Nucleic Acids Research, 2022 **1**
https://doi.org/10.1093/nar/gkac1049

MIBiG 3.0: a community-driven effort to annotate experimentally validated biosynthetic gene clusters

Barbara R. Terlouw [1,†], Kai Blin [2,†], Jorge C. Navarro-Muñoz [1,3], Nicole E. Avalon [4], Marc G. Chevrette [5], Susan Egbert [6], Sanghoon Lee [7], David Meijer [1], Michael J.J. Recchia [7], Zachary L. Reitz [1], Jeffrey A. van Santen [7,8], Nelly Selem-Mojica [9], Thomas Tørring [10], Liana Zaroubi [7], Mohammad Alanjary [1], Gajender Aleti [11], César Aguilar [12], Suhad A.A. Al-Salihi [13], Hannah E. Augustijn [1,14], J. Abraham Avelar-Rivas[15], Luis A. Avitia-Domínguez [14,15], Francisco Barona-Gómez [14,15], Jordan Bernaldo-Agüero [16], Vincent A. Bielinski [17], Friederike Biermann [1,18,19], Thomas J. Booth [2,20], Victor J. Carrion Bravo [14,21,22], Raquel Castelo-Branco[23,24], Fernanda O. Chagas [25], Pablo Cruz-Morales [2], Chao Du [14], Katherine R. Duncan [26], Athina Gavriilidou [27,28], Damien Gayrard [29], Karina Gutiérrez-García [30], Kristina Haslinger [31], Eric J.N. Helfrich [18,19], Justin J.J. van der Hooft [1,32], Afif P. Jati [33], Edward Kalkreuter [34], Nikolaos Kalyvas [3], Kyo Bin Kang [35], Satria Kautsar [34], Wonyong Kim [36], Aditya M. Kunjapur [37], Yong-Xin Li [38], Geng-Min Lin [39], Catarina Loureiro [40], Joris J.R. Louwen [1], Nico L.L. Louwen [1], George Lund [41], Jonathan Parra [42,43,44], Benjamin Philmus [45], Bita Pourmohsenin [27,28], Lotte J.U. Pronk [1], Adriana Rego [23,46], Devasahayam Arokia Balaya Rex [47], Serina Robinson [48], L. Rodrigo Rosas-Becerra [14,15], Eve T. Roxborough [49], Michelle A. Schorn [40], Darren J. Scobie [26], Kumar Saurabh Singh [1], Nika Sokolova [31], Xiaoyu Tang [50], Daniel Udwary [51], Aruna Vigneshwari [52], Kristiina Vind [53,54], Sophie P.J.M. Vromans [1], Valentin Waschulin [55], Sam E. Williams [56], Jaclyn M. Winter [57], Thomas E. Witte [58], Huali Xie [1,59], Dong Yang [60], Jingwei Yu [61], Mitja Zdouc [1], Zheng Zhong [40], Jérôme Collemare [3], Roger G. Linington [7], Tilmann Weber [2,*] and Marnix H. Medema [1,14,*]

[1]Bioinformatics Group, Wageningen University, Droevendaalsesteeg, 6708 PB Wageningen, The Netherlands, [2]The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kgs. Lyngby, Denmark, [3]Westerdijk Fungal Biodiversity Institute, Uppsalalaan 8, 3584 CT Utrecht, The Netherlands, [4]Scripps Institution of Oceanography, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0212, USA, [5]Department of Microbiology and Cell Science, University of Florida, Gainesville, FL 32611, USA, [6]Department of Chemistry, University of Manitoba, 66 Chancellors Cir, Winnipeg, MB R3T 2N2, Canada, [7]Department of Chemistry, Simon Fraser University, 8888 University Drive, Burnaby, British Columbia V5A 1S6, Canada, [8]Unnatural Products, 2161 Delaware Ave. Suite A, Santa Cruz, CA 95060, USA, [9]Centro de Ciencias Matemáticas UNAM, Morelia, México, [10]Department of Biological and Chemical Engineering, Aarhus University, Denmark, [11]Food and Animal Sciences, Department of Agricultural and Environmental Sciences, Tennessee State University, Nashville, TN 37209, USA, [12]Department of Chemistry, Purdue University, West Lafayette, IN, USA, [13]Department of Applied Sciences,

[*]To whom correspondence should be addressed. Tel: +31 317484706; Email: marnix.medema@wur.nl
Correspondence may also be addressed to Tilmann Weber. Tel: +45 24896132; Email: tiwe@biosustain.dtu.dk
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

© The Author(s) 2022. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkac1049/6833236 by guest on 22 November 2022

University of Technology, Iraq, [14]Institute of Biology, Leiden University, Sylviusweg 72, 2333BE Leiden, The Netherlands, [15]Laboratorio Nacional de Genómica para la Biodiversidad-Unidad de Genómica Avanzada, Cinvestav. Km 9.6 Libramiento Norte Carretera Irapuato-León, CP 36824 Irapuato, Gto., México, [16]Departamento de Microbiología Molecular, Instituto de Biotecnología, Universidad Nacional Autónoma de México, Cuernavaca, Morelos, México, [17]Synthetic Biology and Bioenergy Group, J. Craig Venter Institute, La Jolla, CA 92037, USA, [18]Institute of Molecular Bio Science, Goethe-University Frankfurt, D-60438 Frankfurt am Main, Germany, [19]LOEWE Center for Translational Biodiversity Genomics (TBG), Senckenberganlage 25, 60325 Frankfurt am Main, Germany, [20]School of Molecular Sciences, University of Western Australia, Perth, Australia, [21]Departamento de Microbiología, Instituto de Hortofruticultura Subtropical y Mediterránea 'La Mayora', Universidad de Málaga-Consejo Superior de Investigaciones Científicas (IHSM-UMA-CSIC), Universidad de Málaga, Málaga, Spain, [22]Department of Microbial Ecology, Netherlands Institute of Ecology (NIOO-KNAW), Wageningen, The Netherlands, [23]Interdisciplinary Centre of Marine and Environmental Research (CIIMAR), University of Porto, Portugal, [24]Faculty of Sciences, University of Porto, 4150-179 Porto, Portugal, [25]Instituto de Pesquisas de Produtos Naturais Walter Mors, Universidade Federal do Rio de Janeiro, Rio de Janeiro, RJ, 21941-599, Brazil, [26]University of Strathclyde, Strathclyde Institute of Pharmacy and Biomedical Sciences, 141 Cathedral Street, Glasgow, G4 ORE UK, [27]Translational Genome Mining for Natural Products, Interfaculty Institute of Microbiology and Infection Medicine Tübingen (IMIT), University of Tübingen, Tübingen, Germany, [28]Interfaculty Institute for Biomedical Informatics (IBMI), University of Tübingen, Tübingen, Germany, [29]Department of Molecular Microbiology, John Innes Centre, Norwich Research Park, Norwich, NR4 7UH, UK, [30]Department of Embryology, Carnegie Institution for Science, 3520 San Martin Drive, Baltimore, MD 21218, USA, [31]Department of Chemical and Pharmaceutical Biology, Groningen Research Institute of Pharmacy, University of Groningen, Antonius Deusinglaan 1, 9713 AV Groningen, The Netherlands, [32]Department of Biochemistry, University of Johannesburg, Auckland Park, Johannesburg 2006, South Africa, [33]Indonesian Society of Bioinformatics And Biodiversity, Indonesia, [34]Department of Chemistry, University of Florida Scripps Biomedical Research, 110 Scripps Way, Jupiter, FL 33458, USA, [35]College of Pharmacy, Sookmyung Women's University, Seoul, South Korea, [36]Korean Lichen Research Institute, Sunchon National Universtiy, Suncheon, South Korea, [37]Department of Chemical & Biomolecular Engineering, University of Delaware, Newark, DE 19716, USA, [38]Department of Chemistry, The University of Hong Kong, Pokfulam Road, Hong Kong, P.R. China, [39]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA, [40]Laboratory of Microbiology, Wageningen University, Stippeneng 4, 6708WE, Wageningen, The Netherlands, [41]Sustainable Soils and Crops, Rothamsted Research, Harpenden, Hertfordshire, UK, [42]Instituto de Investigaciones Farmacéuticas (INIFAR), Facultad de Farmacia, Universidad de Costa Rica, San José, 11501-2060, Costa Rica, [43]Centro de Investigaciones en Productos Naturales (CIPRONA), Universidad de Costa Rica, San José, 11501-2060, Costa Rica, [44]Centro Nacional de Innovaciones Biotecnológicas (CENIBiot), CeNAT-CONARE, 1174-1200, San José, Costa Rica, [45]Department of Pharmaceutical Sciences, Oregon State University, USA, [46]Institute of Biomedical Sciences Abel Salazar (ICBAS), University of Porto, Portugal, [47]Centre for Integrative Omics Data Science, Yenepoya (Deemed to be University), Mangalore 575018, India, [48]Department of Environmental Microbiology, Eawag: Swiss Federal Institute for Aquatic Science and Technology, Überlandstrasse 133, CH-8600 Dübendorf, Switzerland, [49]School of Chemistry, University of Nottingham, University Park, Nottingham NG7 2RD, UK, [50]Institute of Chemical Biology, Shenzhen Bay Laboratory, Shenzhen 518132, China, [51]DOE Joint Genome Institute, Lawrence Berkeley National Lab, Berkeley, CA, USA, [52]Department of Microbiology, University of Szeged, Hungary, [53]Host-Microbe Interactomics Group, Wageningen University, 6708 WD Wageningen, The Netherlands, [54]NAICONS Srl, 20139 Milan, Italy, [55]School of Life Sciences, The University of Warwick, Coventry CV4 7AL, UK, [56]School of Biochemistry, University of Bristol, University Walk, Bristol BS8 1TD, UK, [57]Department of Medicinal Chemistry, University of Utah, Salt Lake City, UT 84112, USA, [58]Department of Chemistry and Biomolecular Sciences, University of Ottawa, Ottawa, Canada, [59]Key laboratory of Detection for Biotoxins, Ministry of Agriculture and Rural Affairs and Oil Crops Research Institute, Chinese Academy of Agricultural Sciences, Wuhan 430061, China, [60]Department of Chemistry and Natural Products Discovery Center, UF Scripps Biomedical Research, University of Florida, Jupiter, FL 33458, USA and [61]SUSTech-PKU Institute of Plant and Food Science, Department of Biology, School of Life Sciences, Southern University of Science and Technology, Shenzhen, Guangdong 518055, China

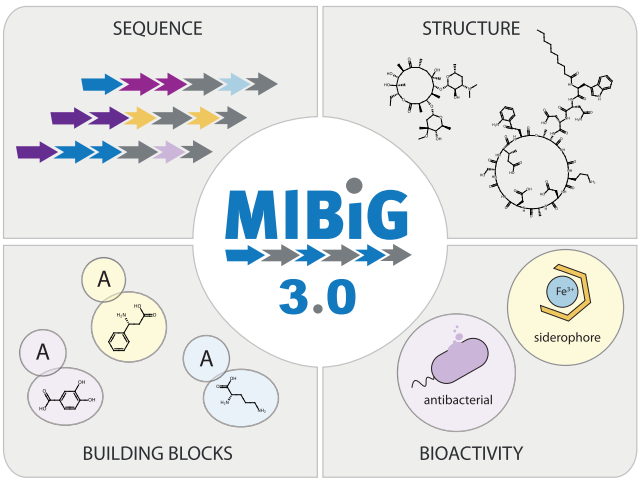
Received September 15, 2022; Revised October 07, 2022; Editorial Decision October 17, 2022; Accepted October 21, 2022

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkac1049/6833236 by guest on 22 November 2022

ABSTRACT

With an ever-increasing amount of (meta)genomic data being deposited in sequence databases, (meta)genome mining for natural product biosynthetic pathways occupies a critical role in the discovery of novel pharmaceutical drugs, crop protection agents and biomaterials. The genes that encode these pathways are often organised into biosynthetic gene clusters (BGCs). In 2015, we defined the Minimum Information about a Biosynthetic Gene cluster (MIBiG): a standardised data format that describes the minimally required information to uniquely characterise a BGC. We simultaneously constructed an accompanying online database of BGCs, which has since been widely used by the community as a reference dataset for BGCs and was expanded to 2021 entries in 2019 (MIBiG 2.0). Here, we describe MIBiG 3.0, a database update comprising large-scale validation and re-annotation of existing entries and 661 new entries. Particular attention was paid to the annotation of compound structures and biological activities, as well as protein domain selectivities. Together, these new features keep the database up-to-date, and will provide new opportunities for the scientific community to use its freely available data, e.g. for the training of new machine learning models to predict sequence-structure-function relationships for diverse natural products. MIBiG 3.0 is accessible online at https://mibig.secondarymetabolites.org/.

GRAPHICAL ABSTRACT



INTRODUCTION

Across all kingdoms of life, organisms produce specialised metabolites: molecules that are produced by bacteria, fungi and plants to gain an advantage over their competitors in challenging environments. Specialised metabolites, also referred to as secondary metabolites or natural products, exhibit a wide variety of biological activities, including many that are useful for pharmaceutical and agricultural applications, e.g. antibiotics, anti-cancer drugs, pesticides and herbicides. The production of specialised metabolites is typically encoded by biosynthetic gene clusters (BGCs): groups of co-localised and co-regulated genes that jointly encode a biosynthetic pathway. Therefore, microbial and plant genomes can be mined for novel specialised metabolite production by detecting BGCs and predicting their encoded products and functions. Similar to how the relationship between DNA, mRNA and protein describes the flow of information in cells, we can define a 'central dogma' of specialised metabolism: a BGC sequence encodes a set of enzymes, which together assemble a compound structure (or a cocktail of structural analogues), which in turn dictates specialised metabolite function. Understanding how information is translated from sequence to structure to function is key to natural product discovery. To address the first stage, sequence information, various tools have been developed that automatically detect BGCs from DNA sequence, including antiSMASH and its siblings fungiSMASH and plantiSMASH (1,2), GECCO (3), DeepBGC (4), RiPP-Miner (5) and PRISM 4 (6).

To facilitate dereplication and comparative analysis of predicted BGCs with known BGCs, and to characterise the interplay between sequence, structure and function, standardised data annotation and storage are essential. To this purpose, we developed the Minimum Information about a Biosynthetic Gene cluster (MIBiG) standard and built a database which contains standardised entries for experimentally validated BGCs of known function (7,8). Each entry minimally contains information about the nucleotide entry and coordinates of the genomic locus involved, the producing organism's taxonomy, biosynthetic class, name of the produced compound(s), and literature reference(s). There are also various optional fields for non-minimal entries, including fields for gene function, product structure and bioactivity, crosslinks to chemical structure databases such as NP Atlas (9) and PubChem (10), and monomer identity. With MIBiG 2.0 containing over 2000 entries, the database has become an important reference for many researchers that mine genomes for natural products. For example, it has been used to estimate the potential for biosynthetic novelty in large-scale microbiome studies (11,12), to identify conserved amino acids playing key roles in catalytic activities across enzyme families (13), to help guide natural product discovery efforts towards high-potential taxa (14), and to train machine-learning algorithms for natural product activity prediction (15).

Here, we present MIBiG 3.0: an update designed to increase the number of non-minimal entries in our database and adding new data entries through a large-scale community annotation effort. We focused on three features: the characterisation and cross-linking of 1188 chemical structures, the annotation of 1002 bioactivities of BGC products, and the validation and annotation of 2020 protein domain substrates of nonribosomal peptide synthetases (NRPSs). In addition, we added 661 novel BGCs to the MIBiG database which were published since the last database update and removed 69 duplicate and low-quality entries (Figure 1). Together, these additions keep the database current, and provide unique opportunities for exploring complex

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkac1049/6833236 by guest on 22 November 2022

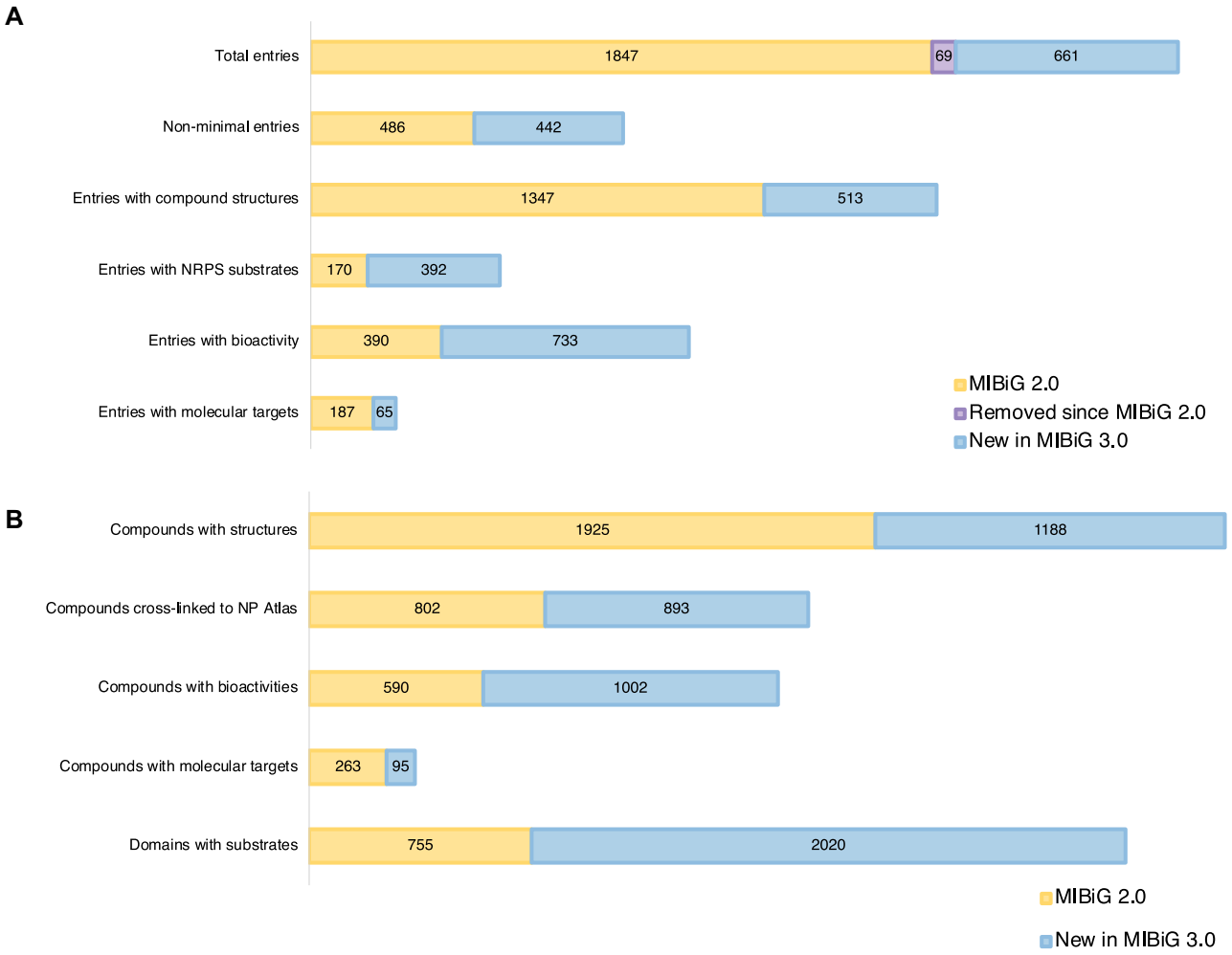


Figure 1. Overview of MIBiG 3.0. (**A**) Added, removed and updated entries since MIBiG 2.0. (**B**) Improvements in the annotation of compounds, bioactivities, molecular targets and NRPS domain substrates.

sequence-structure-function relationships in diverse natural product domains.

METHODS AND IMPLEMENTATION

Manual curation through crowdsourcing and mass online 'annotathons'

As authors themselves typically have the best understanding of the BGC they have studied, we greatly encourage natural product researchers to submit their BGCs to MIBiG during the process of publishing their work. To this purpose, MIBiG supplies an online form through which researchers can request a unique MIBiG identifier and submit their experimentally verified BGCs, pre- or post-publication. Since MIBiG version 2.0, this has yielded 97 manually submitted, high-quality entries which have now been incorporated into MIBiG 3.0. Still, there are far more published BGCs that are not manually submitted to MIBiG.

With an increasing number of papers describing novel BGCs being published every year, manually annotating, validating and adding BGCs to MIBiG has become a mammoth task. Therefore, we took to social media to gauge the community's interest in participating in an online annotation event. We received many positive responses, with 86 people from four different continents volunteering to participate in our MIBiG 'annotathons'. We organised eight three-hour online sessions, accommodating different timezones, with various breakout rooms dedicated to specific annotation tasks: annotating new clusters, annotating and cross-linking compound structures, annotating compound bioactivities, and assigning substrate selectivities to NRPS protein domains. We prepared multiple instruction videos and assigned an expert to each of the breakout rooms who could be directly approached with questions from annotators to ensure that annotation quality was consistent. In addition, one of our annotators at the CINVESTAV research institute mobilised fourteen MSc Integrative Biology students of their 2021 Bacterial Genomics class to annotate compound bioactivities under supervision. Finally, we resolved 125 database issues that were raised by users on our GitHub page, redefining BGC boundaries, correcting biosynthetic classes, adding and removing literature references, fixing compound structures, and removing duplicate entries.

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkac1049/6833236 by guest on 22 November 2022

Annotating and cross-linking compound structures

Since version 2.0, compound structures in MIBiG have been cross-linked to the NP Atlas database: a database containing structures of natural products isolated from bacteria and fungi. During the preparations for version 3.0, we collaborated with the NP Atlas team to (i) add structures for compounds in SMILES format (16), including stereochemical information where possible and (ii) cross-link them to five databases of chemical structures: NP Atlas (9), PubChem, ChemSpider (17), LOTUS (18), and ChEMBL (19). If compound entries were found in multiple databases, SMILES strings from NP Atlas were prioritised. SMILES strings were also collected for existing entries that were already cross-linked to a database but did not report a SMILES string. Correctness of SMILES syntax was validated with PIKAChU (20).

Annotating compound bioactivities

To improve MIBiG as a resource for machine learning models predicting sequence-structure-function relationships, we added bioactivity data for 1002 compounds and chemical target data for 95 compounds. 708 of these annotations were transferred from the dataset assembled by Walker and Clardy, who designed a machine learning model to predict BGC function from sequence (15). To accommodate consistent annotations, we assigned all existing and novel bioactivities to 68 standardised functional categories (Supplementary Table S1).

Annotating NRPS protein domains

To concretise the relationship between NRPS sequence and the structure of its produced nonribosomal peptide (NRP), we annotated and validated the substrate selectivities of 2775 NRPS adenylation (A) domains. A-domains dictate which monomers (predominantly amino acids) are incorporated into (hybrid) NRP scaffolds. Substrate annotation can be performed at different levels: we can define the pre-tailored substrate precursor (e.g. L-aspartic acid); the substrate as recognised by the A-domain (e.g. (3*R*)-3-hydroxy-L-aspartic acid); or the post-tailored integrated monomer that ends up in the final NRP scaffold (e.g. (3*R*)-3-hydroxy-D-aspartic acid). We chose to annotate the substrates as recognised by the A-domain, as this best reflects the biological relationship between A-domain and incorporated monomer. In addition to substrate identity, we also recorded evidence for substrate selectivity in the form of an evidence code and literature references. To this purpose, we added 13 evidence codes to the JSON schema which is used to standardise MIBiG entries (Table 1).

After community annotation, substrate naming was homogenised and each stereochemically ambiguous substrate was manually curated by an expert. Where stereochemistry could be inferred from structure, this is reflected in the substrate name for each stereocenter. Exceptions are amino acid names, which are assumed to be in their L-configuration. To avoid any ambiguity in substrate naming, we also linked each of our 274 unique substrate names to an isomeric SMILES string representing the substrate structure (Figure 2; Supplementary Table S2). SMILES validation and deduplication were handled using PIKAChU (20).

Table 1. Evidence codes for adenylation domain substrate annotations

Evidence code	Accepted as standalone evidence	New in MIBiG 3.0
Activity assay	X	
ACVS assay	X	X
ATP-PPi exchange assay	X	X
Enzyme-coupled assay	X	X
Feeding study	X	
Heterologous expression	X	X
Homology		X
HPLC	X	X
In-vitro experiments	X	X
Knock-out studies	X	X
Mass spectrometry	X	X
NMR	X	X
Radio labelling	X	X
Sequence-based prediction		
Steady-state kinetics	X	X
Structure-based inference	X	
X-ray crystallography	X	X

As indicated, some evidence codes are only accepted as evidence for substrate specificity when combined with a second evidence code that provides further support for a data entry. Thirteen evidence codes were newly introduced in MIBiG 3.0. ACVS assay: δ-(L-*R*-aminoadipyl)-L-cysteinyl-D-valine synthetase assay, specific for measuring penicillin production. HPLC: high-performance liquid chromatography. NMR: nuclear magnetic resonance.

RESULTS AND DISCUSSION

Taking the 'minimal' out of MIBiG

While MIBiG 2.0 serves an important role in the community as a reference database to quickly identify whether a BGC is similar to any known BGCs, its utility as a resource for exploring sequence-structure-function relationships could be improved. This can mainly be explained by the high number of minimal entries in the database: entries that only contain sequence and compound information that could be augmented by adding further standardised annotations. For MIBiG 3.0, we aimed to promote as many existing and novel entries as possible to non-minimal entries by annotating compound structures (1188), bioactivities (1002) and NRPS substrates (2020). In total, we added 661 novel BGCs and 4871 separate data entries to our database, increasing our number of non-minimal entries from 486 to 928 (Figure 1, Supplementary Figure S1). MIBiG 3.0 now contains 2502 entries, spanning 16 phyla across 5 kingdoms of life (Table 2).

Streamlining research into the central dogma of specialised metabolism

With 905 NRPS and modular Type I PKS BGCs in MIBiG 3.0, modular BGCs constitute a substantial part of our database. Modular systems are characterised by enzyme complexes comprising repeating domain architectures, which collectively assemble a natural product scaffold. When the substrate selectivities of the recognition do-

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkac1049/6833236 by guest on 22 November 2022

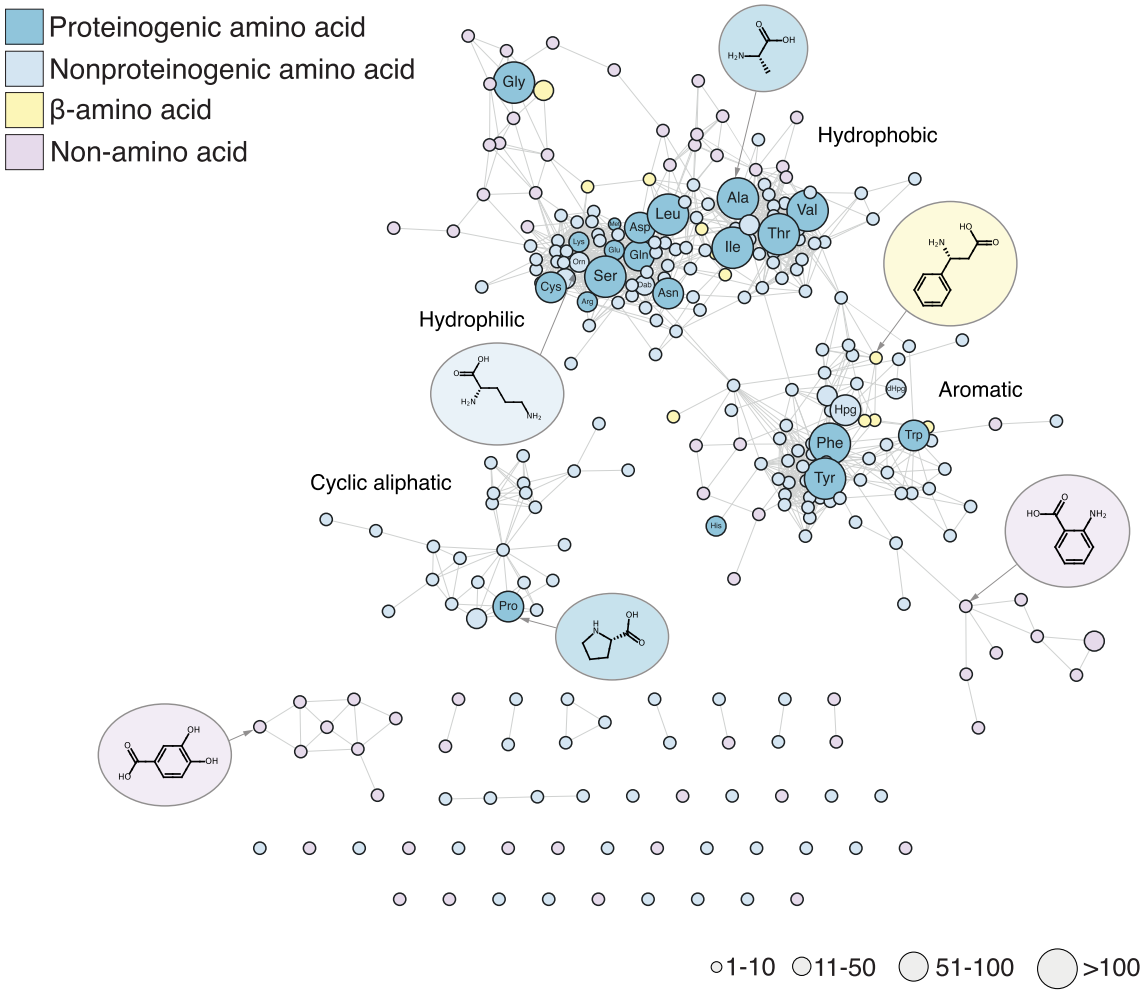


Figure 2. Similarity network of annotated NRPS substrates. Each node represents one of 274 unique NRPS substrate structures in MIBiG 3.0. Colours indicate substrate categories, and node size correlates with the number of annotations for that substrate in the MIBiG database. Substrates were clustered based on Tanimoto similarity of ECFP-4 molecular fingerprints (25) (edge cut-off = 0.46).

Table 2. Entries in MIBiG 3.0 by phylum

Kingdom	Phylum	Number of BGCs in MIBiG 3.0
Bacteria	Actinobacteria	1042
	Proteobacteria	527
	Firmicutes	229
	Cyanobacteria	139
	Bacteroidetes	17
	Candidatus tectomicrobia	6
	Chloroflexi	4
	Verrucomicrobia	3
	Planctomycetes	2
	Kiritimatiellaeota	1
	Unknown	41
Fungi	Ascomycota	415
	Basidiomycota	23
	Unknown	3
Plantae	Streptophyta	43
	Rhodophyta	2
Archaea	Euryarchaeota	3
Chromista	Bacillariophyta	1
	Dinophyceae	1

mains are known (acyltransferase (AT) domains for PKS and A-domains for NRPS), these consistent architectures make it possible to predict the structure of chemical scaffolds with reasonable accuracy. Most AT domains in PKS systems recognise one of two substrates, malonyl-CoA or methylmalonyl-CoA, and excellent bioinformatics tools exist to distinguish between the two (21). However, for A-domains in NRPS systems, which recognise over 500 known substrates (22), substrate prediction is a greater challenge, which will require substantially more data to obtain models of comparably predictive power. Therefore, we decided to make the annotation of the substrate selectivity of NRPS A-domains a major focus of MIBiG 3.0. MIBiG 3.0 now contains annotations for 2775 A-domains (compared to 755 annotations in MIBiG 2.0; Figure 1B), covering 274 unique substrates which are identified by stereochemically curated isomeric SMILES strings (Figure 2; Supplementary Table S2). This makes MIBiG the largest resource for A-domain substrate data, containing 3–4 times as many labelled data points as the training sets used for the A-domain selectivity predictors SANDPUMA (23) and NRPSPredictor2 (24).

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkac1049/6833236 by guest on 22 November 2022

We hope that eventually this dataset will be leveraged to train an improved A-domain substrate predictor, which can in turn be integrated into tools like antiSMASH to improve NRP scaffold structure prediction.

Since version 2.0, we have added structural identifiers of 1188 compounds to our database in SMILES format (16), increasing the number of BGCs with structural data from 1347 to 1860 (Figure 1). By pulling SMILES strings directly from cross-linked databases where possible, we avoid conflicts caused by versioning and SMILES formatting. Additionally, we linked 1002 additional compounds to 51 unique bioactivities, creating opportunities for computationally predicting compound bioactivity from structure. For a further 95 compounds, we were also able to annotate their molecular targets (Figure 1B).

By centering MIBiG 3.0 around the annotation of substrate building blocks, compound structures, and bioactivities, we aspired to streamline future research into all aspects of sequence-structure-function relationships that lie at the heart of natural product research. All data can be easily downloaded and parsed in bulk from our database in JSON and GenBank format or accessed on an entry-by-entry basis through our searchable online repository. As such, we hope that MIBiG 3.0 will prove an important resource for future machine learning endeavours that aim to decode the central dogma of specialised metabolism.

DATA AVAILABILITY

The MIBiG Repository is available at https://mibig.secondarymetabolites.org/. There is no access restriction for academic or commercial use of the repository and its data. The source code components, JSON-formatted data standard, and SQL schema for the MIBiG Repository are available on GitHub (https://github.com/mibig-secmet) under an OSI-approved Open Source licence.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank Simon Shaw and Martin Larralde for validating and fixing numerous existing entries; Andrés G., Andrés L., Antonio, Cristina, Daniel, Luis, Ivón, Diana, Erika, Gabriel, Isamar, Janeth, Rafa and Vanessa from the MSc 2021 Bacterial Genomics class of Integrative Biology at the CINVESTAV research institute for annotating bioactivity information; Caroline Rodenbach, Lhaís Caldas and Yañez-Olvera for contributing to our annotathons; Allison Walker for providing a published dataset of bioactivities which was integrated into MIBiG 3.0.

FUNDING

ERC Starting Grant [948770-DECIPHER to M.H.M.]; Novo Nordisk Foundation [NNF20CC0035580, NNF16OC0021746 to T.W]; Danish National Research Foundation [DNRF137 to T.W]; National Center for Complementary and Integrative Health (NCCIH) of the National Institutes of Health [U24AT010811 to R.L. and F32AT011475 to N.E.A]; Natural Sciences and Engineering Council of Canada Discovery grant [to R.L.]; Netherlands Organization for Scientific Research (NWO) Veni Science Grant [VI.Veni.202.130 to M.A]; European Union Horizon 2020 projects CARTNET [765147], SECRETed [101000794] and MARBLES [101000392]; Horizon 2020 Marie Skłodowska-Curie Actions [893122 to K.H.]; Horizon 2020 Marie Sklodowska-Curie Individual Fellowship [MSCA-IF-EF-ST-897121 to M.A.S.]; U.S. Department of Energy [DE-AC02-05CH11231]; University of Strathclyde PhD Research Excellence Award [to D.S.]; Consejo Nacional de Ciencia y Tecnología (CONACyT) [757173 to L.R.R.-B.]; Portuguese Science and Technology Foundation (FCT) fellowship [SFRH/BD/140567/2018 to A.R.]; U.S. National Science Foundation [CBET-2032243 to A.M.K]; National Research Foundation of Korea [NRF-2022R1C1C2004118 and NRF-2020R1C1C1004046]; National Institutes of Health [GM134688 to E.K. and 1R01AI155694 to J.M.W.]; Netherlands eScience Center (NLeSC) Accelerating Scientific Discoveries Grant [ASDI.2017.030 to J.J.J.v.d.H.]; Deutsche Forschungsgemeinschaft [398967434-TRR 261]; UKRI Biotechnology and Biological Sciences Research Council [BBSRC; BB/R022054/1 and BB/W013959/1]; UK government Department for Environment, Food and Rural Affairs [project DEEPEND: deep ocean resources and biodiscovery]; Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro [E-26/211.314/2019]; Fundaçao para a Ciencia e Tecnologia (FCT) fellowship [SFRH/BD/136367/2018 to R.C.B.]; German Chemical Industry scholarship [to F.B.]; Cooperative Research Centres Projects scheme [CRCPFIVE000119 to T.J.B.]; Consejo Nacional de Ciencia y Tecnología (CONACyT) [735867 to J.B-A.]; Natural Sciences and Engineering Council of Canada PGSD fellowship [to L.Z.]; Natural Sciences and Engineering Council of Canada PGSD fellowship [to M.R.]; Odo van Vloten foundation [to J.N.-M.]; LOEWE Center for Translational Biodiversity Genomics (LOEWE TBG), Funds of the Chemical Industry Germany; Rothamsted Science Initiatives Catalyst Award scheme grant 'Microbial natural product discovery pipeline for next generation fungicides'. Funding for open access charge: European Research Council.

Conflict of interest statement. J.J.J.vdH. is a member of the Scientific Advisory Board of NAICONS Srl., Milano, Italy. A.M.K. is a co-founder of Nitro Biosciences, Inc. M.H.M. is on the scientific advisory board of Hexagon Bio and co-founder of Design Pharmaceuticals.

REFERENCES

1. Blin,K., Shaw,S., Kloosterman,A.M., Charlop-Powers,Z., Van Wezel,G.P., Medema,M.H. and Weber,T. (2021) AntiSMASH 6.0: improving cluster detection and comparison capabilities. *Nucleic Acids Res.*, **49**, W29–W35.
2. Kautsar,S.A., Suarez Duran,H.G., Blin,K., Osbourn,A. and Medema,M.H. (2017) PlantiSMASH: automated identification, annotation and expression analysis of plant biosynthetic gene clusters. *Nucleic Acids Res.*, **45**, W55–W63.
3. Carroll,L.M., Larralde,M., Fleck,J.S., Ponnudurai,R., Milanese,A., Cappio,E. and Zeller,G. (2021) Accurate de novo identification of biosynthetic gene clusters with GECCO. bioRxiv doi:

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkac1049/6833236 by guest on 22 November 2022

https://doi.org/10.1101/2021.05.03.442509, 04 May 2021, preprint: not peer reviewed.
4. Hannigan,G.D., Prihoda,D., Palicka,A., Soukup,J., Klempir,O., Rampula,L., Durcak,J., Wurst,M., Kotowski,J., Chang,D. *et al.* (2019) A deep learning genome-mining strategy for biosynthetic gene cluster prediction. *Nucleic Acids Res.*, **47**, E110.
5. Agrawal,P., Khater,S., Gupta,M., Sain,N. and Mohanty,D. (2017) RiPPMiner: a bioinformatics resource for deciphering chemical structures of ripps based on prediction of cleavage and cross-links. *Nucleic Acids Res.*, **45**, W80–W88.
6. Skinnider,M.A., Johnston,C.W., Gunabalasingam,M., Merwin,N.J., Kieliszek,A.M., MacLellan,R.J., Li,H., Ranieri,M.R.M., Webster,A.L.H., Cao,M.P.T. *et al.* (2020) Comprehensive prediction of secondary metabolite structure and biological activity from microbial genome sequences. *Nat. Commun.*, **11**, 6058.
7. Kautsar,S.A., Blin,K., Shaw,S., Navarro-Muñoz,J.C., Terlouw,B.R., Van Der Hooft,J.J.J., Van Santen,J.A., Tracanna,V., Suarez Duran,H.G., Pascal Andreu,V. *et al.* (2020) MIBiG 2.0: a repository for biosynthetic gene clusters of known function. *Nucleic Acids Res.*, **48**, D454–D458.
8. Medema,M.H., Kottmann,R., Yilmaz,P., Cummings,M., Biggins,J.B., Blin,K., De Bruijn,I., Chooi,Y.H., Claesen,J., Coates,R.C. *et al.* (2015) Minimum information about a biosynthetic gene cluster. *Nat. Chem. Biol.*, **11**, 625–631.
9. Van Santen,J.A., Poynton,E.F., Iskakova,D., Mcmann,E., Alsup,T.A., Clark,T.N., Fergusson,C.H., Fewer,D.P., Hughes,A.H., Mccadden,C.A. *et al.* (2022) The natural products atlas 2.0: a database of microbially-derived natural products. *Nucleic Acids Res.*, **50**, D1317–D1323.
10. Kim,S., Chen,J., Cheng,T., Gindulyte,A., He,J., He,S., Li,Q., Shoemaker,B.A., Thiessen,P.A., Yu,B. *et al.* (2021) PubChem in 2021: new data content and improved web interfaces. *Nucleic Acids Res.*, **49**, D1388–D1395.
11. Paoli,L., Ruscheweyh,H.J., Forneris,C.C., Hubrich,F., Kautsar,S., Bhushan,A., Lotti,A., Clayssen,Q., Salazar,G., Milanese,A. *et al.* (2022) Biosynthetic potential of the global ocean microbiome. *Nature*, **607**, 111–118.
12. Nayfach,S., Roux,S., Seshadri,R., Udwary,D., Varghese,N., Schulz,F., Wu,D., Paez-Espino,D., Chen,I.M., Huntemann,M. *et al.* (2021) A genomic catalog of earth's microbiomes. *Nat. Biotechnol.*, **39**, 499–509.
13. Izoré,T., Candace Ho,Y.T., Kaczmarski,J.A., Gavriilidou,A., Chow,K.H., Steer,D.L., Goode,R.J.A., Schittenhelm,R.B., Tailhades,J., Tosin,M. *et al.* (2021) Structures of a non-ribosomal peptide synthetase condensation domain suggest the basis of substrate selectivity. *Nat. Commun.*, **12**, 2511.
14. Gavriilidou,A., Kautsar,S.A., Zaburannyi,N., Krug,D., Müller,R., Medema,M.H. and Ziemert,N. (2022) Compendium of specialized metabolite biosynthetic diversity encoded in bacterial genomes. *Nat. Microbiol.*, **7**, 726–735.
15. Walker,A.S. and Clardy,J. (2021) A machine learning bioinformatics method to predict biological activity from biosynthetic gene clusters. *J. Chem. Inf. Model.*, **61**, 2560–2571.
16. Weininger,D. (1988) SMILES, a chemical language and information system. *J. Chem. Inf. Model.*, **28**, 31–36.
17. Kelly,R. and Kidd,R. (2015) Editorial: chemspider-a tool for natural products research. *Nat. Prod. Rep.*, **32**, 1163–1164.
18. Rutz,A., Sorokina,M., Galgonek,J., Mietchen,D., Willighagen,E., Gaudry,A., Graham,J.G., Stephan,R., Page,R., Vondrasek,J. *et al.* (2021) The LOTUS initiative for open natural products research. *Elife*, **11**, e70780.
19. Gaulton,A., Hersey,A., Nowotka,M.L., Patricia Bento,A., Chambers,J., Mendez,D., Mutowo,P., Atkinson,F., Bellis,L.J., Cibrian-Uhalte,E. *et al.* (2017) The ChEMBL database in 2017. *Nucleic Acids Res.*, **45**, D945–D954.
20. Terlouw,B.R., Vromans,S.P.J.M. and Medema,M.H. (2022) PIKAChU: a Python-based informatics kit for analysing chemical units. *J. Cheminform.*, **14**, 34.
21. Minowa,Y., Araki,M. and Kanehisa,M. (2007) Comprehensive analysis of distinctive polyketide and nonribosomal peptide structural motifs encoded in microbial genomes. *J. Mol. Biol.*, **368**, 1500–1517.
22. Miller,B.R. and M.,G.A. (2016) Structural biology of non-ribosomal peptide synthetases. *Methods Mol. Biol.*, **1401**, 3–29.
23. Chevrette,M.G., Aicheler,F., Kohlbacher,O., Currie,C.R. and Medema,M.H. (2017) SANDPUMA: ensemble predictions of nonribosomal peptide chemistry reveal biosynthetic diversity across actinobacteria. *Bioinformatics*, **33**, 3202–3210.
24. Röttig,M., Medema,M.H., Blin,K., Weber,T., Rausch,C. and Kohlbacher,O. (2011) NRPSpredictor2 - a web server for predicting NRPS adenylation domain specificity. *Nucleic Acids Res.*, **39**, 362–367.
25. Rogers,D. and Hahn,M. (2010) Extended-connectivity fingerprints. *J. Chem. Inf. Model.*, **50**, 742–754.

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkac1049/6833236 by guest on 22 November 2022