Deliverable D3.5

Title: Final report on phylogenetic tree database for siderophore and Biosurfactant biosynthetic gene clusters containing novel enzymes and their reactions.

Lead Beneficiary	IDENER R&D
Delivery Date	31st May 2023

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.	D3.5
Dissemination Level	Public
Work Package	3
Task	3.1 and 3.2
Lead Beneficiary	IDENER R&D
Contributing beneficiary(ies)	EKUT
Reviewers	
Due date of deliverable	31st May 2023
Actual submission date	
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
v1.0	25/05/2023			

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 can 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 modularise associated metabolic elements to broaden 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 D3.5 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 currently focused on expanding the genetic information related to both the biosurfactants and siderophore previously collected in SECRETed and those with the possibility of being added thanks to the efforts put into expanding the chemical space of the project.

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 presents a summary of improved Biosynthetic gene clusters, informs about the significant challenges and approaches involved in expanding SECRETed biosynthetic space, and describes Gene Cluster Families refinement efforts and their involvement in structure-property-driven retrosynthesis algorithm.

Table of contents

List of Figures

List of Tables

1. Introduction

The pursuit of natural product discovery through genome-directed approaches involves the identification of biosynthetic gene clusters that exhibit similarities to genes responsible for synthesising known secondary metabolites. Despite the broad and diverse range of secondary metabolites, certain compounds, such as non-ribosomal peptides and polyketides, demonstrate remarkable conservation of their biosynthetic pathways among microbial species. This conservation is reflected in the high sequence similarity observed in many core biosynthetic enzymes at the molecular level[1].

In contrast, determining the genetic sequences accountable for synthesising specific metabolites is typically more time-consuming than deciphering their structures and functions. This disparity arises from the reliance on curated and refined genomes, which can take time to acquire in numerous instances. Additionally, establishing the relationship between compounds and genes is time-consuming and requires careful validation[2].

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. Such GCFs can be matched to molecular families (MFs) identified from mass spectrometry (MS) data based on observed/predicted chemical features[3].

For that purpose, more than 1800 molecular structures of siderophores and biosurfactants were refined by cross-referencing other NP-related databases containing their unique data coverage to be further clustered into MFs (see Deliverable D3.2).

This report presents a comprehensive overview of the ongoing endeavours aimed at broadening the genetic space associated with SECRETed compounds, which also have undergone expansion (refer to deliverable D3.4). As a result, there has been an improvement in the count of Biosurfactant and siderophore-producing Biosynthetic Gene Clusters (BGCs) encompassed within the SECRETed space. Furthermore, for specific gene cluster families (GCFs), it has become feasible to identify both core genes and accessory genes that play distinct roles in the synthesis of these compounds.

[1] Ziemert et al., Nat. Prod. Rep. 33, 988–1005. (2016)
[2] Scherlach, K., Hertweck, C., Nat Commun 12, 3864 (2021).
[3] Nguyen et al., *Proc. Natl. Acad. Sci. U.S.A.* **110**:E2611–E2620.(2013)
[6] Nguyen et al., *Proc. Natl. Acad. Sci. U.S.A.* **110**:E2611–E2620.(2013)

2. SECRETed Biosynthetic space extension: Defining the expansion approaches.

Biosynthetic gene clusters (BGCs) are DNA sequences responsible for encoding the production of natural products, particularly secondary metabolites like polyketides and non-ribosomal peptides. This paradigm has revolutionised natural product research by allowing researchers to connect specific genes to the molecules they produce[4]. Nowadays, the process is made more accessible by the usage of bioinformatics tools such as Antismash[5] or DeepBGC[6], which enable the inference of biosynthetic gene clusters (BGCs) within a specific genome. These emerging methodologies significantly contribute to the genetic identification of various secondary metabolites with diverse applications. During this project phase, different tools have played a vital role in expanding the repertoire of newly discovered and well-characterized BGCs, allowing for the genetic characterisation of the SECRETed chemical space by improving genetic information associated with already defined "anchoring points".

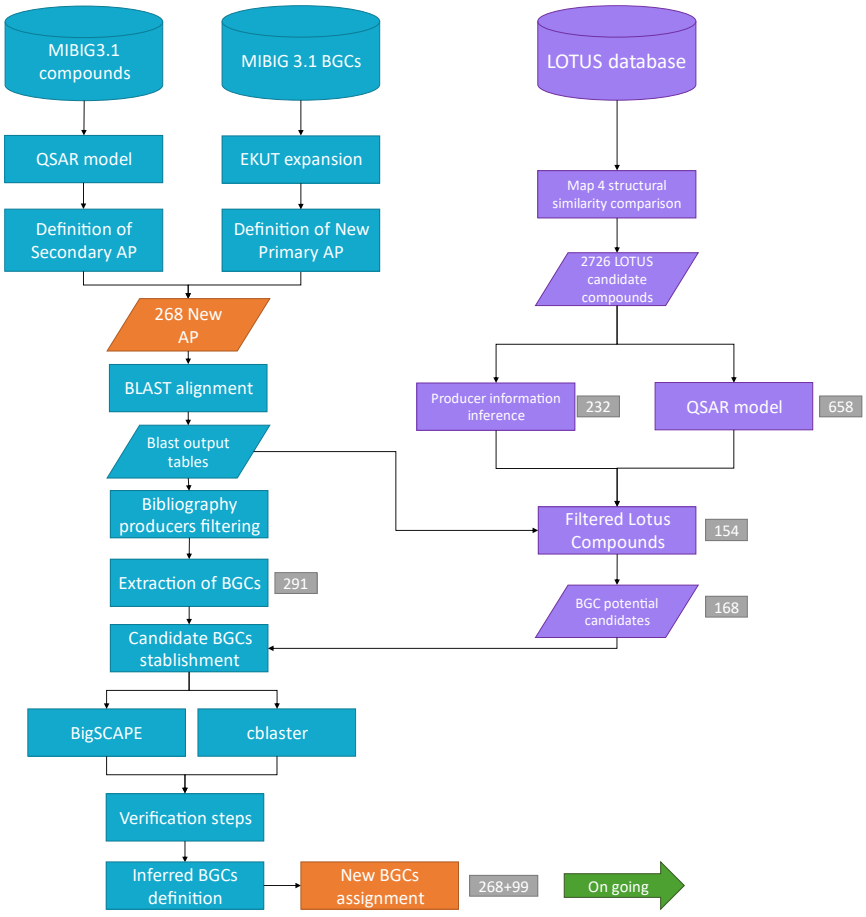


Figure 1. Flow chart covering the two major workflows carried out in SECRETed Biosynthetic space expansion.

[4] Jensen PR., Trends Microbiol. 2016 Dec;24(12):968-977.
[5] Kai Blin et al., Nucleic Acids Research, Volume 49, Issue W1 (W29-W35), 2 July 2021.
[6] Geoffrey D Hannigan et al., Nucleic Acids Research, Volume 47, Issue 18, p e110, 10 October 2019 Oct.

In this way, two complementary approaches have been used (Figure 1)

i. An approach that uses as a starting point the information contained in MIBiG[7] version 3.1[8], where efforts are focused on adapting the latest version of this database to the SECRETed chemical space from a genetic and structural point of view. This part of the workflow has a BGC inference protocol using BLAST-type sequence aligners that have enabled the discovery of inferred BGCs.
ii. An approach based on the chemical expansion mentioned in deliverable 3.4 in which LOTUS is used as a database of possible new biosurfactants or siderophores.

Both approaches had common objectives:

- Expand the amount and genetic information at the BGC level associated with SECRETed compounds.

- Cluster and characterise the expanded BGCs based on the producer of origin and the compound associated with those BGCs.

- Characterize BGCs associated with new compounds derived from SECRETed chemical space expansion.

- 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 redaction rules.

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

[7] R. Terlouw et al., Nucleic Acids Research, Volume 51, Issue D1:D603–D610, (2023)
[8] Terlouw, Barbara R et al.,Nucleic acids research vol. 51,D1,(D603-D610), (2023)

3. SECRETed Biosynthetic space data extension and curation.

3.1. MIBiG 3.1 data obtention and curation.

As explained in the previous section, the first workflow uses the MIBiG 3.1 database. The MIBiG (Minimum Information about a Biosynthetic Gene Cluster) Data Standard and Repository is a project that aims to facilitate the curation and storage of known biosynthetic gene clusters (BGCs)[9]; it is a handy resource in this field and underwent an update in January 2023, aimed at enhancing the information available on biosynthetic gene clusters (BGCs) and their associated compounds.

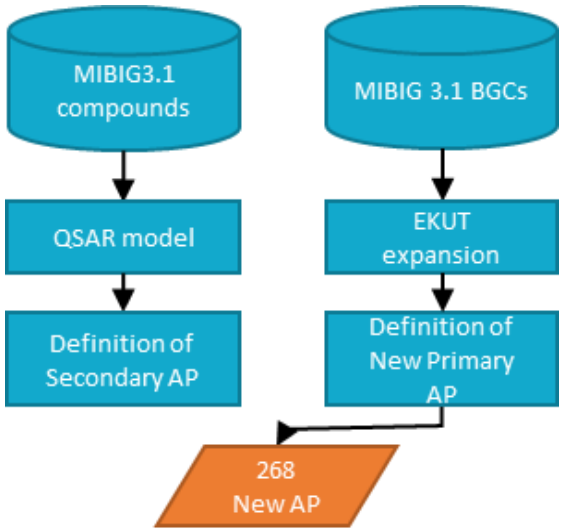


Figure 2. Sub workflow related to MIBiG v3.1-based expansion

It should be recalled that in previous deliverables, the number of reference BGCs was 121; of these 121, 4 BGCs were discarded by the source database for insufficient quality. Therefore, our starting point was those 117 BGCs extracted from MIBiG v2.0 (Figure 3). Therefore, taking advantage of this new update, the SECRETed team has approached this new expansion from two distinct perspectives that highlight different aspects of this database. The first one is based on the use of biosynthetic information. In contrast, the second one approaches this database from a chemical point of view, using the tools already developed during the project.

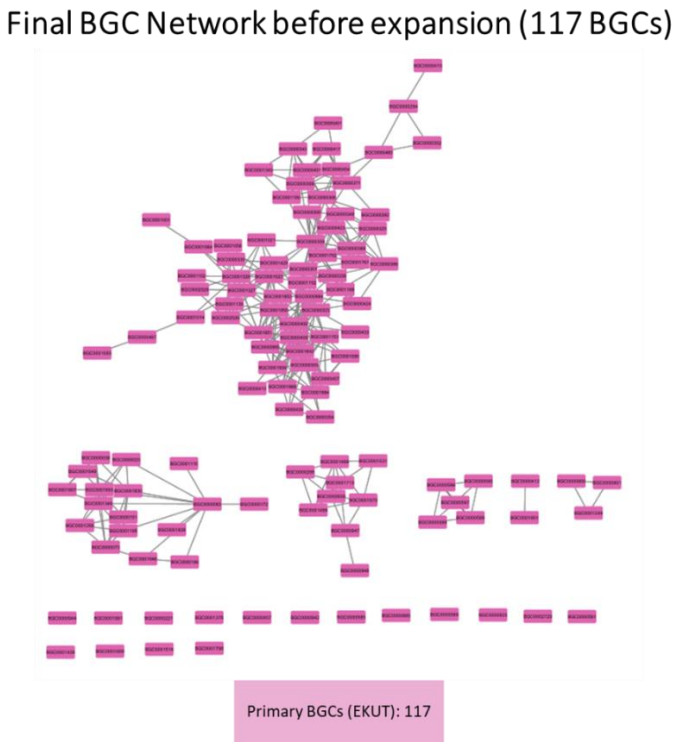


Figure 3. Original BGC network created with BigSCAPE and the original refrence BGC set (117 BGCs).

3.1.1 From Genetic to Structural data: MIBiG 3.1 exploration.

This part of the protocol had three main steps: extraction and curation of the MIBiG database, cross-reference with SECRETed compounds, and verification of interactions by bibliography.

Based only on this first approach, MIBiG v3.1 helped to expand the genetic space from 117 to 255 BGCs encoding the biosynthesis of 361 unique compounds. Among these compounds, 176 siderophores and 185 biosurfactants were already found in the SECRETed chemical space.

3.1.2 From Structural to Genetic data: QSAR approach.

In this approach, we have started with the chemical structures of all compounds collected from MIBiG v3.1. To obtain these structures, we had to extract and curate the database thoroughly. It should be noted that the information corresponding to the structures of the compounds comes from various databases such as PubChem, ChemSpider, or Natural Product Atlas, among others; for this reason, it was necessary to separate the compounds in their databases of origin, verify the information contained in these and download the sdf corresponding to each molecule. There were cases in which a compound was not directly associated with any database; in these cases, a manual search was performed to find the structure of that compound in any database and thus be able to complete the chemical space of MIBiG 3.1 as much as possible.

Once defined, the chemical space of MIBiG (in sdf format) was used as input for the QSAR model mentioned in deliverable 3.4, which can infer whether a given molecule has potential activity as a biosurfactant or siderophore. Once this model was applied, 12 biosurfactants and

47 siderophores were defined as "anchoring points", which were, in turn, associated with 43 BGCs.

Among these additions, the unique contribution of this approach consists of 5 biosurfactants and 36 siderophores defined as "anchoring points", as the rest of the additions were already present in the previous expansion explained (see 3.1.1 section). These unique additions were associated with 30 BGCs.

3.2 MIBiG 3.1-based expansion results.

3.2.1 From Structural to Genetic data: QSAR approach.

Combining both approaches already explained by the usage of data obtained from the MIBiG v3.1 database, the genetic characterisation of SECRETed compounds has been improved for siderophore and biosurfactant cases (Table 1). Using this new data, the existing database has been updated from 117 to 268 new BGCs associated with a total of 507 SECRETed compounds (Figure 2). Recall that this database contains all the available information on these added Gene Clusters, such as links to GenBank files, MIBiG repository[9], NPAtlas[10], and PubChem data[11]. This database also contains visualisations 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 "anchoring point" where further BGC information is inferred in subsequent expansion processes that will be explained later within this deliverable.

Table 1. Simplified statistics of the database status before and after using MIBiG v3.1 as a resource.

	Before MIBiG v3.1	After MIBiG v3.1
Total Siderophores	~560	608
Total Biosurfactants	~1500	1589
Siderophores with BGC	64	292
Biosurfactants with BGC	108	215
Covered Siderophore %	11.43%	48,03%
Covered Biosurfactant %	7.2%	13,53%

[9] Kautsar et al., Nucleic Acids Research, 48(D1), D454–D458.(2020).
[10] van Santen et al., ACS Cent Sci. 5(11):1824-1833. (2019)
[11] Kim et al., Nucleic Acids Res. 44(Database issue): D1202–D1213.(2016)

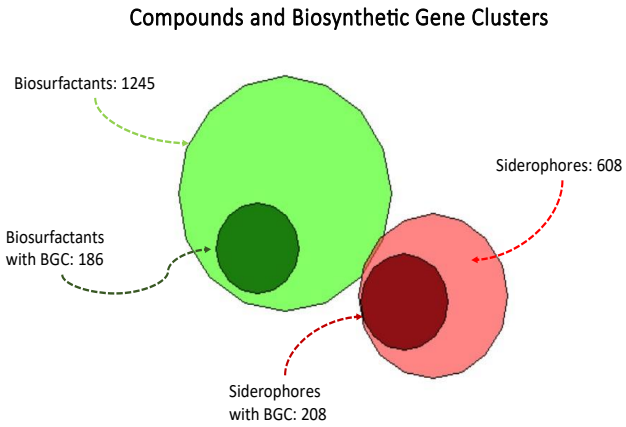


Figure 4. Venn diagram representing the BGCs found for each group of compounds established by the partners after MIBIG 3.1 analysis.

Figure 5. BGC network created with BigSCAPE after MIBiG v3.1-based expansion (Input: 268 BGCs).

3.2 BGC expansion: Inference of new BGCs using sequence alignment-based tools.

In addition to the proposed expansion, a BGCs inference protocol was defined (Figure 3). It uses BLAST as a core tool in combination with other software such as Big-SCAPE and cblaster used in verification steps. The central concept of this approach is simple: using the already expanded reference BGCs as bait to "fish" for new possible BGCs that closely resemble the sequences associated with the Anchoring Points.

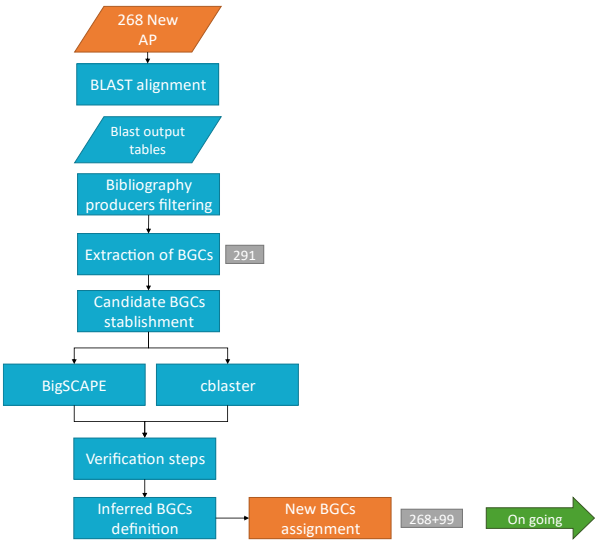


Figure 6. Protocol of BLAST-based protocol to infer new BGCs represented in a flowchart.

To expand and verify the available APs, all "gbk" files related to previously added BGCs were converted into "fasta" files using Bio.SeqIO python package and passed one by one through BLAST. Then the description tables were extracted and unified, and various checks were made to select and check:

- Only matches belonging to one of our producers (previously, our entire database of producers was reviewed). This is done because, in this case, we only want to expand our AP space, not our producer's space.
- Ensure that matches that **are not** the BGCs themselves by checking the sequence ID (It is possible to match with the exact sequence collected in MIBiG, which would be a duplicate).
- Blast Coverage should be higher than 80%. Otherwise, we may collect sequences without functional relation with our original AP.
- We also perform a selection based on the length of the matches. It should be similar to that of a BGC.

Once all these filters were performed, 291 BGC candidates were selected. The next step was collecting the exact matched sequences. In some cases, the target was found within an entire genome sequence. In these cases, it was necessary to extract only the matched portion of that genome. To do that, hit tables resulting from BLAST were consulted, and a Python script was designed to download sequence fragments directly from the NCBI repository.

Once the candidates have been selected, the next step is to verify that they meet the requirements to be considered BGCs associated with the SECRETed chemical space. For this task, the essential tools are BigSCAPE[12] and cblaster[13]. BigSCAPE is a bioinformatics software tool for analysing and clustering biosynthetic gene clusters (BGCs) in microbial genomes. Its

[12] Navarro-Muñoz, Jorge C et al., Nature chemical biology vol. 16,1 (2020)
[13] Gilchrist, Cameron L M et al., Bioinformatics advances vol. 1,1 vbab016. 5 Aug. 2021

primary purpose is to identify and classify BGCs based on their similarity and evolutionary relationships[11]. On the other hand, Cblaster is a bioinformatics tool that allows the analysis of homologous gene clusters in microbial genomes. It compares a user-provided query sequence against a database of genomes to identify similar gene clusters. Cblaster provides information about the location, gene composition, and potential functions of the identified clusters, aiding in the exploration of genetic diversity and evolutionary relationships[12].

The verification process has been done in parallel with both tools, but the filtering steps are the same:

1. The selected BGCs must present a connection with some anchoring point in both the BigSCAPE and cblaster results.
2. Among the genes that form each alignment calculated between reference BGC and inferred BGC, there must be biosynthetic genes, also called core genes. In other words, the relationship between the BGCs cannot be due only to the alignment of accessory genes.
3. The compounds associated with the inferred BGCs must have a similarity ratio greater than 0.597 with those associated with the reference BGC with which it presents a genetic similarity.

Once all these verification steps were carried out, we went from 291 candidates to 99 verified BGCs, which complemented the 268 BGCs already considered as references, thanks to the previous protocols. Therefore, at this stage of the project, it could be said that 367 BGCs are available. However, it is essential to note that these protocols are ongoing and that the process of extending BGCs is iterative, thanks to the defined workflow.

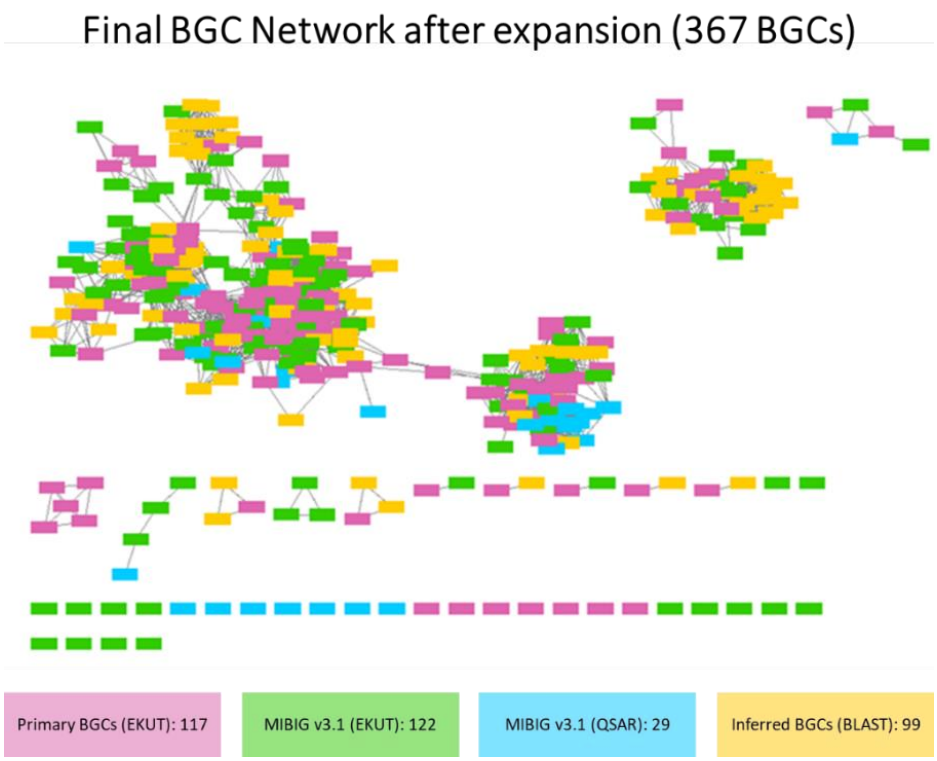


Figure 7. BGC network created with BigSCAPE after BLAST-based BGC inference (Input: 367 BGCs).

Once all these extensions have been applied, their effect on both the BGCs network (Figure 7) and the composite network based on their structural similarity (see Deliverable 3.4) can be appreciated. For example, Figure 8 shows the gain of genetic information after the expansions of the BGC space on cluster 19.

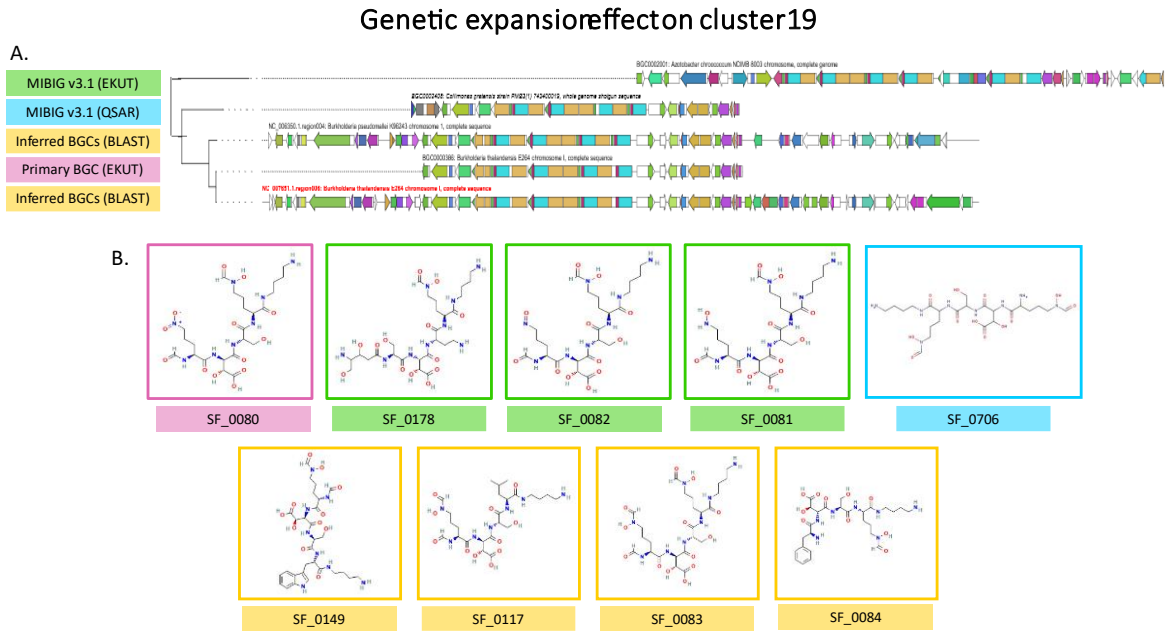


Figure 8. BGC expansion effect on both genetic (A) and structural (B) ambit for cluster 19.

4. SECRETed chemical space expansion: application of chemical expansion to genetic expansion.

In the effort to try to infer as much genetic information as possible from the SECRETed compounds, it was decided to take advantage of the SECRETed chemical space expansion protocol already explained in deliverable D3.4 for the benefit of genetic expansion. By employing this protocol, we have successfully obtained valuable genetic insights. As a reminder, it should be noted that this protocol consisted of 2 main steps: a Structural Similarity comparison between the LOTUS database and the SECRETed database, followed by the application of a QSAR model to identify biosurfactants and siderophores among the LOTUS candidates (see Figure 9).

It was decided to use the extensive data generated from the aforementioned protocols to further enrich the available information on these candidate compounds. Specifically, we have used the tables of BLAST analysis results, which provide details on the genomes/sequences within each match is present. This protocol is further explained below, allowing us to infer genetic information associated with the candidate compounds accurately.

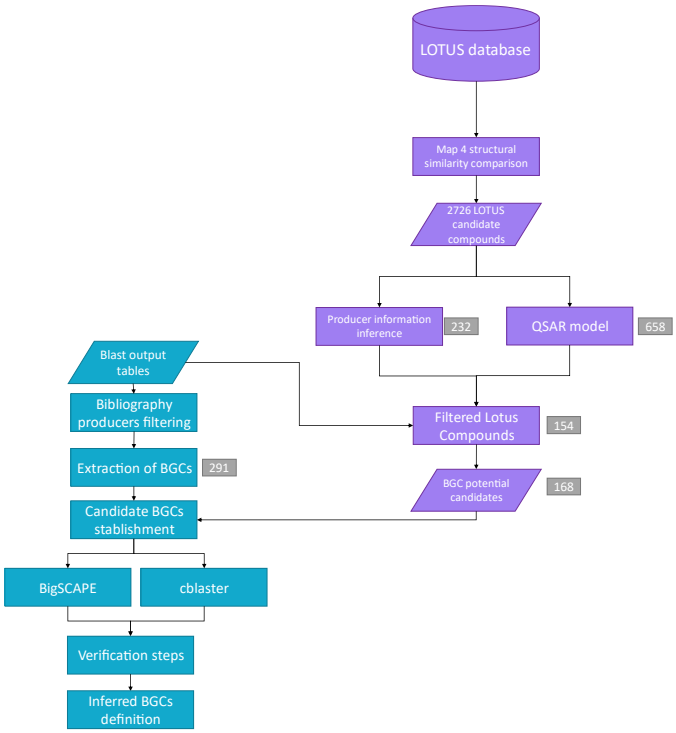


Figure 9. Protocol of BGC inference protocol based on LOTUS chemical expansion.

4.1 Verification of LOTUS candidates using biosynthetic data.

As discussed above, the protocol begins with the selection of compounds that have passed the QSAR model's biosurfactant or siderophore distinction (totalling 658 compounds. Additionally, we obtained descriptive tables from the BLAST nucleotide analysis, which involved querying sequences corresponding to our reference BGCs. However, to effectively utilise this information, we also needed to extract the producers associated with each candidate compound from LOTUS.

To achieve this, we have utilised the PubChem database as an intermediary. PubChem contains a significant portion of the LOTUS compounds and their corresponding registered producers. To ensure precision, the organism-associated data within the LOTUS database was also utilised in this part of the protocol. By integrating these data sources, we identified a total of 232 potential producers associated with the compounds that passed the QSAR model.

Once all the relevant information was acquired, the producers associated with the selected LOTUS compounds were cross-referenced with organisms that contained similar sequences to a given reference BGC, as determined by the BLAST results. Completing this step required two criteria: structural similarity between the LOTUS compound and the compound (anchoring point) produced by the reference BGC and a matching producer strain between the BLAST result and the candidate compound at the strain level. Following this protocol, we identified 154 potential new compounds (Figure 10) associated with 168 BGCs.

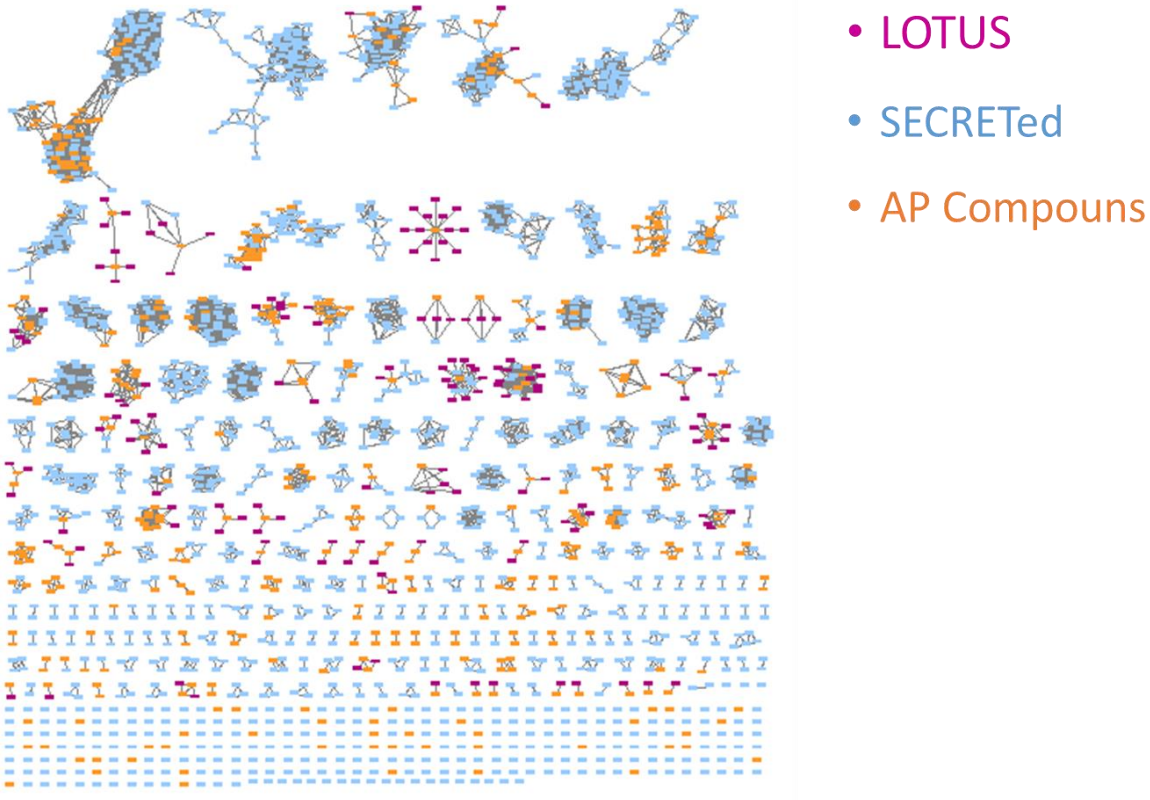


Figure 10. Representation of the effect of LOTUS compounds addition to the molecular families.

5. SECRETed producers space expansion: usage of BLAST protocol results and PubChem database.

In line with the project, one of the objectives of the collaborators is the continuous improvement of the available data. Therefore, although several protocols for the expansion of producers have already been defined in previous deliverables, this desire for continuous improvement has led to testing new ways to expand this space of organisms with the potential to produce certain biosurfactants or siderophores. In this case, we have taken advantage of the efforts of the protocol defined for the inference of new BGCs and the massive consultation of the PubChem database, which contains information on producers associated with specific compounds.

5.1 Blast-based producer space extension.

As summarised in the previous section, the protocol defined for BGC inference provides information on these gene clusters and possible alternative producers to the anchoring points. This can occur when both the Query Cover of the match and the identity exceed 90% and 95%, respectively. In these cases, as much information as possible has been collected on these "alternative producers", and a count has been carried out to determine the potential for expansion of the producer space by taking advantage of this protocol. The steps followed were:

1. Starting from the original outputs obtained from BLAST, we eliminated the matches that met the requirements of Coverage and Identity.
2. We also discarded those matches whose source organism coincided with the strain responsible for producing the original PA (in these cases, it would be considered to have matched with itself).
3. Finally, the names of potential candidate strains were curated and compared to the original producer space (Figure 11).

Initially, the starting point was a space of 12572 candidate sequences. Still, after filtering for coverage and identity, this number was reduced to 8638, which after curation and unification of strain and species names, turned out to be 4464 potential candidate sequences corresponding to 4197 producers (Organism level). Among these producers, 31 were already present in SECRETed data space (Figure 12).

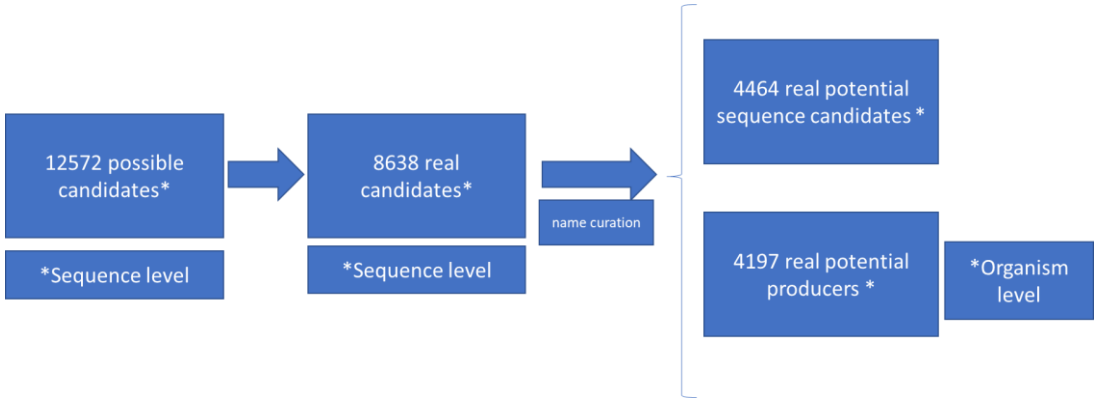


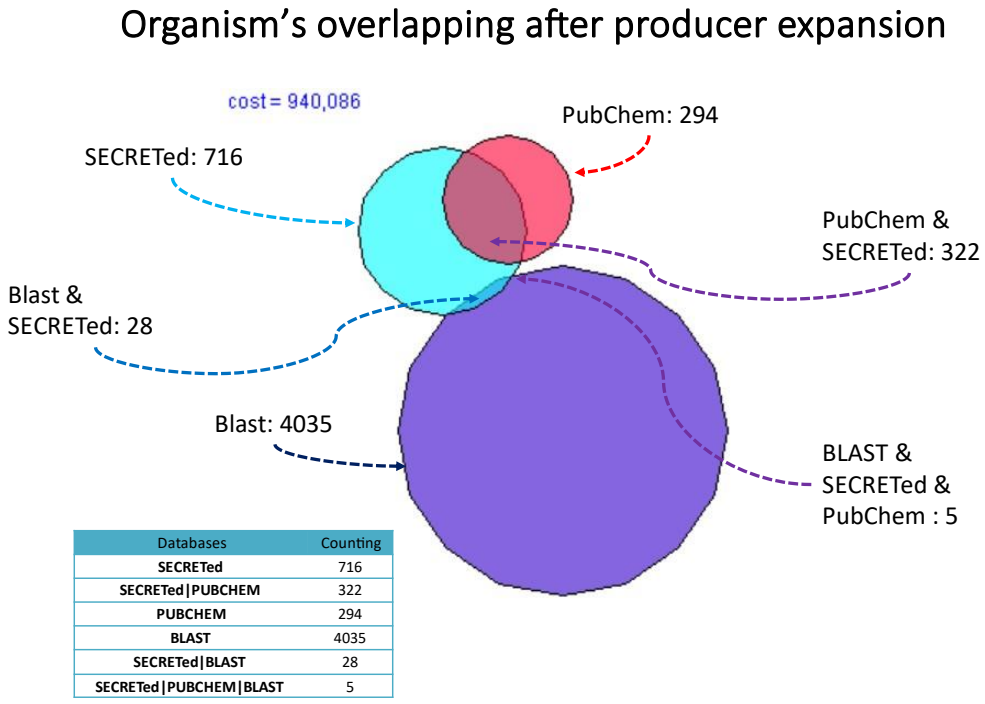
Figure 11. Expansion of producers explained with numbers.

5.2 PubChem-based producer space extension.

The second protocol that is part of this section of the workflow is based on the use of PubChem to infer more producers associated with the chemical space of SECRETed. For this purpose, all available cids of SECRETed compounds (both siderophores and biosurfactants) were used. This protocol also had several steps:

1. Automation of the download of the taxonomy table available for each of the cids.
2. Unification of all the tables
3. Inference of the TaxIDs for each registered organism.
4. Filtering of results, we are only interested in Prokaryotes organisms.
5. Curation and comparison of the organisms obtained concerning the original database (SECRETed producers database).

In this case, after curation and unification of strain and species names turned out to be 621 producers (Organism level). Among them, 322 were already present in the original database, and five were both present in the original space and within BLAST results (Figure 12).



Databases	Counting
SECRETed	716
SECRETed|PUBCHEM	322
PUBCHEM	294
BLAST	4035
SECRETed|BLAST	28
SECRETed|PUBCHEM|BLAST	5

Figure 12. Venn diagram showing the overlapping grade between producers inferred by three different sources: Pubvhem-based protocol, Blast-based protocol and organisms already collected in SECRETed dataset.

Although the results of the protocols described above appear promising, the use of other approaches in the definition of new producer organisms is still under investigation. AntiSMASH (Antibiotics and Secondary Metabolite Analysis SHell) is the most commonly used genome mining tool for BGCs. Previous versions of AntiSMASH[14] could only identify NRPS-Independent siderophores. However, the latest version (antiSMASH 7.0) introduced rules for detecting NRP- metallophores based on the presence of biosynthetic pathways associated

[14] Kai Blin et al., Nucleic Acids Research, Volume 49, Issue W1, 2 July 2021, Pages W29–W35,

with eight different chelating groups. By employing AntiSMASH 7.0, we performed genome mining on all approximately 64,000 species representatives in the Genome Taxonomy Database (GTDB). This ongoing approach aims to identify potential siderophores and biosurfactant producers across the bacterial kingdom from a genome-mining approach. Furthermore, these results were employed to explore the phylogeny of each biosynthetic pathway to uncover their evolutionary history.

6. BGC analysis: Identification of core and accessory genes.

Finally, to enhance our understanding of Siderophores and Biosurfactants, we also analysed the previously investigated Biosynthetic Gene Clusters (BGCs) associated with these compounds from a genetic and functional point of view. For this purpose, we employed various bioinformatic tools such as Clinker[15] (a BGC comparison and visualisation tool), AntiSMASH, and cblaster. Through this analysis, we identified several genes that play a crucial role in the production of different groups of siderophores and biosurfactants. These principal genes were defined as "core genes", and the remaining genes whose function is not purely biosynthetic have been defined as "accessory" genes.

By leveraging this information, it has been possible to create a compound-BGC partnership network that has been posted on the SECRETed web platform. In this platform, it is possible to see not only the network of BGCs with their associated compounds but also the relationship between the genes of a set of BGCs, being able to appreciate the transmission of accessory and core genes (Figure 13).

Moving forward, our research will utilise an expanded database to explore "tailoring genes" that can be employed to modify BGCs and cater to specific industrial chemical requirements. Furthermore, we will focus on investigating condensation starter domains (Cs Domains) to deliberately alter the fatty acid chain length of compounds synthesised by Nonribosomal Peptide Synthetases (NRPS). This aspect is critical as it determines the amphiphilic nature of Biosurfactants and Siderophores.

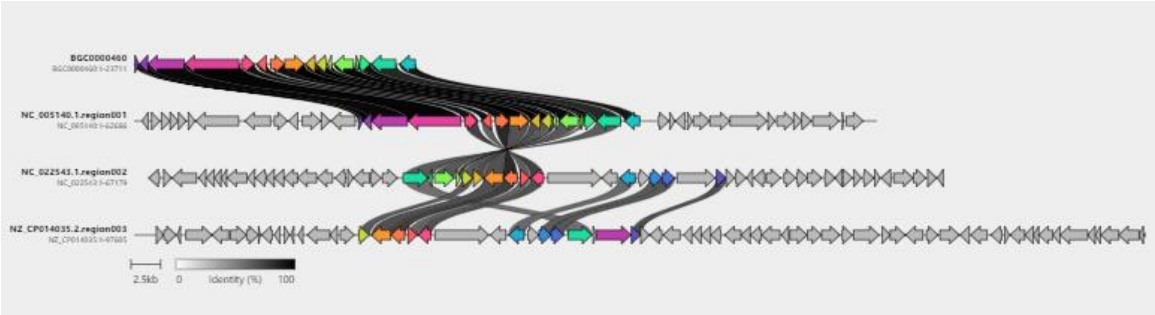


Figure 13. Example of BGC comparison available on SECRETed platform

[15] Breon M Schmidt et al., GigaScience, Volume 7, Issue 7, July 2018.

1. Conclusions

The effective management of information remains a significant challenge in natural products science. Access to comprehensive and well-organized repositories is crucial for researchers to assess existing knowledge, establish connections between discoveries, and gain a broader understanding of natural product diversity and biosynthesis.

In the early 2000s, as the discovery of biosynthetic gene clusters (BGCs) accelerated, the biosynthesis community faced similar challenges encountered by the natural products structure elucidation community three decades earlier. Information regarding BGC discovery was scattered across scientific literature or stored in unstructured formats in genomic databases like NCBI GenBank. This fragmented distribution of knowledge hindered resource interlinking and restricted programmable access, limiting the exploitation of collective wisdom. For this reason, one of SECRETed's 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 report presents a comprehensive overview of the expansion of the biosynthetic gene clusters (BGCs) database, emphasising the growing importance of meta-analysis in this field. Our work encompasses the development of various protocols, leveraging existing databases and incorporating novel approaches to infer BGCs linked to compounds. By combining these protocols, we have achieved an unprecedented extension of SECRETed biosynthetic space. Furthermore, the outcomes of chemical expansion have been integrated into these workflows, highlighting the interconnectedness of both networks.

In addition to expanding the biosynthetic space, significant progress has been made in enhancing the pool of producers associated with SECRETed's chemical space. This advancement enables an increased number of candidates for future genome mining processes. These collective efforts pave the way for further exploration and discovery in the field of natural product research and retrosynthesis.

Altogether This report tries to gather all the updates and improvements applied to the BGCs reference space, which was initially available, and consolidates the bases to define future SECRETed achievements.