Bioinformatic analysis reveals both oversampled and underexplored biosynthetic diversity in nonribosomal peptides

Bo-Siyuan Jian[1], Shao-Lun Chiou[2], Chun-Chia Hsu[2], Josh Ho[2], Yu-Wei Wu[3,4,5]*, John Chu[2]*

[1] Department of Computer Science and Information Engineering, National Taiwan University, Taipei 10617, Taiwan; [2] Department of Chemistry, National Taiwan University, Taipei 10617, Taiwan; [3] Graduate Institute of Biomedical Informatics, College of Medical Science and Technology, Taipei Medical University, Taipei 10675, Taiwan; [4] Clinical Big Data Research Center, Taipei Medical University Hospital, Taipei 10675, Taiwan; [5] TMU Research Center for Digestive Medicine, Taipei Medical University, Taipei 10675, Taiwan

*Correspondence: yuwei.wu@tmu.edu.tw and johnchu@ntu.edu.tw

Clustering identity

Determining the clustering identity cutoff is key to rarefaction analysis. The nonribosomal code, sometimes referred to as the Stachelhaus code, is a compendium of 10 residues that constitute the A domain active site, which in turn dictates the substrate specificity of the A domain. We defined the "cleanliness" of clustering as the proportion of the most abundant predicted substrate BB within each cluster and found that 70% identity cutoff is suitable for the analysis at hand. Specifically, when the cutoff was set at 60, 70, 80, 90, and 100%, the resulting cleanliness were 99, 97, 89, 71, and 36%, respectively (***Figure S2***). Based on this result, we chose to set the clustering identity at 70%, which is in line with previous bioinformatic studies, as well as recent empirical observations (1-3). See the legend of ***Figure S2*** for further discussion.

Categorizing nonribosomal peptide (NRP) building blocks (BB)

The list of all BB that appear in GenBank predictions, MIBiG database, and the SANDPUMA training set can be found in the *Building Blocks (BB) Groups* tab of ***Table S1***; BB are arranged by group and their source (4, 5). GenBank predictions are arranged by bacterial phyla in a separate *GenBank Predictions* tab. Note that the Ω parameter for the **benzoyl** group was calculated based on A domains associated with the SANDPUMA training set (5) as opposed to the MIBiG database; this is because the former has a larger sample size for this BB group than the latter. The Ω parameter for all other BB groups were calculated using the formula described in the manuscript text.

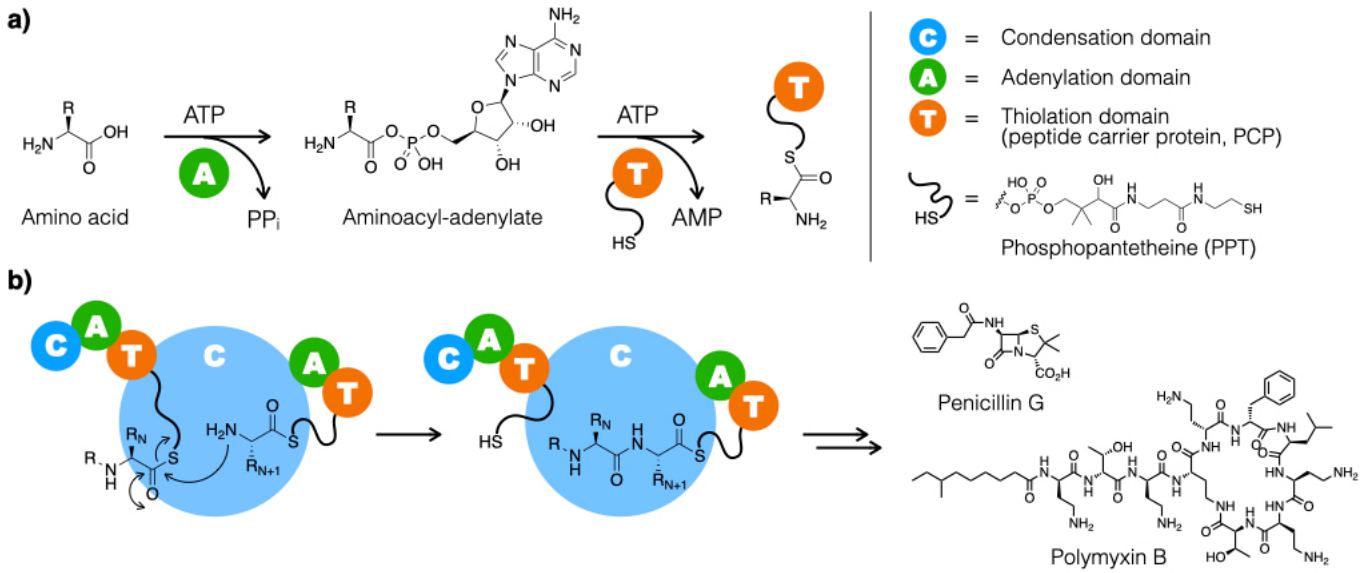
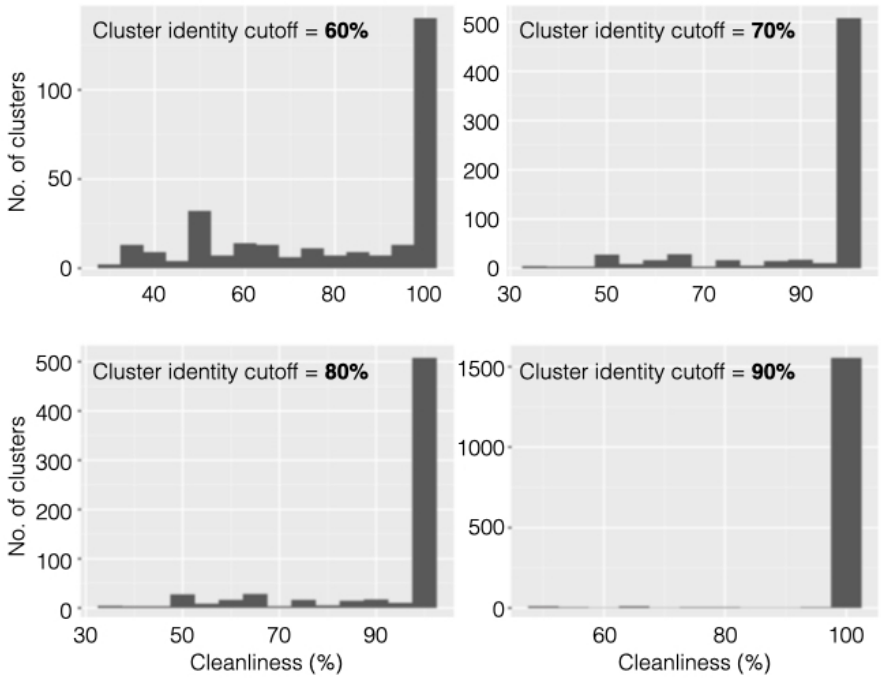


Figure S1. NRP are peptides that are not biosynthesized by the ribosome; they are instead constructed via an enzymatic assembly line. Each module in the assembly line is responsible for incorporating a building block (BB), in most cases an amino acid (AA), into the NRP backbone. A module typically contains multiple semi-autonomously folded domains, each with its own function, including most commonly the condensation (C), adenylation (A), and thiolation (T) domains. The A domain is an enzyme that catalyzes the activation of a substrate BB to form an aminoacyl-adenylate, which is then attached via a thioester bond onto the phosphopantetheine arm of the T domain. Peptide bond formation between BB on neighboring T domains is catalyzed by the C domain in between, wherein the amino group of the BB on the NRP intermediate attacks the activated BB on its *N*-terminal side. This reaction extends the peptide intermediate by one residue and effectively moves it down the assembly line from the N^{th} to the $N+1^{th}$ module. The resulting NRP is colinear to the biosynthetic gene sequences due to such an arrangement.

Setting a high clustering identity cutoff (being too stringent) means that A domains must have nearly identical nonribosomal codes to be placed in the same group. Being too lax has the opposite effect and is not desirable either. Cleanliness, a parameter often used to determine the suitable balance between the two extremes, is defined as the proportion of the most frequently appeared items in a cluster. “Items” are the NRP building blocks in this case. The goal is to identify a clustering identity cutoff with a small number of clusters and a high mean cleanliness.

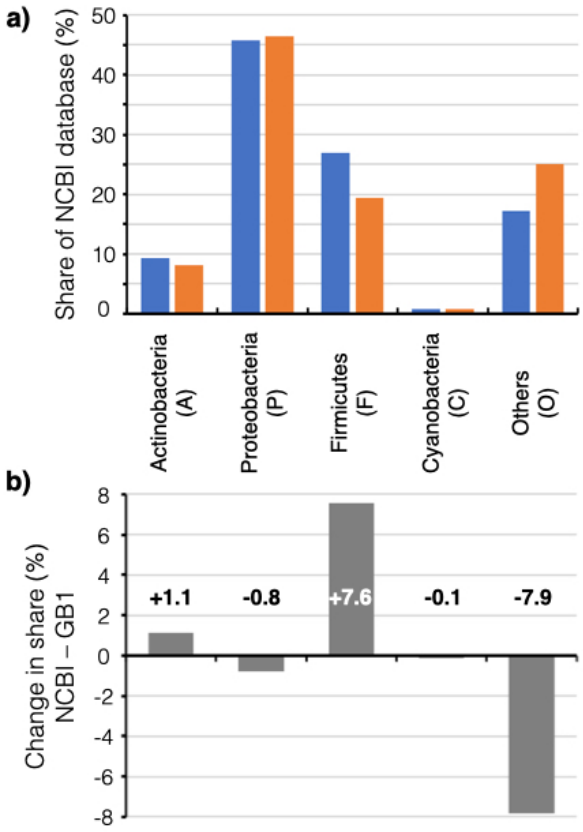
a)



b)

Cutoff	Mean cleanliness	No. of clusters
90%	0.9921	4,591
80%	0.9646	3,022
70%	0.9259	1,894
60%	0.8073	682

Figure S2. a) To determine a suitable clustering identity cutoff, we plotted cleanliness histograms at various cutoffs (90%, 80%, 70%, and 60%). **b)** The mean cleanliness at each threshold was then calculated. We decided to set the clustering identity cutoff as we saw a large drop in both cleanliness and the number of clusters from 70% to 60%.



c)

Species	GenBank enteries	Phylum / Class
Escherichia coli	13,011	P / Gammaproteobacteria
Staphylococcus aureus	8,102	F / Bacilli
Klebsiella pneumoniae	6,172	P / Gammaproteobacteria
Streptococcus pneumoniae	5,724	F / Bacilli
Mycobacterium tuberculosis	5,019	A / Actinomycetia
Acinetobacter baumannii	2,818	P / Gammaproteobacteria
Pseudomonase aeruginosa	2,642	P / Gammaproteobacteria

Figure S3. a) A comparison of the composition at the phylum level of NCBI (GenBank) and GB1. **b)** Changes in phylogenetic composition going from NCBI (GenBank) to GB1. **c)** Species causing the biggest skews are listed.

References

1. T. Stachelhaus, H. D. Mootz, M. A. Marahiel, The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. *Chem. Biol.* **6**, 493-505 (1999).
2. C. Rausch, T. Weber, O. Kohlbacher, W. Wohlleben, D. H. Huson, Specificity prediction of adenylation domains in nonribosomal peptide synthetases (NRPS) using transductive support vector machines (TSVMs). *Nucleic Acids Res.* **33**, 5799-5808 (2005).
3. Z. W. Wei *et al.*, Free Piperazic Acid as a Precursor to Nonribosomal Peptides. *J. Am. Chem. Soc.* **144**, 13556-13564 (2022).
4. S. A. Kautsar *et al.*, MIBiG 2.0: a repository for biosynthetic gene clusters of known function. *Nucleic Acids Res.* **48**, D454-D458 (2020).
5. M. G. Chevrette, F. Aicheler, O. Kohlbacher, C. R. Currie, M. H. Medema, SANDPUMA: ensemble predictions of nonribosomal peptide chemistry reveal biosynthetic diversity across Actinobacteria. *Bioinformatics* **33**, 3202-3210 (2017).