Acta Scientific COMPUTER SCIENCES

Volume 3 Issue 1 January 2021

Editorial

Tag SNP Selection

Sujay Saha*

Department of Computer Science, India

***Corresponding Author:** Sujay Saha, Department of Computer Science, India.

Received: August 21, 2020

Published: November 30, 2020

Abstract

Single Nucleotide Polymorphisms (SNPs) help to identify genetic variants responsible for complex human diseases. It also provides some valuable information on human evolution history. If the number of SNPs is too large then the study related to SNP-based disease identification becomes difficult. Therefore, it is essential to select a small set of representative SNPs, i.e. tag SNPs to represent the rest of the SNPs. Since this problem is already proved to be an NP-hard, therefore some heuristic methods may be useful.

Keywords: SNP; Nitrogen Bases; TagSNP; Linkage-Disequilibrium

Introduction

It is known that DNA is a nucleic acid molecule that contains the genetic instructions used in the development and functioning of all known living organisms. DNA is a long polymer of four simple units called nucleotides or bases. The four bases are Adenine (A), Cytosine (C), Guanine (G), and Thymine (T). There are approximately three billion bases in the human genome and those bases are divided into 23 chromosomes. Diploid organisms have a duplicate set of genetic material consisting of paired chromosomes, one from each parent. Such paired chromosomes which are essentially identical, having only small differences originating from the variability present in the population are called homologous. On the other hand, the organisms with only one copy of each chromosome are called haploid.

As per the Theory of Mendelian Inheritance, the variations in phenotype are due to the differences in genotype, i.e. the observable physical and behavioral characteristics of an organism depends on a set of particular genes of that organism, each of which are responsible for a particular characteristic. For every sexually reproducing organism, each gene is represented by two copies, called alleles-one on each chromosome pair. If both the alleles are same then the gene is said to be homozygous. But if the alleles are different, they are said to be heterozygous. Different alleles may give rise to different phenotypes. Alleles may be classified into two categories - dominant and recessive. Dominant alleles give rise to their corresponding phenotypes when paired with any other allele for the same trait, whereas recessive alleles give rise to their corresponding phenotype only when paired with another copy of the same allele. For dominant allele, the phenotype appears the same in both the heterozygous and homozygous states. But for recessive allele, the phenotype is only seen when both the alleles are same. Let's suppose that a gene exists in two allelic forms A and B. So, three possible genotypes are: AA, AB, BB. If AB individual shows the same characteristic as AA, and BB shows a different phenotype, then A is said to be dominant over B.

Although more than 99% of human DNA sequences are the same across the population, the remaining less than 1% DNA variations can have a major impact on the way humans react to various diseases. One main source of variation comes from Single Nucleotide Polymorphisms (SNPs). A SNP is a mutation at a single nucleotide position where a possible nucleotide type is called an allele. For example, there are two nucleotides in the following two DNA fragments in the fourth position: CCACGTT and CCATGTT. In this case, we say that the SNP has two alleles, C and T. Almost all SNPs have only two alleles, referred to as bi-allelic SNP. The tri-allelic and tetra-allelic SNPs are extremely rare.

When a representative SNP is closely linked to a group of SNPs, it can be chosen as a tag SNP for this group. Efforts have been made on selecting as few tag SNPs as possible for a data set. This set of tag SNPs can predict the remainder of the data set with a high precision. Beside this, we know that the cost of a genetic study is directly influenced by the number of SNPs genotyped. Since the number of SNPs in the human genome is too large, it is impractical to genotype all the SNPs. It suggests to typing a subset of all SNPs because of the high correlation between nearby SNPs. That's why the tag SNP selection problem plays an essential role in SNP data analysis, since the cost and complexity of experiments can be significantly reduced.

Linkage Disequilibrium (LD) is a term normally used for the non-random association of alleles at two or more loci in the population. In most cases, two SNP loci (x, y) share a certain amount of correlation or association, i.e. they are linked. Let's denote that the two alleles of the first SNP as A and a, and the two alleles of the second SNP as B and b. If it is possible to predict the allele at y based on the allele at x in every trial, i.e. if the prediction accuracy is 100%, then these two SNPs are said to be fully linked. If there is no linkage between them then the alleles at positions x and y are independent to each other. But there are occasions where we have some partial information about the allele at y given the allele knowledge at x and this phenomenon is referred to as linkage disequilibrium (LD). The popular metrics of LD, like D, D', r2, are closely related to the problem of tag SNP selections.