Computational and Structural Biotechnology Journal 23 (2024) 1919–1928

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Database article

kmerDB: A database encompassing the set of genomic and proteomic sequence information for each species

Ioannis Mouratidis [a,b,1], Fotis A. Baltoumas [c,1], Nikol Chantzi [a,1], Michail Patsakis [a], Candace S.Y. Chan [d], Austin Montgomery [a], Maxwell A. Konnaris [a,b,e], Eleni Aplakidou [c,f], George C. Georgakopoulos [g], Anshuman Das [a], Dionysios V. Chartoumpekis [h], Jasna Kovac [i], Georgios A. Pavlopoulos [c,j,*], Ilias Georgakopoulos-Soares [a,**]

[a] Institute for Personalized Medicine, Department of Biochemistry and Molecular Biology, The Pennsylvania State University College of Medicine, Hershey, PA, USA
[b] Huck Institutes of the Life Sciences, The Pennsylvania State University, University Park, PA, USA
[c] Institute for Fundamental Biomedical Research, BSRC "Alexander Fleming", Vari, 16672, Greece
[d] Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, CA, USA
[e] Department of Statistics, The Pennsylvania State University, University Park, PA, USA
[f] Department of Basic Sciences, School of Medicine, University of Crete, Heraklion, Greece
[g] National Technical University of Athens, School of Electrical and Computer Engineering, Athens, Greece
[h] Service of Endocrinology, Diabetology and Metabolism, Lausanne University Hospital, Lausanne, Switzerland
[i] Department of Food Science, The Pennsylvania State University, University Park, PA 16802, USA
[j] Center for New Biotechnologies and Precision Medicine, School of Medicine, National and Kapodistrian University of Athens, Athens, 11527, Greece

ARTICLE INFO

Keywords:
K-mer
Nullomer
Quasi-prime
Prime
Genome
Proteome

ABSTRACT

The decrease in sequencing expenses has facilitated the creation of reference genomes and proteomes for an expanding array of organisms. Nevertheless, no established repository that details organism-specific genomic and proteomic sequences of specific lengths, referred to as kmers, exists to our knowledge. In this article, we present kmerDB, a database accessible through an interactive web interface that provides kmer-based information from genomic and proteomic sequences in a systematic way. kmerDB currently contains 202,340,859,107 base pairs and 19,304,903,356 amino acids, spanning 54,039 and 21,865 reference genomes and proteomes, respectively, as well as 6,905,362 and 149,305,183 genomic and proteomic species-specific sequences, termed quasi-primes. Additionally, we provide access to 5,186,757 nucleic and 214,904,089 peptide sequences absent from every genome and proteome, termed primes. kmerDB features a user-friendly interface offering various search options and filters for easy parsing and searching. The service is available at: www.kmerdb.com.

1. Introduction

Rapid advances in high-throughput technologies combined with improvements in modern computer engineering and software development have facilitated the generation of accurate large-scale reference genomes and proteomes across all taxonomic domains of life [40,47,8]. This amount of data has enabled comparisons across organisms to annotate genome and proteomes, define coding regions, discover genes and their functions, and reveal insights from genomic regions that have traditionally been considered functionally irrelevant.

Genomes and proteomes consist of sequences of oligonucleotides and oligopeptides, respectively, which can be partitioned into substrings of a fixed length *k*, known as kmers. Kmers hold significant potential for understanding biological processes, as their patterns and occurrence rates can reveal key aspects of genomic features, including repetitive sequences, areas of biological function, variations in the genome, and the processes of DNA damage and repair [19,23,30,34,44]. Kmers are also used as clinical biomarkers for identifying pathogens and human diseases, as well as for detecting antimicrobial resistance among others [25,36,7].

* Corresponding author at: Institute for Fundamental Biomedical Research, BSRC "Alexander Fleming", Vari, 16672, Greece.
** Corresponding author.
E-mail addresses: pavlopoulos@fleming.gr (G.A. Pavlopoulos), izg5139@psu.edu (I. Georgakopoulos-Soares).
[1] Equally contributing authors

https://doi.org/10.1016/j.csbj.2024.04.050
Received 12 December 2023; Received in revised form 17 April 2024; Accepted 18 April 2024
Available online 21 April 2024

Among these kmers, some are conspicuously absent from a given genome or proteome, and are termed nullomers or nullpeptides [18,27,1,21]. These kmer sequences have been used for various applications including quality control, metagenomics classification, and phylogenetic analysis [11,14,24,32]. Experiments studying a subset of nullpeptides showed they can be highly pathogenic, indicating that certain nullpeptides are absent due to selection constraints [55]. Introduction of nullpeptides in cancer cells resulted in cancer cell killing, indicating putative drug development targets [4]. Additionally, nullpeptides are highly immunogenic and have immunomodulatory effects [41,3,57]. Remarkably, the resurfacing of nullomers in the human genome has been leveraged to detect cancer [17,35,53], demonstrating their potential for disease diagnostics. Similarly, quasi-prime kmers have been defined as a set of sequences that are exclusive to a single species and absent from every other known species with an available reference genome or proteome [38,37].

The first attempt to report such patterns was presented by Koulouras et al. with the creation of a database nullomers.org [27]. However, there are several limitations to consider. The database includes a restricted selection of nullomers and nullpeptides by reporting only peptide and nucleic minimal absent words. Moreover, its coverage, scope, and applicability are constrained by the inclusion of only two reference proteomes and approximately 1500 reference genomes. Another effort, OrthoVenn3, identifies orthologous clusters and detects conserved and variable genomic structures, making it a crucial resource for studying species evolution and genetic diversity [51]. Another database, Telobase, provides telomere motifs across organismal genomes in the tree of life [31]. To our knowledge, no publicly accessible database hosts a comprehensive compilation of the presence and characteristics of each species' peptide and nucleic kmers, all in a user-friendly and queryable format. In the same vein, no established database offers kmers unique to each species (known as quasi-primes) or kmers absent across all species (referred to as primes), despite their potential versatile applications. Consequently, the need for a repository where kmer, nullomer, nullpeptide, quasi-prime, and prime sequences can be queried on a large scale has become increasingly desirable.

In this article, we introduce *kmerDB*, a web-based database built to systematically catalog sets of DNA kmers, nullomers, nullpeptides, quasi-prime, and prime sequences for 54,039 species and 21,865 proteomes spanning all domains of life. The database provides various filter and search options organized in dynamic tables that can be queried and sorted for analysis. Users can investigate kmer patterns across many reference genomes and proteomes and examine kmer composition of various lengths for each organism across different taxonomic levels. Reference genomes and proteomes are linked to established publicly available databases such as the ENA Browser [28], the NCBI Genome Browser [46], the UniProtKB Proteome database [56], and InterPro protein families and domains database [9].

2. Results

2.1. Overall database statistics

Our objective in developing kmerDB was to establish a comprehensive repository of genomic and proteomic kmer data to characterize each species uniquely. We provide the kmer, nullomer, and species-specific (quasi-prime) sequences of each species' genome and proteome as previously outlined by Mouratidis et al. [38]. The current version of kmerDB comprises 54,039 reference genomes and 21,865 reference proteomes. For this dataset, we parsed 202,340,859,107 nucleotides and 19,304,903,356 amino acids across the reference genome and proteome sequences. The total number of kmers in the database is 242,366,914,024 for all reference genomes and 44,019,181,382 for all reference proteomes. Similarly, the total number of nullomers and nullpeptides is 505,812,292,016 and 339,223,621,873, respectively. To clarify, several kmers, nullomers and nullpeptides can be associated with multiple genomes or proteomes and, therefore, may appear multiple times in the dataset. At kmer length sixteen, the number of nucleic quasi-primes is 6,905,362, and at kmer lengths six and seven, the number of peptide quasi-primes is 149,305,183.

Since the kmer space expands exponentially with increasing kmer length, most possible kmers for large values of k are nullomers. This phenomenon is especially pronounced in viruses, which lack many kmers of length greater than seven base pairs (bps), likely due to their smaller genome size. Therefore, we only included kmers and nullomers of length up to seven bps for viral genomes in our database. For eukaryota, archaea, and bacteria, we extracted kmers and nullomers for lengths of six to twelve bps. Finally, we extracted kmers, nullomers, quasi-primes, and primes for lengths of three to seven amino acids for all available proteomes.

We have previously investigated the existence of nucleic quasi-primes, oligonucleotide sequences exclusive to a reference genome of a single species and absent from all others [37]. We have performed a comprehensive search for kmer lengths up to sixteen bps and found the first set of quasi-prime sequences at sixteen base pairs, also provided in the database. Additionally, we have previously examined the occurrence of peptide quasi-primes present in each reference proteome across all species [38]. No peptide quasi-primes were found for kmer lengths below six amino acids. However, we detected peptide quasi-primes at six and seven amino acids kmer length, which are also accessible in the database. Furthermore, we provide the set of nucleic and peptide primes of lengths of sixteen bps and six and seven amino acids. These are sequences absent across all the reference genomes and proteomes, comprising 5,186,757 nucleic primes and 214,904,089 peptide primes.

In kmerDB, each kmer, nullomer, and nullpeptide is associated with a computed probability, for either formation (P_{form}, assigned to kmers) or non-formation ($P_{non\text{-}form}$, assigned to nullomers and nullpeptides). The formation probability (P_{form}) for kmers indicates the likelihood of the kmer occurring by chance. Consequently, higher P_{form} values are generally assigned to kmers likely to form randomly, such as those occurring in multiple genomes or proteomes. Conversely, lower P_{form} values are attributed to rarer kmers, which could serve as distinctive features for a particular genome or proteome. For nullomers and nullpeptides, $P_{non\text{-}form}$ represents the probability of their absence in the genome or proteome. Higher $P_{non\text{-}form}$ values indicate sequences unlikely to be present in a particular genome, while lower values suggest sequences that might not exist by chance, although theoretically possible. The latter are particularly noteworthy, denoting nullomers that could arise through mutation events or polymorphisms, potentially associated with pathological conditions. Fig. 1.

2.2. The kmerDB interface

Users can explore the database by navigating through genomes and proteomes. Access to the data in kmerDB is facilitated via the Browse menu located at the kmerDB navigation bar. This menu allows users to select from the three domains of life (bacteria, archaea, eukaryota) along with viruses. Additionally, users can specify their preference between genomes and proteomes or utilize a combination of both criteria. Upon accessing the kmerDB Browse page, a compilation of genomes and proteomes matching the selected filters is presented (Fig. 2). Further customization of the search is achievable by choosing specific species through the NCBI Taxonomy ID, GenBank/Reference genome accession, UniProt reference proteome ID, or species name. This selection directs the user to the corresponding proteome or genome Entry page (Fig. 3). Furthermore, users can inspect the kmers and nullomers/nullpeptides associated with the chosen genome or proteome. Users can perform queries on kmers or filter them by kmer length for individual species (Fig. 4). For every kmer, nullomer, nullpeptide, and quasi-prime in the database, the computed formation (kmers, quasi-primes) or non-formation (nullomers, nullpeptides) probability is displayed, providing insights into its rarity (see above). In addition, for peptide sequences,

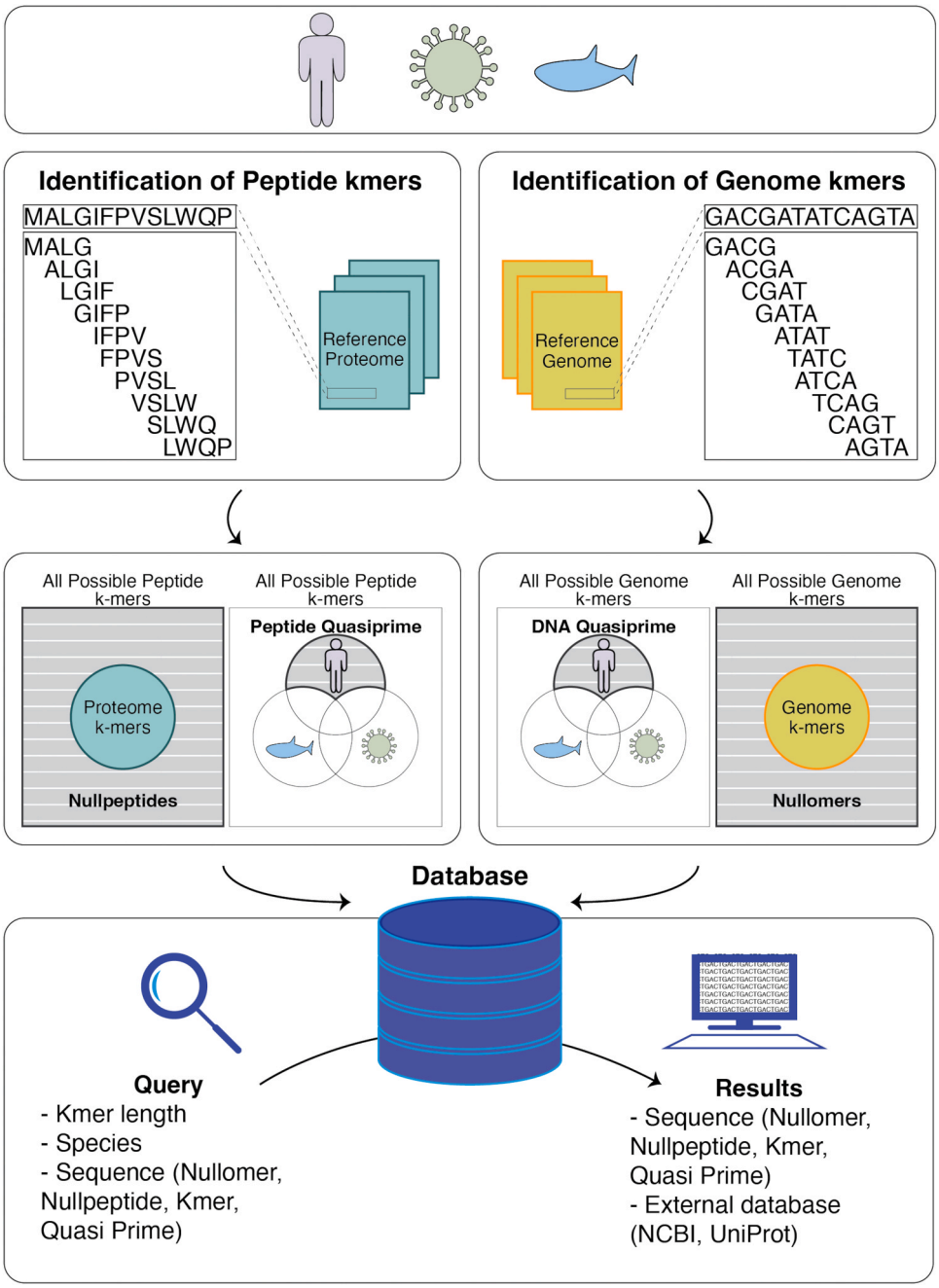

Fig. 1. Illustration of the derivation of kmers, nullomers, and nucleic quasi-primes in reference genomes and kmer peptides, nullpeptides and quasi-prime peptides in reference proteomes. The first step of the process involves cataloging every genome or peptide kmer for each species. The second step involves the derivation of nullomers or nullpeptides. Finally, the set of kmer sequences that are unique to each species are identified. The database encompasses this information for every species and is easily retrievable.

biochemical properties such as polarity, charge, and GRAVY hydrophobicity are computed and displayed. Similarly, for nucleic sequences, kmerDB calculates and presents the % GC content and primer melting temperature (Tm).

The database is also searchable via three search methods, Quick Search, Keyword Search, and Sequence Search (Fig. 5). Using Quick Search, users can quickly retrieve genomes and proteomes of interest using simple keywords. By using Keyword Search, they can perform more refined searches by combining multiple fields, including proteome or genome accessions, taxonomy identifiers, the organism name, domains, and the number of associated kmers/nullomers/nullpeptides or quasi-primes. Finally, through the Sequence Search option, they can directly submit their kmer or nullomer/nullpeptide sequences and retrieve any matching results from kmerDB's subset of statistically significant sequences.

In addition to the above, kmerDB provides links to external genomic and proteomic databases such as the ENA Browser [28], the NCBI Genome Browser [46], the UniProtKB Proteome database [56], and the InterPro protein families and domains database [9].

3. Materials and methods

3.1. Data retrieval and parsing

Reference proteomes were downloaded from UniProt: (Release 2022_03, 19-Sep-2022). These included reference proteomes for

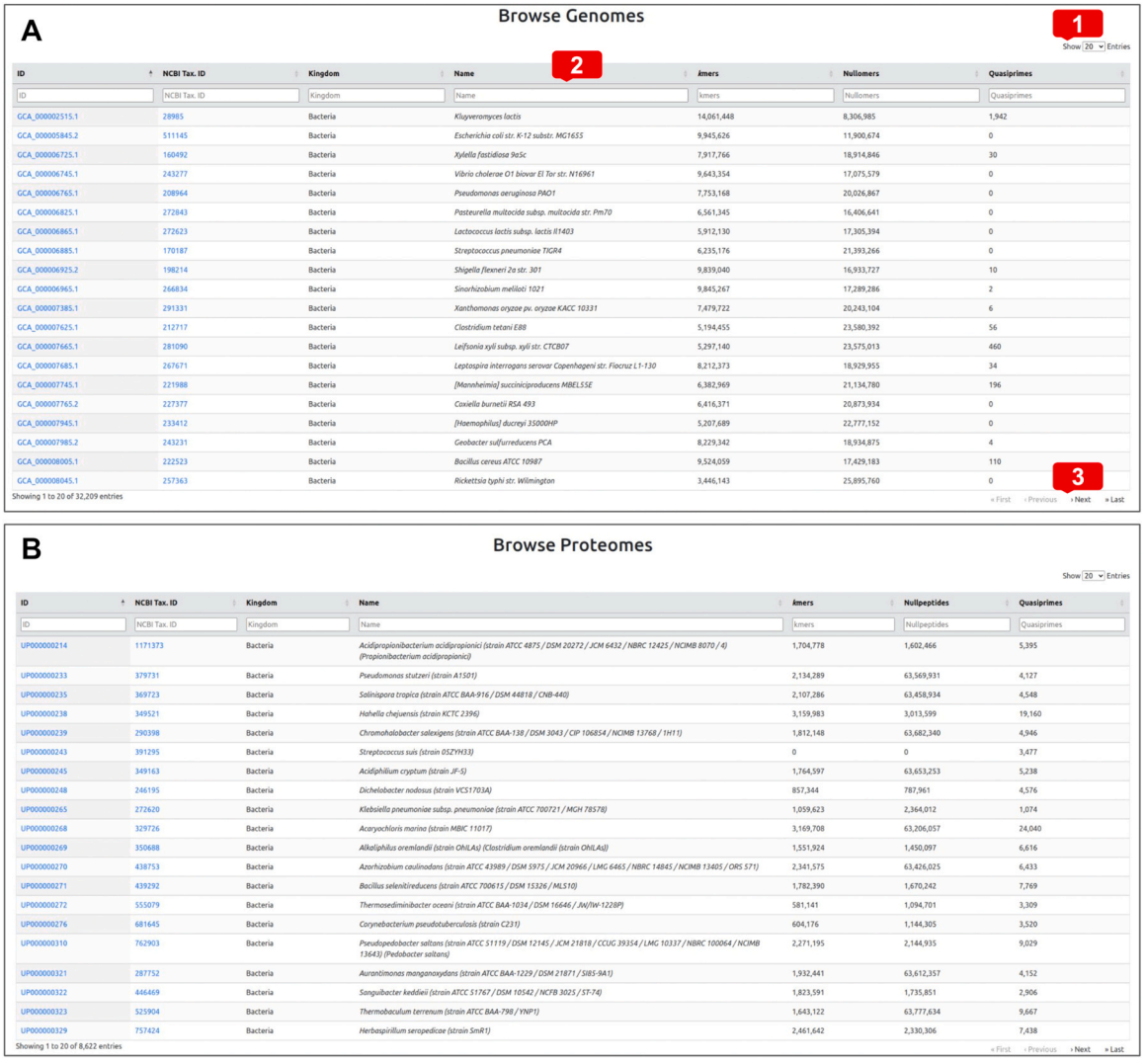

Fig. 2. KmerDB Browse pages for genomes and proteomes. A. The database browser for genomes. The genome identifier (GenBank or RefSeq accession), NCBI Taxonomy ID, organism group, name, and numbers of identified kmers, nullomers and quasi-primes per genome are given. **B.** The database browser for proteomes. The proteome identifier (UniProt proteome ID), NCBI Taxonomy ID, organism group, name, and numbers of identified kmers, nullpeptides and quasi-primes are given. In both tables, the interface includes options to change the number of entries per page (1), column filters to search the displayed items per page (2), and navigation buttons to view the previous or next set of entries.

eukaryota, bacteria, archaea, and viruses (Supplementary Table 1). Only the twenty standard amino acids were used throughout the analyses. Kmer lengths up to and including seven amino acids were studied.

Reference genomes were downloaded from the GenBank and RefSeq databases [40,8] as well as 104 reference genomes from the UCSC genome browser [39] (Supplementary Table 1). Kmer lengths up to and including twelve bps were analyzed to derive kmers and nullomers, whereas sixteen bps was chosen as the kmer length for nucleic quasi-primes. Details on the complexity and runtime execution of the analysis are given in the Supplementary Material (**Supplementary File 1**).

Definitions.

Genomic definitions.

Let us define the alphabet $L = \{A, T, C, G\}$ representing Adenine, Thymine, Cytosine, and Guanine respectively.

We define a ***sequence*** $S = \underline{a_1 a_2 a_3 \ldots a_n}$ where $a_i \in L$ for each $1 \leq i \leq n$.

A ***genome*** consists of a set of sequences over the alphabet L. A kmer refers to a short sequence $s = \underline{b_1 b_2 b_3 \ldots b_k}$ of length k. We define a *kmer* as present in a genome $G = \{S_1, S_2, S_3, \ldots, S_L\}$ if and only if there exists $S_i \in G$ where s is a subsequence of S_i. When a kmer s is present in genome G, then $s \in G$. Kmers of length $k = [6, 12]$ were considered for bacteria, archaea, and eukaryota, while for viruses, Lengths of $k = [3, 7]$ were used, due to the smaller viral genome sizes.

A ***nullomer*** of genome G is defined as a kmer s'that is not present in genome G, meaning $\nexists S_i \in G$ where s' is a subsequence of S_i. Therefore a nullomer for the genome G is any kmer not present in that genome. Similar to kmers, lengths of $k = [6, 12]$ were considered for bacteria, archaea, and eukaryota, and lengths of $k = [3, 7]$ were used for viruses.

Let $P = \{G_1, G_2, G_3, \ldots, G_x\}$the set of all genomes. We define a sequence q as a ***quasi-prime*** if and only if there exists $1 \leq i \leq x$ such that $s \in G_i$ and $s \notin G_j, \forall j \neq i$. Therefore, quasi-primes represent all kmers present in a single genome and absent from every other genome in our database.

Finally, a kmer p is defined as a ***prime*** in our dataset if and only if $\nexists i$ such that $p \in G_i$. Therefore primes represent all theoretically possible kmers that are absent from every genome in our database.

Proteomic definitions.

Similar to DNA sequences, we define an alphabet $L_p = \{G, A, L, M, F, W, K, Q, E, S, P, V, I, C, Y, H, R, N, D, T\}$ representing the common amino acids. A proteome consists of a set of sequences over the alphabet L_p,.

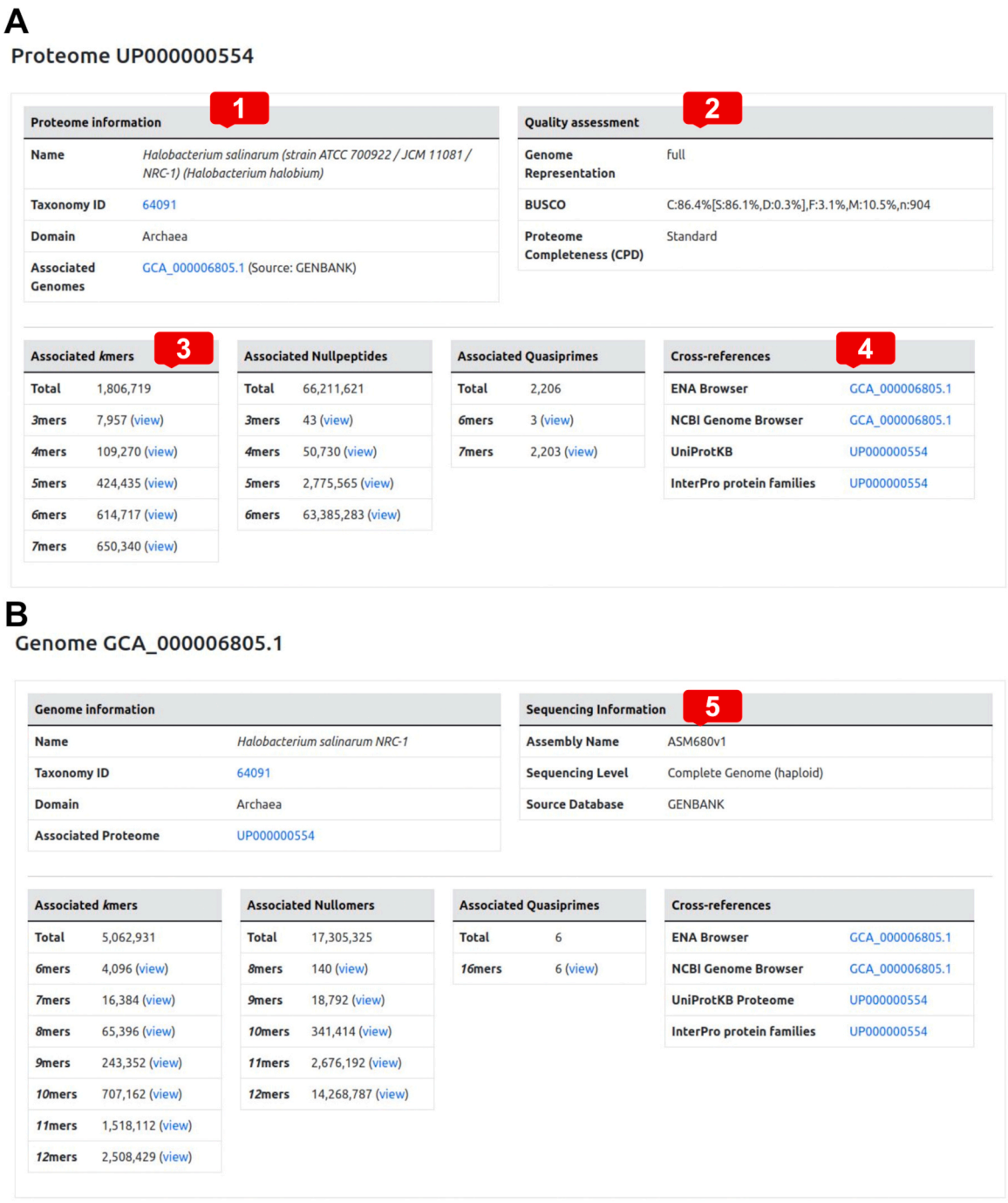

Fig. 3. Proteome and genome entry pages. Examples are shown for the archaeal species *Halobacterium salinarum* NRC-1. **A.** Proteome entry page for *H. salinarum* NRC-1 (ID: UP000000554). The entry page displays the basic annotation of the proteome (1) and a set of quality measurements including the extent of genome representation, proteome completeness (CPD) and, in the case of cell-based species (bacteria, archaea, and eukaryota), the Benchmarking Universal Single-Copy Orthologs (BUSCO) assessment. Access to the proteome's associated kmers, nullpeptides and quasi-primes is given through the tables at the bottom of the page (3). Finally,v cross-reference links to external databases are also offered, including the ENA and NCBI Genome Browsers, UniProtKB, and the InterPro protein family database (4). **B.** Genome entry page for *H. salinarum* NRC-1 (ID: GCA_000006805.1). The entry page follows the same structure as the proteome entry page, with additional information on the genome's sequencing properties, including the assembly name, source database, and sequencing level (5).

Proteomic kmers, nullpeptides, quasi-primes, and primes are defined equivalently to their genomic counterparts. For this study, we considered proteomic kmers and nullpeptides for lengths $k = [3,7]$ and $k = [3,6]$, respectively. Proteomic quasi-primes were studied at lengths $k = [3,6]$.

3.2. Nucleic and peptide kmer and nullpeptide detection

The identification of kmers was performed following previously established definitions defined in [18]. Nullomer and nullpeptide detection were performed as previously described in [18] for each species at each kmer length.

Identification of nucleic and peptide quasi-primes.

DNA quasi-prime identification was performed by identifying kmers

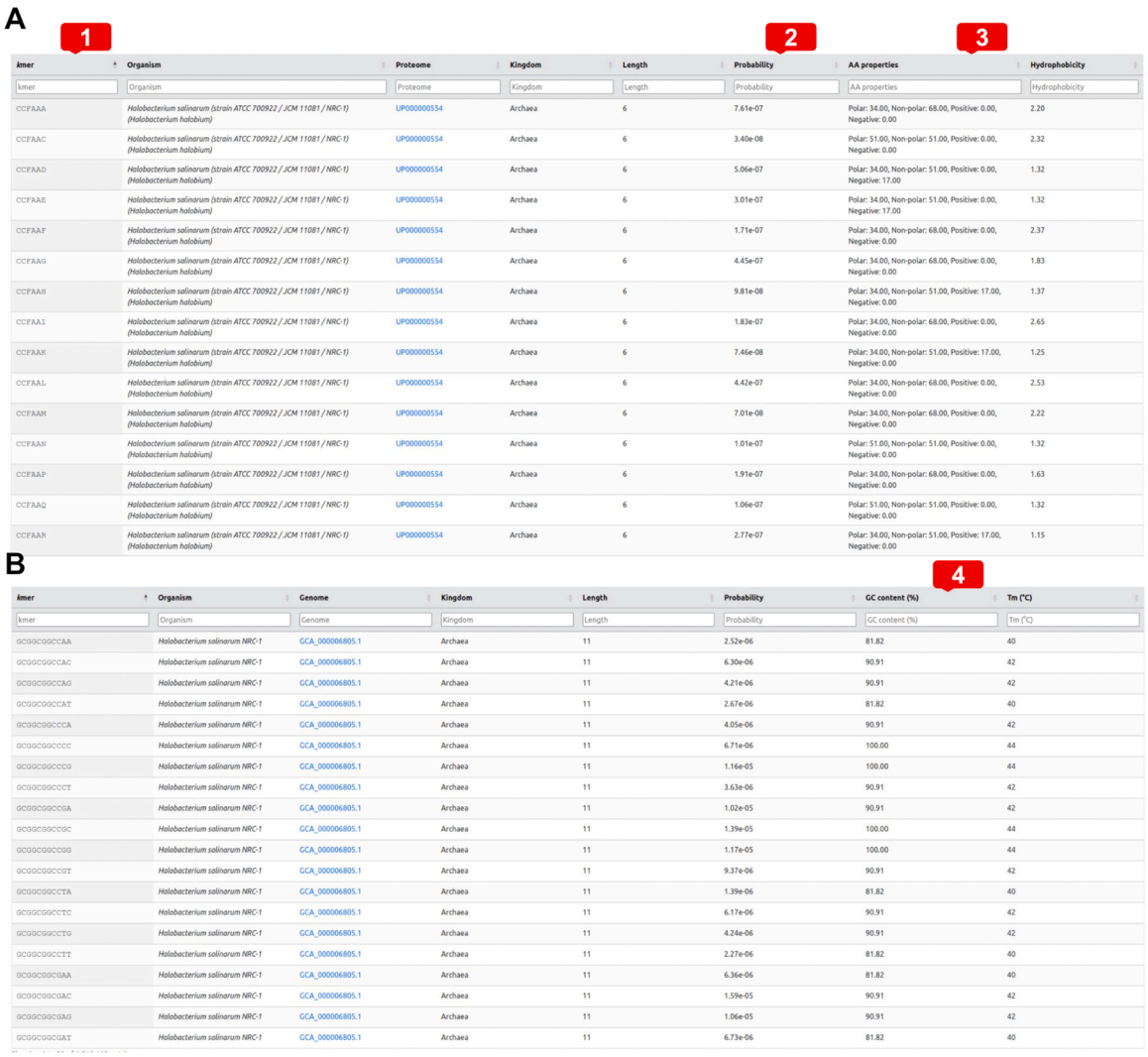

Fig. 4. Kmer search page in individual genomes and proteomes for kmers, nullomers, nullpeptides and quasi-primes. A. Example search for kmer length of six amino acids in *H. salinarum* NRC-1 (ID: UP000000554). The kmer sequence (1), formation probability (2), and sequence features (3), namely, amino acid properties and hydrophobicity are given. **B.** Example search for nullomers with a length of 11 base-pairs in *H. salinarum* NRC-1 (ID: GCA_000006805.1). For DNA sequences, the displayed properties (4) include the % GC content and melting point temperature (Tm).

that were present in each reference genome and nullomers in every other reference genome. Similarly, peptide quasi-prime identification was performed by identifying kmers that were present in each reference proteome and nullomers in every other reference proteome.

Identification of nucleic quasi-primes was performed for kmer length of sixteen bps. This was the shortest kmer length at which we observed DNA quasi-primes. Similarly, for peptide kmers, we performed quasi-prime identification for kmer lengths of six and seven amino acids, since these were the shortest peptide lengths at which we observed quasi-primes.

3.3. Statistical analysis

We used a Markov chain model to determine the formation probability of each kmer, which is the probability of its occurrence by random chance based on the sequence content of its reference genome or proteome. The transition probabilities, indicative of the likelihood of a nucleotide base X following a preceding base Y (where X and Y can be A, T, C, or G), were computed across all reference genomes within our database. Subsequently, we established all 16 possible transition probabilities for each reference genome to ascertain the formation probability of every kmer identified therein. In the context of protein kmers, a similar methodology was adopted. Transition probabilities for each proteome were determined, taking into account the 20 standard amino acids. This set of amino acids led to the calculation of 400 distinct transition probabilities, each reflecting the frequency with which one amino acid is likely to follow another within the protein sequences.

For the observed kmers, the statistical approach to determine their formation probability (P_{form}) was based on multiplying individual transition probabilities, by applying the Markov assumption. This means that for any given kmer, its formation probability was estimated as the product of the probabilities of each sequential transition within the kmer. This method allowed calculating the likelihood of any specific kmer occurring by chance, based on the genomic context.

For both nullomers and nullpeptides, we provided a probabilistic estimate of the nullomer's/nullpeptide's absence in its corresponding reference genome/proteome ($P_{non\text{-}form}$). The formation probability of a nullomer/nullpeptide (P_{form}) is computed and then exponentiated by L, where L represents the total number of potential positions where the nullomer/nullpeptide could be located within the reference genome/proteome. Therefore, P_{form}^{L} yields the expected frequency of the nullomer's occurrence in the reference genome or proteome. Subtracting this value from 1 provides the estimated probability that the nullomer does not appear in the given genome or proteome ($P_{non\text{-}form}$).

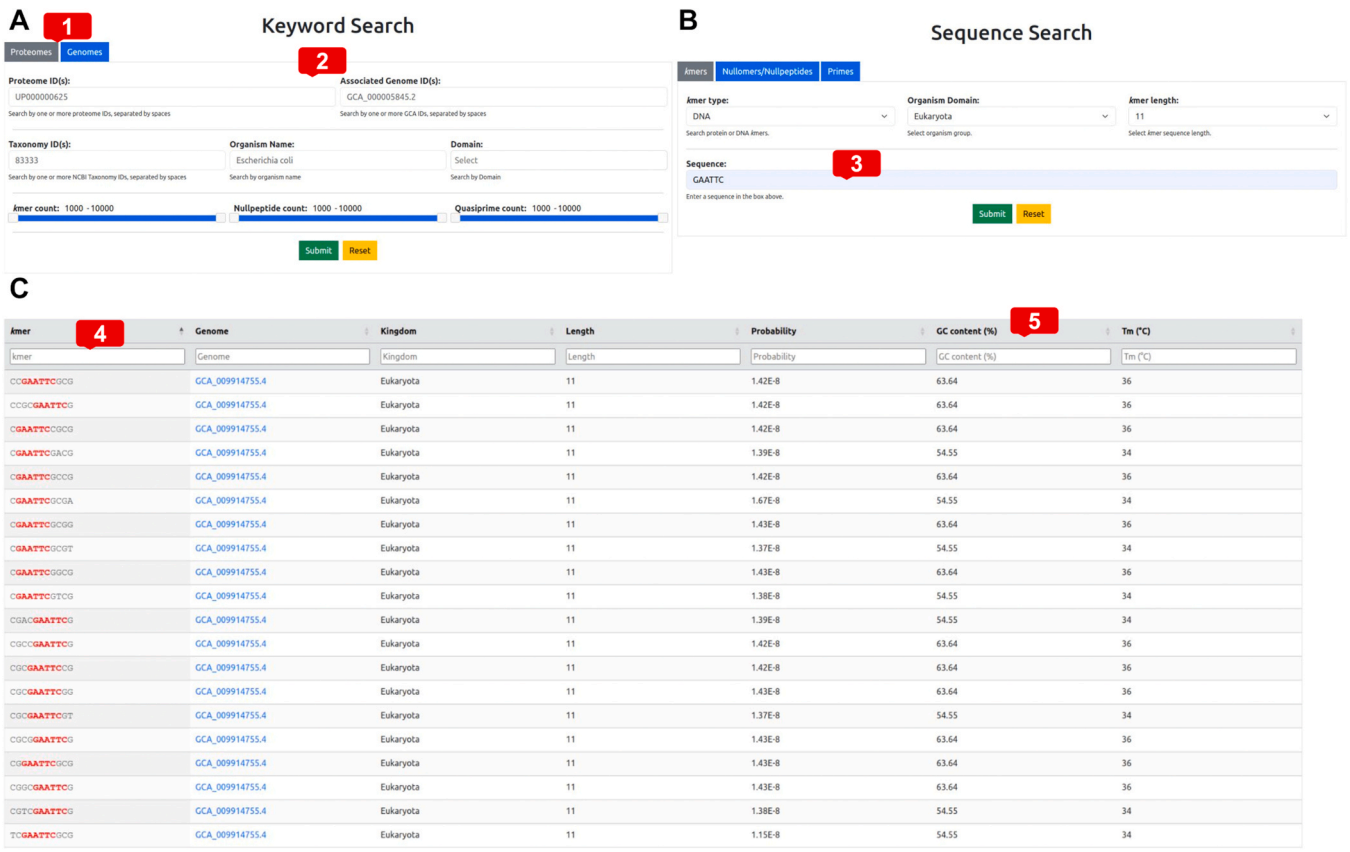

Fig. 5. The search capabilities of kmerDB. A. The keyword search form allows for performing refined searches for genomes and proteomes. The controls at the top of the form (1) select the dataset type (proteome or genome). Multiple fields (2) can be combined to produce exact search results. **B.** The sequence search form allows searching kmers, nullomers, nullpeptides and primes for sequences matching a user-defined query (3). **C.** Example kmer search results for the DNA sequence "GAATTC". The kmer hits are displayed with the matching sequence range highlighted in red. In addition, the kmer properties are also given, including the formation probability, %GC content, and melting point temperature (5).

Following the estimation of the formation/non-formation probability, we sought to estimate the statistical significance of each kmer and nullomer/nullpeptide, by deriving its adjusted P-value (q-value), using the Tarone modification of the Bonferroni adjustment method [52], adapting the approach previously used by Koulouras and Frith [27]. In this approximation, all words of length k (e.g. 7-mers) are ordered in descending order of their Markov chain probability (as described above), and the q-value is calculated as follows:

$$qval = P \cdot \left(a^k - c\right)$$

where P is the Markov probability (P_{form} for kmers, and $P_{non\text{-}form}$ for nullomers and nullpeptides), a is the size of the sequence alphabet (a=4 for DNA nucleotides, a=20 for protein amino acids), k is the word length (e.g. $k = 7$) and c is a counter starting from 0 and increasing by 1 each time a kmer is excluded from testing. The exclusion of a kmer occurs when the computed q-value is above the defined statistical significance threshold (set to 0.01). This filtering produced a subset of statistically significant sequences, which is available for download through the "Downloads" page of the database, and is also used to perform sequence-based queries.

3.4. Database implementation

Kmers, nullomers, nullpeptides, quasi-primes, and primes are organized in prefix tree (trie) data structures, using the Matching Algorithm with Recursively Implemented StorAge (MARISA) Trie implementation and its Python bindings [59]. This particular data structure was chosen as the most performant. Trie hashes produced by MARISA are alphabet-agnostic and can be used to retrieve all contents of an indexed hash table and to perform searches inside that table, either as exact matches or with prefix-based queries. While several kmer-based indexing methods exist in the literature [2,12], such as ssHash [43], ntHash [26], Fulgor [16,26] or Pufferfish [6], they have been implemented as a means to hash existing DNA sequences and produce corresponding dictionaries of k-sized substrings (kmers), which can be subsequently used in several other tasks, such as testing whether an input sequence contains kmers existing in said dictionary. Although such structures are beneficial in sequence feature recognition/prediction (e.g. kmer based taxonomy assignment), they do not serve the purpose of kmerDB, namely, storing kmers in a database-like structure, and retrieving all kmers existing in one or more genomes/proteomes (or, conversely, all nullomers / nullpeptides not appearing in a genome/proteome). At the same time, these structures are geared towards the hashing of DNA kmers, meaning they have been implemented with a 4-letter alphabet (A, T, G, C) hardcoded into their underlying data structure. However, a very large portion of kmerDB concerns protein sequences, which would require the use of a 20-letter alphabet for amino acids.

The current size of the stored kmers and nullomers/nullpeptides is 172 GB and 154 GB, respectively, utilizing the MARISA Trie data structure for storing the sequences of each genome/proteome. By contrast, the initial size of the dataset in uncompressed ASCII format amounts to approximately 2.4 TB. This highlights the efficacy of the MARISA Trie structure as a means of hashing and storing kmer datasets.

The front end of kmerDB is implemented in HTML, CSS, and JavaScript. The back end is supported by the Apache web server and the Slim Framework v. 4.0, with server-side operations handled by PHP and, when required, Python. Genome and proteome metadata are stored in a

MySQL relational database. The kmerDB website layout was designed with the Bootstrap v. 5 framework, jQuery, and the DataTables library. kmerDB is publicly available through http://www.kmerdb.com.

4. Discussion

Here we introduce kmerDB, a novel repository that contains kmer, nullomer, nullpeptide, quasi-prime, and prime sequences for 54,039 reference genomes and 21,865 reference proteomes. While the identification of kmers and nullomers for an individual species can be obtained with bioinformatic tools [33], this, to our knowledge, is the first publicly available database containing all kmers, nullomers, nullpeptides, and quasi-primes for each organism with a reference genome or proteome. The database provides a user-friendly interface that allows users to select species by name, ID, kmer sequence, or kmer length and provides links to other reference databases, including NCBI for genomic kmer sequences [46] and UniProt for peptide kmer sequences [56]. The database incorporates statistical scores for the likelihood of a nucleic or peptide kmer being present/absent from a genome or proteome using Markov models. We note that a previous resource with a similar name (kmer-db) also exists, focusing on computing the evolutionary distance of sequences, but has no association with our work [13]. kmerDB will be updated regularly to incorporate new reference genomes and proteomes as they become available. This is a necessary step, as the database's content (especially nullomers/nullpeptides and quasi-primes) could potentially be altered due to the emergence of additional reference genomes or proteomes, and the possibility of novel variants arising for the existing genomes.

We outline several potential applications of kmerDB across diverse research domains. Previous studies have demonstrated that variations in biological processes can influence the genomic and proteomic composition of an organism, which is reflected in the kmer profile of its genome or proteome [29,48,54,58]. Furthermore, kmers can be associated with specific functional roles, such as transcription factor binding sites [48]. kmerDB facilitates the querying of user-defined kmer sequences against its dataset, enabling investigations into genomic and proteomic kmer disparities across species, including the exploration of kmers with functional significance in genomes or proteomes.

Nullomers and nullpeptides hold utility in evolutionary studies as indicators of negative selection [18,27], for pathogen detection, or as potential candidates for therapeutic drugs [45,49]. For example, there is evidence suggesting the roles of nullpeptides as anti-cancer agents [4,5]. Additionally, nullomers and nullpeptides find applications in cancer detection [35], as vaccine adjuvants [41], or in forensic contexts [20]. Notably, our database incorporates a Markov chain-based statistical score, indicating the likelihood of each nullomer and nullpeptide being absent from a genome or proteome. Nullomers and nullpeptides with lower probabilities of absence are more likely to be subject to selection pressures and can thus be prioritized in subsequent studies.

DNA and peptide quasi-primes serve as universal and concise genomic and proteomic signatures for each organism, presenting potential as detection platforms for pathogens. They offer advantages over traditional methods like cell culturing and colony counting, which are slow and inapplicable to non-culturable species. Nucleic quasi-primes hold promise as biomarkers in metagenomic next-generation sequencing applications, particularly for accurate pathogen detection in clinical settings or ensuring food safety. Peptide quasi-primes hold potential for designing highly specific antibodies to mitigate typical antibody cross-reactivity [15,10]. Quasi-primes also shed light on evolution, serving as sites of accelerated evolution and traits specific to species [22,37]. For instance, human nucleic quasi-primes are linked to brain development and neurological disorders [37]. Consequently, the quasi-primes in the database can advance research on the shortest species-specific nucleic or peptide sequences.

Kmer data from kmerDB can find applications in comparative genomics and evolutionary studies [42,50], aiding sequence specification like identifying highly-specific CRISPR target sites [60]. Prime sequences can serve as genetic barcodes or targetable landing sites in biotechnological applications, facilitating tracking of cells or organisms through genetic tagging. In essence, kmerDB stands as a versatile, rapid, and high-caliber database facilitating convenient access to genomic and proteomic information across species and taxonomies.

Code Availability

The GitHub code is provided at: https://github.com/Georgakopoulos-Soares-lab/kmerdb_stats.

Funding

I.M., N.C., M.P., M.A.K., A.M., and I.G.S., were funded by the startup funds from the Penn State College of Medicine and by the Huck Innovative and Transformational Seed Fund (HITS) award from the Huck Institutes of the Life Sciences at Penn State University. F.A.B. was funded by Fondation Santé and Onassis Foundation. G.A.P. was supported by the Hellenic Foundation for Research and Innovation (H.F.R.I) under the call 'Greece 2.0 - Basic Research Financing Action (Horizontal support of all Sciences), Sub-action II', Grant ID: 16718-PRPFOR. E.A. was supported by the program 'Greece 2.0 - National Recovery and Resilience Plan', Grant ID: TAEDR-0539180. J.K. was supported by the USDA National Institute of Food and Agriculture and Hatch Appropriations under Project PEN04853 and Accession 7005519, and the Multistate project 4666.

CRediT authorship contribution statement

Anshuman Das: Data curation, Formal analysis. **George C. Georgakopoulos:** Data curation, Formal analysis, Validation. **Jasna Kovac:** Data curation, Formal analysis. **Dionysios V. Chartoumpekis:** Data curation, Formal analysis. **Ilias Georgakopoulos-Soares:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Validation, Writing – original draft, Writing – review & editing. **Ioannis Mouratidis:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Validation, Writing – original draft, Writing – review & editing. **Georgios A Pavlopoulos:** Data curation, Formal analysis, Project administration, Supervision, Writing – original draft, Writing – review & editing. **Michail Patsakis:** Data curation, Formal analysis, Methodology, Writing – review & editing. **Nikol Chantzi:** Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Writing – original draft, Writing – review & editing. **Eleni Aplakidou:** Formal analysis. **Fotis A. Baltoumas:** Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Austin Montgomery:** Data curation, Formal analysis, Validation. **Candace S.Y. Chan:** Data curation, Formal analysis. **Maxwell A. Konnaris:** Data curation, Formal analysis, Methodology, Writing – original draft.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

kmerDB is publicly available as a web service at: https://www.kmerdb.com.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2024.04.050.

References

[1] Acquisti C, Poste G, Curtiss D, Kumar S. Nullomers: Really a matter of natural selection? PLoS ONE 2007;2(10):e1022. https://doi.org/10.1371/journal.pone.0001022.

[2] Alanko, Jarno N. , Simon J Puglisi , and Jaakko Vuohtoniemi . 2022. "Succinct K-Mer Sets Using Subset Rank Queries on the Spectral Burrows-Wheeler Transform *." bioRxiv. https://doi.org/10.1101/2022.05.19.492613.

[3] Ali Nilufar, Cody Wolf, Swarna Kanchan, Shivakumar R Veerabhadraiah, Laura Bond, Matthew W Turner, et al. 9S1R nullomer peptide induces mitochondrial pathology, metabolic suppression, and enhanced immune cell infiltration, in triple-negative breast cancer mouse model. Biomed Pharmacother 2024;170(January):115997.

[4] Alileche Abdelkrim, Goswami Jayita, Bourland William, Davis Michael, Hampikian Greg. Nullomer derived anticancer peptides (NulloPs): differential lethal effects on normal and cancer cells in vitro. Peptides 2012;38(2):302–11.

[5] Alileche Abdelkrim, Hampikian Greg. The effect of nullomer-derived peptides 9R, 9S1R and 124R on the NCI-60 panel and normal cell lines. BMC Cancer 2017;17(1):533.

[6] Almodaresi Fatemeh, Sarkar Hirak, Srivastava Avi, Patro Rob. A space and time-efficient index for the compacted colored de bruijn graph. Bioinformatics 2018;34(13):i169–77.

[7] Aun Erki, Brauer Age, Kisand Veljo, Tenson Tanel, Remm Maido. A K-mer-based method for the identification of phenotype-associated genomic biomarkers and predicting phenotypes of sequenced bacteria. PLoS Comput Biol 2018;14(10):e1006434.

[8] Benson Dennis A, Cavanaugh Mark, Clark Karen, Karsch-Mizrachi Ilene, Lipman David J, Ostell James, et al. GenBank. Nucleic Acids Res 2013;41(Database issue):D36–42.

[9] Blum Matthias, Chang Hsin-Yu, Chuguransky Sara, Grego Tiago, Kandasaamy Swaathi, Mitchell Alex, et al. The InterPro Protein Families and Domains Database: 20 Years on. Nucleic Acids Res 2021;49(D1):D344–54.

[10] Bordeaux Jennifer, Welsh Allison, Agarwal Seema, Killiam Elizabeth, Baquero Maria, Hanna Jason, et al. Antibody Validation. BioTechniques 2010;48(3):197–209.

[11] Brandies, Parice, Emma Peel, Carolyn J. Hogg, and Katherine Belov. 2019. "The Value of Reference Genomes in the Conservation of Threatened Species." Genes 10 (11). https://doi.org/10.3390/genes10110846.

[12] Chikhi, Rayan, Jan Holub, and Paul Medvedev. 2019. "Data Structures to Represent a Set of K-Long DNA Sequences." ⟨http://arxiv.org/abs/1903.12312⟩.

[13] Deorowicz Sebastian, Gudys Adam, Dlugosz Maciej, Kokot Marek, Danek Agnieszka. Kmer-Db: instant evolutionary distance estimation. Bioinformatics 2019;35(1):133–6.

[14] Deurenberg Ruud H, Bathoorn Erik, Chlebowicz Monika A, Couto Natacha, Ferdous Mithila, García-Cobos Silvia, et al. Application of next generation sequencing in clinical microbiology and infection prevention. J Biotechnol 2017;243(February):16–24.

[15] Egelhofer, Thea A., Aki Minoda, Sarit Klugman, Kyungjoon Lee, Paulina Kolasinska-Zwierz, Artyom A. Alekseyenko, Ming-Sin Cheung, et al. 2011. "An Assessment of Histone-Modification Antibody Quality." Nature Structural & Molecular Biology 18 (1): 91–93.

[16] Fan, Jason, Noor Pratap Singh, Jamshed Khan, Giulio Ermanno Pibiri, and Rob Patro. 2023. "Fulgor: A Fast and Compact {k-Mer} Index for Large-Scale Matching and Color Queries." In 23rd International Workshop on Algorithms in Bioinformatics (WABI 2023), 18:21. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

[17] Georgakopoulos-Soares, Ilias, Ofer Yizhar-Barnea, Ioannis Mouratidis, Rachael Bradley, Ryder Easterlin, Candace Chan, Emmalyn Chen, John S. Witte, Martin Hemberg, and Nadav Ahituv. 2021. "Leveraging Sequences Missing from the Human Genome to Diagnose Cancer." *medRxiv*.

[18] Georgakopoulos-Soares Ilias, Yizhar-Barnea Ofer, Mouratidis Ioannis, Hemberg Martin, Ahituv Nadav. Absent from DNA and protein: genomic characterization of nullomers and nullpeptides across functional categories and evolution. Genome Biol 2021;22(1):245.

[19] Ghandi Mahmoud, Lee Dongwon, Mohammad-Noori Morteza, Beer Michael A. Enhanced regulatory sequence prediction using gapped K-Mer features. PLoS Comput Biol 2014;10(7):e1003711.

[20] Goswami Jayita, Davis Michael C, Andersen Tim, Alileche Abdelkrim, Hampikian Greg. Safeguarding forensic DNA reference samples with nullomer barcodes. J Forensic Leg Med 2013;20(5):513–9.

[21] Herold J, Kurtz S, Giegerich R. Efficient computation of absent words in genomic sequences. BMC Bioinform. 2008;9:167. https://doi.org/10.1186/1471-2105-9-167.

[22] Hubisz Melissa J, Katherine S Pollard. Exploring the genesis and functions of human accelerated regions sheds light on their role in human evolution. Curr Opin Genet Dev 2014;29(December):15–21.

[23] Iulio Julia di, Bartha Istvan, Wong Emily HM, Yu Hung-Chun, Lavrenko Victor, Yang Dongchan, et al. The human noncoding genome defined by genetic diversity. Nat Genet 2018;50(3):333–7.

[24] Jagadeesan Balamurugan, Gerner-Smidt Peter, Allard Marc W, Leuillet Sébastien, Winkler Anett, Xiao Yinghua, et al. The use of next generation sequencing for improving food safety: translation into practice. Food Microbiol 2019;79(June)):96–115.

[25] Jaillard Magali, Palmieri Mattia, van Belkum Alex, Mahé Pierre. Interpreting K-mer-based signatures for antibiotic resistance prediction. GigaScience 2020;9(10). https://doi.org/10.1093/gigascience/giaa110.

[26] Kazemi Parham, Wong Johnathan, Nikolić Vladimir, Mohamadi Hamid, Warren René L, Birol Inanç. ntHash2: recursive spaced seed hashing for nucleotide sequences. Bioinformatics 2022;38(20):4812–3.

[27] Koulouras Grigorios, Frith Martin C. Significant non-existence of sequences in genomes and proteomes. Nucleic Acids Res 2021;49(6):3139–55.

[28] Leinonen Rasko, Ruth Akhtar Ewan Birney, Lawrence Bower Ana Cerdeno-Tárraga, Ying Cheng Iain Cleland, et al. The European nucleotide archive. *Nucleic Acids Res* 2011:D28–31.

[29] Lerat E. Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs. Heredity 2010;104(6):520–33.

[30] Liu Zian, Samee Md Abul Hassan. Structural underpinnings of mutation rate variations in the human genome. Nucleic Acids Res 2023;51(14):7184–97.

[31] Lyčka Martin, Bubeník Michal, Závodník Michal, Peska Vratislav, Fajkus Petr, Demko Martin, et al. TeloBase: a community-curated database of telomere sequences across the tree of life. Nucleic Acids Res 2024;52(D1):D311–21.

[32] Maljkovic Berry, Irina, Melanie C. Melendrez, Kimberly A. Bishop-Lilly, Wiriya Rutvisuttinunt, Simon Pollett, Eldin Talundzic, Lindsay Morton, and Richard G. Jarman. 2020. "Next Generation Sequencing and Bioinformatics Methodologies for Infectious Disease Research and Public Health: Approaches, Applications, and Considerations for Development of Laboratory Capacity." The Journal of Infectious Diseases 221 (Suppl 3): S292–307.

[33] Marçais Guillaume, Kingsford Carl. A fast, lock-free approach for efficient parallel counting of occurrences of K-mers. Bioinformatics 2011;27(6):764–70.

[34] Mejía-Guerra María Katherine, Buckler Edward S. A K-mer grammar analysis to uncover maize regulatory architecture. BMC Plant Biol 2019;19(1):103.

[35] Montgomery, Austin, Georgios Christos Tsiatsianis, Ioannis Mouratidis, Candace S. Y. Chan, Maria Athanasiou, Anastasios D. Papanastasiou, Verena Kantere, et al. 2023. "Utilizing Nullomers in Cell-Free RNA for Early Cancer Detection." medRxiv. https://doi.org/10.1101/2023.06.10.23291228.

[36] Mouratidis, Ioannis, Nikol Chantzi, Umair Khan, Maxwell A. Konnaris, Candace S. Y. Chan, Manvita Mareboina, and Ilias Georgakopoulos-Soares. 2023. "Frequentmers - a Novel Way to Look at Metagenomic Next Generation Sequencing Data and an Application in Detecting Liver Cirrhosis." bioRxiv. https://doi.org/10.1101/2023.09.19.23295771.

[37] Mouratidis, Ioannis, Maxwell A. Konnaris, Nikol Chantzi, Candace S. Y. Chan, Austin Montgomery, Fotis A. Baltoumas, Michail Patsakis, et al. 2023. "Nucleic Quasi-Primes: Identification of the Shortest Unique Oligonucleotide Sequences in a Species." *bioRxiv*. https://doi.org/10.1101/2023.12.12.571240.

[38] Mouratidis Ioannis, Chan Candace SY, Chantzi Nikol, Tsiatsianis Georgios Christos, Hemberg Martin, Ahituv Nadav, et al. Quasi-prime peptides: identification of the shortest peptide sequences unique to a species. NAR Genom Bioinforma 2023;5(2):lqad039.

[39] Nassar Luis R, Barber Galt P, Benet-Pagès Anna, Casper Jonathan, Clawson Hiram, Diekhans Mark, et al. The UCSC genome browser database: 2023 update. Nucleic Acids Res 2023;51(D1):D1188–95.

[40] O'Leary Nuala A, Mathew W Wright, Rodney Brister J, Ciufo Stacy, Haddad Diana, McVeigh Rich, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 2016;44(D1):D733–45.

[41] Patel Ami, Dong Jessica C, Trost Brett, Richardson Jason S, Tohme Sarah, Babiuk Shawn, et al. Pentamers not found in the universal proteome can enhance antigen specific immune responses and adjuvant vaccines. PLoS One 2012;7(8):e43802.

[42] Perry Scott C, Beiko Robert G. Distinguishing Microbial Genome Fragments Based On Their Composition: Evolutionary And Comparative Genomic Perspectives. Genome Biol Evol 2010;2(January):117–31.

[43] Pibiri Giulio Ermanno. Sparse and skew hashing of K-Mers. Bioinformatics 2022;38(Suppl 1):i185–94.

[44] Poulsgaard Gustav Alexander, Sørensen Simon Grund, Juul Randi Istrup, Nielsen Morten Muhlig, Pedersen Jakob Skou. Sequence dependencies and mutation rates of localized mutational processes in cancer. Genome Med 2023;15(1):63.

[45] Pratas Diogo, Silva Jorge M. Persistent minimal sequences of SARS-CoV-2. Bioinformatics 2021;36(21):5129–32.

[46] Pruitt Kim D, Tatiana Tatusova, William Klimke, Donna R Maglott. NCBI reference sequences: current status, policy and new initiatives. Nucleic Acids Res 2009;37(Database issue)):D32–6.

[47] Schoch Conrad L, Ciufo Stacy, Domrachev Mikhail, Hotton Carol L, Kannan Sivakumar, Khovanskaya Rogneda, et al. NCBI taxonomy: a comprehensive update on curation, resources and tools. Database: J Biol Databases Curation 2020;2020(January). https://doi.org/10.1093/database/baaa062.

[48] Shen Zhen, Bao Wenzheng, Huang De-Shuang. Recurrent neural network for predicting transcription factor binding sites. Sci Rep 2018;8(1):1–10.

[49] Silva Raquel M, Pratas Diogo, Castro Luísa, Pinho Armando J, Ferreira Paulo JSG. Three minimal sequences found in ebola virus genomes and absent from human DNA. Bioinformatics 2015;31(15):2421–5.

[50] Sims Gregory E, Jun Se-Ran, Wu Guohong A, Kim Sung-Hou. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. Proc Natl Acad Sci USA 2009;106(8):2677–82.
[51] Sun Jiahe, Lu Fang, Luo Yongjiang, Bie Lingzi, Xu Ling, Wang Yi. OrthoVenn3: an integrated platform for exploring and visualizing orthologous data across genomes. Nucleic Acids Res 2023;51(W1):W397–403.
[52] Tarone RE. A modified bonferroni method for discrete data. Biometrics 1990;46(2):515–22.
[53] Tsiatsianis Georgios Christos, Chan Candace SY, Mouratidis Ioannis, Chantzi Nikol, Tsiatsiani Anna Maria, Yee Nelson S, et al. Peptide absent sequences emerging in human cancers. Eur J Cancer 2024;196(January):113421.
[54] Tsirigos Aristotelis, Rigoutsos Isidore. A sensitive, support-vector-machine method for the detection of horizontal gene transfers in viral, archaeal and bacterial genomes. Nucleic Acids Res 2005;33(12):3699–707.
[55] Tuller Tamir, Chor Benny, Nelson Nathan. Forbidden penta-peptides. Protein Sci: A Publ Protein Soc 2007;16(10):2251–9.
[56] UniProt Consortium. UniProt: the universal protein knowledgebase in 2023. Nucleic Acids Res 2023;51(D1):D523–31.
[57] Vergni Davide, Gaudio Rosanna, Santoni Daniele. The farther the better: investigating how distance from human self affects the propensity of a peptide to be presented on cell surface by MHC class I molecules, the case of trypanosoma cruzi. PloS One 2020;15(12):e0243285.
[58] Wang Guliang, Karen MVasquez. Dynamic alternative DNA structures in biology and disease. Nat Rev Genet 2023;24(4):211–34.
[59] Yata Susumu "Prefix/Patricia Trie Dictionary Compression by Nesting Prefix/Patricia Tries *Proc 17th Annu Meet Assoc Nat Lang* 2011.
[60] Zhu Jacqueline Jufen, Cheng Albert Wu. JACKIE: fast enumeration of genome-wide single- and multicopy CRISPR target sites and their off-target numbers. CRISPR J 2022;5(4):618–28.