 1H, $H^{\beta'}$ F),

0.45 (d, J = 7Hz, 3H, H^{β} A); ^{13}C NMR (62.8 MHz, DMSO): δ 167.7 (C′F or A), 165.8 (C′F or A), 136.1 (C^{γ} F), 130.4 (C^{δ} F), 128.1 (C^{ε} F), 126.7 (C^{ζ} F), 55.4 (C^{α} F), 49.7 (C^{α} A), 38.3 (C^{β} F), 19.7 (C^{β} A). HRMS (*m/z*): [MH^+] calculated for $C_{12}H_{15}N_2O_2$, 219.1133, found 219.1143.

cFV. ^{1}H NMR (250 MHz, DMSO): δ 8.13 (s, 1H, H^N F or V), 7.92 (s, 1H, H^N F or V), 7.27–7.14 (m, 5H, $H^{\delta\text{-}\varepsilon\text{-}\zeta}$ F), 4.21 (m, 1H, H^{α} F), 3.53 (m, 1H, H^{α}V), 3.15 (dd, J = 13.5, 4.25 Hz, 1H, H^{β} F), 2.86 (dd, J = 13.5, 5 Hz, 1H, $H^{\beta'}$ F), 1.72 (m, 1H, H^{β} V), 0.64 (d, J = 7.1 Hz, 3H, H^{γ} V), 0.24 (d, J = 6.8 Hz, 3H, $H^{\gamma'}$ V); ^{13}C NMR (62.8 MHz, DMSO): δ 166.6 (C′F or V), 166.4 (C′F or V), 136.3 (C^{γ} F), 130.4 (C^{δ} F), 128.0 (C^{ε} F), 126.5 (C^{ζ} F), 59.2 (C^{α} V), 55.0 (C^{α} F), 37.8 (C^{β} F), 31.0 (C^{β} V), 18.3 (C^{γ} V), 16.2 ($C^{\gamma'}$ V). HRMS (*m/z*): [MH^+] calculated for $C_{14}H_{19}N_2O_2$, 247.1446, found 247.1444.

cFT. ^{1}H NMR (250 MHz, DMSO): δ 8.05 (d, J = 2.75 Hz, 1H, H^N F or T), 7.89 (d, J = 2.75 Hz, 1H, H^N F or T), 7.32–7.16 (m, 5H, $H^{\delta\text{-}\varepsilon\text{-}\zeta}$ F), 5.00 (s, 1H, H^{OH} T), 3.93 (m, 1H, H^{α} F), 3.69 (m, 1H, H^{β}T), 3.52 (t, J = ~3 Hz, 1H, H^{α} T), 3.16–3.01 (ABX, 2H, H^{β} F, $H^{\beta'}$ F), 0.96 (d, J = 6.5 Hz, 3H, H^{γ} T); ^{13}C NMR (62.8 MHz, DMSO): δ 167.4 (C′F or T), 166.7 (C′F or T), 137.4 (C^{γ} F), 129.8 (C^{δ} F), 128.2 (C^{ε} F), 126.4 (C^{ζ} F), 67.2 (C^{β} T), 60.3 (C^{α} T), 55.9 (C^{α} F), 40.8 (C^{β} F), 19.7 (C^{γ} T). HRMS (*m/z*): [MH^+] calculated for $C_{13}H_{17}N_2O_2$, 249.1239, found 249.1236.

cFN. ^{1}H NMR (250 MHz, DMSO): δ 8.10 (s, 1H, H^N F or N), 7.62 (s,1H, H^N F or N), 7.32–7.17 (m, 6H, $H^{\delta\text{-}\varepsilon\text{-}\zeta}$ F, H^{NH2} N), 6.94 (s, 1H, $H^{NH2'}$), 4.19 (m 1H, H^{α} F), 4.02 (m, 1H, H^{α} N), 3.10 (dd, J = 13.75, 4.75 Hz, 1H, H^{β} F), 2.92 (dd, J = 13.75, 5.25 Hz, 1H, $H^{\beta'}$ F), 2.26 (dd, J = 16, 4 Hz, 1H, H^{β} N), 1.386 (dd, J = 16, 8.75 Hz, 1H, $H^{\beta'}$ N); ^{13}C NMR (62.8 MHz, DMSO): δ 171.6 (C^{γ} N), 166.9 (C′F or N), 166.4 (C′F or N), 136.4 (C^{γ} F), 130.1 (C^{δ} F), 128.1 (C^{ε} F), 126.7 (C^{ζ} F), 55.2 (C^{α} F), 51.1 (C^{α} N), 38.1 (C^{β} F or N), 38.1 (C^{β} F or N). HRMS (*m/z*): [MH^+] calculated for $C_{13}H_{16}N_3O_3$, 262.1192, found 262.1186.

cFE. ^{1}H NMR (250 MHz, DMSO): δ 11.91 (s,1H, H^{COOH} E), 8.21 (s, 1H, H^N F or E), 8.10 (s,1H, H^N F or E), 7.28–7.14 (m, 5H, $H^{\delta\text{-}\varepsilon\text{-}\zeta}$ F), 4.19 (s, 1H, H^{α} F), 3.69 (m, 1H, H^{α} E), 3.14 (dd, J = 13.25, 3.5 Hz, 1H, H^{β} F), 2.84 (dd, J = 13.25, 4.75 Hz, 1H, $H^{\beta'}$ F), 1.66 (m, 2H, H^{γ} E), 1.32–0.97 (m, 2H, H^{β} E); ^{13}C NMR (62.8 MHz, DMSO): δ 173.9 (C^{δ} E), 166.7 (C′F or E), 166.4 (C′F or E), 136.1 (C^{γ} F), 130.3 (C^{δ} F), 128.1 (C^{ε} F), 126.8 (C^{ζ} F), 55.32 (C^{α} F), 53.0 (C^{α} E), 38.1 (C^{β} F), 28.6 ($C^{\beta \text{ or } \gamma}$ E), 28.5 ($C^{\beta \text{ or } \gamma}$ E). HRMS (*m/z*): [MH^+] calculated for $C_{14}H_{17}N_2O_4$, 277.1188, found 277.1187.

22. Sambrook, J., Fritsch, E.F. & Maniatis, T. *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Laboratory Press, 2001).
23. Altschul, S.F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
24. Schäffer, A.A. *et al.* Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.* **29**, 2994–3005 (2001).
25. Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
26. Gouy, M., Guindon, S. & Gascuel, O. SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol. Biol. Evol.* **27**, 221–224 (2010).
27. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* **23**, 127–128 (2007).
28. Letunic, I. & Bork, P. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res.* **39**, W475–W478 (2011).
29. Remmert, M., Biegert, A., Hauser, A. & Soding, J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods* **9**, 173–175 (2012).
30. Schneider, T.D. & Stephens, R.M. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.* **18**, 6097–6100 (1990).
31. Crooks, G.E., Hon, G., Chandonia, J.M. & Brenner, S.E. WebLogo: a sequence logo generator. *Genome Res.* **14**, 1188–1190 (2004).
32. Studier, F.W. Protein production by auto-induction in high density shaking cultures. *Protein Expr. Purif.* **41**, 207–234 (2005).
33. Braud, S. *et al.* Dual expression system suitable for high-throughput fluorescence-based screening and production of soluble proteins. *J. Proteome Res.* **4**, 2137–2147 (2005).
34. Bajad, S.U. *et al.* Separation and quantitation of water soluble cellular metabolites by hydrophilic interaction chromatography-tandem mass spectrometry. *J. Chromatogr. A* **1125**, 76–88 (2006).
35. Chen, Y.-H., Liou, S.-E. & Chen, C.-C. Two-step mass spectrometric approach for the identification of diketopiperazines in chicken essence. *Eur. Food Res. Technol.* **218**, 589–597 (2004).
36. Falick, A.M., Hines, W.M., Medzihradszly, K.F., Baldwin, M.A. & Gibson, B.W. Low-mass ions produced from peptides by high-energy collision-induced dissociation in tandem mass spectrometry. *J. Am. Soc. Mass Spectrom.* **4**, 882–893 (1993).
37. Johnson, R.S., Martin, S.A., Biemann, K., Stults, J.T. & Watson, J.T. Novel fragmentation process of peptides by collision-induced decomposition in a tandem mass spectrometer: differentiation of leucine and isoleucine. *Anal. Chem.* **59**, 2621–2625 (1987).
38. Papayannopoulos, I.A. The interpretation of collision-induced dissociation tandem mass spectra of peptides. *Mass Spectrom. Rev.* **14**, 49–73 (1995).
39. Roepstorff, P. & Fohlman, J. Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomed. Mass Spectrom.* **11**, 601 (1984).
40. Stark, T. & Hofmann, T. Structures, sensory activity, and dose/response functions of 2,5-diketopiperazines in roasted cocoa nibs (*Theobroma cacao*). *J. Agric. Food Chem.* **53**, 7222–7231 (2005).
41. Armirotti, A., Millo, E. & Damonte, G. How to discriminate between leucine and isoleucine by low energy ESI-TRAP MSn. *J. Am. Soc. Mass Spectrom.* **18**, 57–63 (2007).
42. Jeedigunta, S., Krenisky, J.M. & Kerr, R.G. Diketopiperazines as advanced intermediates in the biosynthesis of Ecteinascidins. *Tetrahedron* **56**, 3303–3307 (2000).

doi:10.1038/nchembio.1868