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

I. Rojas and F. Ortuño (Eds.): IWBBIO 2017, Part I, LNBI 10208, pp. 155–164, 2017.
DOI: 10.1007/978-3-319-56148-6_13

growing databases of genomic sequences also provide an avalanche of resources for improved epidemic surveillance, diagnostics and therapeutics towards promoting healthy living. Furthermore, newer digital signal processing-based bioinformatics methods utilize numerical and/or visual encoding of nucleotide sequences collected from laboratory and environmental surveillances for effective non-alignment analysis [1, 2].

The application of digital signal processing techniques to genomic analysis, coined Genomics Signal Processing (GSP), requires that the nucleotide sequences be encoded numerically or graphically for alignment-free sequence comparison [1, 3, 4]. Next, discriminatory genomic features are extracted from the numeric genome representations to improve on species- or genome-level classification, usually based on a machine learning technique [5]. GSP-based techniques provide alignment-free analyses of the genomes to address the problems of unequal lengths of the sequences, the computational speed and large memory requirements encountered during alignment-based analysis [6, 7]. However for an accurate detection, it is necessary to ensure that the numeric or visual encoding of the nucleotide sequences represents the unique and salient characteristic of the genome as desirable.

Unlike other methods, the Chaos Game Representation (CGR) visually expresses the local patterns of the nucleotide sequences and hence the global structure of the genome in a two-dimensional graphical form [2, 8]. CGR is a scale independent representation developed by Jeffrey [9]. It was derived from the chaos theory, which allows the illustration of frequencies of oligonucleotides in the form of images. With CGR, the oligonucleotides of a genome exhibit the main physiognomies of the whole genome [7]. However, in the original form, CGR is not convenient for processing with a computer, hence, another form of CGR named Frequency Chaos Game Representation (FCGR) was introduced [7, 8, 10]. The CGR pattern of the nucleotide sequences of the same genome are found to be similar but differs quantitatively from the CGR patterns of the genome from another specie. This biological attribute makes the unique genomic signature of CGR and the subsequent features extracted from it, an accurate representation for alignment-free analysis suitable for classification, clustering and identification as proposed in many researches reported recently.

Karamichalis et al. [5] investigated the intra-specie and inter-specie variations of the genomic signatures generated by CGR patterns, using six different distance measures. The study validated the hypothesis that the CGR patterns of the nucleotide sequences of the same genome are similar but differs quantitatively from the CGR patterns of the genome from another specie. The CGR-based genomic signatures also accurately classified the genomic DNA sequences of Homo sapiens and Mus musculus genomes at lower taxonomic levels – class and order.

Messaoudi et al. [11] encoded the genomic sequence of Caenorhabditis elegans (C. elegans) with frequency of CGR patterns, otherwise called Frequency Chaos Game Representation (FCGR), for a time-frequency investigation using the Continuous Wavelet Transform (CWT). The complex Morlet wavelet based CWT revealed significant biological characteristics from the genomic signature of the FCGR patterns.

Kari et al. [12] proposed a molecular distance map developed with the unique genomic signature of CGR suitable for defining relationships between species to identify species, clarify taxonomies and related evolutionary history. Multi-Dimensional Scaling

(MDS) was applied to the distance metrics computed based on the Structural Dissimilarity Index (DSSIM) to produce the map. The map successfully characterized organisms into several taxonomy levels within the Euclidean space that showed the spatial proximity between the nucleotide sequences.

Tanchotsrinon et al. [13] adopted the CGR and Singular Value Decomposition (SVD) for Human Papillomavirus (HPV) genotyping as an approach to fight cervical cancer. Two classes of features were obtained from the SVD-reduced matrices of the original CGR: ChaosCentroid, which captured the structure of the sequences and ChaosFrequency, which represented relevant statistical distribution of nucleotides in the sequences. Their study demonstrated comparative results with no significant difference between their proposed method and the NCBI viral genotyping tool irrespective of the four classification techniques used i.e. Multi-layer Perceptron, Radial Basis Function, K-Nearest Neighbor and Fuzzy K-Nearest Neighbor.

In the current study, we experimentally explored the applicability of FCGR and its appropriate order for classification of viral pathogens from genomic sequences into the right species. This endeavor is aimed at laying a foundation for the development of an alternative, accurate and in silico genomic viral diagnostic tool, which could help in rapid medical interventions in the event of viral pathogens epidemic.

2 Materials and Methods

2.1 Dataset

As shown in Table 1, we extracted the genome sequences of Ebola virus (N = 249), Enterovirus (N = 632), Dengue virus (N = 390), HepatitisC virus (N = 567) and Zika virus (N = 351) from the Virus Pathogen Database and Analysis Resource (ViPR) corpus. This corpus was developed to provide free access to genomic and proteomic sequences of viral pathogens for research and development of vaccines, therapies and diagnostic tools. The Universal Resource Locator (URL) for ViPR as at the time this study was carried out is https://www.viprbrc.org/brc/home.spg?decorator=vipr. The total sample size of the dataset extracted for this study is 2,189. Although, there is a huge collection of pathogenic viral datasets on the corpus, the five viruses were selected due to their prominence as causative agents of diseases that is currently of concern among researchers on a global scale. These viruses are also specifically featured on the home page of the ViPR corpus and the structural diversity of their genomes provide a good basis to investigate the efficacy of FCGR for viral species classification.

Table 1. Extracted dataset for five pathogenic viruses

S/N	Viral species	Number of unique samples
1	Ebola virus	249
2	Enterovirus	632
3	Dengue virus	390
4	HepatitisC virus	567
5	Zika virus	351
	Total	2,189

2.2 FCGR Computation at Different Orders and Naïve Bayes Classifier

FCGR is a numerical matrix in contrast to CGR, which is a graphical representation. Instead of plotting a CGR first and converting it to a FCGR [7, 8]. Wang et al. [10] posits that FCGR can be derived directly from a sequence. Furthermore, Wang et al. [10] introduced the concept of FCGR order, which provides variants in matrix dimensions when FCGR are derived directly from the sequences. For instance, given a sequence S, with f_w representing the frequency of the oligonucleotide w, the matrix structure of a first order FCGR is given as Eq. (1) [10].

$$FCGR_1(S) = \begin{pmatrix} f_C & f_G \\ f_A & f_T \end{pmatrix} \tag{1}$$

The FCGR of (k + 1)th order can be computed by substituting each element f_x in a kth order FCGR with the four elements

$$\begin{pmatrix} f_{CX} & f_{GX} \\ f_{AX} & f_{TX} \end{pmatrix}. \tag{2}$$

Therefore, the matrix structure of a second order FCGR is as shown in Eq. (2) and higher order FCGR can be sequentially computed.

$$FCGR_2(S) = \begin{pmatrix} f_{CC} & f_{GC} & f_{CG} & f_{GG} \\ f_{AC} & f_{TC} & f_{AG} & f_{TG} \\ f_{CA} & f_{GA} & f_{CT} & f_{GT} \\ f_{AA} & f_{TA} & f_{AT} & f_{TT} \end{pmatrix} \tag{3}$$

From Eqs. (1) and (3), it can be seen that a k-th order FCGR is a 2^k x 2^k matrix and it contains 4^k occurrences of the k length oligonucleotides [10, 14]. The direct correspondence of CGR and FCGR, in which a kth order FCGR is equivalent to a CGR of resolution $1/2^k$ was also reported [10]. This makes it possible to observe the major features that are inherent in higher order FCGR (which ordinarily is incomprehensible because of the size) by visual observation of the equivalent CGR. Researchers have also opined that CGR images and correspondingly the FCGR obtained from subsequence of a genome present similar structure as the whole genome [7]. This implies that the CGR image or FCGR of a subsequence is a sufficient genomic signature for species classification rather than the CGR image or FCGR of the whole genome [7, 10]. Therefore, in this study, we ventured to experimentally investigate the efficacy of FCGR at different resolutions or orders ($1 \leq k \leq 7$) for pathogenic virus species classification. We stopped at 7th order because the huge dimension of the matrix elements at 8th order and beyond is computationally expensive without providing any benefits with respect to classification accuracy.

The accurate classification of the viral species from the FCGR-encoded nucleotide sequences was carried out with the Naïve Bayes (NB) classifier, which is a very popular classifier in bioinformatics [15, 16]. NB classifier utilizes a key statistical assumption of conditional independence of the FCGR features within the same class, to

assign a class label to each sequence [15]. The class label in this context refers to the viral species, drawn from Table 1. The NB classifier can be trained using different kernel functions such as uniform, epanechnikov, normal and triangular to detect through classification the most probable viral specie from each encoded sequence.

2.3 Experiments

Experiments were performed in this study to determine the efficacy as well as the appropriate order of FCGR for classifying viral pathogens from genomic sequences. The curated sequences for the five viruses were first converted to their numeric equivalents with 1st to 7th order of FCGR. For each of these orders, the FCGR encoded viral sequences were transmitted to train Naïve Bayes classifier using four different kernel functions, namely; uniform, epanechnikov, normal and triangular. Both the FCGR algorithm and naïve Bayes classifier were implemented in MATLAB R2015a, which also provided the in silico platform to perform all the experiments in this study. The PC on which the experiments were performed contains an Intel Core i5-4210U CPU operating at 2.40 GHz speed, with 8.00 GB RAM and runs 64-bit Windows 8 operating system.

3 Results and Discussion

Figure 1 shows the plot of the number of elements in the computed FCGR matrices against the FCGR order. As illustrated on the graph, a first order FCGR matrix contains 4 elements, a second order contains 16 elements, third order contains 64 elements,

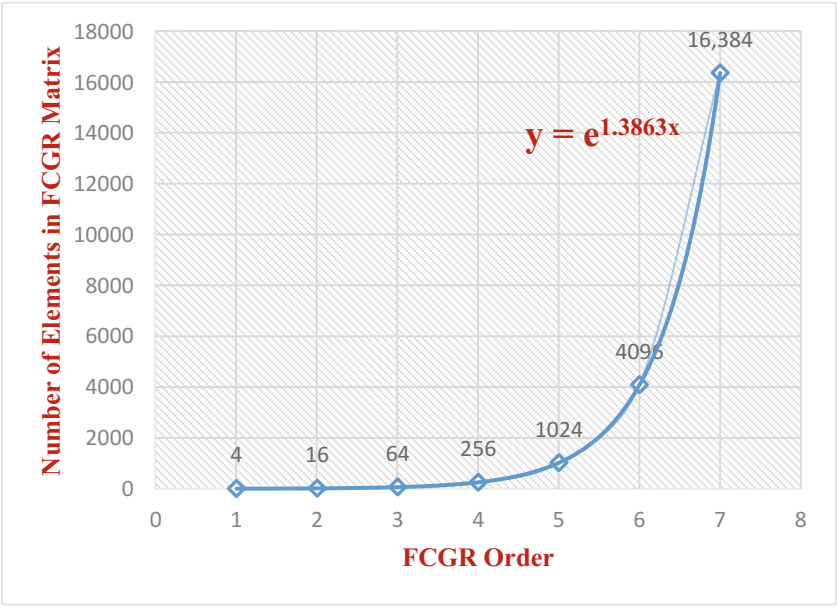

Fig. 1. The number of elements in the FCGR matrix against the FCGR order

fourth order contains 256 elements, fifth order contains 1024 elements, sixth order contains 4096 elements and seventh order contains 16,384 elements. The relationship between the number of elements in the FCGR matrix and the number of FCGR order is clearly an exponential growth which is represented as:

$$y = e^{1.3863x} \tag{4}$$

where y is the number of elements in the FCGR matrix and x is the FCGR order. The results of the experiments, which are hereafter reported provide an insight on the effects of FCGR order vis-à-vis the number of elements in the corresponding FCGR matrices on pathogenic viral classification accuracy.

Tables 2, 3, 4, 5, 6, 7, and 8 show the results we obtained when the first to seventh order FCGR matrices were respectively utilized to encode the viral genomic sequences. We deemed it expedient to compute the average classification accuracies and average misclassification errors for the four different kernels across the FCGR orders. This provides a compact scheme for the comparison of our results based on the FCGR orders. The average classification results for the first order FCGR is shown in Table 2 (Accuracy = 89.8161%, ME = 0.1018). About 7% increase in performance was obtained for the second order FCGR (Accuracy = 96.6281%, ME = 0.0337) over the first order. Furthermore, increase in performance continued with the third order (Accuracy = 98.3651%, ME = 0.0163) up to the fourth order (Accuracy = 98.1607%, ME = 0.0184). There is however a drastic reduction in the performance results for the fifth order (Accuracy = 94.2098%, ME = 0.0579), which continued for the sixth order (Accuracy = 85.7857%, ME = 0.1422) and the lowest performance results in this study was posted for the seventh order FCGR (Accuracy = 78.0654%, ME = 0.2194). It is clearly apparent that approximately, both the third and fourth FCGR order with 64

Table 2. First order FCGR

S/N	Naïve Bayes kernel function	Accuracy	Misclassification error (ME)
1	Uniform	89.7366	0.1026
2	Epanechnikov	89.7366	0.1026
3	Normal	89.8274	0.1017
4	Triangular	89.9637	0.1004
Average		**89.8161**	**0.1018**

Table 3. Second order FCGR

S/N	Naïve Bayes kernel function	Accuracy	Misclassification error (ME)
1	Uniform	96.5486	0.0345
2	Epanechnikov	96.5940	0.0341
3	Normal	96.5032	0.0350
4	Triangular	96.8665	0.0313
Average		**96.6281**	**0.0337**

Table 4. Third order FCGR

S/N	Naïve Bayes kernel function	Accuracy	Misclassification error (ME)
1	Uniform	98.4105	0.0159
2	Epanechnikov	98.4559	0.0154
3	Normal	98.1381	0.0186
4	Triangular	98.4559	0.0154
Average		**98.3651**	**0.0163**

Table 5. Fourth order FCGR

S/N	Naïve Bayes kernel function	Accuracy	Misclassification error (ME)
1	Uniform	98.7284	0.0127
2	Epanechnikov	98.4559	0.0154
3	Normal	97.0027	0.0300
4	Triangular	98.4559	0.0154
Average		**98.1607**	**0.0184**

Table 6. Fifth order FCGR

S/N	Naïve Bayes kernel function	Accuracy	Misclassification error (ME)
1	Uniform	95.5495	0.0445
2	Epanechnikov	96.9119	0.0309
3	Normal	87.5568	0.1244
4	Triangular	96.8211	0.0318
Average		**94.2098**	**0.0579**

Table 7. Sixth order FCGR

S/N	Naïve Bayes Kernel function	Accuracy	Misclassification error (ME)
1	Uniform	85.6494	0.1435
2	Epanechnikov	93.2788	0.0672
3	Normal	71.4805	0.2852
4	Triangular	92.7339	0.0727
Average		**85.7857**	**0.1422**

Table 8. Seventh order FCGR

S/N	Naïve Bayes kernel function	Accuracy	Misclassification error (ME)
1	Uniform	82.8792	0.1712
2	Epanechnikov	80.2906	0.1971
3	Normal	66.6213	0.3338
4	Triangular	82.4705	0.1753
Average		**78.0654**	**0.2194**

and 256 elements respectively, gave the highest performance results (Approximate Accuracy = 98%, Approximate ME = 0.02). The summary of the entire result is graphically represented in Fig. 2.

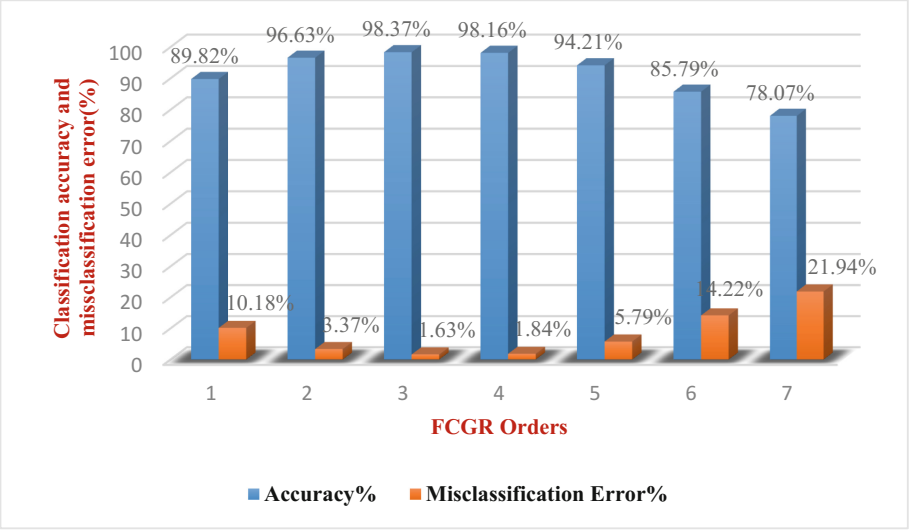

Fig. 2. Summary of the average classification accuracy and misclassification error.

The results in this study clearly agree with the curse of dimensionality philosophy in machine learning, in which too small training data features (first and second order FCGR in this context) may hinder the creation of a reliable classification model for assigning a class to all possible objects in the dataset. Conversely, high dimensions in the training features (sixth and seventh order FCGR in this context) tend to make the contiguity among data points more identical and often lead to lower classification accuracy. Training data with high features has also being reported to lead to high computational cost and memory usage [17], which is the case with the eight order FCGR in this study that motivated its exclusion from the experiments.

Our literature search while undertaking this research yielded few studies that have employed FCGR and other schemes for identification of species ([1, 18–21]. The study by Vijayan et al. [18] utilized a third order FCGR (64 element vector) to encode Eukaryotic organisms with Probabilistic Neural Network (PNN) as a classifier to obtain a classification accuracy of 92.3%. In codicil, the study reported in [1] where a 15-element real Genomic Cepstral Coefficients (GCC) with Radial Basis Function Neural Network (RBFNN) were utilized for identification of four pathogenic viruses gave an accuracy of 97.3%. Obviously, the current result of 98% classification accuracy for five pathogenic viruses with 64 (third order) and 256 (fourth order) elements FCGR and naïve Bayes classifier is comparable to the highlighted similar results in the literature. However, based on the experimental results obtained in this study, the third

order FCGR is recommended as an appropriate genomic feature for pathogenic viral classification. This will appositely culminate in economy of memory space, computational efficiency and acceptable accuracy for viral pathogens classification, which is an important contribution albeit moderate, to GSP and bioinformatics body of knowledge.

4 Conclusion

Thus far, we have been able to achieve the objectives of the current study, which are to determine the efficacy of FCGR and its appropriate order for accurate classification of pathogenic virus from genomic sequences. The 98% classification accuracy obtained with the third order FCGR is clearly promising for developing in silico and accurate diagnostic tool for viral pathogens classification using next generation genomic sequences. In the future, we hope to substantially extend this study by increasing the viral pathogens coverage and further experiment with state-of-the-art machine learning methods like deep learning and hierarchical classifiers.

Acknowledgement. The publication of this study is supported and funded by the Covenant University Centre for Research, Innovation and Development (CUCRID), Covenant University, Canaanland, Ota, Ogun State, Nigeria.

References

1. 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). doi:10.1007/978-3-319-27400-3_25
2. Hoang, T., Yin, C., Yau, S.S.T.: Numerical encoding of DNA sequences by chaos game representation with application in similarity comparison. Genomics **108**(3), 134–142 (2016)
3. Huang, G., Zhou, H., Li, Y., Xu, L.: Alignment-free comparison of genome sequences by a new numerical characterization. J. Theor. Biol. **281**(1), 107–112 (2011)
4. Qi, Z.H., Du, M.H., Qi, X.Q., Zheng, L.J.: Gene comparison based on the repetition of single-nucleotide structure patterns. Comput. Biol. Med. **42**(10), 975–981 (2012)
5. Karamichalis, R., Kari, L., Konstantinidis, S., Kopecki, S.: An investigation into inter-and intragenomic variations of graphic genomic signatures. BMC Bioinform. **16**(1), 1 (2015)
6. Swain, M.T.: Fast comparison of microbial genomes using the Chaos games representation for metagenomic applications. Procedia Comput. Sci. **18**, 1372–1381 (2013)
7. Deschavanne, P.J., Giron, A., Vilain, J., Fagot, G., Fertil, B.: Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. Mol. Biol. Evol. **16**(10), 1391–1399 (1999)
8. Almeida, J.S., Carrico, J.A., Maretzek, A., Noble, P.A., Fletcher, M.: Analysis of genomic sequences by chaos game representation. Bioinformatics **17**(5), 429–437 (2001)
9. Jeffrey, H.J.: Chaos game representation of gene structure. Nucleic Acids Res. **18**, 2163–2170 (1990)

10. Wang, Y., Hill, K., Singh, S., Kari, L.: The spectrum of genomic signatures: from dinucleotides to chaos game representation. Gene **14**(346), 173–178 (2005)
11. Messaoudi, I., Oueslati, A.E., Lachiri, Z.: Wavelet analysis of frequency chaos game signal: a time-frequency signature of the C. elegans DNA. EURASIP J. Bioinform. Syst. Biol. **2014**(1), 1 (2014)
12. Kari, L., Hill, K.A., Sayem, A.S., Karamichalis, R., Bryans, N., Davis, K., Dattani, N.S.: Mapping the space of genomic signatures. PLoS one **10**(5), e0119815 (2015)
13. Tanchotsrinon, W., Lursinsap, C., Poovorawan, Y.: A high performance prediction of HPV genotypes by chaos game representation and singular value decomposition. BMC Bioinform. **16**(1), 1 (2015)
14. Stan, C., Cristescu, C.P., Scarlat, E.I.: Similarity analysis for DNA sequences based on chaos game representation. Case study: the albumin. J. Theoret. Biol. **267**(4), 513–518 (2010)
15. Sandberg, R., Winberg, G., Bränden, C.I., Kaske, A., Ernberg, I., Cöster, J.: Capturing whole-genome characteristics in short sequences using a naive Bayesian classifier. Genome Res. **11**(8), 1404–1409 (2001)
16. 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)
17. Janecek, A., Gansterer, W.N., Demel, M., Ecker, G.: On the relationship between feature selection and classification accuracy. In: FSDM, pp. 90–105, 15 September 2008
18. 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, November 2009
19. Nair, V.V., Nair, A.S.: Combined classifier for unknown genome classification using chaos game representation features. In: Proceedings of the International Symposium on Biocomputing, p. 35. ACM (2010)
20. Yang, L., Tan, Z., Wang, D., Xue, L., Guan, M.X., Huang, T., Li, R.: Species identification through mitochondrial rRNA genetic analysis. Sci. Rep. **4**(4089), 1–11 (2014)
21. Adetiba, E., Olugbara, O.O.: Classification of eukaryotic organisms through cepstral analysis of mitochondrial DNA. In: Mansouri, A., Nouboud, F., Chalifour, A., Mammass, D., Meunier, J., ElMoataz, A. (eds.) ICISP 2016. LNCS, vol. 9680, pp. 243–252. Springer, Cham (2016). doi:10.1007/978-3-319-33618-3_25