3

COMPARISON OF MACHINE LEARNING TECHNIQUES FOR PREDICTING BIOLOGICAL ACTIVITIES OF VOLATILE METABOLITES BASED ON MOLECULAR FINGERPRINTS

Azian Azamimi Abdullah and Khoo Shu Fen

1. INTRODUCTION

Metabolomics is the study of the chemical process which involving small molecule. Those small molecules are the intermediate and result of metabolism which known as metabolites. Typically, there are two groups of metabolite, which are primary metabolite and secondary metabolite. Primary metabolites are essential for supporting life where it directs involved in normal growth and development of the living organism. On the other hand, secondary metabolites are responsible for the survival of the organism in the environment. During the second metabolism, there are some Volatile Organic Compounds (VOCs) or volatile metabolites being produced [1]. VOCs is an organic compound which has low molecular weight in the range of 50-200 Daltons, low boiling point and high vapour pressure [2]. They can serve as the signaling molecules that pass the information between organisms contributing to these characteristics. Along with carbon, they contain another elements as well such as nitrogen, oxygen, hydrogen and etc.

There is a broad diversity of VOCs which including alkanes, alkenes, alcohols, terpenes and so on. VOCs that naturally produced are being documented to plays an important role in human and plants. For instance, the ecological functions that VOCs involved including defense against herbivores, pollinator attraction, plant-plant communication and plant-plant interaction [3]. Other than that, VOCs sampling from human breath, urine

or sweat can function as the biomarker, which can provide an efficient tool for early detection of diseases. Changes of VOCs concentration in the early stage of disease enable it to serve as the biomarker for early detection of disease before the condition become worse. This is very crucial, especially in treatment of cancer, infectious and inflammatory disease [4], [5]. In conclusion, precisely determine the roles of VOCs definitely can provide a better understanding of living systems.

2. PROBLEM STATEMENTS

The advanced analytical technologies such as Mass Spectrometry (MS) and Nuclear Magnetic Resonance (NMR) in metabolomics study had generated a large amount of data and it is common to detect the unknown compounds. Lack of reference MS and NMR spectra for the metabolite compounds become a limitation for the researchers to further examine the biological activities of that particular metabolite compounds [6]. Identification biological activities of VOCs is important for instance researchers able to develop a new suitable method for pest and environment control through predicting the biological activities of VOCs regarding their response to biotic and abiotic stress. By determine the biological activities of VOCs, their potential to serve as disease biomarker had greatly increased. Analysis of those VOCs able to provide an insight into healthy and diseased metabolic state. Despite the identification of biological activities of VOCs able to contribute in term of health care and agriculture, the study regarding the relationship between volatile metabolite and their biological activities is still not common. Thus, this research aims to predict the biological activities of VOCs by using molecular fingerprints and machine learning methods.

3. MOLECULAR FINGERPRINTS

The Structure Data Format (SDF) file of the collected VOCs which contains all the information of those particular compounds are being downloaded from PubChem website. The SDF file is then being imported into ChemDes