Ref. Ares(2023)3763396 - 31/05/2023

Deliverable D3.4

Title: Final report on SECRETed Molecular Networks: Publicly available molecular structures and inference of associated features to molecular fingerprints

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.	D3.4
Dissemination Level	Public
Work Package	3
Task	3.1
Lead Beneficiary	IDENER R&D
Contributing beneficiary(ies)	
Reviewers	IDENER team
Due date of deliverable	31st May 2023
Actual submission date	31st May 2023
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	05-20-2023	IDE R&D	IDENER team	First draft
2.0	05-30-2023	IDE R&D	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, the 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 standardize and modularize 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 D3.4 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 3.1.1: “Molecular Families inspection” focuses on collecting information about structures of compounds, linked properties, and biological activities of biosurfactants and amphiphilic siderophores contained in publicly available databases and available information in scientific literature.

Together with subtask 3.1.2: “Amphiphilic siderophores and biosurfactant biosynthetic gene cluster analysis” 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 how obtained molecular data has been utilized to learn the main attributes conferring surfactant and chelating abilities. Thus, this information has been employed to inspect databases for molecules that may possess such features. Finally,

SECRETed biochemical space has been decomposed into their metabolic intermediates, facilitating both retrosynthetic efforts and further modeling approaches.

Table of contents

List of Figures

List of Tables

1. Introduction

Bacteria are some of the most clever and robust producers of Natural Products (NPs). Their size, generation time, and capacity for horizontal gene transfer allow these organisms to rapidly mutate, evolve, and adapt to their environment[1]. These same characteristics also make them highly amenable to genetic engineering for the study and production of bioactive, and potentially, pharmaceutically relevant compounds.

Moreover, bacteria readily convert simple starting materials into a diverse array of complex molecules with useful bioactivities. The ability of bacteria to survive and thrive in their environment is contingent on these chemistries. SECRETed targeted compounds, biosurfactants and siderophores are within those survival providing molecules.

In addition to that, the need to classify said molecules may mislead their understanding, specially when boundaries are not clear. That is the case of amphiphilic siderophores and chelating biosurfactants.

This fact highlighted the need to learn from a physicochemical point of view what makes a chelating agent and what makes a surface active agent. Thus, said knowledge could be extrapolated to be used to study what other natural products have not been yet tested as those in order to expand SECRETed biochemical space.

To do so, we also needed to fully address SECRETed biochemical space as a sum of building blocks connected by reaction rules. This knowledge opens the possibility to learn when and how the addition of certain bricks leaded to specific physicochemical features.

This report continues the previous deliverables D3.2 and D3.3. Specifically, it shows the advancements regarding chemical data collection and conciliation; its organisation in Molecular Families based on molecular fingerprints and clustering algorithms; the distribution of SECRETed key data along with the different clusters; and the modular comprehensive understanding of biosurfactants and siderophores, which eases further steps of SECRETed project.

[1] Fischbach et al.,Proc Natl Acad Sci USA, 105 4601-4608 (2008)

2. Molecular descriptors and their role intoSECRETed Biochemical space expansion.

As explained in Deliverable D3.2, a molecular descriptor is defined as the "final result of a logical and mathematical procedure, which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardised experiment"[2].

Quantitative structure-property(activity) relationship QSA(P)R models frequently use molecular descriptors. They are studied to predict the activity, toxicity, and other properties resulting from the chemical structures of compounds.

To determine the chemo-mathematical pattern defining the discriminative molecular features that distinguishing biosurfactants and siderophores collected in SECRETed database, a total of 1613 2D descriptors were calculated, using the Mordred[3] software. Molecular descriptors frequently used to describe chemical libraries include molecular weight (MW), number of rotatable bonds (RBs), hydrogen bond acceptors (HBAs), hydrogen bond donors (HBDs), topological polar surface area (TPSA), and the octanol/water partition coefficient (SlogP), which are properties commonly used as descriptors to represent lead-like, drug-like, or medicinally relevant chemical spaces[4]. After performing variable reduction of the descriptors, those showing constant values and those with at least one missing value were discarded. Finally, 800 descriptors were selected to be used in the modelling strategy.

2.1. QSA(P)R approach to classify surfactant molecules.

QSA(P)R models for surfactants have been developed to predict critical micelle concentration (cmc) values, which is a gauge for the surface activity of surfactants[5]. However, to date, no data on discriminative QSA(P)R models with the ability to classify a molecule as an surface active agent is found.

Thus, We started by studying the chemo-mathematical pattern related to SECRETed biosurfactants. The SECRETed database consists of 1309 biosurfactants, so 75% of the data (N=990) is used to train the model and the remaining 25% for its external validation (N=319). As a group of inactives, compounds from the KEGG database of compounds were chosen, choosing those that present a high degree of structural similarity with our biosurfactant database (Figure 1).

[2] Todeschini R, Consonni V Molecular descriptors for chemoinformatics. Wiley-VCH, Weinheim(2009)
[3] Moriwaki et al., J Cheminform. Feb 6;10(1):4.(2018)
[4] Lipinski Drug Discov. Today Technol., 1 pp. 337-341 (2004),
[5] Hu et al., Int J Mol Sci. 11(3): 1020–1047 (2010)

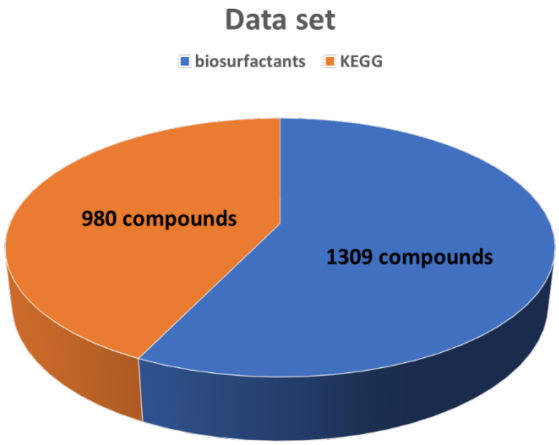

Figure 1 SECRETed Biosurfactant linear discriminative model database

Among the most interesting descriptors composing the QSA(P)R model based on linear discriminative model, we found:

BIC1: bond information content (neighborhood symmetry of 1-order). A higher value of this descriptor is associated with high molecular simetry and complexity in small molecules (Figure 2), which is not a characteristic of surfactant molecules.

Xch-7d descriptor: 7-ordered Chi chain weighted by sigma electrons. It indicates presence of cycles at topological distance 7, which is also an indication of low probability to be a surfactant (Figure 3).

Performed model had a high degree of accuracy to discriminate biosurfactant from similar molecules. Remarkably, the performance values were specially high over synthetic surfactants.

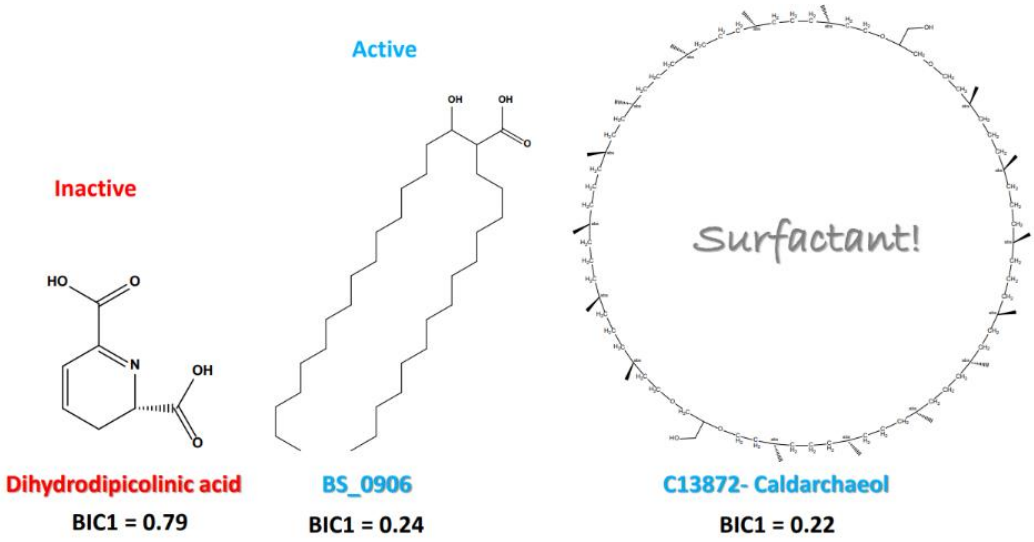

Figure 2 Examples on BIC1 degree of contribution in SECRETed Biosurfactant linear discriminative model.

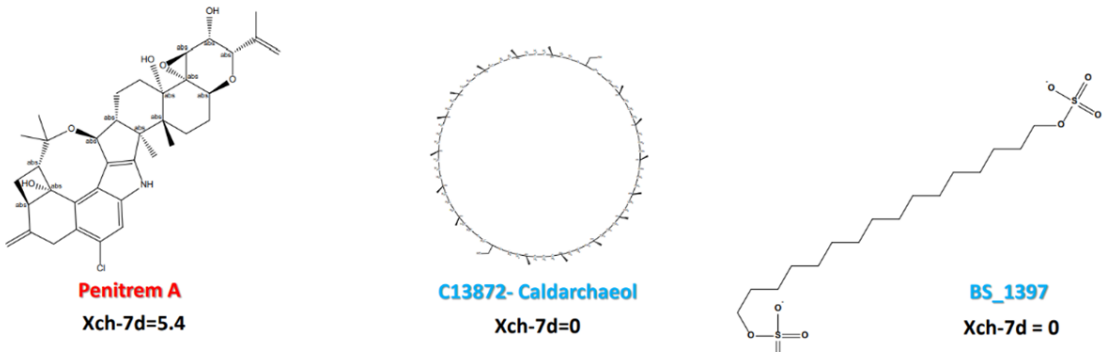

Figure 3 Examples on Xch-7d degree of contribution in SECRETed Biosurfactant linear discriminative model.

Table 1: SECRETed Biosurfactant linear discriminative model performance

Training Set	Percentage of Correct classification	Nº of Biosurfactant	Nª of Non-Biosurfactant
Biosurfactant	76%	752	238
Non-Biosurfactant	76%	152	534
Testing Set			
Biosurfactant	73%	233	86
Non-Biosurfactant	75%	70	204
External Test set		**Synthetic Surfactant**	**Non-Surfactant**
Synthetic Surfactants	99.11%	1116	10

2.2. QSA(P)R approach to classify chelating molecules.

On the other hand, we could not find any QSA(P)R model associated to siderophores in general. The only paper found was related to to identify the chemo-mathematical pattern related to the SECRETed siderophores. The SECRETed database consists of 491 siderophores, so 75% of the data is used to train the model and the remaining 25% for its external validation. As a group of inactives, compounds from the KEGG database of compounds are chosen, choosing those that present a high degree of structural similarity with our database of siderophores.

Among the most interesting descriptors composing the QSA(P)R model based on linear discriminative model, we found:

MATS2p: Moran autocorrelation-lag 2/weighted by atomic polarizabilities. The descriptor MATS2p is correlated with the polarizability of compounds.
EStatevsa1: MOE-type descriptors using Estate indices and surface area contributions.

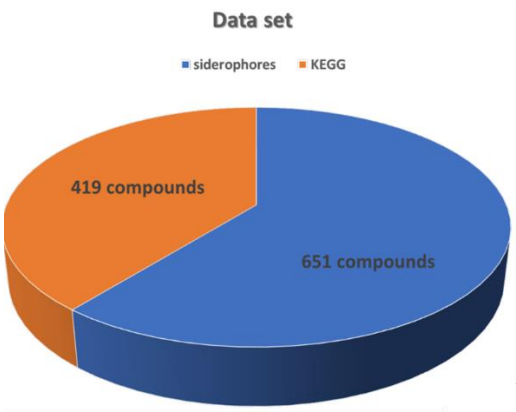

Figure 4 SECRETed Siderophore linear discriminative model database

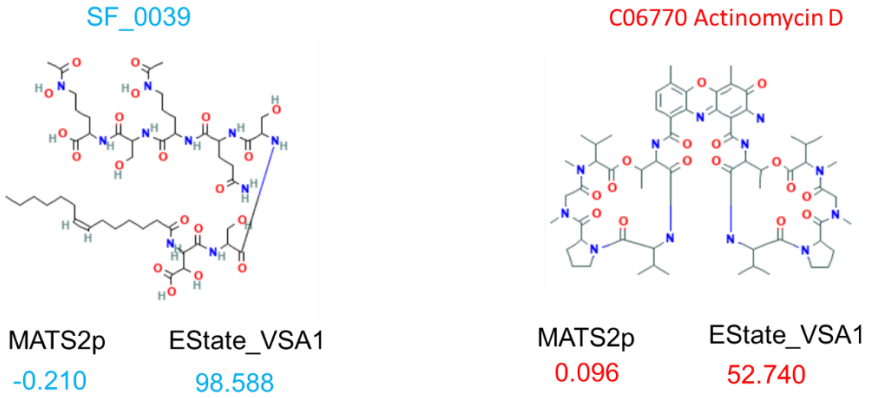

Figure 5 Examples of MATS2p and Estate_VSA1 degree of contribution in SECRETed Biosurfactant linear discriminative model.

Table 2: SECRETed Siderophore linear discriminative model performance

Training Set	Percentage of Correct classification	Nº of Siderophore	Nª of Non-Siderophore
Siderophore	84%	411	78
Non-Siderophore	81%	59	253
Testing Set			
Siderophore	83%	134	28
Non-Siderophore	79%	22	85

2.3. Assessing the dual role of Biosurfactant and siderophores.

Once the models proved their accuracy, we employed them to study which biosurfactants possessed chelating abilities and which siderophores surfactant abilities.

Overall, around 66 of a total of 652 siderophores showed surfactant abilities (of those, 8 had been reported said property). In the case of Biosurfactants, of 1310 compounds, 76 showed chelating abilities (of those, 61 were reported in literature).

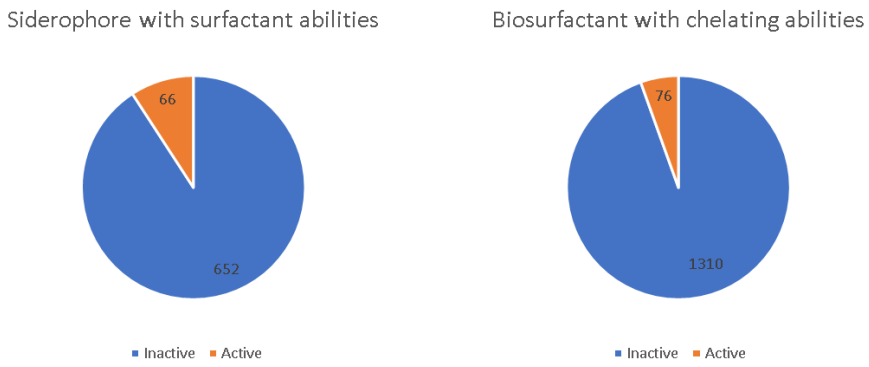

Figure 6 Dual activity in SECRETed siderophores and biosurfactants.

3. Defining the need to expand SECRETed biochemical space database. Informative clusters.

3.1. Molecular Families and informative clusters evolution along project performance.

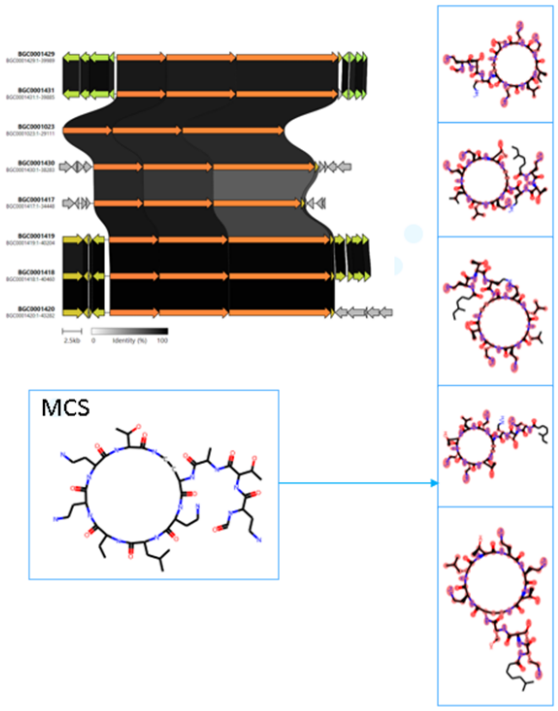

Figure 7 Core gene elements and Maximum common substructure

Deliverables D3.2 and D3.3. explained how EKUT contributed to SECRETed biochemical space with compounds with defined biosynthetic pathways and related Biosynthetic Gene Clusters (BGCs). Majoritarily, said compounds and BGCs were obtained from MiBig[6] database. Based on the high quality data this database provided, said compounds were identified and highlighted to serve as "anchoring points" for the retrosynthesis algorithms.

At that time, we found 240 compounds associated with 117 BGCs.

Specifically, fine-tuned retrosynthetic information coming from said anchoring points was expected to be extrapolated based on both molecular and genetic similarity of potential candidates. That only could be achieved by finding Molecular Families compatible with Gene Cluster Families. Then, BGCs core gene elements could be linked to the core substructure of molecular families. This ultimately helps to unveil the role of BGCs accessory genes and their degree of contribution into the construction of molecular variants (Figure 7).

It was then essential to properly cluster SECRETed biochemical space into MFs, assuring also that the maximum common substructure[7] (MCS) of a cluster is kept as representative as possible.

As explained in deliverable D3.2, MAP4 fingerprint[8] demonstrated to be the appropriate to said task. The suitability of MAP4 fingerprint comes from the fact that it combines the benefits of two complementary approaches: ECFP and Atom-pair. While the first performs well when describing the properties of small molecules, the second can capture the structural features of large molecules. His suitability was potentiated by being combined with a specialized silouete coefficient that maximizes the intra-cluster MCS average similarity.

[6] R. Terlouw et al., Nucleic Acids Research, Volume 51, Issue D1:D603–D610, (2023)
[7] Conte et al. Inter. J. Pattern Recognit. Artif. Intell.18:265–298. (2004).
[8] Capecchi *et al. J Cheminform 12:43 (2020)*

Following this approach, we could define 257 clusters and 298 isolated compounds (Figure 8). Moreover, from 257 MFs with more than two compounds, only 65 presented a compound informative enough to be extrapolated to the surrounding compounds.

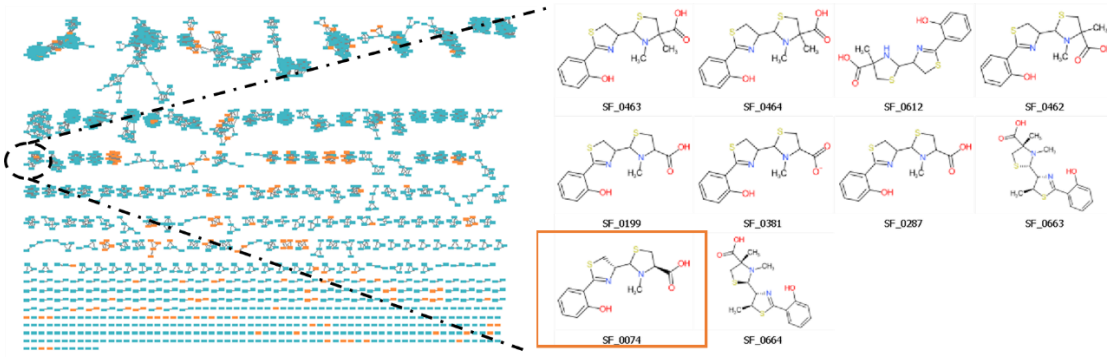

Figure 8 SECRETed biochemical space. Orange nodes represents compounds with associated BGC information.

As explained in deliverable D3.5, EKUT participated actively in expanding the number of compounds with retrosynthetic information. Their work was reflected in the updated MiBig 3.1 version, which provided to SECRETed with a total of 472 compounds associated to 365 BGCs (Figure 9).

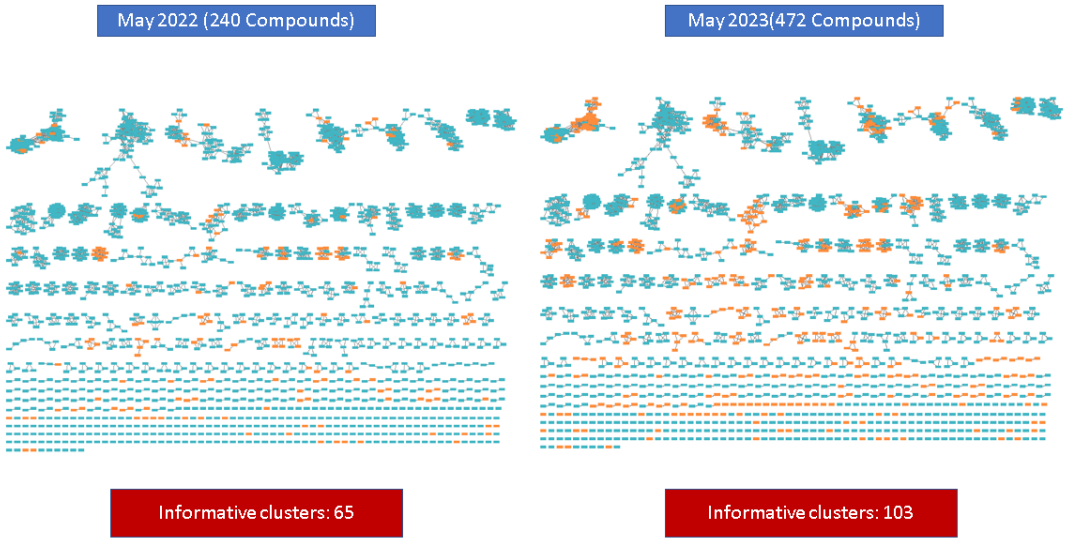

Figure 9 Effect on MiBig updates over SECRETed informative molecular clusters

Still, the increment of informative clusters could not couple the increment of informative compounds. Thus, an alternative approach was needed. Especially when considering that in each informative cluster, not all compounds had a proper compound-producer pair defined, not all the producers were defined at a strain level and not all the strains had their genomes available to inspect for their similar BGCs (see deliverables D3.3 and D3.5).

3.2. Applying the linear discriminative QSA(P)R model to incorporate more informative compounds from MiBig database.

Encouraged by the increment of informative compounds, we evaluated the possibility that some of the high level informative collected MiBig compounds could show surfactant and/or chelating abilities, being also similar enough to SECRETed compounds. To measure the molecular similarity we followed the previous approach using MAP4 fingerprint. Thus, this fingerprint was employed not only to elucidate the molecular similarity of SECRETed MFs, but also to calculate which other compounds from other databases could feed SECRETed biochemical space.

In that way, we could incorporate 59 informative compounds. Of said compounds, 19 formed part of the refined information from MiBig. Then, 40 new informative compounds (5 Biosurfactants and 35 Siderophores) were incorporated into SECRETed biochemical space, conferring 6 new informative clusters. These reference compounds are in addition to those already described in previous deliverables and to the expansion derived from version 3.1 of the MiBIG database explained in Deliverable 3.5 (Figure 10).

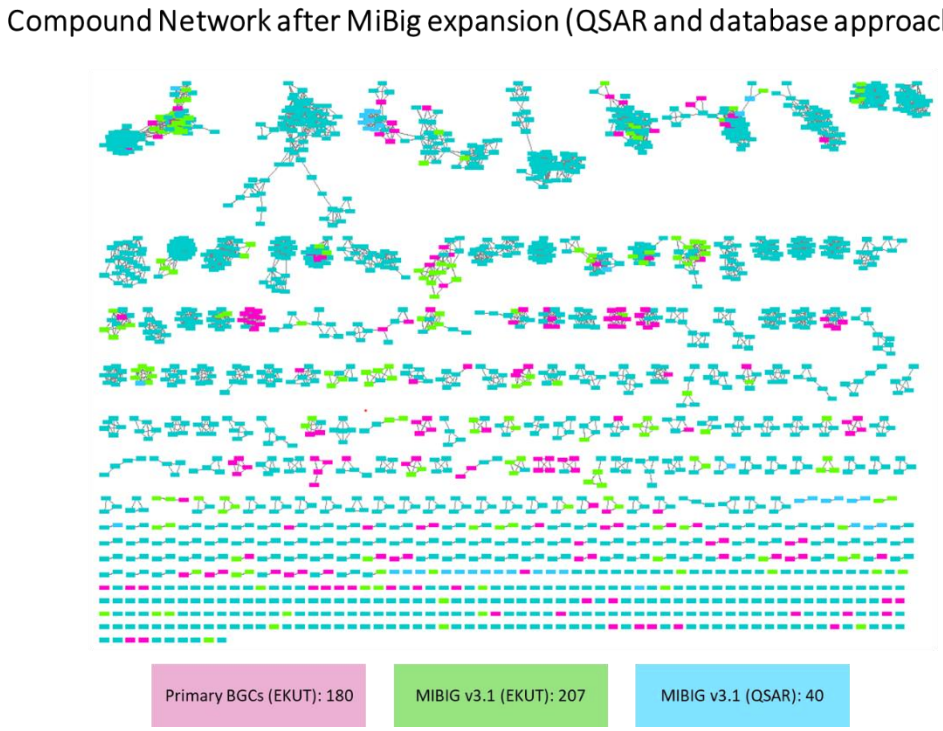

Figure 10. Effect on MiBig and QSAR updates over SECRETed informative molecular clusters.

4. SECRETed biochemical space Expansion.

As stated before, the LOTUS[9] initiative was explored to obtain structure-organism pairs that establish relationships between distinct molecular structures and the living organisms from which they were identified (More details in Deliverable D3.3). Based on their degree of completion, we decided it could be inspected to collect more chemical structures similar enough to already defined MFs, which, in a later stage, could also reflect the surfactant and chelating abilities.

Finally, only those containing sufficiently similar BGCs to the reference compound (see deliverable D3.5)would pass all filters and be considered for the consortium.

First, from the 276.518 natural products available (To date) in said resource, 2726 were sufficiently similar to be considered as MFs members (Figure 11)

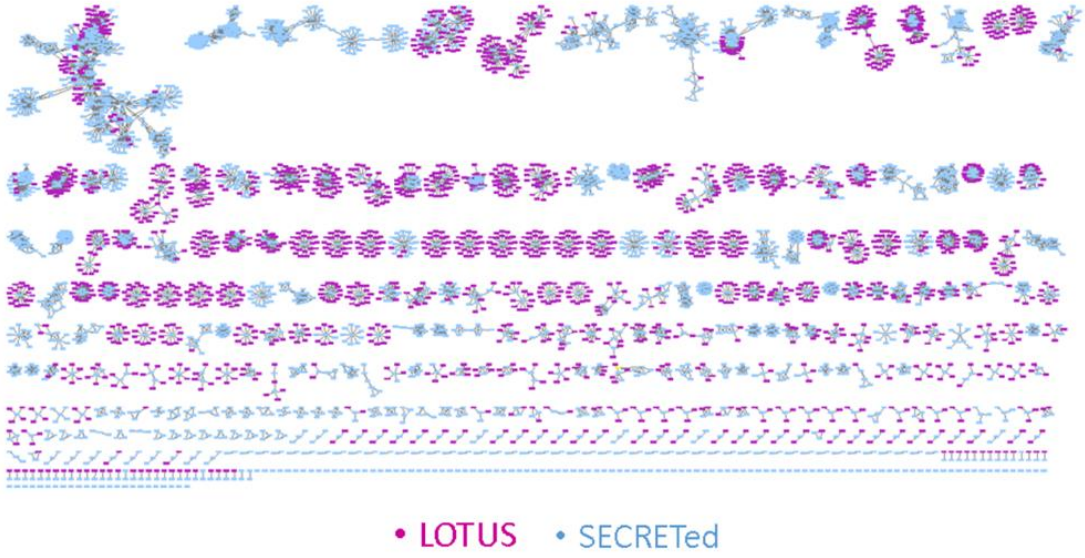

Figure 11 First step into SECRETed expansion: Molecular similarity.

[9] https://lotus.naturalproducts.net/

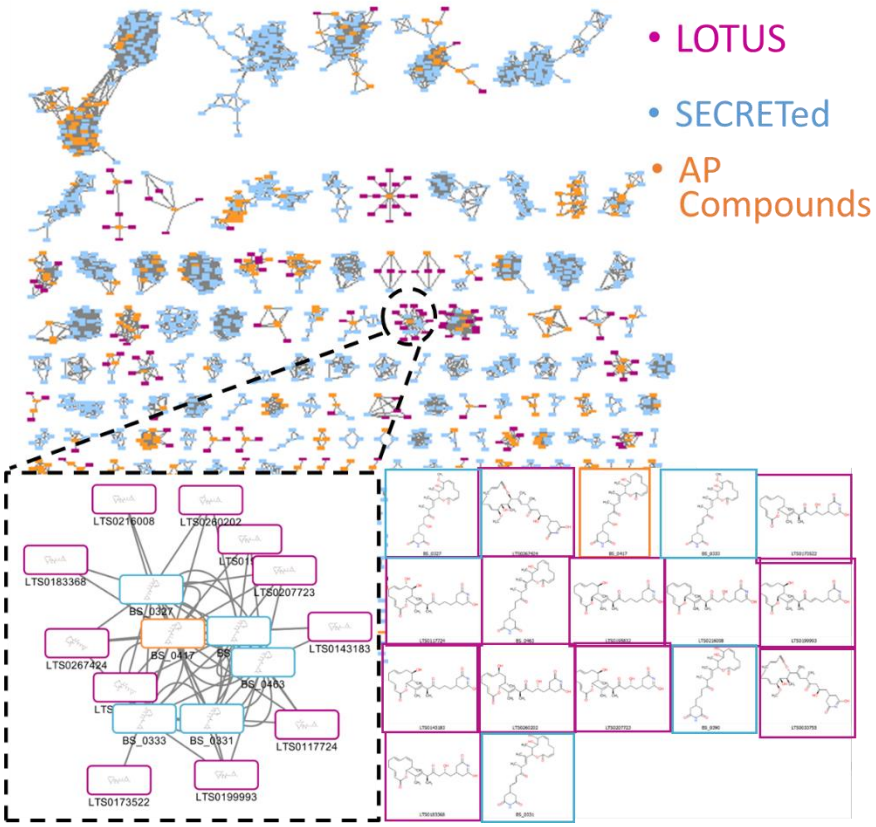

Figure 12 Filtered compounds after chemical expansion.

Of those, 658 compounds passed the QSP(A)R discriminative models. Finally, 154 compounds could provide more BGCs inferences (Figure 12, see Deliverable D3.5).

Finally. All collected compounds were incorporated into the integrative Management platform, in order to facilitate further analysis within SECRETed consortium (Figure 13)

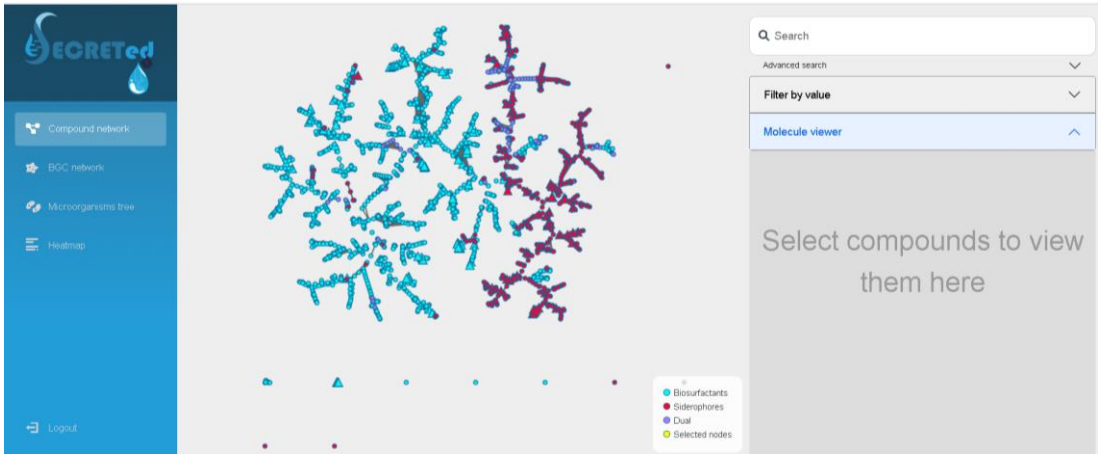

Figure 13 SECRETed updated Biochemical space.

5. SECRETed biochemical space decomposition.

The overarching goal of SECRETed is to unlock the potential of marine and extremophilic bacteria as a source of tailormade amphiphilic molecules by means of i) an extended bioactivity screening in 4 different aquatic microbial collections ii) sequenced based screening of novel marine and extremophilic genomes and metagenomes and iii) the formulation and validation of a 'mix and match' approach where modular genetic elements will be combined to get new tailor-made compounds based on their amphipathic nature.

However, although the chemical structures of a large number of SECRETed compounds are known (around 1900 molecular structures), the metabolic pathways that lead to their synthesis are not yet known (no more than 200 compounds with defined biosynthetic routes).

That is because the lack of knowledge associated to the accessory genes which are crucial to address the molecular variability of most of Natural Products.

Currently, the identification of relevant enzymes in a putative pathway is carried out by trial and error based on knowledge and experience; therefore, an information science-based approach is highly desirable.

A large variety of natural products are produced from a limited variety of starting materials[10]. So, as explained in deliverable D3.2, a good starting point to guide retrosynthesis efforts was to virtually decompose SECRETed biochemical space into their metabolic intermediates.

There are multiple advantages to this approach:

First, this approach determines the possible R-group decompositions of a given target molecule. The query molecule consists of the scaffold and ligand attachment points represented by R-groups. The query molecule can be input as SMARTS, or it can be searched through the implementation of tools such as the Maximum Common Substructure (MCSS)[11]. The advantage of R-group decomposition is that it can be complemented with biological information, which makes it possible to identify structural changes that modify biological activity or other properties.

Second, It also helps into the elucidation of biosynthetic pathways, since the identification of common metabolic intermediates points into common reactions (specially when related to the same MFs). Specially when MFs can also be structurally related molecules based on their mass spectral fragmentation patterns, whereas Gene cluster families are biosynthetic gene clusters that show similar gene cluster organization with a high degree of sequence similarity[12].

[10] Dewick . Medicinal natural products: a biosynthetic approach 3rd ed. Chichester, England: John Wiley & Sons; (2009)
[11] Y. Cao et al., Bioinformatics, 24 pp. i366-i374 (2008),
[12] Nguyen et al., Proc Natl Acad Sci U S A. 110(28):E2611-20. (2013).

Finally, the definition of the basic building blocks that composes complex molecular structures potins to the selection of common metabolic precursors. Classifying biosurfactants and siderophores from this perspective helps the selection of proper microbial host based on their ability to produce massively said precursors under industrial production conditions.

5.1. Rule-base virtual decomposition: Building block database refinement.

As explained in deliverable D3.2, we employed the Metabolic disassembler tool[13]. However, several improvements had to be implemented since the original code of metabolic disassembler was able to disassemble 344 out of 1912, a 17,99%.

This required improving the original code to reduce execution times and memory usage, to favour the computation of large datasets (as in our case, more than 1000 examples), and expand the original reference dataset (Table 3).

Table 3: Number of Metabolic intermediates participating in the decomposition database. Preferential biological units(PBU) and biosynthetic unit library (BUL).

Algorithm	PBU	BUL
Original	7	765
Improved	5006	20035

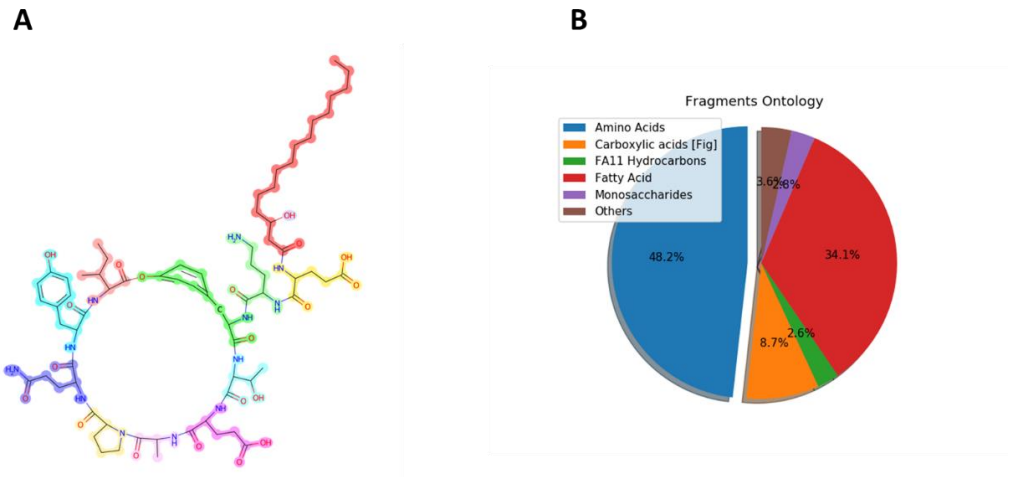

Figure 14. A) Example of molecule modular disassembling. B) SECRETed biochemical space building blocks according KEGG ontology.

[13]Amano et al., BMC Bioinformatics 20(1):728.(2019)

In this way, we could successfully disassemble an overall 86.3 % of a total of 1912 molecules. Figure 14 provides a graphical display of one example of molecule modular disassembling and summarises KEGG ontology terms that represent major building blocks in SECRETed biochemical space.

To encode for the definition of preferential biological units, we iteratively computed the pairwise match of every Kegg molecular structure. Thus, we could identify which metabolic intermediate "explains" the secondary metabolism registered in the KEGG database (Figure 15). In other words, we could automatically identify high "Out degree" molecules (preferential molecular constituents of other molecules) and high "In degree" molecules (Molecules composed of various intermediate molecules).

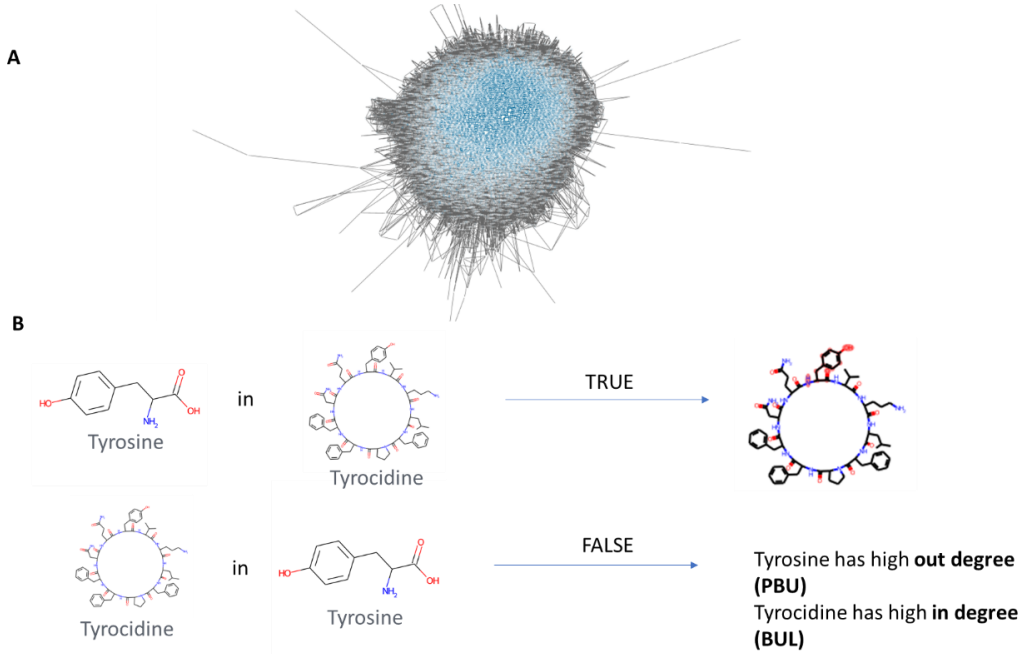

Figure 15. A) KEGG compositional network. B) "In degree" and "Out degree" concepts within KEGG compositional network.

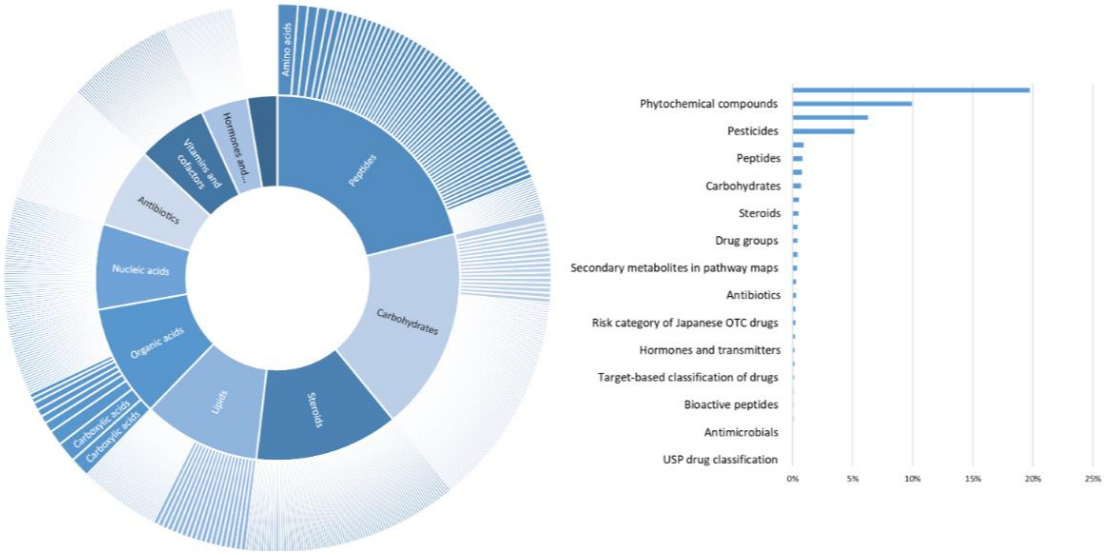

Figure 16. BUL KEGG Compound Brite classification.

Finally, as a final revision stage, we also collected the metabolic intermediates that participated most into KEGG reactions (Figure 17).

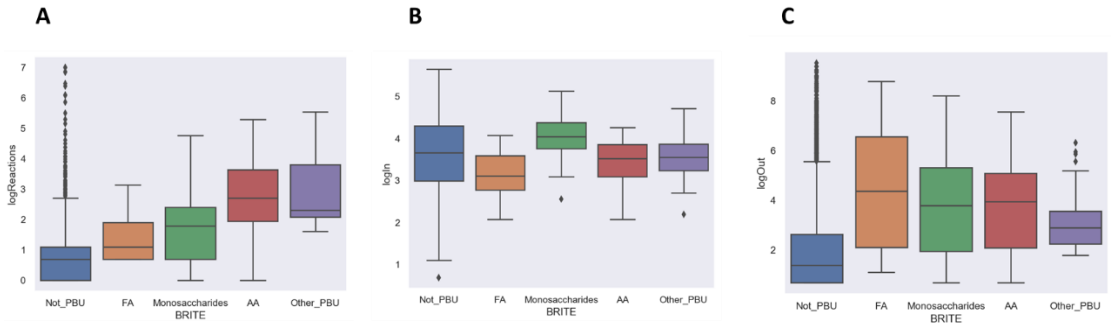

Figure 17. PBU KEGG Compound Brite classification of: A) Number of KEGG reactions (log scale). B) "In degree" KEGG compounds (log scale) C) "Out degree" KEGG compounds.

In this way, we could identify if there were molecules that could not been explained internally based solely on the KEGG database. Then, that implied completing our building block database with LOTUS and Lipidmaps constituents.

5.2. Rule-base virtual decomposition: GEASS algorithm.

Once our building blocks database was completed, we aimed at identifying not only the internal metabolic intermediates that were allocated internally into SECRETed compounds, but also the bonds that connects said intermediates.

Thus, we modified the already improved Metabolic disassembler algorithm into recognizing a set of generalized reaction rules. This would ultimately provide a fragmentation graph that

is “accepted” if generated fragments could be recognized into SECRETed building block database (Figure 18)

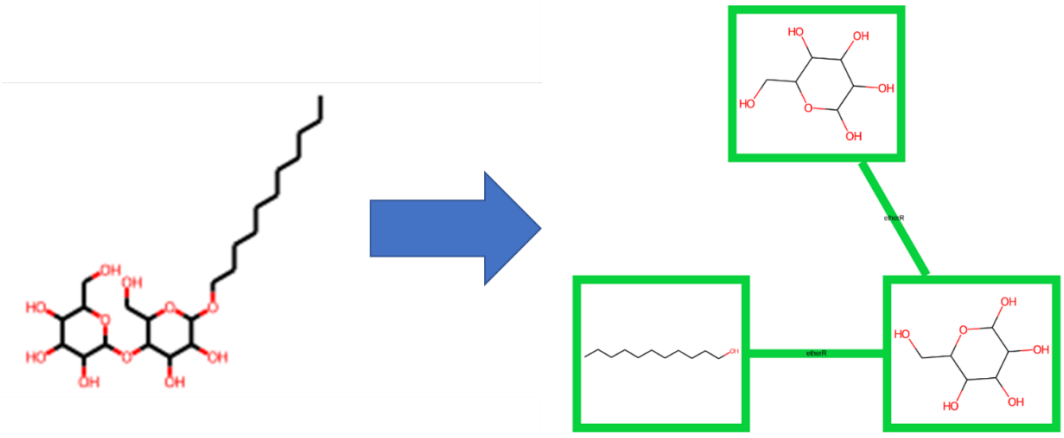

Figure 18. GEASS decomposition example were all fragments are aligned into SECRETed building block database.

More importantly, if a fragment is not aligned (red node, figure 19), the former metabolic disassembler algorithm acts, informing us which constituents conforms said unaligned fragment (connected by red edges, as they are not know yet). This ultimately can guide the development of new reaction rules.

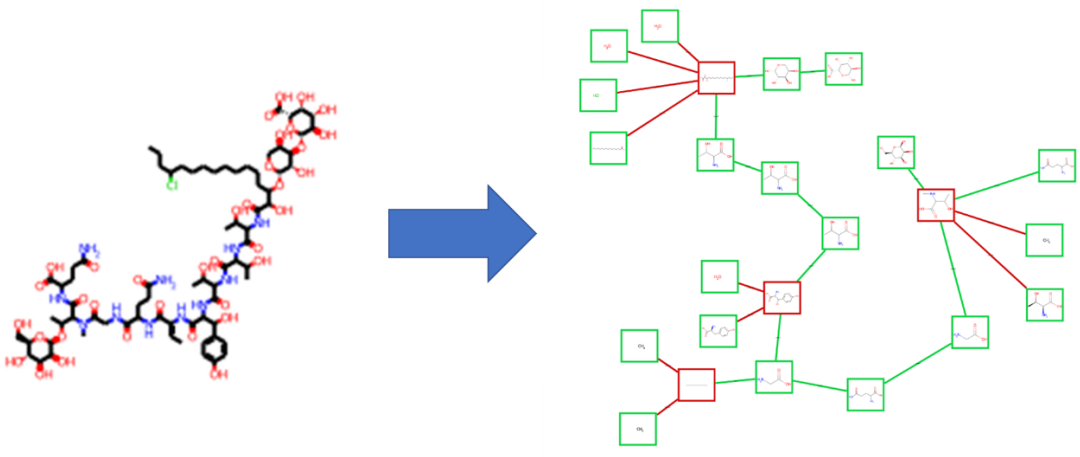

Figure 19. GEASS decomposition example were some fragments are not aligned into SECRETed building block database at first.

Altogether provide us a way to understand and classify SECRETed molecular families, by identifying common building blocks and reaction rules belonging to their participating compounds (Figure 20).

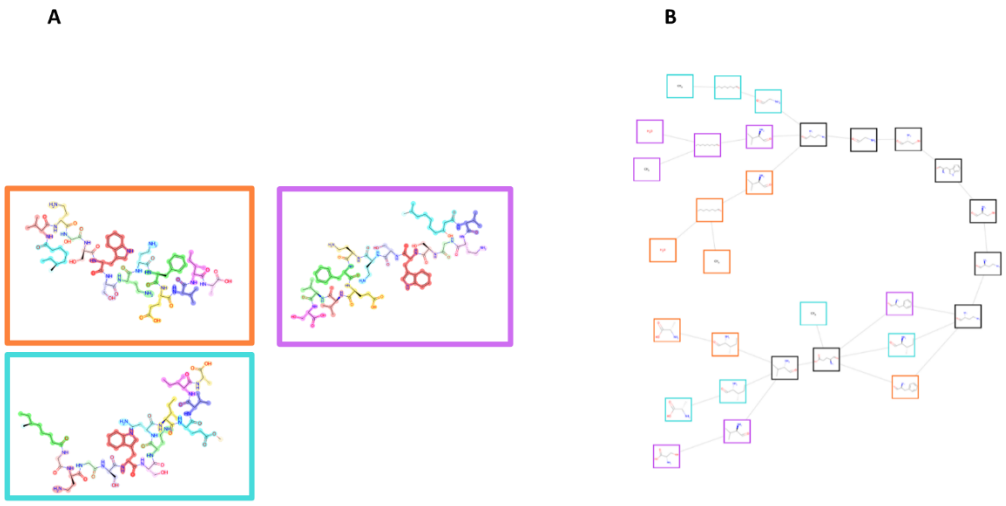

Figure 20. A) Example of compositional logic in three MFs members. B) Common decomposition graph for said MF members.

5.3. Rule-base virtual decomposition: SECRETed compositional statistics.

Once the entire SECRETed database has been metabolically decomposed, we could study the major occurrence of metabolic intermediates.

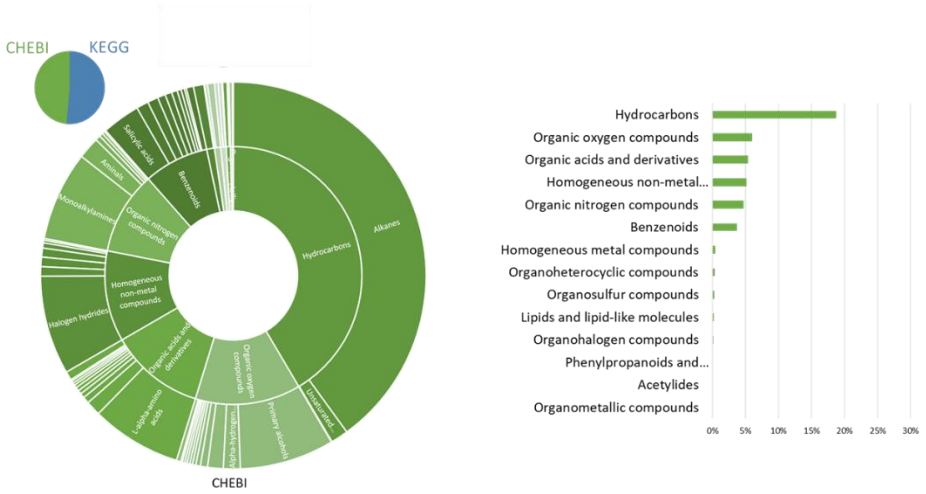

Figure 21. ChEBI classification of SECRETed molecular constituents. Categories have obtained their structural classification through the KEGG and CHEBI databases respectively. The fraction corresponding to each section shows their occurrences in percent within the SECRETED molecules.

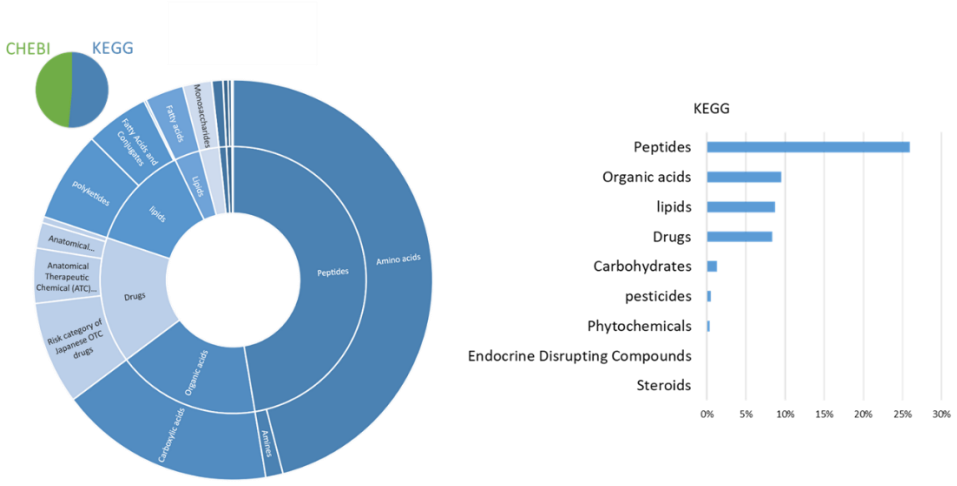

Figure 22. KEGG brite classification of SECRETed molecular constituents. Categories have obtained their structural classification through the KEGG and CHEBI databases respectively. The fraction corresponding to each section shows their occurrences in percent within the SECRETED molecules.

Overall, 398 compounds conformed the entire dataset. Since there are 193 compounds out of 398 compounds (after removing water and phosphate) that did not belong to a KEGG Brite category. it is necessary that we attach some structural information to them to calculate the overall compositional statistics for all of the compounds. To do so, weemployed the CHEBI ontology classification system (Figures 21 and 22).

6. Conclusions.

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.

Previous deliverable D3.2 drawed the baseline for the chemical expansion of SECRETed biochemical space.

In this deliverable we have utilized achieved knowledge to decipher the main features that provides the surfactant and chelating abilities. Thus, we have applied it to find not only which SECRETed compounds presented a dual activity but also which other compounds could be incorporated and facilitate our goals.

Those include Natural Products that were not tested for their surfactant and chelating activities. Among them, informative compounds from MiBig database helped into providing more informative clusters.

Finally, we have refined and improved a new set of tools that paves the way for the development of retrosynthetic algorithms.

Those identify not only which metabolites participate of the synthesis of SECRETed compounds, but also which rules connect them. Moreover, it also can help into elucidate which missing reaction rules are needed to address SECRETed molecular diversity.

In summary, this public report has summarised SECRETed efforts into the expansion of biosurfactants and siderophore databases.