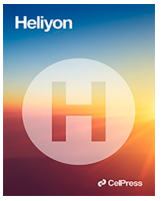
Heliyon 10 (2024) e36914

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

DNA N-gram analysis framework (DNAnamer): A generalized N-gram frequency analysis framework for the supervised classification of DNA sequences

John S. Malamon [a,b]

[a] University of Colorado Anschutz Medical Campus, Department of Surgery, Division of Transplant Surgery, 1635 Aurora Court, Aurora, CO, 80045, USA

[b] Colorado Center for Transplantation Care, Research and Education (CCTCARE), Division of Transplant Surgery, Aurora, CO, 80045, USA

A B S T R A C T

In 1948, Claude Shannon published a mathematical system describing the probabilistic relationships between the letters of a natural language and their subsequent order or syntax structure. By counting unique, reoccurring sequences of letters called N-grams, this language model was used to generate recognizable English sentences from N-gram frequency probability tables. More recently, N-gram analysis methodologies have been successfully applied to address many complex problems in a variety of domains, from language processing to genomics. One such example is the common use of N-gram frequency patterns and supervised classification models to determine authorship and plagiarism. In this paradigm, DNA is a language model where nucleotides are analogous to the letters of a word and nucleotide N-grams are analogous to the words of a sentence. Because DNA contains highly conserved and identifiable nucleotide sequence frequency patterns, this approach can be applied to a variety of classification and data reduction problems, such as identifying species based on unknown DNA segments. Other useful applications of this methodology include the identification of functional gene elements, microorganisms, sequence contamination, and sequencing artifacts. To this end, I present DNAnamer, a generalized and extensible methodological framework and analysis toolkit for the supervised classification of DNA sequences based on their N-gram frequency patterns.

1. Introduction

In 1948, Claude Shannon published "A Mathematical Theory of Communication." [1], which provided a mathematical model defining the maximum information transmission rate and channel capacity that could be achieved given an arbitrarily small probability of error. This groundbreaking work not only laid the foundation for information theory and modern telecommunication systems, but it also ignited the fields of stochastic modeling, data compression [2–4], cryptography [5,6], and machine learning [7,8]. This marked the birth of the Information Age. Moreover, Shannon arrived at the critical conclusion that communication signals and their information sources were statistically independent from the meaning of a given message. This finding is immensely important, as it dispels the notion that more data implies more meaning. Rather, one must sift through layers of structured data and noise to extract meaning.

In this seminal work, Shannon also developed a formal language model describing the probabilistic relationships between the letters in a natural language [9] and their subsequent order, or syntax structure, by counting reoccurring sequences of letters called N-grams. N-grams are unique, iterative sequences of N adjacent symbols. For example, the second-order N-gram set for the word "source" is ["so", "ou", "ur", "rc", "ce"]. These models are referred to as stochastic processes because each N-gram within a set is represented as a random variable governed by a finite set of probabilistic states. Shannon's remarkable ability to produce ordered or

E-mail address: john.malamon@cuanschutz.edu.

https://doi.org/10.1016/j.heliyon.2024.e36914
Received 11 April 2024; Received in revised form 22 August 2024; Accepted 23 August 2024
Available online 24 August 2024

Abbreviations:

ANOVA	analysis of variance
AUC	area under the curve
bat	Miniopterus natalensis
DNA	deoxyribonucleic acid
DNAnamer	DNA N-gram Analysis Framework
dolphin	*Delphinus delphis*
elephant	*Elephas maximus*
hg38	human reference genome version 38
human	*Homo sapiens*
Gb	gigabyte
kb	kilobase
koala	*Phascolarctos cinereus*
Mb	megabase
MB	megabyte
OOB	out-of-bag
PCR	polymerase chain reaction
RF	Random Forest
SVM	Support Vector Machines

grammatically structured English sentences from N-gram probability frequency tables demonstrated that distinct and conserved syntax patterns and structures were clearly embedded within natural languages. To this end, he stated, "It appears then that a sufficiently complex stochastic process will give a satisfactory representation of a discrete [information] source."

The first and most fundamental hypothesis of this approach, H_1, is that the discrete information source of DNA (deoxyribonucleic acid), nucleotide sequences, can be adequately represented as a stochastic process similar to Shannon's language model, where nucleotides ["A","C″,"G","T"] are analogous to the letters of a word and nucleotide sequence N-grams are analogous to the words of a sentence. This methodology also relies on two additional hypotheses. H_2 requires that DNA N-grams such as bi-, tri-, and tetra-nucleotide pairings do not occur randomly, and H_3 requires that DNA N-gram frequency patterns are conserved and identifiable. In fact, it has been demonstrated that di-, tri-, and poly-nucleotide repeats are not randomly distributed and are highly conserved exhibiting complex and distinct patterns [10–12]. Given H_2 and H_3, DNA N-gram frequency patterns can be leveraged to construct a generalized stochastic model to efficiently solve a wide variety of supervised classification and data reduction problems highly relevant to genomics and genetics.

Significant advancements have occurred in the past four decades in the fields of text categorization, authorship [13–18], and plagiarism detection [19,20]. This progress has been achieved through the successful integration of novel N-gram analysis methodologies with supervised machine learning approaches. For example, anti-plagiarism tools based on N-gram analysis are commonplace and often positively regarded in educational institutions [21]. Since the sequencing of the human genome, numerous N-gram analysis methodologies have yielded an impressive range of inventive solutions by addressing relevant problems in genomics, genetics, and proteomics. These solutions include improving sequence alignment algorithms [22–24], describing microbial genetics [25,26], analyzing subcellular proteomes [27], classifying promotor sequences [28], and characterizing protein structures [29–31]. These tasks are highly complex and computationally expensive. The breadth of applications for N-gram analysis is impressive. N-gram analysis coupled with machine learning offers the potential to improve both the detection accuracy and computational efficiency of existing and novel algorithms. Furthermore, DNA barcoding is an essential analysis tool for identifying many microbial species, including, but not limited to, bacteria, fungi, and lichen species. Currently, investigators rely on Sanger and high-throughput DNA sequence analysis to identify most lichen species [32–34].

To this end, I present DNAnamer, or DNA N-gram Analysis Framework, a generalized DNA N-gram frequency analysis methodology and analysis toolkit for the binary and multi-class supervised classification of nucleotide sequences. In summary, DNAnamer automatically pulls the entire reference genome for a given species, calculates second-to fifth-order N-gram frequency tables, and performs binary and multi-class machine learning classification. DNAnamer provides a highly generalized and extensible model by allowing the user to define the training and test sequence sets. DNAnamer is available as an R package.

2. Materials and methods

2.1. Data source and description

The human reference genome (hg38) and four additional reference genomes were randomly selected from the class *Mammalia* and analyzed to demonstrate a specific use case and highlight the broad applicability of this methodology. Each species considered required at least 1000 megabase of nucleotide sequences, or roughly one-third of the human genome. The species used in these *in silico* experiments were *Homo sapiens* (human), *Miniopterus natalensis* (bat), *Elephas maximus* (elephant), *Phascolarctos cinereus* (koala), and

Delphinus delphis (dolphin). The five reference genomes were downloaded from the National Center for Biotechnology Information's genome resource database using the 'biomartr' package [35]. The reference genomes totaled approximately 3.3 (human), 1.8 (bat), 3.4 (elephant), 3.2 (koala), and 2.5 (dolphin) billion nucleotides.

2.2. DNA N-grams and frequency tables

DNA N-grams are unique, iterative sets of adjacent nucleotide sequences. The order of an N-gram is equivalent to the number of adjacent nucleotide sequences. For example, "AA" represents a second-order N-gram. The total number of DNA N-gram sequences in a set is equal to 4^O, where O is the order or the number of consecutive nucleotides. For example, the second-order DNA N-gram nucleotide sequence set is ["AA","AC","AG","AT","CA","CC","CG","CT","GA","GC","GG","GT","TA","TC","TG", "TT"]. Because this study involved analyzing nucleotide frequency patterns, N-gram frequency tables were calculated for all reference genomes using the 'seqinR' package [36]. The analysis of variance (ANOVA) test [37] was used to calculate the difference in N-gram frequency means within and across species. To account for multiple hypothesis testing and control for experiment-wise error, all p-values were adjusted using Tukey's pairwise comparison test. All.

Equation (1) demonstrates a second-order N-gram frequency (f) matrix. The *'no.tests'* was equal to the total DNA sequence length divided by the segment length. The *'no.tests'* was selected to provide at least 200 validation samples per experiment, which provided an ample range of frequency distributions and a minimum performance evaluation resolution of 0.5 %. N-gram frequency tables were calculated for all randomly selected training and validation DNA segment sets and transposed into matrices where the columns were equal to the total cumulative number of N-grams and the rows were equal to the *'no.tests'* for a given experiment. The maximum cumulative sum of N-grams is given in **Equation (2)**.

$$\text{Second-order Frequency Matrix} = \sum_{i=1}^{n=\text{no.tests}} \begin{Bmatrix} f_{AA} & f_{AC} & f_{AG} & f_{AT} \\ f_{CA} & f_{CC} & f_{CG} & f_{CT} \\ f_{GA} & f_{GC} & f_{GG} & f_{GT} \\ f_{TA} & f_{TC} & f_{TG} & f_{TT} \end{Bmatrix} \quad \text{Equation 1}$$

$$\text{Maximum Cumulative Sum of N-grams} = 4^2 + 4^3 + 4^4 + 4^5 \quad \text{Equation 2}$$

2.3. Randomness analysis

This methodology relies on the assumption that DNA N-gram frequencies are not randomly distributed. Specifically, there are two relevant dimensions or definitions of randomness implied. First, in-species randomness was tested to ensure non-randomness across the set of all N-grams within each species. Second, cross-species randomness was tested to ensure nonrandomness between two or more

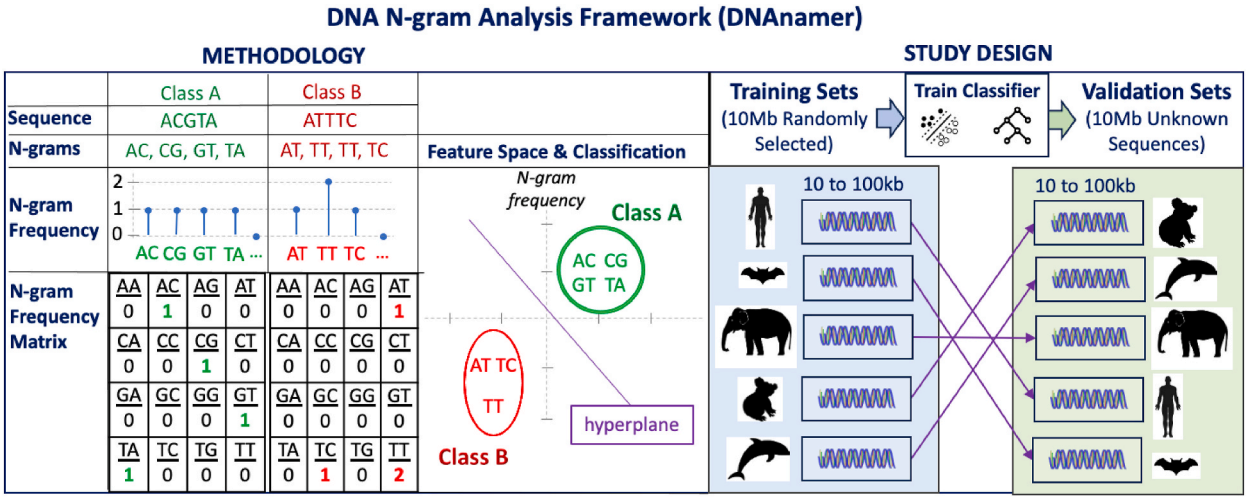


Fig. 1. A graphic abstract of the methodology and experimental design. The DNAnamer (DNA N-gram Analysis Framework) methodology is illustrated to the left, where two simple examples of second-order DNA N-gram sequences are provided. N-gram frequency tables were created, and DNA sequences were classified based on their N-gram frequency patterns. Support Vector Machines were used for binary classification, and the Random Forest methodology was used to classify more than two classes. The experimental study design is depicted to the right, where a range of DNA segments ranging from 10 to 100 kb in length were randomly selected from each of the five species' (human, bat, elephant, koala, and dolphin) reference genomes to develop training sets (blue panel). Validation sets of unknown DNA segments ranging from 1 to 100 kb in length were classified for all five species (light green panel).

given species. Two well-accepted randomness methodologies were applied to directly test these assumptions. First, the Wald-Wolfowitz runs test [38] was used to test the two randomness hypotheses. This approach makes minimal assumptions regarding the underlying N-gram frequency distributions. Second, Bartels' test for randomness [39] was applied. This test is a ranked version of von Neumann's ratio test for randomness [40]. In summary, von Neumann's approach evaluates statistical trends in the successive means of numerical sequences. Bartels' ranked test offers several distinct advantages over von Neumann's ratio test.

2.4. Experimental design

A graphic illustration of DNAnamer and the *in silico* experimental design was provided in Fig. 1, where sets of DNA segments ranging from 10 to 100 kb in length were randomly selected from each of the five species' (human, bat, elephant, koala, and dolphin) reference genomes to develop training and validation sets, each consisting of 20 Mb of continuous DNA sequence. The main objectives of this experiment were to eliminate the aforementioned assumptions, identify N-gram frequency patterns, and measure the classification accuracy and performance of DNAnamer. Training sets were used to calculate priors and predict unknown DNA sequences of varying length. This experiment also examined the relationship between DNAnamer's classification performance as a function of N-gram order and the DNA segment length. Therefore, second, third, fourth, and fifth-order N-grams were analyzed and compared. Relatively small DNA segments up to 100 kb in length were analyzed to discover the upper limits of this model's classification performance. Specifically, unknown human DNA segments totaling 20 Mb and ranging from 10 to 100 kb were classified against each of the four other mammalian species' reference genomes. The five sequence segment lengths tested were 10, 20, 40, 80, and 100 kb. This yielded a total of eighty binary classification experiments. The number of validation tests for each experiment was equal to the total sequence length (20 Mb) divided by the segment length. For example, the 100 kb segment length experiment yielded 200 validation sequences.

2.4.1. Model training and validation

DNA sequences were randomly sampled across the entire reference genome of each species to create balanced training sets each consisting of 20 Mb of continuous sequence. N-gram frequency distribution priors were calculated using only the training sets. The validation sets (20 Mb) consisted of continuous DNA segment sets ranging from 10 to 100 kb and were used to classify unknown DNA sequence segments from all five organisms and test, or validation, sets. Thus, the training and validation sets were split evenly. The AUC was provided as a function of the N-gram order and DNA sequence segment length, yielding twenty multi-classification experiments.

2.4.2. Binary classification

Support Vector Machines (SVM) have proven to be a highly versatile and effective classification approach well-suited for a variety of biomedical and genomic applications [41–44]. Importantly, SVM have been successfully employed to identify species based on their DNA sequences, a technique also known as DNA barcoding [45]. For this reason, the SVM methodology [46] was selected for the binary classification experiments. All eighty experiments were performed using the same parameters. Namely, a radial kernel was selected with a cost of 3 and a sigma value of 0.2. 50 % to.

A radial kernel provides non-linear decision boundaries. Cost is the penalty associated with misclassifications or exceeding the hyperplane's margins. 50 % sample hold-out validation was used to calculate the area under the receiver operating characteristic curve, or AUC. The AUC was provided as a function of the N-gram order (second, third, fourth, and fifth) and DNA segment length (10–100 kb).

2.4.3. Multi-class classification

The Random Forest (RF) methodology [47] has also proven to be a highly versatile and effective classification approach for a variety of biomedical and genomic applications [48–51]. This methodology has been successfully employed to perform DNA barcoding [52–54]. For this reason, the Random Forest classification methodology was selected for multi-class classification. All experiments were performed under the same parameters, or initial conditions. Namely, a maximum of 300 decision trees was specified with 100 permutations and 40 mtry. The number of decision trees represents the number of decision nodes produced for each model. It also presents a trade-off between prediction accuracy and computational efficiency [55]. A sensitivity analysis was conducted to test the stability of the accuracy of the predictions along a wide range of potential values by varying the number of trees from 100 to 500 while holding all other parameters constant. In summary, 300 trees provided the most stable performance without dramatically increasing the processing time. The number of permutations represents the number of out-of-bag (OOB) recursions performed per tree, which was used to assess variable importance and perform bootstrapping to estimate in-sample accuracy and error. The mtry parameter specifies the maximum number of input features or variables available to a decision tree. 40 input features per tree were selected to avoid any chance of overfitting. Variable importance was computed using OOB sample bootstrapping to determine the mean decrease accuracy of each independent variable, which is based on how much the accuracy decreased when the variable was excluded throughout the bootstrapping process.

2.4.4. Benchmark analysis

Finally, computational performance benchmarking was performed on the human reference genome. Total runtimes (seconds), the number of iterations per second (iterations/second), the total memory allocated (GB), and core seconds (runtime/cores) were recorded on a 2.4 GHz 8-Core Intel Core i9 processor with 64 GB of available memory. Performance benchmarks were provided for all necessary

functions and the five DNA segment lengths. 20 Mb of DNA sequences were processed for each of the eight benchmark experiments. Fifth-order N-gram analysis was performed to provide a realistic estimation of the computational resources required to replicate these experiments. All analyses were performed using the R statistical language version 4.3.2 [56].

3. Results

3.1. DNAnamer

The DNA N-gram Analysis Framework (DNAnamer) is a novel N-gram frequency analysis framework for the supervised classification of DNA sequences and is available as an R software package or library. Documentation and vignettes with detailed code demonstrations are available at https://github.com/jmal0403/DNAnamer/wiki. All major classification experiments performed herein can be reproduced using the vignettes and sample code. In summary, DNAnamer provides a highly generalized, efficient, and extensible analytical framework that can be readily applied to take on a variety of classification and data reduction problems relevant to genomics and genetics.

3.2. DNA N-gram frequencies are nonrandomly distributed and conserved

The Wald-Wolfowitz runs test and Bartels' ranked test confirmed in-species and cross-species N-gram frequency nonrandomness. These two randomness tests were repeated for the five N-gram orders and across the five species, yielding adjusted p-values less than 2.2e-16 in all instances. Thus, N-grams were far from randomly distributed within and across species. Fig. 2A–D provided box plots of the second-order N-gram frequencies for 200 randomly selected DNA segments 100 kb in length calculated using human versus bat, elephant, koala, and dolphin reference genomes. ANOVA testing after p-value adjustment revealed many highly statistically significant differences in the N-gram frequency means between and among species (Supplemental Fig. 1), further supporting the nonrandomness hypothesis. Many statistically significant (p-value<0.05, Supplemental Table 1) cross-species N-gram frequency patterns were observed, providing a wealth of independent random variables for classification. The cross-species N-gram frequency correlation matrix was given in Fig. 2E. N-gram frequency patterns were highly conserved across the five species, all yielding correlation coefficients greater than 0.95. Thus, H_2 and H_3 have been directly tested to reveal many unique and identifiable N-gram frequency patterns in and across species.

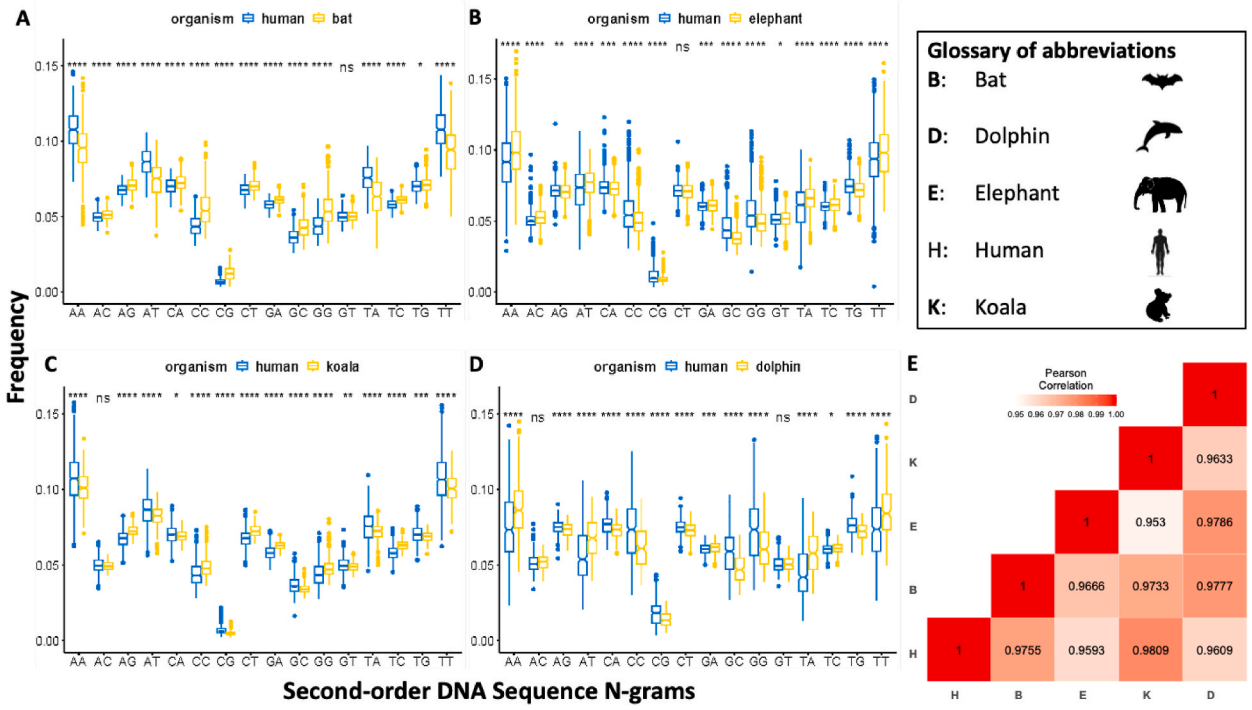


Fig. 2. Second-order DNA N-gram frequencies for human versus bat, elephant, koala, and dolphin sequences. Using training sets of 20 MB and DNA segment lengths of 100 kb, second-order N-gram frequencies were calculated for A) human versus bat; B) human versus elephant; C) human versus koala; and D) human versus dolphin. The ANOVA was used to measure the differences in the group means between the two species. The following p-value notation was used: $p > 0.05$ [13], $p < 0.05$ (*), $p < 0.01$(**), $p < 0.001$ (***), *ns = not significant*. All p-values were adjusted for multi-hypothesis testing. Panel E provided a Pearson's cross-correlation matrix for the second-order N-gram frequencies comparing all five species.

3.3. N-gram variable importance using Random Forest classification

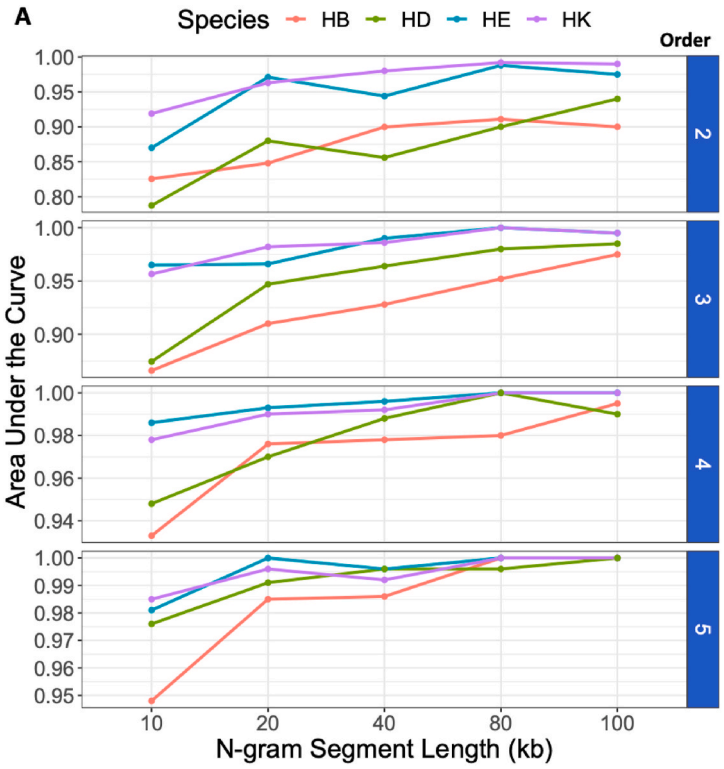
Supplemental Table 2 provided the top 40 most important N-grams for all five species, sorted by their importance in humans. In summary, 36 out of the 40 top N-grams were fifth-order, and the remaining four were fourth-order. Although second- and third-order N-grams were informative to the RF models, fourth- and fifth-order N-grams provided additional statistical discrimination and cross-species variability. Interestingly, variable importance varied significantly among the five species, demonstrating that N-gram frequency patterns were both unique and conserved.

3.4. Classification performance is a positively correlated with N-gram order and segment length

Fig. 3A provided the AUCs grouped by N-gram order and as a function of the DNA segment length for unknown human sequence segments versus the other four species. Fig. 3C provided a table containing a summary of the number of tests performed for each binary classification experiment. In summary, even low-order N-gram analysis performed well in binary classification mode, with second-order analysis averaging AUCs of 0.851, 0.916, 0.919, 0.949, and 0.951 for 10, 20, 40, 80, and 100 kb DNA segments, respectively. Fifth-order analysis averaged AUCs of 0.973, 0.993, 0.993, 0.999, and 1, respectively. In binary mode, all 200 test sequences were correctly classified in the human genome compared to the four other mammalian genomes. For reference, 100 kb is approximately 1/33,000, or 0.00003 %, of the size of the human genome. Fig. 4 provided the AUCs grouped by the N-gram order and as a function of the segment length used to classify the five species. The contingency tables for each multi-class experiment are provided in Supplemental Table 3. Fifth-order classification AUCs ranged from 0.85 to 0.979 and increased with the DNA segment length.

3.5. Computational performance benchmarking

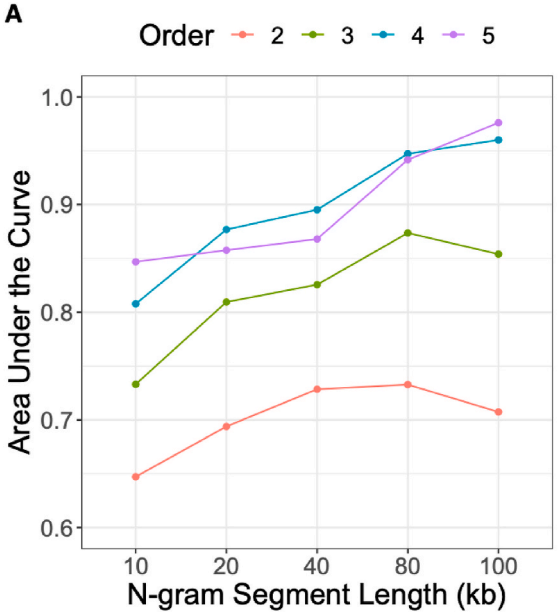
Table 1 summarized the computational performance benchmarks for fifth-order N-gram analysis. Fifth-order N-grams were used to provide an approximation of this model's computational performance using a total of 20 Mb of DNA sequence for each benchmark. Three functions were required to perform supervised classification. They were *getSequence()*, *getFreqDF()*, and *binaryClassification()*.



Segment Length (kb)	Number of Validation Tests	Range of AUCs
10	2,000	0.948 – 0.985
20	1,000	0.985 – 1
40	500	0.986 – 0.996
80	250	0.996 – 1
100	200	1

Fig. 3. Supervised binary classification using Support Vector Machines.
Panel A provided the area under the receiver operating characteristic curve, or AUC, as a function of DNA N-gram segment length and grouped by the N-gram order. Each plot in Panel A was scaled independently. Model performance was assessed at each DNA sequence segment length (10, 20, 40, 80, and 100 kb) and for each order (2, 3, 4, 5). This experiment was repeated for humans versus the four other species, bat (HB), elephant (HE), koala (HK), and dolphin (HD), as depicted in Panel B. Each training and validation set consisted of 20 MB of sequence. Panel C provided the number of validation experiments performed and the range of calculated AUCs at each DNA segment length for fifth-order N-gram analysis.

A



B

Segment Length (kb)	Number of Tests	Range of AUCs
10	2,000	0.647 – 0.847
20	1,000	0.694 – 0.877
40	500	0.729 – 0.895
80	250	0.733 – 0.947
100	200	0.707 – 0.976

Fig. 4. Supervised multi-class classification using the Random Forest methodology. Panel A provided the area under the receiver operating characteristic curve, or AUC, as a function of the DNA N-gram segment length and was colored by the N-gram order. Model performance was assessed at each DNA sequence segment length (10, 20, 40, 80, and 100 kB) and for each order (2, 3, 4, 5). Panel B provided the number of validation experiments performed and the AUC range at each DNA segment length.

The *getSequence()* function randomly selected and constructed the contiguous DNA segments. On average, this function required 19 s to process 20 MB of DNA using 4.7 GB of memory. The *getFreqDF()* function calculated the N-gram frequency tables and required between 30 and 70 s to complete while using 12–32 GB of memory. This function's runtime and memory increased with the number of tests, as expected with a larger matrix. Finally, the *binaryClassification()* function performed SVM classification and required 1–44 s to process using less than 1.4 GB of memory. The average combined processing time for a fifth-order binary classification experiment was only 90 s, or 12 s per processing core.

Table 1
Computational performance benchmarks for fifth-order N-gram analysis.

Function	Segment Length (kB)	Total Runtime (seconds)	Core Seconds	Maximum Memory (GB)	Iterations per second
getSequence	10	19.2	2.40	4.75	0.052
getSequence	20	18.7	2.34	4.77	0.054
getSequence	40	18.8	2.35	4.74	0.053
getSequence	80	18.5	2.31	4.77	0.054
getSequence	100	18.9	2.36	4.76	0.053
getFreqDF	10	69.6	8.70	32.3	0.014
getFreqDF	20	45.4	5.68	16.7	0.022
getFreqDF	40	34.4	4.30	12.7	0.029
getFreqDF	80	32.6	4.08	11.7	0.031
getFreqDF	100	30.8	3.85	11.5	0.033
binaryClassification	10	44.2	5.53	1.4	0.022
binaryClassification	20	10.4	1.30	0.74	0.096
binaryClassification	40	2.57	0.32	0.39	0.389
binaryClassification	80	0.89	0.11	0.22	0.112
binaryClassification	100	0.71	0.09	0.19	0.14

4. Discussion

DNAnamer provides an efficient, accurate, and extensible methodology and statistical analysis toolkit designed to tackle a broad range of complex and computationally expensive classification problems applicable to biomedical sciences, genomics, and genetics. This methodology achieves high levels of classification accuracy using minimal information. This innovative toolkit for stochastic modeling and machine learning has been validated, documented, and is available as a complete and well-tested R package.

To establish theoretical feasibility, a number of fundamental methodological assumptions and hypotheses have been addressed. Specifically, it has been demonstrated that DNA N-gram patterns are highly conserved and exhibit distinct and identifiable signatures, making them excellent for classification. Next, a series of carefully crafted DNA barcoding experiments were designed and executed to offer a concrete example and application for this innovative approach. In completing these classification experiments, it was demonstrated that model performance consistently increased with N-gram order and DNA segment length. Thus, it would seem logical to conclude that classification accuracy could improve by applying higher-order N-gram frequency analysis.

Like all models, this one has both strengths and limitations. One such limitation of this toolkit is that it used 32 GB of memory to calculate the largest N-gram frequency table. This isn't ideal, but it can easily run on a modern laptop computer. The computational performance may be improved by utilizing parallel processing and data reduction techniques. Furthermore, this algorithm's detection limit may be improved by the application of more advanced model optimization and parameter tuning. Another way to improve this model's classification performance will be to perform pre-analysis quality control on the training DNA sequence sets to ensure higher quality input. For example, training models with Sanger-validated sequences should increase classification performance. The small sample size of the species used in this DNA barcoding demonstration was another limitation of this study. Although hundreds of millions of DNA sequences were efficiently analyzed and accurately classified, it will be interesting to see how this approach performs with organisms that exhibit higher levels of DNA sequence variability and toxonomic distance. Because the five species analyzed in this study have relatively high sequence homology, more genetically divergent species may exhibit even more distinct and identifiable N-gram frequency patterns. Among this model's main strengths are efficiency, accuracy, and generality. Efficiency allows for data reduction applications and the analysis of higher-order N-grams. Most importantly, the generality of this methodology makes it ideal for a wide array of problems that we face in the ever-growing and dynamic field of genomics. In an era of 'big data', it is critical to remember that more data does not necessarily imply more meaning.

Broader application and integration were the driving forces behind designing, building, and validating this methodology and analysis toolkit. As such, I will conclude with four specific and promising applications for DNAnamer. First and foremost, this approach is well-suited for classifying functional genetic elements. As previously outlined, N-gram analysis methodologies coupled with machine learning approaches have been highly successful in identifying functional gene elements such as retrotransposons, gene promotor regions, and CpG genomic islands. DNAnamer can be used with modern genotyped variant call sets to discover functional regulatory gene elements in existing and future genetic epidemiological studies. Expanding on these approaches for identifying novel functional and regulatory elements in the human genome holds great potential to increase our knowledge of population genetics, evolutionary genetics, and could improve our understanding of human disease. A second common and practical application for this methodology is the identification of cross-species DNA contamination. Bacterial DNA contamination occurs in DNA extraction kits, PCR (polymerase chain reaction) reagents, and model organisms. This methodology will be applied to detect bacterial and other DNA contaminants, along with sequencing artifacts caused by non-biological processes. The construction of a methodology and analysis toolkit for identifying bacterial and other DNA sequencing artifacts is underway. A third application for DNAnamer is identifying microbial species, such as lichens and fungi. Because lichens must be chemically processed and sequenced for identification, this approach can be used to improve species detection accuracy and speed. Finally, by using supervised learning, this approach allows for the generation of more biologically accurate *in silico* DNA sequences that have been instrumental to improving *in silico* spike-in sensitivity analysis methods. Current *in silico* spike-in sensitivity analysis methodologies assign randomly generated sequences [57]. We will leaverage generative modeling to create spike-in sequences based on biologically derived priors to more rigorously test the detection limits and assess the performance of structural variant and copy number variant calling algorithms. *In silico* spike-in is extremely useful for estimating the sensitivity of a range of variant callers. In conclusion, there are many useful and innovative applications for DNAnamer, both as a tool for data analysis and scientific discovery.

Data availability

All data used to conduct this study can easily be downloaded from the National Center for Biotechnology Information's genome resource database. The software, documentation, and vignettes with detailed code demonstrations are available at https://github.com/jmal0403/DNAnamer/wiki.

Funding

No funding was used in this project.

CRediT authorship contribution statement

John S. Malamon: Writing – original draft, Visualization, Validation, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: John Stephen Malamon reports administrative support was provided by University of Colorado Anschutz Medical Campus School of Medicine. John Stephen Malamon reports a relationship with University of Colorado Anschutz Medical Campus School of Medicine that includes: employment. John Stephen Malamon has patent pending to N/A. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The author has no acknowledgments.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e36914.

References

[1] C.E. Shannon, A mathematical theory of communication, The Bell System Technical Journal 27 (1945). https://ieeexplore.ieee.org/servlet/opac?bknumber=5271069.
[2] Siddique AB Combinatorial Entropy Encoding. ArXiv. 2017;abs/1703.08127.
[3] An algorithm for entropy coding: combinatorial coding, in: S. Bärwolf (Ed.), 2014 World Symposium on Computer Applications & Research (WSCAR), Jan. 2014, 2014 18-20.
[4] Binary combinatorial coding, in: D. Vito, A. Zakhor (Eds.), Data Compression Conference, 2003 Proceedings DCC 2003, 2003, 2003 25-27 March.
[5] C.E. Shannon, A mathematical theory of cryptography, The Bell System Technical Journal 27 (1948), https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.
[6] C.E. Shannon, Communication theory of secrecy systems, The Bell System Technical Journal 28 (4) (1949) 656–715, https://doi.org/10.1002/j.1538-7305.1949.tb00928.x.
[7] I. Ben-Gal, E. Kagan, Information theory: deep ideas, wide perspectives, and various applications, Entropy 23 (2) (2021), https://doi.org/10.3390/e23020232. PubMed PMID: 33671301; PubMed Central PMCID: PMCPMC7922818.
[8] O. Kaynak, The golden age of Artificial Intelligence, Discover Artificial Intelligence 1 (1) (2021) 1, https://doi.org/10.1007/s44163-021-00009-x.
[9] J.M. Zook, N.F. Hansen, N.D. Olson, L. Chapman, J.C. Mullikin, C. Xiao, et al., A robust benchmark for detection of germline large deletions and insertions, Nat. Biotechnol. 38 (11) (2020) 1347–1355, https://doi.org/10.1038/s41587-020-0538-8. PubMed PMID: 32541955; PubMed Central PMCID: PMCPMC8454654.
[10] T.W. Dunlop, R.W. Davies, Conservation of CAG/CTG trinucleotide repeats in developmentally expressed mammalian genes, Mamm. Genome 12 (6) (2001) 475–477, https://doi.org/10.1007/s003350010290. PubMed PMID: 11353398.
[11] L. Fedorova, E.R. Crossley, O.A. Mulyar, S. Qiu, R. Freeman, A. Fedorov, Profound non-randomness in dinucleotide arrangements within ultra-conserved non-coding elements and the human genome, Biology 12 (8) (2023), https://doi.org/10.3390/biology12081125. PubMed PMID: 37627009; PubMed Central PMCID: PMCPMC10452674.
[12] R.L. Stallings, Distribution of trinucleotide microsatellites in different categories of mammalian genomic sequence: implications for human genetic diseases, Genomics 21 (1) (1994) 116–121, https://doi.org/10.1006/geno.1994.1232. PubMed PMID: 8088779.
[13] J. Violos, K. Tserpes, I. Varlamis, T. Varvarigou, Text classification using the N-gram graph representation model over high frequency data streams, Frontiers in Applied Mathematics and Statistics 4 (2018), https://doi.org/10.3389/fams.2018.00041.
[14] M. Jankowska, E.E. Milios, V. Keselj (Eds.), Author Verification Using Common N-Gram Profiles of Text Documents, International Conference on Computational Linguistics, 2014.
[15] T.S. Hugo Jair Escalante, Manuel Montes-y-Gómez, Local Histograms of Character N-Grams for Authorship Attribution, Association for Computational Linguistics, 2011, pp. 288–298.
[16] Z. Volkovich, V. Kirzhner, A. Bolshoy, E. Nevo, A. Korol, The method of N-grams in large-scale clustering of DNA texts, Pattern Recogn. 38 (11) (2005) 1902–1912, https://doi.org/10.1016/j.patcog.2005.05.002.
[17] Koppel M, Schler J, Zigdon K, editors. Automatically Determining an Anonymous Author's Native Language2005; Berlin, Heidelberg: Springer Berlin Heidelberg.
[18] Cavnar WB, Trenkle JM. N-Gram-Based text categorization. Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval; Las Vegas, US. cavnar1994ngrambased1994. p. 161–175.
[19] Kuta M, Kitowski J, editors. Optimisation of Character N-Gram Profiles Method for Intrinsic Plagiarism Detection2014; Cham: Springer International Publishing.
[20] Stamatatos E, editor Intrinsic Plagiarism Detection Using Character N-Gram Profiles 2009.
[21] R.M. Arabyat, B.R. Qawasmeh, S.I. Al-Azzam, M.B. Nusair, K.H. Alzoubi, Faculty members' perceptions and attitudes towards anti-plagiarism detection tools: applying the theory of planned behavior, J Empir Res Hum Res Ethics 17 (3) (2022) 275–283, https://doi.org/10.1177/15562646221078655. PubMed PMID: 35188816; PubMed Central PMCID: PMCPMC9992686.
[22] E. Delibas, A. Arslan, A. Seker, B. Diri, A novel alignment-free DNA sequence similarity analysis approach based on top-k n-gram match-up, J. Mol. Graph. Model. 100 (2020) 107693, https://doi.org/10.1016/j.jmgm.2020.107693. PubMed PMID: 32805559.
[23] M. Ganapathiraju, V. Manoharan, J. Klein-Seetharaman, BLMT: statistical sequence analysis using N-grams, Appl. Bioinf. 3 (2–3) (2004) 193–200, https://doi.org/10.2165/00822942-200403020-00013. PubMed PMID: 15693744.
[24] A. Tomovic, P. Janicic, V. Keselj, n-gram-based classification and unsupervised hierarchical clustering of genome sequences, Comput. Methods Progr. Biomed. 81 (2) (2006) 137–153, https://doi.org/10.1016/j.cmpb.2005.11.007. PubMed PMID: 16423423.
[25] H.U. Osmanbeyoglu, M.K. Ganapathiraju, N-gram analysis of 970 microbial organisms reveals presence of biological language models, BMC Bioinf. 12 (1) (2011) 12, https://doi.org/10.1186/1471-2105-12-12.
[26] S. Pandey, N. Avuthu, C. Guda, StrainIQ: a novel n-gram-based method for taxonomic profiling of human microbiota at the strain level, Genes 14 (8) (2023), https://doi.org/10.3390/genes14081647. PubMed PMID: 37628698; PubMed Central PMCID: PMCPMC10454763.
[27] B.R. King, C. Guda, ngLOC: an n-gram-based Bayesian method for estimating the subcellular proteomes of eukaryotes, Genome Biol. 8 (5) (2007) R68, https://doi.org/10.1186/gb-2007-8-5-r68.

[28] N.Q.K. Le, E.K.Y. Yapp, N. Nagasundaram, H.Y. Yeh, Classifying promoters by interpreting the hidden information of DNA sequences via deep learning and combination of continuous FastText N-grams, Front. Bioeng. Biotechnol. 7 (2019) 305, https://doi.org/10.3389/fbioe.2019.00305. PubMed PMID: 31750297; PubMed Central PMCID: PMCPMC6848157.
[29] S.M.A. Islam, B.J. Heil, C.M. Kearney, E.J. Baker, Protein classification using modified n-grams and skip-grams, Bioinformatics 34 (9) (2018) 1481–1487, https://doi.org/10.1093/bioinformatics/btx823. PubMed PMID: 29309523.
[30] A.K. Sharma, R. Srivastava, Variable length character N-gram embedding of protein sequences for secondary structure prediction, Protein Pept. Lett. 28 (5) (2021) 501–507, https://doi.org/10.2174/0929866527666201103145635. PubMed PMID: 33143605.
[31] J.K. Vries, X. Liu, I. Bahar, The relationship between n-gram patterns and protein secondary structure, Proteins 68 (4) (2007) 830–838, https://doi.org/10.1002/prot.21480. PubMed PMID: 17523186.
[32] L.J. Kelly, P.M. Hollingsworth, B.J. Coppins, C.J. Ellis, P. Harrold, J. Tosh, et al., DNA barcoding of lichenized fungi demonstrates high identification success in a floristic context, New Phytol. 191 (1) (2011) 288–300, https://doi.org/10.1111/j.1469-8137.2011.03677.x. PubMed PMID: 21434928.
[33] M. Kerr, S.D. Leavitt, A custom regional DNA barcode reference library for lichen-forming fungi of the intermountain west, USA, increases successful specimen identification, J Fungi (Basel) 9 (7) (2023), https://doi.org/10.3390/jof9070741. PubMed PMID: 37504730; PubMed Central PMCID: PMCPMC10381598.
[34] R.D. La Torre, D. Ramos, M.D. Mejia, E. Neyra, E. Loarte, G. Orjeda, Survey of lichenized fungi DNA barcodes on king george island (Antarctica): an aid to species discovery, J Fungi (Basel) 9 (5) (2023), https://doi.org/10.3390/jof9050552. PubMed PMID: 37233263; PubMed Central PMCID: PMCPMC10219471.
[35] H.G. Drost, J. Paszkowski, Biomartr: genomic data retrieval with R, Bioinformatics 33 (8) (2017) 1216–1217, https://doi.org/10.1093/bioinformatics/btw821. PubMed PMID: 28110292; PubMed Central PMCID: PMCPMC5408848.
[36] D. Charif, J.R. Lobry, SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis, in: U. Bastolla, M. Porto, H.E. Roman, M. Vendruscolo (Eds.), Structural Approaches to Sequence Evolution: Molecules, Networks, Populations, Springer Berlin Heidelberg, Berlin, Heidelberg, 2007, pp. 207–232.
[37] Girden ER. Anova, Repeated measures, Sage 84 (1992).
[38] A. Wald, J. Wolfowitz, On a test whether two samples are from the same population, Ann. Math. Stat. 11 (2) (1940) 147–162, https://doi.org/10.1214/aoms/1177731909.
[39] R. Bartels, The Rank Version of von Neumann's Ratio Test for Randomness, J. Am. Stat. Assoc. 77 (377) (1982) 40–46, https://doi.org/10.1080/01621459.1982.10477764.
[40] J. von Neumann, Distribution of the ratio of the mean square successive difference to the variance, Ann. Math. Stat. 12 (1941) 367–395, https://doi.org/10.1214/aoms/1177731677.
[41] X.A. Bi, Y. Wang, Q. Shu, Q. Sun, Q. Xu, Classification of autism spectrum disorder using random support vector machine cluster, Front. Genet. 9 (2018) 18, https://doi.org/10.3389/fgene.2018.00018. PubMed PMID: 29467790; PubMed Central PMCID: PMCPMC5808191.
[42] S. Huang, N. Cai, P.P. Pacheco, S. Narrandes, Y. Wang, W. Xu, Applications of support vector machine (SVM) learning in cancer genomics, Cancer Genomics Proteomics 15 (1) (2018) 41–51, https://doi.org/10.21873/cgp.20063. PubMed PMID: 29275361; PubMed Central PMCID: PMCPMC5822181.
[43] J. Li, Z. Weng, H. Xu, Z. Zhang, H. Miao, W. Chen, et al., Support Vector Machines (SVM) classification of prostate cancer Gleason score in central gland using multiparametric magnetic resonance images: a cross-validated study, Eur. J. Radiol. 98 (2018) 61–67, https://doi.org/10.1016/j.ejrad.2017.11.001. PubMed PMID: 29279171.
[44] E. Reynolds, B. Callaghan, M. Banerjee, SVM-CART for disease classification, J. Appl. Stat. 46 (16) (2019) 2987–3007, https://doi.org/10.1080/02664763.2019.1625876. PubMed PMID: 33012942; PubMed Central PMCID: PMCPMC7531767.
[45] T.K. Seo, Classification of nucleotide sequences using support vector machines, J. Mol. Evol. 71 (4) (2010) 250–267, https://doi.org/10.1007/s00239-010-9380-9. PubMed PMID: 20740280.
[46] CaV. Cortes, Vladimir. Support-vector networks, Mach. Learn. 20 (1995) 273–297.
[47] AaW. Liaw, Matthew. Classification and regression by randomForest, R. News 2 (2002) 18–22.
[48] R. Diaz-Uriarte, S. Alvarez de Andres, Gene selection and classification of microarray data using random forest, BMC Bioinf. 7 (2006) 3, https://doi.org/10.1186/1471-2105-7-3. PubMed PMID: 16398926; PubMed Central PMCID: PMCPMC1363357.
[49] B.A. Goldstein, E.C. Polley, F.B. Briggs, Random forests for genetic association studies, Stat. Appl. Genet. Mol. Biol. 10 (1) (2011) 32, https://doi.org/10.2202/1544-6115.1691. PubMed PMID: 22889876; PubMed Central PMCID: PMCPMC3154091.
[50] E. Pellegrino, C. Jacques, N. Beaufils, I. Nanni, A. Carlioz, P. Metellus, et al., Machine learning random forest for predicting oncosomatic variant NGS analysis, Sci. Rep. 11 (1) (2021) 21820, https://doi.org/10.1038/s41598-021-01253-y. PubMed PMID: 34750410; PubMed Central PMCID: PMCPMC8575902.
[51] R. Toth, H. Schiffmann, C. Hube-Magg, F. Buscheck, D. Hoflmayer, S. Weidemann, et al., Random forest-based modelling to detect biomarkers for prostate cancer progression, Clin. Epigenet. 11 (1) (2019) 148, https://doi.org/10.1186/s13148-019-0736-8. PubMed PMID: 31640781; PubMed Central PMCID: PMCPMC6805338.
[52] P.K. Meher, T.K. Sahu, S. Gahoi, R. Tomar, A.R. Rao, funbarRF: DNA barcode-based fungal species prediction using multiclass Random Forest supervised learning model, BMC Genet. 20 (1) (2019) 2, https://doi.org/10.1186/s12863-018-0710-z. PubMed PMID: 30616524; PubMed Central PMCID: PMCPMC6323839.
[53] P.K. Meher, T.K. Sahu, A.R. Rao, Identification of species based on DNA barcode using k-mer feature vector and Random forest classifier, Gene 592 (2) (2016) 316–324, https://doi.org/10.1016/j.gene.2016.07.010. PubMed PMID: 27393648.
[54] L.S. Riza, M.I. Zain, A. Izzuddin, Y. Prasetyo, T. Hidayat, K.A.F. Abu Samah, Implementation of machine learning in DNA barcoding for determining the plant family taxonomy, Heliyon 9 (10) (2023) e20161, https://doi.org/10.1016/j.heliyon.2023.e20161. PubMed PMID: 37767518; PubMed Central PMCID: PMCPMC10520734.
[55] T.M. Oshiro, P.S. Perez, J.A. Baranauskas (Eds.), How Many Trees in a Random Forest?, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.
[56] R.C.R. Team, A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, 2020, Version 4.0. 2.
[57] J.S. Malamon, J.J. Farrell, L.C. Xia, B.A. Dombroski, R.G. Das, J. Way, et al., A comparative study of structural variant calling in WGS from Alzheimer's disease families, Life Sci. Alliance 7 (5) (2024), https://doi.org/10.26508/lsa.202302181. PubMed PMID: 38418088; PubMed Central PMCID: PMCPMC10902710.