BIOINFORMATICS

Mining gene expression data

The use of microarrays to monitor the transcription of thousands of genes under multiple conditions or in multiple cell lines is generating a massive and growing amount of valuable data. But there is a pressing need for more and better analysis tools. Two recent papers report new approaches and show how different methods of data mining can yield new information. Both papers use gene expression data related to cancer biology.

One way of analysing microarray data is to look for groups of genes whose expression patterns are similar across many experiments. The co-regulated genes within such clusters are often found to have related functions. Getz *et al.* started with the idea that some gene clusters might be masked by transcriptional 'noise' from genes outside the cluster, or if the genes are co-regulated in only a subset of the experiments. So the authors developed an algorithm called coupled two-way clustering that breaks down the total dataset into subsets of genes and samples that can reveal significant clusters.

Two previously published datasets were used by Getz *et al.* The first comprised 72 samples of two types of acute leukaemia — acute lymphoblastic leukaemia (ALL) and acute myeloid leukaemia (AML). After applying their analysis, they identified 84 clusters. One of the clusters (comprising 60 genes) separated the samples into AML and ALL. Another cluster (of 28 genes) split the AML patients into those who had received treatment and those who had not. The second dataset used by Getz *et al.* comprised 40 colon cancer samples and 22 controls. Their analysis was able to split the group into the normal and diseased samples using one of the clusters of genes, and another cluster partitioned the samples according to a difference in the methodology used for RNA preparation. Overall, the method does generate meaningful clusters that are not detected when the whole dataset is analysed. The task is now to examine the unexplained clusters to look for biological significance.

Butte *et al.* used an entirely different approach to mine data from two different datasets. The data concerned 60 cell lines established by the National Cancer Institute and used since 1989 to screen anticancer agents. The first dataset comprised the transcript levels for several thousand genes in each cell line. The second dataset comprised the GI50 (the level of anticancer agent required to achieve 50% growth inhibition) for several thousand agents on each cell line. The aim was to look for significant correlation between every possible pair of agents and genes. Correlations were summarized diagrammatically in networks and 202 such networks were found. Many expected associations were found between structurally related anticancer agents, and networks were also identified that linked genes of related function. Only one association was found between an agent and a gene — the GI50 for a thiazolidine carboxylic acid derivative increased with the expression of the gene *LCP1*, which encodes an actin-binding protein. Once again, the networks need to be analysed further to uncover the biological meaning. Among the advantages of this method are that individual genes or agents can be linked more than once, and that negative correlations can be found just as easily as positive correlations.

These two papers expand the range of tools for analysis of transcript profile data, and expose further seams for would be data-miners.

Mark Patterson

References and links

ORIGINAL RESEARCH PAPERS Getz, G. *et al.* Coupled two-way clustering analysis of gene microarray data. *Proc. Natl Acad. Sci. USA* **97**, 12079–12084 (2000) | Butte, A. J. *et al.* Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. *Proc. Natl Acad. Sci. USA* **97**, 12182–12186 (2000)

WEB SITES Computational physics group, Weizmann Institute | Molecular pattern recognition, Whitehead Institute

WEB WATCH

Homophila

Despite Homophila's spooky homepage, human geneticists curious to know what their disease gene does in *Drosophila* have nothing to fear. Ethan Bier and his colleagues at UCSD have compared the gene sequences entered in the Online Mendelian Inheritance in Man (OMIM) database to the genes, EST or genomic sequences in Flybase, the *Drosophila* sequence database. There is a story behind the creation of this web site. After finding that 74.5% of 909 distinct human disease genes have close homologues in fruitflies, the researchers scribbled the gene names and their corresponding syndromes on cards and handed them to a pathologist, who dutifully placed them into piles according to the nature of the disorder. This evolved into Homophila, a site where human disease genes and their fly homologues can be searched according to keyword, human disease name, gene name or OMIM entry number. For instance, typing in 'hypertension' leads to a results table showing, on the left, a description of the human disorder, the human gene symbol (for example, for the angiotensin II receptor) and related online references; on the right, the *Drosophila* protein sequence matches. In case you're daunted by the prospect of trundling through fruitfly data, a link from the fly gene of interest to GadFly (Genome Annotation Database of *Drosophila*) leads you to all that is known about the gene.

There's more to come, as the site curators promise a facility to match disease phenotypes that are common to humans and flies. As molecular signalling pathways are more completely described in flies compared with humans, human disease genes could be cloned on the basis of fly mutant phenotypes that are typical of a particular pathway. The site is updated monthly, so keep a lookout.

Tanita Casci