IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.7, Special Issue.3, Feb. 2019 E-ISSN: 2347-2693

Genome Based Classification of Human Papilloma Virus Using Linear Discriminant Analysis

S. Swain[1*], M.R. Patra[2]

[1,2]Department of Computer Science, Berhampur University, Berhampur, Odisha, India

[*]*Corresponding Author:* sswain1985@gmail.com *Tel.: +91-98615-74715*

Available online at: www.ijcseonline.org

Abstract—Biological classification of Papillomaviridae leads to several hundred different genera (classes) of Human Papilloma Viruses (HPV) that are discriminated on the basis of more than hundred different characteristics. Statistical procedures of classification based on genome and gene size are being applied to biologically define different class labels for HPV. In this paper, Fisher's linear discriminant analysis (LDA) has been used for classification of HPV on the basis of total genome size and gene sizes. Univariate and multivariate modes of classification have been employed to recognize two distinct classes of HPV viz., alpha- papilloma and beta-papilloma that cause cervical cancer in humans. The aim is to build a classification model so as to predict unknown samples. The accuracy of the proposed model has been measured on a sample dataset.

Keywords—Genome, Genes, HPV, LDA, Papillomaviridae, Multivariate analysis, and Univariate analysis

I. INTRODUCTION

Classification is a data mining technique that assigns items in a collection to target categories or classes. A classification task begins with a data set in which the class assignments are known. Statistical classification is a statistical procedure in which individual items are placed into groups based on quantitative information on one or more characteristics inherent in the items and based on a training set of previously labelled items. Early work on statistical classification was undertaken by Fisher (Fisher, 1936) in the context of two-group problems, leading to Fisher's linear discriminant function as the rule for assigning a group to a new observation (Gnanadesikan, 1977) [3]. This early work assumed that data-values within each of the two groups had a multivariate normal distribution. The extension of this same context to more than two-groups has also been considered with a restriction imposed that the classification rule should be linear (Fisher, 1938) [2]. Later work for the multivariate normal distribution allowed the classifier to be nonlinear (McLachlan, 2004) [6]. In the field of biology, Classification of DNA/RNA sequences on the basis of genome sequence information is gaining popularity in order to find the phylogenetic relationship between organisms. Biological databases are now available from which one can easily get the genomic information. Statistical rules can be applied on such databases to classify biological data using certain genome properties.

Classification of different types of Human Papilloma viruses (usually with circular genomes) into separate classes is one of the contemporary problems. The objective of this work is to develop a simple, faster, and effective methodology, that considers Genome size and Gene size as parameters to classify HPV into specific classes of either Alpha-Papilloma or Beta-papilloma using statistical procedure of LDA. Most of the previous works are on linear genomes in relation to Eukaryotes But here, the focus is on genomes that exhibit circular nature in relation to Prokaryotes. Mostly in Prokaryotes the genome is circular, which are also found in organelles like mitochondria and chloroplasts. This topological constraint is the basis for the characteristic properties of closed circular DNA, which have fascinated biologists, physicists, and statisticians. Therefore, we have taken HPV genomes into consideration and try to build classification rules based on genome characteristics of HPV to classify HPVs.

Rest of the paper is organized as follows, Section I contains the introduction to the problem of classification of HPV, Section II contains the related work, Section III contains the methodology with an outline of the experiment conducted, Section IV describes the results and discusses about the classification, Section V concludes research work with future directions.

II. RELATED WORK

A good number of works have been reported in literature on importance of Genome size and gene size. This dramatic growth of biology data and non-biological commercial databases pose a challenge for data mining. Classification is one of the major tools in the analysis of biological data

(Yushan Qui, Xiaoqing Cheng, 2015) [10]. Eukaryotic genome size data are becoming increasingly important both as the basis for comparative research into genome evolution and as direct estimators of the cost and difficulty of genome sequencing programs for an expanding sphere of non-model organisms (T Ryan Gregory James and Barnett D. 2007) [8]. Multivariate Statistical Analysis (MSA) methods have recently been introduced for analysing images of biological macromolecules proposed by M Van Heel (Heel, 1984) [5].

Sihua Peng, Qianghua Xu proposed Simultaneous multiclass classification of tumour types, which is essential for future clinical implementations of microarray-based cancerdiagnosis (2003) [7]. One hundred eighteen papillomavirus (PV) types have been completely described, and a yet higher number of presumed new types have been detected by preliminary data such as sub genomic amplicons. The classification of this diverse group of viruses, which include important human pathogens, has been debated for three decades. (Villiers, Fauque, Broker, and Bernard, 2004) [9]. Dudoit, Fridly and Speed (2002) [1] have compared the performance of different discrimination methods for classifying tumors based on gene expression data. The discrimination methods are applied to datasets from three recently published cancer gene expression studies.

III. METHODOLOGY

Some of the publicly available data sources for biological data are GenBank, Refseq, SwissProt, and PIR. For our work, data has been gathered from GenBank using the Entrez search engine. The GenBank sequence database is an open source, annotated collection of all publicly available nucleotide sequences and their protein translations. Entrez is an integrated database retrieval system that accesses DNA and protein sequence data, *Genome data,* the NCBI taxonomy, and protein structures. All databases indexed by Entrez can be searched via a single query string, supporting Boolean operators (Boolean query) and search term. For our work, we consider data in relation to the circular genome. In the database we first search for bacteria having circular genome, then proceed for viral and mitochondrial ones. Our main objective is to find the organism with common genes and to use gene and genome sizes for classification. Therefore, we have chosen HPV genome dataset (Table 1 and Table 2) from the classes of viruses to be considered as experimental material for the proposed classification.

Table1. The Alpha-papilloma Virus group's gene size and genome size.

Alpha papilloma	**E6**	**E7**	**E1**	**E2**	**L2**	**L1**	**Total(nt)**
HPV– 18	477	318	1974	1098	1389	1707	**7857**
HPV - 2	480	279	1932	1176	1575	1533	**7860**
HPV type 90	447	297	1941	1143	1404	1518	**8033**
HPV - 61	441	288	1959	1149	1380	1518	**7989**
HPV - 54	435	288	1902	1104	1413	1494	**7759**
HPV type 34	447	291	1944	1038	1419	1587	**7723**
HPV type 32	429	315	1929	1185	1431	1512	**7961**
HPV type 26	453	315	1917	1128	1419	1512	**7855**
HPV type 10	447	266	2046	1131	1413	1596	**7919**
HPV type 24	423	291	1824	1404	1572	1539	**7452**
HPV type 7	464	336	1941	1128	1371	1518	**8027**
HPV – 16	477	297	1949	1098	1422	1596	**7904**

Table2. The Beta-papilloma Virus group's gene size and genome size.

Beta papilloma	**E6**	**E7**	**E1**	**E2**	**L2**	**L1**	**Total (nt)**
HPV - cand96	678	300	1854	1407	1566	1539	**7438**
HPV type 92	417	276	1839	1434	1572	1539	**7461**
HPV type 9	447	282	1818	1386	1602	1524	**7434**
HPV 49	417	312	1830	1467	1566	1530	**7560**
HPV – 5	474	312	1821	1545	1557	1560	**7746**
HPV - 107	423	309	1824	1398	1560	1524	**7562**

Using the Entrez toolbox in the NCBI homepage https://www.ncbi.nlm.nih.gov/, by selecting genome and searching for the term *Papillomaviridae*, the given gene and genome information possible to be obtained.

Methodology: Linear discriminant analysis (LDA) is a generalization of Fisher's linear discriminant, a method used in statistics, pattern recognition and machine learning to find a linear combination of features that characterizes or separates two or more classes of objects or events. The resulting combination may be used as a linear classifier or, more commonly, for dimensionality reduction before classification. In linear discrimination, we assume that instances of a class are linearly separable from instances of other classes. This is a discriminant-based approach that estimates the parameters of the linear discriminant directly from a given labeled sample. For classification, here we offered two-classification methods based on LDA. One is univariate classification method that uses genome size as classification measure and the other one is the multivariate classification method that uses common gene sizes as classification measure.

Classification Rule-1(Univariate): We propose the classification rule, which specify that if

$$(\bar{X} - \bar{Y})\, X_0 > \tfrac{1}{2}(\bar{X}^2 - \bar{Y}^2) \quad (1)$$

where $\bar{X}$ = Arithmatic Mean of Group I

$\bar{Y}$ = Arithmatic Mean of Group II

X_0 = The Genome Size of Individual organism

Then, classify X_0 into Group I otherwise classify it into Group II.

Classification Rule-2 (Multivariate): We propose the classification rule, which specify that if

$$[(\bar{X} - \bar{Y})' S^{-1} X_0] > \tfrac{1}{2}[(\bar{X}'\, S^{-1}\, \bar{X}) - (\bar{Y}'\, S^{-1}\, \bar{Y})] \quad (2)$$

where , $\bar{X} = Mean$ Vector of Group I

$\bar{Y} = Mean$ Vector of Group II

$\bar{X}' = Transpose\ Mean$ Vector of Group I

$\bar{Y}' = Transpose\ Mean$ Vector of Group II

S^{-1} = Inverse covariance matrix

X_0 = Gene sizes of individual organism

Then, classify X_0 into Group I otherwise classify it into Group II. This multivariate classification involves matrix operations. In this rule the right-hand side value remains same for every individual case. The classification rule specifies that if the LHS value is greater than the RHS value then classify X_0 into Group I and if the LHS value is less than RHS value then put it in Group II.

Outline of the Experiment:

Step-1: Experimental data (Gene and Genome sizes) collected from the GenBank using Entrez search engine. Raw data collected from GenBank were organized into data table of genome size and gene sizes that are utilized for classification.

Step-2: Then, we define the classification rule as univariate and multivariate classification method and the defined classification rules were implementation through R programming using IDE RStudio. The methods are:

Univariate Classification: The HPV-18 genome size is 7857 nucleotides (nt) in length. By using classification rule-1, we found the LHS value=2577751 and RHS value = 2525435 for HPV-18 genome. The classification rule-1 says that, if the LHS value is greater than the RHS value then put it in Alpha papilloma group. Thus, the classification rule gave true result (Properly grouped in Alpha papilloma) as per biological group classification. Likewise, calculating and comparing LHS to RHS value to group the given data samples into their respective class labels. Results are presented in the tables with graphical representation through pie charts.

Multivariate Classification: Here also, the classification proceeds in the same way for univariate classification rule-1, with calculating and comparing LHS and RHS value to group Resultant tables and their pie chart representations were recorded for future reference. The result of classification is recorded.

Step-3: Now, it is possible to establish the classifier on the basis of classification of training samples. The training samples were tested with the proposed classification rules and are evaluated for their inclusion in respective true classes or not. These training samples are considered to be the testing samples for future unclassified ones. The performance of classification is measured by comparing classification result with GenBank taxonomy classification.

Step-4: Next, we analyse the performance of the classifiers with respect to the predicted results. With this the overall classification procedure is completed. The next step is to check whether the data classified is correct or not which is obtained by comparing previously observed classification with our result.

Step-5: Data distribution check and cross validation technique were used to validate the classifiers that can be used for future classifications. Cross validation technique has been used to check the error rate in the classification rule. The cross-validation procedure considered is Leave One Out Cross Validation (LOOCV).

Finally, we evaluate few biologically unclassified HPVs using the given classification rule.

IV. RESULTS AND DISCUSSIONS

The human papilloma dataset contains eighteen representatives, which are segregated into two groups, such as Alpha-papilloma and Beta-papilloma. We applied the univariate and multivariate classification methods for classifying these eighteen data points. The obtained results are represented in table 3 and 4 where the last column represents the value as True(T) or False(F) indicating whether it is correctly or incorrectly classified:

Table 3: Result Table for HPV classification by Classification Rule-1 (Univariate).

Alphapapilloma	**Total (nt)**	**LHS VALUE**	**RHS VALUE**	**GROUP**	**T/F**
HPV - 18	7857	2577751	2525435	Alpha	T
HPV - 2	7860	2578735	2525435	Alpha	T
HPV type 90	8033	2635493	2525435	Alpha	T
HPV - 61	7989	2621058	2525435	Alpha	T
HPV - 54	7759	2545599	2525435	Alpha	T
HPV type 34	7723	2533788	2525435	Alpha	T
HPV type 32	7961	2611871	2525435	Alpha	T
HPV type 26	7855	2577095	2525435	Alpha	T
HPV type 10	7919	2598092	2525435	Alpha	T
HPV type 24	7452	2444877	2525435	Beta	**F**
HPV type 7	8027	2633525	2525435	Alpha	T
HPV - 16	7904	2593171	2525435	Alpha	T
Betapapilloma	**Total (nt)**	**LHS VALUE**	**RHS VALUE**	**GROUP**	**T/F**
HPV – type 96	7438	2440284	2525435	Beta	T
HPV type 92	7461	2447830	2525435	Beta	T
HPV type 9	7434	2438971	2525435	Beta	T
HPV type 49	7560	2480310	2525435	Beta	T
HPV - 5	7746	2577751	2525435	Alpha	**F**
HPV - 107	7562	2578735	2525435	Beta	T

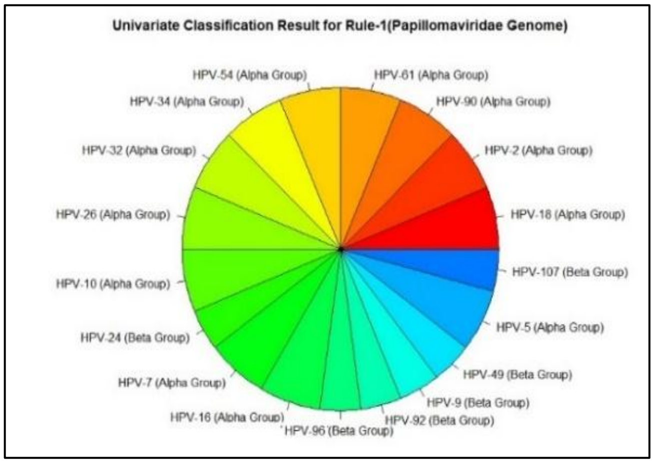

Figure 1: HPV Classification Pie Chart for Univariate Rule-1

Table 4: Result Table for HPV classification by Classification Rule-2 (Multivariate).

Alphapapilloma	**Total (nt)**	**LHS VALUE**	**RHS VALUE**	**GROUP**	**T/F**
HPV - 18	7857	2.0033	0.105086	Alpha	T
HPV - 2	7860	1.5260	0.105086	Alpha	T
HPV type 90	8033	1.9065	0.105086	Alpha	T
HPV - 61	7989	1.8974	0.105086	Alpha	T
HPV - 54	7759	2.0650	0.105086	Alpha	T
HPV type 34	7723	2.7020	0.105086	Alpha	T
HPV type 32	7961	1.7118	0.105086	Alpha	T
HPV type 26	7855	2.0235	0.105086	Alpha	T
HPV type 10	7919	2.3572	0.105086	Alpha	T
HPV type 24	7452	-1.0879	0.105086	Beta	**F**
HPV type 7	8027	2.1938	0.105086	Alpha	T
HPV - 16	7904	2.0529	0.105086	Alpha	T
Betapapilloma	**Total (nt)**	**LHS VALUE**	**RHS VALUE**	**GROUP**	**T/F**
HPV – type 96	7438	-2.0572	0.105086	Beta	T
HPV type 92	7461	-1.3438	0.105086	Beta	T
HPV type 9	7434	-1.0791	0.105086	Beta	T
HPV type 49	7560	-1.4334	0.105086	Beta	T
HPV - 5	7746	-2.6201	0.105086	Beta	T
HPV - 107	7562	-08760	0.105086	Beta	T

Here in univariate classification results, two data points were misclassified, namely, HPV-24 and HPV-5. The classification accuracy is calculated as: T-Classification Percentage (TCP) = (Truly classified / Total sample) *100. T-Classification Percentage = (16/18) *100 = 89%. So, the accuracy level of univariate classification is almost 89%. The multivariate classification rule provides result with almost 95% accurate classification of HPV. One of the simplest and well-known method for evaluating classification performance is Leave One Out Cross Validation (LOOCV). In LOOCV, after excluding the data instance HPV-16 from the data sample, the results of LOOCV were obtained which is as good as expected. When HPV-16 is used as test sample for classification, HPV-16 is also found to be included in its T class label of Alpha papilloma with classification accuracy of almost 95%. Finally, we have evaluated biologically unclassified HPV-78 and it was classified into Alpha-papilloma class.

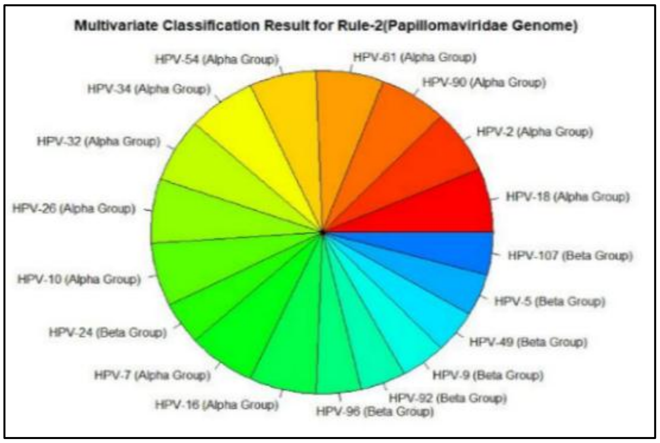

Figure 2: HPV Classification Pie Chart for Multivariate Rule-2

V. CONCLUSION

In this paper, we have explored a statistical procedure based on Fishers Linear discriminant analysis to classify genomes, on the basis of numerical measures as genome and gene sizes. The classification rule or classifier may produce an accurate classification that can be biologically verified. To check the validity of classification rule, a cross validation technique was implemented, which infer the accuracy of the classifier. The study shows that the gene and genome size provide vital information that are biologically significant and their use to classify organism provides very good classification results.

REFERENCE

[1] Dudoit, S., Fridlyand, J. and Speed, T.P. (2002). Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data. Jrnl of the American Statistical Association 97(457), 77-87.

[2] Fisher R.A. (1938) “The statistical utilization of multiple measurements", Annals of Eugenics, 8, 376– 386.

[3] Gnanadesikan, R. (1977). “Methods for Statistical Data Analysis of Multivariate Observations”, Wiley. ISBN 0-471-30845-5 (p. 83–86)

[4] Han, J., Kamber, M. and Pei, J. (2012), Data Mining Concepts and Techniques, Morgan Kaufmann.

[5] Heel, M.V. (1984) “Multivariate statistical classification of noisy images (randomly oriented biological macromolecules)”. Ultramicroscopy Volume 13, Issues 1–2, Pages 165-183.

[6] McLachlan, G. J. (2004). Discriminant Analysis and Statistical Pattern Recognition. Wiley Interscience. ISBN 0-471-69115-1. MR 1190469.

[7] Sihua Peng,Qianghua Xu, Xuefeng Bruce Ling, Xiaoning Peng, Wei Du, Liangbiao Chen. “Molecular classification of cancer types from microarray data using the combination of genetic algorithms and support vector machines”. FEBS letters, 2003 - Wiley Online Library.

[8] T. Ryan Gregory James A. Nicol Heidi Tamm Bellis Kullman Kaur KullmanIlia J. Leitch Brian G. Murray Donald F. Kapraun Johann Greilhuber Michael D. Bennett “Eukaryotic genome size databases”, Nucleic Acids Research, Volume 35, Issue suppl_1, 1 January 2007, Pages D332–D338.

[9] Villiers, E.D., Fauquet, c., Broker, T.R., Bernard HU., “Classification of papillomaviruses”, Virology 324, 2004 – Elsevier pp 17-27.

[10] Yushan Qiu, Xiaoqing Cheng, Wenpin Hou, Wai-Ki Ching. (2015) “On classification of biological data using outlier detection”. 12th International Symposium on Operations Research and its Applications inEngineering, Technology and Management (ISORA 2015).

Authors Profile

Mr. S Swain pursed Bachelor of Science from Berhampur University of Odisha, in 2005, Master of Science in Bioinformatics from Orissa University of Agriculture and Technology, Bhubaneswar, Odisha in year 2007, MCA from IGNOU in the year 2015 and Completed M.Tech in computer Science from Berhampur University in the year 2018. He is currently working as Assistant Professor in Department of Computer Science and Engineering, Vignan Institute of Technology and Management, Berhampur, Odisha since 2011. His main research work focuses on Bioinformatics, Data Analytics, Data Mining, IoT and Computational Intelligence. He has more than 10 years of teaching experience.

Dr M R Pata holds a Ph.D. in Computer Science from the Central University of Hyderabad. He is a Professor in the Department of Computer Science, Berhampur University and has been teaching Computer Science for the last 30 years. He was a United Nations Fellow at the International Institute of Software Technology, United Nations University, Macao. His research interests include Artificial Intelligence, Cloud Computing, Data Mining and E-Governance. He has successfully guided 14 Ph.D. students and has more than 180 research publications to his credit. He has extensively travelled to many countries in USA, Europe, Australia, Africa, and South-East Asia for presenting research papers, chairing technical sessions and delivering invited talks. He has been a member of editorial boards and program committees of many international journals and conferences. He is a member of ACM, and life member of CSI and ISTE.