




Alignment-Free Z-Curve Genomic Cepstral Coefficients and Machine Learning for Classification of Viruses

Emmanuel Adetiba^{1,2} , Oludayo O. Olugbara³,
Tunmike B. Taiwo³, Marion O. Adebisi⁴, Joke A. Badejo¹,
Matthew B. Akanle¹, and Victor O. Matthews¹

¹ Department of Electrical and Information Engineering, College of Engineering,
Covenant University, Ota, Nigeria

emmanuel.adetiba@covenantuniversity.edu.ng

² HRA, Institute for Systems Science, Durban University of Technology,
P.O. Box 1334, Durban, South Africa

³ ICT and Society Research Group, Durban University of Technology,
P.O. Box 1334, Durban 4000, South Africa

⁴ Department of Computer and Information Science,

College of Science and Technology, Covenant University, Ota, Nigeria

Abstract. Accurate detection of pathogenic viruses has become highly imperative. This is because viral diseases constitute a huge threat to human health and wellbeing on a global scale. However, both traditional and recent techniques for viral detection suffer from various setbacks. In codicil, some of the existing alignment-free methods are also limited with respect to viral detection accuracy. In this paper, we present the development of an alignment-free, digital signal processing based method for pathogenic viral detection named Z-Curve Genomic Cepstral Coefficients (ZCGCC). To evaluate the method, ZCGCC were computed from twenty six pathogenic viral strains extracted from the ViPR corpus. Naïve Bayesian classifier, which is a popular machine learning method was experimentally trained and validated using the extracted ZCGCC and other alignment-free methods in the literature. Comparative results show that the proposed ZCGCC gives good accuracy (93.0385%) and improved performance to existing alignment-free methods.

Keywords: Alignment-free · Bayesian · Classifier · Naïve · Pathogenic Virus · ViPR · ZCGCC

1 Introduction

Novel and re-emerging viruses continue to surface and unleash havoc on human health worldwide. Some of these viruses spread rapidly across the globe and they culminate in high morbidity and mortality. For example, the Severe Acute Respiratory Syndrome (SARS) coronavirus caused a global pandemic in 2003, which resulted in approximately 916 deaths and affected around 30 countries [1]. The most recent outbreak of Ebola Virus Disease (EVD), which was the largest in the history of the

disease, started in December 2013 (a decade after the SARS epidemic) and continued until April 2015 in countries like Southern Guinea, Liberia, Nigeria and Sierra Leone. Reports on EVD indicated that there were a total of 15,052 laboratory confirmed cases and 11,169 deaths [2]. Hence, the prompt and unambiguous detection of pathogenic viruses is of critical importance in order to actively control and prevent viral diseases outbreak.

Next Generation Sequencing (NGS) technologies provide unprecedented opportunities to researchers with respect to the development of new methodologies for viral detection. This is because a plethora of viral genomic sequences from NGS based studies are available in the public domain for unrestricted access by researchers. However, researchers have opined that given the abundant NGS data, the analysis of such data is the most challenging aspect of genomic based viral detection [3]. Thus, this opens up a remarkable opportunity for researchers in the bioinformatics and Genomic Signal Processing (GSP) [4, 5] fields. Genomic Signal Processing (GSP) is an emerging branch of bioinformatics, which involves the use of Digital Signal Processing (DSP) techniques for genomic data analysis and the use of the resultant biological facts to develop system based applications [5].

The traditional methods that were mostly in use to identify the origin of genome sequences are pairwise and multiple sequence alignment. However, sequence alignment method is fraught with difficulties for genome-wide comparative analysis of viruses. This is because there is a high rate of divergence between different virus sequences due to gene mutation, horizontal gene transfer as well as gene duplication, insertion and deletion [8]. Likewise, there is currently no universal oligonucleotide that is present in all viruses, which can be used for homologous searches against public databases to detect viruses [3].

To address the problems in the alignment methods, several alignment-free methods have been developed for viral detection using genomic sequences. These include k-mers methods such as G-C content, dinucleotide composition profile and frequency chaos game representation [9–12, 26]. Another category of alignment-free methods which was recently developed by researchers is the genome space based methods [13, 14]. The Natural Vector (NV) representation and its different variants are representative examples of genome space alignment-free methods [13, 15, 16]. However, the performance accuracy using some of the k-mers and NV methods still leave room for improvement [15, 16, 26].

In the study at hand, we developed GSP-based features named Z-Curve Genomic Cepstral Coefficients (ZCGCC), as an alignment-free method that could be applied for the classification of pathogenic viruses. To evaluate the developed features, we extracted the genomic sequences of twenty six pathogenic viral strains from the Virus Pathogen Database and Analysis Resource (ViPR) corpus [5, 6]. The twenty six viral strains belong to four pathogenic viral species (namely - Enterovirus, Dengue, HepatitisC and Ebola), which are currently attracting global attentions due to their causation of deadly diseases [5]. Different configurations of the naïve Bayes classifier were trained and validated with the ZCGCC. Naïve Bayes classifier was selected for this study because of its attractive physiognomies, which have been widely explored for accurate classification of genomic sequences [7].

2 Materials and Methods

2.1 Dataset

Genomic sequences of twenty six viral strains were extracted from the Virus Pathogen Database and Analysis Resource (ViPR) corpus [6] for this study. The extracted strains belong to four pathogenic viral species namely the Ebolavirus, Dengue virus, Hepatitis C and Enterovirus D68, which have been largely responsible for epidemic disease outbreak. The available strains for each of these species are selected for the study at hand to achieve an elaborate and more robust classification than the study in [5]. The distribution of the extracted data presents a challenge known as imbalance dataset, which is addressed with the random oversampling strategy in this study. Furthermore, there are high variations in sequence length even for samples that belong to the same viral strain. For example, the number of sequences for the Ebola Zaire strain varies from 22 to 19,897 while EnterovirusH varies from 20 to 7,374. These huge differences in the length of nucleotides within the same viral strain clearly illustrate the reason why alignment based and some existing alignment free methods cannot offer accurate viral detection [17]. Thus, this provides the rationale for an investigation of a DSP technique in the current study. In total, 1,948 samples of viral strains were extracted. Since each of the viral strains represent a class in the dataset, our experimentation dataset consequently contains twenty six different classes.

2.2 Z-Curve Genomic Cepstral Coefficients

Deoxyribonucleic Acid (DNA) is a biomolecule that stores the digital information that constitute the genetic blueprint of living organisms [9]. Each nucleotide in a DNA is one of Adenine (A), Cytosine (C), Guanine (G), and Thymine (T). DNA sequence analysis using DSP methods requires mapping of nucleotides to appropriate numbers before any other computational operations can be performed. The selection of the representative numbers affects how well the properties of these nucleotides are reflected for the detection of valuable biological characteristics [18]. The Z-Curve genomic mapping method is selected in this study because of its reported strengths over other competing methods [19, 20, 27, 28]. The steps for computing the ZCGCC being proposed are represented in the block diagram shown in Fig. 1 and the computation procedures are presented subsequently.

Step 1: The first block in Fig. 1 involves the computation of Z-curve from the input nucleotide sequences. Z-curve is a three-dimensional space curve, which constitute a unique numerical representation of a given DNA sequence [19]. A vital advantage of the Z-curve representation over the other nucleotide numerical representation methods is its reproducibility property. This implies that once the coordinate of Z-curve are well defined, the corresponding nucleotides can be uniquely reconstructed [20]. Given a nucleotide sequences that is read from the 5' to the 3' – end with N bases that are inspected from the first base to nth base, the cumulative occurring numbers of each of the bases A, C, G and T are represented by A_n , C_n , G_n and T_n respectively. For points Q_i , $\forall i = 0, 1, 2, \dots, n - 1$ in a 3-D coordinate system, the line that connects the

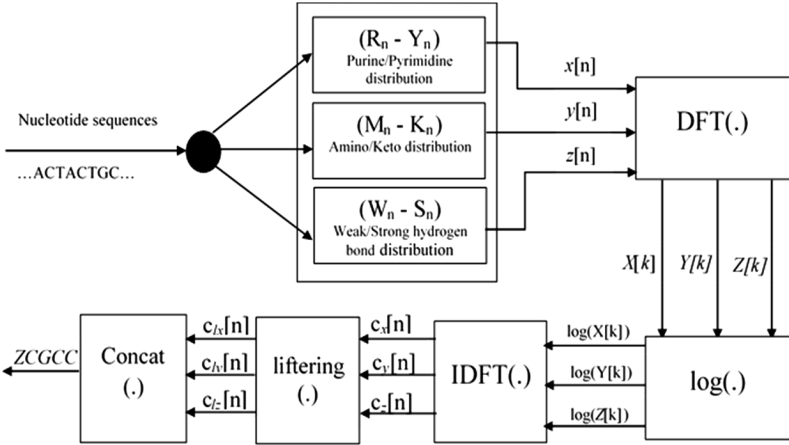


Fig. 1. Functional block diagram of the Z-Curve Genomic Cepstral Coefficients (ZCGCC).

nodes $Q_0(x_0, y_0, z_0)$, $Q_1(x_1, y_1, z_1)$, $Q_2(x_2, y_2, z_2)$, ..., $Q_n(x_n, y_n, z_n)$, in a successive manner is the Z-Curve of the nucleotide sequences being examined. These nodes are mathematically represented as [20, 28]:

$$\begin{cases} x[n] = 2(A_n + G_n) - n & \forall n = 0, 1, 2, \dots, N - 1 \\ y[n] = 2(A_n + C_n) - n \\ z[n] = 2(A_n + T_n) - n \end{cases} \quad (1)$$

where $A_0 = C_0 = G_0 = T_0 = 0$ and $x_0 = y_0 = z_0 = 0$

In order to derive biological meaning from Eq. (1), it is normalized using $A_n + C_n + G_n + T_n = n$, to obtain:

$$\begin{cases} x[n] = (A_n + G_n) - (C_n + T_n) \equiv R_n - Y_n & \forall n = 0, 1, 2, \dots, N - 1 \\ y[n] = (A_n + C_n) - (G_n + T_n) \equiv M_n - K_n \\ z[n] = (A_n + T_n) - (C_n + G_n) \equiv W_n - S_n \end{cases} \quad (2)$$

where R_n , Y_n , M_n , K_n , W_n and S_n are the distributions of the bases of purine, pyrimidine, amino, keto, weak hydrogen bonds and strong hydrogen bonds respectively [21]. The variables $x[n]$, $y[n]$ and $z[n]$ in Eq. 2, which are also illustrated as the outputs of the first block in Fig. 1 are the three independent components of the Z-Curve, with each having distinct biological meaning. Component $x[n]$ represent the distribution of the bases of the purine/pyrimidine (i.e. A or G/C or T) for the first to the n th input nucleotides and it possesses the following attributes:

$$x[n] = \begin{cases} \text{Positive} & \text{if } R_n > Y_n \\ \text{Negative} & \text{if } R_n < Y_n \\ \text{Zero} & \text{if } R_n = Y_n \end{cases} \quad (3)$$

The second component of Z-Curve, which is y_n is the distribution of the bases of the amino/keto group (i.e. A or C/G or T) along the first to n th input nucleotides and it possesses the following attributes:

$$y[n] = \begin{cases} \text{Positive} & \text{if } M_n > K_n \\ \text{Negative} & \text{if } M_n < K_n \\ \text{Zero} & \text{if } M_n = K_n \end{cases} \quad (4)$$

The third component of Z-Curve, z_n is the distribution of the bases of the weak hydrogen bond/strong hydrogen bond (i.e. A or T/C or G) along the first to the n th input nucleotides with the following characteristics:

$$z[n] = \begin{cases} \text{Positive} & \text{if } W_n > S_n \\ \text{Negative} & \text{if } W_n < S_n \\ \text{Zero} & \text{if } W_n = S_n \end{cases} \quad (5)$$

Step 2: The three Z-Curve components computed in the first step, which are streams of digital signals obtained from the input nucleotides are transmitted to the second block in Fig. 1. At this stage, Discrete Fourier Transform (DFT) is applied to the digital signals individually as follows:

$$\begin{cases} X[k] = \sum_{n=0}^{N-1} x[n]e^{-j\frac{2\pi kn}{N}} & \forall k = 0, 1, 2, \dots, N - 1 \\ Y[k] = \sum_{n=0}^{N-1} y[n]e^{-j\frac{2\pi kn}{N}} \\ Z[k] = \sum_{n=0}^{N-1} z[n]e^{-j\frac{2\pi kn}{N}} \end{cases} \quad (6)$$

where $X[k]$, $Y[k]$ and $Z[k]$ are the spectra of the digital signals. The power spectrum, which is a quadratic combination of these spectra were computed for some selected pathogenic viral sequences in this study and the outputs are presented in Sect. 4.

Step 3: Each of the nucleotide spectra computed in the previous step contains peaks which represent the dominant frequency components in the input nucleotide signals. The smooth curve that connects the peaks on a spectrum is referred to as the spectral envelope. The spectral envelope carry the identity of the input nucleotide sequences similar to what obtains in other DSP applications such as speech and mechanical fault diagnosis [22, 23]. The separation of the spectral envelope and spectral details from the spectrum is referred to as cepstral analysis. The required procedure for cepstral analysis are represented with the third, fourth and fifth blocks in Fig. 1 and mathematically depicted as follows:

$$\begin{cases} c_x[n] = \sum_{k=0}^{N-1} \log(X[k]) e^{j\frac{2\pi kn}{N}} \\ c_y[n] = \sum_{k=0}^{N-1} \log(Y[k]) e^{j\frac{2\pi kn}{N}} \\ c_z[n] = \sum_{k=0}^{N-1} \log(Z[k]) e^{j\frac{2\pi kn}{N}} \end{cases} \quad (7)$$

Using Euler's formulae, Eq. (7) becomes:

$$\begin{cases} c_x[n] = \sum_{k=0}^{N-1} \log(X[k]) \cos\left(\frac{2\pi kn}{N}\right) + j \sum_{k=0}^{N-1} \log(X[k]) \sin\left(\frac{2\pi kn}{N}\right) \\ c_y[n] = \sum_{k=0}^{N-1} \log(Y[k]) \cos\left(\frac{2\pi kn}{N}\right) + j \sum_{k=0}^{N-1} \log(Y[k]) \sin\left(\frac{2\pi kn}{N}\right) \\ c_z[n] = \sum_{k=0}^{N-1} \log(Z[k]) \cos\left(\frac{2\pi kn}{N}\right) + j \sum_{k=0}^{N-1} \log(Z[k]) \sin\left(\frac{2\pi kn}{N}\right) \end{cases} \quad (8)$$

real cepstrum *complex cepstrum*

where each of $c_x[n]$, $c_y[n]$ and $c_z[n]$ represents the complex Z-Curve cepstrum of the x [n], y [n] and z [n] components of the Z-Curve for the input nucleotides respectively. The complex cepstrum is a combination of the real and imaginary cepstrum as shown in Eq. (8). The real cepstrum is the log magnitude spectrum of each of the respective signals while the imaginary cepstrum is the phase components. The spectral envelope and spectral details are captured in the real cepstrum. It should be noted that the word ‘‘cepstrum’’ was coined by reversing the first syllable of ‘‘spectrum’’. Hence, in the cepstrum domain, quefrequency also stands for frequency and lifter is used in place of filter [22]. The spectral envelope is the low quefrequency components while the spectral details are the high quefrequency components in the cepstrum domain. Authors in other DSP application domains have reported that the first 15 or 20 coefficients of a cepstrum appositely represent the spectral envelope [24]. As depicted with the fifth block of Fig. 1, the first 15 or 20 coefficients (spectral envelope) of the real cepstrum are liftered using the window:

$$w[n] = \begin{cases} 1, & 0 \leq n \leq L \\ 0, & \textit{elsewhere} \end{cases} \quad (9)$$

where L is the cut off length of the liftering window, which can be either 15 or 20 as earlier stated. The liftering window in Eq. (9) is multiplied with each of the real cepstra sections of Eq. (8) to obtain:

$$\begin{cases} c_{lx}[n] = w[n] \cdot c_x[n] \\ c_{ly}[n] = w[n] \cdot c_y[n] \\ c_{lz}[n] = w[n] \cdot c_z[n] \end{cases} \quad (10)$$

where $c_{lx}[n]$, $c_{ly}[n]$ and $c_{lz}[n]$ are the low quefrequency coefficients of $c_x[n]$, $c_y[n]$ and $c_z[n]$ respectively.

Step 4: In the final step depicted with the last block of Fig. 1, the low quefrequency cepstral coefficients obtained from Step 3 are concatenated to obtain the Z-Curve Genomic Cepstral Coefficients (ZCGCC) in this study. The ZCGCC is a compact genomic feature vector, which represent the distribution of the dominant components of the bases of purine, pyrimidine, amino, keto, weak and strong hydrogen bonds in the input nucleotide sequences. The ZCGCC feature vector is therefore an alignment-free identity of the input nucleotide sequences and it can either be 45 or 60 elements in length depending on if L in Eq. (9) is 15 or 20 respectively. Naïve Bayesian classifier hereafter in this study to determine the discriminatory potency of ZCGCC when it is applied to extract features from the pathogenic viral dataset.

2.3 Experiments

In this study, three experiments were carried out on a PC with an Intel Core i5 CPU, of 2.50 GHz speed, 6.00 GB RAM, and runs 64-bit Windows 8 operating system. In all the experiments, the forty five and sixty element ZCGCC were utilized and their performances were compared using appropriate metrics. In the first experiment, the naïve Bayes classifier was trained with the ZCGCC extracted from the imbalance dataset. In the second experiments, random oversampling was applied to obtain a balanced dataset. The random oversampling strategy involves the addition of instances to the minority class in a random manner [25]. Since the highest number of instances for any class in the dataset is 100 (Table 1), we increased the number of instances for all the minority classes (instances < 100) in the dataset to 100 to obtain the balanced dataset. The ZCGCC feature vectors extracted from the balanced dataset were further used to train the naïve Bayes classifier. The third experiment involved the comparison of the variant of ZCGCC that gave the best result in the second experiment using the balanced dataset with two other alignment free methods in the literature, namely, Electron Ion Interaction Pseudopotential – Genomic Cepstral Coefficient (EIIP-GCC) [5] and Frequency Chaos Game Representation (FCGR) [26].

3 Results and Discussion

3.1 Power Spectrums of the Z-Curve Encoded Viruses

Figure 2 shows the distinct power spectrums of the different strains of Enterovirus, HepatitisC, Dengue and Ebola viruses. Similar to the illustrations in Fig. 2, previous

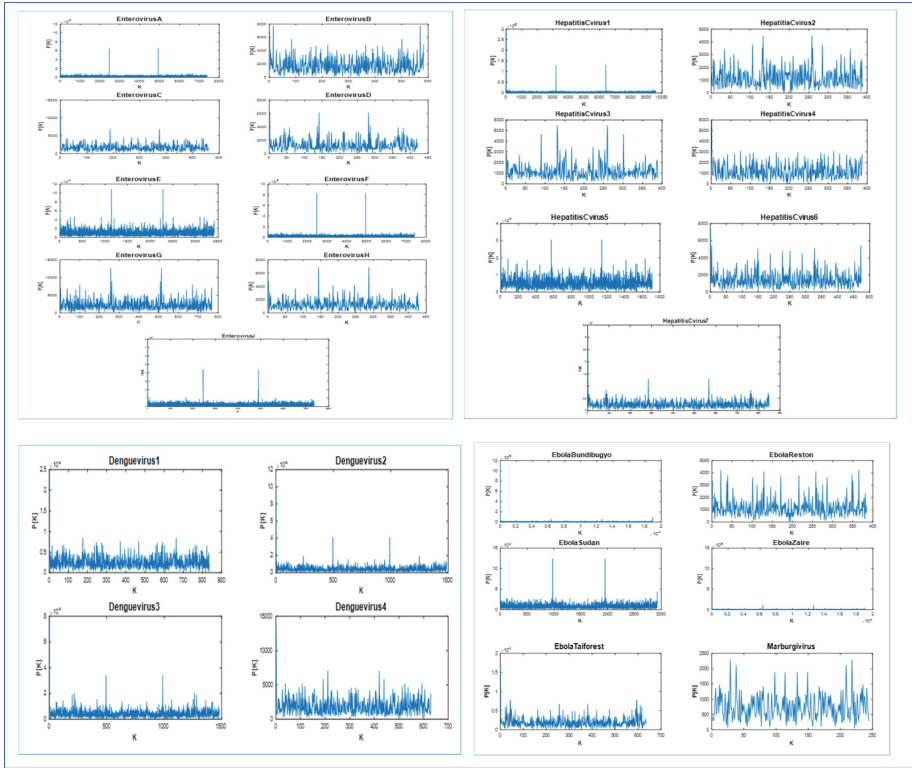


Fig. 2. Power spectrums of Z-Curve encoded Enterovirus, HepatitisC, Dengue and Ebola viruses.

studies have also utilized power spectral of Z-Curve to graphically illustrate the mitochondria DNA of homo sapiens [27] and lung cancer biomarker genes [28, 29].

3.2 Classifier Training Results

The results of the first experiment in which the imbalanced dataset was investigated are shown in Table 1. Four different naïve Bayes kernel functions were tested, namely Gaussian, uniform, epanechnikov and triangular [30]. The sixty element ZCGCC gave higher accuracies and low Misclassification Errors (ME) for each of the kernel functions. Meanwhile, the triangular function ranked best (accuracy = 91.2218%, ME = 0.0878) for the sixty element ZCGCC. Two-sample t-test was further utilized to investigate if the difference between the forty five and sixty element ZCGCC is statistically significant. The test statistic indicates that the null hypothesis of no difference between the mean of the two sets of accuracies is rejected, $p < 0.05$ ($p = 0.0278$) as well as for the two sets of MEs, $p < 0.05$ ($p = 0.0280$). This shows that the performance of the sixty element ZCGCC is significantly better than that of the forty five element ZCGCC for the imbalanced dataset.

Table 1. Experimental results of the imbalanced dataset with ZCGCC.

Kernel function	ZCGCC (45 elements)		ZCGCC (60 elements)	
	Accuracy (%)	ME	Accuracy (%)	ME
Triangular	89.5277	0.1047	91.2218	0.0878
Gaussian	89.0144	0.1099	90.6571	0.0934
Epanechnikov	87.7823	0.1222	90.1437	0.0986
Uniform	87.1150	0.1289	89.3224	0.1068

Table 2 shows the results of the second experiment in which the balanced dataset obtained through random oversampling was used to train the naïve Bayes classifier. The sixty element ZCGCC also gave higher accuracies and lower MEs for all the kernel functions compare to its 45 elements counterpart. Similar to the first experiment, the triangular kernel function gave the best overall performance result for the sixty element ZCGCC (accuracy = 93.0385%, ME = 0.0696).

It is also remarkable that the performance results of the ZCGCC for the balanced dataset in the second experiment are better than the corresponding ZCGCC in the first experiment for all the kernel functions. This shows that random oversampling method positively influenced the performance results of the ZCGCC. Since the sixty element ZCGCC gave superior performances in the first and second experiments over the forty element ZCGCC, we further investigated if the improvement of the sixty element ZCGCC for the balanced dataset (second experiment) over the sixty element ZCGCC for the imbalanced dataset (first experiment) is statistically significant. The null hypothesis of no difference between the two sets of accuracies is rejected because $p < 0.05$ ($p = 0.0122$) and the null hypothesis of no difference between the mean of the two sets of MEs is also rejected, $p < 0.05$ ($p = 0.0122$). Thus, the performance results of the sixty element ZCGCC using the balanced dataset is significantly better than those for the imbalanced dataset.

Thus, the sixty element ZCGCC is proposed as an alignment free method for viral pathogen detection in this study based on its overall best performance.

Table 2. Experimental results of the balanced dataset with ZCGCC

Kernel function	ZCGCC (45 elements)		ZCGCC (60 elements)	
	Accuracy (%)	ME	Accuracy (%)	ME
Triangular	91.9615	0.0804	93.0385	0.0696
Gaussian	91.6538	0.0835	92.7308	0.0727
Uniform	90.6923	0.0937	91.2308	0.0877
Epanechnikov	90.6154	0.0938	92.3462	0.0765

The third experiment was carried out to compare the proposed alignment free method in this study (i.e. sixty element ZCGCC) with two other alignment free methods in the literature, namely EIIP-GCC [6] and FCGR [26]. Table 3 shows the results of the third experiment for EIIP-GCC and FCGR using the balanced dataset.

We deem it adequate to use the balanced dataset for the comparison in this third experiment since it produced the best result for the proposed alignment free method in the second experiment. The performance results of the proposed sixty element ZCGCC in Table 2 for all the kernel functions are better than those of EIIP-GCC in Table 3 for all the corresponding kernel functions. For instance, the triangular kernel function gave the highest accuracy of 93.0385% (ME = 0.0696) for the ZCGCC whereas the accuracy obtained with the triangular kernel function for the EIIP-GCC was 84.5% (ME = 0.1550). Furthermore, the statistical significance of the improvement in the performance of the proposed ZCGCC over EIIP is statistically significant, $p < 0.05$ ($p = 8.82e-06$).

The performance result of the proposed ZCGCC in Table 2, which was obtained using the triangular kernel function is also slightly better than the highest performance result of the FCGR (accuracy = 92.9231%, ME = 0.0708).

Table 3. Experimental results of the balanced dataset with EIIP-GCC and FCGR

Kernel function	EIIP-GCC		FCGR	
	Accuracy (%)	ME	Accuracy (%)	ME
Epanechnikov	84.6154	0.1538	92.9231	0.0708
Triangular	84.5000	0.1550	92.6923	0.0731
Uniform	83.1154	0.1688	92.3846	0.0762
Gaussian	82.7308	0.1727	91.8846	0.0812

It can be inferred from the results obtained in this study that the first 20 elements of the real cepstrum is more representative of the spectral envelope for the genomic signal. A previous study reported the development of ZCURVE_V, which is a gene finding application for viruses using DNA sequences and the Z-Curve mathematical paradigm. The authors reported that ZCURVE_V can accurately predict genes in viral genomes as short as about 1000 nucleotides [19]. However, the alignment free ZCGCC method proposed in this study detect viral genomes of both long and short lengths with accuracy that compares favorably with existing alignment-free methods in the literature.

4 Conclusion

We have successfully reported the development of ZCGCC, which is an alignment-free method for virus detection in this paper. The sixty element ZCGCC gave superior performance to the EIIP-GCC and comparable performance to FCGR. However, ZCGCC provides remarkable advantages such as low dimension, global genome analysis and low computational requirements, which make it a promising method for developing diagnostic tool for detection of pathogenic viral diseases. Future works will include an investigation of the ZCGCC for the detection of other organisms in the prokaryotic and eukaryotic domains of life. We also hope to experiment with other machine learning methods to investigate the possibility of improved performance.

Acknowledgment. Funding to present this work at IWBBIO 2018 was provided by the Covenant University Centre for Research, Innovation and Development, Canaanland, Ota, Nigeria.

References

1. Xie, G., Yu, J., Duan, Z.: New strategy for virus discovery: viruses identified in human feces in the last decade. *Sci. China Life Sci.* **56**(8), 688–696 (2013)
2. Kaushik, A., Tiwari, S., Jayant, R.D., Marty, A., Nair, M.: Towards detection and diagnosis of Ebola virus disease at point-of-care. *Biosens. Bioelectron.* **75**, 254–272 (2016)
3. Mokili, J.L., Rohwer, F., Dutilh, B.E.: Metagenomics and future perspectives in virus discovery. *Curr. Opin. Virol.* **2**(1), 63–77 (2012)
4. Mabrouk, M.S.: A study of the potential of EIPP mapping method in exon prediction using the frequency domain techniques. *Am. J. Biomed. Eng.* **2**(2), 17–22 (2012)
5. Sathish Kumar, S., Duraipandian, N.: An effective identification of species from DNA sequence: a classification technique by integrating DM and ANN. *Int. J. Adv. Comput. Sci. Appl.* **3**(8), 104–114 (2012)
6. Adetiba, E., Olugbara, O.O., Taiwo, T.B.: Identification of pathogenic viruses using genomic cepstral coefficients with radial basis function neural network. In: Pillay, N., Engelbrecht, A.P., Abraham, A., du Plessis, M.C., Snášel, V., Muda, A.K. (eds.) *Advances in Nature and Biologically Inspired Computing. AISC*, vol. 419, pp. 281–291. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-27400-3_25
7. Pickett, B.E., Greer, D.S., Zhang, Y.: Virus pathogen database and analysis resource (ViPR): a comprehensive bioinformatics database and analysis resource for the coronavirus research community. *Viruses* **4**, 3209–3226 (2012)
8. Wang, Q., Garrity, G.M., Tiedje, J.M., Cole, J.R.: Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* **73**(16), 5261–5267 (2007)
9. Li, Y., Tian, K., Yin, C., He, R.L., Yau, S.S.T.: Virus classification in 60-dimensional protein space. *Mol. Phylogenet. Evol.* **99**, 53–62 (2016)
10. Vinga, S., Almeida, J.: Alignment-free sequence comparison—a review. *Bioinformatics* **19**, 513–523 (2003). <https://doi.org/10.1093/bioinformatics/btg005>
11. Kantorovitz, M.R., Robinson, G.E., Sinha, S.: A statistical method for alignment-free comparison of regulatory sequences. *Bioinformatics* **23**(13), i249–i255 (2007)
12. Dai, Q., Yang, Y., Wang, T.: Markov model plus k-word distributions: a synergy that produces novel statistical measures for sequence comparison. *Bioinformatics* **24**(20), 2296–2302 (2008)
13. Sims, G.E., Jun, S.R., Wu, G.A., Kim, S.H.: Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. *Proc. Natl. Acad. Sci.* **106**(8), 2677–2682 (2009)
14. Deng, M., Yu, C., Liang, Q., He, R.L., Yau, S.S.T.: A novel method of characterizing genetic sequences: genome space with biological distance and applications. *PLoS One* **6**(3), e17293 (2011)
15. Yu, C., Liang, Q., Yin, C., He, R.L., Yau, S.S.T.: A novel construction of genome space with biological geometry. *DNA Res.* **17**, 155–168 (2010)
16. Yu, C., Hernandez, T., Zheng, H., Yau, S.C., Huang, H.H., He, R.L., Yau, S.S.T.: Real time classification of viruses in 12 dimensions. *PLoS One* **8**(5), e64328 (2013)

17. Huang, H.H., Yu, C., Zheng, H., Hernandez, T., Yau, S.C., He, R.L., Yau, S.S.T.: Global comparison of multiple-segmented viruses in 12-dimensional genome space. *Mol. Phylogenet. Evol.* **81**, 29–36 (2014)
18. Anastassiou, D.: DSP in genomics: processing and frequency-domain analysis of character strings. In: Proceedings of 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP 2001), vol. 2, pp. 1053–1056. IEEE (2001)
19. Bai Arniker, S., Kwan, H.K.: Advanced numerical representation of DNA sequences. In: International Conference on Bioscience, Biochemistry and Bioinformatics IPCBEE, vol. 3, p. 1 (2012)
20. Guo, F.B., Lin, Y., Chen, L.L.: Recognition of protein-coding genes based on Z-curve algorithms. *Curr. Genomics* **15**(2), 95–103 (2014)
21. Zhang, R., Zhang, C.T.: A brief review: the z-curve theory and its application in genome analysis. *Curr. Genomics* **15**(2), 78–94 (2014)
22. Cornish-Bowden, A.: Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. *Nucleic Acids Res.* **13**(9), 3021 (1985)
23. Randall, R.B.: A history of cepstrum analysis and its application to mechanical problems. In: International Conference at Institute of Technology of Chartres, France, pp. 11–16 (2013)
24. Thakur, S., Adetiba, E., Olugbara, O.O., Millham, R.: Experimentation using short-term spectral features for secure mobile internet voting authentication. *Math. Probl. Eng.* (2015)
25. Sakshat Virtual Labs: Cepstral Analysis of Speech (2011). iitg.vlab.co.in/?sub=59&brch=164&sim=615&cnt=1. Accessed 28 July 2016
26. Adetiba, E., Badejo, J.A., Thakur, S., Matthews, V.O., Adebisi, M.O., Adebisi, E.F.: Experimental investigation of frequency chaos game representation for in silico and accurate classification of viral pathogens from genomic sequences. In: Rojas, I., Ortuño, F. (eds.) IWBBIO 2017. LNCS, vol. 10208, pp. 155–164. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-56148-6_13
27. Vijayan, K., Nair, V.V., Gopinath, D.P.: Classification of organisms using frequency-chaos game representation of genomic sequences and ANN. In: 10th National Conference on Technological Trends (NCTT 2009), pp. 6–7 (2009)
28. Shao, J., Yan, X., Shao, S.: SNR of DNA sequences mapped by general affine transformations of the indicator sequences. *J. Math. Biol.* **67**(2), 433–451 (2013)
29. Adetiba, E., Olugbara, O.O.: Improved classification of lung cancer using radial basis function neural network with affine transforms of Voss representation. *PLoS One* **10**(12), e0143542 (2015)
30. Mathworks, Classification Naive Bayes class. <http://www.mathworks.com/help/stats/classificationnaivebayes-class.html>. Accessed 28 July 2016