BMC Bioinformatics

SOFTWARE Open Access

GRAMEP: an alignment-free method based on the maximum entropy principle for identifying SNPs

Matheus Henrique Pimenta-Zanon[1], André Yoshiaki Kashiwabara[1], André Luís Laforga Vanzela[2] and Fabricio Martins Lopes[1*]

*Correspondence:
fabricio@utfpr.edu.br

[1] Computer Science Department, Universidade Tecnológica Federal do Paraná (UTFPR), Alberto Carazzai, 1640, Cornélio Procópio, Paraná 86300-000, Brazil
[2] Laboratory of Cytogenetics and Plant Diversity, Department of General Biology, Universidade Estadual de Londrina (UEL), Rodovia Celso Garcia Cid, PR-445, Km 380, Londrina, Paraná 86057-970, Brazil

Abstract

Background: Advances in high throughput sequencing technologies provide a huge number of genomes to be analyzed. Thus, computational methods play a crucial role in analyzing and extracting knowledge from the data generated. Investigating genomic mutations is critical because of their impact on chromosomal evolution, genetic disorders, and diseases. It is common to adopt aligning sequences for analyzing genomic variations. However, this approach can be computationally expensive and restrictive in scenarios with large datasets.

Results: We present a novel method for identifying single nucleotide polymorphisms (SNPs) in DNA sequences from assembled genomes. This study proposes GRAMEP, an alignment-free approach that adopts the principle of maximum entropy to discover the most informative k-mers specific to a genome or set of sequences under investigation. The informative k-mers enable the detection of variant-specific mutations in comparison to a reference genome or other set of sequences. In addition, our method offers the possibility of classifying novel sequences with no need for organism-specific information. GRAMEP demonstrated high accuracy in both in silico simulations and analyses of viral genomes, including Dengue, HIV, and SARS-CoV-2. Our approach maintained accurate SARS-CoV-2 variant identification while demonstrating a lower computational cost compared to methods with the same purpose.

Conclusions: GRAMEP is an open and user-friendly software based on maximum entropy that provides an efficient alignment-free approach to identifying and classifying unique genomic subsequences and SNPs with high accuracy, offering advantages over comparative methods. The instructions for use, applicability, and usability of GRAMEP are open access at https://github.com/omatheuspimenta/GRAMEP.

Keywords: Alignment-free methods, SNP mutation identification, Classification of biological sequences, Principle of maximum entropy, Genomic data analysis

Background

The analysis of genomic sequences has been extensively studied to understand species' diversity and evolution [1]. With the advancement of sequencing technology, an ever-increasing amount of genomics data is being generated, opening up new perspectives for research on genetic variants in various organisms [2, 3].

The genome of a species encodes the instructions required for the production of thousands of proteins and RNA molecules [4]. This information is embedded within the DNA sequence, acting as a complex code. Mutations in nucleotides, the DNA building blocks, can be linked to typos in this code, altering DNA sequences and, consequently, the sequence of RNA and synthesized proteins. These variations can affect the organism's phenotype or its observable characteristics. Mutations that affect only a single base pair are known as single nucleotide variants (SNVs) or single nucleotide polymorphisms (SNPs). SNPs are akin to minute variations in the genetic code and can have varying effects on the organism. Some SNPs may be silent, causing no significant changes. Others may have a mild impact, while some can lead to drastic alterations in the phenotype, such as genetic diseases [5, 6].

SNVs can lead to the simultaneous emergence of different phenotypes and result in intraspecific variations. In viruses, such as SARS-CoV-2, several mutations occurred during the 2019 pandemic. Depending on the specific mutations, these variations directly affected the transmission rate, mortality, and infectivity of the virus [7–11]. The study of genomic variation is critical for diagnosing, preventing, and treating diseases. When applied to the study of viral diseases, genomic variation analysis is relevant for epidemiological purposes and for controlling viral spread. Additionally, a systematic understanding of the evolution and taxonomy of various species, including viruses, relies on genomic variation research [12]. Furthermore, organisms like the Dengue virus (DENV) exhibit weak error correction mechanisms, leading to a high mutation rate and significant diversity within their variant populations [13]. The presence of SNPs in viral genomes can have various consequences, such as changes in resistance to antiviral drugs (e.g., Influenza virus and HCV) [13] or enabling immune system evasion (e.g., human immunodeficiency virus type 1 (HIV-1), however, a high mutation rate can also lead to the extinction of certain variants [14, 15].

The precise identification of SNVs in a large number of sequences can be computationally challenging, considering classical alignment methods. The complexity can be polynomial regarding the number and size of the analyzed sequences [16]. Alignment-based approaches are widely used in biological sequence analysis like BLAST [17] to identify regions of similarity between sequences. However, they face significant challenges when dealing with multi-genomic scale data because of computational resources. For instance, considering the existence of gaps, the number of possible alignments for two sequences with 100 base pairs can reach an order of approximately 10^{60}. Finding the best alignment based on a scoring system that provides a value for match, mismatch, and gap penalties is a well-known problem that dynamic programming algorithms, such as Smith-Waterman [18] and Needleman-Wunsch [19], can solve in quadratic time complexity. However, the unfeasibility of these alignment-based approaches for large-scale analysis due to the time complexity [1, 20] underscores further research for alignment-free methods.

It is well known that the high mutation rates are characteristic of viruses and represent a challenge for traditional sequence analysis methods, which rely on aligning sequences to identify similarities [1]. Alignment-free approaches offer an essential alternative for analyzing these highly mutable genomes, mainly because they do not rely on finding regions with high identity [20]. Researchers have adapted alignment-free methods for comparative analyses [1, 20] with many distinct datasets. Furthermore, alignment-free methods are often scalable and computationally efficient, making them a suitable alternative for handling large datasets [21]. Unlike alignment-based approaches, they do not assume collinearity between sequences-meaning they do not require a one-to-one correspondence of residues along the sequence-making them ideal for capturing complex genetic patterns in viruses that undergo frequent recombination [22], horizontal gene transfer [23], duplications [24], and gene losses [25].

In this context, heuristics and alignment-free methods emerge as a viable alternative, providing near-optimal solutions in feasible time and allowing scalability in the analysis of large volumes of data [1, 20]. They offer a more efficient approach for the analysis of complete genomes with lower computational complexity, which is achieved by considering mathematical and computational concepts, such as calculus, information theory, statistics, physics, and linear algebra being more adaptable to different types of sequences, especially those with high mutation rates, such as those found in viruses [20, 26, 27]. Alignment-free methods have applications in broad areas within computational biology, such as a study of viral diversity, identification of genes and regulatory regions, analysis of phylogeny and molecular evolution, and detection of genetic variants associated with diseases [2, 28–30].

Genomic sequence analysis can be conducted using two primary alignment-free approaches: word-based and information theory-based methodologies [1]. Word-based methods focus on analyzing k-mers, short subsequences of fixed-length k extracted from the sequences under study. These k-mers are treated as distinct units, and their frequency and distribution within genomic sequences are used to generate profiles and sequence-specific patterns. Conversely, information theory-based methods employ mathematical and statistical tools from information theory to quantify the information content across genomic sequences. This approach enables the identification of complex, contextual patterns within sequences, aiding in the detection of functional and evolutionary relationships among different genomic regions [31–33].

Several methods adopt the spectrum generated by k-mer occurrence frequency [34]. This technique is based on the premise that different species exhibit unique k-mer patterns, facilitating the identification and classification of sequences according to their k-mer frequency [35]. Alternatively, recent approaches have explored the use of genetic algorithms to identify deterministic subsequences of interest [36].

Incorporating deterministic subsequences as features offers the potential to reduce dimensionality when training machine learning models [36–40]. This approach leverages these subsequences as primary features to characterize input sequences. Additionally, machine learning models trained on such features can develop the ability to recognize complex and subtle patterns, ultimately enabling the classification of novel sequences into distinct species.

Thanos et al. [41] introduces a genomic analysis approach based on information theory, specifically utilizing Shannon entropy applied to non-overlapping blocks of subsequences. This framework adopts Shannon entropy to detect repetitive regions, which often show significant entropy fluctuations compared to other genomic segments. This feature enables the identification of repetitive elements, such as transposons and duplicated sequences, which play critical roles in evolution and genetic regulation.

The GENIES method [32, 33] offers a distinct approach to mutation detection in viral genomes, with a particular emphasis on SARS-CoV-2. It leverages the entropy spectrum, a graphical representation of the relationship between block index and k-mer entropy. Unlike the method proposed by Thanos, GENIES utilizes overlapping subsequences with a calculated step size to evaluate entropy for each k-mer within the genome. Mutations are identified by comparing the entropy ratios of corresponding k-mer blocks between the reference and variant sequences; deviations from a ratio of 1 indicate the presence of mutations within specific k-mers. While this technique is effective for pinpointing specific mutations, a key limitation is its requirement for equal-length reference and variant sequences, rendering GENIES inapplicable to studies involving viral genomes of varying lengths.

MEME (discriminative mode) [42, 43] employs a statistical sequence model based on user-defined parameters, if available, for expected site count and width. It allows the incorporation of prior sequence information and offers two search options: zero or one occurrence per sequence (ZOOPS) and one occurrence per sequence (OOPS). STREME [44] utilizes a generalized suffix tree and evaluates motifs using a unilateral statistical test of enrichment within a specified sequence set compared to a control set.

Lebatteux et al. [40] introduced the CASTOR-KRFE method to identify discriminative subsequences within viral genomes for classification. This method employs a feature selection to identify the most informative k-mers, subsequently using them as features for the classification of clades, species, or sub-variants. An evolution of CASTOR-KRFE, the KEVOLVE method [38, 39], retains this structure but replaces the feature selector with a genetic algorithm for k-mer selection. The latest refinement, KANALYZER [45], leverages these discriminative k-mers to pinpoint specific gene regions containing mutations within the analyzed genomes. It accomplishes this by aligning only in regions where discrepancies occur between the matches of discriminative k-mers and those of the reference sequence. This approach enables the identification of single nucleotide polymorphisms (SNPs) and insertions/deletions (indels), which may impact amino acid changes.

In light of this scenario, we propose a new method leveraging the principle of maximum entropy to pinpoint the most informative deterministic regions unique to each species, clade, or sub-variant within an organism using the k-mers approach. These regions subsequently serve as features for training a robust biological sequence classification model. The proposed approach enables the identification of SNPs within the analyzed organism's genome, providing scalability, allowing vertical expansion through increased hardware capacity or horizontal distribution across multiple processing nodes, ensuring its feasibility and effectiveness for extensive genomic analysis. We show that our approach can analyze the genomes of RNA viruses like SARS-COV, DENV, and HIV [46], which are known for their high recombination rates and reassignment rates.

GRAMEP (Genome Variation Analysis from the Maximum Entropy) is an open software that encompasses its key functionalities in genome analysis.

Implementation

GRAMEP - Genome vaRiation Analysis from the Maximum Entropy, is written mostly in Python with some functions written in Rust. Figure 1 gives an overview of the method. The program contains three main functions: the `get-mutations` function, which identifies the mutations of the input variants relative to the reference; and the `classify` and `predict` functions, which are performed to train the classification model and to test the induced model on unknown sequences, respectively..

Input data

The GRAMEP input data requires two fasta files: one containing a single reference sequence and another containing the sequences of the variant under analysis. Two primary parameters are used to identify exclusive subsequences: 'word' and 'step'.

The 'word' parameter determines the k-mer size, defining the length of the subsequences to be considered during the analysis. On the other hand, the 'step' parameter sets the offset for the sliding window, specifying how far the window moves as it traverses the sequences. These parameters directly influence the size and overlap of the extracted subsequences, allowing flexibility in the selection process according to the analysis needs.

For instance, Fig. 1-Panel A, illustrates the extraction process of 10-length k-mers with a step size of 1, applied to the variant sequences and a sequence of reference. This configuration generates overlapping subsequences of length 10 across the sequence, providing a detailed view of the variation.

In addition to the variant and reference sequences, the k-mer size ('word'), and the step ('step'), it is necessary to specify the location where the analysis results will be saved using the 'save-path' parameter. Other optional parameters in the `get-mutations` function can be set to adjust the analysis as needed. These optional parameters include the annotation file (.gff), allowing mutations to be associated with specific genomic regions such as genes or exons, thereby facilitating the biological interpretation of the results displayed in reports at the end of the execution; the maximum number of allowed SNP mutations per k-mer, which sets the limit of single nucleotide

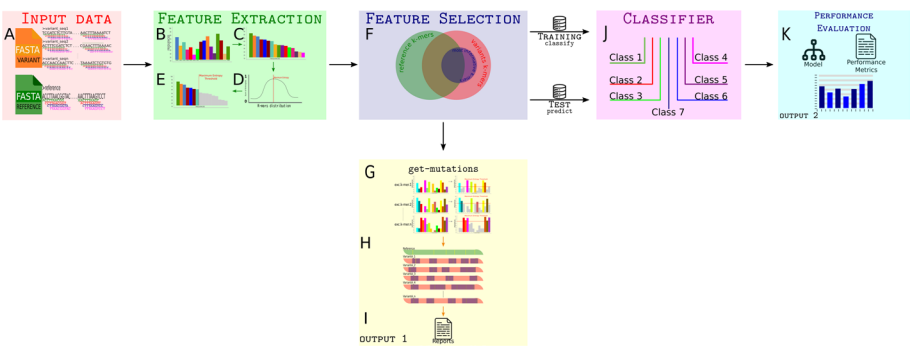

Fig. 1 Overview of the GRAMEP method

polymorphisms (SNPs) permitted for each k-mer; the type of sequence being analyzed, enabling specification of whether the sequences are DNA, RNA, or other types, which influences the preprocessing of sequences during loading. The loading process removes sequences containing characters incompatible with the specified sequence type.

Furthermore, there is the option to generate a complete mutation report, and finally, the chunk size, a parameter that defines the number of sequences to be processed simultaneously in each block (chunk), optimizing parallel analysis. This value depends on the available memory, balancing the efficient use of computational resources with the system's capacity. These additional configurations increase the flexibility and precision of the analysis, allowing customization of parameters according to the objectives and technical limitations of the computational environment.

Feature extraction

In the feature extraction process, extracting relevant details from the sequences is an essential step, as the efficiency and accuracy of predictive models depend heavily on the selection of appropriate features during training [47].

While loading sequences, each sequence is divided into 100 non-overlapping regions to obtain the occurrence frequency of k-mers in each of these regions. In addition to recording the occurrence frequency of each k-mer, this information is stored in a hashmap, which also includes the occurrence frequency of the k-mer in each of the 100 regions. In this way, the hashmap is structured so that the keys correspond to the k-mers, while the values represent both the total occurrence frequency of the k-mer and the occurrence frequency in each specific region.

After loading the occurrence frequencies of the k-mers and the regions, the next step involves calculating the maximum entropy. This calculation is essential for identifying which k-mers and regions are most informative, thus enabling a data-driven selection of the k-mers and regions that hold greater relevance for the analysis at hand.

The entropy permeates diverse scientific domains, encompassing both microscopic and macroscopic scales. In statistical mechanics [48], it reflects the disorder within a system. Similarly, thermodynamics leverages entropy to describe energy exchange and equilibrium states [49]. While Clausius and Kelvin established the second law of thermodynamics, the underlying formula for entropy remained elusive until Boltzmann and Gibbs' groundbreaking work [50]. Their microscopic approach defined the now-renowned Boltzmann-Gibbs entropy:

$$S = -k \sum_{i=1}^{W} p_i \log p_i \tag{1}$$

where k is the Boltzmann constant and p_i represents the probabilities of the W possible microscopic states $\sum_{i=1}^{W} p_i = 1$.

Meanwhile, information theory employs a distinct concept of entropy introduced by Shannon [51]. Here, entropy $H(\Phi)$ quantifies the uncertainty associated with a random variable Φ, formally defined as:

$$H(\Phi) = -\sum_{x \in \mathcal{X}} p(x) \log p(x) \tag{2}$$

where x represents each element in the alphabet of $\mathcal{X}$ and $p(x)$ denotes its probability. Notably, entropy solely depends on the probabilities, not the specific values, of Φ.

Jaynes [52] bridged these realms by connecting the thermodynamic entropy of Boltzmann and Gibbs with Shannon's information-theoretic entropy. The principle of maximum entropy posits that, given available data, one should choose the probability distribution with the highest possible entropy. This essentially implies utilizing only the observed data without incorporating any prior assumptions.

Consider a discrete distribution with n events and observed frequencies $h_1, h_2, \ldots, h_n$. Let $p_i = \frac{h_i}{N}$ (where N is the total number of samples) represent the probability of the i-th outcome. For a distribution with two classes, X and Y, their respective entropies are:

$$H(X) = -\sum_{i=1}^{s} \frac{p_i}{P_X} \log\left(\frac{p_i}{P_X}\right), \tag{3}$$

$$H(Y) = -\sum_{i=s+1}^{n} \frac{p_i}{P_Y} \log\left(\frac{p_i}{P_Y}\right), \tag{4}$$

where P_X and P_Y are the respective probabilities of belonging to class X or Y ($P_X + P_Y = 1$). By maximizing the sum of these class entropies $H(X) + H(Y)$, we achieve the maximum entropy (*ME*):

$$ME = \underset{s=1,2,\ldots,n}{\arg\max}\{H(X) + H(Y)\}. \tag{5}$$

This maximization essentially identifies the point in the distribution where the classes are most separable, signifying maximum uncertainty between them [53]. Leveraging maximum entropy to distinguish class distributions [54] allows us to focus solely on data with high information content, effectively filtering out noise and bias. This approach proves particularly valuable in high-dimensional problems, where it can significantly reduce dimensionality [31].

In this step of the GRAMEP, the k-mer frequencies are sorted in descending order, generating a frequency histogram; Panel C of Fig. 1 illustrates this step. Subsequently, occurrence probabilities are calculated using the maximum entropy principle.

First, the probabilities of each class X and Y, denoted as P_X and P_Y, are estimated. Next, the histogram is iterated from $i = 1$ to s to estimate P_X, and from $s + 1$ to n to estimate P_Y, iteratively for each value of $s = 1, 2, \ldots, n$. Using the estimates of P_X and P_Y, the entropies $H(X)$, Eq. 3, and $H(Y)$, Eq. 4, can be calculated to construct the distribution $H(X) + H(Y)$ and determine the maximum entropy *ME*, Eq. 5. This maximum entropy provides the threshold used to identify the variant's most informative and non-informative k-mers, excluding potential biases and noise present in the sequences.

Following this, an automatic threshold for "informative" subsequences is established. Panels D and E of Fig. 1 graphically illustrate the obtaining of the maximum

entropy value and also the selection of the most informative k-mers from the histogram distribution. K-mers surpassing this threshold are then considered informative and retained to constitute the variant-specific set of informative subsequences.

The entropy calculation is dependent on the data used as input, which can influence the choice of the most informative k-mers in each class. The maximum entropy principle solely considers the probability of event occurrences. This reliance on probability precludes intermediate values from influencing the maximum entropy calculation. Therefore, utilizing multiple sequences per variant is recommended to determine the optimal cutoff point between these classes.

Feature selection

Selecting the informative features is essential for achieving optimal classifier performance, as noisy or irrelevant data can greatly impact the outcomes [55].

Following the identification of informative k-mers specific to the variant, this step accomplishes this by subtracting the set of informative k-mers belonging to the variant from the comprehensive set of k-mers present in the organism's reference sequence, Panel F of Fig. 1 illustrates through Venn diagrams representing the set of k-mers from the reference sequence in green, the set of k-mers from the variant in red and the subset of the most informative k-mers for the variant in blue. This subtraction process effectively filters out k-mers shared between the variant and the reference, ensuring the resulting set solely comprises unique and informative subsequences characteristic of the analyzed variant. The most informative k-mers unique to the variant can be utilized as features for training and validating the classification model, as demonstrated in step J of Fig. 1. Additionally, they can be used to identify the mutations present in the variant, following the workflow outlined in panel G of Fig. 1.

Algorithm 1 provides a summary of how the most informative exclusive k-mers are obtained for each variant under analysis.

Algorithm 1 Obtaining the most informative exclusive k-mers.

Require: Reference.FASTA, Variant.FASTA, word, step
Ensure: Most informative exclusive k-mers
1: **for** sequence in Variant.FASTA sequences **do**
2: Get the frequencies of the k-mers and regions;
3: **end for**
4: Sort the frequencies of occurrence of k-mers in descending order
5: Get the maximum entropy from the frequencies of occurrence
6: Apply the automatic cutoff (threshold) to the k-mers
7: Extract the k-mers from the reference
8: Get the most informative exclusive k-mers for the variant

Output 1: Obtaining mutations and reports

The next step involves identifying the most informative regions where each previously selected k-mer appears within the sequences. To accomplish this, each sequence is divided into 100 regions, and for each most informative exclusive k-mer of the

variant under analysis, a threshold is determined based on the k-mer's frequency in these regions. This procedure is performed similarly to the k-mer selection step, employing the principle of maximum entropy, as illustrated in panel G of Fig. 1.

The subsequent step focuses on detecting mutations and their corresponding positions within the reference genome. This task leverages the established set of exclusive and informative subsequences associated with each variant sequence and their positions. Only subsequences with potential mutation sites within the variant sequence are selected for further analysis. This selection is achieved by considering the intersection of the variant's exclusive subsequences, effectively filtering out non-mutated regions.

The selection of regions relative to the reference sequence where the mutation search will take place is performed by considering the possible variations in size and positions between the reference sequence and the variant sequence. First, the variation between the length of the reference sequence and the variant is calculated. If the variation is significant, we calculate an extra adjustment. Then, we use this adjustment to expand the start and end points of each region, creating a wider range to account for the variation.

These bounds are always adjusted to ensure they do not exceed the length of the reference sequence. This approach allows only the most informative regions to be analyzed, avoiding the processing of the entire sequence. As a result, the strategy reduces the occurrence of false positives by focusing the analysis on areas most relevant (discriminant) to the comparison between the reference and the variant.

Subsequently, for each of these mutation-prone subsequences, the Levenshtein distance is adopted [56], panel H of Fig. 1 illustrates this step. By default, the Levenshtein distance is used to identify SNP mutations. To detect insertion and deletion mutations, the Myers algorithm [57] is employed. The selection of the algorithm depends on the user's objective: the Levenshtein distance generally enables faster identification of SNPs, whereas the Myers algorithm, while slightly slower, facilitates the detection of insertions and deletions. Each most informative unique k-mer and their position is represented in purple in the variant sequences, and SNPs are identified in yellow.

The following step culminates with a comprehensive analysis of mutation abundance and distribution within each variant sequence (Panels H and I in Fig. 1). This step starts with the computation of a frequency table, systematically tallying the occurrence and position of every identified mutation. Furthermore, if feasible, a graphical representation is generated, visually depicting the entirety of mutation locations relative to the reference genome. In the presence of an annotation file in .gff3 format, detailed mutation information can be generated for each sequence. This enriched data encompasses sequence identification, functional annotation, start and end positions of the mutation, type of mutated region, both variant and reference subsequences harboring the mutation, and specific nucleotide changes. Analyzing multiple variants of the same organism enables the identification of shared mutations, offering valuable insights into population-level trends and evolutionary trajectories.

Output 2: Classification and prediction

For classification, a minimum of two classes is required; therefore, if each variant of an organism represents a class, the exclusive and informative k-mers sets for each variant, obtained in the previous steps, are combined to form a single set containing

all unique k-mers for each variant on the reference genome, using the `classify` function. If the most informative exclusive k-mers are not available, they can be obtained via the 'get-kmers' parameter. Subsequently, these k-mers are adopted as features to train a classification and prediction model capable of effectively classifying novel sequences.

The next step involves training the classification and prediction model. This step utilizes each k-mer in the consolidated set as a feature, resulting in a feature matrix, as shown in Panel J in Fig. 1. Each row in this matrix represents a sequence, while each column corresponds to the occurrence frequency of a specific k-mer within that sequence. Following the acquisition of raw data (i.e., occurrence frequencies). The Min–Max rescaling is employed to improve usability in the model. Following data processing, a Random Forest algorithm with default parameters is implemented for training and classification tasks.

Performance evaluation

Performance evaluation is crucial for any statistical machine learning model before deployment in a real-world production environment. The efficiency of a machine learning system is assessed using various metrics. To validate the model, a "10-fold" cross-validation approach is utilized to validate the generated model, demonstrate its robustness, and mitigate overfitting during the training step. The dataset is divided into 10 parts. Nine parts are used for training and one for testing, $S = S_{\text{train}_k} \cup S_{\text{test}_k}, \quad S_{\text{train}_k} \cap S_{\text{test}_k} = \varnothing, \forall\, k \in 1, \ldots, 10$, where S is the complete dataset, S_{train_k} and S_{test_k} are the training and test subsets. The model's performance is evaluated using accuracy, precision, recall (sensitivity), F1 measure, Matthews Correlation Coefficient (MCC), and a confusion matrix.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{7}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{8}$$

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{9}$$

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{10}$$

where *TP*, *TN*, *FP*, and *FN* stand for true positive, true negative, false positive, and false negative, respectively.

After model validation, the trained model, the feature matrix, the confusion matrix, and the metrics obtained are obtained as output, as shown in Panel K in Fig. 1.

Results and discussion

GRAMEP: identification of SNPs mutations

To assess the GRAMEP for accurate SNPs identification, simulations were performed by considering datasets based on HIV and DENV genomes, motivated by other studies [45, 58]. Simulated mutations incorporated parameters reflecting real-world variation, including sequence length (based on average lengths in available datasets), virus-specific mutation rates, sequencing error rates, and genome size variation (based on standard deviation sequence length in available datasets). For each scenario, a fixed-size reference sequence was generated, followed by the creation of 1000 mutated sequences incorporating size variations, sequencing errors, and true mutations.

The adopted parameters for HIV-1 (S-HIV) and DENV (S-DENV) simulations were obtained from literature [13–15, 59]. The S-HIV dataset consisted of strings with a length of 8, 981, a mutation rate of 3×10^{-3}, and a variation rate of 0.0222. These parameters were based on available sequences in the Los Alamos Sequence Database (https://www.hiv.lanl.gov/). The S-DENV dataset, on the other hand, comprised strings with a length of 10, 553, a mutation rate of 1×10^{-3}, and a variation rate of 0.0205. Both simulations employed a uniform error rate of 5×10^{-4}, and the GRAMEP parameters used for simulations were obtained from empirical experiments presented in Supplementary Material - Appendix A.1. In this case, we used the "word" and "step" values of 15 and 1, respectively.

The simulations were performed considering the parameter configuration specified above. Notably, the method achieved a false positive rate (FPR) of zero in both simulations, demonstrating that all identified mutations were confirmed as true positives. Additionally, by providing the frequency of mutation occurrences, GRAMEP allows users to easily check if any mutations have a high occurrence frequency, thereby facilitating detailed analysis of relevant mutations, excluding point mutations arising from sequencing errors or individual sample peculiarities.

Regarding the true positive rate, the GRAMEP achieves performance exceeding 93% in both simulated scenarios. This result indicates a high capacity of the method to identify most of the present mutations, which is further supported by the low rate of false negatives. Furthermore, when evaluating classification metrics such as accuracy, Matthews correlation coefficient (MCC), and F1-score, the effectiveness of GRAMEP in accurately classifying mutations is evident, reinforcing its applicability for genomic variant analysis. Further evaluation metrics are provided in Table 1.

In order to assess GRAMEP in the identification of SNPs in real scenarios, it was adopted a dataset of 20 SARS-CoV-2 strains encompassing 579,053 genomic sequences. The dataset of the SARS-CoV-2 virus was extracted from NCBI. The 20 lineages with the

Table 1 Average and (standard deviation) obtained from performing 1000 simulations on *in silico* data

	TPR	FPR	TNR	FNR	ACC	MCC	F1
S-HIV	93.85 (7.53)	0.0 (0.0)	99.99 (0.0)	6.14 (7.53)	96.73 (2.43)	93.75 (4.53)	96.66 (4.22)
S-DENV	93.70 (7.55)	0.0 (0.0)	99.99 (0.0)	6.30 (7.55)	96.92 (3.76)	94.24 (6.91)	96.58 (4.21)

Adopted metrics: (TPR) True Positive Rate, (FPR) False Positive Rate, (TNR) True Negative Rate, (FNR) False Negative Rate, (ACC) accuracy, (MCC) Matthews Correlation Coefficient and F1-score

highest number of available sequences until January 2024 were selected. The reference genome used was also obtained from NCBI, corresponding to the Wuhan (identification NC_045512.2). Details were presented in Table 2.

For validation, GRAMEP was compared against data from COV2Var [60], a resource analyzing and annotating mutations in over 13 billion SARS-CoV-2 sequences sourced from GISAID. The "word" and "step" parameters were set to 15 and 1. The large number of sequences employed served to demonstrate the scalability of the proposed methodology. In addition, it demonstrates the methodology's scalability, allowing thousands of sequences to be analyzed simultaneously.

To assess the effectiveness of GRAMEP in identifying structural mutations in each variant, results were analyzed based on three different sequence cut-offs (thresholds): 90%, 95%, and 99% of sequences processed by both COV2Var and GRAMEP. Table 3 presents the obtained results, indicating the percentage of mutations identified by GRAMEP that are also present in COV2Var. It can be observed that as the cut-off increases, the concordance between mutations detected by GRAMEP and those cataloged by COV2Var for SARS-CoV-2 variants also increases.

When considering only mutations present in 99% of the analyzed sequences in each variant, the overlap between mutations recorded by COV2Var and those detected by GRAMEP approaches 100% across the evaluated variants, except for two variants. These results indicate that GRAMEP is effective in identifying recurrent structural mutations at high-frequency levels.

Table 2 SARS-CoV-2 adopted dataset to assess the `get-mutations` function

SARS-CoV-2	
Lineages	**Number of sequences**
AY.103	33,450
AY.25	12,961
AY.3	14,237
AY.44	19,658
B.1.526	12,725
B.1	12,478
B.1.1.7	65,875
BA.1.15	20,153
BA.1.18	13,342
BA.1.1	89,471
B.1.2	29,619
BA.1	15,182
BA.2.12.1	76,893
BA.2	49,668
BA.4.6	12,751
BA.5.1	14,064
BA.5.2.1	34,914
BA.5.2	17,994
BA.5.5	16,868
BQ.1.1	16,750
Total	579,053

Table 3 Mutations found for each variant after running GRAMEP on the SARS-CoV-2 dataset compared to the mutations found in 90%, 95%, and 99% of the sequences of the COV2Var database and GRAMEP results

Variant	GRAMEP (90%)	GRAMEP (95%)	GRAMEP (99%)
AY.3	94.28	93.10	100
AY.25	94.11	100	100
AY.44	93.75	95.65	100
AY.103	94.11	96.00	100
B.1	100	100	100
B.1.1.7	78.12	88.46	100
B.1.526	86.66	85.71	100
B.1.2	100	100	100
BA.1	89.74	96.55	100
BA.1.1	88.09	96.77	100
BA.1.15	88.88	97.14	100
BA.1.18	90.00	92.59	90.90
BA.2	87.93	91.11	100
BA.2.12.2	88.73	87.71	100
BA.4.6	85.52	88.40	84.21
BA.5.1	90.62	91.30	100
BA.5.2	91.30	92.72	100
BA.5.2.1	92.64	92.72	100
BA.5.5	89.55	89.47	100
BQ.1.1	90.66	96.15	100
Mean (SD)	90.73 (4.73)	93.58 (4.19)	98.75 (3.98)

The mutations identified by GRAMEP exhibited high concordance with those present in 99% of the sequences belonging to the COV2Var-analyzed variants (based on GISAID [61] data). The identified mutations represented, on average, more than 98% of all variants found in 99% of the sequences. Moreover, none of the identified SNP-type mutations were false positives, further reinforcing the ability to pinpoint mutations truly present in the majority of analyzed sequences. All outputs of the get-mutations function are detailed in the supplementary materials.

An additional potential application of GRAMEP involves identifying mutations common to specific organism variants. This capability stems from the fact that all identified mutations inherently belong to the examined variant. In this context, we explored shared mutations among the SARS-CoV-2 variants, using `get-intersections` function. For example, mutation C14408T was present in all analyzed variants, leading to a proline-to-leucine substitution at amino acid position 4715 [62]. A mutation at position A23403G within the Spike (S) protein region has been observed, leading to the substitution of aspartic acid (D) with glycine (G) at this residue. The remaining identified mutations were predominantly present in most analyzed sequences, suggesting distinctive variations between the variants. Additionally, GRAMEP offers the capability to generate detailed reports for each analyzed sequence, outlining the identified mutations, which facilitates interpretation and comparison of results.

Given its data-driven nature, GRAMEP is well-suited for application across diverse scenarios and organisms. This is because it solely relies on information extracted from

the analyzed sequences, eliminating the need for prior knowledge about the organisms under investigation.

GRAMEP: classification and prediction of sequences

The GRAMEP extends beyond mutation identification and offers the potential for biological sequence classification and prediction. This can be achieved by leveraging the most informative exclusive k-mers associated with each variant within an organism.

To demonstrate this application, a dataset of DENV genomes containing four dengue serotypes was obtained through BV-BRC [46], shown in Table 4. The reference genome adopted for the Dengue virus was extracted from NCBI (identification NC_001477.1).

To evaluate the performance of the GRAMEP in terms of classification, the same methodology of repeated K-fold cross-validation described in [38] was applied. In this process, 100 training datasets were created, each consisting of 250 sequences randomly selected from each serotype, totaling 1000 sequences in the training set. The sequences not selected for the training set comprise the test set, totaling 5051 sequences, ensuring that the model is evaluated on previously unused data. This approach allows for a robust analysis of the method's ability to accurately classify the sequences, ensuring variability in the samples and enhancing the generalization of the results.

The "word" parameter was set to 9, and the "step" parameter was assigned a value of 1, aligning with the values used in other methods analyzed to ensure consistency across the analysis. Following training and validation on dedicated datasets, we employed the trained models for prediction on independent test data.

To assess the performance of the classification models, the standard metrics, including precision, recall, F1-score, Matthews correlation coefficient (MCC), and accuracy, were adopted. For clarity and comprehensiveness, the confusion matrix and metrics for DENV are shown in Fig. 2.

GRAMEP offers a reduction in computational complexity compared to alignment methods. This advantage stems from the requirement of solely analyzing k-mer occurrence frequencies within the sequences. Consequently, our method facilitates the simultaneous analysis of large datasets, even on personal computers.

Comparing GRAMEP to existing state-of-the-art tools

To a broader comparison of the results of the proposed approach with similar methods of classification and identification of mutations from deterministic regions in the

Table 4 Dataset used to execute experiment to asses the GRAMEP `classify` function

Dengue Virus (DENV)	
Serotype	**Number of sequences**
Type 1	2571
Type 2	1756
Type 3	1272
Type 4	462
Total	6061

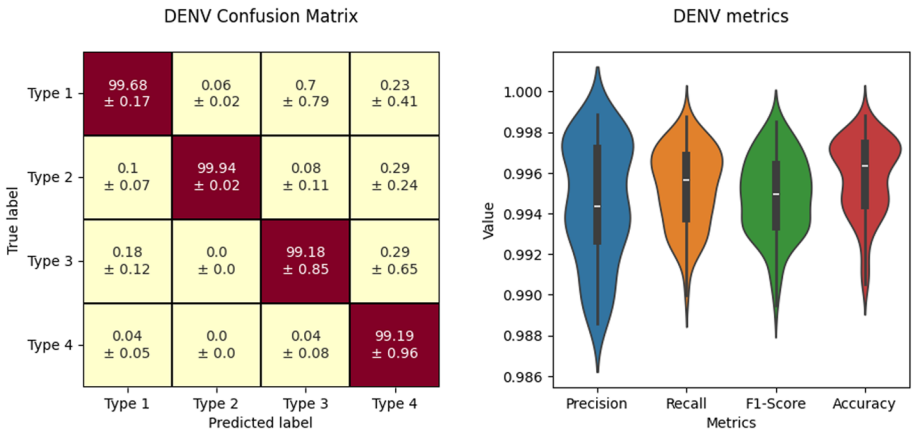

Fig. 2 Confusion matrix and metrics from DENV lineages prediction from GRAMEP

Table 5 Overview of the KEVOLVE dataset [38]

WHO Labe	Pango Lineage	Number of sequences
Alpha	B.1.1.7	175,212
Beta	B.1.351	695
Gamma	P.1	8129
Delta	B.1.617.2	9408
Kappa	B.1.617.1	127
Epsilon	B.1.427/B.1.429	14,674
Iota	B.1.526	19,274
Eta	B.1.525	716
Lambda	C.37	428
Omicron	B.1.1.529/BA.x	106,293
Total number of sequences		334,956

genome in the literature, it was carried out a review and four methods commonly used and studied in problems similar to those presented previously.

The performance of several motif discovery tools in identifying discriminatory sequence regions within SARS-CoV-2 genomes representing different variants was considered. Important methods such as MEME suite (MEME) [42, 43], STREME [44], CASTOR-KRFE [40], and KEVOLVE [38], were considered to compare their results from the previous work by [38].

A dataset of 334, 956 SARS-CoV-2 genomes representing ten World Health Organization (WHO) cataloged variants was analyzed, as shown in Table 5. K-fold cross-validation was performed with a random selection of sequences in 100 iterations. Each fold comprised 2500 sequences for training and the remainder for testing. To ensure balanced representation within the training sets despite varying numbers of available sequences per variant, 250 sequences were allocated to each variant except Kappa (100), Alpha (350), and Omicron (300). Each tool was used to identify discriminatory motifs within the training sets, which were subsequently utilized to train a machine-learning algorithm using the k-mer and step size equal 9 and 1, respectively. We compare the performance of GRAMEP to that achieved by [38] using Fig. 3, which

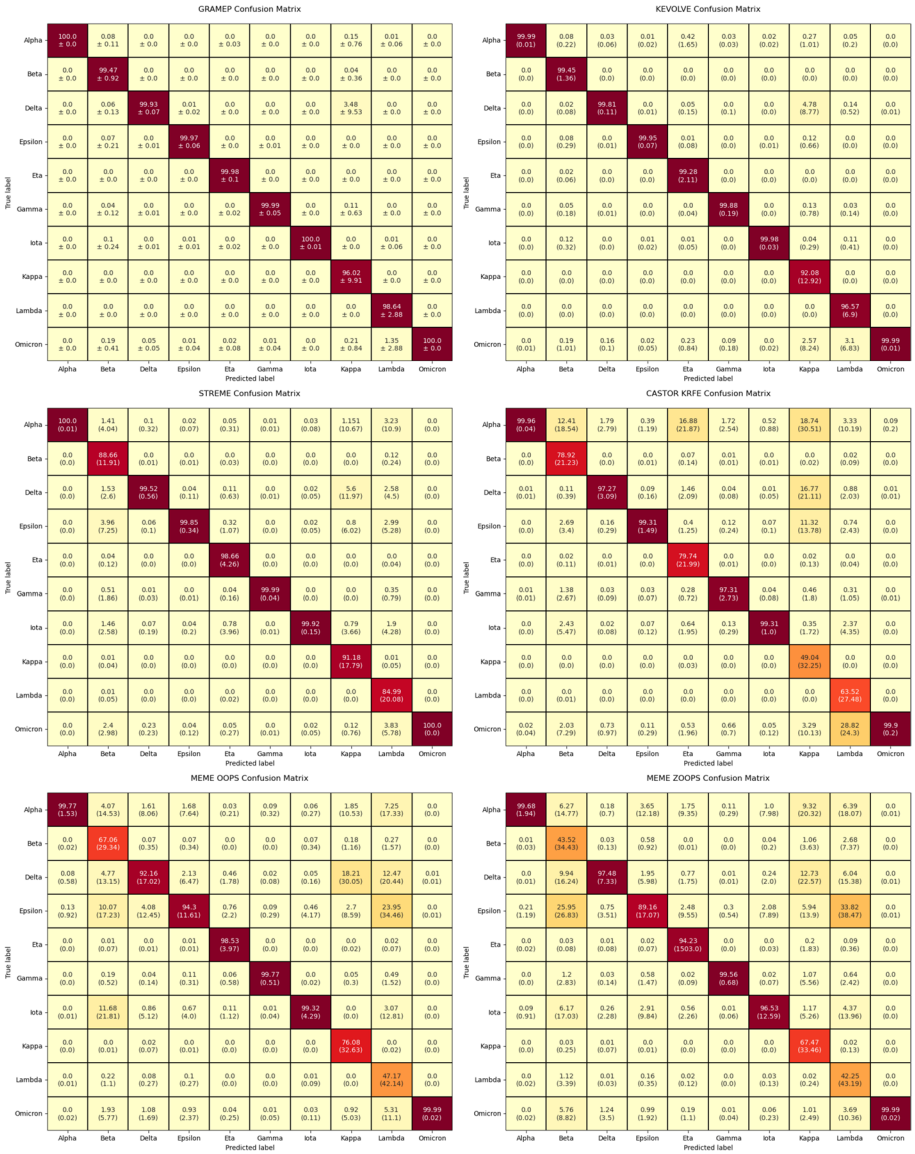

Fig. 3 Confusion matrix from SARS-CoV-2 lineages prediction

presents the confusion matrices considering the prediction accuracy for each variant class and Fig. 4 which presents the metrics obtained by each method.

The analysis of the obtained metrics, alongside the confusion matrix, highlights the GRAMEP's ability to accurately classify new sequences based on a relatively small training dataset. This outcome suggests the method's generalization and indicates that the sub-regions extracted for each variant are indeed discriminative for their respective classes.

The metrics achieved values close to 100%, outperforming the competitor methods, as confirmed by the confusion matrix analysis. Specifically examining the confusion matrix generated by GRAMEP for ten SARS-CoV-2 variant classes, only two classes, Kappa and Lambda, showed values slightly below 99%, with averages of 96% and 98%,

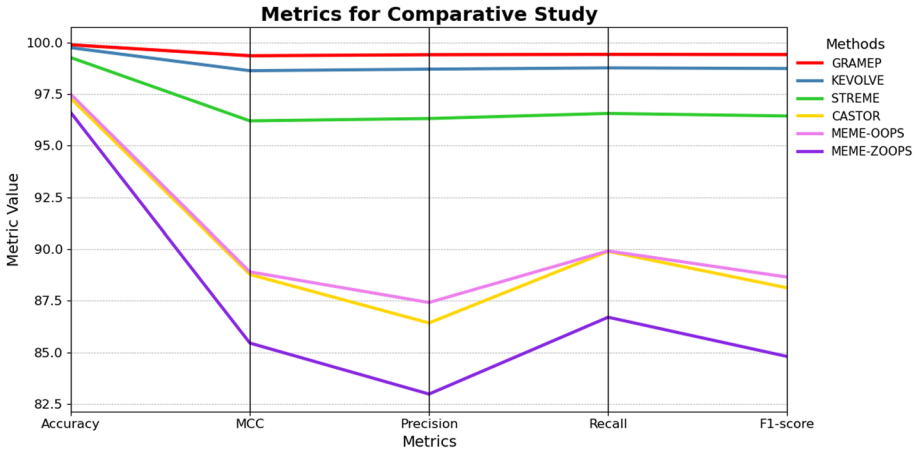

Fig. 4 Results of the Comparative Study

respectively. This trend of lower accuracy in these variant classes was also observed with the other methods, though with a higher rate of misclassification. In particular, the Kappa variant showed some incorrect predictions relative to the Delta variant, which is reasonable given the high degree of genetic similarity between these variants.

As a comparative result, GRAMEP outperformed the other methods, indicating its assertiveness and suitability, offering a contribution as a method and open software for the use and replication of the present study.

Conclusion

Identifying and classifying mutations within genomes are crucial tasks underpinning advancements in public health research, including drug and vaccine development, disease control strategies, and various other areas. However, these tasks present significant challenges. Traditional methods often rely on sequence alignment, which can be computationally expensive, require specific reference information, and potentially generate inaccurate results because of genomic sequences' inherent complexity and variability. Consequently, alignment-free approaches have emerged as a promising alternative for mutation identification. While various methods have been proposed, each employs its unique approach.

This study presents a novel method, called GRAMEP, for selecting the most informative subsequences from genomic data, grounded in the principle of maximum entropy. This approach leverages information theory, particularly Shannon entropy, to identify k-mers that are most discriminative for characterizing different variants of an organism. By prioritizing these informative subsequences, we propose to create unique k-mer signatures for each variant.

Beyond simply selecting the most informative k-mers, the GRAMEP application demonstrates the methodology's potential in genome analysis. The four scenarios analyzed concern obtaining SNPs in silico and viral organisms, identifying mutations in common between variants of the same organism, and classifying sequences from different organisms.

The proposed method is able to discover SNPs with a high reliability rate and also obtain exclusive regions for each variant of the reference, which can be applied as "barcodes" for classifying these organisms. Unlike other methods designed with similar objectives and functionalities, such as KEVOLVE and CASTOR-KRFE, GRAMEP offers a key advantage in requiring only the sequences of the variant under analysis to identify and extract the most discriminative sub-regions. In contrast, tools like KEVOLVE and CASTOR-KRFE generally require data from multiple variants to perform a comparative analysis and identify distinctive regions. This feature enables GRAMEP to be applied more efficiently and with less reliance on external data, making it particularly useful in contexts where access to multiple variants may be limited.

In terms of classification, using an automatic threshold based on maximum entropy reduces the dimensionality of the feature space, maintaining satisfactory accuracy. In addition, our methodology does not depend on prior information; only the input sequences are considered to extract the discrimination features, making it a useful tool in a variety of different scenarios and organisms.

Currently, the GRAMEP primarily focuses on identifying single nucleotide polymorphisms (SNPs) and excludes mutations like insertions or deletions. As a result, real-world applications, such as the SARS-CoV-2 analysis, solely report identified SNP mutations. Future updates could explore this approach further to optimize parametrization for different scenarios. Beyond basic k-mer frequencies, the threshold derived from the maximum entropy principle could be applied to other features extracted from the sequences. These features could encompass physical-chemical and/or biological properties relevant to the specific organism under analysis.

GRAMEP method was implemented in open source and is available at GitHub: https://github.com/omatheuspimenta/GRAMEP under the open-source MIT license. GRAMEP documentation is available from Read the Docs: https://gramep.readthedocs.io/en/latest/. The random sequence generator used during in silico simulations is available at Github: https://github.com/omatheuspimenta/seq_generatoRS.

Appendix A Supplementary Information

A.1 Parameters used

The parameter selection for the scenarios analyzed was conducted through empirical simulations across two distinct scenarios aimed at fine-tuning the parameters. In the first scenario, we used in silico data from HIV, where the k-mer size was varied from 9 to 30 and applied across 100 sets of 1000 sequences each. Performance metrics were then extracted for each simulation, including the true positive rate (TPR), false positive rate (FPR), true negative rate (TNR), false negative rate (FNR), accuracy (ACC), Matthews correlation coefficient (MCC), and F1-score.

In the second scenario, which involved genomes from 20 SARS-CoV-2 variants, 10% of the sequences from each variant were randomly selected. Simulations varied the k-mer size from 9 to 30, comparing the results against data provided by CoV2Var with a cutoff threshold of 95%, as illustrated in Fig. 5.

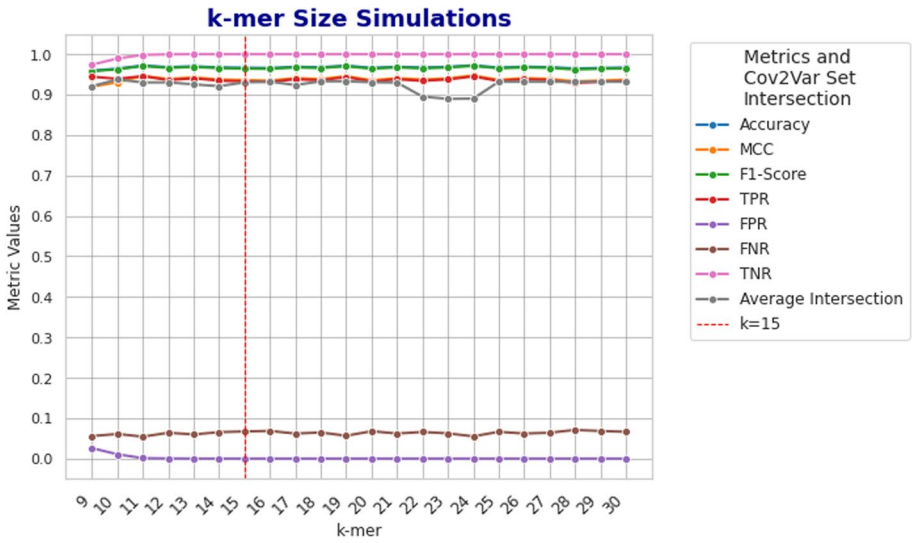

Fig. 5 Plot of Performance Metrics and CoV2Var set intersection for varying *k* values

Table 6 < **variant**>**_FreqExclusiveKmers.csv** example

Position	Reference_value	Variant_value	Frequency
22995	C	A	14234
27874	C	T	14234
8986	C	T	14234
23403	A	G	14233
15451	G	A	14232

The analysis of these results, coupled with the principle of Occam's Razor, enabled the determination of the optimal k-mer size, which was identified as 15. This value was selected to balance simplicity and accuracy in the method's performance.

A.2 gEP outputs

The 'get-mutations' function generates multiple output files and a detailed report if the user opts for its creation. These output files cover various aspects of the method, providing flexibility for different applications. In total, up to 11 files can be generated if graph generation is possible. Below is a description of each file:

1. < **variant**>**_ExclusiveKmers.txt:** a plain text file containing the k-mers unique to the variant relative to the reference sequence.
2. < **variant**>**_ExclusiveKmers.sav:** a binary pickle version of the previous file, suitable for use in Python analyses and program routines.
3. < **variant**>**_FreqExclusiveKmers.csv:** a tabular file listing the identified mutations and their occurrence frequencies. A sample header and rows of the file are shown in Table 6.

4. < **variant**>**_IntersectionKmers.txt:** a text file containing the most informative k-mers common between the variant and the reference.
5. < **variant**>**_IntersectionKmers.sav:** a binary pickle version of the previous file for Python analyses.
6. < **variant**>**_reference.fasta:** a.fasta file containing the reference sequence with all mutations identified in the analysis.
7. < **variant**>**_report.csv:** a comprehensive report of all analyzed sequences, containing information such as sequence ID, mutation position, start and end of the genomic region, gene region type, mutation location, reference k-mer, variant k-mer, and observed mutation. If the user provides a.gff annotation file, this information is enriched. Table 7 provides an example header and rows.
8. < **variant**>**_variations.bed3:** a.bed3 file containing the locations of all mutations identified during the analysis.
9. < **variant**>**_variations.txt:** a text file listing all mutations in 'location:refvar' format.
10. < **variant**>**_variations.sav:** a binary pickle version of '< variant>_variations.txt', for use in Python.
11. **results.pdf:** a generated graph, based in [63], where possible, displaying the reference genome on the x-axis and the detected mutations along with their occurrence frequencies on the y-axis. Figure 6 illustrates a sample graph.

The 'get-intersection' function outputs two files: a text file listing all mutations shared among the selected variants and an upset plot graph illustrating the resulting sets.

The 'classify' function generates two binary files: one containing the trained model and the other containing the intervals for applying the MinMax rescaling. Additionally, three other files are provided: one contains the confusion matrix obtained during

Table 7 < **variant**>**_report.csv example**

Sequence id	Annotation name	Start	End	Type	Modification localization in reference	Reference kmer	Exclusive variant kmer	Reference snp	Variant snp
AY-3_5105	ORF7a	27394	27759	gene	27752	AAAAGA AAGACA GAA	AAAAGA AAGATA GAA	C	T
AY-3_5105	S	21563	25384	gene	22917	ACCTGTATA GATTGT	ACCGGTATA GATTGT	T	G
AY-3_5105	ORF1ab	266	21555	gene	3037	GTATTGTTC TTTCTA	GTATTGTTC TTTTTA	C	T
AY-3_5105	ORF1ab	266	21555	gene	14408	CCCACCTAC AAGTTT	CCCACTTAC AAGTTT	C	T
AY-3_5105	N	28274	29533	gene	28881	AGGGGA ACTTCTCCT	ATGGGA ACTTCT CCT	G	T
AY-3_5105	N	28274	29533	gene	28916	CAATGG CGGTGA TGC	CAATGGCTG TGATGC	G	T
AY-3_5105	ORF1ab	266	21555	gene	14408	TTCCCACCT ACAAGT	TTCCCACTT ACAAGT	C	T

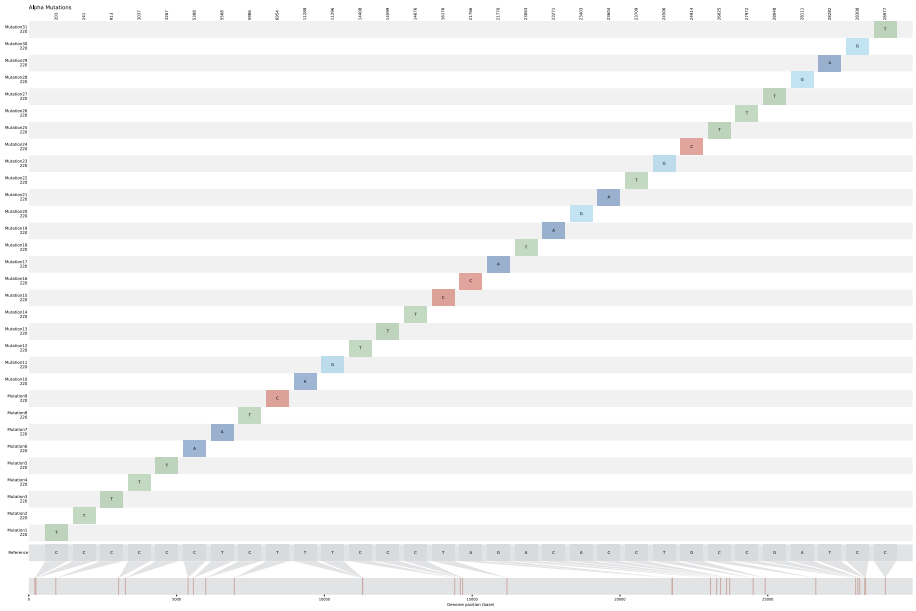

Fig. 6 **results.pdf** example

the model's training, another includes the metrics derived from the training process, and the final file holds the feature matrix extracted during model training.

A.3 `grid-search` function

One approach to determining the appropriate values for the 'word' and 'step' parameters involves utilizing the `grid-search` function.

This method selects parameter values heuristically by exploring a predefined grid containing potential values within specified ranges for the 'word' and 'step' parameters. The parameters are chosen based on their ability to maximize the set of exclusive k-mers. Given the heuristic nature of the method, k-mer values that are close to the maximum are also considered. Fuzzy matching is applied during this process to establish thresholds, enabling the identification and removal of duplicate k-mers.

To perform the grid search for suggested 'word' and 'step' values, two inputs must be provided: a reference sequence for the variant, which is used as a baseline for identifying mutations, and a file containing variant sequences, this file can be a subset of moderate size to ensure computational efficiency.

It is important to note that the values obtained through the grid-search method represent suggested values and may not be optimal since it is a heuristic search. Following the grid search, manual refinement of the parameters may be required, particularly when focusing on a specific variant under study.

Abbreviations

SNVs	Single nucleotide variants
SNPs	Single nucleotide polymorphisms
HIV-1	Human immunodeficiency virus type 1
DENV	Dengue virus
GRAMEP	Genome vaRiation Analysis from the Maximum Entropy

Acknowledgements
This study was financed by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) - Finance Code 001, the Fundação Araucária (Grant Nos. 035/2019, 138/2021 and NAPI - Bioinformática) and CNPq (Grant Nos. 440412/2022-6 and 408312/2023-8).

Author contributions
MHP-Z and FML implemented and tested the computer code and supporting algorithms; conceived and conducted the experiments, analyzed the results, and wrote and reviewed the manuscript. AYK implemented the computer code and supporting algorithms and reviewed the manuscript. ALLV wrote and reviewed the manuscript.

Funding
Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) - Finance Code 001. Fundação Araucária (Grant Nos. 035/2019, 138/2021 and NAPI - Bioinformática). CNPq (Grant Nos. 440412/2022-6 and 408312/2023-8)

Availability of data and materials
Project name: GRAMEP - Genome vaRiation Analysis from the Maximum Entropy Principle ***Project home page:*** https://github.com/omatheuspimenta/GRAMEP ***Operating system(s):*** Linux ***Programming language:*** Python and Rust ***Other requirements:*** Python 3.12 or higher ***License:*** MIT ***Any restrictions to use by non-academics:*** None The data used are available at: https://github.com/omatheuspimenta/GRAMEP/tree/main/data/datasets.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 21 November 2024 Accepted: 6 January 2025
Published online: 25 February 2025

References

1. Zielezinski A, Vinga S, Almeida J, Karlowski WM. Alignment-free sequence comparison: benefits, applications, and tools. Genome Biol. 2017;18(1):186. https://doi.org/10.1186/s13059-017-1319-7.
2. De Pierri CR, Voyceik R, Mattos LGC, Kulik MG, Camargo JO, Oliveira AM, Lima Nichio BT, Marchaukoski JN, Silva Filho AC, Guizelini D, Ortega JM, Pedrosa FO, Raittz RT. SWeeP: representing large biological sequences datasets in compact vectors. Sci Rep. 2020;10(1):91. https://doi.org/10.1038/s41598-019-55627-4.
3. Garg S. Computational methods for chromosome-scale haplotype reconstruction. Genome Biol. 2021. https://doi.org/10.1186/s13059-021-02328-9.
4. Crick F. Central dogma of molecular biology. Nature. 1970;227(5258):561–3. https://doi.org/10.1038/227561a0.
5. Snustad DP, Simmons MJ. Principles of genetics. Hoboken: Wiley; 2015.
6. Garg S. Towards routine chromosome-scale haplotype-resolved reconstruction in cancer genomics. Nat Commun. 2023. https://doi.org/10.1038/s41467-023-36689-5.
7. Tian D, Sun Y, Xu H, Ye Q. The emergence and epidemic characteristics of the highly mutated SARS-CoV-2 Omicron variant. J Med Virol. 2022. https://doi.org/10.1002/jmv.27643.
8. Worobey M. Dissecting the early COVID-19 cases in Wuhan. Science. 2021;374(6572):1202–4. https://doi.org/10.1126/science.abm4454.
9. Perico CP, et al. Genomic landscape of the SARS-CoV-2 pandemic in Brazil suggests an external P.1 variant origin. Front Microbiol. 2022;13:1037455. https://doi.org/10.3389/fmicb.2022.1037455.
10. Franceschi VB, et al. Mutation hotspots and spatiotemporal distribution of SARS-CoV-2 lineages in Brazil, February 2020–2021. Virus Res. 2021;304: 198532. https://doi.org/10.1016/j.virusres.2021.198532.
11. Harvey WT, et al. SARS-CoV-2 variants, spike mutations and immune escape. Nat Rev Microbiol. 2021;19(7):409–24. https://doi.org/10.1038/s41579-021-00573-0.
12. Andersen KG, et al. The proximal origin of SARS-CoV-2. Nat Med. 2020;26(4):450–2. https://doi.org/10.1038/s41591-020-0820-9.
13. Holmes EC. The evolutionary biology of dengue virus. In: Bock G, Goode J, editors. Novartis foundation symposia, vol. 277. 1st ed. Hoboken: Wiley; 2006. p. 177–92. https://doi.org/10.1002/0470058005.ch13.
14. Yeo JY, et al. The determination of HIV-1 RT mutation rate, its possible allosteric effects, and its implications on drug resistance. Viruses. 2020;12(3):297. https://doi.org/10.3390/v12030297.
15. Cuevas JM, et al. Extremely high mutation rate of HIV-1 in vivo. PLoS Biol. 2015;13(9):1002251. https://doi.org/10.1371/journal.pbio.1002251.
16. Lange K. Mathematical and statistical methods for genetic analysis. Statistics for biology and health. New York: Springer; 2002. https://doi.org/10.1007/978-0-387-21750-5.
17. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403–10. https://doi.org/10.1016/S0022-2836(05)80360-2.
18. Smith TF, Waterman MS, et al. Identification of common molecular subsequences. J Mol Biol. 1981;147(1):195–7.

19. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol. 1970;48(3):443–53. https://doi.org/10.1016/0022-2836(70)90057-4.
20. Zielezinski A, et al. Benchmarking of alignment-free sequence comparison methods. Genome Biol. 2019;20(1):144. https://doi.org/10.1186/s13059-019-1755-7.
21. Forsdyke DR. Success of alignment-free oligonucleotide (k-mer) analysis confirms relative importance of genomes not genes in speciation and phylogeny. Biol J Linn Soc. 2019;128(2):239–50.
22. Solis-Reyes S, Avino M, Poon A, Kari L. An open-source k-mer based machine learning tool for fast and accurate subtyping of hiv-1 genomes. PLoS ONE. 2018;13(11):0206409.
23. Cong Y, Chan Y-B, Ragan MA. A novel alignment-free method for detection of lateral genetic transfer based on tf-idf. Sci Rep. 2016;6(1):30308.
24. Song N, Joseph JM, Davis GB, Durand D. Sequence similarity network reveals common ancestry of multidomain proteins. PLoS Comput Biol. 2008;4(5):1000063.
25. Duffy S, et al. Rates of evolutionary change in viruses: patterns and determinants. Nat Rev Genet. 2008;9(4):267–76. https://doi.org/10.1038/nrg2323.
26. Zhang Y, et al. A review on recent computational methods for predicting noncoding RNAs. BioMed Res Int. 2017;2017:1–14. https://doi.org/10.1155/2017/9139504.
27. Murugaiah M, Ganesan M. A novel frequency based feature extraction technique for classification of corona virus genome and discovery of COVID-19 repeat pattern. Braz Arch Biol Technol. 2021;64:21210075. https://doi.org/10.1590/1678-4324-2021210075.
28. Marchet C, Kerbiriou M, Limasset A. BLight: efficient exact associative structure for k-mers. Bioinformatics. 2021;37(18):2858–65. https://doi.org/10.1093/bioinformatics/btab217.
29. Ito EA, et al. BASiNET-BiologicAl Sequences NETwork: a case study on coding and non-coding RNAs identification. Nucleic Acids Res. 2018;46(16):96–96. https://doi.org/10.1093/nar/gky462.
30. Amin N, et al. Evaluation of deep learning in non-coding RNA classification. Nat Mach Intell. 2019;1(5):246–56. https://doi.org/10.1038/s42256-019-0051-2.
31. Barros-Carvalho GA, Van Sluys M-A, Lopes FM. An efficient approach to explore and discriminate anomalous regions in bacterial genomes based on maximum entropy. J Comput Biol. 2017;24(11):1125–33. https://doi.org/10.1089/cmb.2017.0042.
32. Vopson MM. Dynamics of SARS-CoV-2 genetic mutations and their information entropy. preprint, Bioinformatics (June 2022). https://doi.org/10.1101/2022.06.13.495895.
33. Vopson MM, Robson SC. A new method to study genome mutations using the information entropy. Phys A Stat Mech Appl. 2021;584: 126383. https://doi.org/10.1016/j.physa.2021.126383.
34. Kuksa P, Pavlovic V. Efficient alignment-free DNA barcode analytics. BMC Bioinform. 2009;10(S14):9. https://doi.org/10.1186/1471-2105-10-S14-S9.
35. Chor B, Horn D, Goldman N, Levy Y, Massingham T. Genomic DNA k-mer spectra: models and modalities. Genome Biol. 2009;10(10):108. https://doi.org/10.1186/gb-2009-10-10-r108.
36. Fiscon G, Weitschek E, Cella E, Lo Presti A, Giovanetti M, Babakir-Mina M, Ciotti M, Ciccozzi M, Pierangeli A, Bertolazzi P, Felici G. MISSEL: a method to identify a large number of small species-specific genomic subsequences and its application to viruses classification. BioData Mining. 2016;9(1):38. https://doi.org/10.1186/s13040-016-0116-2.
37. Lebatteux D, et al. Machine learning-based approach KEVOLVE efficiently identifies SARS-CoV-2 variant-specific genomic signatures. preprint, Bioinformatics (February 2022). https://doi.org/10.1101/2022.02.07.479343 .
38. Lebatteux D, et al. Machine learning-based approach KEVOLVE efficiently identifies SARS-CoV-2 variant-specific genomic signatures. PLoS ONE. 2024;19(1):0296627. https://doi.org/10.1371/journal.pone.0296627.
39. Lebatteux D, Diallo A.B. Combining a genetic algorithm and ensemble method to improve the classification of viruses. In: 2021 IEEE international conference on bioinformatics and biomedicine (BIBM), pp. 688–693. IEEE, Houston, TX, USA (2021). https://doi.org/10.1109/BIBM52615.2021.9669670
40. Lebatteux D, et al. Toward an alignment-free method for feature extraction and accurate classification of viral sequences. J Comput Biol. 2019;26(6):519–35. https://doi.org/10.1089/cmb.2018.0239.
41. Thanos D, Li W, Provata A. Entropic fluctuations in DNA sequences. Phys A Stat Mech Appl. 2018;493:444–57. https://doi.org/10.1016/j.physa.2017.11.119.
42. Bailey TL, et al. The value of position-specific priors in motif discovery using MEME. BMC Bioinform. 2010;11(1):179. https://doi.org/10.1186/1471-2105-11-179.
43. Bailey TL, Johnson J, Grant CE, Noble WS. The MEME suite. Nucleic Acids Res. 2015;43(W1):39–49. https://doi.org/10.1093/nar/gkv416.
44. Bailey TL. STREME: accurate and versatile sequence motif discovery. Bioinformatics. 2021;37(18):2834–40. https://doi.org/10.1093/bioinformatics/btab203.
45. Lebatteux D, *et al.* KANALYZER: a method to identify variations of discriminative k-mers in genomic sequences. In: 2022 IEEE international conference on bioinformatics and biomedicine (BIBM), pp. 757–762. IEEE, Las Vegas, NV, USA (2022). https://doi.org/10.1109/BIBM55620.2022.9995370.
46. Olson RD, et al. Introducing the bacterial and viral bioinformatics resource center (BV-BRC): a resource combining PATRIC. IRD ViPR Nucleic Acids Res. 2023;51(D1):678–89. https://doi.org/10.1093/nar/gkac1003.
47. Khan S, AlQahtani SA, Noor S, Ahmad N. Pssm-sumo: deep learning based intelligent model for prediction of sumoylation sites using discriminative features. BMC Bioinform. 2024. https://doi.org/10.1186/s12859-024-05917-0.
48. Tsallis C. Nonadditive entropy: the concept and its use. Eur Phys Jo A. 2009;40(3):257. https://doi.org/10.1140/epja/i2009-10799-0.
49. Clausius R. The mechanical theory of heat. Sacramento: Creative Media Partners LLC; 2019.
50. Boltzmann L, McGuinness BF. Theoretical physics and philosophical problems: selected writings. Vienna circle collection. Berlin: Springer; 2012.
51. Shannon CE. A mathematical theory of communication. Bell Syst Tech J. 1948;27(3):379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.

52. Jaynes ET. Information theory and statistical mechanics. Phys Rev. 1957;106(4):620–30. https://doi.org/10.1103/PhysRev.106.620.
53. Guiasu S, Shenitzer A. The principle of maximum entropy. Math Intell. 1985;7(1):42–8. https://doi.org/10.1007/BF03023004.
54. Kapur JN, Sahoo PK, Wong AKC. A new method for gray-level picture thresholding using the entropy of the histogram. Comput Vis Graph Image Process. 1985;29(3):273–85. https://doi.org/10.1016/0734-189X(85)90125-2.
55. Khan S, Khan M, Iqbal N, AmiruddinAbdRahman M, KhalisAbdulKarim M. Deep-pirna: bi-layered prediction model for piwi-interacting rna using discriminative features. Comput Mater Continua. 2022;72(2):2243–58. https://doi.org/10.32604/cmc.2022.022901.
56. Levenshtein VI. Binary codes capable of correcting deletions, insertions, and reversals. Sov Phys Dokl. 1966;10(8):707–10.
57. Myers G. A fast bit-vector algorithm for approximate string matching based on dynamic programming. J ACM (JACM). 1999;46(3):395–415.
58. Struck D, et al. COMET: adaptive context-based modeling for ultrafast HIV-1 subtype identification. Nucleic Acids Res. 2014;42(18):144–144. https://doi.org/10.1093/nar/gku739.
59. Plummer E, et al. Dengue virus evolution under a host-targeted antiviral. J Virol. 2015;89(10):5592–601. https://doi.org/10.1128/JVI.00028-15.
60. Feng Y, et al. COV2Var, a function annotation database of SARS-CoV-2 genetic variation. Nucleic Acids Res. 2024;52(D1):701–13. https://doi.org/10.1093/nar/gkad958.
61. Khare S, et al. GISAID's role in pandemic response. China CDC Wkly. 2021;3(49):1049–51. https://doi.org/10.46234/ccdcw2021.255.
62. Pachetti M, et al. Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. J Transl Med. 2020;18(1):179. https://doi.org/10.1186/s12967-020-02344-6.
63. O'Toole: snipit. GitHub (2020)

Publisher's Note