Escuela Profesional de
Ciencia de la Computación

Bioinformatics

An Analysis of k-Mer Frequency Features with Machine Learning Models for Viral Subtyping Classification

MSc. Vicente Machaca Arceda

Universidad La Salle

2020

Tabla de contenido

1. Introduction
   - What is Bioinformatics?
2. Viral subtype classification
   - Viral subtype
   - K-mer frequency
   - Kameris
   - Castor-KRFE
   - CNN
3. Results
   - Materials and methods
   - Datasets
   - Results
4. Results

vmachacaa@ulasalle.edu.pe

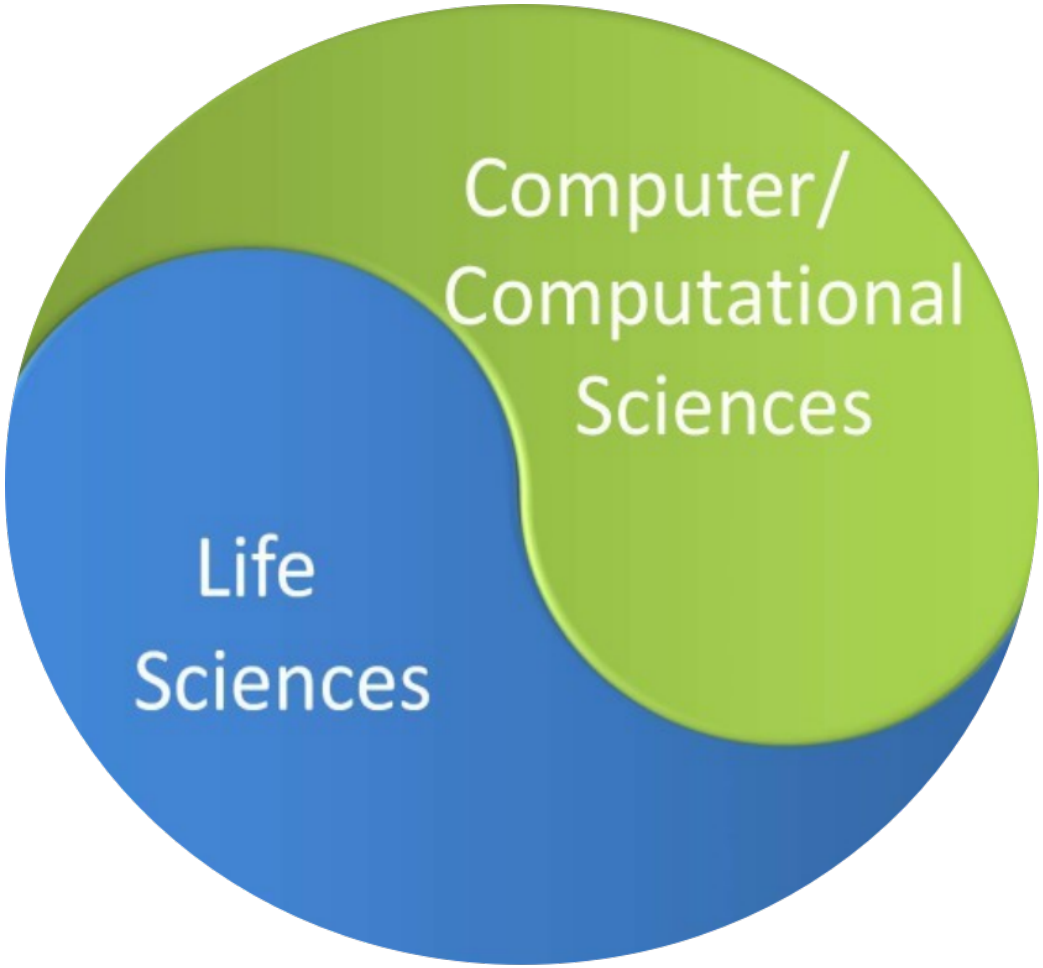
Introduction

What is Bioinformatics?

According to Luscombe et al.: **Bioinformatics** involves the technology that uses computers for storage, retrieval, manipulation, and distribution of information related to biological macromolecules such as DNA, RNA, and proteins [1].

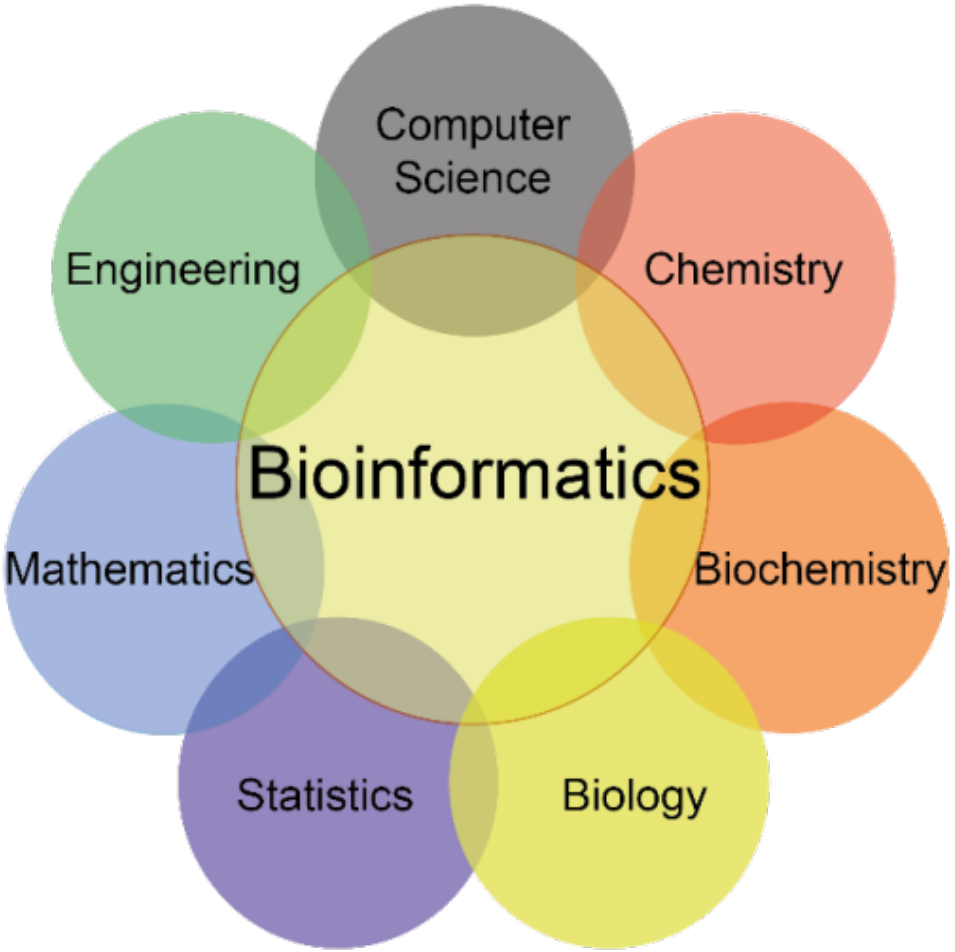
vmachacaa@ulasalle.edu.pe

Bioinformatics



vmachacaa@ulasalle.edu.pe

Bioinformatics



vmachacaa@ulasalle.edu.pe

Tabla de contenido

1. Introduction
   - What is Bioinformatics?
2. Viral subtype classification
   - Viral subtype
   - K-mer frequency
   - Kameris
   - Castor-KRFE
   - CNN
3. Results
   - Materials and methods
   - Datasets
   - Results
4. Results

vmachacaa@ulasalle.edu.pe

Viral subtype classification

Subtype of HIV

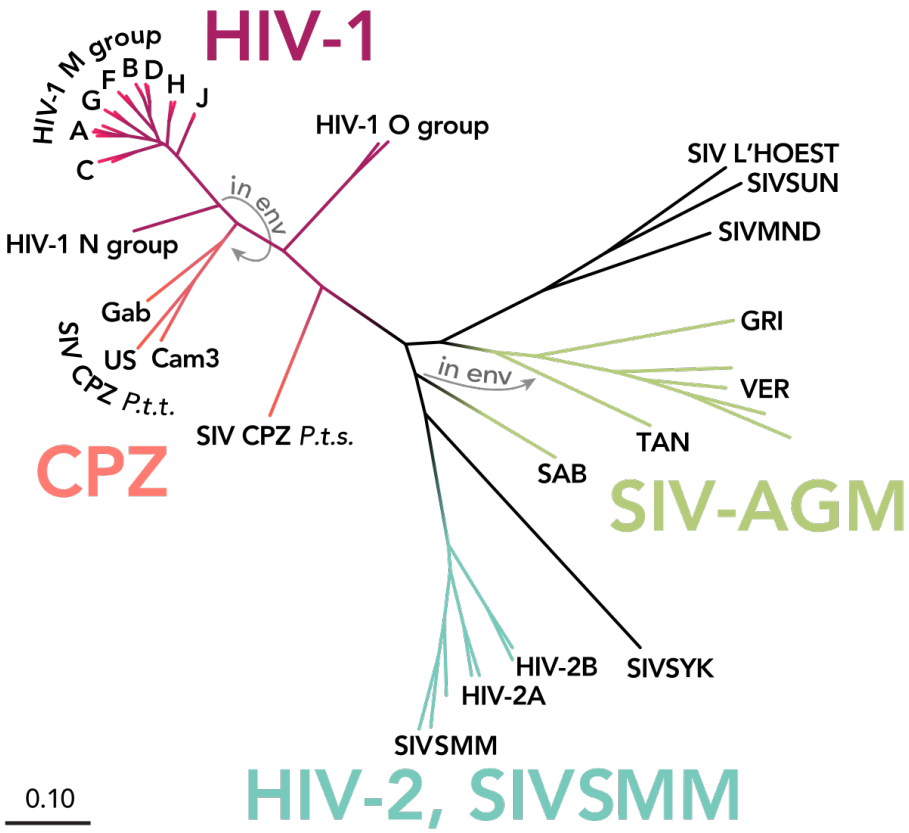


Figure: Phylogenetic tree of the SIV and HIV viruses. Source: [2]

vmachacaa@ulasalle.edu.pe

Viral subtype classification

Example of DNA

```
>J01859.1 Escherichia coli 16S ribosomal RNA, complete sequence
AAATTGAAGAGTTTGATCATGGCTCAGATTGAACGCTGGCGGCAGGCCTAACACATGCAAGTCGAACGGT
AACAGGAAGAAGCTTGCTCTTTGCTGACGAGTGGCGGACGGGTGAGTAATGTCTGGGAAACTGCCTGATG
GAGGGGGATAACTACTGGAAACGGTAGCTAATACCGCATAACGTCGCAAGACCAAAGAGGGGGACCTTCG
GGCCTCTTGCCATCGGATGTGCCCAGATGGGATTAGCTAGTAGGTGGGGTAACGGCTCACCTAGGCGACG
ATCCCTAGCTGGTCTGAGAGGATGACCAGCCACACTGGAACTGAGACACGGTCCAGACTCCTACGGGAGG
CAGCAGTGGGGAATATTGCACAATGGGCGCAAGCCTGATGCAGCCATGCCGCGTGTATGAAGAAGGCCTT
CGGGTTGTAAAGTACTTTCAGCGGGGAGGAAGGGAGTAAAGTTAATACCTTTGCTCATTGACGTTACCCG
CAGAAGAAGCACCGGCTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGGTGCAAGCGTTAATCGGAAT
TACTGGGCGTAAAGCGCACGCAGGCGGTTTGTTAAGTCAGATGTGAAATCCCCGGGCTCAACCTGGGAAC
TGCATCTGATACTGGCAAGCTTGAGTCTCGTAGAGGGGGGTAGAATTCCAGGTGTAGCGGTGAAATGCGT
AGAGATCTGGAGGAATACCGGTGGCGAAGGCGGCCCCCTGGACGAAGACTGACGCTCAGGTGCGAAAGCG
TGGGGAGCAAACAGGATTAGATACCCTGGTAGTCCACGCCGTAAACGATGTCGACTTGGAGGTTGTGCCC
TTGAGGCGTGGCTTCCGGAGCTAACGCGTTAAGTCGACCGCCTGGGGAGTACGGCCGCAAGGTTAAAACT
CAAATGAATTGACGGGGGCCCGCACAAGCGGTGGAGCATGTGGTTTAATTCGATGCAACGCGAAGAACCT
TACCTGGTCTTGACATCCACGGAAGTTTTCAGAGATGAGAATGTGCCTTCGGGAACCGTGAGACAGGTGC
TGCATGGCTGTCGTCAGCTCGTGTTGTGAAATGTTGGGTTAAGTCCCGCAACGAGCGCAACCCTTATCCT
TTGTTGCCAGCGGTCCGGCCGGGAACTCAAAGGAGACTGCCAGTGATAAACTGGAGGAAGGTGGGGATGA
CGTCAAGTCATCATGGCCCTTACGACCAGGGCTACACACGTGCTACAATGGCGCATACAAAGAGAAGCGA
CCTCGCGAGAGCAAGCGGACCTCATAAAGTGCGTCGTAGTCCGGATTGGAGTCTGCAACTCGACTCCATG
AAGTCGGAATCGCTAGTAATCGTGGATCAGAATGCCACGGTGAATACGTTCCCGGGCCTTGTACACACCG
CCCGTCACACCATGGGAGTGGGTTGCAAAAGAAGTAGGTAGCTTAACCTTCGGGAGGGCGCTTACCACTT
TGTGATTCATGACTGGGGTGAAGTCGTAACAAGGTAACCGTAGGGGAACCTGCGGTTGGATCACCTCCTT
```

vmachacaa@ulasalle.edu.pe

Viral subtype classification

Problem

- The most used **alignment-based** method are BLAST and CLUSTALW.

vmachacaa@ulasalle.edu.pe

Viral subtype classification

Problem

- The most used **alignment-based** method are BLAST and CLUSTALW.
- They are slow. For example, it take one hour to align 18 sequences of 18k bp.

vmachacaa@ulasalle.edu.pe

Viral subtype classification

Problem

- The most used **alignment-based** method are BLAST and CLUSTALW.
- They are slow. For example, it take one hour to align 18 sequences of 18k bp.
- DNA sequences increases every day, so **alignment-based** methods get slower every second.

vmachacaa@ulasalle.edu.pe

Viral subtype classification

Objective

Compare **alignment-free** algorithms based on k-mer frequencies.

- Kameris [3].
- Castor-KRFE [4].
- CNN.

We got two publications [5], [6].

vmachacaa@ulasalle.edu.pe

Tabla de contenido

1. Introduction
   - What is Bioinformatics?
2. Viral subtype classification
   - Viral subtype
   - K-mer frequency
   - Kameris
   - Castor-KRFE
   - CNN
3. Results
   - Materials and methods
   - Datasets
   - Results
4. Results

vmachacaa@ulasalle.edu.pe

K-mer frequency

k-mer

For sequence $s = \{A, C, T, G, A, C\}$

- 2-mers set: $\{AC, CT, TG, GA\} \rightarrow \{2, 1, 1, 1\}$
- 3-mers set: $\{ACT, CTG, TGA, GAC\} \rightarrow \{1, 1, 1, 1\}$

vmachacaa@ulasalle.edu.pe

Tabla de contenido

1. Introduction
   - What is Bioinformatics?
2. Viral subtype classification
   - Viral subtype
   - K-mer frequency
   - Kameris
   - Castor-KRFE
   - CNN
3. Results
   - Materials and methods
   - Datasets
   - Results
4. Results

vmachacaa@ulasalle.edu.pe

Kameris

K-mer frequency

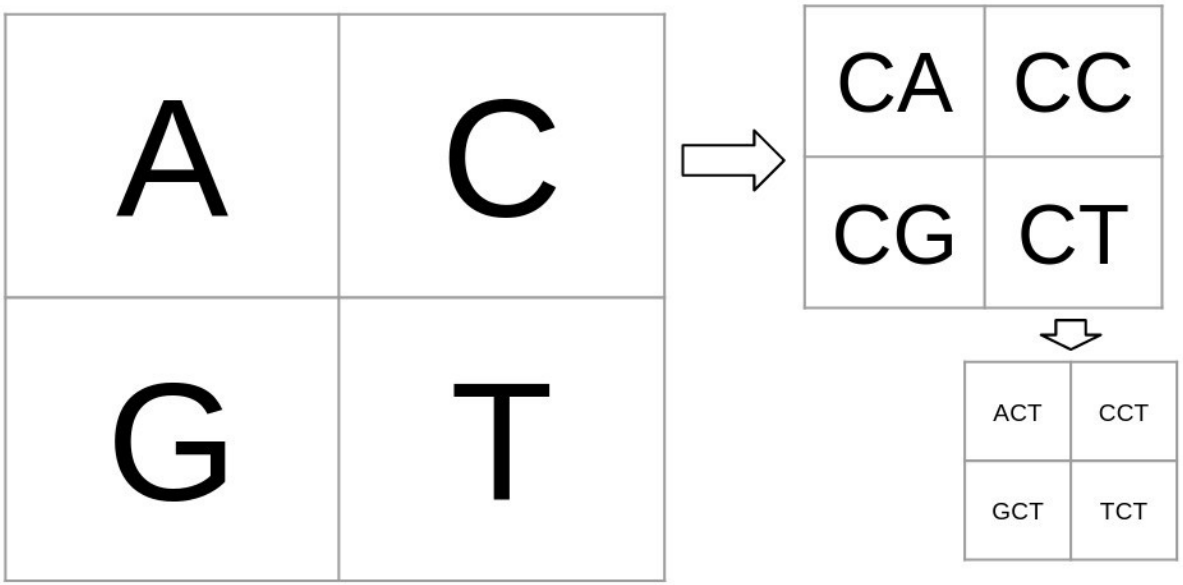


Figure: A FCGR matrix. The *G* quadrant sub-divided into the corresponding G-endings and the *TG* quadrant sub-divided into the corresponding *TG-endings*.

vmachacaa@ulasalle.edu.pe

Kameris

K-mer frequency

aa 5	**ac** 2	**ca** 5	**cc** 1
ag 3	**at** 4	**cg** 0	**ct** 4
ga 2	**gc** 1	**ta** 3	**tc** 5
gg 1	**gt** 2	**tg** 4	**tt** 0

Figure: A FCGR k-mer example, each k-mer is representing as a cell in the matrix, and the frequency of each k-mer is represented as the pixel value.

vmachacaa@ulasalle.edu.pe

Kameris

Method

- Compute k-mer frequencies using FCGR (4^k).
- Dimensionality reduction with Sigle Value Decomposition.
- SVM classifier.

vmachacaa@ulasalle.edu.pe

Tabla de contenido

1. Introduction
   - What is Bioinformatics?
2. Viral subtype classification
   - Viral subtype
   - K-mer frequency
   - Kameris
   - Castor-KRFE
   - CNN
3. Results
   - Materials and methods
   - Datasets
   - Results
4. Results

vmachacaa@ulasalle.edu.pe

Castor-KRFE

Method

- Compute k-mer frequencies, just take into account k-mer presented in DNA sequence.
- Feature elimination with Recursive Feature Elimination.
- SVM classifier.

vmachacaa@ulasalle.edu.pe

Tabla de contenido

1. Introduction
   - What is Bioinformatics?
2. Viral subtype classification
   - Viral subtype
   - K-mer frequency
   - Kameris
   - Castor-KRFE
   - CNN
3. Results
   - Materials and methods
   - Datasets
   - Results
4. Results

vmachacaa@ulasalle.edu.pe

CNN

FCGR

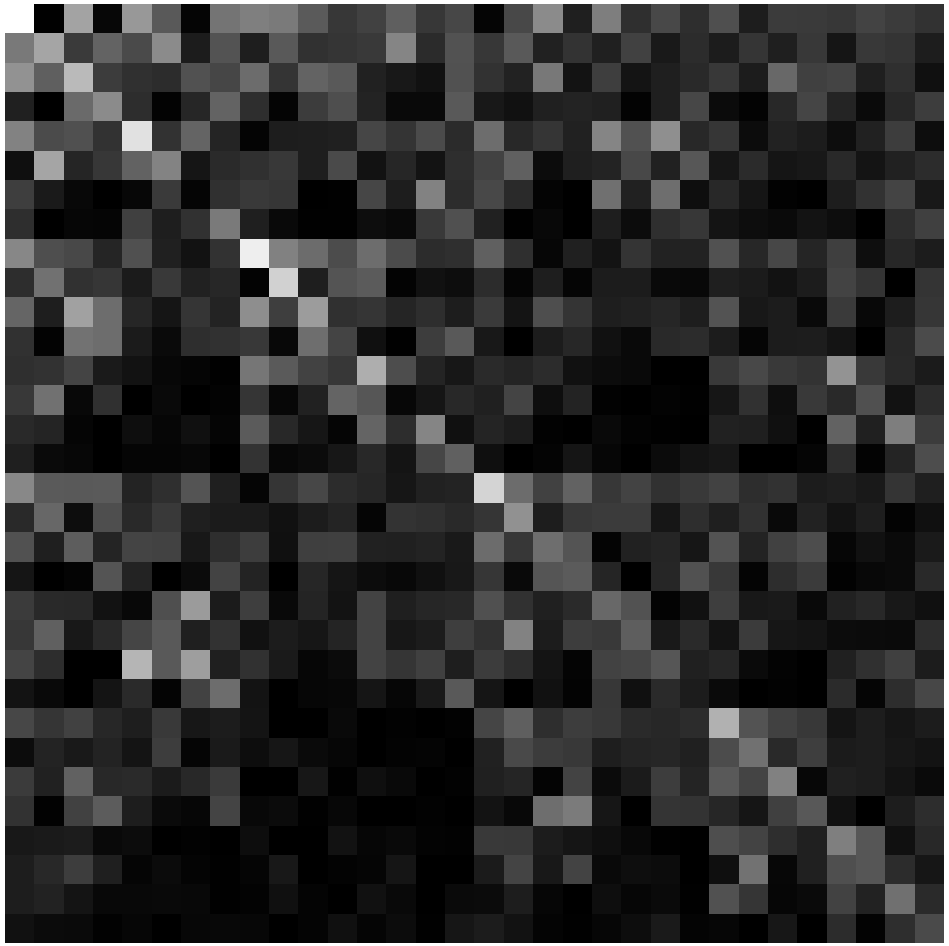


Figure: A FCGR (k=5) of a HIV-1 genome.

vmachacaa@ulasalle.edu.pe

CNN

Method

- Compute FCGR
- Represent the FCGR as an image.
- Train with CNNs.

vmachacaa@ulasalle.edu.pe

Tabla de contenido

1. Introduction
   - What is Bioinformatics?
2. Viral subtype classification
   - Viral subtype
   - K-mer frequency
   - Kameris
   - Castor-KRFE
   - CNN
3. Results
   - Materials and methods
   - Datasets
   - Results
4. Results

vmachacaa@ulasalle.edu.pe

Methods used

Table: Methods used in this research.

Method name	Description
Kameris-SVD	Kameris with dimensionality reduction SVD.
Kameris	Kameris without dimensionality reduction.
Castor-KRFE	Castor with feature elimination RFE.
Castor	Castor without feature elimination.
CNN	The method that used FCGR with CNN (three architectures: CNN-1, CNN-2 and CNN-3).
ML-DSP	The method that process the DNA as a digital signal.

vmachacaa@ulasalle.edu.pe

Tabla de contenido

1. Introduction
   - What is Bioinformatics?
2. Viral subtype classification
   - Viral subtype
   - K-mer frequency
   - Kameris
   - Castor-KRFE
   - CNN
3. Results
   - Materials and methods
   - Datasets
   - Results
4. Results

vmachacaa@ulasalle.edu.pe

Datasets

Table: The datasets used in the experiments.

Data sets	Average seq. length	No. of classes	No. of instances
HBVGENCG	3189	8	230
HIVGRPCG	9164	4	76
HIVSUBCG	8992	18	597
HIVSUBPOL	1211	28	1352
INFSUBHA	1719	2	10825
INFSUBMP	759	2	21421
INSUBFNA	1416	2	10715
EBOSPECG	18917	5	751
RHISPECG	369	3	1316
HPVGENCG	7610	3	125

vmachacaa@ulasalle.edu.pe

Datasets

Table: The datasets used in the experiments.

Data sets	Average seq. length	No. of classes	No. of instances
Primates	16626	2	148
Dengue	10595	4	4721
Protists	31712	3	159
Fungi	49178	3	224
Plants	277931	2	174
Amphibians	17530	3	290
Insects	15689	7	898
3classes	16292	3	2170
Vertebrates	16806	5	4322

vmachacaa@ulasalle.edu.pe

Tabla de contenido

1. Introduction
   - What is Bioinformatics?
2. Viral subtype classification
   - Viral subtype
   - K-mer frequency
   - Kameris
   - Castor-KRFE
   - CNN
3. Results
   - Materials and methods
   - Datasets
   - Results
4. Results

vmachacaa@ulasalle.edu.pe

Results

Comparison between Kameris and Castor

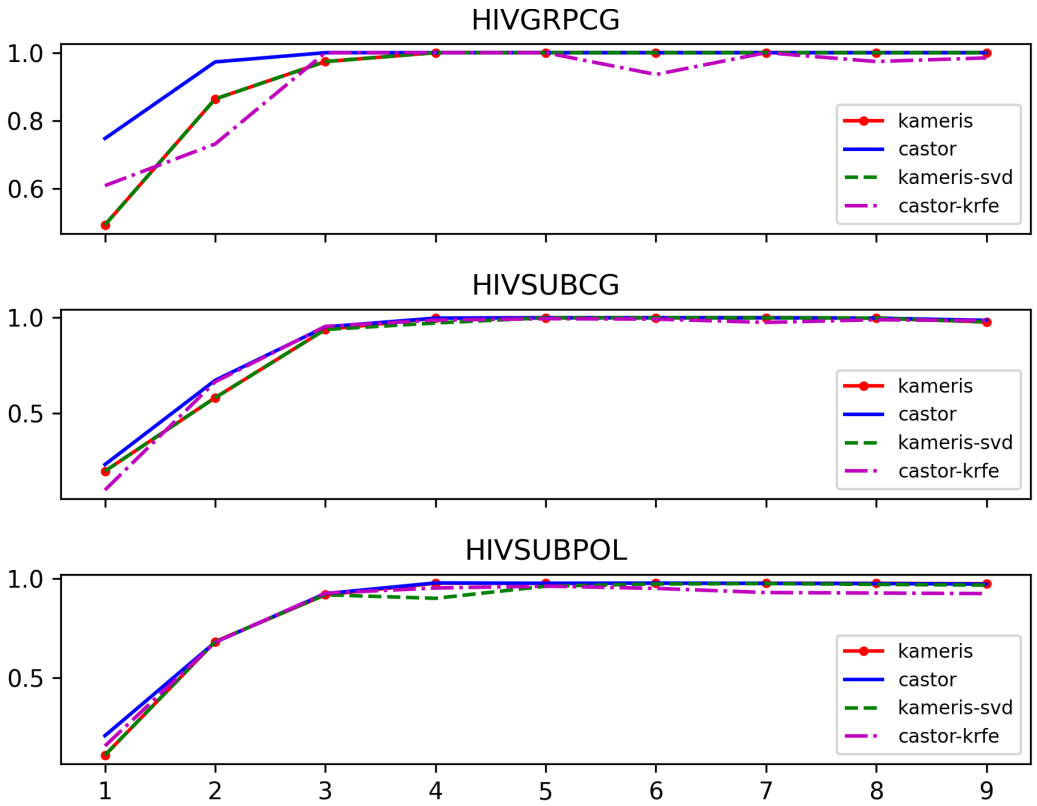


Figure: A comparison of f-score for the datasets HIVGRPCG, HIVSUBCG and HIVSUBPOL. The f-score were computed for different k-mers, ranging from k=1 to k=9.

vmachacaa@ulasalle.edu.pe

Results

Vector size of Kameris and Castor

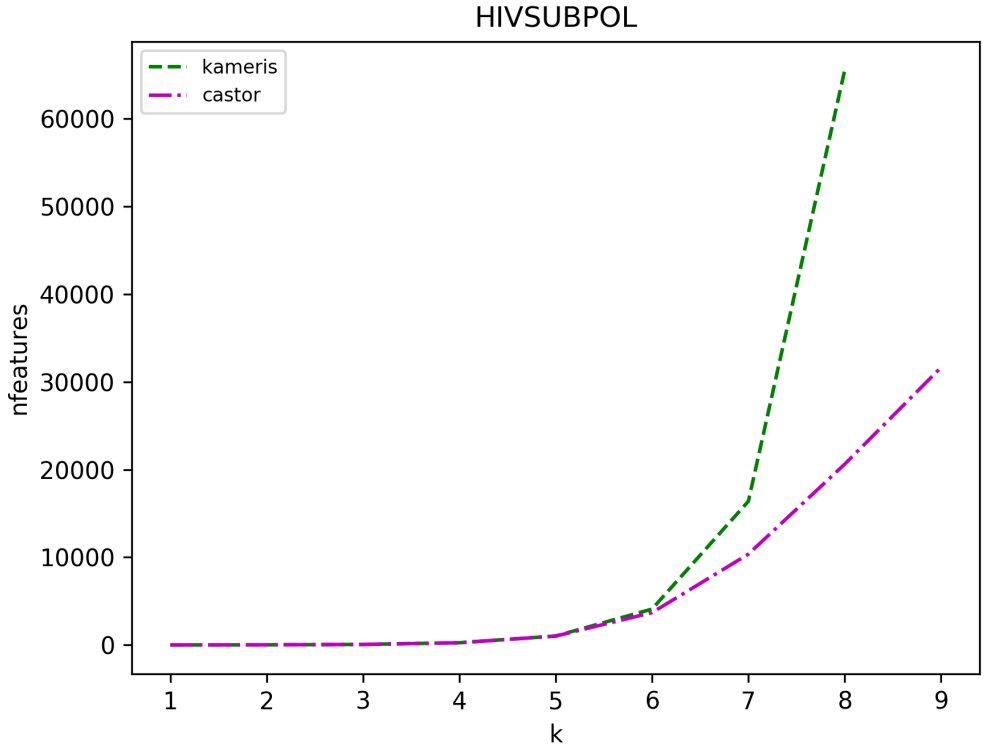


Figure: Size of feature vectors for Castor and Kameris without dimensionality SVD reduction and feature elimination RFE for HIVSUBPOL dataset. X-axis represent *k* value in k-mer.

vmachacaa@ulasalle.edu.pe

Results

Vector size of Kameris and Castor

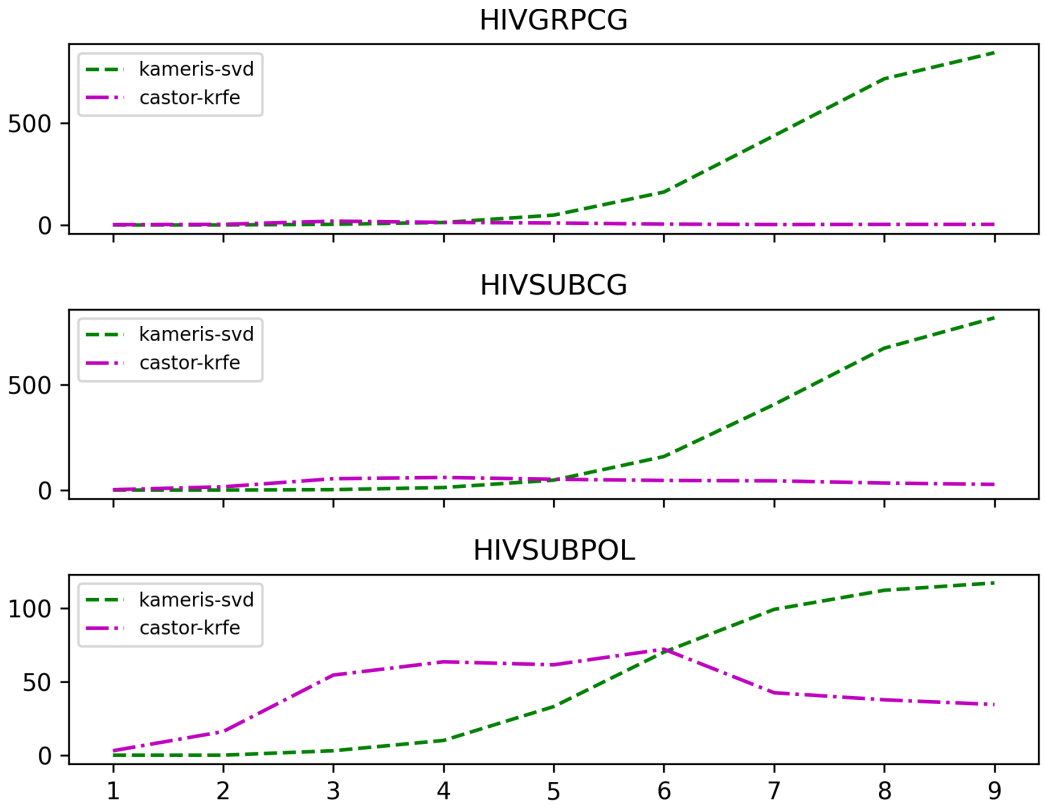


Figure: Size of feature vectors for Kameris-SVD/Kameris and Castor-KRFE/Castor for HIVGRPCG, HIVSUBCG and HIVSUBPOL. X-axis represent *k* value in k-mer.

vmachacaa@ulasalle.edu.pe

Results

Kameris vs Castor

Table: The best f-score with the minimum *k* value in k-mer by each dataset. Also, the number of features is presented.

	Kameris-SVD		
Dataset	(k-mer)	f-score	nfeatures
HIVGRPCG	4	1.0000	12
HIVSUBCG	5	**0.9983**	47
HIVSUBPOL	7	**0.9761**	99
	Castor-KRFE		
Dataset	(k-mer)	f-score	nfeatures
HIVGRPCG	3	1.0000	19
HIVSUBCG	5	0.9937	51
HIVSUBPOL	5	0.9629	65

vmachacaa@ulasalle.edu.pe

Results

Kameris, Castor, ML-DSP and CNN

Table: Accuracy of the three CNN-2 architecture, ML-DSP, Kameris and Castor for each dataset.

Dataset	ML-DSP	CNN-2	Kameris	Castor
EBOSPECG	0.92	1.00	1.00	1.00
HBVGENCG	0.15	1.00	1.00	1.00
HIVGRPCG	0.44	1.00	1.00	1.00
HIVSUBCG	0.05	0.98	**1.00**	**1.00**
HIVSUBPOL	0.01	0.97	**1.00**	**1.00**
INFSUBHA	1.00	1.00	1.00	1.00
INFSUBMP	0.89	0.98	**0.99**	**0.99**
INSUBFNA	1.00	1.00	1.00	1.00
RHISPECG	1.00	1.00	1.00	1.00
HPVGENCG	0.36	1.00	1.00	1.00

vmachacaa@ulasalle.edu.pe

Results

Kameris, Castor, ML-DSP and CNN

Table: Accuracy of the three CNN-2 architecture, ML-DSP, Kameris and Castor for each dataset.

Dataset	ML-DSP	CNN-2	Kameris	Castor
Primates	0.97	1.00	1.00	1.00
Dengue	1.00	1.00	1.00	1.00
Protists	0.50	0.97	**1.00**	**1.00**
Fungi	0.40	1.00	1.00	1.00
Plants	0.69	0.91	0.89	**0.97**
Amphibians	0.60	1.00	1.00	1.00
Insects	0.37	0.97	**0.99**	**0.99**
3classes	0.57	1.00	1.00	1.00
Vertebrates	0.52	1.00	1.00	1.00

vmachacaa@ulasalle.edu.pe

Results

Kameris, Castor, ML-DSP and CNN

We evaluated four methods for viral subtyping classification, based on alignment-free algorithms.

Kameris-SVD outperformed slightly Castor-KRFE. Moreover, if we did not use SVD and RFE for each method, they got the same f1-score.

Castor-KRFE got a smaller feature vector than Kameris-SVD

Kameris and Castor without SVD and RFE got the best accuracy, but they are followed CNNs.

vmachacaa@ulasalle.edu.pe

References I

- N. M. Luscombe, D. Greenbaum, and M. Gerstein, "What is bioinformatics? a proposed definition and overview of the field," *Methods of information in medicine*, vol. 40, no. 04, pp. 346–358, 2001.
- Wikipedia, "Subtypes of hiv," https://en.wikipedia.org/wiki/Subtypes_of_HIV, 2020, accessed: 2020-09-07.
- S. Solis-Reyes, M. Avino, A. Poon, and L. Kari, "An open-source k-mer based machine learning tool for fast and accurate subtyping of hiv-1 genomes," *PloS one*, vol. 13, no. 11, 2018.
- D. Lebatteux, A. M. Remita, and A. B. Diallo, "Toward an alignment-free method for feature extraction and accurate classification of viral sequences," *Journal of Computational Biology*, vol. 26, no. 6, pp. 519–535, 2019.

vmachacaa@ulasalle.edu.pe

References II

- V. M. Arceda, “An analysis of k-mer frequency features with machine learning models for viral subtyping of polyomavirus and hiv-1 genomes,” in *Proceedings of the Future Technologies Conference*. Springer, 2020, pp. 279–290.
- V. E. Machaca Arceda, “An analysis of k-mer frequency features with svm and cnn for viral subtyping classification,” *Journal of Computer Science & Technology*, vol. 20, 2020.

vmachacaa@ulasalle.edu.pe

Questions?

vmachacaa@ulasalle.edu.pe