Making the most of microarray data

Terry Gaasterland & Stefan Bekiranov

Laboratory of Computational Genomics, The Rockefeller University, 1230 York Ave, New York, New York, 10021, USA.
e-mail: gaasterland@genomes.rockefeller.edu and bek@genomes.rockefeller.edu

The impact of microarray technology on biology will depend on computational methods of data analysis. A supervised computer-learning method using support vector machines predicts gene function from expression data—and shows promise.

Microarray assays can measure the transcriptional effects of changes in gene function under different conditions. They can reveal genes that characterize tissue type, developmental stage, or responses to environmental conditions or genetic modifications. Microarray assays will therefore become a general feature of experimental protocols in genetics and cell physiology. As array data burgeon, new questions arise: if we, as a research community, collect all array hybridization data in a central location, can we assign new genes of unknown function to known functional classes? Can we correlate gene expression with gene function? Can we find new classes of co-regulated genes? Can we extract complete gene regulatory networks from microarray gene expression data?

Computation is our only hope, and an article by Michael Brown and colleagues1 in a recent issue of The Proceedings of the National Academy of Sciences describes an approach to microarray data analysis that addresses the first question. The authors use support vector machines (SVMs; Fig. 1), a supervised computer-learning method, to train a 'classification machine' to recognize new genes that are similar in expression pattern to groups of genes known to be co-regulated. In contrast with classical unsupervised clustering methods and pure self-organizing maps, the approach builds on existing knowledge (Fig. 2) and has the potential to refine and correct it.

Assigning new genes to old categories

Brown et al. demonstrate two compelling uses for SVM analysis of microarray data. First, they trained classification machines (SVMs) to recognize five sets of genes in functional classes that were each expected to be co-regulated: those mediating the tricarboxylic acid (TCA) cycle, respiration, cytoplasmic ribosome biosynthesis, proteasome biosynthesis and histone biosynthesis. As a control, they trained SVMs to recognize genes encoding helix-turn-helix proteins, fully expecting them to fail to discriminate this class (which they did). When they asked the trained SVMs to classify each yeast open reading frame of unknown function, 15 genes were assigned: 2 to the TCA class, 4 to respiration, 5 to ribosomes and 4 to the proteasome. One of the respiration candidates was subsequently annotated as encoding a subunit of the mitochondrial ATP synthase complex and another, as a hitherto unknown protein kinase. Of the ribosomal candidates, one is predicted to encode a zinc-finger protein and another, a phosphatase.

To measure performance, Brown et al. trained the SVMs with two-thirds of the positive data (genes known to be in a particular functional class) and with negative data (genes known not to be) for each functional class, and then asked them to assign the remaining one-third of the genes to the respective class. Overall performance was good, and the failures provided new information on each of the five categories: five ‘TCA-cycle’ proteins that were not recognized by any SVM, for example, turned out to be relevant enzymes regulated at the protein level. The SVMs also misclassified several members of the TCA cycle as respiratory proteins, demonstrating that the two classes—both involved in ATP production—are not easily separable.

These results raise the question of which pathways are sufficiently regulated and uniquely regulated at the transcription level to be expanded or confirmed through microarray data. SVMs become uninformative when the noise in the ‘negative’ examples outweighs the the number of ‘positive’ examples. Brown et al. have shown that SVMs can recognize that gene products belong to a particular class. An interesting follow-up will be to train SVMs for every functional class in yeast, Escherichia coli and other organisms for which genomes are complete and annotated, and to see whether they can be used to assess, confirm and extend the functional classifications.

Spoilt for choice: methods of analysis

Unsupervised clustering methods, such as hierarchical and K-means clustering, assume that each gene fits into only one cluster. This, of course, is not necessarily true in biology. Moreover, the fact that genes are in the same cluster does not necessarily mean that they have similar expression patterns. With K-means clustering, for example, the user specifies beforehand the number of clusters to be generated, and the algorithm places each gene in its optimal cluster—which is not always a meaningful one. This requires that clusters be evaluated for quality. The ‘tightness’ of the cluster and the existence of outliers (and, if they exist, their proximity to the next cluster) and a core set of characteristic genes should collectively inform on quality. Most importantly, one should consider: does the cluster make biological sense?

A new unsupervised clustering algorithm implemented in BioC lust™ (ref. 3) has been developed specifically for microarray analysis. Here the user specifies the maximum space that a cluster can occupy, and the algorithm builds suitable clusters. The final number of clusters depends on the space limit. Genes that cluster together are guaranteed to have a certain ‘affinity’ for each other. Users can run BioC lust™ repeatedly to adjust space size so that clusters appear biologically meaningful.

Also independent of prior knowledge, principle components analysis (PCA) reduces a data set with many variables to a smaller number of new uncorrelated variables that explain the majority of the variance. With fewer variables, the data set is easier for a human to understand and visualize—especially if two variables suffice to capture the majority of the variance. At the same time, a certain amount of precision in relationships among genes is lost. As PCA focuses on independent variables, it is especially useful as an initial step for analysis of gene expression under a mix of independent and non-independent conditions.

Learning from supervision

In contrast with unsupervised methods that generate classes for genes, supervised learning methods learn known classes. The trainer must feed the SVM with both positive and negative examples from each class. Brown et al. tested four different types of SVMs in addition to four other machine-learning techniques to compare their relative merits. The SVMs consistently outperformed the other methods in assigning genes to five functional classes defined by yeast microarray data.2

---

**Fig. 1** A support vector machine (SVM) is a computational entity that accepts positive and negative training examples of a topic to be learned. As it ‘learns’, it draws a hyper-plane which maximally separates input data points into two classes, members (green) and non-members (red). Here, input data is shown in three dimensions, which is easily visualized. The data of Brown et al. span 79 dimensions, one for each microarray experiment in the data set.

**Fig. 2** An untrained support vector machine (a) is trained with positive examples (green) and negative examples (red) to build a trained machine (b; Fig. 1) that can take an unknown object (white) and determine whether or not it is similar to the training set. For functional classification of microarray data, positive and negative examples are drawn from annotations of coding regions in genome sequence data (c).
Axin and hepatocellular carcinomas

Hans Clevers

Department of Immunology, University Hospital and Centre for Biomedical Genetics, Utrecht, Netherlands. e-mail: h.clevers@lab.azu.nl

A gene (AXIN1) encoding another component of the Wnt-signalling pathway is reported to be mutated in human tumours, underscoring the oncogenic consequences of inappropriate activation of developmental signalling pathways later in life. Inhibitors of these signals therefore represent potential agents with which to treat cancer.

Two general principles connect the disciplines of developmental biology and molecular oncology. First, crucial decisions during animal development are made through the action of a handful of signal transduction pathways. Second, activating mutations in these pathways are a major cause of human cancer. A classic example is the wingless/Wnt signalling cascade, a critical developmental pathway that is disrupted in several types of tumour. In this issue, Yusuke Nakamura et al. reinforce these principles with the discovery that mutations in the gene encoding axin, which acts in the Wnt-signalling pathway, cause hepatocellular carcinoma.

Polarized with Wnt1

Mouse Wnt1, originally called Int-1, was one of the first proto-oncogenes to be identified—activation of Wnt1 by integration of the Mouse Mammary Tumor Virus causes cancer of the mammary gland in mice. Wnt1 is the mammalian homologue of Wingless, a key regulator of segment polarity in Drosophila embryos. Together with other segment polarity genes, Wingless defines the anterior and posterior regions of each segment of the developing fly embryo. Anterior and posterior regions are distinguished by the presence or absence, respectively, of bristles on the larval cuticle. Analyses of fly mutants with disrupted bristle patterns have uncovered components of the signalling pathways that determine embryonic patterning.

Xenopus laevis offers an equally simple phenotypic assay for elucidating components of the Wnt-signalling pathway in vertebrates. Injection of components of the Wnt cascade into embryos at the 2–8 cell stage induces a secondary body axis, resulting in two-headed tadpoles (Fig. 1).

The SVM results support a tiered method for assessing microarray data. First, for every gene, one should determine its nearest neighbours and ask if its relationships with them are biologically significant. Second, for genes that are known to be regulated together, one should ask: are their expression patterns similar, and if so, which other genes have the same pattern? These can be identified through SVMs or by refining self-organizing maps with supervised phase. Third, one should classify genes through unsupervised learning methods and ask whether clusters are biologically meaningful, and whether they contain outlier genes. Finally, clusters can be tested and refined by training SVMs with genes from the centre of each unsupervised cluster.

Regulons and beyond

One approach to deduce prokaryotic regulons (distinct genes that are part of a single mRNA transcript) from expression data is to mutate a suspected regulator gene and measure associated changes in transcription levels. Another is to compare normal conditions with conditions that invoke the regulatory protein. An interesting exercise for microarray analysis through supervised classification would be to determine whether protein signatures indicative of prokaryotic regulons (as determined by two-dimensional electrophoresis) can be detected through microarray analysis, and whether such signatures can be extended or refined at the transcription level. For example, components of the methionine-biosynthesis pathway of E. coli differ in their relative expression levels during growth in the absence of methionine. Specifically, the level of metE is much higher than that of the other proteins in the pathway. SVM analysis of corresponding microarray data could be used to determine whether metE should be classified with the other genes in the pathway or whether it better correlates with genes in other pathways.

Another interesting protein signature is that for phosphorus limitation, which includes 150 proteins encoded by the ‘phosphorus’ regulon. Among the signature proteins, phoA is induced 1,000-fold while 19 other proteins are repressed, including phoE. Inhibition of phoR, a histidine kinase that induces the phosphorus regulon, is accompanied by phoE expression. So the level of phoE is a useful marker for activity of the regulon. SVMs can be used to determine whether members of the pho-regulon can be predicted based on levels of gene expression—thereby identifying regulons in prokaryotic organisms and representing a first step towards deciphering gene networks.

1. Brazma. A. Report on the first international meeting on microarray gene expression databases (MGED); http://www.ebi.ac.uk/microarray/MGED/.