



Deliverable D3.3

Title: First report on phylogenetic tree database for siderophore and biosurfactant biosynthetic gene clusters and enzymes and their reactions

Lead Beneficiary

IDENER R&D

Delivery Date

31st May 2022



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101000794. This publication reflects only the author's views and the European Union is not liable for any use that may be made of the information contained therein.

Deliverable No.	D3.3
Dissemination Level	Public
Work Package	3
Task	3.1 and 3.2
Lead Beneficiary	IDENER R&D
Contributing beneficiary(ies)	EKUT
Reviewers	IDENER team
Due date of deliverable	31 st May 2022
Actual submission date	31 st May 2022
Main Author/Contributors	IDENER Team

Project Acronym	SECRETed
Project Title	Sustainable Exploitation of bio-based Compounds Revealed and Engineered from naTural sources
Activity	RIA Research and Innovation action
Call	H2020-FNR-2020-2
Funding Scheme	Grant Agreement No: 101000794

Document History				
Version	Date	Beneficiary	Author/Reviwer	Notes
1.0	05-20-2022	IDE R&D, EKUT	IDENER team	First draft
2.0	05-30-2022	IDE R&D, EKUT	IDENER team	Corrections applied.

All the contributors to this deliverable declare that they:

- Are aware that plagiarism and/or literal utilization (copy) of materials and texts from other Projects, works, and deliverables must be avoided and may be subject to disciplinary actions against the related partners and/or the Project consortium by the EU.
- Confirm that all their individual contributions to this deliverable are genuine and their own work or the work of their teams working in the Project, except where is explicitly indicated otherwise.

Have followed the required conventions in referencing the thoughts, ideas, and texts made outside the Project.

Executive Summary

The SECRETed project will fully exploit the potential of Systems and Synthetic Biology toolboxes and their application within aquatic biotechnology to develop novel hybrid compounds for the agrochemical, pharmaceutical, cosmetic, and chemistry sectors. Biosynthetic pathways of marine and extremophilic microorganisms will be reverse engineered to infer the individual roles of their constituent genes, which will be further combined for the production of non-natural biosurfactants and siderophores with tailor-made properties.

Biosurfactants are compounds with a surface-active nature tendency to adsorb at interfaces, while siderophores have the ability to chelate and transport Fe^{3+} ions. The amphiphilic nature of biosurfactants and marine siderophores provides an exciting opportunity to develop methods of biosynthesis that would enable the exchange of their hydrophobic and hydrophilic parts, among other structural changes. The development of hybrid molecules would allow the exploration of new-to-nature compounds endowed with the combination of their respective properties, to address new applications.

Machine Learning algorithms, an inspection of databases, and new experimental and computational-based data will be employed to build a unique microbial amphiphilic compound chemical space to identify the desired genetic mechanisms. Detected genes will be reverse engineered to standardise and modularize associated metabolic elements, with the purpose of broadening their benefits by searching for Industrial-driven formulations based on suitable microbial hosts. The Design-Build-Test-Learn methodological steps will be used to produce new microbial strains that support the selected genetic elements and satisfy sustainable industrial processes.

Deliverable D3.3 is a public report produced in the context of WP3: “Databases integration and Industry-driven designs”. In Task 3.1: “Machine learning approaches to survey databases for siderophore and biosurfactants information”, ML algorithms have been deployed and applied to survey databases and scientific publications. Specifically, subtask 3.1.2: “Amphiphilic siderophores and biosurfactant biosynthetic gene cluster analysis” is focused on collecting information related to biosurfactant and siderophore biosynthetic gene clusters contained in publicly available databases together with available information in scientific literature.

Together with subtask 3.1.2: “Molecular Families inspection”, the aim of Task 3.1 is to construct a unique microbial amphiphilic compound space comprehending molecular structures, physicochemical characteristics, associated bioactivities, and revealed genetic mechanisms responsible for their biosynthesis.

The present report summarises obtained Biosynthetic gene clusters, informs about the major challenges and approaches involved in the generation of SECRETed biosynthetic space, and describes Gene Cluster Families classification efforts and their involvement in structure-property-driven retrosynthesis algorithm.

Table of contents

Executive Summary	3
Table of contents	4
List of Figures	5
List of Tables	5
1. Introduction	6
2. SECRETed Biosynthetic space construction: Defining the Bottom-up and Top-down approaches.	7
3. SECRETed Biosynthetic space data obtention and curation	9
3.1. Bottom-up BGCs and Biosynthetic pathways data obtention and curation.	9
3.2. Top-down Microbial producers data obtention and curation.	10
4. SECRETed Biosynthetic space data expansion.	17
4.1. Bottom-up database expansion.	17
4.2. Top-down database expansion.	18
4.3. Bottom-up and Top-down approaches conciliation.	18
5. Conclusions	21

List of Figures

Figure 1 Flow chart covering the two major workflows carried out in SECRETed Biosynthetic space construction.....	7
Figure 2 Venn diagram representing the BGCs found for each group of compounds established by the partners.....	9
Figure 3. Representative examples of the structure of the database in both visual (A) and tabular (B) formats.	10
Figure 4. Euler diagram showing the SECRETed NPs database consolidation efforts related to natural producers (detailed information in Table 1).....	10
Figure 5. Taxonomy distribution among explored databases.	12
Figure 6. Genera distribution in SECRETed defined microbial producers.....	13
Figure 6. Microbial producer distribution in SECRETed biochemical space. Connected nodes are molecular structures grouped in MFs (see deliverable D3.2).	14
Figure 8. A) Number of taxa with available genome links in NCBI. B) Genome links available when species without link are “downgraded” to their previous taxonomic categories (Genus) . C) Genome links available when strains without link are “downgraded” to their previous taxonomic categories (species or genus).....	15
Figure 9. Number of microbial producers found in each range of genome downloading links per taxa category (Genus, specie and strain).....	16
Figure 10. Example of refined and curated BGC siderophore network	17
Figure 10. Representations taken from the BiG-SliCE and Clinker outputs of the clustering tests performed.	20

List of Tables

Table 1: SECRETed natural origin data from NPs consolidation efforts	11
----------------------------------------------------------------------------	----

1. Introduction

Genome-directed discovery of natural products is pursued through the identification of biosynthetic gene clusters, displaying homology with genes involved in the production of known secondary metabolites. Indeed, although secondary metabolites cover a wide and heterogeneous chemical space, the biosynthetic routes for several classes of compounds, such as non-ribosomal peptides and polyketides, are outstandingly conserved across microbial species: at the molecular level, this translates into high sequence similarity of many core biosynthetic enzymes¹.

On the other hand, the process of mining genetically encoded small molecules is not keeping pace with the rate by which genome sequences are being obtained. In general, refined and curated genome mining is still done one gene cluster at a time and requires many person-years of effort to annotate a single molecule.

Nowadays, to address this issue, bioinformatic efforts are focused on comparing architectural relationships between BGCs in sequence similarity networks and grouping them into gene cluster families (GCFs), each of which contains BGCs across a range of organisms that should be linked to a highly similar natural product chemotype^{2,3}. Such GCFs can be matched to molecular families (MFs) identified from mass spectrometry (MS) data based on observed/predicted chemical features⁴.

For that purpose, more than 1700 molecular structures of siderophores and biosurfactants were collected by cross-referencing other NP-related databases containing their unique data coverage, to be further clustered into MFs (see Deliverable D3.2).

This report collects the efforts in defining SECRETed biosynthesis space by collecting refined and curated BGCs in charge of biosurfactant and siderophore microbial synthesis and expands said information to infer accessory genes in charge of the molecular diversity of SECRETed biochemical space.

¹ Ziemert et al., *Nat. Prod. Rep.* 33, 988–1005. (2016)

² Cimermancic et al., *Cell*. 158(2):412–421.(2014)

³ Doroghazi et al., *Nat Chem Biol*.10(11):963–968.(2014)

⁴ Nguyen et al., *Proc. Natl. Acad. Sci. U.S.A.* 110:E2611–E2620.(2013)

2. SECRETed Biosynthetic space construction: Defining the Bottom-up and Top-down approaches.

Genome mining is a process in which small molecules are discovered by predicting what compound will be genetically encoded based on the sequences of BGCs. The biosynthetic genes encoding the pathways to Natural Products (NPs) are usually co-located on the chromosome in biosynthetic gene clusters (BGCs). This paradigm facilitates their identification and characterization, and advances in next-generation DNA sequencing have created a massive data resource to mine for BGCs encoding novel NP structures and enzymology, a process greatly aided by a wealth of evolving bioinformatics tools. Still, it is necessary to identify well-characterize BGCs and associated NP pathways to identify trustable information or “anchoring points” to compare and infer the function of accessory genes in charge of molecular diversity.

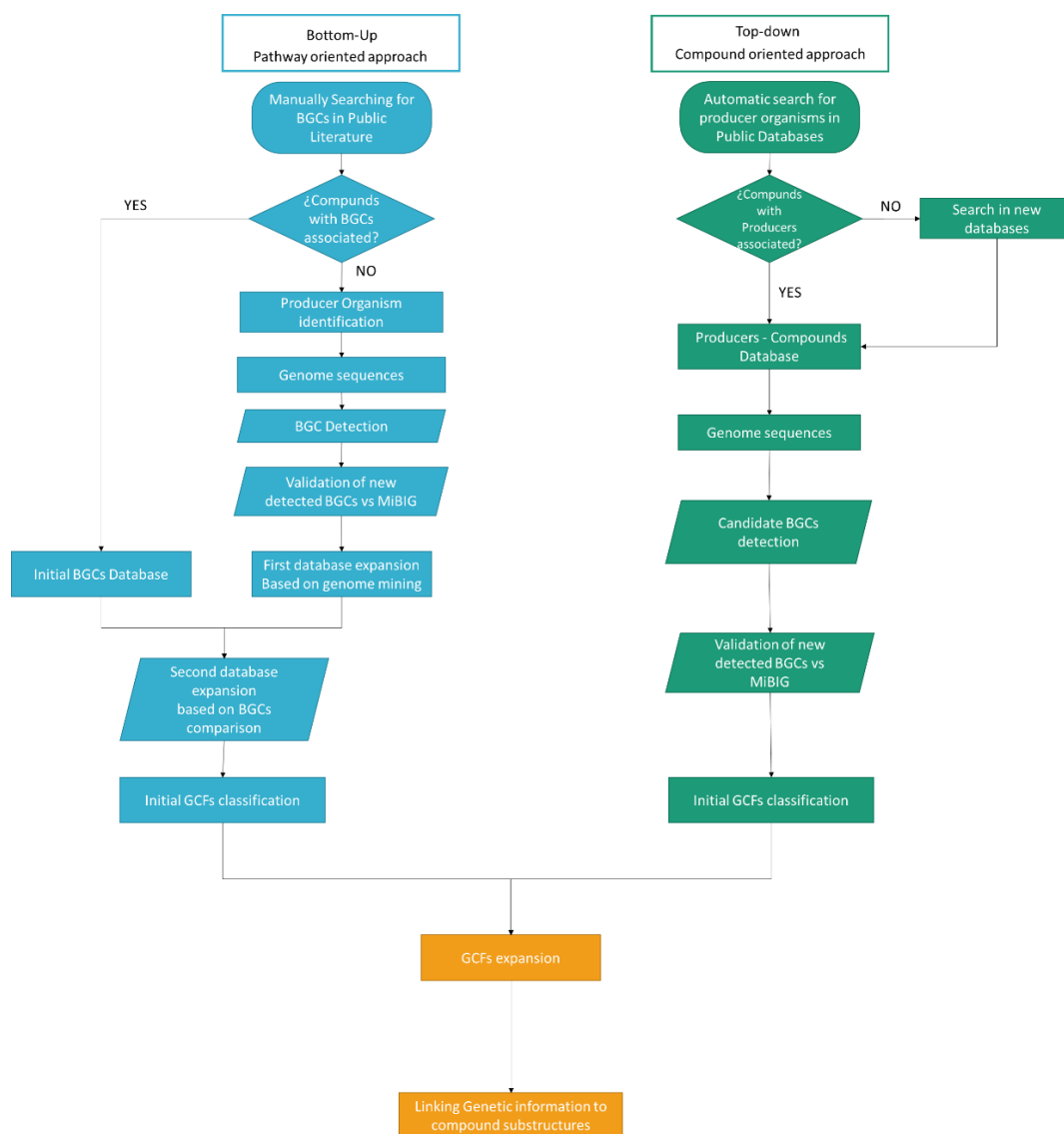


Figure 1 Flow chart covering the two major workflows carried out in SECRETed Biosynthetic space construction.

In this way, two complementary approaches have been used (Figure 1): i) a bottom-up approach, where efforts are centred on collecting refined and curated BGCs biosynthetic pathway information; and ii) a top-down approach, which is focused on providing genomic sequences from microbial producers designated in NPs databases. Both approaches had the common objectives:

- Collecting any single registered BGCs which is associated with SECRETed compounds.
- Clustering characterised known BGCs into Gene Cluster Families (GCFs).
- Inferring unknown BGCs by homology, synteny or domain composition.
- Elucidating the biosynthetic information that connects individual genes to SECRETed compounds.
- Guide retrosynthesis algorithms by constraining the combinatorial retrosynthetic space to defined reaction rules.

The following sections explain SECRETed efforts to achieve the first three objectives.

3. SECRETed Biosynthetic space data obtention and curation.

3.1. Bottom-up BGCs and Biosynthetic pathways data obtention and curation.

As explained in the previous section, the bottom-up approach is based on building on previous efforts to characterise BGCs. Thus, refined and curated BGCs information, where biosynthetic pathways and the function of specific enzymes found in gene clusters are specified, demanded an extensive literature search.

Compounds and Biosynthetic Gene Clusters

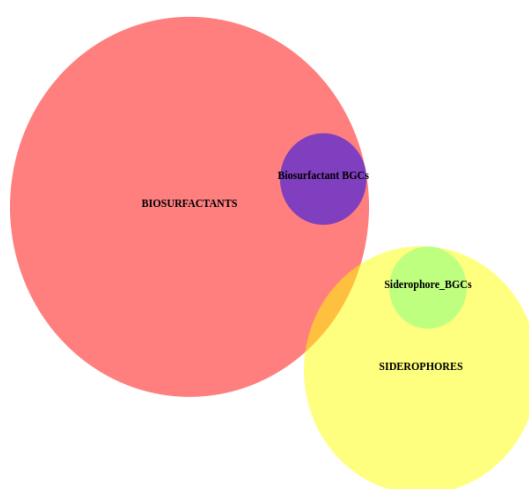


Figure 2 Venn diagram representing the BGCs found for each group of compounds established by the partners.

Three main steps were followed: identification of articles pertaining to microbial siderophores and biosurfactant discovery, extraction of structures, gene clusters and other data from each article, and organisation of these data into a structured format.

Molecular information collected in deliverable D3.2 (~560 Siderophores + ~1500 Biosurfactants) was utilised to inspect scientific literature. This manual search identified the Biosynthetic Gene Clusters for ~74 Siderophores and ~95 Biosurfactants (Figure 2).

Based on the information obtained on the 169 identified BGCs. A database was created containing all the available information on these Gene Clusters (Figure 3), such as links to GenBank files, MiBIG repository⁵, NPAtlas⁶, and PubChem data⁷. This database also contains visualisations (Figure 3, A) of the BGC structure, the class of compounds it encodes and the biosynthetic pathways in which it is involved. It is worth highlighting that this data is the

⁵ Kautsar et al., *Nucleic Acids Research*, 48(D1), D454–D458.(2020).

⁶ van Santen et al., *ACS Cent Sci*. 5(11):1824-1833. (2019)

⁷ Kim et al., *Nucleic Acids Res*. 44(Database issue): D1202–D1213.(2016)

“anchoring point” where further BGC information is inferred, as it was created after a thorough literature search using only actual information based on previous research.

Databases Structure

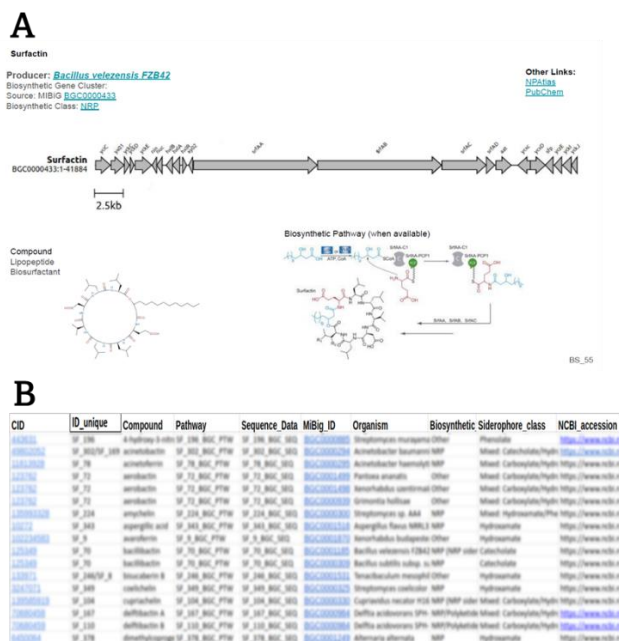


Figure 3. Representative examples of the structure of the database in both visual (A) and tabular (B) formats.

3.2. Top-down Microbial producers data obtention and curation.

As explained in deliverable D3.3, reliance on a disparate set of non-standardized, insular, and specialized databases presents a series of challenges for data access, both within the discipline and for integration and interoperability between related fields. This fact is emphasized when analyzing the individual contribution of each explored database when describing the source of NPs (figure 4, table 1).

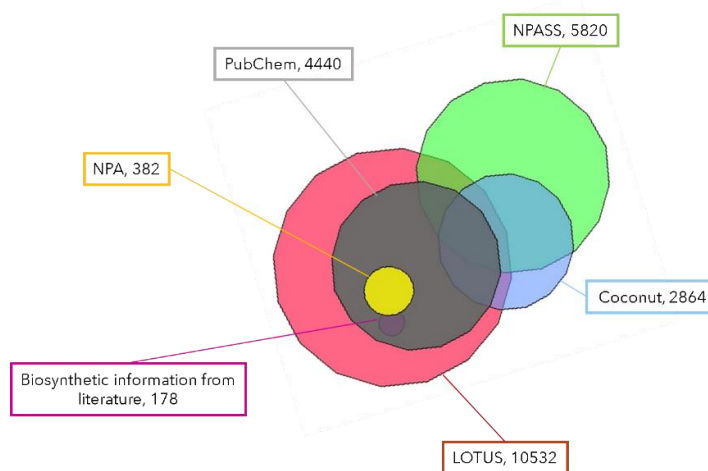


Figure 4. Euler diagram showing the SECRETed NPs database consolidation efforts related to natural producers (detailed information in Table 1).

Table 1: SECRETed natural origin data from NPs consolidation efforts

DATABASE	Organism Counting
NPASS	3185
LOTUS PubChem	3151
LOTUS	3517
Bibliography Coconut LOTUS NPA NPASS	1
LOTUS NPA PubChem	82
Coconut	181
Bibliography	41
Bibliography LOTUS NPA PubChem	18
LOTUS NPASS	171
LOTUS NPA	140
NPA	22
Coconut LOTUS NPA NPASS PubChem	6
Coconut LOTUS NPA	39
LOTUS NPA NPASS	7
Coconut NPA PubChem	2
LOTUS NPA NPASS PubChem	11
LOTUS NPASS PubChem	320
Bibliography LOTUS	23
Coconut LOTUS NPA PubChem	30
Coconut LOTUS	175
Coconut LOTUS NPASS	308
NPA PubChem	4
Bibliography LOTUS PubChem	2
Coconut LOTUS NPASS PubChem	485
Bibliography Coconut LOTUS NPA NPASS PubChem	4
Pubchem	10
Coconut NPA	3
Coconut LOTUS NPA NPASS	3
Bibliography Coconut NPA	1
Bibliography LOTUS NPA NPASS PubChem	3
Bibliography LOTUS NPA	3
Bibliography LOTUS NPASS	1
NPA NPASS	1
Bibliography LOTUS NPASS PubChem	1
Bibliography Coconut LOTUS NPA PubChem	1
Bibliography Coconut LOTUS NPA	1
Bibliography Coconut LOTUS	1
Coconut NPASS	1313
Coconut LOTUS PubChem	310

Data shown relates to the exploration of natural producers associated to SECRETed biosurfactants and siderophores in:

- LOTUS initiative⁸, which has now completed the first steps toward the harmonization, curation, validation and open dissemination of more than 750,000 referenced structure-organism pairs.

⁸ Rutz et al., Elife.11:e70780. (2022).

- COLLEction of Open Natural prodUCts (COCONUT⁹), a data set assembled from 50 open-access databases containing 412,903 compounds and being the largest collection of NP available to this date.
- Natural Product Activity and Species Source (NPASS¹⁰), which complement other databases by providing the experimental activity values and species sources of 35 032 NPs from 25 041 species targeting 5863 targets (2946 proteins, 1352 microbial species and 1227 cell-lines).
- The Natural Products Atlas (NPA¹¹), an Open Access Knowledge Base for Microbial Natural Products Discovery.
- Pubchem¹², a public repository for information on chemical substances and their biological activities, launched in 2004 as a component of the Molecular Libraries Roadmap Initiatives of the US National Institutes of Health (NIH).
- Bibliography, which stands for compiled information in section 3.1, and refers to those microbial producers which not only have BGCs available, but also their biosynthetic pathways are defined in literature.

An important aspect to highlight in this meta-analysis and database conciliation approach is that, majoritarily, most of natural origin information from SECRETed compounds was related to the eukarya superkingdom (figure 5).

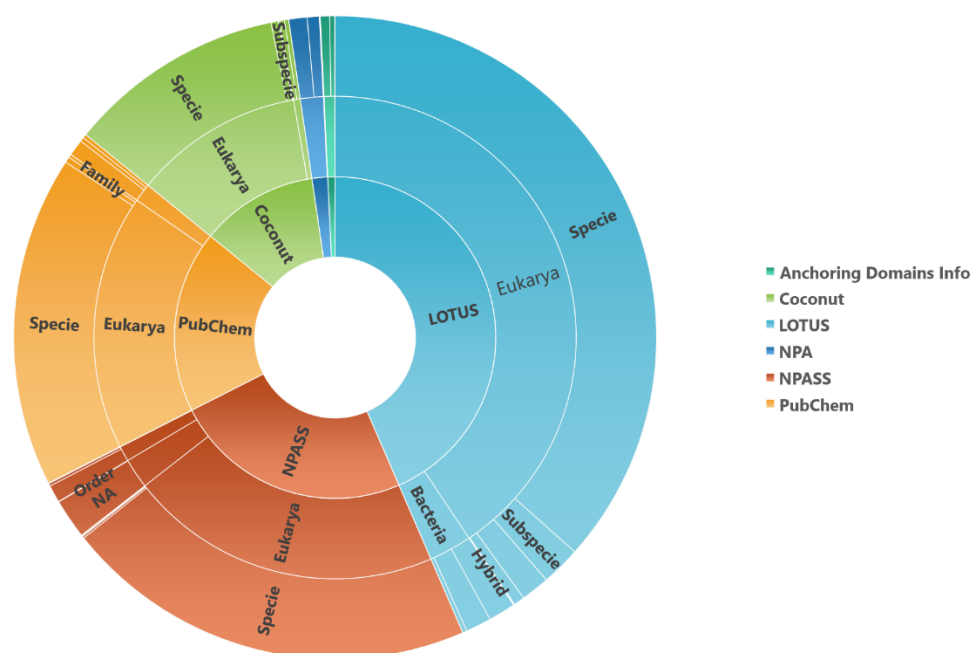


Figure 5. Taxonomy distribution among explored databases.

⁹ Sorokina et al., J Cheminform 10;13(1):2 (2021).

¹⁰ Zeng et al., Nucleic Acids Res 46(D1):D1217-D1222 (2018).

¹¹ van Santen et al., ACS Cent Sci. 5(11):1824-1833. (2019)

¹² Kim et al., Nucleic Acids Res. 44(Database issue): D1202–D1213.(2016)

Data distributed was not homogeneous as, for example, during this analysis, NPASS database contained more than 480 undetermined organisms related to siderophores and biosurfactants registered in SECRETed. Thus, after conciliating and consolidating obtained information, 13576 unique organisms were collected. Of those, 12177 unique eukarya taxa, 914 unique bacterial taxa and 5 unique archaea taxa.

Not surprisingly, the most abundant genera in biosurfactants and siderophores microbial producers (914 bacterial producers) were Streptomyces, Pseudomonas and Bacillus (Figure 6).

Remarkably, said unique 914 bacterial taxa were associated to more than 1100 compounds of 1900 compounds with known molecular structure in SECRETed biochemical space (Figure 7).

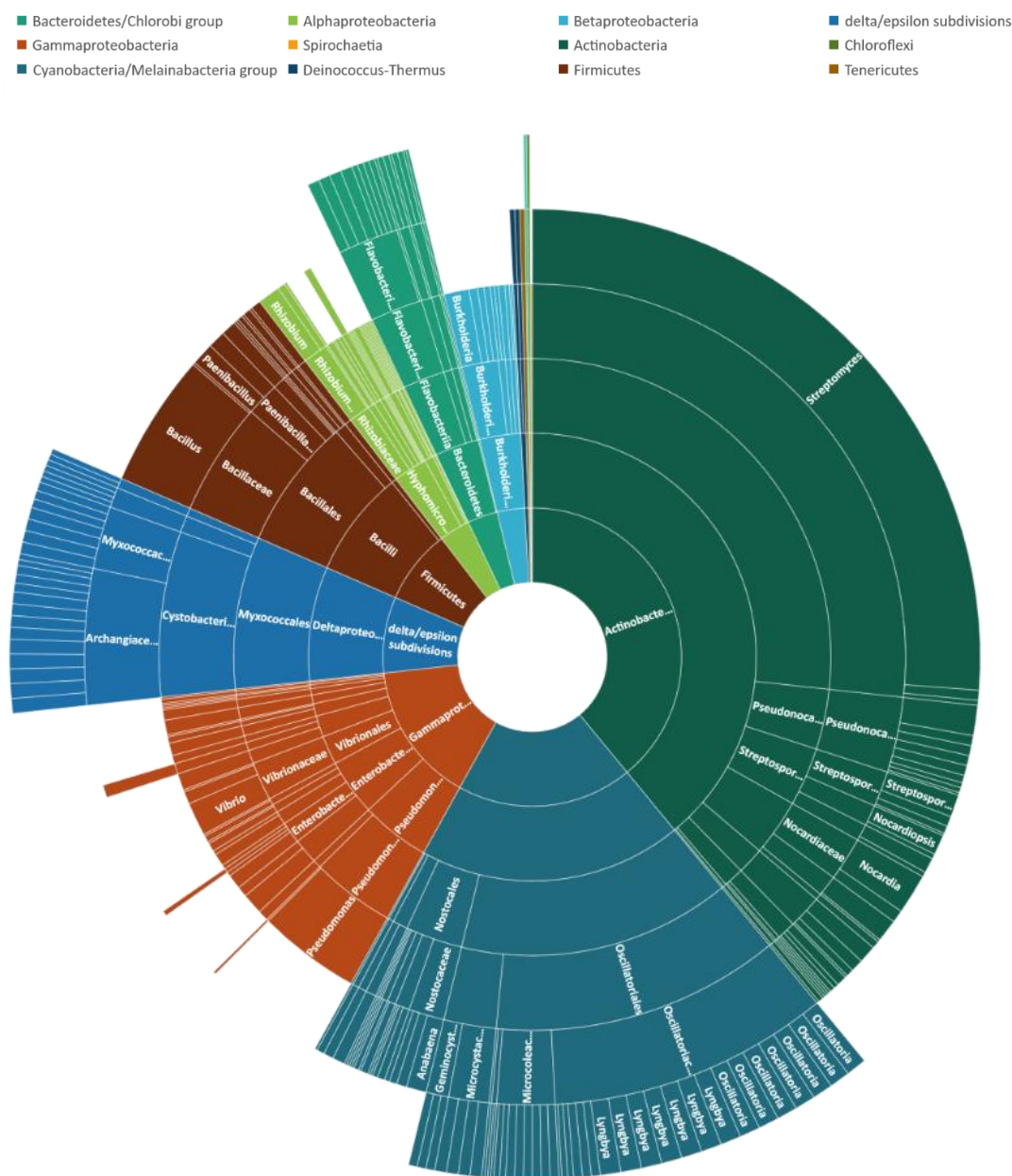


Figure 6. Genera distribution in SECRETed defined microbial producers.

Thus, 31 MFs composed from more than two biosurfactants contained at least one compound with known BGCs and biosynthetic pathway. Similarly, 29 siderophore MFs with more than two members had at least one “anchoring point” whose biosynthetic information can be extrapolated to the rest of the MF.

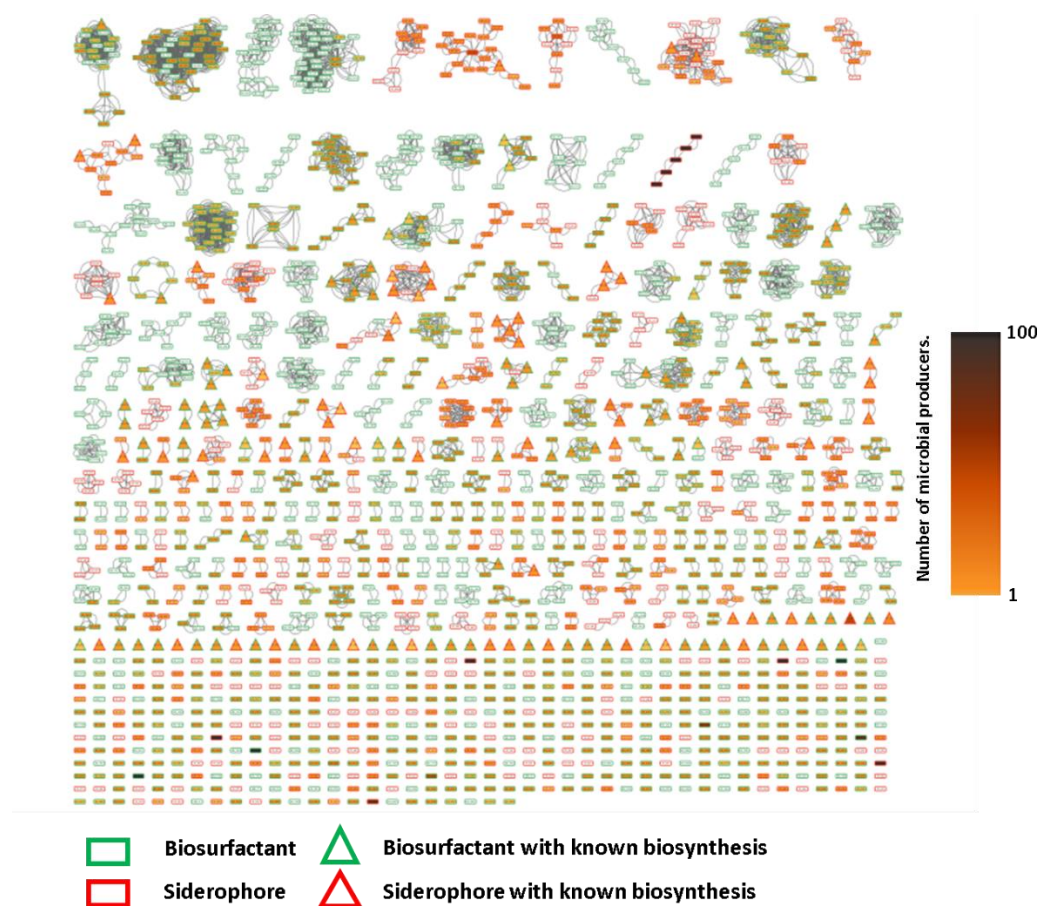


Figure 7. Microbial producer distribution in SECRETed biochemical space. Connected nodes are molecular structures grouped in MFs (see deliverable D3.2).

Once microbial producers are defined, the other major aspect to resolve is genome availability.

Tens of thousands of sequenced microbial genomes or rough drafts of genomes are available at this time, and this number is predicted to grow into the millions over the next decades. Still, this wealth of sequence data may be insufficient when looking for specific strains whose link to molecular structures is defined but sequencing efforts has not being put in place yet.

Moreover, there are plenty of cases where the focus of the research has been dedicated to the compound characterization, leaving the producing organism taxonomical characterization to just identify the species, or even worst, only the genus. Thus, within the broad spectra of unique microbial producers taxa (more than 900), we could define 374 microbial strains producers specified till the strain taxonomy category, 464 microbial producers specified to

their species category and 56 producers specified only in the genus category (Figure 6 and Figure 8A).

Figure 8 shows the degree of genome availability along with the NCBI RefSeq genome collection¹³ (containing more than 400,000 prokaryotic genomes) within each of the different unique microbial producers and their level of taxonomical specification that were compiled and conciliated from above-mentioned databases.

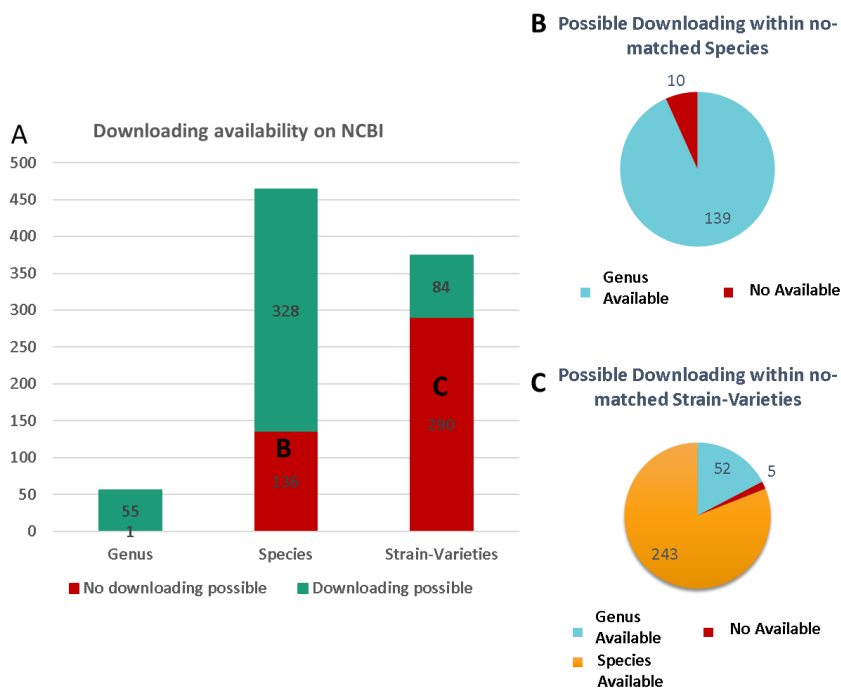


Figure 8. A) Number of taxa with available genome links in NCBI. B) Genome links available when species without link are “downgraded” to their previous taxonomic categories (Genus) . C) Genome links available when strains without link are “downgraded” to their previous taxonomic categories (species or genus).

Overall, 84 microbial strains, 328 microbial species and 55 microbial genus had available genome downloading links (figure 8A).

No need to say that the only reliable structure-organism link is when the microbial strain is defined. Otherwise, further pan-genomic analyses are needed (such as studying in all available genomes the degree of BGC conservation), leading to less reliable results. To further consider that case, we explored the number of genome sequences available when “downgrading” each microbial producer to a broader taxonomic category (Strains>Specie>Genus, Figure 8B and 8C).

Figure 9 categorises several ranges of genome downloading links and classifies the number of microbial producers in each category. Notice that several species such as *Escherichia coli*, had more than 1000 genomes available.

¹³ Tatusova et., Nucleic Acids Res.43(Database issue):D599-605.(2015)

Further steps into Database expansion will be referred to only microbial strains with designated genome sequences, leaving pan-genomic analyses to more advanced stages of the project.

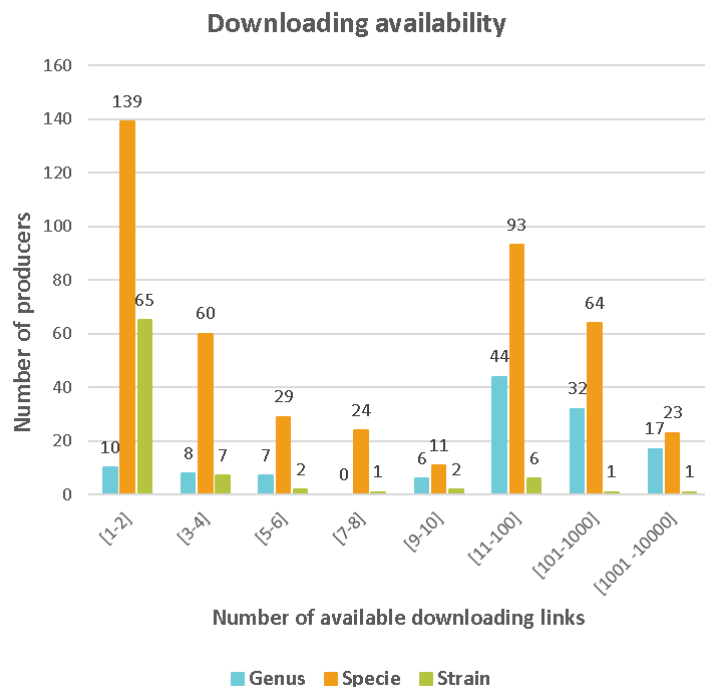


Figure 9. Number of microbial producers found in each range of genome downloading links per taxa category (Genus, specie and strain).

4. SECRETed Biosynthetic space data expansion.

4.1. Bottom-up database expansion.

The next step included in the bottom-up approach (Figure 1) is the expansion of the already created database. This process has been carried out in two consecutive phases:

Manual expansion, which consisted in expanding refined and curated BGCs Database by obtaining the genome sequences of registered microbial producers (see section 3.2) and inferring within their genomes highly similar BGCs using AntiSMASH¹⁴, being this step manually curated. The detected clusters were validated when clustered together with refined information (see section 3.1) using Big-SCAPE tool¹⁵. Under this approach, ~50 Siderophore and ~25 Biosurfactant BGCs were identified.

Automatic expansion, where more potential BGCs were collected based on their gene cluster similarity to highly refined and curated BGCs. In this step, the BiG-FAM database (Kautsar, Blin, et al., 2021) was utilised, providing ~30 gene cluster families (GCFs), containing ~200k BGCs, with similarities above the default threshold. Further pairwise distances to each query BGC calculation using a BiG-SCAPE analysis and its clusterisation in BGC networks is ongoing (Figure 10).

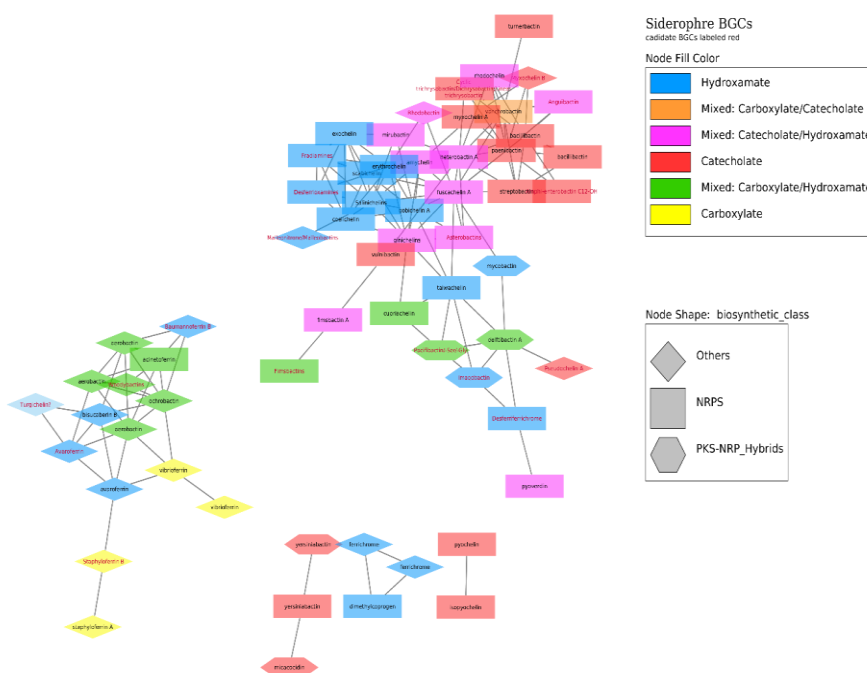


Figure 10. Example of refined and curated BGC siderophore network .

¹⁴ Blin et al., *Nucleic Acids Research*, 49(W1), W29–W35. (2021)

¹⁵ Navarro-Muñoz et al., *Nat Chem Biol*.16(1):60-68.(2020)

The last phase covered by this approach, which is common to the top-down approach, involves the definition of the genes that form the core of the BGCs and the accessory genes. This phase will be explained in more detail at the end of this report once the second workflow being carried out has been defined.

4.2. Top-down database expansion.

The top-down approach (Figure 1) is based on a different pipeline than the one described above. Although it shares commonalities with the bottom-up approach, in this case the focus is on extracting as much information as possible from structure-organisms links. Based on the data available for each compound and taking advantage of the large-scale search of possible producers collected in different compound databases, the sequences of the genomes available for these producers have been used to massively detect BGCs in them.

Several software packages can be used for this last task. Among them, the most developed is AntiSMASH⁵. However, in recent years, new tools have been developed for the same purpose, including GECCO¹⁶ and DeepBGC¹⁷ which are based on machine learning methods instead of rules definition.

Briefly, said approaches identify co-allocated set of genes containing protein domains collected and inferred from known BGCs. These new tools have great potential as their capacity to detect BGCs is much higher than that of AntiSMASH. In other words, for the same genome, they have a greater capacity to detect BGCs. However, the confidence in the detected BGCs is lower as they are not entirely based on known data as it happens with the rules established by AntiSMASH. For this reason, it is important to assess and verify their potential use in this project, considering their results as valuable hypotheses to be verified. It is also remarkable that the project's efforts are also focused on improving these tools in order to increase their reliability.

4.3. Bottom-up and Top-down approaches conciliation.

To conciliate both efforts, clustering and classification algorithms would be put in practice to collect all possible genetic variants to be correlated to obtained MFs and associated features.

For this phase it is important to determine not only the part of the sequence that is found in common in different BGCs but also the accessory genes included in the BGC that will be responsible for the variability of the clusters and therefore of the molecules they encode.

¹⁶ Carroll et al., GECCO. *BioRxiv*, 2021.05.03.442509

¹⁷ Hannigan et al., *Nucleic Acids Research*, 47(18), e110–e110. (2019).

With this goal in mind, BIG-SLiCE¹⁸ has been reported as an ideal tool and is a promising candidate into broadly classify obtained BGCs. With this tool a first classification of the clusters identified is in process to group them into gene cluster families (GCFs).

In this way, convergence between the two approaches is also achieved, conciliating curated and refined expanded information with newly, non-consider to date BGCs hypothesis from Deep Learning approaches.

Finally, fine classification algorithms such as BiG-SCAPE and Clinker are internally identifying core and accessory genes within each obtained GCFs (Figure 10).

The last step of the described pipeline involves the union of the genetic information obtained after the implementation of this workflow with the knowledge acquired in the analysis of the molecular substructures associated with each of the compounds. For this, the first necessary approach will be the automatic annotation of the BGCs obtained and especially of the accessory genes identified in the last two steps of the pipeline defined above.

The definition of core and accessory genes that take part in SECRETed GCFs is essential to address the retrosynthetic pipelines. Thus, by linking accessory genes and molecular variants SECRETed consortium aims to define the extent of molecular modifications that potentially can be addressed in SECRETed biomolecular space.

¹⁸ Kautsar et al., *GigaScience*, 10(1). (2021).

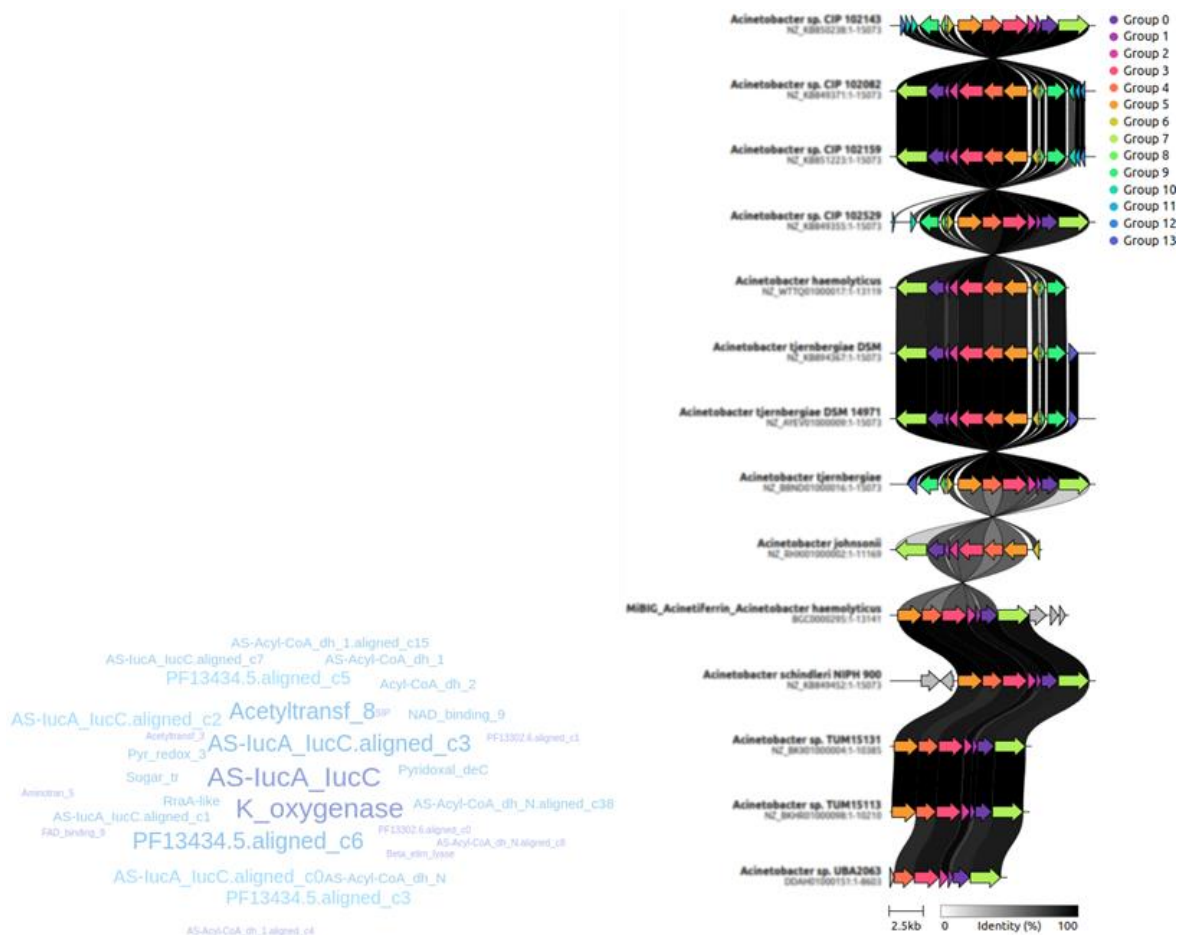


Figure 11. Representations taken from the BiG-SLiCE and Clinker outputs of the clustering tests performed.

5. Conclusions

Information management remains a central limitation in natural products science. Access to comprehensive, structured, freely available repositories containing key data allows researchers to determine what has been found to date, understand how previous discoveries relate to new findings, and identify how new results fit into the broader picture of natural products diversity and biosynthesis.

As the rate of BGC discovery began to accelerate in the early 2000s, the biosynthesis community faced many of the same challenges that had been encountered by the structure of the natural products elucidation community thirty years earlier. In particular, information about BGC discovery was becoming scattered across the scientific literature, or stored in a less structured manner in genomic databases such as NCBI GenBank. As with structure-based discovery, this limited the possibilities for cross-linking between resources and prevented programmable access to exploit the knowledge within.

Central to SECRETed main objectives is to construct a unique microbial amphiphilic compound space comprehending molecular structures, physicochemical characteristics, associated bioactivities, and revealed genetic mechanisms responsible for their biosynthesis.

This work summarizes the BGCs database construction, emphasizing the need of this sort of meta-analyses. In SECRETed, compiled and clustered information will be integrated into Molecular Networks (MNs).

MNs can also combine the layering of multi-informational data, such as taxonomical and biological data to reinterpret SECRETed biochemical space from a different perspective. For example, establishing which genetic clusters and subclusters are responsible for the biosynthesis of the target molecules will also guide SECRETed microbial sequencing efforts facilitating a genetic dereplication platform.

Finally, MNs can help in the elucidation of Molecular families (MFs) and their connection to gene clusters families (GCFs). This ultimately helps to classify enzyme families and their substrate specificities, essential information for SECRETed retrosynthesis algorithms.

Altogether, this report establishes a baseline of what is compiled to date about BGCs and their microbial scaffolds and serves to define future SECRETed achievements which will be summarized into the “Final report on phylogenetic tree database for siderophore and biosurfactant biosynthetic gene clusters containing novel enzymes and their reactions” .