494g New Approaches for Enabling Temporal Expression Profiling Analysis

Eric Yang, Joseph Vitolo, Charles Roth, and Ioannis (Yannis) P. Androulakis

It has been hypothesized that expression profiling using gene arrays can be used to distinguish temporal patterns of change in gene expression in response to a drug in vivo, and that these patterns can be used to identify groups of genes regulated by common mechanisms (Wei, Liu et al. 2004). For such a temporal expression analysis two critical questions need to be addressed:

1. How to establish the potentially complete patterns of co-expression and their non-intuitive temporal relations (Qian, Dolled-Filhart et al. 2001) 2. How to establish the link between genes that exhibit strong correlation in terms of their expression patterns and their underlying regulatory architecture (Qian, Lin et al. 2003) .

A critical component in this, and other similar studies, is how to establish the relationship between the available transcriptional profiles and how to extract in a systematic and unsupervised manner significant expression motifs and the key gene subsets associated with each pattern. In this work we propose a very efficient algorithm for extracting significant patterns of expression (motifs) and begin the assessment of the relationships between gene characterized by similar expression motifs and their underlying regulatory architectures.

Clustering of time series data, of which a subset is the transcriptional data from large-scale microarray experiments, is a very active area of research and a variety of problems have been discussed in the open literature. The fact that this problem persists, in particular as it related to genomic data, is just an indication of the many complexities both computational and interpretational. Among the leading candidates for clustering expression profiles are distance-based methods, with k-means clustering being one of the leading candidates. However, it has been argued recently that distance based methods generate local solutions that are not necessarily meaningful (Lin and Keogh 2004). Furthermore, identifying a priori the number of necessary clusters remains, in general, an open problem. However, significant successes have been identified in the open literature.

The subject of this paper is to explore an alternative approach towards the analysis of temporal expression data in an attempt to better characterize the nature of expression patterns. In this work we focus on the work of (Almon, DuBois et al. 2003), analyzing the corticosteroid effects on rat liver. We will illustrate the application of a novel way for representing the information content of transcriptional profiles. We develop emerging motifs of the expression profiles and by analyzing those identify probes with persistent and overpopulated expression patterns. These in turn can be used to postulate tentative significant expression motifs characteristic of the transcriptional profiles. Our main motivation is to define identifiers that uniquely characterize each transcriptional profile. Our goal is to identify those transcripts that share significant components of their expression patterns. The goal of our approach is to concurrently achieve a characterization of the transcriptional data as well as a significant dimensionality reduction in order to assess the qualitative characteristics of the expression data. In order to do so, we explore the idea proposed by (Lin, Keogh et al. 2002). The algorithm transforms the time series data into a sequence of symbols, which are subsequently hashed to unique (motif-dependent) identifiers. The hashing function explores the concept of proximity preserving hashing (Chin 1994), that is similar structures hashing to similar values. The hash values (motifs) can now be sorted and similar motif values correspond to similar transcriptional profiles.

Finally, we will discuss the associations between emerging stable expression profiles, participating genes and their possible common elements of regulatory elements by analyzing the promoters for the presence of regulatory elements using the ModelInspector utility of the Genomatix software suite

(http://www.genomatix.de). The 374 DNA-protein binding matrices (models) analyzed were collapsed onto 142 families (functionally related binding sites). Preliminary results indicate possible relations between genes of similar motif and their regulatory network but the relations need to be further explored.

References

Almon, R. R., D. C. DuBois, et al. (2003). "Gene arrays and temporal patterns of drug response: corticosteroid effects on rat liver." Funct Integr Genomics 3(4): 171-9. Chin, A. (1994). "Locality-Preserving Hash Functions for General-Purpose Parallel Computation." Algorithmica 12(2-3): 170-181.

Lin, J. and E. Keogh (2004). "Finding or not finding rules in time series." Applications of Artificial Intelligence in Finance and Economics 19: 175-201.

Lin, J., E. Keogh, et al. (2002). Finding Motifs in Time Series. 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, Alberta, Canada.

Qian, J., M. Dolled-Filhart, et al. (2001). "Beyond synexpression relationships: local clustering of time-shifted and inverted gene expression profiles identifies new, biologically relevant interactions." J Mol Biol 314(5): 1053-66.

Qian, J., J. Lin, et al. (2003). "Prediction of regulatory networks: genome-wide identification of transcription factor targets from gene expression data." Bioinformatics 19(15): 1917-26.

Wei, G. H., D. P. Liu, et al. (2004). "Charting gene regulatory networks: strategies, challenges and perspectives." Biochem J 381(Pt 1): 1-12.