Next Generation Sequencing Based Cancer Classification Using Machine Learning

[1]P Jyothirmai, [2]Devar Haribharati S, [3]Diksha Naik, [4]Pooja Naik, [5]Dr. Supriya Patil

[1]Student, [2]Student, [3]Student, [4]Student, [5]Associate Professor
[1]Electronics and Telecommunication,
[1]Padre Conceicao College of Engineering, Verna, Goa, India

Abstract : Next generation sequencing (NGS) is an efficient method used for Deoxyribonucleic acid (DNA) sequencing. Although, with recent advancement in NGS technology, the majority of variants classified using NGS are accurate and reliable but however, a small subset of variants still do require orthogonal confirmation. For this reason, many clinical laboratories confirm NGS results using orthogonal technologies such as Sanger sequencing. Here, we use machine-learning-based model to differentiate between these two types of variants: those that do not require confirmation using an orthogonal technology (high confidence variants), and those that require additional quality testing (low confidence variants). This approach allows identification of few important variants that require orthogonal confirmation.

Keywords **- Next Generation Sequencing (NGS), Deoxyribonucleic acid (DNA), High confidence variants, Low confidence variants**

I. INTRODUCTION

NGS is an efficient method for determining the order of chemical bases in a DNA strand. Human body consists of four chemical bases namely, adenine, thymine, cytosine and guanine. The rules of base pairing (or nucleotide pairing) are: adenine is always paired with the thymine and cytosine is always paired with the guanine. The change in the structure of a DNA strand that is caused by alteration (mispairing), deletion, or insertion of single base units which leads to variation known as mutation that causes cancer. Here, we develop a deterministic machine-learning based model to differentiate between two types of cancer variants: high confidence variants - those that do not require confirmation using additional testing such as Sanger sequencing and low confidence variants - those that require additional testing for confirmation.

.

II. METHODOLOGY

2.1 Block Diagram

DNA is extracted using the Oragene DX 510 saliva collection device. Extracted DNA is compared with the reference DNA (healthy person's DNA) and parameters are determined. These parameters are fed to the neural model as shown in Fig.1.

2.1.1 DNA Extraction from Saliva

Figure 2 shows how DNA is extracted from saliva. Saliva is mixed with cell lysis solution, which breaks down the cell into its components like mitochondria, ribosomes, chromatin etc. After which RNA and other protein impurities are removed and DNA is separated by precipitation.

Saliva when compared to blood collection has the following advantages: it requires no specialized personnel for collection, allows for remote collection by the patient, is painless, well accepted by participants, has decreased risks of disease transmission, does not clot, can be frozen before DNA extraction and possibly has a longer storage time.

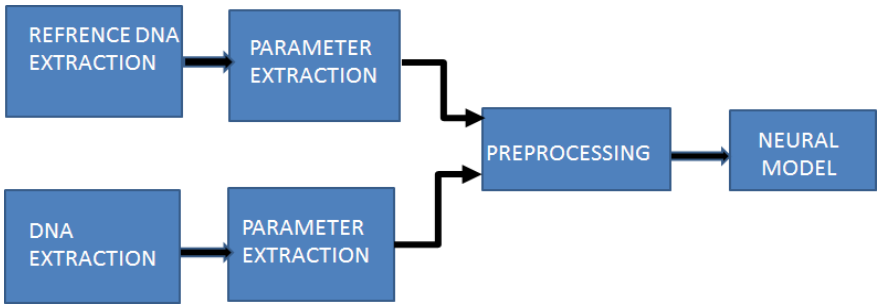

Figure 1 Block Diagram

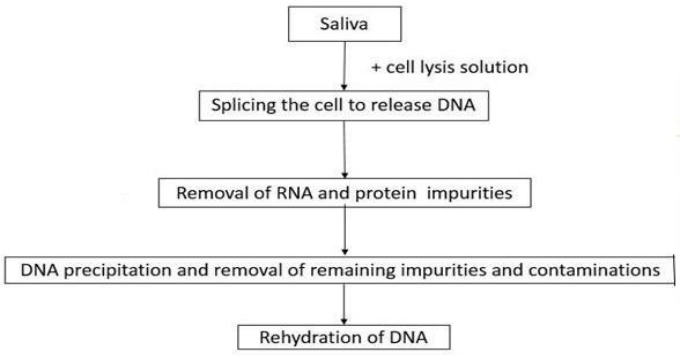

Figure 2 Extraction of DNA from Saliva

2.1.2 Parameters

Following are the parameters determined from the extracted DNA:

2.1.2.1 Depth Position (DP):

NGS read depth at the variant position i.e, number of chemical bases present in a DNA strand.

2.1.2.2 Allele Depth (AD):

Number of reads that support the variant call i.e, number of bases that are affected.

2.1.2.3 Allele Frequency (AF):

Fraction of reads that support the variant call, i.e. AD / DP – fraction of chemical bases affected.

2.1.2.4 Guanine Cytocine (GC):

Fraction of GC content in the bases around the variant position.

2.1.2.5 Mapping Quality (MQ):

Root Mean Square of the mapping quality of the call. It indicates how well a particular strand of DNA is placed with respect to the reference DNA.

2.1.2.6 Genotype Quality (GQ):

Genotype Quality of the call determined by GATK Haplotype Caller. It determines the confidence with which the genotype assigned is correct.

2.1.2.7 Weighted Homo-polymer Rate (WHR):

Weighted Homo-polymer Rate in a window of 20 bases around the variant position, i.e, the sum of squares of the homo-polymer lengths, divided by the number of homo- polymers.

2.1.2.8 Homopolymer Distance (HPL-D):

Distance to the longest homo-polymer within 20 bases from the variant position.

2.1.2.9 Homopolymer Length (HPL-L):

Length of the longest homo-polymer within 20 bases from the variant position.

2.1.2.10 Quality Score (QUAL):

Quality score assigned by the GATK Haplotype Caller to the call.

2.1.2.11 Normalized Quality Score (QD):

QUAL, normalized by DP.

2.1.2.12 FisherStrand (FS):

Phred-scaled p-value using Fisher's exact test, to detect strand bias.

2.2 Implementation using MATLAB

2.2.1 Dataset

Here, we are using 7179 samples, out of which 6666 samples are cancerous and 513 samples are non-cancerous according to Sanger sequencing. Each row of the training dataset represents each parameter extracted and each column represents each of the samples. Each row of the target dataset represents each of the output neuron and each column represents each of the samples.

2.2.2 Training Algorithm - Gradient Descent Algorithm

Gradient descent is an optimization algorithm used to minimize some function by iteratively moving in the direction of steepest descent as defined by the negative of the gradient. In machine learning, we use gradient descent to update the parameters of our model. Parameters refer to coefficients in Linear Regression and weights in neural networks.

2.2.2.1 Gradient Descent Algorithm

This is a gradient descent local search procedure. It measures the output error, calculates the gradient of the error by adjusting the weights in the descending gradient direction.

2.2.2.2 Gradient Descent with Momentum

This algorithm allows a network to respond to the local gradient as well as recent trends in the error surface. It acts like a low-pass filter that means with momentum the network ignores small features in the error surface. A network can get stuck in to a shallow local minimum but with momentum it slides through such local minimum.

2.2.2.3 Gradient Descent with Adaptive Learning Rate

The performance of the gradient descent algorithm is very sensitive to the proper setting of the learning rate. If the learning rate is set too high, the algorithm can oscillate and become unstable. If the learning rate is too small, the algorithm takes too long to converge. It is not practical to determine the optimal setting for the learning rate before training, and, in fact, the optimal learning rate changes during the training process, as the algorithm moves across the performance surface.
The performance of the gradient descent algorithm can be improved if the learning rate is allowed to change during the training process. An adaptive learning rate attempts to keep the learning step size as large as possible while keeping learning stable. The learning rate is made responsive to the complexity of the local error surface.

2.2.2.4 Gradient Descent with Momentum and Adaptive Learning Rate Back-propagation

This algorithm combines adaptive learning rate with momentum training. It is invoked in the same way as gradient descent with adaptive learning rate, except that it has the momentum coefficient as an additional training parameter.

III. Results And Discussions

With gradient descent algorithm, out of all 7179 variants analyzed, the model predicted 7019 variants to be of high confidence and 160 variants to be of low confidence.
With gradient descent with adaptive learning rate algorithm, the model predicted 7120 variants to be of high confidence and 59 variants to be of low confidence.
Using gradient descent with momentum, 7019 variants are predicted as high confidence variants and 160 variants as low confidence variants.
And gradient descent with momentum and adaptive learning rate model predicted 7121 samples to be of high confidence and 58 to be of low confidence.

Table 1 Comparison between different gradient descent algorithms

Hidden nodes under 50		
Algorithm	**No. of hidden nodes**	**Accuracy**
Gradient Descent algorithm	44	97.7713
Gradient Descent with Momentum	44	97.7713
Gradient Descent with Adaptive learning rate	35	99.1782
Gradient Descent algorithm with Momentum and Adaptive learning rate	29	99.1921

Gradient descent with momentum and adaptive learning rate algorithm gave highest accuracy with least number of hidden nodes. Therefore, it can be concluded that gradient descent back-propagation algorithm with momentum and adaptive learning rate gives performance compared to other gradient descent algorithm.

IV. Conclusion

Although, the cost of NGS has dropped dramatically over the past decade, including orthogonal confirmation for all the variants may be adding unnecessary cost to the genetic testing. Our model helps determining only those variants which require orthogonal testing, thereby, reducing the cost.

References

[1] https://www.ncbi.nlm.nih.gov/pmc/articles/pmid/29665779/
[2] https://ml-cheatsheet.readthedocs.io/en/latest/gradient_descent.html.
[3] https://isogg.org/wiki/Next_generation_sequencing
[4] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3841808/
[5] https://www.ebi.ac.uk/training/online/course/ebi-next-generation- sequencing-practical-course/what-you-will-learn/what-next-generation-dna-.
[6] http://www.holehouse.org/mlclass/04_Linear_Regression_with_mult iple_variables.html
[7] https://in.mathworks.com/help/deeplearning/ref/
[8] Variant dataset is taken from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5904977/bin/12864_2018_4659_MOESM2_ESM.xlsx