microorganisms

MDPI

Article

PREGO: A Literature and Data-Mining Resource to Associate Microorganisms, Biological Processes, and Environment Types

Haris Zafeiropoulos [1,2,†], Savvas Paragkamian [1,2,†], Stelios Ninidakis [2], Georgios A. Pavlopoulos [3,4], Lars Juhl Jensen [5] and Evangelos Pafilis [2,*]

1 Department of Biology, University of Crete, Voutes University Campus, P.O. Box 2208, 70013 Heraklion, Crete, Greece; haris-zaf@hcmr.gr (H.Z.); s.paragkamian@hcmr.gr (S.P.)
2 Institute of Marine Biology, Biotechnology and Aquaculture (IMBBC), Hellenic Centre for Marine Research (HCMR), Former U.S. Base of Gournes, P.O. Box 2214, 71003 Heraklion, Crete, Greece; sninidakis@hcmr.gr
3 Institute for Fundamental Biomedical Research, Biomedical Sciences Research Center "Alexander Fleming", 16672 Vari, Greece; pavlopoulos@fleming.gr
4 Center for New Biotechnologies and Precision Medicine, School of Medicine, National and Kapodistrian University of Athens, 11527 Athens, Greece
5 Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, 2200 Copenhagen, Denmark; lars.juhl.jensen@cpr.ku.dk
* Correspondence: pafilis@hcmr.gr or prego@hcmr.gr; Tel.: +30-2810-337748
† These authors contributed equally to this work.

Citation: Zafeiropoulos, H.; Paragkamian, S.; Ninidakis, S.; Pavlopoulos, G.A.; Jensen, L.J.; Pafilis, E. PREGO: A Literature and Data-Mining Resource to Associate Microorganisms, Biological Processes, and Environment Types. *Microorganisms* **2022**, *10*, 293. https://doi.org/10.3390/microorganisms10020293

Academic Editors: Hera Karayanni, Katherine Maria Pappas, Chrysoula Tassou and Danae Venieri

Received: 27 December 2021
Accepted: 20 January 2022
Published: 26 January 2022

Abstract: To elucidate ecosystem functioning, it is fundamental to recognize what processes occur in which environments (where) and which microorganisms carry them out (who). Here, we present PREGO, a one-stop-shop knowledge base providing such associations. PREGO combines text mining and data integration techniques to mine such what-where-who associations from data and metadata scattered in the scientific literature and in public omics repositories. Microorganisms, biological processes, and environment types are identified and mapped to ontology terms from established community resources. Analyses of comentions in text and co-occurrences in metagenomics data/metadata are performed to extract associations and a level of confidence is assigned to each of them thanks to a scoring scheme. The PREGO knowledge base contains associations for 364,508 microbial taxa, 1090 environmental types, 15,091 biological processes, and 7971 molecular functions with a total of almost 58 million associations. These associations are available through a web portal, an Application Programming Interface (API), and bulk download. By exploring environments and/or processes associated with each other or with microbes, PREGO aims to assist researchers in design and interpretation of experiments and their results. To demonstrate PREGO's capabilities, a thorough presentation of its web interface is given along with a meta-analysis of experimental results from a lagoon-sediment study of sulfur-cycle related microbes.

Keywords: text mining; microbiome data; literature-derived associations; comention statistics; biological processes

1. Introduction

Microbes are omnipresent and impact global ecosystem functions [1] through their abundance [2], versatility [3], and interactions [4]. These facts have inspired microbiologists from diverse scientific fields to study their genotype and phenotype [5], their metabolism [6], and their interactions with the environment [7]. All this work has resulted in a wealth of knowledge available in the forms of literature and experimental data. Literature contains vast amounts of information in the free text form that overwhelms researchers. Advanced text mining methods [8] have been developed to assist this issue. Experimental data and their metadata require mining [9] as well for their integration, mostly through metagenomic mining from online repositories. Hence, the combination of this knowledge about microbial

life (who), their metabolic functions (what), and the environment they influence (where) is an important step to study ecosystem function [10].

High Throughput Sequencing (HTS) has turned the page on microbial ecology studies [11]. Over the past 20 years, both the taxonomic and the functional profiles of microbial communities from both local and large-scale regions (e.g., Tara Oceans [12], Earth Microbiome [13]) are being accumulated at a higher and higher rate. Extreme environments, i.e., areas with high salinity, low pH, etc., are being studied, providing us with unprecedented insight [14]. Both amplicon and shotgun metagenomics studies have played a crucial part in this development. Latest technological breakthroughs, such as Metagenome-Assembled Genomes (MAGs) and Single Amplified Genomes (SAGs), are enhancing the assessment of the taxonomic and functional repertoire of microbiomes even further. However, the mass use of these technologies and their consequent data have led to a number of needs and challenges, with metadata curation being among the most crucial ones.

Standards-promoting communities, like Genomic Standards Consortium (GSC) (https://gensc.org/, accessed on 24 December 2021), their efforts, like Minimum Information about any (x) Sequence (MIxS) [15], and projects endorsing those, like National Microbiome Data Collaborative (NMDC) [16,17], offer guidelines and best-practices to assist the annotation of microbial ecology samples. Controlled vocabularies and ontologies contribute to these efforts as they describe each subject area with formal terms [18]. Environment types, for example, are described by the Environment Ontology (ENVO) [19]. Other key biological aspects that have been captured include molecular functions (Gene Ontology Molecular Function (GOmf) [20,21], Enzyme Commission nomenclature [22], etc.), and the pathways carrying out different biological processes (GO Biological Process (GObp), MetaCyc [23], etc.). These knowledge structures, along with taxonomic centralized resources like the National Center for Biotechnology Information (NCBI) Taxonomy [24] and LPSN (List of Prokaryotic names with Standing in Nomenclature) [25], provide the means for a standardized representation of, for example, environments, process-oriented terms, and microbial taxa, respectively. Global-scale public resources (like MGnify [26], JGI/IMG [27], MG-RAST [28]) combine some of the aforementioned resources to support the collection, analysis, and distribution of multiple types of HTS data (e.g., amplicon, metagenomics, metatranscriptomics, etc.).

Besides the data and the analyses *per se*, the related scientific literature stores valuable information in billions of text lines. PubMed [24] and PubMed Central (PMC) [29] are gateways to relationships among microbes (*who*), the environments they live in (*where*) and their associated processes and functions (*what*) hidden in text [30]. Text mining (on both literature and metadata) can serve the extraction of these relationships. Named Entity Recognition (NER) can, for example, locate organism names [31], ENVO and GO terms [32] mentioned in text and map them to their corresponding identifiers. Association statistics, like co-mention analysis, can subsequently suggest ranked association among such entities [33,34]. The new era of omics has been interwoven with data integration [35] by bringing together scattered and fragmented pieces of information.

The time is ripe for tools that integrate all this knowledge and henceforth assist researchers to tackle major challenges like climate change [36], sustainability [37], and synthetic ecology [38]. Many resources have emerged in this realm [39], each one serving a specific purpose, such as BacDive [40]. BacDive is a large-scale curated database with prokaryotic information about phenotypic, morphological, and metabolic information. Other resources like Microbe Directory [41], Web of Microbes (WoM) [42], and Microbial Interaction Network Database (MIND) [43] focus on microbial environmental conditions, metabolite interactions with microbes and microbe-microbe interactions, respectively. In addition, taking advantage of aforementioned resources, novel pipelines, e.g., [44], are emerging with the aim to explore the network associations of who (microbial taxa) is performing what (microbial processes) and where (environments) directly using graph theory [45]. These analyses and resources are important because microbiologists can enrich

their data to explore hypotheses but also to identify potential gaps in knowledge regarding these associations [46].

Here, we present PREGO, a hypothesis generation web resource that is designed to be useful to microbiologists—in particular microbial ecologists and environmental microbiologists. Its specific aims include: (a) the gathering of source data, metadata, and literature followed by the extraction of microorganism, process, environment associations contained therein, (b) making such a mined knowledge base available to life sciences researchers via an easy to use and explore web portal. As such, PREGO can be useful also to system microbiologists and large-scale data analysts through bulk download and programming access. We document the principles, analysis methodology, and contents behind PREGO. Last but not least, we demonstrate PREGO's capabilities for researcher-support related to the above through a case study involving sulfate-reducing microorganisms.

2. Materials and Methods

PREGO is a resource designed to assist molecular ecologists in acquiring a single point overview of *what-where-who* process–environment–organism associations. The system is comprised of two main parts: (a) a server that periodically harvests data and extracts process-environment-organism associations from the scientific literature, environmental samples, and genome annotation sequences (Figure 1, step 1 to 5) and (b) a web-based interface as well as an Application Programming Interface (API) that provides users and programmers with a friendly way to extract and navigate across the process–environment–organism associations (Figure 1, step 6).

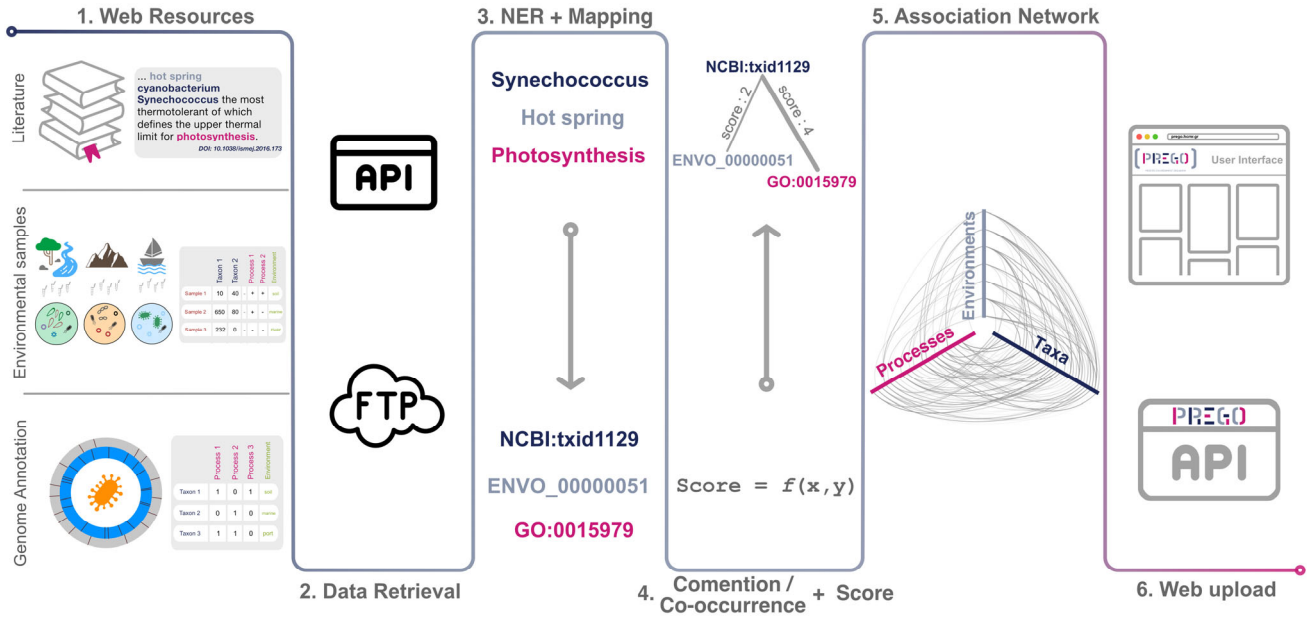

Figure 1. PREGO analysis methodology: PREGO periodically retrieves three distinct types of data from open access resources. Scientific text, environmental sample data, and genomic annotations are handled with respective methodologies in order to standardize their entities. Named Entity Recognition and Comention/Co-occurrence analysis is the common framework in order to build a weighted association network with nodes being the entity identifiers. Lastly, all these associations are available through a Web interface and an API. All these steps have been implemented in an autonomous way with regular cycles of updates (see Appendix B). Icons used from the Noun Project released under CC BY: Books by Shakeel Ch., Bacteria by Maxim Kulikov, ftp by DinosoftLab, Mountain by Diane, Ship on Sea by farra nugraha, River by Chanut is Industries.

2.1. Entity Types, Channels, and Associations

PREGO supports three entity types: *Process*, *Environment*, and *Organism*. For interoperability and consistency, an ontology or taxonomy is adopted for each type of entity. Processes are represented as Gene Ontology (GO) terms and are grouped either as Biological processes (GObp) or as Molecular functions (GOmf). In addition, Environments are represented by terms from the Environmental Ontology. Organisms are represented by the microbial NCBI Taxonomy Ids (Bacteria, Archaea, and unicellular eukaryotes). For the unicellular eukaryotes, a custom list was populated with the unicellular eukaryotic taxa using a curated list.

PREGO's contents are mainly divided into three distinct channels of information based on data origin and format (Figure 1, step 1). The *Literature* channel exploits scientific publications, i.e., abstracts and full text open access scientific publications (Table 1 and Section 2.2). Through the *Annotated Genomes and Isolates* channel, PREGO retrieves genome annotations and their accompanying metadata (Table 1 and Section 2.3). Finally, the *Environmental Samples* channel supports the integration of metagenomic analyses from both amplicon and shotgun studies. These include taxonomic and functional profiles along with their corresponding metadata (Table 1, more details in Section 2.4).

Table 1. Source databases that are integrated in PREGO and the number of items retrieved. The Open Access subset of PubMed Central has a Creative Commons license available for commercial and noncommercial use. JGI has its own license, the same applies for BioProject, MEDLINE®, and PubMed® as well.

Source	# Items	Data Type	Metadata	License
MEDLINE and PubMed	33 million	abstracts (text)	no	NLM Copyright
PubMed Central OA Subset	2.7 million	full article (text)	no	CC for Commercial, non-commercial
JGI IMG	9644	Isolates Annotated genomes	yes	JGI Data Policy
Struo	21,276	Annotated genomes	no	MIT, CC BY-SA 4.0
BioProject	18,752	Annotated genomes with abstracts (text)	yes	INSDC policy
MG-RAST	16,096	marker gene samples	yes	CC0
	7965	metagenomic samples	yes	CC0
MGnify	10,500	marker gene samples	yes	CC-BY, CC0

In cases in which the retrieved data and metadata are in text form, they are standardized to the aforementioned identifiers and taxonomies using Named Entity Recognition (NER) tools, namely the EXTRACT tagger [32,47]. In cases where data contain KEGG Orthology terms and/or Uniref identifiers, they are mapped to the respective GOmf using the mapping files available from the UniProt (see Appendix A). Associations are extracted after the mapping and standardization of the entities from each resource (Figure 1, step 3).

The association extraction pipeline is distinct for each channel and resource because of differences in the data type origin (see *prego_gathering_data* in the Availability of Supporting Source Codes section). By the means of navigation, the large number of associations returned to the user require a type of sorting; ideally, one that ranks the most trustworthy associations at the top. For those reasons, each channel of PREGO has a dedicated scoring scheme bounded within the (0, 5] space for consistency. In Appendix C, the scoring scheme of each channel is elaborated.

2.2. Text Mining of Scientific Literature

PREGO implements a text mining methodology to extract associations of the aforementioned entities from literature. The origin of text mining is a corpus that comprises scientific abstracts and full text articles from MEDLINE® and PubMed® and PubMed Central® Open Access Subset (PMC OA Subset) [48], respectively. The building and periodic update of the corpus is possible through the NCBI File Transfer Protocol (FTP) services. PREGO also has

a dedicated text-mining dictionary (see Availability of Supporting Source Codes section) that contains the entities ids, names, synonyms, and neglected words (stop words). PREGO dictionary incorporates the ORGANISMS [31] and ENVIRONMENTS [49] evaluated dictionaries as well as the experimental dictionaries of Gene Ontology Biological Process and Molecular Function.

Text mining is subsequently performed on the corpus using the dictionary through the EXTRACT tagger [32,47]. The tagger recognizes the entities of the dictionary in each abstract and full text article and assigns their co-mentions with a score. The score is sensitive to the text structural level of co-mention; higher to lower scoring when co-mention appears in the same sentence, then, in the same paragraph, and lastly in the same article. All these are integrated and normalized to a single score for each association, as implemented in STRING 9.1 [34] (see Appendix C for more details). In addition, the tagger extracts each mention in every article to provide the origin of each association it extracts.

2.3. Annotated Genomes and Isolates

Annotated genomes and isolates comprise the most trustworthy data in PREGO's knowledge base because they refer to a single species/strain and also have manually curated metadata. Among other data types, JGI-IMG [27,50] includes millions of genes from isolated genomes (*isolates*), SAGs and MAGs. Such annotations, along with their corresponding metadata, were collected using web-parsing technologies. Their metadata, describing their related environment/ecosystem, were tagged using the EXTRACT tagger to infer *organisms—environments* associations. The annotated KEGG terms were mapped to GOmf terms (see Appendix A). The GOmf terms were then used to extract *organisms—processes* associations.

The Struo pipeline [51] and its outcome when using the Genome Taxonomy DataBase (GTDB) (v.03-RS86) [52] was exploited to enrich *organisms—processes* associations. A set of 21,276 representative genomes, accompanied by UniRef50 annotations, was retrieved using the provided FTP server. The annotations were then mapped to GOmf terms (see Appendix A). Related GTDB genomes were mapped to their corresponding NCBI taxa (see Appendix A). All associations extracted from these resources were assigned arbitrarily a confidence level of four out of five. This score choice reflects the high-quality of these data and metadata.

In addition, BioProject data were integrated to PREGO using the NCBI FTP/e-utils services [48]. The BioProject ids that were integrated are the ones that have been assigned a PubMed abstract, a unicellular taxon, and Genome sequencing as data type. Then, using the text mining pipeline, associations were extracted connecting the assigned taxon with the rest of the entities that appear in the abstracts. This method resulted in associations that were assigned a confidence level of three (out of five) because of the combined method of curated data with text mining.

2.4. Environmental Samples

MGnify [26] and MG-RAST [28] repositories provide a great number of public metagenomic records. In the PREGO framework, both amplicon and shotgun metagenomic analyses are retrieved periodically along with their corresponding metadata. Data retrieval from these resources is possible from their Application Programming Interfaces (APIs). Marker gene analyses are retrieved and by measuring the co-occurrence of taxa present in the various environmental types (e.g., biomes, materials, features, etc.) *organisms—environments* associations are extracted. These associations emerge when a taxon is reported together with a certain environmental type, being mentioned in the metadata of a sample (*metadata based co-occurrence*). Similarly, analyses of metagenomic samples along with their corresponding metadata and annotations are also retrieved and *organisms—environments*, *organisms—processes* and *processes—environments* are extracted. The *processes—environments* associations are possible through co-occurrence of the functional annotations of metagenomes with the environmental metadata of the samples.

In all cases, the EXTRACT tagger is used on the microorganism names and the corresponding metadata of each sample to identify their identifiers (NCBI ids, ENVO terms, GOmf, GObp). All associations in this channel are scored based on the number of samples the entity of interest co-occurs with specific sample metadata (e.g., environmental type) or annotations (functional annotations or taxonomic annotations). The same scoring scheme was implemented across the channel resources (see Appendix C for more details), which ranks these associations with a value in the (0, 5] space.

2.5. Sequence Search

In the case of organisms, PREGO enables sequence-based queries, meaning a sequence (amplicon) can be used as an entry point like it was a taxon name. To this end, a custom database was built using a set of reference custom databases for four commonly used marker genes. For 16S and 18S rRNA, the SILVA database (v.138) [53] and the PR^2 database (version_4.14.0) [54,55] were used. Cytochrome c oxidase I (COI) [56] is another commonly used marker gene; for this reason, Midori 2 (version GB243) [57] was integrated in PREGO's custom database. Finally, for the Internal transcribed spacer (ITS), common in studies focusing on Fungi, the Unite (version 8.3, accessed 10.05.2021) [58] database was added.

2.6. Back-End Server and Front-End Implementation

PREGO is a multi-tier web-based application. It is hosted on a 64 GB RAM DELL R540, 20 core, Debian server. Custom API clients (written in Python) are responsible for retrieving the data and metadata from each source (Figure 1, step 2). These clients as well as the subsequent methodology (Figure 1, step 3 to 6) are updated in regular cycles using custom daemons (see Appendix B, Figure A1). The *mamba/blackmamba* web framework underlies communication to the Postgres association-holding database and the client-side communication. HTML 5, Ajax, JQuery, and custom Javascript enhance the user web experience. PREGO supports widely used browsers (e.g., Chrome, Firefox, Safari, Edge) in various operating systems, such as Windows 10, Linux (Ubuntu 18), and MacOS (10.12, 11).

3. Results

3.1. The PREGO Web Resource

Users can access the PREGO contents through its web User Interface (UI) (Figures 2 and 3), its Application Programming Interface (API) (Figure 4), or bulk download of all associations (Appendix D). The User Interface comes with two search fields: a plain text search and a sequence search (Figure 2a). The latter is used when the user wants to search for a taxon sequence (see Section 2.5 for supported sequence databases). The plain text search supports three types of entry points; the user can search for a taxon name, e.g., *Methanosarcina mazei*, an environmental type, e.g., *lagoon*, or a biological process e.g., *methanogenesis*. In all entry points, PREGO returns an overview page consisting of tabs with associations of the entity of interest with the entities of the two other types (Figure 2b–d) as well as Documents and Downloads tabs (Figure 2e,f).

Regarding the association tabs, when a taxon is used as a query, PREGO returns an overview page consisting of tabs for environments, biological processes, and molecular functions. When an environmental type is used as input, PREGO returns the organisms that have been found to be related to it, as well as the Biological Processes observed in the given environment. Lastly, if a biological process is under study, PREGO returns a tab with the organisms along with another tab with the Environments related to the process. Notably, only the associations with scores higher than 0.5 are presented in the web platform and are sorted in descending order based on their score. The score is visualized with a five-star system (see Appendix C for the scoring scheme).

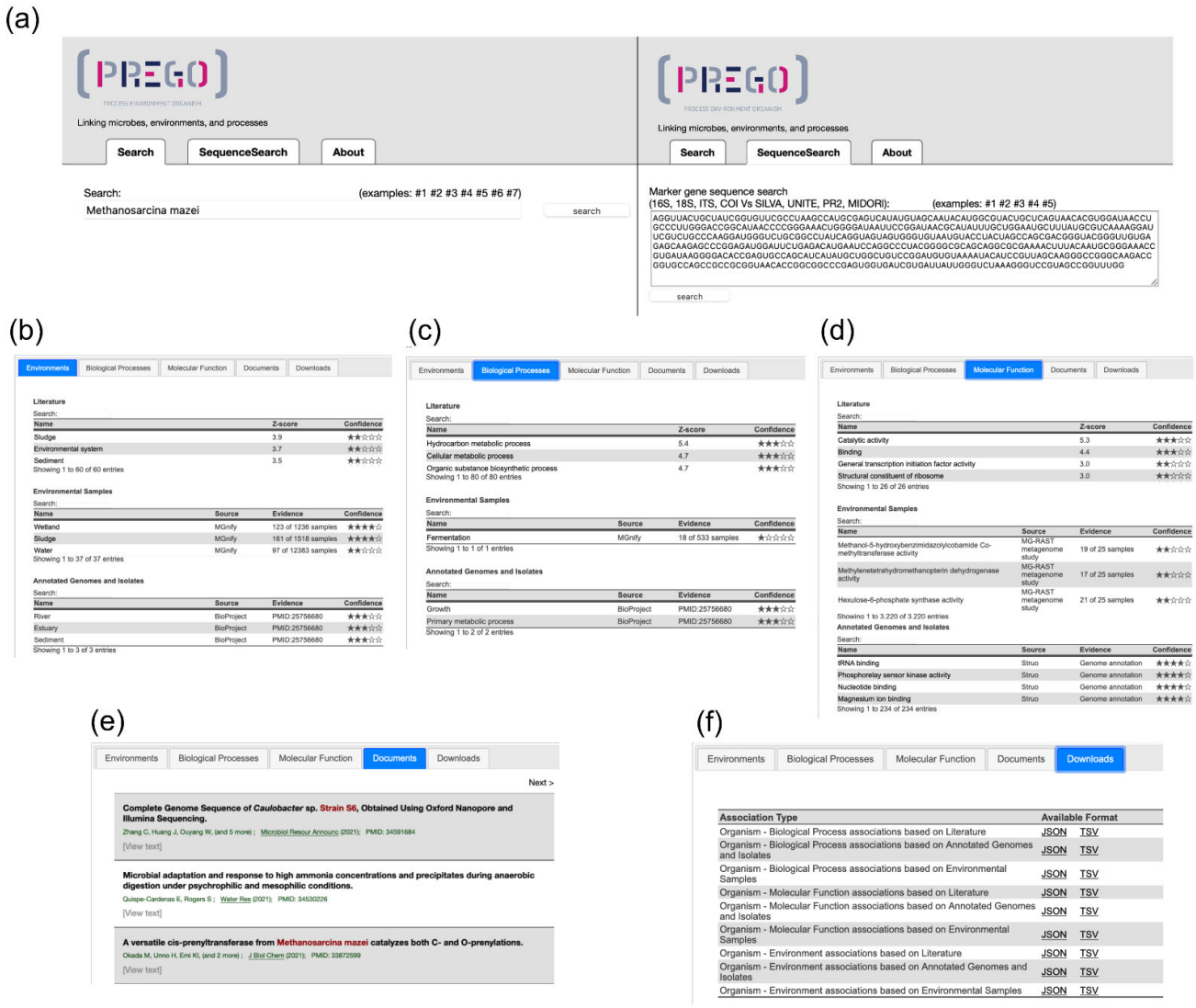

Figure 2. PREGO web user interface: (**a**) There are two search fields for user queries, plain text, and taxa sequences. (**b**–**d**) For the selected query, three association tabs are provided, each one presenting associations of the queried entity with the respective entities, Environments (**b**), Biological Process (**c**), and Molecular Function (**d**). Three channels of information distinguish the associations based on the original data. (**e**) The Documents tab presents the scientific articles that mention the queried entity with a highlighted color. (**f**) The Downloads tab provides the associations of each channel (when available) for download in JSON and TSV format.

Every association tab contains three tables with associations derived from the PREGO channels (see Section 2) along with their supported evidence. The user can both search and scroll through these tables, which makes knowledge extraction easier in cases where, for example, Isolate data contain hundreds of associations. In the *Literature* channel, each association is supported by the scientific articles with text-mining identified co-mentions. When a user clicks on an association, a popup window appears. This window displays abstracts or excerpts of full text with the associated entities highlighted (Figure 3a). Additionally, the *Environmental Samples* and *Genome annotations and Isolates* channels support evidence for each association by providing links to more detailed information. In the former channel, when the users click on an association, they are redirected to pertinent sample pages of MGnify (Figure 3b). Similarly, the latter redirects users to JGI and NCBI genomes when the associations originated from JGI—IMG and Struo, respectively (Figure 3c).

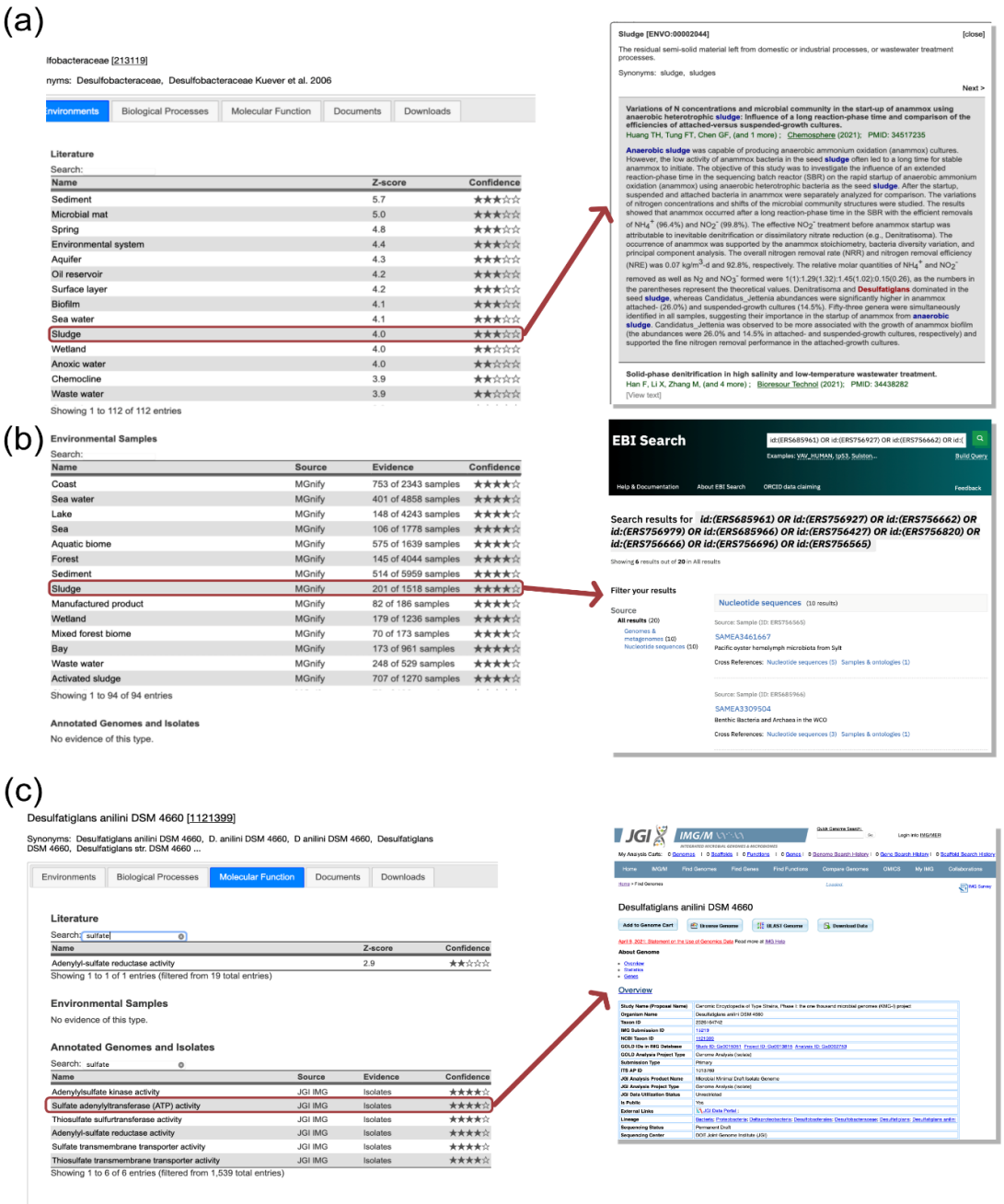

Figure 3. Each association is supported by original data: (**a**) *Literature* channel has a pop-up functionality that displays the scientific articles that support each specific association. The associated entities are highlighted in color. (**b**) *Environmental Samples* channel redirects to the samples that support the specific association (currently MGnify has this functionality). (**c**) *Annotated Genomes and Isolates* channel similarly redirects to the isolates ids that each association is based on (both Struo and JGI IMG are supported).

The *Documents* tab includes a list of scientific publications where the queried entity is mentioned. Through the *Downloads* tab, users are able to get all of the PREGO associations found for their query, per entity type (e.g., all the environments found related to an organism) and per channel (e.g., all the Environments found related to an organism through the *Literature* channel). This data retrieval functionality is also available via the PREGO API (syntax described in Figure 4). Finally, all PREGO associations are available for bulk download from each channel (see Table A2).

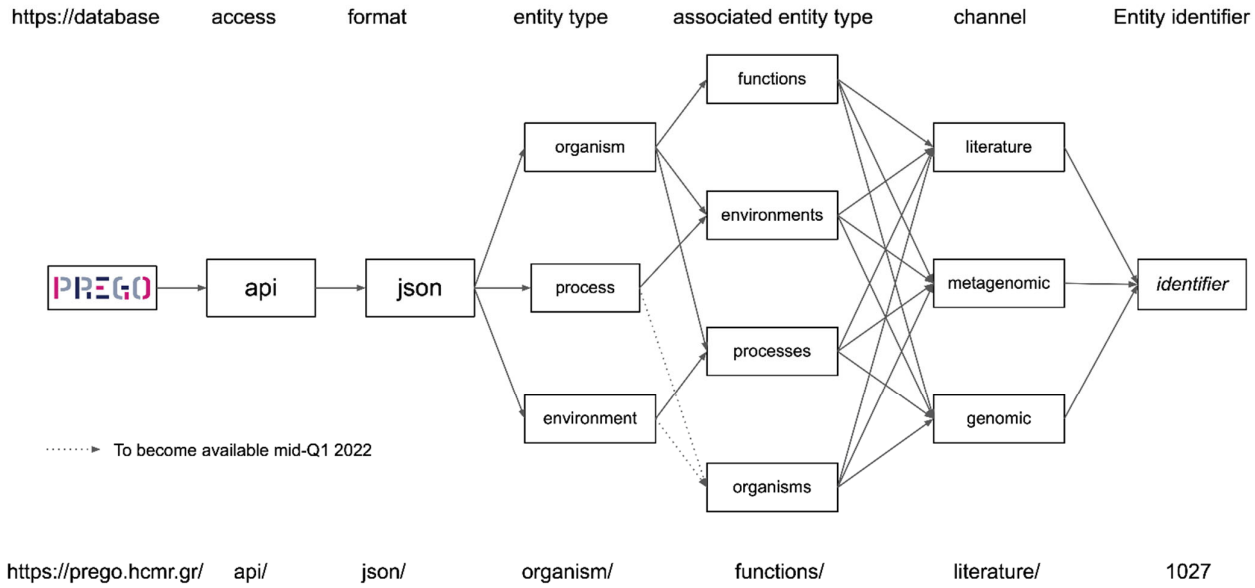

Figure 4. The PREGO API schema.

3.2. PREGO in Action

To demonstrate PREGO's potential, we present four different ways that PREGO can assist molecular ecologists. The demo focuses on the sulfate-reducing microorganisms (SRMs) as well as the processes and environments that relate to sulfate reduction. Through this demo, we highlight how the different channels may provide complementary insights regarding different taxonomic levels and different association types.

3.2.1. Which Environments Are Related to a Taxon?

Based on Pavloudi et al. (2017) [59], several bacterial and archaeal SRM were found in lagoonal sediments, after amplifying and sequencing the dissimilatory sulfite reductase β-subunit (dsrB). Using PREGO for the case of Desulfobacteraceae, the family in which the majority of the observed OTUs of the study belonged to, several environmental types similar to lagoons were retrieved from both the *Literature* and the *Environmental samples* channels (Figure 3a,b). Moreover, most of them had a high z-score, such as "*sediment*", "*sludge*", and "*activated sludge*". Several dissimilar environmental types were associated with Desulfobacteraceae, e.g., "oil reservoir" indicating them as potential environments where sulfate reduction takes place. However, the presence of taxa within that family in different environments, from "sea water" to "forest" and "Wastewater treatment plant", may suggest that this family has ubiquitous representatives in diverse conditions.

Searching for *Desulfatiglans anilini* (https://prego.hcmr.gr/example1, accessed on 24 December 2021) at the species level, the most abundant species in Pavloudi et al. (2017) and, for *Desulfatiglans anilini* DSM 4660 strain (https://prego.hcmr.gr/example2, accessed on 24 December 2021), PREGO provides associations with the "*Anaerobic sediment*", "*Marine oxygen minimum zone*", and "*Anaerobic digester sludge*" terms. These associations further corroborate the relationship between the species and sulfate reduction. More specifically, the "*sulfur spring*" ENVO term was retrieved from the *Environmental samples* channel as well.

3.2.2. Which Biological Processes and Molecular Functions Are Related to a Taxon?

According to Pavloudi et al. (2017), *Desulfatiglans anilini* plays an important role in sulfate reduction. The *Biological Processes* provided by PREGO's Literature channel are the GO terms "*Sulfate reduction*", "*Sulfide oxidation*", and "*Sulfide ion homeostasis*", which support this claim. In addition, the "*Denitrification pathway*" term was also retrieved. This is rather interesting as it is in line with what Pavloudi et al. (2017) discussed about the SRMs and their ability to use various electron acceptors, e.g., nitrate and nitrite.

Furthermore, PREGO's *Molecular Function* tab provides more insight on this example. Several GO terms related to sulfate reduction (e.g., terms related to "*sulfite reductase*") were associated with DSM 4660 strain and *Desulfatiglans anilini* species in multiple channels. Interestingly, in the case of the strain query, the *Annotated Genomes* channel returned many GO terms related to the nitrogen fixation, e.g., "*nitric oxide dioxygenase activity*".

3.2.3. Which Taxa Are Related to a Biological Process?

PREGO can be also used to report organisms that relate to a certain biological process. Searching for "*dissimilatory sulfate reduction*" associations with taxa (https://prego.hcmr.gr/example3, accessed on 24 December 2021) resulted in several taxa that were mentioned in the Pavloudi et al. (2017) study. For example, taxa such as Thermodesulfobacteria and *Thermodesulfovibrio* were found among the entries with the highest score (e.g.,) based on the *Literature* channel. The other two channels did not contain any associations. Using the "*Sulfate assimilation*" (https://prego.hcmr.gr/example4, accessed on 24 December 2021) as the biological process input, PREGO results showed several genera that were missing from PREGO results concerning the "*dissimilatory sulfate reduction*". Hence, manual search of GObp terms that describe the actual biological process of interest is more insightful.

3.2.4. Are There Any Associations between Environments and Biological Processes?

Are there other environmental types, except the lagoonal sediments, in which sulfate assimilation occurs? In that question, and in "*dissimilatory sulfate reduction*" (https://prego.hcmr.gr/example3, accessed on 24 December 2021) in particular, PREGO assigns the highest score to "*sediment*" while, among others, "*anoxic water*", "*oil reservoir*", "*mud volcano*", and "*basalt*" are potentially associated with environments related to sulfate reduction.

Inversely, PREGO is insightful about occurring processes in a specific environmental type. For example, searching for the biological processes that take place in "*basalt*" (https://prego.hcmr.gr/example5, accessed on 24 December 2021), processes like "*Nitrogen fixation*" and "*Reactive nitrogen species metabolic process*" stand out. However, sulfate reduction is not among the associations. However, when asking for "*Mafic lava*" (https://prego.hcmr.gr/example6, accessed on 24 December 2021), both the "*nitrogen fixation*" and "*Sulfur compound metabolic process*" terms are returned. This highlights the suggestions of Pavloudi et al. (2017), regarding the potential use of various electron acceptors from the different strains present in different environmental types.

3.3. PREGO Contents

PREGO contains the literature, environmental samples, and genome annotations of the resources shown in Table 1. The extracted contents of these resources have resulted to a knowledge base with ~364 K distinct taxonomic groups (out of a pool of ~620 K Bacteria, Archaea, and microbial eukaryotes, based on NCBI Taxonomy) from which ~258 K are at the species level (Table 2). These taxa are associated with ~1 K Environment Ontology terms, ~15 K GObp terms, and with ~7.9 K GOmf terms. Combining the above, PREGO maintains a knowledge base of entities and associations between them that form a multipartite network with entities as nodes and scored associations between them as weighted links.

Table 2. The entities of PREGO after the NER and mapping of every source: Counts of distinct entities of Taxa, Environments (ENVO terms), Biological Processes (Gene Ontology Biological process), and Molecular Function (Gene Ontology Molecular Function).

Channel	Source	Taxonomy		Environments	Biological Processes	Molecular Function
Literature	MEDLINE PubMed—PMC OA	Strains	8929	1077	15,079	7318
		Species	240,377			
		Total	342,506			
Environmental samples	MG-RAST amplicon	Strains	1392	162	-	-
		Species	4324			
		Total	5859			
	MG-RAST metagenome	Strains	2522	258	-	3839
		Species	4406			
		Total	7157			
	MGnify amplicon	Strains	2	216	11	-
		Species	1471			
		Total	2955			
Annotated Genomes and Isolates	JGI IMGisolates	Strains	2398	241	-	3670
		Species	11,203			
		Total	13,849			
	STRUO	Strains	6	-	-	2789
		Species	19,289			
		Total	19,325			
	BioProject	Strains	5754	309	626	-
		Species	3373			
		Total	9393			
Total	All	Strains	12,840	1090	15,091	7971
		Species	258,352			
		Total	364,508			

As shown in Figure 5, in its current version (December 2021), PREGO knowledge base covers 157 bacterial phyla (107 are Candidatus), 23 phyla from archaea (18 are Candidatus), and 22 unicellular eukaryotic phyla described in the NCBI Taxonomy database. The number of bacterial taxa present among the associations of each phylum ranges from the order of 10 s, as in the case of Candidatus Coatesbacteria, to hundreds of thousands, e.g., Actinobacteriae. The number of environmental types, found among the PREGO associations for each phylum, ranges from just a few to up to 1000. Similarly, the number of biological processes that have been related to the various phyla may range from less than a dozen, e.g., Yanofskybacteria to up to several thousands, e.g., Bacteroidetes. On the contrary, the number of molecular functions found to be related to taxa of each phylum is rather constant in all three domains.

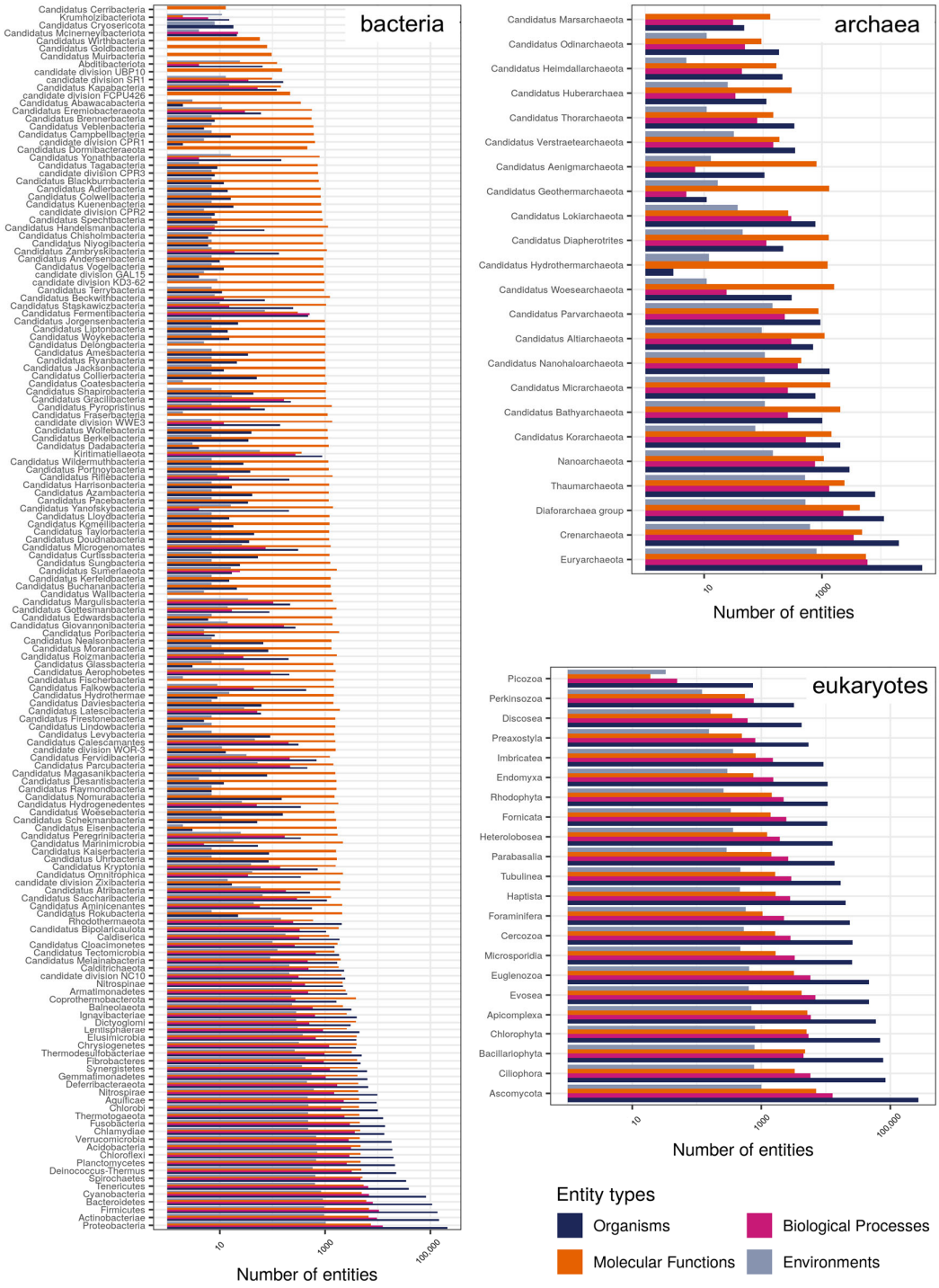

Figure 5. Summary of all the unique entities per phylum for each of the four entity types (in log10 scale) that appear in PREGO. Phyla are grouped based on their superkingdom (in log10 scale). Only phyla for which associations are available in the PREGO platform are mentioned.

4. Discussion

4.1. PREGO Contents

On its current version and according to the NCBI Taxonomy that it is based on, PREGO manages to cover a great range of microbial taxa, as most (if not all phyla) are present in the knowledge base (Figure 5). The different number of organisms' entities per phylum highlights the diverse number of the members of the various phyla. On the contrary, the similar number of molecular functions in all cases indicates the robustness of the main

metabolic processes required for life. With respect to biological processes, their number per phylum varies to some extent, especially for the case of Bacteria and Archaea. That could be observed as, in many cases, phyla that have been recently described using molecular techniques have not been studied extensively yet, e.g., Candidatus Delongbacteria. As expected, the number of environmental types that have been associated with members of each phylum varies, as a phylum may be universally present, while others could be strongly niche-specific (e.g., Hydrothermarchaeota).

Because of its three different channels, PREGO manages to extract associations both in the species and higher taxonomic levels. The *Isolates* channel supports explicit associations at the species level (Table 3 and Figure S3). Interestingly, the number of such genomes seems to have reached a plateau for now, as PREGO-like platforms include the same order of magnitude. The *Literature* channel, on the other hand, promotes the extraction of associations at higher taxonomic levels (Table 3 and Figure S1). This also applies to *environment—organisms* associations derived from the Environmental Samples channel (Table 3 and Figure S2). Associations regarding *biological processes*, though, are strongly enhanced by the *Literature* channel and the massive increase of literature.

Table 3. The associations between entities of PREGO after co-occurrence analysis: The supported entity types of associations are Environments—Biological Processes, Environments—Molecular Functions, Taxa—Environments, Taxa—Biological Processes, Taxa—Molecular Functions.

Channel	Source	Environments—Processes	Environments—Functions	Taxonomy	Taxa—Environments	Taxa—Processes	Taxa—Function
Literature	MEDLINE PubMed—PMC OA	883,997	422,579	Strains	69,968	590,630	384,079
				Species	778,877	3,501,635	1,961,920
				Total	1,669,608	7,969,310	4,613,827
Environmental samples	MG-RAST amplicon	-	-	Strains	13,645	-	-
				Species	39,007		
				Total	53,439		
	MG-RAST metagenome	-	620,846	Strains	262,106	-	8,626,328
				Species	103,913		10,715,548
				Total	372,301		19,950,096
	MGnify amplicon	-	-	Strains	18	-	
				Species	30,122	351	-
				Total	111,976	2097	
Annotated Genomes and Isolates	JGI IMGisolates	-	-	Strains	8229	-	3,461,693
				Species	42,141		13,216,559
				Total	50,888		16,821,850
	STRUO	-	-	Strains	-	-	1803
				Species			4,070,195
				Total			4,079,312
	BioProject	-	-	Strains	3263	7473	
				Species	4187	4294	
				Total	7641	12,169	
Total	All	883,997	1,043,425	Strains	357,229	598,103	12,473,903
				Species	998,247	3,506,280	29,964,222
				Total	2,265,853	7,983,576	45,465,085

Additionally, the text mining methodology of the *Literature* channel has retrieved most of the entities present in PREGO knowledge base (Table 2). A significant contribution to the taxa with associations is due to the PMC OA processing by the text mining pipeline of the Literature channel. This is in-line with reports in other applications of text mining when using full text articles [60]. However, the resulting associations are suggestive because of the text mining nature, and therefore subject for further review by the users.

4.2. Related Tools' Functionality and Content

There is an emerging *niche* for tools similar to PREGO to bring forward microbe associations and metadata. Table 4 summarizes the common and different features of BacDive, WoM, NMDC data portal, and PREGO. All of them commonly share the environmental associations and biological/metabolic processes with the microbes.

BacDive is a well-established platform with a focus on phenotype and cultivation information for about 100,000 prokaryotes, bacteria, and archaea. It has a high level of curation for most of its input types, like literature, internal databases, and personal collections. The NMDC data portal has published the scheme, the user interface, and a demonstrative collection of samples that will be populated later on. Standout features are the spatial visualization with coordinates and the detailed information of the samples, e.g., sequencing instruments and methodology. An alternative approach is facilitated by WoM, which aims to bind chemistry to microbes. An environment, in particular, is defined as the starting metabolite pool that is transformed by an organism. Another tool is The Microbe Directory that contains fully curated metadata for about 8000 microbes from all superkingdoms. This tool focuses on conditions of growth and on host taxa.

Complementary to these tools, PREGO contains associations of bacteria, archaea, and eukaryotes. Distinctive features are the associations of environments with processes/functions and the large-scale literature integration with text mining. Most importantly, most of the tools are complementary to each other with minimum overlap, an indication of the opportunities for further innovative synergies.

Table 4. Feature comparison among platforms that facilitate knowledge discovery and integration of microbial data.

Functionality	BacDive	Web of Microbes	NMDC	PREGO
manual curation	high	high	intermediate	low
literature integration	limited	no	no	yes
environment—taxa associations	yes	yes	yes	yes
environment—process/function associations	no	no	no	yes
process/function—taxa associations	yes	yes	yes	yes
phenotypic data	yes	no	no	no
data origin	original, integration	original	original, integration	integration
spatial coordinates	yes	no	yes	no
application programming interface	yes	no	yes	yes
bulk download	limited	yes	yes	yes

4.3. PREGO Next Steps

PREGO is a user-friendly association mining and sharing platform. Its modular web-architecture grants it the flexibility for further improvements in the aforementioned aspects, namely: source datasets, user interface, entity, and association scope expansion. Regarding datasets, additional data, such as transcriptomes from MGnify and other records annotated with metadata from studies in EuroPMC (https://ebi-metagenomics.github.io/blog/2021/11/17/Publication-Annotations/, accessed on 24 December 2021) [61], could be incorporated. Similarly, the NMDC data platform standards-compliant annotated records (https://data.microbiomedata.org/, accessed on 24 December 2021) could serve as an additional resource with its high-quality metadata [16,17]. Reciprocally, if requested, pertinent literature and association summaries could be programmatically offered to interested third parties.

Furthermore, the entity types supported by the PREGO system could be expanded. For example, GOmf terms could be upgraded as a search-entry point to the system. In addition, disease and tissue describing terms, already supported by the PREGO-underlying EXTRACT system [32], could enter the PREGO ecosystem of associated entities. From a statistics perspective, the calculation of a combined association score, when an association is reported by more than one channel of information, could be another feature to add.

The user interface can be enhanced to support multiple entity and/or sequence queries, instead of single ones. Sequences can be processed by taxonomy assignment pipelines (e.g., PEMA [62]) and be converted into searching PREGO for associations. In addition, network visualization tools, like Arena3D^{web} [63], could allow interactive browsing of associations through multi-layered graphs. Enrichment analyses, like those performed by OnTheFly$^{2.0}$ [64] or Flame [65], can be incorporated. Omics data analysis pipelines, like MiBiOmics [66], environment associations with sequences using SeqEnv [67] and biogeochemical associations with metagenomic data with DiTing [68] could be enabled by comparing the associations pertinent to different groups of entities. The computationally intensive tasks of multiple queries, taxonomy assignments to sequences and enrichment analysis could be offered by our in-house High Performance Computing facility (https://hpc.hcmr.gr/, accessed on 24 December 2021) [69] in synergy with pertinent Research Infrastructures like ELIXIR (https://elixir-europe.org, accessed on 24 December 2021) and LifeWatch ERIC (https://www.lifewatch.eu/, accessed on 24 December 2021).

Availability of Supporting Source Codes: The PREGO software modules are available under BSD 2-Clause "Simplified" License. Scripts, where additional libraries have been used, are subject to their individual licenses. More information on each module can be found as listed below:

- prego_gathering_data https://github.com/lab42open-team/prego_gathering_data
- prego_daemons https://github.com/lab42open-team/prego_daemons
- prego_mappings https://github.com/lab42open-team/prego_mappings
- prego_statistics https://github.com/lab42open-team/prego_statistics

Additional software and curated lists along with their individual license are:

- tagger https://github.com/larsjuhljensen/tagger, BSD 2-Clause "Simplified" License
- mamba https://github.com/larsjuhljensen/mamba, BSD 2-Clause "Simplified" License
- tagger dictionary https://download.jensenlab.org/ and there in: https://download.jensenlab.org/prego_dictionary.tar.gz, CC-BY 4.0 license.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/microorganisms10020293/s1, Figure S1: Summary of all the unique entities per phylum for each of the four entity types (in log10 scale) that appear in PREGO coming from the *Literature* channel. Figure S2: Summary of all the unique entities per phylum for each of the four entity types (in log10 scale) that appear in PREGO coming from the *Environmental samples* channel. Figure S3: Summary of all the unique entities per phylum for each of the four entity types (in log10 scale) that appear in PREGO coming from the *Annotated genomes and Isolates* channel.

Author Contributions: Conceptualization, E.P. and L.J.J.; methodology, H.Z., S.P., E.P. and L.J.J.; software, H.Z., S.P., S.N. and E.P.; validation, H.Z., S.P. and E.P.; formal analysis, H.Z., S.P. and L.J.J.; resources, S.N.; data curation, H.Z., S.P. and E.P.; writing—original draft preparation, H.Z., S.P., E.P. and G.A.P.; writing—review and editing, E.P., G.A.P. and L.J.J.; visualization, S.P. and H.Z.; supervision, E.P.; project administration, E.P.; funding acquisition, E.P. All authors have read and agreed to the published version of the manuscript.

Funding: This project was funded by the Hellenic Foundation for Research and Innovation (HFRI) & the General Secretariat for Research and Innovation (GSRI), under Grant No. 241, PREGO project. S.N. was supported by an EOSC-Life project (PID:14325). G.A.P. was supported by HFRI (1st call of research projects to support faculty members and researchers, Grant:1855-BOLOGNA) and the Marie Skłodowska-Curie Individual Fellowships—MSCA-IF-EF-CAR (Grant ID: 838018-H2020-MSCA-IF-2018). L.J.J. was supported by the Novo Nordisk Foundation [NNF14CC0001].

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The PREGO-mined association-datasets are available under a CC BY 4.0 license. They can be accessed: a. using the Web interface https://prego.hcmr.gr/, b. via the PREGO API (see Figure 4), and c. via bulk download files (See Appendix D).

Acknowledgments: This research was supported in part through computational resources provided by IMBBC (Institute of Marine Biology, Biotechnology and Aquaculture) of the HCMR (Hellenic Centre for Marine Research). Funding for establishing the IMBBC HPC has been received by the MARBIGEN (EU Regpot) project, LifeWatchGreece RI, and the CMBR (Centre for the study and sustainable exploitation of Marine Biological Resources) RI. We would like to thank our colleagues, Lucia Fanini, Christina Pavloudi, Ioulia Santi, George Tsamis, and Miguel Desmarais, for their contribution, great discussions, and their feedback throughout the development of PREGO. We also thank Antonis Potirakis and Dimitris Sidirokastritis for their sysadmin server support. We thank Michail Kouratoras for the design of the PREGO logo. Last but not least, we thank Manolis Badouvas for his administrative feedback and support.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

HTS	High Throughput Sequencing
MAGs	Metagenome-Assembled Genomes
SAGs	Single Amplified Genomes
GSC	Genomic Standards Consortium
NMDC	National Microbiome Data Collaborative
GO	Gene Ontology
GOmf	Gene Ontology (molecular function)
GObp	Gene Ontology (biological process)
ENVO	Environmental Ontology
NCBI	National Center for Biotechnology Information
LPSN	List of Prokaryotic names with Standing in Nomenclature
PMC	PubMed Central
NER	Named Entity Recognition
API	Application Programming Interface
FTP	File Transfer Protocol
GTDB	Genome Taxonomy DataBase
OTU	Operational Taxonomic Unit
SRMs	sulfate-reducing microorganisms

Appendix A

Mappings

PREGO produces entity identifiers either by Named Entity Recognition (NER) with the EXTRACT tagger or by mapping retrieved identifiers to the selected ones. PREGO adopted NCBI taxonomy identifiers for taxa, Environmental Ontology for environments and Gene Ontology as a structure knowledge scheme for Processes (GObp) and Molecular Functions (GOmfs). The latter was for reasons that are two-fold, first Gene Ontology has a Creative Commons Attribution 4.0 License and second there are many resources that have mapped their identifiers to Gene Ontology.

MG-RAST metagenomes and JGI/IMG isolates annotations come with KEGG orthology (KO) terms; Struo-oriented genome annotations, on the other hand, have Uniprot50 ids. The mapping from KO to GOmf and Uniprot50 to GOmf is implemented via UniProtKB mapping files of their FTP server (see idmapping.dat and idmapping_selected.tab files). By using the 3-column mapping file, the initial annotations were mapped to GOmf. As a complement, a list of metabolism-oriented KEGG ORTHOLOGY (KO) terms has been built (see *prego_mappings* in the Availability of Supporting Source Codes section).

Finally, as STRUO annotations refer to GTDB genomes, publicly available mappings (http://ftp.tue.mpg.de/ebio/projects/struo/GTDB_release89/metadata/, accessed on 24 December 2021) were used to link the genomes used with their corresponding NCBI Taxonomy entries.

Appendix B

Daemons

An important component PREGO approach (Figure A1) is the regular updates which keep PREGO in line with the literature and microbiology data advances. The updates are implemented with custom scripts called daemons that are executed regularly spanning from once a month up to six-month cycles. This variation occurs because of the API requirements of each web resource as well as the computational intensity of the association extraction from the retrieved data.

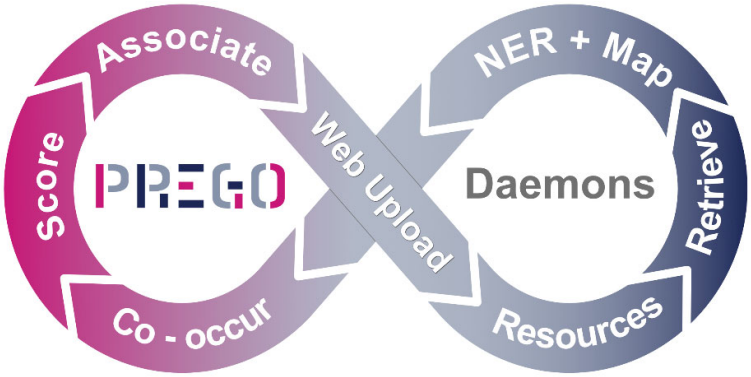

Figure A1. Software daemons perform all steps of the PREGO methodology in a continuous manner similar to the Continuous Development and Continuous Integration method.

Each Daemon is attached to a resource because its data retrieval methods (API, FTP) and following steps, shown in Figure A1, require special handling and multiple scripts (see *prego_daemons* in the Availability of Supporting Source Codes section).

Appendix C

Appendix C.1. Scoring

Scoring in PREGO is used to answer the questions:

- Which associations are more thrustworthy?
- Which associations are more relevant to the user's query?

Relevant, informative, and probable associations are presented to the user through the three channels that were discussed previously. Each channel has its own scoring scheme for the associations it contains and all of them are fit in the interval (0, 5] to maintain consistency. The values of the score are visually shown as stars. The *Genome Annotation and Isolates* channel has fixed values of scores depending on the resource because Genome Annotation is straightforward, and the microbe id is known a priori. On the other hand, *Environmental Samples* channel data are based on samples, which contain metagenomes and OTU tables. Thus, it has two levels of organization, microbes with metadata, and sample identifiers. Each association of two entities is scored based on the number of samples they co-occur. A *Literature* channel scoring scheme is based on the co-mention of a pair of entities in each document, paragraph, and sentence. The differences in the nature of data require different scoring schemes in these channels.

The contingency table (Table A1) of two random variables, X and Y are the starting point for the calculation of scores. The term X = 1 might be a specific NCBI id and Y = 1 a ENVO term. The $c_{1,1}$ is the number of instances that two terms of X = 1 and Y = 1 are co-occurring, i.e., the joint frequency. The marginals are the $c_{1,.}$ and $c_{.,1}$ for x and y, respectively,

which are the backgrounds for each entity type. Different handling of these frequencies leads to different measures. There is not a perfect scoring scheme, just the one that works best on a particular instance. Consequently, scoring attributes require testing different measures and their parameters.

Table A1. Contingency table of co-occurrences between entities X = x and Y = y. This is the basic structure for all scoring schemes. $c_{x,y}$ is the count of the co-occurrence of these entities. $c_{x,.}$ is the count of the x with all the entities of Y type (e.g., Molecular function). Conversely, $c_{.,y}$ is the count of y with all the entities of X type (e.g., taxonomy).

		Y = y		
		Yes	No	Total
X = x	Yes	$c_{x,y}$	$c_{x,0}$	$c_{x,.}$
	No	$c_{0,y}$	$c_{0,0}$	$c_{0,.}$
	Total	$c_{.,y}$	$c_{.,0}$	$c_{.,.}$

Appendix C.2. Literature Channel

Scoring in the Literature channel is implemented as in STRING 9.1 [34] and COMPARTMENTS [70], where the text mining method uses a three-step scoring scheme. First, for each co-mention/co-occurrence between entities (e.g., *Methanosarcina mazei* with Sulfur carrier activity), a weighted count is calculated because of the complexity of the text.

$$C_{x,y} = \sum_{k=1}^{n} w_d \delta_{dk}(\mathrm{x},\mathrm{y}) + w_p \delta_{p,k}(\mathrm{x},\mathrm{y}) + w_s \delta_{sk}(\mathrm{x},\mathrm{y}) \quad \text{(A1)}$$

Different weights are used for each part of the document (k) for which both entities have been co-mentioned, $w_d = 1$ for the weight for the whole document level, $w_p = 2$ for the weight of the paragraph level, and $w_s = 0.2$ for the same sentence weight. Additionally, the delta functions are one (Equation (A1)) in cases the co-mention exists, zero otherwise. Thus, the weighted count becomes higher as the entities are mentioned in the same paragraph and even higher when in the same sentence.

Subsequently, the co-occurrence score is calculated as follows:

$$score_{x,y} = c_{x,y}^{a} \left(\frac{c_{x,y} c_{.,.}}{c_{x,.} c_{.,y}} \right)^{1-a} \quad \text{(A2)}$$

where $a = 0.6$ is a weighting factor, and the $c_{x,.}$, $c_{.,.}$, $c_{.,y}$ are the weighted counts as shown in Table A1 estimated using the same Equation (A1). This value of the weighting factor has been chosen because it has been optimized and benchmarked in various applications of text mining [34,70,71]. The value of Equation (A2) is sensitive to the increasing size of the number of documents (MEDLINE PubMed—PMC OA). Therefore, to obtain a more robust measure, the value of the score is transformed to z-score. This transformation is elaborated in detail in the COMPARTMENTS resource [70]. Finally, the confidence score is the z-score divided by two. Cases in which the scores exceed the (0,4] interval are capped to a maximum of 4 to reflect the uncertainty of the text mining pipeline.

Appendix C.3. Environmental Samples Channel

Data from environmental samples are OTU tables and metagenomes. Thus, for each entity x, the number of samples is calculated as the background and a number of samples of the associated entity (metadata background) $c_{.,y}$ (see Table A1). Each association between entities x, y has a number of samples, $c_{x,y}$ that they co-occur. Note that each resource is independent and the scoring scheme is applied to its entities. This means that the same

association can appear in multiple resources with different scores. The score is calculated with the following formula:

$$score_{x,y} = 2 \times \frac{\sqrt{c_{x,y}}}{c_{.,y}^{0.1}} \quad \text{(A3)}$$

This score is asymmetric because the denominator is the marginal of the associated entity. Thus, the score decreases as the marginal of y is increasing, i.e., the number of samples that y is found. On the other hand, it promotes associations in which the number of samples of the association are similar to the marginal of y. The exponents on the numerator and denominator equal to 0.5 and to 0.1, respectively, in order to reduce the rapid increase of score. Lastly, the value of the score is capped in the range (0, 4].

Appendix D

Bulk Download

Users can also download programmatically all associations per channel through the links that are shown in Table A2. The data are compressed to reduce the download size and md5sum files are provided as well for a sanity check of each download.

Table A2. Bulk download links and md5sum files.

Channel	Link	md5sum	Size (in GB, Zipped)
Literature	https://prego.hcmr.gr/download/literature.tar.gz	literature.tar.gz.md5	5.4
Environmental samples	https://prego.hcmr.gr/download/environmental_samples.tar.gz	environmental_samples.tar.gz.md5	0.69
Annotated genomes and isolates	https://prego.hcmr.gr/download/annotated_genomes_isolates.tar.gz	annotated_genomes_isolates.tar.gz.md5	0.26

References

1. Falkowski, P.G.; Fenchel, T.; Delong, E.F. The Microbial Engines That Drive Earth's Biogeochemical Cycles. *Science* **2008**, *320*, 1034–1039. [CrossRef] [PubMed]
2. Bar-On, Y.M.; Phillips, R.; Milo, R. The Biomass Distribution on Earth. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, 6506–6511. [CrossRef] [PubMed]
3. Delgado-Baquerizo, M.; Maestre, F.T.; Reich, P.B.; Jeffries, T.C.; Gaitan, J.J.; Encinar, D.; Berdugo, M.; Campbell, C.D.; Singh, B.K. Microbial Diversity Drives Multifunctionality in Terrestrial Ecosystems. *Nat. Commun.* **2016**, *7*, 10541. [CrossRef] [PubMed]
4. Röttjers, L.; Faust, K. From Hairballs to Hypotheses–Biological Insights from Microbial Networks. *FEMS Microbiol. Rev.* **2018**, *42*, 761–780. [CrossRef]
5. Morris, A.; Meyer, K.; Bohannan, B. Linking Microbial Communities to Ecosystem Functions: What We Can Learn from Genotype–Phenotype Mapping in Organisms. *Philos. Trans. R. Soc. B Biol. Sci.* **2020**, *375*, 20190244. [CrossRef]
6. Biggs, M.B.; Medlock, G.L.; Kolling, G.L.; Papin, J.A. Metabolic Network Modeling of Microbial Communities. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **2015**, *7*, 317–334. [CrossRef]
7. Hall, E.K.; Bernhardt, E.S.; Bier, R.L.; Bradford, M.A.; Boot, C.M.; Cotner, J.B.; del Giorgio, P.A.; Evans, S.E.; Graham, E.B.; Jones, S.E.; et al. Understanding How Microbiomes Influence the Systems They Inhabit. *Nat. Microbiol.* **2018**, *3*, 977–982. [CrossRef]
8. Jensen, L.J.; Saric, J.; Bork, P. Literature Mining for the Biologist: From Information Retrieval to Biological Discovery. *Nat. Rev. Genet.* **2006**, *7*, 119–129. [CrossRef]
9. Delmont, T.O.; Malandain, C.; Prestat, E.; Larose, C.; Monier, J.-M.; Simonet, P.; Vogel, T.M. Metagenomic Mining for Microbiologists. *ISME J.* **2011**, *5*, 1837–1843. [CrossRef]
10. Raes, J.; Bork, P. Molecular Eco-Systems Biology: Towards an Understanding of Community Function. *Nat. Rev. Microbiol.* **2008**, *6*, 693–699. [CrossRef]
11. Nilsson, R.H.; Anslan, S.; Bahram, M.; Wurzbacher, C.; Baldrian, P.; Tedersoo, L. Mycobiome Diversity: High-Throughput Sequencing and Identification of Fungi. *Nat. Rev. Microbiol.* **2019**, *17*, 95–109. [CrossRef] [PubMed]
12. Pesant, S.; Not, F.; Picheral, M.; Kandels-Lewis, S.; Le Bescot, N.; Gorsky, G.; Iudicone, D.; Karsenti, E.; Speich, S.; Troublé, R.; et al. Open Science Resources for the Discovery and Analysis of Tara Oceans Data. *Sci. Data* **2015**, *2*, 150023. [CrossRef] [PubMed]

13. Gilbert, J.A.; Jansson, J.K.; Knight, R. The Earth Microbiome project: Successes and aspirations. *BMC Biol.* **2014**, *12*, 69. [CrossRef] [PubMed]
14. Shu, W.-S.; Huang, L.-N. Microbial Diversity in Extreme Environments. *Nat. Rev. Microbiol.* **2021**, 1–17. [CrossRef]
15. Yilmaz, P.; Kottmann, R.; Field, D.; Knight, R.; Cole, J.R.; Amaral-Zettler, L.; Gilbert, J.A.; Karsch-Mizrachi, I.; Johnston, A.; Cochrane, G.; et al. Minimum Information about a Marker Gene Sequence (MIMARKS) and Minimum Information about Any (x) Sequence (MIxS) Specifications. *Nat. Biotechnol.* **2011**, *29*, 415–420. [CrossRef]
16. Wood-Charlson, E.M.; Auberry, D.; Blanco, H.; Borkum, M.I.; Corilo, Y.E.; Davenport, K.W.; Deshpande, S.; Devarakonda, R.; Drake, M.; Duncan, W.D.; et al. The National Microbiome Data Collaborative: Enabling Microbiome Science. *Nat. Rev. Microbiol.* **2020**, *18*, 313–314. [CrossRef]
17. Vangay, P.; Burgin, J.; Johnston, A.; Beck, K.L.; Berrios, D.C.; Blumberg, K.; Canon, S.; Chain, P.; Chandonia, J.-M.; Christianson, D.; et al. Microbiome Metadata Standards: Report of the National Microbiome Data Collaborative's Workshop and Follow-On Activities. *mSystems* **2021**, *6*, e01194-20. [CrossRef]
18. Walls, R.L.; Deck, J.; Guralnick, R.; Baskauf, S.; Beaman, R.; Blum, S.; Bowers, S.; Buttigieg, P.L.; Davies, N.; Endresen, D.; et al. Semantics in Support of Biodiversity Knowledge Discovery: An Introduction to the Biological Collections Ontology and Related Ontologies. *PLoS ONE* **2014**, *9*, e89606. [CrossRef]
19. Buttigieg, P.L.; Pafilis, E.; Lewis, S.E.; Schildhauer, M.P.; Walls, R.L.; Mungall, C.J. The Environment Ontology in 2016: Bridging Domains with Increased Scope, Semantic Density, and Interoperation. *J. Biomed. Semant.* **2016**, *7*, 57. [CrossRef]
20. Ashburner, M.; Ball, C.A.; Blake, J.A.; Botstein, D.; Butler, H.; Cherry, J.M.; Davis, A.P.; Dolinski, K.; Dwight, S.S.; Eppig, J.T.; et al. Gene Ontology: Tool for the Unification of Biology. *Nat. Genet.* **2000**, *25*, 25–29. [CrossRef]
21. Gene Ontology Consortium. The Gene Ontology Resource: Enriching a GOld Mine. *Nucleic Acids Res.* **2021**, *49*, D325–D334. [CrossRef] [PubMed]
22. Dixon, H.B.F. IUPAC-IUBMB Joint Commission on Biochemical Nomenclature (JCBN) and Nomenclature Committee of IUBMB (NC-IUBMB), Newsletter 1999. *Eur. J. Biochem.* **1999**, *264*, 607–609.
23. Caspi, R.; Billington, R.; Keseler, I.M.; Kothari, A.; Krummenacker, M.; Midford, P.E.; Ong, W.K.; Paley, S.; Subhraveti, P.; Karp, P.D. The MetaCyc Database of Metabolic Pathways and Enzymes—A 2019 Update. *Nucleic Acids Res.* **2020**, *48*, D445–D453. [CrossRef] [PubMed]
24. Schoch, C.L.; Ciufo, S.; Domrachev, M.; Hotton, C.L.; Kannan, S.; Khovanskaya, R.; Leipe, D.; Mcveigh, R.; O'Neill, K.; Robbertse, B.; et al. NCBI Taxonomy: A Comprehensive Update on Curation, Resources and Tools. *Database J. Biol. Databases Curation* **2020**, *2020*, baaa062. [CrossRef] [PubMed]
25. Parte, A.C.; Sardà Carbasse, J.; Meier-Kolthoff, J.P.; Reimer, L.C.; Göker, M. List of Prokaryotic Names with Standing in Nomenclature (LPSN) Moves to the DSMZ. *Int. J. Syst. Evol. Microbiol.* **2020**, *70*, 5607–5612. [CrossRef] [PubMed]
26. Mitchell, A.L.; Almeida, A.; Beracochea, M.; Boland, M.; Burgin, J.; Cochrane, G.; Crusoe, M.R.; Kale, V.; Potter, S.C.; Richardson, L.J.; et al. MGnify: The Microbiome Analysis Resource in 2020. *Nucleic Acids Res.* **2020**, *48*, D570–D578. [CrossRef]
27. Chen, I.-M.A.; Chu, K.; Palaniappan, K.; Ratner, A.; Huang, J.; Huntemann, M.; Hajek, P.; Ritter, S.; Varghese, N.; Seshadri, R.; et al. The IMG/M Data Management and Analysis System v.6.0: New Tools and Advanced Capabilities. *Nucleic Acids Res.* **2021**, *49*, D751–D763. [CrossRef]
28. Wilke, A.; Bischof, J.; Harrison, T.; Brettin, T.; D'Souza, M.; Gerlach, W.; Matthews, H.; Paczian, T.; Wilkening, J.; Glass, E.M.; et al. A RESTful API for Accessing Microbial Community Data for MG-RAST. *PLoS Comput. Biol.* **2015**, *11*, e1004008. [CrossRef]
29. Roberts, R.J. PubMed Central: The GenBank of the Published Literature. *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 381–382. [CrossRef]
30. Harmston, N.; Filsell, W.; Stumpf, M.P. What the Papers Say: Text Mining for Genomics and Systems Biology. *Hum. Genom.* **2010**, *5*, 17–29. [CrossRef]
31. Pafilis, E.; Frankild, S.P.; Fanini, L.; Faulwetter, S.; Pavloudi, C.; Vasileiadou, A.; Arvanitidis, C.; Jensen, L.J. The SPECIES and ORGANISMS Resources for Fast and Accurate Identification of Taxonomic Names in Text. *PLoS ONE* **2013**, *8*, e65390. [CrossRef] [PubMed]
32. Pafilis, E.; Buttigieg, P.L.; Ferrell, B.; Pereira, E.; Schnetzer, J.; Arvanitidis, C.; Jensen, L.J. EXTRACT: Interactive Extraction of Environment Metadata and Term Suggestion for Metagenomic Sample Annotation. *Database* **2016**, *2016*, baw005. [CrossRef] [PubMed]
33. Von Mering, C.; Jensen, L.J.; Snel, B.; Hooper, S.D.; Krupp, M.; Foglierini, M.; Jouffre, N.; Huynen, M.A.; Bork, P. STRING: Known and Predicted Protein–Protein Associations, Integrated and Transferred across Organisms. *Nucleic Acids Res.* **2005**, *33*, D433–D437. [CrossRef] [PubMed]
34. Franceschini, A.; Szklarczyk, D.; Frankild, S.; Kuhn, M.; Simonovic, M.; Roth, A.; Lin, J.; Minguez, P.; Bork, P.; von Mering, C.; et al. STRING v9.1: Protein-Protein Interaction Networks, with Increased Coverage and Integration. *Nucleic Acids Res.* **2013**, *41*, D808–D815. [CrossRef]
35. Gomez-Cabrero, D.; Abugessaisa, I.; Maier, D.; Teschendorff, A.; Merkenschlager, M.; Gisel, A.; Ballestar, E.; Bongcam-Rudloff, E.; Conesa, A.; Tegnér, J. Data Integration in the Era of Omics: Current and Future Challenges. *BMC Syst. Biol.* **2014**, *8*, I1. [CrossRef]
36. Cavicchioli, R.; Ripple, W.J.; Timmis, K.N.; Azam, F.; Bakken, L.R.; Baylis, M.; Behrenfeld, M.J.; Boetius, A.; Boyd, P.W.; Classen, A.T.; et al. Scientists' Warning to Humanity: Microorganisms and Climate Change. *Nat. Rev. Microbiol.* **2019**, *17*, 569–586. [CrossRef]

37. D'Hondt, K.; Kostic, T.; McDowell, R.; Eudes, F.; Singh, B.K.; Sarkar, S.; Markakis, M.; Schelkle, B.; Maguin, E.; Sessitsch, A. Microbiome Innovations for a Sustainable Future. *Nat. Microbiol.* **2021**, *6*, 138–142. [CrossRef]
38. Conde-Pueyo, N.; Vidiella, B.; Sardanyés, J.; Berdugo, M.; Maestre, F.T.; De Lorenzo, V.; Solé, R. Synthetic Biology for Terraformation Lessons from Mars, Earth, and the Microbiome. *Life* **2020**, *10*, 14. [CrossRef]
39. Baltoumas, F.A.; Zafeiropoulou, S.; Karatzas, E.; Koutrouli, M.; Thanati, F.; Voutsadaki, K.; Gkonta, M.; Hotova, J.; Kasionis, I.; Hatzis, P.; et al. Biomolecule and Bioentity Interaction Databases in Systems Biology: A Comprehensive Review. *Biomolecules* **2021**, *11*, 1245. [CrossRef]
40. Reimer, L.C.; Vetcininova, A.; Carbasse, J.S.; Söhngen, C.; Gleim, D.; Ebeling, C.; Overmann, J. BacDive in 2019: Bacterial Phenotypic Data for High-Throughput Biodiversity Analysis. *Nucleic Acids Res.* **2019**, *47*, D631–D636. [CrossRef]
41. Shaaban, H.; Westfall, D.A.; Mohammad, R.; Danko, D.; Bezdan, D.; Afshinnekoo, E.; Segata, N.; Mason, C.E. The Microbe Directory: An Annotated, Searchable Inventory of Microbes' Characteristics. *Gates Open Res.* **2018**, *2*, 3. [CrossRef] [PubMed]
42. Kosina, S.M.; Greiner, A.M.; Lau, R.K.; Jenkins, S.; Baran, R.; Bowen, B.P.; Northen, T.R. Web of Microbes (WoM): A Curated Microbial Exometabolomics Database for Linking Chemistry and Microbes. *BMC Microbiol.* **2018**, *18*, 115. [CrossRef] [PubMed]
43. Microbial Interaction Network Database. Available online: http://www.microbialnet.org/mind_home.html (accessed on 21 December 2021).
44. Tang, Y.; Dai, T.; Su, Z.; Hasegawa, K.; Tian, J.; Chen, L.; Wen, D. A Tripartite Microbial-Environment Network Indicates How Crucial Microbes Influence the Microbial Community Ecology. *Microb. Ecol.* **2020**, *79*, 342–356. [CrossRef] [PubMed]
45. Koutrouli, M.; Karatzas, E.; Paez-Espino, D.; Pavlopoulos, G.A. A Guide to Conquer the Biological Network Era Using Graph Theory. *Front. Bioeng. Biotechnol.* **2020**, *8*, 34. [CrossRef] [PubMed]
46. Li, K.; Hu, J.; Li, T.; Liu, F.; Tao, J.; Liu, J.; Zhang, Z.; Luo, X.; Li, L.; Deng, Y.; et al. Microbial Abundance and Diversity Investigations along Rivers: Current Knowledge and Future Directions. *Wiley Interdiscip. Rev. Water* **2021**, *8*, e1547. [CrossRef]
47. Jensen, L.J. One Tagger, Many Uses: Illustrating the Power of Ontologies in Dictionary-Based Named Entity Recognition. *bioRxiv* **2016**, 067132. [CrossRef]
48. Sayers, E.W.; Beck, J.; Bolton, E.E.; Bourexis, D.; Brister, J.R.; Canese, K.; Comeau, D.C.; Funk, K.; Kim, S.; Klimke, W.; et al. Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **2021**, *49*, D10–D17. [CrossRef]
49. Pafilis, E.; Frankild, S.P.; Schnetzer, J.; Fanini, L.; Faulwetter, S.; Pavloudi, C.; Vasileiadou, K.; Leary, P.; Hammock, J.; Schulz, K.; et al. ENVIRONMENTS and EOL: Identification of Environment Ontology Terms in Text and the Annotation of the Encyclopedia of Life. *Bioinformatics* **2015**, *31*, 1872–1874. [CrossRef]
50. Mukherjee, S.; Stamatis, D.; Bertsch, J.; Ovchinnikova, G.; Sundaramurthi, J.C.; Lee, J.; Kandimalla, M.; Chen, I.-M.A.; Kyrpides, N.C.; Reddy, T.B.K. Genomes OnLine Database (GOLD) v.8: Overview and Updates. *Nucleic Acids Res.* **2021**, *49*, D723–D733. [CrossRef]
51. De la Cuesta-Zuluaga, J.; Ley, R.E.; Youngblut, N.D. Struo: A Pipeline for Building Custom Databases for Common Metagenome Profilers. *Bioinformatics* **2020**, *36*, 2314–2315. [CrossRef]
52. Parks, D.H.; Chuvochina, M.; Chaumeil, P.-A.; Rinke, C.; Mussig, A.J.; Hugenholtz, P. A Complete Domain-to-Species Taxonomy for Bacteria and Archaea. *Nat. Biotechnol.* **2020**, *38*, 1079–1086. [CrossRef] [PubMed]
53. Quast, C.; Pruesse, E.; Yilmaz, P.; Gerken, J.; Schweer, T.; Yarza, P.; Peplies, J.; Glöckner, F.O. The SILVA Ribosomal RNA Gene Database Project: Improved Data Processing and Web-Based Tools. *Nucleic Acids Res.* **2013**, *41*, D590–D596. [CrossRef] [PubMed]
54. Guillou, L.; Bachar, D.; Audic, S.; Bass, D.; Berney, C.; Bittner, L.; Boutte, C.; Burgaud, G.; de Vargas, C.; Decelle, J.; et al. The Protist Ribosomal Reference Database (PR2): A Catalog of Unicellular Eukaryote Small Sub-Unit RRNA Sequences with Curated Taxonomy. *Nucleic Acids Res.* **2013**, *41*, D597–D604. [CrossRef] [PubMed]
55. Del Campo, J.; Kolisko, M.; Boscaro, V.; Santoferrara, L.F.; Nenarokov, S.; Massana, R.; Guillou, L.; Simpson, A.; Berney, C.; de Vargas, C.; et al. EukRef: Phylogenetic Curation of Ribosomal RNA to Enhance Understanding of Eukaryotic Diversity and Distribution. *PLoS Biol.* **2018**, *16*, e2005849. [CrossRef]
56. Suter, L.; Polanowski, A.M.; Clarke, L.J.; Kitchener, J.A.; Deagle, B.E. Capturing Open Ocean Biodiversity: Comparing Environmental DNA Metabarcoding to the Continuous Plankton Recorder. *Mol. Ecol.* **2021**, *30*, 3140–3157. [CrossRef]
57. Leray, M.; Ho, S.-L.; Lin, I.-J.; Machida, R.J. MIDORI Server: A Webserver for Taxonomic Assignment of Unknown Metazoan Mitochondrial-Encoded Sequences Using a Curated Database. *Bioinformatics* **2018**, *34*, 3753–3754. [CrossRef]
58. Nilsson, R.H.; Larsson, K.-H.; Taylor, A.F.S.; Bengtsson-Palme, J.; Jeppesen, T.S.; Schigel, D.; Kennedy, P.; Picard, K.; Glöckner, F.O.; Tedersoo, L.; et al. The UNITE Database for Molecular Identification of Fungi: Handling Dark Taxa and Parallel Taxonomic Classifications. *Nucleic Acids Res.* **2019**, *47*, D259–D264. [CrossRef]
59. Pavloudi, C.; Oulas, A.; Vasileiadou, K.; Kotoulas, G.; Troch, M.D.; Friedrich, M.W.; Arvanitidis, C. Diversity and Abundance of Sulfate-Reducing Microorganisms in a Mediterranean Lagoonal Complex (Amvrakikos Gulf, Ionian Sea) Derived from DsrB Gene. *Aquat. Microb. Ecol.* **2017**, *79*, 209–219. [CrossRef]
60. Westergaard, D.; Stærfeldt, H.-H.; Tønsberg, C.; Jensen, L.J.; Brunak, S. A Comprehensive and Quantitative Comparison of Text-Mining in 15 Million Full-Text Articles versus Their Corresponding Abstracts. *PLoS Comput. Biol.* **2018**, *14*, e1005962. [CrossRef]
61. Ferguson, C.; Araújo, D.; Faulk, L.; Gou, Y.; Hamelers, A.; Huang, Z.; Ide-Smith, M.; Levchenko, M.; Marinos, N.; Nambiar, R.; et al. Europe PMC in 2020. *Nucleic Acids Res.* **2020**, *49*, D1507–D1514. [CrossRef]

62. Zafeiropoulos, H.; Viet, H.Q.; Vasileiadou, K.; Potirakis, A.; Arvanitidis, C.; Topalis, P.; Pavloudi, C.; Pafilis, E. PEMA: A Flexible Pipeline for Environmental DNA Metabarcoding Analysis of the 16S/18S Ribosomal RNA, ITS, and COI Marker Genes. *GigaScience* **2020**, *9*, giaa022. [CrossRef] [PubMed]
63. Karatzas, E.; Baltoumas, F.A.; Panayiotou, N.A.; Schneider, R.; Pavlopoulos, G.A. Arena3Dweb: Interactive 3D Visualization of Multilayered Networks. *Nucleic Acids Res.* **2021**, *49*, W36–W45. [CrossRef] [PubMed]
64. Baltoumas, F.A.; Zafeiropoulou, S.; Karatzas, E.; Paragkamian, S.; Thanati, F.; Iliopoulos, I.; Eliopoulos, A.G.; Schneider, R.; Jensen, L.J.; Pafilis, E.; et al. OnTheFly2.0: A Text-Mining Web Application for Automated Biomedical Entity Recognition, Document Annotation, Network and Functional Enrichment Analysis. *NAR Genom. Bioinform.* **2021**, *3*, lqab090. [CrossRef] [PubMed]
65. Thanati, F.; Karatzas, E.; Baltoumas, F.A.; Stravopodis, D.J.; Eliopoulos, A.G.; Pavlopoulos, G.A. FLAME: A Web Tool for Functional and Literature Enrichment Analysis of Multiple Gene Lists. *Biology* **2021**, *10*, 665. [CrossRef] [PubMed]
66. Zoppi, J.; Guillaume, J.-F.; Neunlist, M.; Chaffron, S. MiBiOmics: An Interactive Web Application for Multi-Omics Data Exploration and Integration. *BMC Bioinform.* **2021**, *22*, 6. [CrossRef] [PubMed]
67. Sinclair, L.; Ijaz, U.Z.; Jensen, L.J.; Coolen, M.J.L.; Gubry-Rangin, C.; Chroňáková, A.; Oulas, A.; Pavloudi, C.; Schnetzer, J.; Weimann, A.; et al. Seqenv: Linking Sequences to Environments through Text Mining. *PeerJ* **2016**, *4*, e2690. [CrossRef]
68. Xue, C.-X.; Lin, H.; Zhu, X.-Y.; Liu, J.; Zhang, Y.; Rowley, G.; Todd, J.D.; Li, M.; Zhang, X.-H. DiTing: A Pipeline to Infer and Compare Biogeochemical Pathways from Metagenomic and Metatranscriptomic Data. *Front. Microbiol.* **2021**, *12*, 2118. [CrossRef]
69. Zafeiropoulos, H.; Gioti, A.; Ninidakis, S.; Potirakis, A.; Paragkamian, S.; Angelova, N.; Antoniou, A.; Danis, T.; Kaitetzidou, E.; Kasapidis, P.; et al. 0s and 1s in Marine Molecular Research: A Regional HPC Perspective. *GigaScience* **2021**, *10*. [CrossRef]
70. Binder, J.X.; Pletscher-Frankild, S.; Tsafou, K.; Stolte, C.; O'Donoghue, S.I.; Schneider, R.; Jensen, L.J. COMPARTMENTS: Unification and Visualization of Protein Subcellular Localization Evidence. *Database* **2014**, *2014*, bau012. [CrossRef]
71. Pletscher-Frankild, S.; Pallejà, A.; Tsafou, K.; Binder, J.X.; Jensen, L.J. DISEASES: Text mining and data integration of disease–gene associations. *Methods* **2015**, *74*, 83–89. [CrossRef]