CGC International Journal of Contemporary Technology and Research
ISSN: 2582-0486 (online) Vol.-2, Issue-2 DOI: 10.46860/cgcijctr.2020.06.26.96

Survey on usage of Machine Learning in Genomic Medicine

Shanky Goyal[1*], Harsh Sharma, Navleen Kaur

[1]Research Scholar, Department of Computer Science, IKGPTU, Jalandhar
Department of Information Technology
Chandigarh Engineering College, Landran
*shanky.it@cgc.edu.in

***Abstract*: This paper provides a preface of machine learning as a task that can be used to solve some important problems pertaining to genomic medicines. The genomic medicine can determine the risk of different disease in an individual due to the variation in the DNA. Genomic medicine can help to find out the therapies. We, here, concentrate on the ways in which machine learning can aid in determining the link between the DNA as well as number of key molecules present inside a cell with the assumption that the numbers may be related to the disease risks, which can also be referred to as cell variable. The field of the Modern biology allows high rate of production measurement of cell variable which covers up gene expression, splicing, and the procedure of protein binding with nucleic [8] acids, and these all treated as training targets for the predictive models. In today's date, large amount of data sets are available on which we can apply different computational techniques that can help researchers to work hard on solution of genomic medicines**

Keywords: Machine Learning, Medicine, DNA, molecules.

I. INTRODUCTION

Machine learning is wont to solve numerous issues concerning genomic medication. As per UN agency, genetic science is outlined because the study of genes and their functions, and connected techniques. Genetic science is associate knowledge domain field of biology specializing in the structure, function, evolution, mapping, and written material of the genomes. An ordination is associated with the organism's complete collection of DNAs, together with all of its genes. Whereas, Genomic drugs is associated to the development in medical field that involves victimization genomic information concerned with a personal as a part of the clinical care appreciate for diagnostic or therapeutic decision-making and also the health outcomes and policy implications of that clinical use. Also, the genomic drugs are already creating a bearing within the fields of medicine, pharmacological medicine, rare and unknown diseases, and communicable disease. Machine learning is a major chance for the advance of endurance and approach[1] of lifetime of many individuals full of any quite genomic disorder, each at this time and also the future years.

A genome is an organism's complete set of DNAs, which includes all of its genes. The "genome" of any given [2] individual is unique. Each genome consists of all the information required to build that organism and assist it to grow as well as develop. In humans, a copy of the entire genome—more than 3 billion DNA [3] base pairs—is contained in all cells that contain a nucleus. Since 1953, DNAs [4] were thought to be physiological media for transfer of genetic information. However, by 2001, the Human Genome Project demonstrated the model content of a specific human genome. The Human Genome Project was an international scientific research projects with the aim of finding the base pairs that build up the human DNA, and of identifying and tracing all the genes of the human genome from a physical as well as a functional standpoint. It is the world's largest collaborative biological [5] project till date. Mapping of the "human genome" involves arranging a small number of individuals in order to perform gathering these together to get a proper sequence for each of the chromosome. Hence, the final human genome is a mosaic, though not representing any particular individual [6]. The human genome has 20,000 coding genes and 25,000 non-coding genes. While some of them are pivotal for the purpose of life, others for health while few of them are negligibly important and hence can be deleted in their ensemble without harming apparently.

There is some important structural region within a gene which are nucleotide sequences called introns and exons. Introns are considered to be intervening sequences, and exons are expressed sequences. The average of exons is 8.8 and that of the introns is 7.8 per human gene. There are designs in nucleotide sequence where the disease-responsible mutations act through disturbing the designs. The Spinal muscular atrophy (SMA) is a collection of neuromuscular disorders which result in the loss of motor neurons and progressive muscle wasting. There is leading genetic issue in infant mortality called as Spinal muscular atrophy (SMA) that results in deficient creation of Survival motor neuron, if there is missing or damaged SMN1 gene in baby genome. SMN2 gene can manage the production of SMN protein. Various Researches are going on to find out the therapies that can restore function of exons 7 in SMN2. Whereas, for the case of various genetic diseases, process of diagnosis is more complex. For example, Cancer is a heterogeneous disease with multiple seminal pathways, all of which can lead to common symptoms but require various remedy.The diseases such as cancer, genomic data are becoming vital day by day as it can provide the researchers with more detailed information of the disease for the simplified diagnosis and effective treatments at the same time. Doctors have been using blood sample to restrict the transfusion of blood.

The speedily growth in the data for the genomic has helped in the way that it can be collected quickly and cheaply from the patients and from wide community. Here, the focus is on the application of the field of machine learning [7-9] in genome medicine where we can estimate the genomic traits to get the required therapies

and analyze the further disease risks for future preventive measures.

To make the theory of genomic medicine work in real world, it is required to develop a computer system that can precisely depict the texture of the actual genome. The impact of genetic variation and the probable remedy will be depicted and achieved easily, cheaply and accurately with the use of laboratory experiments and model organisms.

At present, protein coding exons are the vital region of study in genome.The most important feature in genome diagnostic channel is the change in amino-acid sequence by the coding mutation. If mutation initiates, a stop codon or non-sense codon into the sequence, then the protein will be curtailed. Withal, the forecast about the disturbance in the structure or the stability of the resulting protein molecule by a mutation is an open dilemma.

Moreover, the coding area makes up only 1.5% of the total human genome, even though there is evidence that at least 5.5% of the positions undergo the process of purifying selection. Disease-causing mutations are rapidly found in the outer area of the protein- coding regions, indicating the fact that only the coding regions are not sufficient. Most of the active noncoding positions are the regulatory sequences i.e. they order to the cells to control essential procedures such as gene expression and decisive exons identification. Thus, to identify and analyze genome regular instructions, there is need of automatic computation model development. The regulatory sequences play a significant role in the complexities of cell biology and it is unable to be alleged by the sheer number of genes or the coding regions.

Now, the question is how we can read the genome. We can detect objects that we can see or also can recognize the voice very easily but it is not easy to read the genome. As a gateway to it, a system having an extraordinary analytical ability is necessary. So, a combination of people expertise in machine learning and genomic biology has made research group and developed some techniques to gather information of genome. Hence, this is providing the machine learning researchers a better field to contribute. In genomic biology, it will be good that researcher utilize their time on developing the techniques not on just data collection.

To support the genomic medicine, computer systems pursuing the tendency to read the texture or design of the genome are crucial. For example, a discovery in the area of ‘‘gene editing’’ is allowing scientists to alter the genomes of existing living cells, with a competence. Gene therapies now include targeted modifications including insertion removal of sequences from the predetermined locations in a genome as per the requirement. Genome editing technology has provided a pathway to remarkable opportunities in the field of genomic medicine making it more essential, we can find effects of these edits in silico (In silico biology refers to computational models of biology). In other words, there’s a lot of difference between the knowledge of how to write and what all is to be written.

II. USE OF MACHINE LEARNING TO ENACT GENOME

Prediction of phenotypes (e.g., external traits and unwellness dangers) from biomarkers similar to order could be supervised using machine learning is a drawback. The inputs to them are strand of DNA sequence (genotype) pertinent to the fundamental biology, and the outputs are the phenotypes. This option isn't ideal as most of the phenotypes and diseases are advanced. This is often because of 2 reasons- 1st is that the sheer quality of relationship between an entire genotype and its constitution. The order shows the way to the state of cell using the varied stages of advanced as well as interlinked bio physiological processes and regulates the mechanisms that are antecedent formed “ad hoc” by evolution. Second, although one might think that such models, it's seemingly that hidden variables of those models wouldn't relate to biological mechanisms that may be worked upon. Focus onto unwellness mechanisms is vital for the aim of developing targeted therapies, however may also give complementary data for makeup screens, that historically identifies chemicals with desired biological results while not information of the precise targets. The most powerful way is the computational model to be trained in two ways- firstly finds out variable of the intermediate cell, after that link all these to phenotype.

The above-mentioned approach has defined two previously defined problems. From genomic sequence, these cell variables are very closely linked and can also detect very easily as compared to phenotypes, learning models that map from DNA to cell variables can be more forthright. So, techniques can use to generate many data sets from these variables under different kin d of condition and then using this data set, we can train the model with good accuracy. From model, we can found out that which cell variables are linked with high risk diseases as compared to healthy person and then doctors can use an usable therapy for restoring the cell variables to its normal condition.

Machine learning [10, 11] plays a central role by turning high-throughput measurements into specialized or general-purpose predictive models for what we referred to as ‘‘cell variables’’ v quantities that are relevant to cell function. Byknowing how mutations affect disease via cell variables, diagnosticians and pharmaco genetics can more easily find direct correlates with disease, develop treatments, and plan targeted therapies for individual patients. The extract of this paper provides that how machine learning based computational models plays a big role to analyse the genetic impetus of disease.

III. CYTOLOGY, MACHINE LEARNING AND GENOMIC MEDICINE

In this segment, we are to discuss the system through which it can be participated toward the objective of genomic medicine. An appraisal to measure the corresponding biological quantity must exist and training data must be collected under different conditions to develop a computational model of a definite cell variable. The biological appraisal into the 1990s needed manual

steps to generate small quantity of the data. These kinds of techniques are quite effective for forming and testing theses, although do not provide data enough to ascertain accurate predictive models of the complicate outcomes. It has now been customary to achieve large numbers of cell variable measurements in a minimal cost experiment with the help of commoditization of high throughput appraisal technologies.

Different technologies are in use for identifying the protein binding site, some are used for making a sequence of different organism genomes. Some more are used for categorized the genomes for medical study to find out the variations.

To measures different genotype, different techniques are in use to calculate cell variables.to help computation model, we provide different kind of inputs like from stretch of DNA, wecan provide availability of some motifs or frequency of nucleotides etc. Through biochemical process, that impact cell variables, we can find out more features like binding of protein between DNA and RNA, nucleosome positioning and occupancy profiles, and RNA secondary structures.

For training model[12], cell variable that are used and the cell on which currently we need results vary for all DNA. So, for this purpose, we train module with features of DNA cells

In case, a model is good at conjecture, mutated DNA sequence can be analysed easily leading to variation in cell variables and it can help in indicating the current state of disease, without requiring any experimental complex measurements from diseased cells. If a model accurately displays the instructions in the genome processed, then it should also have the ability to determine the disease caused by mutation those results in change in the cell variables. If recording of mutation is done based on the rate in which they cause the change in cell variable, then faulty positive result may arise during the large changes caused by mutation on cell variable resulting in no effect on the disease, for example, a mutation causing change in hair colour due to the corresponding change in cell variable.

A. Gene Process:

There is a process called as gene expression in which gene are copied to prepare messenger RNA, then this mRNA is converted in the form of protein. Precursor mRNA is the process to transcribe the DNA sequence having exons and introns into RNA. It helps the development of mature RNA within the nucleus

For splicing out long stretch of sequence known as introns, RNA processing take place which modifies the pre mRNA and then connect with exons. Still at this stage we call RNA molecules as mRNA. After reading three letter wordcodesin the mRNA sequence, translator generate a protein molecule.

For translation purpose, mRNA goes to suitable location, called as mRNA localization. Similarly, proteins also find it specific location in the cell called as protein localization.

There are few steps in RNA processing. In first step, called as splicing, from precursor mRNA, separate the introns and join with exons. Then specific series of adenine base is attached with mRNA at the end. This process is very important and happens after the processing or during the processing of transcription for better interaction between them. After all these steps, from the nucleus, mRNA is transferred to ribosome, from there it reaches to the protein.

B. Splicing model for Computation:

In human body, there are 20000 genes, micro RNAs called as functional molecules, tissues, organs. By implementing different functions on single genes through different techniques will provide information that can help to achieve the complexity.

Based on the mobility framework such as type of tissuehaving the cell and genomics information, some exons may be left out. To follow this process is called as splicing regulation. This function is based on complex combination of genomic elements that exist in DNA and pre-mRNA

Frequency of 10000[13] exons of specific cell type can be used to train the computational model for discovering the instructions that can handle splicing and used to analyse splicing. Research said that various human problems have been affected by mutations in the instructions which regulate splicing.

There is much training data available for sequence interaction between protein cells. These interactions help to make correct model because this interaction affects main functions of the cell. Binding process between DNA and RNA proteins is very selective and careful for the mutations at location of binding. Many diseases occur just due to mutations because it changes binding sites. There is encoding done by human genome of at least~1400 DNA binding proteins and also approx. 1500 RBPs, to make large volume of proteins. During binding, DNA-binding protein called as transcription factor, impact the fluency of copy of special genes to RNA.

Splicing gets input from the computational model of RNA binding proteins. High throughput experiments done based on two processes that are sequencing and microarray-based method. Publically, vast amounts of dataset of both approaches are available.

For finding the reasons that a specific protein bind to, researcher collect specific data from cell fragment using sequencing protein bound or using microarrays by exposing tagged proteins to synthetic fragments[2]

1) Basic Computational Models:

There is assumption that protein should be bind in specific pattern called as consequence sequence. This binding site modelling works on the position- frequency matrix (PFM). The binding site of protein is coordinated by PFM using parameters (i,j) where base i occurs at position j.

2) Beyond PFM-Based Model:

There is problem with PFMs is that they consider that each position participates individually to binding strength. This consideration can work for some of the proteins but not for all. So biologists are working to develop such computational model that can work sequentially on more challenging proteins.

IV. COMPUTATIONAL BIOLOGY USING MACHINE LEARNING

Speech reputation and computer vision has been the most focused high-profile region for the device getting to know researchers recently. Computer vision has a very visceral nature and due to this, human beings are excellent at the tasks mentioned it. Still our learning algorithm is not giving desired results as per our expectations. MNIST is used considerably as a check mattress for newly develop learning algorithms. The main aim of this paper is to demonstrate the concept of issues referring to computational biology in a way this is on hand to system mastering researchers. The scenario in biology is technically distinct from the situation in laptop vision. The visual international is immediately usable to us. Following table 1 explains the implemented approaches for making version the genetic foundation illnesses risks

Table 1: Genetic basis Diseases Risks

Sr.No.	Approaches	Description
1	Genome-Wide Association Studies [2]	➢ find the method through which development inside the population is associated with the variation in region ➢ approach to use computational model to expect diseases risks by giving input as SNP of specific patients. ➢ Then use machine learning algorithm
2	Evolutionary Conservation[2]	➢ Principals of sequence conservation: o Two force-slow accumulation of mutations and to select pressure for those mutations which come due to damage of reproductive fitness of a population o From a common ancestor, the genomes of many species generate.

Table 2[2]: Description of various cell Variables

Cell variable	Brief description	Relevance to disease
Identification of structural and functional regions of the genome	Attaching meaning to, or annotating, different regions of the DNA, such as marking the boundary of introns and exons, identifying parts that have regulatory functions.	Changes in genomic sequences can cause a region which previously served a particular function to become non-functional and vice-versa, or changing its intended function affecting regulation
Binding sites for transcription regulation	Binding of proteins to specific sequence elements of the DNA controls whether transcription can occur, as well as the rate at which it Happens.	Sequence variations to sequence pattern that protein such as transcription factors and complexes that unwind the DNA.
Splicing patterns	Splicing modifies the pre-mRNA by removing introns and selecting the exons to be retained.	Changes to the regulatory elements that control splicing can change the characteristics of the gene products, and in some cases, causes them to be non-functional.
Cleavage site selection and Polyadenylation	The ends of transcripts are cleaved and a stretch of adenine bases are attached before they are ready for Translation	Modification to sequence elements can alter where cleavage occurs. It determines the presence or absence of binding sites for regulatory proteins on the transcript.
RNA structure	The RNA folds into 3D structures influencing its interaction with other molecules in the cell.	The mRNA, beyond the information it contains for encoding protein, has 3D structure. This structure can affect processes that it is involved with, such as transcription, splicing, and translation.
Protein structure	The outcome of translation in a sequence of amino acids that folds into a protein. The protein's 3D structure is important for its function, as it interacts with DNA, RNA and other proteins.	Structure affects function. The ability to predict protein structure from sequences can help in understanding the biological function of a gene, and how mis-folding of proteins contribute to disease.

V. DISCUSSIONS AND FUTURE SCOPES

a. Feedback to Research group:

Main trend in machine learning for researcher is to work on ideas for applications such as voice recognition, computer vision and natural language processing. But focus of researcher can be more divert on computational biology, if and only if they will be able to access the problems that medical field people facing and also for working on problem, data will be easily accessible.

Many enterprises has already started machine learning based projects in field of biology and medicine. They are regularly testing different method to assess the performances of it.

There are some communities who test computational method on predicting impact of variations in genomic and then guide to researchers for future work

b. Dependency on Evolutionary Conservation:

The various features of the tools basically depend upon the conservation features. The only drawback of relying on the conservation features is that the resulting models perform very well even if they rely on the conservation patterns and do not discover any meaningful observation in the genome itself.The computational model is more accurate when given the

conservation features. The most basic problem of the conservation is that the sequences are functional and not all the functional sequences are conserved.We can expect the lack of functional conservation more associated to the disease but this time there is no way to measure the gene products. Further the conservation can only tell us what the particular gene has survived over the years. All in all, conservation is predictive source of knowledge that can never tell you the complete story.

c. Recurrent Neural Networks (RNNs):

The various researchers have observed remarkable change in the results after using the deep recurrent models instead of the Markov models. Similar gains might occur in the sequential prediction applications in the computational biology. We know that the DNA binding proteins are known to arrive by the dynamic process whereas the proteins migrate along the DNA backbone. RNN's has one more useful application that is the imputation of epi-genomic tracks.

d. Interpretability:

This concept in itself is not well defined. Despite the upcoming transformations in machine learning in the early 1990's there has still not been any confirmed definition for the interpretability. Correctness of results totally depends on the framework that is used for communication with the storm cells. There are various machine learning related techniques are available which provide deep insight of the problem that is not in hand of non-expertise. Many examples of machine learning are available where interpretability show useful biological view eg. C-path.

e. Adversarial Data for Genomics:

The various recent researches have shown that when there are adversarial inputs, neural network not able to predict accurately. These inputs are basically designed while keeping in mind the trained or half trained pre-existing neural networks. The main point of concentration is that these adversarial examples do not occur naturally and the values or the inputs are very different at the time of the naturally occurring inputs.

However, one of our major objectives is to use the computational biology and design such model which can analyze the effects of therapies. The resultant genome will obviously be unnatural and hence the question of adversarial inputs arises. In order to address this problem, the adversarial inputs shall be compared to the real-life experiments results, and then validate and do improvement in computational model.

REFERENCES

[1]. M.S., Fernando et al., "Synergy between medical informatics and bioinformatics: facilitating genomic medicine for future health care," Journal of biomedical informatics 37.1,2004, pp.30-42.
[2]. L. Michael KK et al., "Machine learning in genomic medicine: a review of computational problems and data sets," Proceedings of the IEEE 104.1, 2015, pp.176-197.
[3]. B. Linus, and O. Kohlbacher, "Immunoinformatics and epitope prediction in the age of genomic medicine," Genome medicine 7.1 (2015): 119.
[4]. S. Michael C., and B. Langmead, "The DNA data deluge," IEEE Spectrum 50.7 ,2013, pp.28-33.
[5]. M. Vivien, "Biology: The big challenges of big data," ,2013, 255.
[6]. A. Euan A, "The precision medicine initiative: a new national effort," Jama 313.21 ,2015, 2119-2120.
[7]. W. Kiri, "Machine learning that matters," arXiv preprint arXiv:1206.4656, 2012.
[8]. W. James D. and F. HC Crick, "Molecular structure of nucleic acids," Nature 171.4356,1953, pp.737-738.
[9]. S.G. Jenni AM and C. J. Sidey-Gibbons, "Machine learning in medicine: a practical introduction," BMC medical research methodology 19.1, 2019, 64.
[10].C. Federico, R. Rasoini, and G.F.Gensini. "Unintended consequences of machine learning in medicine," Jama 318.6, 2017, pp.517-518
[11].O. Ziad and E. J. Emanuel, "Predicting the future—big data, machine learning, and clinical medicine," The New England journal of medicine 375.13, 2016, 1216.
[12].C. Jonathan H. and S. M. Asch, "Machine learning and prediction in medicine—beyond the peak of inflated expectations," The New England journal of medicine 376.26, 2017, 2507.
[13].K. N. Lam, H. Van Bakel, A. G. Cote, A. Van Der Ven, and T. R. Hughes, ''Sequence specificity is obtained from the majority of modular C2H2 zinc-finger arrays,'' Nucleic Acids Res., vol. 39, no. 11, 2011, pp. 4680–4690.