

Structural clustering of plant secondary metabolite to estimate compound classes reflecting their biosynthetic pathway

Toshiaki Tokimatsu¹
tokimatu@kuicr.kyoto-u.ac.jp

Masaaki Kotera¹
kot@kuicr.kyoto-u.ac.jp

Susumu Goto¹ **Shigehiko Kanaya**² **Minoru Kanehisa**^{1,3}
goto@kuicr.kyoto-u.ac.jp skanaya@gtc.naist.jp kanehisa@kuicr.kyoto-u.ac.jp

¹Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji, Kyoto, 611-0011, Japan

²Graduate School of Information Science, Nara Institute of Science and Technology, Ikoma, Nara, 630-0192, Japan.

³Human Genome Center, Institute of Medical Science, University of Tokyo, Tokyo, 108-8639, Japan.

Keywords: secondary metabolites, compound classification, structural clustering, plant metabolites

1 Introduction

Plants produce over 200,000 metabolites, which are important source of drugs and industrial materials. These plant secondary metabolites can be divided into groups that share the same core substructure, originated from the same biosynthetic pathways. The aim of our research is to understand previously unknown plant secondary metabolic pathway including those metabolites and catalyzing enzymes. In order to overview the secondary metabolic pathway, we have been developing a structural clustering method of metabolites based on fragmented substructures. Here, we present the detailed survey of the resulting metabolite groups. We show the effectiveness of the fragmented fingerprint methods, and examine the strategy of cutting bonds and removing glycoside modifications.

2 Data and Method

Structure data are obtained from the KEGG[1] and KNApSAcK[2] databases. The total 52,950 structures are converted into the KEGG Chemical Function (KCF) format, in which every atomic element (except hydrogen) is discriminated by functional group information (KEGG Atom Type).

Similarity measures between two compounds are defined by fragmented KEGG Atom fingerprints. A compound is represented as a KEGG molecular graph. Multiple fragments (subgraphs) are generated from a compound by cutting every one or two edges of the graph. Every fragment is represented as a vector consisting of the numbers of KEGG atoms. Similarity score of two compounds is defined as the number of the KEGG atoms in the largest fragment shared with both compounds. Following two cutting pattern were tested.

Method 1: cut all types of bonds.

All chemical bonds (except those including hydrogen) are cut to generate fragment vectors.

Method 2: cut only hetero bonds, followed by removing fragments containing glycosides.

Only hetero bonds are cut when generating fragment vectors. We did this hoping to reduce the computational amount. After cutting every one or two hetero bonds, fragment vectors containing acetal or ketal group(s) are considered as glycosides, and are removed in prior to the clustering.

Compounds are subjected to a hierarchical clustering based on the similarity scores defined in Methods 1 or 2. We used the same hierarchical clustering method that was applied to calculate KEGG OC

