A Novel Evolutionary Algorithm for Bi-clustering of Gene Expression Data based on the Order Preserving Sub-Matrix (OPSM) Constraint

Hongchan Roh and Sanghyun Park

Abstract—**Biclustering is a popular method which can reveal unknown genetic pathways. However, even though many algorithms have been suggested, no overwhelming algorithm has been suggested, due to its significant search space, until now. In this respect, several evolutionary algorithms tried to address this problem utilizing the powerful search capability of Evolutionary Computation (EC). However, most algorithms focused on exploiting the Mean Square Residue (MSR) measure which was proposed by Cheng and Church. The Order Preserving Sub-Matrix (OPSM) constraint was rarely considered even though it promises more biologically relevant biclusters than the MSR measure. The goal of this paper is to design an EC algorithm which ensures biologically significant biclusters by using the OPSM constraint and better biclusters than the original OPSM algorithm. We designed a novel encoding method and evolutionary operators suitable for the OPSM constraint. To efficiently explore the search space, we modulized our evolutionary algorithm and applied the co-evolution concept. Through a set of experiments, it was confirmed that our algorithm outperformed a representative EC biclustering algorithm based on CC and the original OPSM algorithm.**

I. Introduction

Recently, biclustering methods have been vigorously researched to discover local patterns in gene expression data. Whereas traditional clustering techniques such as hierarchical clustering [1] and k-means clustering [2] requires clustered genes to behave similarly over all the experimental conditions, biclustering requires genes in the same cluster to behave similarly over a subset of the conditions of gene expression data. This specified clustering concept is useful to uncover genetic pathways that are activated only over some of the conditions. The problem of finding a minimum set of biclusters is a generalization of another problem such as covering a bipartite graph, which has been shown to be NP-hard [15].

Hartigan [3] first introduced the biclustering concept and later, in 2000, Cheng and Church [4] first used this concept in biclustering gene expression data. After Cheng and Church, various methods regarding biclustering gene expression have been suggested. Among them, the Order Preserving Sub-Matrix model (OPSM) [5] is a representative biclustering method concerning the discovery of one or several submatrices in a gene expression matrix in which the expression levels of the selected genes induce the same linear ordering of the selected conditions. Later, [6] extended OPSM by assigning similar expression levels equal ranks.

Evolutionary Computation (EC) performs well in addressing complex optimization problems. It has excellent exploration power that provides the capability of escaping from local optima and working well when solutions to a problem contain complex interacting parts [16]. In addition, EC has been applied for problem solving in various domains such as planning [17], design [18], scheduling [19]-[20], simulation and identification [21], control [22], and classification [23]-[26].

In this respect, EC is very suitable for searching biclusters, and thus it has been applied to several biclustering approaches, combined with the previous well known biclustering methods such as Cheng and Church (CC), and the OPSM.

Several EC-based biclustering algorithms [7]-[10], which utilize CC's mean squared residue (MSR) measure as a fitness function score, have been proposed. An EC-based biclustering algorithm [12] with its own measure which has many common aspects with Sequential Evolutionary Biclustering (SEBI) [10] has been proposed. An EC-based biclustering algorithm [11] which utilizes the OPSM constraint has been also suggested. However, in the true sense of the word, [11] is not a biclustering algorithm because it can operate with not a subset of conditions but all the conditions. It's specially designed for gene expression data whose conditions are time series. Among them, SEBI is a comprehensive algorithm for the other algorithms. It tried to find biclusters which have an MSR score lower than the user-defined threshold while reducing overlapped areas between biclusters, growing bicluster size, and maximizing row variance.

Most of the previous EC algorithms for biclustering have not provided biologically significant biclusters and also cannot fully utilize the search power of EC. Biclustering algorithms' performance depends on how much information about genetic pathways their results, such as biclusters, can provide. Therefore, previous non-EC-based biclustering algorithms are mostly evaluated by biological significance of biclusters they found. However, previous EC algorithms

Manuscript received July 5, 2008. This work was supported by the Korea Science and Engineering Foundation (KOSEF) grant funded by the Korea government (MOST) (No. R01-2006-000-11106-0).

H. Roh is with the Computer Science Department, Yonsei University, Seoul, Korea (corresponding author to provide phone: 82-2-2123-7757, fax: 82-2-365-2579, e-mail: fallsmal@cs.yonsei.ac.kr).

S. Park is with the Computer Science Department, Yonsei University, Seoul, Korea (e-mail: sanghyun@cs.yonsei.ac.kr).