MINING SIGNIFICANT RESISTANT SINGLE NUCLEOTIDE POLYMORPHISM SUBSET USING IMMUNE ANT SWARM OPTIMIZATION FOR ROUGH - APRIORI TECHNIQUE

FATIN NABILAH BINTI ABDUL RAMAN

FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

ACKNOWLEDGEMENT

First and foremost, I offer my sincerity gratitude to my Almighty God for blessing me to complete this project. Then, I would like to thanks and express my sincere appreciation to my supervisor, Dr Choo Yun Huoy, for her guidance and support to this project. Special thanks to her for all the comment and guideline she give to me.

Besides, I would like to thanks my friend and colleagues that give support, opinions that related to my project. Not forgotten to my senior, LustianaPratiwi, I would like to express my sincere appreciation to her for giving guidance in this project.

Last not least, I would like to thanks to my family who give continuous support and encouragement for me to complete this project. Lastly, I would like to express my sincere for those that who has contributes during my project, I appreciate all your kindness and cooperation.

ABSTRACT

Immune Ant Swarm Optimization using Rough Reducts(IASORR) technique is the hybridization of Particle Swarm Optimization (PSO) and Ant Colony Optimization (ACO) using immunity. It is use a rough reducts calculation to identifying an optimal significant obesity attributes set along with immunity to discover a better fitness value in optimizing rough reducts set. Meanwhile, Apriori will also been applied in this project. It is used to find the frequent itemset among the given number of transactions or number of dataset. This paper will be evaluated and applied the mention techniquefor mining the obesity resistant single nucleotide polymorphism (SNPs) subset.

ABSTRAK

Immune Ant Swarm Optimization using Rough Reducts (IASORR) adalah gabungan teknik *Particle Swarm Optimization (PSO)* dan *Ant Colony Optimization (ACO)* bersama sistem *Immunity*. Ia mengguna kiraan teknik *rough reducts* untuk mengenalpasti set *attribute* obesity yang optimum dengan menggunakan system *Immunity* untuk mencari *fitness value* yang lebih baik dalam mengoptimumkan set *rough reducts*. Malah, teknik *Apriori* turut akan digunakan di dalam projek ini kerana ia digunakan untuk mencari item set yang kerap kali muncul anatara bilangan *transaction* atau jumlah dataset yang ada. Selepas itu, kertas projek ini akan dinilai dan teknik yang disebut akan diterapkan untuk *mining the obesity resistant single nucleotide polymorphismset* dalam projek ini.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURE

CHAPTER 1

INTRODUCTION

1.1 Introduction

"Obesity is a medical condition for body in which the excess body fat bas accumulated to the extent that it has an adverse effect on health. It has reached epidemic proportion globally with more than 1 billion adult overweight and is a major contributor to the global burden of chronic disease and disability" in Journal on Obesity And Overweight by World Health Organization. Beside, obesity increases the likelihood of various diseases such as heart disease, type 2 diabetes which is the most common disease and etc. It is also most commonly caused by a combination of excessive food energy intake and lack of physical activity. Besides that, it not only involving the adult but at the same time it also occur among the children. Thus, obesity statistic in Malaysia is getting scarier day by day.

Meanwhile, in 2010 The World Health Organization (WHO) ranked Malaysia as sixth place in Asia with the highest adult obesity rate which showed that 60% of

Malaysian aged between 18 and up, had the BMI over 25 which is means the person is overweight(as shown in Figure 1). Now, Malaysia has been rates as the highest among Asian countries for obesity. As Science Advisor to the Prime Minister, Tan Sri Zakri Abdul Hamid said new findings from British medical journal, The Lancet, it's showed that 49% of women and 44% of men were found to be obese as stated in The Star Online dated June 16, 2014. The statistics have indeed reached alarming proportions and effective measures should be taken immediately to arrest the obesity issue amongst the Malaysia population.

Figure 1:Overweight Populations in Southeast Asia

Meanwhile, some researcher and current knowledge have concluded that *"the genetic factors may be involved in the etiologic of obesity and exclusive on very rare of obesity cases where the gene involved was probably genes that interact with environment factors that is related to the energy and expenditure to the increasing of obesity risk"* in Etiology-Genetic Factors Of Overweight And Obesity. Besides, as we know a single human genome contains huge amount of data and it carries a lot of unnecessary information. Thus with applying a suitable feature selection technique, the significant resistant single nucleotide polymorphism subset in obesity dataset can be identify which will be useful for the future purpose.

Future selection is the most important part in collecting the important knowledge especially in large dataset likely in this research data. Feature selection also can be known as variable section or attribute selection which is involving a process of selecting or retrieving a relevant feature. Thus, the goal of feature selection is to reduce

the number of attribute or feature to be processed without sacrificing class discrimination therefore classification accuracy.

This project will be running in Eclipse platform as Weka 3.6 and below along java run time (JDK). Meanwhile, in this project, Immune Ant Swarm Optimization for Rough Reducts (IASORR) technique is applied. It is to discover a better fitness value in optimizing rough reducts set which is using a rough reducts calculation in order to identify an optimum significant attributes set. This approach is expected to generate a better optimum rough reducts and achieved the significant single nucleotide polymorphism subset. Meanwhile, Apriori will apply to find the frequent item set among the given number of transactions or number of dataset.

1.2 Problem Statement

As we know from some of the research, there prove that obesity also can be the result of an interplay between genetic and environmental factors. While, there also an evidence that numerous study of laboratory show that genetics play an important role in obesity. Besides that, there were too many SNP in each of the gene which hard to recognize as an important specific SNP. Moreover, there are also no specific gene variant have been decided based on it occurrences.

1.3 Objective

1. To identify the suitable gene variants group for obesity resistant diagnosis.
2. To propose IASORR features selection technique and Apriori for mining obesity resistant single nucleotide polymorphism (SNPs) subset.
3. To evaluate the proposed IASORR and Apriori technique on benchmarking dataset.

1.4 Scope

This project will be applying and focusing in using Immune Ant Swarm Optimization for Rough Reducts (IASORR) Technique to analysing and identifying the suitable gene variant group for obesity resistant diagnosis for future used. Meanwhile, Apriori technique will be used if needed that will be hybrid the previous technique. The data will be save and load in Waikato Environment for Knowledge Analysis (WEKA) in .csv file. The obesity dataset will be predict using a full training set. The dataset will be evaluate all on all data for the validation to estimate its classification accuracy. This project will used 30 FTO single nucleotide polymorphism (SNPs).

1.5 Expected Output

In this project, the last output will be evaluate and analyse for future used in obesity resistant diagnosis. This applied approach or method is expected to be able to produce a better optimization technique and produce more accurate result in rough reducts population. Meanwhile, Apriori will applied to find the frequent item set among the given number of transactions or number of dataset.

1.6 Conclusion

As conclusion, IASORR and Apriori will be applied in association tool in Waikato Environment for Knowledge Analysis (WEKA) to identify the suitable gene variant group and find the frequent item in the dataset. This project also requires

obesity dataset to completing the proposed technique for mining obesity resistant single nucleotide polymorphism (SNPs) subset. Thus, to obtain the best obesity dataset, it require a long process of collecting, cleaning of data before used it.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

In this chapter, the topic will be discuss are about the technique, preliminary study on topic that related to the project, and literature review about the project. There will be more explanation on the applied technique; Immune Ant Swarm Optimization for Rough Reducts (IASORR) technique; Apriori technique, and other related bioinformation term; SNPs, allele, gene.

2.2 Biological Term

In this project, there is a lot biological term that is being used especially Obesity, SNPs, genome, allele, chromosome, genotype and FTO will be explain as below.

2.2.1 Obesity

Obesity is a term used to describe somebody whom in a condition of overweight with an excessive amount of fat. It increases the risk of diseases and health problems which will lead to a number of serious and potentially life-threatening condition like high blood pressure, coronary heart disease, type 2 diabetes. It also can affect the quality life and lead to psychological problems like depression. As we know, Malaysian obesity statistics are becoming scarier day by day. At least, 48 percent of Malaysians, of whom 15.2 percent are adults, are obese, based on a National Health and Morbidity Survey (NHMS) in 2011 obese as stated in Astro Awani, Bernama dated January 28, 2015.

2.2.2 Single Nucleotide Polymorphism (SNPs)

Based on Free Online Dictionary, Single Nucleotide Polymorphism (SNP) is a genetic variation in DNA sequence that occurs when a single nucleotide in genome is altered which SNP are usually considered to be a point mutations that have been evolutionarily successful enough to reoccur in a significant proportion of the population of a species.

If more than 1% of a population does not carry the same nucleotide at a specific position in the DNA sequence, then this variation can be classified as a SNP. If a SNP occurs within a gene, then the gene is described as having more than one allele. In

these cases, SNPs may lead to variations in the amino acid sequence. SNPs, however, are not just associated with genes; they can also occur in noncoding regions of DNA.

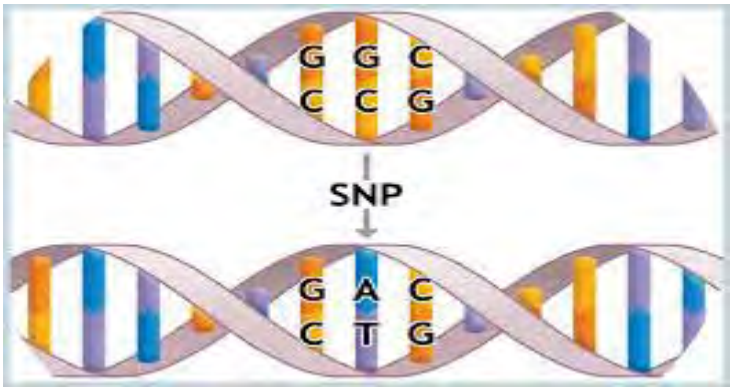

Figure 2: Single Nucleotides Polymorphism (SNP)

2.2.3 Genome

According to Ridley, M. (2006), *"a genome is an organism's complete set of DNA, including all of its genes. For your information, each genome contains all of the information needed to build and maintain that organism which contain a complete set of genetic instructions. In humans, a copy of the entire genome—more than 3 billion DNA base pairs—is contained in all cells that have a nucleus. The genome includes both the genes and the non-coding sequences of the DNA/RNA".*

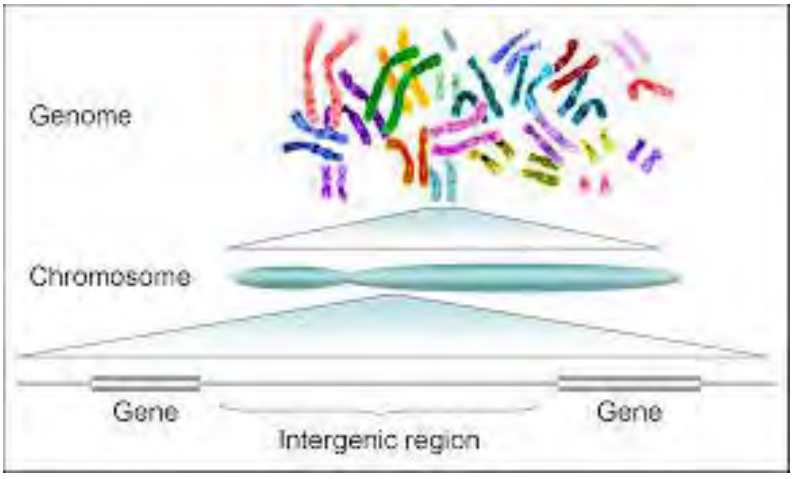

Figure 3: Genome Differentiation

Every cell in your body has a complete set of instructions about how to make your cells and their components, and to direct how these interact. This set of instructions is your genome. Your genome is quite similar to everyone else's genome, which is why we all turn out to be human beings. But every other living thing also has a genome. You could think of the genome as a recipe book that carries all of the instructions to make a human being.

We actually have two genomes each, one from our mother and another one from our father. Thus, it make a fertilised egg containing two genomes with a new set of instructions to make a new person.

2.2.4 Allele / Associate Allele

An allele is an alternative form of a gene (one member of a pair) which it located at a specific position or region on a specific chromosome. These DNA codes will determine the distinct traits that can be passed on or transmitted from parents to offspring through sexual reproduction or known as Mendel's law of segregation.

Allele is any one of two or more genes that may occur alternatively at a given locus on chromosome. Sometimes, different alleles can result in different observable phenotypic traits, such as different pigmentation. However, most genetic variations result in little or no observable variation.

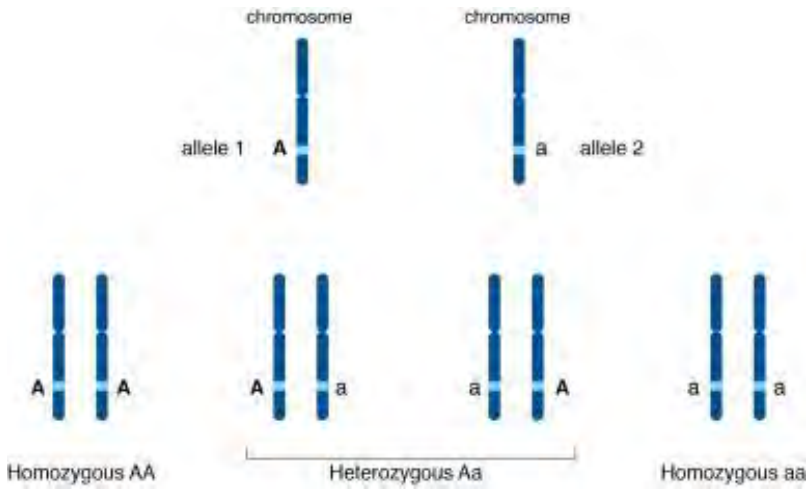

Figure 4:Allele

Most multicellular organisms have two sets of chromosomes or known as diploid. These chromosomes are referred to as homologous chromosomes. Diploid organisms have one copy of each gene or allele on each chromosome. If both alleles are the same, the organism are homozygous which carried the two identical copies of the gene. If the alleles are different, the organism are heterozygous which carried the two different copies of gene.

2.2.5 Chromosome

Chromosomes are the basic building blocks of life where the entire genome of an organism is essentially organized and stored in the form of DNA (deoxyribonucleic acid) which is present inside every cell of organism. In each cell have its own nucleus where the DNA molecule was packaged into a thread-like structures. It is tightly coiled many times around proteins called histones that support its structure

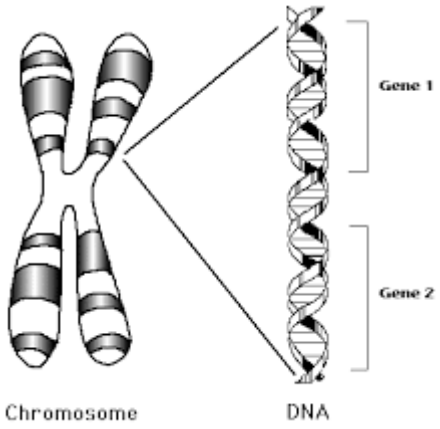

Figure 5: Chromosome

At first, chromosomes are not visible in the cell's nucleus, not even under a microscope, it is because the cell is not dividing. However, the DNA molecule becomes more tightly packed during cell division and it is become visible under a microscope. Most of what researchers know about chromosomes was learned by observing chromosomes during cell division.

Each chromosome has a constriction point called the centromere, which divides the chromosome into two sections, or "arms"; short arm and long arm. The

short arm of the chromosome was labelled the “p arm.” The long arm of the chromosome was labelled the “q arm.” The location of the centromere on each chromosome will gives the chromosome its characteristic shape, and help to describe the location of specific genes.

2.2.6 FTO

As stated by Jia G et al. (2011), *“fat mass and obesity-associated protein also known as alpha-ketoglutarate-dependent dioxygenase FTO is an enzyme that in humans which is encoded by the FTO gene where it is located on chromosome 16. As one homolog in the AlkB family proteins, it is the first mRNA demethylase that has been identified”.* While Loos RJ, Yeo GS (2014) states that *“certain variants of the FTO gene appear to be correlated with obesity in humans”.*

Figure 6: FTO structure

Some researcher state that *“the FTO gene expression was also found to be up regulated in the hypothalamus of rats after food deprivation and strongly negatively correlated to the stimulation of food intake”.* Thus, the increasing of hypothalamic expression for FTO that will affect the regulation of energy intake but not for feeding reward.

2.2.7 Genotype and Phenotype

Genotype is one of a complete heritable genetic identity that would be revealed by personal genome sequencing. It also refer to a particular gene that will carried by an individual. While, phenotype is a description of the actual physical characteristic. It also refer to straightforward visible characteristics like eye colour. Most of phenotype are influenced by both genotype and unique circumstances in the environment.

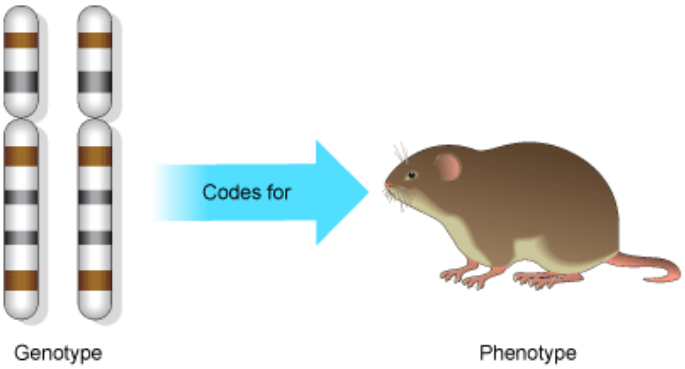

Figure 7: Genotype and Phenotype

2.3 Fact and Finding

This fact and finding is consist of two part. The first part is feature selection where Immune Ant Swarm Optimization for Rough Reducts (IASORR) while in part two is association rule where Apriori technique part.

2.3.1 Feature Selection

"Feature selection is the process of selecting a subset of relevant features for use in model construction" – Feature Selection, Wikipedia entry. It also can be simplified as a process which it will automatically search for the best subset of attribute

in the dataset. Besides, features selection has been an active and fruitful field of research and development for decades in machine learning (Liu et al.,2002b;Robnik-Sikonja and Kononenko,2003) and data mining (Kim et al.,2000;Dash et al.,2002). According to Almuallim and Dietterich (1994), Koller and Sahami (1996) and Blum and Langley (1997) states that *"it has proven in both theory and practice effective in enhancing learning efficiency, increase predictive accuracy and reducing complexity of learned result"*.

Meanwhile, Yang and Pederson (1997) and Xing et al. (2001) in presence of thousands or hundreds of features the researchers notice it is common that a large number of feature is not informative because the features is either irrelevant or redundant. Besides that, learning can also be achieved effectively with a relevant and non-redundant features. Meanwhile, Kohavi and John (1997) had mention that the number of possible feature subsets grows exponentially with the increases of dimensionality.

There exist about two major approaches in features selection; one is individual evaluation and other one is subset evaluation. According to Blum and Langley (1997) and Guyon and Elisseeff (2003) , individual evaluation which also known as feature weighting/ranking where it will assess individual features and assigns them weights according to degree of relevant.

2.3.2 Immune Ant Swarm Optimization for Rough Reducts (IASORR)

According to Lustiana et al. (2011), Immune Ant Swarm Optimization for Rough Reducts (IASORR) is basically a *"hybridization of Particle Swarm Optimization (PSO) and Ant Colony Optimization with immunity to enhance optimization performance. It is rough reduct calculation for identifying an optimum significant attributes set"*. The detailed explanation on this technique will be cover in Chapter 5.

PSO is a population-based stochastic search algorithm which is proposed by Kennedy and Eberhart (1999) and wisely used to solve a broad range of optimization