Software reviews

Expression Profiler: A suite of web-based tools for the analysis of microarray gene expression data

DNA microarray analysis has become one of the most widely used tools for the analysis of gene expression patterns on a genomic scale. Microarray analysis allows the simultaneous interrogation of tens of thousands of genes in a single assay, providing unprecedented quantities of data on temporal, spatial and tissue-specific patterns of expression. Data such as these provide information on the function of unknown genes, novel roles for previously characterised genes, and glimpses of the transcriptional networks that underlie cellular metabolism. Although the laboratory techniques have become widespread and well established, the analysis of data generated in a typical microarray study remains a significant challenge. A typical microarray assay generates data on thousands of genes, and each microarray study tens or hundreds of assays, generating extremely large quantities of data that must then be analysed. Although a number of other approaches have been applied to the analysis of microarray data, \(^5\) the most widely used techniques are those based on clustering. \(^6\)

Cluster analysis techniques can be divided into two broad classes: agglomerative and divisive. Hierarchical clustering is the most widely used agglomerative technique; it creates a set of nested classes that resembles a phylogenetic classification. \(k\)-Means is an example of an divisive clustering technique that divides a data set into a predetermined number \((k)\) of classes where the members within a class are more closely related than the individual classes are to each other. It should be noted that selecting different algorithms, different normalisations, or different distance measures might place different objects into different clusters. While these techniques are extremely powerful for identifying patterns in gene expression data, they can be computationally intensive for large data sets. In addition, the clusters that are produced must be further analysed to determine whether the resulting classification is biologically relevant; a clustering algorithm will always produce clusters. Typically, what one would like to know is whether the genes in a cluster share some biological features: Are they in a related pathway? Do they play a common function? Do they share some identifiable common regulatory sequence?

Expression Profiler is a web-based set of tools for the clustering and analysis of gene expression data developed by Jilo Vaak at the European Bioinformatics Institute (EBI). It consists of a collection of six utilities that, although still under development, represent a very useful means of exploring expression data: EPCLUST, URLMAP, EP:GO, GEMONES, PATMATCH and SPEXS.

EPCLUST is a tool for expression data cluster analysis. It uses a series of web
pages to allow users upload data for analysis, reading a variety of standard formats and reformatting the data appropriately for analysis. The user can store the data on the analysis server at the EBI, as either a public or private dataset. Further, users can use previously stored example data sets provided to learn about features of this and other software included in the system. EPCLUST allows users to partition genes using either hierarchical or k-means clustering, each with a number of variations (including the ability to select genes to seek k-means partitioning). The algorithms are well optimised and the hardware on which the calculations run (64-bit Compaq Alphas) is extremely fast – clustering 12,000 genes across a dozen experiments took less than a minute. The output from these programs is a graphical representation of the clusters (as a GIF) along with information about the calculations (Figure \object="ogf5" kern -1.79mm1). While it might have been better to create a more interactive display, the mapped GIF presentation is a reasonable compromise with a number of features that allow users to further explore the output data. One place where EPCLUST is lacking is that users can only cluster the genes, not the experiments. This is a significant drawback relative to some other array analysis software because in many instances, the main question users want to explore is how their experimental samples are related. While genes can be selected based on their expression properties, such as whether they are over- or under-expressed in a single experiment or the experiments in general, the algorithms use only the expression ratios between conditions, not the absolute hybridisation intensities. This limits what information can be extracted about the genes being analysed, and makes it difficult to use Affymetrix GeneChip™ and filter hybridisation data, but it is the approach most expression analysis programs use.

URLMAP is the second tool in the Expression Profiler suite. It can be used to map for contents, such as the clusters produced by EPCLUST to other web-based tools, including a number of preloaded destinations as well as those in your home laboratory. Consequently, independent of the other tools at EBI, URLMAP represents a nice utility for many laboratories that want to generate hot-linked lists of genomic data.

Linking co-expressed genes to common functions and pathways is a significant challenge because much of the annotation for genes and gene sequences in the public databases is incomplete. The Gene Ontology (GO) project¹⁰ is an attempt to assign genes to biochemical roles, metabolic pathways and subcellular regions using a common, structured vocabulary. EP:GO is a tool that allows users to browse relationships in GO categories, to search GO vocabularies, and to extract genes associated with various assignments in order to assist in the interpretation of expression data.

Once a collection of genes has been selected based on shared patterns of expression and possible biological function, an obvious question is whether there is a biological basis for the observed relationships that can be extracted through an analysis of the genes in the context of their genomic sequence. GENOMES allows the retrieval of full or partial genome sequences, including all or some of the genes annotated in a genome, or, for example, their upstream coding regions. These can then be used as the input to a sequence pattern discovery tool, or to retrieve expression data from the input data to EPCLUST, allowing the user to look for shared expression patterns. PATMATCH is one of the tools that sequence data can be passed to. It matches patterns, such as transcription factor binding consensus sequences, against a sequence database. PATMATCH provides utilities to visualise pattern occurrences along the sequences and to compare matches against expression profiles derived from EPCLUST.

SPExS is the least developed but most ambitious tool in the Expression Profiler
Figure 1: Graphical representation of average linkage clustering results produced by EPCLUST based on an uncentred correlation coefficient distance measure between the gene expression vectors. In the web display, genes that are up-regulated appear as red; down-regulated genes appear green. Clicking on any node on the tree allows the user to zoom into the area defined by its branches and those genes can be used to launch other analysis tools in the Expression Profiler suite.
Software reviews

suite. SPEXS allows users to upload sequences for analysis, searches for patterns occurring most often in the data set, or those over- or under-represented in some subset of the data relative to other sets of sequences, to study the distribution of the generated patterns, and to make profiles based on those motifs discovered. As such, SPEXS allows one to ask precisely the question that one would want to ask in order to test the hypothesis that expression patterns are correlated with sequence – is there support from the genome sequence suggesting that co-expressed genes do share common regulatory sequences, or are observed correlations between sequences and expression merely chance occurrences?

The EBI Expression Profiler suite is a first generation set of tools for the analysis of gene expression from microarrays. While there are areas that could be improved, such as data upload and the addition of more features for exploration of relationships in the data, and more intuitive interaction between the various elements of the system, these things will surely improve with time. Microarray data are inherently more complex than much of the other data types that biologists are used to dealing with. The current system requiring multiple table-delimited files to be uploaded is due to a lack of a standard data exchange format. Vaak and his colleagues at EBI are among leading members of the Microarray Gene Expression Database group (MGED1), which is seeking to establish such standards. Once such standards are developed and put into effect, it should greatly facilitate data import, export and exchange between tools. As for analysis, the protocols and techniques used to analyse gene expression profiles are as diverse and specialised as are the protocols used in the laboratory; the techniques used for the analysis of a time-series experiment is fundamentally different from what one might use to study gene expression in response to an external stressor or in a comparison of normal and tumour tissue. However, these techniques will evolve and begin to become more standardised over time, and as they do, Expression Profiler will probably adapt. Vaak actively solicits suggestions through the web site and a number of features have been added based on user requests.

What one should focus on is the large number of things that Expression Profiler does right. First, it provides access to a suite of tools and the computer hardware necessary to support them that many laboratories would otherwise not be able to take advantage of. Second, it provides the means for investigators to store and manipulate relatively large data sets and to use those data to address questions that go beyond the analysis of expression fingerprints by allowing genes to be explored in at least part of their natural environment – the genome. As the genome sequence of humans and other organisms becomes available and matures into a useful format, having access to tools such as these will become increasingly important. Expression Profiler provides an early version of the tools that will be necessary to explore data from functional genomics in the context of the genome and as such, sets a standard that others must follow.

Everyone interested in exploring microarray gene expression data should, at the least, have a look at Expression Profiler and use their sample data sets to learn something of how expression data analysis can be done.

John Quackenbush
The Institute for Genomic Research
Rockville
MD 20850, USA

References

The Microarray Explorer analysis package

Next to genome sequencing, microarray gene expression technology has become the most important source of high-throughput data for life science researchers. Several software packages have been developed to aid researchers in the task of interpreting microarray-generated data. The Java-based Microarray Explorer (MAExplorer) is one such tool. Designed and implemented by the Laboratory of Experimental and Computational Biology, National Cancer Institute (NCI), in collaboration with the Laboratory of Genetics and Physiology, National Institute of Diabetes and Digestive and Kidney Diseases, it is still under development and officially in ‘beta’ status.

The MAExplorer web site provides full documentation (150-page manual), tutorials, web browser demo applets and the downloadable stand-alone version, which is a free Java application that may be redistributed without restriction. This stand-alone version of MAExplorer can run under Windows, Mac OS (8.1 or greater), Solaris and Linux. There is also a Java applet version available, which runs in a web browser. The applet requires a corresponding data site on a web server, such as the Mammary Genome Anatomy Project (MGAP) and mAdb sites at NCI.

MAExplorer was originally developed for MGAP, for use with a particular type of Research Genetics(TM) array. It can now handle other cDNA arrays with different geometries. MAExplorer can even construct a ‘pseudo’ array geometry for visualisation, given a set of gene intensities and a set of ‘locations’ (eg Incyte data). However, incorporation of new geometries is currently done manually. A Java tool (Cv2Mae) is being developed to aid users in the process by reading and repackaging microarray data files (commercial and one-of-a-kind) to a complete MAExplorer data format.

We tested the stand-alone version (0.89.31-beta, current as of 12th June, 2001) under Windows NT 4.0 on a Dell Optiplex (600 MHz Celeron processor, 512Mb RAM), following the tutorial and using the sample dataset provided (38 experiments, 3,072 clones/experiment). Downloading and installation were quick and clean.

The user begins by selecting a database of experiments (in MAExplorer terminology, hybridization probes or