Thomas D. Wu
is a Senior Bioinformatics Scientist at Genentech, Inc., where he leads research efforts in developing and applying methods for analysing gene expression data.

Large-scale analysis of gene expression profiles

Thomas D. Wu

Received (in revised form): 20th November 2001

Abstract

The accumulation of DNA microarray data has now made it possible to use gene expression profiles to analyse expression data. A gene expression profile contains the expression data for a given gene over various samples, and can be contrasted with an expression signature, which contains the expression data for a single sample. Gene expression profiles are most revealing when samples are grouped appropriately, either by standard clinical or pathological categories or by categories discovered through cluster analysis techniques. Expression profiles can exist at various levels of abstraction, yielding information across various tissues or across diseases within a particular tissue. Hypothesis tests may be applied to expression profiles on a large scale to identify candidate genes of interest.

Keywords: *gene expression, expression profiles, DNA microarrays, bioinformatics, computational biology, data analysis*

INTRODUCTION

Although there are various technologies for measuring gene expression, the resulting data can essentially be thought of as a matrix of expression values. In this framework, as depicted in Figure 1, the vertical axis represents samples and the horizontal axis represents genes. The expression data themselves lie within the matrix, one value for each pairwise combination of gene and sample. To analyse these data, we can study either the rows of the matrix – which we call expression signatures – or the columns –

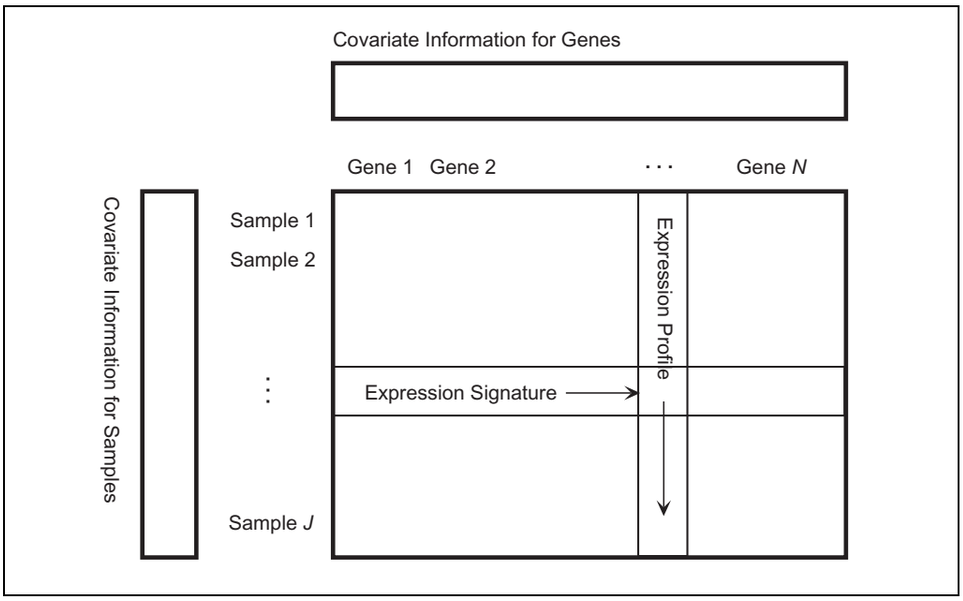

Figure 1: Schematic view of gene expression data. Expression values can be considered as a matrix where the vertical axis represents samples and the horizontal axis represents genes. Expression values for a given sample constitute an expression signature, and expression values for a given gene constitute an expression profile. Covariate information for samples and genes can be represented outside this matrix

Thomas D. Wu, MD, PhD,
Department of Bioinformatics,
Genentech Inc.,
1 DNA Way MS 93,
South San Francisco,
CA 94080,
USA

Tel: +1 650 225 5672
Fax: +1 650 225 5389
E-mail: twu@gene.com

which we call expression profiles. An expression signature represents the values for a single sample; an expression profile, for a single gene.

Early approaches to gene expression analysis emphasised gene clustering

For DNA microarrays,[1] which are the primary generators of gene expression data currently, data for the matrix are gathered one sample, or row, at a time, in the form of an expression signature. Because each microarray measures the expression of thousands of genes for a given sample, early expression value matrices tended to be short and wide. Accordingly, early approaches to gene expression analysis emphasised such methods as gene clustering,[2,3] in large part because the number of genes greatly exceeded the number of samples.

However, although gene clustering has been useful for understanding biology on a global scale, it is less applicable for identifying candidate genes. For that purpose, it appears more straightforward to use the methods of hypothesis testing.[4–6] In hypothesis testing, we attempt to determine whether a given gene is expressed differentially between two conditions. Such differentially expressed genes represent possible targets for therapeutic agents. For example, the anti-cancer drug Trastuzumab (Herceptin®) acts by binding to the protein product of the *HER2* gene, which exhibits over-expression in some breast tumours.[7]

Large numbers of samples have made possible the large-scale analysis of gene expression profiles

In order to perform hypothesis testing adequately, we require a sufficiently large number of samples in our expression value matrix. Over time, as large numbers of samples have been processed on DNA microarrays and organised into databases,[8,9] gene expression matrices have become sufficiently 'tall' to permit the large-scale analysis of gene expression profiles.

Note that we make an important distinction between expression signatures and expression profiles. Most existing work on DNA microarray analysis, even work that has used the term 'expression profile',[10] has focused essentially on what we call expression signatures, or the expression data of a single sample. The data for an expression signature can be derived from a single microarray hybridisation. This paper focuses instead on the expression of each gene across multiple samples. The data for such an expression profile derive from multiple microarray hybridisations and must be gathered computationally. Our interest in expression profiles leads to our data matrix in Figure 1 being arranged differently from the common orientation in the literature, where genes are represented on the vertical axis and samples on the horizontal axis.

In this paper, we show how large-scale analysis of gene expression profiles can be used to identify candidate genes. An initial consideration in constructing expression profiles is to choose an appropriate transformation of the data, if desired, and to perform appropriate scaling and normalisation procedures. Because an expression profile derives from multiple hybridisations, the issues of scaling and normalisation become especially relevant to our discussion. Once the data are normalised, we can then organise the samples in a way that facilitates our task of interest, a step that may require cluster analysis of samples. Then, in order to identify candidate genes, we can perform hypothesis testing to the groups of samples. The final list of candidate genes must be analysed by further bioinformatics techniques to deduce their function.

TRANSFORMATIONS, SCALING AND NORMALISATION

Preprocessing of the expression data is especially critical in profile analysis because each of the expression values in a given profile comes from a different microarray hybridisation. These hybridisations are typically performed by different people at different times, and may involve different laboratory conditions or sample preparations. In order to maximise the interpretability of a gene expression profile, we need to

Various problems with microarray data affect standardisation

standardise data across microarray hybridisations.

Unfortunately, such standardisation is often difficult because of various problems with microarray data. Specifically, these problems include: (1) the large dynamic range of microarray data; (2) expression values that are reported to be negative; (3) variability across microarrays; (4) non-linearity between observed intensity values and the underlying gene expression; and (5) variability that is dependent upon the expression level.

We can see these problems graphically if we compare the data from two hybridisations, as shown in Figure 2(a). Such a comparison may represent the results from two differently labelled

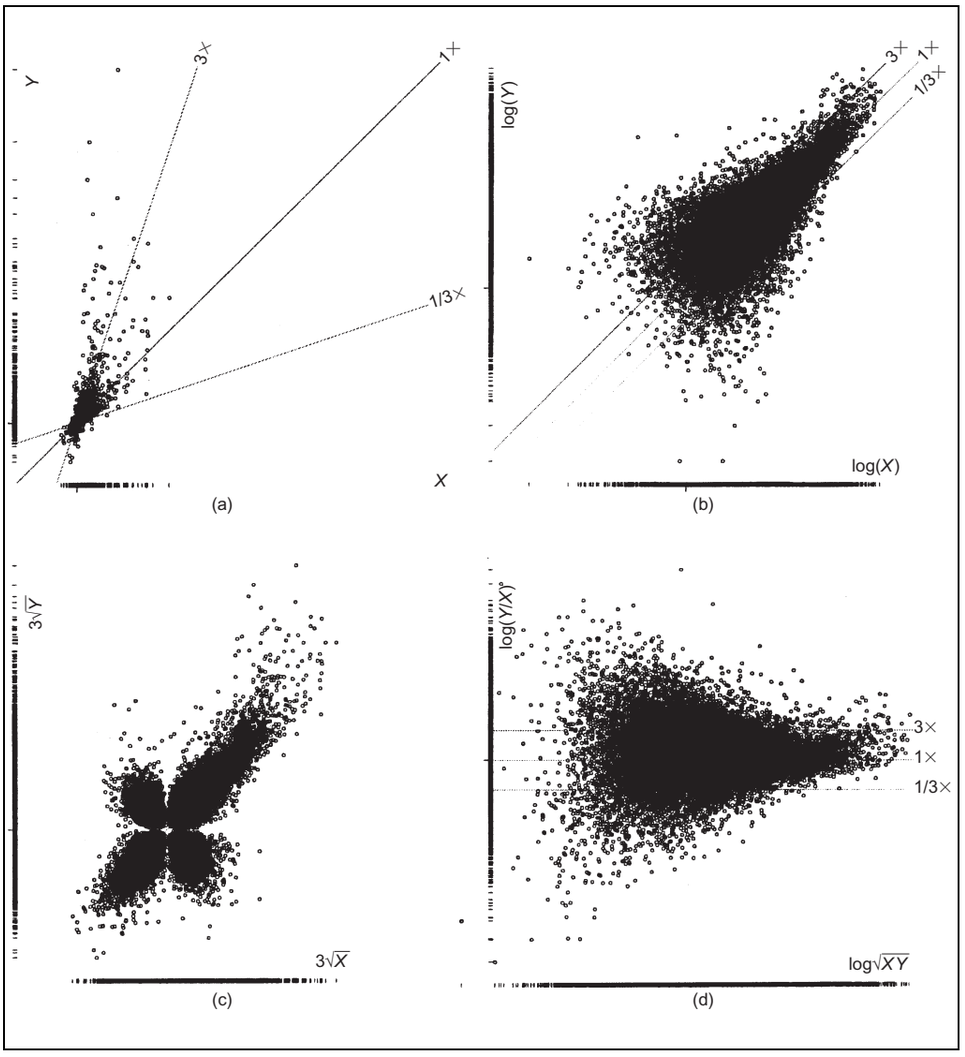

Figure 2: Scatterplot analysis of two sets of gene expression data. Scatterplot (a) shows the data on a standard scale. Scatterplot (b) shows the non-negative data on a logarithmic scale. Scatterplot (c) shows the data using a cube root transformation. Scatterplot (d) shows the log ratio as a function of the log geometric mean expression, and can be obtained by rotating scatterplot (b) by 45° clockwise. For scatterplots (a), (b) and (d), the sloped lines indicate ratios of a three-fold, one-fold and one-third-fold ratio. The three-fold and one-third-fold ratios are not shown for scatterplot (c), because they cannot be represented by a straight line. The scatterplots show that data are more variable at low-intensity values, an indication of heteroscedasticity

mRNA mixtures on a single microarray, as in the case of cDNA microarrays,[11] or the results from two distinct oligo or cDNA microarrays.

The graph shown here represents the latter case, with both samples having been hybridised to distinct oligo microarrays from Affymetrix Inc.[12,13] The scatterplot shows the enormous dynamic range of the microarrays, with the vast majority of genes exhibiting low expression values, and a smaller fraction exhibiting expression values at much higher orders of magnitude.

Microarray data have a large dynamic range

The large dynamic range causes problems in various ways. For example, if we try to fit a least-squares line to the scatterplot, the resulting line estimate will be affected almost entirely by the most extreme values. This behaviour of least-squares fitting is well known in the statistical literature, and one way to address this problem is to use a method for robust linear estimation.[14]

Another way to handle the dynamic range of microarray data is to transform the data, most commonly with a logarithmic scale. This scale spreads the data more evenly and reduces the leverage of the most extreme values. However, a logarithmic scale sometimes gives too much coverage to the low end of expression values, and it cannot handle negative values at all, which are reported in some microarray technologies, including those from Affymetrix. (The next release of Affymetrix Microarray Suite software will ensure that all reported expression values are positive.) To handle negative values, other transformations have been proposed, such as a cube root transformation.[15] This transformation reduces the effect of extreme values, although to a lesser extent than the logarithmic transformation. The cube root transform can also represent negative values. The logarithmic and cube root transformations are shown in Figures 2(b) and 2(c), respectively.

Microarray data can also exhibit non-linearity and non-constant variance (heteorscedasticity)

We would expect, and in practice generally find, that most genes are expressed at approximately the same level in two comparable samples, such that the data lie close to the one-fold ratio line. However, as we see in Figure 2(a), much of the data appear to deviate from this line. One reason for such a deviation is that the two microarrays are on different scales. If we wish to compare data across microarrays, we will need to place them on comparable scales, a process called scaling. The simplest approach to scaling is to multiply each set of microarray data such that their resulting means or medians equal some predetermined value, such as 100. Another more sophisticated approach would be to fit a straight line through the scatterplot and use the resulting slope for scaling.

However, in this particular case, the logarithmic and cube root plots show that the data do lie predominantly on the one-fold ratio line, except for a tail at the higher gene expression values. This indicates that the problem is one of non-linearity, rather than scaling. Such non-linearity can occur because of such phenomena as saturation of the observed intensities at high intensity levels. The degree of non-linearity can be assessed by rotating the scatterplot of Figure 2(b) by 45° clockwise. The resulting plot in Figure 2(d) shows the logarithm of the ratio on the vertical axis and the logarithm of the geometric mean of the expression values on the horizontal axis. If two sets of expression data are related linearly, the data would lie predominantly on a horizontal line, possibly shifted up or down if simple linear scaling were involved. On the other hand, if the data are related non-linearly, then the scatterplot would show a curve, such as the upward tail shown here.

Note that the rotated scatterplot demonstrates not only non-linearity, but also more variability at low intensities than at high intensities. Unfortunately, this feature of the data violates one of the basic assumptions underlying most statistical tests, namely, that the variability of the data is independent of the actual values of the data. In other words, most statistical tests, such as the *t*-test, assume

that the variance or standard deviation of the data is constant – the same for both low intensities and high intensities. But with the microarray data here, the variance decreases as the intensity increases.

This phenomenon of a non-constant variance is called *heteroscedasticity* in statistics, and its presence in microarray data means that we must interpret differences or ratios differently, depending on the overall intensity value. A two-fold increase at the high end of the range may be significant, but a two-fold increase at the low end may simply reflect noise. In order to address this issue, we can either try to standardise the data at this stage in order to equalise the variability, or defer the problem to the later step of hypothesis testing, by measuring the amount of variability and modifying the computations accordingly. Bioinformatics techniques for handling the issues of non-linearity and heteroscedasticity are being investigated actively.

Samples can be clustered using the same methods as for gene clustering

SAMPLE CLUSTERING

As the science of medicine has progressed, we have come to understand that diseases fall into categories and subtypes. We know, for example, that there are several different types of lung tumours. Our classification of disease to date relies primarily on clinicopathological grounds, particularly on visual characteristics, as determined by observations of gross anatomy or microscopic histology.

However, if we are interested in understanding disease at the molecular level, we need a corresponding classification of disease that is also molecular. Microarrays can also help in this regard, because they can measure an expression signature for each sample. We can therefore use these expression signatures as a molecular description of each sample, in place of the visual description used currently.[16,17]

Clustering samples by their tissue of origin yields a tissue profile

These gene expression signatures can be subjected to cluster analysis to discover subtypes of disease at the molecular level. There are various types of cluster analysis that have been used in microarray data analysis, including hierarchical clustering,[3,18–20] *k*-means analysis,[21–23] principal components analysis[24] and self-organising maps.[25,26] Cluster analysis has been performed primarily on expression profiles, for the task of classifying genes.[27] But for analysing expression profiles themselves, the most relevant cluster analysis would be that performed on samples through their expression signatures. The resulting classification of samples can then be used for subsequent hypothesis testing.

In order to cluster samples, we can apply the same methods that have been used for clustering genes.[28,29] However, for both types of cluster analysis, the results depend greatly upon the clustering method chosen. In particular, the results depend on the way in which distance or similarity is measured between samples, such as Pearson or Spearman correlation measures or Euclidean types of metrics. Furthermore, many clustering methods require that the user specify the number of clusters in advance, and this choice will also determine the final results. The task of cluster analysis is especially difficult because it is difficult or impossible to know whether a given clustering is correct. In some applications, we may have a known classification to consult, to see whether a given clustering is correct. But for determining the molecular classification of disease, we will probably not have such a gold standard to rely on.

Although ideally we would like to classify samples on the basis of their molecular characteristics, clinicopathological descriptions are nevertheless revealing in many cases. As an overview, we may simply classify samples according to their organ or system of origin, and according to whether the sample is normal or diseased. We call this representation a *tissue profile*, and we show an example in Figure 3. In a tissue profile, we show each sample as a separate point, and the points are grouped according to the chosen classification.

We can also classify samples at a more

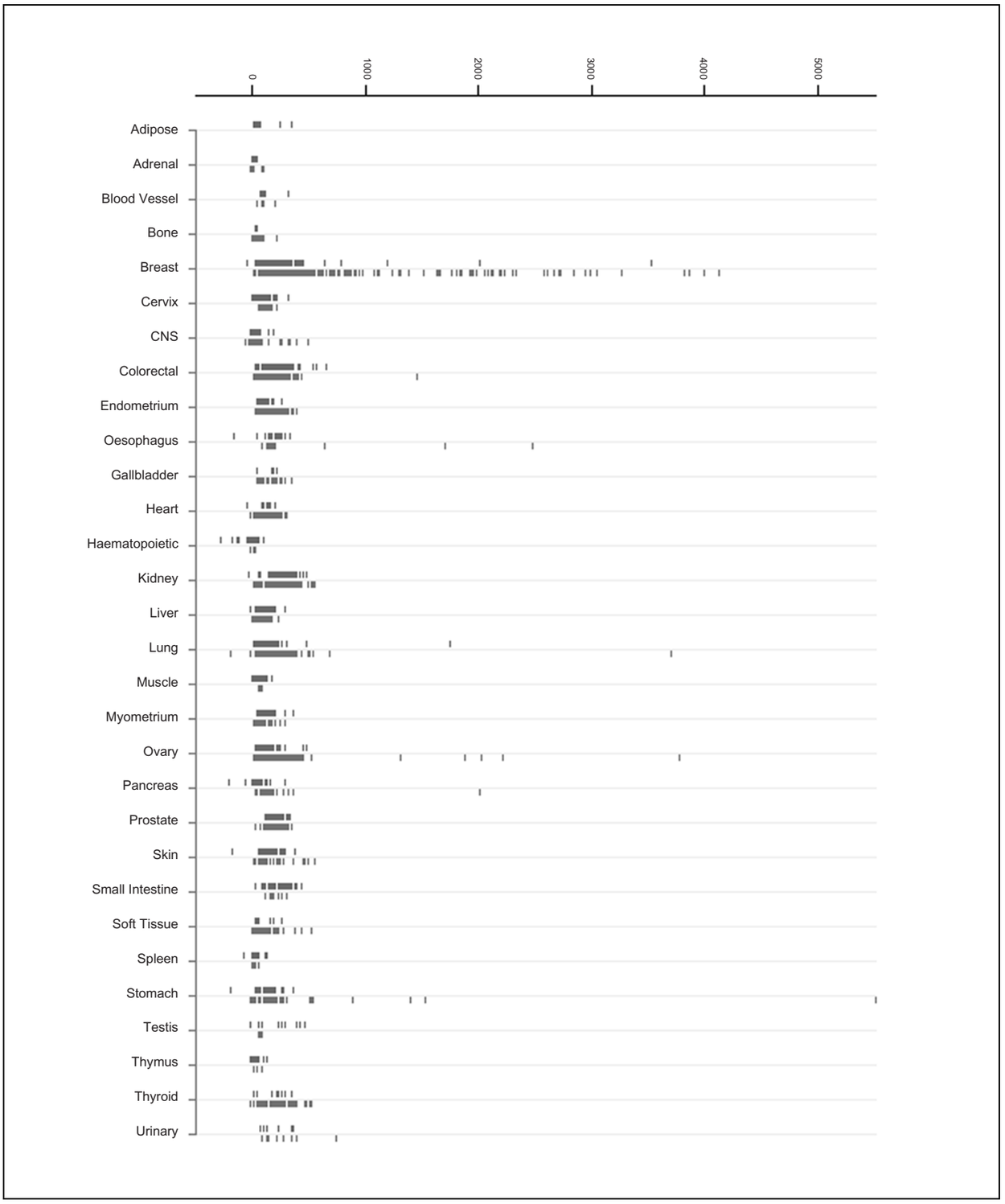

Figure 3: Tissue profile for the *HER2* gene. This figure shows the expression profile of the *HER2* gene in 3,417 samples of human tissue. The expression data were obtained from hybridisations onto Affymetrix GeneChip® probe arrays. The horizontal axis measures the average difference metric produced by Affymetrix Microarray Suite 4.0 software. The expression values are organised according to tissue, and further subdivided into normal and diseased samples, shown respectively as the points above and below each horizontal line. The data show that *HER2* is over-expressed primarily in diseased breast tissue

Clustering samples by their disease category yields a disease profile

detailed level. For instance, breast samples can be classified according to their disease category. When tissue samples are organised in this way, such as in Figure 4, we can often gain a more refined understanding of how gene expression differs among different disease types. We call this display a *disease profile*. A disease profile can be used either for mining purposes or for a retrospective analysis of a given candidate gene.[30]

The way in which we classify samples reflects the overall goal of the analysis. For example, if we were interested in discovering markers for atherosclerosis, we might classify samples according to their degree of atherosclerotic disease.[31] Likewise, if we were interested in markers of inflammation, we would create separate categories for inflamed samples and their normal counterparts.[32,33]

HYPOTHESIS TESTING

The tissue and disease profiles shown in the previous section can often reveal promising candidates by visual inspection.

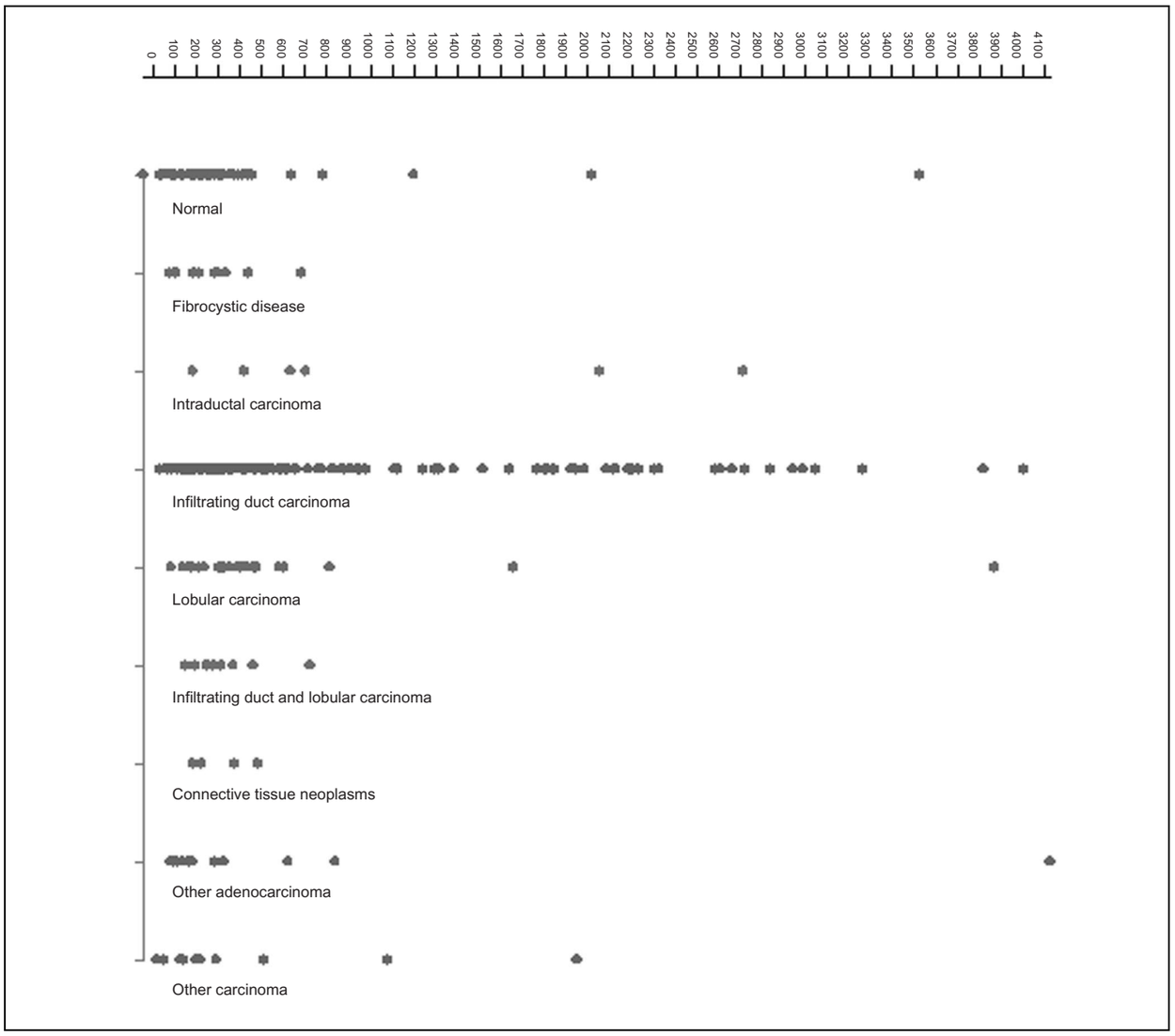

Figure 4: Disease profile for the *HER2* gene in breast samples. This figure shows the expression profile of *HER2* gene in 389 samples of human breast tissue. The samples are organised according to pathological disease categories. The data show that *HER2* is over-expressed primarily in infiltrating duct carcinomas

However, in order to evaluate thousands of gene candidates, automated methods are needed. In particular, an automated method must measure how different the gene expression is in two or more conditions.

Replicate samples permit the measurement of variability needed for hypothesis testing

One way to carry out this measurement is hypothesis testing. Hypothesis testing not only compares the difference between two groups, but also accounts for the underlying variability that is observed. In order to measure the amount of variability, replicated samples for each of the groups are needed.[34] The amount of variability observed in these replicates will affect how we interpret any differences between the groups. If the variability observed is small, then the difference becomes more statistically significant. Conversely, if the variability observed is large, then an observed difference becomes less significant.

There are standard formulae in statistics that account for both the observed difference and variability. One well-known measure is the *t* statistic, in which the difference in the group means is divided by the pooled standard deviation in the two groups. If we wish to extend the analysis to more than two groups, the analogous measure is the *F* statistic. For either statistic, we can quantify its significance as a *P* value, and therefore identify candidate genes as those with a sufficiently significant *P* value.

Inappropriate clustering of samples can lead to an overestimate of variability (*overdispersion*)

One of the driving forces for classifying samples into appropriate subtypes is the fact that hypothesis testing depends critically on the measurement of variability, which in turn depends on identifying homogeneous subsets of samples. If we did not classify our samples appropriately but rather pooled a heterogeneous group of samples together, we would observe an artificially high amount of variability. This phenomenon is known in statistics as *overdispersion*, and it represents a particular problem for hypothesis testing.[35] Most standard statistical tests, such as the *t*-test and *F*-test, assume that values in the different groups are distributed according to a normal, or Gaussian, distribution. The presence of heterogeneity violates this assumption, and we must therefore find alternatives to the standard tests or process the data to satisfy the assumption. The most appropriate type of statistical test or data processing for expression profiles remains a subject for further bioinformatics research.

Another avenue for research involves how best to use paired samples. Often, we have both normal and tumour tissue samples from the same donor, and such paired samples should be ideal for data analysis because they essentially eliminate the problem of donor-to-donor variation. Classical statistical theory has tests for such situations, such as a paired *t*-test, but such tests have not been applied widely in microarray data analysis.

Finally, large-scale applications of hypothesis testing face the problem of multiple hypothesis testing.[36] When we apply a given statistical test repeatedly to different data, even if those data are random, we are bound to obtain some fraction of results that appear to be statistically significant. This problem applies especially to gene expression analysis, where we may apply a given statistical test to thousands of expression profiles. A statistical significance cutoff that would be adequate for a single hypothesis test, such as $P < 0.01$, would yield many false positive results when the test is applied repeatedly. The simplest way to handle this problem is to adjust the *P* threshold by a Bonferroni correction, but this method proves to be too conservative. There are other approaches to this problem, including stepdown adjustments of *P* values,[37] false discovery rates[38] and resampling approaches,[39] and these approaches are being applied to microarray data analysis.

DETERMINING GENE FUNCTION

Hypothesis testing typically yields a set of candidate cDNA fragments or oligomers from a microarray. Because these fragments represent only part of a given gene, a further round of sequence analysis

is required to make sense of expression data. The goal of this sequence analysis is to determine more fully the function of the candidate gene, a task sometimes called *functional annotation*.

Candidate genes can be characterised further by sequence analysis

In sequence analysis, we begin by trying to extend a cDNA fragment or oligomer to the full-length gene. Extensions can be performed at the transcriptional level, with expressed sequence tags (ESTs) or assemblies of ESTs, or at the genomic level, with genomic sequence. In both cases, we would like ultimately to identify the protein product, since most bioinformatics programs for identifying function work best at the protein level. For EST data, the most likely candidate for the protein should be a long open reading frame (ORF), but errors in sequencing can sometimes complicate the process. For genomic data, we would have to make a prediction of the intron–exon structure in order to infer the correct protein.

Protein sequences can then be subjected to various bioinformatics programs to determine their function. Many such programs exist, including PFAM,[40] BLOCKS,[41] EMOTIF[42] and EMATRIX.[43,44] The decision to proceed on a particular gene as a therapeutic target can hinge upon a proper characterisation of protein function. If the protein has been studied previously, information from the literature may also prove useful.

Gene expression data from different parts of a gene may be compared

The full-length gene can also be used to discover related cDNA fragments or oligomers on the microarray. Many microarrays will represent different parts of a gene, sometimes because of a deliberate attempt to measure various splice forms, and sometimes because the full-length gene was simply not known at the time of microarray design. These other cDNA fragments or oligomers can then be studied to corroborate or contradict the initial expression data. When the expression profiles conflict, there may be evidence of alternative splice forms with differing expression patterns.

DISCUSSION

Gene expression data provide a comprehensive and high-throughput method for identifying candidate genes for further investigation. However, analysis of the data requires a new set of bioinformatics approaches and methods. Bioinformatics has traditionally dealt with sequence information, which can be viewed computationally as strings of characters. In contrast, expression data consist primarily of matrices of numbers. Whereas sequence-based bioinformatics has relied upon the fact that the genome is essentially the same across cells and among individuals, expression-based bioinformatics must deal with the fact that expression patterns vary according to cell, disease and environment. The raw sequence data for traditional bioinformatics has a relatively low rate of errors. On the other hand, expression data have a relatively high degree of experimental variability.

Accordingly, bioinformatics approaches to analysing expression data are still in the developmental stage and they are evolving as the type and amount of expression data change. In this paper, we have outlined some approaches to analysing gene expression data based on gene expression profiles, which have become practical only somewhat recently through the accumulation of a large volume of DNA microarray data. Of course, in order to handle such large volumes of data, we also require a significant amount of bioinformatics infrastructure,[45,46] including databases and visualisation and analysis tools.

Recently, the microarray community has been making proposals to pool microarray data into a common, shared database.[47] The most well known of these proposals has been that of the MGED (Microarray Gene Expression Database) group. One of the possible benefits of such a common database would be the construction of gene expression profiles such as those described in this paper. However, as we have discussed, the success of such an endeavour will depend

greatly on our ability to interpret expression levels uniformly across different microarray experiments.

In our experience, microarray data can provide promising leads, but interpreting them requires careful consideration. The results obtained from microarray data depend in large part upon the methods used in their processing and analysis. When they are analysed appropriately, gene expression data can provide us with a new level of insight into cell and molecular biology. Gene expression profiles constitute one approach for analysing large amounts of gene expression data and provide a methodology for investigating biological function on a global scale.

Acknowledgments
I would like to thank Hilary Clark, David Eberhard, William Forrest, Steve Guerrero, Kenneth Hillan, Michael Ostland, Paul Polakis, Victoria Smith, Michael Ward, Mickey Williams, William Wood and Zemin Zhang for stimulating discussions about sample classification and microarray data analysis.

References

1. Lockhart, D. J. and Winzeler, E. A. (2000), 'Genomics, gene expression and DNA arrays', *Nature*, Vol. 405, pp. 827–836.
2. DeRisi, J. L., Iyer, V. R. and Brown, P. O. (1997), 'Exploring the metabolic and genetic control of gene expression on a genomic scale', *Science*, Vol. 278, pp. 680–686.
3. Eisen, M. B., Spellman, P. T., Brown, P. O. and Botstein, D. (1998), 'Cluster analysis and display of genome-wide expression patterns', *Proc. Natl Acad. Sci. USA*, Vol. 95, pp. 14863–14868.
4. Golub, T. R., Slonim, D. K., Tamayo, P. *et al.* (1999), 'Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring', *Science*, Vol. 286, pp. 531–537.
5. Jin, H., Yang, R., Awad, T. A. *et al.* (2001), 'Effects of early ACE inhibition on cardiac gene expression following acute myocardial infarction', *Circulation*, Vol. 103, pp. 736–742.
6. Wu, T. D. (2001), 'Analysing gene expression data from DNA microarrays to identify candidate genes', *J. Pathol.*, Vol. 195, pp. 53–65.
7. Mendelsohn, J. and Baselga, J. (2000), 'The EGF receptor family as targets for cancer therapy', *Oncogene*, Vol. 19, pp. 6550–6565.
8. Ringwald, M., Eppig, J. T., Kadin, J. A. and Richardson, J. E. (2000), 'GXD: a Gene Expression Database for the laboratory mouse: Current status and recent enhancements', *Nucleic Acids Res.*, Vol. 28, pp. 115–119.
9. Scherf, U., Ross, D. T., Waltham, M. *et al.* (2000), 'A gene expression database for the molecular pharmacology of cancer', *Nature Genet.*, Vol. 24, pp. 236–244.
10. Hughes, T. R., Marton, M. J., Jones, A. R. *et al.* (2000), 'Functional discovery via a compendium of expression profiles', *Cell*, Vol. 102, pp. 109–126.
11. Schena, M., Shalon, D., Davis, R. W. and Brown, P. O. (1995), 'Quantitative monitoring of gene expression patterns with a complementary DNA microarray', *Science*, Vol. 270, pp. 467–470.
12. Fodor, S., Rava, R., Huang, X. *et al.* (1993), 'Multiplexed biochemical assays with biological chips', *Nature*, Vol. 364, pp. 555–556.
13. Lipshutz, R. J., Fodor, S. P. A., Gingeras, T. R. and Lockhart, D. J. (1999), 'High density synthetic oligonucleotide arrays', *Nature Genet.*, Vol. 21, pp. 20–24.
14. Weisberg, S. (1985), 'Applied Linear Regression', 2nd edn, John Wiley and Sons, New York.
15. Tusher, V. G., Tibshirani, R. and Chu, G. (2001), 'Significance analysis of microarrays applied to the ionizing radiation response', *Proc. Natl Acad. Sci. USA*, Vol. 98, pp. 5116–5121.
16. Perou, C. M., Jeffrey, S. S., van de Rijn, M. *et al.* (1999), 'Distinctive gene expression patterns in human mammary epithelial cells and breast cancers', *Proc. Natl Acad. Sci. USA*, Vol. 96, pp. 9212–9217.
17. Perou, C. M., Sørlie, T., Eisen, M. B. *et al.* (2000), 'Molecular portraits of breast tumours', *Nature*, Vol. 406, pp. 747–752.
18. Jain, A. K. and Dubes, R. C. (1988), 'Algorithms for Clustering Data', Prentice Hall, Englewood Cliffs, NJ.
19. Chu, S., DeRisi, J., Eisen, M. and Mulholland, J. (1998), 'The transcriptional program of sporulation in budding yeast', *Science*, Vol. 282, pp. 699–705.
20. Eickhoff, H., Schuchhardt, J., Ivanov, I. *et al.* (2000), 'Tissue gene expression analysis using arrayed normalized cDNA libraries', *Genome Res.*, Vol. 10, pp. 1230–1240.
21. Hartigan, J. A. and Wong, M. A. (1979), 'A *K*-means clustering algorithm', *Appl. Statistics*, Vol. 28, pp. 100–108.
22. Tavazoie, S., Hughes, J. D., Campbell, M. J.

et al. (1999), 'Systematic determination of genetic network architecture', *Nature Genet.*, Vol. 22, pp. 281–285.

23. Herwig, R., Poustka, A. J., Müller, C. *et al.* (1999), 'Large-scale clustering of cDNA fingerprinting data', *Genome Res.*, Vol. 9, pp. 1093–1105.
24. Mardia, K. V., Kent, J. T. and Bibby, J. M. (1979), 'Multivariate Analysis', Academic Press, San Diego, CA.
25. Kohonen, T. (1997), 'Self-organizing Maps', Springer, New York.
26. Tamayo, P., Slonim, D., Mesirov, J. *et al.* (1999), 'Interpreting patterns of gene expression with self-organising maps: Methods and application to hematopoietic differentiation', *Proc. Natl Acad. Sci. USA*, Vol. 96, pp. 2907–2912.
27. Brazma, A. and Vilo, J. (2001), 'Gene expression data analysis', *FEBS Lett.*, Vol. 480, pp. 17–24.
28. Alon, U., Barkai, N., Notterman, D. A. *et al.* (1999), 'Broad patterns of gene expression revealed by clustering analysis of tumour and normal colon tissues probed by oligonucleotide arrays', *Proc. Natl Acad. Sci. USA*, Vol. 96, pp. 6745–6750.
29. Alizadeh, A. A., Eisen, M. B., Davis, R. E. *et al.* (2000), 'Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling', *Nature*, Vol. 403, pp. 503–511.
30. Gerritsen, M. E., Peale, F. and Wu, T. D. (2002), 'Gene expression profiling in silico: Relative expression of candidate angiogenesis-associated genes in renal cell carcinomas', *Exp. Nephrol.*, Vol. 10, in press.
31. McCaffrey, T. A., Fu, C., Du, B. *et al.* (2000), 'High-level expression of *Egr-1* and *Egr-1*-inducible genes in mouse and human atherosclerosis', *J. Clin. Invest.*, Vol. 105, pp. 653–662.
32. Heller, R. A., Schena, M., Chai, A. *et al.* (1997), 'Discovery and analysis of inflammatory disease-related genes using cDNA microarrays', *Proc. Natl Acad. Sci. USA*, Vol. 94, pp. 2150–2155.
33. Kaminski, N., Allard, J. D., Pittet, J. F. *et al.* (2000), 'Global analysis of gene expression in pulmonary fibrosis reveals distinct programs regulating lung inflammation and fibrosis', *Proc. Natl Acad. Sci. USA*, Vol. 97, pp. 1778–1783.
34. Lee, M. L. T., Kuo, F. C., Whitmore, G. A. and Sklar, J. (2000), 'Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations', *Proc. Natl Acad. Sci. USA*, Vol. 97, pp. 9834–9839.
35. Rice, J. A. (1995), 'Mathematical Statistics and Data Analysis', 2nd edn, Duxbury Press, Pacific Grove, CA.
36. Shaffer, J. P. (1995), 'Multiple hypothesis testing', *Ann. Rev. Psychol.*, Vol. 46, pp. 561–584.
37. Lunneborg, C. E. (2000), 'Data analysis by resampling: Concepts and applications', Duxbury Press, Pacific Grove, CA.
38. Benjamini, Y. and Hochberg, Y. (1995), 'Controlling the false discovery rate: A practical and powerful approach to multiple testing', *J. R. Stat. Soc. Ser. B*, Vol. 57, pp. 289–300.
39. Westfall, P. H. and Young, S. S. (1993), 'Resampling-based Multiple Testing: Examples and Methods for *p*-value Adjustment', Wiley, New York.
40. Bateman, A., Birney, E., Durbin, R. *et al.* (2000), 'The Pfam protein families database', *Nucleic Acids Res.*, Vol. 28, pp. 263–266.
41. Henikoff, J. G., Greene, E. A., Pietrokovski, S. and Henikoff, S. (2000), 'Increased coverage of protein families with the Blocks database servers', *Nucleic Acids Res.*, Vol. 28, pp. 228–230.
42. Nevill-Manning, C. G., Wu, T. D. and Brutlag, D. L. (1998), 'Highly specific protein sequence motifs for genome analysis', *Proc. Natl Acad. Sci. USA*, Vol. 95, pp. 5865–5871.
43. Wu, T. D., Nevill-Manning, C. G. and Brutlag, D. L. (1999), 'Minimal-risk scoring matrices for sequence analysis', *J. Comput. Biol.*, Vol. 6, pp. 219–235.
44. Wu, T. D., Nevill-Manning, C. G. and Brutlag, D. L. (2000), 'Fast probabilistic assessment of sequence function using scoring matrices', *Bioinformatics*, Vol. 16, pp. 233–244.
45. Ermolaeva, O., Rastogi, M., Pruitt, K. D. *et al.* (1998), 'Data management and analysis for gene expression arrays', *Nature Genet.*, Vol. 20, pp. 19–23.
46. Aach, J., Rindone, W. and Church, G. M. (2000), 'Systematic management and analysis of yeast expression data', *Genome Res.*, Vol. 10, pp. 431–445.
47. Brazma, A., Robinson, A., Cameron, G. and Ashburner, M. (2000), 'One-stop shop for microarray data', *Nature*, Vol. 403, pp. 699–700.