*Journal of Computer Science & Computational Mathematics, Volume 9, Issue 1, March 2019*
DOI: 10.20967/jcscm.2019.01.001

# A Study on Frequent Itemset Mining for Identifying Associated Multiple SNPs

Sofianita Mutalib*, Azlinah Mohamed, Shuzlina Abdul-Rahman

Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia
*Corresponding author email: sofi@tmsk.uitm.edu.my

***Abstract*:** Genome-wide association studies (GWAS) have gained a lot of interest in public health research to investigate the correlations of genetic variants and traits. Mostly, GWAS use standard statistical tests for each genetic variant to capture main genetic effects. Machine learning and data mining approaches are also promising enough to complement single and multiple genetic variants in understanding the general association of complex human disease. This paper explores a data mining approach to discover patterns of multiple genetic variants associated with a disease. Frequent itemset mining method was applied and the intersection algorithm in a row enumeration strategy was chosen to discover itemsets from genetic variants, which is known as Single Nucleotide Polymorphism (SNP). We chose the intersection algorithm because it is more suitable to mine high dimensional and sparse dataset. The found itemsets could be used by scientists to study associated with multiple genes in multifactorial disease.

***Keywords*:** frequent itemset mining, genetic variants, intersection algorithm, single nucleotide polymorphism, row enumeration.

## 1. Introduction

Data mining has shown many significant results in discovering patterns and potentially useful knowledge in many domains. The life sciences and biological domains are considered as one of critical fields that always seek for useful knowledge including relationships and correlation between factors. One of active studies in life sciences and biological domain is genome wide association study (GWAS), that is to identify the genetic variants associated with physical traits, diseases or phenotypes. GWAS collect dense marker sets, known as Single Nucleotide Polymorphisms (SNPs) and SNPs are the most abundant type of sequence variants in the human genome. SNPs occur approximately once in 100 to 300 base-pairs [1, 2]. Commonly, SNPs are normally to be found at least 5% in the population. Thorough studies need to be conducted to develop an appropriate method in identifying association of the genetic variants. The algorithms in data mining are expected to discover hidden knowledge in the genetic association that compliments the scientists perform further validation.

Most of GWAS focus on detection of main effects of SNPs by using statistical on single SNP test [3, 4]. Testing each SNP for the effect is computationally infeasible for high dimensional data. Moreover, complex diseases are most probably not caused by a single SNP alone, but it is resulted from complex non-linear interactions among genetic factors [2]. The conventional single SNP testing did not provide a good solution for disease prediction and there is a need to develop a solution for multiple SNP analyses from a very high number of SNPs dataset [5]. Nonetheless, multiple SNPs test by using machine learning or data mining approach could be performed in discovering combinations of SNPs in identifying risk factors especially in multifactorial disease. Even though many researchers have done tremendous findings on identifying loci, dimensionality reduction and also detecting and modeling gene-gene (SNP-SNP) interaction in certain diseases through GWAS, but the scientists had also shown more interest in understanding the produced result. They had also highlighted the lack of transparency and limited biological interpretation in the result of existing methods [4, 6].

Frequent itemset mining (FIM) is a data mining approach to find frequent patterns occur in specific data set. The frequent patterns are discovered after running the specific algorithm with pre-determined support value, as the threshold. Additionally, the produced patterns encapsulate efficient and useful knowledge discovered from high dimensional data sets such as a genome wide dataset. Next, the frequent patterns will be used to generate rules. A rule is a structured representation of knowledge that easily be understood and a knowledge base could be developed to represent the relationship between genetic variants (multiple SNPs) to complex disease, such as for diabetes and cancers. A rule normally represented as in the form of "if-then rules" could provide better understanding and interpretation to the result of data mining [7, 8]. The relationship is presented using "if-then rules" could motivate the scientists and inspire them to start further investigation and experiments. Indirectly, the outcome of the data mining result will be useful in the field of health, in order to identify the cause and effect of genetic variants of the selected disease. The study can benefit other researches in producing intervention for better living.

This paper would present a study in exploring FIM method that could mine itemset and to find meaningful knowledge in SNP data. With the motivation of studies that explore FIM methods in microarray and several benchmark datasets, this study applied an existing FIM algorithm and analyze the itemset. Among the existing FIM algorithms, row enumeration strategies are suitable for high dimensional data sets [9, 10], which mostly employed transposed table operations or intersection operations. The rest of the paper is organized as follows. The description of genetic variants and genome wide association studies are given in Section 2. Section 3 is a summary of related studies. Section 4 describes the experimental design for frequent itemset mining. The results of experiments are presented in Section 5. Finally, conclusions are presented in Section 6.

## 2. Genetic Variation

Genetic variation describes the genetic differences that may occurring in individuals of the same population. As scientists obtain and analyze genomic sequence information, more and more genetic variants are identified. Genotype is part of the cell and it carries genetic codes consists of inheritance instruction for the appearance or behavior that show genetic differences. Genotype can be defined by many markers such as polymorphisms, microsatellites and copy number variations [11, 12] and there are several types of genetic polymorphisms [1], namely Restriction Fragment Length Polymorphism (RFLP), Simple Sequence Repeats, Tandem Repeats and Single Nucleotide Polymorphism (SNP). SNP genotyping is the measurement of genetic variations of SNPs between members of a species. SNP occurs in a genome sequence when a single nucleotide, A, T, C or G, differs between members of biological species or a population. SNPs in populations are found with minor allele frequency or the lowest allele frequency at a locus. SNP could be the key marker for genetic association with heritable traits such as complex diseases. A marker with a SNP could be extended to an interesting relationship through studies of multiple SNPs instead. The trend of research currently shows encouraging technologies and methods to perform multiple SNPs (multi-genes) by incorporating variation from the entire genome of a person in the population. Among the most awaited results is the cause and effect of genetic variants in multi-genes to diseases. The growth of research interests in this area is shown by the increasing number of GWAS for many different purposes.

The first published paper in GWAS was by DeWan and collegues in Chinese population [13] for age-related macular degeneration (AMD) with result on chromosome 10q26 as major genetic factor for wet AMD. This is followed by Klein et al. [14] for AMD in H factor polymorphism and Wellcome Trust Case Control Consortium (WTCCC) in seven diseases [15]. The data collection would normally gather 100 to 300 thousand SNPs. The relationship with disease is generally evaluated for each SNP using a trend test across the number of minor alleles. Through the SNPs test, GWAS proved association, but only for small portion of genetic variants and the test requires of a step for fine mapping in the locus to determine the most critical variants, which means numbers of experiments are to be replicated many times for different locus. It is a tedious repetitive process to be done for millions of SNPs, but result probably reveals only a small portion of SNPs are significant, and however, it is considered a good enough result.

Normally, the method used to identify single SNP association is significant test using regression [3, 16]. However, several researchers argued if it is a single SNP caused gene disorder, it is always true for Mendelian diseases, meanwhile the risk of complex disease mostly due to the interaction of multiple SNPs, rather than a single SNP [7, 11, 17]. The analysis of GWAS data can also test multimarker combinations of SNPs, haplotypes, or interactions for their association with disease. With multiple SNPs test, the result gained could be used in predicting and classification purposes. Unfortunately, multiple SNPs test is complicated and need more computational resources. The problem also related to the issue of curse of dimensionality [18], when large number attributes compare to records in GWAS database. Thus, it is important to integrate computational intelligent methods that can perform efficiently to discover related knowledge using these biological datasets.

## 3. Related Studies in FIM

Studies of multiple SNPs effect have been done using machine learning methods including classification [3, 4]. Classification is referred to a task to categorize a set of data into more than a class. The classifier is built using a set of variables or features and mapped to predetermined classes, with 'yes' or 'no' label. For example, a record, X, is represented by an n-dimensional attribute vector, $X = (x_1, x_2, x_3, \ldots, x_n)$ and X has a specific class label value that determine the category of the record. Normally the tasks involved in classification are training and testing the classifier algorithm as a supervised method. In the context genomics data, the purpose of the studies in classification could be ranged from prediction, pattern recognition and so forth. Kelemen et al. [16] identified popular supervised learning algorithms for SNPs are random forest, multifactor dimensionality reduction and support vector machine.

Other than classification, FIM is a popular method for discovering interesting relationship between items in large databases. An itemset is a collection of related items that occur together in a specified database. FIM firstly is applied in customer transaction database, and the algorithm is known as Apriori [19]. It is also known as market basket analysis, which the results could be used by the retailers to develop marketing strategies and the frequency of occurrence products bought together is analyzed. It helps in understanding the shopping behaviors and the results could be used in marketing. The relationship between these items is also known as frequent patterns (FPs). FIM critical process would be the searching for the itemset of frequent patterns. It always correlated with the amount of data to be processed in memory and engines load. Having a set of $n$ items per basket, there would be $n!/r!(n-r)!$ combination of items generated, and $r$ is an integer number. The Apriori algorithm attempts to find frequent patterns with common $k$-itemsets to at least a minimum support threshold. Apriori uses a bottom up approach, where the frequent itemsets are extended one item at a time and groups of candidates are tested against the data. The stopping criteria is when no further successful extentions are found. It would have to store all these combinations of the transactions and secondly, to iterate through all of them to find the frequent pairs. The process involves multiple scanning and normally results bottleneck in candidate-generation-and-test. Problems of memory consumption normally occur in producing $k$-itemsets and when $k$ equal to 2, as the number of candidates increase exponentially.

When the number of candidates is huge, the workload of counting the support value of each candidate becomes really tedious. So the improvement of FIM algorithms could be done in reducing the number of database scans in searching the FPs and in facilitating support counting of the candidates. Another algorithm, FP-growth algorithm that constructs frequent tree

based on conditional base and generate frequent patterns without candidate. This algorithm reduces the number of multiple databases scanning. Apriori, FP-growth and its variants, such as CHARM, CLOSET and MAFIA, are operating by enumerating candidate itemsets, and this is referred to column enumeration strategy. These algorithms use breadth first search and tree structure to count candidate itemsets efficiently. These algorithms work very well in few items with many transactions. More algorithms have been developed in frequent itemset mining for high dimensional datasets. The dataset is considered as high dimensional when the number of items are higher than number of samples. Itemset enumerating algorithms would be inefficient for these data type. The later algorithm scheme is enumerating transaction sets. These algorithms construct transposed table and implement intersection of items by enumerating sets of transactions. The algorithms are CARPENTER, RERII, MaxConf and ISTA [9, 10, 20, 21].

Regardless the algorithm schemes applied in finding frequent patterns, if we do not limit or put a good threshold, the produced frequent patterns and the number of association rules would be very high. So, the searching of FPs can be restricted to closed or maximal frequent patterns only [22]. A closed frequent pattern is if there is an itemset and the item set does not have a superset with same support. Meanwhile, a maximal frequent pattern is when the itemset is frequent and there does not exist any superset that is frequent. The frequent closed patterns are useful in removing some redundant association rules, because they are subset of frequent patterns. Meanwhile maximal frequent patterns are the smallest representation of frequent patterns, but they do not contain the support of their subsets and we might be losing some of the information. As the consequences, frequent closed patterns are the most popular form of compressing the frequent patterns.

In relation to genetic studies, FIM algorithms are useful to associate features/genes and CARPENTER, RERII and ISTA were tested on microarray datasets. Many of the genes and gene polymorphisms implicated in complications to date are linked to protein and trait changes, and hence may form a common genetic group that influences the development of disease complications. In order to allow scientists to target the complications more effectively, future research needs to not only identify new genetic variants that contribute to each particular condition, but also look at the association that may occur in the multiple in the known variants. Because of that, we implemented FIM to discover the association in the multiple variants.

## 4. Experimental Design

### 4.1 Source of data

The data source for our experiment is GWAS datasets from WTCCC generated by Affymetrix GeneChop 500K Mapping Array Set for Type 2 Diabetes (T2D). T2D is reported to have number of increasing cases in many countries. T2D is a metabolic disorder and it is developed when the pancreas has problem in producing sufficient insulin. Previous studies have suggested various putative T2D susceptibility SNP variants in various genes including TCF7L2, PPARG, KCNJ11, CDKN2A/B, FTO, CDKAL1 and so on [15, 23, 24]. To perform experiments for FIM, this paper focuses on Chromosome 11 and 16 only, and the related genes are KCNJ11 and FTO only.

### 4.2 Data preprocessing

The SNPs datasets are treated as categorical data, and each SNP has one of three possible values. An example is given in table 1. Let $X = \{x_1, x_2, \ldots, x_m\}$ be a set of samples and $I = \{i_1, i_2, \ldots, i_k\}$ be a set of *k* elements called items in the sample. The set of individuals, X, from GWAS is stored in a table T defined by a set of "SNP" as items, $i_1, i_2, \ldots, i_k$. Each item $i_k$, is associated with a defined genotype. Genotypes of each attribute are categorical such that no genetic model assumptions are incorporated.

**Table 1.** Samples and items representation.

| **Individual** | **$i_1$ is SNP 1 =**{AA, AC, CC} | **$i_2$ is SNP 2 =**{CC, CT, TT} | **….** | **$i_k$ is SNP k =** {AA, AG, GG} |
|---|---|---|---|---|
| $x_1$ | $i_1$ = {AA, AC or CC} | $i_2$ = {CC, CT or TT} | …. | $i_k$ = {AA, AG or GG} |
| $x_2$ | $i_1$ = {AA, AC or CC} | $i_2$ = {CC, CT or TT} | …. | $i_k$ = {AA, AG or GG} |
| . | | | | |
| . | | | | |
| . | | | | |
| $x_m$ | $i_1$ = {AA, AC or CC} | $i_2$ = {CC, CT or TT} | …. | $i_k$ = {AA, AG or GG} |

Table 1 shows sample and items representation for GWAS dataset and the dataset has *k* number of SNPs. As an example, consider SNP 1, as an item, and it has two alleles "A" and "C" and the possible genotypes of SNP 1 are "AA", "AC" and "CC", which "AA" and "CC" are homozygous and "AC" is heterogonous genotype. Every individual, X, has exactly the same set of SNPs $\{i_1, i_2, \ldots, i_k\}$ with one of possible genotype values. We represent the individuals and SNPs in a transactional table, as shown in table 2 for example. Table 2 shows a sample of dataset with five SNPs (items) and six individuals (samples).

The dataset for FIM experiments consists of 1000 SNPs and 515 individuals for each Chromosome 11 and Chromosome 16. A proper feature selection based on filtering method was applied and the measure used is information gain for ranking only 1000 SNPs. Meanwhile, a random selection tool was used to reduce the number of samples to 515 only out of 1999.

**Table 2.** Individuals with five SNPs.

| Individual | Itemset |
|---|---|
| $x_1$ | $i_{1=AC}, i_{2=GT}, i_{3=CT}, i_{4=AA}, i_{5=CC}$ |
| $x_2$ | $i_{1=AC}, i_{2=GT}, i_{3=TT}, i_{4=AG}, i_{5=CT}$ |
| $x_3$ | $i_{1=AC}, i_{2=GT}, i_{3=TT}, i_{4=AG}, i_{5=CC}$ |
| $x_4$ | $i_{1=AC}, i_{2=GG}, i_{3=CT}, i_{4=AA}, i_{5=CT}$ |
| $x_5$ | $i_{1=AA}, i_{2=GT}, i_{3=CC}, i_{4=AG}, i_{5=CC}$ |
| $x_6$ | $i_{1=AA}, i_{2=GG}, i_{3=CT}, i_{4=AA}, i_{5=CT}$ |

### 4.3 Identified SNP variants

The information of SNPs and genes that reported with the risk of T2D is collected and it is used for further result analysis. The documented gene relationships are also gathered from Online Mendelian inheritance in Man (OMIM). The OMIM is an international standard to annotate genes in molecular function [25]. Based on the literatures, several studies have been conducted on WTCCC datasets in T2D and the identified SNPs in Chromosome 11 and 16, are shown in table 3.

**Table 3.** SNP variants with T2D risks.

| dbSNP | Ch. no | Gene | Reference |
|---|---|---|---|
| rs5215(similar to rs5219) | 11 | KCNJ11 | Scott et al. [24] |
| rs8050136 | 16 | Fat Mass and Obesity-associated Gene (FTO) | Zeggini et al. [23]<br>Scott et al. [24] |
| rs9939609 | 16 | Fat Mass and Obesity-associated Gene (FTO) | Wellcome, 2007 [15] |

### 4.4 Mining the Frequent closed itemsets

All experiments were performed on a computer with Intel CORE i7, 2.0 GHz CPU, 8GB RAM running on Windows 7 Home. This paper aims to investigate the generated itemsets using sample enumeration algorithm. We applied ISTA algorithm by Borgelt et al. [10]. We monitored the numbers of generated itemsets and analyzed the generated frequent patterns according to identified SNPs.

## 5. Result and Discussion

We collected the number of FPs mined in both datasets, Chromosome 11 and Chromosome 16, for T2D group by varying minimum support (*min_sup*) values.

### 5.1 Generated itemsets

After running experiments on different support values, we found that as the *min_sup* value increases, the number of frequent closed itemsets decreases and this situation is quite common in itemset mining. Table 4 shows the generated itemsets for two support values, which are 40 and 80, in both chromosomes. When the support value is 80, the total of itemsets generated is smaller than when the support value is 40. It happened to both chromosomes. The identified SNPs, as in table 2, are found when *min_sup* is set as 40 percent and lower. No identified SNPs found in the frequent closed itemsets if the support value is more than 40. For Chromosome 11, 2740886 is generated when support value is 40 with 1686 itemsets with rs5215. Meanwhile for Chromosome 16, 1652408 itemsets generated and 6369 (0.39%) itemsets are found with identified SNPs, rs8050136 and rs9939609. It shows that the identified SNPs could be occur in small portion of FPs but in low support values.

**Table 4.** Support Values and Generated Itemsets for Chromosome 11 and Chromosome 16

| Chromosome | Support value | Total of generated itemsets | Total of generated itemsets with identified SNPs |
|---|---|---|---|
| 11 | 40 | 2740886 | 1686 |
| 11 | 80 | 76814 | none |
| 16 | 40 | 1652408 | 6369 |
| 16 | 80 | 101451 | none |

Figure 1 shows the *k*-itemsets for identified SNPs and the highest value of *k* for Chromosome 11 is when *k* = 24, with 314 itemsets. The shortest combination of SNPs has 16 items and the longest itemsets is 35. Meanwhile Figure 2, is for Chromosome 16. The highest total of itemsets, is when *k* = 30 with 1062 itemsets. The shortest combination is 24 items and the longest combination is 40 items which occurs once. Mostly the *k*-itemsets were concentrated within *k*= 21 and *k*= 25 in Chromosome 11. Meanwhile for Chromosome 16, the concentration of itemsets can be found when *k*=28 and *k*=34.

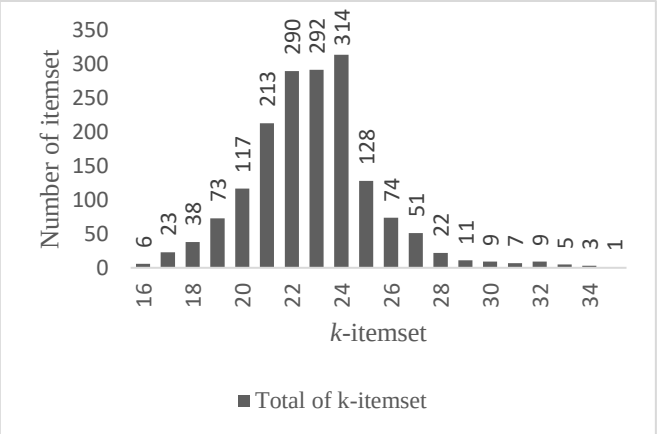


**Figure 1.** *k*-itemsets for Chromosome 11

In Chromosome 11, the itemsets were found more in shorter *k*-itemsets. Meanwhile for Chromosome 16, the itemsets were found more in longer *k*-itemset.

### 5.2 Genotype of the identified SNPs

We further studied the genotype value for the identified SNPs in the highest itemsets total, 24-itemset for Chromosome 11 and 30-itemset for Chromosome 16. For Chromosome 11, the genotypes for the identified SNP = rs5215 are CC, TT and CT, and mostly CT is occurred, which is 286 out of 314. Meanwhile for Chromosome 16, the identified SNP for rs9939609 is with genotype; AT, AA or TT, and 100% from 1062 itemsets are having AT. Next, the identified rs8050136,

has AC, AA or CC, 100% of the itemsets are AC. Figure 3 shows the percent value of the itemsets with identified SNPs and genotype value. Apparently, SNP feature "rs9939609 = AT" is always found in pairs with SNP feature "rs8050136 = AC", and it affects the *k* value of the itemsets and produce longer meaningful itemsets. It shows the variants in gene *FTO*, could be highly interrelated and this piece of information could be an important knowledge and a complementary to the identified SNPs found in the previous studies [15, 23, 24].

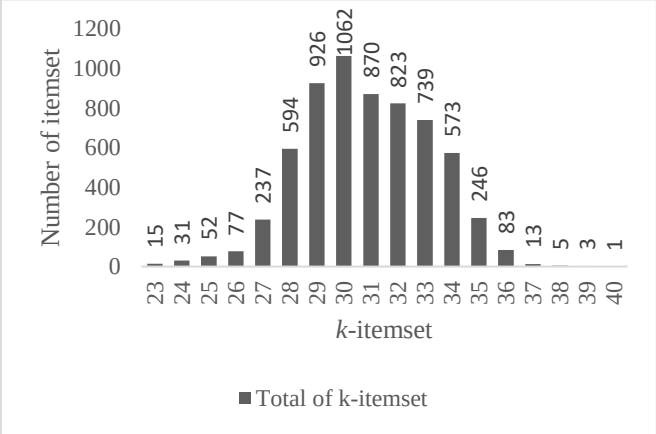


**Figure 2.** *k*-itemsets for Chromosome 16

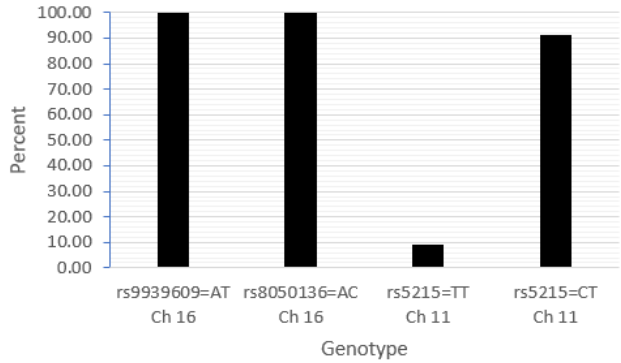


**Figure 3.** Identified SNPs and the Genotype Values

## 6. Conclusion

Complex diseases have their own complexity due to interaction among genetic factors and, also with environmental factors. In this paper we applied one of intersecting algorithms, ISTA, in mining frequent patterns from SNPs dataset, that represents genetic factors in disease. We had implemented two chromosomes with selected SNPs. Frequent closed itemsets with identified SNPs are normally found in low minimum support (*min_sup*) and in this study, support value is 40 or less. Next, we analyzed the itemsets found for the combination of the SNPs and we also studied the genotype value. We found that rs9939609 and rs8050136 are frequently occurring together. The study works on sets of high dimensional data is significant in discovering hidden knowledge. Through the frequent itemset mining in biological data, the results show interesting combinations and meaningful for knowledge discovery. We are optimist with FIM algorithm capability specifically in SNPs dataset and further development and enhancement can contribute in healthcare advancement system.

## Acknowledgement

The authors would like to thank to Research Management Centre, Universiti Teknologi MARA, Shah Alam, Malaysia for the research support and Ministry of Higher Education, Malaysia for the research grant: FRGS 156/2013.

## References

[1] B. R. Korf, *Human Genetics and Genomics*, Third ed.: Blackwell Publishing, 2007.

[2] M. I. McCarthy, J. N. Hirschhorn, "Genome-wide association studies: past, present and future", *Human Molecular Genetics*, vol. 17, no. R2, pp. R100-R101, 2008.

[3] S. Szymczak, J. M. Biernacka, H. J. Cordell, O. González-Recio, I. R. König, H. Zhang, Y. V. Sun, "Machine learning in genome-wide association studies", *Genetic Epidemiology*, vol. 33, no. S1, pp. S51-S57, 2009.

[4] H. J. Cordell, "Detecting gene-gene interactions that underlie human diseases", *Nat Rev Genet*, vol. 10, no. 6 pp. 392-404, 2009.

[5] G. Atluri, R. Gupta, G. Fang, G. Pandey, M. Steinbach, V. Kumar, "Association Analysis Techniques for Bioinformatics Problems", in *Bioinformatics and Computational Biology*, vol. 5462, S. Rajasekaran, Ed., ed: Springer Berlin / Heidelberg, pp. 1-13, 2009.

[6] J. H. Moore, F. W. Asselbergs, S. M. Williams, "Bioinformatics challenges for genome-wide association studies", *Bioinformatics*, vol. 26, no. 4, pp. 445-455, 2010.

[7] D. B. Kell, "Genotype–phenotype mapping: genes as computer programs", *TRENDS in Genetics*, vol. 18, no. 11, pp. 555-559, 2002.

[8] V. Aguiar, J. A. Seoane, A. Freire, L. Guo, "GA-Based Data Mining Applied to Genetic Data for the Diagnosis of Complex Diseases", *Soft Computing Methods for Practical Environment Solutions: Techniques and Studies.*, pp. 219-239, 2010.

[9] F. Pan, G. Cong, A. K. H. Tung, J. Yang, M. J. Zaki, "Carpenter: finding closed patterns in long biological datasets", *In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, Washington D.C., pp. 637-642, 2003.

[10] C. Borgelt, X. Yang, R. Nogales-Cadenas, P. Carmona-Saez, A. Pascual-Montano, "Finding closed frequent item sets by intersecting transactions", *In Proceedings of the 14th International Conference on Extending Database Technology*, Uppsala, Sweden, pp. 367-376, 2011.

[11] M. R. Barnes, "SNP and Mutation Data on the Web – Hidden Treasures for Uncovering", *Comp Funct Genomics.*, vol. 3, no. 1, pp. 67–74, 2002.

[12] T.-H. Wang, H.-S. Wang, "A Genome-Wide Association Study Primer for Clinicians", *Taiwanese Journal of*

*Obstetrics and Gynecology*, vol. 48, no. 2, pp. 89-95, 2009.

[13] A. DeWan, M. Liu, S. Hartman, S. S.-M. Zhang, D. T. L. Liu, C. Zhao, P. O. S. Tam, W. M. Chan, D. S. C. Lam, M. Snyder, C. Barnstable, C. P. Pang, J. Hoh, "HTRA1 Promoter Polymorphism in Wet Age-Related Macular Degeneration", *Science*, vol. 314, no. 5801, pp. 989-992, 2006.

[14] R. J. Klein, C. Zeiss, E. Y. Chew, J.-Y. Tsai, R. S. Sackler, C. Haynes, A. K. Henning, J. P. SanGiovanni, S. M. Mane, S. T. Mayne, M. B. Bracken, F. L. Ferris, J. Ott, C. Barnstable, J. Hoh, "Complement Factor H Polymorphism in Age-Related Macular Degeneration", *Science*, vol. 308, no. 5720, pp. 385-389, 2005.

[15] WTCCC, "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls", *Nature*, vol. 447, no. 7145, pp. 661-678, 2007.

[16] A. Kelemen, A. V. Vasilakos, L. Yulan, "Computational Intelligence in Bioinformatics: SNP/Haplotype Data in Genetic Association Study for Common Diseases", *IEEE Transactions on Information Technology in Biomedicine*, vol. 13, no. 5, pp. 841-847, 2009.

[17] J. N. Hirschhorn, M. J. Daly, "Genome-wide association studies for common diseases and complex traits", Nature Reviews Genetics , vol. 6, no. 2, pp. 95-108, 2005.

[18] J. Han, H. Cheng, D. Xin, X. Yan, "Frequent pattern mining: current status and future directions", *Data Mining and Knowledge Discovery*, vol. 15, no. 1, pp. 55-86, 2007.

[19] R. Agrawal, T. Imielinski, A. Swani, "Mining Association Rules between Sets of Items in Large Databases", in *ACM SIGMOD Conference*, pp. 207-216, 1993.

[20] G. Cong, K.-L. Tan, A. K. H. Tung, F. Pan, "Mining Frequent Closed Patterns in Microarray Data", *In Proceedings of the Fourth IEEE International Conference on Data Mining*, pp. 363-366, 2004.

[21] T. McIntosh, S. Chawla, "On discovery of maximal confident rules without support pruning in microarray data", *In Proceedings of the 5th international workshop on Bioinformatics*, Chicago, Illinois, pp. 37-45, 2005.

[22] C. Borgelt, "Frequent item set mining", *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 2, no. 6, pp. 437-456, 2012.

[23] E. Zeggini, M. Weedon, C. Lindgren, T. Frayling, K. Elliott, H. Lango, N. Timpson, J. Perry, N. Rayner, R. Freathy, "Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes", *Science*, vol. 316, no. 5829, pp. 1336-1341, 2007.

[24] L. J. Scott, K. L. Mohlke, L. L. Bonnycastle, C. J. Willer, Y. Li, W. L. Duren, M. R. Erdos, H. Stringham, P. Chines, A. Jackson, "A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants", *Science*, vol. 316, no. 5829, pp. 1341-1345, 2007.

[25] V. McKusick, "Mendelian Inheritance in Man and Its Online Version, OMIM", *American Journal of Human Genetics*, vol. 80, no. 4, pp. 588-604, 2007.