

Clustering Gene Expression Patterns

AMIR BEN-DOR,¹ RON SHAMIR,² and ZOHAR YAKHINI³

ABSTRACT

Recent advances in biotechnology allow researchers to measure expression levels for thousands of genes simultaneously, across different conditions and over time. Analysis of data produced by such experiments offers potential insight into gene function and regulatory mechanisms. A key step in the analysis of gene expression data is the detection of groups of genes that manifest similar expression patterns. The corresponding algorithmic problem is to cluster multicondition gene expression patterns. In this paper we describe a novel clustering algorithm that was developed for analysis of gene expression data. We define an appropriate stochastic error model on the input, and prove that under the conditions of the model, the algorithm recovers the cluster structure with high probability. The running time of the algorithm on an n -gene dataset is $O(n^2[\log(n)]^c)$. We also present a practical heuristic based on the same algorithmic ideas. The heuristic was implemented and its performance is demonstrated on simulated data and on real gene expression data, with very promising results.

Key words: clustering algorithms, gene expression analysis, DNA arrays, probabilistic analysis, tissue classification.

1. INTRODUCTION

IN ANY LIVING CELL that undergoes a biological process, different subsets of its genes are expressed in different stages of the process. The particular genes expressed at a given stage and their relative abundance are crucial to the cell's proper function. Measuring gene expression levels in different developmental stages, different body tissues, different clinical conditions, and different organisms is instrumental in understanding biological processes. Such information can help the characterization of gene function, the determination of effects of experimental treatments, and the understanding of many other molecular biological processes.

Current approaches to measuring gene expression profiles include SAGE (Velculescu *et al.*, 1997), RT/PCR (Somogyi *et al.*, 1995), and hybridization-based assays. In the latter, a set of oligonucleotides, or a set of appropriate cDNA molecules, is immobilized on a surface to form the hybridization array. When a labeled target DNA (or RNA) mixture is introduced to the array, target sequences hybridize to complementary immobilized molecules. The resulting hybridization pattern (detected, for example, by fluorescence) is indicative of the mixture's content. Hybridization arrays are thus used as molecular recognition tools for nucleic acids (see Blanchard and Hood, 1996; Drmanac *et al.*, 1991; Eisen and Prown, 1999; Khrapko *et al.*, 1991; Lennon and Lehrach, 1991; Lin *et al.*, 1996; Lysov *et al.*, 1995; Pevzner *et al.*, 1991).

¹Department of Computer Science and Engineering, University of Washington, Seattle, Washington.

²Department of Computer Science, Sackler Faculty of Exact Sciences, Tel-Aviv University, Tel-Aviv, Israel.

³Hewlett Packard Laboratories Israel, Haifa, Israel.