Experimental Investigation of Frequency Chaos Game Representation for in Silico and Accurate Classification of Viral Pathogens from Genomic Sequences

Emmanuel Adetiba[1,4(✉)], Joke A. Badejo[1], Surendra Thakur[3],
Victor O. Matthews[1], Marion O. Adebiyi[2,4], and Ezekiel F. Adebiyi[2,4]

[1] Department of Electrical and Information Engineering, College of Engineering,
Covenant University, Ota, Nigeria
emmanueladetiba@gmail.com
[2] Department of Computer and Information Science, College of Science
and Technology, Covenant University, Ota, Nigeria
[3] KZN e-Skills CoLab, Durban University of Technology, Durban, South Africa
[4] Covenant University Bioinformatics Research (CUBRe), Ota, Nigeria
emmanuel.adetiba@covenantuniversity.edu.ng

Abstract. This paper presents an experimental investigation to determine the efficacy and the appropriate order of Frequency Chaos Game Representation (FCGR) for accurate and *in silico* classification of pathogenic viruses. For this study, we curated genomic sequences of selected viral pathogens from the virus pathogen database and analysis resource corpus. The viral genomes were encoded using the first to seventh order FCGRs so as to produce training and testing genomic data features. Thereafter, four different kernels of naïve Bayes classifier were experimentally trained and tested with the generated FCGR genomic features. The performance result with the highest average classification accuracy of 98% was returned by the third and fourth order FCGRs. However, due to consideration for memory utilization, computational efficiency vis-à-vis classification accuracy, the third order FCGR is deemed suitable for accurate classification of viral pathogens from genome sequences. This provides a promising foundation for developing genomic based diagnostic toolkit that could be used to promptly address the global incidence of epidemics from pathogenic viruses.

Keywords: Classification · FCGR · Genome · GSP · Naïve Bayes · Pathogens · Sequences · Virus

1 Introduction

Automatic detection of diverse species of viral pathogens associated with emerging deadly ailments within human populations cannot be over-emphasized as they remain a big threat to both personal and public health. Recent advances in molecular biology, next generation sequencing and online bioinformatics platforms offer a vast computational ecosystem for accurate identification of causative viral pathogens associated with the deadly human diseases. While allowing for extensive analysis, the rapidly