Analysis of large-scale gene expression data

Gavin Sherlock

The advent of cDNA and oligonucleotide microarray technologies has led to a paradigm shift in biological investigation, such that the bottleneck in research is shifting from data generation to data analysis. Hierarchical clustering, divisive clustering, self-organizing maps and k-means clustering have all been recently used to make sense of this mass of data.

Addresses
Department of Genetics, Stanford University Medical Center, 300 Pasteur Drive, Stanford, California 94306-5120, USA; e-mail: sherlock@genome.stanford.edu

Current Opinion in Immunology 2000, **12**:201–205

0952-7915/00/$ – see front matter

Abbreviation
SOM self-organizing map

Introduction

Microarrays [1–5] may generate tens of thousands of data points for every experiment performed. A study may consist of many experiments, such that in a single study the order of a million datapoints may be generated (e.g. see [6•]). Such volumes of data are too large to analyze by simple sorting in spreadsheets, or plotting on a single or few graphs. For sense to be made of the data, systematic methods for their organization are required. I review some of the recent developments in algorithms and tools for the analysis and organization of large-scale expression data, including clustering methods and methods for correlating expression data to other biological data.

Suitable metrics

Obviously, a metric to quantify whether two expression profiles are similar to each other is needed. In this regard, it is useful to consider the values that make up the expression profile for a single gene as a series of coordinates that define a vector. One distance metric that can thus be used is the Pearson correlation, which is essentially a measure as to how similar are the directions in which two expression vectors point. The Pearson correlation treats the vectors as though they were the same (unit) length, and is thus insensitive to the amplitude of changes that may be seen in the expression profiles. A second distance measure that can be used is the Euclidean distance, which measures the absolute distance between two points in space, which in this case are defined by two expression vectors. The Euclidean distance thus takes into account both the direction and the magnitude of the vectors. The Pearson correlation metric has been widely used and, as pointed out by Heyer *et al.* [7•], attributes high scores to expression patterns that are visually similar. However, the Pearson correlation may also give rise to false positives, that is, it may attribute an artificially high score to patterns that are not necessarily that similar. The investigation by Heyer *et al.* [7•] of this effect suggests that it may be caused by outliers, such that if the expression levels of two patterns are unrelated in all but one of the time points, and in that time point there is a significant peak or trough, then a high correlation may still result. They therefore propose a new correlation measure, called the jackknife correlation, that is robust to single outliers, thus reducing the number of false positives but continuing to give high scores to expression patterns that are similar over all conditions.

Agglomerative clustering

A useful approach to analyzing gene expression data has been the use of clustering [8••,9]. It is a bottom up (agglomerative) approach, whereby single expression profiles are successively joined to form nodes, which in turn are then joined further. The process continues until all individual profiles and nodes have been joined to form a single hierarchical tree. The advantage of this approach is that it is simple, and the end result can be easily visualized, from which coordinately regulated patterns can be relatively easily discerned by eye (Figure 1).

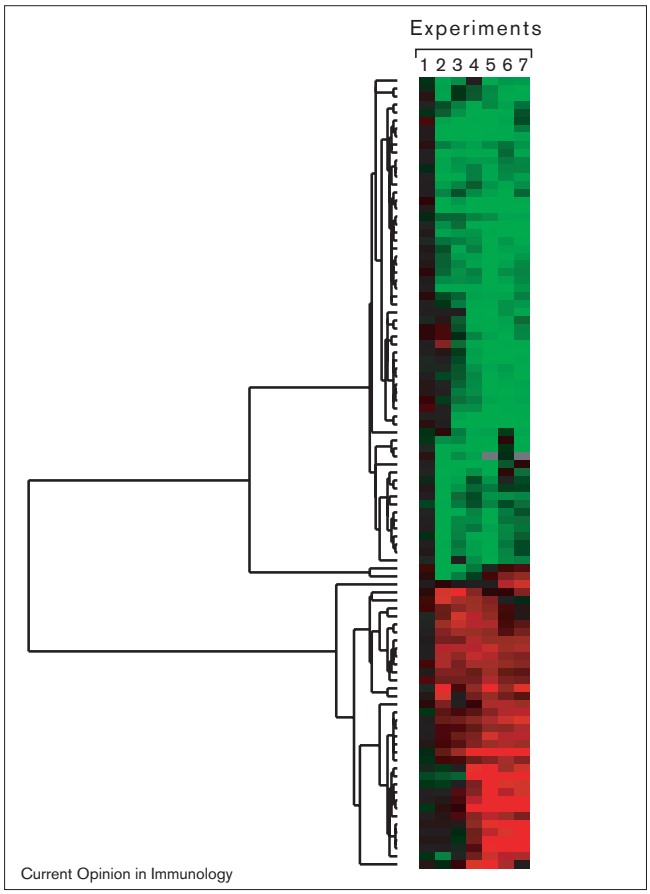
Divisive clustering

Recently, an alternative, divisive, clustering method has been applied to gene expression data [10••]. This approach is the opposite of that taken with the agglomerative method by Eisen *et al.* [8••], in that it could be considered top down, rather than bottom up. Two vectors are initialized randomly, and each gene is assigned to one of the two vectors using a probability function. The vectors are iteratively recalculated to form the centroids of two clusters. Each cluster is successively split in the same fashion, until each cluster consists of a single profile. The history of the data splitting is used to construct a binary tree. Alon *et al.* [10••] also include a node-switching algorithm to order the branches in a somewhat optimal manner that is similar, in concept, to the algorithm implemented in [11•] for agglomerative clustering. Alon *et al.* [10••] also introduce the notion of two-dimensional clustering, in which not only the genes but also the arrays are organized by clustering; that is, both the rows and columns of the expression data matrix are rearranged. If the arrays correspond to different cell types, this two-way clustering serves as a method for distinguishing cell types from one another. This notion is also used by Perou *et al.* [12] in their comparison of gene expression patterns in human mammary epithelial cells and breast cancers.

Partitioning of data

Hierarchical clustering can lead to artifacts. With the agglomerative method, as clusters become larger the expression profile that represents that cluster, which is the average of all profiles that belong to the cluster, may not

Figure 1



An example cluster. Each row corresponds to a single gene, and each column corresponds to a single array or experiment. The branch lengths indicate the correlation with which genes/nodes were joined, with longer branches indicating a lower correlation.

reflect accurately any of the contained profiles. Hence, the higher up in the tree one looks, the less relevant the genes within a cluster may be to each other. In addition, if a 'bad' decision is made early on during tree construction, it cannot be corrected later. The divisive method reasonably avoids these two problems, but both methods suffer a potential drawback when using two-way clustering. When clustering by arrays, the similarity between each array is typically calculated over the total number of genes within the dataset, and is therefore only an 'on average' measurement. If tissue A is most similar to tissue B on average, the fact that for a subset of genes tissue A is most similar to tissue C will be ignored. Hence, some very real biology may be discarded.

These problems may be avoided by first partitioning the data into reasonably homogeneous groups. These groups can then be individually clustered, in both dimensions if desired. Partitioning expression data before clustering is akin to organizing protein sequences on the basis of their similarity. An evolutionary tree would not be derived initially from a set of 6000 protein sequences; instead, the proteins would first be partitioned into smaller families depending on their BLAST scores [13], then clustered using a program such as CLUSTALW [14]. To partition expression data into groups, self-organizing maps, k-means clustering and the quality cluster algorithm have all been used [7•,15•,16,17].

Self-organizing maps

A self-organizing map (SOM) [18] has a series of partitions, each with a reference vector that contains the same number of data points as there are experiments being considered. The partitions are in a pre-defined geometrical configuration, such as a two-dimensional grid, and initially their reference vectors are random. To assign genes to the partitions, a gene is picked at random, and it is determined to which reference vector the gene's expression vector is most similar. That reference vector is then adjusted so that it is more similar to the randomly picked gene's expression vector. Then the reference vectors of each 'partition' that are close (on the two-dimensional grid) to the moved reference vector are also adjusted, so that they too are more similar to the gene's expression vector. These steps are repeated several thousand times, decreasing the amount by which reference vectors are adjusted and making the definition of 'close' (above) more stringent. Thus, fewer reference vectors are moved by smaller amounts as time goes on. Finally, the genes are mapped to the relevant partitions depending on which reference vector they are most similar to, thus partitioning the data.

K-means clustering

K-means clustering [19] is similar to the method of self-organizing maps, in that it uses partitions with reference vectors attached, but one partition does not directly influence another. k-means clustering may therefore be considered as one dimensional. First, the reference vectors are initialized randomly, and genes are partitioned to their most similar reference vector. Second, each reference vector is recalculated as the average of the genes that mapped to it. Last, these steps are repeated until convergence, that is, all genes map to the same partition on consecutive iterations.

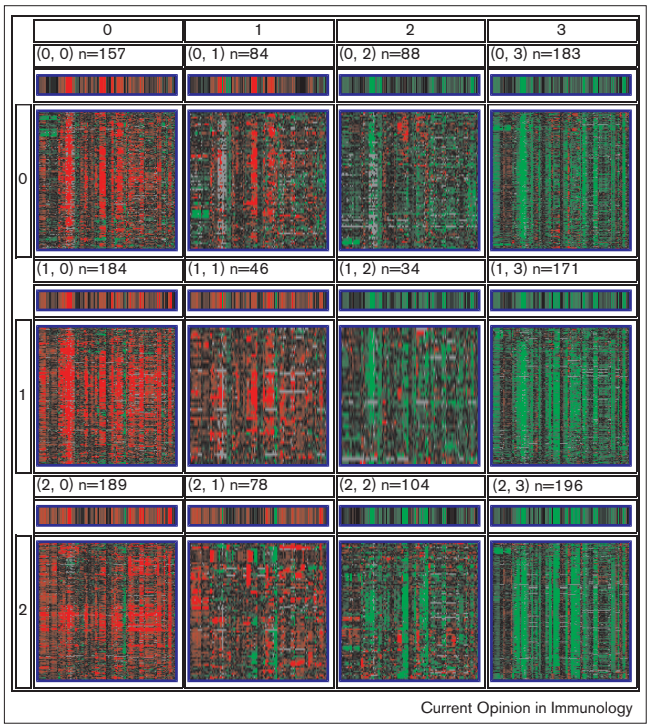
For both k-means and SOMs, it makes intuitive sense to further organize the members of each cluster, using, for instance, hierarchical clustering (as implemented in [11•]). Because the contents of each partition are reasonably homogenous, the drawbacks of hierarchical clustering are of little concern. Figure 2 shows 1500 yeast genes under various conditions separated into 12 partitions using a SOM. It is easy to see, using this display method [11•], that correlated expression profiles are put into the same partition, whereas anti-correlated ones are put into partitions in opposing corners of the grid. This partitioning effectively avoids many of the problems noted above that occur with hierarchical clustering. It should also be noted that in SOMs the partitions that are similar to each other are adjacent, which is in contrast to k-means clustering. This

reflects the manner in which the artificial expression vectors are manipulated to model the data. Because of this difference, SOMs tend to be able to model the complexity within a dataset somewhat more effectively than k-means clustering. It should be noted that while hierarchical clustering produces the same result every time with a given set of data (i.e., it is deterministic), both SOMs and k-means are nondeterministic, owing to the random initialization and, in the case of SOMs, owing to the random order in which genes are used to move the reference vectors. The partitioning is mostly the same each time, however, and could probably be made more robust by using the first two principle components of the data and vectors spanning the space between them, to seed the partitions initially, as has been suggested by Kohonen [18].

An important consideration is how many partitions to actually make. Currently, guesswork is needed from the researcher to decide how many significant patterns there may be in the data (but see [20] for discussion). A novel clustering technique proposed by Heyer *et al.* [7•] — the quality clustering algorithm — requires no such pre-definition of the number of clusters. First, the diameter of a cluster is defined as the lowest pairwise (jackknife) correlation between any of the genes that lie within that cluster, subtracted from 1. Second, a candidate cluster is formed by taking the first gene and adding to it the gene that minimizes the increase in cluster diameter. This process continues until no further genes can be added to the cluster without exceeding a defined diameter threshold. Third, a second candidate cluster is formed by starting with the second gene and repeating the process. All genes are available to this second cluster. Fourth, this process is repeated for each gene, such that there are as many candidate clusters as there are genes. At this point, the largest candidate cluster is selected and retained, and the genes that it contains are removed from consideration. Fifth, steps one to four are then repeated using the remaining set of genes. Last, this whole process repeats until some defined termination condition, for example, the largest cluster must contain a certain number of genes, which could be two or more.

This technique has some useful attributes. It avoids the problems associated with hierarchical clustering, while still being deterministic, and it also avoids the problem associated with k-means and SOMs of how many clusters to define. In addition, it can have orphan expression profiles that may not belong to any cluster, which is not a feature of k-means or SOMs (although it would be trivial to implement k-means and SOMs such that inclusion into a cluster required a correlation between a gene and the reference vector above a certain threshold). Of course, it should be noted that instead of setting the number of clusters, a diameter threshold must be set. As with setting the number of clusters, setting the diameter could be considered arbitrary, although setting the minimum allowed correlation within a cluster certainly has a more intuitive appeal.

Figure 2



A self-organizing map. Each partition may contain a different number of genes, although the image for each partition is the same size for display purposes. The contents of each partition have been rearranged by clustering. Red indicates induction of a gene and green indicates repression, and the magnitude of the change is indicated by the intensity of the color. The smaller-sized color bars associated with each partition indicate the final contents of the reference vector associated with that partition (see text for details).

In addition, the distribution of all pairwise correlations within the dataset can be used as a guide for determining the threshold.

Correlating expression data to other information

Obviously, the large quantities of expression data being generated do not exist in a vacuum; that is, there are already a great deal of non-expression data about the samples that are being investigated. These data may be of the form of gene annotation, or may, for instance, be clinical data about different cell lines, or may be experimental data relating how a microarray experiment was carried out. It is of clear interest and importance to develop tools to correlate such supporting data with the expression data. An obvious example is whether the expression data contain information that can be used as a diagnostic or prognostic tool. Golub *et al.* [21••] have done exactly this, analyzing the expression of 6817 human genes in 27 acute lymphoblastic leukemias and 11 acute myeloid leukemias. They set out to use these data to build a class predictor, which, when challenged with expression data from an acute leukemia sample of unknown origin, would be able

to accurately classify it into either of the previously seen classes, or potentially define a new third class. To do this, they first identified genes whose expression correlated to the class distinction to be predicted, and then tested whether the number of genes that appeared to have predictive value was greater than would be expected by chance. They identified roughly 1100 genes of such a nature, and conservatively (and somewhat arbitrarily) chose the 50 with the highest scores to be used as part of a predictor. With these genes in hand, they characterized an unknown sample as follows: each gene has a 'weighted vote' for whether it would assign the sample to one class or the other. The vote is weighted, such that how well the gene can distinguish the two classes is taken into account; in other words, a gene that can perfectly distinguish the two classes has a greater say than one that distinguishes them less well. The weighted votes for all the genes are then summed to see to which class the predictor would assign the new sample. In addition, the margin of victory allows a confidence to be attributed to the assignment. To test their technique, Golub *et al.* [21••] took their 38 samples, built a predictor from 37 of them, and used the predictor to assign the missing sample. They repeated this for all 38 samples. Their technique was remarkably robust in this cross validation test, correctly assigning 36 of the 38 samples, with the other 2 having uncertain assignment. On an independent set of 34 leukemias, the technique made strong and correct predictions about 29 of them. Furthermore, Golub *et al.* [21••] they showed that these predictions were robust using predictors derived from between 10 and 200 genes.

Tavazoie *et al.* [17], using clusters of genes generated by k-means clustering of the cell-cycle data of Cho *et al.* [22], devised a technique where they determined whether the functional categories into which the genes within each cluster fell was significant. To do this, they used the 199 functional categories curated by the MIPS database [23], and asked the question "is the overlap between the genes in a functional class and the genes in a particular expression cluster greater than would be expected by chance?", using the hypergeometric distribution statistic. Although they were able to show that some categories show significant overlap with some of the clusters, it should be noted that the number of clusters that they used for the k-means clustering was essentially a subjective number, and the significance scores would change with different numbers of clusters. The technique is however of obvious utility.

Conclusions and future prospects

Clearly, analysis of expression data is still in its infancy, and although many techniques of obvious utility exist, there are still a great many areas in which research will improve these tools and invent new ones. Development of visualization tools to allow the biologist to intuit information from the organized data are also of great importance. False color representations such as those used by Wen *et al.* [9] and Eisen *et al.* [8••] greatly aid interpretation of the data. The majority of the clustering techniques currently in use rely on similarity metrics that do not have a measure of whether such a correlation is higher than would be expected by chance, and hence have no indication of the significance of a cluster. 'Porting' of Karlin–Altschul like statistics [24] from BLAST to be used on gene expression data would be of obvious benefit, because what we really want to know is whether these patterns are more similar than we would expect by chance, as opposed to whether they are just similar.

In addition, algorithms that are akin to position-specific iterated (PSI)-BLAST [25] would also be useful for gene expression data. As the number of experiments being clustered increases, the likelihood of finding significant correlations decreases. Two genes may be similar across 30 arrays, but dissimilar across another 200. We don't want to lose the information about that former similarity, which we would capture if we had the gene expression equivalent of an all versus all PSI-BLAST program to find 'expression domains'. This would really help us get to the heart of the combinatorial nature of transcriptional control. Development of such tools, and their free availability — at least in the academic community — is paramount. With such tools in hand, great strides in the correlation of clinical, functional and promoter data with expression data will follow these initial promising steps.

Update

Since the original submission of this review, Brown *et al.* [26] described the application of support vector machines to microarray expression data. Support vector machines constitute a supervised computer learning method and were used by Brown *et al.* to predict the functions of genes of unknown function, based on the similar expression patterns to genes of known function.

Acknowledgements

The author would like to thank Paul Spellman in particular for reading and commenting on the manuscript. In addition all the members of the David Botstein and Patrick Brown laboratories have provided a stimulating and exciting environment in which many of these ideas were generated.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. DeRisi JL, Iyer VR, Brown PO: **Exploring the metabolic and genetic control of gene expression on a genomic scale.** *Science* 1997, **278**:680-686.
2. Lashkari DA, DeRisi JL, McCusker JH, Namath AF, Gentile C, Hwang SY, Brown PO, Davis RW: **Yeast microarrays for genome wide parallel genetic and gene expression analysis.** *Proc Natl Acad Sci USA* 1997, **94**:13057-13062.
3. Schena M, Shalon D, Heller R, Chai A, Brown PO, Davis RW: **Parallel human genome analysis: microarray-based expression monitoring of 1000 genes.** *Proc Natl Acad Sci USA* 1996, **93**:10614-10619.
4. Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MU, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H *et al.*: **Expression monitoring by hybridization to high-density oligonucleotide arrays.** *Nat Biotechnol* 1996, **14**:1675-1680.

5. Wodicka L, Dong H, Mittmann M, Ho MH Lockhart DJ: **Genome-wide expression monitoring in Saccharomyces cerevisiae.** *Nat Biotechnol* 1997, **15**:1359-1367.

6. • Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D Futcher B: **Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization.** *Mol Biol Cell* 1998, **9**:3273-3297.
Four independent sets of experiments (one from Cho *et al.* [22]) are analysed to provide a comprehensive picture of the cell-cycle-regulated transcripts in yeast.

7. • Heyer LJ, Kruglyak S, Yooseph S: **Exploring expression data: identification and analysis of coexpressed genes.** *Genome Res* 1999, **9**:1106-1115.
The authors propose a novel correlation measurement, the jackknife correlation, that reduces the likelihood of false positives. They also describe a new clustering method, which doesn't suffer from some of the problems associated with other partitioning methods.

8. •• Eisen MB, Spellman PT, Brown PO Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proc Natl Acad Sci USA* 1998, **95**:14863-14868.
The use of an agglomerative hierarchical clustering method is described, with a useful false color representation of the resulting tree based structures. In addition software, for both analysis and visualization, is described that is freely available to academic researchers at http://rana.stanford.edu/software.

9. Wen X, Fuhrman S, Michaels GS, Carr DB, Smith S, Barker JL, Somogyi R: **Large-scale temporal gene expression mapping of central nervous system development.** *Proc Natl Acad Sci USA* 1998, **95**:334-339.

10. •• Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ: **Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays.** *Proc Natl Acad Sci USA* 1999, **96**:6745-6750.
This paper introduces a divisive clustering method for the analysis of colon cancer data. The notion of two dimensional clustering is also introduced, as a method whereby different cell lines and tissues can be compared with one another.

11. • Clustering Software on World Wide Web URL: http://genome-www.stanford.edu/~sherlock/cluster.html
Software for hierarchical clustering, using a novel node-switching algorithm to somewhat optimize node order, and making self-organizing maps and k-means clustering (with automatic clustering of each of the partitions) is freely available to academic researchers. In addition, a web-based SOM and k-means viewing program is also available.

12. Perou CM, Jeffrey SS, van de Rijn M, Rees CA, Eisen MB, Ross DT, Pergamenschikov A, Williams CF, Zhu SX, Lee JC *et al.*: **Distinctive gene expression patterns in human mammary epithelial cells and breast cancers.** *Proc Natl Acad Sci USA* 1999, **96**:9212-9217.

13. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215**:403-410.

14. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucleic Acids Res* 1994, **22**:4673-4680.

15. • Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, Golub TR: **Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation.** *Proc Natl Acad Sci USA* 1999, **96**:2907-2912.
This paper introduces the use of self-organizing maps as a partitioning method for gene expression data.

16. Toronen P, Kolehmainen M, Wong G, Castren E: **Analysis of gene expression data using self-organizing maps.** *FEBS Lett* 1999, **451**:142-146.

17. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM: **Systematic determination of genetic network architecture.** *Nat Genet* 1999, **22**:281-285.

18. Kohonen T: *Self Organizing Maps.* Berlin: Springer; 1995.

19. Everitt B: *Cluster Analysis 122.* London: Heinemann; 1974.

20. Milligan GW, Cooper MC: **An examination of procedures for determining the number of clusters in a data set.** *Psychometrika* 1985, **50**:159-179.

21. •• Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA *et al.*: **Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.** *Science* 1999, **286**:531-537.
A technique for classifying tumor samples based on expression data is described. The technique is robust and also provides a measurement to indicate the confidence of the assignment.

22. Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, Wodicka L, Wolfsberg TG, Gabrielian AE, Landsman D, Lockhart DJ, Davis RW: **A genome-wide transcriptional analysis of the mitotic cell cycle.** *Mol Cell* 1998, **2**:65-73.

23. Mews HW, Frishman D, Gruber C, Geir B, Haase D, Kaps A, Lemcke K, Mannhaupt G, Pfeiffer F, Schuller C *et al.*: **MIPS: a database for genomes and protein sequences.** *Nucleic Acids Res* 2000, **28**:37-40.

24. Karlin S, Altschul SF: **Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes.** *Proc Natl Acad Sci USA* 1990, **87**:2264-2268.

25. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25**:3389-3402.

26. Brown MP, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, Ares M Jr, Haussler D: **Knowledge-based analysis of microarray gene expression data by using support vector machines.** *Proc Natl Acad Sci USA* 2000, **97**:262-267.