Identification of Single Nucleotide Polymorphisms and associated Disease Genes using NCBI resources

Navreet Kaur
M.Tech Student
Department of Computer Engineering.
University College of Engineering, Punjabi University,
Patiala, Punjab, India
navreet_9@yahoo.co.in

Dr. Amardeep Singh
Professor
Department of Computer Engineering.
University College of Engineering, Punjabi University,
Patiala, Punjab, India
amardeep_dhiman@yahoo.com

Abstract — Single Nucleotide Polymorphisms (SNP) are the most common and simplest form of variation associated with genetic diseases. SNPs are responsible for more than half of the genetic polymorphisms known to cause inherited diseases. The SNPs are single base changes in the DNA sequence which are responsible for genetic diversity among individuals to the occurrence of diseases. The identification and analysis of the effects of SNP in relation to diseases can be helpful in early detection of diseases, preventing the onset of diseases, reducing the ill effects of diseases and in personalized drug development. The paper focuses on the method of detection of SNPs and associated disease genes using the National Center for Biotechnology Information (NCBI) resources. The BLAST, NCBI Map Viewer, dbSNP and Cn3D tools have been used to identify the SNPs and associated genes aligning the given Expressed Sequence Tag (EST). The results show the presence of a SNP in the used EST. The SNP is responsible for Hemochromatosis disease caused by variation in the HFE gene. The method can be generalized and used to identify the SNPs and disease genes present in a given DNA sequence.

Keywords- mutations, SNPs, genetic diseases, dbSNP

I. INTRODUCTION

The Human Genome Project (HGP) completed in 2003 has sequenced the complete genome of humans. It has triggered lot of interest in understanding the variability of human genome. The HGP shows that nearly 99.9% of the genome among individuals is same, only 0.1 % is different. This difference arises due to mutations, the most common being Single Nucleotide Polymorphisms (SNPs). Around 17 million SNPs have been identified on the human genome as per National Institute of Health (NIH).The SNPs involve the substitution of one base for another in the long string of DNA sequence built over the set of chemical letters {Adenine (A), Cytosine(C), Guanine (G), and Thymine (T)}.The SNPs are considered to play a critical role in determining the susceptibility of an individual to inherited diseases. The inherited diseases are caused by a defective or abnormal gene inherited from parents [1]. The study and analysis of SNPs and diseases is important because the variations in genome sequences underlie differences in our susceptibility to, or protection from, all kinds of diseases, in the age of onset and severity of illness, and in the way our bodies respond to treatment [2]. In recent years, the application of clinicogenetic knowledge has revolutionized the ability to understand the effects of nucleotide substitutions and genetic basis of several complex and common disorders [3]. In this paper, gauging the importance of SNPs in relation to diseases, a method for identification of SNPs and disease genes in a given DNA sequence has been presented.

The SNPs can change the amino acid of the protein and structure of a protein, impair/alter its function and lead to disease. Such SNPs which have the ability to alter the protein product are termed as non synonymous SNPs. There are various other classes of SNPs as listed in Table 1, but out of all, the non synonymous SNPs are mainly associated with inherited diseases. The non synonymous SNPs disrupt the protein structure, leading to diseases through various ways like breaking of disulphide bond, disruption of protein- protein interactions, size changes in the hydrophobic core, introduction of buried charged residues, disruption of hydrogen bonding network etc. [4].

TABLE I. FUNCTIONAL CLASSES OF SNPS

Functional Class of SNP	***Sub Type of SNP***	***Region***	***Abbreviation***	***Effect of SNP***
Non Synonmous SNP		Coding	nsSNP	SNPs in exons that cause an amino acid substitution
	Missense		mSNP	Change a codon to one that encodes a different amino acid and cause a small change in the protein produced
	Nonsense		nSNP	Change an amino-acid-coding codon to a single "stop" codon and cause an incomplete protein. This can have serious effects since the incomplete protein probably won't function
Synonmous SNP		Coding	sSNP	SNPs in exons that do not change the codon to substitute an amino acid.
Intronic SNP		Non Coding	iSNP	SNP fall in the non coding region (introns).
Regulatory SNP		Non Coding	rSNP	SNP fall in the regulatory region of gene.

The non synonymous SNPs affect the protein structure either through an amino acid substitution (Missense) or through an introduction of a stop codon (Nonsense). The missense substitutions are the most common SNPs associated with many genetic diseases like Sicke Cell Anaemia, Tay Sachs Disease, Hemochromatosis etc. Out of the 17 million known SNPs 1 million are missense SNPs. Usually, a large number of missense SNPs have little effect on the structure of protein, but the “functional” missense SNPs have a significant impact on protein structure. These functional missense SNPs may exert their effect on human physiology through various mechanisms, including modification of splice sites; inactivation of protein functional sites, such as catalytic, ligand-binding and post-translational modification sites; alteration of protein solubility and stability; or affecting the interactions of proteins - thereby perturbing protein functions, such as, the kinetic parameters of enzymes, signal transduction activities of transmembrane receptors, and architectural roles of structural proteins [5] The information related to SNPs and associated diseases is available from many public online databases like dbSNP, Human Genome Mutation database (HGMD), OMIM etc. These databases store SNPs belonging to all the functional classes. A method is needed to identify the SNPs associated with diseases out of the large pool of SNP data present in databases. The proceeding section explains the method for identification of disease associated SNPs in a given DNA sequence.

II. MATERIALS AND METHODS

The method mentioned below for detecting disease associated SNPs in a given DNA sequence has used the resources provided by National Center for Biotechnology Information (NCBI), MutDB, Cn3D and SNPeffect. All the resources are open source and available online.

A. Resources used in the study

The detailed explanation of the resources and the method of usage of the various software’s is presented below. The links for accessing/downloading the software’s are provided in Table II.

1) NCBI Basic Local Alignment Search Tool (BLAST) :

The Blast tool is used to find regions of local similarity between sequences [6]. The program compares nucleotide and protein sequences to sequence database. The BLAST provides genome sequences of various species like humans, mouse, rat etc for comparison. The BLAST has several variations like blastn, blastp, blastx, tblastn, tblastx to chose from for specialized narrowed searching and comparison.

2) NCBI Map Viewer :

The map viewer provides a variety of genome mapping and sequencing data. It is used for viewing and searching an organism’s complete genome [7]. It shows integrated views of a collection of genetic, physical, and sequence maps for annotated genes, expressed sequences, SNPs, and other features, and, thus, is a valuable tool for the identification and localization of genes that contribute to human disease.

3) NCBI dbSNP:

It is the database of SNPs. It consists of information about SNPs and other common mutations like insertions and deletions. It provides various ways of searching the database depending upon requirement like clinical source SNPs, SNPs in gene etc. All the details related to SNPs which include the functional class, function, associated disease, chromosome positions etc are reflected by the dbSNP.

4) NCBI Cn3D:

It stands for "see in 3 Dimensional". It is used to view 3D structures from the NCBI Entrez database. It can simultaneously display structure, sequence and alignment. It also has powerful alignment and annotation editing features [8]. The Cn3D needs to be downloaded for visualizing the structures. The protein structure and the effect of SNPs can be studied using the tool.

5) MutDB:

The MutDB is is an initiative of the Mooney Laboratory. The goal of the MutDB is to annotate human variation data with protein structural information and other functionally relevant information, if available. The mutations are organized by gene [9]. The gene entries can be browsed by disease, gene name, SNP. The search for a gene lists all the transcript variants of the gene and provides detailed information about each one of them.

6) SNPEffect:

It is a database for phenotyping the human SNPs. It primarily focuses on the molecular characterization and annotation of disease and polymorphism variants in the human proteome [10]. The software presents a detailed phenotypic summary of the searched variation using the tools which include:

- TANGO to predict aggregation prone regions.
- WALTZ to predict amylogenic regions.
- LIMBO to predict hsp70 chaperone binding sites.
- FoldX to analyze the effect of structure stability.

TABLE II. WEBLINKS FOR TOOLS

S.No	***Name of Tool***	***Link***
1	BLAST	blast.ncbi.nlm.nih.gov/
2	Map Viewer	www.ncbi.nlm.nih.gov/mapview/
3	dbSNP	www.ncbi.nlm.nih.gov/SNP/
4	Cn3D	http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml
5	MutDB	http://mutdb.org/
6	SNPEffect	http://snpeffect.switchlab.org/

B. Steps for Identification of SNPs and associated diseases

The SNPs can be identified in a provided DNA sequence and analyzed by integrating the resources discussed above. The steps are outlined below:

1) Compare Expressed Sequence Tag (EST) to Human Genome:

The EST is compared against the human genome using the blastn variation of BLAST. The BLAST compares the EST against all the sequences present in the database to find if it matches any already existing sequence. The output of BLAST of lists all the matching sequences along with percentage of similarity. The EST may be 100% similar or may have variations. The match with the highest degree of similarity is picked for analysis. The variation reflects the difference at nucleotide position. This difference between the EST and the matched sequence may be due to a sequencing error in EST or because of the presence of a SNP in EST.

2) Identify the Chromosome and Genes expressing the EST:

The chromosome to which the EST aligns is found by clicking the "Genome View" on the Blast Output page. The chromosome corresponding to the EST is shown marked in the genome view. To view the gene to which the EST aligns, the "Map Element" link is followed given at the bottom of the genome view window. It opens the NCBI Map Viewer.

3) Determining whether the ESTs contain known SNPs:

The Map viewer shows all the exons of the gene aligned to the EST in thick blue bars whereas the thin blue lines represent introns. The specific exon(s) to which the EST aligns is shown highlighted as a thick red bar. The Map viewer provides various links like sequence viewer, associated diseases (OMIM link) etc to study the details of the gene. As mentioned in Step 1 the EST may be 100% similar or may have variations. If The variation(s) found in EST could be due to the following reasons:

- It could be a variation from the normal genome sequence due to DNA sequencing error of EST.
- It could be a known SNP present in the human genome sequence.
- It could be an unknown SNP.

To view the existing SNPs in the gene, the variation map is added in Map viewer by clicking the “Maps and Options” link. It lists all the SNPs present in the gene. The listed SNPs are compared with variation found using BLAST. The comparison can be done either using sequence alignment in MATLAB or by manual analysis if lesser number of SNPs are aligned on the gene.

If the variation found using BLAST matches any of the SNP, it means that the variation is a known SNP else either of the other two cases prevails”.

4) *Determine whether the SNP is known to cause a disease:*

The SNP found above is analyzed using the dbSNP. The dbSNP by default shows the SNPs in the coding region of all the transcript variants. In order to view the disease associated SNPs, click clinical source. The SNPs which are responsible for a disease consist of an OMIM link. Check if the found SNP has an OMIM link. Follow the link to study the disease caused by SNP and other details.

5) *Studying the Structural and Functional Effects of SNP:*

The effects of the SNP on the protein structure and function can be studied through Cn3D and SNPEffect. The mutDB can be used to study the gene and SNP in details. It shows information regarding all the transcript variants. The Cn3D allows editing features to modify the representation of the protein structure.

Fig 1. displays the flow of the above described steps.

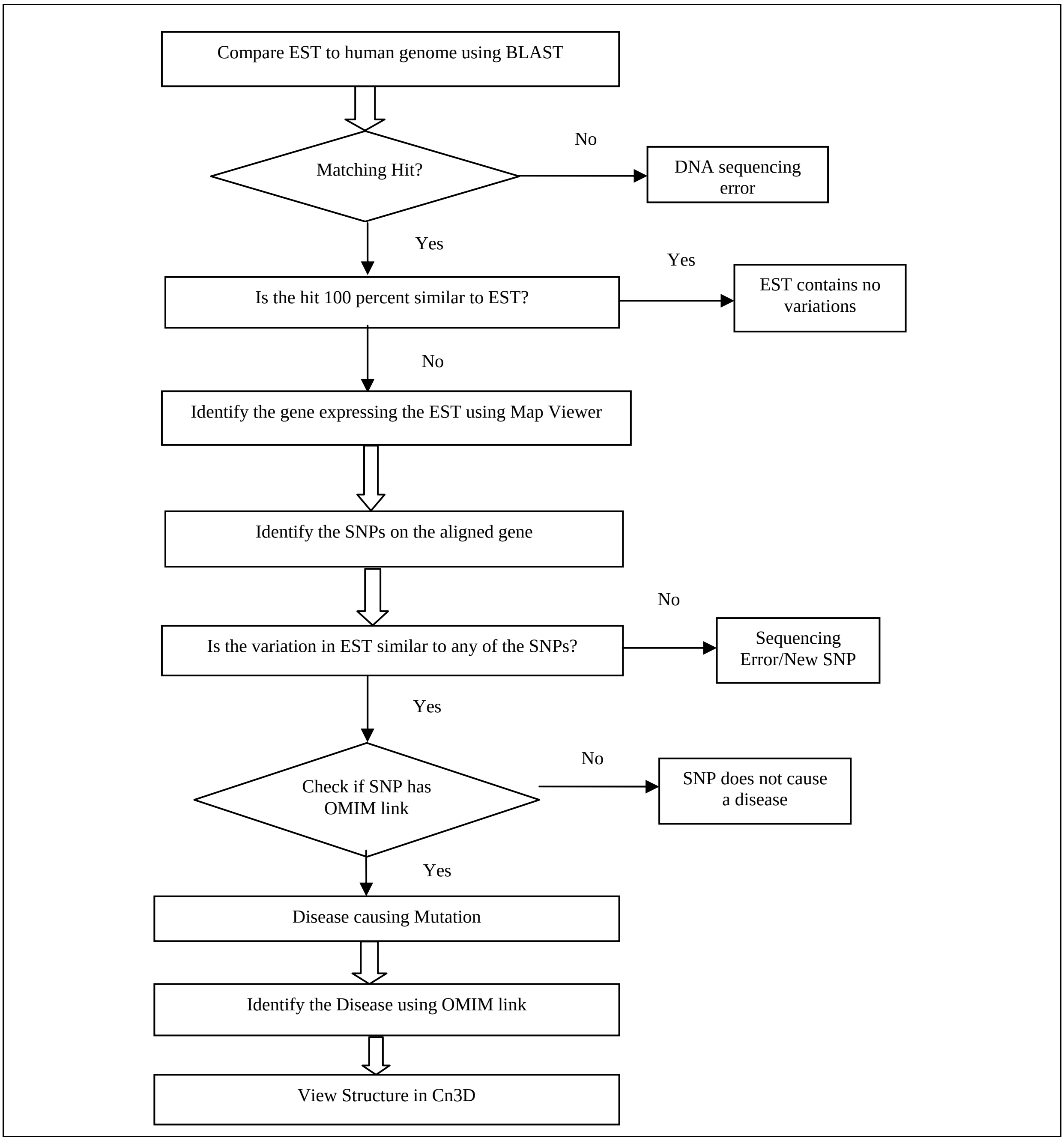
Compare EST to human genome using BLAST
Matching Hit?
No
DNA sequencing error
Yes
Is the hit 100 percent similar to EST?
Yes
EST contains no variations
No
Identify the gene expressing the EST using Map Viewer
Identify the SNPs on the aligned gene
Is the variation in EST similar to any of the SNPs?
No
Sequencing Error/New SNP
Yes
Check if SNP has OMIM link
No
SNP does not cause a disease
Yes
Disease causing Mutation
Identify the Disease using OMIM link
View Structure in Cn3D

Fig 1. Flowchart for identification of SNPs and associated diseases

III. Results And Discussions

The results show the presence of a SNP in the EST sequence. The SNP was found on chromosome 6 in the HFE gene. The SNP causes a change in the amino acid cystine to tyrosine at the 282nd position of the chain of amino acids. This change results in a breakage in the disulphide bond and hence affects the absorption of iron leading to the hemochromatosis disease. Fig 2. shows the cystine to tyrosine change position highlighted in green, viewed in Cn3D.

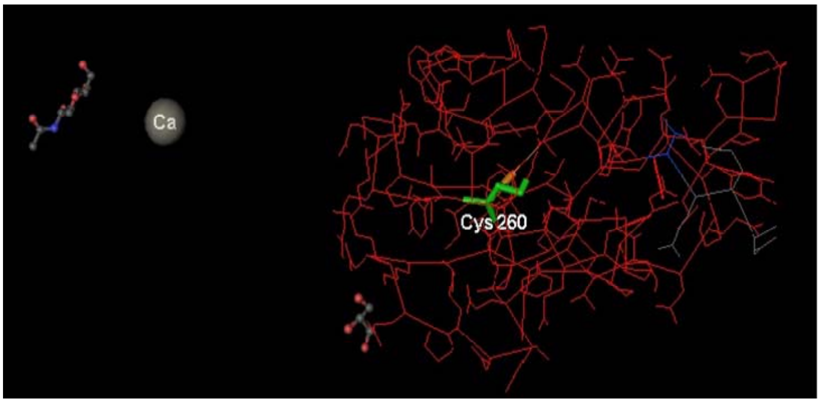

Fig. 2 Cystine to Tyrosine Amino acid Change Position

Similarly the above steps can be used to find the SNPs, if any, present in the DNA sequence. The Cn3D and SNPEffect can be used to understand the various structural and functional details of the protein structure and function. The identification of SNPs in the DNA sequence of a person proves to be beneficial in various ways. The identification can lower the effects of disease by early detection or in some cases prevent the onset of disease. Also sometimes people suffer from adverse drug reactions, the SNPs can help in development of personalized drugs.

ACKNOWLEDGMENT

N. K. thanks the developers of the course “Identification of Disease Genes” by NCBI which provided the foundation for the study.

REFERENCES

[1] S. C. Rastogi, N. Mendiratta and P. Rastogi. Bioinformatics: Methods and Applications: Genomics, Proteomics and Drug Discovery, 3rd ed., PHI Learning Pvt. Ltd., New Delhi, pp.2–3.

[2] A. Chakravarti. “Single Nucleotide Polymorphisms:… to a future of genetic medicine” , Nature 2001;409: pp 822-823.

[3] B.S. Shastry “SNPs: impact on gene function and phenotype”, Humana Press 2009: pp 3-22

[4] R. E. Steward, M. W. MacArthur, R. A. Laskowski and J. M. Thornton. “Molecular basis of inherited diseases : a structural perspective” Trends in Genetics 2003;19: pp 505-513

[5] T. R. Rebbeck, et al. “SNPs, haplotypes, and cancer: applications in molecualr epidemiology.” Cancer Epidemiology Biomarkers & Prevention 2004;13: pp 681-687

[6] BLAST: Basic Local Alignment Search Tool, Available at: http://blast.ncbi.nlm.nih.gov/Blast.cgi

[7] NCBI Entrez Map Viewer Help Document, 2004, Available at: http://www.ncbi.nlm.nih.gov/projects/mapview/static/MapViewerHelp.html

[8] Cn3D macromolecular structure viewer, 2011, Available at: http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml

[9] MutDb, 2008, Available at: http://mutdb.org/

[10] SNPEffect: Phenotyping Human Mutations ,2013, Available at: http://snpeffect.switchlab.org/