OXFORD

Briefings in Bioinformatics, 2025, **26(1)**, bbae607

https://doi.org/10.1093/bib/bbae607

Problem Solving Protocol

DiMA: sequence diversity dynamics analyser for viruses

Shan Tharanga[1], Eyyüb Selim Ünlü[2,3], Yongli Hu[1,4,‡], Muhammad Farhan Sjaugi[1,§], Muhammet A. Çelik[5,§], Hilal Hekimoğlu[6], Olivo Miotto[7,8], Muhammed Miran Öncel[6], Asif M. Khan[1,6,9,*]

[1]Centre for Bioinformatics, School of Data Sciences, Perdana University, MAEPS Building, Jalan MAEPS Perdana, Serdang, Kuala Lumpur 50490, Malaysia
[2]Istanbul Faculty of Medicine, Istanbul University, Turgut Özal Millet St, Topkapi, Istanbul 34093, Türkiye
[3]Genome Surveillance Unit, Wellcome Sanger Institute, Mill Ln, Hinxton, Saffron Walden CB10 1SA, United Kingdom
[4]Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, 8 Medical Drive, Singapore 117597, Singapore
[5]Celik Sarayı, Yeni Elektrik Santral St. No:29/2, Meram, Konya 42090, Türkiye
[6]Beykoz Institute of Life Sciences and Biotechnology, Bezmialem Vakif University, Ali Ihsan Kalmaz St., No.10 Beykoz, Istanbul 34820, Türkiye
[7]Nuffield Department of Clinical Medicine, University of Oxford, Old Road, Old Road Campus, Oxford OX3 7LF, United Kingdom
[8]Mahidol-Oxford Tropical Medicine Research Unit, Faculty of Tropical Medicine, Mahidol University, 420/6 Ratchawithi Rd., Ratchathewi District, Bangkok 10400, Thailand
[9]College of Computing and Information Technology, University of Doha for Science and Technology, Jelaiah Street, Duhail North, Doha, Qatar
*Corresponding authors. Asif M. Khan, College of Computing and Information Technology, University of Doha for Science and Technology, Jelaiah Street, Duhail North, Doha, Qatar. E-mail: asif.khan@udst.edu.qa
‡Present Address: Beyond Limits SG Pte Ltd, 13 Stamford Road, Capitol Singapore, 178905, Singapore
§Present Address: Birunisoft PLT, Kelana Square,Jalan SS 7/26, Kelana Jaya, Petaling Jaya 47301, Selangor, Malaysia
§Present Address: Biomedical Center (BMC), Physiological Chemistry, Faculty of Medicine, LMU Munich, Planegg-Martinsried, Germany

Abstract

Sequence diversity is one of the major challenges in the design of diagnostic, prophylactic, and therapeutic interventions against viruses. DiMA is a novel tool that is big data-ready and designed to facilitate the dissection of sequence diversity dynamics for viruses. DiMA stands out from other diversity analysis tools by offering various unique features. DiMA provides a quantitative overview of sequence (DNA/RNA/protein) diversity by use of Shannon's entropy corrected for size bias, applied via a user-defined *k-mer* sliding window to an input alignment file, and each *k-mer* position is dissected to various diversity motifs. The motifs are defined based on the probability of distinct sequences at a given *k-mer* alignment position, whereby an index is the predominant sequence, while all the others are (total) variants to the index. The total variants are sub-classified into the major (most common) variant, minor variants (occurring more than once and of incidence lower than the major), and the unique (singleton) variants. DiMA allows user-defined, sequence metadata enrichment for analyses of the motifs. The application of DiMA was demonstrated for the alignment data of the relatively conserved Spike protein (2,106,985 sequences) of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and the relatively highly diverse *pol* gene (2637) of the human immunodeficiency virus-1 (HIV-1). The tool is publicly available as a web server (https://dima.bezmialem.edu.tr), as a Python library (via PyPi) and as a command line client (via GitHub).

Keywords: viruses; proteome, genome; sequence diversity; diversity dynamics

Introduction

Viral infectious diseases are a major public health threat. There are more than 200 viruses that are considered infectious to humans [1]. Viral infections have the potential to be of pandemic proportion, as demonstrated by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), responsible for the recent global health crisis, COVID-19 [2]. The first infection was reported in Wuhan, China, which eventually spread to the world at large, infecting more than 776 million people, with over seven million deaths (as of October 2024; https://data.who.int/dashboards/covid19). There is a need to better understand viruses to help mitigate these devastating effects, in particular for pathogens that are of global priority, such as *Filoviruses and Flaviviruses,* among others (https://www.niaid.nih.gov/research/emerging-infectious-diseases-pathogen).

The SARS-CoV-2 pandemic exemplified the value of sequence data [3] in developing intervention strategies against viruses [4, 5], such as diagnostics, therapeutics, and prophylactics. Sequence data are a treasure trove to better understand viral evolution and interaction with the host. It allows the study of viral diversity, which can help identify the changes and evolutionary pressures. Sequence change of even a single amino acid can affect the recognition of a virus by the host immune system [6]. Viral diversity is a result of sequence substitutions that are primarily an outcome of evolutionary forces, such as mutation, recombination, and/or reassortment [7]. Among these, mutations have the most significant effect, the rate of which can range from 10^{-8} to 10^{-6} and 10^{-6} to 10^{-4} nucleotide sites, per cell infection for DNA and RNA viruses, respectively [8]. The variation within a single host can result in a spectrum of viral variants that are described as viral quasi-species [9].

Various studies have looked into the analyses of viral sequence variability and conservation. This has been facilitated primarily by alignment-dependent and to some extent alignment-free

Received: May 21, 2024. **Revised:** September 22, 2024. **Accepted:** November 13, 2024

approaches. The latter is used when data is big, or the diversity is prohibitively large to allow for a reliable alignment [10]. This approach, however, is not applicable if the goal is to study site-dependent substitutions [11]. An alignment-dependent approach allows us to study conservation and variability across the sequences of a viral species and enables comparative analysis across related species. There are various alignment-based approaches to study diversity, such as through sequence similarity search [12], phylogeny [13], and biological patterns [14] and profiles [15], among others.

A continuing goal is a greater understanding of viral diversity and an effective strategy to overcome the diversity for intervention applications [16]. There is a need for a more detailed and quantitative analysis of the extent of viral substitutions, including the composition and frequency (incidence) or probability of the different variants of the viral genome/proteome [17, 18]. Towards this, Hu *et al.* [19] presented a novel approach to dissect the dynamics of viral variability by use of diversity motifs for the human immunodeficiency virus (HIV)-1 clade B. This was subsequently expanded and applied by Abd Raman et al. [20] to the influenza A (H5N1) virus. The authors used Shannon's entropy [21, 22], determined for overlapping aligned peptides of length *k*, via a sliding window, for a panoramic overview of the viral proteome diversity landscape, and further, as the novelty, each *k-mer* position was dissected to various diversity motifs. The motifs quantify the diversity dynamics for each of the overlapping *k-mer* positions by evaluating the incidence of the distinct sequences present at the position, which are classified broadly into four diversity motifs: index, major variant, minor variants, and unique variants (Fig. 1A). The index is the predominant sequence at a given aligned position. It represents the prevalent wild-type sequence and if highly conserved and species-specific, may be an attractive target for a vaccine, drug, and diagnostic designs. The remaining ranked sequences at the position are variants to the index, with at least one amino acid (or nucleotide) difference. The major variant is the second most predominant sequence. The minor variants are distinct sequences that occur more than once and are of incidence lower than the major variant. Unique variants are observed only once in the alignment, and may be sequencing artefacts.

Herein, we present DiMA (Diversity Motif Analyser), a tool designed to facilitate the quantification and dissection of sequence diversity dynamics for any virus. It is publicly available both as a web server (https://dima.bezmialem.edu.tr) and as a standalone command-line interface (CLI), client tool (Python Package Index: https://pypi.org/project/dima-cli), useful for datasets of size beyond the web server limit. DiMA is big data ready, allows for user-defined sequence metadata enrichment, and enables comparative protein and DNA/RNA sequence diversity dynamics analyses, within and between sequences of viral species. While the entropy function is also offered by various other tools, the novel key distinguishing feature of DiMA is the use of diversity motifs to study sequence diversity dynamics, besides integrating entropy and metadata.

Materials and methods

Overall workflow

DiMA workflow is illustrated in Fig. 1B. The workflow can be divided into three parts, input, process, and output. The input to DiMA is simply a multiple sequence alignment (MSA) file of either protein or DNA/RNA sequences (in aligned FASTA format). The sequences can be obtained from publicly available primary/derived databases, such as NCBI Nucleotide/Protein/Virus databases [23] and ViPR [24], or specialist databases, such as Influenza Research Database (IRD) [25], among others. Both full-length and partial sequences are recommended for a comprehensive survey of diversity. Primary databases are prone to errors [26–28], and thus derived databases, which are integrated, value-added resources, help minimize these issues. NCBI Virus is recommended for its ease of use and data quality; nonetheless, sequence data records of interest should be reviewed for any irrelevant sequences and possible issues, such as errors, discrepancies, and anomalies [29]. Sequence duplicates (full-length or partial match) are common in public databases, which may be an artefact (multiple uploads of the same isolate) or an actual indication of virus occurrence in nature (multiple uploads of different isolates with the same sequence). Thus, it may be desired to analyse both the redundant and non-redundant datasets [30].

There are many tools available for MSA (see the next section for more details). Given the less optimal, heuristic nature of MSA [31–33], all resulting alignments should be manually inspected and corrected for any misalignments; short partial sequences can be a common cause of spurious alignments. The tool EMBOSS Seqret (https://www.ebi.ac.uk/jdispatcher/sfc/emboss_seqret) may be used to convert an aligned file to an aligned FASTA format for use in DiMA. DiMA provides a quantitative measure of sequence diversity by use of Shannon's entropy, applied via a user-defined *k-mer* sliding window on the alignment. The entropy value is corrected for sample size bias by applying a statistical adjustment. DiMA interrogates the diversity dynamics by dissecting each *k-mer* alignment position to various diversity motifs. Collectively, DiMA outputs plots of the entropy values, diversity motifs, and user-enriched metadata (such as year, host, and country) for each of the *k-mer* positions, providing a holistic view of the diversity and its dynamics.

Input and various considerations

Input MSA

MSA is an important tool in bioinformatics that enables an evolutionary assessment of sequence changes, providing key insights into genetic relationships and functional conservation across different species [34]. A detailed, step-by-step workflow for generating high-quality MSA can be found in the DiMA user manual (https://dima.readthedocs.io/en/latest/). This workflow begins with sequence cleaning, a crucial step involving the deduplication of identical sequences, as well as the identification and removal of contaminant sequences and regions.

The next critical step is selecting an appropriate MSA tool tailored to the specific dataset. Several MSA tools are available, employing different algorithms to handle the complexity of aligning diverse and large numbers of sequences. Prominent examples include MUSCLE [35], Clustal Omega [36], MAFFT [37], and MAGUS [38], among others. MAGUS, in particular, is designed for large-scale alignments and can handle datasets containing up to one million sequences [38]. Recent benchmarks comparing the performance of these tools on various datasets [39, 40] can provide valuable guidance in selecting the most suitable MSA tool for specific research needs.

Aligning large sample sizes (e.g., over 20 000 sequences) poses significant computational challenges, often requiring substantial memory and processing power. To mitigate these demands, many MSA tools offer faster, though potentially less accurate, algorithms. Additionally, leveraging parallelized versions of these tools on high-performance computing systems can significantly reduce

Figure 1. DiMA at a glance. **A**. Diversity motif definition for a sample *nonamer* position of 20 sequences. **B**. DiMA workflow. **C**. DiMA server results page. Submission summary (top panel): Overview of information about the input alignment, such as the query data name, alignment length, the support threshold, and the support at the selected *k-mer* position, among others. *K-mer* position entropy (middle panel): Entropy value of the selected *k-mer* position indicates the level of variability among the sequences at the position, with zero representing a completely conserved position. The entropy plot provides a panoramic view of the diversity and is responsive and interactive. Motif distribution (bottom panel): All sequences at the selected *k-mer* alignment position are quantified for distinct sequences and are ranked classified into diversity motifs based on their incidences (probability). *K-mer* position sequences (bottom panel): Distinct sequences at the selected aligned *k-mer* position are listed according to their corresponding incidences. Metadata (bottom panel): If the header tags are provided, DiMA will show a pie chart for each type of metadata.

computational time. For ultra-large-scale alignments involving over 100 000 sequences, where computational resources are a limiting factor, reference-based alignments provide a more feasible solution. However, it is important to note that this approach may lead to a loss of diversity representation due to its reliance on a reference sequence [41].

The accuracy and reliability of an MSA are influenced by various factors, including the number of sequences, their similarity levels, the frequency of insertions and deletions (indels), and the lengths of the sequences involved [42]. The accuracy tends to decrease with the inclusion of partial sequences, low complexity regions, and large number of sequences, highlighting

the need for careful inspection and validation of alignment outputs. Notably, all MSA tools employ heuristic methods to generate alignments, which are not guaranteed to be optimal. As a result, manual inspection and correction of alignments are often necessary, especially when dealing with highly diverse sequences and large datasets [32, 43]. Several auxiliary tools, such as GUIDANCE [44], SATe [45], HoT [46], Gblocks [43], and SuiteMSA [47] have been developed to facilitate this process. These tools assist in improving alignment quality by providing metrics for alignment confidence or by refining the alignment to reduce errors. However, they do not entirely eliminate the need for manual inspection.

Based on our experience with viral sequence data, we have observed that the performance of MSA tools can vary significantly depending on the organism and the dataset in question. A tool that performs well on one dataset may not necessarily be the best choice for another, underscoring the importance of empirical testing and parameter optimization. Therefore, while any of the commonly accepted MSA tools with default parameters can serve as a starting point, it is advisable to use alignment correction tools for refining the alignment. A final manual review is essential to identify and correct any obvious errors that may have been missed by the automated tools.

The input MSA file (up to 100 MB; larger files are possible with the CLI version) is accepted only in an aligned FASTA format (sample input files are provided on the web server), and the analysis can be customized through various parameter settings and features:

i) k-mer length.

DNA/RNA or protein multiple sequence alignments are typically analysed by evaluating the sequence composition within individual alignment positions, known as *1-mer* analysis. This approach provides a basic understanding of sequence diversity, but does not capture more intricate patterns that are crucial for understanding the functional and structural nuances of the sequences. In contrast, *k-mer* entropy offers a more comprehensive analysis by examining overlapping subsequences of length k (e.g., $k=2$ for dinucleotides or dipeptides). This method captures additional contextual information about sequence neighborhoods that *1-mer* analysis overlooks [14, 48]. In biological systems, nucleotides or amino acids do not function in isolation; they interact with neighboring elements, contributing collectively to the structural conformation and functional properties of the entire molecule. While *1-mer* analysis can provide insights into individual sequence positions, analysing sequences in terms of *k-mer*s reveals complex patterns such as motifs or domains that are integral to the biological function and are missed at the *1-mer* level.

DiMA enhances the ability to analyse sequence alignments by allowing users to choose any *k-mer* window length, including *1-mer*, enabling a thorough examination of the alignment across various *k-mer* window sizes. When selecting an appropriate *k-mer* window length (ranging from *1-mer* up to the full length of the alignment), several factors should be considered, including the length of the alignment, the number of sequences, the type of sequence (DNA/RNA or protein), and the specific objectives of the analysis.

The choice of *k-mer* length can be guided by the length of the shortest and longest known motifs or domains relevant to the sequence type being studied. For example, the shortest biologically relevant protein *k-mer* size can be as small as three amino acids, corresponding to β-turn motifs [49]. This can be further refined by incorporating knowledge about the expected motifs or domains in the aligned sequences. Protein sequences typically provide more detailed information compared to DNA/RNA sequences due to the diversity of 20 amino acids versus four nucleotides (where three nucleotides encode a single amino acid). Consequently, equivalent corresponding *k-mer* windows for DNA/RNA sequences would need to be longer to capture similar levels of complexity. For alignments with a long length of sequences, a larger *k-mer* size may be preferable to achieve data smoothing. DiMA defaults to a window size of nine amino acids for protein alignments and 27 nucleotides for DNA/RNA alignments, but users can adjust the window size to fit their specific analysis requirements. The purpose of the diversity study should also guide the choice of *k-mer* length. Genetic diversity is defined by variations at the nucleotide level; however, not all genetic changes translate into protein-level variations, as only non-synonymous substitutions lead to amino acid changes. Furthermore, not every amino acid change affects antigenic properties or immune recognition. For example, a single amino acid alteration might not disrupt the recognition of an existing epitope or create a new one. In our study of antigenic diversity within the context of the cellular immune response, we used a *k-mer* length of nine (nonamer) because human leukocyte antigen (HLA)-I molecules typically bind peptides ranging from eight to 15 amino acids, with nine being the most common length [50]. HLA class II molecules can accommodate longer peptides, up to 22 amino acids, with a binding core of nine amino acids [30, 51, 52].

ii) Metadata Parsing.

The description/definition line or headers of the input-aligned sequences in FASTA format can be utilized to annotate the sequence with meaningful metadata, which can provide additional dimensions to viral diversity studies. Such metadata are often available in public databases or could be from in-house findings. The NCBI Virus database, e.g. provides up to as many as 23 standard metadata annotations to each sequence record, such as host, biosample, geolocation, and year of isolation, among others. The input file can be formatted to include metadata as desired by the user, whether taking advantage of information from the public database records and/or from in-house records. DiMA provides a feature to parse and tag such metadata to the sequence headers of the input alignment file. It currently provides six pre-defined, commonly used header tags, such as 'Accession', 'Strain', 'Species', 'Year', 'Geolocation', and 'Host'; and users can also define their own custom tags. The pipe ('|') character is used as a delimiter between the tags, while the order of the metadata in the input file and those of the tags should match. As the number of available metadata can vary between sequences in an input file and since the order is important, empty pipes ('||') that reflect unknown/unavailable metadata, must be added to preserve the order of the header tags. Two examples of sequence headers are 'AGN52936.1 |*Homo sapiens*|United Arab Emirates|2013' and '>YP_009047204.1 |*H. sapiens*||2012', which correspond to the header tags as follows: 'Accession', 'Host', 'Geolocation', and 'Year', in that order. The NCBI Virus database allows customization of the FASTA header line, with the inclusion of empty pipes where applicable. The user-defined header tags in DiMA would be accordingly ascribed to all the *k-mers* generated from the input-aligned sequences.

iii) Support Threshold.

The number of sequences in an input alignment file affects the reliability of insights gained from the data. Ideally, large sample size is preferred to mitigate the effect of any potential data biases. Herein, 'support' is defined as the number of sequences at a given *k-mer* position that do not harbor a gap and/or an

unknown/ambiguous nucleotide base or amino acid residue. The *k-mer* positions that are below a user-defined 'support' threshold (default is 100 sequences, arbitrarily defined) are referred to as 'low support' (tagged as 'LS'). A position of 'no support' (tagged as 'NS') is possible when all of the aligned sequences at a given *k-mer* position have a combination of gap(s) and/or unknown/ambiguous character(s). Support may vary between *k-mer* positions due to the inclusion of incomplete or partial sequences into the alignment.

Process

Shannon's entropy

Shannon's entropy, originally introduced as a theory of communication [21, 22], is used to measure the disorder for a given variable and has been widely adopted in biology as a diversity index, including for sequences [30, 52]. DiMA calculates the entropy, *H(x)* for each overlapping, *k-mer* alignment position *x* of the input alignment file by applying Shannon's formula:

$$H(x) = -\sum_{i=1}^{n(x)} p(i,x) \log_2 p(i,x)$$

where *p(i,x)* is the relative incidence of a given *k-mer* sequence *i* at the alignment position *x* (start position of the aligned *k-mer*) and *n(x)* is the number of distinct *k-mer* sequences at the position. These two factors affect the entropy value, whereby it is high for a position with a large number of distinct *k-mer* sequences and is low when a sequence exhibits a clear high probability. The minimum possible entropy value is 0, which means complete conservation of a given *k-mer* position, whereby only one distinct *k-mer* sequence is observed in all (100%) of the aligned sequences analysed. The maximum entropy value would be dependent on the *k-mer* size of choice and is applicable when all possible outcomes (permutations with repetition) are observed with equal probability. The maximum peptide entropy value for a *k-mer* window size of nine is ~ 39 (i.e., $\log_2 20^9$). This, however, is theoretical for biological sequences given that conservation is expected among homologs, and some combination of *k-mers* may not be sterically possible. The highest nonamer entropy value that we have thus far reported is 9.2 for HIV-1 clade B envelope protein [19], and it may be used as a benchmark for comparative proteome diversity dynamics analyses.

The entropy computation for a given *k-mer* position is dependent on the number of sequences in the input alignment. Only *k-mer* sequences that do not harbor a gap and/or unknown/ambiguous nucleotide base or amino acid residue are used for the entropy computation. Entropy is calculated irrespective of the support tag ('LS' or 'NS') for a position. NS positions are maintained to keep the input alignment length intact, with a default assigned pseudo entropy value of zero, which is cautioned with the inclusion of the NS tag in the results page and the downloadable output files.

Each *k-mer* alignment position's raw entropy value is corrected for sample size bias [53] by applying a statistical adjustment to the input alignment, which estimates entropy values for infinitely sized resampled alignments with the analogous *k-mer* distribution [30]. As illustrated in Fig. 2, this is done by resampling the sequences that can be analysed (*N*) at a *k-mer* position and creating alignments of varying sizes (number of sequences), determined through a systematic random sampling approach. The entropy values are measured for each sample size, plotted against the respective sample 1/*N*, and a linear regression is used to extrapolate the entropy estimate for the *k-mer* position when $N \to \infty$. This algorithm was tested through resampling of an alignment with a wide range of sample sizes (from 100 to 100 000 number of sequences). The corrected and uncorrected entropy results were benchmarked against the uncorrected entropy calculation of 1 000 000 sample size (Supplement Fig. S1 available online at http://bib.oxfordjournals.org/). Resampling for an infinite-size set estimate is only done for positions when N is higher than the support threshold (*T*); otherwise, all of *N* is used for a direct computation of the entropy. An exceptional low support ('ELS') tag is used in the results page and downloadable output files when $N = T$, simply as a way to distinguish from when $N < T$. In the case of N only slightly higher than T and where regression may potentially result in a negative entropy value due to small number of resamples, all of N is used for a direct compute of the entropy.

Motif classification

DiMA interrogates the diversity dynamics at each *k-mer* alignment position by classifying the distinct sequences into diversity motifs, based on their incidence (Fig. 1A). The most frequent *k-mer* sequence is termed as the 'index', with 'major variant' as the second most common and 'unique variants' as the least, occurring only once, while 'minor variants' are in between these two variant motifs. As per the motif definition, certain *k-mer* positions may not exhibit specific motifs, such as no index, major or minor, and only singletons are observed. Also, in some instances, a position may exhibit more than one index or major variant, which is observed when two or more distinct *k-mer* sequences are of the same incidence. The term 'total variants' encompasses all the variants that are at least one nucleotide/amino acid different from the index. The term 'distinct variants' is the count of the different *k-mer* sequences among the total variants (major, minor, and unique variants). The relative frequency of each motif's distinct sequences, expressed in percentage (as incidence) is computed by DiMA for each of the *k-mer* positions. The denominator used for the incidence computes is *N* (for each of the motifs), except for the term 'distinct variants', where the total variants count is used. As such, a value of 100% for 'distinct variants' will only be attained when all the *k-mers* are unique variants (no major and/or minor variants). Additionally, DiMA identifies historically conserved sequences (HCSs) across the input alignment length, selected based on a user-defined index incidence threshold (default: 100%). Selected index sequences that overlap or are immediately adjacent in their *k-mer* positions are concatenated as longer sequences. An HCS of reasonable length of 80% or higher incidence may be attractive for further consideration as an intervention target.

Stratification by metadata

At each *k-mer* alignment position, an individual distinct sequence is populated with the corresponding metadata header tags from all the input alignment sequences that share the same distinct sequence. Accordingly, the relative incidence for each of the different constituent items of each header tag is calculated and plotted.

Output

Interactive interface

DiMA outputs a result page with 12 different panels, organized into three parts (top, middle and bottom) that offer multiple facets of sequence diversity (Fig. 1C). The multiple panels at the top are akin to a dashboard that provides the user with a quick summary of information general to the input alignment and specific to a given *k-mer* position: i) query name, ii) sequence type, iii) *k-mer*

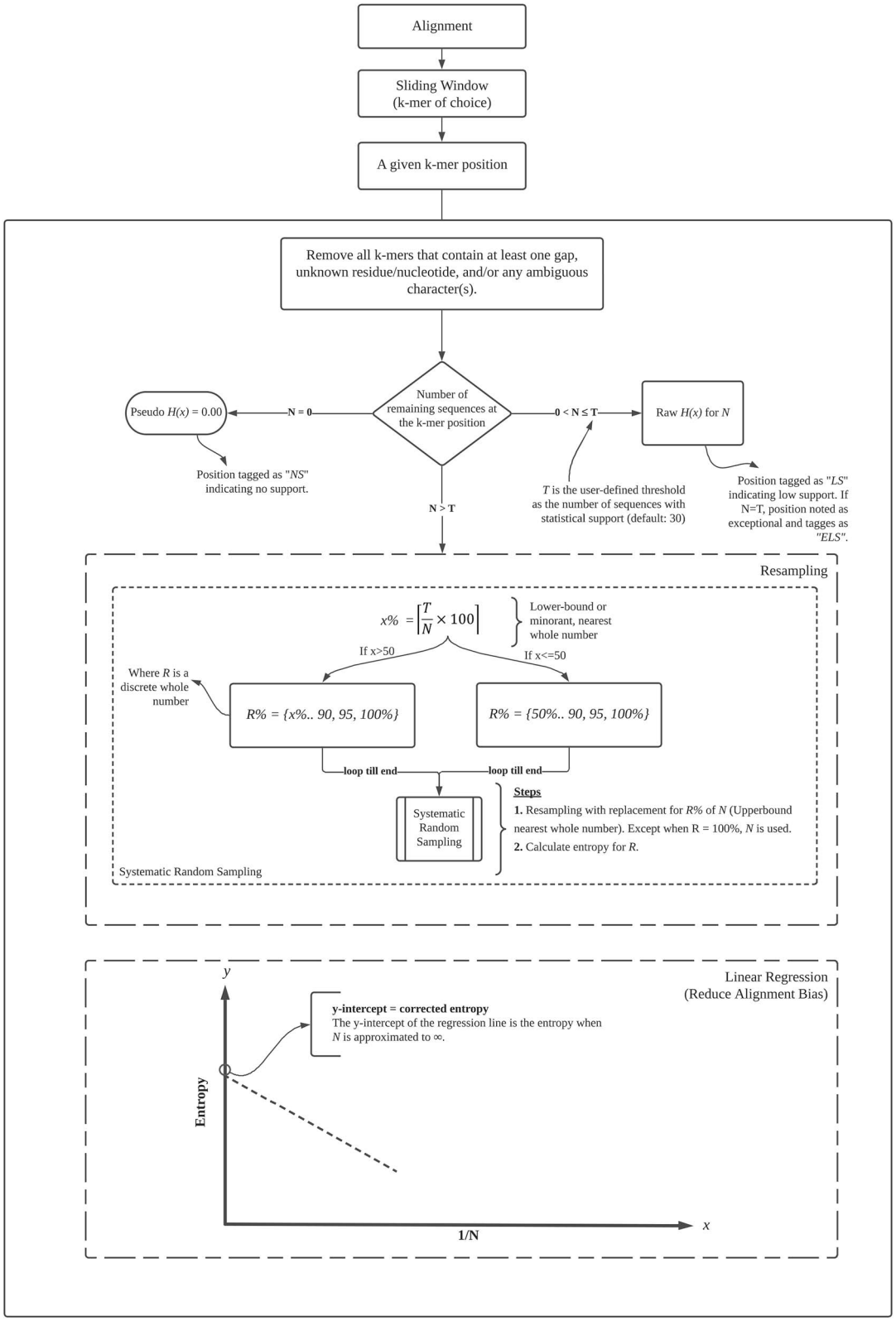

Figure 2. Workflow of entropy correction algorithm to address sample size bias.

length, iv) support threshold, v) alignment length, vi) sequence metadata, vii) position, viii) position's entropy, support and distinct variant count and ix) download results. The single middle panel (x) showcases the entropy plot for all the *k-mer* positions of the input alignment, providing an overview of the sequence diversity, with low entropy values representing conservation and high indicating variability; hovering the cursor over a specific position on the plot will reveal the applicable entropy value, and clicking a position will switch to that position. Positions with 'NS' and 'LS' support tags have been highlighted with different colors

on the plot. The three panels at the bottom provide information for the selected position, namely xi) motif distribution and xii) *k-mer* position distinct sequences, whereby the selection of a specific distinct *k-mer* sequence would illustrate all the relevant xiii) sequence metadata, with a drop-down menu option for the different header tags defined by the user. Changing the selected *k-mer* position (in panel vii) would dynamically update the information in panels (xi to xiii).

Download

The top panel (ix) of the DiMA output allows for downloading of the analyses results in JSON and XLSX formats. The JSON file provides the complete analyses results as key-value pairs, which can be viewed using a public JSON viewer tool (such as https://jsonformatter.org/json-viewer). Separately, the XLSX file provides for easier viewing through the MS Office Excel application or an equivalent that supports the format. A concatenated list of HCS based on a user-defined index incidence threshold can also be downloaded (in JSON format) and viewed in a text editor or JSON viewer.

DiMA implementation

DiMA is designed using the Python language, with bindings written in Rust. This allows DiMA to remain highly performant while being accessible to the larger Python user base. Designed with big data in mind, DiMA can handle large datasets while still maintaining a relatively small memory footprint. Furthermore, DiMA implements data parallelism where each *k-mer* position is handled in parallel, making it efficient and scalable. In addition, the dataset uploaded by the user on the web service is validated for MSA and header consistency using functions written in Rust, compiled down to WebAssembly. This allows for efficient validation of large datasets and mitigates user's submission from failing once sent to our server for processing. DiMA is compatible with common web browsers, including Chrome 98.0, Firefox 56.0, Edge 94.0, and Safari 15.0.

Results

Application

We demonstrate two applications on the use of the DiMA web server for viral sequence diversity dynamics studies. One is using the Spike protein of *severe acute respiratory syndrome coronavirus-2* (SARS-CoV-2), with raw sequence data and metadata downloaded from the GISAID [54] database. The second is the DNA polymerase (*pol*) gene of human immunodeficiency virus type 1 (HIV-1) group M, with curated alignment sourced from the HIV Sequence Compendium 2021 [55], made available on the Los Alamos HIV sequence database [56]. SARS-CoV-2 Spike protein plays a pivotal role in host cell entry by binding to angiotensin-converting enzyme 2 (ACE2) receptors, and thus is a key target for developing therapeutics to block viral entry [57]. Spike protein is composed of two subunits (S1 and S2), with multiple protein domains within each. HIV-1 *pol* gene is translated into a protein that provides essential enzymatic functions critical for viral replication and has been a major target for drug design [58, 59].

Spike protein of SARS-CoV-2

All 11,808,363 available Spike protein sequence records were downloaded (as of July 2022) and deduplicated by use of CD-HIT v4.8.1 [60]. The resulting 2,106,985 non-redundant sequences were multiple sequence aligned by use of an in-house reference guided MSA tool, RefAligner (using the GISAID accession ID 'EPI_ISL_402124' sequence as a reference). The aligned dataset was analysed using the command-line version of DiMA (DiMA-CLI), which enables big data analysis, with metadata parsing enabled and minimum incidence for HCS concatenation set to 99.95% (to retain a reasonable number of HCS), while other parameters were set to the default. The analysis results are made available on the DiMA web server (https://dima.bezmialem.edu.tr/results/6cfb645e-eefa-40db-ac4c-43a6712b2760).

A total of 1265 aligned overlapping nonamer (9-mer) positions, covering the full-length of the alignment, were analysed for antigenic diversity. The protein was generally conserved (Fig. 3A) with a nonamer entropy range of < 0.01 to ∼1.9 bits (compared to a theoretical maximum of 38.8 bits) for the full-length protein, and ∼0.01-1.91 and < 0.01-1.59 bits for the subunits S1 and S2, respectively. The average entropy for the full-length protein was ∼0.28 bits, while the average for subunits S1 and S2, were ∼0.40 and ∼0.18 bits, respectively. The high conversation is consistent with the relatively recent evolutionary emergence and, thus, the short history of the virus. However, S2 showed even higher conservation given the need for functional preservation [61].

The nonamer position 675 *(the number refers to the starting alignment position of the k-mer; i.e., 675-683)* peaked as the most diverse, with the highest observed entropy value of ∼1.97 bits. This position had a total support of 2 208 211 sequences, with 343 distinct nonamers. This nonamer region (675-683) overlaps the furin cleavage site (681-684), a site known to facilitate the virus-cell membrane fusion. Variations in this region affect viral infectivity and transmissibility [62]. The predominant distinct nonamer (QTQTNSRRR) or the index at this position occurred in 876 048 of the sequences analysed, and thus, exhibited an incidence of ∼39.67%. The remaining 342 nonamers at this position were variants to the index and occurred across the 1 332 163 remaining sequences analysed. The total variants (∼60.33%) of the position comprised of a major variant (QTQTKSHRR; 833,415; ∼37.74%), which corresponded to the Wuhan reference (Genbank ID: YP_009724390.1); 234 distinct minor variants (∼22.58%); and 108 unique variants (∼0.005%). Among the total variants, the total number of 'distinct variants' was 343 (∼0.03%). The index sequence has been reported to be a putative T-epitope [63] and exhibited reduced binding affinity to MHC-I, relative to the wildtype Wuhan reference, which had diminished in incidence and became the major variant. This reflects an evolutionary pressure at the position facilitating immune escape. Sequence metadata visualization for the nonamer position 675 showed that the index was observed primarily in *H. sapiens* (∼99.84%), with the remaining (∼0.16%) in other host species such as *Canis lupus familiaris*, *Panthare leo*, *Neovison vison*, *Fells catus*, *Odocoileus virginianus*, *Mustela lutreola*, and environmental sources. While the index was widespread in 29 countries, it was most commonly detected in two geographical locations, predominantly the United States (∼35.33%), followed by the United Kingdom (∼13.73%). Such a distribution may suggest regional adaptation and selective pressures favoring the specific 9-mer. The index nonamer was most common among sequences collected in September 2021 (∼14.07%), and with a decreasing frequency trend over the following three months (∼12.5, ∼11.4, ∼7.8). During the same period, the major variant exhibited an increasing frequency trend, suggesting a temporal shift in fitness.

DiMA identified five HCSs of lengths ranging from nine to 23 amino acids which merit further investigation as potential targets for vaccine design, inhibitory drug analysis, and diagnostic studies. All the five HCSs were observed to match reported SARS-CoV-2 T and B-cell epitopes in the human host. Separately, Fuente *et al.* [64] described four of the HCSs and assessed

Figure 3. Applications of DiMA for the diversity dynamics studies of SARS-CoV-2 spike protein and HIV-1 *pol* gene. **A**. The highest entropy observed was ~1.97 for SARS-CoV-2 spike protein at position 675 *(the number refers to the starting alignment position of the k-mer)*, which was selected for motif distribution and metadata view of each distinct *k-mer* at the position. **B**. HIV-1 *pol* gene exhibited various conserved and highly diverse positions. **C-D**. Motif distributions of the highly diverse position (2177, highest entropy) and the highly conserved position (3046, lowest total variants), respectively, in HIV-1 *pol*.

their immunogenicity to design epitopes stable against future mutations. Similarly, two of the HCSs matched known drug targets reported in public repositories, indicating their potential for drug design [65, 66]. Diagnostically, all the five HCSs were broadly conserved (revealed by BLAST searches) and can be used to detect other species members of the *Coronaviridae* family. A lower HCS threshold may be used to identify those that are SARS-CoV-2 species-specific for diagnostic purposes.

Pol gene of HIV-1 group M

The *pol* gene data was downloaded (as of July 2024) as a pre-curated alignment of 2637 sequences, without any metadata, from the HIV Sequence Compendium 2021, and thus, did not require any data processing. DiMA analysis of the gene was performed with the following parameters: overlapping *k-mers* of 27 nucleotides each (corresponding to nine amino acids) for antigenic diversity, default support threshold (100), HCS threshold arbitrarily set to 70% (one HCS was observed at ~88%), and metadata parsing disabled. The analysis results are made available on the DiMA web server (https://dima.bezmialem.edu.tr/results/89163c56-70a3-456e-8c71-5f3276804311).

A total of 3364 aligned 27-mer positions were analysed to assess genetic diversity corresponding to antigenic diversity across the *pol* gene. Of these positions, 564 (~16.76%) were classified as low support (LS), and 680 (~20.21%) as no support (NS). Excluding LS and NS positions, the gene exhibited entropy values ranging from ~0.86 to ~9.57 bits, with an average entropy of ~5.35 bits (Fig. 3B), compared to a theoretical maximum of 54 bits. This range of entropy values reflects the spectrum of genetic diversity present within the gene, indicating that some regions are highly variable while others are more conserved. The relatively high average entropy suggests that the gene maintains

considerable sequence diversity throughout its length, which is consistent with the high mutation rate and genetic variability characteristic of HIV-1. Such diversity is often a hallmark of viruses that must rapidly adapt to host immune responses [67]. However, despite this overall high genetic diversity, the protein nonamer entropy (data not shown) is notably lower, with an average of ∼1.91 bits (out of ∼ 39). This lower entropy at the protein level suggests strong purifying selection on the protein's sequence, likely due to its essential role in the virus's replication cycle. The disparity between gene and protein entropy indicates that a significant proportion of the nucleotide variations are synonymous substitutions. These substitutions do not alter the amino acid sequence of the encoded protein, which could provide a survival advantage by allowing the virus to evade immune detection without compromising the functional integrity of its essential proteins, such as transcriptional and translation efficiency [68], regulation of gene expression [69], immune system evasion through RNA structure [70], and maintenance of essential functions while allowing diversity [7]. This observation aligns with the understanding that HIV-1 utilizes both high genetic variability and selective pressures to balance adaptation and maintain vital functions.

The *k-mer* position 2177 of the gene exhibited the highest entropy of ∼9.57 bits (Fig. 3C), illustrating the complex composition of the viral variant population (conversely, Fig. 3D illustrates for the highly conserved position). These 27-mer region overlap with the RNase H domain (aa positions 681-684 of the HXB2 reference sequence), a critical site for viral replication. Variation in this region can enhance viral replication and/or enhance resistance to antiretroviral therapies [71]. The index 27-mer at position 2177 was only observed at an incidence of ∼4.44%. The minor variants were instead the principal variant with a collective incidence of ∼54.06% at the position. This suggests that this position of the *pol* gene from the reported viral population is under selective pressure, potentially driven by the enhancement of replication efficiency and/or the ability to evade drug pressure.

Two HCSs were identified of lengths 36 and 34 nucleotides, both of which corresponded to the integrase protein domains, essential for the insertion of viral DNA into the chromosome of the host cell [72]. The amino acid translation of one of the HCS was observed to match reported HIV-1 T-cell epitope in the human host. The epitope (KRKGGIGGY) is presented by the HLA-B∗27:05 allele, which is associated with a more effective host response to HIV-1 and slower disease progression [73].

The above two applications are cursory analyses of a protein and a gene alignment by DiMA. Output data from DiMA can be analysed further in various ways, such as i) scatter plot of the relationship between entropy and incidence of total variants; ii) scatter plot of motif incidence (for each diversity motif) against total variants; iii) incidence distribution violin plots of the diversity motifs; and iv) distribution of conservation level of index incidence for *k-mer* positions. This can be done for individual proteins/genes of a virus and further pooling of the results can allow for a genome/proteome-wide analyses. Examples of such analyses are demonstrated in Hu *et al.* (2013) and Abd Raman *et al.* (2020). Further, a notable finding from these studies was motif switching, a phenomenon where the fitness change of one or more nucleic or amino acids, changed the incidence, and thus the motif of a given *k-mer* distinct sequence across its *k-mer* positions. Motif switching has been observed to involve all the diversity motifs studied herein. Hence, DiMA results also serve as a starting point to unravel and understand the complexity of motif switching.

Advancing viral sequence diversity and evolution studies with DiMA

DiMA is an enabler of research questions that could not be answered easily in the past due to the lack of a robust tool that could handle the large number of viral species and their large number of sequences for comparative analyses. DiMA addresses these challenges through its novel features (Table 1), which can be summarized as follows: i) ability to analyse diverse data types and integration with big data; ii) corrected entropy calculations and capturing local (neighboring) effects through *k-mer* sliding window; ii) comprehensive analysis of viral diversity; iii) facilitate interpretation of sequence variability and evolutionary dynamics; and iv) cataloguing of HCSs. These features open the door for the study of any viral species with sufficient data, allowing viral sequence diversity analysis within a species (intra- and inter-genes/proteins) and between species (inter-genomes/proteomes). The insights gained from the analysis of a species, as illustrated in the application section above, can serve as a catalogue for the species, allowing for comparative analyses to others. Such a comparison, as demonstrated in our earlier work [19, 20, 30, 74, 75] provided a holistic view and deeper understanding of the viral diversity landscape, providing insights in terms of viral evolution, with implications for the design of diagnostic, therapeutic, and prophylactic interventions. Compared to existing studies, a key novel feature of DiMA is its ability to provide granular insights into, underlying distinct *k-mer* sequences, represented as diversity motifs. These motifs reflect the overall diversity, measured as entropy, and are enriched with metadata that capture spatial and temporal dimensions. The effects of these diversity motifs were correlated to the biology of the virus, highlighting the dynamic roles of the index and variants in influencing the evolutionary fitness. This approach also has the potential to identify candidate targets (HCS) for developing intervention strategies. DiMA paves the way for application to all viruses, both well-known and understudied, offering opportunities to corroborate existing findings and/or uncover new insights.

Discussion

DiMA stands out from other diversity analysis tools, such as Protein Variability Server (PVS) [76] and Los Alamos (LANL) (http://www.hiv.lanl.gov/) virus databases Entropy tool, by offering various unique features. A comparison table between these tools and DiMA can be found in our help page (https://dima.readthedocs.io/en/latest/#novel-features) and also provided herein (Table 1). Notably, DiMA is capable of handling both nucleic acid and amino acid sequences. Shannon's entropy is calculated for a user-defined *k-mer* size of a sliding window, which is better suited to capture the local (neighboring) effect of sequence substitutions, in particular when dealing with longer sequences. A statistical adjustment is applied to the computed entropy values for sample size bias correction. Distinctively, the key feature of DiMA is that it interrogates the diversity dynamics by dissecting each *k-mer* alignment position to various diversity motifs, based on the incidence of distinct sequences. Moreover, it allows for metadata enrichment of the motifs. Additionally, DiMA identifies HCSs across the input alignment length, selected based on a user-defined index incidence threshold, which are concatenated when overlapping or adjacent. DiMA outputs a result page with 12 different panels, organized into three parts (top, middle and bottom) that offer multiple facets of sequence diversity and are largely interactive. The input to DiMA is simply a multiple

Table 1. Novel features of DiMA in comparison with other web servers for viral sequence variation analysis.

Features	DiMA[a]	PVS[b]	LANL[c]	BV-BRC[d]
Analysis of nucleic acid - amino acid sequences	✓ - ✓	× - ✓	✓ - ✓	✓ - ✓
Shannon Entropy on user-defined sliding window	✓	×	✓	×
Entropy correction for size bias	✓	×	×	×
k-mer/variant frequency calculation	✓	×	×	✓
k-mer diversity motifs classification	✓	×	×	×
Metadata inclusion	✓	×	×	×
Identification of historically conserved sequences	✓	✓	×	×
Multiple interactive visualizations	✓	×	×	×
Web service input size limit	100 MB[e]	~0.2 MB	Not defined	Not defined

[a] https://dima.bezmialem.edu.tr [b] http://imed.med.ucm.es/PVS/ [c] https://www.hiv.lanl.gov/content/sequence/ENTROPY/entropy.html [d] https://www.bv-brc.org/app/MSA [e] Analysis of larger files possible with CLI version (https://github.com/BVU-BILSAB/DiMA) which there is no limit.

sequence alignment file of size up to 100 MB, while larger files can be accommodated by use of the command line (CLI) version.

DiMA is a tool that enables sequence diversity dynamics analysis for any virus of interest. It is big data ready and applicable to DNA/RNA/protein sequence data, with a user-friendly interface and a detailed user manual provided for a broader appeal. DiMA enables comparative sequence diversity dynamics analyses for a better understanding and insight within and between DNA/RNA/protein sequences of a virus species or genomes/proteomes of different viral species, whether at the genus, family, or higher lineage taxonomy rank (limited by the feasibility of a reliable alignment). The diversity motifs represent inherent patterns in the organization of the large number of sequences that facilitate cooperative virus fitness selection, whereby the spectrum of variants allow the virus population to explore changes in selection pressure for maximum reproductive fitness [77] and support long-term evolvability [77, 78]. In some cases, there may be functional constraints [79], such as multiple host dependency limiting the virus population from a fitness peak [20, 30] or in contrast, the presence of extremely variable hotspots contributing to the general collapse of the early immunity or facilitating immune escape [19]. Examining viral variants, leveraging on various metadata dimensions of interest, such as spatio-temporal, host, and clinical phenotypes, among others, may provide important key insights into the evolution of the virus.

Sequence diversity at highly variable positions appears to be embodied in minor and unique variants, which can exist as many different (distinct) sequences. DiMA can help elucidate variant sequence structure and incidence with increased total variants. Thus, provides a compendium of the possible spectrum of sequence variants for a virus of interest. The viruses can range from being highly conserved, such as West Nile virus [80] to extremely variable and highly plastic, such as HIV-1 [19]. DiMA analyses of different viral species can provide a catalogue of HCS targets for a comparative and rational design of new intervention strategies. DiMA can potentially be used for non-viral pathogens, such as bacteria and fungi, among others.

Limitations

DiMA's effectiveness is dependent on the quality of the alignment, which is constrained by the heuristic nature of MSA. This process relies on the subjective manual inspection and correction of misalignments by the user. Additionally, DiMA currently focuses only on sequence diversity in the context of substitutions and does not consider other structural sequence variations, such as copy number changes. Shannon's entropy is just one of the many diversity metrics available; other metrics may be explored in future iterations of DiMA. The input file size (up to 100 MB) limitation of the web server is overcome through the offering of the GUI limited, CLI version. Despite these limitations, DiMA is a significant step forward toward the study of viral sequence diversity dynamics.

Key Points

- DiMA is a novel, big data-ready tool designed to facilitate the dissection of sequence diversity dynamics for viruses.
- DiMA provides a quantitative overview of sequence (DNA/RNA/protein) diversity using Shannon's entropy, corrected for size bias, with a user-defined *k-mer* sliding window and options for metadata enrichment, dissecting each *k-mer* position into various diversity motifs for dynamics analyses.
- DiMA enables robust comparative analyses across viral species, supporting the design of more effective diagnostic, prophylactic, and therapeutic interventions.
- DiMA can potentially be used for non-viral pathogens, such as bacteria and fungi, among others.

Acknowledgement

We gratefully acknowledge all data contributors, i.e., the authors and their originating laboratories responsible for obtaining the specimens, and their submitting laboratories for generating the genetic sequence and metadata and sharing via the GISAID Initiative and HIV LANL, on which this research is based. We thank all those who used and/or evaluated DiMA and/or its earlier iterations, directly or indirectly, during the research and development phase (https://dima.readthedocs.io/en/latest/#acknowledgement).

Supplementary data

Supplementary data are available at *Briefings in Bioinformatics* online.

Conflict of interest: None declared.

Funding

The computational resources and services used in this work were provided by Perdana University School of Data Sciences,

Malaysia and Bezmialem Vakif University, Turkey. AMK was supported by University of Doha for Science and Technology, Qatar, Perdana University, Malaysia, Bezmialem Vakif University, Turkey, and The Scientific and Technological Research Council of Turkey (TÜBİTAK). This publication/paper has been produced benefiting from the 2232 International Fellowship for Outstanding Researchers Program of TÜBİTAK (Project No: 118C314). However, the entire responsibility of the publication/paper belongs to the owner of the publication/paper. The financial support received from TÜBİTAK does not mean that the content of the publication is approved in a scientific sense by TÜBİTAK.

Data and materials availability

DiMA web server is implemented in Rust and Python version 3.10. and is accessible through https://dima.bezmialem.edu.tr. It is also freely available as a standalone command-line interface (CLI) client tool at https://pypi.org/project/dima-cli and the source code is accessible on the GitHub repository at https://github.com/PU-SDS/DiMA. DiMA documentation is available at: https://dima.readthedocs.io/en/latest/. The findings of this study are based on data associated with 2,583,449 sequences available on GISAID (as of March 18th, 2022), and accessible at http://doi.org/10.55876/gis8.230207sb, and 24,985 sequences available on HIV LANL, accessible at https://www.hiv.lanl.gov/content/sequence/NEWALIGN/align.html.

References

1. Forni D, Cagliani R, Clerici M, *et al.* Disease-causing human viruses: novelty and legacy. *Trends Microbiol* 2022;**30**:1232–42. https://doi.org/10.1016/j.tim.2022.07.002.
2. Neumann G, Kawaoka Y. Which virus will cause the next pandemic? *Viruses* 2023;**15**:199–207. https://doi.org/10.3390/v15010199.
3. Chong LC, Khan AM. Historical milestone in 42 years of viral sequencing-impetus for a community-driven sequencing of global priority pathogens. *Front Microbiol* 2022;**13**:1020148. https://doi.org/10.3389/fmicb.2022.1020148.
4. Koyama T, Platt D, Parida L. Variant analysis of SARS-CoV-2 genomes. *Bull World Health Organ* 2020;**98**:495–504. https://doi.org/10.2471/BLT.20.253591.
5. Lau KA, Horan K, Gonçalves da Silva A, *et al.* Proficiency testing for SARS-CoV-2 whole genome sequencing. *Pathology* 2022;**54**:615–22. https://doi.org/10.1016/j.pathol.2022.04.002.
6. Walker BD, Goulder PJ, AIDS. Escape from the immune system. *Nature* 2000;**407**:313–4. https://doi.org/10.1038/35030283.
7. Domingo E, Sheldon J, Perales C. Viral quasispecies evolution. *Microbiol Mol Biol Rev* 2012;**76**:159–216. https://doi.org/10.1128/MMBR.05023-11.
8. Peck KM, Lauring AS. Complexities of viral mutation rates. *J Virol* 2018;**92**:1–8. https://doi.org/10.1128/JVI.01031-17.
9. Domingo E, Perales C. Viral quasispecies. *PLoS Genet* 2019;**15**:e1008271. https://doi.org/10.1371/journal.pgen.1008271.
10. Chong LC, Lim WL, Ban KHK, *et al.* An alignment-independent approach for the study of viral sequence diversity at any given rank of taxonomy lineage. *Biology (Basel)* 2021;**10**:1–15. https://doi.org/10.3390/biology10090853.
11. Zhang Q, Jun SR, Leuze M, *et al.* Viral Phylogenomics using an alignment-free method: a three-step approach to determine optimal length of k-mer. *Sci Rep* 2017;**7**:40712. https://doi.org/10.1038/srep40712.
12. Pearson WR. An introduction to sequence similarity ("homology") searching. *Curr Protoc Bioinformatics* 2013;**Chapter 3**:3.1.1–8. https://doi.org/10.1002/0471250953.bi0301s42.
13. Liu D, Shi W, Shi Y, *et al.* Origin and diversity of novel avian influenza a H7N9 viruses causing human infection: phylogenetic, structural, and coalescent analyses. *Lancet* 2013;**381**:1926–32. https://doi.org/10.1016/S0140-6736(13)60938-1.
14. Olsen LR, Kudahl UJ, Simon C, *et al.* BlockLogo: visualization of peptide and sequence motif conservation. *J Immunol Methods* 2013;**400-401**:37–44. https://doi.org/10.1016/j.jim.2013.08.014.
15. Villamor DEV, Ho T, Al Rwahnih M, *et al.* High throughput sequencing for plant virus detection and discovery. *Phytopathology* 2019;**109**:716–25. https://doi.org/10.1094/PHYTO-07-18-0257-RVW.
16. Poirier EZ, Vignuzzi M. Virus population dynamics during infection. *Curr Opin Virol* 2017;**23**:82–7. https://doi.org/10.1016/j.coviro.2017.03.013.
17. Lauring AS. Within-host viral diversity: a window into viral evolution. *Annu Rev Virol* 2020;**7**:63–81. https://doi.org/10.1146/annurev-virology-010320-061642.
18. Pybus OG, Rambaut A. Evolutionary analysis of the dynamics of viral infectious disease. *Nat Rev Genet* 2009;**10**:540–50. https://doi.org/10.1038/nrg2583.
19. Hu Y, Tan PT, Tan TW, *et al.* Dissecting the dynamics of HIV-1 protein sequence diversity. *PloS One* 2013;**8**:e59994. https://doi.org/10.1371/journal.pone.0059994.
20. Abd Raman HS, Tan S, August JT, *et al.* Dynamics of influenza a (H5N1) virus protein sequence diversity. *PeerJ* 2020;**7**:e7954. https://doi.org/10.7717/peerj.7954.
21. Shannon CE. A mathematical theory of communication. *Bell Syst Tech J* 1948;**27**:379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.
22. Schneider TD, Stormo GD, Gold L, *et al.* Information content of binding sites on nucleotide sequences. *J Mol Biol* 1986;**188**:415–31. https://doi.org/10.1016/0022-2836(86)90165-8.
23. Coordinators NR. Database resources of the National Center for biotechnology information. *Nucleic Acids Res* 2018;**46**:D8–13. https://doi.org/10.1093/nar/gkx1095.
24. Pickett BE, Sadat EL, Zhang Y, *et al.* ViPR: an open bioinformatics database and analysis resource for virology research. *Nucleic Acids Res* 2012;**40**:D593–8. https://doi.org/10.1093/nar/gkr859.
25. Zhang Y, Aevermann BD, Anderson TK, *et al.* Influenza research database: an integrated bioinformatics resource for influenza virus research. *Nucleic Acids Res* 2017;**45**:D466–74. https://doi.org/10.1093/nar/gkw857.
26. Schnoes AM, Brown SD, Dodevski I, *et al.* Annotation error in public databases: Misannotation of molecular function in enzyme superfamilies. *PLoS Comput Biol* 2009;**5**:e1000605. https://doi.org/10.1371/journal.pcbi.1000605.
27. Chen Q, Zobel J, Verspoor K. Benchmarks for measurement of duplicate detection methods in nucleotide databases. *Database (Oxford)* 2017;**2023**:1–17. https://doi.org/10.1093/database/baw164.
28. Subramaniy V, Pandian SC. A complete survey of duplicate record detection using data mining techniques. *Information Technology Journal* 2012;**11**:941–5. https://doi.org/10.3923/itj.2012.941.945.
29. B. National Research Council Board on. In: Pool R, Esnayra J (eds). *Bioinformatics: Converting Data to Knowledge: Workshop Summary*. National Academies Press (US). Copyright © 2000. Washington (DC): National Academy of Sciences, 2000.
30. Khan AM, Miotto O, Nascimento EJM, *et al.* Conservation and variability of dengue virus proteins: implications for vaccine

design. *PLoS Negl Trop Dis* 2008;**2**:e272. https://doi.org/10.1371/journal.pntd.0000272.
31. Thompson JD, Linard B, Lecompte O, *et al.* A comprehensive benchmark study of multiple sequence alignment methods: current challenges and future perspectives. *PloS One* 2011;**6**:e18093. https://doi.org/10.1371/journal.pone.0018093.
32. Nuin PA, Wang Z, Tillier ER. The accuracy of several multiple sequence alignment programs for proteins. *BMC Bioinformatics* 2006;**7**:471. https://doi.org/10.1186/1471-2105-7-471.
33. Hosseininasab A, Hoeve W-JV. Exact multiple sequence alignment by synchronized decision diagrams. *INFORMS Journal on Computing* 2021;**33**:721–38.
34. Edgar RC, Batzoglou S. Multiple sequence alignment. *Curr Opin Struct Biol* 2006;**16**:368–73. https://doi.org/10.1016/j.sbi.2006.04.004.
35. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 2004;**32**:1792–7. https://doi.org/10.1093/nar/gkh340.
36. Sievers F, Wilm A, Dineen D, *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal omega. *Mol Syst Biol* 2011;**7**:539. https://doi.org/10.1038/msb.2011.75.
37. Katoh K, Misawa K, Kuma K, *et al.* MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* 2002;**30**:3059–66. https://doi.org/10.1093/nar/gkf436.
38. Smirnov V, Warnow T. MAGUS: multiple sequence alignment using graph clUStering. *Bioinformatics* 2021;**37**:1666–72. https://doi.org/10.1093/bioinformatics/btaa992.
39. Long H, Li M, Fu H. Determination of optimal parameters of MAFFT program based on BAliBASE3.0 database. *Springerplus* 2016;**5**:736. https://doi.org/10.1186/s40064-016-2526-5.
40. Sievers F, Higgins DG. QuanTest2: benchmarking multiple sequence alignments using secondary structure prediction. *Bioinformatics* 2020;**36**:90–5. https://doi.org/10.1093/bioinformatics/btz552.
41. Moshiri N. ViralMSA: massively scalable reference-guided multiple sequence alignment of viral genomes. *Bioinformatics* 2021;**37**:714–6. https://doi.org/10.1093/bioinformatics/btaa743.
42. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 1994;**22**:4673–80. https://doi.org/10.1093/nar/22.22.4673.
43. Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* 2000;**17**:540–52. https://doi.org/10.1093/oxfordjournals.molbev.a026334.
44. Penn O, Privman E, Ashkenazy H, *et al.* GUIDANCE: a web server for assessing alignment confidence scores. *Nucleic Acids Res* 2010;**38**:W23–8. https://doi.org/10.1093/nar/gkq443.
45. Liu K, Warnow T. Large-scale multiple sequence alignment and tree estimation using SATe. *Methods Mol Biol* 2014;**1079**:219–44. https://doi.org/10.1007/978-1-62703-646-7_15.
46. Landan G, Graur D. Local reliability measures from sets of co-optimal multiple sequence alignments. *Pac Symp Biocomput* 2008;**13**:15–24.
47. Anderson CL, Strope CL, Moriyama EN. SuiteMSA: visual tools for multiple sequence alignment comparison and molecular sequence simulation. *BMC Bioinformatics* 2011;**12**:184. https://doi.org/10.1186/1471-2105-12-184.
48. Olsen LR, Zhang GL, Keskin DB, *et al.* Conservation analysis of dengue virus T-cell epitope-based vaccine candidates using peptide block entropy. *Front Immunol* 2011;**2**:69. https://doi.org/10.3389/fimmu.2011.00069.
49. Marcelino AM, Gierasch LM. Roles of beta-turns in protein folding: from peptide models to protein engineering. *Biopolymers* 2008;**89**:380–91. https://doi.org/10.1002/bip.20960.
50. Trolle T, McMurtrey CP, Sidney J, *et al.* The length distribution of class I-restricted T cell epitopes is determined by both peptide supply and MHC allele-specific binding preference. *J Immunol* 2016;**196**:1480–7. https://doi.org/10.4049/jimmunol.1501721.
51. Chang ST, Ghosh D, Kirschner DE, *et al.* Peptide length-based prediction of peptide-MHC class II binding. *Bioinformatics* 2006;**22**:2761–7. https://doi.org/10.1093/bioinformatics/btl479.
52. Yang OO. Candidate vaccine sequences to represent intra- and inter-clade HIV-1 variation. *PloS One* 2009;**4**:e7388. https://doi.org/10.1371/journal.pone.0007388.
53. Paninski L. Estimation of entropy and mutual information. *Neural Comput* 2003;**15**:1191–253. https://doi.org/10.1162/089976603321780272.
54. Shruti K, *et al.* GISAID's role in pandemic response. *China CDC Weekly* 2021;**3**:1049–51. https://doi.org/10.46234/ccdcw2021.255.
55. Apetrei CH. In: Rambaut B, Wolinsky A, Brister S. *et al.* (eds.), *HIV Sequence Compendium* 2021LA-UR-23-22840, p. 2021. NM: Theoretical Biology and Biophysics Group, Los Alamos National Laboratory.
56. Kuiken C, Korber B, Shafer RW. HIV sequence databases. *AIDS Rev* 2003;**5**:52–61.
57. Du L, *et al.* MERS-CoV spike protein: a key target for antivirals. *Expert Opin Ther Targets* 2017;**21**:131–43. https://doi.org/10.1080/14728222.2017.1271415.
58. Frankel AD, Young JA. HIV-1: fifteen proteins and an RNA. *Annu Rev Biochem* 1998;**67**:1–25. https://doi.org/10.1146/annurev.biochem.67.1.1.
59. Hill M, Tachedjian G, Mak J. The packaging and maturation of the HIV-1 pol proteins. *Curr HIV Res* 2005;**3**:73–85. https://doi.org/10.2174/1570162052772942.
60. Fu L, Niu B, Zhu Z, *et al.* CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 2012;**28**:3150–2. https://doi.org/10.1093/bioinformatics/bts565.
61. Huang Y, Yang C, Xu XF, *et al.* Structural and functional properties of SARS-CoV-2 spike protein: potential antivirus drug development for COVID-19. *Acta Pharmacol Sin* 2020;**41**:1141–9. https://doi.org/10.1038/s41401-020-0485-4.
62. Zhang L, Mann M, Syed ZA, *et al.* Furin cleavage of the SARS-CoV-2 spike is modulated by O-glycosylation. *Proc Natl Acad Sci U S A* 2021;**118**:1–7. https://doi.org/10.1073/pnas.2109905118.
63. Gomari MM, Tarighi P, Choupani E, *et al.* Structural evolution of Delta lineage of SARS-CoV-2. *Int J Biol Macromol* 2023;**226**:1116–40. https://doi.org/10.1016/j.ijbiomac.2022.11.227.
64. De la Fuente IM, *et al.* Stability of SARS-CoV-2 spike antigens against mutations. *medRxiv* 2022;1–44. https://doi.org/10.1101/2022.10.13.22280980.
65. Stincarelli MA, Quagliata M, di Santo A, *et al.* SARS-CoV-2 inhibitory activity of a short peptide derived from internal fusion peptide of S2 subunit of spike glycoprotein. *Virus Res* 2023;**334**:199170. https://doi.org/10.1016/j.virusres.2023.199170.
66. Zhu Y, Yu D, Yan H, *et al.* Design of Potent Membrane Fusion Inhibitors against SARS-CoV-2, an emerging coronavirus with high Fusogenic activity. *J Virol* 2020;**94**:1–12. https://doi.org/10.1128/JVI.00635-20.
67. Mansky LM, Temin HM. Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the

fidelity of purified reverse transcriptase. *J Virol* 1995;**69**:5087–94. https://doi.org/10.1128/jvi.69.8.5087-5094.1995.
68. Plotkin JB, Kudla G. Synonymous but not the same: the causes and consequences of codon bias. *Nat Rev Genet* 2011;**12**:32–42. https://doi.org/10.1038/nrg2899.
69. Chamary JV, Hurst LD. Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. *Genome Biol* 2005;**6**:R75. https://doi.org/10.1186/gb-2005-6-9-r75.
70. Pijlman GP, Funk A, Kondratieva N, *et al.* A highly structured, nuclease-resistant, noncoding RNA produced by flaviviruses is required for pathogenicity. *Cell Host Microbe* 2008;**4**:579–91. https://doi.org/10.1016/j.chom.2008.10.007.
71. Delviks-Frankenberry KA, Nikolenko GN, Pathak VK. The "connection" between HIV drug resistance and RNase H. *Viruses* 2010;**2**:1476–503. https://doi.org/10.3390/v2071476.
72. Chiu TK, Davies DR. Structure and function of HIV-1 integrase. *Curr Top Med Chem* 2004;**4**:965–77. https://doi.org/10.2174/1568026043388547.
73. Payne RP, Kløverpris H, Sacha JB, *et al.* Efficacious early antiviral activity of HIV gag- and pol-specific HLA-B 2705-restricted CD8+ T cells. *J Virol* 2010;**84**:10543–57. https://doi.org/10.1128/JVI.00793-10.
74. Tan S, Sjaugi MF, Fong SC, *et al.* Avian influenza H7N9 virus adaptation to human hosts. *Viruses* 2021;**13**:871–90. https://doi.org/10.3390/v13050871.
75. Chong LC, Khan AM. Vaccine target discovery. In: Shoba Ranganathan, Michael Gribskov, Kenta Nakai, Christian Schönbach (eds). *Encyclopedia of Bioinformatics and Computational Biology*, Vol. **3**, pp. 241–51. NX Amsterdam, The Netherlands: Elsevier, 2019.
76. Garcia-Boronat M, Diez-Rivero CM, Reinherz EL, *et al.* PVS: a web server for protein sequence variability analysis tuned to facilitate conserved epitope discovery. *Nucleic Acids Res* 2008;**36**:W35–41. https://doi.org/10.1093/nar/gkn211.
77. Eigen M, Schuster P. The hypercycle. A principle of natural self-organization. Part a: emergence of the hypercycle. *Naturwissenschaften* 1977;**64**:541–65. https://doi.org/10.1007/BF00450633.
78. Dennehy JJ. Evolutionary ecology of virus emergence. *Ann N Y Acad Sci* 2017;**1389**:124–46. https://doi.org/10.1111/nyas.13304.
79. Saakian DB, Munoz E, Hu CK, *et al.* Quasispecies theory for multiple-peak fitness landscapes. *Phys Rev E Stat Nonlin Soft Matter Phys* 2006;**73**:041913. https://doi.org/10.1103/PhysRevE.73.041913.
80. Koo QY, Khan AM, Jung KO, *et al.* Conservation and variability of West Nile virus proteins. *PloS One* 2009;**4**:e5352. https://doi.org/10.1371/journal.pone.0005352.

Briefings in Bioinformatics, 2025, **26(1)**, bbae607
https://doi.org/10.1093/bib/bbae607
Problem Solving Protocol