Computational gene expression analysis; correlating data and knowledge

Torgeir R. Hvidsten
The Linnaeus Centre for Bioinformatics
Uppsala University
Husargatan 3
SE-751 24 Uppsala
SWEDEN
Torgeir.Hvidsten@lcb.uu.se

Introduction

Physiological and pathophysiological responses are associated with specific changes in cellular gene expression and gaining insight into these specific patterns enable hypotheses about gene function, understanding of gene regulation and promising new possibilities for molecular medical diagnosis. Microarray technology is a fundamental research vehicle in this context, allowing simultaneous readouts of relative gene expression levels for thousands of pre-selected genes in a tissue sample (Figure 1). Conduction several experiment using samples from different time points, clinical states and/or different treatments constitute a powerful setup for studying the genome-wide changes in cellular gene expression. However, this setup also raises non-trivial computational challenges not only related to storage and analysis of data, but also to evaluation and interpretation of the results according to current biological knowledge.

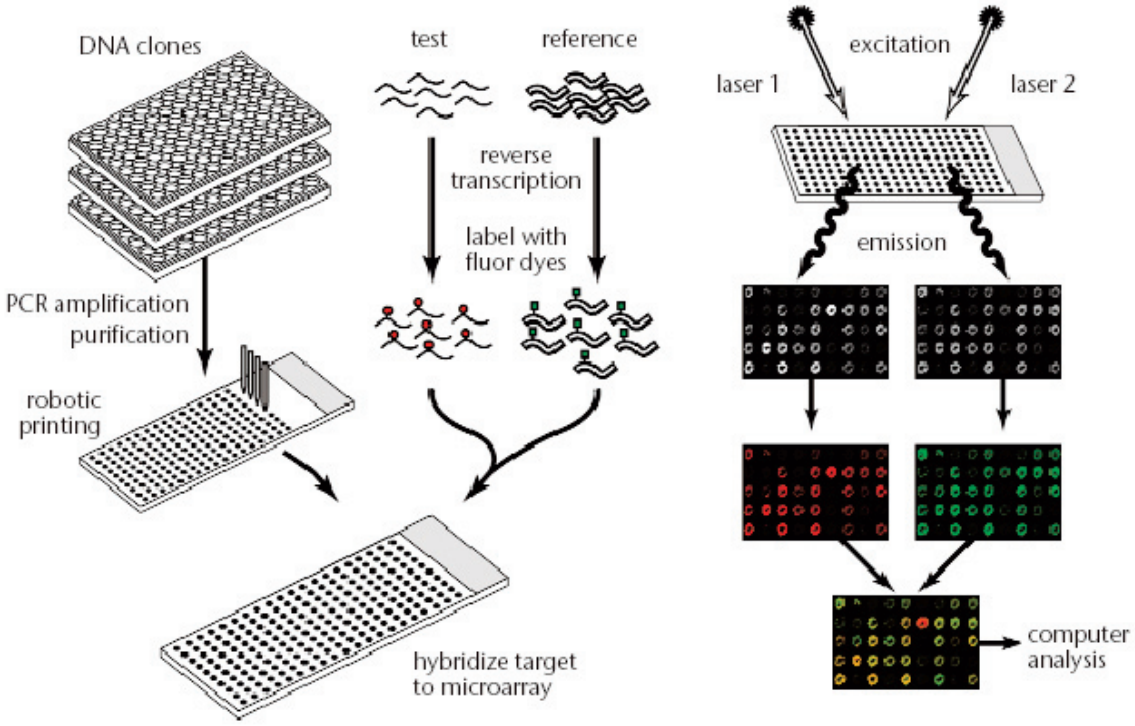

Figure 1 Microarray experiment. Gene probes of interest are printed on glass microscope slides. RNA from both test and reference sample are first fluorescently labeled with Cye3 (red) or Cye5 (green) and then hybridized to the clones on the array. The arrays are scanned and the resulting images are pseudo-colored and merged. Data from hybridization experiment is commonly viewed as normalized ratios Cye3/Cye5, indicating increased (>1), decreased (<1) or unchanged (=1) expression level in the test sample relative to the reference sample. Conduction several experiment using samples from different time points, clinical states and/or different treatments constitute a powerful setup for studying the genome-wide changes in cellular gene expression in living systems (D. J. Duggan et al., Nature Genetics 21, 10-14, 1999).

Gene expression data analysis

The typical intermediate result of a microarray study is the genes ◊ samples – matrix where each row holds the expression profile of one gene over all samples and each column hold the expression profile of one sample over all genes. Common for most data analysis methods is that they are based on representing genes/samples as points in a multi-dimensional expression spaced spanned by the samples/genes (obviously, the exception is studies in which samples are taken from different time point and genes can be represented as functions of time in two dimensions).

Data analyses strategies typically try to reduce the complexity by assume that there are underlying, known or unknown, classes of genes or samples that corresponds to biological or medical phenomena. Such classes can, for examples, be groups of genes that code for proteins with the same molecular function or are involved in the same biological process, or groups of tissue samples that are taken from patients with the same disease or that underwent the same medical treatment. There are two conceptually different analysis strategies. The first strategy is called class discovery or clustering (1,2). Such methods try to discover biological related genes or samples by assuming that they will appear close to each other in the expression space. Hierarchical clustering is one commonly used clustering method in the context of gene expression analysis (Figure 2). The second strategy take advantage of already well studied examples and try to learn models that can be used both to gain insight into the parameters discriminating different classes and for prediction/diagnosis. These methods are called supervised learning or machine learning (3). Models can range from simply identifying genes that are differentially expressed in, for example, samples from different cancer types, through IF-THEN rules combining such discriminatory genes, to complex illegible models such as artificial neural networks and hyper-planes from support vector machines.

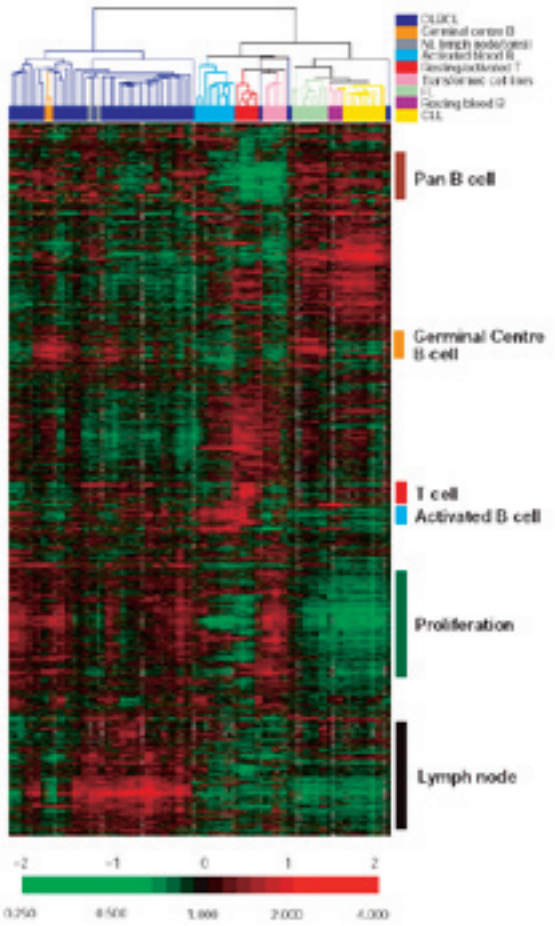

Figure 2 Hierarchical clustering. Hierarchical clustering organize samples in a tree that reflects their level of (expression) similarity. The figure shows the hierarchical clustering of 96 samples from normal and malignant lymphoma using expression profiles from almost 20 000 selected human genes. Diffuse large B-cell lymphoma (DLBCL) is known to be clinically heterogeneous in that only about 40% respond well to current treatment. The figure shows how this subtype is almost perfect clustered separately from the other samples. Furthermore, it was shown that the two sub-clusters of DLBCL (found using clustering) include patients with significantly different expected survival time indicating two distinct molecular types of DLBCL (Alizadeh et al., Nature 403, 503-511, 2000).

Common to both clustering and machine learning methods is the need to validate results in order to assess their biological or medical relevance. While early studies on the relation between gene function and gene expression only visually hinted on expression similarity within groups on functional related genes, newer studies typically test the

statistical significance of overlap between expression clusters and functional classes or even build models capable of hypothesizing novel functions for genes. Such computerized use of large amount of biological knowledge, in this case knowledge of gene function, requires structured databases and controlled vocabularies (6). Building such databases again requires reading thousands of articles in order to extract the relevant information. As a consequence, developing new tools for text-mining (automatic extraction of knowledge from text) has become one of the most important research areas in Bioinformatics (7).

Selected data analysis tools and further reading

The research community early recognized the need for standardization to facilitate data management, processing, transfer, publication and reproducibility of microarray data. This has resulted in Laboratory Information Management Systems (LIMS) sharing common data formats (MIAME: Minimum Information about a Microarray Experiment), and data warehouses such as ArrayExpress (4) and the Gene Expression Omnibus (5) publishing microarray studies. There is a wide variety of freely available systems and services applicable to gene expression data analysis on the web. Some recommended sites and directions for further reading are indicated here:

References

1. Cluster and TreeView: various clustering algorithms including a much used visualization tool for hierarchical clustering (http://rana.lbl.gov/EisenSoftware.htm).
2. J-Express: cluster and statistical analysis of gene expression data (http://www.ii.uib.no/~bjarted/jexpress/).
3. The ROSETTA system: a general toolkit for IF-THEN rule-based supervised learning (http://www.idi.ntnu.no/~aleks/rosetta). Case study using ROSETTA for functional gene expression analysis: http://www.lcb.uu.se/~hvidsten/fibroblast/.
4. ArrayExpress: public database for microarray studies by the European Bioinformatics Institute (EBI) (http://www.ebi.ac.uk/arrayexpress).
5. Gene Expression Omnibus: public database for microarray studies by the National Center for Biotechnology Information at NIH (http://www.ncbi.nlm.nih.gov/geo/).
6. Gene Ontology: a dynamic controlled vocabulary of gene and protein roles in cells (http://www.geneontology.org/).
7. PubGene: gene networks automatically extracted from literature (http://www.pubgene.uio.no/).
8. The Chipping Forecast II, freely available supplement to Nature Genetics, 32: 461 – 552, 2002 (http://www.nature.com/ng/web _ specials/)

Announcements

BEN - the Belgian EMBnet node (http://www.be.embnet.org) - has been invited to give a workshop in Bioinformatics, the 3 and 4 April 2003 at the "Université de la Mediterranée" (http://mediterranee.univ-mrs.fr/), in Marseille, France. Guy Bottu and Valérie Ledent ran the workshop in front of an attendance of 24 bioinformaticians and IT specialists.

This workshop (http://www.esil.univ-mrs.fr/%7Ethieffry/BEN/index.html) was organised in the frame of the DESS CCI (Compétences Complémentaires en Bioinformatique; http://www.dil.univ-mrs.fr/desscci/). It focused on the Software Suites Staden and EMBOSS, the EMBOSS web interface WEMBOSS developed by BEN and the Sequence Retrieval System SRS. The focus was on the overview of the functionalities of these software packages, as well as their installation and maintenance on Linux machines.

Other courses and meetings

See our web site:
http://www.embnet.org/activities/dyn _ schedule.php

You can even use this page to announce your own courses and meetings (OBS: this service is moderated).