Classification of Eukaryotic Organisms Through Cepstral Analysis of Mitochondrial DNA

Emmanuel Adetiba(✉) and Oludayo O. Olugbara

ICT and Society Research Group, Durban University of Technology,
Durban 4001, South Africa
{emmanuelal, oludayoo}@dut.ac.za

Abstract. Accurate classification of organisms into taxonomical hierarchies based on genomic sequences is currently an open challenge, because majority of the traditional techniques have been found wanting. In this study, we employed mitochondrial DNA (mtDNA) genomic sequences and Digital Signal Processing (DSP) for accurate classification of Eukaryotic organisms. The mtDNA sequences of the selected organisms were first encoded using three popular genomic numerical representation methods in the literature, which are Atomic Number (AN), Molecular Mass (MM) and Electron-Ion Interaction Pseudopotential (EIIP). The numerically encoded sequences were further processed with a DSP based cepstral analysis to obtain three sets of Genomic Cepstral Coefficients (GCC), which serve as the genomic descriptors in this study. The three genomic descriptors are named AN-GCC, MM-GCC and EIIP-GCC. The experimental results using the genomic descriptors, backpropagation and radial basis function neural networks gave better classification accuracies than a comparable descriptor in the literature. The results further show that the accuracy of the proposed genomic descriptors in this study are not dependent on the numerical encoding methods.

Keywords: Atomic number · Cepstral · EIIP · Genomic · Molecular mass · mtDNA

1 Introduction

Taxonomy is a hierarchical system that is employed to group organisms up to the species level. It is the principal method used to estimate organism's diversity and makes the study of living organisms highly convenient [1]. DNA based classification of organisms into groups within taxonomical hierarchies has applications in areas such as evolutionary characterization, bio-diversity research, forensic studies, food and meat authentication, detection of relationship within and between organisms as well as species identification [2]. Some of the traditional methods for nuclear DNA based classification of organisms include sequence alignment and analysis of compositional bias. According to [3], sequences that may not seem to have resemblance using sequence alignment may be found to be similar using compositional bias. This is because biases within the nuclear DNA genomes of the same organism are smaller than

A. Mansouri et al. (Eds.): ICISP 2016, LNCS 9680, pp. 243–252, 2016.
DOI: 10.1007/978-3-319-33618-3_25

that between different organisms. The availability of more genomic data in recent times has however presented evidences of huge variation in sequences within the same category of organisms, which has impacted negatively on the efficacy of these two traditional methods. Thus, a number of other DNA based methods have been developed to replace traditional methods. One of the most reliable, sensitive and specific of these modern methods is mitochondrial DNA (mtDNA) sequencing. The mtDNA sequencing generate mitochondria genomic sequences from the cells of eukaryotic species. The mtDNA sequences represent a minute fraction of the total DNA sequences in the eukaryotic cells. mtDNA sequences have a lot of attributes that make them suitable for taxonomic classification of eukaryotic species. One of these attributes is the ease with which the sequences can be isolated from organisms even with degraded or low amount of samples. Another vital attribute is the substantial variation in the mtDNA sequences of organisms belonging to different species [4].

Given the abovementioned attractive attributes, some studies in the literature have utilized bioinformatics and Genomic Signal Processing (GSP) techniques to process mtDNA so as to solve species classification or identification problems. Vijayan et al. [5] extracted Frequency Chaos Game Representation (FCGR) features from the chaos game representation images of mtDNA sequences of eight eukaryotic organisms to train Artificial Neural Networks (ANN) classifier. The authors reported a classification accuracy of 92.3 % with Probabilistic Neural Network (PNN) using 64 element feature vector. When the feature vector was reduced to 16 elements by exploiting the fractal nature of the mtDNA, the authors reported that the network complexity was radically reduced with no appreciable reduction in classification accuracy since an average accuracy of 90.1 % was obtained. Rastogi et al. [6] carried out a study on species identification of various animal samples using mtDNA and nuclear sequences obtained from their tissues. The bioinformatics tools utilized for the study are BLAST, Molecular Evolutionary Genetic Analysis (MEGA) v3.1 and ClustalW program. The result of this study showed that the mtDNA sequences are more efficient for species identification and authentication than nuclear sequences. According to the authors, the superiority of mtDNA sequences emanates from their relatively rapid evolution at the sequence level due to the mitochondrial inability to repair damages in the DNA. Kitpipit et al. [7] undertook a study on Tiger species identification using mtDNA from two individuals of four of the five subspecies of Tiger. The authors successfully sequenced and processed a total of 7891 bp which represent 46.4 % of the total tiger mtDNA using FINCH TV 1.4.0 and ClustalX bioinformatics tools. Based on the result in this study, there was no sequence variation within the 7891 bp that can be used to reliably differentiate the Tiger subspecies. This study further validates the low variability in intra-species mtDNA which make it highly potent for interspecies differentiation.

Genomic sequences other than mtDNA have also been employed for taxonomic classification in the literature. A study was carried out in [8] to classify the genomic sequences of four pathogenic viruses which include ebolavirus, enterovirus D68, dengue and hepatitis c viruses. The authors used Genomic Cepstral Coefficients (GCC) and Gaussian Radial Basis Function (RBF) to achieve classification rate of 97.3 %.

The compact and highly discriminatory GCC utilized in [8] and the strong inter-species variability of mtDNA as illustrated in [5] provided the motivation for the study at hand. Firstly, we acquired the genome of eight eukaryotic organisms from the

National Center for Biotechnology Information (NCBI) organelle database and numerically encoded them using three Physico-Chemical Property Based Mapping (PCPBM) schemes, which are Atomic Number (AN), Molecular Mass (MM) and Electron Ion Interaction Potential (EIIP) [9, 10]. Secondly, we computed Genomic Cepstral Coefficients (GCC) from the encoded genomes to obtain three different set of descriptors from the eukaryotic organisms. These descriptors are named in this study as AN-GCC, MM-GCC and EIIP-GCC. Thirdly, experiments were performed using each of the three descriptors as well as Back Propagation Neural Network (BPNN) and Radial Basis Function Neural Network (RBFNN) as utilized in [5]. The classification results obtained from the experiments were compared with the FCGR descriptor reported in [5]. Furthermore, we compared the results obtained using the three descriptors in this study so as to determine the most discriminatory of them.

The rest of this paper is organized as follows. Section 2 contains the materials and methods, Sect. 3 contains the results and discussion, while the paper is concluded in Sect. 4.

2 Materials and Methods

2.1 Dataset

We extracted mitochondrial DNA (mtDNA) genomes of Eukaryotic organisms, which belong to eight different taxonomical categories as the dataset for this study [5]. These data are from NCBI organelle database (www.ncbi.nlm.nih.gov/genome/browse/?report=5) and were extracted on the 31st of October, 2015. On this database, a search by organism query was carried out using the names of the Eukaryotic organisms (Table 1) to obtain the mtDNA genome accession numbers of the organisms in each category. The statistics of the dataset in the current study are shown in Table 1. As shown in the Table, the cumulative size of the dataset is 1,249, with Protostomia and Vertebrata each having the highest number of organisms (198) while Porifera has the lowest number of organisms (60). The size ranges of the genome for each of the Eukaryotic organism is also shown in Table 1. Notably, an organism in Plant category has the smallest genome size of 288 and another organism in the same category has the highest genome size of 1,555,935. As earlier established, the huge variation in genome lengths within the same class of organisms as shown in the Table is bound to impact negatively on the outcome of traditional organism classification methods such as sequence alignment and compositional bias analysis.

2.2 Numerical Representation of the mtDNA Genomes

Genome sequences are biologically represented with the collection of the four nucleotides, which are adenine, thymine, cytosine and guanine. The sequences are symbolized using character strings that consist of the letters A for Adenine, T for Thymine, C for Cytosine and G for Guanine. The use of Digital Signal Processing (DSP) in the literature to solve some critical problems in genomics has been possible through numerical representation of genome sequences and this has given birth to a

Table 1. Eukaryotic organisms extracted for this study

S/N	Eukaryotic organisms	Number of organisms	Range of genome size
1	Acoelomata (Flatworms)	83	13,387 – 27,133
2	Cnidaria	124	2,811 – 22,015
3	Fungi	196	1,136 – 235,849
4	Plant	271	288 – 1,555,935
5	Porifera	60	5,596 – 28,958
6	Protostomia	198	8,118 – 48,161
7	Pseudocoelomata(Nematodes)	119	12,626 – 26, 194
8	Vertebrata	198	3,427 – 22,184

branch of bioinformatics named Genomic Signal Processing (GSP) [11]. The methods for numerical representation of sequences in the GSP literature are classified into two, which are Fixed Mapping (FM) and Physico Chemical Property Based Mapping (PCPBM). FM methods use binary, real or complex number to transform genome sequences into a series of arbitrary numerical sequences while the PCPBM methods numerically transform genome sequences such that the biological principals and structures in the sequences can be detected [10]. The PCPMB methods is highly relevant for the study at hand because our goal is to use the inherent biology structures in the mtDNA sequences to classify unknown organism into the appropriate Eukaryotic category. Other attributes that are paramount for the numerical representation of the mtDNA sequences in this study are (i) single and non-redundant representation (ii) fixed magnitude representation for each nucleotide (iii) non-derivation from other numerical representation methods and (iv) accessibility to DSP analysis. All these attributes are essential for low computational overhead, memory conservation and detection of the inherent periodicity in genome sequences [10]. The three PCPBM methods in the literature, which fully satisfy the foregoing criteria are Atomic Number (AN), Molecular Mass (MM) and Electron-Ion Interaction Pseudopotential (EIIP) [9–11]. Hence, these methods were nominated for numerical representation of the mtDNA dataset in this study. Table 2 shows the nucleotides and their corresponding AN, MM and EIIP values. In this study, all the sequences of the organisms shown in Table 1 were numerically transformed based on the values of each nucleotide in the respective methods shown in Table 2.

Table 2. Numerical representation of the four nucleotides using AN, MM and EIIP

Nucleotide	Atomic Number (AN)	Molecular Mass (MM)	Electron-Ion Interaction Potential (EIIP)
A	70	134	0.1260
G	78	150	0.0806
C	58	110	0.1340
T	66	125	0.1335

2.3 Signal Cepstral Analysis

The application of the principle used in Fourier Transform (FT) for the detection of periodicity components in a Fourier spectrum is referred to as cepstrum analysis [12]. Given a numerically represented mtDNA sequence, which is a discrete signal denoted as $\hat{x}(n)$, with a spectrum denoted as $X(w)$, the cepstrum can be computed as the inverse FT of the logarithmic spectrum as follows:

$$\hat{x}(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log(X(w)) e^{jwt} dw \tag{1}$$

Since $X(w)$ is a complex and even function, $\hat{x}(n)$ is usually referred to as a complex cepstrum even if the input signal is real. However, a real cepstrum can be computed by considering the spectrum magnitude $|X(w)|$ as:

$$c_x(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log(|X(w)|) e^{jwt} dw \tag{2}$$

As reported in [8], the real cepstrum in Eq. (2) consistently outperformed the complex cepstrum for all the experiments carried out in the previous study. Furthermore, it is a usual practice to restrict the number of cepstral coefficients that preserves the spectral envelope while removing the fine spectrum information. Retaining the first fifteen coefficients to represent the signal envelope gave better performance than the lower coefficients experimented in the aforementioned previous study [8]. Hence, for the study at hand, the first fifteen coefficients of the real cepstrum of the numerically encoded genome sequences form the GCC descriptors for each of the organisms. Based on the three different PCPBM methods earlier selected, three sets of descriptors, which are AN-GCC, MM-GCC and EIIP-GCC, were computed for the Eukaryotic organisms in this study. The algorithms for these descriptors were implemented in MATLAB R2014a programming environment. Each of the descriptors contains 15 element vector per organism and culminates in a 15 × 1,249 data matrix for the selected Eukaryotic organisms. One of the major benefits of utilizing the 15 element GCCs is that the complexity of the succeeding supervised classifier is drastically reduced.

2.4 Supervised Classification

In supervised classification, the training dataset is represented as $\{(x_j, c_j)\}$, with $j \in \{1, \ldots, N\}$ where each x_j contains n features and the class labels $c_j \in \{1, \ldots, l\}$ where l is the number of classes in the data. The supervised classification task involves the development of a model based on the set of N instances (i.e. the training data). The developed model is thereafter used to assign class labels to unknown instances using the values of the n features. Adapting the supervised classification paradigm to this study, symbols $N = 1{,}249$, $n = 15$ and $l = 8$. One of the most popularly used supervised classification methods in the bioinformatics and species classification literature is

Artificial Neural Network (ANN) [2, 10]. Since it is not known a priori which ANN topologies is more suitable for our dataset in this study, we experimented with the Backpropagation Neural Network (BPNN) and Radial Basis Function Neural Network (RBFNN) as was done in a closely related study reported in [5].

BPNN is reputed to be very good at learning various patterns [13]. In this study, the BPNN was tested for one, two and three number of layers and different number of neurons in each of the layers respectively [5]. The input layer contains 15 neurons, which is equal to the number of features in the training dataset and the output layer contains 8 neurons since there are 8 classes of Eukaryotic organisms in the dataset. The linear activation functions were selected for both the input and the output layers [10], the tansigmoid function was selected for the hidden layers while Levenberg-Marquardt training algorithm with a learning rate of 0.1 and Mean Square Error (MSE) goal of 0, was used for all the configurations [5].

RBFNN is suitable when there is a large training dataset and its design is very fast. RBFNN comprises of the input layer, which contains 15 neurons in this study, only one hidden layer, whose number of neurons are determined and created during training and the output layer, which is configured with 8 neurons. The two additional parameters that are supplied to RBFNN are the MSE goal and the spread factor.

The performances of the BPNN and RBFNN supervised classifiers were captured using accuracy and training time [5]. The implementations of the two classifiers in the Neural Network Toolbox of MATLAB R2014a were used in this study.

2.5 Experiments

The experiments in this study were performed on a computer system that contains an Intel Core i5-3210 M CPU, which operates at 2.50 GHz speed, 6.00 GB RAM, and runs 64-bit Windows 8 operating system. 70 % of the experimental dataset was used for training, 15 % for testing and the remaining 15 % for validation. The first experiment involved training each of the configurations of the BPNN five times using the AN-GCC, MM-GCC and EIIP-GCC descriptors one after the other. We needed to carry out five different trainings and obtained the average accuracy and training time for each BPNN configuration because the network normally begins with random weights. These initial random weights often culminate in the same network configuration, with the same training dataset, generating different accuracies when trained at different times. In the second experiment, we configured the RBFNN with an MSE goal of 0.0 and a spread factor of 1 and trained the network using AN-GCC, MM-GCC and EIIP-GCC respectively to obtain the accuracies and training times. All the results we obtained in the foregoing experiments are hereafter reported and discussed.

3 Results and Discussion

The results of the first experiment are shown in Table 1. As shown in the Table, the highest average accuracy after the BPNN was trained with AN-GCC is 88.04 % when the BPNN was configured with two hidden layers of 25 neurons in the first hidden layer and 15 neurons in the second hidden layer. The highest average accuracy was 88.70 % when the BPNN was trained using MM-GCC and configured with three hidden layers having 30, 20 and 10 neurons in the first, second and third hidden layers respectively. When the BPNN was trained with EIIP-GCC, the highest average accuracy was 88.66 % with two hidden layers of 50 and 30 neurons respectively. Although, the MM-GCC gave marginally higher average accuracy (88.70 %) compared to AN-GCC (88.04 %) and EIIP-GCC (88.66 %) trained classifiers, this little improvement is not significantly better, given the complexity of the BPNN configuration (three hidden layers) that generated this level of accuracy.

It is noteworthy that across all the configurations of the BPNN, the performances of the three descriptors are very similar. Hence, it is difficult to claim that any of them is the best. The similarity in the performance accuracies of the three descriptors is also illustrated in Fig. 1. As shown in the figure, there is a strong overlap in the average classification accuracies obtained using the three descriptors. Figure 1 also shows the plot of the average classification accuracies obtained using the FCGR descriptor proposed in [5] to train the same BPNN configurations. It is clearly shown in the figure that all the three descriptors in this study gave better average classification accuracies than the FCGR descriptor. As expected, the training times for BPNN configurations with 20 to 60 neurons in one hidden layer require less than 2 min while the configurations with two or three hidden layers took approximately 3 min for all the three descriptors (Table 3). Even though the BPNN with the FCGR descriptor took shorter time for training as shown in Fig. 2, the better performance obtained using our proposed descriptors is a justification for the relatively longer training time.

Table 3. Results of the first experiment

Number of hidden layer	Number of neurons in the hidden layer	Average accuracy (%)			Training time (min:sec)		
		AN-GCC	MM-GCC	EIIP-GCC	AN-GCC	MM-GCC	EIIP-GCC
1	20	83.49	81.73	85.73	0:32	0:55	0:72
1	40	81.52	86.98	84.23	1:04	0:88	0:81
1	60	83.49	86.44	88.22	1:66	1:98	1:82
1	80	78.50	87.05	81.06	2:72	3:01	2:42
2	[20 10]	84.32	86.61	85.41	2:81	3:02	2:14
2	[25 15]	**88.04**	86.76	87.81	3:00	2:39	2:74
2	[30 15]	85.24	85.88	85.25	2:56	2:84	2:54
2	[50 25]	86.87	82.83	86.84	3:01	3:00	3:01
2	[50 30]	83.92	84.13	**88.66**	3:02	3:02	3:01
3	[20 15 10]	86.36	86.10	80.34	3:00	2:83	3:00
3	[25 20 15]	87.05	84.88	85.14	3:00	3:00	3:00
3	[30 15 5]	81.32	84.40	85.08	3:00	3:00	3:00
3	[30 20 10]	83.97	**88.70**	82.13	3:01	3:00	3:00
3	[30 25 20]	83.99	87.45	83.64	3:00	3:00	3:01

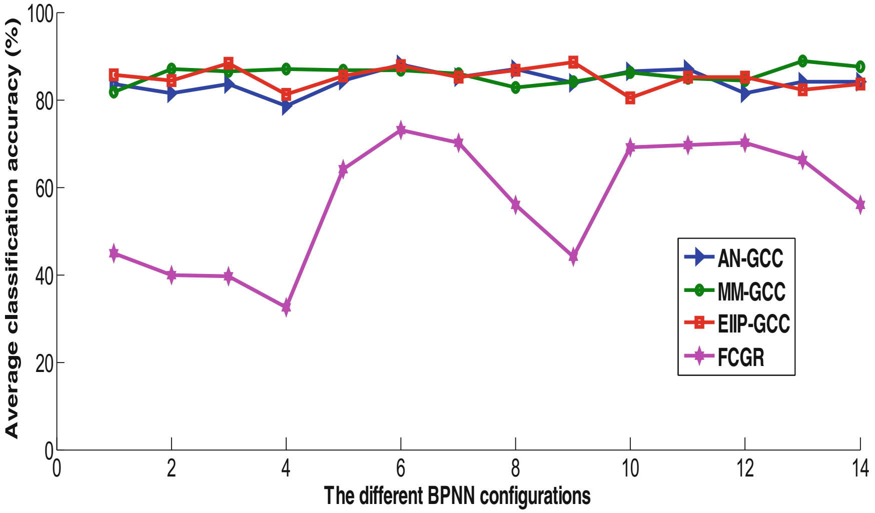

Fig. 1. Plots of the average classification accuracies for the first experiment

We carried out the second experiment so as to determine if RBFNN can give better classification accuracies than the result we obtained in the first experiment [10]. The RBFNN configurations earlier described, was trained with AN-GCC, MM-GCC and EIIP-GCC descriptors. Higher classification accuracy of 98.8 % and training time of approximately 3 min were obtained for each of the three descriptors respectively. This accuracy is better than the results in the first experiment and the results that was obtained using the FCGR descriptor and RBFNN in [5].

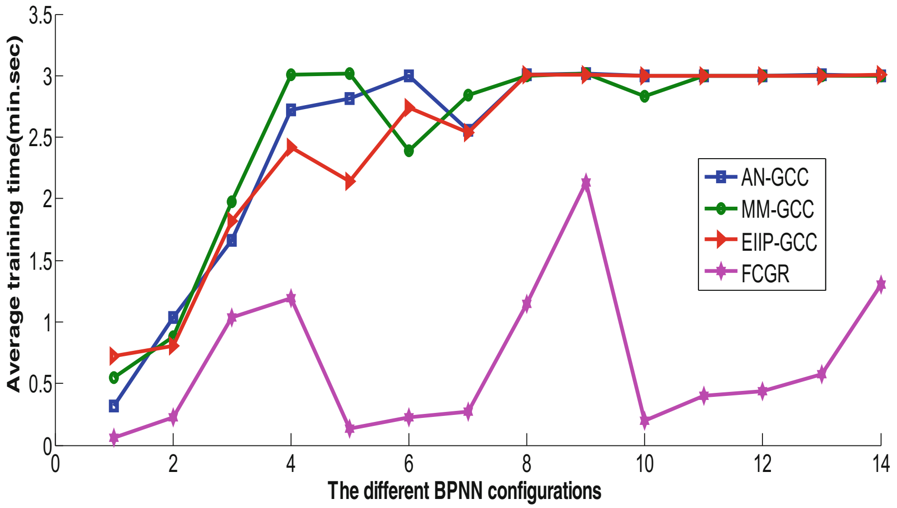

Fig. 2. Plots of the training time for the first experiment

The foregoing experimental results clearly show that the AN-GCC, MM-GCC and EIIP-GCC genomic descriptors have better efficacy than the comparable descriptor in the literature [5]. The results obtained from the two classifiers (BPNN and RBFNN) further implies that the three genomic descriptors in this study are not dependent on any of the three PCPBM method. It can therefore be unequivocally stated that, using any of the three descriptors in this study with RBFNN for classification of Eukaryotic organisms in real time will produce acceptable performance.

4 Conclusion

In this paper, we have been able to successfully obtain three highly discriminatory genomic descriptors for Eukaryotic organism classification based on mtDNA, cepstral analysis and RBFNN classifier. The descriptors were also shown to be independent of the numerical encoding methods utilized, since each of them produced 98.8 % accuracy with RBFNN. These descriptors have high prospect of being applicable for taxonomical classification of organisms in fields as diverse as bio-diversity study, food authentication, forensics, clinical diagnosis and host of others. In the future, we hope to utilize other genomic numerical encoding methods so as to determine their efficacy for the development of discriminatory genomic descriptors. We also hope to utilize other signal processing techniques such as power cepstral, linear predictive coding and higher order spectrum for enhanced genomic based taxonomical classification of organisms.

Acknowledgements. Emmanuel Adetiba is on postdoctoral fellowship at the ICT and Society (ICTAS) Research Group, funded by the Durban University of Technology, South Africa. He is on postdoctoral research leave from the Department of Electrical & Information Engineering, College of Engineering, Covenant University, Ota, Ogun State, Nigeria.

References

1. Komarek, J., Kastovsky, J., Mares, J., Johansen, J.R.: Taxonomic classification of cyanoprokaryotes (cyanobacterial genera) 2014, using a polyphasic approach. Preslia **86**(4), 295–335 (2014)
2. 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)
3. Zanoguera, F., De Francesco, M.: Protein classification into domains of life using Markov chain models. In: IEEE Computational Systems Bioinformatics Conference, CSB 2004, pp. 517–519 (2004)
4. 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)
5. 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)

6. Rastogi, G., Dharne, M.S., Walujkar, S., Kumar, A., Patole, M.S., Shouche, Y.S.: Species identification and authentication of tissues of animal origin using mitochondrial and nuclear markers. Meat Sci. **76**, 666–674 (2007)
7. Kitpipit, T., Linacre, A., Tobe, S.S.: Tiger species identification based on molecular approach. Forensic Sci. Int Genet. Suppl. Ser. **2**, 310–312 (2009)
8. 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, Heidelberg (2016)
9. Kwan, J.Y.Y., Kwan, B.Y.M., Kwan, H.K.: Spectral analysis of numerical exon and intron sequences. In: 2010 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW), pp. 876–877 (2010)
10. Adetiba, E., Olugbara, O.O.: Lung Cancer Prediction Using Neural Network Ensemble with Histogram of Oriented Gradient Genomic Features. The Scientific World Journal, pp. 1–17 (2015)
11. Abo-Zahhad, M., Ahmed, S.M., Abd-Elrahman, S.A.: Genomic analysis and classification of exon and intron sequences using DNA numerical mapping techniques. Int. J. Inf. Technol. Comput. Sci. (IJITCS) **4**(8), 22 (2012)
12. Liang, B., Iwnicki, S.D., Zhao, Y.: Application of power spectrum, cepstrum, higher order spectrum and neural network analyses for induction motor fault diagnosis. Mech. Syst. Sign. Proces. **39**, 342–360 (2013)
13. Yu, S.N., Chou, K.T.: Combining independent component analysis and backpropagation neural network for ECG beat classification. In: 28th Annual International Conference EMBS 2006, pp. 3090–3093 (2006)