Template-Based Gene Expression Analysis

Torgeir R. Hvidsten[1] Torgeir.Hvidsten@idi.ntnu.no
Tor-Kristian Jenssen[1] Tor-Kristian.Jenssen@idi.ntnu.no
Jan Komorowski[1] Jan.Komorowski@idi.ntnu.no
Astrid Lægreid[2] Astrid.Lagreid@medisin.ntnu.no
Arne Sandvik[2] Arne.Sandvik@medisin.ntnu.no
Dyre Tjeldvoll[1] Dyre.Tjeldvoll@idi.ntnu.no

[1] Department of Computer and Information Science, Norwegian University of Science and Technology, 7491 Trondheim, Norway
[2] Department of Physiology and Biomedical Engineering, Norwegian University of Science and Technology, Trondheim, Norway

Keywords: gene expression analysis, templates

1 Introduction

Analysis of gene expression profiles of genome-wide data created in cDNA-microarray experiments is a complex task that seems to require an inclusion of biological knowledge. We have attempted to express this knowledge by using an approach of the so-called templates. An analysis method has been defined and implemented. It has been tested on the fibroblast data set [3]. Preliminary results suggests that the method generates biologically interpretable classes of related genes.

2 Methods and Results

We have approached the problem of gene expression analysis from two different angles. In the first approach we used a clustering method that blindly bundles the time series of gene expressions. Similarity of the profiles that emphasizes an intuitively appealing concept of equal gradients was defined by means of a Haar wavelet transform [1]. A discernibility relation was constructed and gene profiles which were indiscernible were clustered to the same classes. The method was tested on the fibroblast data that are publicly available [3]. The data set contains measurements of gene expression levels for about 8600 human genes in response of fibroblasts to serum. 517 genes showed a significant* change in the expression levels during a 24 hour experiment. Of these, about 340 genes are non-EST's. Following Iyer et al. [3] we applied the afore mentioned method to the measurements over the entire 24 hour period. Our clustering results were similar although no detailed comparison analysis was performed. The results were presented in [2].

We have later found indications that the requirement of similarity over the entire period of 24-hours may be too restrictive and, supposedly, obscured the underlying biological knowledge. Although we could have relaxed the 24-hour constraint to be time-windows of varying length, it seemed more plausible to also pre-define which patterns of the gradients of the functions were to be searched for. So motivated, we decided to explore a second approach in which we have encoded background knowledge as a set of *templates*. A template is a prototypical pattern of expression level profiles restricted to a predefined time sub-interval. After initial experiments with several combinations of gradients we arrived at five templates that capture the basic features of all functions: *increasing–decreasing, decreasing–increasing, increasing, decreasing* and *constant*.

*For the definition of significance see Note 7 in [3]

The template approach has been tested on the same data set. The obtained clusters were then manually partially validated using the current biological knowledge.

The preliminary analyzes showed that the cluster of genes with increasing expression levels at an early interval (2–6 hours) contained a large amount of genes coding for proteins involved in protein synthesis. The cluster of genes with increasing expression levels at a later interval (4–16 hours) contained almost no protein synthesis associated genes, but instead a substantial number of genes involved in cell cycle processes was found.

We are now in the process of annotating all the known genes in the 517 gene data set according to Ashburner's Gene Ontology [6]. A comparison of clustering according to Gene Ontology annotation with clustering of the dynamics of gene expression levels will be presented at the meeting.

3 Discussion

By defining templates restricted in duration and form we essentially restrict the family of profiles under consideration to a well-defined class. The temporal restriction, motivated by knowledge about phases of the fibroblast serum responses, helps illuminate how a single gene may play different roles at different stages. The restriction to a pre-defined set of time courses is appealing, as we think it is possible to relate a pre-defined time-course to a biologically plausible function or role in a process.

The issue of finding the best clusters is in part a problem of feature extraction and feature selection. In our case it translates into finding appropriate intervals and their combinations. One way of achieving this is by employing Boolean reasoning to find minimal sets of intervals from all the possible combinations. The ROSETTA system [5] will be used to this end.

Several major challenges remain. Ultimately a good classification should be informative and relevant to the type of questions asked (e.g. molecular mechanisms versus cell biological or physiological processes). It should take into account that the same gene may participate in distinct processes at different time-points in the fibroblast response.

Acknowledgments

We wish to thank Aleksander Øhrn for discussions on the indiscernibility relations and Boolean reasoning. This research has been supported in part by the Norwegian Cancer Society, Norwegian Research Council and the Norwegian University of Science and Technology.

References

[1] Holschneider, M., *Wavelets – An Analysis Tool*, Oxford Science Publications, 1995.

[2] Hvidsten, T. R., Jenssen, T-K., Lægreid, A., Øhrn, A., Tjeldvoll, D. and Komorowski, J., Boolean Reasoning in the Analysis of Gene Expression Data, Poster at *The Data Mining for Bio-informatics – Towards In Silico Biology conference*, Hinxton–Cambridge, UK, 1999.

[3] Iyer, V.R., Eisen, M.B., Ross, D.T., Schuler, G., Moore, T., Lee, J.C.F., Trent, J.M., Staudt, L.M., Jr, J.H., Boguski, M.S., Lashkari, D., Shalon, D., Botstein, D. and Brown, P.O., The transcriptional program in the response of human fibroblasts to serum, *Science*, 283(5398):83–87, 1999.

[4] Klösgen, W. and Żytkow, J., (Eds.), *Handbook of Data Mining and Knowledge Discovery*, Oxford University Press, 2000.

[5] Komorowski, J., Øhrn, A. and Skowron, A., The ROSETTA Software System for Rough Sets, to appear in: [4].

[6] Ashburner, M. *et al.*, The Gene Ontology Database, `http://genome-www.stanford.edu/GO/`