

Visualizing Genetic Patterns: A Comparative Analysis of DNA Sequences Through Image Processing

Probir Mondal, Debranjana Pal, Krishnendu Basuli, and Pratyay Banerjee

Cite as: Mondal, P., Pal, D., Basuli, K., & Banerjee, P. (2024). Visualizing Genetic Patterns: A Comparative Analysis of DNA Sequences Through Image Processing. International Journal of Microsystems and IoT, 2(2), 586-590. <https://doi.org/10.5281/zenodo.10803111>




© 2024 The Author(s). Published by Indian Society for VLSI Education, Ranchi, India



Published online: 20 February 2024.



Submit your article to this journal: 




Article views: 



View related articles: 



View Crossmark data: 

DOI: <https://doi.org/10.5281/zenodo.10803111>



Visualizing Genetic Patterns: A Comparative Analysis of DNA Sequences Through Image Processing

Probir Mondal¹, Debranjana Pal², Krishnendu Basuli³ and Pratyay Banerjee⁴

¹Department of Computer Science, P. R. Thakur Government College, West Bengal, India

²Department of Computer Science and Engineering, IIT Kharagpur, West Bengal, India

³Department of Computer Science, West Bengal State University, India

⁴Department of Physics, P. R. Thakur Government College, West Bengal, India

ABSTRACT

A comparative analysis of DNA sequence is investigated through Image Processing. The underlying algorithm transforms, in a novel way, genetic data into images. The information is encoded by using the pixel intensities representing the four constituent nucleotide bases viz. A, T, C and G. These sequences are then employed to generate visual representations, facilitating an intuitive understanding of complex genetic information. Our study integrates machine learning techniques to compare and cluster these DNA sequence-based images, offering a powerful tool for classification. By leveraging machine learning algorithms, we enable the automated recognition of genetic similarities/dissimilarities within genomes which, in turn, streamline the time-consuming process of traditional sequence comparison.

KEYWORDS

DNA sequence, Sequence alignment, Image analysis, Clustering, Machine Learning

1. INTRODUCTION

The field of genomics has undergone a profound transformation in recent years, driven by advances in data analysis techniques and the convergence of multiple disciplines. One particularly innovative avenue of exploration is alignment-based sequence comparison [1-6], a process that lies at the heart of understanding genetic diversity and evolutionary relationships. Later, alignment free methods were employed to compare very long sequences [7-17]. Here, in this article, we introduce a multifaceted approach that combines numerical representation [18], image conversion from numeric array and machine learning for the purpose of sequence clustering [19, 20].

Traditionally, the comparison of DNA or protein sequences has relied on complex algorithms that operate on raw genetic data. However, these methods can be computationally intensive and challenging to interpret visually. Our approach begins by encoding genetic sequences into numerical representations where different pixel intensities for the four nucleotide bases, viz. A, T, C and G provide a concise and standardized format for genetic data.

To make this genetic information more accessible and intuitive, we take the additional step of converting these DNA sequences into images [21, 22]. Each nucleotide corresponds to a pixel intensity in the resulting image, thereby creating a visual representation of the genetic

sequence. This transformation enables researchers and scientists to harness the power of image analysis techniques for the comparison and interpretation of genetic data.

Moreover, we leverage machine learning algorithms to cluster these image-based representations of genetic sequences [23-25]. Several clustering methods, such as k-means [26], Gaussian mixture model [27], hierarchical clustering [28] and spectral clustering [29], are employed to automatically group similar sequences together. This fusion of image-based sequence representations and machine learning offers an efficient approach to explore genetic diversity.

In this article, we highlight the potential applications of our approach in genomics research, personalized medicine, and biodiversity conservation. By bridging the gap between computational analysis, visualization, and machine learning [30, 31], we aim to empower researchers with a powerful tool that unlocks deeper insights into genetic patterns and evolutionary relationships.

In Sec. 2, we provide a comprehensive exposition of our methodology providing the details of the numerical representation and the subsequent image generation process. Following this, we proceed to extract features from these images which serve as the foundation for our Machine Learning model development. In Sec. 3, we furnish our experimental results.

To rigorously assess the accuracy and robustness of the algorithm, we employ it on full genome sequence of fish mtDNA taken from benchmark dataset [6]. Finally, we conclude summarizing the key findings and insights derived from our result. Additionally, we outline potential avenues for future exploration and development in this field.

2. ALGORITHMIC WORKFLOW

Fig. 1 depicts the flowchart of our novel technique and the components of the algorithm workflow. The major steps of our algorithm are described next.

2.1 Numerical Array Representation:

Here we represent each DNA sequence through a numeric array with four uniformly distributed intensities. To do so we associate the pixel values with each of the four nucleotides in a DNA as follows: Adenine (A) \rightarrow 1, Cytosine (C) \rightarrow 63, Guanine (G) \rightarrow 127, and Thymine (T) \rightarrow 255. As an example, following this association, a DNA sequence ATTGCCAT may be represented as the numeric array {1, 255, 255, 127, 63, 63, 1, 255}.

2.2 Converting a Numerical Array into an Image:

To represent this numerical array as an image, we need to fix the dimension of the image. To this end, we calculate the length of the array and maintain 3:1 aspect ratio while setting the value of height and width of an image. The array is then converted into an image where each value represents an intensity level of a pixel. For example, 1 represents black while 255 represents white etc. The image will have a grid-like appearance where the number of rows and columns in the grid matches the calculated size. This resizing process ensures that the array is represented in a visually meaningful way as an image, and the format is often used for convenience and visual clarity. The resulting image can be analyzed or displayed more easily than a long one-dimensional array.

2.3 Feature Extractions from Image:

The statistical method is a powerful approach for analyzing the spatial distribution of grey-level values within an image. These methods involve calculating local characteristics at each point in the image and extracting a set of statistics from the distribution of these local characteristics. One common approach is using histogram-based features, which are considered as first-order statistics [13]. These features are derived directly from the original image without considering the relationships between neighboring pixels.

In the context of texture analysis, various statistical measures can be employed to capture important characteristics of

texture patterns beyond histogram-based features. Here, we will consider only two such features: skewness and entropy and show that they suffice, to a good approximation, to generate the expected phylogeny amongst closely related organisms.

Skewness: It is a measure of asymmetry of a distribution. Skewness is defined as

$$\mu_3 = \frac{1}{\sigma^3} \sum_{i=0}^{G-1} (i - \mu)^3 p_i, \quad (1)$$

Where P_i is the probability of occurrence of a pixel with intensity level i .

Here, μ is the average intensity level while σ^2 represents variance, i.e., average fluctuation of intensities about the mean value. In equation 1, G is the number of distinct intensity levels in a typical image. In this article we choose $G=4$ and consequently I run over the set {0,1,2,3}. $\mu_3=0$ if the histogram is symmetrical about the average value of the intensity. Otherwise, positive skewness leads to a longer tail on the right while negative skewness indicates a longer tail on the left of the distribution. The data points corresponding to DNA sequences of closely related species is expected to show, in a natural way, togetherness in terms of skewness values amalgamating within a cluster.

Entropy: Apart from skewness, we compute entropy of an image constructed from the DNA sequence. Entropy H is a measure of the amount of information contained in an image. It is defined as

$$H = - \sum_{i=0}^{G-1} p_i \log_2 (p_i), \quad (2)$$

In the context of images, entropy is often used to characterize the amount of uncertainty or disorder in the distribution of pixel values which is essentially used to capture textural similarities among DNA sequences. It is the central moment that measures sharpness of the histogram peaks in the frequency distribution of pixels. Here we consider the entropy value computed from a DNA sequence as a characteristic measure of phylogenetic nearness. The method mentioned here applies to the principles of texture analysis based on these moments (statistics) to identify and compute features of a DNA sequence.

2.4 Clustering with Statistical Features:

Machine learning clustering is a powerful technique used to uncover hidden patterns and group similar data points together. In this context, statistical features play a vital role in characterizing the underlying structure of the data. Utilizing above mentioned statistical features (skewness and entropy) in machine learning clustering algorithms such as K-Means, Hierarchical Clustering, or DBSCAN enables us to group data points with similar statistical properties. These methods consider the distribution of features within each cluster to

find natural groupings and patterns within the data. Here we use unsupervised clustering technique i.e., K-Means clustering.

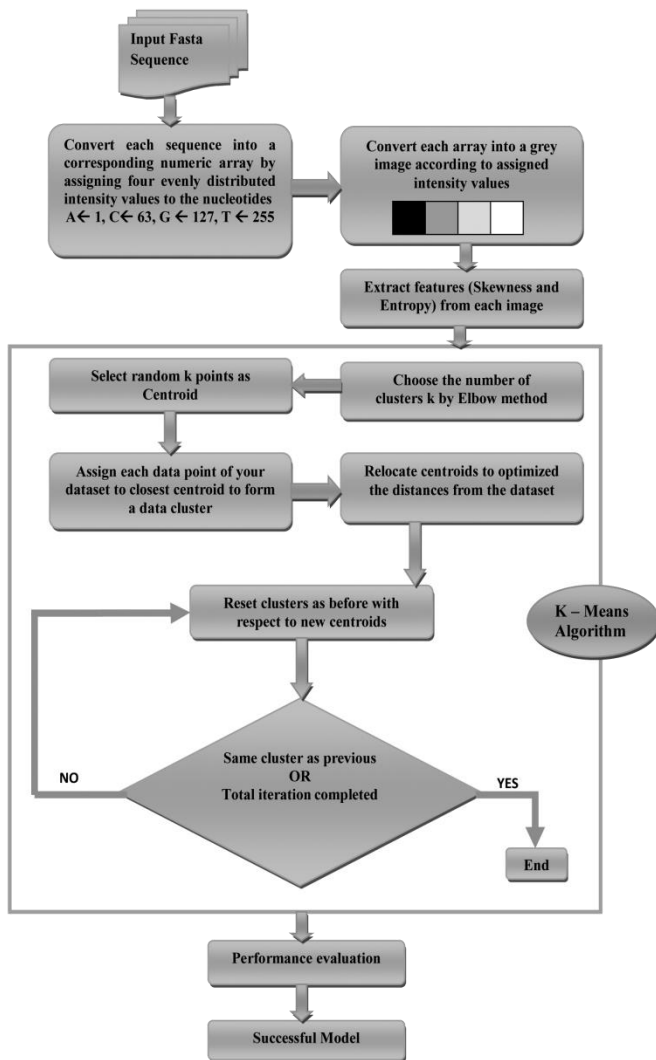


Fig. 1. Flowchart depicting the stepwise conversion of DNA sequence into an image consisting of four different intensities corresponding to different pixel values and subsequent generation of a successful model through the extraction of features from the image

3. EXPERIMENTAL RESULTS



Fig. 2 Grey image of the numeric array constructed from the DNA sequence of Baboon.

First, we read the fasta sequence of mammals mtDNA and convert the DNA sequence into a numerical array. Then we use this array to create the image shown in Fig. 2 by inserting the pixel intensities at the corresponding array elements. Using equations 1 and 2 we extract two statistical features viz., skewness and entropy from the image as shown in Table. 1. The primary output of unsupervised learning is the discovered structure or grouping within the data. This may take the form of cluster assignments for individual data points, centroids for each cluster and visual representations of the data's inherent structure, such as scatter plots, heatmaps or dendrograms. Here we choose clustering technique to visualize the data.

In Fig. 3, the graph illustrates the relationships among distinct species, employing the K-Means clustering method. This graph visually conveys the degree of similarity/dissimilarity among these species, offering insights into their evolutionary connections.

In the context of testing efficiency of our method we have made a comparative study with the benchmark dataset pertaining to fish mtDNA. We find the result of our analysis to be quite promising. Indeed, to quantify the amount of similarity, we obtain the normalized Robinson-Foulds (nRF) metric between our dendrogram and that in the esteemed AF-Project [3]. Having published our method in the benchmark, the nRF turns out to be 0.77. In Fig. 4 we obtain the corresponding phylogenetic tree for fish mtDNA sequence.

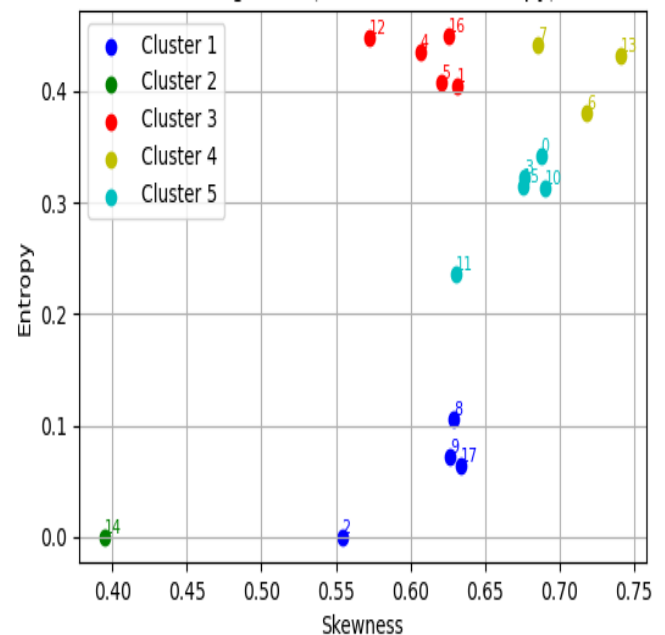


Fig. 3 K-Means clustering in 2D on the data set for mammals

Table. 1 Two features viz., skewness and entropy are obtained from the images of 18 species of mammals.

Sl No.	Species	Skewness	Entropy
0	Baboon	0.68746	0.34233
1	Blue_whale	0.63108	0.40532
2	Cat	0.55412	0.32023
3	Common_chimpanzee	0.67651	0.32198
4	Cow	0.60677	0.43494
5	Fin_whale	0.62116	0.40825
6	Gibbon	0.71840	0.38087
7	Gorilla	0.68515	0.44225
8	Gray_seal	0.62889	0.10560
9	Harbour_seal	0.62657	0.07239
10	Homosepiens	0.69009	0.31283
11	Horse	0.63063	0.23573
12	Mouse	0.57241	0.44780
13	Orangutan	0.74093	0.43136
14	Platypus	0.39506	0.32062
15	Pygmy_chimpanzee	0.67535	0.31427
16	Rat	0.62592	0.44948
17	White_rhinos	0.63357	0.06434

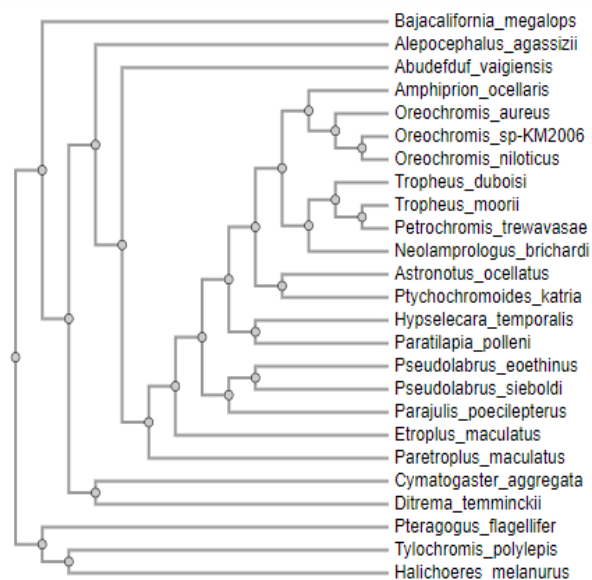


Fig. 4 Dendrogram for fish mtDNA sequence having 25 species using our algorithm.

4. CONCLUSION

In this article we have adopted an interdisciplinary approach to unravel the genetic diversity. By combining principles of image processing and machine learning, we have retrieved genetic information. The synergy among image processing, encoding and machine learning yields a comprehensive study that enhances the accessibility of genetic data. Here, we have been able to place in a novel way the different species in the right cluster by applying lesser number of physically relevant statistics compared to other well-known algorithms in the context of phylogenetic

reconstruction. Moreover, our method expedites the identification of evolutionary relationships for huge number of genetic data. We believe that in future it is possible to extend our existing algorithm to achieve superior results through the incorporation of deep learning.

REFERENCES

- Menlove K.J., Clement M. & Crandall K.A. (2009). Similarity Searching Using BLAST, Journal of Bioinformatics for DNA Sequence Analysis, Methods in Molecular Biology 537, 1–22. https://doi.org/10.1007/978-1-59745-251-9_1
- Pearson W.R. & Lipman D.J. (1988). Improved tools for biological sequence comparison. Proc Natl Acad Sci USA 85: 2444-2448. <https://doi.org/10.1073/pnas.85.8.2444>
- Needleman S.B. & Wunsch C.D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol. 48(3):443-53. [https://doi.org/10.1016/0022-2836\(70\)90057-4](https://doi.org/10.1016/0022-2836(70)90057-4)
- Smith, T. F. & Waterman, M. S. (1981). Identification of common molecular subsequences. Journal of Molecular Biology, 147(1), 195-197. [https://doi.org/10.1016/0022-2836\(81\)90087-5](https://doi.org/10.1016/0022-2836(81)90087-5)
- Thompson J.D., Plewniak F. & Poch O. (1999) A comprehensive comparison of multiple sequence alignment programs, *Nucleic Acids Research*, 27(13), 2682–2690. <https://doi.org/10.1093/nar/27.13.2682>
- Morgenstern B., Dress A. & Werner T. (1996). Multiple DNA and protein sequence alignment based on segment-to-segment comparison. Proc Natl Acad Sci U S A, 93(22):12098-103. <https://doi.org/10.1073/pnas.93.22.12098>
- Zielezinski A., Girgis H.Z, G. & Bernard et al. (2019). Benchmarking of alignment-free sequence comparison methods, *Genome Biol.*,20, 144. <https://doi.org/10.1186/s13059-019-1755-7>
- Zielezinski A., Vinga S., Almeida J. & Karlowski W.M. (2017) Alignment-free sequence comparison: benefits, applications, and tools, *Genome Biol.*, 18, 186. <https://doi.org/10.1186/s13059-017-1319-7>
- Bohnsack K.S., Kaden M., Abel J. & Villmann T. (2023) Alignment-Free Comparison: A Systematic Survey from a Machine Learning Perspective, in *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 20(1),119-135. <https://doi.org/10.1109/TCBB.2022.3140873>
- Vinga S. & Almeida J. (2003). Alignment-free sequence comparison - a review, *Bioinformatics*, 19, 513–523.. <https://doi.org/10.1093/bioinformatics/btg005>
- Blaisdell B. E. (1989). Average values of a dissimilarity measure not requiring sequence alignment are twice the averages of conventional mismatch counts requiring sequence alignment for a computer-generated model system, *J. Mol. Evol.*, 29(6), 538–547. <https://doi.org/10.1007/BF02602925>
- Bonham-Carter O., Steele J. & Bastola D. (2013). Alignment-free genetic sequence comparisons: a review of recent approaches by word analysis", *Brief. Bioinform.* 15, 890–905. <https://doi.org/10.1093/bib/bbt052>
- Vinga S. (2014). Information theory applications for biological sequence analysis, *Brief. Bioinform.*, 15, 376–389. <https://doi.org/10.1093/bib/bbt068>
- Anjum N., Nabil R.L., Rafi R.I., Bayzid Md. S. & Rahman M. S. (2023). CDMAWS: An alignment-free phylogeny estimation method using cosine distance on minimal absent word sets. *IEEE/ACM Trans. Comput. Biol. Bioinform.*20(1), 196-205. <https://doi.org/10.1109/TCBB.2021.3136792>
- Yi H. & Jin L. (2013). Co-phylog: an assembly-free phylogenomic approach for closely related organisms, *Nucleic Acids Res.*, 41(7), e75. <https://doi.org/10.1093/nar/gkt003>
- Gardner S.N., Slezak T. & Hall B. G. (2015). kSNP3.0: SNP detection and phylogenetic analysis of genomes without genome alignment or reference genome, *Bioinform.*, 31(17), 2877–2878. <https://doi.org/10.1093/bioinformatics/btv271>

17. Luczak B. B., James B. T. & Girgis H. Z. (2019). A survey and evaluations of histogram-based statistics in alignment-free sequence comparison", *Briefings Bioinf.*, 20(4), 1222-1237.
<https://doi.org/10.1093/bib/bbx161>
18. Hoang T., Yin C. & Yau SS. (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>
19. Yang A., Zhang W., Wang J., Yang K., Han Y. & Zhang L. (2020) Review on the Application of Machine Learning Algorithms in the Sequence Data Mining of DNA. *Front Bioeng Biotechnol.*8:1032.
<https://doi.org/10.3389/fbioe.2020.01032>
20. Zou Q., Lin G., Jiang X., Liu X. & Zeng X. (2020). Sequence clustering in bioinformatics: an empirical study, *Briefings in Bioinformatics*, 21(1), 1–10.
<https://doi.org/10.1093/bib/bby090>
21. Kobori Y. & Mizuta S. (2016) Similarity Estimation Between DNA Sequences Based on Local Pattern Histograms of Binary Images. *Genomics Proteomics Bioinformatics*. 14(2):103-12.
<https://doi.org/10.1016/j.gpb.2015.09.007>
22. Chen W., Liao B. & Li W. (2018). Use of image texture analysis to find DNA sequence similarities. *Journal of Theoretical Biology*. 455, 1-6.
<https://doi.org/10.1016/j.jtbi.2018.07.001>
23. Delibas E., & Arslan A. (2020). DNA sequence similarity analysis using image texture analysis based on first-order statistics. *J Mol Graph Model*, 99, 107603.
<https://doi.org/10.1016/j.jmgm.2020.107603>
24. Shamir L., Wolkom C.A. & Goldberg I.G. (2009). Quantitative measurement of aging using image texture entropy, *Bioinformatics*, 25(23), 3060–3063.
<https://doi.org/10.1093/bioinformatics/btp571>
25. Cussi, D.P. & Arceda, V.E M. (2023). DNA Genome Classification with Machine Learning and Image Descriptors. In: Arai, K. (eds) *Advances in Information and Communication. FICC 2023. Lecture Notes in Networks and Systems*, vol 652. Springer, Cham. 652. 39–58.
https://doi.org/10.1007/978-3-031-28073-3_4
26. Coates A. & Ng A.Y. (2012). Learning Feature Representations with K-Means. In: Montavon, G., Orr, G.B., Müller, KR. (eds) *Neural Networks: Tricks of the Trade. Lecture Notes in Computer Science*. Springer, Berlin, Heidelberg. 7700, 561–580.
https://doi.org/10.1007/978-3-642-35289-8_30
27. Permuter, H., Francos, J., & Jermyn, I. (2006). A study of Gaussian mixture models of color and texture features for image classification and segmentation. *Pattern Recognition*, 39(4), 695-706.
<https://doi.org/10.1016/j.patcog.2005.10.028>
28. Ward J. H. (1963) Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301), 236-244.
<https://doi.org/10.1080/01621459.1963.10500845>
29. Xu Y., Srinivasan A. & Xue, L. A. (2021). Selective Overview of Recent Advances in Spectral Clustering and Their Applications. In: Zhao, Y., Chen, (.DG. (eds) *Modern Statistical Methods for Health Research. Emerging Topics in Statistics and Biostatistics* . Springer, Cham., 247-277.
https://doi.org/10.1007/978-3-030-72437-5_12
30. F. Pedregosa et al. (2011). Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning, Research*, 12, 2825–2830.
<https://dl.acm.org/doi/10.5555/1953048.2078195>
31. Zhang F., Du B., Zhang L. & Zhang L. (2016). Hierarchical feature learning with dropout k-means for hyperspectral image classification, *Neurocomputing*, 187, 75-82.
<https://doi.org/10.1016/j.neucom.2015.07.132>

AUTHORS



Probir Mondal received his BS. degree in Computer Science from University of Calcutta, Kolkata, India in 2008 and Master of Science in Computer Science from the same institution in 2010. He was awarded MTech. in Computer Science & Engineering from the same institution in 2013. He is currently pursuing PhD at the Department of Computer Science in West Bengal State University, West Bengal, India. His area of interest includes graph theory, analysis of

algorithm, image processing, artificial intelligence, and bioinformatics.

Email: probir.mondal@prtgc.ac.in



Debranjana Pal received MTech degree in Computer Science and Engineering from University of Calcutta, India in 2013. He is currently pursuing PhD from Department of Computer Science and Engineering, Indian Institute of Technology Kharagpur, India.

His research interests are in the areas of machine learning, cryptography, and security.

Email: debranjana.crl@gmail.com



Krishnendu Basuli received his master's degree in computer science from SRTM University, Maharashtra. He was awarded PhD from University of Calcutta, West Bengal, India in 2015. His area of interest includes graph theory, logic design, operating system, analysis of algorithms, computer programming methodology, bioinformatics, computer architecture and organization, system programming and soft computing.

Email: krishnendu.basuli@gmail.com



Pratyay Banerjee received his bachelor's degree in physics from Jadavpur University, Kolkata, India in 2004 and Master of Science in Physics from the same institution in 2007. He was awarded PhD from Saha Institute of Nuclear Physics, Kolkata in 2015. His area of interest includes quantum integrable systems, mathematical physics, artificial intelligence, image processing and bioinformatics.

Corresponding author Email:

pratyay.banerjee@prtgc.ac.in