BMC Genomics

RESEARCH

Open Access

sORFdb – a database for sORFs, small proteins, and small protein families in bacteria

Julian M. Hahnfeld[1*], Oliver Schwengers[1], Lukas Jelonek[1], Sonja Diedrich[1], Franz Cemič[2] and Alexander Goesmann[1]

Abstract

Small proteins with fewer than 100, particularly fewer than 50, amino acids are still largely unexplored. Nonetheless, they represent an essential part of bacteria's often neglected genetic repertoire. In recent years, the development of ribosome profiling protocols has led to the detection of an increasing number of previously unknown small proteins. Despite this, they are overlooked in many cases by automated genome annotation pipelines, and often, no functional descriptions can be assigned due to a lack of known homologs. To understand and overcome these limitations, the current abundance of small proteins in existing databases was evaluated, and a new dedicated database for small proteins and their potential functions, called 'sORFdb', was created. To this end, small proteins were extracted from annotated bacterial genomes in the GenBank database. Subsequently, they were quality-filtered, compared, and complemented with proteins from Swiss-Prot, UniProt, and SmProt to ensure reliable identification and characterization of small proteins. Families of similar small proteins were created using bidirectional best BLAST hits followed by Markov clustering. Analysis of small proteins in public databases revealed that their number is still limited due to historical and technical constraints. Additionally, functional descriptions were often missing despite the presence of potential homologs. As expected, a taxonomic bias was evident in over-represented clinically relevant bacteria. This new and comprehensive database is accessible via a feature-rich website providing specialized search features for sORFs and small proteins of high quality. Additionally, small protein families with Hidden Markov Models and information on taxonomic distribution and other physicochemical properties are available. In conclusion, the novel small protein database sORFdb is a specialized, taxonomy-independent database that improves the findability and classification of sORFs, small proteins, and their functions in bacteria, thereby supporting their future detection and consistent annotation. All sORFdb data is freely accessible via https://sorfdb.computational.bio.

Keywords Small proteins, Protein families, Short open reading frames, SORF, Database, Bacteria

Background

A significant portion of bacterial proteins are well studied today, broadly available in public databases, and routinely annotated in newly sequenced genomes [1–3]. Despite these advancements, the exploration of small proteins of up to 100 amino acids (AAs), encoded by short open reading frames (sORFs), has been largely neglected, and they often have been disregarded as noise in eukaryotic and bacterial genomes [1, 4]. Following, we consider small proteins to be functional proteins with a length of 100 AA or fewer.

*Correspondence:
Julian M. Hahnfeld
julian.hahnfeld@computational.bio.uni-giessen.de
[1] Bioinformatics and Systems Biology, Justus Liebig University Giessen, Heinrich-Buff-Ring, Giessen 35392, Hesse, Germany
[2] Department of Computer Science, University of Applied Sciences Giessen, Gutfleischstrasse, Giessen 35390, Hesse, Germany

The application of various length cutoffs for the prediction and identification of protein sequences has led to an inconsistent definition of small proteins. Historically, these cutoffs have resulted from limitations in laboratory protocols and gene prediction tools to reliably detect proteins of such a small size. Gene prediction tools exhibit higher false-positive rates for smaller proteins, which are addressed by implementing strict length limits [5, 6]. Due to the high number of false-positive small proteins in early annotated genomes, minimum length cutoffs were implemented in genome databases [7], and previous small proteins thought to be coding had to be removed later [8].

However, in recent years, the development of experimental ribosome profiling techniques and improvements in mass spectroscopy have resulted in the detection of numerous small proteins [9–11]. Following the identification of small proteins, the elucidation of their purpose has revealed essential cellular functions, including regulatory proteins, membrane-associated or secreted proteins, toxin-antitoxin systems, stress response proteins, and various virulence factors [1, 12–18]. These prominent roles emphasize that the largely unexplored space of small proteins provides essential functions in bacteria. The best-studied bacterial organisms containing small proteins are model organisms and clinically relevant species like *Escherichia coli* and *Salmonella enterica*, in which many new proteins and their encoding sORFs have been reported [10, 18]. The most recently identified proteins in *E. coli* belonged to small proteins with up to 100 AA, particularly with 50 AA or fewer [19].

Consequently, genetic origins and underlying evolutionary mechanisms of small proteins still need to be better understood [17] as they tend to exhibit features differing from genes encoding for proteins longer than 100 AA in bacteria. In particular, start codon usage, ribosomal binding sites (RBSs), and composition biases can differ from longer coding genes [12, 17, 19–21]. These differences could stem from small proteins being developed through de novo gene origination, and their comparatively young evolutionary age is insufficient to show the typical organism-specific features of longer coding genes [17, 22]. Pervasive translation of sORFs is also possible [22, 23], although sORFs encoding functional small proteins are probably subject to codon bias [10, 21, 23]. Because of these differences, sORFs and small proteins in bacteria have been overlooked for a long time and are still underrepresented in public databases.

Clustering and identifying new protein families from protein sequences of average length has become a standard bioinformatic procedure. The Markov clustering algorithm [24] has been proven to be reliable for identification in general [25, 26]. However, small proteins challenge existing clustering approaches and tools due to their short length. A metagenomic study has shown vast numbers of hitherto unknown small proteins in human microbiomes and protein families identified by clustering [13].

Addressing these issues, we present sORFdb, to our knowledge the first dedicated database for small proteins and sORF sequences in bacteria. It is a high-quality repository for known sORF and small protein sequences. In addition to protein sequences, physicochemical features are provided to support the search for small protein groups of interest. Furthermore, it offers small protein families and hidden Markov models, enabling the consistent identification and annotation of these families and providing entry points for further research. All data of sORFdb are publicly available for download and can be accessed via an interactive website at https://sorfdb.computational.bio.

Methods and implementation

Creation of a small protein and sORF database for bacteria

For the creation of the sORFdb database, genomes and protein sequences from various data sources were downloaded and processed. From GenBank (Release 256) [27], the 269,214 latest annotated genomes with an assembly level of "complete genome", "chromosome", or "scaffold" were downloaded and used as the primary source for sORFs and small proteins. Small proteins up to 100 AA in length were retrieved from the UniProt database (v2023_03) [2]. Curated and non-fragmented small proteins were downloaded from Swiss-Prot [2], and non-fragmented ones with evidence of existence at the protein, transcript, or homology level were downloaded from UniProtKB [2]. In addition, small proteins of the SmProt database (v2.0) [28] were retrieved and stored with the entries from the UniProt databases in a dataset of verified small proteins. These were directly added to the sORFdb database. For filtering and identification steps of hypothetical proteins and small proteins from an unknown annotation source, the UniRef100 entries [2] of the proteins with evidence and the SmProt proteins were used. Hidden Markov models (HMMs) from AntiFam (v7.0) [8] and Pfam (v35.0) [29] were downloaded, compressed with HMMER (v3.3.2) [30] and used for filtering and scanning for protein domains and motifs.

To extract sORFs and small proteins, several filtering and processing steps were applied to the aforementioned annotated bacterial genomes from GenBank. Complete, unambiguous, and unfragmented sORF and protein sequences and their functional product descriptions were extracted from the annotated genomes. False-positive small proteins were filtered out using PyHMMER (v0.9.0) [31] and AntiFam HMMs [8] with gathering cutoffs and

an additional upper E-value threshold of 1E-5. RBSs of sORFs were detected using Pyrodigal (v2.1.0) [32]. In addition, extracted small proteins were filtered according to whether their annotation source was from a known trusted source, i.e., a reference, representative or NCBI Prokaryotic Genome Annotation Pipeline (PGAP) annotated genome, or an unknown source. Non-hypothetical proteins from trusted annotation sources were stored in the sORFdb database. Additionally, hypothetical proteins and proteins from unknown annotation sources were compared against the dataset of verified small proteins from the Swiss-Prot, UniProt, and SmProt databases using Diamond (v2.1.8) [33].

To examine homology, BLAST Score Ratio Values (SRV), as proposed by Lerat et al. [34], were calculated by normalizing bit scores of the best-observed alignment hits with the maximum bit scores of protein self-hits (*Observed score/Maximum score*). This normalization is used because common E-value or bit score thresholds are often too strict for small proteins due to their short length. As a relative metric, SRVs contain sequence length information, while SRV thresholds, are independent of sequence length. In contrast to E-value thresholds, SRV thresholds generalize well across proteins of different lengths and do not need to be more lenient for small proteins. All homology-filtered small proteins with an SRV of 0.7 or higher were stored in the sORFdb database. To detect very small proteins with only up to 50 AA length, potentially missed by the original genome annotation, a combined approach using Pyrodigal [32], and a homology search with a minimum SRV threshold of 0.7 was employed. A subsequent filtering step excludes all hits overlapping with existing annotations and filters for canonical start codons.

For all small proteins, physicochemical properties were calculated using Biopython (v1.8.1) [35] and Peptides.py (v0.3.1) [36]. Additionally, they were screened for Pfam families and domains with gathering cutoffs and an additional upper E-value threshold of 1E-5. The taxonomy of all small proteins was adapted to the nomenclature for phyla described by Oren and Garrity [37].

The workflow for the creation of sORFdb was implemented in Nextflow (v23.04.1) [38] to achieve an automated and reproducible procedure (Fig. 1). The supplemental materials provide all software packages and tools used, along with their respective versions (Suppl. Tab. S1-3).

Clustering of potential small protein families

sORFdb provides potential small protein families along with corresponding HMMs. Due to their short length and understudied clustering properties, a custom graph-based clustering approach was developed to find potential protein families (Suppl. Fig. S1). To reduce the graph size and focus on less studied small proteins, only non-redundant small proteins with 50 AA or fewer were clustered.

During the initial step of the clustering approach, an all-against-all BLAST search was performed using BLAST+ (v2.14.1) [39]. SRVs were calculated for all BLAST hits, and a lower limit of 0.3, as proposed by Lerat et al. [34], was applied to identify possible homologs. In addition, a minimum mutual alignment coverage of at least 70.0 % was required to obtain sequence alignments of higher quality and to exclude artifacts of small proteins only sharing a few AA. Afterward, singletons comprising proteins without homologs and distant hits with only one alignment with another protein were excluded. The SRVs of the BLAST hits were transformed into a symmetric undirected graph. For this purpose, SRVs were averaged with their reverse BLAST hit.

Small proteins represent nodes and SRVs weighted edges in the graph. To reduce the node degrees in the graph and improve the clustering performance, only k best edges of a node were kept. For this purpose, k was chosen as the minimum number of edges of a node without creating singletons in the graph. The previous pruning steps exclude distant proteins that otherwise lead to singletons at high values of k. Therefore, k was chosen as the smallest possible value for which no singletons were reported.

Afterward, to remove edges that could lead to incorrect clusters, a heuristic proposed by Apeltsin et al. [40] was applied to every graph component, consisting of a connected subgraph of proteins unconnected to every other connected subgraph.

The pruned graph was then split into batches of components with similar properties depending on the component's mean node degree and mean edge weights. This was done to improve the selection of the inflation parameter value, which controls the granularity of the clustering. For all batches, a Markov clustering with different inflation values, between 1.2 and 4.0, was computed using MCL (v22-282) [24]. The inflation value was chosen based on the efficiency criterion [41]. A visualization of the clusters used for family identification is available in the supplemental materials (Suppl. Fig. S2).

For all clusters with more than five members, multiple sequence alignments were computed using MUSCLE (v5.1) [42]. Based on these alignments, HMMs were built using PyHMMER (v0.10.2) [31]. Gathering cutoff values were computed for all HMMs, and where possible, a protein product was assigned by a major voting decision based on the annotated protein functions in sORFdb.

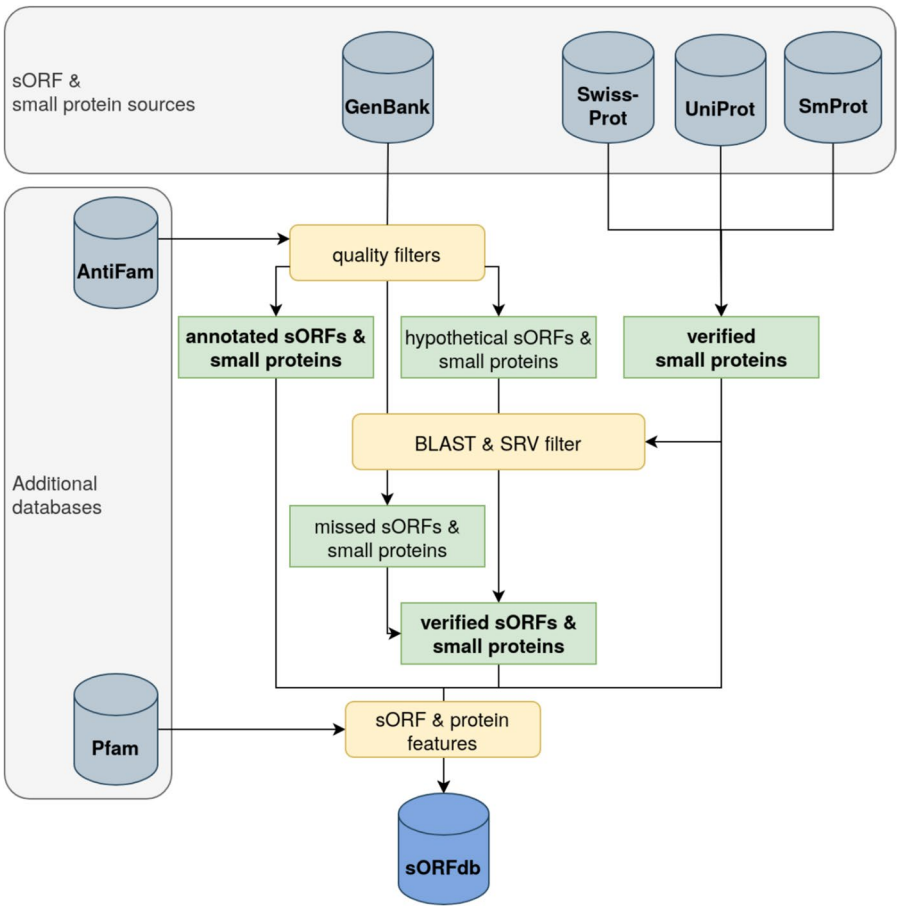

Fig. 1 Scheme of the data processing and sORFdb compilation workflow

Annotated genomes from GenBank were quality-filtered for complete and unambiguous sORFs and their annotation source. Small proteins with evidence were retrieved from the Swiss-Prot, UniProt, and SmProt databases and used for sORFdb and additional quality filtering steps. Hypothetical proteins and small proteins from an unknown source were filtered using Score Ratio Value cutoffs based on normalized bit scores. Similarly, missing sORFs were identified. Spurious small proteins were filtered out using AntiFam. Pfam families and domains were assigned, and physicochemical properties were calculated for all small proteins

sORFdb website

The data of sORFdb is stored in an Elasticsearch cluster [43]. Access to the data is provided via a REST API that was implemented in Java with the Vert.x framework [44]. The website's graphical user interface was implemented using Vite, Vue, and Typescript [45–47]. It provides a function for an exact sequence and an ID search using the API above to match the queries against the sequences and IDs stored within the Elasticsearch cluster. The alignment-based search uses a BLAST SequenceServer [48] with all stored small proteins as a database returning the IDs of the matching subject sequences which then are used for a search in the Elasticsearch cluster. The small protein family search is performed on the server-side using HMMER (v3.3.2), and the IDs of matching HMMs are also used to search for the HMM entries in the Elasticsearch cluster. For all entries cross-links to the original data sources are provided.

The web-frontend, the Elasticsearch server, and the BLAST SequenceServer are deployed on a scalable Kubernetes cluster, which is hosted in the de.NBI consortium's cloud computing infrastructure.

Results

Small proteins encoded by sORFs have long been overlooked in bacteria due to laboratory and computational limitations. With the advent of new laboratory protocols, many small proteins with essential functions have been reported. Despite these advancements and improvements in gene prediction tools, sORFs and small proteins still need to be explored. There are no dedicated databases

that focus solely on small proteins in bacteria. sORFdb was created to provide a taxon-independent collection of high-quality bacterial sORFs, small proteins and their assignments to protein families in a comprehensive database to address this issue. This resource is accessible via https://sorfdb.computational.bio.

A large-scale collection of small proteins

To capture bacterial sORFs and small proteins in a comprehensive, taxonomically independent, and standardized manner, including potential unannotated sequences, they were collected from the public data sources GenBank, Swiss-Prot, UniProt, and SmProt, enriched with additional information on taxonomy and RBS usage and processed into a dedicated database. As the public databases contain sequences of varying-quality, all sequences were quality-filtered in a workflow before acceptance.

A total of 31,653,437 sORFs and 34,007,166 small proteins were collected from public databases. 269,214 annotated bacterial genomes from GenBank were systematically screened for sORFs and small proteins, and different filtering steps were applied to extracted sequences. The filtered proteins were split into two groups. The first consists of non-hypothetical small proteins stemming from a trusted annotation source. The second group contains hypothetical ones or ones stemming from an unknown annotation source. From the first group, 22,846,872 annotated small proteins were included in sORFdb. An additional homology-based filtering step was applied to the second group. 8,596,036 small proteins were kept after applying a strict SRV filter since they possessed a homolog with an SRV of 0.7 or higher to sequences from Swiss-Prot, UniProt, or SmProt. For 2,722,346 small proteins previously annotated as "hypothetical protein", the product description could be updated using information from well-annotated homologs. From UniProt, 2,322,213 small proteins with evidence on transcript, protein, or homology level were collected.

Since automated annotation pipelines rely primarily on computational gene prediction tools, they are limited by hard length cutoffs, and sORFs encoding small proteins may be missed despite possible homologs. To collect these in the annotated genomes, they were detected using a combined approach comprising Pyrodigal, a homolog search, and an overlap and start codon filter. Based on homology alone, 1,363,907 potentially missing small proteins could be detected. After applying the filter mentioned above, a further 198,723 were stored in the sORFdb database. As a result, sORFdb contains 5,073,415 non-redundant small protein sequences and 5,640,450 non-redundant sORF sequences. Despite the absence of sORF sequences for some of the small proteins collected, the total number of non-redundant proteins is smaller than that of non-redundant sORFs due to the use of synonymous codons. Detailed information on the numbers of the total and unique sORF and small protein sequences, as well as related database sources, are shown in Table 1.

The group of proteins for which the most new proteins have been reported in recent years is the group of small proteins with up to 50 AA [19]. To investigate the length distribution of the total and non-redundant small proteins in sORFdb, their lengths were compared with entries in the UniRef100 database.

In line with expectations, the number of known small proteins in the sORFdb and the UniRef100 database tremendously declines with decreasing length (Fig. 2). The historical and the default gene length cutoffs of standard gene prediction tools and databases (30 AA, 38 AA, and 60 AA) are visible for the predicted UniRef100 entries [5, 7]. In contrast, these cutoffs do neither occur for UniRef100 entries with evidence nor sORFdb entries. While less numerous than all predicted UniRef100 entries, sORFdb provides a considerably higher number of non-redundant small proteins, especially with fewer than 50 AA, than the UniRef100 entries with evidence.

Taxonomic distribution

Based on the literature and reports for newly identified small proteins in clinically relevant species, we suspected a bias in the taxonomic distribution in our sORFdb database. For this reason, the taxonomy information of sORFs and small proteins was extracted from source genomes and databases to assess the spread and conservation across different bacterial taxa. Phyla were standardized to use a consistent nomenclature based on Oren and Garrity [37].

In line with our expectation, the taxonomic distribution of small proteins in the sORFdb database showed

Table 1 Number of sORFs and small proteins in sORFdb and the used public data sources

Database		sORFs	Small proteins
GenBank	total	31,641,552	31,641,552
	non-redundant	5,628,909	4,366,039
Swiss-Prot	total	-	30,520
	non-redundant	-	19,718
UniProt	total	-	2,322,213
	non-redundant	-	1,612,347
SmProt	total	11,885	12,881
	non-redundant	11,858	12,419
sORFdb	total	**31,653,437**	**34,007,166**
	non-redundant	**5,640,450**	**5,073,415**

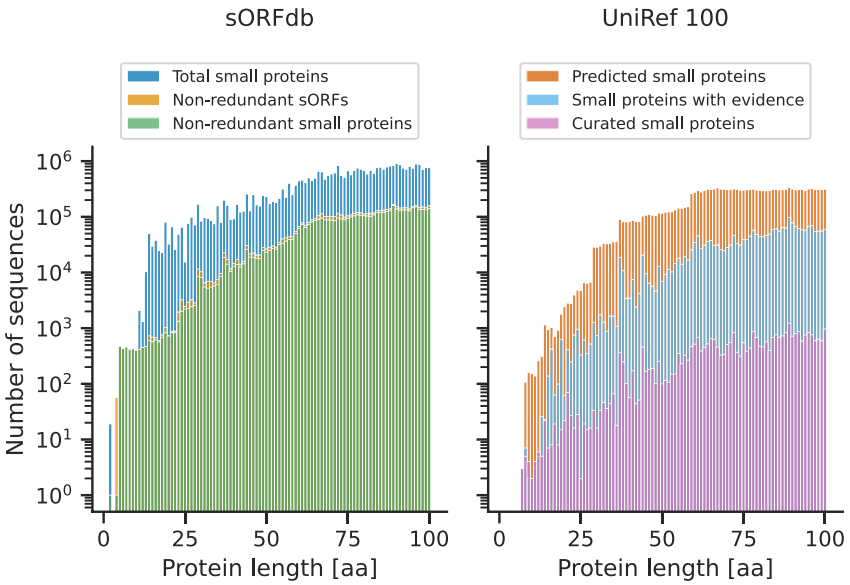

Fig. 2 Length distribution of sORFs and small proteins in sORFdb and UniRef100

The number of known small proteins decreases with decreasing sequence length. sORFdb provides more non-redundant small proteins than the UniRef100 database with evidence. Especially for sORFs encoding small proteins with few AA, sORFdb provides more entries

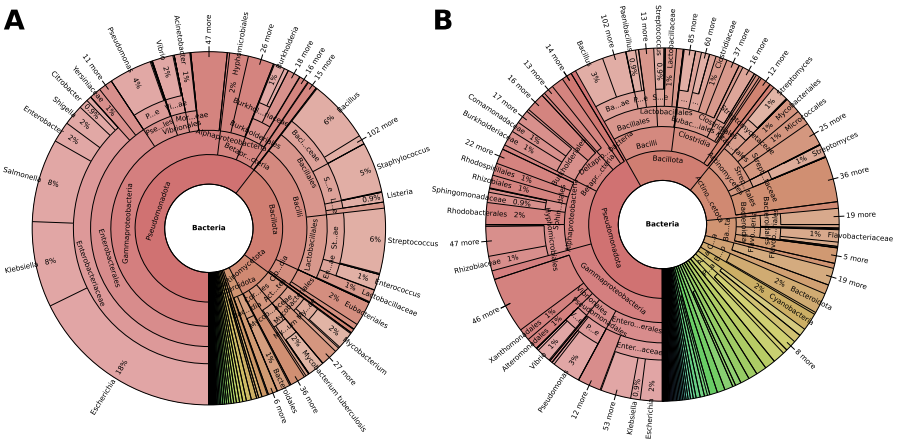

Fig. 3 Taxonomic distribution of redundant and non-redundant small proteins

A Most known small proteins in sORFdb stem from Pseudomonadota , model and clinically relevant organisms. **B** The taxonomic distribution of non-redundant small proteins showed a much less pronounced bias in comparison. The figure was created with Krona (v2.8.1) [49]

a clear overrepresentation of clinically relevant species and model organisms (Fig. 3A). 60.0 % of all small proteins belonged to the phylum of *Pseudomonadota*, formerly known as *Proteobacteria*. Within this phylum, 34.0 % of all protein entries belonged to *Escherichia*, *Klebsiella*, and *Salmonella* genera. Other dominantly represented genera were *Pseudomonas*, *Bacillus*, *Staphylococcus*, and *Streptococcus*, each accounting for 4–6 %. This bias is much less prominent in the taxonomic distribution of non-redundant small proteins (Fig. 3B). While 42 % of all non-redundant small proteins are also found in *Pseudomonadota*, there are no overrepresented genera, as is the case for all entries in the database.

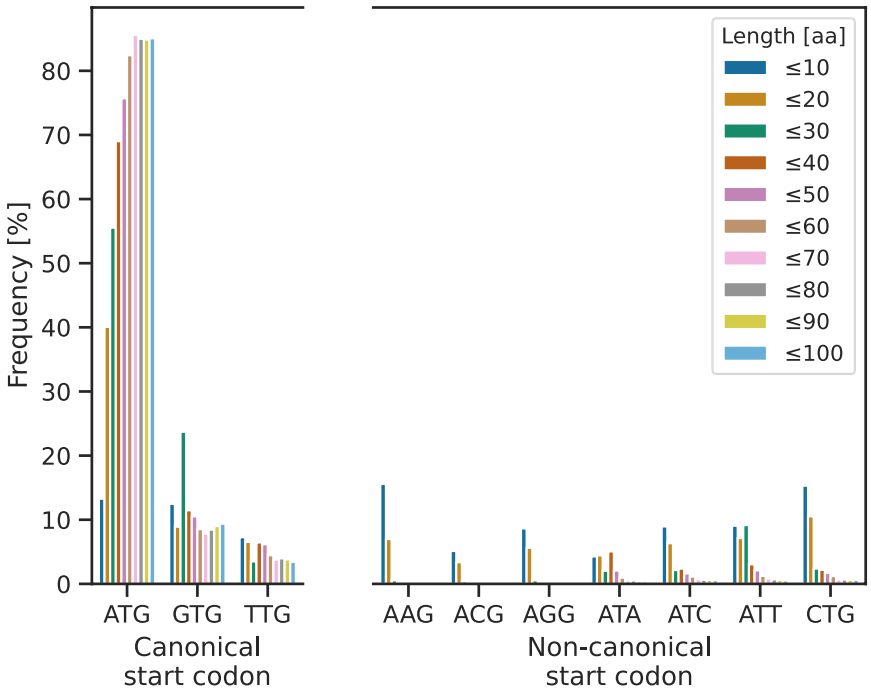

Fig. 4 Distribution of canonical and non-canonical start codons in sORFs

With decreasing sORF length, the frequency of non-canonical start codons increases. Shorter sORFs have a higher frequency of the alternative canonical start codons GTG and TTG. The ones encoding small proteins with 20 AA or fewer have the highest frequency of non-canonical start codons. In addition, they show a shift towards different start codons compared to the non-canonical ones used in the group with more than 20 AA

Differing genetic properties between sORFs and longer genes

sORFs are known to have non-canonical start codons more often than genes encoding proteins with more than 100 AA [20]. To determine their start codon usage, these were extracted from all non-redundant sORF sequences in sORFdb. With decreasing length, there was a frequency increase in non-canonical start codons, while canonical start codons were most frequently used for all sORFs encoding small proteins of more than 20 AA (Fig. 4). Although ATG was the most frequent canonical start codon, the frequency of the alternative canonical start codons GTG and TTG increased with decreasing sequence length. sORFs encoding small proteins of 20 AA or fewer had a high proportion of non-canonical start codons compared to longer ones. The codons AAG, ACG, and AGG occur much more frequently in these than in sORFs encoding for small proteins with more than 20 AA. The non-canonical start codons ATA, ATC, ATT, and CTG occurred primarily in these longer sORFs. Regarding their source databases and genera, 73.4 % of the sORFs encoding small proteins with 20 AA or fewer belonged to the genus *Escherichia*. 68.9 % of these sORFs were collected from SmProt, while the remaining sequences were obtained from the GenBank database. In the group with up to 10 AA, 99.8 % of the sequences were annotated in the genus *Escherichia*, and 99.6 % of them were extracted from the SmProt database.

sORFs are known to be pervasively transcribed, which can happen through leaderless translation [22, 23]. To analyze potential leaderless translation, the usage of RBSs of the non-redundant sORFs in the annotated genomes from GenBank was examined using Pyrodigal [32]. RBS could be detected in 73.8 % of all non-redundant sORFs. Despite this fact, the existence of an RBS varies enormously depending on the sequence length. With decreasing size, the detection of an RBS decreased (Fig. 5). This can be observed for sORFs encoding small proteins with 60 AA or fewer. For all sORFs encoding small proteins with 10 AA or fewer, no RBSs could be detected, and 87.5 % of the ones encoding small proteins with 20 AA or fewer also did not have a predicted RBS.

Functions of small proteins

Functional characterizations of small proteins have revealed essential roles in bacteria. However, homologs and functional descriptions are often unavailable for newly discovered small proteins. For this reason, all small proteins were filtered during the sequence collection process, and hypothetical protein products were

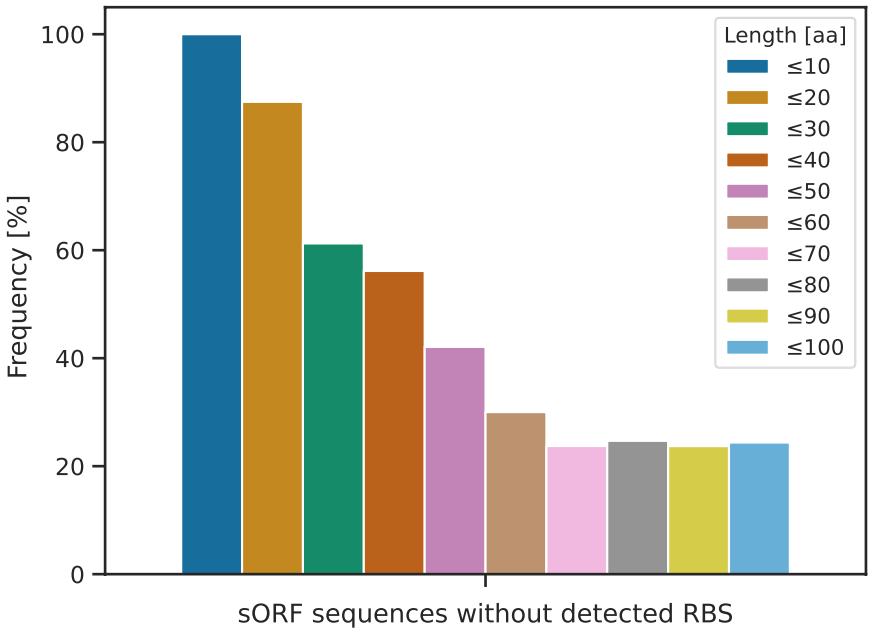

Fig. 5 Frequency of sORFs without a detected ribosomal binding site

With decreasing sORF length, the frequency of detected RBSs also decreased, and for sORFs encoding proteins with 10 AA or fewer, no RBSs were found at all

re-annotated with functional descriptions of homologs, if available. In addition, all small proteins were queried against the Pfam database to assign protein families and domains.

The most common functional descriptions and Pfam hits were analyzed to investigate whether the functions of the collected small proteins were consistent with the literature. For 74.0 % of the non-redundant small proteins, a Pfam family or domain could be assigned, with the number of known assigned Pfam domains decreasing with decreasing sequence length. Most small proteins in sORFdb are structural proteins of ribosomes that are highly conserved and well-studied. Besides these, essential functions of regulatory proteins, stress response proteins, and toxin-antitoxin systems are predominant. This is in line with previous reported findings [1, 14, 15, 17, 19]. Most regulatory proteins are denoted as helix-turn-helix containing transcriptional regulators. Cold-shock proteins are the most abundant stress response proteins, and the three most common toxin-antitoxin systems are Type II toxin-antitoxin systems of the RelE/ParE, HicA or Phd/YeFM families. The top 20 small protein product annotations are available in the Supplementary Material Table S4. Small proteins with 50 AA or fewer also frequently possess the functional description for helix-turn-helix containing transcriptional regulators. However, many of these are membrane-associated, like the yjcZ family sporulation protein, lmo0937 family protein ATPase subunits, and others. The most common toxin-antitoxin systems are entericidin family proteins. This has also been reported in previous studies [1, 12, 17, 19]. In addition, small proteins with the domain of unknown function DUF3265, DUF2256, or DUF1127 are also frequently found in the annotations of sORFdb and the assigned Pfam domains.

Families of small proteins in bacteria

While small proteins in bacteria are a rapidly evolving field of research and the number and deduced functions of novel identified proteins are the subject of current studies, their families are still understudied [13]. Clusters of similar proteins can be used as a starting point to identify conservation within and across taxonomic groups and the evolution of beneficial functions of these proteins. To address this, potential families were inferred using a custom graph-based clustering approach on the non-redundant collection of bacterial small proteins.

Small proteins with up to 50 AA are the most rapidly growing protein category [19], and longer proteins have been better studied since they were not affected by historical cutoffs [5–7]. For this reason, the clustering and the small protein families focused on the 309,042 less-studied small proteins with up to 50 AA. The clustering approach was developed to handle small protein sequences' properties better and make minimal assumptions about their clustering behavior. After applying

different pruning strategies, 272,018 small proteins remained for the clustering with MCL. 16,518 clusters were assigned in total by the graph-based clustering approach, of which 4,073 were singletons. Clusters with at least five members were used as the basis for the small protein families. Many of the separate graph components were completely assigned to one cluster. There were also large complex-structured graph components, consisting of proteins such as ribosomal or small proteins sharing a domain of unknown function, for which many clusters were determined (Suppl. Fig. S1).

Clusters with at least five members were denoted as small protein families to distinguish between technical sequence clusters and potential families. Based on this, 8,884 novel small protein families were created. These families had a mean of 27.7 and a median of 11 members. The most prominent family consisted of 363 members. Most families had members with a length between 40 and 50 AA. Additionally, there were more families with sequences of approximately 38 AA in length. While there was a slightly increased number of families with members of around 30 AA in length, families with shorter members were scarce (Fig. 6). HMMs were built for all small protein families, and accompanying gathering cutoffs were calculated to foster the detection and annotation of small proteins belonging to the identified families. For 8,798 of the 8,884 families, a functional description could be assigned using a majority voting approach based on the existing functional annotation of cluster members. For example, these families shared simple protein motifs, such as a domain of unknown function or a functional description. The most abundant functional descriptions assigned to the families stemmed from ribosomal proteins and regulatory proteins.

Interactive web-based access to sORFdb

An interactive website was developed to provide a user-friendly interface for the sORFdb database and to integrate additional services for sequence and family search and data exploration. It makes the collected sORFs, small proteins, small protein families, and related information easily accessible to the scientific community. To accomplish this, it offers various functions for the interaction with the collected data. The database, protein and sORF

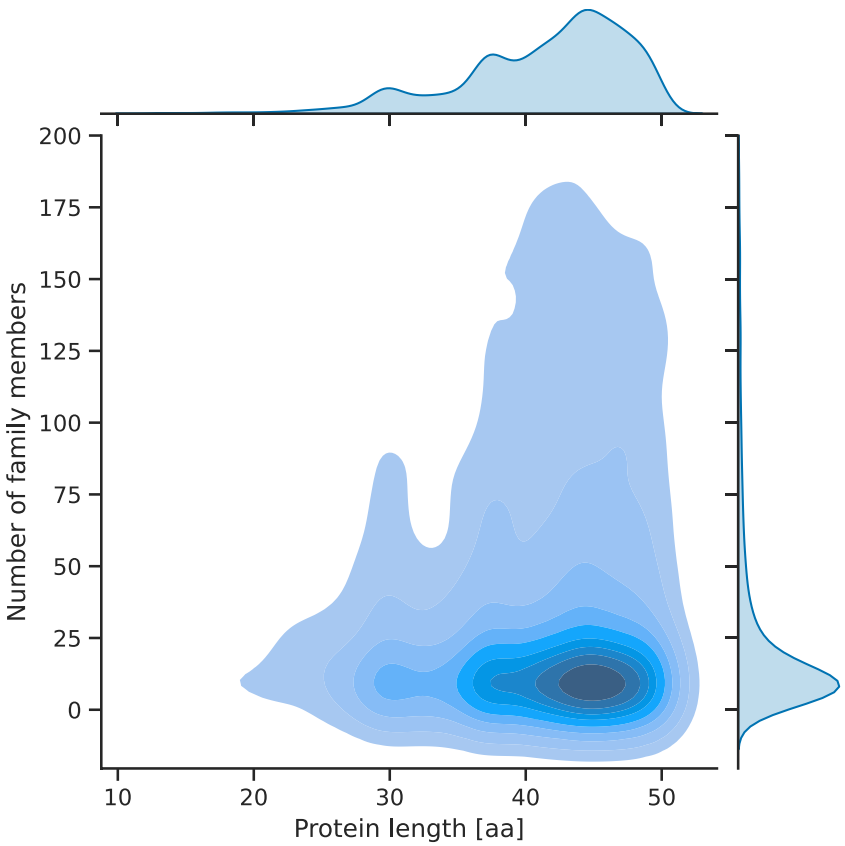

Fig. 6 Distribution of cluster size and sequence length of small protein families

Most small protein families had an average member length between 40 and 50 AA or around 30 AA. Most of the families with shorter member proteins had members with a length of about 30 AA

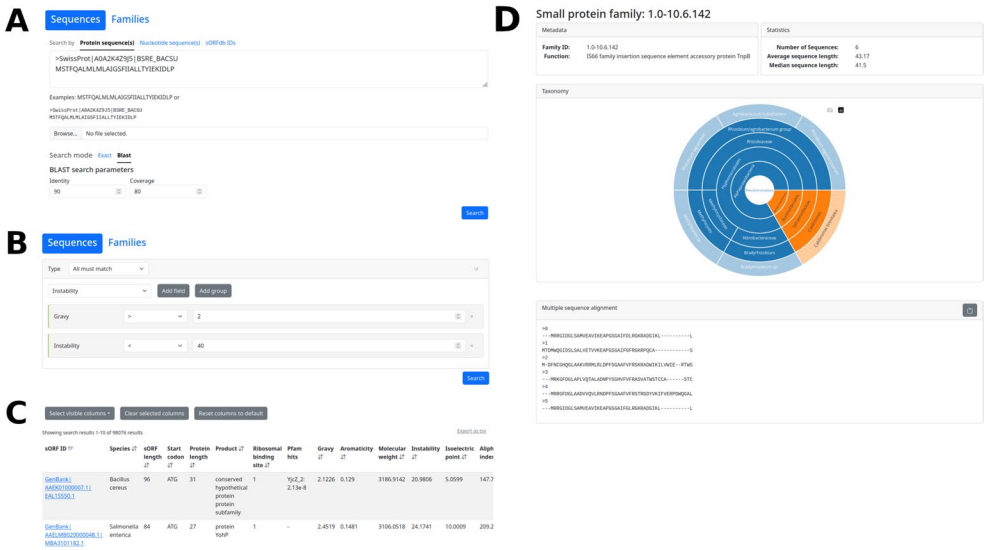

Fig. 7 Screenshots of the sORFdb website

A The sORFdb website' search page offers a fast exact and a BLAST-based sequence search for all sORFs and small proteins in the database. **B** The browse page provides an interactive selection for taxonomy, sequence-based features, and physicochemical properties to view matching protein entries in the database. **C** The matching entries of a search or browse query are made available in a downloadable table. The identifier of each entry links to a detail page with additional information. **D** The small protein family detail page provides information about family function, member sequences, taxonomy and the multiple sequence alignment

sequence data, small protein families and the corresponding HMMs are available for download.

To enable researchers to find homologous sORF and small protein sequences, a sequence-based search function is provided with a fast, exact search and a similarity-based BLAST search (Fig. 7A). In addition, a highly sensitive search for small proteins belonging to known protein families is available. All sequences and families are findable and accessible via unique IDs. Besides sequence-based search functions, browse functions are provided to view small proteins and sORFs matching user selected criteria. These criteria can be based on taxonomy, sequence features, functional description, or physicochemical properties (Fig. 7B). Search results and filtered sORFdb entries can be downloaded for local processing (Fig. 7C). To provide further information, links to original resources are provided on a detailed page for each database entry. Similarly, the small protein families can be browsed and inspected in a detail view (Fig. 7D).

Discussion

Small proteins of 100 AA or fewer encoded by sORFs have long been overlooked in bacteria [1, 4]. However, the advent of ribosome profiling, improvements in mass spectrometry, and metagenomics have led to the identification of numerous sORFs and small proteins [9–11, 13]. Despite this, sORFs and small proteins with evidence on transcript or protein level are often missing from public databases and newly annotated genomes. To address these issues, the landscape of publicly available bacterial sORFs and small proteins in annotated genomes and protein databases was captured and analyzed to provide a unified, dedicated database for sORFs, small proteins, and their families.

To the best of our knowledge, sORFdb is currently the largest and most comprehensive sequence database for bacterial sORFs, small proteins, and related families. The combination of different data sources, particularly the integration of GenBank and the application of filtering steps, provides access to a broad collection of sequences and a higher number of sequences than individual data sources used. This aggregation also enables the access to sORFs and small proteins from databases which do not provide direct access for these or have a different focus. These include GenBank, which is a database of nucleotide sequences with supporting biological annotations, as well as Swiss-Prot and UniProt, which contain protein descriptions including functional descriptions, but do not focus on high-quality small proteins. Due to this approach, sORFdb is taxonomically independent and not focused on specific species compared to the specialized small protein database SmProt, which focuses on *E. coli* and ribosome profiling [28]. In addition to the small protein and encoding sORF sequences, information on RBS usage and physicochemical properties are provided. Most importantly, sORFdb defines families for small bacterial

proteins to facilitate the consistent identification of these hard-to-predict proteins as a starting point for further studies. Regarding the functions of small proteins, the annotated functions of the non-redundant proteins in sORFdb largely coincide with the literature. As expected, the most frequently found proteins are ribosomal. The following annotated top functions of regulatory proteins, membrane-associated proteins, stress response proteins and toxin-antitoxin systems of small proteins are also often described in the literature [1, 12–15, 18]. Besides these functions, the Pfam domains DUF3265, DUF2256, and DUF1127 occur with high frequency in small proteins with up to 50 AA. Recent studies show that proteins with the DUF1127 domain serve essential functions concerning the sRNA maturation and RNA turnover as well as the phosphate and carbon metabolism [50, 51]. Based on the DUF1127 small proteins, 37 different families with DUF1127 were identified. In contrast, the 41,097 singletons and 3,561 clusters with up to 4 small proteins identified with the clustering approach show the existence of less conserved or understudied groups. This is also consistent with reports of small proteins conserved in only a few organisms [1]. While the functional annotations of well-studied small proteins in sORFdb align with the known literature, further investigation is needed for less conserved and understudied groups.

The taxonomic distribution of the available small proteins in sORFdb shows a clear bias towards Pseudomonadota, particularly towards *E. coli* (Fig. 3A). This bias is due to the historical over-representation of these bacteria in sequenced genomes and the fact that most ribosome profiling experiments are conducted in this organism [10, 18]. Although this bias is not as evident for the non-redundant small proteins (Fig. 3B), it still imposes limitations. Combining different databases did not reduce this known bias, and SmProt, which only contains data from *E. coli*, further reinforced this effect. Nevertheless, the taxonomic distribution of non-redundant small proteins shows that sORFdb covers a broad range of non-clinically relevant bacteria despite this bias, allowing a taxon-independent search for small proteins.

Since automated annotation pipelines and gene prediction tools are limited in their ability to predict sORFs, a homology-based approach was used to detect potential missing sORFs in annotated genomes from GenBank. Identifying missing small proteins was based on assumptions derived from the current knowledge of sORFs encoding functional proteins, which is biased towards *E. coli* and related bacteria. Therefore, only potentially missing small proteins with canonical start codons, a known homolog and prediction with Pyrodigal were included [10, 12, 21, 23]. Applying this search filter, 198,723 likely non-spurious small proteins were identified from the 1,363,907 candidates found by homology search. This comparatively low number, in conjunction with the filters applied, and the matching taxonomic bias of the database, indicates that more than a homology search is needed for identifying small proteins. Another limitation is the computational prediction of sORFs since the used tools are not optimized for sequences of such short lengths.

The distribution of start codons in non-redundant sORFs contrasts with the criterion for using canonical start codons that was applied for missing sORFs in the GenBank genomes. As the length decreases, the number of non-canonical start codons increases and shifts towards different non-canonical start codons compared to sORFs that encode small proteins with more than 20 AA (Fig. 4). A possible reason could be that most sORFs of this length were collected from SmProt and identified by ribosome profiling in *E. coli* alone. Therefore, they might include sORFs encoding non-functional transcripts expressed by pervasive translation [21, 22]. Alternatively, these sORFs may be evolutionary young, stemming from de novo gene origination, and therefore may not exhibit the typical start codon and RBS usage observed in *E. coli* [17]. To address this, further studies on the codon usage of sORFs encoding functional small proteins are needed to distinguish spurious sORFs expressed by pervasive translation from sORFs stemming from de novo gene origination and conserved sORFs.

The selection of an appropriate identity threshold is critical for the clustering of homologous protein sequences. If prior knowledge is available, a suitable threshold can be chosen depending on the evolutionary distance or available information about the composition of the protein families to be clustered. Otherwise, a 30 % identity threshold or the application of E-value or bit score thresholds have been shown to capture more distant homologs [52]. However, this approach cannot be applied to small proteins because for short sequences, even self-hits can have values outside these established thresholds [52]. For this reason, a custom graph-based clustering approach was used to identify small protein families. This approach was chosen to minimize assumptions about the clustering behavior of small proteins as much as possible since their clustering properties are not well known. SRVs based on normalized bit scores were used as a relative similarity metric to address the possibility of insignificant bit scores for shorter small proteins, which can occur even for self-hits. Here, the lower threshold of 0.3 allows the detection of distant homologs. The clustering granularity is automatically selected based on the inflation value with the highest efficiency score. The various pruning steps aim to improve the clustering

by excluding singletons and barely matching small proteins beforehand.

A total of 8,884 small protein families with at least five members were identified using this clustering approach. Despite the successful identification of protein families, only a few families covering sequences with fewer than 30 AAs could be identified (Fig. 6). This is due to several limitations, including the small number of collected sequences of this length in sORFdb (Fig. 2), too few small proteins reported in the literature being included in databases, and possible low sequence conservation [1]. Most families cover small proteins with a length between 40 to 50 amino acids and around 30 amino acids. This is likely due to the fact that there is an increased number of non-redundant small proteins of about 30 AA length in the sORFdb database which is caused by historical and technical length cutoffs (Fig. 2).

A functional description could be assigned to nearly all protein families, and HMMs with gathering cutoffs providing high accuracy were built accordingly. The various filtering steps employed during database creation and clustering reduced the number of false positives in the database and subsequently in the small protein families. For this reason the HMMs can be used to accurately predict small proteins for genome annotation. The most common functional descriptions of the small protein families were consistent with those reported in the literature, such as toxin-antitoxin systems, membrane-associated systems, and regulatory proteins [1, 12, 15, 18]. Despite this consistency of the identified high-quality small protein families, 37,024 small proteins were a priori excluded from the clustering by filtering strategies, and another 4,073 were reported as singletons. It is unclear whether these are true positives or false positives. They could be false positives that slipped through the extensive filtering steps during the database creation. This could be the case for pervasively translated non-functional small proteins predicted with ribosome profiling [21–23] or for false positives sORFs detected by gene prediction tools [5, 6]. Another possibility is that they may be small proteins without homologs in the protein databases or are underrepresented in bacterial genomes due to difficult detection.

The development of the global microbial smORFs catalog (GMSC), published during the submission of our work, highlights the diversity of small proteins and that there is still a large number of unknown small proteins in prokaryotes. Compared to sORFdb, which focuses on high-quality sORFs and small proteins in bacterial genomes, the GMSC resource contains a large catalog of sORFs and small proteins of varying quality, mostly identified from metagenomes. The focus on the microbial metagenome revealed a higher proportion of small proteins in archaea than in bacteria in the GSMC [53].

sORFdb provides a comprehensive resource for information on practically all currently known high-quality sORFs and small proteins. Due to the understudied nature of these targets, there is still room for improvement in their detection and identification of their functions. Improved computational gene prediction and laboratory protocols for the identification of non-spurious small proteins and the elucidation of sORF and small protein properties are still open fields that need further research.

Conclusion

To the best of our knowledge, sORFdb is the first comprehensive, taxonomically independent database dedicated to sORF and small protein sequences and related information in bacteria. For this purpose, high-quality information from protein and genome databases enriched with physicochemical properties was combined. Furthermore, small protein families identified by a custom graph clustering approach accompanied by HMMs are provided to foster detection and consistent annotation.

In conclusion, the sORFdb database aims to serve as a high-quality primary resource for researchers studying sORFs and short proteins. It will help to improve the functional annotation of sORFs and small proteins, as well as the future detection of novel short proteins in bacteria.

Abbreviations

AA	Amino acid
HMM	Hidden Markov model
RBS	Ribosomal binding site
sORF	Short open reading frame
SRV	Score ratio value

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-025-11301-w.

Supplementary Material 1.

Acknowledgements

The authors would like to thank Dr. Jochen Blom for his helpful advice on the score ratio values. We acknowledge provision of computing resources by the Bioinformatics Core Facility (BCF) at Justus Liebig University Giessen.

Authors' contributions

JH and OS designed the study. JH designed, implemented, and tested the database creation workflow and clustering approach. JH, FC and SD conceived and designed the clustering approach. JH and LJ developed the website. LJ developed the server. JH conducted data analyses and interpreted the data. JH wrote the manuscript. AG, OS and SD substantially revised the manuscript. AG was responsible for funding. All authors read and approved the manuscript.

Funding
Open Access funding enabled and organized by Projekt DEAL. This work was supported by the Darwin Doctoral Scholarship provided by the Justus Liebig University Giessen, Germany and by the BMBF-funded de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI) (031A532B, 031A533A, 031A533B, 031A534A, 031A535A, 031A537A, 031A537B, 031A537C, 031A537D, 031A538A).

Data availability
The website can be accessed at https://sorfdb.computational.bio. All data is available for download at https://zenodo.org/records/10688271. The source code of the sORFdb workflow and the clustering approach is available at https://github.com/ag-computational-bio/sorfdb.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 31 October 2024 Accepted: 29 January 2025
Published online: 05 February 2025

References

1. Storz G, Wolf YI, Ramamurthi KS. Small Proteins Can No Longer Be Ignored. Ann Rev Biochem. 2014;83:753–77. https://doi.org/10.1146/annurev-biochem-070611-102400.
2. The UniProt Consortium. UniProt: The Universal Protein Knowledgebase in 2023. Nucleic Acids Res. 2023;51(D1):D523–31. https://doi.org/10.1093/nar/gkac1052.
3. Haft DH, Badretdin A, Coulouris G, DiCuccio M, Durkin AS, Jovenitti E, et al. RefSeq and the Prokaryotic Genome Annotation Pipeline in the Age of Metagenomes. Nucleic Acids Res. 2024;52(D1):D762–9. https://doi.org/10.1093/nar/gkad988.
4. Orr MW, Mao Y, Storz G, Qian SB. Alternative ORFs and Small ORFs: Shedding Light on the Dark Proteome. Nucleic Acids Res. 2020;48(3):1029–42. https://doi.org/10.1093/nar/gkz734.
5. Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: Prokaryotic Gene Recognition and Translation Initiation Site Identification. BMC Bioinformatics. 2010;11(1):1–11. https://doi.org/10.1186/1471-2105-11-119.
6. Lomsadze A, Gemayel K, Tang S, Borodovsky M. Modeling Leaderless Transcription and Atypical Genes Results in More Accurate Gene Prediction in Prokaryotes. Genome Res. 2018;28(7):1079–89. https://doi.org/10.1101/gr.230615.117.
7. National Center for Biotechnology Information. What Kind of Data Can Be Submitted to GenBank? In: The GenBank Submissions Handbook [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2014. p. 3.
8. Eberhardt RY, Haft DH, Punta M, Martin M, O'Donovan C, Bateman A. AntiFam: A Tool to Help Identify Spurious ORFs in Protein Annotation. Database. 2012;2012:bas003. https://doi.org/10.1093/database/bas003.
9. Ingolia NT, Hussmann JA, Weissman JS. Ribosome Profiling: Global Views of Translation. Cold Spring Harb Perspect Biol. 2019;11(5):a032698. https://doi.org/10.1101/cshperspect.a032698.
10. Vazquez-Laslop N, Sharma CM, Mankin A, Buskirk AR. Identifying Small Open Reading Frames in Prokaryotes with Ribosome Profiling. J Bacteriol. 2022;204(1):e00294–21. https://doi.org/10.1128/JB.00294-21.
11. Ahrens CH, Wade JT, Champion MM, Langer JD. A Practical Guide to Small Protein Discovery and Characterization Using Mass Spectrometry. J Bacteriol. 2022;204(1):e00353–21. https://doi.org/10.1128/jb.00353-21.
12. Hemm MR, Paul BJ, Schneider TD, Storz G, Rudd KE. Small Membrane Proteins Found by Comparative Genomics and Ribosome Binding Site Models. Mol Microbiol. 2008;70(6):1487–501. https://doi.org/10.1111/j.1365-2958.2008.06495.x.
13. Sberro H, Fremin BJ, Zlitni S, Edfors F, Greenfield N, Snyder MP, et al. Large-Scale Analyses of Human Microbiomes Reveal Thousands of Small, Novel Genes. Cell. 2019;178(5):1245–1259.e14. https://doi.org/10.1016/j.cell.2019.07.016.
14. Khitun A, Ness TJ, Slavoff SA. Small Open Reading Frames and Cellular Stress Responses. Mol Omics. 2019;15(2):108–16. https://doi.org/10.1039/C8MO00283E.
15. Fozo EM, Hemm MR, Storz G. Small Toxic Proteins and the Antisense RNAs That Repress Them. Microbiol Mol Biol Rev. 2008;72(4):579–89. https://doi.org/10.1128/MMBR.00025-08.
16. Venturini E, Svensson SL, Maaß S, Gelhausen R, Eggenhofer F, Li L, et al. A Global Data-Driven Census of Salmonella Small Proteins and Their Potential Functions in Bacterial Virulence. microLife. 2020;1(uqaa002). https://doi.org/10.1093/femsml/uqaa002.
17. Gray T, Storz G, Papenfort K. Small Proteins; Big Questions. J Bacteriol. 2022;204(1):e00341–21. https://doi.org/10.1128/JB.00341-21.
18. Simoens L, Fijalkowski I, Van Damme P. Exposing the Small Protein Load of Bacterial Life. FEMS Microbiol Rev. 2023;47(6):fuad063. https://doi.org/10.1093/femsre/fuad063.
19. Hemm MR, Weaver J, Storz G. Escherichia Coli Small Proteome. EcoSal Plus. 2020;9(1). https://doi.org/10.1128/ecosalplus.ESP-0031-2019.
20. Gvozdjak A, Samanta MP. Genes Preferring Non-AUG Start Codons in Bacteria. arXiv. 2020. https://doi.org/10.48550/arXiv.2008.10758.
21. Stringer A, Smith C, Mangano K, Wade JT. Identification of Novel Translated Small Open Reading Frames in Escherichia Coli Using Complementary Ribosome Profiling Approaches. J Bacteriol. 2022;204(1):e00352–21. https://doi.org/10.1128/JB.00352-21.
22. Wade JT, Grainger DC. Pervasive Transcription: Illuminating the Dark Matter of Bacterial Transcriptomes. Nat Rev Microbiol. 2014;12(9):647–53. https://doi.org/10.1038/nrmicro3316.
23. Smith C, Canestrari JG, Wang AJ, Champion MM, Derbyshire KM, Gray TA, et al. Pervasive Translation in Mycobacterium Tuberculosis. eLife. 2022;11:e73980. https://doi.org/10.7554/eLife.73980.
24. Van Dongen S. Graph Clustering Via a Discrete Uncoupling Process. SIAM J Matrix Anal Appl. 2008;30(1):121–41. https://doi.org/10.1137/040608635.
25. Enright AJ, Van Dongen S, Ouzounis CA. An Efficient Algorithm for Large-Scale Detection of Protein Families. Nucleic Acids Res. 2002;30(7):1575–84. https://doi.org/10.1093/nar/30.7.1575.
26. Pavlopoulos GA, Baltoumas FA, Liu S, Selvitopi O, Camargo AP, Nayfach S, et al. Unraveling the Functional Dark Matter through Global Metagenomics. Nature. 2023;1–9. https://doi.org/10.1038/s41586-023-06583-7.
27. Sayers EW, Cavanaugh M, Clark K, Pruitt KD, Sherry ST, Yankie L, et al. GenBank 2024 Update. Nucleic Acids Res. 2023;52(D1):D134–7. https://doi.org/10.1093/nar/gkad903.
28. Li Y, Zhou H, Chen X, Zheng Y, Kang Q, Hao D, et al. SmProt: A Reliable Repository with Comprehensive Annotation of Small Proteins Identified from Ribosome Profiling. Genomics Proteomics Bioinforma. 2021. https://doi.org/10.1016/j.gpb.2021.09.002.
29. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, et al. Pfam: The Protein Families Database in 2021. Nucleic Acids Res. 2021;49(D1):D412–9. https://doi.org/10.1093/nar/gkaa913.
30. Eddy SR. Accelerated Profile HMM Searches. PLoS Comput Biol. 2011;7(10):e1002195. https://doi.org/10.1371/journal.pcbi.1002195.
31. Larralde M, Zeller G. PyHMMER: A Python Library Binding to HMMER for Efficient Sequence Analysis. Bioinformatics. 2023;39(5):btad214. https://doi.org/10.1093/bioinformatics/btad214.
32. Larralde M. Pyrodigal: Python Bindings and Interface to Prodigal, an Efficient Method for Gene Prediction in Prokaryotes. J Open Source Softw. 2022;7(72):4296. https://doi.org/10.21105/joss.04296.
33. Buchfink B, Reuter K, Drost HG. Sensitive Protein Alignments at Tree-of-Life Scale Using DIAMOND. Nat Methods. 2021;18(4):366–8. https://doi.org/10.1038/s41592-021-01101-x.
34. Lerat E, Daubin V, Moran NA. From Gene Trees to Organismal Phylogeny in Prokaryotes: The Case of the γ-Proteobacteria. PLoS Biol. 2003;1(1):e19. https://doi.org/10.1371/journal.pbio.0000019.

35. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics. Bioinformatics. 2009;25(11):1422–3. https://doi.org/10.1093/bioinformatics/btp163.
36. Larralde M. Peptides.Py. Python; 2023. https://github.com/althonos/peptides.py.
37. Oren A, Garrity GM. Valid Publication of the Names of Forty-Two Phyla of Prokaryotes. Int J Syst Evol Microbiol. 2021;71(10):005056. https://doi.org/10.1099/ijsem.0.005056.
38. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow Enables Reproducible Computational Workflows. Nat Biotechnol. 2017;35(4):316–9. https://doi.org/10.1038/nbt.3820.
39. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic Local Alignment Search Tool. J Mol Biol. 1990;215(3):403–10. https://doi.org/10.1016/S0022-2836(05)80360-2.
40. Apeltsin L, Morris JH, Babbitt PC, Ferrin TE. Improving the Quality of Protein Similarity Network Clustering Algorithms Using the Network Edge Weight Distribution. Bioinformatics. 2011;27(3):326–33. https://doi.org/10.1093/bioinformatics/btq655.
41. van Dongen S. Performance Criteria for Graph Clustering and Markov Cluster Experiments. Tech Report. Information Systems [INS]; 2000. https://ir.cwi.nl/pub/4461.
42. Edgar RC. Muscle5: High-accuracy Alignment Ensembles Enable Unbiased Assessments of Sequence Homology and Phylogeny. Nat Commun. 2022;13(1):6968. https://doi.org/10.1038/s41467-022-34630-w.
43. Elasticsearch. Elasticsearch: The Official Distributed Search & Analytics Engine. Elastic. https://www.elastic.co/elasticsearch. Accessed 17 Mar 2024.
44. Eclipse. Eclipse Vert.x. https://vertx.io/. Accessed 17 Mar 2024.
45. You E. Vite. TypeScript. 2020. Reprint, vite. 2023. https://github.com/vitejs/vite.
46. You E. Vuejs/Core. TypeScript. 2018. Reprint, vuejs. 2023. https://github.com/vuejs/core.
47. Microsoft. Typescript Is JavaScript With Syntax For Types. https://www.typescriptlang.org/. Accessed 18 Oct 2023.
48. Priyam A, Woodcroft BJ, Rai V, Moghul I, Munagala A, Ter F, et al. Sequenceserver: A Modern Graphical User Interface for Custom BLAST Databases. Mol Biol Evol. 2019;36(12):2922–4. https://doi.org/10.1093/molbev/msz185.
49. Ondov BD, Bergman NH, Phillippy AM. Interactive Metagenomic Visualization in a Web Browser. BMC Bioinforma. 2011;12(1):1–10. https://doi.org/10.1186/1471-2105-12-385.
50. Grützner J, Billenkamp F, Spanka DT, Rick T, Monzon V, Förstner KU, et al. The Small DUF1127 Protein CcaF1 from Rhodobacter Sphaeroides Is an RNA-binding Protein Involved in sRNA Maturation and RNA Turnover. Nucleic Acids Res. 2021;49(6):3003–19. https://doi.org/10.1093/nar/gkab146.
51. Kraus A, Weskamp M, Zierles J, Balzer M, Busch R, Eisfeld J, et al. Arginine-Rich Small Proteins with a Domain of Unknown Function, DUF1127, Play a Role in Phosphate and Carbon Metabolism of Agrobacterium Tumefaciens. J Bacteriol. 2020;202(22). https://doi.org/10.1128/jb.00309-20.
52. Pearson WR. An Introduction to Sequence Similarity ("Homology") Searching. Current protocols in bioinformatics / editoral board, Andreas D Baxevanis [et al]. 2013. https://doi.org/10.1002/0471250953.bi0301s42.
53. Duan Y, Santos-Júnior CD, Schmidt TS, Fullam A, de Almeida BLS, Zhu C, et al. A Catalog of Small Proteins from the Global Microbiome. Nat Commun. 2024;15(1):7563. https://doi.org/10.1038/s41467-024-51894-6.

Publisher's Note