Journal of Biology and Today's World
ISSN 2322-3308
http://www.journalbio.com

Received: 24 November 2012 • Accepted: 06 December 2012

Research

doi:10.15412/J.JBTW. 01020103

Use of Haplotype assembly problem in eliminating SNPs from ApoE4 gene of the human genome

Monica Shekhar[1], Harleen Jabbal[1*], Jasleen Kaur[1], R.Selvakumar[2]

[1]School of Computing Sciences and Engineering, VIT University, India

[2]School of Advanced Sciences, VIT University, India

*correspondence should be addressed to Harleen Jabbal, School of Computing Sciences and Engineering, VIT University, India; Tell: +919566812830; Fax: +91; Email: harleenjabbal@gmail.com.

ABSTRACT

The human genomic sequence comprises of chromosomal DNA molecules. A DNA sequence is a double helical structure consisting of molecules of sugar, phosphate and nucleotides A, T, G, and C. A Single Nucleotide Polymorphism (SNP) is defined as a difference of a nucleotide between the genomic sequences of any biological species. The haplotype assembly problem was used as a solution methodology to eliminate SNPs in the ApoE gene of the human genomic sequence.

Key words: SNP, ApoE gene, Haplotype, C++ technology

1. INTRODUCTION

The genetic material of a human being consists of 23 pairs of chromosomes, where each chromosome pair is a set of two haploid chromosomes, together forming a diploid chromosome. Each chromosome is formed of an organized structure of DNA and protein present in the nucleus of every cell (1, 2). DNA, the basic building block of a chromosome, is a nucleic acid containing the genetic information of an individual. The DNA forms a double helical structure consisting of 2 long polymers made of nucleotides, sugar and phosphate. Nucleotides are of two types, purines (Adenine and Guanine) and pyrimidines (Thiamine and Cytosine). Normally, a purine of one strand bonds with a pyrimidine of the other with a bonding to T and G bonding to C (3). This is called complimentary base-pairing. Sugar and phosphate form the backbone of the helical strands, while the bases lie horizontally between them. Stretches of DNA, code for some hereditary characteristic of an individual. They hold information to build and maintain an organism's cells and pass hereditary characteristics to offspring. These units, located on a specific locus of the chromosome, are referred to as genes. Genes are located on each copy of the chromosome. Individuals have many genes which are responsible for various biological traits, for example, eye colour, number of limbs, blood type, etc. The variations in these genes account for the varied characteristics in individuals. The gene is, thus, the basic instruction while an allele is a variant of that gene. All individuals usually have a gene for a particular characteristic but different people have a specific allele for that gene which is responsible for the individual's unique set of biological traits. An allele present on each chromosome can be of different variants of a gene. If the two alleles are of the same type, they are said to be homozygotes and if they are of different types, they are heterozygotes. A species of organisms include multiple alleles. For instance, the gene for the ABO blood type recognizes 3 alleles namely I^A, I^B and I^O. These alleles determine the compatibility of blood transfusions. The set made up of one allele of each gene in a single chromosome is called a haplotype. These combinations of alleles are normally transported together. So a haplotype may be one locus, several loci or an entire chromosome depending on the type of recombination that occurred. In other words, it is a set of SNPs on a single chromosome (4, 5). A Single Nucleotide Polymorphism (SNP) occurs when a single nucleotide differs between chromosomes of members of a biological species or paired chromosomes of an individual. For example, two sequenced fragments from the DNA of two individuals AGAATC and AGATTC differ by a single nucleotide. So we can say that there are two alleles A and T. Two fragments taken from the chromosomal sequence of an individual are said to be in conflict when there exists a

SNP between them. The problem considers the apolipoprotein E gene from the gene of the human being. The protein ApoE is mapped to chromosome 19 and has three allelic forms, ApoE2, ApoE3 and ApoE4. ApoE4 is found to have physiological consequences in Alzheimer's disease. ApoE is technically defined by two SNPs, one of them being rs429358 located in the 4th axon of the ApoE gene. The more common allele is T, but if the allele is C, and the same chromosome also holds the rs7412 C allele, the combination forms the ApoE4 allele, which is a major factor responsible for the Alzheimer's disease. Two sequenced DNA fragments from the same or different individual may contain a difference in a single nucleotide. In this case, we say that there exists a SNP. The portion of the sequence that contains the SNP is called a fragment. The haplotype assembly problem creates bipartite graphs to eliminate conflicts in the fragments of the ApoE gene (6, 7). It was noticed that a set of fragments taken from the ApoE gene sequence forms a feasible haplotype assembly problem and by eliminating minimum number of conflicting fragments, a maximal bipartite graph was obtained. The conflict free sequence was found to be applicable in the field of pharmacogenomics, where customized drugs are manufactured based on the genomic sequence of an individual. This solution methodology was automated using C++ technology to enable similar inferences for a larger input set of fragments. The importance of studying SNPs in gene sequences lies in how they affect the way humans develop diseases and the way their genes respond to drugs. SNPs can be used for comparing the sequences of individuals with and without a particular disease, and thus, enables researchers to develop personalized drugs for them. In other words, SNPs are potentially useful in the field of pharmacogenomics, which focuses on manufacturing customized drugs for a particular individual by using their genetic information.

2. MATERIALS AND METHODS

2.1. Obtaining the ApoE4 gene sequence

The open source website owned by the National Centre for Biotechnology Information (NCBI) provides access to biomedical and genomic information for the betterment of health sciences. It provides a free tool BLAST, Basic Local Alignment Search Tool, which performs carries out several functions using a vast set of genomic databases. The experiment involved a BLAST search of the ApoE genomic sequence. The tool provided a graphical display to review several alignments of this gene. The following steps highlight the extraction process

Step1. We watched locating the ApoE gene in Figure 1.

Step2. We watched specifying the particular sequence for the search criteria in Figure 2. Extracting the ApoE4 sequence from the entire sequence using BLAST

Step3. We watched the BLAST tool displaying the SNP's with the aligned fragments in Figure 3.

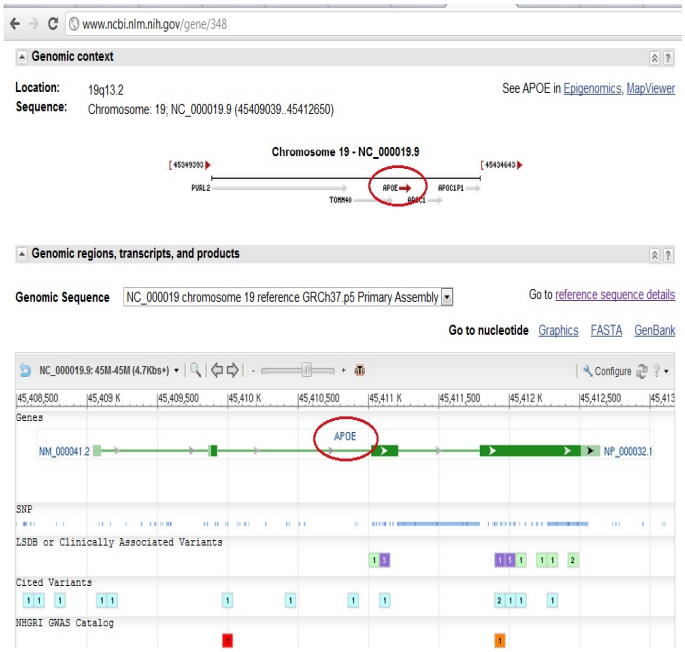

Figure 1. The BLAST tool showing the location of the ApoE gene sequence

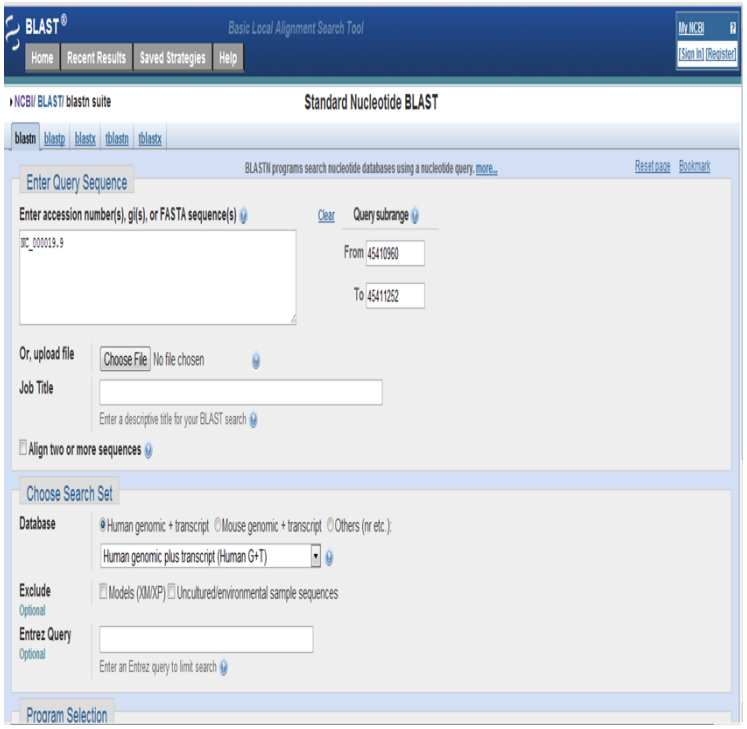

Figure 2. Extracting the ApoE4 sequence from the entire sequence using BLAST

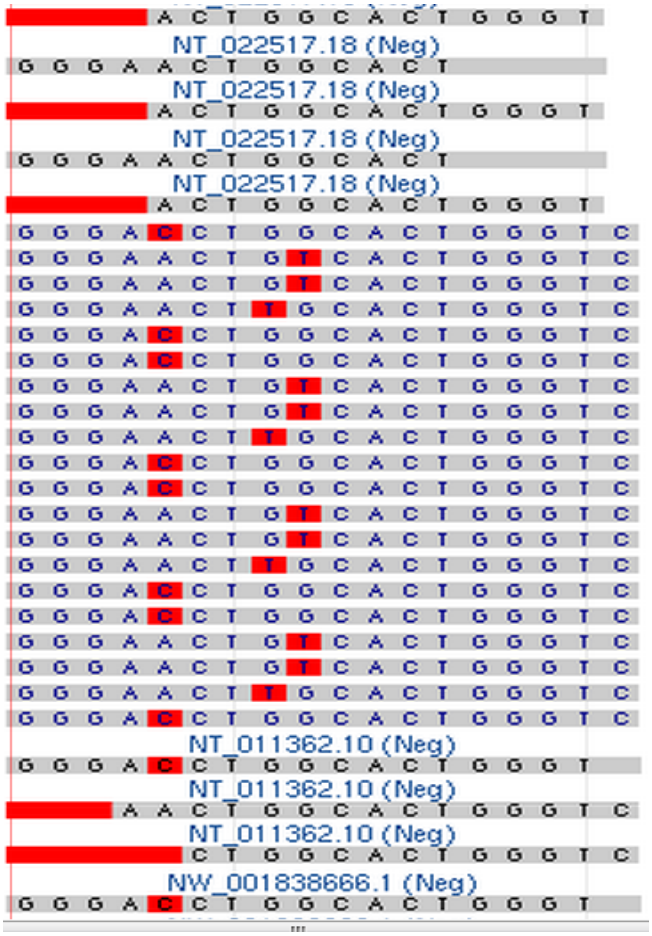

Figure 3. BLAST tool displayed the fragments of the ApoE gene. The highlighted regions represent the occurrence of a SNP

2.2. Solution methodology

A set of m fragments, F, a set of n SNPs, S, are identified from the ApoE sequence. A relation R: S×F= {0, if no SNP occurs; A or B, if a SNP s_i occurs on a fragment f_j where A and B denote the heterozygous alleles on the chromosome}. This tipple {S, F, and R} is known as an SNP assembly. An SNP assembly is feasible when the set F can be partitioned into two blocks H_1 and H_2 called haplotypes. Depending on the occurrence of SNPs on a fragment, F is used to create conflict graph G_F as shown in Figure 4. The set of vertices are the fragments and there is an edge between them if and only if the two fragments are in conflict, i.e. if they contain a SNP. Such a graph can be plotted by using the matrix S x F or F x S, where $A_{i,j}$ takes values as defined by the relation R on S and F. The Table 1 shows the matrix for 10 fragments chosen from Figure 3.

The problem is to remove minimal number of fragments so as to render the graphs conflict free. For G_F, we find a maximal bipartite graph. This process involves the removal of the fewest possible vertices to make a bipartite graph as shown in Figure 5.

Table 1. Matrix representation of the 1st 10 fragments taken from Figure 2, displaying the set of fragments and the set of SNPs

	s1	*s2*	*s3*
f1	A	A	A
f2	B	A	A
f3	A	A	B
f4	A	A	B
f5	A	B	A
f6	B	A	A
f7	B	A	A
f8	A	A	B
f9	A	A	B
f10	A	B	A

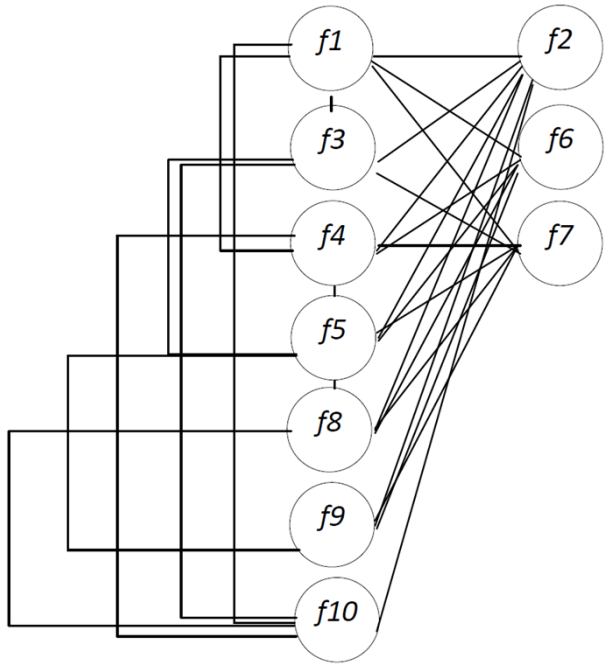

Figure 4. The conflict Graph G_F showing the fragments as the vertices and edges are present between fragments only if they are in conflict

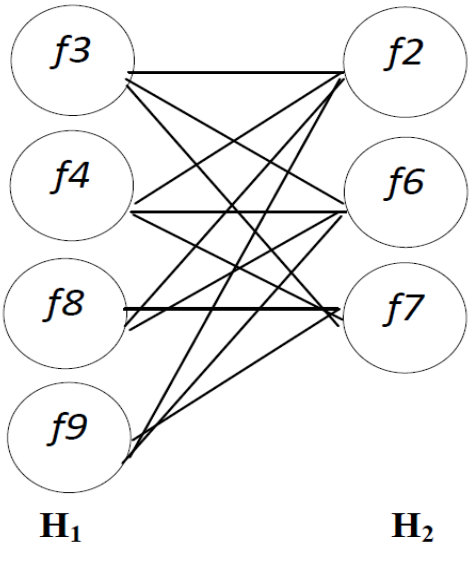

Figure 5. A bipartite graph which shows the Haplotypes H_1 and H_2 corresponding to Figure 4

2.3. Tools

A program is created using C++ technology which converts graphs into maximal bipartite graphs. It takes as input the vertices and edges of the conflict graph G_F and gives as output the vertices and edges of the maximal bipartite graph. The following is the algorithm that is used to create a maximal bipartite graph:

1. Input number of vertices n in G
2. For i from 1to n
   - 2.1 For j from 1to n
     - 2.1.1 Input into A [i] [j] 1 if there is an edge from i to j, 0 otherwise
3. Do for all i from 1 to n
   - 3.1 Do for all k from 1 to n
   - 3.2 Assign group 1 or 2 to each vertex
   - 3.3 The Grouping Graph is thus obtained
4. For i from 1 to n
   - 4.1 For j from 1 to n
     - 4.1.1 If group of vertex i is 1
       - 4.1.1.1 G_1[i] = vertex i;
     - 4.1.2 If group of vertex i is 2
       - 4.1.2.1 G_2 [i] = vertex i;
5. For all i in G_1 [0] to G_1 [n]
   - 5.1.1.1 Consider the adjacency matrix for vertices in G_1
   - 5.1.1.2 Do
   - 5.1.1.3 Eliminate the vertex with max number of 1's in a row
   - 5.1.1.4 While (Matrix becomes a NULL matrix)
6. For all i in G_1 [0] to G_1 [n]
   - 6.1.1.1 Consider the adjacency matrix for vertices in G_1
   - 6.1.1.2 Do
   - 6.1.1.3 Eliminate the vertex with max number of 1's in a row
   - 6.1.1.4 While (Matrix becomes a NULL matrix)
7. The remaining vertices after elimination in step 5 and step 6 form the bipartite graph according to the respective groups (Figure 6).

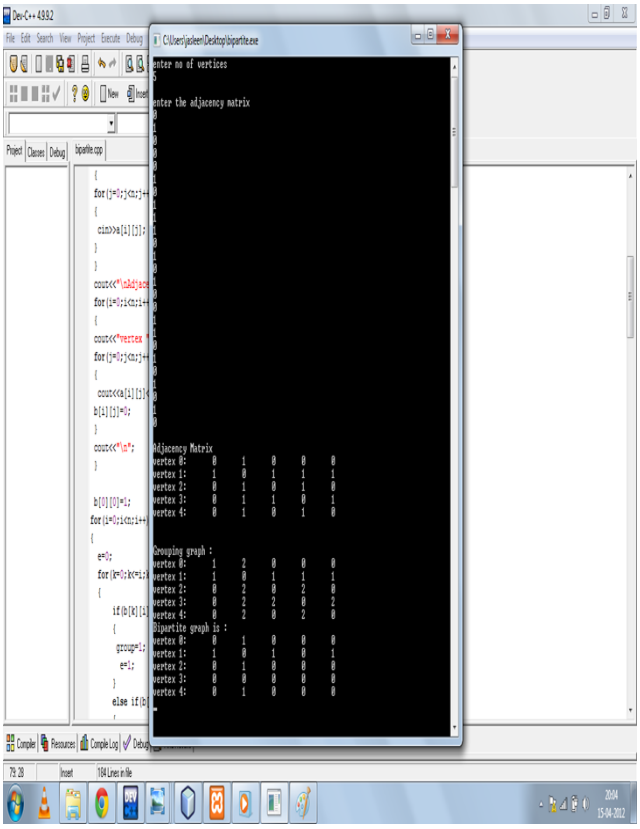

Figure 6. Screenshot showing the adjacency matrix as input and the adjacency matrix of the maximal bipartite graph as output

3. RESULTS AND DISCUSSION

It was observed that by removing fragments 1, 5 and 10, a conflict-free group of fragments was obtained, which could be partitioned into two haplotypes. From this, we can determine the two sets of fragments which have been inherited from the parents' set of haplotypes. Further, it can be seen that fragments 3, 4, 8 and 9 belong to H1, while fragments 2, 6 and 7 belong to H2. This kind of analysis is useful in the creation and study of haplotype maps of the human genome. A haplotype map is made with the purpose of relating genetic variation among individuals with its corresponding disease susceptibility. By studying the occurrences of SNPs in the ApoE gene, the haplotype map of that gene can be created. This map can contribute to the study of the variations of that gene among individuals. The usefulness of haplotype maps is in the field of pharmacogenomics, which focuses on customized health care. Pharmacogenomics is the study of an individual's genetic design and their response to drug treatment. Since each individual is genetically different, drugs that work well on some may not work well for others, and hence, specific drugs can be manufactured in accordance to each individual's genetic design. A pharmacogenomics experiment on Alzheimer's disease requires the identification of candidate genes (one of them being ApoE). And it then involves the identification of polymorphisms in the ApoE gene, which use the corresponding maps.

4. CONCLUSION

The study has used the open source tool, BLAST, provided

by NCBI to identify SNPs in the ApoE gene responsible for Alzheimer's disease. A feasible assembly problem has been identified on the given set of fragments and a conflict graph has been constructed using the concept of matrices. The haplotypes have been identified corresponding to the haplotype map of the ApoE gene using the concept of bipartite graphs. Haplotype maps are invaluable in the field of pharmacogenomics and with the increasing interest to tailor made treatments to patients based on their genomic profiles, their significance will only increase with time.

ACKNOWLEDGMENT

No mentioned any acknowledgment by authors.

AUTHORS CONTRIBUTION

This work was carried out in collaboration between all authors.

CONFLICT OF INTEREST

The authors declared no potential conflicts of interests with respect to the authorship and/or publication of this article.

REFERENCES

1. Lippert R, Schwartz R, Lancia G, Istrail S. Algorithmic strategies for the single nucleotide polymorphism haplotype assembly problem. Briefings in bioinformatics. 2002;3(1):23-31.
2. Smith K. Genetic Polymorphism and SNPs. Disponible sur http://www cs mcgill ca/~ kaleigh/compbio/snp/snp_summary html. 2002.
3. Fallin D, Cohen A, Essioux L, Chumakov I, Blumenfeld M, Cohen D, et al. Genetic analysis of case/control data using estimated haplotype frequencies: application to APOE locus variation and Alzheimer's disease. Genome research. 2001;11(1):143-51.
4. Peila R, Rodriguez BL, Launer LJ. Type 2 diabetes, APOE gene, and the risk for dementia and related pathologies The Honolulu-Asia Aging Study. Diabetes. 2002;51(4):1256-62.
5. Cladaras C, Hadzopoulou-Cladaras M, Felber B, Pavlakis G, Zannis V. The molecular basis of a familial apoE deficiency. An acceptor splice site mutation in the third intron of the deficient apoE gene. Journal of Biological Chemistry. 1987;262(5):2310-5.
6. Pennacchio LA, Rubin EM. Comparative genomic tools and databases: providing insights into the human genome. Journal of Clinical Investigation. 2003;111(8):1099-106.
7. Syvänen A-C. Accessing genetic variation: genotyping single nucleotide polymorphisms. Nature Reviews Genetics. 2001;2(12):930-42.