Discovering Modules in Time-Series Gene Expression Data Using Biclustering

Sara C. Madeira[123] and Arlindo L. Oliveira[12]

[1] INESC-ID, Lisbon, Portugal
[2] Technical University of Lisbon, IST, Lisbon, Portugal
[3] University of Beira Interior, Covilhã, Portugal

Abstract. Several non-supervised machine learning methods have been used in the analysis of gene expression data. Recently, biclustering, a non-supervised approach that performs simultaneous clustering on the row and column dimensions of the data matrix, has been shown to be remarkably effective in a variety of applications. The advantages of biclustering (when compared to clustering) in the discovery of local expression patterns has been extensively studied and documented [1]. These expression patterns can be used to identify relevant biological processes possibly involved in regulatory mechanisms. Although, in its general form, biclustering is NP-complete, in the case of time-series expression data the interesting biclusters can be restricted to those with contiguous columns leading to a tractable problem.

In this context, we have recently proposed CCC-Biclustering [2], an algorithm that finds and reports all maximal contiguous column coherent biclusters (CCC-biclusters) in time linear on the size of the expression matrix by processing a discretized matrix using string processing techniques based on suffix trees. Each expression pattern shared by a group of genes in a contiguous subset of time points is a potentially relevant biological process (module). However, discretization may limit the ability of the algorithm to discover biologically relevant patterns due to the noise inherent to most Microarray experiments. To overcome this problem we present a new algorithm that finds CCC-biclusters with up to a given number of errors per gene in the expression pattern that identifies the CCC-bicluster. These errors can, in general, be substitutions of a symbol in the expression pattern by other symbols in the alphabet (identifying measurement errors), or restricted to the lexicographically closer discretization symbols (identifying discretization errors).

References

1.S. C. Madeira and A. L. Oliveira. Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1(1):24–45, 2004.

2.S. C. Madeira and A. L. Oliveira. A linear time algorithm for biclustering time series expression data. In *Proc. of 5th Workshop on Algorithms in Bioinformatics*, pages 39–52. Springer, LNCS/LNBI 3692, 2005.

Keywords

BICLUSTERING WITH ERRORS, TIME-SERIES EXPRESSION DATA, EXPRESSION PATTERNS, BIOLOGICAL PROCESSES, MODULES