World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:15, No:09, 2021

Frequent Pattern Mining for Digenic Human Traits

Authors : Atsuko Okazaki, Jurg Ott

Abstract : Some genetic diseases ('digenic traits') are due to the interaction between two DNA variants. For example, certain forms of Retinitis Pigmentosa (a genetic form of blindness) occur in the presence of two mutant variants, one in the ROM1 gene and one in the RDS gene, while the occurrence of only one of these mutant variants leads to a completely normal phenotype. Detecting such digenic traits by genetic methods is difficult. A common approach to finding disease-causing variants is to compare 100,000s of variants between individuals with a trait (cases) and those without the trait (controls). Such genome-wide association studies (GWASs) have been very successful but hinge on genetic effects of single variants, that is, there should be a difference in allele or genotype frequencies between cases and controls at a disease-causing variant. Frequent pattern mining (FPM) methods offer an avenue at detecting digenic traits even in the absence of single-variant effects. The idea is to enumerate pairs of genotypes (genotype patterns) with each of the two genotypes originating from different variants that may be located at very different genomic positions. What is needed is for genotype patterns to be significantly more common in cases than in controls. Let Y = 2 refer to cases and Y = 1 to controls, with X denoting a specific genotype pattern. We are seeking association rules, 'X → Y', with high confidence, P(Y = 2|X), significantly higher than the proportion of cases, P(Y = 2) in the study. Clearly, generally available FPM methods are very suitable for detecting disease-associated genotype patterns. We use fpgrowth as the basic FPM algorithm and built a framework around it to enumerate high-frequency digenic genotype patterns and to evaluate their statistical significance by permutation analysis. Application to a published dataset on opioid dependence furnished results that could not be found with classical GWAS methodology. There were 143 cases and 153 healthy controls, each genotyped for 82 variants in eight genes of the opioid system. The aim was to find out whether any of these variants were disease-associated. The single-variant analysis did not lead to significant results. Application of our FPM implementation resulted in one significant ($p < 0.01$) genotype pattern with both genotypes in the pattern being heterozygous and originating from two variants on different chromosomes. This pattern occurred in 14 cases and none of the controls. Thus, the pattern seems quite specific to this form of substance abuse and is also rather predictive of disease. An algorithm called Multifactor Dimension Reduction (MDR) was developed some 20 years ago and has been in use in human genetics ever since. This and our algorithms share some similar properties, but they are also very different in other respects. The main difference seems to be that our algorithm focuses on patterns of genotypes while the main object of inference in MDR is the 3 × 3 table of genotypes at two variants.

Keywords : digenic traits, DNA variants, epistasis, statistical genetics
Conference Title : ICFPMA 2021 : International Conference on Frequent Pattern Mining and Algorithms
Conference Location : Toronto, Canada
Conference Dates : September 20-21, 2021