International Journal of Innovative
Computing, Information and Control
Volume **7**, Number **10**, October **2011**

ICIC International ©2011 ISSN 1349-4198
pp. 6041–6053

DISCOVERING INDIRECT GENE ASSOCIATIONS BY FILTERING-BASED INDIRECT ASSOCIATION RULE MINING

Yu-Cheng Liu[1], J. W. Shin[2] and Vincent S. Tseng[1,*]

[1]Department of Computer Science and Information Engineering
[2]Department of Parasitology
National Cheng Kung University
No. 1, University Road, Tainan City 701, Taiwan
uchenliu@gmail.com; hippo@mail.ncku.edu.tw
[*]Corresponding author: tsengsm@mail.ncku.edu.tw

Received April 2010; revised August 2010

Abstract. *Data mining is a popular technology used for microarray analysis. Using this technique, biologists can effectively elucidate gene expression data. In this research, we propose the FIARM (Filtering-Based Indirect Association Rule Mining) algorithm to analyze gene microarray data. The form $\langle X, Y|M\rangle$ is used to present the indirect relation of X and Y, which depends on M. This signifies that both gene X and gene M are likely involved in a given biological activity. Furthermore, both gene Y and gene M likely join together to carry out another biological activity. As gene M is the necessary factor in these different biological activities, it can help biologists determine gene relationships in diverse activities. We use semantic similarity of Gene Ontology to verify the accuracy of discovered gene relations. Under experimental evaluation, the proposed method can discover the relationship dissimilated by association rules to effectively assist biologists in complicated genetic research.*

Keywords: Data mining, Microarray, Gene expression analysis, Indirect association rule

1. **Introduction.** Bioinformatics has become an important topic of computer science. Computer methods allow biologists to analyze and evaluate their ideas more easily [10]. Examples of such use include: multiple gene alignment [5], motif identification [4], microarray analysis [11], protein structure prediction [20] and pathway analysis [19]. Previously, biologists wanting to elucidate gene expression were forced to perform laboratory experiments; however, manipulation of multiple genes is very costly. Microarray technology, on the other hand, can elucidate many genes simultaneously making microarray of genes an important topic.

Initially, the analysis of gene expression by researchers was primarily focused on statistical methods [10]. However, the use of statistical methods is time-consuming when data are large. Data mining technologies [7,12,16] are useful to discover information from large datasets. Thus, growing numbers of researchers use data mining technologies to elucidate genetic information efficiently.

Past researches concerning data mining technologies generally used clustering [8,16,17], classification [3] and association rules [9,11] to analyze gene expression data. These methods allow biologists to quickly discover information of interest from large pools of data. Association rules are characteristically suited to recognizing the relationships between genes. Every item in such rules can state whether a given gene is expressed or repressed to narrate the expression relationship in a cellular environment. For example, the rule {cancer}⇒{gene X↑, gene Y↓, gene Z↑} can be mined from gene expression data. This