Machine learning techniques in biological data classification and clustering: Initiation of a scientific voyage

Amit Kumar Banerjee*[1], Neelima Arora[2]

[1]Biology Division, CSIR-Indian Institute of Chemical Technology, Uppal Rd, IICT Colony, Tarnaka, Hyderabad, Telangana 500007, India.[2]Institute of Science and Technology, Jawaharlal Nehru Technological University, Kukatpally, Hyderabad, Telangana 500085, India.

Abstract: Machine learning (ML) techniques have revolutionized the way of data classification, clustering, segregation, and novel element identification. ML techniques are having tremendous impetus for biological complex data classification. A number of studies reported novel data classification methods, complex biological element classification, and clustering. The present article briefs our experience in classifying biological species based on the biomarker genes and important proteins using state-of-the-art machine learning algorithms including artificial neural networks, support vector machines, decision trees, Bayesian methods, etc. Increased complexity warranted thorough human investigations and inspection to have a better classification on a case-by-case basis. Obtained outcomes were satisfactory and yielded novel strategies along with identifying the comparative superiority of specific algorithms for the specific datasets. However, obtaining a universal method or strategy remains the future objective. Automation of the process and precision increment for classification and clustering of the multi-parametric complex biological datasets are the other future goals.

Keywords: Machine-learning, Intelligent techniques, Biological classification, Deep learning, Artificial neural network, Support vector machines.

Citation: Amit Kumar Banerjee and Neelima Arora (2020) Machine learning techniques in biological data classification and clustering: Initiation of a scientific voyage. Journal of PeerScientist 2(1): e1000011.

Received February 09, 2020; **Accepted** February 27, 2020; **Published** March 04, 2020.

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Competing Interests: The authors have declared that no competing interests exist.

* **E-mail:** amitk_b@yahoo.co.in | **Phone:** +91-9885774127

I. INTRODUCTION

Human always seeks to unravel nature's mysteries through comprehensible logical and scientific explanations. Thus, the evolution of science had an extraordinary journey with amazing tales of experiments, discoveries, and brave voyages of explorations. In life sciences, a logical pattern or grouping of entities was scientifically initiated by Linnaeus, and his efforts were appreciated and summarized marvelously by acknowledging that "God created the world, Linnaeus put it in order."[1].

Biological classification seemed to be the only logical method to group and understand the enormous flora and fauna that inhabit this planet. The progress of systematic classification paved the way for understanding species. This endless journey is still continuing. Biological designs of the organism from the kingdom to the molecular level have been the richest tapestry of nature which we are attempting to appreciate and unfold. With growing knowledge, the complexity increased manifolds. Hence, the process of having a concluding solution at the moment is arduous. The growing amount of data and parameter complexities also pose a great number of challenges. Therefore, understanding biological and physical patterns is an enormous challenge to our race.

Why understanding such phenomena is so important to us? Such a phenomenon is associated with our very existence, and will definitely direct our future survival as well. For instance, understanding, or at least predicting the future evolution of pathogens might save us from upcoming epidemics. Human evolution pattern analysis can possibly allow us to know the course of our own future evolution. Information about all organisms can help us in accounting all the living beings systematically; crucial life cycle data may allow us to unveil the hidden switches that may trigger the change in the pattern of our life. Therefore, individual small steps taken today towards comprehending mysteries of life forms may have a great impact on us tomorrow. Classifying or data segregating has been the primary task in these aspects.

For example, molecular taxonomy claimed its own crown for successfully identifying, and annotating a novel species and aid in the journey of natural flora or fauna discovery. Several modern species biomarkers such as Internal Transcribed Spacer (ITS) [2-3], and Cyclooxygenase or COX genes [4] proved their efficiency over the time. Similarly, disease-specific markers [5-6] offered tremendous support for identification of a disease condition and differentiating them from healthy individuals or controls. Therefore, classification and clustering is multidimensional today and directly associated with our day to day life.

Mystery of patterns in nature:

Nature is known as the master artist that created logical and scientific patterns. Recognizing and finding missing pieces of this intriguing puzzle is our responsibility. Interestingly, such information also corresponds to one or the other important scientific reasons which are essential for survival and growth on this very planet. For instance, the existence of phi code, Fibonacci sequence, and the golden ratio in uncountable natural aspects astonishes us even today [7].

As rightly mentioned by Richard P. Feynman, "Nature uses only the longest threads to weave her patterns, so that each small piece of her fabric reveals the organization of the entire tapestry." The existence of elements and living beings in this universe strictly follow physical laws and organized sequential behavior. Even the most apparently disorganized event follows a physical law and organized behavior at its core. Often, our limited knowledge refrains us from scientifically understanding a spontaneous event. Progress of science, improvement in mathematical calculations, and an improved understanding of chaotic systems allowed us to logically explain numerous such processes.

Evolution was once considered a chaotic and discrete natural game without having any logical explanation. Even though, several intricate facts still require explanations, yet, we know that evolution follows strict mathematical calculations in many aspects. Our growing abilities to unwrap the cause-and-effects with scientific explanations can help in achieving success. Biological classification is one such topic that requires extensive high-dimensional mathematical, computational, and statistical exercises to comprehend the convoluted non-linearity. The modern era of interdisciplinary scientific practices has such aids and tools that may allow us to understand the existing complexity in an easy and better way.

Pattern recognition by classification and clustering for the ocean of data:

High-throughput scientific data generation techniques such as Next Generation Sequencing (NGS) techniques, MicroArray, Molecular Dynamics, Genomics, transcriptomics, epigenomics, metabolomics, proteomics [8], matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) [9] provided us unlimited opportunities to generate extraordinary base-by-base information that may help us to minutely inspect each data source and compare with others. Gene, protein, RNA sequence and structure-based classification methods have become acceptable in this context (Figure 1). However, involved technical shortcomings and artifacts are yet to be managed for cent percent reliable outputs.

On the other hand, a revolution in the big data storage and analytic techniques simultaneously yielded multiple sophisticated options to preprocess, curate, and segregate gigantic data volumes efficiently. Therefore, hidden designs could be traced out to understand the evolution of the data source and possible future changes. However, technical limitations in data handling remain a major hurdle in the implementation of such advanced classification and clustering methodology. Two different aspects have become vital in this context, generation and standardization of primary and secondary data, and selection of the most important attributes that can truly differentiate the biological entities with respect to the specific context.

Our experience:

Earlier, we made attempts to understand the classification of different protein families and groups through advanced computational approaches. In this journey, we tried to understand the most specific important features that may be crucial in discriminating the protein molecules for various protein families and groups. Our experiments included different kinases such as CaMK/CAM kinase [10-11], AGC Kinase [12-15], histidine kinase [16], industrially important proteins such as xylanase [17] and pyruvate dehydrogenase [18], important complex structural proteins such as keratin [19], and so on. All these classification and clustering exercises enhanced our understanding of the intricacy of biological datasets, the requirement of case-by-case observations, optimization of parameters, and intelligent choice and application of the available algorithms.

Similar to the complex protein features, we made an effort to understand and discriminate gene and biomarker regions of disease vectors using state-of-the-art artificial intelligent system [20-21]. Even though we were able to generalize the individual cases for an acceptable

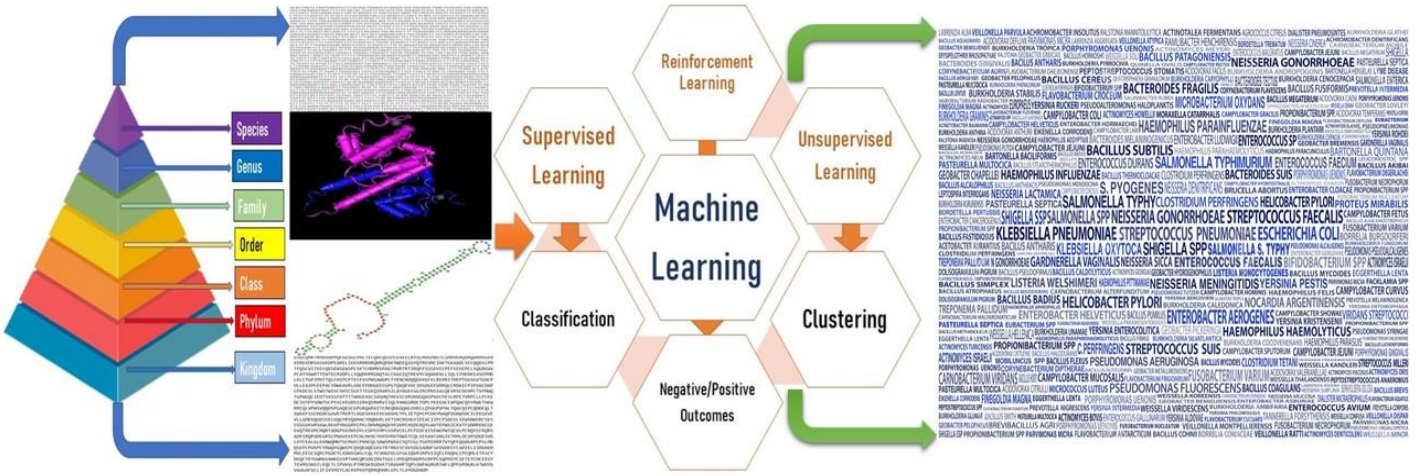

Figure1: A thematic presentation of understanding machine learning technique implementation in sequence and structure-based species identification.

outcome but our experience suggests that a lot of standardization and manual support is still required to improve the outcomes.

During the course of our studies, we utilized most of the available advanced machine learning-based classification and clustering techniques including Artificial Neural Network (ANN), Support Vector Machines (SVM), Decision tree, Classification And Regression Tree (CART), Self-Organizing Maps (SOM), Random Forest, and so on. Rigorous experiments were conducted for each case towards optimization of the parameters, iterations, error reduction, prediction accuracy improvement, and ROC values. For individual cases, we were successful in developing the best possible classification or clustering strategies in these studies. In addition, we identified the important discriminating biophysical parameters that were found statistically important in segregating the classes. However, our long term goal is to attain a common solution for this kind of sequence-based group identification and discrimination. As part of data integrity analysis, we were also successful in identifying specific sequence and secondary structural differences in highly conserved biomarker sequences [22].

Future perspectives:

Though for a long time, we have been experimenting with several other sequence datasets that are having highly overlapping features and biased sample numbers, yet we are quite far from developing a generalized strategy to classify or cluster protein and gene sequences based on their inherent or derived properties. Our decade long experience in this specific direction made us conclude that making AI a black box system to understand nature is still a cherished dream. However, researchers have been successful in developing important AI and machine learning-based applications that are being used extensively and successfully from product suggestions to customer target, face recognition, image analysis, and many other sectors. Recognizing natural hidden patterns at the molecular level with confirmed accuracy and explanation may require some more time.

II. Conclusion

Research efforts from all over the globe have already directed towards understanding the woven intricacy of nature. A distinct strategically sound and focused effort may help in attaining this goal. As Karl Pearson mentioned: “The classification of facts, the recognition of their sequence and relative significance is the function of science, and the habit of forming a judgment upon these facts unbiased by personal feeling is characteristic of what may be termed the scientific frame of mind”, dedicated, collaborative, global efforts in this direction may yield us success. Advances in technologies for effective and true data generation and improvement in pattern recognition techniques may reveal the hidden science behind the natural organization system one day.

Author’s Contribution: AKB and NA equally contributed in designing the study, reviewing the literature and writing the manuscript. Both authors have read and approved the final manuscript.

References

1. Gustafsson, Åke. "Linnaeus' peloria: the history of a monster." *Theoretical and Applied Genetics* 54.6 (1979): 241-248.
2. Kumari, Shipra, et al. "Internal transcribed spacer-based CAPS marker development for Lilium hansonii identification from wild Lilium native to Korea." *Scientia Horticulturae* 236 (2018): 52-59.
3. Minamoto, Toshifumi, et al. "Nuclear internal transcribed spacer-1 as a sensitive genetic marker for environmental DNA studies in

common carp Cyprinus carpio." *Molecular ecology resources* 17.2 (2017): 324-333.

4. Zubov, A. S., et al. "Description of a new species of Chrysina Kirby, 1828 (Coleoptera: Scarabaeidae: Rutelinae) from optima group, based on morphological characters and mtDNA COX I molecular marker." *Acta Biologica Sibirica* 5.4 (2019): 150-155.
5. Li, Jie, Dong Wang, and Yadong Wang. "IBI: Identification of Biomarker Genes of Individual Tumor Sample." *Frontiers in Genetics* 10 (2019): 1236.
6. Sokouti, Massoud, Mohsen Sokouti, and Babak Sokouti. "The Role of Biomarker Genes in the Diagnosis and Treatment of Non-small Cell Lung Cancer." *Current Respiratory Medicine Reviews* 14.3 (2018): 142-148.
7. Akhtaruzzaman, Md, and Amir A. Shafie. "Geometrical substantiation of Phi, the golden ratio and the baroque of nature, architecture, design and engineering." *International Journal of Arts* 1.1 (2011): 1-22.
8. D'Argenio, Valeria. "The high-throughput analyses era: are we ready for the data struggle?." *High-throughput* 7.1 (2018): 8.
9. Stanssens, Patrick, et al. "High-throughput MALDI-TOF discovery of genomic sequence polymorphisms." *Genome research* 14.1 (2004): 126-133.
10. Banerjee, A. Kumar, Neelima Arora, and U. S. N. Murty. "Classification and regression tree (CART) analysis for deriving variable importance of parameters influencing average flexibility of CaMK kinase family." *Electronic Journal of Biology* 4.1 (2008): 27-33.
11. Murty, U. S. N., Amit Kumar Banerjee, and Neelima Arora. "An in silico approach to cluster CAM kinase protein sequences." *J Proteomics Bioinform* 2 (2009): 97-107.
12. Banerjee, Amit Kumar, et al. "Exploring the interplay of sequence and structural features in determining the flexibility of AGC kinase protein family: a bioinformatics approach." *Journal of Proteomics and Bioinformatics* 1 (2008): 77-89.
13. Murty, U. S. N., Amit Kumar Banerjee, and Neelima Arora. "Application of Kohonen maps for solving the classification puzzle in AGC kinase protein sequences." *Interdisciplinary Sciences: Computational Life Sciences* 1.3 (2009): 173-178.
14. Banerjee, Amit Kumar, B. Poorna Manasa, and Upadhyayula Suryanarayana Murty. "Assessing the relationship among physicochemical properties of proteins with respect to hydrophobicity: A case study on AGC kinase superfamily." *Indian J Biochem Biophys* 47.6 (2010):370-377.
15. Banerjee, Amit Kumar, et al. "Towards classifying organisms based on their protein physicochemical properties using comparative intelligent techniques." *Applied Artificial Intelligence* 25.5 (2011): 426-439.
16. Banerjee, Amit Kumar, et al. "Application of intelligent techniques for classification of bacteria using protein sequence-derived features." *Applied biochemistry and biotechnology* 170.6 (2013): 1263-1281.
17. Arora, Neelima, et al. "Comparative characterization of commercially important xylanase enzymes." *Bioinformation* 3.10 (2009): 446.
18. Banerjee, Amit Kumar, et al. "Classification and clustering analysis of pyruvate dehydrogenase enzyme based on their physicochemical properties." *Bioinformation* 4.10 (2010): 456.
19. Banerjee, Amit Kumar, et al. "Keratin protein property based classification of mammals and non-mammals using machine learning techniques." *Computers in biology and medicine* 43.7 (2013): 889-899.
20. Banerjee, Amit Kumar, et al. "Classification and identification of mosquito species using artificial neural networks." *Computational Biology and Chemistry* 32.6 (2008): 442-447.
21. Banerjee, Amit Kumar, Neelima Arora, and U. S. N. Murty. "Stability of ITS2 secondary structure in Anopheles: what lies beneath." *International Journal of Integrative Biology* 1.3 (2007): 232-238.
22. Banerjee, Amit Kumar, Neelima Arora, and U. S. Murty. "How far is ITS2 reliable as a phylogenetic marker for the mosquito genera." *Electronic Journal of Biology* 3.3 (2007): 61-68.