Western Graduate&PostdoctoralStudies

Western University
Scholarship@Western

Electronic Thesis and Dissertation Repository

6-1-2020 10:00 AM

Machine Learning with Digital Signal Processing for Rapid and Accurate Alignment-Free Genome Analysis: From Methodological Design to a Covid-19 Case Study

Gurjit Singh Randhawa, *The University of Western Ontario*

Supervisor: Kari, Lila, *The University of Waterloo*

A thesis submitted in partial fulfillment of the requirements for the Doctor of Philosophy degree in Computer Science

© Gurjit Singh Randhawa 2020

Follow this and additional works at: https://ir.lib.uwo.ca/etd

Part of the Artificial Intelligence and Robotics Commons, Bioinformatics Commons, Computational Biology Commons, and the Genomics Commons

Recommended Citation

Randhawa, Gurjit Singh, "Machine Learning with Digital Signal Processing for Rapid and Accurate Alignment-Free Genome Analysis: From Methodological Design to a Covid-19 Case Study" (2020). *Electronic Thesis and Dissertation Repository*. 7007.
https://ir.lib.uwo.ca/etd/7007

This Dissertation/Thesis is brought to you for free and open access by Scholarship@Western. It has been accepted for inclusion in Electronic Thesis and Dissertation Repository by an authorized administrator of Scholarship@Western. For more information, please contact wlswadmin@uwo.ca.

Abstract

In the field of bioinformatics, taxonomic classification is the scientific practice of identifying, naming, and grouping of organisms based on their similarities and differences. The problem of taxonomic classification is of immense importance considering that nearly 86% of existing species on Earth and 91% of marine species remain unclassified. Due to the magnitude of the datasets, the need exists for an approach and software tool that is scalable enough to handle large datasets and can be used for rapid sequence comparison and analysis. We propose ML-DSP, a stand-alone alignment-free software tool that uses Machine Learning and Digital Signal Processing to classify genomic sequences. ML-DSP uses numerical representations to map genomic sequences to discrete numerical series (genomic signals), Discrete Fourier Transform (DFT) to obtain magnitude spectra from the genomic signals, Pearson Correlation Coefficient (PCC) as a dissimilarity measure to compute pairwise distances between magnitude spectra of any two genomic signals, and supervised machine learning for the classification and prediction of the labels of new sequences. We first test ML-DSP by classifying 7396 full mitochondrial genomes at various taxonomic levels, from kingdom to genus, with an average classification accuracy of $> 97\%$. We also provide preliminary experiments indicating the potential of ML-DSP to be used for other datasets, by classifying 4271 complete dengue virus genomes into subtypes with 100% accuracy, and 4710 bacterial genomes into phyla with 95.5% accuracy. Second, we propose another tool, MLDSP-GUI, where additional features include: a user-friendly Graphical User Interface, Chaos Game Representation (CGR) to numerically represent DNA sequences, Euclidean and Manhattan distances as additional distance measures, phylogenetic tree output, oligomer frequency information to study the under- and over-representation of any particular sub-sequence in a selected sequence, and inter-cluster distances analysis, among others. We test MLDSP-GUI by classifying 7881 complete genomes of *Flavivirus* genus into species with 100% classification accuracy. Third, we provide a proof of principle that MLDSP-GUI is able to classify newly discovered organisms by classifying the novel COVID-19 virus.

Keywords: Machine Learning, Digital Signal Processing, Discrete Fourier Transform, taxonomic classification, whole genome analysis, genomic signature, Chaos Game Representation, alignment-free sequence analysis

Summary

Sequence classification is the scientific practice of identifying, naming, and grouping organisms based on their differences and similarities. Considering that most of the existing species (nearly 86% of species on Earth and 91% of marine species) remain unclassified, the problem of sequence classification is of immense importance. Due to the magnitude of the datasets, the problem of sequence comparison and analysis for the purpose of classification remains challenging. Sequence (dis)similarity analysis has multiple possible applications including taxonomic classification (classify organisms on the basis of shared characteristics), virus-subtype classification (assign viral sequences to their subtypes), disease classification (classify human genomic sequences on the basis of disease type), human haplogroup classification (assign human mitochondrial on the basis of maternal lineage), etc. The need exists for an approach and software tool that is scalable enough to handle large datasets and is able to provide accurate classifications within a short time period. We propose a machine learning-based methodology, ML-DSP, that is effective in the classification of newly discovered organisms, in distinguishing genomic signatures and identifying their mechanistic determinants, and in evaluating genome integrity. We also propose MLDSP-GUI, an extension of ML-DSP with multiple additional valuable features. Lastly, we show the applicability of our approach to taxonomy classification, virus-subtype classification and provide a proof of principle that our approach is able to classify newly discovered organisms by classifying the previously unclassified novel coronavirus (COVID-19 virus) sequences.

Co-Authorship Statement

This thesis consists of three published articles. The article in Chapter 3 was published in the journal *BMC Genomics*, while the article in Chapter 4 was published in the journal *Bioinformatics*, and the article in Chapter 5 in the journal *PLOS ONE*. The papers in Chapters 3, 4 and 5 have two senior authors (K.A.H, L.K.). The major individual contributions are listed below.

Chapter 3 contains the article "ML-DSP: Machine Learning with Digital Signal Processing for ultrafast, accurate, and scalable genome classification at all taxonomic levels" by Gurjit S. Randhawa, Kathleen A. Hill, and Lila Kari. The individual contributions are as follows. G.S.R. and L.K. conceived the study and wrote the manuscript. G.S.R. designed and tested the software. G.S.R., L.K. and K.A.H. conducted the data analysis and edited the manuscript, with K.A.H. contributing biological expertize.

Chapter 4 contains the article "MLDSP-GUI: an alignment-free standalone tool with an interactive graphical user interface for DNA sequence comparison and analysis" by Gurjit S. Randhawa, Kathleen A. Hill, and Lila Kari. The individual contributions are as follows. G.S.R. and L.K. conceived the study and wrote the manuscript. G.S.R. designed and tested the software. G.S.R., L.K. and K.A.H. conducted the data analysis and edited the manuscript, with K.A.H. contributing biological expertize.

Chapter 5 contains the article "Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: COVID-19 case study" by Gurjit S. Randhawa, Maximillian P.M. Soltysiak, Hadi El Roz, Camila P.E. de Souza, Kathleen A. Hill, and Lila Kari. The individual contributions are as follows: Conceptualization, G.S.R.; methodology, G.S.R. and L.K.; software, G.S.R. and L.K.; validation, G.S.R.; formal analysis, G.S.R., M.P.M.S, H.E., C.P.E.d.S. and K.A.H.; investigation, G.S.R., M.P.M.S and K.A.H.; resources, G.S.R.; data curation, G.S.R.; writing—original draft preparation, G.S.R and M.P.M.S writing—review and editing, G.S.R., M.P.M.S., H.E., C.P.E.d.S., K.A.H. and L.K.; visualization, G.S.R.; supervision, K.A.H and L.K.; project administration, K.A.H. and L.K.; funding acquisition, K.A.H. and L.K.

Acknowlegements

First and foremost, I would like to express my gratitude to my supervisor, Dr. Lila Kari for her guidance and encouragement. Her enthusiasm, independent thought, meticulousness, and dedication are an inspiration I hope to learn from and emulate in my career. I am very thankful for all her tremendous support in all aspects of my professional development.

I would also like to thank Dr. Kathleen A. Hill for being an extraordinary mentor. She introduced me, and ignited my interest in knowledge areas I was less familiar with than I care to admit. She has been a great collaborator and has helped me a lot in expanding the outreach of my research, and introduced me to the immensely valuable research community.

I offer special thanks to my friend and colleague Maximillian P.M. Soltysiak. We had to set challenging deadlines and tirelessly work to honor them for the time-sensitive COVID-19 case study. I must also thank my current and former team-mates: Zihao Wang, Pablo Millan Arias, Fatemeh Alipour, Dr. Rallis Karamichalis, and Stephen Solis-Reyes for fostering a constructive research environment. I also want to acknowledge the research group working out of Dr. Hill's lab, with whom I have had the pleasure of collaborating with on multiple projects.

Lastly, I would like to thank my family and friends who have been a bedrock of support, all through my life. I am what I am because of how my parents have influenced me. Most of all, I would like to thank my wife Taran for her love, support, and understanding.

Contents

Abstract **ii**

Summary **iv**

Co-Authorship Statement **v**

Acknowlegements **vi**

List of Figures **xi**

List of Tables **xviii**

List of Appendices **xx**

1 Introduction **1**

2 Literature review **6**

- 2.1 Biological background 6
- 2.2 Genomic sequence analysis methods 8
  - 2.2.1 Alignment-based methods 8
  - 2.2.2 Alignment-free methods 10
- 2.3 Our approach 14
  - 2.3.1 DNA numerical representations 14
  - 2.3.2 Discrete Fourier Transform 19
  - 2.3.3 Distance measures 19

2.3.4 Multi-dimensional scaling 20
2.3.5 Supervised learning classification models 22

3 ML-DSP: Machine Learning with Digital Signal Processing 38
3.1 Background 38
3.1.1 Numerical representations of DNA sequences 40
3.1.2 Digital Signal Processing 41
3.1.3 Supervised Machine Learning 42
3.2 Methods and Implementation 43
3.2.1 DNA numerical representations 44
3.2.2 Discrete Fourier Transform (DFT) 44
3.2.3 Pearson Correlation Coefficient (PCC) 46
3.2.4 Supervised Machine Learning 49
3.2.5 Classical Multidimensional Scaling (MDS) 50
3.2.6 Software implementation 50
3.2.7 Datasets 51
3.3 Results and Discussion 51
3.3.1 Analysis of distances and of length normalization approaches 52
3.3.2 Analysis of various numerical representations of DNA sequences 52
3.3.3 ML-DSP for three classes of vertebrates 55
3.3.4 Classifying genomes with ML-DSP, at all taxonomic levels 55
3.3.5 MoDMap visualization vs. ML-DSP quantitative classification results . 59
3.3.6 Applications to other genomic datasets 61
3.3.7 Comparison of ML-DSP with state-of-the-art alignment-based and alignment-free tools 61
3.3.8 Discussion 65
3.4 Conclusions 68
3.5 Availability and Requirements 69

4 MLDSP-GUI: ML-DSP with Graphical User Interface 77

4.1 Introduction 77

4.2 Materials and methods 78

4.3 Software description 79

5 COVID-19 case study 82

5.1 Introduction 82

5.2 Materials and methods 86

5.3 Results 88

5.4 Discussion 101

5.5 Conclusion 109

6 Conclusion 122

A Copyright Releases 124

B Software description 126

C MLDSP-GUI: Supplementary Material 129

C.A Interactive MLDSP-GUI features 129

C.A.1 Left panel 135

C.A.2 Center panel 138

C.A.3 Right panel 142

C.B Provided datasets 145

C.C Availability 145

D COVID-19 case study: Supplementary Material 148

D.A Software availability 148

D.B Spearman's rank correlation coefficient test results 149

D.C Dataset availability 149

E Addendum **174**

Curriculum Vitae **179**

List of Figures

2.1 Flowchart showing MLDSP methodology. 15

2.2 The Chaos Game Representation (CGR) of the DNA sequence CGGTAT. . . . 17

2.3 The Chaos Game Representation (CGR) of the mtDNA sequence of (a) Canadian Beaver (*Castor canadensis*), NCBI accession NC_007011.1, 16767 bp length and (b) Canada goose (*Branta canadensis*), NCBI accession KY311838.1, 16760 bp length. 18

2.4 A MoDMap generated by applying multi-dimensional scaling to the pairwise distances between six most populated Canadian cities. 21

3.1 Canada goose (blue) vs European beaver (red): comparison of the DFT magnitude spectra of the first 100 bp of their mtDNA genomes. (**a**): Graphical illustration of the discrete digital signals of the respective DNA sequences, obtained using the "PP" representation. (**b**): DFT magnitude spectra of the signals in (**a**). 47

3.2 Canada goose (blue, 16,760 bp) vs. European beaver (red, 16,722 bp) - comparison between the DFT phase spectra of their full mtDNA genomes. 47

3.3 MoDMap of 3,740 full mtDNA genomes in subphylum Vertebrata, into three classes: Birds (blue, Aves: 553 genomes), fish (red, Actinopterygii 2,176 genomes, Chondrichthyes 130 genomes, Coelacanthiformes 2 genomes, Dipnoi 5 genomes), and mammals (green, Mammalia: 874 genomes). The accuracy of the ML-DSP classification into three classes, using the Quadratic SVM classifier, with the "PP" numerical representation, and PCC between magnitude spectra of DFT, was 100%. 57

3.4 MoDMap of family Cyprinidae and its genera. (**a**): Genera *Acheilognathus* (blue, 10 genomes), *Rhodeus* (red, 11 genomes), *Schizothorax* (green, 19 genomes), *Labeo* (black, 19 genomes), *Acrossocheilus* (magenta, 12 genomes), *Onychostoma* (yellow, 10 genomes); (**b**): Genera *Acheilognathus* and *Rhodeus*, which overlapped in (**a**), are visually separated when plotted separately in (**b**). The classification accuracy with Quadratic SVM of the dataset in (**a**) was 91.8%, and of the dataset in (**b**) was 100%. 60

3.5 MoDMap of the superorder Ostariophysi, and the confusion matrix for the Quadratic SVM classification of this superorder into orders. (**a**): MoDMap of orders Cypriniformes (blue, 643 genomes), Characiformes (red, 31 genomes), Siluriformes (green, 107 genomes). (**b**): The confusion matrix generated by Quadratic SVM, illustrating its true class vs. predicted class performance (top-to-bottom and left-to-right: Cypriniformes, Characiformes, Siluriformes). The numbers in the squares on the top-left to bottom-right diagonal (blue) indicate the numbers of correctly classified DNA sequences, by order. The off-diagonal pink squares indicate that 6 mtDNA genomes of the order Characiformes have been erroneously predicted to belong to the order Cypriniformes (center-left), and 2 mtDNA genomes of the order Siluriformes have been erroneously predicted to belong to the order Cypriniformes (bottom-left). The Quadratic SVM that generated this confusion matrix had a 99% classification accuracy. 62

3.6 (**a**) MoDMap of 4,271 dengue virus genomes. The colours represent virus subtypes DENV-1 (blue, $2,008$ genomes), DENV-2 (red, $1,349$ genomes), DENV-3 (green, $1,010$ genomes), DENV-4 (black, 354 genomes); The classification accuracy of the Quadratic SVM classifier for this dataset was 100%. (**b**) MoDMap of 4,710 bacterial genomes. The colours represent bacterial phyla: Spirochaetes (blue, 437 genomes), Firmicutes (red, 1,129 genomes), Proteobacteria (green, 3,144 genomes). The accuracy of the Quadratic SVM classifier for this dataset was 95.5%. 63

3.7 MoDMaps of the influenza virus dataset from Table 3.5, based on the four methods. The points represent viral genomes of subtypes H1N1 (red, 13 genomes), H2N2 (black, 3 genomes), H5N1 (blue, 11 genomes), H7N3 (magenta, 5 genomes), H7N9 (green, 6 genomes); ModMaps are generated using distance matrices computed with (**a**) FFP; (**b**) MEGA7(MUSCLE); (**c**) MEGA7(CLUSTALW); (**d**) ML-DSP. 66

3.8 Phylogenetic tree comparison: FFP with ML-DSP. The phylogenetic tree generated for 38 influenza virus genomes using (**a**): FFP (**b**): ML-DSP. 67

3.9 Phylogenetic tree comparison: MEGA7(MUSCLE/CLUSTALW) with ML-DSP. The phylogenetic tree generated for 38 influenza virus genomes using (**a**): MEGA7(MUSCLE/CLUSTALW) (**b**): ML-DSP. 67

4.1 Screenshot of MLDSP-GUI showing a MoDMap3D of 7,881 full mtDNA genomes of the *Flavivirus* genus, classified into species. More details in Supplementary Material. 79

5.1 MoDMap3D of (a) 3273 viral sequences from Test-1 representing 11 viral families and realm Riboviria, (b) 2779 viral sequences from Test-2 classifying 12 viral families of realm Riboviria, (c) 208 *Coronaviridae* sequences from Test-3a classified into genera. 89

5.2 MoDMap3D of (a) 124 *Betacoronavirus* sequences from Test-4 classified into sub-genera, (b) 153 viral sequences from Test-5 classified into 4 sub-genera and COVID-19 virus, (c) 76 viral sequences from Test 6 classified into *Sarbecovirus* and COVID-19 virus. 93

5.3 The UPGMA phylogenetic tree using the Pearson Correlation Coefficient generated pairwise distance matrix shows COVID-19 virus (Red) sequences proximal to the bat *Betacoronavirus RaTG13* (Blue) and bat SARS-like coronaviruses *ZC45/ZXC21* (Green) in a distinct lineage from the rest of *Sarbecovirus* sequences (Black). 95

5.4 The neighbor-joining phylogenetic tree using the Pearson Correlation Coefficient generated pairwise distance matrix shows COVID-19 virus (Red) sequences proximal to the bat *Betacoronavirus RaTG13* (Blue) and bat SARS-like coronaviruses *ZC45/ZXC21* (Green) in a distinct lineage from the rest of *Sarbecovirus* sequences (Black). 96

5.5 Chaos Game Representation (CGR) plots at $k = 7$ of (a) COVID-19 virus / Wuhan seafood market pneumonia virus isolate Wuhan-Hu-1/*MN908947.3*, (b) *Betacoronavirus* / CoV / Bat / Yunnan / *RaTG13* / *EPI_ISL_402131*, (c) *Betacoronavirus* / Bat SARS-like coronavirus isolate *bat-SL-CoVZC45* / *MG772933.1*, (d) *Betacoronavirus* / Bat SARS-like coronavirus isolate *bat-SL-CoVZXC21* / *MG772934.1*, (e) *Alphacoronavirus* /*DQ811787 PRCV ISU−1*, (f) *Gammacoronavirus* / Infectious bronchitis virus NGA /*A116E7* / *2006* / *FN430415*, and (g) *Deltacoronavirus* / PDCoV / USA / Illinois121 /2014/*KJ481931*. Chaos plot vertices are assigned top left Cytosine, top right Guanine, bottom left Adenine and bottom right Thymine. 97

5.6 Hexbin scatterplots of the proportional k-mer ($k = 7$) frequencies of the COVID-19 virus sequences vs. the four genera: (a) *Alphacoronavirus*, $\rho = 0.7$; (b) *Betacoronavirus*, $\rho = 0.74$; (c) *Gammacoronavirus*, $\rho = 0.63$ and (d) *Deltacoronavirus*, $\rho = 0.6$. The color of each hexagonal bin in the plot represents the number of points (in natural logarithm scale) overlapping at that position. All ρ values resulted in p-values $< 10^{-5}$ for the correlation test. By visually inspecting each hexbin scatterplot, the degree of correlation is displayed by the variation in spread between the points. Hexagonal points that are closer together and less dispersed as seen in (b) are more strongly correlated and have less deviation. 99

5.7 Hexbin scatterplots of the proportional k-mer ($k = 7$) frequencies of the COVID-19 virus sequences vs. the four sub-genera: (a) *Embecovirus*, $\rho = 0.59$; (b) *Merbecovirus*, $\rho = 0.64$; (c) *Nobecovirus*, $\rho = 0.54$ and (d) *Sarbecovirus*, $\rho = 0.72$. The color of each hexagonal bin in the plot represents the number of points (in natural logarithm scale) overlapping at that position. All ρ values resulted in p-values $< 10^{-5}$ for the correlation test. By visually inspecting each hexbin scatterplot, the degree of correlation is displayed by the variation in spread between the points. Hexagonal points that are closer together and less dispersed as seen in (d) are more strongly correlated and have less deviation. . . 100

5.8 Time performance of MLDSP-GUI for Test1 to Test-6 (in seconds). 108

C.S1 MLDSP-GUI implements a four-step pipeline for data transformation from genomic sequences to taxonomic classification. 130

C.S2 MLDSP-GUI can be viewed as a combination of 3-vertical panels (Left panel, Center panel, and Right panel). Each panel has multiple sub-panel components. 134

C.S3 Left panel components: Input parameters, progress status, dataset statistics, and logos. 135

C.S4 Chaos Game Representation (CGR) of (a): *Homo sapiens* chromosome 1, first 100,000 bp segment, NCBI accession: *NC*_000001.11 (b): Bacterium (*Intrasporangium flavum*) complete genome, NCBI accession: *MLJO*01000003.1 (c): *Dengue* virus 1 complete genome, NCBI accession: *AB*608789.1 (d): *Pseudomonas* phage *Andromeda* complete genome, NCBI accession: *NC*_031014.1. 137

C.S5 Center panel components: MoDMap3D, selected sequence statistics, inter-cluster distances, and *k*-mer frequencies of the selected sequence. Export buttons for: saving 3D plot, distance matrix, UPGMA tree and inter-cluster distances. 138

C.S6 “Zooming in" a ModMap3D, by re-plotting a subset of its dataset, can sometimes clarify cluster separations (separations can also be independently confirmed by the output of the supervised machine learning classifiers). Here, subfigures (a) and (c) are each obtained by re-plotting clusters which appear to be overlapping in the ModMap3D of the dataset of human mtDNA genomes from subfigure (b), as follows: **(a)** ModMap3D of 350 complete human mitochondrial genomes from the dataset in Table S1, line 13 (subset of dataset in line 12); **(b)** ModMap3D of 1,150 human mitochondrial genomes from the dataset in Table S1, line 12; **(c)** ModMap3D of 250 human mitochondrial genomes from the dataset in Table S1, line 14 (subset of dataset in line 12). 140

C.S7 Right panel components: Digital signal representation, classification accuracies, confusion matrix, and classify a new sequence. 142

C.S8 MLDSP-GUI test run for the 7,881 *Flavivirus* genomes in the dataset in Table S1, line 10 using the “purine/pyrimidine” representation with length normalization to median length. The Digital Signal Representation component (top right panel) shows the magnitude spectrum of the selected point/sequence. Note that even though this is the same dataset as the one in Figure C.S2, the visual shape of clusters is different and the classification accuracy is lower for the Linear Discriminant classifier. The visual differences in the clusters are due to the different numerical representations used. In general, the choice of numerical representation, supervised classifier, and other parameters depend on the specific dataset, and one should choose those that achieve the best numerical classification accuracy or confusion matrix. 144

E.S1 MoDMap3D representing a dataset comprising of *COXI* gene of 2630 vertebrates (bats: 819, birds: 1199, fish: 612) computed using ML-DSP with two different numerical representation, (a) Chaos Game Representation (CGR) at k-value 6 and, (b) Purine/Pyrimidine (PP) representation. 177

List of Tables

2.1 Rules for numerical representations of DNA sequences. 16

3.1 Numerical representations of DNA sequences. 45

3.2 Maximum classification accuracy scores when using Euclidean vs. Pearson's correlation coefficient (PCC) as a distance measure. 53

3.3 Average classification accuracies for 13 numerical representations. 54

3.4 Maximum classification accuracy (of the accuracies obtained with each of the six classifiers) of ML-DSP, for datasets at different taxonomic levels, from 'domain into kindgoms' down to 'family into genera'. 56

3.5 Comparison of classification accuracy and processing time for the distance matrix computation with MEGA7(MUSCLE), MEGA7(CLUSTALW), FPP, and ML-DSP. 64

5.1 Classification accuracy scores of viral sequences at different levels of taxonomy. 90

5.2 Predicted taxonomic labels of 29 COVID-19 virus sequences. 91

5.3 Genus to sub-genus classification accuracy scores of *Betacoronavirus*. 94

5.4 Spearman's rank correlation coefficient (ρ) values from Fig 5.6 and 5.7, for which all p-values $< 10^{-5}$. The strongest correlation value was found between *Betacoronavirus* and *Sarbecovirus* when using the data sets from Test 3a from Table 5.2 and Test 4 from Table 5.3, respectively. 98

C.S1 Additional datasets provided. 146

D.S1 Spearman's rank correlation coefficient (ρ) value for $k = 1$ to $k = 7$. 149

D.S2 Accession IDs of the sequences downloaded from the GISAID. 151
D.S3 Accession IDs of sequences used in Test-1 to Test-6. 173

E.S1 Performance comparison of CLUSTALW, MUSCLE, and ML-DSP. 177

List of Appendices

Appendix A Copyright Releases 124
Appendix B Software description 126
Appendix C MLDSP-GUI: Supplementary Material 129
Appendix D COVID-19 case study: Supplementary Material 148
Appendix E Addendum 174

Chapter 1

Introduction

Organism classification is important to better understand and preserve biodiversity, considering that approximately 86% of existing species on Earth and 91% of marine species are still unclassified [1, 2]. Taxonomy, the science of naming, defining, and classifying biological organisms, groups the organisms on the basis of their shared characteristics. Besides morphology-based and functionality-based taxonomy, DNA-based approaches have been employed in modern times to analyze genomic DNA sequences and classify organisms based on their sequence similarities. Sequence analysis methods can be alignment-based or alignment-free. The traditional alignment-based methods [3, 4, 5, 6] look for correspondence of individual bases that are in the same order in two or more sequences and as a result, are generally computationally demanding. These methods are further categorized on the basis of global alignment (alignment over the entire length of the sequence) and local alignment (focus is to identify widely divergent regions) [7]. The alignment-free methods provide an alternative while addressing the limitations and the challenges of the alignment-based approaches [8, 9]. These methods bypass altogether the base-to-base comparisons and classify the organisms on the basis of their genomic signatures, a specific quantitative characteristic of a DNA genomic sequence that is pervasive along the genome of the same organism while being dissimilar for DNA sequences of different organisms [10]. The detailed discussion on existing alignment-based and alignment-free methods is given in Section 2.2. Though existing alignment-free methods address most of the limitations of the alignment-based methods, they often lack software implementations and are tested on very small datasets [9]. Hence, a novel method is required that is open source, publicly available, fast, scalable, and proven to achieve satisfactory classification accuracy using a variety of large real-world datasets.

Our goal is to develop an ultra-fast, scalable, and highly accurate DNA sequence analysis method, which we accomplish by proposing a general-purpose alignment-free method ML-DSP (Machine Learning with Digital Signal Processing) [11]. ML-DSP implements a four-step pipeline for genomic sequences analysis comprising: One-dimensional numerical representations of DNA sequences to map genomic sequences to genomic signals, Discrete Fourier Transform (DFT) to obtain magnitude spectra from genomic signals, Pearson Correlation Coefficient (PCC) as a dissimilarity measure for pair-wise distance calculation between magnitude spectra of any two genomic signals, and supervised machine learning classification for classification and prediction of new sequences. For visualization of classification results, Multi-Dimensional Scaling (MDS) is used for dimensionality reduction and the three most significant dimensions are used to produce a three-dimensional Molecular Distance Map (MoDMap3D) [12].

Our research findings are organized in the following way. Chapter 3 contains the article "ML-DSP: Machine Learning with Digital Signal Processing for ultrafast, accurate, and scalable genome classification at all taxonomic levels" [11] in which we propose our alignment-free method ML-DSP and perform genome classification at different taxonomic levels using complete mitochondrial (mtDNA) sequences. This comprehensive analysis also shows the method's applicability to the classification of bacterial sequences and virus-subtypes. ML-DSP shows the potential for filling in the gaps in the field of taxonomy by suggesting taxonomy labels for unclassified sequences. Chapter 4 contains the article "MLDSP-GUI: an alignment-free standalone tool with an interactive graphical user interface for DNA sequence comparison and analysis" [13]. MLDSP-GUI is an extension of ML-DSP with the addition of a user-friendly interactive Graphical User Interface (GUI), of a two-dimensional Chaos Game

Representation (CGR) [14] to numerically represent DNA sequences, of Euclidean and Manhattan distances as additional distance measures, of the option of a phylogenetic tree output in Newick-formatted file, of oligomer (sub-word) frequency information to study the under-and-over representation of any particular sub-sequence in a selected sequence, and of inter-cluster distances analysis. ML-DSP and MLDSP-GUI are stand-alone tools and hence they also address data-security and data-privacy concerns that could arise in the health-science applications, because they eliminate the need of transferring the private data to the remote servers. Chapter 5 contains the article "Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: COVID-19 case study" [15]. This article shows our method's ability to accurately identify the taxonomy of novel unclassified sequences. The recent COVID-19 viral outbreak that originated in Wuhan, China raises a question about the scalability and the speed of the existing methods for comparing a novel sequence with thousands of known viral sequences. Our alignment-free approach not only provides rapid taxonomic identification of the novel viral sequence by comparing it against the thousands of known species, but also bypasses altogether the complexity involved in the annotations and additional biological information that are necessary requirements for alignment-based methods or clinical analyses.

We conclude this thesis in Chapter 6, which contains a discussion about possible extensions of current work, including the investigation of the environmental impact on genomic signatures, disease classification and how diseases compromise genomic integrity, and identification of the bacterial origin of mitochondrial DNA and chloroplast DNA in eukaryotes. Lastly, we discuss potential uses of our approach in studying genotyping data to investigate the genetic makeup of an organism.

Bibliography

[1] Mora C, Tittensor DP, Adl S, *et al.* How many species are there on earth and in the ocean? PLoS Biology. 2011; 9(8): e1001127.

[2] May RM. Why worry about how many species and their loss? PLoS Biology. 2011; 9(8): e1001130.

[3] Hebert PDN, Cywinska A, Ball SL, *et al.* Biological identifications through DNA barcodes. Proceedings of the Royal Society of London Series B: Biological Sciences. 2003; 270(1512): 313–321.

[4] Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004; 32(5): 1792–7.

[5] Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 1994; 22(22): 4673–80.

[6] Larkin MA, Blackshields G, Brown NP, *et al.* CLUSTAL W and CLUSTAL X version 2.0. Bioinformatics. 2007; 23(21): 2947–8.

[7] Polyanovsky VO, Roytberg MA, Tumanyan VG. Comparative analysis of the quality of a global algorithm and a local algorithm for alignment of two sequences. Algorithms for Molecular Biology. 2011; 6(1): 25.

[8] Vinga S, Almeida J. Alignment-free sequence comparison–a review. Bioinformatics. 2003; 19(4): 513–523.

[9] Zielezinski A, Vinga S, Almeida J, Karlowski WM. Alignment-free sequence comparison: benefits, applications, and tools. Genome Biology. 2017, 18: 186.

[10] Karlin S, Burge C. Dinucleotide relative abundance extremes: a genomic signature. Trends in Genetics. 1995; 11(7): 283–290.

[11] Randhawa GS, Hill KH, Kari L. ML-DSP: Machine Learning with Digital Signal Processing for ultrafast, accurate, and scalable genome classification at all taxonomic levels. BMC Genomics. 2019; 20: 267.

[12] Karamichalis R, Kari L, Konstantinidis S, Kopecki S. An investigation into inter- and intragenomic variations of graphic genomic signatures. BMC Bioinformatics. 2015; 16: 246.

[13] Randhawa GS, Hill KH, Kari L. MLDSP-GUI: an alignment-free standalone tool with an interactive graphical user interface for DNA sequence comparison and analysis. Bioinformatics. 2019; btz918.

[14] Jeffrey HJ. Chaos game representation of gene structure. Nucleic Acids Res. 1990; 18: 2163–2170.

[15] Randhawa GS, Soltysiak MPM, El Roz H, Hill KA, de Souza CPE, Kari L. Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: COVID-19 case study. PLoS ONE. 2020; 15(4): e0232391;

Chapter 2

Literature review

2.1 Biological background

Earth is home to a great diversity of life forms, estimated at nearly 8.7 million (±1.3 million) species [1, 2]. The naming and categorization of these organisms date back to the origin of human languages, as it has always been essential to communicate information about poisonous or edible plants to other people [3]. One of the earliest documents *Divine Husbandman's Materia Medica* containing 365 Chinese medicines derived from minerals, plants, and animals, is believed to be the work of Shen Nung (2737 *BC* – 2697 *BC*), compiled by multiple authors between *AD* 25 – *AD* 220 [4]. As illustrated in ancient wall paintings, the naming of medicinal plants was in use around 1500 *BC* in Egypt [3]. In the West, ancient work on taxonomy (naming and categorization of organisms) was done by Greeks and Romans [3]. The Greek philosopher Aristotle (384 *BC* – 322 *BC*) attempted the first systematic classification (animals with and without blood) of living organisms, followed by his student Theophrastus (370 *BC* – 285 *BC*) who classified 480 plant species based on their growth form [3]. Caesalpino extended the work of Theophrastus and wrote *De plantis* in the year 1583 that contained a classification of 1500 plant species based on their fruit and seed form together with the growth form [5]. The foundation of modern taxonomy was laid out by Carl Linnaeus who formulated and published the first nomenclature rules in 1735 [6]. After Charles Darwin proposed the evolutionary theory in 1858, Ernst Harckel established the term phylogeny to study evolutionary history using similarities and differences among different groups of organisms [7]. In 1965, Willi Henning founded the modern cladistic method that categorizes organisms based on shared characteristics [8]. Early taxonomy focused on the shared morphological characteristics to categorize the group of biological organisms, whereas modern taxonomy extended the characteristics use from merely morphological to molecular [9]. Deoxyribonucleic Acid (DNA), and Ribonucleic Acid (RNA) are a natural choice of molecules that can be used in sequence analyses for various purposes, including taxonomy.

Deoxyribonucleic Acid (DNA) is a molecule that encodes the genetic information that allows all known living organisms to function, grow and reproduce. DNA is a directed polymer made from monomeric units called nucleotides. The four different nucleotides of DNA are Adenine(A), Cytosine (C), Guanine (G), Thymine (T). A DNA strand can be represented as a string over a four-letter alphabet consisting of letters A, C, G, and T. In a double-stranded DNA molecule, the bases on one strand pair with the complementary bases on another strand, A with T and C with G, to form units called base pairs. The two strands comprising the DNA double strand run in opposite directions to each other, and thus each strand is the reverse complement of the other. DNA may be present in different parts of a cell. Prokaryotes (bacteria and archaea) store their DNA in the cytoplasm. Eukaryotic organisms (animals, plants, fungi, and protists) store most of their DNA inside the cell nucleus as nuclear DNA, and some in the mitochondria as mitochondrial DNA or in chloroplasts as chloroplast DNA. Viruses may have single- or double-stranded DNA or RNA (Ribonucleic Acid) as their genetic material. In Section 2.2, we discuss existing DNA sequence analysis methods and in Chapter 3, we explore DNA sequence classification at all taxonomic levels using our proposed method.

2.2 Genomic sequence analysis methods

In the field of bioinformatics, DNA sequence classification is the scientific practice of identifying, naming, and grouping of organisms based on their differences and similarities. The problem of species classification is of immense importance considering that nearly 86% of existing species on Earth and 91% of marine species, of the estimated 8.7 million (±1.3 million) species, remain unclassified [1, 2]. With advancements in techniques such as Next Generation Sequencing (NGS), the tremendous growth in the quantity of genomic data makes real-time sequence analysis quite challenging [10]. In addition to taxonomic classification, sequence (dis)similarity analysis has multiple possible applications including virus-subtype classification (assign viral sequences to their subtypes), disease classification (classify human genomic sequences on the basis of disease type), human-haplogroup classification (assign human mtDNA sequences on the basis of maternal lineage), etc.

Sequence comparison and analysis methods are broadly categorized into two groups: (i) alignment-based, and (ii) alignment-free methods. Alignment-based methods search for base-to-base correspondences in two or more sequences and it requires the sequences to be more or less conserved. Sequence similarity is measured by computing a score based on the number of matches, mismatches, and insertions/deletions between compared sequences. These methods can accurately align closely related sequences, but it is difficult to compute a reliable alignment for divergent sequences. Alignment-free methods provide an alternative by bypassing base-to-base comparisons altogether. The sequence similarity analysis is based on the concept of genomic signatures. The next subsections discuss a variety of these methods proposed and developed in the literature.

2.2.1 Alignment-based methods

The development of sequence analysis methods started around four decades ago [11]. Initially, algorithms were mostly borrowed from existing computer science methodologies such as string processing [12], a natural choice considering the availability of a limited amount of genomic data. Alignment-based methods search for a correspondence between individual bases that are in the same order in two or more sequences [11]. The sequence similarity is quantitatively measured by computing an alignment-score based on the number of matches, mismatches, and indels (insertions/deletions) [13]. Many alignment-based tools have been developed such as, BLAST [14], FASTA [15], MUSCLE [16], ClustalW [17], ClustalX [18], MAFFT [19], etc. Though alignment-based methods have been successfully used for genome classification, they are not applicable when one needs to compare sequences originating from different regions of various genomes. Some limitations of alignment-based methods are [11, 20, 21]:

(i) Alignment-based methods assume sequences to be continuous and homologous (more or less conserved sequence fragments that have remained essentially unchanged throughout evolution). Sequences with great variation and high mutation rates, such as viral sequences, usually don't strictly follow this assumption. Moreover, the long-range interactions resulting from recombination (with shuffling) of conserved segments are overlooked [22, 23].

(ii) The accuracy of sequence alignment depends on the amount of sequence identity (amount of exact matches between two sequences). When sequence identity falls below a threshold value, the accuracy can rapidly drop off.

(iii) Alignment-based methods are generally computationally demanding. As the number and lengths of sequences grow, so does the demand for computation time and memory.

(iv) Computationally, it is not possible to solve multiple-sequence alignment, (which is an NP-hard problem) for thousands of complete genomes in a feasible time.

(v) The alignment score depends on multiple a priori assumptions. The selection of input parameters e.g. gap penalty, match/mismatch scores, etc., may often change the results.

2.2.2 Alignment-free methods

The alignment-free methods have been proposed as an alternative to address situations where alignment-based methods are computationally inefficient or fail [11, 20, 21]. They have following advantages: (i) alignment-free methods are capable of recognizing homology even when the loss of contiguity is beyond the possibility of alignment [20]. (ii) With alignment-free methods, similarities can be found that can't be discovered through edit distances (counting the minimum number of operations required to transform one string into the other), which are used in alignment-based methods [24]. (iii) ability to compare unrelated sequences. There are a variety of alignment-free methods proposed over the last few decades.

Random walk [25, 26] was one of the first alignment-free methods that were proposed. It generates two-dimensional graphical representations of genomic sequences and compares them using Manhattan and Euclidean distances. More specifically, the four nucleotides *T, A, C, G* are encoded by four possible moves corresponding to the directions *up, down, left, right* respectively, to generate a graphical representation in a plane. Susceptible to degeneracy, initially this method was considered unsuitable for genomic analysis. The method was later improved [27, 28] by using the geometric center of the points in the walk for sequence comparison. Modified versions of the random walk technique have been used to produce the similarity matrices from the first exon of the β-globin gene of several mammals [29, 30, 31, 32, 33] and to generate the phylogenetic trees for primate mitochondrial DNA [30], coronaviruses [34], etc. The random walk technique has also been used to analyze proteins [35, 36, 37], bacteria [38] and yeast [39]. In the random walk technique, the plotting of the current point depends on the preceding points. Randic *et al.* [40, 41] proposed an alternative representation, called "cell" representation, where the plotting of points is independent of the preceding points. They proposed the construction of a 12-component vector by using the leading eigenvalues of the L/L matrix (Length by Length matrix) for the comparison of the first exon of β-globin region of 11 mammals. The elements of the L/L matrix are defined as the quotient of the Euclidean distance between a pair of dots of the plotted curve and the sum of distances between the same pair of dots measured along the curve. Various modifications were proposed following this study [42, 43, 44], but these techniques failed to receive attention because the representation construction is computationally inefficient.

Qi *et al.* proposed a graph theory based method [45], where for each DNA sequence a weighted directed graph with four vertices (one vertex for each nucleotide) is constructed. Each edge of the graph represents a unique dinucleotide and graph has sixteen edges in total. The edge weights are updated based on both ordering and frequency of nucleotides, and an adjacency matrix of size 4×4 corresponding to the edge weights is constructed. The dissimilarity between any two DNA sequences is measured by computing a distance between their respective adjacency matrices.

Over the years, other alignment-free methods have been proposed which used different approaches. Markov models have been used to cluster coding DNA sequences [46], to study intra-genomic variations for viruses and some animals [47], and to build phylogenies of *S. flexneri*, *E. Coli* [48], Hepatitis-E virus [49] and HIV-1 [50]. Thermal melting profiles have been used to classify several mammalian species using β-globin and αchain class II MHC genes [51]. Lempel-Ziv complexity has been used to cluster protein families into functional subtypes [52]. This method has also been used to build phylogenetic trees of fungi using ribosomal DNA sequences [53], perennial plant genus *Galanthus* using nuclear and chloroplast DNA sequences [54], and HEV and mammals using DNA sequences [55, 56, 57, 58] .

Another popular category of alignment-free methods makes use of word frequencies [59, 60, 61]. The difference between the two sequences can be obtained by computing the k-mer (subsequences of length k) frequencies first and then distance between them. The word-based alignment-free technique was first used to construct accurate phylogenetic trees for mammalian alpha- and beta-globin genes [62]. Bao *et al.* [63] proposed a Category-Position-Frequency (CPF) model, which utilized word frequency and position information of nucleotides in DNA sequences. The main disadvantage of this method is that the adjacent word matches are dependent on each other. Leimeister *et al.* [64] proposed a method based on spaced-words frequencies to address the problem of dependency on adjacent word matches. This method used spaced-words, defined by patterns of ‘match’ and ‘don't care’ positions, for alignment-free sequence comparison. Sims *et al.* [65] proposed a k-mer vectors based method called Feature Frequency Profiles (FFP). FFP has been used for phylogenetic analysis using a variety of sequences including intron sequences of mammals [65], mitochondrial DNA sequences of primates and nuclear DNA sequences of plants [66], and bacterial genomes [67]. Many authors [68, 69, 70, 71, 72, 73, 74, 75, 76, 77] have used Chaos game Representation (CGR) [78] for k-mer-based sequence analysis. CGR is a two-dimensional graphical representation of DNA sequence, and the details of the CGR construction are given in Section 2.3.1. CGR has been used in literature on a variety of sequences e.g. to build phylogenies using mitochondrial DNA sequences [71, 72], nuclear DNA sequences [73, 75], bacterial sequences [76], and viral sequences [77, 70].

In recent years, Genomic Signal Processing (GSP) [79] based alignment-free methods have also been proposed. GSP-based methods apply techniques of Digital Signal Processing (DSP) to genomic data. GSP-based methods have been successfully used for a variety of applications, e.g., to distinguish introns from exons [80, 81, 82], for complete genome phylogenetic analysis of primates, bacteria and influenza [83], and for classification of whole bacterial genomes [84]. Borraya *et al.* [85] proposed a GSP-based method for the computation of alignment-free distances between DNA sequences, where DNA sequences were mapped to numerical sequences based on the nucleotide doublet values ($0 - 15$ for all possible 16 combinations). The analysis was done on relatively small dataset composed of the ribosomal $S18$ subunit gene. Yin *et al.* [86] proposed another alignment-free method that encoded each DNA sequence to four binary indicator sequences and applied Discrete Fourier Transform (DFT) to compute the power spectra. The Euclidean distance of full DFT power spectra of the DNA sequences was used as a dissimilarity measure. Other DSP techniques have also been used for genome similarity analysis, e.g. comparing the phase spectra of the DFT of digital signals of full mtDNA genomes [87, 88].

Though existing alignment-free methods have successfully addressed most of the limitations of the alignment-based methods, they have some disadvantages of their own. Zielezinski *et al.* [11] reviewed the majority of existing alignment-free methods and highlighted the following limitations:

(i) A majority of existing alignment-free methods are still exploring the technical foundations and lack software implementation, so it is not possible to compare their performance on common datasets. Without comparison or existing proven results, it is difficult for users to pick one method for their specific application.

(ii) Most of the existing alignment-free methods that have software implementations available are tested using very small real-world datasets or simulated sequences. Their applicability to a variety of applications is untested.

(iii) Though alignment-free methods have lower time-complexity, their memory consumption is still an issue, at least for k-mer based methods. The use of longer k-mers for multigenome data can cause possible memory overhead.

We propose a novel alignment-free GSP-based methodology that addresses the limitations of the existing alignment-free methods in addition to the alignment-based methods, see Section 2.3 for details.Though our proposed approach addresses the previously identified limitations of both alignment-based and alignment-free algorithms, high memory use remains an issue when CGR, a k-mer dependent numerical representation, is used. The high memory use is because of the length of sequences, and large size of datasets. In particular, high memory use is unavoidable if the required analysis demands the use of full genomes. Another notable limitation of our methodology is inherited from the use of supervised machine learning algorithms. More specifically, our approach can only predict the label of an unknown new sequence by assigning a label from the available labels in the training set. In case the actual label is missing from the training set, our approach assigns a closest available label (the label of the most similar sequence in the training set).

2.3 Our approach

Any DNA sequence can be represented as a string over a four-letter alphabet consisting of letters A, C, G, and T. Consequently, by using an appropriate numerical encoding, a DNA sequence can be encoded as a discrete numerical sequence using DNA numerical representations such as the ones in [89, 90, 91], and hence treated as a digital signal. These digital signals (discrete numerical sequences) generated from the genomic sequences are called genomic signals [92]. The genomic signals can be analyzed using various Digital Signal Processing (DSP) [93, 94] techniques, and the whole process can be termed Genomic Signal Processing (GSP) [85, 79].

Our objective is to develop a GSP-based alignment-free method in combination with machine learning, and use it for sequence analysis and comparison. We propose and test a GSP-based pipeline that maps genomic sequences to genomic signals, computes magnitude spectra by applying DFT to genomic signals, computes a pairwise distance matrix by evaluating the dissimilarities between pairs of magnitude spectra of any two genomic signals, and uses supervised machine learning algorithms to classify genomic sequences based on these distances. The proposed methodology is outlined in the flowchart shown in Figure 2.1. Various components of the proposed methodology are discussed in sub-sections 2.3.1-2.3.5.

2.3.1 DNA numerical representations

We tested our approach on 14 DNA numerical representations, of which 13 are one-dimensional representations and the last one is a two-dimensional representation. The thirteen different one-dimensional numerical representations for DNA sequences are grouped as: Fixed mappings DNA numerical representations (Table 2.1 representations #1, #2, #3, #6, #7, see [89], and representations #10, #11, #12, #13 - which are one-dimensional variants of the binary representation proposed in [89]), mappings based on some physio-chemical properties of nucleotides (Table 2.1 representation #4, see [89, 95], and representation #5, see [89, 95, 96]), and map-

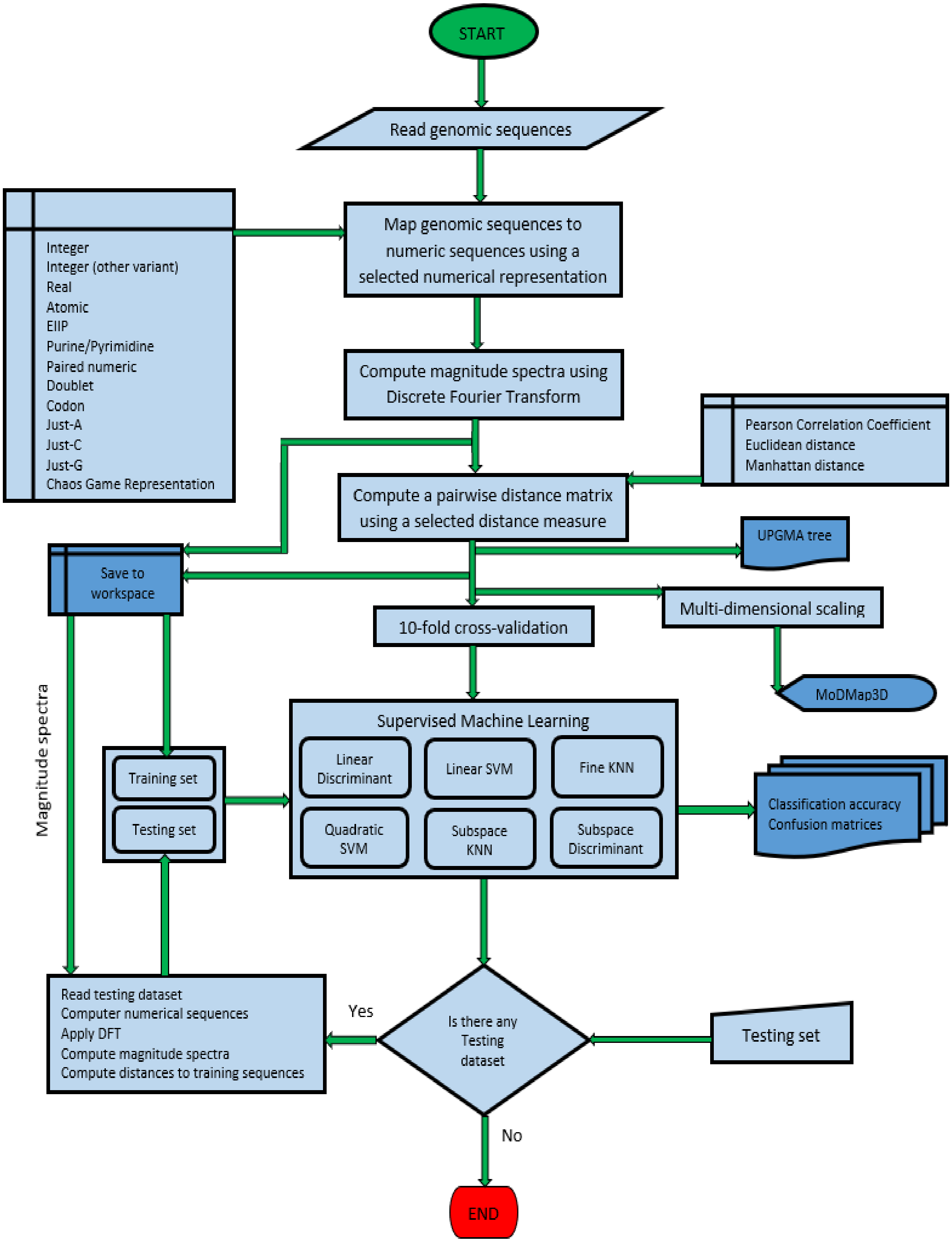


Figure 2.1: Flowchart showing MLDSP methodology.

pings based on the nearest-neighbour values (Table 2.1 representations #8, #9, see [85]). Table 2.1 gives the rules for constructing genomic signals from DNA sequences using the 13 one-dimensional representations. For example, if the numerical representation is Integer (#1 in Table 2.1), then for the sequence $S = CGGTAT$, the corresponding numerical representation is $N = (1, 3, 3, 0, 2, 0)$. The comparison analysis of 13 one-dimensional representation is given in sub-section 3.3.2.

Table 2.1: Rules for numerical representations of DNA sequences.

#	**Representation**	**Rules**	**Output for S = CGGTAT**
1	Integer	T=0, C=1, A=2, G=3	[1 3 3 0 2 0]
2	Integer (other variant)	T=1, C=2, A=3, G=4	[2 4 4 1 3 1]
3	Real	T=−1.5, C=0.5, A=1.5, G=−0.5	[0.5 −0.5 −0.5 −1.5 1.5 −1.5]
4	Atomic	T=6, C=58, A=70, G=78	[58 78 78 6 70 6]
5	EIIP (electron-ion interaction potential)	T=0.1335, C=0.1340, A=0.1260, G=0.0806	[0.1340 0.8060 0.8060 0.1335 0.1260 0.1335]
6	PP (purine/pyrimidine)	T/C=1, A/G=−1	[1 −1 −1 1 −1 1]
7	Paired numeric	T/A=1, C/G=−1	[−1 −1 −1 1 1 1]
8	Nearest-neighbor based doublet	0−15 for all possible doublets	[14 11 10 2 1 7]
9	Codon	0−63 for all possible 64 Codons	[6 51 11 56 22 44]
10	Just-A	A=1, rest=0	[0 0 0 0 1 0]
11	Just-C	C=1, rest=0	[1 0 0 0 0 0]
12	Just-G	G=1, rest=0	[0 1 1 0 0 0]
13	Just-T	T=1, rest=0	[0 0 0 1 0 1]

Numerical representations of DNA sequences used in genomic classification. The second column lists the numerical representation name, the third column describes the rule it uses, and the fourth is the output of this rule for the input DNA sequence $S = CGGTAT$. For the nearest-neighbor based doublet representation and codon representation, the DNA sequence is considered to be wrapped (the last position is followed by the first).

In addition to 13 one-dimensional numerical representation, we also used a two-dimensional representation, called Chaos Game Representation (CGR) [78]. CGR was suggested as a good candidate for the role of genomic signature by Deschavanne *et al.* [73, 74]. CGR is a square-shaped graphical representation with four corners labeled as A, C, G, T respectively (representing four different DNA nucleotides). For every letter in the DNA sequence, a dot is plotted within the square. The first dot is plotted in the middle of the segment defined by the square. For each consecutive nucleotide, a dot is plotted in the middle of the last plotted dot and the corner labelled by that nucleotide. Figure 2.2 shows the steps involved in creating the CGR plot of the DNA sequence CGGTAT. Figure 2.3a shows the CGR plot of the complete mtDNA sequences of Canadian beaver (*Castor canadensis*), NCBI accession *NC*_007011.1, 16767 bp long and Figure 2.3b shows the CGR plot of the complete mtDNA sequence of Canada goose *Branta canadensis*, NCBI accession *KY*311838.1, 16760 bp long. The use of CGR as a numerical representation for our method is given in Section 5.3.

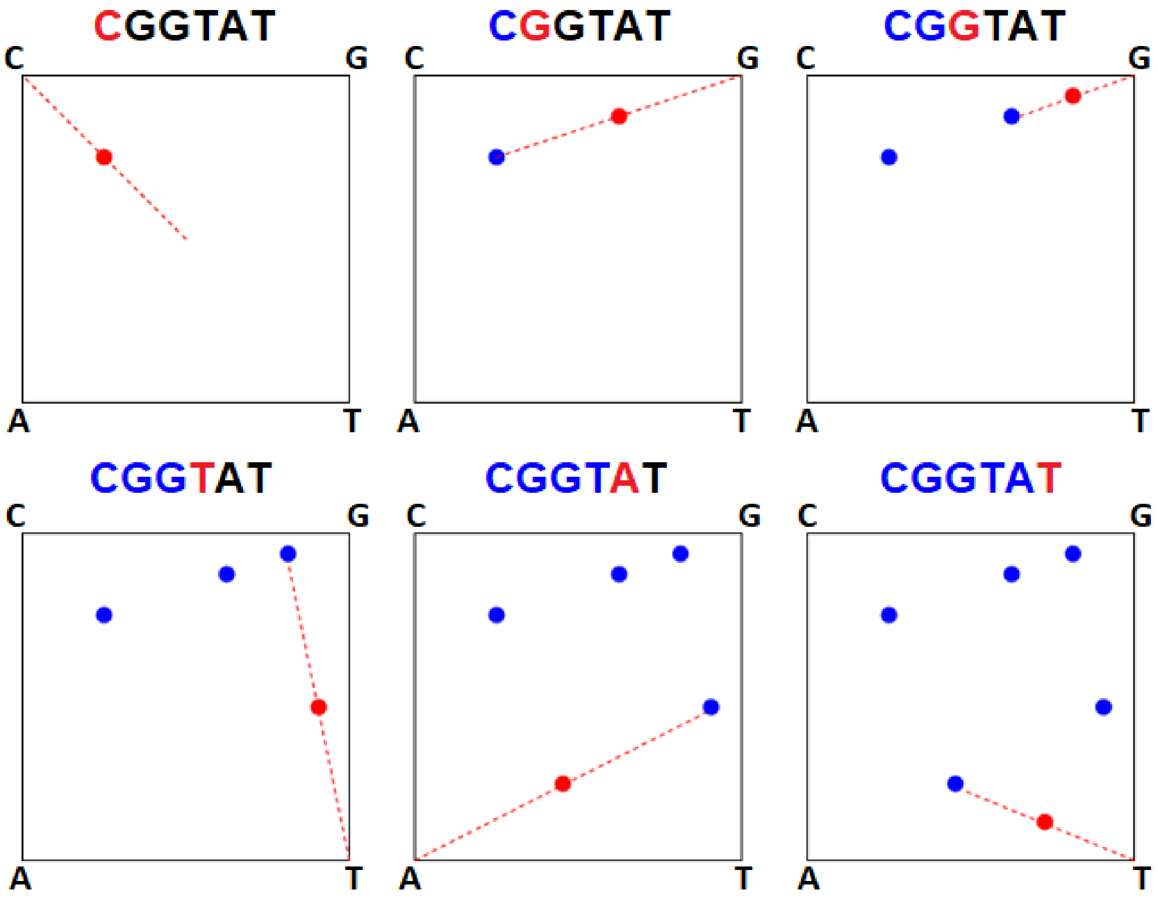


Figure 2.2: The Chaos Game Representation (CGR) of the DNA sequence CGGTAT.

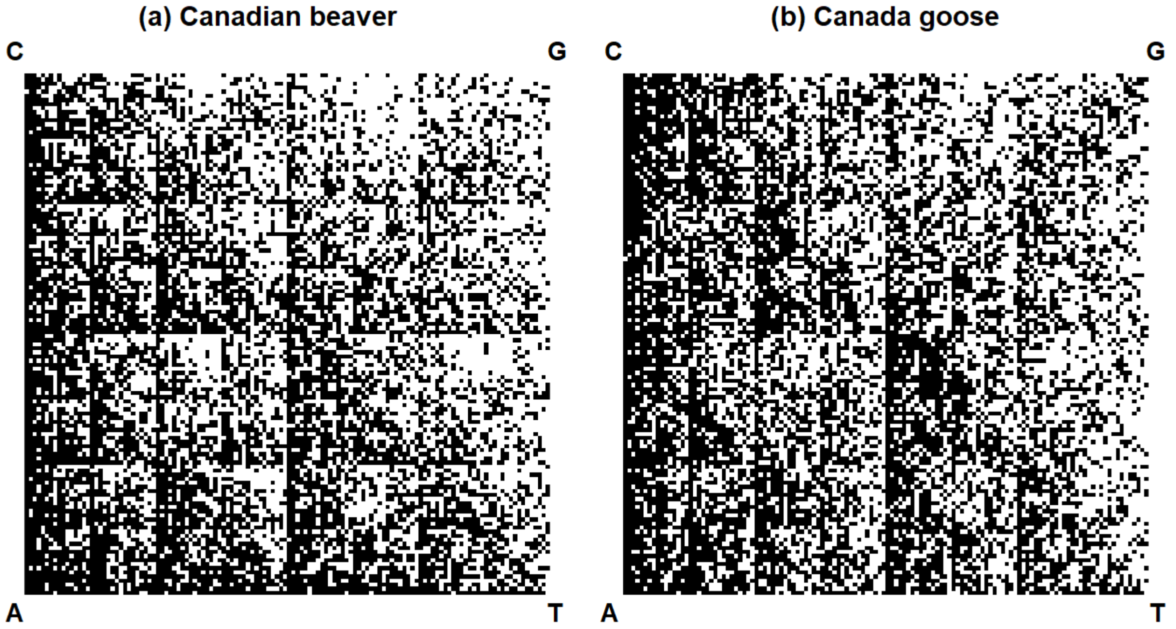


Figure 2.3: The Chaos Game Representation (CGR) of the mtDNA sequence of (a) Canadian Beaver (*Castor canadensis*), NCBI accession *NC*_007011.1, 16767 bp length and (b) Canada goose (*Branta canadensis*), NCBI accession *KY*311838.1, 16760 bp length.

2.3.2 Discrete Fourier Transform

Discrete Fourier Transform (DFT) [97] is applied to the genomic signals (discrete numerical representations of the genomic sequences) to compute the magnitude spectra. Suppose we have a dataset of n sequences. For CGR numerical representation, columns of each $2D$ vector are concatenated to reshape it as a $1D$ vector similar to the outcome of $1D$ numerical representations. For selected k value (k being the length of k-mers), CGR of any sequence i ($0 \leq i \leq n-1$) will be of size $2^k \times 2^k$ and its corresponding $1D$ vector will of size p, where $p = 2^k \times 2^k$. Then, the DFT of an i^{th} ($0 \leq i \leq n-1$) genomic signal $N_i = N_i(0), N_i(1), ..., N_i(p-1)$ results in another sequence of complex numbers, $F_i(k) = F_i(0), F_i(1), ..., F_i(p-1)$ where, for $0 \leq k \leq p-1$ we have:

$$F_i(k) = \sum_{j=0}^{p-1} N_i(j) \cdot e^{(-\iota 2\pi/p)kj} \tag{2.1}$$

The magnitude spectrum of a genomic signal N_i is the absolute value of the vector F_i.

2.3.3 Distance measures

In this thesis, there are three different dissimilarity measures being used: Euclidean distance [98], Manhattan distance [99], and Pearson Correlation Coefficient (PCC) [100, 101].

The Euclidean distance d_{EUC} between two magnitude spectra X and Y, each of length p, is computed as:

$$d_{EUC} = \sqrt{\sum_{i=0}^{p-1}(X_i - Y_i)} \tag{2.2}$$

The Manhattan distance d_{MAN} between two magnitude spectra X and Y, each of length p, is computed as:

$$d_{MAN} = \sum_{i=0}^{p-1} |X_i - Y_i| \tag{2.3}$$

The Pearson Correlation Coefficient r_{XY} between two magnitude spectra X and Y, each of length p, is computed as:

$$r_{XY} = \frac{\sum_{i=0}^{p-1}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=0}^{p-1}(X_i - \overline{X})^2} \times \sqrt{\sum_{i=0}^{p-1}(Y_i - \overline{Y})^2}} \tag{2.4}$$

where the average $\overline{X}$ is defined as $(\sum_{i=0}^{p-1} X_i)/p$ and similarly for Y. The results are normalized by taking $(1 - r_{XY})/2$, to obtain dissimilarity values between 0 and 1. It should be noted that $1 - r_{XY}$ is not a metric, whereas $\sqrt{1 - r_{XY}}$ is a metric.

2.3.4 Multi-dimensional scaling

Multi-Dimensional Scaling (MDS) is a means of visualizing the degree of similarity between individual objects in a given dataset. Classical multidimensional scaling takes a pairwise distance matrix ($n \times n$ matrix, for n objects) as input, and produces n points in a q-dimensional Euclidean space, where $q \leq n - 1$. More specifically, the output is an $n \times q$ coordinate matrix, where each row corresponds to one of the n input objects, and that row contains the q coordinates of the corresponding object-representing point [102]. The Euclidean distance between each pair of points is meant to approximate the distance between the corresponding two objects in the original distance matrix. These points can then be simultaneously visualized in a 2- or 3-dimensional space by taking the first 2, respectively 3, coordinates (out of q) of the coordinate matrix. The result is a Molecular Distance Map (MoDMap) [103], and the MoDMap of a genomic dataset represents a visualization of the simultaneous interrelationships among all

DNA sequences in the dataset. Figure 2.4 shows a MoDMap generated by applying MDS to the pairwise distances between six most populated Canadian cities (Toronto, Montreal, Vancouver, Calgary, Edmonton, and Ottawa). A 6 × 6 pairwise distance matrix D is created, where any element d_{ij}, $1 \leq i \leq 6$, of matrix D is the distance in kilometers between the i^{th} and j^{th} city. The MDS algorithm takes matrix D as input, and produce the set of coordinates (two dimensional for this example) of six cities as output. Figure 2.4 shows a MoDMap produced by plotting the output of MDS as points, and the placement of points represents the estimated distances between the cities.

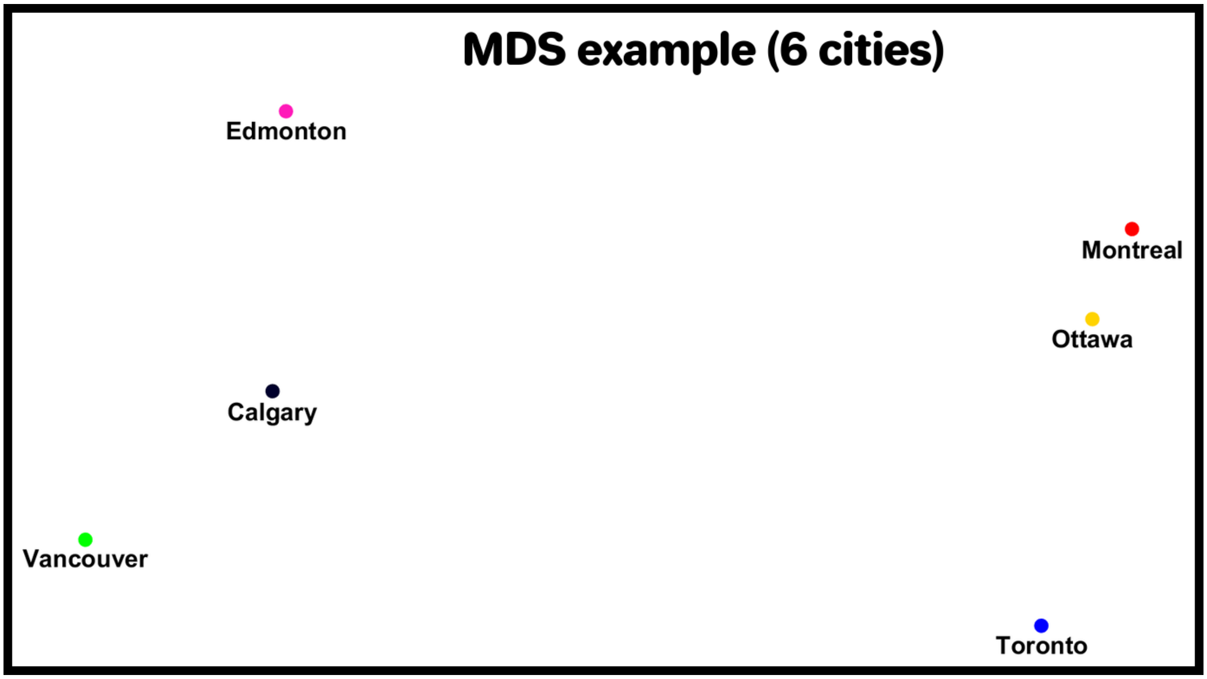


Figure 2.4: A MoDMap generated by applying multi-dimensional scaling to the pairwise distances between six most populated Canadian cities.

2.3.5 Supervised learning classification models

Supervised learning classification algorithms learn from the labelled training data and classify the new observations (testing data) into the training classes (a class is a group of similar observations). In this thesis, we used six classification models: Linear Discriminant, Linear SVM, Quadratic SVM, Fine KNN, Subspace Discriminant, and Subspace KNN. The 10-fold cross-validation score is used to assess the classification performance. In this approach, the dataset is randomly partitioned into ten equal-sized subsets. The classification model is trained using 9 of the subsets with available class labels, and the prediction accuracy is measured by testing the remaining subset. The process is repeated 10 times, and the accuracy score of the classification model is then computed as the average of the accuracies obtained in the 10 separate runs.

(i) **Linear Discriminant:** Linear discriminant analysis [104] is a fast classification method, and its memory usage is small. The space of X data points divides into K regions (number of classes). For linear discriminant analysis, the regions are separated by straight lines. This model assumes that the data in each class has a Gaussian mixture distribution. The model has different means, but the same covariance matrix for each class. The sample mean is computed first for each class. Then the sample covariance is computed by taking the empirical covariance matrix of the difference between the sample mean of each class and the observations of that class. The prediction function used to classify the observations is based on three factors: posterior probability, prior probability, and cost. The multi-objective minimization function used to predict the class $\widehat{y}$ of any observation x is:

$$\widehat{y} = \underset{y=1,\ldots,K}{arg\ min} \sum_{c=1}^{K} \widehat{P}(c \mid x)\, C(y \mid c) \tag{2.5}$$

where $\widehat{P}(c \mid x)$ is the posterior probability that an observation x belongs to class c and $C(y \mid c)$ is the cost of classifying an observation as y when its true class is c. Cost C is 0 if $y = c$, and 1 otherwise.

The posterior probability $\widehat{P}(c \mid x)$ is computed by Bayes' rule taking the product of prior probability $P(c)$ and the multivariate Gaussian (or normal) distribution:

$$\widehat{P}(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)} \tag{2.6}$$

where, $P(x)$ is the normalization constant equal to the sum over c of $P(x \mid c)P(c)$. The prior probability $P(c)$ of class c is computed by dividing the number of training samples of that class by the total number of training samples. The density function of the multivariate Gaussian with mean and covariance at an observation x is:

$$P(x \mid c) = \frac{1}{(2\pi|\sum_c|)^{\frac{1}{2}}} exp(-\frac{1}{2}(x-\mu_c)^T \sum\nolimits_c^{-1} (x-\mu_c)) \tag{2.7}$$

where $|\sum_c|$ is the determinant of $\sum_c$, and $\sum_c^{-1}$ is the inverse matrix.

(ii) **Linear Support Vector Machine:** Linear Support Vector Machine (SVM) [105, 106] makes a linear separation between classes. The SVM model finds the best hyperplane that separates all data points of one class from the data points of the other class. For binary classification, the best hyperplane means the one that has the largest distance to the nearest data points of any class i.e. the largest margin between the two classes. For three or more classes, multiple binary SVMs are used with Error-Correcting Output Codes (ECOC) classifier. An ECOC model reduces the problem of classification with three or more classes to a set of binary classification problems. For n classes, $n(n-1)/2$ one-versus-one binary classifiers are constructed.

(iii) **Quadratic Support Vector Machine:** It is not always possible to get a linear separation between the clusters (classes). The Quadratic SVM [105, 106] uses a quadratic function instead of a linear function to gain separation between the clusters. The data points are then mapped to a higher dimensional space to get linear separation. Quadratic SVM has slow prediction speed and large memory usage for multi-class classification.

(iv) **Fine KNN (*K*-Nearest Neighbours):** Fine KNN [107, 108] classifier performs a proximity search that typically has good predictive accuracy in low dimensions. The testing data points are categorized based on their distance to data points (neighbors) in a training dataset. In the Fine KNN classification model, the number of neighbors (K) is set to 1. The model calculates the Euclidean distance between the feature vectors of the testing data point and of the training data points. Given a set X of n data points, the Fine KNN model finds the K closest points in X to a testing data point or set of points. The testing data point is assigned a predictive class the same as of its closest neighbor (data point).

(v) **Subspace Discriminant:** The subspace discriminant is an ensemble model that uses a combination of linear discriminant weak learners [109]. We used the default 30 linear discriminant learners. Suppose n is the number of weak learners and d is the number of dimensions (features) in the data, an ensemble model chooses without replacement a random set of m predictors from d possible features (where, $m = |d/n|$) for each weak learner. The weak learners are trained on their respective sets of m predictors. The prediction is made by taking the average of prediction scores of all the weak learners. The class with the highest average score is assigned to the testing data point.

(vi) **Subspace KNN:** The subspace KNN is an ensemble model that uses a combination of Fine KNN weak learners [109]. We used the default 30 Fine KNN learners. The use of multiple learners makes the classification process slower. It has been shown that the combined (average) accuracies of the ensemble models typically increase with the increasing number of component classifiers, and with an appropriate subspace dimensionality, the ensemble methods can be superior to the individual learner models. Subspace ensembles also have the advantage of using less memory than ensembles with all predictors.

Linear discriminant and linear SVM models are more suitable if linear boundaries are expected between the classes. The linear discriminant model is the most popular because it is simple and fast. The discriminant analysis assumes that different classes generate data based on different Gaussian distributions and are linearly separable. Linear SVM model tries to find linear separability between data points that are most difficult to separate. For more than two classes, a classification problem is reduced to a set of binary classification sub-problems, and one SVM learner is used for each sub-problem. For higher-dimensional data, where it is challenging to linearly separate the variables, quadratic SVM gives better results than the linear SVM, with a little compromise on the time performance. Fine-KNN works well with a small number of data points but doesn't scale well to large input data. The ensemble models (Subspace Discriminant, and Subspace KNN) comprise several supervised learning models. The constituting models are individually trained and the final prediction is achieved by merging the results of individual models. This gives higher predictive power to the ensemble models, than any of their constituting learning algorithms independently. The higher predictive power comes at the cost of poor time performance and more memory usage.

Bibliography

[1] Mora C, Tittensor DP, Adl S, *et al.* How many species are there onearth and in the ocean? PLoS Biology. 2011; 9(8): e1001127.

[2] May RM. Why worry about how many species and their loss? PLoS Biology. 2011; 9(8): e1001130.

[3] Holley D. General Biology II: Organisms and Ecology. Dog Ear Publishing, Indianapolis, USA. 2017.

[4] Zhao Z, Guo P, Brand E. A concise classification of bencao (materia medica). Chin Med 13. 2018; 18.

[5] Cesalpino A. *De Plantis libri XVI.* Apud Georgium Marescottum. 1583.

[6] Linnaeus C. *Systema naturae per regna tria naturae :secundum classes, ordines, genera, species, cum characteribus, differentiis, synonymis, locis*. Stockholm: Laurentius Salvius. 1758.

[7] Dayrat B. The Roots of Phylogeny: How Did Haeckel Build His Trees? Systematic Biology. 2003; 52(4): 515–527.

[8] Hennig W. Phylogenetic Systematics. Annual Review of Entomology. 1965; 10(1): 97–116.

[9] Padial JM, Miralles A, De la Riva I, Vences M. The integrative future of taxonomy. Front Zool. 2010; 7: 16.

[10] Schmidt B, Hildebrandt A. Next-generation sequencing: big data meets high performance computing. Drug Discovery Today. 2017; 22(4): 712–717.

[11] Zielezinski A, Vinga S, Almeida J, Karlowski WM. Alignment-free sequence comparison: benefits, applications, and tools. Genome Biology. 2017; 18: 186.

[12] Gusfield D. Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology. Cambridge University Press, USA. 1997.

[13] Polyanovsky VO, Roytberg MA, Tumanyan VG. Comparative analysis of the quality of a global algorithm and a local algorithm for alignment of two sequences. Algorithms for Molecular Biology. 2011; 6(1): 25.

[14] Altschul SF, Madden TL, Schäffer AA, *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997; 25: 3389–402.

[15] Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A. 1988; 85: 2444–8.

[16] Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004; 32(5): 1792–7.

[17] Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 1994; 22(22): 4673–80.

[18] Larkin MA, Blackshields G, Brown NP, *et al.* CLUSTAL W and CLUSTAL X version 2.0. Bioinformatics. 2007; 23(21): 2947–8.

[19] Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 2002; 30: 3059–66.

[20] Vinga S, Almeida J. Alignment-free sequence comparison–a review. Bioinformatics. 2003; 19(4): 513–523.

[21] Song K, Ren J, Reinert G, *et al.* New developments of alignment-free sequence comparison: measures, statistics and next-generation sequencing. Briefings in Bioinformatics. 2014; 15(3): 343–353.

[22] Zhang YX, Perry K, Vinci VA, *et al.* Genome shuffling leads to rapid phenotypic improvement in bacteria. Nature. 2002; 415: 644–646.

[23] Lynch M. Intron evolution as a population-genetic process. Proc. Natl Acad. Sci. USA. 2002; 99: 6118–6123.

[24] Schwende I, Pham TD. Pattern recognition and probabilistic measures in alignment-free sequence analysis. Briefings in Bioinformatics. 2014; 15(3): 354–368.

[25] Hamori E, Ruskin J H curves, a novel method of representation of nucleotide series especially suited for long DNA sequences. Journal of Biological Chemistry. 1983; 258(2): 1318–1327.

[26] Gates MA A simple way to look at DNA. Journal of theoretical biology. 1986; 119(3): 319–328.

[27] Yau SST, Wang J, Niknejad A, *et al.* DNA sequence representation without degeneracy. Nucleic Acids Research. 2003; 31(12): 3078–3080.

[28] Liao B. A 2D graphical representation of DNA sequence. Chemical Physics Letters. 2005; 401(1–3), 196–199.

[29] Liao B, Tan M, Ding K. A 4D representation of DNA sequences and its application. Chemical Physics Letters. 2005; 402(4–6): 380–383.

[30] Liao B, Tan M, Ding K. Application of 2-D graphical representation of DNA sequence. Chemical Physics Letters. 2005; 414(4-6): 296–300.

[31] Yao YH, Nan XY, Wang TM. A new 2D graphical representation-Classification curve and the analysis of similarity/dissimilarity of DNA sequences. Journal of Molecular Structure: THEOCHEM. 2006; 764(1–3): 101–108.

[32] Tang XC, Zhou PP, Qiu WY. On the similarity/dissimilarity of DNA sequences based on 4D graphical representation. Chinese Science Bulletin. 2010; 55(8): 701–704.

[33] Qi ZH, Li L, Qi XQ. Using Huffman coding method to visualize and analyze DNA sequences. Journal of Computational Chemistry. 2011; 32(15): 3233–3240.

[34] Zheng WX, Chen LL, Ou HY, *et al*. Coronavirus phylogeny based on a geometric approach. Molecular phylogenetics and evolution. 2005; 36(2): 224–232.

[35] Yau SST, Yu C, He R. A protein map and its application. DNA and cell biology. 2008; 27(5): 241–250).

[36] Wen J, Zhang Y. A 2D graphical representation of protein sequence and its numerical characterization. Chemical Physics Letters. 2009; 476(4-6): 281–286.

[37] Yu C, Liang Q, Yin C, *et al*. A novel construction of genome space with biological Geometry. DNA Research. 2010; 17(3): 155–168.

[38] Lobry JR. A simple vectorial representation of DNA sequences for the detection of replication origins in bacteria. Biochimie. 1996; 78(5): 323–326.

[39] Zhang CT, Wang J. Recognition of protein coding genes in the yeast genome at better than 95% accuracy based on the Z curve. Nucleic acids research. 2000; 28(14): 2804–2814.

[40] Randic M, Vracko M, Lers N, Plavsic D. Analysis of similarity/dissimilarity of DNA sequences based on novel 2-D graphical representation. Chemical Physics Letters. 2003; 371(1–2): 202-207.

[41] Randic M, Vracko M, Lers N, Plavsic D. Novel 2-D graphical 269 representation of DNA sequences and their numerical characterization. Chemical Physics Letters. 2003; 368(1–2): 1–6.

[42] Liao B, Li R, Zhu W, Xiang X. On the similarity of DNA primary sequences based on 5-D representation. Journal of Mathematical Chemistry. 2007; 42(1): 47–57.

[43] Qi Z, Qi X. Novel 2D graphical representation of DNA sequence based on dual nucleotides. Chemical Physics Letters. 2007; 440(1–3): 139–144.

[44] Qi XQ, Wen J, Qi ZH. New 3D graphical representation of DNA sequence based on dual nucleotides. Journal of theoretical biology. 2007; 249(4): 681–690.

[45] Qi X, Wu Q, Zhang Y, *et al*. A novel model for DNA sequence similarity analysis based on graph theory. Evolutionary Bioinformatics Online. 2011; 7: 149–158.

[46] Blaisdell B. A measure of the similarity of sets of sequences not requiring sequence alignment. Proceedings of the National Academy of Sciences of the United States of America. 1986; 83(14): 5155–5159.

[47] Churchill G. Hidden Markov chains and the analysis of genome structure. Computers & Chemistry. 1992; 16(2): 107–115.

[48] Pham TD, Zuegg J. A probabilistic measure for alignment-free sequence comparison. Bioinformatics. 2004; 20(18): 3455–3461.

[49] Chang G, Wang T. Weighted relative entropy for alignment-free sequence comparison based on Markov model. Journal of Biomolecular Structure and Dynamics. 2011; 28(4): 545–555.

[50] Chang G, Wang H, Zhang T. A novel alignment-free method for whole genome analysis: Application to HIV-1 subtyping and HEV genotyping. Information Sciences. 2014; 279:776–784.

[51] Reese E, Krishnan VV. Classification of DNA sequences based on thermal melting profiles. Bioinformation. 2010; 4(10): 463–467.

[52] Albayrak A, Otu HH, Sezerman UO. Clustering of protein families into functional subtypes using Relative Complexity Measure with reduced amino acid alphabets. BMC Bioinformatics. 2010; 11: 428.

[53] Bastola DR, Otub HH, Doukas SE, *et al*. Utilization of the relative complexity measure to construct a phylogenetic tree for fungi. Mycological Research. 2004; 108(2): 117–125.

[54] Bakış Y, Otu HH, Taşçı N, *et al*. Testing robustness of relative complexity measure method constructing robust phylogenetic trees for *Galanthus L.* using the relative complexity measure. BMC Bioinformatics. 2013; 14: 20.

[55] Huang Y, Yang L, Wang T. Phylogenetic analysis of DNA sequences based on the generalized pseudo-amino acid composition. Journal of Theoretical Biology. 2011; 269(1): 217–223.

[56] Li B, Li YB, He HB. LZ complexity distance of DNA sequences and its application in phylogenetic tree reconstruction. Genomics, Proteomics & Bioinformatics. 2005; 3(4): 206–212.

[57] Liu J, Li D. Conditional LZ complexity of DNA sequences analysis and its application in phylogenetic tree reconstruction. Proceedings of the International Conference on BioMedical Engineering and Informatics. 2008; pp. 111–116.

[58] Otu HH, Sayood K. A new sequence distance measure for phylogenetic tree construction. Bioinformatics. 2003; 19(16): 2122–2130.

[59] Almeida JS. Sequence analysis by iterated maps, a review. Briefings in Bioinformatics. 2014; 15(3): 369–375.

[60] Vinga S. Information theory applications for biological sequence analysis. Briefings in Bioinformatics. 2014; 15(3): 376–389.

[61] Bonham-Carter O, Steele J, Bastola D. Alignment-free genetic sequence comparisons: a review of recent approaches by word analysis. Briefings in Bioinformatics. 2014; 15(6): 890–905.

[62] Blaisdell BE. Effectiveness of measures requiring and not requiring prior sequence alignment for estimating the dissimilarity of natural sequences. Journal of molecular evolution. 1989; 29(6): 526–537.

[63] Bao J, Yuan R, Bao Z. An improved alignment-free model for DNA sequence similarity metric. BMC Bioinformatics. 2014; 15(1): 321.

[64] Leimeister CA, Boden M, Horwege S, *et al.* Fast alignment-free sequence comparison using spaced-word frequencies. Bioinformatics. 2014; 30(14): 1991–1999.

[65] Sims GE, Jun SR, Wu GA, Kim SH. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. In: Proceedings of the National Academy of Sciences of the USA. 2009; 106(8):2677–2682.

[66] Wu G, Jun SR, Sims G, Kim SH. Whole-proteome phylogeny of large dsDNA virus families by an alignment-free method. Proceedings of the National Academy of Sciences of the United States of America. 2009; 106(31):12826–12831.

[67] Vliet AHMV, Kusters J. Use of alignment-free phylogenetics for rapid genome sequence-based typing of Helicobacter pylori virulence markers and antibiotic susceptibility. Journal of Clinical Microbiology. 2015; 53(9):2877–2888.

[68] Almeida J, Carriço JA, Maretzek A, *et al.* Analysis of genomic sequences by Chaos Game Representation. Bioinformatics. 2001; 17(5): 429–37.

[69] Karamichalis R, Kari L. MoDMaps3D: an interactive webtool for the quantification and3D visualization of interrelationships in a dataset of DNA sequences. Bioinformatics. 2017; 33(19): 3091–3093.

[70] Solis-Reyes S, Avino M, Poon A, Kari L. An open-source k-mer based ma-chine learning tool for fast and accurate subtyping of HIV-1 genomes. PLoS One. 2018; 13(11): e0206409.

[71] Wang Y, Hill K, Singh S, Kari L. The spectrum of genomic signatures: From dinucleotides to chaos game representation. Gene. 2005; 346:173–185.

[72] Hatje K, Kollmar M. A phylogenetic analysis of the brassicales clade based on an alignment-free sequence comparison method. Frontiers in Plant Science. 2012; 3.

[73] Deschavanne P, Giron A, Vilain J, Fagot G, Fertil B. Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. Molecular Biology and Evolution. 1999; 16(10):1391–1399.

[74] Deschavanne P, Giron A, Vilain J, Dufraigne C, Fertil B. Genomic signature is preserved in short DNA fragments. Proceedings IEEE International Symposium on Bio-Informatics and Biomedical Engineering. 2000; 161–167.

[75] Edwards S, Fertil B, Giron A, Deschavanne P. A genomic schism in birds revealed by phylogenetic analysis of DNA strings. Systematic Biology. 2002; 51(4):599–613.

[76] Deschavanne P, DuBow M, Regeard C. The use of genomic signature distance between bacteriophages and their hosts displays evolutionary relationships and phage growth cycle determination. Virology Journal. 2010; 7:163.

[77] Pandit A, Sinha S. Using genomic signatures for HIV-1 sub-typing. BMC Bioinformatics. 2010; 11: S26.

[78] Jeffrey HJ. Chaos game representation of gene structure. Nucleic Acids Res. 1990; 18: 2163–2170.

[79] Shmulevich I, Dougherty ER. Genomic Signal Processing. Princeton University Press, Princeton, New Jersey, USA. 2007.

[80] Chakravarthy N, Spanias A, Iasemidis LD, Tsakalis K. Autoregressive modeling and feature analysis of DNA sequences. EURASIP Journal on Applied Signal Processing. 2004; 2004: 13–28.

[81] Yu Z, Anh VV, Zhou Y, Zhou LQ. Numerical sequence representation of DNA sequences and methods to distinguish coding and non-coding sequences in a complete genome. In:Proceedings 11th World Multi-Conference on Systemics, Cybernetics and Informatics. 2007; p. 171–176.

[82] Abo-Zahhad M, Ahmed S, Abd-Elrahman S. Genomic analysis and classification of exon and intron sequences using DNA numerical mapping techniques. International Journal of Information Technology and Computer Science. 2012; 4(8): 22–36.

[83] Yin C, Yau SST. An improved model for whole genome phylogenetic analysis by Fourier transform. Journal of Theoretical Biology. 2015; 382: 99–110.

[84] Skutkova H, Vitek M, Sedlar K, Provaznik I. Progressive alignment of genomic signals by multiple dynamic time warping. Journal of Theoretical Biology. 2015; 385: 20–30.

[85] Borrayo E, Mendizabal-Ruiz EG, Vélez-Pérez H, *et al.* Genomic signal processing methods for computation of alignment-free distances from DNA sequences. PLoS One. 2014; 9(11): e110954.

[86] Yin C, Chen Y, Yau STS. A measure of DNA sequence similarity by Fourier Transform with applications on hierarchical clustering. Journal of Theoretical Biology. 2014; 359: 18–28.

[87] Cristea PD. Large scale features in DNA genomic signals. Signal Processing. 2003; 83(4): 871–888.

[88] Skutkova H, Vitek M, Babula P, *et al.* Classification of genomic signals using dynamic time warping. BMC Bioinformatics. 2013; 14(10): S1.

[89] Kwan HK, Arniker SB. Numerical representation of DNA sequences. In: 2009 IEEE International Conference on Electro/Information Technology. 2009; p. 307–310.

[90] Hoang T, Yin C, Yau SS. Numerical encoding of DNA sequences by Chaos Game Representation with application in similarity comparison. Genomics. 2016; 108(3): 134–142.

[91] Mendizabal-Ruiz G, Román-Godínez I, Torres-Ramos S, *et al.* On DNA numerical representations for genomic similarity computation. PLoS One. 2017; 12(3): e0173288.

[92] Karlin S, Burge C. Dinucleotide relative abundance extremes: a genomic signature. Trends in Genetics. 1995; 11(7): 283–290.

[93] Cristea PD. Conversion of nucleotide sequences into genomic signals. Journal of Cellular and Molecular Medicine. 2002; 6(2): 279–303.

[94] Lorenzo-Ginori JV, Rodriguez-Fuentes A, Grau Abalo R, Sanchez Rodriguez R. Digital signal processing in the analysis of genomic sequences. Current Bioinformatics. 2009; 4(1): 28–40.

[95] Adetiba E, Olugbara OO. Classification of eukaryotic organisms through cepstral analysis of mitochondrial DNA. In: International Conference on Image and Signal Processing. 2016; vol.9680, p. 243–252.

[96] Adetiba E, Olugbara OO, Taiwo TB. Identification of pathogenic viruses using genomic cepstral coefficients with radial basis function neural network. In: Proceedings of the 7th World Congress on Nature and Biologically Inspired Computing. 2016; vol. 419, p. 281–290.

[97] Strang, G. Wavelets. American Scientist. 1994; 82 (3): 250–255.

[98] Iversen GR, Gergen M, Gergen MM. Statistics: The Conceptual Approach. Berlin Heidelberg: Springer; 1997.

[99] Krause EF. Taxicab Geometry: An Adventure in Non-Euclidean geometry. Courier Dover Publications, New York, USA. 2012.

[100] Asuero AG, Sayago A, González AG. The correlation coefficient: an overview. Critical Reviews in Analytical Chemistry. 2006; 36(1): 41–59.

[101] El-Badawy IM, Aziz AM, Omar Z, Malarvili MB. Correlation between different DNAperiod-3 signals: An analytical study for exons prediction. In: 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference. 2017; p. 1123–1128.

[102] Kari L, Hill KA, Sayem AS, *et al*. Mapping thespace of genomic signatures. PLoS One. 2015; 10(5): e011981

[103] Karamichalis R, Kari L, Konstantinidis S, Kopecki S. An investigation into inter- and intragenomic variations of graphic genomic signatures. BMC Bioinformatics. 2015; 16: 246.

[104] Fisher RA. The Use of Multiple Measurements in Taxonomic Problems. Annals of Eugenics. 1936; 7(2): 179–188

[105] Christianini N, Shawe-Taylor JC. An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods. Cambridge University Press, Cambridge, UK. 2000.

[106] Scholkopf B, Smola A. Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond. Adaptive Computation and Machine Learning. The MIT Press, Cambridge, MA. 2002.

[107] Friedman JH, Bentely J, Finkel RA. An Algorithm for Finding Best Matches in Logarithmic Expected Time. ACM Transactions on Mathematical Software 3. 1977; 3: 209–226.

[108] Altman NS. An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician. 1992; 46(3): 175–185.

[109] Ho TK. The random subspace method for constructing decision forests. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1998; 20(8): 832–844.

Chapter 3

ML-DSP: Machine Learning with Digital Signal Processing for ultrafast, accurate, and scalable genome classification at all taxonomic levels

3.1 Background

Of the estimated 8.7 million (±1.3 million) species existing on Earth [1], only around 1.5 million distinct eukaryotes have been catalogued and classified so far [2], leaving 86% of existing species on Earth and 91% of marine species still unclassified. To address the grand challenge of all species identification and classification, a multitude of techniques have been proposed for genomic sequence analysis and comparison. These methods can be broadly classified into alignment-based and alignment-free. Alignment-based methods and software tools are numerous, and include, e.g., MEGA7 [3] with sequence alignment using MUSCLE [4], or CLUSTALW [5, 6]. Though alignment-based methods have been used with significant success for genome classification, they have limitations [7] such as the heavy time/memory computational cost for multiple alignment in multigenome scale sequence data, the need for continuous homologous sequences, and the dependence on a priori assumptions on, e.g., the gap penalty and threshold values for statistical parameters [8]. In addition, with next-generation sequencing (NGS) playing an increasingly important role, it may not always be possible to align many short reads coming from different parts of genomes [9]. To address situations where alignment-based methods fail or are insufficient, alignment-free methods have been proposed [10], including approaches based on Chaos Game Representation of DNA sequences [11, 12, 13], random walk [14], graph theory [15], iterated maps [16], information theory [17], category-position-frequency [18], spaced-words frequencies [19], Markov-model [20], thermal melting profiles [21], word analysis [22], among others. Software implementations of alignment-free methods also exist, among them COMET [23], CASTOR [24], SCUEAL [25], REGA [26], KAMERIS [27], and FFP (Feature Frequency Profile) [28]. While alignment-free methods address some of the issues faced by alignment-based methods, [7] identified the following challenges they face:

(i) **Lack of software implementation:** Most of the existing alignment-free methods are still exploring technical foundations and lack software implementation, which is necessary for methods to be compared on common datasets.

(ii) **Use of simulated sequences or very small real world datasets:** The majority of the existing alignment-free methods are tested using simulated sequences or very small real-world datasets. This makes it hard for experts to pick one tool over the others.

(iii) **Memory overhead:** Scalability to multigenome data can cause memory overhead in word-based methods, especially when long *k*-mers are used.

To overcome these challenges, we propose ML-DSP, a novel combination of supervised **M**achine **L**earning with **D**igital **S**ignal **P**rocessing of the input DNA sequences, as a general-purpose alignment-free method and software tool for genomic DNA sequence classification at all taxonomic levels.

The main contribution of ML-DSP is the *feature vector* that we propose to be used by the supervised learning algorithms. Given a genomic DNA sequence, its feature vector consists of the pairwise Pearson Correlation Coefficient (PCC) between (a) the magnitude spectrum of the Discrete Fourier Transform (DFT) of the digital signal obtained from the given sequence by some suitable numerical encoding of the letters A, C, G, T into numbers, and (b) the magnitude spectra of the DFT of all the other genomic sequences in the training set. The use of this new feature vector, which has not previously been used in conjunction with machine learning algorithms, allows ML-DSP to significantly outperform existing methods in terms of speed, while achieving an average classification accuracy of $> 97\%$. This substantial performance improvement allows ML-DSP to scale up and successfully classify much larger datasets than existing studies. Indeed, in contrast with previous benchmark datasets, each comprising less than fifty sequences, this study accurately classifies thousands of genomes from a variety of species: eukaryotic (7,396 complete mitochondrial genomes), viral (4,271 genomes), and bacterial (4,710 genomes). In addition, this study provides the first comprehensive analysis and comparison of all thirteen one-dimensional numerical representations of DNA sequences used in the Genomic Signal Processing (GSP: digital signal processing applied to genomes) literature for classification purposes. We conclude that the “Purine/Pyrimidine (PP)”, “Just-A”, and “Real” numerical representations are the top three performers in terms of classification accuracy of ML-DSP for our main dataset. This is surprising given that these three numerical representations do not appear to contain sufficient biological information for the accuracy attained. For example, the numerical representation “Just-A” (encoding A as “1”, and G, C, T as “0”) retains the incidence and spacing for A, but not individually for the other three nucleotides.

3.1.1 Numerical representations of DNA sequences

Digital Signal Processing (DSP) can be employed in the context of comparative genomics because genomic sequences can be numerically represented as discrete numerical sequences and hence treated as digital signals. Several numerical representations of DNA sequences, that use numbers assigned to individual nucleotides, have been proposed in the literature [29], e.g., based on a fixed mapping of each nucleotide to a number, without biological significance; using mappings of nucleotides to numerical values deduced from their physio-chemical properties; or using numerical values deduced from the doublets or codons that the individual nucleotide was part of [29, 30]. In [31, 32] three physio-chemical based representations of DNA sequences (atomic, molecular mass, and Electron-Ion Interaction Potential, EIIP) were considered for genomic analysis, and the authors concluded that the choice of numerical representation did not have any effect on the results. A recent study comparing different numerical representation techniques on a small dataset [33] concluded that multi-dimensional representations (such as Chaos Game Representation) yielded better genomic comparison results than some one-dimensional representations. However, in general there is no agreement on whether or not the choice of numerical representation for DNA sequences makes a difference on the genome comparison results, or on which numerical representations are best suited for analyzing genomic data. We address this issue by providing a comprehensive analysis and comparison of thirteen one-dimensional numerical representations, for suitability in genome analysis.

3.1.2 Digital Signal Processing

Following the choice of a suitable numerical representation for DNA sequences, DSP techniques can be applied to the resulting discrete numerical sequences, and the whole process has been termed Genomic Signal Processing (GSP) [30]. DSP techniques have previously been used for DNA sequence comparison, e.g., to distinguish coding regions from non-coding regions [34, 35, 36], to align genomic signals for classification of biological sequences [37], for whole genome phylogenetic analysis [38], and to analyze other properties of genomic sequences [39]. In our approach, genomic sequences are represented as discrete numerical sequences, treated as digital signals, transformed via DFT into corresponding magnitude spectra, and compared via Pearson Correlation Coefficient (PCC) to create a pairwise distance matrix.

3.1.3 Supervised Machine Learning

Machine learning has been used in small-scale genomic analysis studies [40, 41, 42], and classification analyses associated with microarray gene expression data [43, 44, 45]. In this vein, ML-DSP focusses on the use of the primary DNA sequence data for taxonomic classification, and is based on a novel combination of supervised machine learning with feature vectors consisting of the pairwise distances between the magnitude spectrum of the DFT obtained from the digital signal generated from a DNA sequence, and the magnitude spectra of the DFT of the digital signals generated from all other sequences in the training set. The taxonomic labels of sequences are provided for training purposes. Six supervised machine learning classifiers (Linear Discriminant, Linear SVM, Quadratic SVM, Fine KNN, Subspace Discriminant, and Subspace KNN) are trained on these pairwise distance vectors, and then used to classify new sequences. Independently, classical MultiDimensional Scaling (MDS) generates a $3D$ visualization, called Molecular Distance Map (MoDMap) [46], of the interrelationships among all sequences.

For our computational experiments, we used a large dataset of $7,396$ complete mtDNA sequences, and six different classifiers, to compare one-dimensional numerical representations for DNA sequences used in the literature for classification purposes. For this dataset, we concluded that the "PP", "Just-A", and "Real" numerical representations were the best numerical representations. We analyzed the performance of ML-DSP in classifying the aforementioned genomic mtDNA sequences, from the highest level (domain into kingdoms) to lower level (family into genera) taxonomical ranks. The average classification accuracy of ML-DSP was $> 97\%$ when using the "PP", "Just-A", and "Real" numerical representations.

To evaluate our method, we compared its performance (accuracy and speed) on three datasets: two previously used small benchmark datasets [47], and a large real world dataset of $4,322$ complete vertebrate mtDNA sequences. We found that ML-DSP had significantly better accuracy scores than the alignment-free method FFP on all datasets. When compared to the state-of-the-art alignment-based method MEGA7 (with alignment using MUSCLE or CLUSTALW), ML-DSP achieved similar accuracy but superior processing times (2,250 to 67,600 times faster) for the small benchmark dataset of 41 mammalian genomes. The contrast in running time was even more extreme for the large dataset of 4,322 mtDNA genomes, where ML-DSP took 28 seconds, while MEGA7(MUSCLE/CLUSTALW) could not complete the computation after 2 hours/6 hours and had to be terminated.

Lastly, we provide preliminary computational experiments that indicate the potential of ML-DSP to successfully classify viral genomes (4,271 complete dengue virus genomes into four subtypes) and bacterial genomes (4,710 complete bacterial genomes into three phyla).

3.2 Methods and Implementation

The main idea behind ML-DSP is to combine supervised machine learning techniques with digital signal processing, for the purpose of DNA sequence classification. More precisely, for a given set $S = \{S_1, S_2, \ldots, S_n\}$ of n DNA sequences, ML-DSP uses:

- DNA numerical representations to obtain a set $N = \{N_1, N_2, \ldots, N_n\}$ where N_i is a discrete numerical representation of the sequence S_i, $1 \leq i \leq n$.

- Discrete Fourier Transform (DFT) applied to the length-normalized digital signals N_i, to obtain the frequency distribution; the magnitude spectrum M_i of this frequency distribution is then obtained.

- Pearson Correlation Coefficient (PCC) to compute the distance matrix of all pairwise distances for each pair of magnitude spectra (M_i, M_j), where $1 \leq i, j \leq n$.

- Supervised Machine Learning classifiers which take the pairwise distance matrix for a set of sequences, together with their respective taxonomic labels, in a training set, and output the taxonomic classification of a new DNA sequence. To measure the performance of such a classifier, we use the 10-fold cross-validation technique.

- Independently, Classical Multidimensional Scaling (MDS) takes the distance matrix as input and returns an $(n \times q)$ coordinate matrix, where n is the number of points (each point represents a unique sequence from set S) and q is the number of dimensions. The first three dimensions are used to display a MoDMap, which is the simultaneous visualization of all points in $3D$-space.

3.2.1 DNA numerical representations

To apply digital signal processing techniques to genomic data, genomic sequences are first mapped into discrete numerical representations of genomic sequences, called *genomic signals* [48]. In our analysis of various numerical representations for DNA sequences (Table 3.1), we considered only $1D$ numerical representations, that is, those which produce a single output numerical sequence, called also *indicator sequence*, for a given input DNA sequence.

We did not consider other numerical representations, such as binary [29], or nearest dissimilar nucleotide [49], because those generate four numerical sequences for each genomic sequence, and would thus not be scalable to classifications of thousands of complete genomes.

3.2.2 Discrete Fourier Transform (DFT)

Our alignment-free classification method of DNA sequences makes use of the DFT magnitude spectra of the discrete numerical sequences (discrete digital signals) that represent DNA sequences. In some sense, these DFT magnitude spectra reflect the nucleotide distribution of the originating DNA sequences.

To start with, assuming that all input DNA sequences have the same length p, for each DNA sequence $S_i = (S_i(0), S_i(1), \ldots, S_i(p-1))$, in the input dataset, where $1 \le i \le n$, $S_i(k) \in$

Table 3.1: Numerical representations of DNA sequences.

#	Representation	Rules	Output for $S_1 = CGAT$
1	Integer	$T = 0,\ C = 1,\ A = 2,\ G = 3$	[1 3 2 0]
2	Integer (other variant)	$T = 1,\ C = 2,\ A = 3,\ G = 4$	[2 4 3 1]
3	Real	$T = -1.5,\ C = 0.5,\ A = 1.5,\ G = -0.5$	[0.5 − 0.5 1.5 − 1.5]
4	Atomic	$T = 6,\ C = 58,\ A = 70,\ G = 78$	[58 78 70 6]
5	EIIP (electron-ion interaction potential)	$T = 0.1335,\ C = 0.1340,\ A = 0.1260,\ G = 0.0806$	[0.1340 0.8060 0.1260 0.1335]
6	PP (purine/pyrimidine)	$T/C = 1,\ A/G = -1$	[1 − 1 − 1 1]
7	Paired numeric	$T/A = 1,\ C/G = -1$	[−1 − 1 1 1]
8	Nearest-neighbor based doublet	0 – 15 for all possible doublets	[14 8 1 7]
9	Codon	0 – 63 for all possible 64 Codons	[2 35 22 44]
10	Just-A	$A = 1,\ rest = 0$	[0 0 1 0]
11	Just-C	$C = 1,\ rest = 0$	[1 0 0 0]
12	Just-G	$G = 1,\ rest = 0$	[0 1 0 0]
13	Just-T	$T = 1,\ rest = 0$	[0 0 0 1]

Numerical representations of DNA sequences analyzed for usability in genomic classification with ML-DSP. The second column lists the numerical representation name, the third column describes the rule it uses, and the fourth is the output of this rule for the input DNA sequence $S_1 = CGAT$. For the nearest-neighbor based doublet representation and codon representation, the DNA sequence is considered to be wrapped (the last position is followed by the first).

$\{A, C, G, T\}$, $0 \leq k \leq p - 1$, we calculate its corresponding discrete numerical representation (discrete digital signal) N_i defined as

$$N_i = (f(S_i(0)), f(S_i(1)), \ldots, f(S_i(p-1)))$$

where, for each $0 \leq k \leq p - 1$, the quantity $f(S_i(k))$ is the value under the numerical representation f of the nucleotide in the position k of the DNA sequence S_i.

Then, the DFT of the signal N_i is computed as the vector F_i where, for $0 \leq k \leq p - 1$ we have

$$F_i(k) = \sum_{j=0}^{p-1} f(S_i(j)) \cdot e^{(-2\pi i/p)kj} \tag{3.1}$$

The magnitude vector corresponding to the signal N_i can now be defined as the vector M_i where, for each $0 \leq k \leq p - 1$, the value $M_i(k)$ is the absolute value of $F_i(k)$, that is, $M_i(k) = |F_i(k)|$. The magnitude vector M_i is also called the magnitude spectrum of the digital signal N_i

and, by extension, of the DNA sequence S_i. For example, if the numerical representation f is Integer (row 1 in Table 3.1), then for the sequence $S_1 = CGAT$, the corresponding numerical representation is $N_1 = (1, 3, 2, 0)$, the result of applying DFT is $F_1 = (6,\ -1 - 3i,\ 0,\ -1 + 3i)$ and its magnitude spectrum is $M_1 = (6,\ 3.1623,\ 0,\ 3.1623)$.

Fig 3.1a shows the discrete digital signal (using the "PP" numerical representation, row 6 of Table 3.1) of the DNA sequence consisting of the first 100 bp of the mtDNA genome of *Branta canadensis* (Canada goose, NCBI accession number NC_007011.1), and of the DNA sequence consisting of the first 100 bp of the mtDNA genome of *Castor fiber* (European beaver; NCBI accession number NC_028625.1). Fig 3.1b shows the DFT magnitude spectra of the same two signals/sequences. As can be seen in Fig 3.1b, these mtDNA sequences exhibit different DFT magnitude spectrum patterns, and this can be used to distinguish them computationally by using. e.g., the Pearson Correlation Coefficient, as described in the next subsection. Other techniques have also been used for genome similarity analysis, for example comparing the phase spectra of the DFT of digital signals of full mtDNA genomes, as seen in Fig 3.2 and [50, 51].

Note that, with the exception of the example in Fig 3.1, all of the computational experiments in this paper use full genomes.

3.2.3 Pearson Correlation Coefficient (PCC)

Consider two variables X and Y (here X and Y are the magnitude spectra M_i and M_j of two signals), each of length p, that is, $X = \{X_0, \ldots, X_{p-1}\}$ and $Y = \{Y_0, \ldots, Y_{p-1}\}$. The Pearson Correlation Coefficient r_{XY} between X and Y is the ratio of their covariance (measure of how much X and Y vary together) to the product of their standard deviations [52, 53], that is,

$$r_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \tag{3.2}$$

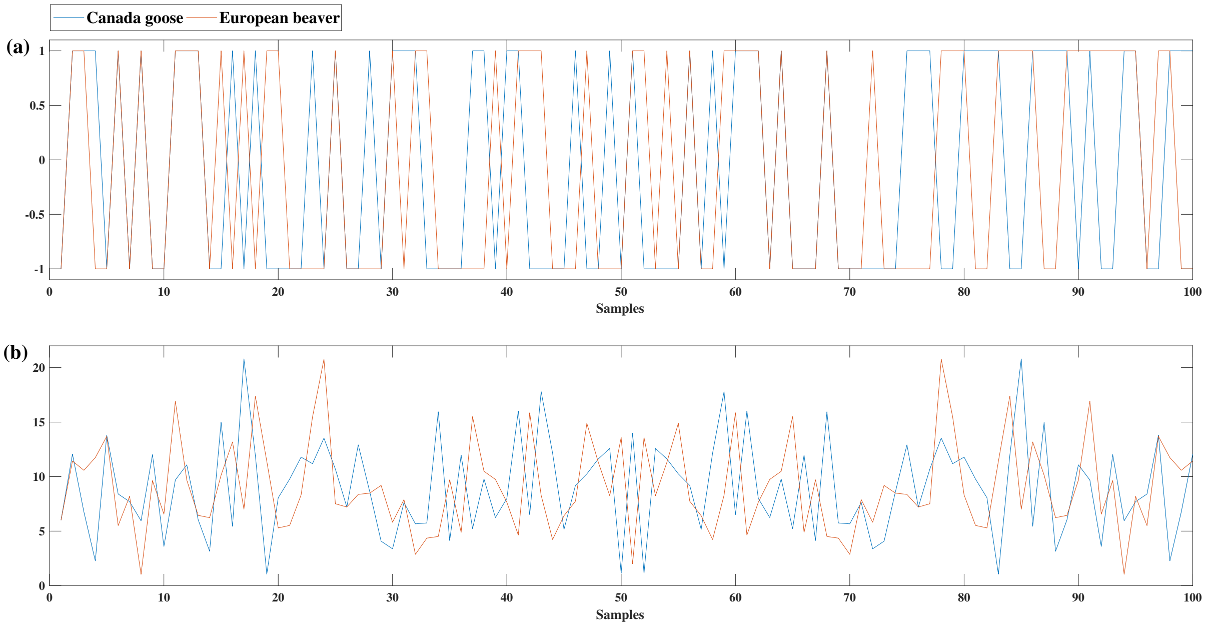


Figure 3.1: Canada goose (blue) vs European beaver (red): comparison of the DFT magnitude spectra of the first 100 bp of their mtDNA genomes. (**a**): Graphical illustration of the discrete digital signals of the respective DNA sequences, obtained using the “PP” representation. (**b**): DFT magnitude spectra of the signals in (**a**).

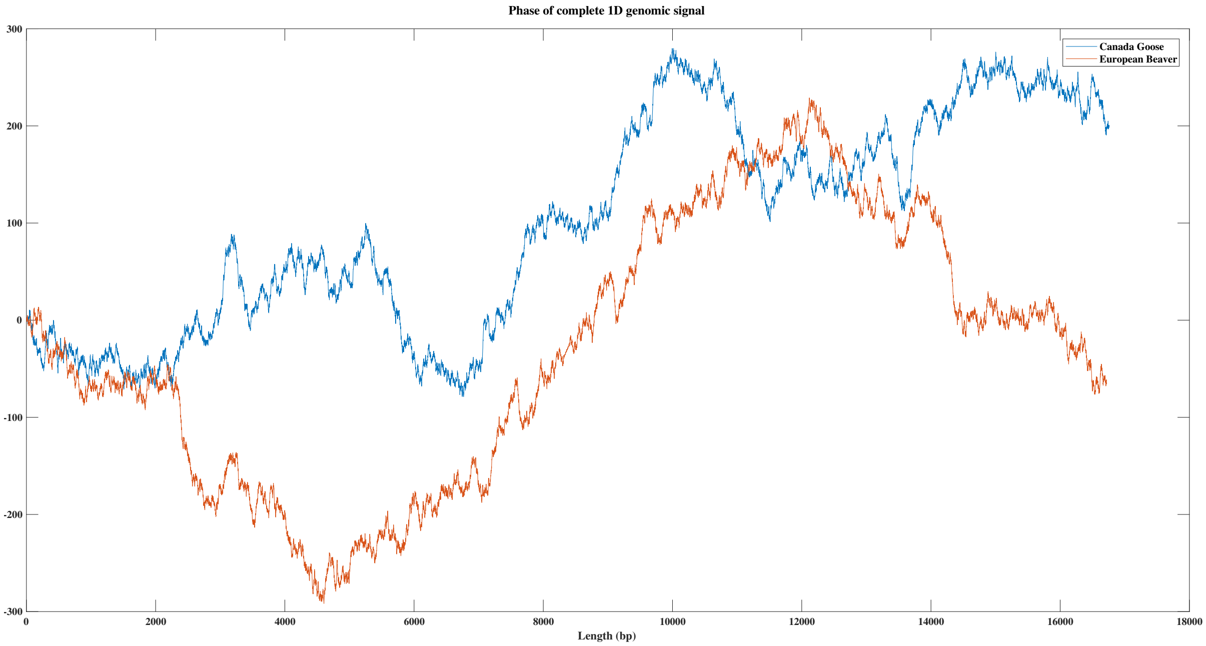


Figure 3.2: Canada goose (blue, 16,760 bp) vs. European beaver (red, 16,722 bp) - comparison between the DFT phase spectra of their full mtDNA genomes.

where the covariance of X and Y is $\sigma_{XY} = \sum_{i=0}^{p-1}(X_i - \overline{X})(Y_i - \overline{Y})/(p-1)$, and the standard deviation is $\sigma_X = \sqrt{\sum_{i=0}^{p-1}(X_i - \overline{X})^2/(p-1)}$, and similarly for σ_Y, where the average is defined as $\overline{X} = (\sum_{i=0}^{p-1} X_i)/p$ and similarly for Y. Now the formula for the Pearson Correlation Coefficient becomes

$$r_{XY} = \frac{\sum_{i=0}^{p-1}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=0}^{p-1}(X_i - \overline{X})^2} \times \sqrt{\sum_{i=0}^{p-1}(Y_i - \overline{Y})^2}} \tag{3.3}$$

The Pearson Correlation Coefficient between X and Y is a measure of their linear correlation, and has a value between +1 (total positive linear correlation) and −1 (total negative linear correlation); 0 is no linear correlation. We normalized the results, by taking $(1 - r_{XY})/2$, to obtain distance values between 0 and 1 (value 0 for identical signals, and 1 for negatively correlated signals). For our data sets, the PCC values between any two digital signals of DNA sequences ranged between 0 and 0.6.

For each pairwise distance calculation, the Pearson Correlation Coefficient requires the input variables (that is, the magnitude spectra of the two sequences) to have the same length. The length of a magnitude spectrum is equal to the length of corresponding numerical digital signal, which in turn is equal to the length of the originating DNA sequence. Given that genome sequences are typically of different lengths, it follows that their corresponding digital signals need to be length-normalized, if we are to be able to use the Pearson Correlation Coefficient. Hoang et al. avoided normalization and considered only the first few mathematical moments constructed from the power spectra for comparison, after applying DFT [54]. The limitation of this method is that one loses information that may be necessary for a meaningful comparison. This is especially important when the genomes compared are very similar to each other.

Different methods for length-normalizing digital signals were tested: down-sampling [55], up-sampling to the maximum length using zero padding [30], even scaling extension [56], periodic extension, symmetric padding, or anti-symmetric padding [57]. For example, zero-padding, which adds zeroes to all of the sequences shorter than the maximum length, was used in [30], e.g., for taxonomic classifications of ribosomal S18 subunit genes from twelve organisms. While this method may work for datasets of sequences of similar lengths, it is not suitable for datasets of sequences of very different lengths (our study: fungi mtDNA genomes dataset - 1,364 bp to 235,849 bp; plant mtDNA genomes dataset - 12,998 bp to 1,999,595 bp; protist mtDNA genomes dataset - 5,882 bp to 77,356 bp). In such cases, zero-padding acts as a tag and may lead to inadvertent classification of sequences based on their length rather than based on their sequence composition. Thus, we employed instead anti-symmetric padding, whereby, starting from the last position of the signal, boundary values are replicated in an anti-symmetric manner. We also considered two possible ways of employing anti-symmetric padding: normalization to the maximum length (where shorter sequences are extended to the maximum sequence length by anti-symmetric padding) vs. normalization to the median length (where shorter sequences are extended by anti-symmetric padding to the median length, while longer sequences are truncated after the median length).

3.2.4 Supervised Machine Learning

In this paper we used the Linear discriminant, Linear SVM, Quadratic SVM, Fine KNN, Subspace discriminant and Subspace KNN classifiers from the Classification Learner application of MATLAB (Statistics and Machine Learning Toolbox). The default MATLAB parameters were used.

To assess the performance of the classifiers, we used 10-fold cross validation. In this approach, the dataset is randomly partitioned into 10 equal-size subsets. The classifier is trained using 9 of the subsets, and the accuracy of its prediction is tested on the remaining subset. As part of the supervised learning, taxonomic labels are supplied for the DNA sequences in the 9 subsets used for training. The process is repeated 10 times, and the accuracy score of the classifier is then computed as the average of the accuracies obtained in the 10 separate experiments. The standard algorithms were modified so that no information about sequences in the testing set (that is, no distance matrix entries containing distances to/from any sequence in the testing set to any other sequence) was available during the training stage.

3.2.5 Classical Multidimensional Scaling (MDS)

Classical multidimensional scaling takes a pairwise distance matrix ($n \times n$ matrix, for n input items) as input, and produces n points in a q-dimensional Euclidean space, where $q \leq n - 1$. More specifically, the output is an $n \times q$ coordinate matrix, where each row corresponds to one of the n input items, and that row contains the q coordinates of the corresponding item-representing point [11]. The Euclidean distance between each pair of points is meant to approximate the distance between the corresponding two items in the original distance matrix.

These points can then be simultaneously visualized in a 2- or 3-dimensional space by taking the first 2, respectively 3, coordinates (out of q) of the coordinate matrix. The result is a Molecular Distance Map [46], and the MoDMap of a genomic dataset represents a visualization of the simultaneous interrelationships among all DNA sequences in the dataset.

3.2.6 Software implementation

The algorithms for ML-DSP were implemented using the software package MATLAB R2017A, license no. 964054, as well as the open-source toolbox Fathom Toolbox for MATLAB [58] for distance computation. All software can be downloaded from https://github.com/grandhawa/MLDSP. The user can use this code to reproduce all results in this paper, and also has the option to input their own dataset and use it as training set for the purpose of classifying new genomic DNA sequences.

All experiments were performed on an ASUS ROG G752VS computer with 4 cores (8 threads) of a 2.7GHz Intel Core i7 6820HK processor and 64GB DD4 2400MHz SDRAM.

3.2.7 Datasets

All datasets in this paper can be found at https://github.com/grandhawa/MLDSP in the "DataBase" directory. The mitochondrial dataset comprises all of the 7,396 complete reference mtDNA sequences available in the NCBI Reference Sequence Database RefSeq on June 17, 2017. We performed computational experiments on several different subsets of this dataset. The bacteria dataset comprises all 4,710 complete bacterial genomes with lengths between 20,000 bp and 500,000 bp, available in the aforementioned NCBI database on the same date. The dengue virus dataset contained all 4,721 dengue virus genomes available in the NCBI database on August 10, 2017. Note that any letters "N" in these DNA sequences were deleted.

For the performance comparison between ML-DSP and other alignment-free and alignment-based methods we also used the benchmark datasets of 38 influenza virus sequences, and 41 mammalian complete mtDNA sequences from [47].

3.3 Results and Discussion

Following the design and implementation of the ML-DSP genomic sequence classification tool prototype, we investigated which type of length-normalization and which type of distance were most suitable for genome classification using this method. We then conducted a comprehensive analysis of the various numerical representations of DNA sequences used in the literature, and determined the top three performers. Having set the main parameters (length-normalization method, distance, and numerical representation), we tested ML-DSP's ability to classify mtDNA genomes at taxonomic levels ranging from the domain level down to the genus level, and obtained average levels of classification accuracy of $> 97\%$. Finally, we compared ML-DSP with other alignment-based and alignment-free genome classification methods, and showed that ML-DSP achieved higher accuracy and significantly higher speeds.

3.3.1 Analysis of distances and of length normalization approaches

To decide which distance measure and which length normalization method were most suitable for genome comparisons with ML-DSP, we used nine different subsets of full mtDNA sequences from our dataset. These subsets were selected to include most of the available complete mtDNA genomes (Vertebrates dataset of 4,322 mtDNA sequences), as well as subsets containing similar sequences, of similar length (Primates dataset of 148 mtDNA sequences), and subsets containing mtDNA genomes showing large differences in length (Plants dataset of 174 mtDNA sequences).

The classification accuracy scores obtained using the two considered distance measures (Euclidean and Pearson Correlation Coefficient) and two different length-normalization approaches (normalization to maximum length and normalization to median length) on several datasets are listed in Table 3.2. The classification accuracy scores are slightly higher for PCC, but sufficiently close to those obtained when using the Euclidean distance to be inconclusive.

In the remainder of this paper we chose the Pearson Correlation Coefficient because it is scale independent (unlike the Euclidean distance, which is, e.g., sensitive to the offset of the signal, whereby signals with the same shape but different starting points are regarded as dissimilar [59]), and the length-normalization to median length because it is economic in terms of memory usage.

3.3.2 Analysis of various numerical representations of DNA sequences

We analyzed the effect on the ML-DSP classification accuracy of thirteen different one-dimensional numeric representations for DNA sequences, grouped as: Fixed mappings DNA numerical representations (Table 3.1 representations #1, #2, #3, #6, #7, see [29], and representations #10, #11, #12, #13 - which are one-dimensional variants of the binary representation proposed in [29]), mappings based on some physio-chemical properties of nucleotides (Table 3.1 representation #4, see [29, 32], and representation #5, see [29, 31, 32]), and mappings based on the nearest-neighbour values (Table 3.2 representations #8, #9, see [30]).

Table 3.2: Maximum classification accuracy scores when using Euclidean vs. Pearson's correlation coefficient (PCC) as a distance measure.

Data Set	**No. of Seq.**	**Max Length (bp)**	**Min Length (bp)**	**Median Length (bp)**	**Maximum Accuracy**			
					Euclidean		**PCC**	
					Norm. to Max Length (a)	**Norm. to Median Length (b)**	**Norm. to Max Length (c)**	**Norm. to Median Length (d)**
Primates (Haplorrhini: 115, Strepsirrhini: 33)	148	17531	15467	16554	98.6%	100%	100%	100%
Protists (Alveolata: 34, Rhodophyta: 46, Stramenopiles: 79)	159	77356	5882	35660	89.3%	90.6%	96.2%	91.2%
Fungi (Basidiomycota: 30, Pezizomycotina: 104, Saccharomycotina:92)	226	235849	1364	39154	70.1%	82.6%	87.9%	89.3%
Plants (Chlorophyta: 44, Streptophyta: 130)	174	1999595	12998	128211	95.4%	94.8%	90.2%	91.4%
Amphibians (Anura: 161, Caudata:95, Gymnophiona: 34)	290	28757	15757	17271	95.2%	97.6%	98.3%	99.0%
Mammals (Xenarthrans: 30, Bats: 54, Carnivores: 135, Even-toed Ungulates: 242, Insectivores: 40, Marsupials: 34, Primates: 148, Rodents and Rabbits: 147)	830	17734	15289	16537	95.2%	96.1%	97.8%	97.1%
Insects (Coleoptera: 95, Dictyptera: 77, Diptera: 149, Hemiptera: 126, Hymenoptera: 47, Lepidoptera:294, Orthoptera: 110)	898	20731	10662	15529	87.9%	90.0%	91.3%	94.2%
3 classes (Amphibians: 290, Mammals: 874, Insects: 1006)	2170	28757	8118	16361	99.9%	99.7%	99.8%	99.7%
Vertebrates (Amphibians: 290, Birds: 553, Fish: 2313, Mammals: 874, Reptiles: 292)	4322	28757	14935	16616	99.6%	99.8%	99.6%	99.7%
Table Average Accuracy	——	——	——	——	92.4%	94.6%	95.7%	95.7%

(a)(c) Genomes normalized to the maximum genome sequence length; (b)(d) Genomes normalized to the median genome sequence length

The datasets used for this analysis were the same as those in Table 3.2. The supervised machine learning classifiers used for this analysis were the six classifiers listed in the Methods and Implementation section, with the exception of the datasets with more than 2,000 sequences where two of the classifiers (Subspace Discriminant and Subspace KNN) were omitted as being too slow. The results and the average accuracy scores for all these numerical representations, classifiers and datasets are summarized in Table 3.3.

As can be observed from Table 3.3, for all numerical representations, the table average accuracy scores (last row: average of averages, first over the six classifiers for each dataset,

Table 3.3: Average classification accuracies for 13 numerical representations.

DataSet/ Classification Model	**Numerical Representation**												
	Integer	Integer (Other)	Real	Atomic	EIIP	PP	Paired Num.	NN based doublet	Codon	Just-A	Just-C	Just-G	Just-T
Primates (148 sequences)													
Linear Discriminant	97.3%	98.0%	99.3%	98.6%	99.3%	99.3%	97.3%	97.3%	98.0%	98.0%	97.3%	96.6%	96.6%
Linear SVM	97.3%	95.9%	98.6%	96.6%	97.3%	98.0%	95.9%	97.3%	94.6%	98.0%	96.6%	96.6%	95.3%
Quadratic SVM	97.3%	95.9%	98.6%	93.2%	95.9%	98.6%	96.6%	98.6%	95.9%	98.0%	98.0%	97.3%	95.9%
Fine KNN	98.0%	98.0%	100.0%	98.0%	96.6%	100.0%	99.3%	99.3%	98.0%	100.0%	98.6%	100.0%	98.6%
Subspace Discriminant	98.0%	97.3%	99.3%	98.0%	99.3%	98.6%	95.3%	97.3%	95.9%	98.0%	97.3%	98.0%	95.3%
Subspace KNN	98.0%	97.3%	98.6%	96.6%	95.9%	98.0%	100%	98.0%	98.0%	99.3%	97.3%	98.6%	98.6%
Average	**97.7%**	**97.1%**	**99.1%**	**96.8%**	**97.4%**	**98.8%**	**97.4%**	**98.0%**	**96.7%**	**98.6%**	**97.5%**	**97.9%**	**96.7%**
Protists (159 sequences)													
Linear Discriminant	83.6%	84.9%	85.5%	86.2%	86.2%	84.3%	85.5%	83.0%	85.5%	84.3%	83.6%	83.0%	83.6%
Linear SVM	84.3%	83.0%	83.6%	83.0%	83.0%	71.7%	82.4%	83.0%	83.6%	83.6%	83.6%	83.6%	83.0%
Quadratic SVM	84.9%	84.9%	83.6%	82.4%	83.0%	81.1%	85.5%	84.9%	86.2%	83.0%	84.3%	83.0%	86.2%
Fine KNN	86.8%	86.2%	81.8%	84.3%	88.1%	78.0%	89.9%	88.7%	91.8%	86.8%	88.7%	93.7%	92.5%
Subspace Discriminant	85.5%	84.9%	88.1%	86.8%	85.5%	86.8%	83.6%	83.0%	85.5%	84.9%	83.6%	83.0%	83.6%
Subspace KNN	88.7%	87.4%	91.8%	85.5%	88.1%	91.2%	89.9%	88.1%	93.1%	86.8%	88.1%	92.5%	93.7%
Average	**85.6%**	**85.2%**	**85.7%**	**84.7%**	**85.7%**	**82.2%**	**86.1%**	**85.1%**	**87.6%**	**84.9%**	**85.3%**	**86.5%**	**87.1%**
Fungi (226 sequences)													
Linear Discriminant	76.3%	76.8%	82.1%	50.9%	57.1%	80.4%	75.4%	68.8%	77.7%	81.7%	70.5%	71.9%	79.0%
Linear SVM	66.5%	58.0%	76.8%	49.1%	46.0%	73.7%	73.2%	66.1%	71.0%	75.9%	64.7%	66.1%	75.4%
Quadratic SVM	58.9%	59.8%	82.6%	33.9%	37.9%	79.9%	71.4%	67.4%	63.4%	71.0%	67.9%	71.4%	64.3%
Fine KNN	61.6%	56.7%	84.4%	49.6%	54.9%	85.7%	72.3%	65.2%	58.0%	68.8%	61.6%	68.8%	67.9%
Subspace Discriminant	74.6%	75.0%	78.6%	46.0%	55.4%	79.0%	75.0%	71.4%	78.1%	79.9%	68.8%	69.2%	78.6%
Subspace KNN	63.4%	58.9%	89.3%	51.8%	58.0%	89.3%	68.3%	63.8%	59.8%	67.9%	65.6%	72.8%	64.3%
Average	**66.9%**	**64.2%**	**82.3%**	**46.9%**	**51.6%**	**81.3%**	**72.6%**	**67.1%**	**68.0%**	**74.2%**	**66.5%**	**70.0%**	**71.6%**
Plants (174 sequences)													
Linear Discriminant	96.0%	95.4%	76.4%	92.5%	93.7%	91.4%	95.4%	96.0%	95.4%	96.0%	96.0%	96.0%	96.0%
Linear SVM	96.0%	96.0%	85.6%	96.0%	96.0%	87.9%	94.8%	96.0%	96.0%	96.0%	96.0%	96.0%	96.0%
Quadratic SVM	96.0%	96.0%	86.8%	96.0%	96.0%	88.5%	94.3%	96.0%	96.0%	96.0%	96.0%	96.0%	96.0%
Fine KNN	93.1%	94.8%	91.4%	94.3%	94.3%	90.8%	86.8%	93.1%	94.3%	93.7%	91.4%	93.1%	93.1%
Subspace Discriminant	96.0%	95.4%	87.4%	94.8%	95.4%	87.9%	94.8%	96.0%	96.0%	96.0%	96.0%	96.0%	96.0%
Subspace KNN	93.7%	94.3%	90.2%	94.3%	94.3%	90.2%	92.5%	92.5%	94.8%	93.7%	94.3%	94.8%	94.3%
Average	**95.1%**	**95.3%**	**86.3%**	**94.7%**	**95.0%**	**89.5%**	**93.1%**	**94.9%**	**95.4%**	**95.2%**	**95.0%**	**95.3%**	**95.2%**
Amphibians (290 sequences)													
Linear Discriminant	92.1%	91.4%	95.5%	89.0%	89.3%	99.0%	94.5%	93.4%	91.4%	96.2%	93.4%	93.8%	91.7%
Linear SVM	91.0%	90.0%	89.0%	88.3%	88.6%	93.1%	89.0%	91.4%	90.0%	93.1%	92.1%	92.4%	90.3%
Quadratic SVM	90.3%	89.0%	92.4%	59.3%	83.4%	96.6%	91.0%	93.1%	86.9%	94.1%	93.1%	93.4%	90.7%
Fine KNN	90.0%	86.9%	96.6%	83.8%	83.4%	98.3%	87.9%	92.1%	89.7%	93.4%	91.7%	94.8%	89.7%
Subspace Discriminant	90.7%	90.3%	90.0%	89.3%	89.3%	96.6%	90.3%	91.7%	90.3%	95.2%	92.8%	92.1%	91.0%
Subspace KNN	88.3%	86.6%	94.1%	85.2%	84.5%	98.3%	89.7%	92.8%	87.2%	94.5%	90.0%	94.8%	90.3%
Average	**90.4%**	**89.0%**	**92.9%**	**82.5%**	**86.4%**	**97.0%**	**90.4%**	**92.4%**	**89.3%**	**94.4%**	**92.2%**	**93.6%**	**90.6%**
Mammals (830 sequences)													
Linear Discriminant	98.3%	97.6%	97.7%	97.0%	96.0%	97.1%	96.6%	97.2%	96.7%	98.0%	96.9%	96.3%	96.3%
Linear SVM	90.6%	89.6%	88.9%	84.5%	85.3%	91.6%	86.5%	91.2%	88.8%	90.8%	90.0%	88.2%	88.1%
Quadratic SVM	92.4%	89.9%	91.0%	32.9%	41.7%	93.4%	88.0%	93.4%	89.9%	90.7%	92.5%	89.8%	90.5%
Fine KNN	94.1%	92.3%	96.0%	79.9%	81.0%	96.6%	93.9%	93.7%	91.7%	96.3%	96.3%	94.8%	95.5%
Subspace Discriminant	92.3%	91.9%	92.3%	88.3%	87.7%	94.0%	90.2%	91.7%	90.4%	92.3%	93.4%	91.9%	91.3%
Subspace KNN	92.8%	90.8%	95.5%	78.2%	79.2%	96.4%	91.2%	93.3%	89.2%	94.8%	94.3%	94.9%	92.2%
Average	**93.4%**	**92.0%**	**93.6%**	**76.8%**	**78.5%**	**94.9%**	**91.1%**	**93.4%**	**91.1%**	**93.8%**	**93.9%**	**92.7%**	**92.3%**
Insects (898 sequences)													
Linear Discriminant	92.2%	92.7%	90.1%	91.6%	92.2%	94.2%	93.3%	92.4%	89.2%	93.1%	92.1%	94.4%	90.4%
Linear SVM	86.9%	82.6%	85.9%	66.7%	69.5%	85.3%	86.4%	90.0%	80.5%	89.4%	87.4%	88.4%	86.2%
Quadratic SVM	85.0%	81.8%	86.7%	24.4%	21.3%	87.1%	85.7%	89.6%	82.6%	89.5%	88.0%	89.6%	85.3%
Fine KNN	82.0%	79.3%	80.0%	62.5%	68.0%	93.2%	83.3%	87.9%	80.8%	85.6%	83.6%	87.9%	83.0%
Subspace Discriminant	85.7%	83.9%	88.3%	77.5%	79.3%	89.1%	88.0%	88.2%	82.1%	87.1%	87.6%	88.2%	86.4%
Subspace KNN	80.4%	77.3%	90.5%	61.0%	67.6%	92.0%	81.4%	86.9%	77.4%	85.4%	86.0%	89.3%	81.4%
Average	**85.4%**	**82.9%**	**86.9%**	**64.0%**	**66.3%**	**90.2%**	**86.4%**	**89.2%**	**82.1%**	**88.4%**	**87.5%**	**89.6%**	**85.5%**
3Classes (2170 sequences; Subspace Discriminant & Subspace KNN omitted)													
Linear Discriminant	99.9%	99.9%	99.6%	99.4%	99.7%	99.7%	99.7%	99.7%	99.8%	99.8%	99.9%	99.9%	99.6%
Linear SVM	94.1%	90.2%	99.4%	89.8%	89.3%	99.6%	99.2%	98.1%	94.6%	99.1%	97.3%	99.3%	97.9%
Quadratic SVM	97.5%	92.5%	99.4%	66.6%	78.8%	99.7%	99.5%	98.7%	97.6%	99.4%	98.4%	99.5%	98.8%
Fine KNN	95.9%	95.2%	97.6%	93.3%	94.4%	95.9%	97.6%	97.7%	96.4%	98.9%	98.0%	99.2%	98.4%
Average	**96.9%**	**94.5%**	**99.0%**	**87.3%**	**90.6%**	**98.7%**	**99.0%**	**98.6%**	**97.1%**	**99.3%**	**98.4%**	**99.5%**	**98.7%**
Vertebrates (4322 sequences; Subspace Discriminant & Subspace KNN omitted)													
Linear Discriminant	99.7%	99.7%	99.6%	99.3%	99.5%	99.7%	99.2%	99.3%	99.3%	99.3%	99.4%	99.5%	99.2%
Linear SVM	98.3%	98.2%	98.5%	96.3%	96.8%	97.9%	98.0%	98.4%	98.2%	98.2%	98.5%	98.8%	98.4%
Quadratic SVM	98.1%	96.6%	99.0%	40.6%	34.0%	98.7%	98.4%	98.2%	96.7%	98.5%	98.7%	98.8%	98.6%
Fine KNN	97.1%	96.1%	98.4%	88.3%	91.7%	97.9%	96.4%	96.3%	95.3%	96.4%	97.5%	97.6%	97.2%
Average	**98.3%**	**97.7%**	**98.9%**	**81.1%**	**80.5%**	**98.6%**	**98.0%**	**98.1%**	**97.4%**	**98.1%**	**98.5%**	**98.7%**	**98.4%**
Table Average	**90.0%**	**88.7%**	**91.6%**	**79.4%**	**81.3%**	**92.3%**	**90.5%**	**90.7%**	**89.4%**	**91.9%**	**90.5%**	**91.5%**	**90.7%**

and then over all datasets), are high. Surprisingly, even using a single nucleotide numerical representation, which treats three of the nucleotides as being the same, and singles out only one of them ("Just-A"), results in an average accuracy of 91.9%. The best accuracy, for these datasets, is achieved when using the "PP" representation, which yields an average accuracy of 92.3%.

For subsequent experiments we selected the top three representations in terms of accuracy scores: "PP", "Just-A", and "Real" numerical representations.

3.3.3 ML-DSP for three classes of vertebrates

As an application of ML-DSP using the "PP" numerical representation for DNA sequences, we analyzed the set of vertebrate mtDNA genomes (median length 16,606 bp). The MoDMap, i.e., the multi-dimensional scaling 3D visualization of the genome interrelationships as described by the distances in the distance matrix, is illustrated in Fig 3.3. The dataset contains 3,740 complete mtDNA genomes: 553 bird genomes, 2,313 fish genomes, and 874 mammalian genomes. Quantitatively, the classification accuracy score obtained by the Quadratic SVM classifier was 100%.

3.3.4 Classifying genomes with ML-DSP, at all taxonomic levels

We tested the ability of ML-DSP to classify complete mtDNA sequences at various taxonomic levels. For every dataset, we tested using the "PP", "Just-A", and "Real" numerical representations.

The starting point was domain Eukaryota ($7,396$ sequences), which was classified into kingdoms, then kingdom Animalia was classified into phyla, etc. At each level, we picked the cluster with the highest number of sequences and then classified it into the next taxonomic level sub-clusters. The lowest level classified was family Cyprinidae (81 sequences) into its six genera. For each dataset, we tested all six classifiers, and the maximum of these six classification accuracy scores for each dataset are shown in Table 3.4.

Table 3.4: Maximum classification accuracy (of the accuracies obtained with each of the six classifiers) of ML-DSP, for datasets at different taxonomic levels, from 'domain into kindgoms' down to 'family into genera'.

Test	No. of Seq.	Max Length	Min Length	Median Length	Mean Length	Numerical Representation Maximum Accuracy				
						PP	Real	Just-A	Random3*	Random13**
Domain to Kingdom Domain:Eukaryota Kingdoms: Plants:,254, Animals: 6697, Fungi: 267, Protists :178	7396	1999595	1136	16580	25434	96.2%	97.3%	96.1%	95.5%	92.8%
Domain to Kingdom (No Protists) Domain:Eukaryota Kingdoms: Plants:254, Animals: 6697, Fungi: 267	7218	1999595	1136	16573	25254	97.9%	98.4%	97.9%	97.4%	94.4%
Kingdom to Phylum Kingdom: Animalia Phylum: Chordata:4367, Cnidaria: 127, Ecdysozoa: 1572, Porifera: 60, Echinodermata: 44, Lophotrochozoa: 403, Platyhelminthes: 100	6673	48161	5596	16553	16474	96.2%	95.9%	95.3%	93.6%	85.6%
Phylum to SubPhylum Phylum:Chordata SubPhylum:Cephalochordata:9, Craniata: 4334, Tunicata:24	4367	28757	13424	16615	16791	99.7%	99.8%	99.8%	99.5%	99.7%
SubPhylum to Class SubPhylum:Vertebrata Class: Amphibians(Amphibia):290, Birds(Aves): 553, Fish(Actinopterygii, Chondrichthyes, Dipnoi, Coelacanthiformes): 2313, Mammals(Mammalia): 874, Reptiles(Crocodylia, Sphenodontia, Squamata, Testudines): 292	4322	28757	14935	16616	16806	99.7%	99.6%	99.3%	99.2%	86.2%
Class to SubClass Class:Actinopterygii SubClass: Chondrostei: 24, Cladistia: 11, Neopterygii: 2141	2176	22217	15534	16589	16656	100%	99.9%	99.9%	99.8%	99.2%
SubClass to SuperOrder SubClass: Neopterygii SuperOrder: Osteoglossomorpha:23, Elopomorpha: 60, Clupeomorpha: 75, Ostariophysi: 792, Protacanthopterygii: 66, Paracanthopterygii: 46, Acanthopterygii:426	1488	22217	15534	16597	16669	96.2%	96.4%	95.4%	94.4%	78.8%
SuperOrder to Order SuperOrder:Ostariophysi Order: Cypriniformes: 643, Characiformes: 31, Siluriformes: 107	781	17859	16123	16597	16621	99.0%	98.7%	98.8%	97.6%	92.2%
Order to Family Order: Cypriniformes Family: Balitoridae: 25, Catostomidae:12, Cobitidae: 51, Cyprinidae: 502, Nemacheilidae: 47	635	17859	16411	16601	16627	98.9%	97.8%	98.3%	97.3%	85.7%
Family to Genus Family: Cyprinidae Genus: *Schizothorax*: 19, *Labeo*: 19, *Acrossocheilus*: 12, *Acheilognathus*: 10, *Rhodeus*: 11, *Onychostoma*: 10	81	17155	16563	16597	16630	91.8%	92.6%	91.4%	85.2%	66.7%
Table Average Accuracy	—	—	—	—	—	97.6%	97.6%	97.2%	96.0%	88.1%

At each level, the cluster with the highest number of sequences was chosen as the next dataset to be classified into its sub-taxa. *Random3: each sequence is represented by a random representation among PP, Real, or Just-A. **Random13: each sequence is represented by random representation among 13 representations (Integer, Integer(Other), Real, Atomic, EIIP, PP, Paired Numeric, Nearest neighbor based doublet, Codon, Just-A, Just-C, Just-G or Just-T).

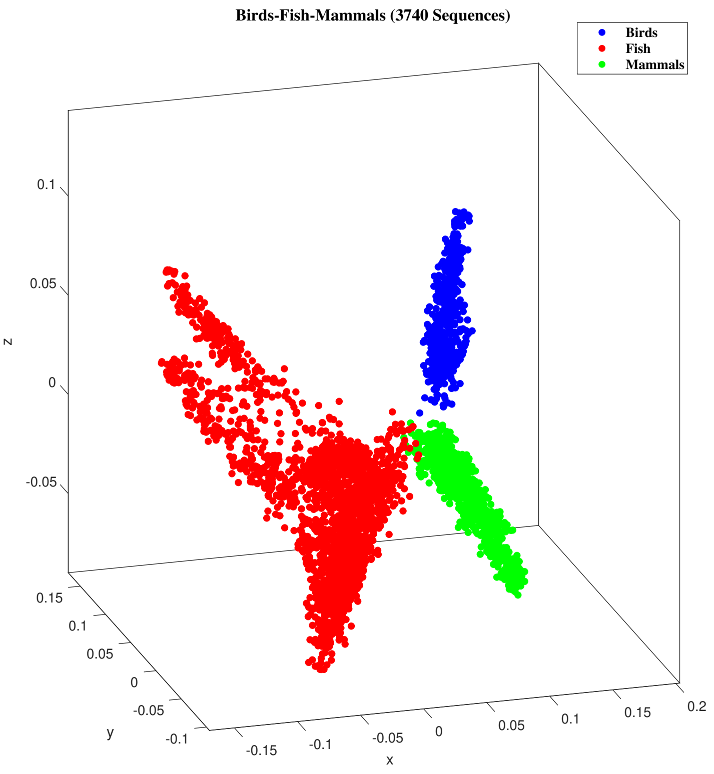


Figure 3.3: MoDMap of 3,740 full mtDNA genomes in subphylum Vertebrata, into three classes: Birds (blue, Aves: 553 genomes), fish (red, Actinopterygii 2,176 genomes, Chondrichthyes 130 genomes, Coelacanthiformes 2 genomes, Dipnoi 5 genomes), and mammals (green, Mammalia: 874 genomes). The accuracy of the ML-DSP classification into three classes, using the Quadratic SVM classifier, with the "PP" numerical representation, and PCC between magnitude spectra of DFT, was 100%.

Note that, at each taxonomic level, the maximum classification accuracy scores (among the six classifiers) for each of the three numerical representations considered are high, ranging from 91.4% to 100%, with only three scores under 95%. As this analysis also did not reveal a clear winner among the top three numerical representations, the question then arose whether the numerical representation we use mattered at all. To answer this question, we performed two additional experiments, that exploit the fact that the Pearson correlation coefficient is scale independent, and only looks for a pattern while comparing signals. For the first experiment we selected the top three numerical representations ("PP", "Just-A", and "Real") and, for each sequence in a given dataset, a numerical representation among these three was randomly chosen, with equal probability, to be the digital signal that represents it. The results are shown under the column “Random3” in Table 3.4: The maximum accuracy score over all the datasets is 96%. This is almost the same as the accuracy obtained when one particular numerical representation was used (1% lower, which is well within experimental error). We then repeated this experiment, this time picking randomly from any of the thirteen numerical representations considered. The results are shown under the column “Random13” in Table 3.4, with the table average accuracy score being 88.1%.

Overall, our results suggest that all three numerical representations “PP”, “Just-A”, and “Real” have very high classifications accuracy scores (average >97%), and even a random choice of one of these representations for each sequence in the dataset does not significantly affect the classification accuracy score of ML-DSP (average 96%).

We also note that, in addition to being highly accurate in its classifications, ML-DSP is ultrafast. Indeed, even for the largest dataset in Table 3.2, subphylum Vertebrata (4,322 complete mtDNA genomes, average length 16,806 bp), the distance matrix computation (which is the bulk of the classification computation) lasted under 5 seconds. Classifying a new primate mtDNA genome took 0.06 seconds when trained on 148 primate mtDNA genomes, and classifying a new vertebrate mtDNA genome took 7 seconds when trained on the 4,322 vertebrate mtDNA genomes. The result was updated with an experiment whereby QSVM was trained on the 4,322 complete vertebrate genomes in Table 3.2, and querried on the 694 new vertebrate mtDNA genomes uploaded on NCBI between June 17, 2017 and January 7, 2019. The accuracy of classification was 99.6%, with only three reptile mtDNA genomes mis-classified as amphibian genomes: *Bavayia robusta*, robust forest bavayia - a species of gecko, NC_034780, *Mesoclemmys hogei*, Hoge’s toadhead turtle, NC_036346, and *Gonatodes albogularis*, yellow-headed gecko, NC_035153.

3.3.5 MoDMap visualization vs. ML-DSP quantitative classification results

The hypothesis tested by the next experiments was that the quantitative accuracy of the classification of DNA sequences by ML-DSP would be significantly higher than suggested by the visual clustering of taxa in the MoDMap produced with the same pairwise distance matrix.

As an example, the MoDMap in Fig 3.4a, visualizes the distance matrix of mtDNA genomes from family Cyprinidae (81 genomes) with its genera *Acheilognathus* (10 genomes), *Rhodeus* (11 genomes), *Schizothorax* (19 genomes), *Labeo* (19 genomes), *Acrossocheilus* (12 genomes), *Onychostoma* (10 genomes); only the genera with at least 10 genomes are considered. The MoDMap seems to indicate an overlap between the clusters *Acheilognathus* and *Rhodeus*, which is biologically plausible as these genera belong to the same sub-family Acheilognathinae. However, when zooming in by plotting a MoDMap of only these two genera, as shown in Fig 3.4b, one can see that the clusters are clearly separated visually. This separation is confirmed by the fact that the accuracy score of the Quadratic SVM classifier for the dataset in Fig 3.4b is 100%. The same quantitative accuracy score for the classification of the dataset in Fig 3.4a with Quadratic SVM is 91.8%, which intuitively is much better than the corresponding MoDMap would suggest. This is likely due to the fact that the MoDMap is a three-dimensional approximation of the positions of the genome-representing points in a multi-dimensional space (the number of dimensions is $(n-1)$, where n is the number of sequences).

This being said, MoDMaps can still serve for exploratory purposes. For example, the MoDMap in Fig 3.4a suggests that species of the genus *Onychostoma* (subfamily listed “unknown” in NCBI) (yellow), may be genetically related to species of the genus *Acrossocheilus* (subfamily Barbinae) (magenta). Upon further exploration of the distance matrix, one finds that indeed the distance between the centroids of these two clusters is lower than the distance between each of these two cluster-centroids to the other cluster-centroids. This supports the hypotheses, based on morphological evidence [60], that genus *Onychostoma* belongs to the subfamily Barbinae, respectively that genus *Onychostoma* and genus *Acrossocheilus* are closely

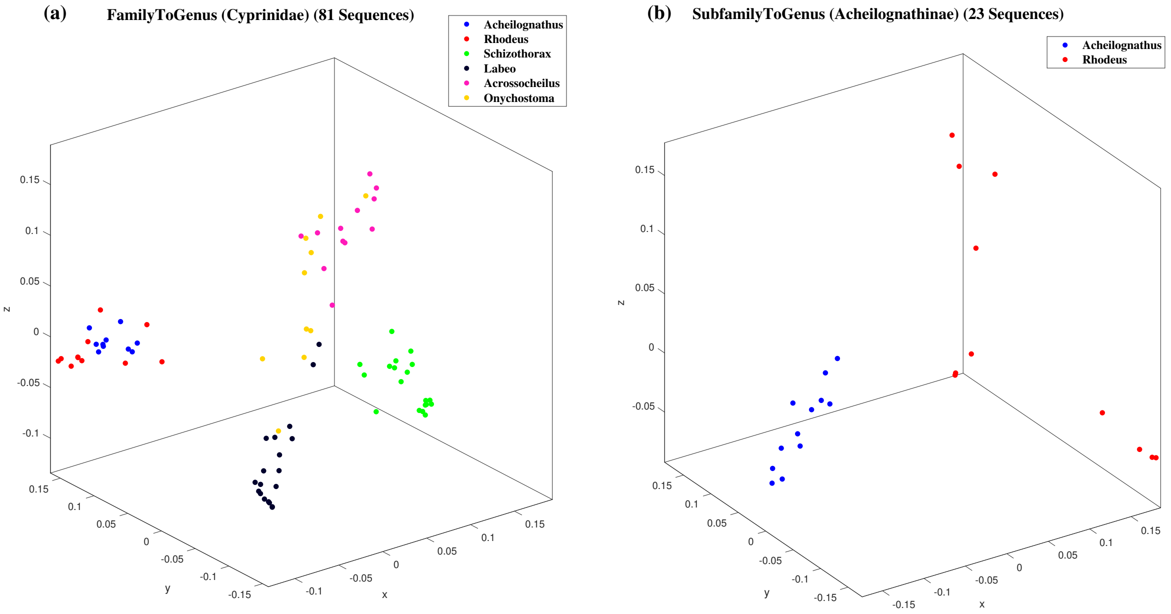


Figure 3.4: MoDMap of family Cyprinidae and its genera. (**a**): Genera *Acheilognathus* (blue, 10 genomes), *Rhodeus* (red, 11 genomes), *Schizothorax* (green, 19 genomes), *Labeo* (black, 19 genomes), *Acrossocheilus* (magenta, 12 genomes), *Onychostoma* (yellow, 10 genomes); (**b**): Genera *Acheilognathus* and *Rhodeus*, which overlapped in (**a**), are visually separated when plotted separately in (**b**). The classification accuracy with Quadratic SVM of the dataset in (**a**) was 91.8%, and of the dataset in (**b**) was 100%.

related [61]. Note that this exploration, suggested by MoDMap and confirmed by calculations based on the distance matrix, could not have been initiated based on ML-DSP alone (or other supervised machine learning algorithms), as ML-DSP only predicts the classification of new genomes into one of the taxa that it was trained on, and does not provide any other additional information.

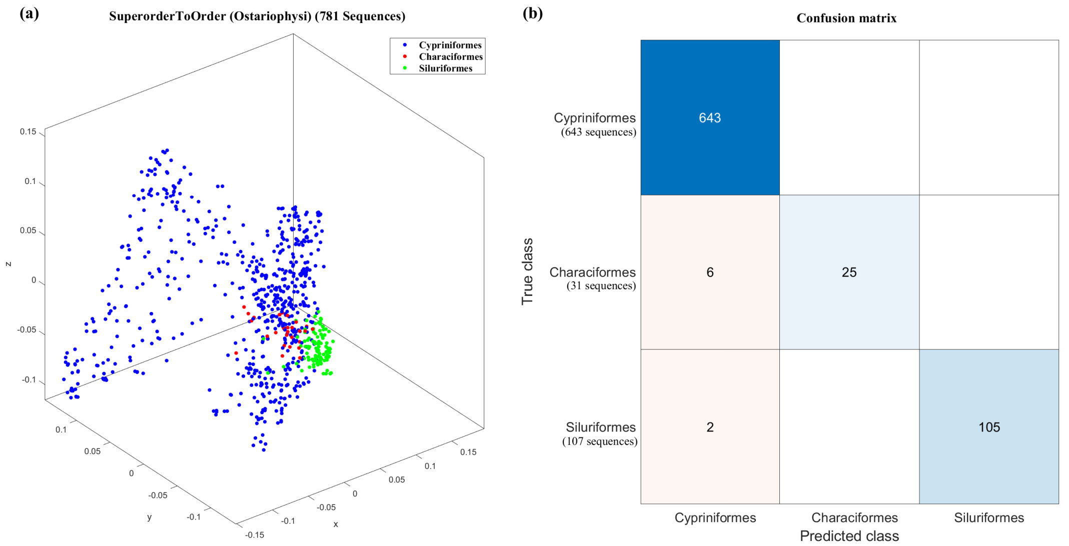
As another comparison point between MoDMaps and supervised machine learning outputs, Fig 3.5a shows the MoDMap of the superorder Ostariophysi with its orders Cypriniformes (643 genomes), Characiformes (31 genomes) and Siluriformes (107 genomes). The MoDMap shows the clusters as overlapping, but the Quadratic SVM classifier that quantitatively classifies these genomes has an accuracy of 99%. Indeed, the confusion matrix in Fig 3.5b shows that Quadratic SVM mis-classifies only 8 sequences out of 781 (recall that, for m clusters, the $m \times m$ confusion matrix has its rows labelled by the true classes and columns labelled by the predicted classes; the cell (i, j) shows the number of sequences that belong to the true class i, and have been predicted to be of class *j*). This indicates that when the visual representation in a MoDMap shows cluster overlaps, this may only be due to the dimensionality reduction to three dimensions, while ML-DSP actually provides a much better quantitative classification based on the same distance matrix.

3.3.6 Applications to other genomic datasets

The two experiments in this section indicate that the applicability of our method is not limited to mitochondrial DNA sequences. The first experiment, Fig 3.6a, shows the MoDMap of all 4,721 complete dengue virus sequences available in NCBI on August 10, 2017, classified into the subtypes DENV-1 (2,008 genomes), DENV-2 (1,349 genomes), DENV-3 (1,010 genomes), DENV-4 (354 genomes). The average length of these complete viral genomes is 10,595 bp. Despite the dengue viral genomes being very similar, the classification accuracy of this dataset into subtypes, using the Quadratic SVM classifier, was 100%. The second experiment, Fig 3.6b, shows the MoDMap of 4,710 bacterial genomes, classified into three phyla: Spirochaetes (437 genomes), Firmicutes (1,129 genomes), and Proteobacteria (3,144 genomes). The average length of these complete bacterial genomes is 104,150 bp, with the maximum length being 499,136 bp and the minimum length being 20,019 bp. The classification accuracy of the Quadratic SVM classifier for this dataset was 95.5%.

3.3.7 Comparison of ML-DSP with state-of-the-art alignment-based and alignment-free tools

The computational experiments in this section compare ML-DSP with three state-of-the-art alignment-based and alignment-free methods: the alignment-based tool MEGA7 [3] with alignment using MUSCLE [4] and CLUSTALW [5, 6], and the alignment-free method FFP (Feature Frequency Profiles) [28].



	Cypriniformes	Characiformes	Siluriformes
Cypriniformes (643 sequences)	643		
Characiformes (31 sequences)	6	25	
Siluriformes (107 sequences)	2		105

Figure 3.5: MoDMap of the superorder Ostariophysi, and the confusion matrix for the Quadratic SVM classification of this superorder into orders. (**a**): MoDMap of orders Cypriniformes (blue, 643 genomes), Characiformes (red, 31 genomes), Siluriformes (green, 107 genomes). (**b**): The confusion matrix generated by Quadratic SVM, illustrating its true class vs. predicted class performance (top-to-bottom and left-to-right: Cypriniformes, Characiformes, Siluriformes). The numbers in the squares on the top-left to bottom-right diagonal (blue) indicate the numbers of correctly classified DNA sequences, by order. The off-diagonal pink squares indicate that 6 mtDNA genomes of the order Characiformes have been erroneously predicted to belong to the order Cypriniformes (center-left), and 2 mtDNA genomes of the order Siluriformes have been erroneously predicted to belong to the order Cypriniformes (bottom-left). The Quadratic SVM that generated this confusion matrix had a 99% classification accuracy.

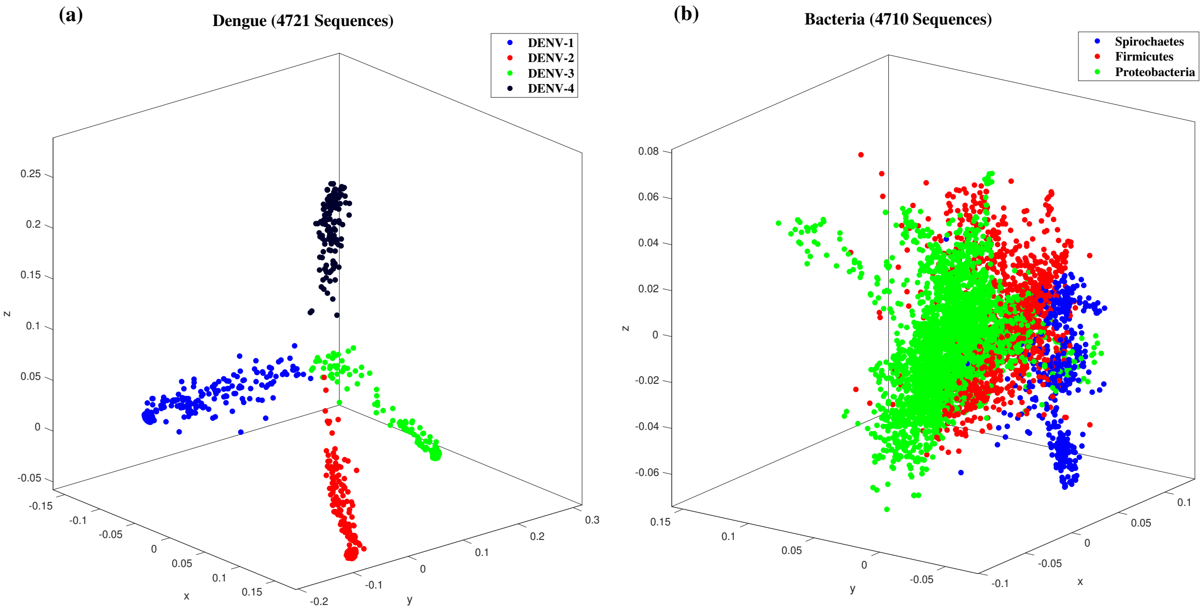


Figure 3.6: (**a**) MoDMap of 4,271 dengue virus genomes. The colours represent virus subtypes DENV-1 (blue, 2, 008 genomes), DENV-2 (red, 1, 349 genomes), DENV-3 (green, 1, 010 genomes), DENV-4 (black, 354 genomes); The classification accuracy of the Quadratic SVM classifier for this dataset was 100%. (**b**) MoDMap of 4,710 bacterial genomes. The colours represent bacterial phyla: Spirochaetes (blue, 437 genomes), Firmicutes (red, 1,129 genomes), Proteobacteria (green, 3,144 genomes). The accuracy of the Quadratic SVM classifier for this dataset was 95.5%.

For this performance analysis we selected three datasets. The first two datasets are benchmark datasets used in other genetic sequence comparison studies [47]: The first dataset comprises 38 influenza viral genomes, and the second dataset comprises 41 mammalian complete mtDNA sequences. The third datase, of our choice, is much larger, consisting of 4, 322 vertebrate complete mtDNA sequences, and was selected to compare scalability.

For the alignment-based methods, we used the distance matrix calculated in MEGA7 from sequences aligned with either MUSCLE or CLUSTALW. For the alignment-free FFP, we used the default value of $k = 5$ for k-mers (a k-mer is any DNA sequence of length k; any increase in the value of the parameter k, for the first dataset, resulted in a lower classification accuracy score for FFP). For ML-DSP we chose the Integer numerical representation and computed the average classification accuracy over all six classifiers for the first two datasets, and over all classifiers except Subspace Discriminant and Subspace KNN for the third dataset.

Table 3.5 shows the performance comparison (classification accuracy and processing time) of these four methods. The processing time included all computations, starting from reading the datasets to the completion of the distance matrix - the common element of all four methods. The listed processing times do not include the time needed for the computation of phylogenetic trees, MoDMap visualizations, or classification.

Table 3.5: Comparison of classification accuracy and processing time for the distance matrix computation with MEGA7(MUSCLE), MEGA7(CLUSTALW), FPP, and ML-DSP.

DataSet	**Parameter**	**MEGA7 (MUSCLE)**	**MEGA7 (CLUSTALW)**	**FFP**	**ML-DSP**
Influenza Virus	Maximum Classification Accuracy	97.4%	97.4%	68.4%	100%
(38 sequences)	Average Classification Accuracy	93.4%	95.6%	57.0%	94.7%
Average Length: 1407bp	**Processing Time**	**7.44 sec**	**2 min 14 sec**	**0.2 sec**	**0.2 sec**
Mammalia	Maximum Classification Accuracy	95.1%	95.1%	49.6%	92.7%
(41 sequences)	Average Classification Accuracy	89.7%	90.7%	41.5%	87.8%
Average Length: 16647bp	**Processing Time**	**11 min 15sec**	**5 hr 38 min**	**0.3 sec**	**0.3 sec**
Vertebrates	Maximum Classification Accuracy	——	——	61.7%	99.7%
(4322 sequences)	Average Classification Accuracy	——	——	48.3%	98.3%
Average Length: 16806bp	**Processing Time**	**>2 hours**	**>6 hours**	**94 sec**	**28 sec**

As seen in Table 3.5 (columns 3, 4, and 6) ML-DSP overwhelmingly outperforms the alignment-based software MEGA7(MUSCLE/CLUSTALW) in terms of processing time. In terms of accuracy, for the smaller virus and mammalian benchmark datasets, the average accuracies of ML-DSP and MEGA7(MUSCLE/CLUSTALW) were comparable, probably due to the small size of the training set for ML-DSP. The advantage of ML-DSP over the alignment-based tools became more apparent for the larger vertebrate dataset, where the accuracies of ML-DSP and the alignment-based tools could not even be compared, as the alignment-based tools were so slow that they had to be terminated. In contrast, ML-DSP classified the entire set of 4,322 vertebrate mtDNA genomes in 28 seconds, with average classification accuracy 98.3%. This indicates that ML-DSP is significantly more scalable than the alignment-based MEGA7(MUSCLE/CLUSTALW), as it can speedily and accurately classify datasets which alignment-based tools cannot even process.

As seen in Table 3.5 (columns 5 and 6), ML-DSP significantly outperforms the alignment-free software FFP in terms of accuracy (average classification accuracy 98.3% for ML-DSP vs.

48.3% for FFP, for the large vertebrate dataset), while at the same time being overall faster.

This comparison also indicates that, for these datasets, both alignment-free methods (ML-DSP and FFP) have an overwhelming advantage over the alignment-based methods (MEGA7 (MUSCLE/CLUSTALW)) in terms of processing time. Furthermore, when comparing the two alignment-free methods with each other, ML-DSP significantly outperforms FFP in terms of classification accuracy.

As another angle of comparison, Fig 3.7 displays the MoDMaps of the first benchmark dataset (38 influenza virus genomes) produced from the distance matrices generated by FFP, MEGA7 (MUSCLE), MEGA7 (CLUSTALW), and ML-DSP respectively. Fig 3.7a shows that with FFP it is difficult to observe any visual separation of the dataset into subtype clusters. Fig 3.7b, MEGA7 (MUSCLE), and Fig 3.7c MEGA7 (CLUSTALW) show overlaps of the clusters of points representing subtypes H1N1 and H2N2. In contrast, Fig 3.7d, which visualizes the distance matrix produced by ML-DSP, shows a clear separation among all subtypes.

Finally Figures 3.8 and 3.9 display the phylogenetic trees generated by each of the four methods considered. Fig 3.8a, the tree generated by FFP, has many misclassified genomes, which was expected given the MoDMap visualization of its distance matrix in Fig 3.7a. Fig 3.9a displays the phylogenetic tree generated by MEGA7, which was the same for both MUSCLE and CLUSTALW: It has only one incorrectly classified H5N1 genome, placed in middle of H1N1 genomes. Fig 3.8b and Fig 3.9b display the phylogenetic tree generated using the distance produced by ML-DSP (shown twice, in parallel with the other trees, for ease of comparison). ML-DSP classified all genomes correctly.

3.3.8 Discussion

The computational efficiency of ML-DSP is due to the fact that it is alignment-free (hence it does not need multiple sequence alignment), while the combination of 1D numerical representations, Discrete Fourier Transform and Pearson Correlation Coefficient makes it extremely computationally time efficient, and thus scalable.

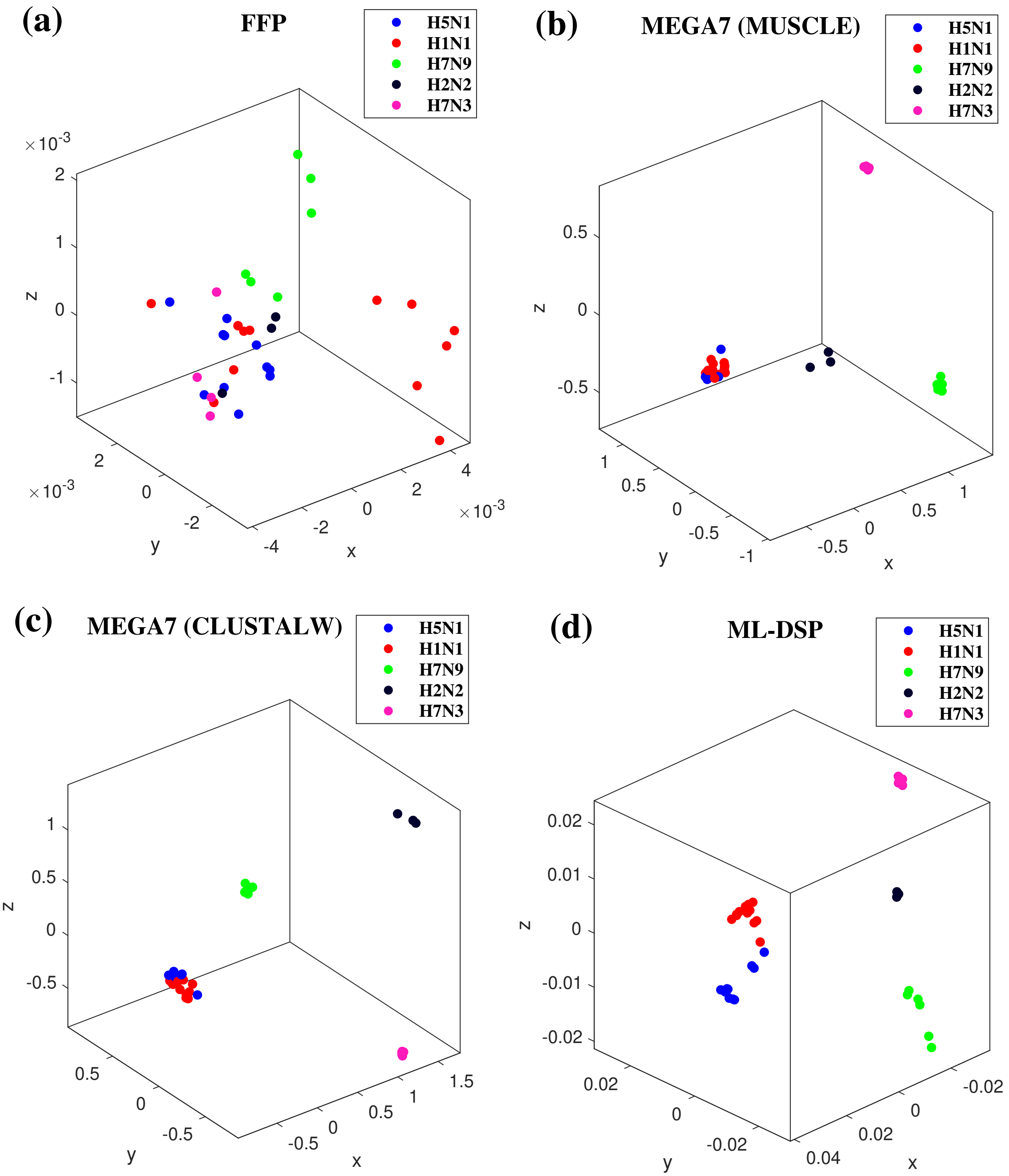


Figure 3.7: ModMaps of the influenza virus dataset from Table 3.5, based on the four methods. The points represent viral genomes of subtypes H1N1 (red, 13 genomes), H2N2 (black, 3 genomes), H5N1 (blue, 11 genomes), H7N3 (magenta, 5 genomes), H7N9 (green, 6 genomes); ModMaps are generated using distance matrices computed with (**a**) FFP; (**b**) MEGA7(MUSCLE); (**c**) MEGA7(CLUSTALW); (**d**) ML-DSP.

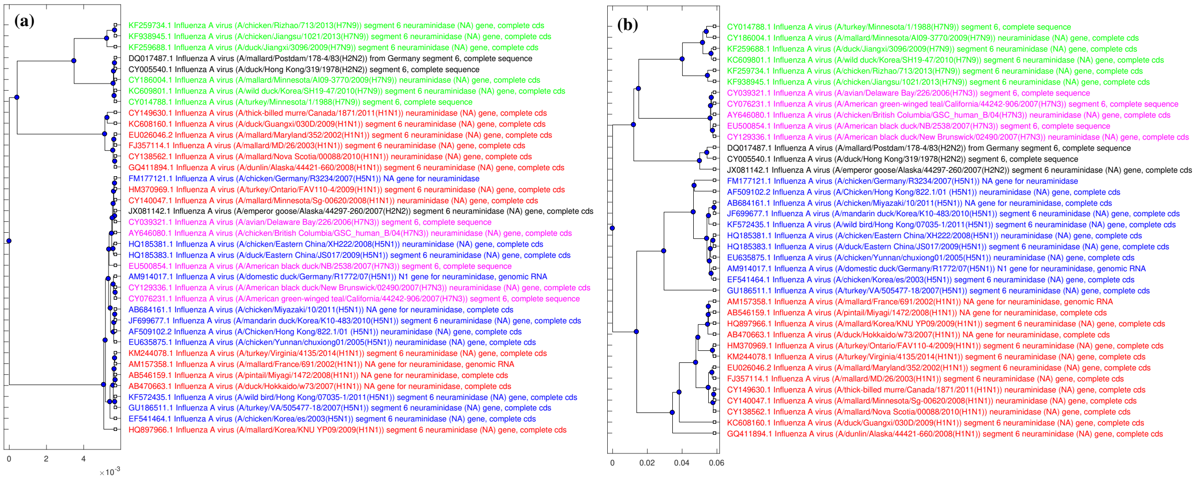

(a)
KF259734.1 Influenza A virus (A/chicken/Rizhao/713/2013(H7N9)) segment 6 neuraminidase (NA) gene, complete cds
KF938945.1 Influenza A virus (A/chicken/Jiangsu/1021/2013(H7N9)) segment 6 neuraminidase (NA) gene, complete cds
KF259688.1 Influenza A virus (A/duck/Jiangxi/3096/2009(H7N9)) segment 6 neuraminidase (NA) gene, complete cds
DQ017487.1 Influenza A virus (A/mallard/Postdam/178-4/83(H2N2)) from Germany segment 6, complete sequence
CY005540.1 Influenza A virus (A/duck/Hong Kong/319/1978(H2N2)) segment 6, complete sequence
CY186004.1 Influenza A virus (A/mallard/Minnesota/AI09-3770/2009(H7N9)) neuraminidase (NA) gene, complete cds
KC609801.1 Influenza A virus (A/wild duck/Korea/SH19-47/2010(H7N9)) segment 6 neuraminidase (NA) gene, complete cds
CY014788.1 Influenza A virus (A/turkey/Minnesota/1/1988(H7N9)) segment 6, complete sequence
CY149630.1 Influenza A virus (A/thick-billed murre/Canada/1871/2011(H1N1)) neuraminidase (NA) gene, complete cds
KC608160.1 Influenza A virus (A/duck/Guangxi/030D/2009(H1N1)) segment 6 neuraminidase (NA) gene, complete cds
EU026046.2 Influenza A virus (A/mallard/Maryland/352/2002(H1N1)) segment 6 neuraminidase (NA) gene, complete cds
FJ357114.1 Influenza A virus (A/mallard/MD/26/2003(H1N1)) segment 6 neuraminidase (NA) gene, complete cds
CY138562.1 Influenza A virus (A/mallard/Nova Scotia/00088/2010(H1N1)) neuraminidase (NA) gene, complete cds
GQ411894.1 Influenza A virus (A/dunlin/Alaska/44421-660/2008(H1N1)) segment 6 neuraminidase (NA) gene, complete cds
FM177121.1 Influenza A virus (A/chicken/Germany/R3234/2007(H5N1)) NA gene for neuraminidase
HM370969.1 Influenza A virus (A/turkey/Ontario/FAV110-4/2009(H1N1)) segment 6 neuraminidase (NA) gene, complete cds
CY140047.1 Influenza A virus (A/mallard/Minnesota/Sg-00620/2008(H1N1)) neuraminidase (NA) gene, complete cds
JX081142.1 Influenza A virus (A/emperor goose/Alaska/44297-260/2007(H2N2)) segment 6 neuraminidase (NA) gene, complete cds
CY039321.1 Influenza A virus (A/avian/Delaware Bay/226/2006(H7N3)) segment 6, complete sequence
AY646080.1 Influenza A virus (A/chicken/British Columbia/GSC_human_B/04(H7N3)) neuraminidase (NA) gene, complete cds
HQ185381.1 Influenza A virus (A/chicken/Eastern China/XH222/2008(H5N1)) neuraminidase (NA) gene, complete cds
HQ185383.1 Influenza A virus (A/duck/Eastern China/JS017/2009(H5N1)) segment 6 neuraminidase (NA) gene, complete cds
EU500854.1 Influenza A virus (A/American black duck/NB/2538/2007(H7N3)) segment 6, complete sequence
AM914017.1 Influenza A virus (A/domestic duck/Germany/R1772/07(H5N1)) N1 gene for neuraminidase, genomic RNA
CY129336.1 Influenza A virus (A/American black duck/New Brunswick/02490/2007(H7N3)) neuraminidase (NA) gene, complete cds
CY076231.1 Influenza A virus (A/American green-winged teal/California/44242-906/2007(H7N3)) segment 6, complete sequence
AB684161.1 Influenza A virus (A/chicken/Miyazaki/10/2011(H5N1)) NA gene for neuraminidase, complete cds
JF699677.1 Influenza A virus (A/mandarin duck/Korea/K10-483/2010(H5N1)) segment 6 neuraminidase (NA) gene, complete cds
AF509102.2 Influenza A virus (A/Chicken/Hong Kong/822.1/01 (H5N1)) neuraminidase (NA) gene, complete cds
EU635875.1 Influenza A virus (A/chicken/Yunnan/chuxiong01/2005(H5N1)) neuraminidase (NA) gene, complete cds
KM244078.1 Influenza A virus (A/turkey/Virginia/4135/2014(H1N1)) segment 6 neuraminidase (NA) gene, complete cds
AM157358.1 Influenza A virus (A/mallard/France/691/2002(H1N1)) NA gene for neuraminidase, genomic RNA
AB546159.1 Influenza A virus (A/pintail/Miyagi/1472/2008(H1N1)) NA gene for neuraminidase, complete cds
AB470663.1 Influenza A virus (A/duck/Hokkaido/w73/2007(H1N1)) NA gene for neuraminidase, complete cds
KF572435.1 Influenza A virus (A/wild bird/Hong Kong/07035-1/2011(H5N1)) segment 6 neuraminidase (NA) gene, complete cds
GU186511.1 Influenza A virus (A/turkey/VA/505477-18/2007(H5N1)) segment 6 neuraminidase (NA) gene, complete cds
EF541464.1 Influenza A virus (A/chicken/Korea/es/2003(H5N1)) segment 6 neuraminidase (NA) gene, complete cds
HQ897966.1 Influenza A virus (A/mallard/Korea/KNU YP09/2009(H1N1)) segment 6 neuraminidase (NA) gene, complete cds
0
2
4
$\times 10^{-3}$
(b)
CY014788.1 Influenza A virus (A/turkey/Minnesota/1/1988(H7N9)) segment 6, complete sequence
CY186004.1 Influenza A virus (A/mallard/Minnesota/AI09-3770/2009(H7N9)) neuraminidase (NA) gene, complete cds
KF259688.1 Influenza A virus (A/duck/Jiangxi/3096/2009(H7N9)) segment 6 neuraminidase (NA) gene, complete cds
KC609801.1 Influenza A virus (A/wild duck/Korea/SH19-47/2010(H7N9)) segment 6 neuraminidase (NA) gene, complete cds
KF259734.1 Influenza A virus (A/chicken/Rizhao/713/2013(H7N9)) segment 6 neuraminidase (NA) gene, complete cds
KF938945.1 Influenza A virus (A/chicken/Jiangsu/1021/2013(H7N9)) segment 6 neuraminidase (NA) gene, complete cds
CY039321.1 Influenza A virus (A/avian/Delaware Bay/226/2006(H7N3)) segment 6, complete sequence
CY076231.1 Influenza A virus (A/American green-winged teal/California/44242-906/2007(H7N3)) segment 6, complete sequence
AY646080.1 Influenza A virus (A/chicken/British Columbia/GSC_human_B/04(H7N3)) neuraminidase (NA) gene, complete cds
EU500854.1 Influenza A virus (A/American black duck/NB/2538/2007(H7N3)) segment 6, complete sequence
CY129336.1 Influenza A virus (A/American black duck/New Brunswick/02490/2007(H7N3)) neuraminidase (NA) gene, complete cds
DQ017487.1 Influenza A virus (A/mallard/Postdam/178-4/83(H2N2)) from Germany segment 6, complete sequence
CY005540.1 Influenza A virus (A/duck/Hong Kong/319/1978(H2N2)) segment 6, complete sequence
JX081142.1 Influenza A virus (A/emperor goose/Alaska/44297-260/2007(H2N2)) segment 6 neuraminidase (NA) gene, complete cds
FM177121.1 Influenza A virus (A/chicken/Germany/R3234/2007(H5N1)) NA gene for neuraminidase
AF509102.2 Influenza A virus (A/Chicken/Hong Kong/822.1/01 (H5N1)) neuraminidase (NA) gene, complete cds
AB684161.1 Influenza A virus (A/chicken/Miyazaki/10/2011(H5N1)) NA gene for neuraminidase, complete cds
JF699677.1 Influenza A virus (A/mandarin duck/Korea/K10-483/2010(H5N1)) segment 6 neuraminidase (NA) gene, complete cds
KF572435.1 Influenza A virus (A/wild bird/Hong Kong/07035-1/2011(H5N1)) segment 6 neuraminidase (NA) gene, complete cds
HQ185381.1 Influenza A virus (A/chicken/Eastern China/XH222/2008(H5N1)) neuraminidase (NA) gene, complete cds
HQ185383.1 Influenza A virus (A/duck/Eastern China/JS017/2009(H5N1)) segment 6 neuraminidase (NA) gene, complete cds
EU635875.1 Influenza A virus (A/chicken/Yunnan/chuxiong01/2005(H5N1)) neuraminidase (NA) gene, complete cds
AM914017.1 Influenza A virus (A/domestic duck/Germany/R1772/07(H5N1)) N1 gene for neuraminidase, genomic RNA
EF541464.1 Influenza A virus (A/chicken/Korea/es/2003(H5N1)) segment 6 neuraminidase (NA) gene, complete cds
GU186511.1 Influenza A virus (A/turkey/VA/505477-18/2007(H5N1)) segment 6 neuraminidase (NA) gene, complete cds
AM157358.1 Influenza A virus (A/mallard/France/691/2002(H1N1)) NA gene for neuraminidase, genomic RNA
AB546159.1 Influenza A virus (A/pintail/Miyagi/1472/2008(H1N1)) NA gene for neuraminidase, complete cds
HQ897966.1 Influenza A virus (A/mallard/Korea/KNU YP09/2009(H1N1)) segment 6 neuraminidase (NA) gene, complete cds
AB470663.1 Influenza A virus (A/duck/Hokkaido/w73/2007(H1N1)) NA gene for neuraminidase, complete cds
HM370969.1 Influenza A virus (A/turkey/Ontario/FAV110-4/2009(H1N1)) segment 6 neuraminidase (NA) gene, complete cds
KM244078.1 Influenza A virus (A/turkey/Virginia/4135/2014(H1N1)) segment 6 neuraminidase (NA) gene, complete cds
EU026046.2 Influenza A virus (A/mallard/Maryland/352/2002(H1N1)) segment 6 neuraminidase (NA) gene, complete cds
FJ357114.1 Influenza A virus (A/mallard/MD/26/2003(H1N1)) segment 6 neuraminidase (NA) gene, complete cds
CY149630.1 Influenza A virus (A/thick-billed murre/Canada/1871/2011(H1N1)) neuraminidase (NA) gene, complete cds
CY140047.1 Influenza A virus (A/mallard/Minnesota/Sg-00620/2008(H1N1)) neuraminidase (NA) gene, complete cds
CY138562.1 Influenza A virus (A/mallard/Nova Scotia/00088/2010(H1N1)) neuraminidase (NA) gene, complete cds
KC608160.1 Influenza A virus (A/duck/Guangxi/030D/2009(H1N1)) segment 6 neuraminidase (NA) gene, complete cds
GQ411894.1 Influenza A virus (A/dunlin/Alaska/44421-660/2008(H1N1)) segment 6 neuraminidase (NA) gene, complete cds
0
0.02
0.04
0.06

Figure 3.8: Phylogenetic tree comparison: FFP with ML-DSP. The phylogenetic tree generated for 38 influenza virus genomes using (**a**): FFP (**b**): ML-DSP.

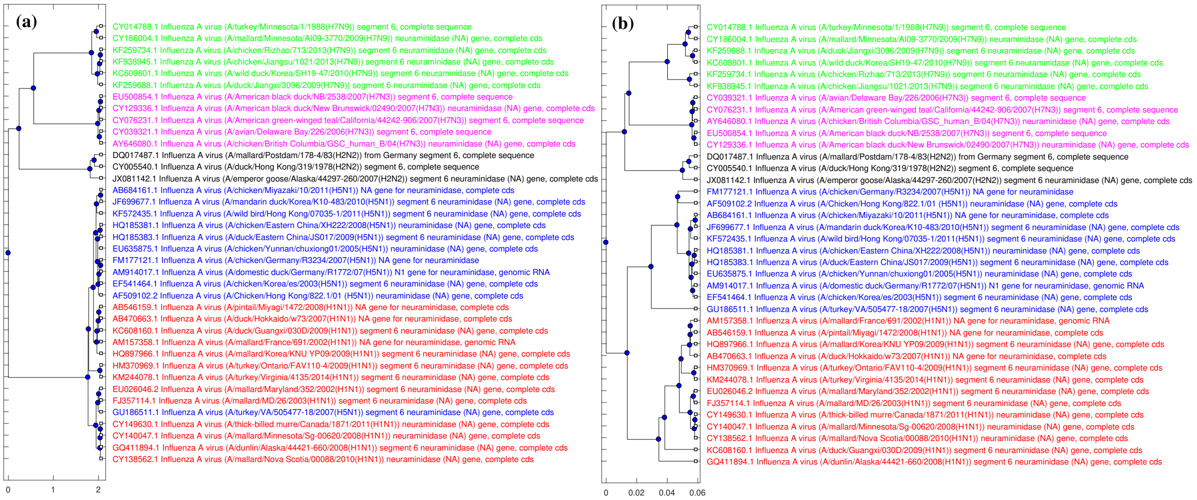

(a)
CY014788.1 Influenza A virus (A/turkey/Minnesota/1/1988(H7N9)) segment 6, complete sequence
CY186004.1 Influenza A virus (A/mallard/Minnesota/AI09-3770/2009(H7N9)) neuraminidase (NA) gene, complete cds
KF259734.1 Influenza A virus (A/chicken/Rizhao/713/2013(H7N9)) segment 6 neuraminidase (NA) gene, complete cds
KF938945.1 Influenza A virus (A/chicken/Jiangsu/1021/2013(H7N9)) segment 6 neuraminidase (NA) gene, complete cds
KC609801.1 Influenza A virus (A/wild duck/Korea/SH19-47/2010(H7N9)) segment 6 neuraminidase (NA) gene, complete cds
KF259688.1 Influenza A virus (A/duck/Jiangxi/3096/2009(H7N9)) segment 6 neuraminidase (NA) gene, complete cds
EU500854.1 Influenza A virus (A/American black duck/NB/2538/2007(H7N3)) segment 6, complete sequence
CY129336.1 Influenza A virus (A/American black duck/New Brunswick/02490/2007(H7N3)) neuraminidase (NA) gene, complete cds
CY076231.1 Influenza A virus (A/American green-winged teal/California/44242-906/2007(H7N3)) segment 6, complete sequence
CY039321.1 Influenza A virus (A/avian/Delaware Bay/226/2006(H7N3)) segment 6, complete sequence
AY646080.1 Influenza A virus (A/chicken/British Columbia/GSC_human_B/04(H7N3)) neuraminidase (NA) gene, complete cds
DQ017487.1 Influenza A virus (A/mallard/Postdam/178-4/83(H2N2)) from Germany segment 6, complete sequence
CY005540.1 Influenza A virus (A/duck/Hong Kong/319/1978(H2N2)) segment 6, complete sequence
JX081142.1 Influenza A virus (A/emperor goose/Alaska/44297-260/2007(H2N2)) segment 6 neuraminidase (NA) gene, complete cds
AB684161.1 Influenza A virus (A/chicken/Miyazaki/10/2011(H5N1)) NA gene for neuraminidase, complete cds
JF699677.1 Influenza A virus (A/mandarin duck/Korea/K10-483/2010(H5N1)) segment 6 neuraminidase (NA) gene, complete cds
KF572435.1 Influenza A virus (A/wild bird/Hong Kong/07035-1/2011(H5N1)) segment 6 neuraminidase (NA) gene, complete cds
HQ185381.1 Influenza A virus (A/chicken/Eastern China/XH222/2008(H5N1)) neuraminidase (NA) gene, complete cds
HQ185383.1 Influenza A virus (A/duck/Eastern China/JS017/2009(H5N1)) segment 6 neuraminidase (NA) gene, complete cds
EU635875.1 Influenza A virus (A/chicken/Yunnan/chuxiong01/2005(H5N1)) neuraminidase (NA) gene, complete cds
FM177121.1 Influenza A virus (A/chicken/Germany/R3234/2007(H5N1)) NA gene for neuraminidase
AM914017.1 Influenza A virus (A/domestic duck/Germany/R1772/07(H5N1)) N1 gene for neuraminidase, genomic RNA
EF541464.1 Influenza A virus (A/chicken/Korea/es/2003(H5N1)) segment 6 neuraminidase (NA) gene, complete cds
AF509102.2 Influenza A virus (A/Chicken/Hong Kong/822.1/01 (H5N1)) neuraminidase (NA) gene, complete cds
AB546159.1 Influenza A virus (A/pintail/Miyagi/1472/2008(H1N1)) NA gene for neuraminidase, complete cds
AB470663.1 Influenza A virus (A/duck/Hokkaido/w73/2007(H1N1)) NA gene for neuraminidase, complete cds
KC608160.1 Influenza A virus (A/duck/Guangxi/030D/2009(H1N1)) segment 6 neuraminidase (NA) gene, complete cds
AM157358.1 Influenza A virus (A/mallard/France/691/2002(H1N1)) NA gene for neuraminidase, genomic RNA
HQ897966.1 Influenza A virus (A/mallard/Korea/KNU YP09/2009(H1N1)) segment 6 neuraminidase (NA) gene, complete cds
HM370969.1 Influenza A virus (A/turkey/Ontario/FAV110-4/2009(H1N1)) segment 6 neuraminidase (NA) gene, complete cds
KM244078.1 Influenza A virus (A/turkey/Virginia/4135/2014(H1N1)) segment 6 neuraminidase (NA) gene, complete cds
EU026046.2 Influenza A virus (A/mallard/Maryland/352/2002(H1N1)) segment 6 neuraminidase (NA) gene, complete cds
FJ357114.1 Influenza A virus (A/mallard/MD/26/2003(H1N1)) segment 6 neuraminidase (NA) gene, complete cds
GU186511.1 Influenza A virus (A/turkey/VA/505477-18/2007(H5N1)) segment 6 neuraminidase (NA) gene, complete cds
CY149630.1 Influenza A virus (A/thick-billed murre/Canada/1871/2011(H1N1)) neuraminidase (NA) gene, complete cds
CY140047.1 Influenza A virus (A/mallard/Minnesota/Sg-00620/2008(H1N1)) neuraminidase (NA) gene, complete cds
GQ411894.1 Influenza A virus (A/dunlin/Alaska/44421-660/2008(H1N1)) segment 6 neuraminidase (NA) gene, complete cds
CY138562.1 Influenza A virus (A/mallard/Nova Scotia/00088/2010(H1N1)) neuraminidase (NA) gene, complete cds
0
1
2
(b)
CY014788.1 Influenza A virus (A/turkey/Minnesota/1/1988(H7N9)) segment 6, complete sequence
CY186004.1 Influenza A virus (A/mallard/Minnesota/AI09-3770/2009(H7N9)) neuraminidase (NA) gene, complete cds
KF259688.1 Influenza A virus (A/duck/Jiangxi/3096/2009(H7N9)) segment 6 neuraminidase (NA) gene, complete cds
KC609801.1 Influenza A virus (A/wild duck/Korea/SH19-47/2010(H7N9)) segment 6 neuraminidase (NA) gene, complete cds
KF259734.1 Influenza A virus (A/chicken/Rizhao/713/2013(H7N9)) segment 6 neuraminidase (NA) gene, complete cds
KF938945.1 Influenza A virus (A/chicken/Jiangsu/1021/2013(H7N9)) segment 6 neuraminidase (NA) gene, complete cds
CY039321.1 Influenza A virus (A/avian/Delaware Bay/226/2006(H7N3)) segment 6, complete sequence
CY076231.1 Influenza A virus (A/American green-winged teal/California/44242-906/2007(H7N3)) segment 6, complete sequence
AY646080.1 Influenza A virus (A/chicken/British Columbia/GSC_human_B/04(H7N3)) neuraminidase (NA) gene, complete cds
EU500854.1 Influenza A virus (A/American black duck/NB/2538/2007(H7N3)) segment 6, complete sequence
CY129336.1 Influenza A virus (A/American black duck/New Brunswick/02490/2007(H7N3)) neuraminidase (NA) gene, complete cds
DQ017487.1 Influenza A virus (A/mallard/Postdam/178-4/83(H2N2)) from Germany segment 6, complete sequence
CY005540.1 Influenza A virus (A/duck/Hong Kong/319/1978(H2N2)) segment 6, complete sequence
JX081142.1 Influenza A virus (A/emperor goose/Alaska/44297-260/2007(H2N2)) segment 6 neuraminidase (NA) gene, complete cds
FM177121.1 Influenza A virus (A/chicken/Germany/R3234/2007(H5N1)) NA gene for neuraminidase
AF509102.2 Influenza A virus (A/Chicken/Hong Kong/822.1/01 (H5N1)) neuraminidase (NA) gene, complete cds
AB684161.1 Influenza A virus (A/chicken/Miyazaki/10/2011(H5N1)) NA gene for neuraminidase, complete cds
JF699677.1 Influenza A virus (A/mandarin duck/Korea/K10-483/2010(H5N1)) segment 6 neuraminidase (NA) gene, complete cds
KF572435.1 Influenza A virus (A/wild bird/Hong Kong/07035-1/2011(H5N1)) segment 6 neuraminidase (NA) gene, complete cds
HQ185381.1 Influenza A virus (A/chicken/Eastern China/XH222/2008(H5N1)) neuraminidase (NA) gene, complete cds
HQ185383.1 Influenza A virus (A/duck/Eastern China/JS017/2009(H5N1)) segment 6 neuraminidase (NA) gene, complete cds
EU635875.1 Influenza A virus (A/chicken/Yunnan/chuxiong01/2005(H5N1)) neuraminidase (NA) gene, complete cds
AM914017.1 Influenza A virus (A/domestic duck/Germany/R1772/07(H5N1)) N1 gene for neuraminidase, genomic RNA
EF541464.1 Influenza A virus (A/chicken/Korea/es/2003(H5N1)) segment 6 neuraminidase (NA) gene, complete cds
GU186511.1 Influenza A virus (A/turkey/VA/505477-18/2007(H5N1)) segment 6 neuraminidase (NA) gene, complete cds
AM157358.1 Influenza A virus (A/mallard/France/691/2002(H1N1)) NA gene for neuraminidase, genomic RNA
AB546159.1 Influenza A virus (A/pintail/Miyagi/1472/2008(H1N1)) NA gene for neuraminidase, complete cds
HQ897966.1 Influenza A virus (A/mallard/Korea/KNU YP09/2009(H1N1)) segment 6 neuraminidase (NA) gene, complete cds
AB470663.1 Influenza A virus (A/duck/Hokkaido/w73/2007(H1N1)) NA gene for neuraminidase, complete cds
HM370969.1 Influenza A virus (A/turkey/Ontario/FAV110-4/2009(H1N1)) segment 6 neuraminidase (NA) gene, complete cds
KM244078.1 Influenza A virus (A/turkey/Virginia/4135/2014(H1N1)) segment 6 neuraminidase (NA) gene, complete cds
EU026046.2 Influenza A virus (A/mallard/Maryland/352/2002(H1N1)) segment 6 neuraminidase (NA) gene, complete cds
FJ357114.1 Influenza A virus (A/mallard/MD/26/2003(H1N1)) segment 6 neuraminidase (NA) gene, complete cds
CY149630.1 Influenza A virus (A/thick-billed murre/Canada/1871/2011(H1N1)) neuraminidase (NA) gene, complete cds
CY140047.1 Influenza A virus (A/mallard/Minnesota/Sg-00620/2008(H1N1)) neuraminidase (NA) gene, complete cds
CY138562.1 Influenza A virus (A/mallard/Nova Scotia/00088/2010(H1N1)) neuraminidase (NA) gene, complete cds
KC608160.1 Influenza A virus (A/duck/Guangxi/030D/2009(H1N1)) segment 6 neuraminidase (NA) gene, complete cds
GQ411894.1 Influenza A virus (A/dunlin/Alaska/44421-660/2008(H1N1)) segment 6 neuraminidase (NA) gene, complete cds
0
0.02
0.04
0.06

Figure 3.9: Phylogenetic tree comparison: MEGA7(MUSCLE/CLUSTALW) with ML-DSP. The phylogenetic tree generated for 38 influenza virus genomes using (**a**): MEGA7(MUSCLE/CLUSTALW) (**b**): ML-DSP.

ML-DSP is not without limitations. We anticipate that the need for equal length sequences and use of length normalization could introduce issues with examination of small fragments of larger genome sequences. Usually genomes vary in length and thus length normalization always results in adding (up-sampling) or losing (down-sampling) some information. Although the Pearson Correlation Coefficient can distinguish the signal patterns even in small sequence fragments, and we did not find any considerable disadvantage while considering complete mitochondrial DNA genomes with their inevitable length variations, length normalization may cause issues when we deal with the fragments of genomes, and the much larger nuclear genome sequences.

Lastly, ML-DSP has two drawbacks, inherent in any supervised machine learning algorithm. The first is that ML-DSP is a black-box method which, while producing a highly accurate classification prediction, does not offer a (biological) explanation for its output. The second is that it relies on the existence of a training set from which it draws its "knowledge", that is, a set consisting of known genomic sequences and their taxonomic labels. ML-DSP uses such a training set to "learn" how to classify new sequences into one of the taxonomic classes that it was trained on, but it is not able to assign it to a taxon that it has not been exposed to.

3.4 Conclusions

We proposed ML-DSP, an ultrafast and accurate alignment-free supervised machine learning classification method based on digital signal processing of DNA sequences (and its software implementation). ML-DSP successfully addresses the limitations of alignment-free methods identified in [7], as follows:

(i) **Lack of software implementation**: ML-DSP is supplemented with freely available source-code. The ML-DSP software can be used with the provided datasets or any other custom dataset and provides the user with any (or all) of: pairwise distances, 3D sequence interrelationship visualization, phylogenetic trees, or classification accuracy scores. A quantitative comparison showed that ML-DSP significantly outperforms state-of-the-art alignment-based MEGA7 (MUSCLE/CLUSTALW) and alignment-free (FFP) software in terms of speed and classification accuracy.

(ii) **Use of simulated sequences or very small real-world datasets**: ML-DSP was successfully tested on a variety of large real-world datasets, comprizing thousands of complete genomes, such as all complete mitochondrial DNA sequences available on NCBI at the time of this study, and similarly large sets of viral genomes and bacterial genomes. ML-DSP was tested in different evolutionary scenarios such as different levels of taxonomy (from domain to genus), small dataset (38 sequences), large dataset (4,322 sequences), short sequences (1,136 bp), long sequences (1,999,595 bp), benchmark datasets of influenza virus and mammalian mtDNA genomes etc.

(iii) Memory overhead: ML-DSP uses neither *k*-mers nor any compression algorithms. Thus, scalability does not cause an exponential memory overhead, and a high classification accuracy is preserved with large datasets.

In addition, we provided a comprehensive quantitative analysis of all 13 one-dimensional numerical representations of DNA sequences used in the Genomic Signal Processing literature and found that, on average, the "PP", "Just-A", and "Real" representations performed better than others. We also showed that the classification accuracy of ML-DSP was significantly higher than the corresponding MoDMap visualizations of the dataset would indicate, likely due to the inherent dimensionality limitations of the latter. Lastly, we showed the potential for ML-DSP to be used for classifications of other DNA sequence genomic datasets, such as large datasets of complete viral or bacterial genomes.

3.5 Availability and Requirements

Project name: ML-DSP

Project home page: https://github.com/grandhawa/MLDSP

Operating system(s): Microsoft Windows

Programming language: MATLAB R2017A, license no. 964054

License: Creative Commons Attribution License

Any restrictions to use by non-academics: MATLAB license required

Bibliography

[1] Mora C, Tittensor DP, Adl S, Simpson AGB, Worm B. How many species are there on earth and in the ocean? PLoS Biology. 2011 Aug;9(8):e1001127.

[2] May RM. Why worry about how many species and their loss? PLoS Biology. 2011 Aug;9(8):e1001130.

[3] Kumar S, Stecher G, Tamura K. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Molecular Biology and Evolution. 2016;33(7):1870–1874.

[4] Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research. 2004;32(5):1792–1797.

[5] Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research. 1994 Nov;22(22):4673–4680.

[6] Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, et al. Clustal W and Clustal X version 2.0. Bioinformatics. 2007;23(21):2947–2948.

[7] Zielezinski A, Vinga S, Almeida J, Karlowski WM. Alignment-free sequence comparison: benefits, applications, and tools. Genome Biology. 2017 Oct;18(1):186.

[8] Vinga S, Almeida J. Alignment-free sequence comparison—a review. Bioinformatics. 2003;19(4):513–523.

[9] Schwende I, Pham TD. Pattern recognition and probabilistic measures in alignment-free sequence analysis. Briefings in Bioinformatics. 2014;15(3):354–368.

[10] Song K, Ren J, Reinert G, Deng M, Waterman MS, Sun F. New developments of alignment-free sequence comparison: measures, statistics and next-generation sequencing. Briefings in Bioinformatics. 2014;15(3):343–353.

[11] Kari L, Hill KA, Sayem AS, Karamichalis R, Bryans N, Davis K, et al. Mapping the space of genomic signatures. PLoS One. 2015 May;10(5):e0119815.

[12] Hoang T, Yin C, Yau SS. Numerical encoding of DNA sequences by Chaos Game Representation with application in similarity comparison. Genomics. 2016;108(3):134–142.

[13] Almeida J, Carriço JA, Maretzek A, Noble PA, M F. Analysis of genomic sequences by Chaos Game Representation. Bioinformatics. 2001;17 5:429–37.

[14] Yao YH, Dai Q, Nan XY, He PA, Nie ZM, Zhou SP, et al. Analysis of similarity/dissimilarity of DNA sequences based on a class of 2D graphical representation. Journal of Computational Chemistry. 2008 Jul;29(10):1632–1639.

[15] Qi X, Wu Q, Zhang Y, Fuller E, Zhang CQ. A novel model for DNA sequence similarity analysis based on graph theory. Evolutionary Bioinformatics Online. 2011 Oct;7:149–158.

[16] Almeida JS. Sequence analysis by iterated maps, a review. Briefings in Bioinformatics. 2014 May;15(3):369–375.

[17] Vinga S. Information theory applications for biological sequence analysis. Briefings in Bioinformatics. 2014;15(3):376–389.

[18] Bao J, Yuan R, Bao Z. An improved alignment-free model for DNA sequence similarity metric. BMC Bioinformatics. 2014 Sep;15(1):321.

[19] Leimeister CA, Boden M, Horwege S, Lindner S, Morgenstern B. Fast alignment-free sequence comparison using spaced-word frequencies. Bioinformatics. 2014;30(14):1991–1999.

[20] Chang G, Wang H, Zhang T. A novel alignment-free method for whole genome analysis: Application to HIV-1 subtyping and HEV genotyping. Information Sciences. 2014;279:776–784.

[21] Reese E, Krishnan VV. Classification of DNA sequences based on thermal melting profiles. Bioinformation. 2010 Apr;4(10):463–467.

[22] Bonham-Carter O, Steele J, Bastola D. Alignment-free genetic sequence comparisons: a review of recent approaches by word analysis. Briefings in Bioinformatics. 2014;15(6):890–905.

[23] Struck D, Lawyer G, Ternes AM, Schmit JC, Bercoff DP. COMET: adaptive context-based modeling for ultrafast HIV-1 subtype identification. Nucleic Acids Research. 2014 Oct;42(18):e144–e144.

[24] Remita MA, Halioui A, Malick Diouara AA, Daigle B, Kiani G, Diallo AB. A machine learning approach for viral genome classification. BMC Bioinformatics. 2017 Apr;18:208.

[25] Kosakovsky Pond SL, Posada D, Stawiski E, Chappey C, Poon AF, Hughes G, et al. An evolutionary model-based algorithm for accurate phylogenetic breakpoint mapping and subtype prediction in HIV-1. PLoS Computational Biology. 2009 Nov;5(11):e1000581.

[26] de Oliveira T, Deforche K, Cassol S, Salminen M, Paraskevis D, Seebregts C, et al. An automated genotyping system for analysis of HIV-1 and other microbial sequences. Bioinformatics. 2005;21(19):3797–3800.

[27] 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):e0206409.

[28] Sims GE, Jun SR, Wu GA, Kim SH. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. In: Proceedings of the National Academy of Sciences of the USA. vol. 106; 2009. p. 2677–2682.

[29] Kwan HK, Arniker SB. Numerical representation of DNA sequences. In: 2009 IEEE International Conference on Electro/Information Technology; 2009. p. 307–310.

[30] Borrayo E, Mendizabal-Ruiz EG, Vélez-Pérez H, Romo-Vázquez R, Mendizabal AP, Morales JA. Genomic signal processing methods for computation of alignment-free distances from DNA sequences. PLoS One. 2014 Nov;9(11):e110954.

[31] Adetiba E, Olugbara OO, Taiwo TB. Identification of pathogenic viruses using genomic cepstral coefficients with radial basis function neural network. In: Advances in Nature and Biologically Inspired Computing,Proceedings of the 7th World Congress on Nature and Biologically Inspired Computing. vol. 419; 2016. p. 281–290.

[32] Adetiba E, Olugbara OO. Classification of eukaryotic organisms through cepstral analysis of mitochondrial DNA. In: International Conference on Image and Signal Processing. vol. 9680; 2016. p. 243–252.

[33] Mendizabal-Ruiz G, Román-Godínez I, Torres-Ramos S, Salido-Ruiz RA, Morales JA. On DNA numerical representations for genomic similarity computation. PLoS One. 2017 Mar;12(3):e0173288.

[34] Chakravarthy N, Spanias A, Iasemidis LD, Tsakalis K. Autoregressive modeling and feature analysis of DNA sequences. EURASIP Journal on Applied Signal Processing. 2004 Jan;2004:13–28.

[35] Yu Z, Anh VV, Zhou Y, Zhou LQ. Numerical sequence representation of DNA sequences and methods to distinguish coding and non-coding sequences in a complete genome. In: Proceedings 11th World Multi-Conference on Systemics, Cybernetics and Informatics; 2007. p. 171–176.

[36] Abo-Zahhad M, Ahmed S, Abd-Elrahman S. Genomic analysis and classification of exon and intron sequences using DNA numerical mapping techniques. International Journal of Information Technology and Computer Science. 2012 Jul;4(8):22–36.

[37] Skutkova H, Vitek M, Sedlar K, Provaznik I. Progressive alignment of genomic signals by multiple dynamic time warping. Journal of Theoretical Biology. 2015;385:20–30.

[38] Yin C, Yau SST. An improved model for whole genome phylogenetic analysis by Fourier transform. Journal of Theoretical Biology. 2015;382:99–110.

[39] Lorenzo-Ginori JV, Rodriguez-Fuentes A, Grau Abalo R, Sanchez Rodriguez R. Digital signal processing in the analysis of genomic sequences. Current Bioinformatics. 2009;4(1):28–40.

[40] Weitschek E, Cunial F, Felici G. LAF: Logic alignment free and its application to bacterial genomes classification. BioData Mining. 2015 Dec;8:39.

[41] Fiscon G, Weitschek E, Cella E, Lo Presti A, Giovanetti M, Babakir-Mina M, et al. MISSEL: a method to identify a large number of small species-specific genomic subsequences and its application to viruses classification. BioData Mining. 2016 Dec;9:38.

[42] Remita MA, Halioui A, Malick Diouara AA, Daigle B, Kiani G, Diallo AB. A machine learning approach for viral genome classification. BMC Bioinformatics. 2017 Apr;18:208.

[43] Lu H, Yang L, Yan K, Xue Y, Gao Z. A cost-sensitive rotation forest algorithm for gene expression data classification. Neurocomputing. 2017;228:270–276.

[44] Lu H, Meng Y, Yan K, Gao Z. Kernel principal component analysis combining rotation forest method for linearly inseparable data. Cognitive Systems Research. 2018;53:111–122.

[45] Liu Y, Lu H, Yan K, Xia H, An C. Applying cost-sensitive extreme learning machine and dissimilarity integration to gene expression data classification. Computational Intelligence and Neuroscience. 2016;2016.

[46] Karamichalis R, Kari L. MoDMaps3D: an interactive webtool for the quantification and 3D visualization of interrelationships in a dataset of DNA sequences. Bioinformatics. 2017;33(19):3091–3093.

[47] Li Y, He L, Lucy He R, Yau SST. A novel fast vector method for genetic sequence comparison. Scientific Reports. 2017;7(1).

[48] Cristea PD. Conversion of nucleotide sequences into genomic signals. Journal of Cellular and Molecular Medicine. 2002 04;6(2):279–303.

[49] Afreixo V, Bastos CAC, Pinho AJ, Garcia SP, Ferreira PJSG. Genome analysis with distance to the nearest dissimilar nucleotide. Journal of Theoretical Biology. 2011;275(1):52–58.

[50] Cristea PD. Large scale features in DNA genomic signals. Signal Processing. 2003;83(4):871–888.

[51] Skutkova H, Vitek M, Babula P, Kizek R, Provaznik I. Classification of genomic signals using dynamic time warping. BMC Bioinformatics. 2013 Aug;14(10):S1.

[52] Asuero AG, Sayago A, González AG. The correlation coefficient: an overview. Critical Reviews in Analytical Chemistry. 2006;36(1):41–59.

[53] El-Badawy IM, Aziz AM, Omar Z, Malarvili MB. Correlation between different DNA period-3 signals: An analytical study for exons prediction. In: 2017 Asia-Pacific Signal

and Information Processing Association Annual Summit and Conference; 2017. p. 1123–1128.

[54] Hoang T, Yin C, Zheng H, Yu C, He RL, Yau SST. A new method to cluster DNA sequences using Fourier power spectrum. Journal of Theoretical Biology. 2015;372:135–145.

[55] Sedlar K, Skutkova H, Vitek M, Provaznik I. Set of rules for genomic signal downsampling. Computers in Biology and Medicine. 2016;69:308–314.

[56] Yin C, Chen Y, Yau SST. A measure of DNA sequence similarity by Fourier transform with applications on hierarchical clustering. Journal of Theoretical Biology. 2014;359:18–28.

[57] Strang G, Nguyen T. Wavelets and Filter Banks. Wellesley, MA: Wellesley-Cambridge Press; 1996.

[58] Jones DL. Fathom Toolbox for Matlab: software for multivariate ecological and oceanographic data analysis; 2015. [Online]. Available from: `http://www.marine.usf.edu/user/djones/`.

[59] Lee S, Kwon D, Lee S. Efficient similarity search for time series data based on the minimum distance. In: International Conference on Advanced Information Systems Engineering. vol. 2348; 2002. p. 377–391.

[60] Taki Y. Cyprinid fishes of the genera Onychostoma and Scaphiodonichthys from Upper Laos, with remarks on the dispersal of the genera and their allies. Japanese Journal of Ichthyology. 1975;22(3):143–150.

[61] Zheng L, Yang J, Chen X. Molecular phylogeny and systematics of the Barbinae (Teleostei: Cyprinidae) in China inferred from mitochondrial DNA sequences. Biochemical Systematics and Ecology. 2016;68:250–259.

Chapter 4

MLDSP-GUI: An alignment-free standalone tool with an interactive graphical user interface for DNA sequence comparison and analysis

4.1 Introduction

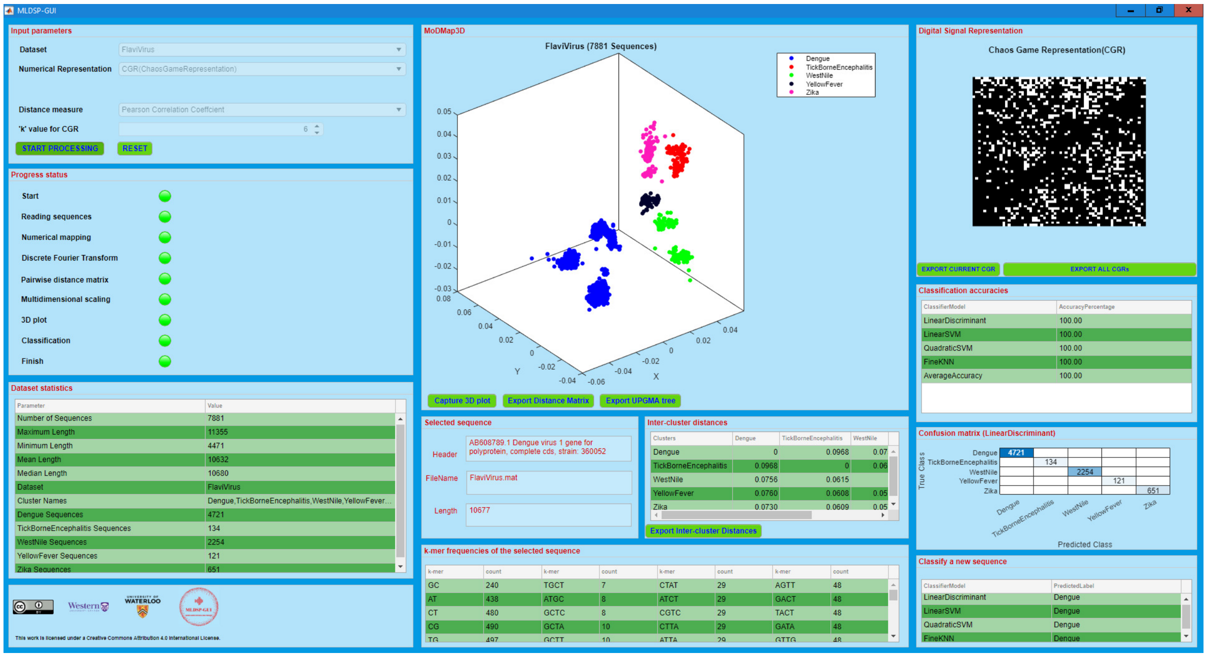
Alignment-based methods have been successfully used for genome classification, but their use has limitations such as the need for contiguous homologous sequences, the heavy memory/time computational cost, and the dependence on *a priori* assumptions about, e.g., the gap penalty and threshold values for statistical parameters. To address these challenges, alignment-free methods have been proposed. [7] defined two categories of alignment-free methods: those that use fixed-length word (oligomer) frequencies, and those that do not require finding fixed-length segments. MLDSP-GUI (Machine Learning with Digital Signal Processing and Graphical User Interface) combines both approaches in that it can use one-dimensional numerical representations of DNA sequences that do not require calculating *k*-mer (oligomers of length *k*) frequencies, see [5] but, in addition, it can also use *k*-mer dependent two-dimensional Chaos Game Representation (CGR) of DNA sequences, see [1, 3].

While alignment-free methods address some of the limitations of alignment-based methods, they still face some challenges. First, most of the existing alignment-free methods lack software implementations, which is necessary for methods to be compared on common datasets. Second, among methods that have software implementations available, the majority have been tested only on simulated sequences or on small real-world datasets. Third, the scalability issue in the form of, e.g., excessive memory overhead and execution time, still remains unsolved for large values of *k*, in the case of *k*-mer based methods.

MLDSP-GUI is a software tool that addresses all of these major challenges and introduces novel features and applications such as: An interactive graphical user interface; Output as either a 3D plot or phylogenetic tree in Newick format; Inter-cluster distance calculation; *k*-mer frequency calculation ($k = 2, 3, 4$) for analysis of under- and over-representation of oligomers; Visualisation of DNA sequences as two-dimensional CGRs; Use of Pearson Correlation Coefficient (PCC), Euclidean or Manhattan distances; Success in classifying large, real-world, datasets. The use of *k*-mer independent one-dimensional numerical representations and Discrete Fourier Transform make MLDSP-GUI ultrafast, memory-economical and scalable, while the use of supervised machine learning leads to classification accuracies over 92%. Lastly, MLDSP-GUI is user-friendly and thus ideally designed for cross-disciplinary applications.

4.2 Materials and methods

MLDSP-GUI is an interactive software tool which implements and significantly augments the ML-DSP approach proposed in [5] for the classification of genomic sequences. It is a pipeline which consists of: *(i)* Computing numerical representations of DNA sequences, *(ii)* applying Discrete Fourier Transform (DFT), *(iii)* calculating pairwise distances, and *(iv)* classifying using supervised machine learning (see Supplementary Material). More precisely, numeri-



Input parameters

Dataset: FlaviVirus
Numerical Representation: CGR(ChaosGameRepresentation)
Distance measure: Pearson Correlation Coeffcient
'k' value for CGR: 6

Progress status

Start ●
Reading sequences ●
Numerical mapping ●
Discrete Fourier Transform ●
Pairwise distance matrix ●
Multidimensional scaling ●
3D plot ●
Classification ●
Finish ●

Dataset statistics

Parameter	Value
Number of Sequences	7881
Maximum Length	11355
Minimum Length	4471
Mean Length	10632
Median Length	10680
Dataset	FlaviVirus
Cluster Names	Dengue,TickBorneEncephalitis,WestNile,YellowFever...
Dengue Sequences	4721
TickBorneEncephalitis Sequences	134
WestNile Sequences	2254
YellowFever Sequences	121
Zika Sequences	651

Selected sequence

Header: AB608789.1 Dengue virus 1 gene for polyprotein, complete cds, strain: 360052
FileName: FlaviVirus.mat
Length: 10677

Inter-cluster distances

Clusters	Dengue	TickBorneEncephalitis	WestNile
Dengue	0	0.0968	0.07
TickBorneEncephalitis	0.0968	0	0.06
WestNile	0.0756	0.0615	
YellowFever	0.0760	0.0608	0.05
Zika	0.0730	0.0609	0.05

k-mer frequencies of the selected sequence

k-mer	count	k-mer	count	k-mer	count	k-mer	count
GC	240	TGCT	7	CTAT	29	AGTT	48
AT	438	ATGC	8	ATCT	29	GACT	48
CT	480	GCTC	8	CGTC	29	TACT	48
CG	490	GCTA	10	CTTA	29	GATA	48
TG	497	GCTT	10	ATTA	29	GTTG	48

Classification accuracies

ClassifierModel	AccuracyPercentage
LinearDiscriminant	100.00
LinearSVM	100.00
QuadraticSVM	100.00
FineKNN	100.00
AverageAccuracy	100.00

Confusion matrix (LinearDiscriminant)

True Class \ Predicted Class	Dengue	TickBorneEncephalitis	WestNile	YellowFever	Zika
Dengue	4721				
TickBorneEncephalitis		134			
WestNile			2254		
YellowFever				121	
Zika					651

Classify a new sequence

ClassifierModel	PredictedLabel
LinearDiscriminant	Dengue
LinearSVM	Dengue
QuadraticSVM	Dengue
FineKNN	Dengue

Figure 4.1: Screenshot of MLDSP-GUI showing a MoDMap3D of 7,881 full mtDNA genomes of the *Flavivirus* genus, classified into species. More details in Supplementary Material.

cal representations are used to represent genomic sequences as discrete numerical sequences that can be treated as digital signals. The corresponding magnitude spectra are then obtained by applying DFT to the numerically represented sequences. A distance measure (PCC, Euclidean, or Manhattan distance) is used to calculate pairwise distances between magnitude spectra. Lastly, supervised machine learning classifiers are trained on feature vectors (consisting of the columns of the pairwise distance matrix of the training set), and then used to classify new sequences. We use 10-fold cross-validation to verify the classification accuracy. Independently, classical multidimensional scaling, see [2, 4, 6], generates a visualization of the classification results in the form of a 3D Molecular Distance Map (MoDMap3D) that displays the dissimilarity-based inter-sequence relationships.

4.3 Software description

MLDSP-GUI not only gives the user the option to visualize an approximation of the interrelationships among sequences in three-dimensional space, but also provides precise quantitative information for further analysis. The distance matrix provides the quantitative dissimilarity between any two points/sequences, while the classification accuracy scores and confusion matrix give a measure of the classification success for each individual classifier. Figure 1 shows a screenshot of MLDSP-GUI used to classify a dataset of 7,881 full mtDNA genomes of the *Flavivirus* genus. The computation of the distance matrix took 12 seconds (PCC, CGR, $k = 6$), the one-time training of the four classifiers and 10-fold cross-validation accuracy computation took 22 mins, and the classification of a new sequence 1 min.

MLDSP-GUI takes DNA sequences in fasta file format as input. Users can select any of the provided datasets, or can input their own dataset. The tool is capable of processing a variety of DNA sequences including natural, simulated, or synthetic sequences. The 3D interactive plot can be rotated, zoomed in/out, and explored by clicking on any of the points. It auto-updates the selected point/sequence statistics such as sequence length, k-mer frequencies, name of parent fasta file, accession number, etc. The supervised machine learning component gives MLDSP-GUI the capability to predict the taxon of any new sequence, provided that it has been trained on a dataset containing that taxon. MLDSP-GUI is implemented using MATLAB R2019a App Designer, license no. 964054. A single executable platform-independent file is provided that can be used to install and run the software tool. The Supplementary Material file provides additional information on MLDSP-GUI features, as well as the provided datasets.

Bibliography

[1] Jeffrey H.J. (1990) Chaos game representation of gene structure. *Nucleic Acids Res.*, 18, 2163-2170.

[2] Karamichalis R. *et al.* (2015) An investigation into inter- and intragenomic variations of graphic genomic signatures, *BMC Bioinformatics*, 16, 246.

[3] Kari L. *et al.* (2015) Mapping the space of genomic signatures, *PLoS ONE*, 10, e0119815.

[4] Kruskal J. (1964) Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29, 1-27.

[5] Randhawa G.S. *et al.* (2019) ML-DSP: Machine Learning with Digital Signal Processing for ultrafast, accurate, and scalable genome classification at all taxonomic levels, *BMC Genomics*, 20, 267.

[6] Solis-Reyes S. *et al.* (2018) An open-source k-mer based machine learning tool for fast and accurate subtyping of HIV-1 genomes, *PLoS ONE*, 13, e0206409.

[7] Zielezinski A. *et al.* (2017) Alignment-free sequence comparison: benefits, applications, and tools, *Genome Biology*, 18, 186.

Chapter 5

Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: COVID-19 case study

5.1 Introduction

Coronaviruses are single-stranded positive-sense RNA viruses that are known to contain some of the largest viral genomes, up to around 32 kbp in length [1, 2, 3, 4, 5]. After increases in the number of coronavirus genome sequences available following efforts to investigate the diversity in the wild, the family *Coronaviridae* now contains four genera (International Committee on Taxonomy of Viruses, [6]). While those species that belong to the genera *Alphacoronavirus* and *Betacoronavirus* can infect mammalian hosts, those in *Gammacoronavirus* and the recently defined *Deltacoronavirus* mainly infect avian species [4, 7, 8, 9]. Phylogenetic studies have revealed a complex evolutionary history, with coronaviruses thought to have ancient origins and recent crossover events that can lead to cross-species infection [8, 10, 11, 12]. Some of the largest sources of diversity for coronaviruses belong to the strains that infect bats and birds, providing a reservoir in wild animals for recombination and mutation that may enable cross-species transmission into other mammals and humans [4, 7, 8, 10, 13].

Like other RNA viruses, coronavirus genomes are known to have genomic plasticity, and this can be attributed to several major factors. RNA-dependent RNA polymerases (RdRp) have high mutation rates, reaching from 1 in 1000 to 1 in 10000 nucleotides during replication [7, 14, 15]. Coronaviruses are also known to use a template switching mechanism which can contribute to high rates of homologous RNA recombination between their viral genomes [9, 16, 17, 18, 19, 20]. Furthermore, the large size of coronavirus genomes is thought to be able to accommodate mutations to genes [7]. These factors help contribute to the plasticity and diversity of coronavirus genomes today.

The highly pathogenic human coronaviruses, Severe Acute Respiratory Syndrome coronavirus (SARS-CoV) and Middle East respiratory syndrome coronavirus (MERS-CoV) belong to lineage B (sub-genus *Sarbecovirus*) and lineage C (sub-genus *Merbecovirus*) of *Betacoronavirus*, respectively [9, 21, 22, 23]. Both result from zoonotic transmission to humans and lead to symptoms of viral pneumonia, including fever, breathing difficulties, and more [24, 25]. Recently, an unidentified pneumonia disease with similar symptoms caused an outbreak in Wuhan and is thought to have started from a local fresh seafood market [26, 27, 28, 29, 30]. This was later attributed to a novel coronavirus (the COVID-19 virus), and represents the third major zoonotic human coronavirus of this century [31]: On February 28, 2020, the World Health Organization set the COVID-19 risk assessment for regional and global levels to “Very High” [32].

From analyses employing whole genome to viral protein-based comparisons, the COVID-19 virus is thought to belong to lineage B (*Sarbecovirus*) of *Betacoronavirus*. From phylogenetic analysis of the RdRp protein, spike proteins, and full genomes of the COVID-19 virus and other coronaviruses, it was found that the COVID-19 virus is most closely related to two bat SARS-like coronaviruses, *bat-SL-CoVZXC21* and *bat-SL-CoVZC45*, found in Chinese horseshoe bats *Rhinolophus sinicus* [12, 33, 34, 35, 36, 37]. Along with the phylogenetic data, the genome organization of the COVID-19 virus was found to be typical of lineage B (*Sar-*

becovirus) *Betacoronaviruses* [33]. From phylogenetic analysis of full genome alignment and similarity plots, it was found that the COVID-19 virus has the highest similarity to the bat coronavirus *RaTG13* [38]. Close associations to bat coronavirus *RaTG13* and two bat SARS-like CoVs (*ZC45* and *ZXC21*) are also supported in alignment-based phylogenetic analyses [38]. Within the COVID-19 virus sequences, over 99% sequence similarity and a lack of diversity within these strains suggest a common lineage and source, with support for recent emergence of the human strain [12, 31]. There is ongoing debate whether the COVID-19 virus arose following recombination with previously identified bat and unknown coronaviruses [39] or arose independently as a new lineage to infect humans [38]. In combination with the identification that the angiotensin converting enzyme 2 (ACE2) protein is a receptor for COVID-19 virus, as it is for SARS and other *Sarbecovirus* strains, the hypothesis that the COVID-19 virus originated from bats is deemed very likely [12, 33, 35, 38, 40, 41, 42, 43, 44].

All analyses performed thus far have been alignment-based and rely on the annotations of the viral genes. Though alignment-based methods have been successful in finding sequence similarities, their application can be challenging in many cases [45, 46]. It is realistically impossible to analyze thousands of complete genomes using alignment-based methods due to the heavy computation time. Moreover, the alignment demands the sequences to be continuously homologous which is not always the case. Alignment-free methods [47, 48, 49, 50, 51] have been proposed in the past as an alternative to address the limitations of the alignment-based methods. Comparative genomics beyond alignment-based approaches have benefited from the computational power of machine learning. Machine learning-based alignment-free methods have also been used successfully for a variety of problems including virus classification [49, 50, 51]. An alignment-free approach [49] was proposed for subtype classification of HIV-1 genomes and achieved $\sim$ 97% classification accuracy. MLDSP [50], with the use of a broad range of $1D$ numerical representations of DNA sequences, has also achieved very high levels of classification accuracy with viruses. Even rapidly evolving, plastic genomes of viruses such as *Influenza* and *Dengue* are classified down to the level of strain and subtype, respectively with 100% classification accuracy. MLDSP-GUI [51] provides an option to use $2D$ Chaos Game Representation (CGR) [52] as numerical representation of DNA sequences. CGR's have a longstanding use in species classification with identification of biases in sequence composition [48, 51, 52]. MLDSP-GUI has shown 100% classification accuracy for *Flavivirus* genus to species classification using $2D$ CGR as numerical representation [51]. MLDSP and MLDSP-GUI have demonstrated the ability to identify the genomic signatures (a species-specific pattern known to be pervasive throughout the genome) with species level accuracy that can be used for sequence (dis)similarity analyses. In this study, we use MLDSP [50] and MLDSP-GUI [51] with CGR as a numerical representation of DNA sequences to assess the classification of the COVID-19 virus from the perspective of machine learning-based alignment-free whole genome comparison of genomic signatures.Using MLDSP and MLDSP-GUI, we confirm that the COVID-19 virus belongs to the *Betacoronavirus*, while its genomic similarity to the sub-genus *Sarbecovirus* supports a possible bat origin.

This paper demonstrates how machine learning using intrinsic genomic signatures can provide rapid alignment-free taxonomic classification of novel pathogens. Our method delivers accurate classifications of the COVID-19 virus without *a priori* biological knowledge, by a simultaneous processing of the geometric space of all relevant viral genomes. The main contributions are:

- Identifying intrinsic viral genomic signatures, and utilizing them for a real-time and highly accurate machine learning-based classification of novel pathogen sequences, such as the COVID-19 virus;
- A general-purpose bare-bones approach, which uses raw DNA sequences alone and does not have any requirements for gene or genome annotation;
- The use of a “decision tree" approach to supervised machine learning (paralleling taxonomic ranks), for successive refinements of taxonomic classification.

- A comprehensive and "in minutes" analysis of a dataset of 5538 unique viral genomic sequences, for a total of 61.8 million bp analyzed, with high classification accuracy scores at all levels, from the highest to the lowest taxonomic rank;
- The use of Spearman's rank correlation analysis to confirm our results and the relatedness of the COVID-19 virus sequences to the known genera of the family *Coronaviridae* and the known sub-genera of the genus *Betacoronavirus*.

5.2 Materials and methods

The Wuhan seafood market pneumonia virus (COVID-19 virus/SARS-CoV-2) isolate Wuhan-Hu-1 complete reference genome of 29903 bp was downloaded from the National Center for Biotechnology Information (NCBI) database on January 23, 2020. All of the available 28 sequences of COVID-19 virus and the bat *Betacoronavirus RaTG13* from the GISAID platform, and two additional sequences (*bat-SL-CoVZC45*, and *bat-SL-CoVZXC21*) from the NCBI, were downloaded on January 27, 2019. All of the available viral sequences were downloaded from the Virus-Host DB (14688 sequences available on January 14, 2020). Virus-Host DB covers the sequences from the NCBI RefSeq (release 96, September 9, 2019) and GenBank (release 233.0, August 15, 2019). All sequences shorter than 2000 bp and longer than 50000 bp were ignored to address possible issues arising from sequence length bias. Accession numbers for all the sequences used in this study can be found in supplementary tables D.S2 and D.S3.

MLDSP [50] and MLDSP-GUI [51] were used as the machine learning-based alignment-free methods for complete genome analyses. As MLDSP-GUI is an extension of the MLDSP methodology, we will refer to the method hereafter as MLDSP-GUI. Each genomic sequence is mapped into its respective genomic signal (a discrete numeric sequence) using a numerical representation. For this study, we use a two-dimensional k-mer (oligomers of length k) based numerical representation known as Chaos Game Representation (CGR) [52]. The k-mer value 7 is used for all the experiments. The value $k = 7$ achieved the highest accuracy scores for the HIV-1 subtype classification [49] and this value could be relevant for other virus related analyses. The magnitude spectra are then calculated by applying Discrete Fourier Transform (DFT) to the genomic signals [50]. A pairwise distance matrix is then computed using the Pearson Correlation Coefficient (PCC) [53] as a distance measure between magnitude spectra. The distance matrix is used to generate the 3D Molecular Distance Maps (MoDMap3D) [54] by applying the classical Multi-Dimensional Scaling (MDS) [55]. MoDMap3D represents an estimation of the relationship among sequences based on the genomic distances between the sequences. The feature vectors are constructed from the columns of the distance matrix and are used as an input to train six supervised-learning based classification models (Linear Discriminant, Linear SVM, Quadratic SVM, Fine KNN, Subspace Discriminant, and Subspace KNN) [50]. A 10-fold cross-validation is used to train, and test the classification models and the average of 10 runs is reported as the classification accuracy. The trained machine learning models are then used to test the COVID-19 virus sequences. The unweighted pair group method with arithmetic mean (UPGMA) [56] and neighbor-joining [57] phylogenetic trees are also computed using the pairwise distance matrix.

In this paper, MLDSP-GUI is augmented by a decision tree approach to the supervised machine learning component and a Spearman's rank correlation coefficient analysis for result validation. The decision tree parallels the taxonomic classification levels, and is necessary so as to minimize the number of calls to the supervised classifier module, as well as to maintain a reasonable number of clusters during each supervised training session. For validation of MLDSP-GUI results using CGR as a numerical representation, we use Spearman's rank correlation coefficient [58, 59, 60, 61], as follows. The frequency of each k-mer is calculated in each genome. Due to differences in genome length between species, proportional frequencies are computed by dividing each k-mer frequency by the length of the respective sequence. To determine whether there is a correlation between k-mer frequencies in COVID-19 virus genomes and specific taxonomic groups, a Spearman's rank correlation coefficient test is conducted for $k = 1$ to $k = 7$.

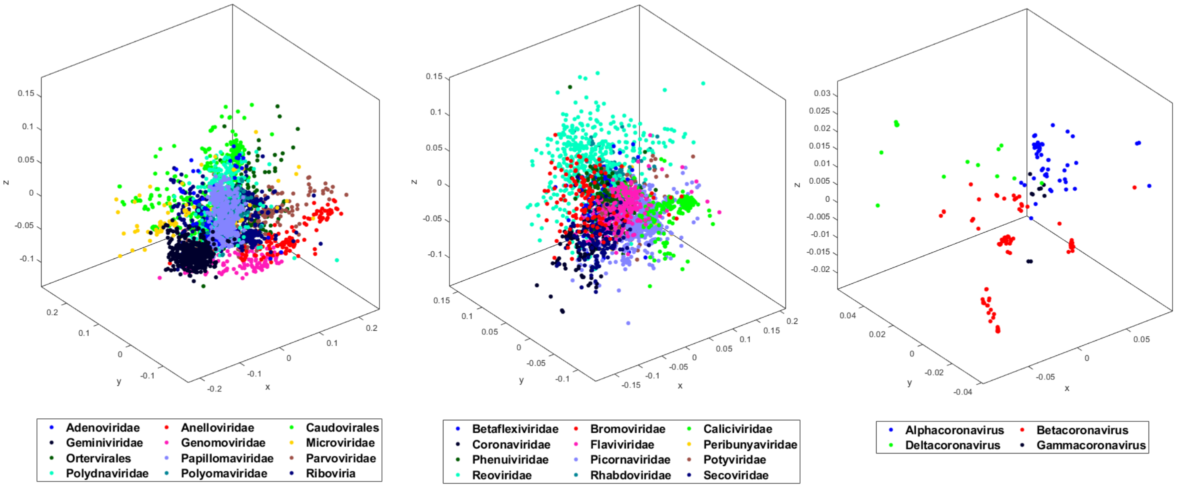
5.3 Results

Table 5.1 provides the details of three datasets Test-1, Test-2, Test-3a and Test-3b used for analyses with MLDSP-GUI. Each dataset's composition (clusters with number of sequences), the respective sequence length statistics, and results of MLDSP-GUI after applying 10-fold cross-validation as classification accuracy scores are shown. The classification accuracy scores for all six classification models are shown with their average, see Table 5.1.

As shown in Table 5.1, for the first test (Test-1), we organized the dataset of sequences into 12 clusters (11 families, and Riboviria realm). Only the families with at least 100 sequences were considered. The Riboviria cluster contains all families that belong to the realm Riboviria. For the clusters with more than 500 sequences, we selected 500 sequences at random. Our method can handle all of the available 14668 sequences, but using imbalanced clusters, in regard to the number of sequences, can introduce an unwanted bias. After filtering out the sequences, our pre-processed dataset is left with 3273 sequences organized into 12 clusters (*Adenoviridae*, *Anelloviridae*, *Caudovirales*, *Geminiviridae*, *Genomoviridae*, *Microviridae*, *Ortervirales*, *Papillomaviridae*, *Parvoviridae*, *Polydnaviridae*, *Polyomaviridae*, and Riboviria). We used MLDSP-GUI with CGR as the numerical representation at $k = 7$. The maximum classification accuracy of 94.9% is obtained using the Quadratic SVM model. The respective MoDMap3D is shown in Fig 5.1(a). All six classification models trained on 3273 sequences were used to classify (predict the labels of) the 29 COVID-19 virus sequences. All of our machine learning-based models correctly predicted and confirmed the label as Riboviria for all 29 sequences (Table 5.2).

Test-1 classified the COVID-19 virus as belonging to the realm Riboviria. The second test (Test-2) is designed to classify the COVID-19 virus among the families of the Riboviria realm. We completed the dataset pre-processing using the same rules as in Test-1 and obtained a dataset of 2779 sequences placed into the 12 families (*Betaflexiviridae*, *Bromoviridae*, *Caliciviridae*, *Coronaviridae*, *Flaviviridae*, *Peribunyaviridae*, *Phenuiviridae*, *Picornaviridae*, *Potyviridae*, *Reoviridae*, *Rhabdoviridae*, and *Secoviridae*), see Table 5.1. MLDSP-GUI with

CGR at $k = 7$ as the numerical representation was used for the classification of the dataset in Test-2. The maximum classification accuracy of 93.1% is obtained using the Quadratic SVM model. The respective MoDMap3D is shown in Fig 5.1(b). All six classification models trained on 2779 sequences were used to classify (predict the label of) the 29 COVID-19 virus sequences. All of our machine learning-based models predicted the label as *Coronaviridae* for all 29 sequences (Table 5.2) with 100% classification accuracy. Test-2 correctly predicted the family of the COVID-19 virus sequences as *Coronaviridae*. Test-3 performs the genus-level classification.



(a) **(b)** **(c)**

Figure 5.1: MoDMap3D of (a) 3273 viral sequences from Test-1 representing 11 viral families and realm Riboviria, (b) 2779 viral sequences from Test-2 classifying 12 viral families of realm Riboviria, (c) 208 *Coronaviridae* sequences from Test-3a classified into genera.

Table 5.1: Classification accuracy scores of viral sequences at different levels of taxonomy.

Dataset	**Clusters**	**Number of sequences**	**Classification model**	**Classification accuracy (in %)**
Test-1: 11 families and Riboviria; 3273 sequences; Maximum length: 49973 Minimum length: 2002 Median length: 7350 Mean length: 13173	*Adenoviridae* *Anelloviridae* *Caudovirales* *Geminiviridae* *Genomoviridae* *Microviridae* *Ortervirales* *Papillomaviridae* *Parvoviridae* *Polydnaviridae* *Polyomaviridae* Riboviria	198 126 500 500 115 102 233 369 182 304 144 500	LinearDiscriminant LinearSVM QuadraticSVM FineKNN SubspaceDiscriminant SubspaceKNN AverageAccuracy	91.7 90.8 95 93.4 87.6 93.2 92
Test-2: Riboviria families; 2779 sequences; Maximum length: 31769 Minimum length: 2005 Median length: 7488 Mean length: 8607	*Betaflexiviridae* *Bromoviridae* *Caliciviridae* *Coronaviridae* *Flaviviridae* *Peribunyaviridae* *Phenuiviridae* *Picornaviridae* *Potyviridae* *Reoviridae* *Rhabdoviridae* *Secoviridae*	121 122 403 210 222 166 107 437 196 470 192 133	LinearDiscriminant LinearSVM QuadraticSVM FineKNN SubspaceDiscriminant SubspaceKNN AverageAccuracy	91.2 89.2 93.1 90.3 89 90.4 90.5
Test-3a: *Coronaviridae*; 208 sequences; Maximum length: 31769 Minimum length: 9580 Median length: 29704 Mean length: 29256	*Alphacoronavirus* *Betacoronavirus* *Deltacoronavirus* *Gammacoronavirus*	53 126 20 9	LinearDiscriminant LinearSVM QuadraticSVM FineKNN SubspaceDiscriminant SubspaceKNN AverageAccuracy	98.1 94.2 95.2 95.7 97.6 96.2 96.2
Test-3b: *Coronaviridae*; 60 sequences; Maximum length: 31429 Minimum length: 25402 Median length: 28475 Mean length: 28187	*Alphacoronavirus* *Betacoronavirus* *Deltacoronavirus*	20 20 20	LinearDiscriminant LinearSVM QuadraticSVM FineKNN SubspaceDiscriminant SubspaceKNN AverageAccuracy	100 93.3 93.3 95 95 95 95.3

All classifiers trained on Test-1, Test-2, Test-3a, and Test-3b datasets were used to predict the labels of 29 COVID-19 virus sequences. All classifiers predicted the correct labels for all of the sequences (Riboviria when trained using Test-1, *Coronaviridae* when trained using Test-2, and *Betacoronavirus* when trained using Test-3a and Test-3b).

Table 5.2: Predicted taxonomic labels of 29 COVID-19 virus sequences.

Training dataset	**Testing dataset**	**Classification models**	**Prediction accuracy (%)**	**Predicted label**
Test-1	29 COVID-19 virus sequences	Linear Discriminant	100	Riboviria
		Linear SVM	100	Riboviria
		Quadratic SVM	100	Riboviria
		Fine KNN	100	Riboviria
		Subspace Discriminant	100	Riboviria
		Subspace KNN	100	Riboviria
Test-2	29 COVID-19 virus sequences	Linear Discriminant	100	*Coronaviridae*
		Linear SVM	100	*Coronaviridae*
		Quadratic SVM	100	*Coronaviridae*
		Fine KNN	100	*Coronaviridae*
		Subspace Discriminant	100	*Coronaviridae*
		Subspace KNN	100	*Coronaviridae*
Test-3(a\b)	29 COVID-19 virus sequences	Linear Discriminant	100	*Betacoronavirus*
		Linear SVM	100	*Betacoronavirus*
		Quadratic SVM	100	*Betacoronavirus*
		Fine KNN	100	*Betacoronavirus*
		Subspace Discriminant	100	*Betacoronavirus*
		Subspace KNN	100	*Betacoronavirus*

The third test (Test-3a) is designed to classify the COVID-19 virus sequences at the genus level. We considered 208 *Coronaviridae* sequences available under four genera (*Alphacoronavirus*, *Betacoronavirus*, *Deltacoronavirus*, *Gammacoronavirus*) (Table 5.1). MLDSP-GUI with CGR at $k = 7$ as the numerical representation was used for the classification of the dataset in Test-3a. The maximum classification accuracy of 98.1% is obtained using the Linear Discriminant model and the respective MoDMap3D is shown in Fig 5.1(c). All six classification models trained on 208 sequences were used to classify (predict the label of) the 29 COVID-19 virus sequences. All of our machine learning-based models predicted the label as *Betacoronavirus* for all 29 sequences (Table 5.2). To verify that the correct prediction is not an artifact of possible bias because of larger *Betacoronavirus* cluster, we did a secondary Test-3b with cluster size limited to the size of smallest cluster (after removing the *Gammacoronavirus* because it just had 9 sequences). The maximum classification accuracy of 100% is obtained using the Linear Discriminant model for Test-3b. All six classification models trained on 60 sequences were used to classify the 29 COVID-19 virus sequences. All of our machine learning-based models predicted the label as *Betacoronavirus* for all 29 sequences (Table 5.2). This secondary test showed that the possible bias is not significant enough to have any impact on the classification performance.

Given confirmation that the COVID-19 virus belongs to the *Betacoronavirus* genus, there now is a question of its origin and relation to the other viruses of the same genus. To examine this question, we preprocessed our dataset from our third test to keep the sub-clusters of the *Betacoronavirus* with at least 10 sequences (Test-4). This gives 124 sequences placed into four clusters (*Embecovirus*, *Merbecovirus*, *Nobecovirus*, *Sarbecovirus*) (Table 5.3). The maximum classification accuracy of 98.4% with CGR at $k = 7$ as the numerical representation is obtained using the Quadratic SVM model. The respective MoDMap3D is shown in Fig 5.2(a). All six classifiers trained on 124 sequences predicted the label as *Sarbecovirus*, when used to predict the labels of 29 COVID-19 virus sequences. For Test-5, we added the COVID-19 virus with 29 sequences as the fifth cluster, see Table 5.3. The maximum classification accuracy of 98.7%

with CGR at $k = 7$ as the numerical representation is obtained using the Subspace Discriminant model. The respective MoDMap3D is shown in Fig 5.2(b). In the MoDMap3D plot from Test-5, COVID-19 virus sequences are placed in a single distinct cluster, see Fig 5.2(b). As visually suggested by the MoDMap3D (Fig 5.2(b)), the average inter-cluster distances confirm that the COVID-19 virus sequences are closest to the *Sarbecovirus* (average distance 0.0556), followed by *Merbecovirus* (0.0746), *Embecovirus* (0.0914), and *Nobecovirus* (0.0916). The three closest sequences based on the average distances from all COVID-19 virus sequences are *RaTG13* (0.0203), *bat-SL-CoVZC45* (0.0418), and *bat-SL-CoVZXC21* (0.0428).

For Test-6, we classified *Sarbecovirus* (47 sequences) and COVID-19 virus (29 sequences) clusters and achieved separation of the two clusters visually apparent in the MoDMap3D, see Fig 5.2(c). Quantitatively, using 10-fold cross-validation, all six of our classifiers report 100% classification accuracy. We generated phylogenetic trees (UPGMA and neighbor-joining) based on all pairwise distances for the dataset in Test-6 that show the separation of the two clusters and relationships within the clusters (Fig 5.3 and 5.4). As observed in Test-5, the phylogenetic trees show that the COVID-19 virus sequences are closer to the *Betacoronavirus RaTG13* sequence collected from a bat host.

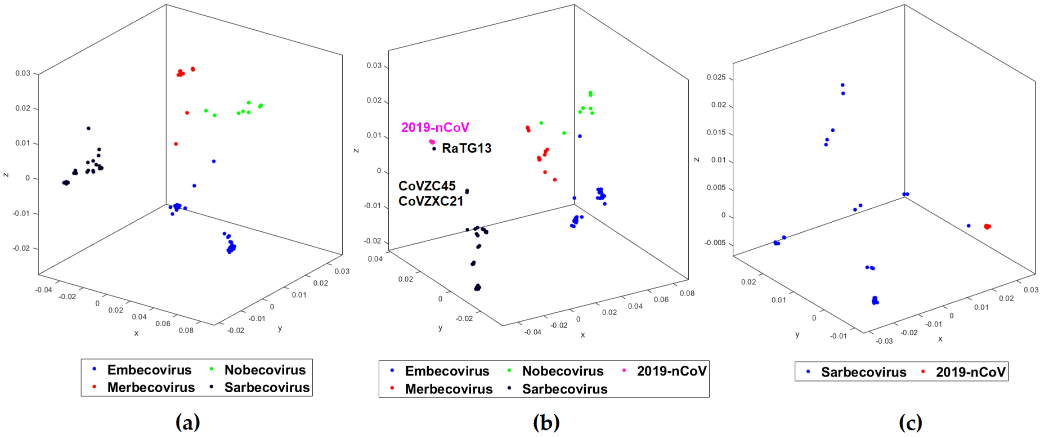


Figure 5.2: MoDMap3D of (a) 124 *Betacoronavirus* sequences from Test-4 classified into sub-genera, (b) 153 viral sequences from Test-5 classified into 4 sub-genera and COVID-19 virus, (c) 76 viral sequences from Test 6 classified into *Sarbecovirus* and COVID-19 virus.

Table 5.3: Genus to sub-genus classification accuracy scores of *Betacoronavirus*.

Dataset	**Clusters**	**Number of sequences**	**Classification model**	**Classification accuracy (in %)**
Test-4: *Betacoronavirus*; 124 sequences; Maximum length: 31526 Minimum length: 29107 Median length: 30155 Mean length: 30300	*Embecovirus* *Merbecovirus* *Nobecovirus* *Sarbecovirus*	49 18 10 47	LinearDiscriminant LinearSVM QuadraticSVM FineKNN SubspaceDiscriminant SubspaceKNN AverageAccuracy	97.6 98.4 98.4 97.6 98.4 97.2 97.6
Test-5: *Betacoronavirus* and COVID-19 virus; 153 sequences; Maximum length: 31526 Minimum length: 29107 Median length: 29891 Mean length: 30217	*Embecovirus* *Merbecovirus* *Nobecovirus* *Sarbecovirus* COVID-19 virus	49 18 10 47 29	LinearDiscriminant LinearSVM QuadraticSVM FineKNN SubspaceDiscriminant SubspaceKNN AverageAccuracy	98.6 97.4 97.4 97.4 98.7 96.1 97.5
Test-6: *Sarbecovirus* and COVID-19 virus; 76 sequences; Maximum length: 30309 Minimum length: 29452 Median length: 29748 Mean length: 29772	*Sarbecovirus* COVID-19 virus	47 29	LinearDiscriminant LinearSVM QuadraticSVM FineKNN SubspaceDiscriminant SubspaceKNN AverageAccuracy	100 100 100 100 100 100 100

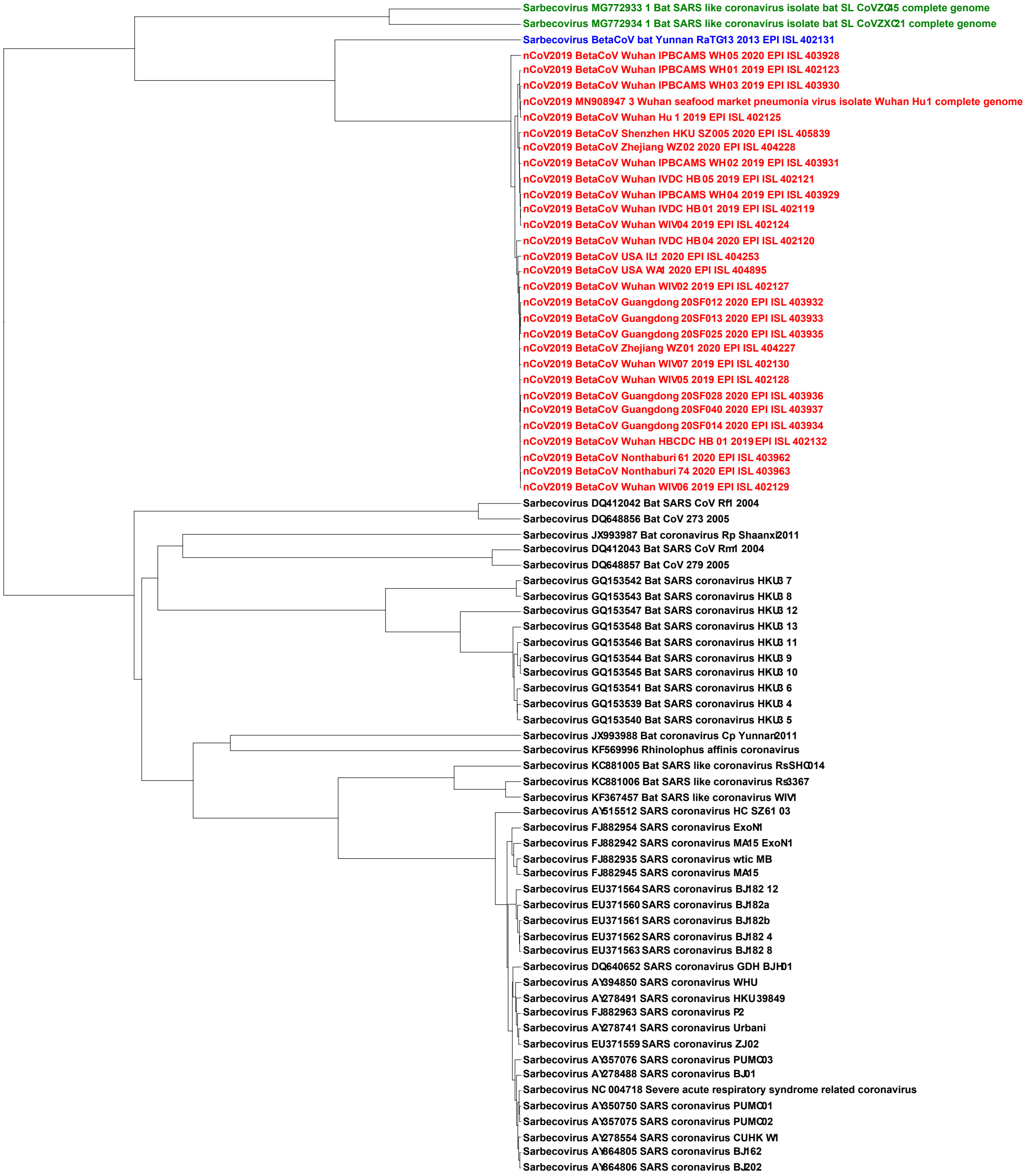

Sarbecovirus MG772933 1 Bat SARS like coronavirus isolate bat SL CoVZC45 complete genome
Sarbecovirus MG772934 1 Bat SARS like coronavirus isolate bat SL CoVZXC21 complete genome
Sarbecovirus BetaCoV bat Yunnan RaTG13 2013 EPI ISL 402131
nCoV2019 BetaCoV Wuhan IPBCAMS WH05 2020 EPI ISL 403928
nCoV2019 BetaCoV Wuhan IPBCAMS WH01 2019 EPI ISL 402123
nCoV2019 BetaCoV Wuhan IPBCAMS WH03 2019 EPI ISL 403930
nCoV2019 MN908947 3 Wuhan seafood market pneumonia virus isolate Wuhan Hu1 complete genome
nCoV2019 BetaCoV Wuhan Hu 1 2019 EPI ISL 402125
nCoV2019 BetaCoV Shenzhen HKU SZ005 2020 EPI ISL 405839
nCoV2019 BetaCoV Zhejiang WZ02 2020 EPI ISL 404228
nCoV2019 BetaCoV Wuhan IPBCAMS WH02 2019 EPI ISL 403931
nCoV2019 BetaCoV Wuhan IVDC HB05 2019 EPI ISL 402121
nCoV2019 BetaCoV Wuhan IPBCAMS WH04 2019 EPI ISL 403929
nCoV2019 BetaCoV Wuhan IVDC HB01 2019 EPI ISL 402119
nCoV2019 BetaCoV Wuhan WIV04 2019 EPI ISL 402124
nCoV2019 BetaCoV Wuhan IVDC HB04 2020 EPI ISL 402120
nCoV2019 BetaCoV USA IL1 2020 EPI ISL 404253
nCoV2019 BetaCoV USA WA1 2020 EPI ISL 404895
nCoV2019 BetaCoV Wuhan WIV02 2019 EPI ISL 402127
nCoV2019 BetaCoV Guangdong 20SF012 2020 EPI ISL 403932
nCoV2019 BetaCoV Guangdong 20SF013 2020 EPI ISL 403933
nCoV2019 BetaCoV Guangdong 20SF025 2020 EPI ISL 403935
nCoV2019 BetaCoV Zhejiang WZ01 2020 EPI ISL 404227
nCoV2019 BetaCoV Wuhan WIV07 2019 EPI ISL 402130
nCoV2019 BetaCoV Wuhan WIV05 2019 EPI ISL 402128
nCoV2019 BetaCoV Guangdong 20SF028 2020 EPI ISL 403936
nCoV2019 BetaCoV Guangdong 20SF040 2020 EPI ISL 403937
nCoV2019 BetaCoV Guangdong 20SF014 2020 EPI ISL 403934
nCoV2019 BetaCoV Wuhan HBCDC HB 01 2019 EPI ISL 402132
nCoV2019 BetaCoV Nonthaburi 61 2020 EPI ISL 403962
nCoV2019 BetaCoV Nonthaburi 74 2020 EPI ISL 403963
nCoV2019 BetaCoV Wuhan WIV06 2019 EPI ISL 402129
Sarbecovirus DQ412042 Bat SARS CoV Rf1 2004
Sarbecovirus DQ648856 Bat CoV 273 2005
Sarbecovirus JX993987 Bat coronavirus Rp Shaanxi2011
Sarbecovirus DQ412043 Bat SARS CoV Rm1 2004
Sarbecovirus DQ648857 Bat CoV 279 2005
Sarbecovirus GQ153542 Bat SARS coronavirus HKU3 7
Sarbecovirus GQ153543 Bat SARS coronavirus HKU3 8
Sarbecovirus GQ153547 Bat SARS coronavirus HKU3 12
Sarbecovirus GQ153548 Bat SARS coronavirus HKU3 13
Sarbecovirus GQ153546 Bat SARS coronavirus HKU3 11
Sarbecovirus GQ153544 Bat SARS coronavirus HKU3 9
Sarbecovirus GQ153545 Bat SARS coronavirus HKU3 10
Sarbecovirus GQ153541 Bat SARS coronavirus HKU3 6
Sarbecovirus GQ153539 Bat SARS coronavirus HKU3 4
Sarbecovirus GQ153540 Bat SARS coronavirus HKU3 5
Sarbecovirus JX993988 Bat coronavirus Cp Yunnan2011
Sarbecovirus KF569996 Rhinolophus affinis coronavirus
Sarbecovirus KC881005 Bat SARS like coronavirus RsSHC014
Sarbecovirus KC881006 Bat SARS like coronavirus Rs3367
Sarbecovirus KF367457 Bat SARS like coronavirus WIV1
Sarbecovirus AY515512 SARS coronavirus HC SZ61 03
Sarbecovirus FJ882954 SARS coronavirus ExoN1
Sarbecovirus FJ882942 SARS coronavirus MA15 ExoN1
Sarbecovirus FJ882935 SARS coronavirus wtic MB
Sarbecovirus FJ882945 SARS coronavirus MA15
Sarbecovirus EU371564 SARS coronavirus BJ182 12
Sarbecovirus EU371560 SARS coronavirus BJ182a
Sarbecovirus EU371561 SARS coronavirus BJ182b
Sarbecovirus EU371562 SARS coronavirus BJ182 4
Sarbecovirus EU371563 SARS coronavirus BJ182 8
Sarbecovirus DQ640652 SARS coronavirus GDH BJH01
Sarbecovirus AY394850 SARS coronavirus WHU
Sarbecovirus AY278491 SARS coronavirus HKU39849
Sarbecovirus FJ882963 SARS coronavirus P2
Sarbecovirus AY278741 SARS coronavirus Urbani
Sarbecovirus EU371559 SARS coronavirus ZJ02
Sarbecovirus AY357076 SARS coronavirus PUMC03
Sarbecovirus AY278488 SARS coronavirus BJ01
Sarbecovirus NC 004718 Severe acute respiratory syndrome related coronavirus
Sarbecovirus AY350750 SARS coronavirus PUMC01
Sarbecovirus AY357075 SARS coronavirus PUMC02
Sarbecovirus AY278554 SARS coronavirus CUHK W1
Sarbecovirus AY864805 SARS coronavirus BJ162
Sarbecovirus AY864806 SARS coronavirus BJ202

Figure 5.3: The UPGMA phylogenetic tree using the Pearson Correlation Coefficient generated pairwise distance matrix shows COVID-19 virus (Red) sequences proximal to the bat *Betacoronavirus RaTG13* (Blue) and bat SARS-like coronaviruses *ZC45/ZXC21* (Green) in a distinct lineage from the rest of *Sarbecovirus* sequences (Black).

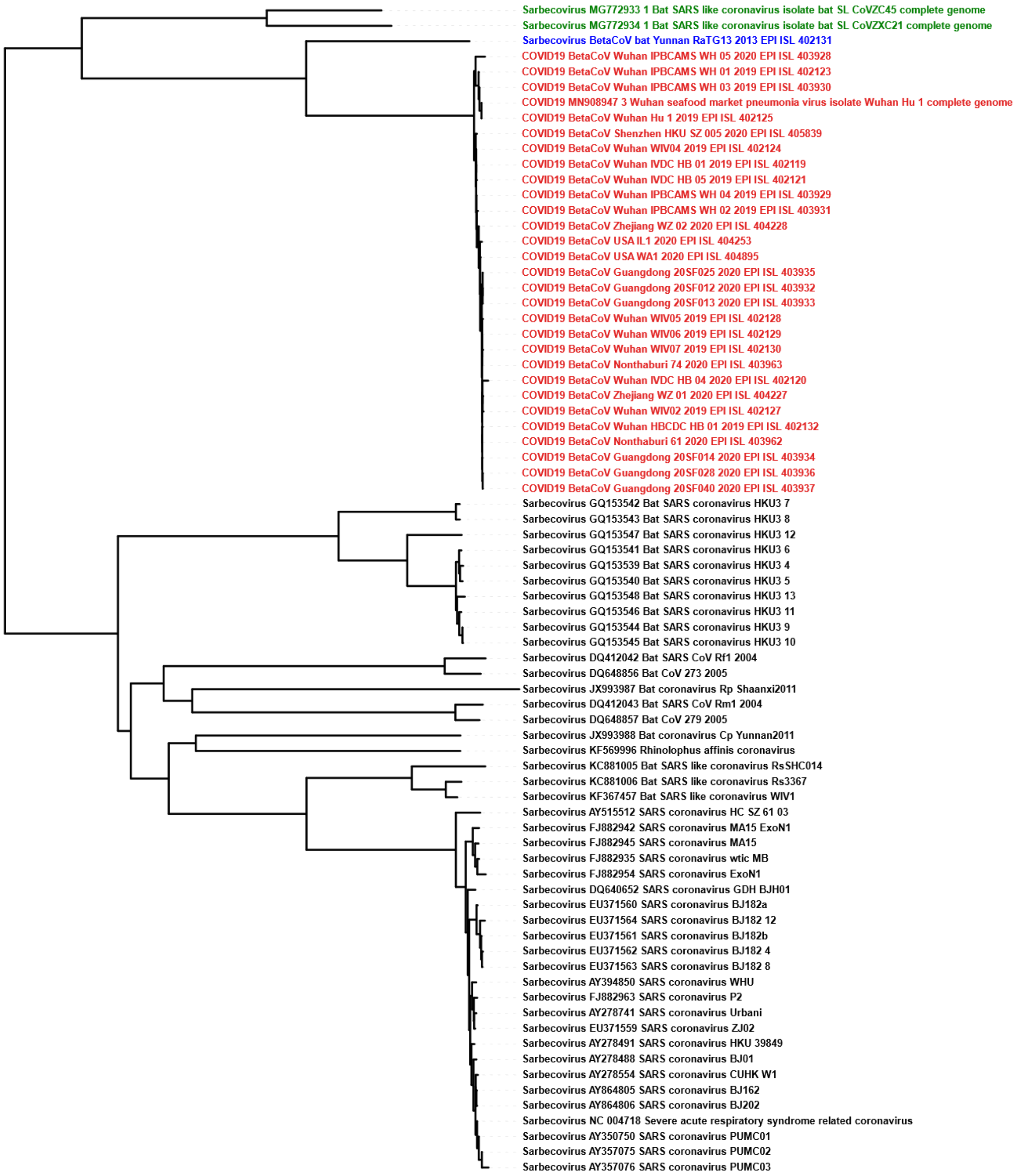
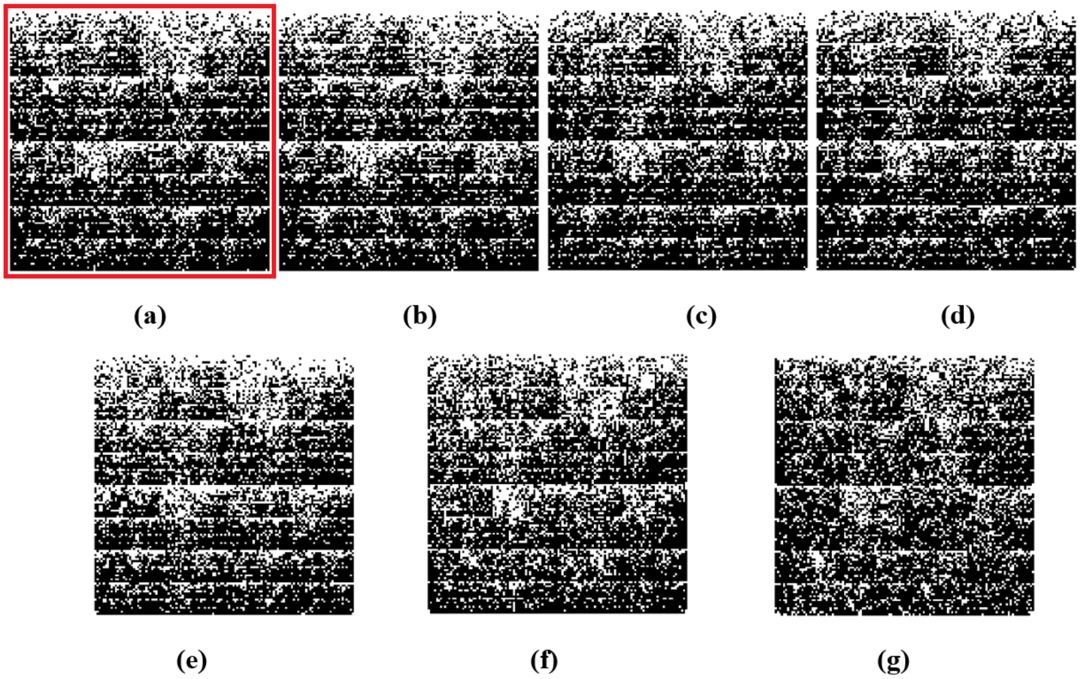

Sarbecovirus MG772933 1 Bat SARS like coronavirus isolate bat SL CoVZC45 complete genome
Sarbecovirus MG772934 1 Bat SARS like coronavirus isolate bat SL CoVZXC21 complete genome
Sarbecovirus BetaCoV bat Yunnan RaTG13 2013 EPI ISL 402131
COVID19 BetaCoV Wuhan IPBCAMS WH 05 2020 EPI ISL 403928
COVID19 BetaCoV Wuhan IPBCAMS WH 01 2019 EPI ISL 402123
COVID19 BetaCoV Wuhan IPBCAMS WH 03 2019 EPI ISL 403930
COVID19 MN908947 3 Wuhan seafood market pneumonia virus isolate Wuhan Hu 1 complete genome
COVID19 BetaCoV Wuhan Hu 1 2019 EPI ISL 402125
COVID19 BetaCoV Shenzhen HKU SZ 005 2020 EPI ISL 405839
COVID19 BetaCoV Wuhan WIV04 2019 EPI ISL 402124
COVID19 BetaCoV Wuhan IVDC HB 01 2019 EPI ISL 402119
COVID19 BetaCoV Wuhan IVDC HB 05 2019 EPI ISL 402121
COVID19 BetaCoV Wuhan IPBCAMS WH 04 2019 EPI ISL 403929
COVID19 BetaCoV Wuhan IPBCAMS WH 02 2019 EPI ISL 403931
COVID19 BetaCoV Zhejiang WZ 02 2020 EPI ISL 404228
COVID19 BetaCoV USA IL1 2020 EPI ISL 404253
COVID19 BetaCoV USA WA1 2020 EPI ISL 404895
COVID19 BetaCoV Guangdong 20SF025 2020 EPI ISL 403935
COVID19 BetaCoV Guangdong 20SF012 2020 EPI ISL 403932
COVID19 BetaCoV Guangdong 20SF013 2020 EPI ISL 403933
COVID19 BetaCoV Wuhan WIV05 2019 EPI ISL 402128
COVID19 BetaCoV Wuhan WIV06 2019 EPI ISL 402129
COVID19 BetaCoV Wuhan WIV07 2019 EPI ISL 402130
COVID19 BetaCoV Nonthaburi 74 2020 EPI ISL 403963
COVID19 BetaCoV Wuhan IVDC HB 04 2020 EPI ISL 402120
COVID19 BetaCoV Zhejiang WZ 01 2020 EPI ISL 404227
COVID19 BetaCoV Wuhan WIV02 2019 EPI ISL 402127
COVID19 BetaCoV Wuhan HBCDC HB 01 2019 EPI ISL 402132
COVID19 BetaCoV Nonthaburi 61 2020 EPI ISL 403962
COVID19 BetaCoV Guangdong 20SF014 2020 EPI ISL 403934
COVID19 BetaCoV Guangdong 20SF028 2020 EPI ISL 403936
COVID19 BetaCoV Guangdong 20SF040 2020 EPI ISL 403937
Sarbecovirus GQ153542 Bat SARS coronavirus HKU3 7
Sarbecovirus GQ153543 Bat SARS coronavirus HKU3 8
Sarbecovirus GQ153547 Bat SARS coronavirus HKU3 12
Sarbecovirus GQ153541 Bat SARS coronavirus HKU3 6
Sarbecovirus GQ153539 Bat SARS coronavirus HKU3 4
Sarbecovirus GQ153540 Bat SARS coronavirus HKU3 5
Sarbecovirus GQ153548 Bat SARS coronavirus HKU3 13
Sarbecovirus GQ153546 Bat SARS coronavirus HKU3 11
Sarbecovirus GQ153544 Bat SARS coronavirus HKU3 9
Sarbecovirus GQ153545 Bat SARS coronavirus HKU3 10
Sarbecovirus DQ412042 Bat SARS CoV Rf1 2004
Sarbecovirus DQ648856 Bat CoV 273 2005
Sarbecovirus JX993987 Bat coronavirus Rp Shaanxi2011
Sarbecovirus DQ412043 Bat SARS CoV Rm1 2004
Sarbecovirus DQ648857 Bat CoV 279 2005
Sarbecovirus JX993988 Bat coronavirus Cp Yunnan2011
Sarbecovirus KF569996 Rhinolophus affinis coronavirus
Sarbecovirus KC881005 Bat SARS like coronavirus RsSHC014
Sarbecovirus KC881006 Bat SARS like coronavirus Rs3367
Sarbecovirus KF367457 Bat SARS like coronavirus WIV1
Sarbecovirus AY515512 SARS coronavirus HC SZ 61 03
Sarbecovirus FJ882942 SARS coronavirus MA15 ExoN1
Sarbecovirus FJ882945 SARS coronavirus MA15
Sarbecovirus FJ882935 SARS coronavirus wtic MB
Sarbecovirus FJ882954 SARS coronavirus ExoN1
Sarbecovirus DQ640652 SARS coronavirus GDH BJH01
Sarbecovirus EU371560 SARS coronavirus BJ182a
Sarbecovirus EU371564 SARS coronavirus BJ182 12
Sarbecovirus EU371561 SARS coronavirus BJ182b
Sarbecovirus EU371562 SARS coronavirus BJ182 4
Sarbecovirus EU371563 SARS coronavirus BJ182 8
Sarbecovirus AY394850 SARS coronavirus WHU
Sarbecovirus FJ882963 SARS coronavirus P2
Sarbecovirus AY278741 SARS coronavirus Urbani
Sarbecovirus EU371559 SARS coronavirus ZJ02
Sarbecovirus AY278491 SARS coronavirus HKU 39849
Sarbecovirus AY278488 SARS coronavirus BJ01
Sarbecovirus AY278554 SARS coronavirus CUHK W1
Sarbecovirus AY864805 SARS coronavirus BJ162
Sarbecovirus AY864806 SARS coronavirus BJ202
Sarbecovirus NC 004718 Severe acute respiratory syndrome related coronavirus
Sarbecovirus AY350750 SARS coronavirus PUMC01
Sarbecovirus AY357075 SARS coronavirus PUMC02
Sarbecovirus AY357076 SARS coronavirus PUMC03

Figure 5.4: The neighbor-joining phylogenetic tree using the Pearson Correlation Coefficient generated pairwise distance matrix shows COVID-19 virus (Red) sequences proximal to the bat *Betacoronavirus RaTG13* (Blue) and bat SARS-like coronaviruses *ZC45/ZXC21* (Green) in a distinct lineage from the rest of *Sarbecovirus* sequences (Black).

Fig 5.5 shows the Chaos Game Representation (CGR) plots of different sequences from the four different genera (*Alphacoronavirus*, *Betacoronavirus*, *Deltacoronavirus*, *Gammacoronavirus*) of the family *Coronaviridae*. The CGR plots visually suggest and the pairwise distances confirm that the genomic signature of the COVID-19 virus Wuhan-Hu-1 (Fig 5.5(a)) is closer to the genomic signature of the *BetaCov-RaTG13* (Fig 5.5(b); distance: 0.0204), followed by the genomic signatures of *bat-SL-CoVZC45* (Fig 5.5(c); distance: 0.0417), *bat-SL-CoVZXC21*(Fig 5.5(d); distance: 0.0428), *Alphacoronavirus*/*DQ*811787 *PRCV ISU*-1 (Fig 5.5(e); distance: 0.0672), *Gammacoronavirus*/Infectious bronchitis virus NGA/*A*116*E*7/2006/*FN*430415 (Fig 5.5(f); distance: 0.0791), and *Deltacoronavirus* /PDCoV/USA/Illinois121/2014/*KJ*481931 (Fig 5.5(g); distance: 0.0851).


(a) **(b)** **(c)** **(d)**

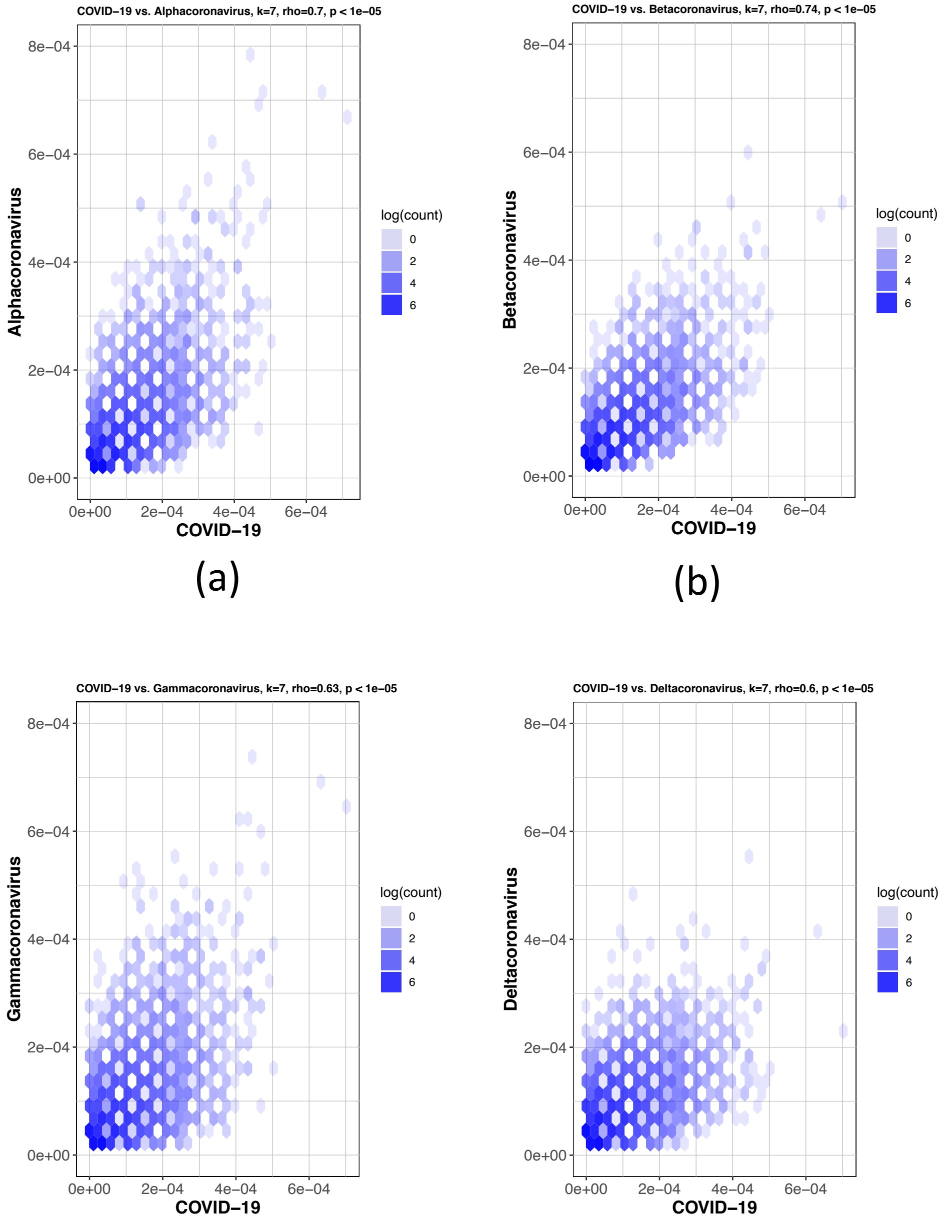
(e) **(f)** **(g)**

Figure 5.5: Chaos Game Representation (CGR) plots at $k = 7$ of (a) COVID-19 virus / Wuhan seafood market pneumonia virus isolate Wuhan-Hu-1/*MN908947.3*, (b) *Betacoronavirus* / CoV / Bat / Yunnan / *RaTG13* / *EPI_ISL_402131*, (c) *Betacoronavirus* / Bat SARS-like coronavirus isolate *bat-SL-CoVZC45* / *MG772933.1*, (d) *Betacoronavirus* / Bat SARS-like coronavirus isolate *bat-SL-CoVZXC21* / *MG772934.1*, (e) *Alphacoronavirus* /*DQ811787 PRCV ISU−1*, (f) *Gammacoronavirus* / Infectious bronchitis virus NGA /*A116E7* / *2006* / *FN430415*, and (g) *Deltacoronavirus* / PDCoV / USA / Illinois121 /2014/*KJ481931*. Chaos plot vertices are assigned top left Cytosine, top right Guanine, bottom left Adenine and bottom right Thymine.

The Spearman's rank correlation coefficient tests were used to further confirm the MLDSP findings. The first test in Fig 5.6 shows COVID-19 virus being compared to the four genera; *Alphacoronavirus*, *Betacoronavirus*, *Gammacoronavirus* and *Deltacoronavirus*. The COVID-19 virus showed the highest k-mer frequency correlation to *Betacoronavirus* at $k = 7$ (Table 5.4), which is consistent with the MLDSP results in Test-3 (Table 5.2). The COVID-19 virus was then compared to all sub-genera within the *Betacoronavirus* genus: *Embecovirus*, *Merbecovirus*, *Nobecovirs* and *Sarbecovirus* seen in Fig 5.7. The Spearman's rank test was again consistent with the MLDSP results seen in Table 5.3, as the k-mer frequencies at $k = 7$ showed the highest correlation to the sub-genus *Sarbecovirus* (Table 5.4). These tests confirm the findings in MLDSP and are consistent with the COVID-19 virus as part of the sub-genus *Sarbecovirus*.

Table 5.4: Spearman's rank correlation coefficient (ρ) values from Fig 5.6 and 5.7, for which all p-values $< 10^{-5}$. The strongest correlation value was found between *Betacoronavirus* and *Sarbecovirus* when using the data sets from Test 3a from Table 5.2 and Test 4 from Table 5.3, respectively.

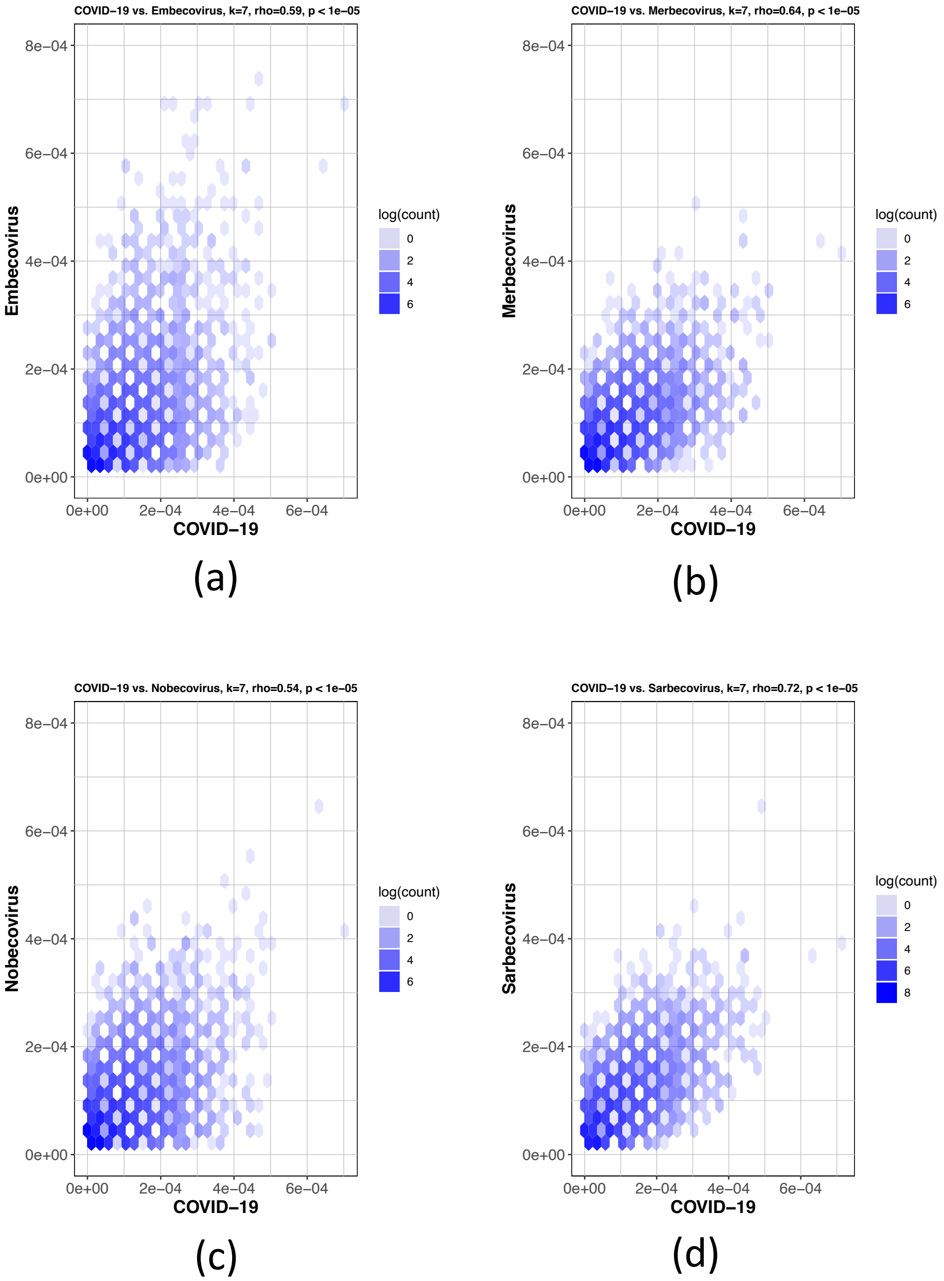
Dataset	**Comparison Groups COVID-19 virus vs.**	**ρ value**
Test-3a	*Alphacoronavirus*	0.70
	Betacoronavirus	**0.74**
	Gammacoronavirus	0.63
	Deltacoronavirus	0.60
Test-4	*Embecovirus*	0.59
	Merbecovirus	0.64
	Nobecovirus	0.54
	Sarbecovirus	**0.72**



(a) (b)

(c) (d)

Figure 5.6: Hexbin scatterplots of the proportional k-mer ($k = 7$) frequencies of the COVID-19 virus sequences vs. the four genera: (a) *Alphacoronavirus*, $\rho = 0.7$; (b) *Betacoronavirus*, $\rho = 0.74$; (c) *Gammacoronavirus*, $\rho = 0.63$ and (d) *Deltacoronavirus*, $\rho = 0.6$. The color of each hexagonal bin in the plot represents the number of points (in natural logarithm scale) overlapping at that position. All ρ values resulted in p-values $< 10^{-5}$ for the correlation test. By visually inspecting each hexbin scatterplot, the degree of correlation is displayed by the variation in spread between the points. Hexagonal points that are closer together and less dispersed as seen in (b) are more strongly correlated and have less deviation.



(a) (b)

(c) (d)

Figure 5.7: Hexbin scatterplots of the proportional k-mer ($k = 7$) frequencies of the COVID-19 virus sequences vs. the four sub-genera: (a) *Embecovirus*, $\rho = 0.59$; (b) *Merbecovirus*, $\rho = 0.64$; (c) *Nobecovirus*, $\rho = 0.54$ and (d) *Sarbecovirus*, $\rho = 0.72$. The color of each hexagonal bin in the plot represents the number of points (in natural logarithm scale) overlapping at that position. All ρ values resulted in p-values $< 10^{-5}$ for the correlation test. By visually inspecting each hexbin scatterplot, the degree of correlation is displayed by the variation in spread between the points. Hexagonal points that are closer together and less dispersed as seen in (d) are more strongly correlated and have less deviation.

5.4 Discussion

Prior work elucidating the evolutionary history of the COVID-19 virus had suggested an origin from bats prior to zoonotic transmission [12, 33, 35, 38, 41, 62]. Most early cases of individuals infected with the COVID-19 virus had contact with the Huanan South China Seafood Market [26, 27, 28, 29, 30, 31]. Human-to-human transmission is confirmed, further highlighting the need for continued intervention [33, 62, 63, 64]. Still, the early COVID-19 virus genomes that have been sequenced and uploaded are over 99% similar, suggesting these infections result from a recent cross-species event [12, 31, 40].

These prior analyses relied upon alignment-based methods to identify relationships between the COVID-19 virus and other coronaviruses with nucleotide and amino acid sequence similarities. When analyzing the conserved replicase domains of ORF1ab for coronavirus species classification, nearly 94% of amino acid residues were identical to SARS-CoV, yet overall genome similarity was only around 70%, confirming that the COVID-19 virus was genetically different [64]. Within the RdRp region, it was found that another bat coronavirus, *RaTG13*, was the closest relative to the COVID-19 virus and formed a distinct lineage from other bat SARS-like coronaviruses [38, 40]. Other groups found that two bat SARS-like coronaviruses, *bat-SL-CoVZC45* and *bat-SL-CoVZXC21*, were also closely related to the COVID-19 virus [12, 33, 34, 35, 36, 37]. There is a consensus that these three bat viruses are most similar to the COVID-19 virus, however, whether or not the COVID-19 virus arose from a recombination event is still unknown [38, 39, 40].

Regardless of the stance on recombination, current consensus holds that the hypothesis of the COVID-19 virus originating from bats is highly likely. Bats have been identified as a reservoir of mammalian viruses and cross-species transmission to other mammals, including humans [4, 7, 8, 10, 13, 65, 66, 67]. Prior to intermediary cross-species infection, the coronaviruses SARS-CoV and MERS-CoV were also thought to have originated in bats [24, 25, 34, 68, 69, 70]. Many novel SARS-like coronaviruses have been discovered in bats across China, and even in European, African and other Asian countries [34, 71, 72, 73, 74, 75, 76, 77]. With widespread geographic coverage, SARS-like coronaviruses have likely been present in bats for a long period of time and novel strains of these coronaviruses can arise through recombination [4]. Whether or not the COVID-19 virus was transmitted directly from bats, or from intermediary hosts, is still unknown, and will require identification of the COVID-19 virus in species other than humans, notably from the wet market and surrounding area it is thought to have originated from [30]. While bats have been reported to have been sold at the Huanan market, at this time, it is still unknown if there were intermediary hosts involved prior to transmission to humans [27, 31, 33, 39, 78]. Snakes had been proposed as an intermediary host for the COVID-19 virus based on relative synonymous codon usage bias studies between viruses and their hosts [39], however, this claim has been disputed [79]. China CDC released information about environmental sampling in the market and indicated that 33 of 585 samples had evidence of the COVID-19 virus, with 31 of these positive samples taken from the location where wildlife booths were concentrated, suggesting possible wildlife origin [80, 81]. Detection of SARS-CoV in Himalyan palm civets and horseshoe bats identified 29 nucleotide sequences that helped trace the origins of SARS-CoV isolates in humans to these intermediary species [13, 24, 38, 77]. Sampling additional animals at the market and wildlife in the surrounding area may help elucidate whether intermediary species were involved or not, as was possible with the SARS-CoV.

Viral outbreaks like COVID-19 demand timely analysis of genomic sequences to guide the research in the right direction. This problem being time-sensitive requires quick sequence similarity comparison against thousands of known sequences to narrow down the candidates of possible origin. Alignment-based methods are known to be time-consuming and can be challenging in cases where homologous sequence continuity cannot be ensured. It is challenging (and sometimes impossible) for alignment-based methods to compare a large number of sequences that are too different in their composition. Alignment-free methods have been used successfully in the past to address the limitations of the alignment-based methods [48, 49, 50, 51]. The alignment-free approach is quick and can handle a large number of sequences. Moreover, even the sequences coming from different regions with different compositions can be easily compared quantitatively, with equally meaningful results as when comparing homologous/similar sequences. We use MLDSP-GUI (a variant of MLDSP with additional features), a machine learning-based alignment-free method successfully used in the past for sequence comparisons and analyses [50]. The main advantage alignment-free methodology offers is the ability to analyze large datasets rapidly. In this study we confirm the taxonomy of the COVID-19 virus and, more generally, propose a method to efficiently analyze and classify a novel unclassified DNA sequence against the background of a large dataset. We namely use a “decision tree" approach (paralleling taxonomic ranks), and start with the highest taxonomic level, train the classification models on the available complete genomes, test the novel unknown sequences to predict the label among the labels of the training dataset, move to the next taxonomic level, and repeat the whole process down to the lowest taxonomic label.

Test-1 starts at the highest available level and classifies the viral sequences to the 11 families and Riboviria realm (Table 5.1). There is only one realm available in the viral taxonomy, so all of the families that belong to the realm Riboviria are placed into a single cluster and a random collection of 500 sequences are selected. No realm is defined for the remaining 11 families. The objective is to train the classification models with the known viral genomes and then predict the labels of the COVID-19 virus sequences. The maximum classification accuracy score of 95% was obtained using the Quadratic SVM model. This test demonstrates that MLDSP-GUI can distinguish between different viral families. The trained models are then used to predict the labels of 29 COVID-19 virus sequences. As expected, all classification models correctly predict that the COVID-19 virus sequences belong to the Riboviria realm, see Table 5.2. Test-2 is composed of 12 families from the Riboviria, see Table 5.1, and the goal is to test if MLDSP-GUI is sensitive enough to classify the sequences at the next lower taxonomic level. It should be noted that as we move down the taxonomic levels, sequences become much more similar to one another and the classification problem becomes challenging. MLDSP-GUI is still able to distinguish between the sequences within the Riboviria realm with a maximum classification accuracy of 91.1% obtained using the Linear Discriminant classification model. When the COVID-19 virus sequences are tested using the models trained on Test-2, all of the models correctly predict the COVID-19 virus sequences as *Coronaviridae* (Table 5.2). Test-3a moves down another taxonomic level and classifies the *Coronaviridae* family to four genera (*Alphacoronavirus*, *Betacoronavirus*, *Deltacoronavirus*, *Gammacoronavirus*), see Table 5.1. MLDSP-GUI distinguishes sequences at the genus level with a maximum classification accuracy score of 98%, obtained using the Linear Discriminant model. This is a very high accuracy rate considering that no alignment is involved and the sequences are very similar. All trained classification models correctly predict the COVID-19 virus as *Betacoronavirus*, see Table 5.2. Test-3a has *Betacoronavirus* as the largest cluster and it can be argued that the higher accuracy could be a result of this bias. To avoid bias, we did an additional test removing the smallest cluster *Gammacoronavirus* and limiting the size of remaining three clusters to the size of the cluster with the minimum number of sequences i.e. 20 with Test-3b. MLDSP-GUI obtains 100% classification accuracy for this additional test and still predicts all of the COVID-19 virus sequences as *Betacoronavirus*. These tests confirm that the COVID-19 virus sequences are from the genus *Betacoronavirus*.

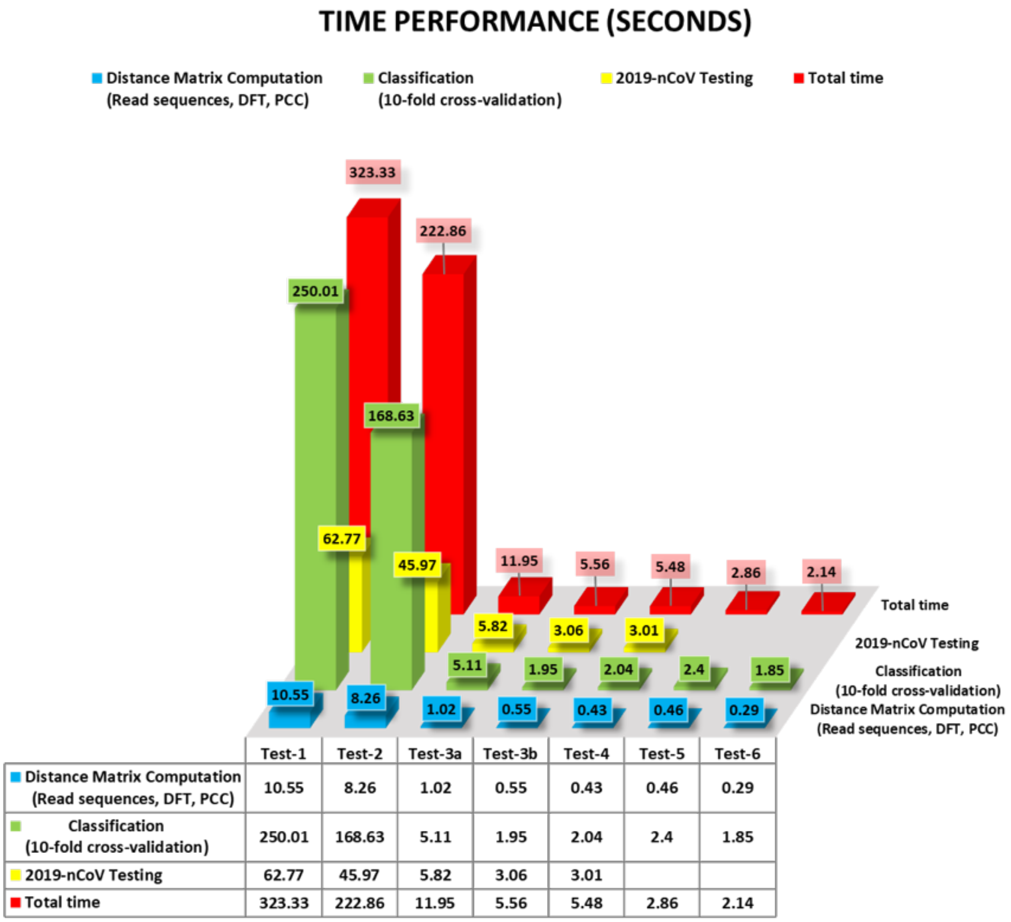
Sequences become very similar at lower taxonomic levels (sub-genera and species). Test-4, Test-5, and Test-6 investigate within the genus *Betacoronavirus* for sub-genus classification. Test-4 is designed to classify *Betacoronavirus* into the four sub-genera (*Embecovirus*, *Merbecovirus*, *Nobecovirus*, *Sarbecovirus*), see Table 5.3. MLDSP-GUI distinguishes sequences at the sub-genus level with a maximum classification accuracy score of 98.4%, obtained using the Quadratic SVM model. All of the classification models trained on the dataset in Test-4 predicted the label of all 29 COVID-19 virus sequences as *Sarbecovirus*. This suggests substantial similarity between the COVID-19 virus and the *Sarbecovirus* sequences. Test-5 and Test-6 (see Table 5.3) are designed to verify that the COVID-19 virus sequences can be differentiated from the known species in the *Betacoronavirus* genus. MLDSP-GUI achieved a maximum classification score of 98.7% for Test-5 and 100% for Test-6 using Subspace Discriminant classification model. This shows that although the COVID-19 virus and *Sarbecovirus* are closer on the basis of genomic similarity (Test-4), they are still distinguishable from known species. Therefore, these results suggest that the COVID-19 virus may represent a genetically distinct species of *Sarbecovirus*. All the COVID-19 virus sequences are visually seen in MoDMap3D generated from Test-5 (see Fig 5.2(b)) as a closely packed cluster and it supports a fact that there is 99% similarity among these sequences [12, 31]. The MoDMap3D generated from the Test-5 (Fig 5.2(b)) visually suggests and the average distances from COVID-19 virus sequences to all other sequences confirm that the COVID-19 virus sequences are most proximal to the *RaTG13* (distance: 0.0203), followed by the *bat-SL-CoVZC45* (0.0418), and *bat-SL-CoVZX21* (0.0428). To confirm this proximity, UPGMA and neighbor-joining phylogenetic trees are computed from the PCC-based pairwise distance matrix of sequences in Test-6, see Fig 5.3 and 5.4. Notably, the UPGMA model assumes that all lineages are evolving at a constant rate (equal evolution rate among branches). This method may produce unreliable results in cases where the genomes of some lineages evolve more rapidly than those of the others. To further verify the phylogenetic relationships, we also produced a phylogenetic tree using the neighbor-joining method that allows different evolution rates among branches and obtained a highly similar output. The phylogenetic trees placed the *RaTG13* sequence closest to the COVID-19 virus sequences, followed by the *bat-SL-CoVZC45* and *bat-SL-CoVZX21* sequences. This closer proximity represents the smaller genetic distances between these sequences and aligns with the visual sequence relationships shown in the MoDMap3D of Fig 5.2(b).

We further confirm our results regarding the closeness of the COVID-19 virus with the sequences from the *Betacoronavirus* genus (especially sub-genus *Sarbecovirus*) by a quantitative analysis based on the Spearman's rank correlation coefficient tests. Spearman's rank correlation coefficient [58, 59, 60, 61] tests were applied to the frequencies of oligonucleotide segments, adjusting for the total number of segments, to measure the degree and statistical significance of correlation between two sets of genomic sequences. Spearman's ρ value provides the degree of correlation between the two groups and their k-mer frequencies. The COVID-19 virus was compared to all genera under the *Coronaviridae* family and the k-mer frequencies showed the strongest correlation to the genus *Betacoronavirus*, and more specifically *Sarbecovirus*. The Spearman's rank tests corroborate that the COVID-19 virus is part of the *Sarbecovirus* sub-genus, as shown by CGR and MLDSP. When analyzing sub-genera, it could be hard to classify at lower k values due to the short oligonucleotide frequencies not capturing enough information to highlight the distinctions. Therefore despite the Spearman's rank correlation coefficient providing results for $k = 1$ to $k = 7$, the higher k-mer lengths provided more accurate results, and $k = 7$ was used.

Attributes of the COVID-19 virus genomic signature are consistent with previously reported mechanisms of innate immunity operating in bats as a host reservoir for coronaviruses. Vertebrate genomes are known to have an under-representation of CG dinucleotides in their genomes, otherwise known as CG suppression [82, 83]. This feature is thought to have been due to the accumulation of spontaneous deamination mutations of methyl-cytosines over time [82]. As viruses are obligate parasites, evolution of viral genomes is intimately tied to the biology of their hosts [84]. As host cells develop strategies such as RNA interference and restriction-modification systems to prevent and limit viral infections, viruses will continue to counteract these strategies [83, 84, 85]. Dinucleotide composition and biases are pervasive across the genome and make up a part of the organism's genomic signature [84]. These host genomes have evolutionary pressures that shape the host genomic signature, such as the pressure to eliminate CG dinucleotides within protein coding genes in humans [83]. Viral genomes have been shown to mimic the same patterns of the hosts, including single-stranded positive-sense RNA viruses, which suggests that many RNA viruses can evolve to mimic the same features of their host's genes and genomic signature [82, 83, 84, 85, 86]. As genomic composition, specifically in mRNA, can be used as a way of discriminating self vs non-self RNA, the viral genomes are likely shaped by the same pressures that influence the host genome [83]. One such pressure on DNA and RNA is the APOBEC family of enzymes, members of which are known to cause G to A mutations [86, 87, 88]. While these enzymes primarily work on DNA, it has been demonstrated that these enzymes can also target RNA viral genomes [87]. The APOBEC enzymes therefore have RNA editing capability and may help contribute to the innate defence system against various RNA viruses [86]. This could therefore have a direct impact on the genomic signature of RNA viruses. Additional mammalian mechanisms for inhibiting viral RNA have been highlighted for retroviruses with the actions of zinc-finger antiviral protein (ZAP) [82]. ZAP targets CG dinucleotide sequences, and in vertebrate host cells with the CG suppression in host genomes, this can serve as a mechanism for the distinction of self vs non-self RNA and inhibitory consequences [82]. Coronaviruses have A/U rich and C/G poor genomes, which over time may have been, in part, a product of cytidine deamination and selection against CG dinucleotides [89, 90, 91]. This is consistent with the fact that bats serve as a reservoir for many coronaviruses and that bats have been observed to have some of the largest and most diverse arrays of APOBEC genes in mammals [67, 68]. The Spearman's rank correlation data and the patterns observed in the CGR images from Fig 5.5, of the coronavirus genomes, including the COVID-19 virus identify patterns such as CG underepresentation, also present in vertebrate and, importantly, bat host genomes.

With human-to-human transmission confirmed and concerns for asymptomatic transmission, there is a strong need for continued intervention to prevent the spread of the virus [32, 33, 62, 63, 64]. Due to the high amino acid similarities between the COVID-19 virus and SARS-CoV main protease essential for viral replication and processing, anticoronaviral drugs targeting this protein and other potential drugs have been identified using virtual docking to the protease for treatment of COVID-19 [29, 43, 44, 92, 93, 94, 95]. The human ACE2 receptor has also been identified as the potential receptor for the COVID-19 virus and represents a potential target for treatment [41, 42].

MLDSP-GUI is an ultra-fast, alignment-free method as is evidenced by the time-performance of MLDSP-GUI for Test-1 to Test-6 given in Fig 5.8. MLDSP-GUI took just 10.55 seconds to compute a pairwise distance matrix (including reading sequences, computing magnitude spectra using DFT, and calculating the distance matrix using PCC combined) for the Test-1 (largest dataset used in this study with 3273 complete genomes). All of the tests combined (Test-1 to Test-6) are doable in under 10 minutes including the computationally heavy 10-fold cross-validation, and testing of the 29 COVID-19 virus sequences.



	Test-1	Test-2	Test-3a	Test-3b	Test-4	Test-5	Test-6
🟦 Distance Matrix Computation (Read sequences, DFT, PCC)	10.55	8.26	1.02	0.55	0.43	0.46	0.29
🟩 Classification (10-fold cross-validation)	250.01	168.63	5.11	1.95	2.04	2.4	1.85
🟨 2019-nCoV Testing	62.77	45.97	5.82	3.06	3.01		
🟥 Total time	323.33	222.86	11.95	5.56	5.48	2.86	2.14

Figure 5.8: Time performance of MLDSP-GUI for Test1 to Test-6 (in seconds).

The results of our machine learning-based alignment-free analyses using MLDSP-GUI support the hypothesis of a bat origin for the COVID-19 virus and classify COVID-19 virus as sub-genus *Sarbecovirus*, within *Betacoronavirus*.

5.5 Conclusion

This study provides an alignment-free method based on intrinsic genomic signatures that can deliver highly-accurate real-time taxonomic predictions of yet unclassified new sequences, *ab initio*, using raw DNA sequence data alone and without the need for gene or genome annotation. We use this method to provide evidence for the taxonomic classification of the COVID-19 virus as *Sarbecovirus*, within *Betacoronavirus*, as well as quantitative evidence supporting a bat origin hypothesis. Our results are obtained through a comprehensive analysis of over 5000 unique viral sequences, through an alignment-free analysis of their two-dimensional genomic signatures, combined with a “decision tree" use of supervised machine learning and confirmed by Spearman’s rank correlation coefficient analyses. This study suggests that such alignment-free approaches to comparative genomics can be used to complement alignment-based approaches when timely taxonomic classification is of the essence, such as at critical periods during novel viral outbreaks.

Bibliography

[1] Enjuanes L, Brian D, Cavanagh D, Holmes K, Lai MMC, Laude H, et al. Coronaviridae. In: Regenmortel MV, Fauquet CM, Bishop DHL, Carstens EB, Estes MK, Lemon SM, et al., editors. Virus Taxonomy. Seventh Report of the International Committee on Taxonomy of Viruses, Academic Press; 2000. pp. 835–849.

[2] Weiss SR, Navas-Martin S. Coronavirus Pathogenesis and the Emerging Pathogen Severe Acute Respiratory Syndrome Coronavirus. Microbiol. Mol. Biol. 2005; Rev. 69: 635–664.

[3] Su S, Wong G, Shi W, Liu J, Lai ACK, Zhou J, et al. Epidemiology, Genetic Recombination, and Pathogenesis of Coronaviruses. Trends in Microbiology. 2016; 24: 490–502.

[4] Cui J, Li F, Shi ZL. Origin and evolution of pathogenic coronaviruses. Nature Reviews Microbiology. 2019; 17: 181–5192.

[5] Schoeman D, Fielding BC. Coronavirus envelope protein: Current knowledge. Virology Journal. 2019; 16.

[6] de Groot RJ, Baker SC, Baric R, Enjuanes L, Gorbalenya AE, Holmes KV, et al. Family Coronaviridae. In: King AMQ, Adams MJ, Carstens EB, Lefkowitz EJ, editors. Virus taxonomy. Ninth report of the international committee on taxonomy of viruses, Elsevier Academic Press; 2012. pp. 806–828.

[7] Woo PCY, Lau SKP, Huang Y, Yuen KY. Coronavirus diversity, phylogeny and interspecies jumping. Experimental Biology and Medicine. 2009; 234: 1117–1127.

[8] Wertheim JO, Chu DKW, Peiris JSM, Kosakovsky Pond SL, Poon LLM. A Case for the Ancient Origin of Coronaviruses. J. Virol. 2013; 87: 7039–7045.

[9] Luk HKH, Li X, Fung J, Lau SKP, Woo PCY. Molecular epidemiology, evolution and phylogeny of SARS coronavirus. Infection, Genetics and Evolution. 2019; 71: 21–30.

[10] Vijaykrishna D, Smith GJD, Zhang JX, Peiris JSM, Chen H, Guan Y. Evolutionary Insights into the Ecology of Coronaviruses. J. Virol. 2007; 81: 4012–4020.

[11] Lau SK, Li KS, Tsang AK, Shek CT, Wang M, Choi GK, et al. Recent Transmission of a Novel Alphacoronavirus, Bat Coronavirus HKU10, from Leschenault's Rousettes to Pomona Leaf-Nosed Bats: First Evidence of Interspecies Transmission of Coronavirus between Bats of Different Suborders. J. Virol. 2012; 86: 11906–11918.

[12] Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet. 2020; doi:10.1016/S0140-6736(20)30251-8.

[13] Li W, Shi Z, Yu M, Ren W, Smith C, Epstein JH, et al. Bats are natural reservoirs of SARS-like coronaviruses. Science. 2005; 310: 676–679.

[14] Duffy S, Shackelton LA, Holmes EC. Rates of evolutionary change in viruses: Patterns and determinants. Nature Reviews Genetics. 2008; 9: 267–276.

[15] Jenkins GM, Rambaut A, Pybus OG, Holmes EC. Rates of molecular evolution in RNA viruses: A quantitative phylogenetic analysis. J. Mol. Evol. 2002; 54: 156–165.

[16] Nagy PD, Simon AE. New insights into the mechanisms of RNA recombination. Virology. 1997; 235: 1–9.

[17] Rowe CL, Fleming JO, Nathan MJ, Sgro JY, Palmenberg AC, Baker SC. Generation of coronavirus spike deletion variants by high-frequency recombination at regions of predicted RNA secondary structure. J. Virol. 1997; 71: 6183–90.

[18] Cavanagh D. Coronaviridae: a review of coronaviruses and toroviruses. In: Schmidt A, Wolff MH, Weber O, editors. Coronaviruses with Special Emphasis on First Insights Concerning SARS. Birkhäuser-Verlag, 2005; pp. 1–54.

[19] Lai MMC. RNA recombination in animal and plant viruses. Microbiological Reviews. 1992; 56: 61–79.

[20] Pasternak AO, Spaan WJM, Snijder EJ. Nidovirus transcription: How to make sense...? Journal of General Virology. 2006; 87: 1403–1421.

[21] Drosten C, Günther S, Preiser W, van der Werf S, Brodt HR, Becker S, et al. Identification of a Novel Coronavirus in Patients with Severe Acute Respiratory Syndrome. N. Engl. J. Med. 2003; 348: 1967–1976.

[22] Ksiazek TG, Erdman D, Goldsmith CS, Zaki SR, Peret T, Emery S, et al. A Novel Coronavirus Associated with Severe Acute Respiratory Syndrome. N. Engl. J. Med. 2003; 348: 1953–1966.

[23] Zaki AM, van Boheemen S, Bestebroer TM, Osterhaus ADME, Fouchier RAM. Isolation of a Novel Coronavirus from a Man with Pneumonia in Saudi Arabia. N. Engl. J. Med. 2012; 367: 1814–1820.

[24] Guan Y, Zheng BJ, He YQ, Liu XL, Zhuang ZX, Cheung CL, et al. Isolation and characterization of viruses related to the SARS coronavirus from animals in Southern China. Science. 2003; 302: 276–278.

[25] Alagaili AN, Briese T, Mishra N, Kapoor V, Sameroff SC, de Wit E, et al. Middle east respiratory syndrome coronavirus infection in dromedary camels in Saudi Arabia. MBio. 2014; 5.

[26] Zhu N, Zhang D, Wang W, Li X, Yang Bo, Song J, et al. A Novel Coronavirus from Patients with Pneumonia in China, 2019. N. Engl. J. Med. 2020; doi:10.1056/NEJMoa2001017.

[27] Lu H, Stratton CW, Tang Y. Outbreak of Pneumonia of Unknown Etiology in Wuhan China: the Mystery and the Miracle. J. Med. Virol. 2020; doi:10.1002/jmv.25678.

[28] Hui DS, I Azhar E, Madani TA, Ntoumi F, Kock R, Dar O, et al. The continuing 2019-nCoV epidemic threat of novel coronaviruses to global health — The latest 2019 novel coronavirus outbreak in Wuhan, China. International Journal of Infectious Diseases. 2020; 91: 264–266.

[29] Liu T, Hu J, Kang M, Lin L, Zhong H, Xiao J, et al. Transmission dynamics of 2019 novel coronavirus (2019-nCoV). BioRxiv [Preprint]. 2020 bioRxiv 919787 [posted 2020 January 25; cited 2020 January 31]. Available from: https://www.biorxiv.org/content/10.1101/2020.01.25.919787v1 doi:10.1101/2020.01.25.919787.

[30] Perlman S. Another Decade, Another Coronavirus. N. Engl. J. Med. 2020; doi:10.1056/NEJMe2001126.

[31] Gralinski LE, Menachery VD. Return of the Coronavirus: 2019-nCoV. Viruses. 2020; 12: 135.

[32] Coronavirus disease 2019 (COVID-19) Situation Report - 39. 2020 February 28 [cited 28 February 2020]. In: WHO website [Internet]. Available from: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200228-sitrep-39-covid-19.pdf

[33] Chan JFW, Yuan S, Kok KH, To KKW, Chu H, Yang J, et al. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. Lancet. 2020; doi:10.1016/S0140-6736(20)30154-9.

[34] Hu B, Zeng LP, Yang XL, Ge XY, Zhang W, Li B, et al. Discovery of a rich gene pool of bat SARS-related coronaviruses provides new insights into the origin of SARS coronavirus. PLoS Pathog. 2017; 13.

[35] Dong N, Yang X, Ye L, Chen K, Chan EWC, Yang M, Chen S. Genomic and protein structure modelling analysis depicts the origin and infectivity of 2019-nCoV, a new coronavirus which caused a pneumonia outbreak in Wuhan, China. BioRxiv [Preprint]. 2020 bioRxiv 913368 [posted 2020 January 22; cited 2020 January 31]. Available from: https://www.biorxiv.org/content/10.1101 /2020.01.20.913368v2 doi:10.1101/2020.01.20.913368.

[36] Guo Q, Li M, Wang C, Wang P, Fang Z, Tan J, et al. Host and infectivity prediction of Wuhan 2019 novel coronavirus using deep learning algorithm. BioRxiv [Preprint]. 2020 bioRxiv 914044 [posted 2020 January 22; cited 2020 January 31]. Available from: https://www.biorxiv.org/content/10.1101 /2020.01.21.914044v2 doi:10.1101/2020.01.21.914044.

[37] Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, et al. A new coronavirus associated with human respiratory disease in China. Nature. 2020; 579: 265–269.

[38] Paraskevis D, Kostaki EG, Magiorkinis G, Panayiotakopoulos G, Sourvinos G, Tsiodras S. Full-genome evolutionary analysis of the novel corona virus (2019-nCoV) rejects the hypothesis of emergence as a result of a recent recombination event. Infection, Genetics and Evolution. 2020; 79: 104212.

[39] Ji W, Wang W, Zhao X, Zai J, Li X. Homologous recombination within the spike glycoprotein of the newly identified coronavirus may boost cross species transmission from snake to human. J. Med. Virol. 2020; doi:10.1002/jmv.25682.

[40] Zhou P, Yang X, Wang X, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. 2020; 579: 270–273.

[41] Letko M, Marzi A, Munster V. Functional assessment of cell entry and receptor usage for SARS-CoV-2 and other lineage B betacoronaviruses. Nat. Microbiol. 2020; 5: 562–569.

[42] Zhao Y, Zhao Z, Wang Y, Zhou Y, Ma Y, Zuo W. Single-cell RNA expression profiling of ACE2, the putative receptor of Wuhan 2019-nCov. BioRxiv [Preprint]. 2020 bioRxiv 919985 [posted 2020 January 26; cited 2020 January 31]. Available from: https://www.biorxiv.org/content/10.1101/2020.01.26.919985v1 doi:10.1101/2020.01.26.919985.

[43] Li Y, Zhang J, Wang N, Li H, Shi Y, Gui G, et al. Therapeutic Drugs Targeting 2019-nCoV Main Protease by High-Throughput Screening. BioRxiv [Preprint]. 2020 bioRxiv 922922 [posted 2020 January 30; cited 2020 January 31]. Available from: https://www.biorxiv.org/content/10.1101/2020.01.28.922922v2 doi:10.1101/2020.01.28.922922.

[44] Liu X, Wang XJ. Potential inhibitors against 2019-nCoV coronavirus M protease from clinically approved medicines. Journal of Genetics and Genomics. 2020; 47(2): 119–121.

[45] Vinga S, Almeida J. Alignment-free sequence comparison–a review. Bioinformatics. 2003; 19(4): 513–523.

[46] Zielezinski A, Vinga S, Almeida J, Karlowski WM. Alignment-free sequence comparison: benefits, applications, and tools. Genome Biology. 2017, 18: 186.

[47] Kari L, Hill KA, Sayem AS, Karamichalis R, Bryans N, Davis K, Dattani NS. Mapping the space of genomic signatures. PLoS ONE. 2015; 10: e0119815.

[48] Karamichalis R, Kari L, Konstantinidis S, Kopecki S. An investigation into inter- and intragenomic variations of graphic genomic signatures. BMC Bioinformatics. 2015; 16: 246.

[49] Solis-Reyes S, Avino M, Poon A. An open-source k-mer based machine learning tool for fast and accurate subtyping of HIV-1 genomes. PLoS ONE. 2018; 13: e0206409.

[50] Randhawa GS, Hill KH, Kari L. ML-DSP: Machine Learning with Digital Signal Processing for ultrafast, accurate, and scalable genome classification at all taxonomic levels. BMC Genomics. 2019; 20: 267.

[51] Randhawa GS, Hill KH, Kari L. MLDSP-GUI: an alignment-free standalone tool with an interactive graphical user interface for DNA sequence comparison and analysis. Bioinformatics. 2019; btz918.

[52] Jeffrey HJ. Chaos game representation of gene structure. Nucleic Acids Res. 1990; 18: 2163–2170.

[53] Asuero AG, Sayago A, González AG. The correlation coefficient: an overview. Crit Rev Anal Chem. 2006; 36(1): 41–59.

[54] Karamichalis R, Kari L. MoDMaps3D: an interactive webtool for the quantification and 3D visualization of interrelationships in a dataset of DNA sequences. Bioinformatics. 2017; 33(19): 3091–3.

[55] Kruskal J. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika. 1964; 29: 1–27.

[56] Sokal RR, Michener CD. A statistical method for evaluating systematic relationships. University of Kansas Science Bulletin. 1958; 38: 1409–1438.

[57] Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution. 1987; 4(4): 406–425.

[58] Carneiro RL, Requião RD, Rossetto S, Domitrovic T, Palhano FL. Codon stabilization coefficient as a metric to gain insights into mRNA stability and codon bias and their relationships with translation. Nucleic acids research. 2019; 47(5): 2216–2228.

[59] Karumathil S, Raveendran NT, Ganesh D, Kumar NS, Nair RR, Dirisala VR. Evolution of Synonymous Codon Usage Bias in West African and Central African Strains of Monkeypox Virus. Evolutionary Bioinformatics Online. 2018; 14: doi:10.1177/1176934318761368.

[60] Vinogradov AE, Anatskaya OV. DNA helix: the importance of being AT-rich. Mammalian Genome. 2017; 9(10): 455–464.

[61] Hollander M, Wolfe DA, Chicken E. Nonparametric statistical methods, 3rd Edition, John Wiley & Sons; 2013.

[62] Zhao S, Lin Q, Ran J, Musa SS, Yang G, Wang W, et al. Preliminary estimation of the basic reproduction number of novel coronavirus (2019-nCoV) in China, from 2019 to 2020: A data-driven analysis in the early phase of the outbreak. International Journal of Infectious Diseases. 2020; 92: 214–217.

[63] Shao P, Shan Y. Beware of asymptomatic transmission: Study on 2019-nCoV prevention and control measures based on extended SEIR model. BioRxiv [Preprint]. 2020 bioRxiv 923169 [posted 2020 January 28; cited 2020 January 31]. Available from: https://www.biorxiv .org/content/10.1101/2020.01.28.923169v1 doi:10.1101/2020.01.28.923169.

[64] Chen Z, Zhang W, Lu Y, Guo C, Guo Z, Liao C, et al. From SARS-CoV to Wuhan 2019-nCoV Outbreak: Similarity of Early Epidemic and Prediction of Future Trends. BioRxiv [Preprint]. 2020 bioRxiv 919241 [posted 2020 January 27; cited 2020 January 31]. Available from: https://www.biorxiv.org/content /10.1101/2020.01.24.919241v3 doi:10.1101/2020.01.24.919241.

[65] Hayward JA, Tachedjian M, Cui J, Field H, Holmes EC, Wang L, Tachedjian G. Identification of diverse full-length endogenous betaretroviruses in megabats and microbats. Retrovirology. 2013; 10.

[66] Cui J, Tachedjian G, Wang LF. Bats and Rodents Shape Mammalian Retroviral Phylogeny. Sci. Rep. 2015; 5.

[67] Hayward JA, Tachedjian M, Cui J, Cheng AZ, Johnson A, Baker ML, et al. Differential evolution of antiretroviral restriction factors in pteropid bats as revealed by APOBEC3 gene complexity. Mol. Biol. Evol. 2018; 35: 1626–1637.

[68] Wong A, Li X, Lau S, Woo P. Global Epidemiology of Bat Coronaviruses. Viruses. 2019; 11(2): 174.

[69] Yang XL, Hu B, Wang B, Wang MN, Zhang Q, Zhang W, et al. Isolation and Characterization of a Novel Bat Coronavirus Closely Related to the Direct Progenitor of Severe Acute Respiratory Syndrome Coronavirus. J. Virol. 2016; 90: 3253–3256.

[70] Lau SK, Li KS, Tsang AK, Lam CS, Ahmed S, Chen H, et al. Genetic Characterization of Betacoronavirus Lineage C Viruses in Bats Reveals Marked Sequence Divergence in the Spike Protein of Pipistrellus Bat Coronavirus HKU5 in Japanese Pipistrelle: Implications for the Origin of the Novel Middle East Respiratory Syndrome Coronavirus. J. Virol. 2013; 87: 8638–8650.

[71] Lacroix A, Duong V, Hul V, San S, Davun H, Omaliss K, et al. Genetic diversity of coronaviruses in bats in Lao PDR and Cambodia. Infect. Genet. Evol. 2017; 48: 10–18.

[72] Drexler JF, Gloza-Rausch F, Glende J, Corman VM, Muth D, Goettsche M, et al. Genomic Characterization of Severe Acute Respiratory Syndrome-Related Coronavirus in European Bats and Classification of Coronaviruses Based on Partial RNA-Dependent RNA Polymerase Gene Sequences. J. Virol. 2010; 84: 11336–11349.

[73] Rihtarič D, Hostnik P, Steyer A, Grom J, Toplak I. Identification of SARS-like coronaviruses in horseshoe bats (Rhinolophus hipposideros) in Slovenia. Arch. Virol. 2010; 155: 507–514.

[74] He B, Zhang Y, Xu L, Yang W, Yang F, Yun Feng, et al. Identification of Diverse Alphacoronaviruses and Genomic Characterization of a Novel Severe Acute Respiratory Syndrome-Like Coronavirus from Bats in China. J. Virol. 2014; 88: 7070–7082.

[75] Wacharapluesadee S, Duengkae P, Rodpan A, Kaewpom T, Maneeorn P, Kanchanasaka B, et al. Diversity of coronavirus in bats from Eastern Thailand Emerging viruses. Virol. J. 2015; 12: 1–7.

[76] Tong S, Conrardy C, Ruone S, Kuzmin IV, Guo X, Tao Y, et al. Detection of novel SARS-like and other coronaviruses in bats from Kenya. Emerg. Infect. Dis. 2009; 15: 482–485.

[77] Lau SKP, Woo PCY, Li KSM, Huang Y, Tsoi H, Wong BHL, et al. Severe acute respiratory syndrome coronavirus-like virus in Chinese horseshoe bats. Proc. Natl. Acad. Sci. 2005; National Academy of Sciences, U. S. A., 102: 14040–14045.

[78] Virologists weigh in on novel coronavirus in China's outbreak. 2020 January 08 [cited 31 January 2020]. In: University of Minnesota [Internet]. Available from: http://www.cidrap.umn.edu/news-perspective/2020/01/virologists-weigh-novel-coronavirus-chinas-outbreak.

[79] nCoV's relationship to bat coronaviruses & recombination signals (no snakes) - no evidence the 2019-nCoV lineage is recombinant. 2020 January 31 [cited 31 January 2020]. In: Virological blog [Internet]. Available from: http://virological .org/t/ncovs-relationship-to-bat-coronaviruses-recombination-signals-no-snakes-no-evidence-the-2019-ncov-lineage-is-recombinant/331.

[80] Experts: nCoV spread in China's cities could trigger global epidemic. 2020 January 27 [cited 31 January 2020]. In: University of Minnesota [Internet]. Available from: http://www.cidrap.umn.edu/news-perspective/2020/01/experts -ncov-spread-chinas-cities-could-trigger-global-epidemic.

[81] China detects large quantity of novel coronavirus at Wuhan seafood market. 2020 January 27 [cited 31 January 2020]. In: Xinhuanet News [Internet]. Available from: http://www.xinhuanet.com/english/2020-01/27/c_138735677.htm.

[82] Takata MA, Gonçalves-Carneiro D, Zang TM, Soll SJ, York A, Blanco-Melo D, Bieniasz PD. CG dinucleotide suppression enables antiviral defence targeting non-self RNA. Nature. 2017; 550(7674): 124–127.

[83] Greenbaum BD, Levine AJ, Bhanot G, Rabadan R. Patterns of evolution and host gene mimicry in influenza and other RNA viruses. PLoS Pathogens. 2008; 4(6): doi:10.1371/journal.ppat.1000079.

[84] Lobo FP, Mota BEF, Pena SDJ, Azevedo V, Macedo AM, Tauch A, et al. Virus-host coevolution: Common patterns of nucleotide motif usage in Flaviviridae and their hosts. PLoS ONE. 2009; 4(7): 10.1371/journal.pone.0006282.

[85] Kindler E, Thiel V. To sense or not to sense viral RNA-essentials of coronavirus innate immune evasion. Current Opinion in Microbiology. 2014; 20: 68–75.

[86] Milewska A, Kindler E, Vkovski P, Zeglen S, Ochman M, Thiel V, et al. APOBEC3-mediated restriction of RNA virus replication. Scientific Reports. 2018; 8(1): doi:10.1038/s41598-018-24448-2.

[87] Bishop KN, Holmes RK, Sheehy AM, Malim MH. APOBEC-mediated editing of viral RNA. Science. 2004; 305(5684): 645.

[88] Pyrc K, Jebbink MF, Berkhout B, Van der Hoek L. Genome structure and transcriptional regulation of human coronavirus NL63. Virology Journal. 2004; 1(1): 7.

[89] Berkhout B, Van Hemert F. On the biased nucleotide composition of the human coronavirus RNA genome. Virus Research. 2015; 202: 41–47.

[90] Woo PCY, Lau SKP, Huang Y, Yuen KY. Coronavirus diversity, phylogeny and interspecies jumping. Experimental Biology and Medicine. 2009; 234(10): 1117–1127.

[91] Woo PCY, Huang Y, Lau SKP, Yuen KY. Coronavirus Genomics and Bioinformatics Analysis. Viruses. 2010; 2(8): 1804–1820.

[92] Xue X, Yu H, Yang H, Xue F, Wu Z, Shen W, et al. Structures of Two Coronavirus Main Proteases: Implications for Substrate Binding and Antiviral Drug Design. J. Virol. 2008; 82: 2515–2527.

[93] Anand K, Ziebuhr J, Wadhwani P, Mesters JR, Hilgenfeld R. Coronavirus main proteinase (3CLpro) Structure: Basis for design of anti-SARS drugs. Science. 2003; 300: 1763–1767.

[94] Nukoolkarn V, Lee VS, Malaisree M, Aruksakulwong O, Hannongbua S. Molecular dynamic simulations analysis of ritronavir and lopinavir as SARS-CoV 3CLpro inhibitors. J. Theor. Biol. 2008; 254: 861–867.

[95] Xu Z, Peng C, Shi Y, Zhu Z, Mu K, Wang X, Zhu W. Nelfinavir was predicted to be a potential inhibitor of 2019-nCov main protease by an integrative approach combining homology modelling, molecular docking and binding free energy calculation. BioRxiv [Preprint]. 2020 bioRxiv 921627 [posted 2020 January 28; cited 2020 January 31]. Available from: https://www.biorxiv.org/content /10.1101/2020.01.27.921627v1 doi: 10.1101/2020.01.27.921627.

Chapter 6

Conclusion

In this thesis, we show that certain genomic signatures can be used to measure the quantitative dissimilarity between any two genomic sequences. In Chapter 3, we show that, irrespective of the one-dimensional numerical representation used, the intergenomic dissimilarity is strong enough to classify sequences at different taxonomic levels. The successful classification of virus subtypes shows that the concept of genomic signature holds even for sequences at the subspecies level where they become very similar. These genomic signatures, coupled with supervised machine learning lead to highly accurate classification. In Chapter 4, we show that the two-dimensional Chaos Game Representation (CGR) can also be used as a genomic signatures, for superior classification results. In addition, the software tool that we developed is open-source, ultra-fast, scalable, stand-alone with a user-friendly graphical user interface, and it provides an assurance to the users that their private data is safe and secure. In Chapter 5, we show the importance of the proposed methodology when a timely analysis of novel unclassified sequences is required. We demonstrate the use of a "decision tree" approach (paralleling taxonomic ranks) to supervised machine learning, for successive refinements of taxonomic classification. This study provides a proof of concept that alignment-free methods can deliver highly-accurate real-time taxonomic predictions of yet unclassified new sequences, *ab initio*, using raw DNA sequence data alone, and without the need for gene or genome annotation.

In this thesis, we used six different classifiers covering a wide gamut from low to high time complexity. In the near future, multi-factor selection criteria can be established to include an even wider variety of classifiers such as Random Forest, Decision tree, etc. Also, there is a possibility to include more dissimilarity measures, especially the ones which are considered a natural choice for spectral analysis in the field of signal processing wherein, e.g. coherence is widely used to examine the relation between two signals.

Future directions of the research should explore factors beyond genetic relatedness and include potential impacts of environmental influences upon genomic signatures. It is presumed that genomic signatures are a product of the complex interactions of genetic relatedness and environment.

An important future direction is the examination of genomic signature diversity with genomic instability and disease phenotypes. Applications of this type of classification may exist for cancer phenotypes and a wide spectrum of inherited diseases. Whereas entire genome sequences have been the focus of this classification approach, it is well worth testing capabilities at lower genome sequence resolution. There is a wealth of human single nucleotide genotyping data publicly available and with extensive phenotypic data. Genotypes for sequence classification may prove useful and efficient materials for discovery of as yet unrecognized genetic variants associated with classes of phenotypic information like cancer type and inherited disease type.

Appendix A

Copyright Releases

Chapter 3 contains the article "ML-DSP: Machine Learning with Digital Signal Processing for ultrafast, accurate, and scalable genome classification at all taxonomic levels" from BMC Genomics. According to their website, https://bmcgenomics.biomedcentral.com/submission-guidelines/copyright

"Copyright on any open access article in a journal published by BioMed Central is retained by the author(s). Authors grant BioMed Central a license to publish the article and identify itself as the original publisher. Authors also grant any third party the right to use the article freely as long as its integrity is maintained and its original authors, citation details and publisher are identified. The Creative Commons Attribution License 4.0 formalizes these and other terms and conditions of publishing articles."

Chapter 4 contains the article "MLDSP-GUI: an alignment-free standalone tool with an interactive graphical user interface for DNA sequence comparison and analysis" from Bioinformatics. According to their website, https://academic.oup.com/journals/pages/open_access/funder_policies/chorus/standard_publication_model

"Journals content which is published and distributed under the Oxford University Press (OUP) Journals, Standard Publication Model is published on an exclusive licence or copyright assignment basis, and reuse rights are retained and controlled by OUP. Individuals accessing

OUP journals content may view, reproduce or store copies of such content provided that it is used only for personal, non-commercial use. Any use of such content in whole or in part must include the bibliographic citation, including author attribution, date, article title, journal title, and URL and MUST include a copy of the copyright notice."

Chapter 5 contains the article "Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: COVID-19 case study" from PLoS ONE. According to their website http://journals.plos.org/plosone/s/content-license

"PLOS applies the Creative Commons Attribution (CC BY) license to works we publish. This license was developed to facilitate open access namely, free immediate access to, and unrestricted reuse of, original works of all types. Under this license, *authors agree to make articles legally available for reuse, without permission or fees, for virtually any purpose*. Anyone may copy, distribute or reuse these articles, as long as the author and original source are properly cited."

"Using PLOS Content: *No permission is required from the authors or the publishers to reuse or repurpose PLOS content provided the original article is cited.* In most cases, appropriate attribution can be provided by simply citing the original article. If the item you plan to reuse is not part of a published article (e.g., a featured issue image), then indicate the originator of the work, and the volume, issue, and date of the journal in which the item appeared. For any reuse or redistribution of a work, you must also make clear the license terms under which the work was published."

Appendix B

Software description

ML-DSP (Machine Learning with Digital Signal Processing) implements a four-step pipeline for the analysis of the genomic sequences comprising: (i) One-dimensional numerical representations of DNA sequences to map genomic sequences to the genomic signals, (ii) Discrete Fourier Transform (DFT) to get magnitude spectra from the genomic signals, (iii) pair-wise distance calculation using Pearson Correlation Coefficient (PCC) between the magnitude spectra of any two genomic signals, and (iv) supervised machine learning classification to obtain quantitative classification accuracies scores and to predict the labels of new sequences. The algorithms for ML-DSP were implemented using the software package MATLAB $R2017a$, license no. 964054. MATLAB license is required to run and use ML-DSP. The source code used for the results in this thesis was optimized by implementing most of the components designed to run in parallel. The source code (1015 lines of MATLAB code) of ML-DSP is available at the following link: https://github.com/grandhawa/MLDSP

We further refined the ML-DSP code by adding more flexibility to accept sequences contained in the '.fna', and '.txt' files, in addition to the default '.fasta' files. Our code cleans the sequences by removing all the unrecognized characters and keeping only the occurrences of A ,C, G, and T. Our preprocessing code omits the shorter sequences based on the minimum sequence length parameter entered as input by the user. The user can also alter the value of the maximum sequence length parameter, which is used to select the random fragments of the selected length for the longer sequences. The default value ‘0’ selects the complete sequences with their original length. For the longer sequences, the user can also select the option to select multiple non-overlapping fragments per sequence. Another valuable feature is the balancing of the clusters, where the user can select any value for the maximum cluster size parameter. This parameter puts an upper limit on the size of clusters, and the clusters smaller than the selected value remain unchanged. For the clusters larger than the selected value, random sequences (number of sequences are equal to the selected value) are chosen to capture the diversity of the whole dataset. The default value ‘0’ selects the balanced clusters, where every cluster has an equal number of sequences limited by the size of the smallest cluster. The improved script also provides an option to test multiple sequences at once using already trained classification models.

The software package also includes a script to download a customized dataset from the National Center for Biotechnology Information (NCBI) database. The default bulk download feature of NCBI demands a list of accession numbers and has restrictions on the number of sequences, number of download requests per minute, download data limit, etc. Moreover, the only way to download contigs of Whole Genome Shotgun (WGS) sequences is the manual browsing and saving individual contigs one at a time. This cumbersome manual approach may take hours just to download a few hundred sequences. Our script reads a list of accession numbers, identifies if any accession number is of a WGS sequence, parses the NCBI webpages for the available contigs of the WGS sequences, and downloads the sequences in parallel. Our script handles the NCBI restriction on the number of download requests by introducing a delay and re-requesting a sequence if the NCBI server produces an error. The exception handling block keeps on adjusting the delay parameter and generating the download requests until the download succeeds. The script also cleans the sequences by keeping only the occurrences of A, C, G, and T before writing the downloaded data to the ‘.fasta’ files.

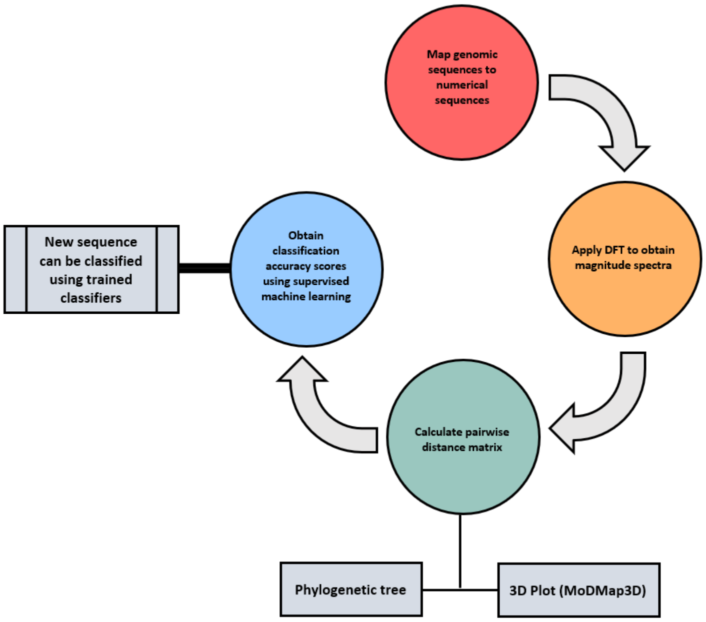
MLDSP-GUI is an extension of ML-DSP that comes with multiple valuable additions of (i) user-friendly interactive Graphical User Interface (GUI), (ii) two-dimensional Chaos Game Representation (CGR) to numerically represent DNA sequences, (iii) Euclidean and Manhattan distances as additional distance measures, (iv) Phylogenetic tree output in Newick-formatted file, (v) oligomer (sub-word) frequency information to study the under-and-over representation of any particular sub-sequence in a selected sequence, (vi) Inter-cluster distances analysis. MLDSP-GUI gives users an option to export and save to the disk the customized results such as distance matrix, inter-cluster distances, oligomer frequencies, 3D molecular distance map, CGR plots of all sequences, etc. MLDSP-GUI is implemented using MATLAB $R2019a$ App Designer, license no. 964054. A single executable platform-independent file is provided that can be used to install and run the software tool. Though MLDSP-GUI is a MATLAB application, MATLAB license is not required to run and use this tool. The MATLAB source code is 2296 lines of code long. MLDSP-GUI is an open-source tool with Graphical User Interface that is publicly available for download at the following link: https://sourceforge.net/projects/mldsp-gui/

Appendix C

MLDSP-GUI: Supplementary Material

C.A Interactive MLDSP-GUI features

MLDSP-GUI implements a four-step pipeline that takes as input a set of genomic DNA sequences and outputs their taxonomic classification. It consists of: *(i)* computing numerical representation of DNA sequences, *(ii)* applying Discrete Fourier Transform (DFT), *(iii)* calculating pairwise distances (Pearson Correlation Coefficient PCC, Euclidean, or Manhattan), and *(iv)* classifying using supervised machine learning, see Figure C.S1. Independently, multidimensional scaling uses the pairwise distance matrix to display an interactive 3D molecular distance map. The user also has the option to generate a phylogenetic tree from the pairwise distance matrix. A new sequence can be classified using the trained classifiers.


Map genomic sequences to numerical sequences
Apply DFT to obtain magnitude spectra
Calculate pairwise distance matrix
Obtain classification accuracy scores using supervised machine learning
New sequence can be classified using trained classifiers
Phylogenetic tree
3D Plot (MoDMap3D)

Supplementary Figure C.S1: MLDSP-GUI implements a four-step pipeline for data transformation from genomic sequences to taxonomic classification.

The methods used in the MLDSP-GUI pipeline are discussed below:

(i) *Numerical representations*: Genomic sequences are mapped into discrete numerical representations. Users can pick one of the 14 available numerical representations. MLDSP-GUI implements 13 one-dimensional numerical representations (Integer, Integer-other variant, Real, Atomic, EIIP-electron-ion interaction potential, PP – purine/pyrimidine, Paired numeric, Nearest-neighbor based doublet, Codon, Just-A, Just-C, Just-G, Just-T), see [4] and 1 two-dimensional representation (Chaos Game Representation - CGR), see [2]. One-dimensional representations replace every ‘A, C, G, T’ in a genomic sequence with a specific numeric value (depending on the choice of the representation) to compute a one-dimensional discrete numerical vector. CGR computes a *k*-mer (subword of length *k*) dependent two-dimensional plot for each genomic sequence by using the method described in [2]. These discrete numerical sequences computed from the genomic sequences can be treated as digital signals and have been called in the literature “genomic signals", see [1]. The whole process of applying the Digital Signal Processing (DSP) techniques to genomic data (numerical sequences in our case) has been termed Genomic Signal Processing (GSP), see [3]. For any genomic sequence S_i of length p, its corresponding one-dimensional numerical sequence (genomic signal) N_i will be of length p. If CGR is selected as a numerical representation, then a two-dimensional plot of the size $2^k \times 2^k$ will be generated for a selected k-mer value k, that is the length of the corresponding genomic signal N_i will be $2^k \times 2^k$.

(ii) *Discrete Fourier Transform (DFT)*:

Discrete Fourier Transform (DFT) is applied to the genomic signals (discrete numerical representations of the genomic sequences) to compute the magnitude spectra. Suppose we have a dataset of n sequences. Then, the DFT of an i^{th} $(0 \leq i \leq n-1)$ genomic signal $N_i = N_i(0), N_i(1), ..., N_i(p-1)$ results in another sequence of complex numbers, $F_i(k) = F_i(0), F_i(1), ..., F_i(p-1)$ where, for $0 \leq k \leq p-1$ we have:

$$F_i(k) = \sum_{j=0}^{p-1} N_i(j) \cdot e^{(-\imath 2\pi/p)kj} \tag{C.1}$$

The magnitude spectrum of a genomic signal N_i is the absolute value of the vector F_i.

(iii) *Pairwise distance calculation*:

MLDSP-GUI implements three distance measures: the Pearson Correlation Coefficient (PCC), Euclidean distance and Manhattan distance.

The Pearson Correlation Coefficient r_{XY} between two magnitude spectra X and Y, each of length p, is computed as:

$$r_{XY} = \frac{\sum_{i=0}^{p-1}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=0}^{p-1}(X_i - \overline{X})^2} \times \sqrt{\sum_{i=0}^{p-1}(Y_i - \overline{Y})^2}} \tag{C.2}$$

where the average $\overline{X}$ is defined as $(\sum_{i=0}^{p-1} X_i)/p$ and similarly for Y. The results are normalized by taking $(1 - r_{XY})/2$, to obtain distance values between 0 and 1.

The Euclidean distance d_{EUC} between two magnitude spectra X and Y, each of length p, is computed as:

$$d_{EUC} = \sqrt{\sum_{i=0}^{p-1}(X_i - Y_i)} \tag{C.3}$$

The Manhattan distance d_{MAN} between two magnitude spectra X and Y, each of length p, is computed as:

$$d_{MAN} = \sum_{i=0}^{p-1} |X_i - Y_i| \tag{C.4}$$

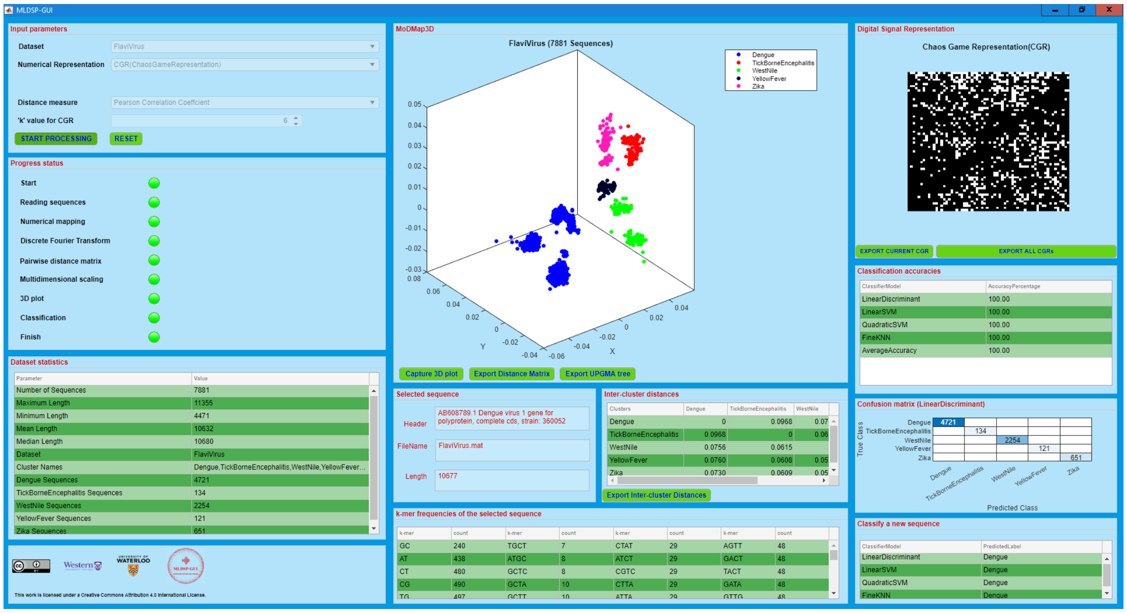
(iv) *Supervised Machine Learning classification*:

Supervised machine learning algorithms train the classification models using given input-output pairs consisting of the feature vector corresponding to a genomic sequence as the input, and the label of the sequence (taxon) as the output. In our case, the feature vector for any given sequence consists of the pairwise distances between (a) the magnitude spectrum obtained from the given sequence, and (b) the magnitude spectra obtained from all the other genomic sequences in the training set. The trained classification models can then be used to predict the labels of testing sequences. MLDSP-GUI implements the 10-fold cross-validation technique, and gives the choice of six classifiers (Linear Discriminant, Linear SVM, Quadratic SVM, Fine KNN, Subspace Discriminant, and Subspace KNN) for performing the task of supervised machine learning. Subspace Discriminant and Subspace KNN are omitted if the dataset contains more than 2000 sequences, because they are computationally heavy. 10-fold cross-validation consists in dividing the dataset randomly into

10 equal sets (9 used for training and 1 used for testing). The feature vectors of the sequences in the training set are constructed using columns of the pairwise-distance matrix, as follows: All columns and rows which correspond to the sequences in the testing set are omitted, and the remaining columns are used as feature vectors for the training. The trained models are then used to predict the labels of the sequences from the testing test. The whole process is repeated 10 times, and the average classification accuracy scores (prediction accuracy) are reported as the output. MLDSP-GUI also has the option whereby a novel input sequence can be tested (its label can be predicted) using the trained classifiers.

MLDSP-GUI displays results as three vertical panels, each panel subdivided into multiple sub-panel components. Figure S2 shows a test run of MLDSP-GUI on the *Flavivirus* dataset. The 7,881 complete genomes of the *Flavivirus* genus (average length 10,632 bp - the right panel shows the CGR representation of one of the *Dengue* virus genomes) are clustered into the virus species of *Dengue* (blue, 4,721 sequences), *Tick-Borne Encephalitis* (red, 134 sequences), *West Nile* (green, 2,254 sequences), *Yellow Fever* (black, 121 sequences), and *Zika* (magenta, 651 sequences). The classification accuracy using any of the four classifiers (Linear Discriminant, Linear SVM, Quadratic SVM, or Fine KNN) is 100%. MLDSP-GUI is also able to suggest classification of some virus species into subtypes, e.g., the four blue clusters correspond to the *Dengue* virus subtypes *Dengue*-1, *Dengue*-2, *Dengue*-3, and *Dengue*-4.

The next subsections of this Supplementary Material discuss the three panels (Left panel, Center panel, and Right panel) and their components in detail.



Input parameters

Dataset FlaviVirus
Numerical Representation CGR(ChaosGameRepresentation)
Distance measure Pearson Correlation Coeffcient
'k' value for CGR 6

Progress status

- Start ●
- Reading sequences ●
- Numerical mapping ●
- Discrete Fourier Transform ●
- Pairwise distance matrix ●
- Multidimensional scaling ●
- 3D plot ●
- Classification ●
- Finish ●

Dataset statistics

Parameter	Value
Number of Sequences	7881
Maximum Length	11355
Minimum Length	4471
Mean Length	10632
Median Length	10680
Dataset	FlaviVirus
Cluster Names	Dengue,TickBorneEncephalitis,WestNile,YellowFever...
Dengue Sequences	4721
TickBorneEncephalitis Sequences	134
WestNile Sequences	2254
YellowFever Sequences	121
Zika Sequences	651

This work is licensed under a Creative Commons Attribution 4.0 International License.

Selected sequence

Header AB608789.1 Dengue virus 1 gene for polyprotein, complete cds, strain: 360052
FileName FlaviVirus.mat
Length 10677

Inter-cluster distances

Clusters	Dengue	TickBorneEncephalitis	WestNile
Dengue	0	0.0968	0.07
TickBorneEncephalitis	0.0968	0	0.06
WestNile	0.0756	0.0615	
YellowFever	0.0760	0.0608	0.05
Zika	0.0730	0.0609	0.05

k-mer frequencies of the selected sequence

k-mer	count	k-mer	count	k-mer	count	k-mer	count
GC	240	TGCT	7	CTAT	29	AGTT	48
AT	438	ATGC	8	ATCT	29	GACT	48
CT	480	GCTC	8	CGTC	29	TACT	48
CG	490	GCTA	10	CTTA	29	GATA	48
TG	497	GCTT	10	ATTA	29	GTTG	48

Classification accuracies

ClassifierModel	AccuracyPercentage
LinearDiscriminant	100.00
LinearSVM	100.00
QuadraticSVM	100.00
FineKNN	100.00
AverageAccuracy	100.00

Confusion matrix (LinearDiscriminant)

True Class \ Predicted Class	Dengue	TickBorneEncephalitis	WestNile	YellowFever	Zika
Dengue	4721				
TickBorneEncephalitis		134			
WestNile			2254		
YellowFever				121	
Zika					651

Classify a new sequence

ClassifierModel	PredictedLabel
LinearDiscriminant	Dengue
LinearSVM	Dengue
QuadraticSVM	Dengue
FineKNN	Dengue

Supplementary Figure C.S2: MLDSP-GUI can be viewed as a combination of 3-vertical panels (Left panel, Center panel, and Right panel). Each panel has multiple sub-panel components.

All experiments were performed on an ASUS ROG G752VS computer with 4 cores (8 threads) of a 2.7GHz Intel Core i7 6820HK processor and 64GB DD4 2400MHz SDRAM.

C.A.1 Left panel

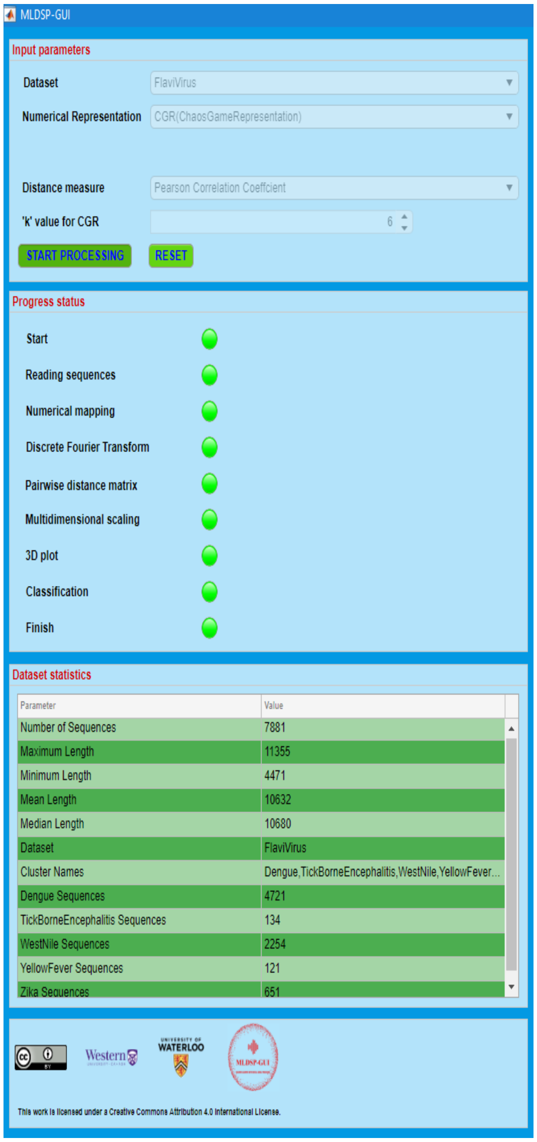
The left panel components are shown in Figure C.S3.

1. *Input parameters*:

   The user can select a dataset among one of the provided datasets, or “browse” to select a user-defined dataset. Some additional datasets are also provided, see Table C.S1.

   The user has the option to select one of the 13 one-dimensional numerical representations of DNA sequences (Integer, Integer-other variant, Real, Atomic, EIIP, purine/pyrimidine, Nearest neighbor based doublet, Codon, Just-A, Just-C, Just-G, Just-T) or the two-dimensional Chaos Game Representation (CGR).

   For example, the one-dimensional numerical representation “purine/pyrimidine" assigns A/G the value -1, and C/T the value +1, whereby the DNA sequence ACGTTAGC is represented as the numerical sequence [-1 1 -1 1 1 -1 -1 1]. If the user selects any of the one-dimensional representations, then a value for the length normalization parameter (maximum, minimum, mean or median) can be selected. The default is the length normalization using the median length.



Dataset: FlaviVirus
Numerical Representation: CGR(ChaosGameRepresentation)
Distance measure: Pearson Correlation Coeffcient
'k' value for CGR: 6

Progress status	
Start	●
Reading sequences	●
Numerical mapping	●
Discrete Fourier Transform	●
Pairwise distance matrix	●
Multidimensional scaling	●
3D plot	●
Classification	●
Finish	●

Parameter	Value
Number of Sequences	7881
Maximum Length	11355
Minimum Length	4471
Mean Length	10632
Median Length	10680
Dataset	FlaviVirus
Cluster Names	Dengue,TickBorneEncephalitis,WestNile,YellowFever...
Dengue Sequences	4721
TickBorneEncephalitis Sequences	134
WestNile Sequences	2254
YellowFever Sequences	121
Zika Sequences	651

This work is licensed under a Creative Commons Attribution 4.0 International License.

Supplementary Figure C.S3: Left panel components: Input parameters, progress status, dataset statistics, and logos.

   Alternatively, given a fixed value of the parameter k, the two-dimensional CGR representation of a DNA sequence simultaneously represents its k-mer frequencies as a two-dimensional plot (see Figure C.S4 for examples; for details on how to generate the CGR of a DNA sequence see Jeffrey H.J., 1990 *Nucleic Acids Res., 18*, 2163 – 2170). If the user selects CGR, then a k-value (k is the length of k-mers to be considered when constructing the CGR) can be selected. The default value is $k = 9$ (the computations for this value could be somewhat slower), and the recommended value for a larger dataset (more than two thousand sequences) is $k = 6$.

The user can also select a distance measure: Pearson Correlation Coefficient (PCC, the default distance), Euclidean distance, or Manhattan distance.

After selecting the input parameters, the user can click on the Start Processing button to start the computation.

A RESET button to reset all parameters to default is also available.

2. *Progress status*:

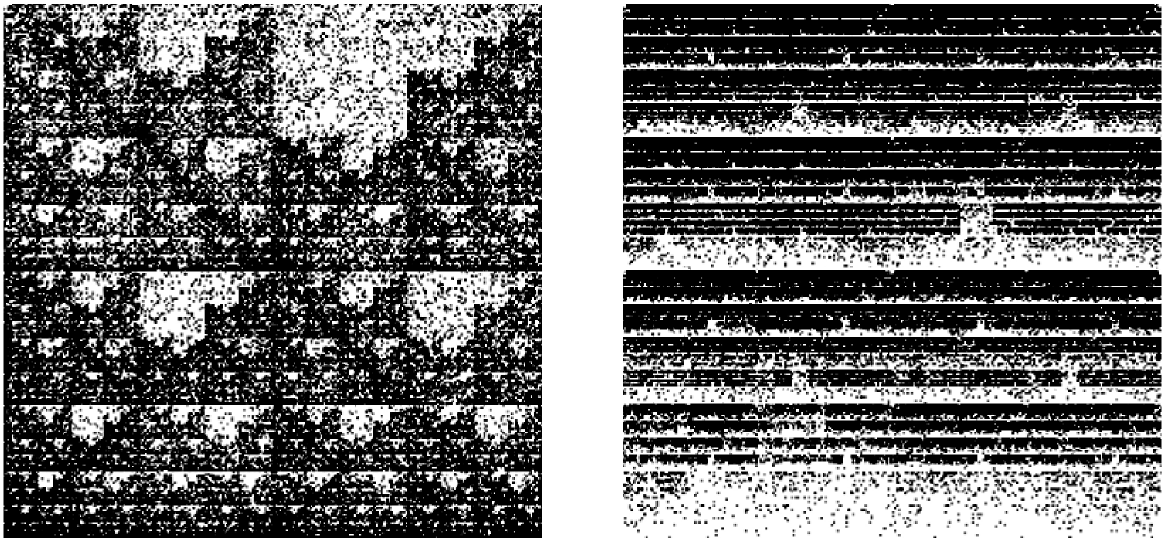
   This sub-panel dynamically lists all the processing steps of a MLDSP-GUI computation. Each step has a colored lamp to highlight their respective status: Red means not started, yellow means in process, and green means completed.

3. *Dataset statistics*:

   This sub-panel shows some statistics of the selected dataset: number of sequences, length statistics (maximum length, minimum length, mean length, and median length), the selected dataset name, cluster names, and the size of clusters.

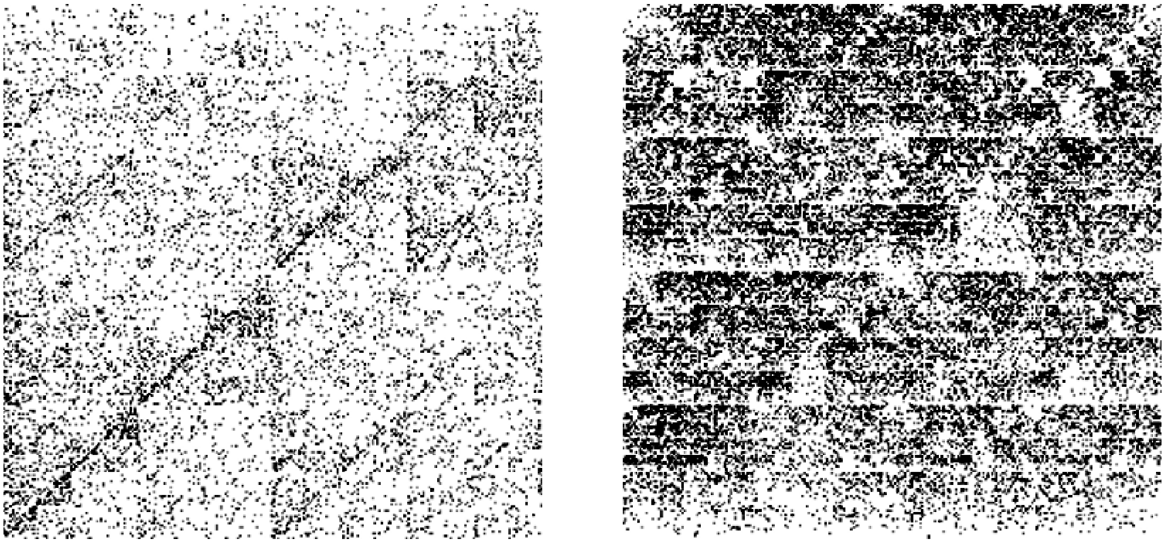
4. *Logos*:

   MLDSP-GUI is licensed under a Creative Commons Attribution 4.0 International License. This sub-panel contains the logos for Creative Commons, authors' affiliated institutions (The University of Western Ontario, and University of Waterloo), and MLDSP-GUI.



(a) Human (*Homo sapiens*)

(b) Bacterium (*Intrasporangium flavum*)



(c) Dengue virus

(d) Pseudomonas virus

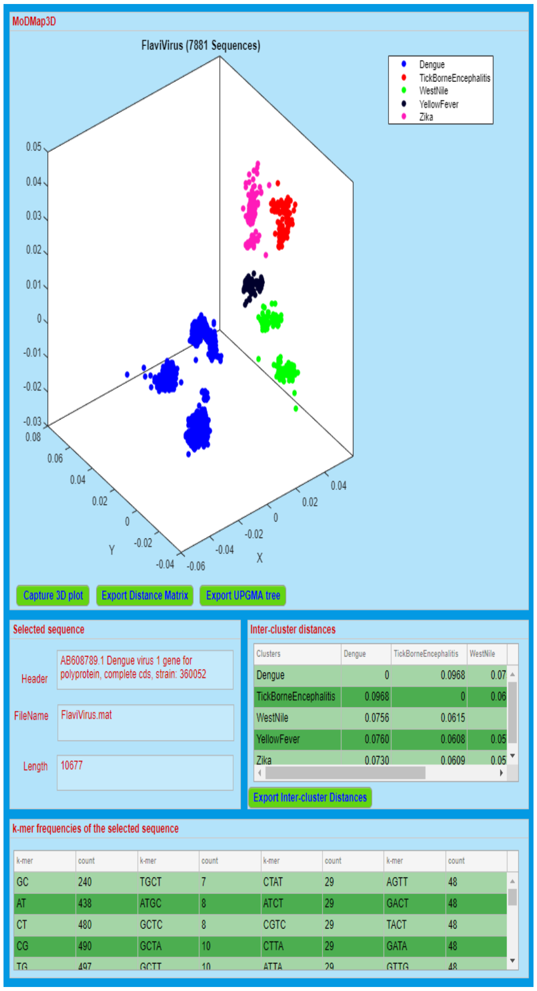
Supplementary Figure C.S4: Chaos Game Representation (CGR) of (a): *Homo sapiens* chromosome 1, first 100,000 bp segment, NCBI accession: *NC*_000001.11 (b): Bacterium (*Intrasporangium flavum*) complete genome, NCBI accession: *MLJO*01000003.1 (c): *Dengue* virus 1 complete genome, NCBI accession: *AB*608789.1 (d): *Pseudomonas* phage *Andromeda* complete genome, NCBI accession: *NC*_031014.1.

C.A.2 Center panel

The center panel components are shown in Figure C.S5.

1. *MoDMap3D*:

   This sub-panel shows the interactive three-dimensional Molecular Distance Map (MoDMap3D) visual representation of the interrelationships among the DNA sequences in the dataset. Each point represents a DNA sequence, and the positioning of points indicates the inter-sequence relationships based on the distance used (Pearson Correlation Coefficient, Euclidean, Manhattan). Clicking on a point results in information about the selected point/sequence being displayed in the panel Selected sequence. The user also has the option to Export Distance Matrix as an excel spreadsheet, to Export UPGMA tree (UPGMA = Unweighted Pair Group Method with Arithmetic mean) in Newick phylogenetic tree format, and to Capture 3D plot of the visualized molecular distance map, as a .png file, by clicking the respective buttons.



Selected sequence

Header: AB608789.1 Dengue virus 1 gene for polyprotein, complete cds, strain: 360052

FileName: FlaviVirus.mat

Length: 10677

Inter-cluster distances

Clusters	Dengue	TickBorneEncephalitis	WestNile
Dengue	0	0.0968	0.07
TickBorneEncephalitis	0.0968	0	0.06
WestNile	0.0756	0.0615	
YellowFever	0.0760	0.0608	0.05
Zika	0.0730	0.0609	0.05

k-mer frequencies of the selected sequence

k-mer	count	k-mer	count	k-mer	count	k-mer	count
GC	240	TGCT	7	CTAT	29	AGTT	48
AT	438	ATGC	8	ATCT	29	GACT	48
CT	480	GCTC	8	CGTC	29	TACT	48
CG	490	GCTA	10	CTTA	29	GATA	48
TG	497	GCTT	10	ATTA	29	GTTG	48

Supplementary Figure C.S5: Center panel components: MoDMap3D, selected sequence statistics, inter-cluster distances, and *k*-mer frequencies of the selected sequence. Export buttons for: saving 3D plot, distance matrix, UPGMA tree and inter-cluster distances.

   Note that the MoDMap3D should only be viewed as a visualization tool, and is not necessarily indicative of the classification accuracy of MLDSP-GUI. This is because

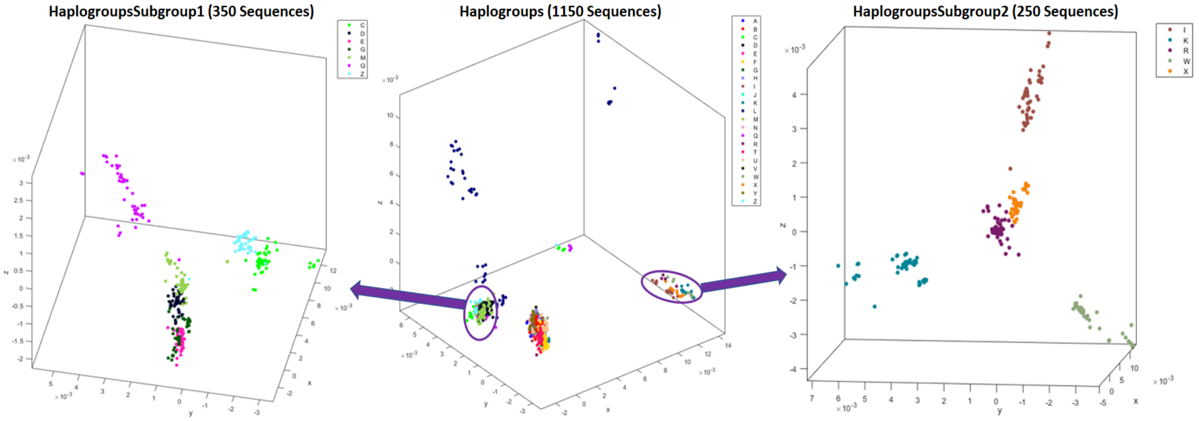
MoDMap3D is based on multidimensional scaling and it tries to map a multi-dimensional space onto a three-dimensional space. As such, the visual information it conveys may be imperfect (depending on the real dimensionality of the dataset that is visualized). In other words, clusters that appear to be overlapping in a MoDMap3D could in fact be perfectly separated by MLDSP-GUI, and the quantitative separability of clusters can only be accurately ascertained by looking at the accuracy scores of classifiers and at the confusion matrix.

As an example, Figure C.S6b shows some overlapping clusters (which indicates poor classification accuracy) in the MoDMap3D of 1,150 randomly chosen complete human mtDNA haplogroups (A, B, C, D, E, F, G, H, I, J, K, L, M, N, Q, R, T, U, V, W, X, Y, Z) sequences. However, the classification accuracy of the Linear Discriminant classifier for this dataset is reported to be 99%. The high accuracy of the quantitative classification is further confirmed by the clear visual separation obtained if we “zoom in" into the overlapping clusters of Figure C.S6b. Indeed Figure C.S6a, which displays human mtDNA haplogroups C, D, E, G, M, Q, Z, and Figure C.S6c which displays human mtDNA haplogroups I, K, R, W, X, both show clear separation.

As a concluding remark, when there is a discrepancy between MoDMap3D and the classification results of supervised machine learning, the latter is usually much better and also is the reliable quantitative result that should be used.

2. *Selected sequence*:

   Any point in a MoDMap3D can be selected by clicking on it. This sub-panel displays information about a selected point/sequence: Header (accession number, scientific name or other information available in the fasta file), FileName (name of its fasta file), and Length (in base pairs) of the selected sequence.



(a) (b) (c)

Supplementary Figure C.S6: "Zooming in" a ModMap3D, by re-plotting a subset of its dataset, can sometimes clarify cluster separations (separations can also be independently confirmed by the output of the supervised machine learning classifiers). Here, subfigures (a) and (c) are each obtained by re-plotting clusters which appear to be overlapping in the ModMap3D of the dataset of human mtDNA genomes from subfigure (b), as follows: **(a)** ModMap3D of 350 complete human mitochondrial genomes from the dataset in Table S1, line 13 (subset of dataset in line 12); **(b)** ModMap3D of 1,150 human mitochondrial genomes from the dataset in Table S1, line 12; **(c)** ModMap3D of 250 human mitochondrial genomes from the dataset in Table S1, line 14 (subset of dataset in line 12).

3. *Inter-cluster distances*:

   Inter-cluster distances are shown in this sub-panel. For n clusters, the inter-cluster distances are shown as an $n \times n$ matrix as follows. If M_i is the number of sequences in the cluster i, and $dist(a_s, b_t)$ gives the distance between any two sequences a_s, b_t, then the inter-cluster distance between any two clusters i and j where, $0 \leq i, j \leq n$, $1 \leq s \leq M_i$, $1 \leq t \leq M_j$, is computed as:

$$C(i, j) = \frac{\sum_{s=1}^{M_i} \sum_{t=1}^{M_j} dist(a_s, b_t)}{M_i \cdot M_j} \tag{C.5}$$

   The user also has the option to Export Inter-cluster Distances as an excel spreadsheet.

4. *k-mer frequencies of the selected sequence*:

   This sub-panel shows the k-mer frequencies (counts) for $2 \leq k \leq 4$, listed, for each k, in increasing order. This information can serve to analyze under-representation or over-representation of the respective oligomers.

C.A.3 Right panel

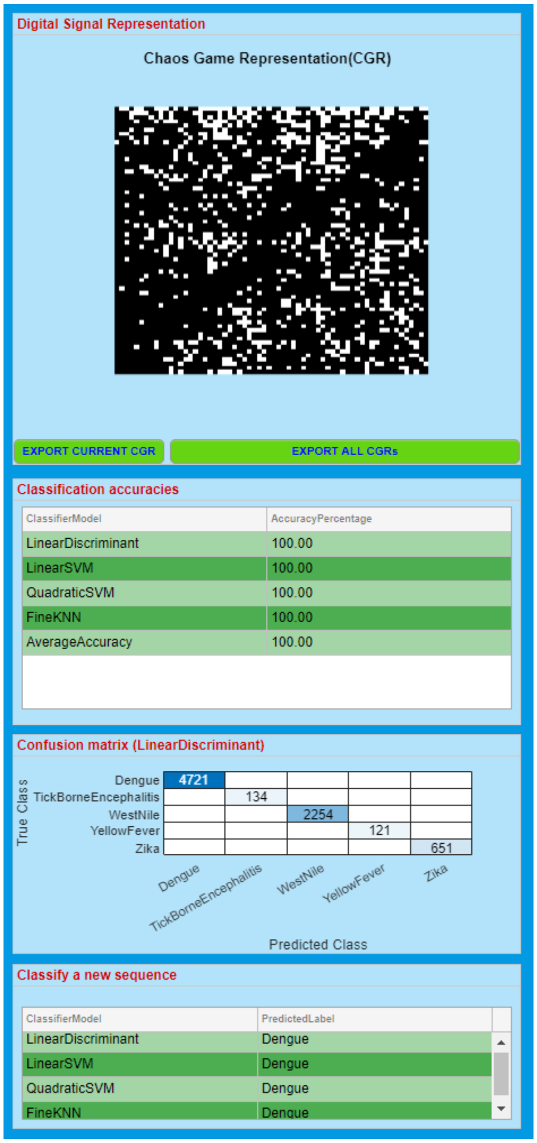
The right panel components are shown in Figure C.S7.

1. *Digital Signal Representation*:

   This sub-panel displays either the magnitude spectrum of the Discrete Fourier Transform applied to the numerical representation of a DNA sequence (if the one-dimensional representation was selected, Figure C.S8), or the CGR image of the DNA sequence (if the two-dimensional representation was selected, Figure C.S7).

2. *Classification accuracy*:

   The classification accuracies of six supervised machine learning classifiers (Linear Discriminant, Linear SVM, Quadratic SVM, Fine KNN, Subspace Discriminant, and Subspace KNN) using 10-fold cross validation is shown. Subspace Discriminant and Subspace KNN are omitted if the dataset has more than two thousand sequences. The average accuracy over all classifiers is also displayed.


Digital Signal Representation
Chaos Game Representation(CGR)
EXPORT CURRENT CGR
EXPORT ALL CGRs
Classification accuracies
Confusion matrix (LinearDiscriminant)
Classify a new sequence

ClassifierModel	AccuracyPercentage
LinearDiscriminant	100.00
LinearSVM	100.00
QuadraticSVM	100.00
FineKNN	100.00
AverageAccuracy	100.00

True Class \ Predicted Class	Dengue	TickBorneEncephalitis	WestNile	YellowFever	Zika
Dengue	4721				
TickBorneEncephalitis		134			
WestNile			2254		
YellowFever				121	
Zika					651

ClassifierModel	PredictedLabel
LinearDiscriminant	Dengue
LinearSVM	Dengue
QuadraticSVM	Dengue
FineKNN	Dengue

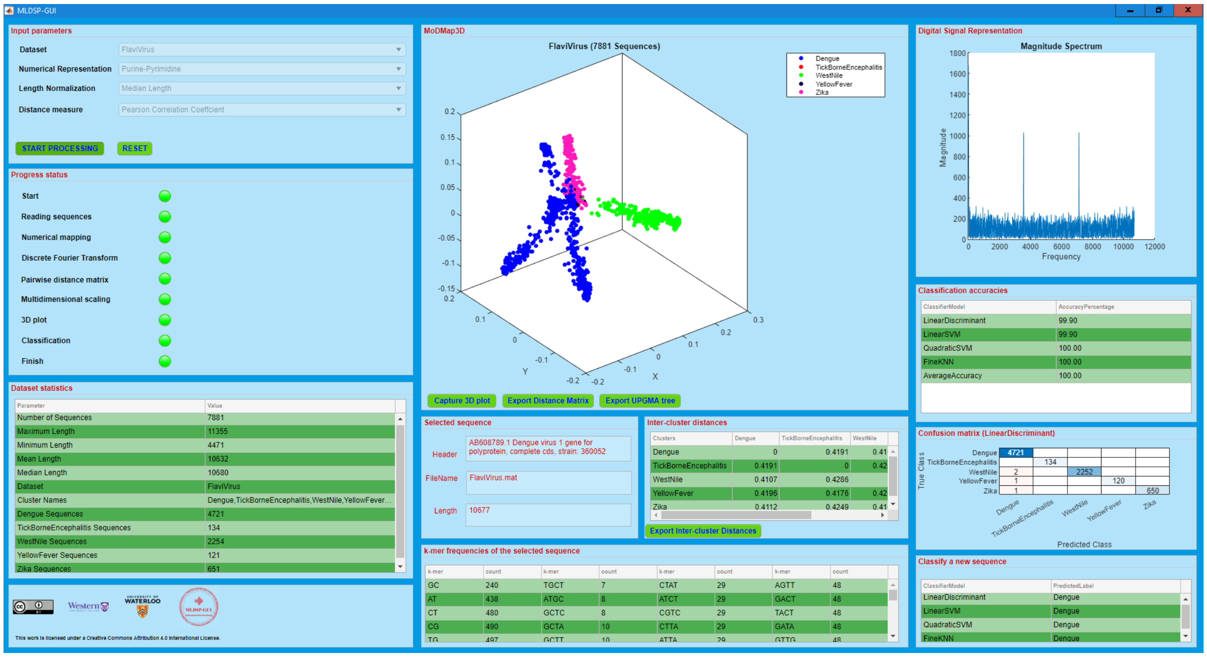
Supplementary Figure C.S7: Right panel components: Digital signal representation, classification accuracies, confusion matrix, and classify a new sequence.

3. *Confusion matrix*:

   A confusion matrix is displayed in this sub-panel, which changes dynamically depending on the classifier that is selected in the sub-panel above. For m clusters, the $m \times m$ confusion matrix has its rows labeled by the true classes and columns labeled by the predicted classes; the cell (i, j) shows the number of sequences that belong to the true class i, and have been predicted by the classifier to be of class j.

4. *Classify a new sequence*:

   MLDSP-GUI gives the option to predict the label of a new sequence, using all of the classifiers trained on a given dataset. The user can browse for a sequence (fasta file), and obtain the predicted label(s) as a result. Note that the new sequence will not be displayed in the MoDMap3D. Note also that any new sequence will be classified into one of the clusters that are displayed in the current MoDMap3D. This is an inherent limitation of supervised machine learning, in that a supervised machine learning classifier can only classify a new sequence into one of the clusters it has been trained on (it therefore classifies erroneously if the new sequence does not belong to any of the clusters that the classifier has previously “learned").



Dataset: FlaviVirus
Numerical Representation: Purine-Pyrimidine
Length Normalization: Median Length
Distance measure: Pearson Correlation Coeffcient

Dataset statistics

Parameter	Value
Number of Sequences	7881
Maximum Length	11355
Minimum Length	4471
Mean Length	10632
Median Length	10680
Dataset	FlaviVirus
Cluster Names	Dengue,TickBorneEncephalitis,WestNile,YellowFever...
Dengue Sequences	4721
TickBorneEncephalitis Sequences	134
WestNile Sequences	2254
YellowFever Sequences	121
Zika Sequences	651

This work is licensed under a Creative Commons Attribution 4.0 International License.

Selected sequence

Header: AB608789.1 Dengue virus 1 gene for polyprotein, complete cds, strain: 360052
FileName: FlaviVirus.mat
Length: 10677

Inter-cluster distances

Clusters	Dengue	TickBorneEncephalitis	WestNile
Dengue	0	0.4191	0.41
TickBorneEncephalitis	0.4191	0	0.42
WestNile	0.4107	0.4286	
YellowFever	0.4196	0.4176	0.42
Zika	0.4112	0.4249	0.41

k-mer frequencies of the selected sequence

k-mer	count	k-mer	count	k-mer	count	k-mer	count
GC	240	TGCT	7	CTAT	29	AGTT	48
AT	438	ATGC	8	ATCT	29	GACT	48
CT	480	GCTC	8	CGTC	29	TACT	48
CG	490	GCTA	10	CTTA	29	GATA	48
TG	497	GCTT	10	ATTA	29	GTTG	48

Classification accuracies

ClassifierModel	AccuracyPercentage
LinearDiscriminant	99.90
LinearSVM	99.90
QuadraticSVM	100.00
FineKNN	100.00
AverageAccuracy	100.00

Confusion matrix (LinearDiscriminant)

True Class \ Predicted Class	Dengue	TickBorneEncephalitis	WestNile	YellowFever	Zika
Dengue	4721				
TickBorneEncephalitis		134			
WestNile	2		2252		
YellowFever	1			120	
Zika	1				650

Classify a new sequence

ClassifierModel	PredictedLabel
LinearDiscriminant	Dengue
LinearSVM	Dengue
QuadraticSVM	Dengue
FineKNN	Dengue

Supplementary Figure C.S8: MLDSP-GUI test run for the 7,881 *Flavivirus* genomes in the dataset in Table S1, line 10 using the “purine/pyrimidine” representation with length normalization to median length. The Digital Signal Representation component (top right panel) shows the magnitude spectrum of the selected point/sequence. Note that even though this is the same dataset as the one in Figure C.S2, the visual shape of clusters is different and the classification accuracy is lower for the Linear Discriminant classifier. The visual differences in the clusters are due to the different numerical representations used. In general, the choice of numerical representation, supervised classifier, and other parameters depend on the specific dataset, and one should choose those that achieve the best numerical classification accuracy or confusion matrix.

C.B Provided datasets

Besides the datasets provided in the executable file (primates' mtDNA, influenza virus subtypes, *Flavivirus* viruses, mitochondrial disease genomes), MLDSP-GUI provides additional datasets that can be downloaded separately and imported into the already installed tool. All datasets were obtained from the NCBI Reference Sequence Database RefSeq on July 11, 2019, with the exception of the Disease-classification dataset (Table S1, line 6), which was obtained from Human Mitochondrial Database hmtDB on November 13, 2018. The additional datasets' details are given in Table C.S1.

C.C Availability

MLDSP-GUI is open-source, cross-platform compatible, and is available under the terms of the Creative Commons Attribution 4.0 International license (http://creativecommons.org/licenses/by/4.0/). The executable and dataset files are available at https://sourceforge.net/projects/mldsp-gui/.

Supplementary Table C.S1: Additional datasets provided.

S.No.	**Dataset**	**Number of sequences**	**Clusters**
1	**3classes**	**3,200**	**Amphibians: 264, Mammals: 1,133, Insects:1,803**
2	**Amphibians**	**264**	**Anura: 142, Caudata: 89, Gymnophiona: 33**
3	**Birds-Fish-Mammals**	**4,565**	**Birds (Aves): 698, Mammals (Mammalia): 2,734 Fish (Actinopterygii, Chondrichthyes, Coelacanthiformes, Dipnoi): 1,133**
4	**ClassToSubclass (Actinopterygii)**	**2,566**	**Chondrostei: 28, Cladistia: 11, Neopterygii: 2,527**
5	**Dengue**	**4,721**	**DENV-1: 2,008, DENV-2: 1,349, DENV-3: 1,010, DENV-4: 354**
6	**Disease-Classification**	**102**	**Epilepsy: 81, Glaucoma: 21**
7	**DomainToKingdom (Eukaryota)**	**9,727**	**Plants: 265, Animals: 8,825, Fungi: 393, Protists: 244**
8	**DomainToKingdom (Eukaryota_noProtists)**	**9,483**	**Plants: 265, Animals: 8,825, Fungi:393**
9	**FamilyToGenus (Cyprinidae)**	**92**	**Schizothorax: 24, Labeo: 21, Acrossocheilus: 15, Acheilognathus: 11, Rhodeus: 11, Onychostoma: 10**
10	**Flavivirus**	**7,881**	**Dengue: 4,721, TickBorneEncephalitis: 134, WestNile: 2,254, YellowFever: 121, Zika: 651**
11	**Fungi**	**340**	**Basidiomycota: 77, Pezizomycotina: 160, Saccharomycotina: 103**
12	**Human haplogroups**	**1,150**	**A:50, B:50, C:50, D:50, E:50, F:50, G:50, H:50, I:50, J:50, K:50, L:50, M:50, N:50, Q:50, R:50, T:50, U:50, V:50, W:50, X:50, Y:50, Z:50**
13	**Human haplogroups subgroup1**	**350**	**C:50, D:50, E:50, G:50, M:50, Q:50, Z:50**
14	**Human haplogroups subgroup2**	**250**	**I:50, K:50, R:50, W:50, X:50**
15	**Influenza**	**38**	**H1N1: 13, H2N2: 3, H5N1: 11, H7N3: 5, H7N9: 6**
16	**Insects**	**1636**	**Coleoptera: 196, Dictyptera: 235, Diptera: 253, Hemiptera: 272, Hymenoptera: 71, Lepidoptera: 442, Orthoptera: 167**
17	**KingdomToPhylum (Animalia)**	**8,792**	**Chordata: 5,224, Cnidaria: 157, Ecdysozoa: 2,585, Porifera: 64, Echinodermata: 67, Lophotrochozoa: 567, Platyhelminthes: 128**
18	**Mammalia**	**1,075**	**Xenarthrans: 36, Bats: 90, Carnivores: 145, Even-toed Ungulates: 271, Insectivores: 45, Marsupials: 35, Primates: 211, Rodents and Rabbits: 242**
19	**OrderToFamily (Cypriniformes)**	**756**	**Balitoridae: 29, Catostomidae: 14, Cobitidae: 55, Cyprinidae: 597, Nemacheilidae: 61**
20	**PhylumToSubphylum (Chordata)**	**5,224**	**Cephalochordata: 9, Craniata: 5,189, Tunicata:26**
21	**Plants**	**265**	**Chlorophyta: 66, Streptophyta: 199**
22	**Primates**	**211**	**Haplorrhini: 127, Strepsirrhini: 84**
23	**Protists**	**222**	**Alveolata: 38, Rhodophyta: 80, Stramenopiles: 104**
24	**SubclassToSuperorder (Neopterygii)**	**1,759**	**Osteoglossomorpha: 23, Elopomorpha: 63, Clupeomorpha: 92, Ostariophysi: 953, Protacanthopterygii: 76, Paracanthopterygii: 48, Acanthopterygii: 504**
25	**SubfamilyToGenus (Acheilognathinae)**	**26**	**Acheilognathus: 15, Rhodeus: 11**
26	**SubphylumToClass (Vertebrata)**	**5,176**	**Amphibians (Amphibia): 264, Birds (Aves): 698, Fish (Actinopterygii, Chondrichthyes, Dipnoi, Coelacanthiformes): 2,734, Mammals (Mammalia): 1,133, Reptiles (Crocodylia, Sphenodontia, Squamata, Testudines): 347**
27	**SuperorderToOrder (Ostariophysi)**	**942**	**Cypriniformes: 768, Characiformes: 40, Siluriformes: 134**

Bibliography

[1] P.D. Cristea, "Conversion of nucleotide sequences into genomic signals", *J Cell Mol Med.*, 6(2):279–303 (2002).

[2] H.J. Jeffrey, "Chaos game representation of gene structure", *Nucleic Acids Res.*, 18, 2163-2170 (1990).

[3] H.K. Kwan, and S.B. Arniker, "Numerical representation of DNA sequences", *IEEE International Conference on Electro/Information Technology*, 307–310 (2009).

[4] Gurjit S. Randhawa, Kathleen A. Hill, and Lila Kari, "ML-DSP: Machine Learning with Digital Signal Processing for ultrafast, accurate, and scalable genome classification at all taxonomic levels", *BMC Genomics*, **20**, 267 (2019).

Appendix D

Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: COVID-19 case study Supplementary Material

D.A Software availability

MLDSP-GUI is an open-source alignment-free tool with Graphical User Interface is publicly available for download at the following link (No license is required to download and use the tool):

https://sourceforge.net/projects/mldsp-gui/

MLDSP is an open-source alignment-free tool (MATLAB license required to run this program) available at the following link:

https://github.com/grandhawa/MLDSP

D.B Spearman's rank correlation coefficient test results

The ρ values from the Spearman's correlation coefficient test for $k = 1$ to $k = 7$ are given in Supplementary Table S1. The P-value is $< 1e - 5$ for $k = 2$ to $k = 6$, 0.0833 for $k = 1$ with an exception of 0.3333 in case of *Deltacorovavirus*.

COVID-19 vs.	**k=1**	**k=2**	**k=3**	**k=4**	**k=5**	**k=6**	**k=7**
Alphacoronavirus	1	0.97	0.96	0.96	0.94	0.88	0.70
Betacoronavirus	1	0.95	0.95	0.95	0.94	0.89	0.74
Gammacoronavirus	1	0.93	0.91	0.92	0.90	0.83	0.63
Deltacoronavirus	0.8	0.98	0.94	0.92	0.90	0.81	0.60
Embecovirus	1	0.85	0.88	0.88	0.86	0.79	0.59
Merbecovirus	1	0.96	0.95	0.94	0.92	0.84	0.64
Nobecovirus	1	0.89	0.83	0.83	0.80	0.73	0.54
Sarbecovirus	1	0.98	0.97	0.97	0.95	0.88	0.72

Supplementary Table D.S1: Spearman's rank correlation coefficient (ρ) value for $k = 1$ to $k = 7$.

D.C Dataset availability

All sequences downloaded from NCBI and Virus-Host-DB are uploaded to the SourceForge as fasta files. Accession numbers of 29 sequences downloaded from GISAID (28 COVID19 and a bat betacoronavirus RaTG13) are provided in 'GISAIDsequences.txt'.

https://sourceforge.net/projects/mldsp-gui/files/COVID19Dataset

The sequences are downloaded from three databases: Virus-Host-DB, NCBI, and GISAID.

Virus-Host-DB

All the viral sequences (apart from the ones downloaded from NCBI and GISAID) used in this study are obtained from the Virus-Host-DB available at:

https://www.genome.jp/virushostdb/

NCBI

Wuhan-Hu-1 complete genome (Accession: NC_045512.2), bat-SL-CoVZC45 (Accession: MG772933.1), and bat-SL-CoVZXC21 (Accession: MG772934.1) are obtained from the Virus-Host-DB available at:

https://www.ncbi.nlm.nih.gov/

GISAID

28 COVID-19 sequences and Beta-CoV-RaTG13 sequence (from Bat) are downloaded from the GISAID. We gratefully acknowledge the Authors, the Originating and Submitting Laboratories for their sequence and metadata shared through GISAID, which is used in this research. All submitters of 28 COVID-19 sequences and 1 BetaCoV/Bat/RaTG13 sequence may be contacted directly via

www.gisaid.org

Accession ID	Originating lab	Submitting lab	Authors
EPI_ISL_404227	Zhejiang Provincial Center for Disease Control and Prevention	Department of Microbiology, Zhejiang Provincial Center for Disease Control and Prevention	Yin Chen, Yanjun Zhang, Haiyan Mao, Junhang Pan, Xiuyu Lou, Yiyu Lu, Juying Yan, Hanping Zhu, Jian Gao, Yan Feng, Yi Sun, Hao Yan, Zhen Li, Yisheng Sun, Liming Gong, Qiong Ge, Wen Shi, Xinying Wang, Wenwu Yao, Zhangnv Yang, Fang Xu, Chen Chen, Enfu Chen, Zhen Wang, Zhiping Chen, Jianmin Jiang, Chonggao Hu
EPI_ISL_404228	Zhejiang Provincial Center for Disease Control and Prevention	Department of Microbiology, Zhejiang Provincial Center for Disease Control and Prevention	Yanjun Zhang, Yin Chen, Haiyan Mao, Junhang Pan, Xiuyu Lou, Yiyu Lu, Juying Yan, Hanping Zhu, Jian Gao, Yan Feng, Yi Sun, Hao Yan, Zhen Li, Yisheng Sun, Liming Gong, Qiong Ge, Wen Shi, Xinying Wang, Wenwu Yao, Zhangnv Yang, Fang Xu, Chen Chen, Enfu Chen, Zhen Wang, Zhiping Chen, Jianmin Jiang, Chonggao Hu
EPI_ISL_402132	Wuhan Jinyintan Hospital	Hubei Provincial Center for Disease Control and Prevention	Bin Fang, Xiang Li, Xiao Yu, Linlin Liu, Bo Yang, Faxian Zhan, Guojun Ye, Xixiang Huo, Junqiang Xu, Bo Yu, Kun Cai, Jing Li, Yongzhong Jiang
EPI_ISL_402127 EPI_ISL_402128 EPI_ISL_402129 EPI_ISL_402130 EPI_ISL_402124	Wuhan Jinyintan Hospital	Wuhan Institute of Virology, Chinese Academy of Sciences	Peng Zhou, Xing-Lou Yang, Ding-Yu Zhang, Lei Zhang, Yan Zhu, Hao-Rui Si, Zhengli Shi
EPI_ISL_403963 EPI_ISL_403962	Bamrasnaradura Hospital	1. Department of Medical Sciences, Ministry of Public Health, Thailand 2. Thai Red Cross Emerging Infectious Diseases - Health Science Centre 3. Department of Disease Control, Ministry of Public Health, Thailand	Pilailuk, Okada; Siripaporn, Phuygun; Thanutsapa, Thanadachakul; Supaporn, Wacharapluesadee; Sittiporn,Parnmen; Warawan,Wongboot; Sunthareeya, Waicharoen; Rome, Buathong; Malinee, Chittaganpitch; Nanthawan, Mekha
EPI_ISL_402120 EPI_ISL_402119 EPI_ISL_402121	National Institute for Viral Disease Control and Prevention, China CDC	National Institute for Viral Disease Control and Prevention, China CDC	Wenjie Tan, Xiang Zhao, Wenling Wang, Xuejun Ma, Yongzhong Jiang, Roujian Lu, Ji Wang, Weimin Zhou, Peihua Niu, Peipei Liu, Faxian Zhan, Weifeng Shi, Baoying Huang, Jun Liu, Li Zhao, Yao Meng, Xiaozhou He, Fei Ye, Na Zhu, Yang Li, Jing Chen, Wenbo Xu, George F. Gao, Guizhen Wu
EPI_ISL_402123	Institute of Pathogen Biology, Chinese Academy of Medical Sciences & Peking Union Medical College	Institute of Pathogen Biology, Chinese Academy of Medical Sciences & Peking Union Medical College	Lili Ren, Jianwei Wang, Qi Jin, Zichun Xiang, Zhiqiang Wu, Chao Wu, Yiwei Liu
EPI_ISL_402125	unknown	National Institute for Communicable Disease Control and Prevention (ICDC) Chinese Center for Disease Control and Prevention (China CDC)	Zhang,Y.-Z., Wu,F., Chen,Y.-M., Pei,Y.-Y., Xu,L., Wang,W., Zhao,S., Yu,B., Hu,Y., Tao,Z.-W., Song,Z.-G., Tian,J.-H., Zhang,Y.-L., Liu,Y., Zheng,J.-J., Dai,F.-H., Wang,Q.-M., She,J.-L. and Zhu,T.-Y.
EPI_ISL_403931 EPI_ISL_403928 EPI_ISL_403930 EPI_ISL_403929	Institute of Pathogen Biology, Chinese Academy of Medical Sciences & Peking Union Medical College	Institute of Pathogen Biology, Chinese Academy of Medical Sciences & Peking Union Medical College	Lili Ren, Jianwei Wang, Qi Jin, Zichun Xiang, Zhiqiang Wu, Chao Wu, Yiwei Liu
EPI_ISL_403937 EPI_ISL_403936 EPI_ISL_403935 EPI_ISL_403934 EPI_ISL_403933 EPI_ISL_403932	Guangdong Provincial Center for Diseases Control and Prevention; Guangdong Provincial Public Health	Department of Microbiology, Guangdong Provincial Center for Diseases Control and Prevention	Min Kang, Jie Wu, Jing Lu, Tao Liu, Baisheng Li, Shujiang Mei, Feng Ruan, Lifeng Lin, Changwen Ke, Haojie Zhong, Yingtao Zhang, Lirong Zou, Xuguang Chen, Qi Zhu, Jianpeng Xiao, Jianxiang Geng, Zhe Liu, Jianxiong Hu, Weilin Zeng, Xing Li, Yuhuang Liao, Xiujuan Tang, Songjian Xiao, Ying Wang, Yingchao Song, Xue Zhuang, Lijun Liang, Guanhao He, Huihong Deng, Tie Song, Jianfeng He, Wenjun Ma
EPI_ISL_404895	Providence Regional Medical Center	Division of Viral Diseases, Centers for Disease Control and Prevention	Queen,K., Tao,Y., Li,Y., Paden,C.R., Lu,X., Zhang,J., Gerber,S.I., Lindstrom,S.
EPI_ISL_404253	IL Department of Public Health Chicago Laboratory	Pathogen Discovery, Respiratory Viruses Branch, Division of Viral Diseases, Centers for Dieases Control and Prevention	Ying Tao, Krista Queen, Clinton R. Paden, Jing Zhang, Yan Li, Anna Uehara, Xiaoyan Lu, Brian Lynch, Senthil Kumar K. Sakthivel, Brett L. Whitaker, Shifaq Kamili, Lijuan Wang, Janna' R. Murray, Susan I. Gerber, Stephen Lindstrom, Suxiang Tong
EPI_ISL_405839	The University of Hong Kong - Shenzhen Hospital	Li Ka Shing Faculty of Medicine, The University of Hong Kong	Chan,J.F.-W., Yuan,S., Kok,K.H., To,K.K.-W., Chu,H., Yang,J., Xing,F., Liu,J., Yip,C.C.-Y., Poon,R.W.-S., Tsai,H.W., Lo,S.K.-F., Chan,K.H., Poon,V.K.-M., Chan,W.M., Ip,J.D., Cai,J.P., Cheng,V.C.-C., Chen,H., Hui,C.K.-M., Yuen,K.Y.
EPI_ISL_402131 (Bat RaTG13)	Wuhan Institute of Virology, Chinese Academy of Sciences	Wuhan Institute of Virology, Chinese Academy of Sciences	Yan Zhu, Ping Yu, Bei Li, Ben Hu, Hao-Rui Si, Xing-Lou Yang, Peng Zhou, Zheng-Li Shi

Supplementary Table D.S2: Accession IDs of the sequences downloaded from the GISAID.

The accession numbers and the sources for all the used sequences are given below:

Test-1; Source: Virus-Host-DB
Adenoviridae
AB448767,AC_000007,JN880453,JN880454,JN880455,JN880456,JN935766,JQ326209,JQ776547,KC529648,KC693021,KF268207,AC_000008,KF279629,KF528688,KF802426,KF906413,KM591901,KM591902,KM591903,NC_000899,NC_000942,NC_001405,AC_000009,NC_001454,NC_001460,NC_001720,NC_001734,NC_001813,NC_001876,NC_001958,NC_002501,NC_002513,NC_002685,AC_000010,NC_002702,NC_003266,NC_004037,NC_006144,NC_006879,NC_009989,NC_010956,NC_011202,NC_011203,NC_012584,AC_000011,NC_012959,NC_014564,NC_014899,NC_014969,NC_015225,NC_015323,NC_015455,NC_015932,NC_016437,NC_016895,AC_000012,NC_017825,NC_017979,NC_020074,NC_020485,NC_020487,NC_021168,NC_021221,NC_022266,NC_022612,NC_022613,AC_000013,NC_024150,NC_024474,NC_024486,NC_024684,NC_025678,NC_025962,NC_027705,NC_027708,NC_028103,NC_028105,AC_000014,NC_028107,NC_028113,NC_029898,NC_029899,NC_029902,NC_030116,NC_030792,NC_030860,NC_030874,NC_031503,AC_000016,NC_031948,NC_032105,NC_034382,NC_034626,NC_034834,NC_035072,NC_035207,NC_035619,NC_038332,NC_038333,AC_000017,NC_038334,NC_039032,NC_040811,NC_043094,NC_043405,NC_043696,U46933,X73487,Y09598,AB724351,AC_000018,AC_000019,AC_000020,AC_000189,AC_000190,AC_000191,AF036092,AF083975,AF108105,AM749299,AB765926,AP012285,AP012302,AY458656,AY737797,AY737798,AY803294,AY849321,AY875648,DQ086466,DQ315364,AC_000001,DQ393829,DQ792570,DQ900900,DQ923122,EF121005,EF564601,FJ025899,FJ025900,FJ025901,FJ025902,AC_000002,FJ025903,FJ025904,FJ025905,FJ025906,FJ025907,FJ025908,FJ025909,FJ025910,FJ025911,FJ025912,AC_000003,FJ025913,FJ025914,FJ025915,FJ025916,FJ025917,FJ025918,FJ025919,FJ025920,FJ025921,FJ025922,AC_000004,FJ025923,FJ025924,FJ025925,FJ025926,FJ025927,FJ025928,FJ025929,FJ025930,FJ349096,FJ404771,AC_000005,FJ597732,FJ643676,FJ824826,GQ384080,GU191019,HM770721,HQ241818,HQ241820,HQ883276,JF964962,AC_000006,JN860676,JN860677,JN860678,JN860679,JN860680,JN880448,JN880449,JN880450,JN880451,JN880452
Anelloviridae
AM711976,AM712003,AM712004,AM712030,AM712031,AM712032,AM712033,AM712034,FR823283,GU450331,HQ335082,HQ335083,HQ335084,HQ335085,JN704611,KJ194622,KM262781,KM262785,NC_001427,NC_002076,NC_002195,NC_007013,NC_007014,NC_009225,NC_012126,NC_014068,NC_014069,NC_014070,NC_014071,NC_014072,NC_014073,NC_014074,NC_014075,NC_014076,NC_014077,NC_014078,NC_014079,NC_014080,NC_014081,NC_014082,NC_014083,NC_014084,NC_014085,NC_014086,NC_014087,NC_014088,NC_014089,NC_014090,NC_014091,NC_014092,NC_014093,NC_014094,NC_014095,NC_014096,NC_014097,NC_014480,NC_015212,NC_015396,NC_015783,NC_017091,NC_018401,NC_020498,NC_022788,NC_022789,NC_024890,NC_024891,NC_024908,NC_025215,NC_025726,NC_025727,NC_025966,NC_026138,NC_026662,NC_026663,NC_026664,NC_026764,NC_026765,NC_027059,NC_027430,NC_030297,NC_030650,NC_034978,NC_035135,NC_035136,NC_035192,NC_038336,NC_038337,NC_038338,NC_038339,NC_038340,NC_038341,NC_038342,

NC_038343,NC_038344,NC_038345,NC_038346,NC_038347,NC_038348,NC_038349,NC_038350,NC_038351,NC_038352,NC_038353,NC_038354,NC_038355,NC_038356,NC_038357,NC_038358,NC_038359,NC_038360,NC_038361,NC_038362,NC_038363,NC_040531,NC_040546,NC_040547,NC_040617,NC_040618,NC_040668,NC_040686,NC_040687,NC_040720, NC_040801,NC_043413,NC_043414,NC_043415
Caudovirales
AB626963,AB746912,AB757801,AF527608,AP011956,AY526908,AY526909,CP000711,CP008753,DQ113772,DQ121662,DQ222851,DQ289556,DQ394806,DQ394807,DQ394808,DQ394809,DQ394810,DQ426905,DQ838728,EU056923,EU568876,EU622808,FQ482084,GQ303261,GQ478082,GQ478083,GQ478085,GQ478087,GU196281,HE614282,HE956707,HE983844,HG428758,HG793132,HG796219,HG796220,HG796221,HM152765,HQ110083,HQ634152,HQ641341,HQ641343,HQ641344,HQ641346,JF314845,JF767210,JF773396,JN175269,JN254801,JN255163,JN699002,JN811560,JQ067085,JQ267518,JQ691610,JQ740790,JQ740791,JQ740792,JQ740793,JQ740794,JQ740795,JQ740796,JQ740797,JQ740798,JQ740799,JQ740800,JQ740801,JQ740802,JQ740803,JQ740805,JQ740806,JQ740807,JQ740808,JQ740809,JQ740810,JQ740811,JQ740812,JQ740814,JQ780163,JQ957925,JQ965700,JQ965701,JQ965702,JQ965703,JX000007,JX174275,JX274646,JX274647,JX403939,JX409894,JX409895,JX421753,JX483873,JX483874,JX483875,JX483879,JX483880,JX564242,JX570703,JX570707,JX570708,JX570711,JX681814,KC182543,KC182544,KC182545,KC182548,KC182549,KC182550,KC330681,KC333879,KC348598,KC348599,KC348600,KC348601,KC348602,KC348603,KC348604,KC413987,KC413988,KC522412,KC542353,KC556893,KC556894,KC556895,KC556896,KC556898,KC787107,KC787108,KC821615,KC821627,KC911856,KC911857,KC969441,KF030445,KF302032,KF302033,KF302035,KF302036,KF302037,KF591601,KF669657,KF676640,KF751793,KF751794,KF751795,KF751796,KF751797,KF771236,KF800937,KJ018210,KJ021043,KJ417497,KJ502657,KJ545483,KJ572844,KJ578763,KJ578764,KJ578766,KJ578769,KJ578771,KJ578775,KJ578777,KJ617393,KJ725374,KM058087,KM091442,KM091443,KM091444,KM233455,KM591905,KM612260,KM612261,KM612262,KM612263,KM612265,KM923970,KP017310,KP209285,KP296794,KP791807,KP869108,KR131710,LK985321,LN610580,LN681534,M11813,NC_000871,NC_000872,NC_000896,NC_000929,NC_000935,NC_001271,NC_001317,NC_001416,NC_001604,NC_001609,NC_001629,NC_001697,NC_001706,NC_001825,NC_001835,NC_001895,NC_001900,NC_001901,NC_001902,NC_001909,NC_001978,NC_002072,NC_002166,NC_002167,NC_002185,NC_002214,NC_002321,NC_002371,NC_002486,NC_002515,NC_002519,NC_002628,NC_002649,NC_002661,NC_002666,NC_002667,NC_002668,NC_002669,NC_002670,NC_002671,NC_002703,NC_002730,NC_002747,NC_002796,NC_003050,NC_003085,NC_003157,NC_003216,NC_003278,NC_003288,NC_003291,NC_003298,NC_003313,NC_003315,NC_003356,NC_003390,NC_003444,NC_003524,NC_003907,NC_004066,NC_004112,NC_004165,NC_004166,NC_004167,NC_004302,NC_004303,NC_004305,NC_004313,NC_004333,NC_004348,NC_004456,NC_004466,NC_004584,NC_004585,NC_004586,NC_004587,NC_004588,NC_004589,NC_004615,NC_004616,NC_004617,NC_004664,NC_004665,NC_004678,NC_004679,NC_004740,NC_004745,NC_004746,NC_004775,NC_004777,NC_004814,

NC_004821,NC_004827,NC_004831,NC_004902,NC_004996,NC_005045,NC_005056,NC_005069,NC_005178,NC_005263,NC_005294,NC_005340,NC_005342,NC_005344,NC_005345,NC_005354,NC_005355,NC_005356,NC_005357,NC_005822,NC_005833,NC_005841,NC_005879,NC_005882,NC_005884,NC_005886,NC_005887,NC_005891,NC_005893,NC_006356,NC_006548,NC_006557,NC_006882,NC_006936,NC_006940,NC_006949,NC_006953,NC_007019,NC_007046,NC_007047,NC_007048,NC_007049,NC_007050,NC_007051,NC_007052,NC_007053,NC_007054,NC_007055,NC_007056,NC_007057,NC_007058,NC_007059,NC_007060,NC_007061,NC_007062,NC_007063,NC_007064,NC_007065,NC_007145,NC_007149,NC_007291,NC_007456,NC_007458,NC_007497,NC_007501,NC_007603,NC_007637,NC_007709,NC_007710,NC_007734,NC_007804,NC_007805,NC_007806,NC_007807,NC_007808,NC_007814,NC_007924,NC_007967,NC_008152,NC_008193,NC_008201,NC_008202,NC_008265,NC_008363,NC_008364,NC_008367,NC_008370,NC_008371,NC_008376,NC_008583,NC_008617,NC_008689,NC_008694,NC_008695,NC_008717,NC_008721,NC_008722,NC_008723,NC_008798,NC_008799,NC_009014,NC_009016,NC_009018,NC_009232,NC_009234,NC_009235,NC_009236,NC_009237,NC_009382,NC_009514,NC_009526,NC_009531,NC_009540,NC_009541,NC_009542,NC_009543,NC_009551,NC_009552,NC_009554,NC_009603,NC_009604,NC_009643,NC_009737,NC_009761,NC_009762,NC_009763,NC_009799,NC_009810,NC_009812,NC_009813,NC_009814,NC_009815,NC_009818,NC_009819,NC_009875,NC_009935,NC_009936,NC_009990,NC_010147,NC_010179,NC_010275,NC_010325,NC_010326,NC_010342,NC_010353,NC_010363,NC_010463,NC_010495,NC_010807,NC_010808,NC_010945,NC_011038,NC_011040,NC_011042,NC_011043,NC_011045, NC_011046,NC_011048,NC_011085,NC_011104,NC_011107,NC_011142,NC_011201,NC_011216,NC_011222,NC_011267,NC_011291,NC_011308,NC_011318,NC_011344,NC_011373,NC_011534,NC_011551,NC_011589,NC_011611,NC_011612,NC_011613,NC_011614,NC_011645,NC_011646,NC_011801,NC_011802,NC_011976,NC_012223,NC_012418,NC_012419,NC_012662,NC_012742,NC_012753,NC_012756,NC_012784,NC_012788,NC_012884,NC_013055,NC_013059,NC_013152,NC_013153,NC_013154,NC_013155,NC_013195,NC_013594,NC_013597,NC_013598,NC_013599,NC_013600,NC_013638,NC_013643,NC_013644,NC_013645,NC_013646,NC_013647,NC_013648, NC_013649,NC_013651,NC_013696,NC_014229,NC_014460,NC_014900,NC_015158,NC_015159,NC_015208
Geminiviridae
KF229718,KF229722,KF652077,KJ628309,KM189819,L14460,L14461,L39638,NC_000869,NC_000870,NC_000882,NC_001346,NC_001359,NC_001369,NC_001412,NC_001438,NC_001439,NC_001466,NC_001467,NC_001468,NC_001478,NC_001507,NC_001508,NC_001647,NC_001828,NC_001868,NC_001917,NC_001928,NC_001929,NC_001930,NC_001931,NC_001932,NC_001933,NC_001934,NC_001935,NC_001936,NC_001937,NC_001938,NC_001939,NC_001983,NC_001984,NC_002046,NC_002047,NC_002048,NC_002049,NC_002510,NC_002543,NC_002555,NC_002556,NC_002817,NC_002981,NC_002984,NC_002985,NC_003199,NC_003326,NC_003357,NC_003379,NC_003418,NC_003434,NC_003493,NC_003504,NC_003505,NC_003556,NC_003609,NC_003664,NC_003665,NC_003708,NC_003709,NC_003722,NC_003744,NC_003803,

NC_003804,NC_003822,NC_003825,NC_003828,NC_003830,NC_003831,NC_003856,
NC_003857,NC_003860,NC_003861,NC_003862,NC_003865,NC_003866,NC_003867,
NC_003868,NC_003887,NC_003891,NC_003896,NC_003897,NC_003898,NC_004005,
NC_004042,NC_004043,NC_004044,NC_004071,NC_004090,NC_004091,NC_004096,
NC_004097,NC_004098,NC_004099,NC_004100,NC_004101,NC_004147,NC_004153,
NC_004192,NC_004300,NC_004356,NC_004558,NC_004559,NC_004569,NC_004580,
NC_004581,NC_004582,NC_004583,NC_004607,NC_004608,NC_004609,NC_004611,
NC_004612,NC_004613,NC_004614,NC_004618,NC_004625,NC_004626,NC_004627,
NC_004628,NC_004630,NC_004634,NC_004635,NC_004637,NC_004638,NC_004639,
NC_004640,NC_004641,NC_004642,NC_004644,NC_004645,NC_004646,NC_004647,
NC_004648,NC_004650,NC_004651,NC_004654,NC_004655,NC_004656,NC_004657,
NC_004658,NC_004659,NC_004660,NC_004661,NC_004662,NC_004673,NC_004674,
NC_004675,NC_004676,NC_004732,NC_004755,NC_004824,NC_004825,NC_005031,
NC_005032,NC_005319,NC_005320,NC_005321,NC_005330,NC_005331,NC_005338,
NC_005347,NC_005348,NC_005635,NC_005636,NC_005807,NC_005811,NC_005812,
NC_005842,NC_005843,NC_005844,NC_005845,NC_005846,NC_005850,NC_005851,
NC_005852,NC_005853,NC_005855,NC_006358,NC_006359,NC_006384,NC_006631,
NC_006874,NC_006876,NC_006995,NC_007210,NC_007211,NC_007290,NC_007338,
NC_007339,NC_007638,NC_007723,NC_007724,NC_007726,NC_007727,NC_007730,
NC_007965,NC_007966,NC_008056,NC_008057,NC_008058,NC_008059,NC_008236,
NC_008267,NC_008283,NC_008284,NC_008299,NC_008304,NC_008305,NC_008316,
NC_008317,NC_008329,NC_008373,NC_008374,NC_008377,NC_008492,NC_008493,
NC_008494,NC_008495,NC_008517,NC_008559,NC_008779,NC_008780,NC_008793,
NC_008794,NC_009030,NC_009031,NC_009088,NC_009354,NC_009451,NC_009490,
NC_009491,NC_009545,NC_009546,NC_009547,NC_009548,NC_009549,NC_009550,
NC_009553,NC_009605,NC_009606,NC_009607,NC_009612,NC_009644,NC_009645,
NC_009646,NC_009647,NC_010238,NC_010293,NC_010294,NC_010307,NC_010313,
NC_010352,NC_010417,NC_010435,NC_010439,NC_010440,NC_010441,NC_010618,
NC_010647,NC_010648,NC_010713,NC_010714,NC_010791,NC_010792,NC_010797,
NC_010799,NC_010812,NC_010818,NC_010833,NC_010834,NC_010835,NC_010836,
NC_010837,NC_010838,NC_010839,NC_010840,NC_010946,NC_010947,NC_010948,
NC_010949,NC_010950, NC_010951,NC_010952,NC_010953,NC_011024,NC_011052,
NC_011058,NC_011096,NC_011135,NC_011181,NC_011182,NC_011268,NC_011309,
NC_011346,NC_011347,NC_011348,NC_011583,NC_011584,NC_011804,NC_011805,
NC_011919,NC_012041,NC_012118,NC_012120,NC_012137,NC_012206,NC_012481,
NC_012482,NC_012492,NC_012553,NC_012554,NC_012664,NC_012665,NC_012786,
NC_012787,NC_013017,AB007990,AB110218,AB162141,AB192965,AB192966,
AB236321,AB236323,AB236325,AB306314,AB439841,AB439842,AB921568,AF105975,
AF112352,AF112353,AF141897,AF141922,AF173555,AF173556,AF241479,AF291705,
AF291706,AF329886,AF329888,AF329889,AF379637,AF416741,AF416742,AF428255,
AF490004,AF491306,AJ132574,AJ132575,AJ223505,AJ224504,AJ311031,AJ314739,
AJ314740,AJ319674,AJ420316,AJ420317,AJ420318,AJ457823,AJ457824,AJ457985,
AJ457986,AJ496286,AJ496287,AJ512761,AJ512762,AJ543429,AJ564742,AJ564743,

AJ566744,AJ579307,AJ579308,AJ781302,AJ810156,AJ810157,AJ849916,AJ865337, AJ971263,AM181683,AM183224,AM230634,AM230635,AM261326,AM421522, AM691745,AM712436,AM940137,AM980883,AM989927,AY036009,AY036010, AY090555,AY090556,AY090557,AY090558,AY190290,AY190291,AY650283,AY754814, D00200,D00201,D00940,DQ845787,DQ868525,EF011559,EF015778,EF190217, EF536859,EF536861,EF536868,EF536873,EF536876,EF536878,EF536886,EU024118, EU024119,EU024120,EU273816,EU273817,EU273818,EU365686,EU856366,FJ176701, FJ560719,FJ751234,FM179613,FM210034,FM877474,FN252890,FN256256,FN256257, FN256258,FN256259,FN256260,FN256261,FN256292,FN297834,FN401520,FN436001, GU001879,GU076440,GU076443,GU076445,GU076447,GU076449,GU076451, GU076452,GU076454,GU180085,GU256531,GU440580,GU732203,HE580236,HE616777,HE659516,HE659517,HE793429,HM140364,HM140365,HM140366,HM140368, HM140369,HM140370,HM140371,HM626516,HM626517,HM859902,HM859903, JF451352,JN676150,JN676151,JN680352,JN680353,JN989417,JN989425,JN989441, JN989446,JQ247188,JQ303121,JQ303122,JQ621843,JX082259,JX448368,JX911332, K02029,K02030,KC149941,KC172700,KC427995,KC476655,KF176552
Genomoviridae
JN704610,KF371641,KF371642,KP133076,KP133077,KP133078,KP133079,KP133080, NC_013116,NC_023844,NC_023870,NC_023871,NC_023872,NC_024689,NC_024690, NC_024691,NC_024909,NC_025728,NC_025729,NC_025730,NC_025731,NC_025732, NC_025733,NC_025734,NC_025735,NC_025736,NC_025737,NC_025738,NC_025741, NC_026144,NC_026161,NC_026162,NC_026163,NC_026164,NC_026165,NC_026166, NC_026167,NC_026168,NC_026169,NC_026254,NC_026261,NC_026806,NC_026807, NC_026808,NC_026809,NC_026810,NC_026817,NC_026818,NC_027776,NC_027820, NC_027821,NC_028459,NC_028460,NC_030138,NC_030139,NC_030140,NC_030141, NC_030142,NC_030143,NC_030144,NC_030145,NC_030146,NC_030147,NC_030447, NC_030448,NC_030887,NC_033270,NC_033736,NC_033742,NC_033743,NC_033747, NC_035137,NC_035138,NC_035139,NC_035197,NC_035477,NC_037062,NC_038479, NC_038480,NC_038481,NC_038482,NC_038483,NC_038484,NC_038485,NC_038486, NC_038487,NC_038488,NC_038489,NC_038490,NC_038491,NC_038492,NC_038493, NC_038494,NC_038495,NC_038496,NC_038497,NC_038498,NC_038499,NC_038501, NC_038502,NC_040317,NC_040326,NC_040327,NC_040330,NC_040338,NC_040339, NC_040340,NC_040346,NC_040347,NC_040348,NC_040351,NC_040370,NC_040371, NC_040372,NC_040379
Microviridae
AJ550635,AY751298,DQ079873,DQ079878,DQ079880,DQ079881,DQ079882,DQ079883, DQ079884,DQ079885,DQ079886,DQ079887,DQ079888,DQ079889,DQ079890, DQ079891,DQ079892,DQ079893,DQ079894,DQ079895,DQ079896,DQ079897, DQ079898,DQ079899,DQ079900,DQ079901,DQ079902,DQ079903,DQ079904, DQ079905,DQ079906,DQ079907,DQ079908,DQ079909,KC237308,KC821628, KC821631,KF044309,KF044310,KJ997912,M14428,NC_001330,NC_001420, NC_001422,NC_001730,NC_001741,NC_001998,NC_002180,NC_002194,NC_002643, NC_003438,NC_007461,NC_007817,NC_007818,NC_007819,NC_007820,NC_007821,

NC_007822,NC_007823,NC_007824,NC_007825,NC_007827,NC_007856,NC_012868, NC_015785,NC_021797,NC_021805,NC_022790,NC_026013,NC_026665,NC_027633, NC_027634,NC_027635,NC_027636,NC_027637,NC_027638,NC_027639,NC_027640, NC_027641,NC_027642,NC_027643,NC_027644,NC_027645,NC_027646,NC_027647, NC_027648,NC_028993,NC_028994,NC_029012,NC_029014,NC_030458,NC_030472, NC_030476,NC_040328,NC_040329,NC_040341,NC_040342,NC_040349,NC_040350, NC_040373,NC_040374,NC_040375
Ortervirales
AB187566,AB611707,AF033809,AF053745,AF126467,AF221065,AF411814,DQ093792, DQ241301, DQ241302,DQ365814,DQ399707,DQ451009,DQ822073,EF133960,EF494181 ,EU293537,EU523109,FJ195346,FN692043,HM210570,HQ154630,HQ246218,HQ540591, HQ540595,J01998,JF274252,JF908815,JN032736,JN134185,JQ303225,JX245014, KC802224,KF029431,KF313137,KJ668270,KJ668271,KP284572,M11841,M14008, M16605,M23385,NC_000858,NC_001343,NC_001362,NC_001364,NC_001402,Y13051 NC_001403,NC_001407,NC_001408,NC_001413,NC_001414,NC_001436,NC_001450, NC_001452,NC_001463,NC_001482,NC_001488,NC_001494,NC_001497,NC_001499, NC_001500,NC_001501,NC_001502,NC_001503,NC_001506,NC_001511,NC_001514, NC_001549, NC_001550,NC_001574,NC_001618,NC_001634,NC_001648,NC_001654, NC_001702,NC_001722, NC_001724,NC_001725,NC_001739,NC_001802,NC_001815, NC_001831,NC_001839,NC_001866,NC_001867,NC_001885,NC_001914,NC_001940, NC_002201,NC_003031,NC_003059,NC_003138,NC_003323,NC_003378,NC_003381, NC_003382,NC_003498,NC_003554,NC_004036,NC_004324,NC_004450,NC_004455, NC_004540,NC_004994,NC_005947,NC_006934,NC_006955,NC_007002,NC_007003, NC_007015,NC_007654,NC_007815,NC_008017,NC_008018,NC_008034,NC_008094, NC_009010,NC_009424,NC_009889,NC_010737,NC_010738,NC_010820,NC_010955, NC_011097,NC_011546,NC_011592,NC_011800,NC_011920,NC_012728,NC_013262, NC_013455,NC_014474, NC_014648,NC_015116,NC_015228,NC_015328,NC_015502, NC_015503,NC_015504,NC_015505,NC_015506,NC_015507,NC_015655,NC_015784, NC_017830,NC_018105,NC_018505,NC_018616,NC_018858,NC_020999,NC_022365, NC_022517,NC_022518,NC_023153,NC_023485,NC_024301,NC_026020,NC_026238, NC_026472,NC_026819,NC_027117,NC_027131,NC_027924,NC_028462,NC_029303, NC_029852,NC_029853,NC_030205,NC_030462,NC_031326,NC_033738,NC_033739, NC_034252,NC_035472,NC_038378,NC_038379,NC_038380,NC_038381,NC_038382, NC_038512,NC_038669,NC_038858,NC_038922,NC_038923,NC_038986,NC_038987, NC_038995,NC_039022,NC_039023,NC_039024,NC_039025,NC_039026,NC_039027, NC_039028,NC_039029,NC_039030,NC_039031,NC_039085,NC_039228,NC_039238, NC_039242,NC_040461,NC_040462,NC_040552,NC_040622,NC_040635,NC_040692, NC_040693,NC_040712,NC_040807,NC_040808,NC_040809,NC_040841,NC_043194, NC_043195,NC_043382,NC_043445,NC_043491,NC_043523,NC_043534,NC_043535, U04327,U21247,U85505,U85506,U94692,X00255,X13744,X54482,X57540,Y07725,
Papillomaviridae
AB027020,AB027021,AB211993,AB331650,AB331651,AB361563,AB543507, AB793779,AF020905,AF092932,AF151983,AF293960,AF349909,AJ400628,AJ620205,

AJ620208,AJ620209,AJ620211,AY395706,AY904722,AY904723,AY904724,D21208,
D90252,D90400,DQ080079,DQ080080,DQ080082,DQ080083,DQ090005,DQ098913,
DQ098917,DQ180494,DQ344807,EF028290,EF117891,EF422221,EF558838,EF558839,
EF558840,EF558841,EF558842,EF558843,EF591300,EU240895,EU360723,EU410347,
EU410348,EU410349,EU490515,EU490516,EU493091,EU918769,FJ492742,FJ492743,
FJ947080,FM955837,FM955838,FM955839,FM955840,FM955841,FM955842,FN547152,
FN598907,FN677755,FN677756,FR751039,GQ180114,GQ227670,GQ244463,GQ246950,
GQ246951,GQ845441,GQ845442,GQ845444,GQ845445,GQ845446,GU117620,
GU117624,GU117629,GU117630,GU117633,GU129016,HE963025,HG939559,
HM999990,HM999991,HM999993,HM999994,HM999995,HM999997,HM999998,
HM999999,J04353,JF304766,JF304767,JF304768,JF800658,JF906559,JN171845,
JN709469,JN709470,JN709471,JN709472,JQ798171,JX174438,JX899359,KC138720,
KC470240,KC858265,KF006398,KF006400,KJ145795,KM085343,KM983393,KP276343,
L41216,M12732,M12737,M14119,M20219,M32305,M62877,M73236,M74117,
NC_001352,NC_001354,NC_001355,NC_001356,NC_001357,NC_001457,NC_001458,
NC_001522,NC_001523,NC_001524,NC_001526,NC_001531,NC_001541,NC_001576,
NC_001583,NC_001586,NC_001587,NC_001591,NC_001593,NC_001595,NC_001596,
NC_001605,NC_001619,NC_001676,NC_001678,NC_001690,NC_001691,NC_001693,
NC_001694,NC_001789,NC_002232,NC_003348,NC_003748,NC_003973,NC_004068,
NC_004104,NC_004195,NC_004197,NC_004500,NC_004765,NC_005134,NC_006563,
NC_006564,NC_006951,NC_007150,NC_007612,NC_008032,NC_008184,NC_008188,
NC_008189,NC_008297,NC_008298,NC_008519,NC_008582,NC_010107,NC_010226,
NC_010329,NC_010739,NC_010817,NC_011051,NC_011109,NC_011280,NC_011530,
NC_011765,NC_012123,NC_012213,NC_012485,NC_012486,NC_013035,NC_013117,
NC_013237,NC_014143,NC_014185,NC_014326,NC_014469,NC_014952,NC_014953,
NC_014954,NC_014955,NC_014956,NC_015267,NC_015268,NC_015325,NC_015691,
NC_015692,NC_016013,NC_016014,NC_016074,NC_016075,NC_016157,NC_016898,
NC_017716,NC_017862,NC_017993,NC_017994,NC_017995,NC_017996,NC_017997,
NC_018074,NC_018075,NC_018076,NC_018575,NC_019023,NC_019852,NC_020084,
NC_020085,NC_020500,NC_020501,NC_021472,NC_021483,NC_021930,NC_022095,
NC_022253,NC_022373,NC_022647,NC_022892,NC_023178,NC_023496,NC_023852,
NC_023873,NC_023882,NC_023891,NC_023894,NC_023895,NC_024300,NC_024893,
NC_026640,NC_026946,NC_027528,NC_027779,NC_028125,NC_028126,NC_028267,
NC_028492,NC_030151,NC_030795,NC_030796,NC_030797,NC_030798,NC_030799,
NC_030800,NC_030801,NC_030839,NC_031756,NC_033740,NC_033745,NC_033781,
NC_034616,NC_035193,NC_035199,NC_035201,NC_035208,NC_035478,NC_035479,
NC_037059,NC_037061,NC_037064,NC_037067,NC_037069,NC_038516,NC_038517,
NC_038518,NC_038519,NC_038520,NC_038521,NC_038522,NC_038523,NC_038524,
NC_038525,NC_038526,NC_038527,NC_038531,NC_038889,NC_038914,NC_039036,
NC_039037,NC_039038,NC_039039,NC_039040,NC_039041,NC_039042,NC_039086,
NC_039089,NC_040548,NC_040550,NC_040569,NC_040578,NC_040579,NC_040580,
NC_040583,NC_040604,NC_040619,NC_040620,NC_040640,NC_040655,NC_040688,
NC_040691,NC_040709,NC_040727,NC_040728,NC_040785,NC_040787,NC_040803,
NC_040804, NC_040805,NC_040806,NC_040818,NC_040827,U06714,U21941,U31778,

U31779,U31780,U31781, U31782,U31783,U31784,U31785,U31786,U31787,U31788, U31791,U31793,U31794,U37537,U83595,X05015,X05817,X55964,X55965,X70829, X74462,X74466,X74467,X74468,X74469,X74470,X74471, X74473,X74478,X74479, X74481,X74483,X77858
Parvoviridae
AF028704,AF028705,DQ196318,DQ196319,DQ335246,DQ778300,DQ898166,EF441262, EF584447, FJ375127,FJ375128,FJ375129,FJ440683,FJ441297,FJ445512,GQ368252, HM053694,JF429836,JN681175,JQ268283,JQ268284,JX645345,JX827169,KC580640, KF170373,KF214638,KF214640,KF214645,KJ486491,KJ634207,KM105951,KM390024, KM598414,KM598419,KP280068,M81888,NC_000883,NC_000936,NC_001401, NC_001510,NC_001539,NC_001540,NC_001662,NC_001701,NC_001718,NC_001729, NC_001829,NC_001899,NC_002077,NC_002190,NC_003346,NC_004284,NC_004285, NC_004286,NC_004287,NC_004288,NC_004289,NC_004290,NC_004295,NC_004442, NC_004828,NC_005040,NC_005041,NC_005341,NC_005889,NC_006147,NC_006148, NC_006152,NC_006259,NC_006260,NC_006261,NC_006263,NC_006555,NC_007018, NC_007218,NC_007455,NC_011317,NC_011545,NC_012042,NC_012564,NC_012636, NC_012685,NC_012729,NC_014357,NC_014358,NC_014468,NC_014665,NC_015115, NC_015718,NC_016031,NC_016032,NC_016647,NC_016744,NC_016752,NC_017823, NC_018399,NC_018450,NC_019492,NC_020499,NC_022089,NC_022104,NC_022564, NC_022748,NC_022800,NC_023020,NC_023673,NC_023842,NC_023860,NC_024452, NC_024453,NC_024454,NC_024888,NC_025825,NC_025891,NC_025965,NC_026251, NC_026815, NC_026943,NC_027429,NC_028136,NC_028650,NC_028973,NC_029133, NC_029300,NC_029797,NC_030296,NC_030402,NC_030837,NC_030873,NC_031450, NC_031670,NC_031695,NC_031751,NC_031959,NC_032097,NC_034445,NC_034532, NC_035180,NC_035185,NC_035186,NC_037053,NC_038532,NC_038533,NC_038534, NC_038535,NC_038536,NC_038537,NC_038538,NC_038539,NC_038540,NC_038541, NC_038542,NC_038543,NC_038544,NC_038545,NC_038546,NC_038547,NC_038883, NC_038895,NC_038898,NC_039043,NC_039044,NC_039045,NC_039046,NC_039047, NC_039048,NC_039049,NC_039050,NC_040533,NC_040562,NC_040603,NC_040623, NC_040626, NC_040652,NC_040671,NC_040672,NC_040694,NC_040695,NC_040843, NC_043446,U12469,X01457
Polydnaviridae
AY651828,AY651829,AY651830,DQ075354,DQ075355,DQ075356,DQ075357,DQ075358 ,DQ075359,DQ075360,EF067319,EF067320,EF067321,EF067322,EF067323,EF067324, EF067325,EF067326,EF067327,EF067328,EF067329,EF067330,EF067331,EF067332, NC_005165,NC_006633,NC_006634,NC_006635,NC_006636,NC_006637,NC_006638, NC_006639,NC_006640,NC_006641,NC_006642,NC_006643,NC_006644,NC_006645, NC_006646,NC_006647,NC_006648,NC_006649,NC_006650,NC_006651,NC_006652, NC_006653,NC_006654,NC_006655,NC_006656,NC_006657,NC_006658,NC_006659, NC_006660,NC_006661,NC_006662,NC_007028,NC_007029,NC_007030,NC_007031, NC_007032,NC_007033,NC_007034,NC_007035,NC_007036,NC_007037,NC_007038, NC_007039,NC_007040,NC_007041,NC_007044,NC_007985,NC_007986,NC_007987, NC_007988,NC_007989,NC_007990,NC_007991,NC_007992,NC_007993,NC_007994,

NC_007995,NC_007996,NC_007998,NC_007999,NC_008000,NC_008001,NC_008002,NC_008003,NC_008004,NC_008005,NC_008006,NC_008007,NC_008008,NC_008847,NC_008848,NC_008849,NC_008850,NC_008851,NC_008852,NC_008853,NC_008854,NC_008855,NC_008856,NC_008857,NC_008858,NC_008859,NC_008860,NC_008861,NC_008862,NC_008863,NC_008864,NC_008865,NC_008866,NC_008867,NC_008868,NC_008869,NC_008870,NC_008871,NC_008872,NC_008873,NC_008874,NC_008875,NC_008876,NC_008877,NC_008878,NC_008879,NC_008880,NC_008881,NC_008882,NC_008883,NC_008884,NC_008885,NC_008886,NC_008887,NC_008888,NC_008889,NC_008890,NC_008891,NC_008892,NC_008893,NC_008894,NC_008895,NC_008896,NC_008897,NC_008898,NC_008899,NC_008900,NC_008901,NC_008902,NC_008903,NC_008904,NC_008905,NC_008906,NC_008907,NC_008908,NC_008909,NC_008910,NC_008911,NC_008912,NC_008913,NC_008914,NC_008915,NC_008916,NC_008917,NC_008918,NC_008919,NC_008920,NC_008921,NC_008922,NC_008923,NC_008924,NC_008925,NC_008926,NC_008927,NC_008928,NC_008929,NC_008930,NC_008931,NC_008932,NC_008933,NC_008934,NC_008935,NC_008936,NC_008937,NC_008938,NC_008939,NC_008940,NC_008941,NC_008946,NC_008947,NC_008948,NC_008949,NC_008950,NC_008951,NC_008952,NC_008953,NC_008954,NC_008955,NC_008956,NC_008957,NC_008958,NC_008959,NC_008960,NC_008961,NC_008962,NC_008963,NC_008964,NC_008965,NC_008966,NC_008967,NC_008968,NC_008969,NC_008970,NC_008971,NC_008972,NC_008973,NC_008976,NC_008977,NC_008978,NC_008979,NC_008980,NC_008981,NC_008982,NC_008983,NC_008984,NC_008985,NC_008986,NC_008987,NC_008988,NC_008989,NC_008990,NC_008991,NC_008992,NC_008993,NC_008994,NC_008995,NC_008996,NC_008997,NC_008998,NC_008999,NC_009000,NC_009001,NC_009002,NC_009003,NC_043261,NC_043262,NC_043263,NC_043264,NC_043266,NC_043267,NC_043270,NC_043271,NC_043273,NC_043307,NC_043308,NC_043309,NC_043310,NC_043311,NC_043312,NC_043315,NC_043316,NC_043318,NC_043319,NC_043320,NC_043321,NC_043322,NC_043323,NC_043324,NC_043325,NC_043326,NC_043327,NC_043328,NC_043329,NC_043330,NC_043331,NC_043332,NC_043333,NC_043334,NC_043335,NC_043336,NC_043337,NC_043338,NC_043339,NC_043340,NC_043341,NC_043342,NC_043343,NC_043344,NC_043345,NC_043346,NC_043347,NC_043348,NC_043349,NC_043350,NC_043351,NC_043352,NC_043354,NC_043356,NC_043357,NC_043358,NC_043359,NC_043360,NC_043361,NC_043362
Polyomaviridae
AB767295,AF118150,DQ192570,DQ192571,EF127906,EF127907,EF127908,FR823284,HG764413, HQ385747,HQ385750,J02288,JX259273,JX262162,KJ577598,KM496323,KM496324,KM496325,M30540,NC_001442,NC_001505,NC_001515,NC_001538,NC_001663,NC_001669,NC_001699,NC_004763,NC_004764,NC_004800,NC_007611,NC_007922,NC_007923,NC_009238,NC_009539,NC_009951,NC_010277,NC_011310,NC_013439,NC_013796,NC_014361,NC_014406,NC_014407,NC_014743,NC_015150,NC_017085,NC_017982,NC_018102,NC_019844,NC_019850,NC_019851,NC_019853,NC_019854,NC_019855,NC_019856,NC_019857,NC_019858,NC_020065,NC_020066,NC_020067,NC_020068,NC_020069,NC_020070,NC_020071,NC_020106,NC_020890,NC_022519,NC_023008,NC_023845,NC_024118,NC_025259,NC_025368,NC_025370,

NC_025380,NC_025790,NC_025800,NC_025811,NC_025892,NC_025894,NC_025895,NC_025896,NC_025898,NC_025899,NC_026012,NC_026015,NC_026141,NC_026244,NC_026473,NC_026762,NC_026766,NC_026767,NC_026768,NC_026769,NC_026770,NC_026942,NC_026944,NC_027531,NC_027532,NC_028117,NC_028119,NC_028120,NC_028121,NC_028122,NC_028123,NC_028127,NC_028635,NC_030148,NC_030838,NC_031757,NC_032005,NC_032120,NC_033737,NC_034218,NC_034219,NC_034220,NC_034221,NC_034251,NC_034253,NC_034378,NC_034456,NC_035181,NC_038554,NC_038555, NC_038556,NC_038557,NC_038558,NC_038559,NC_039051,NC_039052,NC_039053,NC_040538,NC_040566,NC_040573,NC_040598,NC_040600,NC_040607,NC_040634,NC_040638,NC_040676,NC_040677,NC_040705,NC_040714,NC_040715,NC_040821,NC_040822
Riboviria
AB032553,AB042808,AB050936,AB073912,AB090161,AB187514,AB194796,AB205396,AB220921,AB252582,AB365435,AB426611,AB447427,AB447428,AB447429,AB447430,AB447431,AB447432,AB447433,AB447434,AB447435,AB447436,AB447437,AB447438,AB447439,AB447440,AB447441,AB447442,AB447443,AB447444,AB447445,AB447446,AB447447,AB447448,AB447449,AB447450,AB447451,AB447452,AB447453,AB447454,AB447455,AB447456,AB447457,AB447458,AB447459,AB447460,AB447461,AB447462,AB447463,AB541201,AB541202,AB541203,AB541204,AB541205,AB543808,AB558119,AB593690,AB614356,AB678778,AB690461,AB795432,AC_000192,AF002227,AF039205,AF046869,AF057136,AF059242,AF059243,AF070476,AF079457,AF081485,AF083069,AF086833,AF091605,AF091736,AF093797,AF103734,AF123432,AF123433,AF145896,AF162711,AF201929,AF227250,AF230973,AF241359,AF260508,AF274010,AF309418,AF311056,AF311938, AF311939,AF316321,AF326963,AF327920,AF327921,AF327922,AF338106,AF352027,AF361253, AF389115,AF389116,AF389117,AF389452,AF389453,AF389454,AF389455,AF389456,AF389462,AF389463,AF389464,AF389465,AF389466,AF407339,AF457102,AF524867,AF525933,AJ005695,AJ132961,AJ132997,AJ276479,AJ276480,AJ276481,AJ577589,AJ781401,AJ880277,AJ889866,AJ889867,AJ889868,AJ889918,AM113988,AM157175,AM235750,AM404308,AM498051,AM498052,AM498053,AM744987,AM744988,AM744989,AM744997,AM744998,AM744999,AM745007,AM745008,AM745009,AM745017,AM745018,AM745019,AM745027,AM745028,AM745029,AM745035,AM745037,AM745038,AM745039,AM745047,AM745048,AM745049,AM745057,AM745058,AM745059,AM745067,AM745068,AM745069,AM745077,AM745078,AM745079,AM910652,AY010722,AY032605,AY134748,AY260942,AY260943,AY260944,AY260949,AY260950,AY260951,AY278488,AY278491,AY278554,AY278741,AY297819,AY302539,AY302540,AY302541,AY302542,AY302543,AY302544,AY302545,AY302546,AY302547,AY302548,AY302549,AY302550,AY302551,AY302552,AY302553,AY302554,AY302555,AY302556,AY302557,AY302559,AY302560,AY350750,AY353550,AY357075,AY357076,AY394850,AY429470,AY462107,AY485642,AY486084,AY508697,AY515512,AY518894,AY554397,AY556057,AY556070,AY575773,AY588319,AY593765,AY593796,AY593805,AY593806,AY593808,AY593809,AY593840,AY593847,AY593851,AY646283,AY646511,AY685920,AY685921,AY686687,AY729016,AY741811,AY743910,AY751783,AY772730,AY773285,AY800279,AY842931,

AY843297,AY843298,AY843299,AY843300,AY843301,AY843302,AY843303,AY843304,AY843305,AY843306,AY843307,AY843308,AY859526,AY863002,AY864805,AY864806,AY876912,AY876913,AY898809,CY011117,CY011118,CY011119,CY011125,CY011126,CY011127,CY011133,CY011134,CY011135,CY011141,CY011142,CY011143,D00239,D00435,D00507, D00538,D00627,D00820,D13096,D90457,DQ011234,DQ011855,DQ028633,DQ058829,DQ070852,DQ217792,DQ238861,DQ256132,DQ256133,DQ256134,DQ286292,DQ294633,DQ315670,DQ328874,DQ328875,DQ358078,DQ369797,DQ399290,DQ412042,DQ412043,DQ447649,DQ447652,DQ447657,DQ456824,DQ473486,DQ473488,DQ473489,DQ473490,DQ473491,DQ473492,DQ473493,DQ473494,DQ473497,DQ473499,DQ473500,DQ473504,DQ473505,DQ473506,DQ473507,DQ473508,DQ473510,DQ473511,DQ480514,DQ640652,DQ648794,DQ648856,DQ648857,DQ658413,DQ811787,DQ812092,DQ812093,DQ848678,DQ851494,DQ902712,DQ902713,DQ911368,DQ915164,DQ995634,DQ995640,DQ995647,EF011023,EF014462,EF015886,EF017707,EF065505,EF065506,EF065507,EF065508,EF065509,EF065510,EF065511,EF065512,EF065513,EF065514,EF065515,EF065516,EF067923,EF067924,EF107097,EF108464,EF173414,EF173415,EF173420,EF173423,EF173425,EF424615,EF424616,EF424617,EF424618,EF424619,EF424620,EF424621,EF424622,EF424623,EF424624,EF424625,EF424626,EF424627,EF424628,EF424629,EF429197,EF429198,EF429199,EF429200,EF446132,EF446615,EF552688,EF552689,EF552690,EF552691,EF552692,EF552693,EF552694,EF552695,EF552696,EF552697,EF555644,EF555645,EF558545,EF667343,EF667344,EU004663,EU004664,EU004665,EU004666,EU004667,EU004668,EU004669,EU004670,EU004671,EU004672,EU004673,EU004674,EU004675,EU004676,EU004677,EU004678,EU004679,EU004680,EU004681,EU004682,EU004683,EU020009,EU037962,EU140838,EU143843,EU155216,EU155260,EU371559,EU371560,EU371561,EU371562,EU371563,EU371564,EU420137,EU420138,EU439428,EU563512,EU627591,EU716175,EU755009,EU779803,EU815052,EU854589,FJ009367,FJ355929,FJ355930,FJ376620,FJ387164,FJ415324,FJ425184,FJ425185,FJ425186,FJ425187,FJ425188,FJ425189,FJ434664,FJ445112,FJ445113,FJ445114,FJ445116,FJ445118,FJ445119,FJ445120,FJ445121,FJ445122,FJ445123,FJ445124,FJ445125,FJ445126,FJ445127,FJ445128,FJ445129,FJ445130,FJ445131,FJ445132,FJ445133,FJ445134,FJ445135,FJ445136,FJ445138,FJ445140,FJ445141,FJ445142,FJ445143,FJ445144,FJ445145,FJ445146,FJ445147,FJ445148,FJ445149,FJ445150,FJ445151,FJ445152,FJ445153,FJ445154,FJ445155,FJ445156,FJ445157
Test-2; Source: Virus-Host-DB
Betaflexiviridae
AF057136,NC_001946,NC_038324,NC_038325,NC_038966,NC_039087,NC_040545,NC_040554,NC_040564,NC_040568,NC_040616,NC_040627,NC_001948,NC_040630,NC_040643,NC_040689,NC_040703,NC_040797,NC_040800,NC_043081,NC_043082,NC_043086,NC_043087,NC_002468,NC_043088,NC_043412,NC_002500,NC_002552,NC_002729,NC_002795,NC_003462,NC_003499,NC_003557,AY646511,NC_003602,NC_003604,NC_003689,NC_003870,NC_003877,NC_005138,NC_005343,NC_006550,NC_006946,NC_007289,EU020009,NC_008020,NC_008266,NC_008292,NC_008552,NC_009087,NC_009383,NC_009759,NC_009764,NC_009892,NC_009991,FJ009367,

NC_010305,NC_010538,NC_011062,NC_011106,NC_011525,NC_011540,NC_011552, NC_012038,NC_012210,NC_012519,JF320811,NC_012869,NC_013006,NC_013527, NC_014730,NC_014821,NC_015220,NC_015395,NC_015782,NC_016080,NC_016404, JX559646,NC_016440,NC_017859,NC_018175,NC_018448,NC_018458,NC_018714, NC_019025,NC_019029,NC_019030,NC_020996,NC_001361,NC_023295,NC_023892, NC_024449,NC_024686,NC_025388,NC_025468,NC_025469,NC_026248,NC_026616, NC_027527,NC_001409,NC_028111,NC_028868,NC_028975,NC_029085,NC_029086, NC_029087,NC_029088,NC_029089,NC_029301,NC_030657,NC_001749,NC_030926, NC_031089,NC_034264,NC_034376,NC_034377,NC_034833,NC_035202,NC_035203, NC_035462, NC_037058
Bromoviridae
AJ276479,AJ276480,AJ276481,NC_001440,NC_001495,NC_002024,NC_002025, NC_002026,NC_002027,NC_002028,NC_002034,NC_002035,NC_002038,NC_002039, NC_002040,NC_003451,NC_003452,NC_003453,NC_003464,NC_003465,NC_003480, NC_003541,NC_003542,NC_003543,NC_003546,NC_003547,NC_003548,NC_003568, NC_003569,NC_003570,NC_003649,NC_003650,NC_003651,NC_003671,NC_003673, NC_003674,NC_003808,NC_003809,NC_003810,NC_003833,NC_003834,NC_003835, NC_003836,NC_003837,NC_003838,NC_003842,NC_003844,NC_003845,NC_004006, NC_004007,NC_004008,NC_004120,NC_004121,NC_004122,NC_004362,NC_004363, NC_005848,NC_005849,NC_005854,NC_006064,NC_006065,NC_006566,NC_006567, NC_006568,NC_006999,NC_007000,NC_007001,NC_008037,NC_008038,NC_008039, NC_008706,NC_008707,NC_008708,NC_009536,NC_009537,NC_009538,NC_011553, NC_011554,NC_011555,NC_011807,NC_011808,NC_011809,NC_012134,NC_012135, NC_012136,NC_013266,NC_013267,NC_013268,NC_018402,NC_018403,NC_018404, NC_022127,NC_022128,NC_022129,NC_022250,NC_022251,NC_022252,NC_025477, NC_025478,NC_025481,NC_025482,NC_025483,NC_025484,NC_027928,NC_027929, NC_027930,NC_038776,NC_038777,NC_039074,NC_039075,NC_039076,NC_040389, NC_040390,NC_040391,NC_040392,NC_040393,NC_040394,NC_040435,NC_040436, NC_040437,NC_040469,NC_040471
Caliciviridae
AB042808,AB187514,AB220921,AB365435,AB447427,AB447428,AB447429,AB447430, AB447431,AB447432,AB447433,AB447434,AB447435,AB447436,AB447437,AB447438, AB447439,AB447440,AB447441,AB447442,AB447443,AB447444,AB447445,AB447446, AB447447,AB447448,AB447449,AB447450,AB447451,AB447452,AB447453,AB447454, AB447455,AB447456,AB447457,AB447458,AB447459,AB447460,AB447461,AB447462, AB447463,AB541201,AB541202,AB541203,AB541204,AB541205,AB543808,AB614356, AF091736,AF093797,AF145896,AY032605,AY134748,AY485642,AY741811,AY772730, DQ058829,DQ369797,DQ456824,DQ658413,DQ911368,EF014462,EU004663,EU004664, EU004665,EU004666,EU004667,EU004668,EU004669,EU004670,EU004671,EU004672, EU004673,EU004674,EU004675,EU004676,EU004677,EU004678,EU004679,EU004680, EU004681,EU004682,EU004683,EU854589,FJ355929,FJ355930,FJ387164,FJ514242, FJ515294,FJ537135,FJ537136,FJ537137,FJ537138,GQ475301,GQ475302,GU594162, GU980585,GU991353,GU991354,GU991355,HF952119,HF952120,HF952121,HF952122,

HF952123,HF952124,HF952125,HF952126,HF952127,HF952128,HF952129,HF952130, HF952131,HF952132,HF952133,HF952134,HF952135,HM002617,HQ009513,HQ392821, HQ449728,HQ664990,JF320644,JF320645,JF320646,JF320647,JF320648,JF320649, JF320650,JF320651,JF320652,JF320653,JF781268,JN400599,JN400600,JN400601, JN400602,JN400603,JN400604,JN400605,JN400606,JN400607,JN400608,JN400609, JN400610,JN400611,JN400612,JN400613,JN400614,JN400615,JN400616,JN400617, JN400618,JN400619,JN400620,JN400621,JN400622,JN400623,JN400624,JN400625, JN400626,JN595867,JQ388274,JQ613567,JQ613568,JQ613569,JQ613570,JQ622197, JQ798158,JQ911594,JQ911595,JQ911596,JQ911597,JQ911598,JX018212,JX023285, JX023286,JX047864,JX126912,JX126913,JX439815,JX439816,JX439817,JX439818, JX439819,JX448566,JX459900,JX459901,JX459902,JX459903,JX459904,JX459905, JX459906,JX459907,JX459908,JX846924,JX846927,JX989073,JX989074,JX989075, JX993277,KC013592,KC175323,KC175342,KC175343,KC175344,KC175345,KC175346, KC175347,KC175348,KC175349,KC175350,KC175351,KC175352,KC175353,KC175354, KC175355,KC175356,KC175357,KC175358,KC175359,KC175360,KC175361,KC175362, KC175363,KC175364,KC175365,KC175366,KC175367,KC175368,KC175369,KC175370, KC175371,KC175372,KC175373,KC175374,KC175375,KC175376,KC175377,KC175378, KC175379,KC175380,KC175381,KC175382,KC175383,KC175384,KC175385,KC175386, KC175387,KC175388,KC175389,KC175390,KC175391,KC175392,KC175393,KC175394, KC175395,KC175396,KC175397,KC175398,KC175399,KC175400,KC175401,KC175402, KC175403,KC175404,KC175405,KC175406,KC175407,KC175408,KC175409,KC175410, KC409301,KC409302,KC463910,KC464496,KC464497,KC464498,KC464499,KC464500, KC577174,KC577175,KC631827,KC792553,KC894731,KC894942,KC894943,KC960615, KF204570,KF306212,KF306213,KF306214,KF429760,KF429761,KF429765,KF429766, KF429768,KF429770,KF429773,KF429774,KF429776,KF429777,KF429778,KF429783, KF429787,KF429789,KF429790,KF712491,KF712496,KF712497,KF712498,KF712499, KF712501,KF712502,KF712504,KF712510,KJ196276,KJ196277,KJ196278,KJ196279, KJ196280,KJ196281,KJ196282,KJ196283,KJ196284,KJ196285,KJ196286,KJ196287, KJ196288,KJ196289,KJ196293,KJ196294,KJ196295,KJ196296,KJ196297,KJ196298, KJ196299,KJ407072,KJ407073,KJ407074,KJ407075,KJ407076,KJ508818,KJ541743, KJ649705,KJ685403,KJ685405,KJ685408,KJ685412,KJ685413,KJ685414,KJ685415, KJ685417,KM272334,NC_000940,NC_001481,NC_001543,NC_001959,NC_002551, NC_002615,NC_004064,NC_004541,NC_004542,NC_006269,NC_006554,NC_006875, NC_007916,NC_008311,NC_008580,NC_010624,NC_011050,NC_011704,NC_012699, NC_017936,NC_019712,NC_024031,NC_024078,NC_025676,NC_027026,NC_027122, NC_029645,NC_029646,NC_029647,NC_030793,NC_031324,NC_033081,NC_033776, NC_034444,NC_035675,NC_039475,NC_039476,NC_039477,NC_039897,NC_040674, NC_040876,NC_043512,NC_043516,NC_044045,NC_044046,NC_044047,U15301, U54983,X86557
Coronaviridae
MG772933.1,MG772934.1,AC_000192,AF201929,AY278488,AY278491,AY278554, AY278741, AY350750,AY357075,AY357076,AY394850,AY515512,AY518894,AY646283, AY864805,AY864806,D13096,DQ011855,DQ412042,DQ412043,DQ640652,DQ648794,

DQ648856,DQ648857,DQ811787,DQ848678,DQ915164,EF065505,EF065506,EF065507,EF065508,EF065509,EF065510,EF065511,EF065512,EF065513,EF065514,EF065515,EF065516,EF424615,EF424616,EF424617,EF424618,EF424619,EF424620,EF424621,EF424622,EF424623,EF424624,EF446615,EU371559,EU371560,EU371561,EU371562,EU371563,EU371564,EU420137,EU420138,FJ376620,FJ415324,FJ425184,FJ425185,FJ425186,FJ425187,FJ425188,FJ425189,FJ647218,FJ647219,FJ647220,FJ647221,FJ647222,FJ647223,FJ647224,FJ647225,FJ647226,FJ647227,FJ882935,FJ882942,FJ882945,FJ882954,FJ882963,FJ884686,FJ938051,FJ938052,FJ938053,FJ938054,FJ938055,FJ938056,FJ938057,FJ938058,FJ938059,FJ938060,FJ938061,FJ938062,FJ938063,FJ938064,FJ938065,FJ938066,FJ938067,FN430414,FN430415,GQ153539,GQ153540,GQ153541,GQ153542,GQ153543,GQ153544,GQ153545,GQ153546,GQ153547,GQ153548,GU553361,GU553362,HM211098,HM211099,HM211100,HM211101,HM245926,HQ392469, HQ392470,HQ392471,HQ392472,JF705860,JF792616,JN183882,JN183883,JQ173883,JQ410000,JQ989272,JX169867,JX860640,JX869059,JX993987,JX993988,KC667074,KC776174,KC881005,KC881006,KF367457,KF569996,KF793824,KF906249,KJ473821,KJ481931,KJ567050,KJ601777,KJ601778,KJ601779,KJ601780,KJ769231,KM820765,KP981395,LM645057,LN610099,NC_001451,NC_001846,NC_002306,NC_002645,NC_003045,NC_003436,NC_004718,NC_005831,NC_006213,NC_006577,NC_009019,NC_009020,NC_009021,NC_009657,NC_009988,NC_010437,NC_010438,NC_010646,NC_010800,NC_011547,NC_011549,NC_011550,NC_012936,NC_014470,NC_016991, NC_016992,NC_016993,NC_016994,NC_016995,NC_016996,NC_017083,NC_018871,NC_019843,NC_022103,NC_023760,NC_025217,NC_026011,NC_028752,NC_028806,NC_028811,NC_028814,NC_028824,NC_028833,NC_030292,NC_030886,NC_032107,NC_032730,NC_034440,NC_034972,NC_035191,NC_038294,NC_038861,NC_039207,NC_039208,BetaCoV/bat/Yunnan/RaTG13/2013|EPI_ISL_402131
Flaviviridae
NC_027819,NC_027998,NC_027999,NC_028137,NC_028377,NC_029054,NC_029055,NC_030289,NC_030290,NC_030291,NC_030400,NC_030401,NC_030653,NC_030791,NC_031327,NC_031916,NC_031947,NC_031950,NC_032088,NC_033693,NC_033694,NC_033697,NC_033698,NC_033699,NC_033715,NC_033721,NC_033723,NC_033724,NC_033725,NC_033726,NC_034007,NC_034017,NC_034018,NC_034151,NC_034204,NC_034222,NC_034223,NC_034224,NC_034225,NC_034242,NC_034442,NC_035071,NC_035118,NC_035187,NC_035432,NC_035889,NC_038425,NC_038426,NC_038427,NC_038428,NC_038429,NC_038430,NC_038431,NC_038432,NC_038433,NC_038434,NC_038435,NC_038436,NC_038437,NC_038882,NC_038912,NC_038964,NC_039218,NC_039219, NC_039237,NC_040555,NC_040589,NC_040610,NC_040645,NC_040682,NC_040776,NC_040788,NC_040815,NC_043110,U70263,Z46258,AB690461,AB795432,AF002227,AF070476,AF091605,AF311056,AF326963,AF407339,AJ132997,AM404308,AM910652,AY554397,AY842931,AY859526,AY863002,AY898809,DQ480514,EF424625,EF424626,EF424627,EF424628,EF424629,EF429197,EF429198,EF429199,EF429200,EU155216,EU155260,FJ654700,GQ275355,HQ231763,JN704144,JN860200,JQ289550,JQ920421,JX196334,JX227952,JX227953,JX227954,JX227955,JX227958,JX227960,JX227962,JX227963,JX227965,JX227967,JX227970,JX227972,JX227979,JX477686,

KC815310,KC815311,KC990542,KF907503,KF917538,KJ469370,KJ660072,KM225263, KM225264,KM225265,KM408491,M91671,NC_000943,NC_001437,NC_001461, NC_001474,NC_001475,NC_001477,NC_001563,NC_001564,NC_001655,NC_001672, NC_001710,NC_001809,NC_001837,NC_002031,NC_002640,NC_002657,NC_003635, NC_003675,NC_003676,NC_003678,NC_003679,NC_003687,NC_003690,NC_003996, NC_004102,NC_004119,NC_005039,NC_005062,NC_005064,NC_006551,NC_007580, NC_008604,NC_008718,NC_008719,NC_009026,NC_009028,NC_009029,NC_009823, NC_009824,NC_009825,NC_009826,NC_009827,NC_009942,NC_012532,NC_012533, NC_012534,NC_012671,NC_012735,NC_012812,NC_012932,NC_015843,NC_016997, NC_017086,NC_018705,NC_018713,NC_020902,NC_021069,NC_021153,NC_021154, NC_023176,NC_023424,NC_023439,NC_024017,NC_024018,NC_024077,NC_024111, NC_024112,NC_024113,NC_024114,NC_024299,NC_024377,NC_024805,NC_024806, NC_024889,NC_025672,NC_025673,NC_025677,NC_025679,NC_026620,NC_026623, NC_026624,NC_026797,NC_027709,NC_027817
Peribunyaviridae
NC_001925,NC_001926,NC_004108,NC_004109,NC_005775,NC_005776,NC_009894, NC_009895,NC_018459,NC_018461,NC_018463,NC_018465,NC_018466,NC_018467, NC_018476,NC_018478,NC_021242,NC_021243,NC_022038,NC_022039,NC_022595, NC_022596,NC_024074,NC_024076,NC_026281,NC_026283,NC_026617,NC_026618, NC_026619,NC_027715,NC_027717,NC_031135,NC_031136,NC_031221,NC_031222, NC_031287,NC_031288,NC_031291,NC_031292,NC_034459,NC_034460,NC_034461, NC_034468,NC_034475,NC_034477,NC_034479,NC_034482,NC_034487,NC_034488, NC_034489,NC_034490,NC_034491,NC_034492,NC_034493,NC_034495,NC_034497, NC_034499,NC_034500,NC_034504,NC_034505,NC_034506,NC_034631,NC_034633, NC_038713,NC_038714,NC_038715,NC_038717,NC_038718,NC_038720,NC_038723, NC_038724,NC_038727,NC_038728,NC_038729,NC_038730,NC_038733,NC_038734, NC_038735,NC_038736,NC_038738,NC_038739,NC_038741,NC_038742,NC_038942, NC_039183,NC_039184,NC_039186,NC_039187,NC_043036,NC_043037,NC_043546, NC_043548,NC_043550,NC_043551,NC_043552,NC_043553,NC_043555,NC_043556, NC_043559,NC_043560,NC_043561,NC_043563,NC_043564,NC_043565,NC_043567, NC_043568,NC_043570,NC_043571,NC_043573,NC_043575,NC_043577,NC_043578, NC_043579,NC_043580,NC_043583,NC_043584,NC_043586,NC_043587,NC_043588, NC_043589,NC_043591,NC_043592,NC_043594,NC_043595,NC_043597,NC_043599, NC_043600,NC_043602,NC_043603,NC_043605,NC_043607,NC_043608,NC_043612, NC_043614,NC_043615,NC_043617,NC_043618,NC_043619,NC_043621,NC_043623, NC_043627,NC_043629,NC_043630,NC_043632,NC_043633,NC_043634,NC_043637, NC_043638,NC_043639,NC_043641,NC_043645,NC_043646,NC_043651,NC_043652, NC_043653,NC_043655,NC_043674,NC_043675,NC_043687,NC_043688,NC_043690, NC_043691,NC_043692,NC_043694,NC_043697,NC_043699
Phenuiviridae
NC_002323,NC_002324,NC_002325,NC_002326,NC_002327,NC_002328,NC_003753, NC_003754,NC_003755,NC_003776,NC_005214,NC_005220,NC_006319,NC_006320, NC_014396,NC_014397,NC_015373,NC_015374,NC_015411,NC_015412,NC_015450,

NC_015451,NC_018136,NC_018138,NC_022630,NC_022631,NC_023633,NC_023635,NC_024494,NC_024495,NC_027140,NC_027141,NC_029082,NC_029127,NC_029128,NC_029901,NC_029903,NC_031138,NC_031139,NC_031295,NC_031298,NC_031313,NC_031316,NC_031317,NC_031318,NC_031320,NC_031321,NC_032158,NC_032159,NC_032257,NC_032276,NC_032277,NC_032278,NC_032280,NC_032282,NC_033830,NC_033835,NC_033836,NC_033838,NC_033840,NC_033841,NC_033842,NC_033844,NC_033846,NC_033847,NC_033848,NC_033849,NC_036597,NC_036598,NC_036602,NC_036604,NC_036605,NC_037612,NC_037614,NC_037616,NC_038257,NC_038258,NC_038261,NC_038262,NC_038748,NC_038750,NC_038751,NC_038752,NC_038754,NC_038757,NC_038934,NC_039191,NC_039192,NC_040450,NC_040493,NC_040494,NC_043045,NC_043046,NC_043049,NC_043051,NC_043450,NC_043451,NC_043477,NC_043481,NC_043482,NC_043509,NC_043510,NC_043609,NC_043611,NC_043679,NC_043680,X89628
Picornaviridae
AB090161,AB205396,AB252582,AB426611,AB678778,AF039205,AF081485,AF083069,AF123432,AF123433,AF162711,AF230973,AF241359,AF274010,AF311938,AF311939,AF316321,AF327920,AF327921,AF327922,AF352027,AF361253,AF524867,AJ005695,AJ132961,AJ577589,AJ889918,AM235750,AY302539,AY302540,AY302541,AY302542,AY302543,AY302544,AY302545,AY302546,AY302547,AY302548,AY302549,AY302550,AY302551,AY302552,AY302553,AY302554,AY302555,AY302556,AY302557,AY302559,AY302560,AY429470,AY462107,AY508697,AY556057,AY556070,AY593765,AY593796,AY593805,AY593806,AY593808,AY593809,AY593840,AY593847,AY593851,AY686687,AY751783,AY773285,AY843297,AY843298,AY843299,AY843300,AY843301,AY843302,AY843303,AY843304,AY843305,AY843306,AY843307,AY843308,AY876912,AY876913,D00239,D00435,D00538,D00627,D00820,D90457,DQ256132,DQ256133,DQ256134,DQ294633,DQ315670,DQ358078,DQ473486,DQ473488,DQ473489,DQ473490,DQ473491,DQ473492,DQ473493,DQ473494,DQ473497,DQ473499,DQ473500,DQ473504,DQ473505,DQ473506,DQ473507,DQ473508,DQ473510,DQ473511,DQ812092,DQ812093,DQ902712,DQ902713,DQ995634,DQ995640,DQ995647,EF015886,EF067923,EF067924,EF107097,EF173414,EF173415,EF173420,EF173423,EF173425,EF552688,EF552689,EF552690,EF552691,EF552692,EF552693,EF552694,EF552695,EF552696,EF552697,EF555644,EF555645,EF667343,EF667344,EU140838,EU716175,EU755009,EU815052,FJ445112,FJ445113,FJ445114,FJ445116,FJ445118,FJ445119,FJ445120,FJ445121,FJ445122,FJ445123,FJ445124,FJ445125,FJ445126,FJ445127,FJ445128,FJ445129,FJ445130,FJ445131,FJ445132,FJ445133,FJ445134,FJ445135,FJ445136,FJ445138,FJ445140,FJ445141,FJ445142,FJ445143,FJ445144,FJ445145,FJ445146,FJ445147,FJ445148,FJ445149,FJ445150,FJ445151,FJ445152,FJ445153,FJ445154,FJ445155,FJ445156,FJ445157,FJ445160,FJ445161,FJ445162,FJ445163,FJ445164,FJ445165,FJ445167,FJ445168,FJ445169,FJ445170,FJ445171,FJ445172,FJ445173,FJ445174,FJ445175,FJ445176,FJ445178,FJ445179,FJ445180,FJ445181,FJ445182,FJ445183,FJ445185,FJ445186,FJ445187,FJ445188,FJ445189,FJ445190,FM955278,GQ122332,GQ249161,GQ323774,GQ485310,GQ485311,GQ865517,HM185056,HM777023,HQ400942,HQ654774,HQ702854,HQ728260,HQ728261,HQ728262,

HQ875059,JF905564,JN088541,JN379039,JN710381,JQ277724,JQ814852,JQ818253,JQ898342,JQ911763,JQ975417,JX050181,JX174177,JX262382,JX491648,JX961709,JX982257,KC663628,KC811837,KF312882,KF422142,KF831027,KF874626,KF958308,KF990476,KJ857508,KM203656,KM609480,KP036483,L24917,LK021688,M12197,M16560,M20301,NC_001366,NC_001430,NC_001472,NC_001479,NC_001489,NC_001490,NC_001612,NC_001617,NC_001859,NC_001897,NC_001918,NC_002058,NC_003976,NC_003983,NC_003985,NC_003987,NC_003988,NC_003990,NC_004421,NC_004441,NC_004451,NC_006553,NC_008250,NC_008714,NC_009448,NC_009891,NC_009996,NC_010354,NC_010415,NC_010810,NC_011349,NC_011829,NC_012798,NC_012800,NC_012801,NC_012802,NC_012957,NC_012986,NC_013695,NC_014411,NC_014412,NC_014413,NC_015626,NC_015934,NC_015936,NC_015940,NC_015941,NC_016156,NC_016403,NC_016769,NC_018226,NC_018400,NC_018506,NC_018668,NC_021178,NC_021201,NC_021220,NC_021482,NC_022332,NC_022802,NC_023162,NC_023422,NC_023858,NC_023861,NC_023984,NC_023985,NC_023987,NC_023988,NC_024070,NC_024073,NC_024120,NC_024765,NC_024766,NC_024767,NC_024768,NC_024769,NC_024770,NC_025114,NC_025432,NC_025474,NC_025675,NC_025890,NC_025961,NC_026249,NC_026314,NC_026315,NC_026316,NC_026470,NC_026921,NC_027054,NC_027214,NC_027818,NC_027918,NC_027919,NC_028363,NC_028364,NC_028365,NC_028366,NC_028380,NC_028964,NC_028970,NC_028981,NC_029854,NC_029905,NC_030454,NC_030843,NC_031105,NC_031106,NC_032126,NC_033695,NC_033793,NC_033818,NC_033819,NC_033820,NC_034206,NC_034245,NC_034267,NC_034381,NC_034385,NC_034453,NC_034617,NC_034971,NC_035110,NC_035198,NC_035779,NC_036588,NC_037654,NC_038303,NC_038304,NC_038305,NC_038306,NC_038307,NC_038308,NC_038309,NC_038310,NC_038311,NC_038312,NC_038313,NC_038314,NC_038315,NC_038316,NC_038317,NC_038318,NC_038319,NC_038878,NC_038880,NC_038957,NC_038961,NC_038989,NC_039004,NC_039209,NC_039210,NC_039211,NC_039212,NC_039235,NC_040605,NC_040611,NC_040642,NC_040673,NC_040684,NC_043071,NC_043072,NC_043544,U05876,U16283,U22521,V01149,X00925,X05690, X56019,X67706,X77708,X84981,X92886
Potyviridae
AB194796,AJ889866,AJ889867,AJ889868,AM113988,AM157175,AY010722,AY575773,D00507,DQ851494,EF017707,EF558545,EU563512,HE608963,HE608964,HF585099,HF585103,HM590055,JQ924285,JQ924286,NC_000947,NC_001445,NC_001517,NC_001555,NC_001616,NC_001671,NC_001768,NC_001785,NC_001814,NC_001841,NC_001886,NC_002349,NC_002350,NC_002509,NC_002600,NC_002634,NC_002990,NC_002991,NC_003224,NC_003377,NC_003397,NC_003398,NC_003399,NC_003482,NC_003483,NC_003492,NC_003501,NC_003536,NC_003537,NC_003605,NC_003606,NC_003742,NC_003797,NC_004010,NC_004011,NC_004013,NC_004016,NC_004017,NC_004035,NC_004039,NC_004047,NC_004426,NC_004573,NC_004752,NC_005028,NC_005029,NC_005136,NC_005288,NC_005304,NC_005778,NC_005903,NC_005904,NC_006262,NC_006941,NC_007147,NC_007180,NC_007216,NC_007433,NC_007728,NC_007913,NC_008028,NC_008393,NC_008558,NC_008824,NC_009741,NC_009742,NC_009744,NC_009745,NC_009805,NC_009994,NC_009995,NC_010521,NC_010735,

NC_010736,NC_010954,NC_011541,NC_011560,NC_011918,NC_012698,NC_012799,NC_013261,NC_014037,NC_014038,NC_014064,NC_014252,NC_014325,NC_014327,NC_014536,NC_014742,NC_014790,NC_014791,NC_014898,NC_014905,NC_015393,NC_015394,NC_016044,NC_016159,NC_016441,NC_017824,NC_017967,NC_017970,NC_017977,NC_018093,NC_018176,NC_018455,NC_018572,NC_018833,NC_018872,NC_019031,NC_019409,NC_019412,NC_019415,NC_020072,NC_020105,NC_020896,NC_021065,NC_021197,NC_021786,NC_022745,NC_023014,NC_023175,NC_023628,NC_024471,NC_025250,NC_025254,NC_025821,NC_026615,NC_026759,NC_027210,NC_027706,NC_028144,NC_028145,NC_029051,NC_029076,NC_030118,NC_030236,NC_030293,NC_030391,NC_030794,NC_030840,NC_030847,NC_031339,NC_032912,NC_034208,NC_034273,NC_034835,NC_035134,NC_035458,NC_035459,NC_035461,NC_036802,NC_037051,NC_038560,NC_038561,NC_038562,NC_038920,NC_038984,NC_039002,NC_039088,NC_040507,NC_040508,NC_040650,NC_040802,NC_040836,NC_043133,NC_043141,NC_043149,NC_043165,NC_043168,NC_043171,NC_043172,NC_043424,NC_043532,NC_043536, NC_043537,U05771
Reoviridae
AF389452,AF389453,AF389454,AF389455,AF389456,AF389462,AF389463,AF389464,AF389465,AF389466,AM498051,AM498052,AM498053,AM744987,AM744988,AM744989,AM744997,AM744998,AM744999,AM745007,AM745008,AM745009,AM745017,AM745018,AM745019,AM745027,AM745028,AM745029,AM745035,AM745037,AM745038,AM745039,AM745047,AM745048,AM745049,AM745057,AM745058,AM745059,AM745067,AM745068,AM745069,AM745077,AM745078,AM745079,FN563984,HG513046,NC_002557,NC_002558,NC_002559,NC_002560,NC_002567,NC_003006,NC_003007,NC_003008,NC_003009,NC_003010,NC_003016,NC_003017,NC_003018,NC_003019,NC_003020,NC_003654,NC_003655,NC_003656,NC_003657,NC_003658,NC_003659,NC_003696,NC_003697,NC_003698,NC_003699,NC_003700,NC_003701,NC_003702,NC_003703,NC_003728,NC_003729,NC_003730,NC_003734,NC_003735,NC_003736,NC_003737,NC_003749,NC_003750,NC_003751,NC_003752,NC_003759,NC_003761,NC_003762,NC_003771,NC_003772,NC_003773,NC_003774,NC_004181,NC_004182,NC_004183,NC_004184,NC_004185,NC_004186,NC_004187,NC_004188,NC_004210,NC_004211,NC_004212,NC_004213,NC_004214,NC_004217,NC_004218,NC_004219,NC_005166,NC_005167,NC_005168,NC_005169,NC_005170,NC_005171,NC_005986,NC_005989,NC_005990,NC_005996,NC_005997,NC_005998,NC_005999,NC_006000,NC_006013,NC_006014,NC_006017,NC_006021,NC_006023,NC_007154,NC_007155,NC_007157,NC_007158,NC_007159,NC_007160,NC_007163,NC_007524,NC_007525,NC_007533,NC_007534,NC_007535,NC_007536,NC_007546,NC_007547,NC_007548,NC_007549,NC_007550,NC_007551,NC_007559,NC_007560,NC_007561,NC_007562,NC_007563,NC_007572,NC_007574,NC_007582,NC_007583,NC_007584,NC_007586,NC_007592,NC_007656,NC_007657,NC_007658,NC_007666,NC_007667,NC_007668,NC_007669,NC_007670,NC_007736,NC_007737,NC_007738,NC_007739,NC_007748,NC_007749,NC_007750,NC_008171,NC_008172,NC_008173,NC_008174,NC_008175,NC_008729,NC_008730,NC_008731,NC_008732,NC_008733,NC_008735,NC_008736,NC_009243,NC_009244,NC_009247,NC_009248,

NC_009249,NC_010584,NC_010585,NC_010586,NC_010587,NC_010588,NC_010589,NC_010666,NC_010667,NC_010668,NC_010669,NC_010670,NC_010743,NC_010744,NC_010745,NC_010746,NC_010747,NC_010748,NC_011506,NC_011507,NC_011508,NC_011510,NC_012535,NC_012536,NC_012537,NC_012538,NC_012539,NC_012754,NC_012755,NC_013225,NC_013226,NC_013227,NC_013228,NC_013229,NC_013230,NC_013396,NC_013397,NC_013398,NC_014236,NC_014237,NC_014238,NC_014239,NC_014240,NC_014241,NC_014511,NC_014512,NC_014513,NC_014514,NC_014522,NC_014523,NC_014598,NC_014599,NC_014600,NC_014601,NC_014602,NC_014708,NC_014709,NC_014710,NC_014714,NC_014715,NC_014716,NC_014717,NC_015126,NC_015127,NC_015128,NC_015129,NC_015130,NC_015877,NC_015878,NC_015879,NC_015880,NC_015881,NC_016874,NC_016875,NC_016876,NC_016879,NC_016880,NC_016881,NC_020439,NC_020440,NC_020441,NC_020442,NC_020447,NC_021541,NC_021543,NC_021545,NC_021551,NC_021580,NC_021581,NC_021589,NC_021590,NC_021625,NC_021626,NC_021630,NC_021631,NC_022553,NC_022554,NC_022555,NC_022620,NC_022626,NC_022627,NC_022633,NC_022634,NC_022639,NC_023420,NC_023486,NC_023487,NC_023488,NC_023491,NC_023492,NC_023813,NC_023814,NC_023815,NC_023816,NC_023819,NC_023820,NC_024503,NC_024504,NC_024505,NC_024506,NC_024507,NC_024916,NC_024917,NC_024918,NC_024919,NC_025485,NC_025486,NC_025487,NC_025488,NC_025493,NC_025801,NC_025802,NC_025803,NC_025804,NC_025808,NC_025845,NC_025846,NC_025847,NC_025848,NC_025849,NC_025850,NC_025851,NC_026825,NC_026826,NC_026827,NC_026828,NC_027533,NC_027534,NC_027535,NC_027539,NC_027553,NC_027554,NC_027567,NC_027568,NC_027569,NC_027572,NC_027574,NC_027803,NC_027808,NC_027809,NC_027811,NC_027812,NC_027816,NC_028465,NC_029904,NC_029911,NC_029912,NC_029913,NC_029914,NC_029917,NC_029918,NC_030158,NC_030159,NC_030160,NC_030161,NC_030162,NC_030163,NC_030405,NC_030406,NC_030412,NC_030413,NC_030414,NC_030415,NC_033782,NC_033783,NC_033784,NC_034168,NC_034169,NC_034170,NC_034171,NC_034172,NC_035935,NC_035936,NC_036468,NC_036469,NC_036470,NC_036471,NC_036476,NC_036477,NC_037570,NC_037571,NC_037572,NC_037573,NC_037574,NC_037578,NC_037579,NC_037580,NC_037581,NC_037582,NC_037583,NC_038564,NC_038565,NC_038568,NC_038570,NC_038574,NC_038575,NC_038582,NC_038584,NC_038588,NC_038592,NC_038594,NC_038595,NC_038600,NC_038604,NC_038605,NC_038610,NC_038614,NC_038615,NC_038620,NC_038624,NC_038625,NC_038629,NC_038630,NC_038634,NC_038635,NC_038636,NC_038637,NC_038640,NC_038641,NC_038645,NC_038648,NC_038649,NC_038652,NC_038657,NC_038660,NC_038661,NC_038662,NC_038664,NC_038665,NC_038945,NC_038948,NC_040408,NC_040409,NC_040413,NC_040414,NC_040440,NC_040443,NC_040444,NC_040445,NC_040447,NC_040472,NC_040473,NC_040476,NC_040478,NC_040479,NC_040499,NC_040501,NC_040502,NC_040503,NC_040504,NC_040506,NC_043180,NC_043182,NC_043183,NC_043184,NC_043185,NC_043190,NC_043368,NC_043369, NC_043370
Rhabdoviridae
KC519324,KC685626,KP688373,NC_000855,NC_000903,NC_001542,NC_001560,NC_001615,NC_001652,NC_002251,NC_002526,NC_002803,NC_003243,NC_003746,

NC_005093,NC_005974,NC_005975,NC_006429,NC_006942,NC_007020,NC_007642, NC_008514,NC_009527,NC_009528,NC_009608,NC_009609,NC_011532,NC_011558, NC_011568,NC_011639,NC_013135,NC_013955,NC_016136,NC_017685,NC_017714, NC_018381,NC_018629,NC_020803,NC_020804,NC_020805,NC_020806,NC_020807, NC_020808,NC_020809,NC_020810,NC_022580,NC_022581,NC_022755,NC_024473, NC_025251,NC_025253,NC_025255,NC_025340,NC_025341,NC_025342,NC_025353, NC_025356,NC_025358,NC_025359,NC_025362,NC_025364,NC_025365,NC_025371, NC_025376,NC_025377,NC_025378,NC_025382,NC_025384,NC_025385,NC_025387, NC_025389,NC_025391,NC_025393,NC_025394,NC_025395,NC_025396,NC_025397, NC_025399,NC_025400,NC_025401,NC_025405,NC_025406,NC_025408,NC_028230, NC_028231,NC_028232,NC_028234,NC_028236,NC_028237,NC_028239,NC_028241, NC_028244,NC_028246,NC_028255,NC_028266,NC_028867,NC_030451,NC_031079, NC_031083,NC_031093,NC_031215,NC_031216,NC_031225,NC_031227,NC_031232, NC_031236,NC_031240,NC_031268,NC_031272,NC_031273,NC_031276,NC_031278, NC_031282,NC_031283,NC_031301,NC_031303,NC_031305,NC_031690,NC_031691, NC_031955,NC_031957,NC_031958,NC_031988,NC_033701,NC_033705,NC_034240, NC_034443,NC_034447,NC_034448,NC_034449,NC_034450,NC_034451,NC_034454, NC_034508,NC_034529,NC_034530,NC_034531,NC_034533,NC_034534,NC_034535, NC_034536,NC_034537,NC_034538,NC_034539,NC_034540,NC_034541,NC_034542, NC_034543,NC_034544,NC_034545,NC_034546,NC_034548,NC_034549,NC_034550, NC_034551,NC_036390,NC_038236,NC_038275,NC_038276,NC_038277,NC_038278, NC_038279,NC_038280,NC_038281,NC_038282,NC_038283,NC_038284,NC_038285, NC_038286,NC_038287,NC_038755,NC_038756,NC_039020,NC_039021,NC_039200, NC_039201,NC_039202,NC_039206,NC_040532,NC_040599,NC_040602,NC_040664, NC_040669,NC_040786,NC_043065,NC_043066,NC_043067,NC_043525,NC_043538, NC_043648,NC_043649,Z93414
Secoviridae
NC_001632,NC_003003,NC_003004,NC_003445,NC_003446,NC_003495,NC_003496, NC_003502,NC_003509,NC_003544,NC_003545,NC_003549,NC_003550,NC_003615, NC_003621,NC_003622,NC_003623,NC_003626,NC_003628,NC_003693,NC_003694, NC_003738,NC_003741,NC_003785,NC_003786,NC_003787,NC_003788,NC_003791, NC_003792,NC_003799,NC_003800,NC_003839,NC_003840,NC_004439,NC_004440, NC_005096,NC_005097,NC_005266,NC_005267,NC_005289,NC_005290,NC_006056, NC_006057,NC_006271,NC_006272,NC_006964,NC_006965,NC_008182,NC_008183, NC_009013,NC_009032,NC_010709,NC_010710,NC_010987,NC_010988,NC_011189, NC_011190,NC_013075,NC_013076,NC_013218,NC_013219,NC_015414,NC_015415, NC_015492,NC_015493,NC_016443,NC_016444,NC_017938,NC_017939,NC_018383, NC_018384,NC_020897,NC_020898,NC_022004,NC_022006,NC_022798,NC_022799, NC_023016,NC_023017,NC_025479,NC_025480,NC_027915,NC_027926,NC_027927, NC_028139,NC_028146,NC_029036,NC_029038,NC_031763,NC_031766,NC_032270, NC_032271,NC_033492,NC_033493,NC_034214,NC_034215,NC_035214,NC_035215, NC_035218,NC_035219,NC_035220,NC_035221,NC_038320,NC_038744,NC_038759, NC_038760,NC_038761,NC_038762,NC_038763,NC_038764,NC_038765,NC_038766,

NC_038767,NC_038768,NC_038862,NC_038863,NC_039072,NC_039073,NC_039077,NC_039078,NC_040399,NC_040400,NC_040416,NC_040417,NC_040586,NC_043076,NC_043385,NC_043388, NC_043411,NC_043447,NC_043448,NC_043684,NC_043685
Test-3a; Source: Virus-Host-DB; NCBI; GISAID
Alphacoronavirus
AY518894,FJ938054,FJ938055,FJ938056,FJ938057,FJ938058,FJ938059,FJ938060,FJ938061,FJ938062,GU553361,D13096,GU553362,HM245926,HQ392469,HQ392470,HQ392471,HQ392472,JN183882,JN183883,JQ410000,JQ989272,DQ811787,LM645057,NC_002306,NC_002645,NC_003436,NC_005831,NC_009657,NC_009988,NC_010437,NC_010438,NC_018871,DQ848678,NC_022103,NC_023760,NC_028752,NC_028806,NC_028811,NC_028814,NC_028824,NC_028833,NC_030292,NC_032107,EU420137,NC_032730,NC_034972,NC_035191,NC_038861,EU420138,FJ938051,FJ938052,FJ938053
Betacoronavirus
EU371561,EU371562,EU371563,EU371564,FJ415324,FJ425184,FJ425185,FJ425186,FJ425187,FJ425188,FJ425189,FJ647218,FJ647219,FJ647220,FJ647221,FJ647222,FJ647223,FJ647224,FJ647225,FJ647226,FJ647227,FJ882935,FJ882942,FJ882945,FJ882954,FJ882963,FJ884686,FJ938063,FJ938064,FJ938065,FJ938066,FJ938067,GQ153539,GQ153540,GQ153541,GQ153542,GQ153543,GQ153544,GQ153545,GQ153546,GQ153547,GQ153548,HM211098,HM211099,HM211100,HM211101,JF792616,JQ173883,JX169867,JX860640,JX869059,JX993987,JX993988,KC667074,KC776174,KC881005,KC881006,KF367457,KF569996,KF906249,KJ473821,NC_001846,NC_003045,NC_004718,NC_006213,NC_006577,NC_009019,NC_009020,NC_009021,NC_012936,NC_017083,NC_019843,NC_025217,NC_026011,NC_030886,NC_038294,NC_039207,AC_000192,AF201929,AY278488,AY278491,AY278554,AY278741,AY350750,AY357075,AY357076,AY394850,AY515512,AY864805,AY864806,DQ011855,DQ412042,DQ412043,DQ640652,DQ648794,DQ648856,DQ648857,DQ915164,EF065505,EF065506,EF065507,EF065508,EF065509,EF065510,EF065511,EF065512,EF065513,EF065514,EF065515,EF065516,EF424615,EF424616,EF424617,EF424618,EF424619,EF424620,EF424621,EF424622,EF424623,EF424624,EF446615,EU371559,EU371560,MG772933.1,MG772934.1,EPI_ISL_402131
Deltacoronavirus
FJ376620,KJ481931,KJ567050,KJ601777,KJ601778,KJ601779,KJ601780,KJ769231,KM820765,KP981395,NC_011547,NC_011549,NC_011550,NC_016991,NC_016992,NC_016993,NC_016994,NC_016995,NC_016996,NC_039208
Gammacoronavirus
AY646283,FN430414,FN430415,JF705860,KF793824,LN610099,NC_001451,NC_010646,NC_010800
Test-3b; Source: Virus-Host-DB; GISAID
Alphacoronavirus
JQ989272,JQ410000,DQ811787,FJ938058,NC_022103,NC_028752,EU420137,NC_038861,FJ938051, FJ938056,FJ938059,NC_034972,NC_028811,HQ392471,FJ938057,NC_028824,NC_028814,FJ938060,HM245926,NC_028833

Betacoronavirus
EF065513,FJ882942,FJ425185,HM211100,GQ153540,NC_006213,GQ153543,EF424624,FJ647220,FJ938065,EPI_ISL_402131,FJ938066,AY278554,DQ915164,DQ011855,FJ882945,FJ647225,FJ425184, FJ415324,FJ882935
Deltacoronavirus
FJ376620,KJ481931,KJ567050,KJ601777,KJ601778,KJ601779,KJ601780,KJ769231,KM820765,KP981395,NC_011547,NC_011549,NC_011550,NC_016991,NC_016992,NC_016993,NC_016994,NC_016995,NC_016996,NC_039208
Test-4; Source: Virus-Host-DB; NCBI; GISAID
Embecovirus
AC_000192,EF424620,EF424621,EF424622,EF424623,EF424624,EF446615,FJ415324,FJ425184,FJ425185,FJ425186,AF201929,FJ425187,FJ425188,FJ425189,FJ647218,FJ647219,FJ647220,FJ647221,FJ647222,FJ647223,FJ647224,DQ011855,FJ647225,FJ647226,FJ647227,FJ884686,FJ938063,FJ938064,FJ938065,FJ938066,FJ938067,JF792616,DQ915164,JQ173883,JX169867,JX860640,KF906249,NC_001846,NC_003045,NC_006213,NC_006577,NC_012936,NC_026011,EF424615,EF424616,EF424617,EF424618, EF424619
Merbecovirus
DQ648794,EF065505,EF065506,EF065507,EF065508,EF065509,EF065510,EF065511,EF065512,JX869059,KC667074,KC776174,KJ473821,NC_009019,NC_009020,NC_019843,NC_038294,NC_039207
Nobecovirus
EF065513,EF065514,EF065515,EF065516,HM211098,HM211099,HM211100,HM211101,NC_009021,NC_030886
Sarbecovirus
MG772933.1,MG772934.1,EPI_ISL_402131,FJ882935,FJ882942,FJ882945,FJ882954,FJ882963,GQ153539,GQ153540,GQ153541,GQ153542,GQ153543,GQ153544,GQ153545,GQ153546,GQ153547,GQ153548,JX993987,JX993988,KC881005,KC881006,KF367457,KF569996,NC_004718,AY278488,AY278491,AY278554,AY278741,AY350750,AY357075,AY357076,AY394850,AY515512,AY864805,AY864806,DQ412042,DQ412043,DQ640652,DQ648856,DQ648857,EU371559,EU371560,EU371561,EU371562,EU371563,EU371564
Test-5; Source: Virus-Host-DB; NCBI; GISAID
Embecovirus; Merbecovirus; Nobecovirus; Sarbecovirus: same as Test-4
2019-nCoV
EPI_ISL_402119,EPI_ISL_402130,EPI_ISL_402132,EPI_ISL_403928,EPI_ISL_403929,EPI_ISL_403930,EPI_ISL_403931,EPI_ISL_403932,EPI_ISL_403933,EPI_ISL_403934,EPI_ISL_403935,EPI_ISL_402120,EPI_ISL_403936,EPI_ISL_403937,EPI_ISL_403962,EPI_ISL_403963,EPI_ISL_404227,EPI_ISL_404228,MN908947.3,EPI_ISL_402121,EPI_ISL_402123,EPI_ISL_402124,EPI_ISL_402125,EPI_ISL_402127,EPI_ISL_402128,EPI_ISL_402129,EPI_ISL_404253,EPI_ISL_404895,EPI_ISL_405839
Test-6; Source: Virus-Host-DB; NCBI; GISAID
Sarbecovirus; 2019-nCoV: same as Test-5

Supplementary Table D.S3: Accession IDs of sequences used in Test-1 to Test-6.

Appendix E

Addendum

Alignment-based methods require regions of contiguous homologous sequences to be able to compare the (dis)similarities between sequences. DNA is a double-stranded molecule, with two strands complementary to one another, and sometimes complementary sequences are deposited to the databases. In our proposed method, a few numerical representations, such as Purine/Pyrimidine representation can process the sequences from different strands without resulting in erroneous classification. In contrast, for alignment-based methods, the sequences to be compared must be from the same strand. Moreover, mitochondrial DNA is circular in nature, and most often authors deposit the corresponding linear sequences to the databases with different starting positions. Alignment-free methods can handle sequences with different starting positions, but this puts the alignment-based methods at undue disadvantage, when comparing the performance of alignment-free methods with that of alignment-based methods. In this addendum, we describe a new experiment that is intended to address this issue.

In Chapter 3, we compared the performance of our proposed alignment-free methodology with two alignment-based methods (MUSCLE, CLUSTALW), and one alignment-free method (FFP). We used three datasets for comparison, two benchmark datasets (38 Influenza sequences and 41 Mammalian sequences), and one larger dataset of 4322 complete mtDNA sequences. The two curated benchmark datasets have been used in the past for sequence analysis and are free from the possible anomalies that may cause difficulties for alignment-based methods. The third dataset used for the comparison was not curated to verify if all the sequences belong to the same DNA strand and start from the same position. Due to the large dataset size, alignment-based methods (MUSCLE, CLUSTALW) were unable to complete the processing on the third dataset and hence no classification accuracy scores were reported for this dataset. However, one could argue that, in this comparison (Chapter 3), alignment-based methods were not utilized optimally.

To address this issue, we performed a new test that utilizes alignment-based methods in the way they were intended, by using a benchmark dataset of cytochrome c oxidase subunit I (*COXI*, also known as *COI*) gene of 3089 vertebrates (bats: 840, birds: 1623, fish: 626), previously used for DNA Barcoding analysis [1]. We curated the dataset by removing all of the unrecognized characters and keeping only the occurrences of *'A', 'C', 'G', 'T'*. We discarded all the sequences with length less than 600 after removing the unrecognized letters to have a curated dataset of 2630 vertebrates (bats: 819, birds: 1199, fish: 612). The performance of ML-DSP [2] was compared with two state-of-the-art alignment-based methods, CLUSTALW [3], and MUSCLE [4], both available as part of MEGAX [5]. CLUSTALW was tested using a default *'slow and accurate mode'*, as well as, *'fast and approximate mode'* with the respective default parameters. MUSCLE was tested with default parameters. ML-DSP was tested using two numerical representations, Chaos Game Representation (CGR) at *k*-value 6, and Purine/Pyrimidine (PP) representation with the sequences normalized to the median length [2]. For both representations, Pearson Correlation Coefficient (PCC) was used as a dissimilarity measure to compute a pairwise distance matrix. The dataset details and the results of performance comparison are given in Table E.S1. The reported processing time included all computations, starting from reading the datasets to the completion of the distance matrix - the common element of all three methods. All experiments were performed on an ASUS ROG $G752VS$ computer with 4 cores (8 threads) of a 2.7 GHz Intel Core $i7$ 6820 HK processor and 64 GB DD4 2400 MHz SDRAM.

ML-DSP achieved an average classification accuracy (over six classifiers used in this thesis) of 99.7% using CGR at k-value 6 (see the respective MoDMap3D in Figure E.S1(a)). An average classification accuracy of 100% is achieved when ML-DSP is used with PP as numerical representation (see the respective MoDMap3D in Figure E.S1(b)). MUSCLE achieved similar average accuracy score of 99.8%. CLUSTALW achieved slightly lower average classification accuracy of 98.5% when tested using *'fast and approximate mode'*.

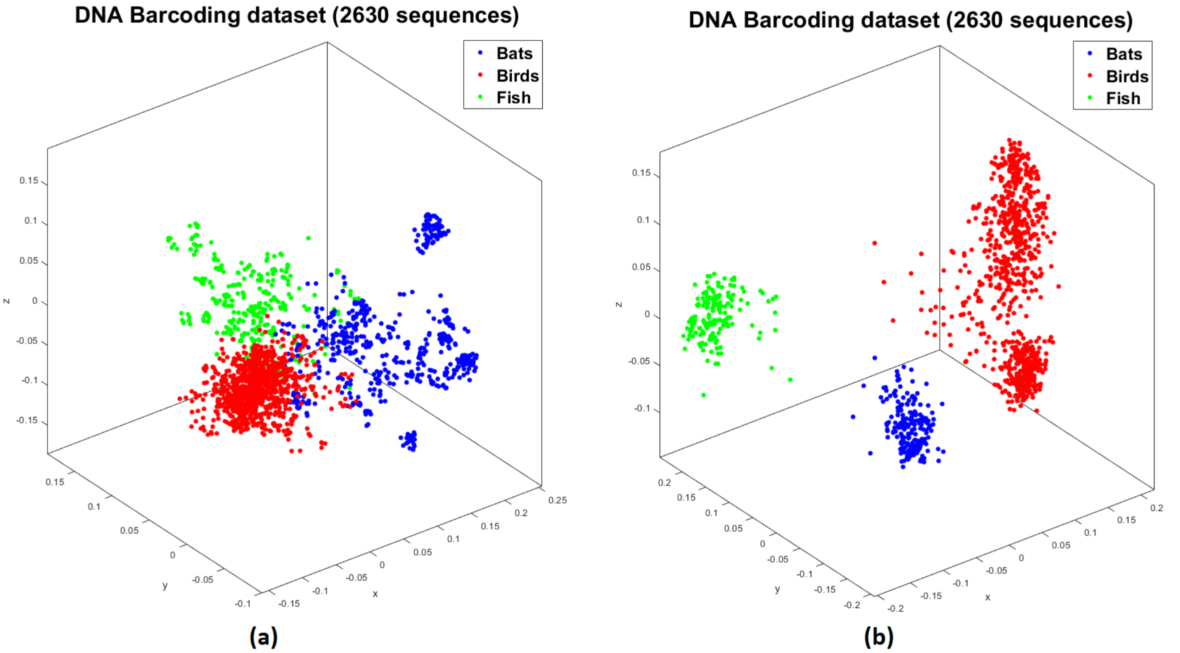
The selected dataset of 2630 sequences does not pose a challenge to the multi-sequence alignment methods, because this is a relatively smaller dataset with an average sequence length of only 650 bp, with each sequence representing the same region of the genome. ML-DSP completed the dataset processing in under 3 seconds, whereas MUSCLE took 3 minutes to complete. CLUSTALW completed the processing in 38 minutes when tested using *'fast and approximate mode'*. With *'slow and accurate mode'*, CLUSTALW was unable to complete the processing in 2 hours 30 minutes and was terminated.

To summarize, ML-DSP overwhelmingly outperformed the alignment-based methods MUSCLE and CLUSTALW in terms of processing time. ML-DSP and MUSCLE achieved (near-) perfect classification accuracy scores. Our results show that while the ML-DSP can easily adapt to the short and conserved sequences (suitable and sometimes strictly required for the alignment), it is challenging for alignment-based methods to process larger datasets of complete genomes. A few alignment-based methods, such as MUSCLE, can process the smaller datasets quickly, but that involves a lot of manual effort and biological expertize on data curation, which is often ignored and not included in the computational cost.

Supplementary Table E.S1: Performance comparison of CLUSTALW, MUSCLE, and ML-DSP.

Dataset	**Parameter**		**CLUSTALW**		**MUSCLE**	**ML-DSP**	
			Slow mode	**Fast mode**		**CGR(k=6)**	**PP**
DNA Barcoding dataset (COXI gene, 2630 sequences) Bats: 819, Birds: 1199, Fish: 612 Length statistics: Maximum: 678, Mean: 650 Minimum: 600, Median: 653	Processing time		>2 hr 30 min	38 min	3 min	2.33 sec	1.61 sec
	Classification Accuracy (%)	Linear Discriminant	—-	95.8	98.9	99.9	100
		Linear SVM	—-	98.9	100	99.8	100
		Quadratic SVM	—-	99.7	100	99.9	100
		Fine KNN	—-	99.6	100	99.9	100
		Subspace Discriminant	—-	97.6	100	98.9	100
		Subspace KNN	—-	99.5	100	99.9	100
		Average	—-	98.5	99.8	99.7	100

Performance of CLUSTALW, MUSCLE, and ML-DSP is compared using a dataset comprising cytochrome c oxidase subunit I (*COXI*) gene of 2630 vertebrates (bats: 819, birds: 1199, fish: 612). ML-DSP shows superior processing time and similar accuracy scores in comparison with MUSCLE and CLUSTALW.



Supplementary Figure E.S1: MoDMap3D representing a dataset comprising of *COXI* gene of 2630 vertebrates (bats: 819, birds: 1199, fish: 612) computed using ML-DSP with two different numerical representation, (a) Chaos Game Representation (CGR) at k-value 6 and, (b) Purine/Pyrimidine (PP) representation.

Bibliography

[1] Weitschek E, Fiscon G, Felici G. Supervised DNA Barcodes species classification: analysis, comparisons and results. BioData Mining. 2014; 7(4).

[2] Randhawa GS, Hill KH, Kari L. ML-DSP: Machine Learning with Digital Signal Processing for ultrafast, accurate, and scalable genome classification at all taxonomic levels. BMC Genomics. 2019; 20: 267.

[3] Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004; 32(5): 1792–7.

[4] Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 1994; 22(22): 4673–80.

[5] Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. Molecular Biology and Evolution. 2018; 35(6): 1547–1549.

Curriculum Vitae

Name:	Gurjit Singh Randhawa
Post-Secondary Education and Degrees:	The University of Western Ontario London, ON 2015 - 2020 Ph.D. (Computer Science)
	The University of Western Ontario London, ON 2013 - 2014 M.Sc. (Computer Science)
	Guru Nanak Dev University Amritsar, India 2008 - 2010 M.C.A.
	Guru Nanak Dev University Amritsar, India 2005 - 2008 B.C.A.
Honours and Awards:	Student and New Investigator Travel Award (×2) Environmental Mutagenesis and Genomics Society (EMGS) 2018, 2019
	First prize in Software Engineering and Programming Languages, Bioinformatics and Theory of Computer Science UWO Research in Computer Science (UWORCS) 2019
	First prize in Bioinformatics & Distributed Systems UWO Research in Computer Science (UWORCS) 2018
	Western Graduate Research Scholarship (WGRS) 2015-19
Related Work Experience:	Graduate Teaching Assistant The University of Western Ontario 2015 - 2020

Publications:

1. Gurjit S. Randhawa, Maximillian P.M. Soltysiak, Hadi El Roz, Camila P.E. de Souza, Kathleen A. Hill, and Lila Kari,“Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: COVID-19 case study”, PLoS ONE. 2020; 15(4): e0232391;
2. Gurjit S. Randhawa, Kathleen A. Hill, and Lila Kari,“MLDSP-GUI: An alignment-free standalone tool with an interactive graphical user interface for DNA sequence comparison and analysis”, Bioinformatics. 2019; btz918.
3. Gurjit S. Randhawa, Kathleen A. Hill, and Lila Kari,“ML-DSP: Machine Learning with Digital Signal Processing for ultrafast, accurate, and scalable genome classification at all taxonomic levels”, BMC Genomics. 2019; 20: 267.
4. Sunny Sharma, and Gurjit S. Randhawa, “Optimization of Online Job Shop Partitioning and Scheduling for Heterogeneous Systems using Genetic Algorithm”, International Journal of Computer Trends and Technology (IJCTT). 2016; 34(3): 144–149.
5. Palak Sharma, Shelza, Gurjit S. Randhawa, and Rajinder S. Virk, “Cost Optimization of Pipeline Systems Using Genetic Algorithm”, International Journal of Computer Engineering and Technology (IJCET). 2013; 4(4).
6. Sookham R.P. Singh, Gurjit S. Randhawa, and Rajinder S. Virk, “Efficacy of Genetic Algorithms in Staging Cervical Cancer”, International Journal of Cancer Research. 2013; 47(2): 1164–1168.
7. Gurvinder Singh, Rajinder S. Virk, Gurjit S. Randhawa, “Enhancing Computational Capabilities in Higher Education by use of GA’s”, International Conference on Role of Technology in Enhancing the Quality of Higher Education (ICRT-12), Oct 26-27, 2012; Kanya Maha Vidyalaya, Jalandhar, India.