International Journal of Engineering Trends and Technology
ISSN: 2231 – 5381 /doi:10.14445/22315381/IJETT-V68I10P210

Volume 68 Issue 10, 58-62, October 2020
© 2020 Seventh Sense Research Group®

A Review of Data Mining Optimization Techniques for Bioinformatics Applications

Preeti Thareja[1], Rajender Singh Chhillar[2]

[1]*Research Scholar, Department of Computer Science and Applications, M.D. University, Rohtak, Haryana, India*
[2]*Professor, Department of Computer Science and Applications, M.D. University, Rohtak, Haryana, India*

[1]preetithareja10@gmail.com, [2]chhillar02@gmail.com

Abstract — *Geneticists are scaling up their attempts by using a range of investigational and genomics methodologies to understand biological functions. It has ended in a torrent of biomedical and clinical data, that can be daunting for scientists to manage with no adequate resources for information managing and examing, particularly when there is a lack of practice or coding, statistical and simulation expertise. Custom analytics tools have, therefore become highly essential in bioinformatics and can help speed up the research process. This paper provides a comprehensive overview of data mining techniques, methods of optimization and the evolving state of the bioinformatics industry in India.*

Keywords — *Bioinformatics, Data Mining, Optimization, Validation Metrics, India.*

I. INTRODUCTION

Geneticists are scaling up their research in the therapeutic setting to consider the genetic mechanisms driving disease passageways. As a result, genetic and therapeutic information is overloaded from gene expression and protein patterns, DNA grids, protein behaviours, clinical images, disease passageways, and electronic medical records. There are basic analytics challenges that need to be addressed to manipulate these details and uncover relevant information that can be converted into therapeutic interventions. Technical challenges such as managing noisy and fragmented results, performing the computer-intensive job and combining different data channels are current obstacles confronting post-genome geneticists.

Based on the genomic research study and individualized medicines, the worldwide bioinformatics market is expected to hit $19.8 billion by 2025. According to a study conducted by Modor Intelligence, the compound annual growth rate of the global bioinformatics market is around 20.75%. The bioinformatics market is expected to drive a large number of healthcare investments in the Asia Pacific area. In addition, high IoT technology implementation in this market is forecasted to have a significant effect on growth in the area. India is projected to be the main driver of the bioinformatics industry in the world. DNA, RNA and peptide sequencing demands government and private initiatives. This accelerates the development of molecular biology and bioinformatics, thereby increasing fieldwork on proteomics and therapeutics. These are the key drivers for the worldwide bioinformatics industry growth, which is expected to undergo a raise over the forecast period.

This paper includes the following sections: Recent work in the field of bioinformatics applications has been mentioned in Section 2. In Section 3, the growth of bioinformatics in India has been discussed. Finally, in rest Section, Importance and Applications of Optimization and Data mining in the Bioinformatics Sector have been discussed.

II. RECENT TECHNICAL ADVANCES FOR OPTIMIZING BIOINFORMATICS DATA

Numerous researches related to bioinformatics data has been performed using different techniques of optimization and data mining. The author's methods, algorithms and observations performed over the last nine years are addressed below:

Han Luo et al. (2019) proposed NIHO network-based architecture for refining and finalizing genetic networks by combining 6 genomic network systems. NIHO is capable of learning genes high-level characteristics from a highly diverse network, consisting of the completed matrix and neural network. In the predicted area, the acquired low-dimensional descriptions are then used to measure the geometric distance of the genes. Ultimately, by reviewing the proximity ratings, NIHO implies the relationships between genes and adds those relationships that did not originally exist within genetic networks. The work has been verified with accuracy and network performance [1].

Ko-Wei Huang et al. (2019) proposed an optimal search and quick convergence clustering algorithm based on PSO and GS algorithm, called the Memetic Particle Gravitation Optimization (MPGO). MPGO's key processes are integrated activity and enhancement of diversity. Integrated activity constitutes the interchange of individuals from two sitepopulations following a predetermined amount of function analysis. Diversity enhancement constitutes an operator called an enhancement, related to the differential evolution method of crossover, to enrich the multiplicity of the individual system [2].

Bin Hu et al. (2018) suggested methodology that examines the space of potential subsets to achieve a feature set that optimizes predictive precision and reduce unnecessary features in biomedical high-dimensional information and 1-nearest neighbour is evaluator. The authors enhanced SFL procedure that adds a weight feature of the disorder, an absolute group approach of balance and an adaptive transfer feature and proves better classification accuracy [3].

Jesse M. Zhang et al. (2018) proposed an immersive analysis and description algorithm for addressing the single-cell RNA sequence clustering challenge. It is called DendroSplit. This aims at commercially generating several clusters for the dataset. It provides reasons for all choices made during the cluster generation process. The technique includes preprocessing data, measuring pair distances, doing hierarchical Clustering, dividing and finally merging to obtain improved outcomes of clustering problems [4].

G. Surya Narayana et al. (2017) suggested a similarity-based categorical data clustering strategy in which the inter- and intra-attribute similarity are simultaneously measured and thus merged to enhance performance. The similarity-based clustering method of K-medoids is implemented to cluster the features depending on the Euclidean distance to reduce the overhead. The optimization strategy for the bee colony (BC) is implemented to pick the best functionality for users. The metrics used for claiming the results are Accuracy, Adjusted Rand Index, Clustering Error, Convergence Time, and Data Dimensionality [5].

Huang et al. (2017) presented a novel genetic clustering algorithm to pick the number of clusters automatically. The method employs harmonious mating and crossover operators to direct the population towards a stronger convergence direction and thereby improving the efficiency of Clustering on genuine-life and synthetic datasets. The new single-point crossover operator known as the variable-length-and-gender-balance crossover, is designed to ensure a probabilistic balance in both population gender ratio and chromosome length dynamics [6].

Rami Al-Dalky et al. (2016) proposed Monte Carlo Genetic Network Simulator (MCforGN). It is a system of classifiers that develops genetic networks, recognizes functionally linked genes, and predicts interactions between gene-diseases. It examines findings using a set of mathematical methods to determine the probability of the co-occurrences of genes involved. MCforGN can diagnose correlations between gene-diseases by using a mixture of core steps (to classify the key genes of disorder-specific genetic networks) and Monte Carlo simulation [7].

N. N. R. Ranga Suri et al. (2015) suggested a new technique for grouping categorical data mainly to identify outliers, by changing the traditional k-modes algorithm. The confusion about the clustering method is resolved by taking a rough sets-based approach to soft computing. The sensitivity study conducted as part of this study shows the effect of different parameters on the efficiency of the proposed algorithm [8].

Feng Jiang et al. (2015) portrayed two separate k-modes algorithms for initialization, one centered on the conventional distance-dependent outlier identification strategy and the other centered on the entropy dependent outlier identification strategy for partitions. The methods can ensure that the selected preliminary cluster cores are not outliers through use of the outlier identification strategies to measure the extent of outlierness of each entity. A new distance variable, weighted matching, is introduced in the initialize step, to measure the distance between two entities defined by categorical values [9].

Ujjwal Maulik et al. (2013) formulated a multi-objective genetic algorithm-based biclustering strategy which at once preserves three objective functions for obtaining dense biclusters with strong interaction strengths. The suggested biclustering approach is used to classify specific association modules on the databases of scientifically confirmed and expected associations among a group of HIV-1 proteins as well as a collection of human proteins. For this, all the knowledge about the association is understood as a bipartite graph. Such human proteins could be potential targets for anti-HIV drugs [10].

Ching-Seh Wu et al. (2012) implemented an evolutionary algorithm for the dynamic optimization of biomedical content from dispersed and heterogeneous contexts, which can greatly improve diagnosis and therapy decision-making. A multi-agent system is introduced to define, track and address any inconsistencies between the configured task sequence and real biomedical records [11].

Razwan Andonie et al. (2011) propose two artificial intelligence estimation methods for small training sets at the cost of computing overhead. Both methods are FAMR-based. The first method, named GA-FAMR, consists of two phases, wherein the genetic algorithm is used in the first stage to optimize the importance of the training data assigned. Thereby, it strengthens FAMR's generalization ability. Enhanced relevance is used in the second stage to give inputs to FAMR. Another method named, Ordered FAMR comes from a verified algorithm of Dagher et al. All proposed strategies are used to forecast recently made HIV-1 protease inhibitors [12].

Aaron M Newman et al. (2010) combined three strategies: Kohonen Self-organizing map from machine learning, density-equalizing cartography algorithm from cartography and minimal spanning tree from graph theory to a new method of automated self-organizing network ensemble (AutoSOME). The three strategies combine to distinguish important node groups and anomalies from high-dimensional information. SOM is used to minimize the broad and unfiltered source dataset. DE helps transform every node lattice group to spatial point grouping. MST detects multi-geometry clusters. A further significant function is an average ensemble, increasing performance reliability and cluster efficiency [13].

TABLE 1 lists the features of the techniques used in the literature survey studied. Along with the features, it lists the datasets where all these techniques are applied.

TABLE I: STATE OF THE ART TECHNIQUES AND DATASETS USED

Techniques	Features	Datasets Used
Genetic Algorithms with Neural Networks [1]	It increases network forecasting efficiency.	Heterogeneous disease gene sets.
Hybrid PSO and GSA memetic clustering algorithm [2]	It contributes to quick convergence and an effective search.	Homogeneous datasets and image segmentation datasets.
k-Nearest Neighbors with improved shuffled frog leaping algorithm [3]	It increases the recognition of related subsets with greater accuracy in the classification.	High-Dimension biomedical datasets.
Hierarchical Clustering with a split based on separation score [4]	This improves interactivity and subjectivity clustering and distinguishes various levels of biologically relevant samples in the results.	Single-cell RNA sequence datasets.
k-medoids with association rule mining [5]	This increases the efficiency of resource allocation and reduces overheads for computations.	Spatio-Textual datasets, Categorical datasets.
Genetic Clustering with eugenic theory [6]	This increases clustering efficiency by dynamically assessing cluster numbers and their cores.	Homogeneous datasets.
Text mining with Monte-Carlo simulation [7]	This helps to build new gene pathways and will identify any incomplete genes.	Genome datasets.
Rough k modes with ranking based outlier analysis and detection [8]	This answers confusion about the inclusion of an outlier entity in the cluster.	Categorical data.
k modes with distance-based and partition entropy-based outlier detection [9]	This assures the discovery of original cluster centres, which are not outliers.	Categorical data.
Multiobjective Biclustering [10]	The primary objective is to classify small clusters with strong strengths of interaction.	PPI datasets.
Intelligent agent based on JAVA with evolutionary computing [11]	This addresses the problem of collecting data and maximizing the output of decentralized static and dynamic information.	Heterogeneous medical datasets.
A fuzzy neural network with genetic computing [12]	It is used with greater accuracy for assessing biological behaviours of protease inhibitors.	Biological molecular datasets.
Self-organizing maps with minimum spanning tree clustering [13]	This easily recognizes distinct and fuzzy clusters of data without prior knowledge of the amount or arrangement of clusters.	High dimensional microarray data

III. TECHNIQUES FOR HANDLING BIOINFORMATICS DATA

Data mining techniques are needed to address complex data processing problems in accordance with optimization techniques, where chaotic and fragmented data and high computational-intensive challenges have to be managed. This allows scientists to deliver useful results and observations from a variety of biochemical, scientific and pharmacological treatments. Validations are vital to prove work and meet the researcher's needs.

A. Data Mining Techniques

Data mining is based on two data types, structured data and unstructured data. Data with the predefined format are classified under structured data, while data with no specified manner are unstructured. There are different strategies to manage this type of data, such as classification clustering, association rules, regression, etc. TABLE 2 lists several important methods used in the current work.

TABLE II: SUMMARY OF DATA MINING TECHNIQUES USED IN THE CURRENT WORK

Data Mining Techniques	References
K-modes	[5], [8], [9]
K-means	[6], [12]
Graph Clustering	[10], [13]
k-Nearest Neighbors	[3]
Hierarchical Clustering	[4]
Text Mining	[7]
Partitional Clustering	[2]

B. Optimization Techniques

For improve solutions to problems, optimization techniques come into play, genetic algorithms, swarm intelligence, memetic algorithm, gravitational search, etc. Among which genetic algorithms are in demand, as seen in TABLE 3. Genetic algorithms are inspired by human evolutionary biology, including reproduction, crossover and mutation operators.

TABLE III: SUMMARY OF OPTIMIZATION TECHNIQUES USED IN THE CURRENT WORK

Optimization Techniques	References
Genetic Algorithm/Evolutionary Algorithm	[1], [6], [10]–[12]
Swarm Intelligence	[2], [5]
Ensemble Learning	[13]
Frog Leaping	[3]

C. Validation Techniques

Validation is the function of checking the performance of data mining models against actual data. Before deploying into a production environment, it is critical to validate analytical models by analyzing their reliability and functionality. TABLE 4 shows some of the metrics covered in the current work.

TABLE IVV: SUMMARY OF METRICS USED IN THE CURRENT WORK

Metrics	References
Accuracy	[1]–[3], [5], [7]–[9], [11], [12]
Recall, Precision	[7], [9]
Adjusted Rand Index	[4], [5]
DBI, NMI, VRC, CS	[6]
Purity	[8]
Wilcoxon Rank Test	[13]
z-score	[10]
*DBI: Davis Bouldin Index, VRC: Variance Ration Criterion, CS: Cluster Similarity, NMI: Normalised Mutual Information	

IV. BIOINFORMATICS GROWTH IN INDIA

India was one of the world's first countries to develop a network of national bioinformatics companies. In 1986 the Biotechnology Department (DBT) initiated a curriculum on bioinformatics. The DBT branch, the Biotechnology Information System Network (BTIS), now links 57 primary research centres, serving the entire world. More than 100 Therapeutics databases have been created. Furthermore, several major global databases were established under the National Jai Vigyan Project, with implications for proteomics and genomics.

The central strength of R&D in Indian biotechnology is its well-trained staff, which is strongly supported by experts who are excellently proficient in mathematics, chemistry and physics. The nation, therefore, lacks the technical ability to manage all facets of the collection, production, study and evaluation of biological knowledge.

India's popular technological talents are another significant asset in the field of bioinformatics. There is the liberty to work and analyze the data relating to gene sequencing, operational proteomics and genomics. Bioinformatics is among today's fastest-growing biotechnology areas in India. More than 200 firms are active in bioinformatics in Bangalore, Hyderabad, Pune, Chennai and Delhi.

V. BIOINFORMATICS APPLICATIONS

Bioinformatics has become extremely relevant, as large amounts of data collected by methods including protein sequencing and nucleic acid require data mining analytical techniques for potential research and therapeutic applications. The bioinformatics market is therefore predicted to fuel increased competition in the forecast period. The analytical techniques appear suitable for bioinformatics because they are rich in statistics. Some of the important applications of bioinformatics are shown in Figure 1.

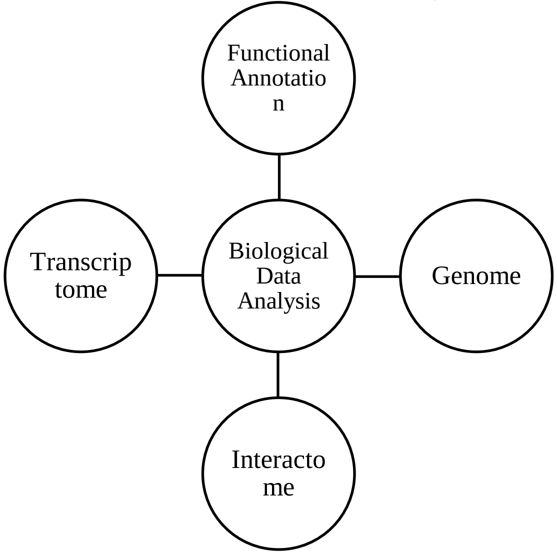

Fig. 1 Applications of Bioinformatics

1) *Transcriptome:* For the efficient diagnosis and evaluation of biological markers, a broad range of data formats from RNA-Seq, DNA sequencing, miRNA screening, ChIP-Seq, 4C-Seq, and spectrometry tests are targeted. For rigorous analysis and study of chemical processes and pathways from actual observations using techniques such as graph theories, empirical support is given.

2) *Interactome:* It deals with the study of both the relationships and the implications of those interactions between proteins and other cell molecules. It covers all protein-protein interactions occurring within a cell.

3) *Genome:* An organism's genome is all of its inherited details preserved in its DNA (or RNA). This covers both the genes and the non-coding DNA sequences.

4) *Functional Annotation:* It consists of adding biological data to genomics. In simple terms, it is referred to as the functional interpretation of genes.

Dale et al. (2019) established genetic-level consequences of disease by exploring patterns of gene expression in species utilizing biclustering in accordance with a genetic algorithm [14]. Tapan et al. (2016) took advantage of fuzzy auto-organizing maps to explore DNA motifs [15]. There are many studies already addressed in section 2 concerning bioinformatics applications. Figure 2 demonstrates the research of different authors for bioinformatics applications over 9 years from 2010 to 2019.

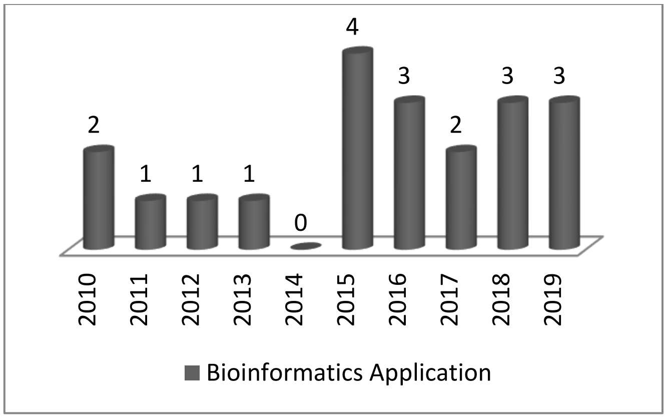

Fig. 2 Bioinformatics application usage during the period 2010-2019

TABLE 5 lists the applications of bioinformatics covered in the current work.

TABLE V: SUMMARY OF METRICS USED IN THE CURRENT WORK

Applications	References
Functional Annotation	[1], [3]–[9], [11], [13], [14]
Genome	[1], [7], [12], [15]
Transcriptome	[4], [13], [15]
Interactome	[10]

As can be seen, the interest resides in the field of functional annotation.

Empirical methods have been built and extended to a broad variety of genetic fields of study, such as coding and monitoring, systematic analysis of covariance as well as multiple regression analysis, extensive analytics in Medicare, and so on. Some of such methods are analysis, correlation, classification, preprocessing, Clustering, error rates, product testing, and anomaly detection. Visual Analytics is important to explain and disseminate information after the acquisition.

Genomics ability to help improve soil quality and crop yield is a new area of interest and hope in agriculture. If genetic information can be used to forecast crop yield or health (and subsequent effects on soil), farmers may adequately predict and maximize yields. The complete human ordering, the whole collection of genetic information among each human cell, has now been established. Recognizing these genetic paths ensures that scientists raise awareness of the trait of diseases and their cures, detect the strategies underlying evolutionary processes such as growth and ageing, and monitor our early development and connection with alternative species. The main barrier among investigators and the evidence they get is the vast amount of knowledge given. It's where data mining, along with optimizations plays a role.

VI. CONCLUSION

Most of the scientists are predominantly trained to glean new data. Biology historically lacks the means to evaluate massive databases such as the archive of the human genome. In conjunction with biology, the computer science discipline, colloquially known as bioinformatics, has been creating techniques and methods that assist many biologists in handling and examining the incredible portions of information that pledge to strengthen the human condition. One such technology is data mining. This paper offers a research study of biomedical data mining and optimizing techniques. Latest research relying on the past ten years was taken up and listed as to the challenges addressed, the data types used, and the frameworks used. The paper also gives an insight into bioinformatics research in India. India is growing faster in research projects to make a new history in biomedical data that would help the healthcare sector and drug discovery sector in several ways. This paper will help academics and data scientists to know the current scenario of data mining and optimization in the bioinformatics from an Indian outlook. Many drug companies minimize drug discovery expenditures as they turn their attention from outsourcing the technology to designing in-house techniques, enabling them to tailor their necessary programs to serve their needs and increase the efficiency of their workflow.

REFERENCES

[1] J. Liu, "*Databases for the Completion of Gene Networks,*" *IEEE Access*, vol. 7, pp. 168859–168869, 2019.

[2] K. Huang, Z. Wu, H. Peng, and M. Tsai, "*Memetic Particle Gravitation Optimization Algorithm for Solving Clustering Problems,*" IEEE Access, vol. 7, no. 2, pp. 80950–80968, 2019.

[3] B. Hu *et al.*, "*Feature Selection for Optimized High-Dimensional Biomedical Data Using an Improved Shuffled Frog Leaping Algorithm*," vol. 15, no. 6, pp. 1765–1773, 2018.

[4] J. M. Zhang, J. Fan, H. C. Fan, D. Rosenfeld, and D. N. Tse, "*An interpretable framework for clustering single-cell RNA-Seq datasets,*" pp. 35–39, 2018.

[5] G. S. Narayana and D. Vasumathi, "*An Attributes Similarity-Based K - Medoids Clustering Technique in Data Mining,*" *Arab. J. Sci. Eng.*, 2017.

[6] F. Huang, X. Li, S. Zhang, and J. Zhang, "*Harmonious Genetic Clustering,*" pp. 1–16, 2017.

[7] R. Al-dalky, K. Taha, D. Homouz, and M. Qasaimeh, "*Applying Monte Carlo Simulation to Biomedical Literature to Approximate Genetic Network,*" vol. 5963, no. c, pp. 1–10, 2015.

[8] R. Suri, "Detecting outliers in categorical data through rough clustering," 2015.

[9] F. Jiang, G. Liu, J. Du, and Y. Sui, "*Initialization of K -modes clustering using outlier detection techniques,*" *Inf. Sci. (Ny).*, 2015.

[10] U. Maulik *et al.*, "Mining Quasi-Bicliques from HIV-1 – Human Protein Interaction Network : A Multiobjective Biclustering Approach," pp. 1–14, 2012.

[11] C. Wu, I. Khoury, and H. Shah, *"Optimizing Medical Data Quality Based on Multiagent Web Service Framework,"* vol. 16, no. 4, pp. 745–757, 2012.

[12] R. Andonie, "*Fuzzy ARTMAP Prediction of Biological Activities for Potential HIV-1 Protease Inhibitors Using a Small Molecular Data Set,*" vol. 8, no. 1, pp. 80–93, 2011.

[13] A. M. Newman and J. B. Cooper, "*AutoSOME : a clustering method for identifying gene expression modules without prior knowledge of cluster number,*" 2010.

[14] J. Dale, "*Multi-objective Optimization Approach to find Biclusters in Gene Expression Data,*" no. 1.

[15] S. Tapan, D. Wang, S. Member, and A. Self-organizing, "*A Further Study on Mining DNA Motifs Using Fuzzy Self-Organizing Maps,*" vol. 27, no. 1, pp. 113–124, 2016.