wjert, 2020, Vol. 6, Issue 3, 479-487. *Review Article* ISSN 2454-695X

World Journal of Engineering Research and Technology

WJERT

www.wjert.org SJIF Impact Factor: 5.924

AN IMPLEMENTATION OF PROTOTYPE TO ANALYZE CLINICAL EXOME SEQUENCE FOR CANCER PREDICTION USING PYTHON

Dr. Krupa Mehta*[1], Dr. Devarshi Mehta[2], Dr. Rakesh Rawal[3], Dr. Maulik Patel[4] and Dr. Vishal Dahiya[5]

[1]GLSFCT, GLS University, Ahmedabad, Gujarat, India.

[2]GLSFCT, GLS University, Ahmedabad, Gujarat, India.

[3]Dept. of Life Science, Gujarat University, Ahmedabad, Gujarat, India.

[4]Bioinnovations, Gujarat University, Ahmedabad, Gujarat, India.

[5]IICT, Indus University, Ahmedabad, Gujarat, India.

Article Received on 02/04/2020 Article Revised on 23/04/2020 Article Accepted on 13/05/2020

***Corresponding Author**

Dr. Krupa Mehta

GLSFCT, GLS University, Ahmedabad, Gujarat, India.

ABSTRACT

The proteins are the coding regions of the genome which stores the biological detail of human body. The information of mutation causing any disease is also available in protein that makes protein coding region i.e. exome sequence capable to predict the disease. The study of exome sequence helps to predict the disease easily compare to that of DNA sequence. The disease prediction requires passing multiple complex commands on different tools. The implementation of proposed prototype reduces the work load of clinician from passing set of commands to a single click. Moreover, there is absence of a generalized mechanism to predict cancer from exome sequence. This research work is an effort towards this direction. The proposed prototype produces difference between source exome sequence and reference exome sequence with respect to their SNP (Single Nucleotide Polymorphism) location and Indels (INsertion DELition). The difference in SNP location is useful for cancer prediction. At present, the clinician is supposed to perform various analysis steps by using different tools making the process of analysis very complex.

KEYWORDS: Biopython, cancer, exome, exome sequence, protein coding, python, Tkinter.

INTRODUCTION OF EXOME SEQUENCE

The human genome is collection of 3 billion nucleotides forming a DNA. A small portion of genome, approximately 1.5 percentage, is actually translated into coding region i.e. proteins. Protein is the base for human body structure, traits and diseases. The genome is composed of Exons and Introns. The term Exon was derived from "EXpressed regiON," since these are the regions that get translated, or expressed as proteins.[1] The term Intron is derived from "INTRagenic regiON" which is not represented in the final protein. The "exome" consists of all the genome's exons, which are the coding portions of genes.

Capturing the exomes from the living being helps to predict and identify the diseases. The exome represents less than 2% of the genome, but contains ~85% of known disease related variants, making whole-exome sequencing a cost-effective alternative compare to Whole Genome Sequencing (WGS).[2] This method allows variations in the protein-coding region of any gene to be identified, rather than selecting only a few genes. The mutations that take place in exons are responsible for diseases.

Importance of exome sequence

The exome is a source of rare disease related variants. Exome sequencing is used to identify genes and mutations that influence risk for human diseases. One of the immediate applications of exome sequencing is facilitating the accurate diagnosis of individuals with Mendelian disorders that are difficult to confirm using clinical or laboratory criteria alone. It is well justified strategy for discovering rare alleles underlying Mendelian phenotypes and perhaps complex traits like positional cloning, protein coding disorder, functional disorder. The exome sequencing predicts the functional consequences that can cause the damage by focusing on a large fraction of rare, protein-altering variants.

Exome sequencing is rapidly proving to be a powerful new strategy for finding the cause of known or suspected Mendelian disorders for which the genetic basis has yet to be discovered. The drop in per-base sequencing price is expected to drive the generation of immense amount of exome sequence data, creating a big data challenge in bioinformatics. Exome sequencing experiments produce millions to billions of short sequence reads at a high speed. The generated exome sequences need to be analyzed to identify the disease. Computer science comes into the picture to mange and analyzes such bulky and complex sequences.

Role of exome sequence in predicting cancer disease

Study of exome sequences allows early prediction of disease in human. Exome sequencing is the powerful and accurate method of predicting diseases. The exome represents less than 2% of the genome, but contains ~85% of known disease related variants, making whole-exome sequencing a cost-effective alternative to whole-genome sequencing. Sequencing the cancer exome provides useful information about the coding mutations that contribute to tumor progression.

Exome sequence analysis is more simple and cost effective than that of Whole Genome Sequencing (WGS). The size of WES data per sample is approximately a sixth of that of WGS data which not only reduces the storage burden but also reduces the processing time of analysis. However, the incomplete coverage of functional elements and low sensitivity for structural variant detection are considered as limitations of WES. Due to the limitations of WES, its utility was highly debatable when it was first introduced. In spite of such limitations, WES is more preferred tool for researchers as it is more viable practically and produces more robust results for large number of samples.[3] Exome sequencing is revolutionizing Mendelian disease gene identification. This results in improved clinical diagnosis, more accurate genotype-phenotype correlations and new insights into the role of rare genomic variation in disease.

The advent of cost effective and accurate Next Generation Sequencing (NGS) technologies, the approach to identify and cure disease is changing with the analysis of biological sequences.[4] The research in biomedical field has taken a magical turn after the successful implementation of NGS technologies. Capturing the biological sequencing and presenting was never such easy process before the introduction of NGS technologies. Next-generation sequencing (NGS) is now a popular technique for identifying novel and rare variants that are potentially associated with diseases. The analysis of NGS data often requires the integration of various resources.

In 2009, the Whole Exome Sequencing (WES) was introduced supporting the NGS technologies and making the process of disease identification much easier. The WES was merged with the NGS platforms enabling the analysis of rare and novel disease by selectively capturing and sequencing only the coding regions of the human genome.

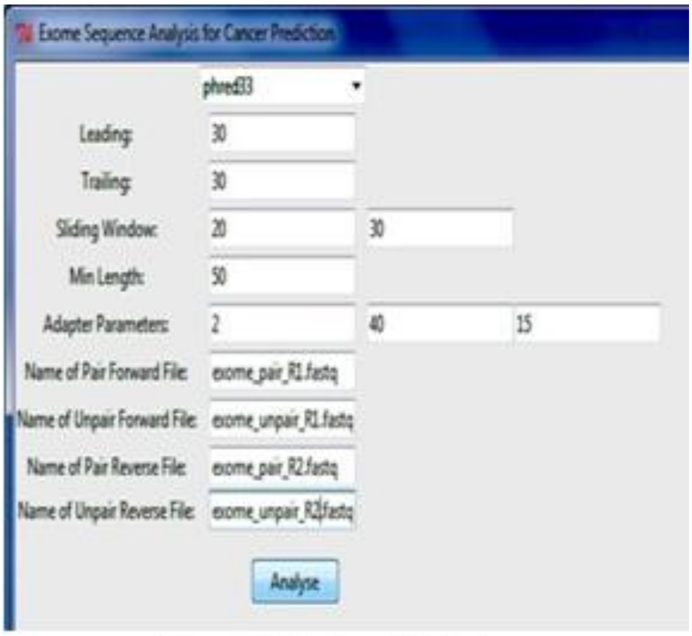
Cancer is one of the leading causes of death worldwide. Early detection and prevention of cancer becomes very important to reduce deaths caused by it. In 1986, with the ongoing Human Genome Project, it was predicted that the completion of HGP will produce reference cancer genes that will help to predict the cancer. The HGP was competed in 2003, fulfilling its promise providing a research and development platform to illuminate the pathogenesis of cancer.[5]

Computational analysis is one of the most challenging issues for cancer prediction through Whole Genome Sequence (WGS). WGS produces more than 90-150 GB of cancer and same amount of data for normal DNA that approximately reaches to terabyte of raw data. To process such massive amount of data, sophisticated computer resources are required. Cloud computing can provide solution to this problem, although it requires high speed network and bandwidth to transfer such bulky data.

Collaborating the power of computer with the exome sequence analysis can contribute effectively in cancer prediction. The conversion of biological pipeline into the algorithm serves the purpose. The task of automating the pipeline is very well supported by the existing computational tools. Computer science has a rich library of programming languages which can be used to design an interface for clinician. Various programming languages like C, C++, Java, Python, etc are available. The programming language, Python is having rich set of libraries supporting biological operations. Python is also capable of integrating third party tools effectively.

Proposed prototype for cancer prediction

This research work provides a base to predict the cancer disease by employing NGS technologies and exome sequence. NGS technologies are producing biological sequences like RNA, DNA, Protein sequences with accuracy. These generated sequences are main source of information for any type of disease analysis. This research work utilizes the power of exome sequences to predict the dreadful cancer disease. The analysis of captured exome sequence can only be fruitful and effective when it passes through multiple steps of an algorithm.



Exome Sequence Analysis for Cancer Prediction

phred33

Leading: 30

Trailing: 30

Sliding Window: 20 30

Min Length: 50

Adapter Parameters: 2 40 15

Name of Pair Forward File: exome_pair_R1.fastq

Name of Unpair Forward File: exome_unpair_R1.fastq

Name of Pair Reverse File: exome_pair_R2.fastq

Name of Unpair Reverse File: exome_unpair_R2.fastq

Figure 1 Developed Prototype

Algorithms can effectively and accurately be implemented in computers. The storage and processing capability of computer allows an algorithm to analyze the given information quickly and accurately. Computing contributes in processing and storage of data, processing of mathematically sophisticated calculations, processing of the complex logic of pipelines, finding the similarities between the sequences and identifying the new sequences.

Bioinformatics would not be possible without the advances in computer science hardware and software. The sequence analysis is playing vital role in prediction of diseases and for the designing of personalized medicines. The designed algorithm will take exome sequence as an input and performs various operations to predict cancer disease. The designed algorithm is:

Step – 1 Clean the captured exome sequence

Step – 2 Map the sequence

Step – 3 Calculate the quality score

Step – 4 Perform the SNP check

Step – 5 SNP/Variant Recalibration

Step – 6 Filter SNP(s) and Indel(s)

Step – 7 Identify the SNP location

Step – 8 Compare the SNP location in given sequence with the reference sequence.

Step – 8.1 Identify the SNP location on genome and their disease

Step – 8.2 Identification of pathogenic SNP

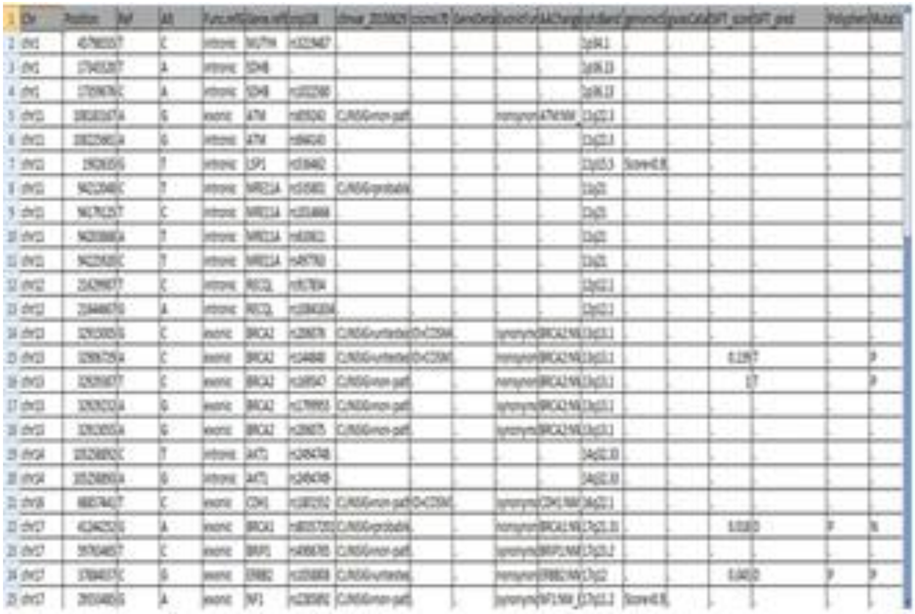


Figure 2: Generated VCF File.

The proposed cost effective algorithm to predict cancer at early stage is based on exome sequence analysis. The exome sequence analysis identifies the presence of the mutation responsible for cancer risk. The algorithm is implemented using Python language. The implementation is discussed below. The work of prototype starts after capturing the raw sequence. The raw sequence is provided as an input in prototype. The first screen of prototype is shown in Figure – 1.

According to the provided exome sequence and the inputs, a comparative VCF (Variant Calling Format) file is generated. The VCF file mentions the mutation of given exome sequence. This VCF file is compared with the GFF (General Feature Format) file containing the reference sequence. The comparison of both the files reveals the difference in SNPs and from this difference; a clinician can diagnose the cancer. The generated VCF is as Figure – 2.

The generated VCF file is compared with GFF file. The comparison of chromosome and position of VCF is compared with that of GFF file. When the chromosomes are same and position of VCF file falls between the start and end positions of the GFF file, the suspected mutation position is reported to the clinician.

The process of comparing VCF file with GFF file can be summarized in below steps:
Step – 1 Sort VCF file by chromosome
Step – 2 Sort GFF file by chromosome

```
Step – 3 While (end of VCF file not reached)
          VCFPosition = selected position from a row of VCF file
          startPosition = start position from a row of   GFF file
          endPosition = end position from a row of GFF file
          VCFchr = chromosome from VCF file
          GFFchr = chromosome from GFF file

If (VCFchr = GFFchr) then
If (VCFPosition >startPosition AND VCFPosition <endPosition) then
Report the mutation
End if
End if
End While
```

Importance of proposed prototype

The proposed algorithm to predict cancer at early stage is implemented using Python language. The implementation integrates various tools to perform steps of exome sequence analysis like cleaning, mapping, variant identification and annotation. The results produced by these tools are taken as base for further analysis using biological package of Python i.e. BioPython. Tkinter is used to design easy to use graphical interface which is presented to the clinician.

The implementation of proposed algorithm reduces the work load of clinician from passing set of commands to a single click. Moreover, there is absence of a generalized mechanism to predict cancer from exome sequence. This research work is an effort towards this direction.

The proposed prototype produces difference between source exome sequence and reference exome sequence with respect to their SNP (Single Nucleotide Polymorphism) location and Indels (INsertion DELition). The difference in SNP location will be useful for cancer prediction. At present, the clinician is supposed to perform various analysis steps by using different tools making the process of analysis very complex. This research work also provides a platform to further predict the disease and to design personalized medicine.

CONCLUSION

The exome sequence contains just 2% of the DNA sequence but contains 85% of known disease related variants. This characteristic of exome sequence makes it cost-effective alternate to the whole genome sequence. The exome sequence contains only protein coding region of the human genome so it is easy to perform analysis for exome sequence compare to that of whole human genome.

The current trend in clinical medicine is to analyze the exome sequence to predict rare or novel disease and design the remedy accordingly. The proposed algorithm to predict cancer at early stage is implemented using Python language. The implementation integrates various tools to perform steps of exome sequence analysis like cleaning, mapping, variant identification and annotation.

The results produced by these tools are taken as base for further analysis using biological package of Python i.e. BioPython. The graphical user interface is developed using Tkinter. It might seem novel approach now but will become the common practice of clinicians.

REFERENCES

1. Choi M, Scholl UI, Ji W, Liu T, Tikhonova IR, Zumbo P, et al, "Genetic diagnosis by whole exome capture and massively parallel DNA sequencing", Proc Natl Acad Sci U S A, 2009; 106: 19096-19101.
2. https://www.broadinstitute.org [30 Jan 16].
3. Sarah B. Ng1, Kati J. Buckingham, Choli Lee, Abigail W. Bigham, Holly K. Tabor, Karin M. Dent, Chad D. Huff, Paul T. Shannon, Ethylin Wang Jabs, Deborah A. Nickerson, Jay Shendure and Michael J. Bamshad, "Exome sequencing identifies the cause of a Mendelian disorder", PMC, 2010 July 01.
4. Sboner A, Mu X, Greenbaum D, Auerbach R, Gerstein M, "The real cost of sequencing: higher than you think!", Genome Biology, 2011; 12: 125.
5. Ada, Rajneet Kaur, "Using Some Data Mining Techniques to Predict the Survival Year of Lung Cancer Patient", IJCSMC, April 2013; 2(4): 1 – 6.
6. Gerald Goh, Murim Choi, "Application of Whole Exome Sequencing to Identify Disease-Causing Variants in Inherited Human Diseases", pISSN 1598-866X eISSN 2234-0742, Genomics Inform, 2012; 10(4): 214-219.
7. Hidewaki Nakagawa, Masashi Fujita, "Whole genome sequencing analysis for cancer genomics and precision medicine", Wiley Cancer Science, January, 2018.

8. Francis S. Collins, “SHATTUCK LECTURE — Medical And Societal Consequences Of The Human Genome Project”, The New England Journal of Medicine, July 1, 1999.
9. Francis S. Collins, Ari Patrinos, Elke Jordan, Aravinda Chakravarti, Raymond Gesteland, LeRoy Walters, and the members of the DOE and NIH planning groups, “New Goals for the U.S. Human Genome Project: 1998-2003”, SCIENCE, October 1998; 282,23.
10. Jeffery P. Struewing, Patricia Hartge, Sholom Wacholder, Sonya M. Baker, Martha Berlin, Mary Mcadams, Michelle M. Timmerman, Lawrence C. Brody, and Margaret A. Tucker, “The Risk Of Cancer Associated With Specific Mutations Of Brca1 And Brca2 Among Ashkenazi Jews”, The New England Journal Of Medicine,15, May 1997.
11. Zuoheng Wang, Xiangtao Liu, Bao-Zhu Yang and Joel Gelernter, “The role and challenges of exome sequencing in studies of human diseases”, Frontier in Genetics, August 2013; 4: 160.
12. Michael J. Bamshad, Sarah B. Ng, Abigail W. Bigham, Holly K. Tabor, Mary J. Emond, Deborah A. Nickerson and Jay Shendure, “Exome sequencing as a tool for Mendelian disease gene discovery”, Nature Reviews: Genetics, Translational Genetics, November 2011; 12.
13. Gerald Goh, Murim Choi, “Application of Whole Exome Sequencing to Identify Disease-Causing Variants in Inherited Human Diseases”, pISSN 1598-866X eISSN 2234-0742, Genomics Inform, 2012; 10(4): 214-219.
14. Krupa Mehta, Dr. Devarshi Mehta and Dr. Vishal Dahiya, “Exome Sequence Analysis Using Computational Tools For Clinical Diagnosis”, International Journal of Technological Innovation in Modern Engineering & Science, June-2018; 04(06).
15. Stephan Pabinger, Andreas Dander, Maria Fischer, Rene Snajder, Michael Sperk, Mirjana Efremova, Birgit Krabichler, Michael R. Speicher, Johannes Zschocke and Zlatko Trajanoski, “A survey tool for variant analysis of next generation genome sequencing data”, Briefings in Bioinformatics, January 2013; 15(2): 256-278.