Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Detection of intra-family coronavirus genome sequences through graphical representation and artificial neural network

Tirthankar Paul\textsuperscript{a},\textsuperscript{*}, Seppo Vainio\textsuperscript{b}, Juha Roning\textsuperscript{a}

\textsuperscript{a} InfoTech Oulu, Faculty of Information Technology and Electrical Engineering, Biomimetics and Intelligent Systems Group (BISG), University of Oulu, Oulu, Finland
\textsuperscript{b} Infotech Oulu and Kvantum Institute, Faculty of Biochemistry and Molecular Medicine, Disease Networks, University of Oulu, Oulu, Finland

Abstract

In this study, chaos game representation (CGR) is introduced for investigating the pattern of genome sequences. It is an image representation of the genome for the overall visualization of the sequence. The CGR representation is a mapping technique that assigns each sequence base into the respective position in the two-dimension plane to portray the DNA sequence. Importantly, CGR provides one to one mapping to nucleotides as well as sequence. A coordinate of the CGR plane can tell the corresponding base and its location in the original genome. Therefore, the whole nucleotide sequence (until the current nucleotide) can be restored from the one point of the CGR. In this study, CGR coupled with artificial neural network (ANN) is introduced as a new way to represent the genome and to classify intra-coronavirus sequences. A hierarchy clustering study is done to validate the approach and found to be more than 90\% accurate while comparing the result with the phylogenetic tree of the corresponding genomes. Interestingly, the method makes the genome sequence significantly shorter (more than 99\% compressed) saving the data space while preserving the genome features.

Keywords:
Chaos game representation
Artificial neural network
Coronavirus

1. Introduction

Representations of genomic data into numerical, graphical or audio form have gained importance in bioinformatics research. In recent years, many studies showed different ways of genome representations to find their various DNA significances. A graphical representation of DNA is convenient to achieve a visual analysis of the distribution of nucleotide bases A, C, G and T of DNA. The mapping model was introduced as an H-curve, a three-dimensional space representation of the DNA sequence by Hamori and Ruskin in 1983 (Hamori and Ruskin, 1983). Later on, researchers found many exciting ways of graphical representation in different methods (Bielinska-Waż and Wąż, 2017; Mo et al., 2018; Randić et al., 2006; Hoang et al., 2016; Touati et al., 2021; Poor and Yaghoobi, 2019; Sun et al., 2020). Chaos Game Representation (CGR) is one of the graphical techniques, which assigns each DNA bases into the respective position in the two-dimension plane to portray the DNA sequence (Hoang et al., 2016). This 2-D representation was introduced to show the depiction of the local and global pattern of DNA by using iterated function systems based on chaotic dynamics techniques (IFS) (Jeffrey, 1990). The CGR technique was recently shown as an efficient classifier for helitrons families in Caenorhabditis elegans genomes (Touati et al., 2021). Also, the CGR tool is useful to make a comparative study among genome sequences (Hoang et al., 2016). A genome sequence can be characterized in both graphical and numerical form through Chaos Game Representation. Importantly, CGR provides one to one mapping to nucleotides as well as sequence mapping (Jeffrey, 1990). From one coordinate of the CGR plane, it can tell the corresponding base and its location in the original genome. Therefore, the whole nucleotide sequence (until the current nucleotide) can be restored just from a point location of the CGR. Additionally, a CGR portrays whole-genome data or parts of the sequence in a single plane, making the CGR tool more practical for a comparative study (Deschavanne et al., 1999). Recently, CGR was used for capturing the recurrence features from SARS-Cov-2 datasets, and the results were proposed for clustering (Olyaei et al., 2020). Mostly, the CGR plot were made in square in shape but it can be presented in n-vertex polygon CGR (Xiaohui et al., 2014). The year 2020 started with a threat of a pandemic where a virus badly hit the human life. Gradually, the virus spread almost all around the world. The human community is presently suffering from a pandemic caused by a positive RNA strand virus, Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) (Yan et al., 2020). Moreover, there are several viral diseases, as mentioned in the world health...
organization (WHO) pandemic list, such as the Middle East respiratory syndrome (MERS), Disease X, severe acute respiratory syndrome (SARS), Ebola and Influenzas. These viruses have shown their impact on human in several years (WHO 0000). Initially, WHO suspected the source of the present pandemic as Disease X. However, a novel coronavirus (2019-nCoV) is found for the COVID-19 pandemic (Xia et al., 2020; Wu et al., 2020; Zhu et al., 2020). Sometimes, a new virus comes from its variants or after mutations. For example, SARS-CoV-2 is 79.5% similar in nucleotide sequence with the previous SARS-CoV family and shows 96% similarity with another coronavirus, SL-Cov-RatTG13 (Zhou et al., 2020). More than these two examples, there are also many more cases for similarities in nucleotide sequences, as discussed in the study (Zhou et al., 2020).

Genome studies of sequence analysis in the laboratory are always expensive and time-consuming work. Therefore, the computing process using an artificial neural network (ANN) can be an efficient tool in studying sequence analysis in segmentation (Cheng et al., 2012) and computing of DNA sequences (Zhong et al., 2020). An ANN network consists of an input layer, hidden layer and output layer. The smallest processing unit of an ANN is neurons, connecting the different layers in the network (Giarro et al., 2016). Each neuron is responsible for carrying the input from the previous layer and computing the sum or part of the input with individual weight. The simplified human brain inspires the interconnections, nodes and layers of a neural network to compute different recognition algorithms. ANN is used when there is plenty of data and no algorithmic solution possible with them. These data are used to train the network during the learning process of the network (Hoang et al., 2020). Multilayer perception (MLP) with one or more hidden layers is more accurate with given sufficient training data (Hoang et al., 2020). A multilayer ANN is capable of finding a correlation among the input data. This machine learning tool is efficient in pattern recognition that can be applied in gene detection, sequence classification, disease detection and many other aspects in bioinformatics. The ANN method was used in the detection of high dimensional complex cancer datasets (Lancashire et al., 2009). The technique is one of the deep learning methods applied to solve data analysis and computer vision. It was efficient to classify DNA damages based on comet assay images (Atila et al., 2020). A neural network tool can be beneficial to organize the virus sequences into their family in this context. The virus sequences are positive RNA strand, mainly thirty thousand bases in size. A classification and clustering study were made among the different virus sequences using a support vector machine (SVM) and haar wavelet (Paul et al., 2021). This study explores the combine performance of CGR and ANN to detect the virus sequences from the same family. For the purpose, the Coronavirus family, including SARS-CoV, SARS-CoV2, MERS and Alphacoronavirus (Alpha CoV), were taken in the analysis.

2. Material and methods

The CGR method converts the one-dimensional sequence into a two-dimensional graphical form. First, the DNA nucleotides were assigned to the four vertices in a unit square Euclidean plane. The coordinates of the nucleotide vertices were $A = (0,0); C = (0,1); T = (1,0)$ and $G = (1,1)$. The mapping process plots a dot starting from the center $P_0 = (0.5,0.5)$ of the plane. It reads the DNA string ($S_n$) and checks the first character ($S_{n-0}$) to plot in the plane. This stage points the position ($P_1$) to the half distance between the last dot (center) and the vertex matching to the first nucleotide. For the next point ($P_2$), it is precisely at the halfway of $P_1$ (now acts as the initial point) and corresponding nucleotide vertex. Similarly, the process continues to the Nth or last character ($S_{n-N}$) of the DNA sequence.

Mathematically it can be presented as Eq. (1)

$$P_n = 0.5*(P_{n-1} + V_{n})$$  \hspace{1cm} (1)

Where $n = 0,1,2,3…N$ is the position of the ‘N’ bases DNA string ($S_n$) and $V_n$ is the corner coordinate of the unit square; $V_0 = (0.0), V_1 = (0.1), V_2 = (1.1)$ and $V_3 = (1.0)$. A short sequence ‘ACTGAAATGA’ ($S_n where length = 9$) was taken as an illustration for the CGR process shown in Fig. 1a. Inversely, from a given CGR coordinate, the original sequence can be traced back in Fig. 1b. The genomic position of the sequence governs the geometric coordinates of the CGR plane. Each coordinate of the CGR plane acts as a memory of the source sequence for the specific nucleotide, pointed by the coordinate (Jonas et al., 2001). For example, the third nucleotide (from left to right, $P_3 = T$) in $S_n$ can be tracked to recover the other nucleotides (first and second, $P_1 = A, P_2 = C$) subsequence until the third position in Fig. 1b. For example, the third nucleotide (from left to right, $P_3 = T$) in $S_n$ can be tracked to recover the other nucleotides (first and second, $P_1 = A, P_2 = C$) subsequence until the third position in Fig. 1b. First, $P_3$ was found in the fourth quadrant of the CGR plane in the giant square where the correspondent nucleotide vertex is ‘T’. Now, CGR plane needs to zoom only in the mentioned side square. Similarly, in the next step, the point $P_2$ lies in the ‘A’ vertex side (yellow square). If the vertexes of these squares are arranged in ascending order based on the size, the controlling vertexes ‘ACT’ can be retrieved as the recovered sequences.

2.1. Frequency Chaos Game representation (FCGR)

The Frequency Chaos Game Representation (FCGR) is an inherited
method of CGR. It converts a DNA sequence into a 2-D plot based on a k-mer occurrence of the DNA. In the FCGR method, the CGR unit square plot is divided into $2^k \times 2^k$ squares (Lichtblau, 2019). Each sub-square captures the k-mers according to the geometric position of the corresponding nucleotide of the DNA string. For $K = 1$ or one order FCGR ($\text{FCGR}_1$), CGR is divided four sub-square. Each sub-square holds the $K$ subsequent nucleotide, i.e. A, C, T and G (Poor and Yaghoobi, 2019). Fig. 2a shows the $\text{FCGR}_1$ plot of the sequence $S_n$ and the distributions of monomers are $A = 3$, $C = 1$, $G = 2$, and $T = 4$. Similarly, there are sixteen sub-square with dinucleotide feature for $\text{FCGR}_2$, which is shown in Fig. 2b. Here, the dimer distributions are like $AA = 1$, $AC = 1$, $GA = 1$, $AT = 1$, $CT = 1$, $TG = 2$, $GT = 1$, $TT = 1$ and the rest are zero. The same way $\text{FCGR}_3$, $\text{FCGR}_4$ or $\text{FCGR}_n$ can be determined.

### 2.2. FCGR to numerical coding

Each k-mer occurrence can be found from the $\text{FCGR}_n$ process. This frequency matrix of FCGR can be restructured with the occurrence of k-mer. The dimer ($k = 2$), trimer ($k = 3$) or k-mer can be determined from the nucleotide sequences. And the total number of dimers are $4^2 = 16$, trimers $4^3 = 64$ and so on ($4^k$ for k-mer). Here, $\text{FCGR}_3(K = 3)$ is shown as an example, similarly any $\text{FCGR}_n(K = 3)$ can be found. The smallest unit square of $\text{FCGR}_3$ plane was mapped based on the occurrence of the trimers. As an example, the SARS-CoV sequence (accession number: AG278741) of length 29,727 base pairs was represented in the form of $\text{FCGR}_3$ (Drosten et al., 2003). The DNA 2-D image of the corresponding sequence is shown in Fig. 3. The color indicates the presence of codons in the genome sequence. The darker color represents the fewer presence, and the lighter color is for the most significant number of the existence of the trinucleotides.

### 2.3. Classification approach by neural network

Here, a feed forward MLP model was designed based on supervised learning algorithms. Generally, in the beginning, the weight of the model was fed with random values during the training phase. Later, each feeding of training data or epoch, the model follows some specific
learning protocol with the adjustment of weight to minimize the error between desired and actual output. Here, the learning process used scaled conjugate gradient method. The output $y_k(x)$ is defined by Eq. (2).

$$y_k(x) = \sum_{i=1}^{n} x_i w_i + b$$  \hspace{1cm} (2)

The FCGR$_3$ matrix elements were taken as input of the neural network as shown in Fig. 4. There are four virus family members as an output to detect through the ANN. The ANN was designed with 64 input neurons for the 64 datasets of FCGR$_3$ matrix and 4 output for the classified result of the four types of viruses. In the hidden layer, 10 neurons were taken for the ANN model in the Neural network pattern recognition application in Matlab 2019b.

2.4. Datasets

The different coronavirus sequences are available in the National Center for Biotechnology Information (NCBI) (NCBI-Database, 2020). The data for the present study was downloaded with Genbank accession number, location and the published date. A total number of 1787 sequences were taken with FCGR$_2$ and FCGR$_3$ related sequences as a supplementary file (S1).

3. Result and discussion

The CGR method is capable of translating the genome sequence into graphical and numeric data at the same time. Here, all the sequences are mapped into FCGR$_2$ and FCGR$_3$ formats. Then, they are classified through ANN. Out of 1787 sequences, randomly selected fifteen percentage were taken for each validation and testing separately. The rest, seventy percentage sequences were used as training data. The four different coronavirus types were classified, and a confusion matrix was plotted for training, validation, test data as shown in Fig. 5a and Fig. 5b for the FCGR$_2$ and FCGR$_3$ formats, respectively. The target classes were marked 1, 2, 3 & 4 for MERS, SARS-CoV-2, Alpha CoV & SARS-CoV, respectively.

The size of FCGR$_2$ dataset is sixteen and for the FCGR$_3$ formats, it is sixty-four. Here, the virus genome sequences (approximately 30,000 bases) were mapped into the CGR formats (64 and 16 lengths). Fig. 5a and Fig. 5b show that FCGR$_3$ data performs better than the FCGR$_2$ dataset for data classification. FCGR$_3$ achieved 99.8% overall accuracy and, on the hand, FCGR$_2$ was 90.4% accurate to accomplish the classification target. The mapped datasets (FCGR$_2$ and FCGR$_3$) achieved the desired outcomes. Therefore, the other FCGR$_k$ ($k =$ anything other than 2 and 3) sequence was not prepared as the dataset would have been very small or very long in length.

The dendrogram plot is a hierarchy clustering representing a tree. It was made from the CGR data and corresponding eleven genomes sequences as leaf nodes, in Fig. 6a and Fig. 6b. A random eleven number of sequences were taken as a representative for the easy readability of the plot, however, the programming is adaptable for all 1787 sequences in one tree (some examples given in supplementary file). Randomly, 2–4 genomes were taken from each species of the 1787 coronavirus dataset. Hierarchical binary clustering method was applied to make four distinguish cluster for the groups (SARS-CoV, SARS-CoV2, MERS and
Alpha CoV) of the virus family. The branch colors are indicating different groups. To verify the result, a phylogenetic tree in Fig. 7 was plotted with the same genome sequences that were used in Figs. 6 (a and b). The Dendrogram tree of CGR shows that the genome features are preserved even after transforming the long genome sequences into short FCGR3 and FCGR2 sequences.
The dendrogram tree (Fig. 6a and Fig. 6b) and phylogenetic tree (Fig. 7) were made from the same genome sequences, which were selected randomly from the dataset (given in the supplementary files). The input sequence length for the phylogenetic tree is nearly thirty thousand bases (coronavirus sequence length). On the other hand, only 16 and 64 length datasets were represented the virus genome in the dendrogram trees. It is evident that the phylogenetic tree shows the desired result as total length of the sequence without applying any mapping or transformation of the data (Deng et al., 2006). Strikingly, the dendrogram tree or the hierarchical cluster tree from CGR data were nested with the series of subsets which were defined as SARS-CoV, SARS-CoV-2, MERS and Alpha CoV. In contrast, both the trees, dendrogram tree and phylogenetic tree are not identical (Fig. 6 and Fig. 7), but importantly, the subsets of both the trees are having the same group members. The only difference, which was made by Alpha CoV subset, made the difference in the evolutionary distances for both the trees.

The CGR process is one-to-one sequential mapping in the numerical representation of genome data. As a result, the nature of the genome could be preserved and sustained after the transformation. The great advantage of the transformation is the possibility to reconstruct the original genome data up to a given point of the CGR sequence. Moreover, the CGR behaves as a genome signature in a unique two-dimensional plot (Hoang et al., 2016; Deschavanne et al., 1999). The CGR pattern of the genome sequences is unique to each species. The species go through a slight variation along the whole genome. The variation in CGR pattern among the species is primarily significant on few factors, i.e. the base concentration, unusual repetition and the stretches of the bases of the genome (Deschavanne et al., 1999). Here, for the coronavirus, there is a very small sequential difference in the virus strain at different locations of the same virus. Therefore 2-D representation of CGR plane shows a mild pattern (color) difference in some smallest unit squares (Fig. 3b). However, there is a difference in the protein structure for different viruses even from the same family, that gives a more diverse 2-D CGR plane compared to the same genome.

Fig. 6b. Dendrogram tree from the FCGR$_2$ data.

Fig. 7. Phylogenetic tree for randomly selected coronavirus genomes.
variant. An ANN is an efficient categorizing tool, which can classify the 2-D CGR plane matrix data by using the learning algorithms.

4. Conclusion

In this study, a virus classification method was introduced in a novel way. Firstly, a virus protein sequences were translated into CGR plane. There are sixty-four codons and twenty amino acids, which are responsible to make proteins. Coincidently, CGR plane consists of sixty-four smallest unit square box. Therefore, each codon was distributed on the specific smallest box on the CGR plane. All the genome sequences were plotted into CGR plane that gave an $2^4 \times 2^4$ matrix for each sequence. The protein features were mapped into the CGR 2-D plane. Visualization of the CGR plot gives a primary idea about the sequences. Here, the matrix was created with various genome of different coronavirus to show the CGR plane is efficient to classify the group of viruses. From the result, it can be concluded that the clustering was done with high accuracy rate. Therefore, the method will be more efficient for diversified virus sequences for the inter-family clustering. The classification method was designed based on the CGR method and ANN tool was used for higher accuracy. Besides, the technique was also used in distance analysis. Virus classification using the CGR and phylogenetic analysis was compared for genome sequences and found to be comparable.

The work resulted in an artificial intelligence-based algorithm to detect virus protein sequences. The main benefit of the method is encoding the genome sequence (big data) into an organized, well represented small data. The $FCGR_2$ and $FCGR_3$ have 0.053% and 0.213% data size, respectively, compared to the size of their actual genome sequence. Two more dendrogram trees were plotted and given as supplementary files (S2 and S3). The sequences were taken randomly from the dataset (S1) for the plot S2 and S3. There is a small clustering error (1 out of 36 samples for Fig. 6a, S2 and S3 in plot S3. Where a SARS-CoV2 virus to show the CGR plane is efficient to classify the group of viruses. Coincidently, using this method will reduce the processing time and memory occupancy of a genome database. This will improve the efficiency of the virus detecting tool, as the training can be done with a larger number of genome samples compared to earlier methods. The virus protein sequence will be converted into image sequences which will be efficient for categorizing the dataset into virus families.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.eswa.2022.116559.

References

Hamori, E., & Rubink, J. (1983). H curves, a novel method of representation of nucleotide series especially suited for long DNA sequences. J. Biol. Chem., 258(2), 1318–1327.

Bielinska-Waj, D., & Waj, P. (2017). Spectral-dynamic representation of DNA sequences. J. Biomed. Inform., 72, 1–7. https://doi.org/10.1016/j.jbi.2017.06.001

Mo, Z., Zhu, W., Sun, Y., Xiang, Q., Zheng, M., Chen, M., & Li, Z. (2018). One novel representation of DNA sequence based on the global sequence and local position information. Sci. Rep., 8, 1–7. https://doi.org/10.1038/s41598-018-26005-3

Randić, M., Zupan, J., Vikić-Topić, D., & Plavšić, D. (2006). A novel unexpected use of a graphical representation of DNA: Graphical alignment of DNA sequences. Chem. Lett., 43(4(4)), 375–379. https://doi.org/10.1246/cl.2006.09.044

Hoang, T., Yin, C., & Yau, S.-T. (2016). Numerical encoding of DNA sequences by chaos game representation with application in similarity comparison. Genomics., 108(3–4), 134–142. https://doi.org/10.1016/j.ygeno.2016.08.002

Touati, R., Massouidi, I., Oueslati, A. E., Lachiri, Z., & Khararat, M. (2021). New intraclass Helitrons Classification Using DNA-Image Sequences and Machine Learning Approaches. Irfeb., 42(3), 154–164. https://doi.org/10.1016/j.irfeb.2019.12.004

Poor, N. R., & Yaghoubi, M. (2019). A new approach in DNA sequence compression: Fast DNA sequence compression using parallel chaos game representation. Expert Syst. Appl., 116, 487–493. https://doi.org/10.1016/j.eswa.2018.09.012

Sun, Z., Pei, S., He, R. L., & Yau, S. T. T. (2020). A novel numerical representation for proteins: Three-dimensional Chaos Game Representation and its Extended Natural Vector. Comput. Struct. Biotechnol. J., 18, 1904–1913. https://doi.org/10.1016/j.csbj.2020.07.004

Jeffrey, H. J. (1990). Chaos game representation of gene structure. Nucleic Acids Res., 18 (8), 2163–2170. https://doi.org/10.1093/nar/18.8.2163

Deschavanne, P. J., Giron, A., Vilain, J., Fagot, G., & Fertil, B. (1999). Genomic signature: Characterization and classification of species assessed by chaos game representation of sequences. Mol. Biol. Evol., 16(10), 1391–1399. https://doi.org/10.1093/oxfordjournals.molbev.a026048

Olyaei, M. H., Pirgizi, J., Khalifeh, K., & Khatennemoor, A. (2020). RCOVID19: Recurrence-based SARS-CoV-2 features using chaos game representation. Data Br., 32, 106144. https://doi.org/10.1016/j.dib.2020.106144

Xiaohui, N., Feng, S., Xuehai, H., Jingbo, X., & Nana, L. (2014). Predicting the protein solubility by integrating chaos games representation and entropy in information theory. Expert Syst. Appl., 41(4), 1672–1679. https://doi.org/10.1016/j.eswa.2013.08.064

R Yan, Y. Zhang, Y. Li, L. Xia, Y. Guo, Q. Zhou, Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2, Science (80–). 367 (2020) 1444–1448. https://doi.org/10.1126/science.abd2762

World Health Organization., visited on (n.d.). https://www.who.int/activities/ prioritizing-diseases-for-research-and-development-in-emergency-contexts

Xia, S., Liu, M., Wang, C., Xu, W., Lan, Q., Feng, S., … Lu, L. (2020). Inhibition of SARS-CoV-2 (previously 2019-nCoV) infection by a highly potent pan-coronavirus fusion inhibitor targeting its spike protein that harbors a high capacity to mediate membrane fusion. Cell Res., 30(4), 343–355. https://doi.org/10.1038/s41422-020-0305-x

Wu, F., Zhao, S.-u., Yu, B., Chen, Y.-M., Wang, W., Song, Z.-G., … Zhang, Y.-Z. (2020). A new coronavirus associated with human respiratory disease in China. Nature., 579 (7798), 265–269. https://doi.org/10.1038/s41586-020-2008-3

Zhu, N.A., Zhang, D., Wang, W., Li, X., Yang, B.o., Song, J., … Tan, W. (2020). A novel coronavirus from patients with pneumonia in China. N. Engl. J. Med., 382(8), 727–733. https://doi.org/10.1056/NEJMoa2001017

Zhou, P., Yang, X.-L., Wang, X.-G., Hu, B., Zhang, L., Zhang, W., … Shi, Z.-L. (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature., 579(7798), 270–273. https://doi.org/10.1038/s41586-020-2021-3

Cheng, W. C., Huang, J. C., & Liou, C. Y. (2012). Segmentation of DNA using simple recurrent neural network. Knowledge-Based Syst., 26, 271–280. https://doi.org/10.1016/j.knosys.2011.09.001

Zong, G., Li, T., Jiao, W., Wang, L. N., Dong, J., & Liu, C. L. (2020). DNA computing inspired deep networks design. Neurocomputing, 382, 140–147. https://doi.org/10.1016/j.neucom.2019.11.098

Garro, B. A., Rodríguez, K., & Vázquez, R. A. (2016). Classification of DNA microarrays using artificial neural networks and ABC algorithm. Appl. Soft Comput. J., 38, 548–566. https://doi.org/10.1016/j.asoc.2015.10.022

Hoang, T., Yin, C., & Yau, S.-T. (2020). Splice sites detection using chaos game representation and neural network. Genomics., 112(2), 1847–1852. https://doi.org/10.1016/j.genom.2019.10.018

Lancashire, L. J., Lemetre, C., & Ball, G. R. (2009). An introduction to artificial neural networks in bioinformatics - Application to complex microarray and mass spectrometry datasets in cancer studies. Brief. Bioinform., 10, 315–329. https://doi.org/10.1093/bib/bbp012

Atila, Ü., Baydilli, Y. Y., Sehirli, E., & Turan, M. K. (2020). Classification of DNA damages on segmented comet assay images using convolutional neural network. Comput. Methods Programs Biomed., 186, 105192. https://doi.org/10.1016/j.cmpb.2019.105192

CRediT authorship contribution statement

Tirthankar Paul: Conceptualization, Methodology, Software, Writing – original draft, Visualization. Seppo Vainio: Supervision, Funding acquisition. Juha Roning: Supervision, Funding acquisition, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work is supported by Infotech Oulu doctoral program.
Paul, T., Vainio, S., & Roning, J. (2021). Clustering and classification of virus sequence through music communication protocol and wavelet transform. Genomics., 113(1), 778–784. https://doi.org/10.1016/j.ygeno.2020.10.009
Jonas, M. F., Almeida, S., Carriço, J. A., Maretzek, A., & Noble, P. A. (2001). Analysis of genomic sequences by. Chaos Game Representation Jonas, 17, 429–437.
Lichtblau, D. (2019). Alignment-free genomic sequence comparison using FCGR and signal processing. BMC Bioinformatics., 20, 1–17. https://doi.org/10.1186/s12859-019-3330-3
Drosten, C., Günther, S., Preiser, W., van der Werf, S., Brodt, H.-R., Becker, S., … Doerr, H. W. (2003). Identification of a novel coronavirus in patients with severe acute respiratory syndrome. N. Engl. J. Med., 348(20), 1967–1976. https://doi.org/10.1056/NEJMoa030747
Deng, R., Huang, M., Wang, J., Huang, Y., Yang, J., Feng, J., & Wang, X. (2006). PTreeRec: Phylogenetic Tree Reconstruction based on genome BLAST distance. Comput. Biol. Chem., 30(4), 300–302. https://doi.org/10.1016/j.compbiolchem.2006.04.003
NCBI-Database, (2020). https://www.ncbi.nlm.nih.gov/ [visited on 30.11.2020].