New approach to gene expression analysis

Martina Habeck, freelance writer

A team of Penn State scientists has developed a new tool to mine the massive amounts of data that is generated by microarray experiments [1]. Their information-theoretic approach is reportedly superior to the method most commonly used to date, which is hierarchical clustering with a standard correlation coefficient.

Making sense of microarray data

Gene-expression microarrays have become an important tool for drug discovery. These experiments create a vast amount of data – with the currently available technologies it is possible to assay 10,000+ genes at a time. Scientists have only just begun to learn how to interpret and apply these data.

One method for data analysis is clustering, a mathematical approach to group together genes that have similar temporal expression patterns. How can this approach lead to new therapeutics? In many cases, genes within such a cluster are co-regulated because they share promoter elements and are controlled via a single transcription factor. A better understanding of these interactions could lead to the identification of new and better-quality targets and could thus enable the prediction of responses to particular gene-specific drugs.

Finding the right mathematical tool

Clustering microarray data involves two steps: first, the investigator needs to define a mathematical description for similarity – the Pearson correlation *r* is often used for this purpose; second, genes with similar expression patterns are grouped together using a cluster algorithm. These algorithms fall into two categories: hierarchical and non-hierarchical.

Hierarchical agglomerative cluster analysis is the strategy that is most commonly used for microarray data analysis. It starts by assuming that all the elements of a data set form their own individual cluster. The clustering algorithm then merges (agglomerates) the two most similar cluster pairs together to form a single cluster. This is repeated until all elements are clustered into a single cluster. The model can be graphically represented by a dendrogram, which is comparable to the phylogenetic tree diagram.

Although this algorithm is quick and easy to perform and the visualized result has a clear structure, there is some debate over whether this technique is of any use for the analysis of gene expression. Hierarchical clustering forces data points into a strict hierarchy of nested subsets. However, this structure does not necessarily make biological sense; the recognition of fundamental patterns of gene expression is left to the observer.

Trying alternatives

Bioinformaticists are, therefore, working increasingly with non-hierarchical algorithms. Raj Acharya and Jyotsna Kasturi at Penn State University (http://www.psu.edu) use self-organizing maps (SOMs). 'The main aim here is to identify clusters of genes with similar expression profiles, using no *a priori* knowledge (unsupervised learning),' explains Kasturi. SOM algorithms work with a fixed number of clusters; that number corresponds to the number of the most prominent gene expression patterns. The algorithm sorts data into each cluster according to how closely its expression pattern matches the average expression pattern of the cluster. 'Finally, we end up with a set of well separated and 'dense' gene clusters,' says Kasturi.

SOMs have proved to be valuable in identifying the predominant gene expression patterns during yeast growth and during the process of haematopoietic differentiation [2]. However, the Penn State scientists believed that there was room for further improvement because SOMs are commonly used in combination with the Pearson correlation (PC). The PC essentially measures the similarity between two gene expression profiles by comparing all the time points of the first gene to those of the second gene and evaluating how well the data points fall along an imaginary line. Acharya and Kasturi took a fundamentally different approach: They used SOM in combination with an information-theoretic measure, called Kullback-Leibler (KL) divergence. It compares the shapes of two gene expression profiles and measures their dissimilarity.

According to Eric Neumann, Vice President of Informatics at Beyond Genomics (http://www.beyondgenomics.com), swapping PC for KL divergence is an important achievement: Because of the way it works, the PC is sensitive to single outliers. And 'outliers can really distort your findings,' reminds Neumann. He argues that an information-theoretic approach, such as KL divergence, is better suited to discover and remove or correct outliers because it considers them as unreliable information that needs to be weighted against the overall distribution. A recent study published by Kasturi *et al*. proves his point [1].

1359-6446/03/$ – see front matter PII: S1359-6446(03)02705-3

Testing the information-theoretic approach

In collaboration with Murali Ramanathan at the University at Buffalo (http://www.buffalo.edu), the two Penn State scientists tested their approach by applying it to two previously published data sets:

- a set of 517 genes that were expressed after stimulation of serum with human fibroblasts [3]; and
- a set of 6108 cell-cycle regulated yeast genes [4].

Using visual inspection and statistical analysis, the investigators demonstrated that SOM plus KL divergence is superior to hierarchical clustering plus PC, or even to SOM plus PC [2]: the SOM algorithm was able to detect clusters with distinct patterns of temporal gene expression but the cluster plots obtained with KL divergence were better separated and more dense than those obtained with the PC distance measure. By contrast, hierarchical clustering produced a large amount of false positives, and many clusters contained only a single gene. Acharya and Kasturi are now working on improving their method (a more user-friendly version of the programme will be available soon on Kasturi's website (http://www.cse.psu.edu/~jkasturi).

More meaningful information

An important future direction would be to modify the approach to take into account slight phase shifts in transcript production, suggests Neumann. 'Genes can be controlled the same way as another gene, but they may take a little longer to get expressed,' he explains. 'You would supposedly see profiles that look very similar but have a bit of a time shift; people refer to this as a phase shift. Unless you explicitly take that into account in your analysis [and only a few papers have done so to date], all these tools are doing time-to-same-time comparisons; they don't address the whole problem.'

But the SOM plus KL divergence approach will already be an important addition to the array of tools that scientists can use for the analysis of gene expression data: 'If you use our method, you can come up with well separated clusters that probably won't be obtained by other methods,' concludes Acharya. He predicts that this could translate into more meaningful information about gene function – and could thus improve the drug discovery process.

References

1 Kasturi J *et al.* (2003) An information theoretic approach for analyzing temporal patterns of gene expression. *Bioinformatics* 19, 449–458
2 Tamayo, P. *et al.* (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. U. S. A.* 96, 2907–2912
3 Iyer VR *et al.* (1999) The transcriptional program in the response of human fibroblasts to serum. *Science* 283, 83–87
4 Alter O *et al.* (2000) Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci. U. S. A.* 97, 10101–1016

Spinach makes a safer anthrax vaccine

Hillary E. Sussman, freelance writer

The current form of the vaccine against anthrax infection has been licensed for human use since 1970 but has recently been characterized as sub-optimal. In response, a resurgence of research to produce a purer, safer vaccine has ensued. One such effort, by scientists at Thomas Jefferson University (http://www.tju.edu), has revealed that spinach can be used as a vehicle for the production of an edible vaccine against anthrax [1].

The anthrax bacterium

Anthrax is caused by the spore-forming bacterium *Bacillus anthracis*, which exists as spores in the soil and, therefore, commonly affects grazing animals such as cattle and sheep. Human infection, although rare, occurs following direct skin contact with infected animals or their wool, hides or tissues, by ingestion of contaminated meat, or via inhalation of the spores. Left untreated, inhalation anthrax is almost always fatal and early intervention with antibiotics, such as ciprofloxacin, is essential.

The status quo

Vaccination is recommended for persons at risk of exposure to anthrax spores. The current vaccine is based on cell-free culture supernatants of an attenuated strain of *B. anthracis* adsorbed on aluminium hydroxide (in the USA) or precipitated with aluminium phosphate (in the UK) – aluminium acts as an adjuvant. It is incompletely characterized and difficult to standardize and, therefore, exhibits inconsistency between lots. It is also relatively reactogenic, with side effects including a possible link to Gulf War Syndrome (whose symptoms include chronic fatigue, depression, skin rashes and gastrointestinal disorders [2]),

1359-6446/03/$ – see front matter PII: S1359-6446(03)02706-5