Trials, Skills, and Future Standpoints of AI Based Research in Bioinformatics

Mubina Malik, Jaimin N Undavia

Abstract: ***In recent times, computer field has entered in all types of business and industries. Recent advancements in the information technology field, has open up many possibilities in multidisciplinary research. Machine learning, deep learning, convolution neural network, etc. are recent development in computer fields which has change the way of development of algorithms. Such algorithms can learn over a period of time while in execution and improves its performance and continue learning. Bioinformatics is the recent example of the science which strives to use such recent technologies of computer science for betterment in its own field. This article reviews Artificial Intelligence subset such as Machine learning and Deep learning in the genomics and proteomics domain. This article provides profound insights of various AI techniques which can be incorporated in the field of bioinformatics. The paper also highlighted the future research potential of this field. Computational biology, genomics, proteomics, Drug designing, gene expression level analysis are the major research areas in bioinformatics. These areas are also discussed in the paper.***

Keywords: ***Artificial Intelligence, Bioinformatics, Deep Learning, Genomics, Machine Learning, Proteins.***

I. INTRODUCTION

In recent era, multidisciplinary studies bring more attractions among researchers. Bioinformatics is very important field which has capacity to map the whole human genome computationally. At a very abstract level, bioinformatics can be defined as “collection, classification, storage, and analysis of biochemical and biological information using computers especially as applied to molecular genetics and genomics.”. Although, mapping of genomes of living organisms was done by scientist but the greatest challenge for them is how to compile and store the huge biological data. For this, computer had to be used to store this huge data. Bioinformatics can give a great opportunity to the researcher to store the data in the form of database. Also, biological data can be observed, analyzed and interpreted using computer.

If the data is in the raw form or not compiled, then the researchers cannot use them. It is easy to store the data but it is also complex and challenges the researcher to extract the required information from that huge data. Standalone biological data are good but they have not any potential until they are digitized. To carryout biological research it is important to make biological tool which extract the information from the database and can be analyzed for research. This biological tool can be used to understand the biology of an organism. The main challenges in computational biology, which require the development of tools and methods capable of transforming all these heterogeneous data into biological knowledge about the underlying mechanism [1]. For example, determine the protein sequence from DNA sequence which will help to understand the protein function in living organism. Research in Bioinformatics and computational biology include abstraction to implementation of new algorithms for data analysis to the development of database and web tools to access them. These tools and methods provide knowledge in the form of testable model. Bioinformatics is a new area of science where a combination of statistics, molecular biology, and computational methods is used for analyzing and processing biological information like gene, DNA, RNA, and proteins.

As mentioned in previous paragraph, information technology has to refine biological field by inculcating an efficient combination. Artificial Intelligence such as machine learning and deep learning plays a vital role in bioinformatics field. Genomics and proteomics technologies have enabled biology to enter the era of big data [2]. The field of machine learning, which aims to develop computer algorithms that improve with experience, holds promise to enable computers to assist humans in the analysis of large & complex data sets [3]. Machine learning and deep learning can be used to analyze and convert the huge/complex data set into the knowledge, however it also has some challenges which are discussed further. As mentioned above, due to the faster growth in advanced technology with practical applications in the computer field, bioinformatics can be considered by the programmer/computer technology professional to identify the future diseases which will shape the future of humanity.

Revised Manuscript Received on April 25, 2020.

Mubina Malik , Assistant Professor, Smt. Chandaben Mohanbhai Patel Institute of Computer Applications Faculty of Computer Science and Applications, Charotar University of Science and Technology, Changa.

Dr. Jaimin N Undavia, Assistant Professor, Smt. Chandaben Mohanbhai Patel Institute of Computer Applications Faculty of Computer Science and Applications, Charotar University of Science and Technology, Changa.

II. LITERATURE SURVEY

A Significant amount of work is found in this area. Here we have reviewed and used following references for this article.

Machine learning applicability has been shown by Pedro Larranag et al. in their article titled "Machine learning in bioinformatics" in year 2005. In this article, categorizations of machine learning methods are found and their versatile applications in the field of bioinformatics have been noted. Further, machine learning's capability in the field of genomics, proteomics, system biology, etc. is also discovered. [1]

In the article "Deep Learning in Bioinformatics" Seonwoo Min, Byunghan Lee, and Sungroh Yoon gave a review on input data, research objective, limitations and future directions of bioinformatics research with deep learning architecture. Moreover, they have discussed about the impact of Deep Learning techniques such as DNN, CNN, RNN and Emergent architectures in bioinformatics domain specifically in omics, biomedical imaging and biomedical signal processing with research avenues. Also, they have concluded that deep learning can be applied in the current discussed domain but it will not provide a great result and further studies required to exploit the capability of deep learning in bioinformatics. [4]

Manisha Mathur, Research fellow, Advanced Milk Testing Research Laboratory wrote research paper named "Bioinformatics challenges: A review" in the year 2018. The review highlights few challenges such as data management and organization, data mining, statistical analysis, software and algorithm development with its future perspectives in the field of bioinformatics. Unsolved age-old mysterious problems of biology such as nerve function, behavior and aging etc. are mentioned in the article. Important parameter which leads to problems in handling of queries for integration of biological database as a statistician and Bioinformatician was also listed out by the author. [5]

Alejandra J. Magana, Manaz Taleyarkhan, Daniela Rivera Alvarado, Michael Kane, John Springer and Kari Clase discussed the growing importance of bioinformatics in the education as well as in the research field in the paper "A Survey of Scholarly Literature Describing the Field of Bioinformatics Education and Bioinformatics Educational Research" published in 2014. Opportunities and challenges for integrating computing and biology were also stated. Authors have analyzed various research papers from 1998 to 2013 and represented the graphical view which shows the total 38 papers published for the related fields of bioinformatics such as technology education or computer science education, biology and biotechnology category, engineering education, education or science education in the respective years. [6]

Challenges of single cell genomics in the field of bioinformatics were discussed by Luwen Ning, Geng Liu, Guibo Li,Yong Hou,YinTong, Jiankui He in the article "Current Challenges in the Bioinformatics of Single Cell Genomics" published in the year 2014. In this paper authors have discussed the bioinformatics tools which are available for single cell as well as bulk cell genomics and stated that current bioinformatics tools developed for bulk cell sequencing do not work well with single cell sequencing data. Major challenges for single cell DNA/RNA sequencing data such as low genome coverage and high amplification bias are discussed with the suggestions to overcome the challenges. Also, authors mention that the primary nucleic acid sequence analysis of single cell genomic DNAs and RNAs will be solved in a few years. [7]

Research article "Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics" published in 2015 by Ehsaneddin Asgari, Mohammad R.K. Mofrad gave a brief idea about the feature extraction methods: BioVec, ProtVec and GeneVec. It is clearly mentioned in the article that machine learning techniques in bioinformatics can widely benefit from ProtVec and GeneVec representation. These representations can be considered as pre-training for various application in deep learning and machine learning in bioinformatics. Moreover, ProtVec can be used in protein interaction predictions, structure prediction, and protein data visualization. [8]

Authors Maxwell W. Libbrecht and William Stafford Noble gave the introduction of Machine Learning applications for the analysis of genomics dataset including epigenetic, proteomics and metabolomic data in the paper entitled "Machine Learning applications in genetics and genomics" in the year 2015. Authors gave outline of the challenges in applying machine learning methods to practical problems in genomics. Mainly they focused on heterogeneous data because classification of machine learning methods uses fixed-length vectors of real number, so the methods cannot be applied directly to heterogeneous data. In this article researcher discussed and specify the links of few collaborative projects which can be used to avail the facility of large data set of genomics. Moreover, guideline for the selection of machine learning methods & practical applications for the analysis of genomic data set was also mentioned. [3]

DeepPrime2Sec model for predicting protein secondary structure from primary structure with most challenging dataset Q8 (8 classes) on CullPDB/CB513 was developed by Ehsaneddin Asgari, Nina Poerner , Alice C. McHardy, and Mohammad R.K. Mofrad in the year 2019 which was published in the article "DeepPrime2Sec: Deep Learning for Protein Secondary Structure Prediction from the Primary Sequences". Authors have given different protein sequence representations: one-hot vectors, biophysical features, protein sequence embedding (ProtVec), deep amino acid contextualized embedding (ELMo), and the Position Specific Scoring Matrix (PSSM) & different deep-learning architectures: convolutional neural networks (CNN), recurrent neural networks (in particular Bi-LSTM), use of highway connection, attention mechanism, and multi-scale CNN and showed that PSSM and its combination with one-hot vectors gave 0.699 accuracy and CNN-BiLSTM architecture model gave 69.9% accuracy which was the best performance in protein secondary structure prediction. They also mentioned that by ignoring the boundary amino acids from the evaluation, the Q8 accuracy would increase for an extra 20%. [9]

A comparison using updated benchmark datasets were provided by Felipe Kenji Nakano, Mathias Lietaert and Celine Vens in the research article "Machine learning for discovering missing or wrong protein function annotations" in 2019.

Authors presented 24 new datasets for predicting protein function annotations. Prediction probability of annotations which changed between the 2007 and 2018 versions were also highlighted by the authors. Different methods such as ClusEnsemble, HMC-GA, HMC-LMLP and AWX compared with different datasets. Although, prediction task is more challenging for machine learning, authors suggest to instigate deep learning techniques which are used without the need of extracting features. [10]

Daniel Quang and Xiaohui Xie proposed a novel hybrid convolutional and bi-directional long short-term memory recurrent neural network framework for predicting non-coding DNA functions from the sequence alone and allows it to simultaneously learn motifs and a complex regulatory grammar between the motifs. Source code is available on GitHub. [11]

III. RESEARCH CHALLENGES IN COMPUTATION OF BIOINFORMATICS

Analysis of whole batch of genes at once in DNA sequence on a broad scale can be done by bioinformatics from lab techniques to computer programs. Following paragraphs will discuss the research challenges.

A. Biological data in computational biology:

Although myriads of data are available for genomics and proteomics analysis, still it is the most challenging to retrieve, manage and check the correctness of this data. Firstly, we discussed about some of the most challenging parts of data management in terms of computational biology and bioinformatics. The main challenge for data management is the data itself, here we observed some of the critical things to access and select the dataset for the above challenge which are listed below.

1. To identify which dataset needs to be considered or to be discarded.
2. It's difficult to get a user-friendly web interface to retrieve genomics data which can be used by the computer professional.
3. A dataset that can relate the genome sequence to other species.
4. Not all dataset has full access to download the data.

Another challenge faced by the biologist in the post-genome era is handling noisy and incomplete data, multiple comparison issues, small n and large p problem, processing compute-intensive task and integrating various data sources. Moreover, to develop new data mining methods for scalable and effective analysis which means to design new algorithms for software development is also challenging. Data mining can be defined as "search for hidden trends within large sets of data or extracting patterns from data". Data mining involved four classes of the task that are Classification, Clustering, Regression, Association rule learning. Genomics and proteomics analysis require data mining approaches at all levels. Proper techniques will create wealth of information from the biological specimens such as healthy and diseased tissues.

The data format for genomics datasets is also challenging in computational biology. The majority of machine learning and statistic methods for classification assume that data are in fixed-length vectors of real numbers, such methods are cannot be applied to genomics data [3]. Furthermore, the integration of molecular data with clinical medical information is also required, which helps to design a particular drug. [5]

B. Gene expression level analysis:

Gene expression level analysis means "formation of a gene product from its coding gene". In a biological activity it is very crucial part where change in gene expression pattern is reflected in a change of biological process. This analysis can solve the existing problems such as to solve the folding pathway of a protein given its amino acids sequence, deduce biochemical pathway given collection of RNA expression profiles, Protein structure prediction, Homology searches, Multiple alignment and phylogeny construction, Genomic sequence analysis and gene finding [5]. Gene expression differs according to developmental stage, tissue, age, and environmental conditions and because of that it's a big challenge to identify the pattern for some diseases. Some of the issues to measuring gene expression are 1) The amount of mRNA produced is not always directly proportional to known function processes such as translation into protein or regulation of another gene and 2) Analysis of gene expression patterns are dependent on the amount of mRNA detected for certain genes [12].

C.Selection of an appropriate architecture and hyper

D.parameters:

Proper selection of an architecture and hyper parameters also play a vital role for computational biology. Awareness of the capability of each architecture to obtain a robust result is also a challenging task for the researchers. Artificial Intelligence is very vast field which has numbers of architectures with number of techniques to solve the bioinformatics problems. Whereas benefits of these architectures are roughly understood [4], choosing among AI, ML or DL with proper architecture and techniques is the initial stage of practical applications in any respective field. In next section we have highlighted the various methods of AI in bioinformatics. Selection of hyper parameters for the analytics purpose is another critical step in which bias values, initial weight values, proper layers, learning iterations, etc. are considered. Value selection must be adequate of these hyper parameters. If value changed in hyper parameter, it results in significant performance deviation.

E. Proper study design and efficient & realistic interpretation of information:

Before researcher can create a model, the accurate study and interpretation of the information are required. There are several issues that can be considered for the study which include type of DNA microarray platform selected, mRNA preparation and data analysis [13]. Main crucial thing in microarray study is large m and small n problems where m is the number of variables and gene measurements and n is the number of observations from which those measurements are obtained.

Apart from the highlighted technical challenges, some literature also narrated crucial challenges which are not purely under this technicality. Such identified challenges are listed below:

1. how the integration of genomic data with existing and future experimental results?
2. whether bioinformaticians can effectively formulate new hypotheses before experimental work takes place?
3. How computational algorithms are retrieving informative features for prediction?

IV. AI TECHNIQUES IN BIOINFORMATICS

In the previous section we have highlighted some of the challenges and which can be focused in the bioinformatics domain research. In this section we are elaborating Artificial Intelligence techniques which can be used in bioinformatics domain.

As discussed in the above section that heterogeneous biological data is a big challenge, in context with that two main issues emerge as a common problem in bioinformatics domain which are large input and small sample size. To deal with these problems feature selection methods such as univariate and multivariate can be applied. Moreover, researcher Shahid Ashrafi Esfahani University, Esfahan, Iran has mentioned in his research article "Feature selection techniques in bioinformatics" that for bioinformatics community multivariate feature selection methods such as ensemble FS approaches is the most promising for the future perspectives [14]. Another author Khawla Tadist et al. discussed six types of feature selection method such as Filter, Wrapper, Embedded, hybrid, Ensemble, Itegrative methods with its previous research work and probable advantages [15]. Howbeit, Integration of multiple datatypes is one of the main concerns which required translating data from one format to another format such as XML. Changing data format will improve the performance of data mining and computational methods. Author Marco Mesiti et al. discussed some of the methods to translate the biological data in XML format in the article "XML-based approaches for the integration of heterogeneous bio-molecular data" which shows the issues and new perspective to integrate datatypes in XML format [16].

Artificial Intelligence subset such as Machine Learning and Deep Learning can be useful for the research in bioinformatics. Machine learning facing a problem in bioinformatics for small sample size, individual variability, high dimensionality and complex relations between data where as Deep learning will work fine with the large number of data. Here we have identified some of the popular algorithms which can be used initially for the genomics, proteomics and system biology such as Support vector machine [SVM], random forest, hidden Markov models, Bayesian networks, Gaussian networks. Whereas, Deep learning is most popular AI approach for image processing, speech and natural language processing in bioinformatics. Mainly, Deep Neural Networks and Recurrent Neural Networks are used for predicting protein structure, protein classification and analysis of gene expression regulation, while convolutional neural networks are used for image analysis and to analyze gene regulation.

Below table demonstrate few possible methods which is discuss in the paper and can be used in the bioinformatics domain:

V. RESULT AND DISCUSSION

Above sections emphasized the trials and skills of AI based research in the field of bioinformatics specifically to genomics and proteomics. Below table demonstrate the methods which can be used to resolve the challenges of bioinformatics research using AI techniques and will give the prominent outcome with accuracy for the given field.

Table- I: Methods used for genomics & proteomics domain

Methods		References
Feature Selection Methods	Multivariant Feature Selection Methods such as ensemble FS approaches.	[14]
Feature Extraction Methods	BioVec, ProtVec, GeneVec	[8]
Deep Learning Methods	Convolutional Neural Networks (CNN) Deep Neural Networks (DNN), Recurrent Neural Networks (RNN)	[4,9]
Machine Learning Methods	Bayesian Networks. Gaussian Networks. Support Vector Machine [SVM]. Random Forest. Hidden Markov Models.	[1,3]

Feature selection methods are used to eliminate duplicate data from the dataset. Feature extraction methods are used to create new feature that have most significant information. Machine Learning and Deep Learning algorithms which can give the prediction result based on the implementation on biological sequences such as DNA, RNA, Protein, Amino Acid.

VI. RESEARCH POTENTIAL

After the extensive study of the technology, various research domain can be focused to enhance the current way they are working. As far as, Bioinformatics is concerned it has huge makeover with IT industry especially in Proteomics, Genomics, and Drug Designing and Disease prediction from protein structure.

Human gene is associated with specific organs and the computational analysis of defect in these gene with disease will help to identify drug targets. Furthermore, various AI techniques for protein structure & function analysis as well as protein structure & function modification can be applied which can be used to predict certain diseases. Here, we have highlighted some of the future aspects which can be focused further in research such as drug designing by fragments, Protein Translation Modification (PTM) analysis and prediction, Generation of database with Amino Acid sequence and its related disease.

VII. CONCLUSION

Bioinformatics has become a very essential interdisciplinary field which helping to 'Omics' fields such as proteomics, genomics. Overall, we have discussed about the importance, challenges and future perspective of bioinformatics in the field of computer science and its applications. Concluding this paper bioinformatics is the field where biology and computer science works together to collect, retrieve, organize and analyze the heterogeneous biological data. In this survey we have given general fundamental of Machine Learning and Deep learning that can be used in the bioinformatics domain. we have discussed technical as well as non-technical challenges in bioinformatics research. Also, we have highlighted possible solutions and techniques which can be implement to overcome given challenges. Moreover, impact of Artificial Intelligence subset such as machine learning and deep learning in bioinformatics with its various techniques and specific usage were explored in the paper. Apart from the challenges and solutions we also draw a special attention to the future perspective of research such as Drug designing, Protein structure and function prediction which can be further focused in the field of bioinformatics.

REFERENCES

1. Pedro Larranaga, Borja Calvo, Roberto Santana, Concha Bielza, Josu Galdiano, Inaki Inza, Jose A. Lozano, Ruben Armananzas, Guzman Santafe, Aritz Perezand Victor Robles, "Machine learning in bioinformatics", Briefings In Bioinformatics, Volume 7. Issue 1, 2005, (pp. 86-112)
2. Ma C, Zhang HH, Wang X," Machine learning for Big Data analytics in plants", Trends Plant Science, 19(12):798-808, 2014 December, Epub 2014 Sep 12. Review. PubMed PMID: 25223304.
3. Libbrecht MW, Noble WS,"Machine learning applications in genetics and genomics", Nature Reviews Genetics, 2015,16(6):321-32.
4. Seonwoo Min, Byunghan Lee, and Sungroh Yoon , "Deep Learning in Bioinformatics"
5. Manisha Mathur, "Bioinformatics challenges: A review", International Journal of Advanced Scientific Research, Volume 3, Issue 6, November 2018. (pp.29-33)
6. Alejandra J. Magana, Manaz Taleyarkhan, Daniela Rivera Alvarado, Michael Kane, John Springer, and Kari Clase, "A Survey of Scholarly Literature Describing the Field of Bioinformatics Education and Bioinformatics Educational Research", CBE—Life Sciences Education, Volume. 13, Winter 2014.(pp. 607–623)
7. Ning Luwen, Geng Liu,Guibo Li, Yin Tong, and Jiankui He, "Current Challenges in the Bioinformatics of Single Cell Genomics" ,Frontiers in Oncology – Review Article, 27th January,2014.
8. Ehsaneddin Asgari ,Mohammad R.K.Mofrad, "Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics", PLOSONE, 10th November,2015.
9. Ehsaneddin Asgari, Nina Poerner, Alice C. McHardy, and Mohammad R.K. Mofrad , "DeepPrime2Sec: Deep Learning for Protein Secondary Structure Prediction from the Primary Sequences", bioRxiv preprint first posted online, 18th July 2019.
10. Felipe Kenji Nakano, Mathias Lietaert and Celine Vens, "Machine learning for discovering missing or wrong protein function annotations", BMC Informatics, 2019. (pp. 1-32)
11. D. Quang and X. Xie, "DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences", Nucleic Acids Res., volume 44, Issue 11, June 2016.
12. Winston Patrick Kuo, Eun-Young Kim, Jeff Trimarchi, Tor-Kristian Jenssen, Staal A. Vinterbo, Lucila Ohno-Machado, "A primer on gene expression and microarrays for machine learning researchers", Journal of Biomedical Informatics, Science Direct, 2004. (pp. 293 -303)
13. Jacques Nicolas , "Artificial Intelligence and Bioinformatics", 2018. hal-01850570.
14. Mohamad Reza Hosseini , Naser Nematbakhsh , Motahareh Nadimi , "Feature selection techniques in bioinformatics", National Conference on Modern Research in Electrical, Computer and Medical Engineering, 7th August,2017.
15. Tadist, K., Najah, S., Nikolov, N.S. et al. "Feature selection methods and genomic big data: a systematic review", J Big Data 6, 79 (2019).
16. Marco Mesiti, Ernesto Jiménez-Ruiz, Ismael Sanz, Rafae Berlanga-Llavori, Paolo Perlasca , Giorgio Valentini1 and David Manset, "XML-based approaches for the integration of heterogeneous bio-molecular data", BMC Bioinformatics 2009.

AUTHORS PROFILE

Mubina Malik working as an Assistant Professor in Smt. Chandaben Mohanbhai Patel Institute of Computer Applications Faculty of Computer Science and Applications, Charotar University of Science and Technology, Changa. She pursed Master of Computer Applications from Sardar Patel University, Gujarat, India and Bachelor Degree in Environment Science from Sardar Patel University. She has published 3 research papers in reputed international journals and it's also available online. Her total citation is 38. Her main research work focuses on Public Key Infrastructure, Machine Learning, Deep Learning, Bioinformatics She has 10 years of teaching experience.

Dr. Jaimin N Undavia, working as an Assistant Professor in Smt. Chandaben Mohanbhai Patel Institute of Computer Applications Faculty of Computer Science and Applications, Charotar University of Science and Technology, Changa. He got his doctorate from CHARUSAT University. He has published 19 international paper, 1 national, 1 international book chapter and 1 international book. He possesses 16 years of extensive teaching experience. His research area is Big Data Analytics, Robotics, IoT and Machine learning. He is serving 5 international journal as a reviewer and looking for more advancements in the field of robotics.