Title

NMPFamsDB: a database of novel protein families from microbial metagenomes and metatranscriptomes

Permalink

https://escholarship.org/uc/item/4j6231sz

Journal

Nucleic Acids Research, 52(D1)

ISSN

0305-1048

Authors

Baltoumas, Fotis A

Karatzas, Evangelos

Liu, Sirui

et al.

Publication Date

2024-01-05

DOI

10.1093/nar/gkad800

Peer reviewed

Nucleic Acids Research, 2023, 1–11
https://doi.org/10.1093/nar/gkad800
Database issue

NMPFamsDB: a database of novel protein families from microbial metagenomes and metatranscriptomes

Fotis A. Baltoumas [1,*], Evangelos Karatzas [1], Sirui Liu [2], Sergey Ovchinnikov [2], Yorgos Sofianatos [1], I-Min Chen [3], Nikos C. Kyrpides [3] and Georgios A. Pavlopoulos [1,3,4,*]

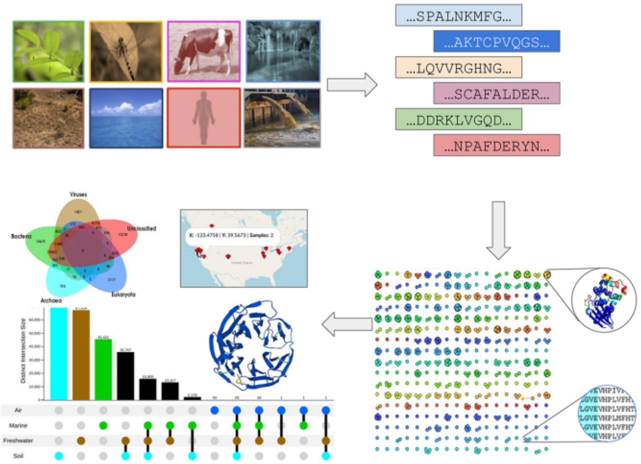
[1] Institute for Fundamental Biomedical Research, BSRC "Alexander Fleming", Vari, 16672, Greece
[2] John Harvard Distinguished Science Fellowship Program, Harvard University, Cambridge, MA 02138, USA
[3] DOE Joint Genome Institute, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720-8150, USA
[4] Center for New Biotechnologies and Precision Medicine, School of Medicine, National and Kapodistrian University of Athens, 75 Mikras Asias Street, Athens 11527, Greece

[*] To whom correspondence should be addressed. Tel: +30 210 9656310; Fax: +30 210 9653934; Email: baltoumas@fleming.gr
Correspondence may also be addressed to Georgios A. Pavlopoulos. Email: pavlopoulos@fleming.gr
Present address: Georgios A. Pavlopoulos, Institute for Fundamental Biomedical Research, BSRC "Alexander Fleming", 34 Fleming Street, Vari 16672, Greece.

Abstract

The Novel Metagenome Protein Families Database (NMPFamsDB) is a database of metagenome- and metatranscriptome-derived protein families, whose members have no hits to proteins of reference genomes or Pfam domains. Each protein family is accompanied by multiple sequence alignments, Hidden Markov Models, taxonomic information, ecosystem and geolocation metadata, sequence and structure predictions, as well as 3D structure models predicted with AlphaFold2. In its current version, NMPFamsDB hosts over 100 000 protein families, each with at least 100 members. The reported protein families significantly expand (more than double) the number of known protein sequence clusters from reference genomes and reveal new insights into their habitat distribution, origins, functions and taxonomy. We expect NMPFamsDB to be a valuable resource for microbial proteome-wide analyses and for further discovery and characterization of novel functions. NMPFamsDB is publicly available in http://www.nmpfamsdb.org/ or https://bib.fleming.gr/NMPFamsDB.

Graphical abstract



Introduction

Metagenomics is the study of metagenomes, defined as the total amount of genomic material in an environmental sample. Metagenomic (DNA) and metatranscriptomic (RNA) shotgun sequencing of complex biological samples has emerged as a prevalent source of information in the study and classification of microorganisms, as well as a treasure trove of novel sequence data (1). Advances in high-throughput shotgun sequencing technologies have improved the quality and reduced the cost of the method, resulting in a significant increase in the available metagenome and metatranscriptome sequences. All of the above has led to the application of metagenomics and metatranscriptomics to various biological fields, from ecology and biotechnology to disease diagnosis and treatment (2).

In order to extract the genetic composition of metagenomic samples, researchers usually follow two distinct methods: in

Received: August 14, 2023. **Editorial Decision:** August 22, 2023. **Accepted:** September 19, 2023

the first, sequence reads are accurately mapped to a known, annotated set of reference genome sequences to provide a quick overview of the presence of known organisms, genes and potential functions. The MG-RAST system (3) is one of the most popular implementations of this analysis type. Conversely, the second method employs the massive *de novo* assembly of the reads into contigs/scaffolds, which can provide invaluable insights into the presence of novel organisms and their genetic makeup in the analyzed samples. Advances in assembly and binning tools (4) have led to a significant increase in the assembled fraction of the average metagenome, coupled with a parallel exponential increase in the number of metagenome-assembled genomes (MAGs). Popular repositories employing this approach include the Integrated Microbial Genomes with Microbiome Samples (IMG/M) (5) and MGnify (6).

Despite their respective strengths, both methods share the same limitation with regards to gene annotation, which primarily relies on predicting gene function by searching for homologs in fully sequenced reference genomes or in gene and protein classification databases such as Pfam (7), InterPro (8), COG (9) or KEGG Orthology (10). While this approach can help annotate sequences that map to reference families, any genes predicted in assembled metagenomic data with no hits to any of the aforementioned resources are typically ignored and dropped from any subsequent analysis. As a result, the vast majority of existing metagenomic sequence data remains unexplored, limiting the ability to investigate the true diversity of the so-called *functional dark matter*.

In this study we present NMPFamsDB, a publicly available database of novel metagenome protein families, with no similarity to known protein domains or reference genomes and proteomes. Data in NMPFamsDB has been derived from microbial metagenome and metatranscriptome sequences from IMG/M (5). Novel sequences with no hits to Pfam (7) or reference genomes were clustered into families (NMPFs) based on sequence identity. For each NMPF, multiple sequence alignments (MSAs) and Hidden Markov Models (HMMs) were calculated. NMPFs are also accompanied by environmental and taxonomic metadata along with high-quality protein structure and topology predictions, including 3D structure models.

Materials and methods

Data retrieval and clustering

Protein sequences (longer than 35 amino acids (aa)) were collected from the IMG/M platform (5) for all public reference (isolate) genomes and assembled metagenomes and metatranscriptomes, and were filtered to remove low complexity regions. The reference genome dataset was represented by 89 412 bacterial, 9202 viral, 3073 archaeal and 804 eukaryotic genomes, which corresponded to 87 084 214 bacterial, 221 027 viral, 2 464 569 archaeal and 4 902 193 eukaryotic proteins, resulting in a final reference dataset of 94 672 003 protein sequences. Similarly, protein sequences with length ≥35aa and from scaffolds longer than 500 bp were collected from 20 759 metagenomes and 6172 metatranscriptomes and were translated to a non-redundant set of 8 364 611 943 predicted protein sequences. To identify the microbial functional dark matter, protein sequences with hits to Pfam-A (7) (using each Pfam profile's trusted cutoff) or to any sequences from the reference genome set (30% identity, 70% alignment length) were discarded, resulting in an unexplored metagenomic protein space consisting of 1 171 974 849 protein sequences.

After generating an all-versus-all sequence similarity network (SSN) (70% identity, 70% alignment length), the HipMCL clustering algorithm (11) was utilized to generate protein clusters (families). With the use of 2500 compute nodes (170 000 compute cores), HipMCL clustered the SSN in 3.5 h and generated 113 752 protein clusters with 100 members or more. Each cluster was represented by a full multiple sequence alignment (MSA), containing all cluster members, as well as a non-redundant 'seed' MSA using the definitions dictated by the Pfam database (max. 80% sequence identity, min. 75% alignment coverage). The seed MSAs were also used to produce Hidden Markov Model (HMM) profiles in the HMMER/Pfam (12) and HH-suite (13) formats. The consensus sequences of the MSAs were further searched against the reference genomes and Pfam-A to identify and discard distant homologs, as well as against the reference proteomes of RefSeq (14) and the unannotated clusters of Pfam-B. The final dataset (henceforth referred to as Environmental Dataset—ED) consisted of 19 986 348 non-redundant sequences from metagenomes and metatranscriptomes that were organized in 106 198 protein clusters (henceforth referred to as Novel Metagenome Protein Families (NMPFs)).

Structural predictions and protein fold recognition

The query sequences for structural analysis of NMPFs were determined by taking the central (pivot) sequence of each seed MSA. Central sequences were defined by performing pairwise distance calculations, creating an all-against-all distance matrix and selecting the sequence with the minimum Hamming distance from the alignment's average sequence. Calculations were performed using Python and the TensorFlow (15), SciKit (16) and Biopython (17) libraries. Positions not aligned to the query sequence were removed before further analysis. The resulting pivot sequences were submitted to a number of prediction methods in order to obtain as much structural and topological annotation as possible. Secondary structure predictions were performed using Porter v.5 (18). Disordered regions were predicted using the MobiDB-lite consensus algorithm (19). Signal peptides were predicted using SignalP-6.0 (20). Finally, protein topology prediction was performed in two stages: in the first stage, the sequences were submitted to PRED-CLASS (21), a neural network-based algorithm capable of discriminating protein topology for four categories (globular, transmembrane, fibrous and mixed globular-fibrous). In the second stage, sequences predicted as transmembrane by PRED-CLASS were submitted to Phobius (22) and PRED-TMBB2 (23) for the topology prediction of α-helical transmembrane proteins and transmembrane β-barrels, respectively. Results for structure and topology analysis are given in Supplementary Table S1.

3D structural predictions were performed using AlphaFold2 (24), resulting in 80 585 3D models. For the NMPF clusters with available 3D models, the secondary structure predictions of Porter5 were replaced by the secondary structure assignments of the models, calculated using DSSP (25,26). The generated 3D models were evaluated based on their pLDDT and predicted TM (pTM) scores; 13096 high-quality models (with pTM > 0.7) were identified and were subsequently searched against the SCOPe (27) and PDB (28) databases to detect structural homologs, using the TM-align (29) and MM-align (30) structure-based alignment methods. Models with no homologs to either database (TM-score < 0.5) were considered as potential novel structural folds. In addition to the above, models were clustered using TM-align and an all vs all approach, with a TM-score cutoff of 0.5 (Supplementary Table S2).

Taxonomic, environmental and geographical annotation

The metadata obtained from the ED dataset was used to annotate each NMPF cluster with regards to their associated ecosystems and geographical distribution. For ecosystems, the GOLD (31) ecosystem classification was used to organize datasets into ecosystem groups. Additional environmental annotation was provided by mapping the ED datasets to the Environment Ontology (ENVO) (32) and Earth Microbiome Project Ontology (EMPO) (33) schemes, based on the metadata of their associated GOLD studies. Each NMPF was then assigned to one or more ecosystems based on the ecosystem information of the ED dataset in which its sequences were found. Similarly, the geographical location metadata of the ED dataset, when available, was retrieved and mapped to their corresponding NMPFs.

Similar to the ecosystem metadata, NMPF taxonomic annotation was performed by assigning the taxonomy of the source sequencing scaffolds to each family's contained sequences. Initial taxonomic metadata was retrieved from the IMG/M (5) records of each sequencing scaffold, where available. Notably, the majority of the scaffolds used were too short and thus remained taxonomically unclassified. In addition, there was very little information on the taxonomy of viral scaffolds, or the existence of potentially unidentified eukaryotic sequences. To alleviate these issues, annotations for scaffolds >5 kb that had previously been identified as viral and included in IMG/VR v. 3.0 (34) were used. In addition, scaffolds of length 1–5 kb were analyzed with DeepVirFinder v1.0 (35) and the generated p-values were subsequently converted to q-values to obtain estimates of the false-discovery rate. Scaffolds with $q \leq 0.001$ were retained as putative viral scaffolds. Unclassified scaffolds were further analyzed to identify potential eukaryotic sequences using two eukaryotic sequence detection tools, Whokaryote (36) and EukRep (37). Furthermore, the NMPF clusters were searched against the Tara Oceans collection of eukaryotic MAGs (38). Finally, all remaining unclassified scaffolds were analyzed using the MMseqs2 taxonomy tool (39), performing six-frame translation searches against UniRef50 (40) and assigning each analyzed scaffold to the lowest common ancestor of the best hits for each frame.

NMPF quality metrics

In order to evaluate the quality of each NMPF, a number of criteria have been established, based on the different analyzed aspects of the families. These metrics include the following: (i) Transcriptomic evidence, i.e. the existence of metatranscriptome-derived sequences, (ii) the most common taxonomic group, (iii) the percentage of genes with valid ribosome-binding site (RBS) motifs, (iv) the percentage of genes near scaffold ends, (v) the percentage of genes coming from short (<2 kb) scaffolds, (vi) the number of associated GOLD sequencing projects, (vii) 3D structure prediction with AlphaFold2 and (viii) the predicted TM-score (pTM) of the 3D model. Transcriptomic evidence (i) refers to the existence of actively expressed genes in an NMPF; indicating on the validity of the contained sequences. The most common taxonomic group (ii) refers to the taxonomy shared by the majority of the NMPF's sequences at the kingdom level (bacteria, archaea, eukaryota, viruses or unclassified). The existence of RBS motifs (an indicator of a valid start codon) (iii), percentage of sequences near scaffold ends (iv) and percentage of sequences from short scaffolds (v) are all indicators of whether the NMPF is composed by complete (valid start or stop sites) or potentially truncated sequences; generally speaking, a high percentage of RBS-containing genes and low percentages of sequences from short scaffolds and/or near scaffold ends indicates that most sequences in the family are complete. The number of associated GOLD projects (vi) is an indication of whether the sequences come from distinct sequencing projects, rather than different analyses of the same project (potential technical replicates). Finally, the existence of an AlphaFold2 3D model (vii) shows that the NMPF's MSAs are robust enough to produce a 3D structure through co-evolution patterns, while the pTM score (viii) is an indicator of the overall structural integrity of that structure.

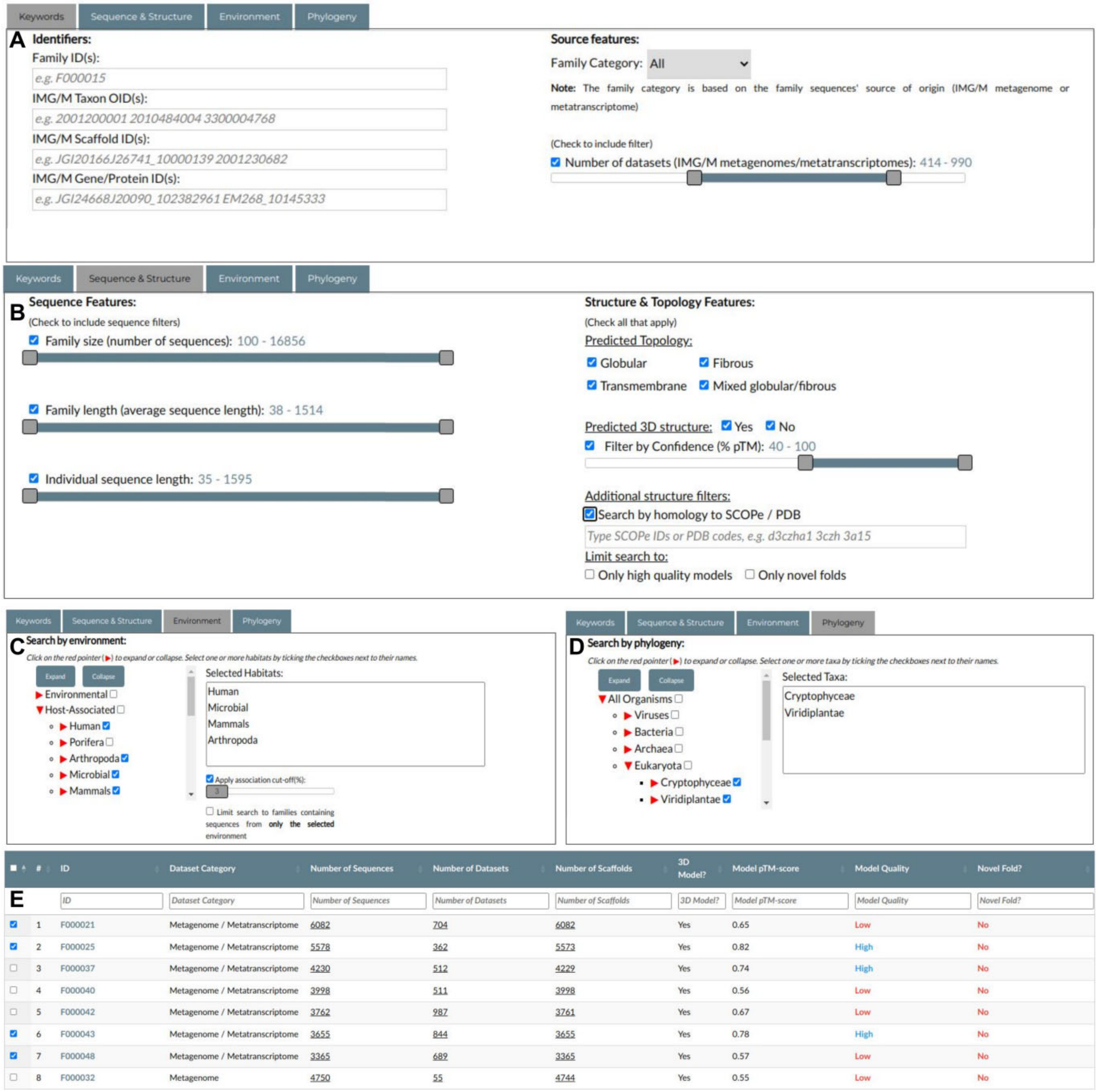
Implementation

The frontend of NMPFamsDB is implemented in HTML, CSS and JavaScript. The backend is supported by an Apache web server and a MySQL relational database. Server-side programming is mainly handled by PHP, while additional operations are implemented using Python and R/Shiny. The NMPFamsDB Application Programming Interface (API) was implemented using the Slim Framework (https://www.slimframework.com/). Sequence logos are rendered using Skylign (41). MSAs are visualized using MSAviewer (42). Structure and topology predictions are visualized using the SIB/nextProt feature viewer (43). 3D models are rendered with the Molstar (Mol*) structure viewer (44). Maps are rendered using the OpenLayers API (https://openlayers.org/). Sequence queries in the database are performed using LAST (45) for pairwise sequence alignments and HMMER v. 3.2 (12) for HMM-based searches. Interactive tables are generated using the DataTables package (https://datatables.net/). Plots are generated using the JavaScript/ApexCharts, Processing/P5, R/ggplot2 (46), R/plotly (47), R/chorddiag and R/upsetjs packages.

Results

Database components

In its current version, NMPFamsDB contains a total of 106198 NMPFs, each with 100 members or more. In the database, each NMPF has been assigned a unique, 7-digit identifier (F000001 to F106198). A distribution of NMPFs based on their average sequence length and number of clusters is given in Supplementary Figure S1. An analysis of all NMPFs based on the established quality metrics (see *Methods* section '*NMPF quality metrics*') shows that the majority of NMPFs (n = 67906) contain at least a fraction of metatranscriptome-derived sequences; specifically, 64186 NMPFs comprise a mixture of metagenome and metatranscriptome sequences (Mixed Metagenome/Metatranscriptome), while 3720 clusters come exclusively from metatranscriptomes (Metatranscriptome-



Identifiers:

Family ID(s): e.g. F000015

IMG/M Taxon OID(s): e.g. 2001200001 2010484004 3300004768

IMG/M Scaffold ID(s): e.g. JGI20166J26741_10000139 2001230682

IMG/M Gene/Protein ID(s): e.g. JGI24668J20090_102382961 EM268_10145333

Source features:

Family Category: All

Note: The family category is based on the family sequences' source of origin (IMG/M metagenome or metatranscriptome)

(Check to include filter)

☑ Number of datasets (IMG/M metagenomes/metatranscriptomes): 414 - 990

Sequence Features:

(Check to include sequence filters)

☑ Family size (number of sequences): 100 - 16856

☑ Family length (average sequence length): 38 - 1514

☑ Individual sequence length: 35 - 1595

Structure & Topology Features:

(Check all that apply)

Predicted Topology:

☑ Globular ☑ Fibrous

☑ Transmembrane ☑ Mixed globular/fibrous

Predicted 3D structure: ☑ Yes ☑ No

☑ Filter by Confidence (% pTM): 40 - 100

Additional structure filters:

☑ Search by homology to SCOPe / PDB

Type SCOPe IDs or PDB codes, e.g. d3czha1 3czh 3a15

Limit search to:

☐ Only high quality models ☐ Only novel folds

Search by environment:

Click on the red pointer (▶) to expand or collapse. Select one or more habitats by ticking the checkboxes next to their names.

- ▶ Environmental ☐
- ▼ Host-Associated ☐
  - ▶ Human ☑
  - ▶ Porifera ☐
  - ▶ Arthropoda ☑
  - ▶ Microbial ☑
  - ▶ Mammals ☑

Selected Habitats:

Human
Microbial
Mammals
Arthropoda

☑ Apply association cut-off(%): 3

☐ Limit search to families containing sequences from **only the selected** environment

Search by phylogeny:

Click on the red pointer (▶) to expand or collapse. Select one or more taxa by ticking the checkboxes next to their names.

- ▼ All Organisms ☐
  - ▶ Viruses ☐
  - ▶ Bacteria ☐
  - ▶ Archaea ☐
  - ▼ Eukaryota ☐
    - ▶ Cryptophyceae ☑
    - ▶ Viridiplantae ☑

Selected Taxa:

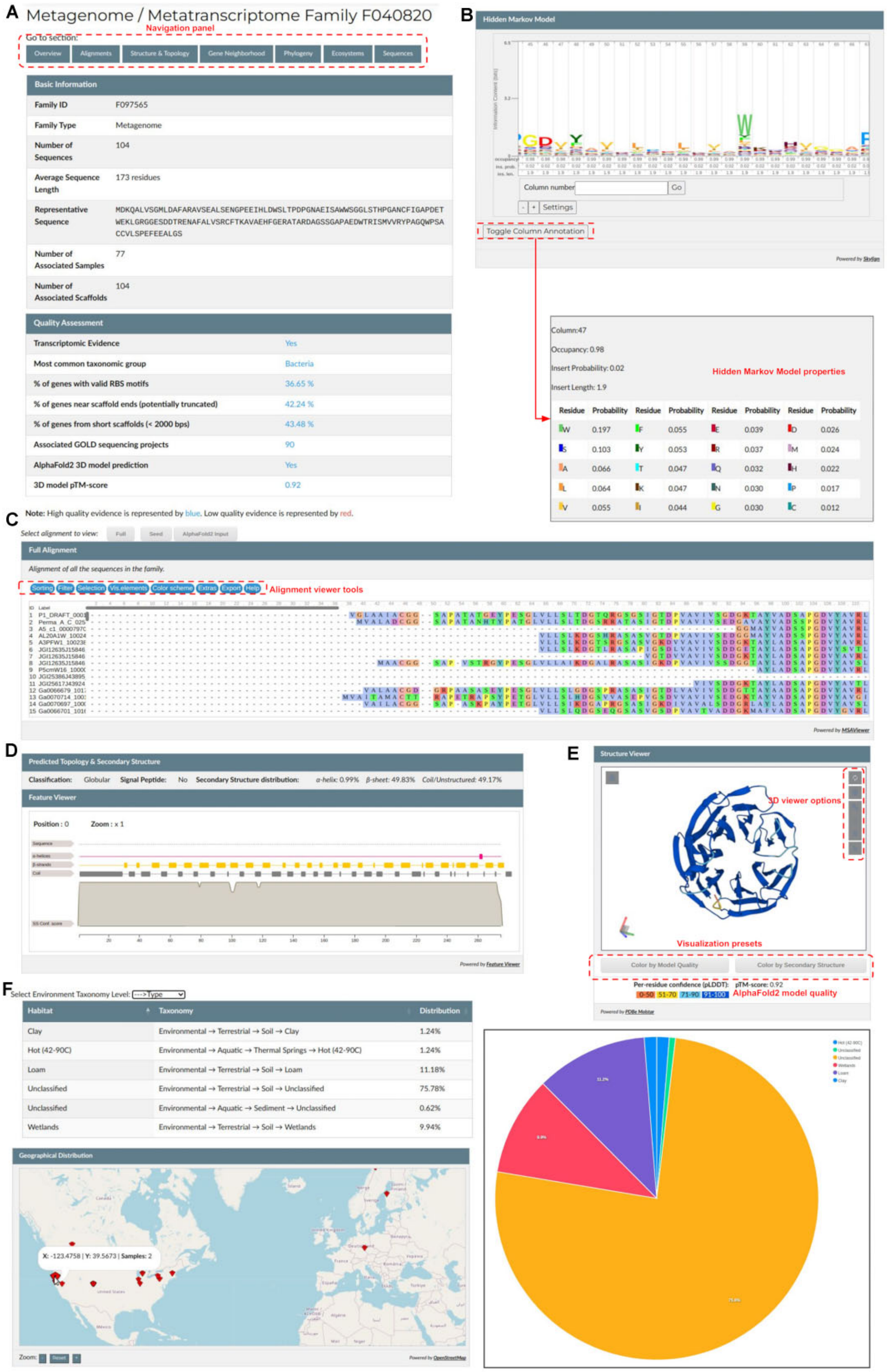
Cryptophyceae
Viridiplantae

■	#	ID	Dataset Category	Number of Sequences	Number of Datasets	Number of Scaffolds	3D Model?	Model pTM-score	Model Quality	Novel Fold?
		ID	Dataset Category	Number of Sequences	Number of Datasets	Number of Scaffolds	3D Model?	Model pTM-score	Model Quality	Novel Fold?
☑	1	F000021	Metagenome / Metatranscriptome	6082	704	6082	Yes	0.65	Low	No
☑	2	F000025	Metagenome / Metatranscriptome	5578	362	5573	Yes	0.82	High	No
☐	3	F000037	Metagenome / Metatranscriptome	4230	512	4229	Yes	0.74	High	No
☐	4	F000040	Metagenome / Metatranscriptome	3998	511	3998	Yes	0.56	Low	No
☐	5	F000042	Metagenome / Metatranscriptome	3762	987	3761	Yes	0.67	Low	No
☑	6	F000043	Metagenome / Metatranscriptome	3655	844	3655	Yes	0.78	High	No
☑	7	F000048	Metagenome / Metatranscriptome	3365	689	3365	Yes	0.57	Low	No
☐	8	F000032	Metagenome	4750	55	4744	Yes	0.55	Low	No

Figure 1. The NMPFamsDB family browser interface. A search panel exists at the top of the page, with search options grouped into four categories and accessible through tab buttons at the top of the panel. (**A**) Keyword options include searching by various identifiers, family classification based on its contained sequence types, and the number of associated datasets. (**B**) Sequence & Structure options include filters for the number and average length of sequences, the predicted protein topology, the existence of a predicted 3D structure model and, for 3D models, the associated confidence score (% pTM-score). (**C**) Environment search contains the database's ecosystems, hierarchically organized in an interactive tree structure. Users can select one or more ecosystems, which will appear at the right of the panel. They can also limit their search by setting an association cut-off, or by selecting the families that belong only to the chosen environments (100% association). (**D**) Phylogeny search, similar to environment, contains a list of taxonomy entities, hierarchically classified. (**E**) The results of a search are presented in an interactive table that can be further filtered using the column filters below the labels.

only) and 38 292 from metagenomes (Metagenome-only). With regards to each NMPF's most prevalent taxonomy, 59 780 NMPFs are classified as primarily bacterial, 2843 as archaeal, 7930 as eukaryotic, 13 963 as viral and 21 682 as unclassified. 63 820 NMPFs contain genes with valid RBS motifs, while 71 710 NMPFs are primarily composed from sequences that are likely to be complete (i.e. have low percentages of members from short scaffolds or from near scaffold ends). 95375 NMPFs are derived from multiple GOLD sequencing projects (50 or more per family). Finally, most NMPFs ($n = 80\,585$; 75.88%) have 3D structure models predicted with AlphaFold2, while 13 096 of these predictions can be considered high-confidence (pTM-score > 0.70).

Apart from the NMPFs themselves, NMPFamsDB also hosts the metadata of the ED datasets and their associated sequencing scaffolds, from which the analyzed protein



Metagenome / Metatranscriptome Family F040820

Basic Information	
Family ID	F097565
Family Type	Metagenome
Number of Sequences	104
Average Sequence Length	173 residues
Representative Sequence	MDKQALVSGMLDAFARAVSEALSENGPEEIHLDWSLTPDPGNAEISAWWSGGLSTHPGANCFIGAPDET WEKLGRGGESDDTRENAFALVSRCFTKAVAEHFGERATARDAGSSGAPAEDWTRISMVVRYPAGQWPSA CCVLSPEFEEALGS
Number of Associated Samples	77
Number of Associated Scaffolds	104

Quality Assessment	
Transcriptomic Evidence	Yes
Most common taxonomic group	Bacteria
% of genes with valid RBS motifs	36.65 %
% of genes near scaffold ends (potentially truncated)	42.24 %
% of genes from short scaffolds (< 2000 bps)	43.48 %
Associated GOLD sequencing projects	90
AlphaFold2 3D model prediction	Yes
3D model pTM-score	0.92

Note: High quality evidence is represented by blue. Low quality evidence is represented by red.

Column:47

Occupancy: 0.98

Insert Probability: 0.02

Insert Length: 1.9

Residue	Probability	Residue	Probability	Residue	Probability	Residue	Probability
W	0.197	F	0.055	E	0.039	D	0.026
S	0.103	Y	0.053	R	0.037	M	0.024
A	0.066	T	0.047	Q	0.032	H	0.022
L	0.064	K	0.047	N	0.030	P	0.017
V	0.055	I	0.044	G	0.030	C	0.012

Full Alignment

Alignment of all the sequences in the family.

Predicted Topology & Secondary Structure

Classification: Globular **Signal Peptide:** No **Secondary Structure distribution:** *α-helix:* 0.99% *β-sheet:* 49.83% *Coil/Unstructured:* 49.17%

Per-residue confidence (pLDDT): 0-50 51-70 71-90 91-100 **pTM-score:** 0.92

Habitat	Taxonomy	Distribution
Clay	Environmental → Terrestrial → Soil → Clay	1.24%
Hot (42-90C)	Environmental → Aquatic → Thermal Springs → Hot (42-90C)	1.24%
Loam	Environmental → Terrestrial → Soil → Loam	11.18%
Unclassified	Environmental → Terrestrial → Soil → Unclassified	75.78%
Unclassified	Environmental → Aquatic → Sediment → Unclassified	0.62%
Wetlands	Environmental → Terrestrial → Soil → Wetlands	9.94%

Figure 2. An example of an NMPFamsDB entry (F040820). (**A**) The top of the entry page contains a navigation browser and an overview panel; clicking each button will redirect users to the respective part of the entry. The overview panel contains basic NMPF annotation. (**B**) Interactive viewer for the family's HMM profile in sequence logo representation. Clicking on the logo or the *Toggle Column Annotation* button will open a window showing that particular position's properties. (**C**) Interactive multiple sequence alignment viewer for the family's MSAs. (**D**) The NMPF's structure and topology predictions are presented in an interactive feature viewer. (**E**) If the NMPF has a predicted 3D structure model, it is presented through an interactive molecular viewer. (**F**) The NMPF's ecosystem metadata are presented in an interactive table and pie chart. In addition, the geographical distribution of the family is shown in an interactive map.

sequences were derived. Specifically, the database contains 19 326 ED datasets (14 913 metagenomes and 4413 metatranscriptomes). Each dataset is represented by its IMG/M-assigned Taxon OID. NMPFamsDB also holds 17280119 scaffolds, each represented by its IMG/M assigned Taxon and Scaffold OIDs. Of these scaffolds, approximately 36.2% ($n = 6\ 257\ 223$) remained unclassified with regards to their taxonomy; the rest were classified as bacterial ($n = 80\ 49\ 154$), archaeal ($n = 382\ 761$), eukaryotic ($n = 1184393$) and viral ($n = 1\ 406\ 588$). Finally, NMPFamsDB contains 1972 distinct ecosystems, organized using a hierarchical system derived from GOLD Ecosystems and grouped into three main categories: *Environmental* ($n = 1132$), *Host-associated* ($n = 459$) and *Engineered* ($n = 228$). Each of these categories was further divided into subcategories (*Terrestrial, Aquatic*, *Human, Mammal, Plants* etc.). The metadata of the datasets, scaffolds and ecosystems, was also used to annotate their associated NMPFs; as such, each NMPF contains all annotations derived from its associated datasets and scaffolds, including taxonomic associations, ecosystems and geolocation information.

The NMPFamsDB interface

The data contained in the NMPFamsDB web page can be accessed through the *Browse* menu at the NMPFamsDB navigation bar. Through the *Browse Families* page users can navigate the database's NMPFs and perform both simple and complex queries (Figure 1). Specifically, users can search NMPFs using NMPFamsDB identifiers or the keywords of their related IMG/M metagenome datasets (Taxon OID), scaffolds (Scaffold ID) or sequences (Gene ID). They can also choose to retrieve NMPFs based on the category of their sequences (Metagenome-only, Metatranscriptome-only, Mixed, or all families), or limit their searches to families with a selected number of datasets (Figure 1A). In addition, users can perform searches by applying filters to the NMPFs' sequence and structure features, namely, the number or average length of sequences in the family, the predicted protein topology, the existence of a 3D structure model and its confidence (Figure 1B). Finally, they can perform searches based on the ecosystem (Figure 1C) or taxonomic metadata (Figure 1D) of the families.

All of the above can be accessed through the panels at the top of the *Browse Families* page, with search options grouped into four categories (*Keyword*, *Sequence & Structure*, *Environment* and *Phylogeny*). Notably, multiple search parameters can be combined to create complex queries, e.g. a search can be performed for metagenome-only families, with an available 3D structure having a confidence of at least 70%, that have been associated with Host-associated ecosystems. The results of the queries are presented in an interactive table at the bottom of the search panel (Figure 1E) in paginated form. The results can be filtered using the search fields located below the column labels. Each individual NMPF entry can be accessed by clicking the link of its identifier. Finally, one or more entries can be selected using the checkboxes at the left side and exported in comma- or tab-delimited format.

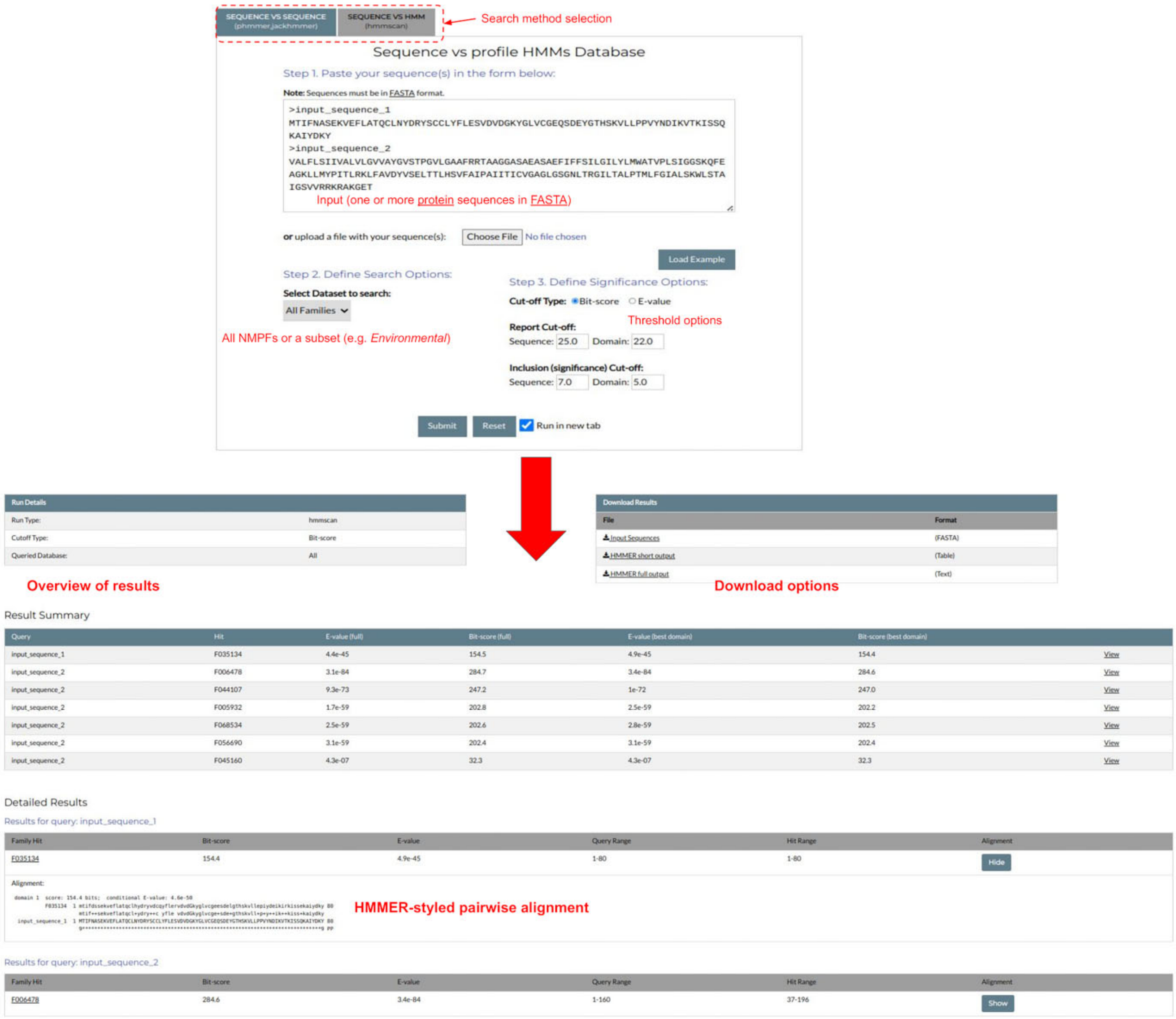
The data of each individual NMPF can be explored by visiting its respective *Family Entry* page (Figure 2). The family entry is organized into distinct sections, accessible through a navigation browser at the top of the entry. The *Overview* section contains basic NMPF information (Figure 2A), including its category, the number of associated sequences, datasets and scaffolds, the family's average sequence length, and a representative sequence, derived from the family's MSA. In addition, the section lists the NMPF's quality metric values and displays its most common taxonomic group and most abundant ecosystem classification. It also contains an interactive Sequence Logo viewer for the family's *HMM profile* (Figure 2B). The logo can be navigated with the mouse; clicking on any logo position, or clicking the *Toggle Column Annotation* button will open a window showing that particular position's residue probabilities in the HMM. The profile itself is also available for download in the HMMER and HH-suite formats, accessible through the download button at the top of the entry.

The *Alignments* section (Fig. 2C) contains an interactive alignment viewer, through which the NMPF's *Full* and *Seed MSAs* can be inspected. The viewer offers a number of options for visualizing and filtering the alignments, accessible through a menu at the top of its panel. These include different coloring schemes for amino acids, filtering alignment positions based on sequence identity or column occupancy, searching the MSAs using sequence patterns or regular expressions, and visualizing elements such as conservation histograms, sequence logos or the family consensus on top of the MSA. The MSAs are available for download in the FASTA format through the download button at the top of the entry. In addition, the sequences themselves can also be found in the *Sequences* section at the bottom of the entry page.

The structural and topology annotation of the NMPF can be found in the *Structure & Topology* section (Fig. 2D). Annotations are mapped in the family's pivot sequence and are shown in graphical format through an interactive feature viewer; these include the predicted topology of the NMPF, its secondary structure and per-residue structure confidence score and, based on the family's topology, the presence or absence of specific topological features, such as a signal peptide or transmembrane segments. If a 3D model is available for the family, it is also shown in an interactive 3D viewer at the right of the section (Figure 2E). Finally, if the 3D model has significant similarity ($TM > 0.5$) to known protein domains, derived from the SCOPe database, the top 5 hits are shown in a list at the bottom of the section.

Apart from the above, each NMPF is further annotated with metadata, based on the associated ED datasets and sequencing scaffolds. Taxonomic annotation is provided from the sequencing scaffolds and is available in the *Phylogeny* section. In addition, if, through the scaffolds' annotation, the NMPF's sequences have been found to be in close proximity to protein-coding genes with known structural or functional domains (Pfam families), the latter are presented in a distinct section called *Gene Neighborhood*. Finally, the *Environmental Properties* section shows the ecosystem annotation and geographical distribution of the NMPF based on its datasets (Figure 2F). The associated ecosystems are presented in an interactive table and associated pie chart graph, while the geographical distribution of the family is shown in an interactive world map, with each point corresponding to the longitude and latitude coordinates of an ED dataset.

Similar to NMPFs, the database's protein sequences, ED datasets, scaffolds and ecosystems can be accessed through the *Browse Sequences, Datasets, Scaffolds* and *Ecosystems* pages respectively (Supplementary Figures S2–S5). Queries can be


SEQUENCE VS SEQUENCE (phmmer,jackhmmer)
SEQUENCE VS HMM (hmmscan)
Search method selection
Choose File
Load Example
Submit
Reset
Input (one or more protein sequences in FASTA)
All NMPFs or a subset (e.g. *Environmental*)
Threshold options
Overview of results
Download options
HMMER-styled pairwise alignment
Hide
Show

Sequence vs profile HMMs Database

Step 1. Paste your sequence(s) in the form below:

Note: Sequences must be in FASTA format.

```
>input_sequence_1
MTIFNASEKVEFLATQCLNYDRYSCCLYFLESVDVDGKYGLVCGEQSDEYGTHSKVLLPPVYNDIKVTKISSQ
KAIYDKY
>input_sequence_2
VALFLSIIVALVLGVVAYGVSTPGVLGAAFRRTAAGGASAEASAEFIFFSILGILYLMWATVPLSIGGSKQFE
AGKLLMYPITLRKLFAVDYVSELTTLHSVFAIPAIITICVGAGLGSGNLTRGILTALPTMLFGIALSKWLSTA
IGSVVRRKRAKGET
```

or upload a file with your sequence(s): No file chosen

Step 2. Define Search Options:

Select Dataset to search: All Families

Step 3. Define Significance Options:

Cut-off Type: ● Bit-score ○ E-value

Report Cut-off:
Sequence: 25.0 Domain: 22.0

Inclusion (significance) Cut-off:
Sequence: 7.0 Domain: 5.0

☑ Run in new tab

Run Details	
Run Type:	hmmscan
Cutoff Type:	Bit-score
Queried Database:	All

Download Results

File	Format
Input Sequences	(FASTA)
HMMER short output	(Table)
HMMER full output	(Text)

Result Summary

Query	Hit	E-value (full)	Bit-score (full)	E-value (best domain)	Bit-score (best domain)	
input_sequence_1	F035134	4.4e-45	154.5	4.9e-45	154.4	View
input_sequence_2	F006478	3.1e-84	284.7	3.4e-84	284.6	View
input_sequence_2	F044107	9.3e-73	247.2	1e-72	247.0	View
input_sequence_2	F005932	1.7e-59	202.8	2.5e-59	202.2	View
input_sequence_2	F068534	2.5e-59	202.6	2.8e-59	202.5	View
input_sequence_2	F056690	3.1e-59	202.4	3.1e-59	202.4	View
input_sequence_2	F045160	4.3e-07	32.3	4.3e-07	32.3	View

Detailed Results

Results for query: input_sequence_1

Family Hit	Bit-score	E-value	Query Range	Hit Range	Alignment
F035134	154.4	4.9e-45	1-80	1-80	Hide

Alignment:

```
domain 1  score: 154.4 bits;  conditional E-value: 4.6e-50
         F035134  1 mtifdssekveflatqclhydryvdcqyflervdvdGkyglvcgeesdelgthskvllepiydeikirkissekaiydky 80
                    mtif++sekveflatqcl+ydry++c yfle vdvdGkyglvcge+sde+gthskvll+p+y++ik++kiss+kaiydky
  input_sequence_1  1 MTIFNASEKVEFLATQCLNYDRYSCCLYFLESVDVDGKYGLVCGEQSDEYGTHSKVLLPPVYNDIKVTKISSQKAIYDKY 80
                    g++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++g pp
```

Results for query: input_sequence_2

Family Hit	Bit-score	E-value	Query Range	Hit Range	Alignment
F006478	284.6	3.4e-84	1-160	37-196	Show

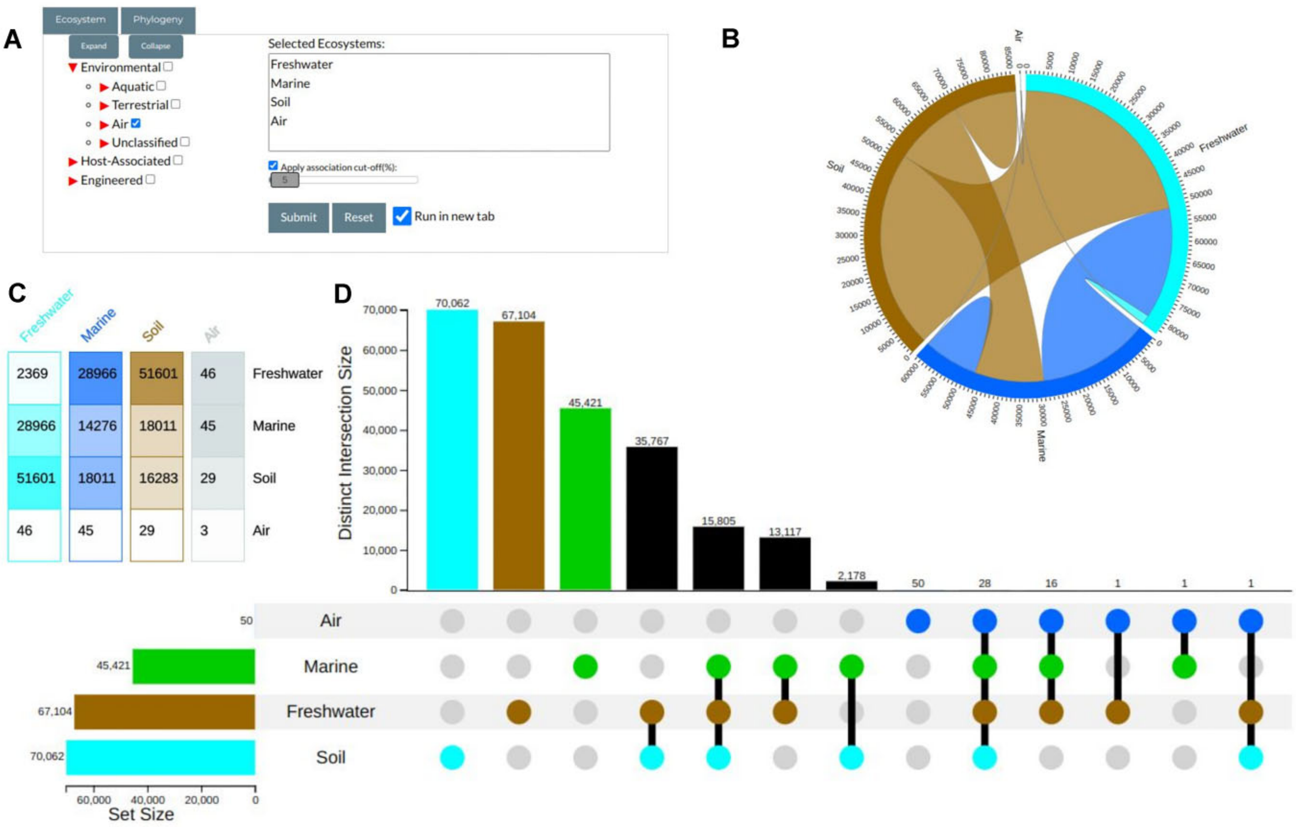
Figure 3. Input form (top) and example results (bottom) for the NMPFamsDB HMMER search tool. In the input form, the user can select which search method to run, either Sequence versus Sequence (using phmmer and jackhmmer) or Sequence versus HMM (using hmmscan). One or more query sequences can be submitted in the FASTA format. The user can also choose the reference database to run against (either the entire NMPFamsDB or one of its subsets) and define search parameters, including the threshold types and cutoff values. The results include a summary table of all sequence hits, with information such as the bit-score and *E*-value, as well as the pairwise alignments for each hit. All results are available for download in text files for further analysis.

performed in a manner similar to NMPFs by using the search panel at the top of each page, with search options including family, dataset, scaffold or sequence identifiers, dataset category, taxonomic assignments and ecosystem associations. In addition, ED datasets can be queried based on their sequencing center and sampling location, while scaffolds are based on contig length. Search results are returned in table format and can be exported in a tab- or comma-delimited file; in addition, sequence search results can be exported in the standardized FASTA format for further analysis.

NMPFamsDB holds the detailed records of the analyzed ED datasets and scaffolds, accessible through their dedicated *Dataset* and *Scaffold Entry* pages. Examples are shown in Supplementary Figures S6 and S7. For each ED dataset and scaffold, the reported metadata includes its category (Metagenome or Metatranscriptome), sequencing information (sequencing center and status) and properties (genome size, number of associated scaffolds and genes), taxonomic annotation and geographic metadata, including the sampling location, coordinates in longitude and latitude (in degrees) and, where applicable, altitude or length (in meters). In addition, lists of the associated NMPFs and sequences, as well as hyperlinks to IMG/M are provided.

Sequence search and data visualization

NMPFamsDB offers a number of analysis tools that are grouped into two categories: *Sequence Search* and *Data Visualization*. The *Sequence Search* tools include interfaces to *HMMER* (Figure 3) and *LAST* (Supplementary Figure S8), allowing users to upload their query sequences and HMM-based queries or perform pairwise alignments against NMPFs.



▼ Environmental □
○ ▶ Aquatic □
○ ▶ Terrestrial □
○ ▶ Air ☑
○ ▶ Unclassified □
▶ Host-Associated □
▶ Engineered □

Selected Ecosystems:
Freshwater
Marine
Soil
Air

☑ Apply association cut-off(%): 5

Submit Reset ☑ Run in new tab

Freshwater	Marine	Soil	Air	
2369	28966	51601	46	Freshwater
28966	14276	18011	45	Marine
51601	18011	16283	29	Soil
46	45	29	3	Air

Figure 4. The Ecosystem & Phylogeny Visualization tool allows users to create interactive plots showing the distribution of NMPFs in ecosystems or taxonomic groups. (**A**) Through the tool input form, users can select ecosystem or taxonomy categories and retrieve the NMPFs associated with them. (**B–D**) Examples of the plots generated by the analysis. (B) Circos plot, (C) color-coded matrix, (D) upset plot. All results can be downloaded as high-resolution images, as well as in tab-delimited format, for further analysis.

In all cases, users can perform searches against the entire NMPFamsDB or one of its subsets, based on their taxonomic or ecosystem associations. In addition, they can adjust sequence search parameters including cut-off types (Bit-score or *E*-value) and values, substitution matrices and gap costs for pairwise alignments, as well as inclusion and report threshold types and values for HMM-based searches. The results include the pairwise alignments between the query sequences and their hits, as well as their stats (Bit-score, *E*-value, alignment gaps etc.) and can be downloaded in parsable text format for further analysis. In addition, NMPFamsDB offers a *Pattern Search* tool (Supplementary Figure S9) enabling queries using sequence motifs, either in the established PROSITE (48) format or via regular expressions.

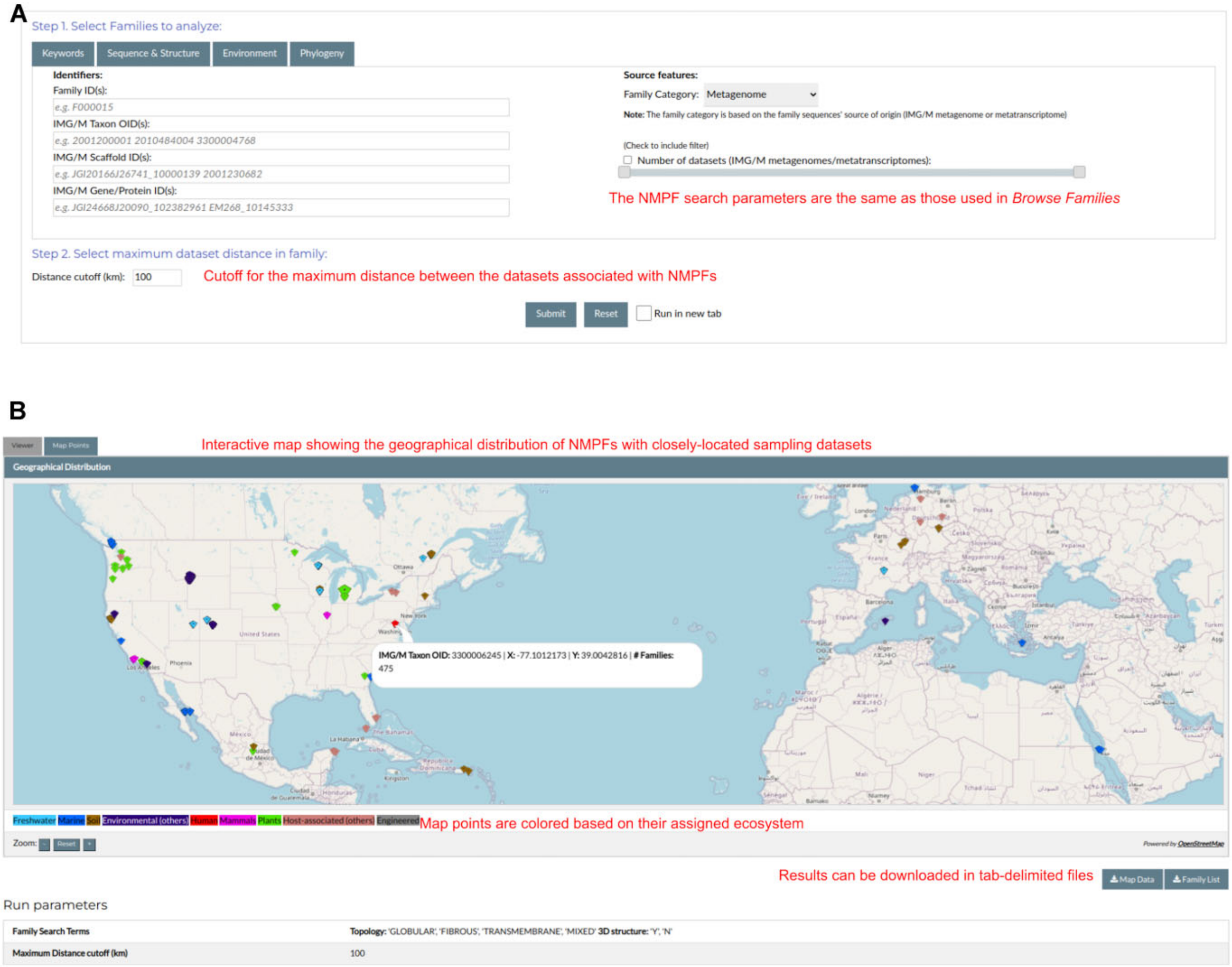
The *Data Visualization* tools offer interfaces for *Ecosystem & Phylogeny* and *Geographical Distribution visualization*, allowing users to explore the taxonomic, ecosystem and geographical associations and relationships of NMPFs based on their metadata. In the *Ecosystem & Phylogeny* tool (Figure 4), users can select NMPFs based on their association with organism categories or ecosystems at various levels, and create various types of interactive charts including color-coded matrices, Venn diagrams, pie charts, bar plots and Upset charts. Finally, the *Geographical Distribution* tool (Figure 5) offers the ability to select and visualize NMPFs based on the maximum distance among their associated metagenomic samples; this way, users can identify NMPFs that are limited to specific geographical locations.

Programmatic access

In addition to the web interface, NMPFamsDB offers an Application Programming Interface (API) for the automated retrieval of database components. The API can be accessed with both GET and POST requests and offers tools to search and retrieve data and metadata on NMPFs, metagenome datasets, sequencing scaffolds and 3D models. Results are returned in JSON format. A detailed description is offered in the *Programmatic Access* section of the NMPFamsDB website.

Discussion

Herein we have presented the initial version of the NMPFamsDB repository which hosts novel protein clusters from IMG metagenomes and metatranscriptomes with no hits to reference genomes or Pfams. Future versions of NMPFamsDB will contain NMPFs with <100 members as well as the equivalent clusters from the reference genomes for comparison purposes. Pipelines and services will be implemented for continuous sequence updates as they originate from the IMG/M platform and to support updates of viral scaffolds through the IMG/VR repository. While at the moment statistics and plots are generated on-the-fly, in future NMPFamsDB versions, several components will be replaced with precomputed results for speed optimizations. Overall, the growing number of metagenomic datasets and the continuous detection of new viral con-



Step 1. Select Families to analyze:

Identifiers:

Family ID(s): *e.g. F000015*

IMG/M Taxon OID(s): *e.g. 2001200001 2010484004 3300004768*

IMG/M Scaffold ID(s): *e.g. JGI20166J26741_10000139 2001230682*

IMG/M Gene/Protein ID(s): *e.g. JGI24668J20090_102382961 EM268_10145333*

Source features:

Family Category: Metagenome

Note: The family category is based on the family sequences' source of origin (IMG/M metagenome or metatranscriptome)

(Check to include filter)

□ Number of datasets (IMG/M metagenomes/metatranscriptomes):

The NMPF search parameters are the same as those used in *Browse Families*

Step 2. Select maximum dataset distance in family:

Distance cutoff (km): 100

Cutoff for the maximum distance between the datasets associated with NMPFs

□ Run in new tab

Interactive map showing the geographical distribution of NMPFs with closely-located sampling datasets

IMG/M Taxon OID: 3300006245 | **X:** -77.1012173 | **Y:** 39.0042816 | **# Families:** 475

Map points are colored based on their assigned ecosystem

Results can be downloaded in tab-delimited files

Run parameters

Family Search Terms	**Topology:** 'GLOBULAR', 'FIBROUS', 'TRANSMEMBRANE', 'MIXED' **3D structure:** 'Y', 'N'
Maximum Distance cutoff (km)	100

Figure 5. The geographical distribution tool allows users to explore the distribution of NMPFs with datasets located within a specific distance from each other. (**A**) Through the input form, users can select the set of families they wish to analyze and provide a cut-off (in kilometers) for the distance between the ED datasets of each family (here set at 100 km). Searches can be performed using the same options as in *Browse Families*. (**B**) The results are presented in an interactive world map, with map points corresponding to ED datasets. Each map point is colored based on its assigned ecosystem. The map points and a list of the analyzed families can also be downloaded for further analysis.

tigs together with the ongoing development of analysis and search capabilities within the IMG/M and IMG/VR systems will render NMPFamsDB a critical community resource for the study of microbial functional dark matter.

Data availability

NMPFamsDB is publicly available as a web service at http://www.nmpfamsdb.org/ or https://bib.fleming.gr/NMPFamsDB. The associated data of each NMPF, namely, its sequences, MSAs, HMM and, where applicable, 3D structure model, are accessible through a download form at the top of its respective *Family Entry* page. Bulk collections of the above data formats are also available for download through the database's *Downloads* page; these include the entire NMPFamsDB dataset as well as smaller subsets, based on sequence origin (*Metagenome-only*, *Metatranscriptome-only*, *Mixed*), ecosystem association (*Environmental*, *Host-associated*, *Engineered*) and taxonomic annotation (*Bacterial*, *Archaeal*, *Eukaryotic*, *Viral*, *Unclassified*).

Supplementary data

Supplementary Data are available at NAR Online.

Acknowledgements

This work used resources from the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the U.S. Department of Energy.

Author contributions: N.C.K. and G.A.P. conceived and supervised the project and performed sequence analysis and clustering. F.A.B. participated in sequence, structure, and metadata analysis and implemented the database. E.K. assisted with scripting, data visualization, and the implementation of the database. D.P.E ran the phylogenetic analyses. S.L. and S.O. performed 3D structure predictions with AlphaFold2 and analyzed the results. Y.S helped with sequence taxonomic assignment and classification. I.M.C. participated in data analysis and visualization and provided data and metadata from IMG/M. All authors participated in writing the manuscript. All authors have read and approved the final version of the manuscript.

Funding

HFRI (1st call of research projects to support faculty members and researchers, Grant: 1855-BOLOGNA); EU Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie [838018]; Fondation Sante; U.S. Department of Energy Joint Genome Institute (https://ror.org/04xm1d337), a DOE Office of Science User Facility, supported by the Office of Science of the U.S. Department of Energy, operated under Contract No. DE-AC02-05CH11231; Funding for open access charge: U.S. Department of Energy Joint Genome Institute (https://ror.org/04xm1d337), a DOE Office of Science User Facility, supported by the Office of Science of the U.S. Department of Energy, operated under Contract No. DE-AC02-05CH11231.

Conflict of interest statement

None declared.

References

1. Rinke,C., Schwientek,P., Sczyrba,A., Ivanova,N.N., Anderson,I.J., Cheng,J.-F., Darling,A., Malfatti,S., Swan,B.K., Gies,E.A., *et al.* (2013) Insights into the phylogeny and coding potential of microbial dark matter. *Nature*, **499**, 431–437.
2. Oulas,A., Pavloudi,C., Polymenakou,P., Pavlopoulos,G.A., Papanikolaou,N., Kotoulas,G., Arvanitidis,C. and Iliopoulos,l. (2015) Metagenomics: tools and insights for analyzing next-generation sequencing data derived from biodiversity studies. *Bioinform. Biol. Insights*, **9**, BBI.S12462.
3. Meyer,F., Bagchi,S., Chaterji,S., Gerlach,W., Grama,A., Harrison,T., Paczian,T., Trimble,W.L. and Wilke,A. (2019) MG-RAST version 4—Lessons learned from a decade of low-budget ultra-high-throughput metagenome analysis. *Briefings Bioinf.*, **20**, 1151–1159.
4. Ayling,M., Clark,M.D. and Leggett,R.M. (2020) New approaches for metagenome assembly with short reads. *Briefings Bioinf.*, **21**, 584–594.
5. Chen,I.-M.A., Chu,K., Palaniappan,K., Ratner,A., Huang,J., Huntemann,M., Hajek,P., Ritter,S.J., Webb,C., Wu,D., *et al.* (2023) The IMG/M data management and analysis system v.7: content updates and new features. *Nucleic Acids Res.*, **51**, D723–D732.
6. Richardson,L., Allen,B., Baldi,G., Beracochea,M., Bileschi,M.L., Burdett,T., Burgin,J., Caballero-Pérez,J., Cochrane,G., Colwell,L.J., *et al.* (2023) MGnify: the microbiome sequence data analysis resource in 2023. *Nucleic Acids Res.*, **51**, D753–D759.
7. Mistry,J., Chuguransky,S., Williams,L., Qureshi,M., Salazar,G.A., Sonnhammer,E.L.L., Tosatto,S.C.E., Paladin,L., Raj,S., Richardson,L.J., *et al.* (2021) Pfam: the protein families database in 2021. *Nucleic Acids Res.*, **49**, D412–D419.
8. Paysan-Lafosse,T., Blum,M., Chuguransky,S., Grego,T., Pinto,B.L., Salazar,G.A., Bileschi,M.L., Bork,P., Bridge,A., Colwell,L., *et al.* (2023) InterPro in 2022. *Nucleic Acids Res.*, **51**, D418–D427.
9. Galperin,M.Y., Wolf,Y.I., Makarova,K.S., Vera Alvarez,R., Landsman,D. and Koonin,E.V. (2021) COG database update: focus on microbial diversity, model organisms, and widespread pathogens. *Nucleic Acids Res.*, **49**, D274–D281.
10. Kanehisa,M., Sato,Y., Kawashima,M., Furumichi,M. and Tanabe,M. (2016) KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.*, **44**, D457–D462.
11. Azad,A., Pavlopoulos,G.A., Ouzounis,C.A., Kyrpides,N.C. and Buluç,A. (2018) HipMCL: a high-performance parallel implementation of the Markov clustering algorithm for large-scale networks. *Nucleic Acids Res.*, **46**, e33.
12. Eddy,S.R. (2011) Accelerated profile HMM searches. *PLoS Comput. Biol.*, **7**, e1002195.
13. Steinegger,M., Meier,M., Mirdita,M., Vöhringer,H., Haunsberger,S.J. and Söding,J. (2019) HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinf.*, **20**, 473.
14. Li,W., O'Neill,K.R., Haft,D.H., DiCuccio,M., Chetvernin,V., Badretdin,A., Coulouris,G., Chitsaz,F., Derbyshire,M.K., Durkin,A.S., *et al.* (2021) RefSeq: expanding the Prokaryotic Genome Annotation Pipeline reach with protein family model curation. *Nucleic Acids Res.*, **49**, D1020–D1028.
15. TensorFlow Developers (2022) TensorFlow.
16. Pedregosa,F., Varoquaux,G., Gramfort,A., Michel,V., Thirion,B., Grisel,O., Blondel,M., Prettenhofer,P., Weiss,R., Dubourg,V., *et al.* (2011) Scikit-learn: machine learning in Python. *J. Mach. Learn Res.*, **12**, 2825–2830.
17. Cock,P.J.A., Antao,T., Chang,J.T., Chapman,B.A., Cox,C.J., Dalke,A., Friedberg,I., Hamelryck,T., Kauff,F., Wilczynski,B., *et al.* (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423.
18. Torrisi,M., Kaleel,M. and Pollastri,G. (2019) Deeper profiles and cascaded recurrent and convolutional neural networks for state-of-the-art protein secondary structure prediction. *Sci. Rep.*, **9**, 12374.
19. Necci,M., Piovesan,D., Clementel,D., Dosztányi,Z. and Tosatto,S.C.E. (2020) MobiDB-lite 3.0: fast consensus annotation of intrinsic disorder flavors in proteins. *Bioinformatics*, **36**, 5533–5534.
20. Teufel,F., Almagro Armenteros,J.J., Johansen,A.R., Gíslason,M.H., Pihl,S.I., Tsirigos,K.D., Winther,O., Brunak,S., von Heijne,G. and Nielsen,H. (2022) SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nat. Biotechnol.*, **40**, 1023–1025.
21. Pasquier,C., Promponas,V.J. and Hamodrakas,S.J. (2001) PRED-CLASS: cascading neural networks for generalized protein classification and genome-wide applications. *Proteins*, **44**, 361–369.
22. Käll,L., Krogh,A. and Sonnhammer,E.L.L. (2007) Advantages of combined transmembrane topology and signal peptide prediction–the Phobius web server. *Nucleic Acids Res.*, **35**, W429–W432.
23. Tsirigos,K.D., Elofsson,A. and Bagos,P.G. (2016) PRED-TMBB2: improved topology prediction and detection of beta-barrel outer membrane proteins. *Bioinformatics*, **32**, i665–i671.
24. Bryant,P., Pozzati,G. and Elofsson,A. (2022) Improved prediction of protein-protein interactions using AlphaFold2. *Nat. Commun.*, **13**, 1265.
25. Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
26. Touw,W.G., Baakman,C., Black,J., te Beek,T.A.H., Krieger,E., Joosten,R.P. and Vriend,G. (2015) A series of PDB-related databanks for everyday needs. *Nucleic Acids Res.*, **43**, D364–D368.
27. Chandonia,J.-M., Guan,L., Lin,S., Yu,C., Fox,N.K. and Brenner,S.E. (2022) SCOPe: improvements to the structural classification of proteins - extended database to facilitate variant interpretation and machine learning. *Nucleic Acids Res.*, **50**, D553–D559.
28. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
29. Zhang,Y. and Skolnick,J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.*, **33**, 2302–2309.

30. Mukherjee,S. and Zhang,Y. (2009) MM-align: a quick algorithm for aligning multiple-chain protein complex structures using iterative dynamic programming. *Nucleic Acids Res.*, **37**, e83.
31. Mukherjee,S., Stamatis,D., Li,C.T., Ovchinnikova,G., Bertsch,J., Sundaramurthi,J.C., Kandimalla,M., Nicolopoulos,P.A., Favognano,A., Chen,I.-M.A., *et al.* (2023) Twenty-five years of Genomes OnLine Database (GOLD): data updates and new features in v.9. *Nucleic Acids Res.*, **51**, D957–D963.
32. Buttigieg,P.L., Pafilis,E., Lewis,S.E., Schildhauer,M.P., Walls,R.L. and Mungall,C.J. (2016) The environment ontology in 2016: bridging domains with increased scope, semantic density, and interoperation. *J. Biomed. Semantics*, **7**, 57.
33. Thompson,L.R., Sanders,J.G., McDonald,D., Amir,A., Ladau,J., Locey,K.J., Prill,R.J., Tripathi,A., Gibbons,S.M., Ackermann,G., *et al.* (2017) A communal catalogue reveals Earth's multiscale microbial diversity. *Nature*, **551**, 457–463.
34. Roux,S., Páez-Espino,D., Chen,I.-M.A., Palaniappan,K., Ratner,A., Chu,K., Reddy,T.B.K., Nayfach,S., Schulz,F., Call,L., *et al.* (2021) IMG/VR v3: an integrated ecological and evolutionary framework for interrogating genomes of uncultivated viruses. *Nucleic Acids Res.*, **49**, D764–D775.
35. Ren,J., Song,K., Deng,C., Ahlgren,N.A., Fuhrman,J.A., Li,Y., Xie,X., Poplin,R. and Sun,F. (2020) Identifying viruses from metagenomic data using deep learning. *Quant. Biol.*, **8**, 64–77.
36. Pronk,L.J.U. and Medema,M.H. (2022) Whokaryote: Distinguishing Eukaryotic and Prokaryotic Contigs in Metagenomes Based on Gene Structure. *Microb. Genom.*, **8**, mgen000823.
37. West,P.T., Probst,A.J., Grigoriev,I.V., Thomas,B.C. and Banfield,J.F. (2018) Genome-reconstruction for eukaryotes from complex natural microbial communities. *Genome Res.*, **28**, 569–580.
38. Delmont,T.O., Gaia,M., Hinsinger,D.D., Frémont,P., Vanni,C., Fernandez-Guerra,A., Eren,A.M., Kourlaiev,A., Agata,L., Clayssen,Q., *et al.* (2022) Functional repertoire convergence of distantly related eukaryotic plankton lineages abundant in the sunlit ocean. *Cell Genomics*, **2**, 100123.
39. Mirdita,M., Steinegger,M., Breitwieser,F., Söding,J. and Levy Karin,E. (2021) Fast and sensitive taxonomic assignment to metagenomic contigs. *Bioinformatics*, **37**, 3029–3031.
40. Suzek,B.E., Wang,Y., Huang,H., McGarvey,P.B. and Wu,C.H. (2015) UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, **31**, 926–932.
41. Wheeler,T.J., Clements,J. and Finn,R.D. (2014) Skylign: a tool for creating informative, interactive logos representing sequence alignments and profile hidden Markov models. *BMC Bioinf.*, **15**, 7.
42. Yachdav,G., Wilzbach,S., Rauscher,B., Sheridan,R., Sillitoe,I., Procter,J., Lewis,S.E., Rost,B. and Goldberg,T. (2016) MSAViewer: interactive JavaScript visualization of multiple sequence alignments. *Bioinformatics*, **32**, 3501–3503.
43. Zahn-Zabal,M., Michel,P.-A., Gateau,A., Nikitin,F., Schaeffer,M., Audot,E., Gaudet,P., Duek,P.D., Teixeira,D., Rech de Laval,V., *et al.* (2020) The neXtProt knowledgebase in 2020: data, tools and usability improvements. *Nucleic Acids Res.*, **48**, D328–D334.
44. Sehnal,D., Bittrich,S., Deshpande,M., Svobodová,R., Berka,K., Bazgier,V., Velankar,S., Burley,S.K., Koča,J. and Rose,A.S. (2021) Mol* viewer: modern web app for 3D visualization and analysis of large biomolecular structures. *Nucleic Acids Res.*, **49**, W431–W437.
45. Kiełbasa,S.M., Wan,R., Sato,K., Horton,P. and Frith,M.C. (2011) Adaptive seeds tame genomic sequence comparison. *Genome Res.*, **21**, 487–493.
46. Wickham,H. (2016) In: *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag, NY.
47. Sievert,C. (2021) Interactive Web-based Data Visualization with R, plotly, and shiny. *J. R. Stat. Soc.*, **184**, 1150.
48. Sigrist,C.J.A., de Castro,E., Cerutti,L., Cuche,B.A., Hulo,N., Bridge,A., Bougueleret,L. and Xenarios,I. (2013) New and continuing developments at PROSITE. *Nucleic Acids Res.*, **41**, D344–D347.

Received: August 14, 2023. **Editorial Decision:** August 22, 2023. **Accepted:** September 19, 2023