BIOINFORMATICS
Vol. 17 Suppl. 1 2001
Pages S115–S122

GEST: a gene expression search tool based on a novel Bayesian similarity metric

Lawrence Hunter[1], *Ronald C. Taylor*[1], *Sonia M. Leach*[1] *and Richard Simon*[2]

[1]*Center for Computational Pharmacology, Department of Pharmacology, School of Medicine, C236, University of Colorado Health Sciences Center, 4200 E. Ninth Avenue, Denver CO, 80206, USA and* [2]*Biometric Research Branch, National Cancer Institute, 9000 Rockville Pike, Executive Plaza North, Bethesda MD, 20892, USA*

Received on February 6, 2001; revised and accepted on March 30, 2001

ABSTRACT

Gene expression array technology has made possible the assay of expression levels of tens of thousands of genes at a time; large databases of such measurements are currently under construction. One important use of such databases is the ability to search for experiments that have similar gene expression levels as a query, potentially identifying previously unsuspected relationships among cellular states. Such searches depend crucially on the metric used to assess the similarity between pairs of experiments. The complex joint distribution of gene expression levels, particularly their correlational structure and non-normality, make simple similarity metrics such as Euclidean distance or correlational similarity scores suboptimal for use in this application. We present a similarity metric for gene expression array experiments that takes into account the complex joint distribution of expression values. We provide a computationally tractable approximation to this measure, and have implemented a database search tool based on it. We discuss implementation issues and efficiency, and we compare our new metric to other standard metrics.

Contact: larry.hunter@uchsc.edu

INTRODUCTION

The advent of high throughput gene expression assays have made it possible to quantitatively assess the expression levels of tens of thousands of genes at a time. The results of such assays are being accumulated in databases, both public and private, and these databases are growing quickly. Although publicly available gene expression databases are still relatively small, there are plans to generate very large collections of expression profiles, see, e.g., (Abbott, 1999). When such databases become available, algorithms for searching these databases will grow in significance.

These databases embody information about the large scale transcriptional response of a range of tissues from different organisms to various growth conditions, genetic perturbations, drug treatments, and other experimental manipulations. The phenomenology of these transcriptional responses is still largely unknown. One valuable method for unraveling the significance of these measurements is to determine the conditions under which similar expression profiles are generated. As (Bassett *et al.*, 1999) observed, "expression profiles can be aligned with one another to identify similar cellular responses."

Gene expression is a complex process, and searching databases of expression measurements to identify related cellular states is a nontrivial task. In this paper, we explore the importance of the similarity metric used for gene expression database search and argue that traditional metrics fail to address some important characteristics of the data with respect to identifying cellular states. We propose a novel similarity metric we believe to be better suited to this task.

Searching gene expression databases

In the metaphor suggested in (Bassett *et al.*, 1999), searching expression databases can be compared to searching the sequence databases. The goal is to find similarities that are biologically significant. In comparing polypeptide sequences, not all differences are treated equally; we use substitution cost matrices to identify differences which are (or are not) likely to be biologically meaningful. It is entirely possible for a database sequence to be more similar to a query than another database sequence that actually has more identical residues, because the differences in the first sequence are more conservative than the differences in the second. The substitution costs used to calculate these similarities are derived from empirical estimates of the likelihood of seeing particular substitutions in homologous proteins. To carry the metaphor back to the gene expression comparison, it would be desirable to find a similarity met-

ric for gene expression data that reflected the biological significance of the observed differences in a manner analogous to the substitution cost matrices in sequence searching.

Although expression assay techniques vary significantly, all generate quantitative information about large numbers of genes. A database of such experiments can be generally conceived of as a table of numbers with as many rows as there are experiments, and as many columns as there are genes. In that conception, the gene expression search task is: given a particular row in the table, find other rows that are "similar" to it. In order for such searches to be useful in understanding the transcriptional responses of tissues, the similarity metric over the transcript levels must be biologically sound.

There are a variety of technical and methodological challenges to ensuring that data gathered by different laboratories is comparable at all, including sensitivity to the exact probes used, the environment in which the probes are applied, and particularly the control against which they are compared. We do not address these issues here, other than to note that it would be of great community value for investigators using spotted arrays to publish the results of a hybridization of their internal reference to some universal reference, since such a result could be used to indirectly compute the ratio of each gene on the array to the universal reference.

Similarity metrics

Given the above characterization of the gene expression database as a table, it is natural to think of the elements of the database as points in k-dimensional space, where k is the number of genes assayed in the experiments. The similarity metric we need is then a function of two k-dimensional vectors. Although many different vector distance[†] metrics have been proposed (see, e.g., (Wilson & Martinez, 1997)), most applications in continuous spaces use either the Euclidean distance function or some type of correlational similarity function (see Eq. 1 and Eq. 2, below).

$$\text{EuclideanDistance}(X, Y) = \sqrt{\sum_{g \in \text{genes}} (X_g - Y_g)^2} \qquad (1)$$

$$\text{CorrelationSimilarity}(X, Y) = \frac{\sum_{g \in \text{genes}} [(X_g - \overline{X_g}) \cdot (Y_g - \overline{Y_g})]}{\sqrt{\sum_{g \in \text{genes}} (X_g - \overline{X_g})^2 \cdot \sum_{g \in \text{genes}} (Y_g - \overline{Y_g})^2}} \qquad (2)$$

[†] Distance metrics can be trivially transformed into similarity metrics by inversion.

where $\overline{X_g}$ is the mean of the gene expression values in experiment X.

Previous work has generally compared pairs of genes using a set of experiments, rather than pairs of experiments using a set of genes, but the issues are similar. For example, (Eisen *et al.*, 1998) rescale the log gene expression ratios to mean 0 and variance 1 and use normalized dot product as a similarity metric for pairs of genes over a set of experiments, and then uses those distances as the basis for a hierarchical clustering.

One variant metric takes the absolute value or square of each element in the sum of the correlation (or dot product) metric so that anticorrelated genes contribute positively to similarity. Another variant, the Mahalanobis distance, uses the covariance matrix to adjust the contributions of differences for each gene based on the variance observed for that gene. In most applications of the Mahalanobis distance, the covariance matrix is assumed to be diagonal, so that it effectively normalizes by the variance of each gene independently.

Qualities of gene expression data

In order to assess the applicability of the above similarity metrics to gene expression database searching, it is important to understand some of the formal characteristics of the data. Most important for this purpose are: variability and noise, correlational structure, and the distributional form of the expression data. These characteristics can in part be deduced from existing understanding of biological systems, and can also be estimated from the existing expression data.

The following empirical calculations use a publicly available collection of yeast expression data. Transcript levels for 6024 genes are measured in a combined set of 92 experiments: 73 observations from (Spellman *et al.*, 1998), nine observations from (DeRisi *et al.*, 1997) and ten observations from (Chu *et al.*, 1998).

Variation and Noise. One important quality of this data is that it is noisy. Even identifying which gene expression values genuinely vary at all in these experiments is nontrivial. Many investigators identify the subset of genes that appear to vary significantly in the experimental conditions by setting some threshold, such as 2- or 3-fold change, to define significance. However, the range of values (or log ratio values) seen has a statistical distribution that depends on the number of experiments. As the number of experiments increases, the percentage of genes identified as showing significant variation in expression using any fixed threshold criterion will increase just due to chance. Alternatively, setting a very conservative threshold is likely to exclude genes that are genuinely varying.

If we had an accurate estimate of the variance due to measurement error, we could use it to determine which

genes varied in a statistically significant manner. Ideally, such a variance estimate would come from a measured standard error based on replicate samples. Since such information is not always available, it is worth considering various alternative approaches. One such approach would be to estimate that variance based on median variation among genes that are expected to be invariant across a set of measurements, i.e., housekeeping genes.

For the yeast data considered here, we have neither replicate samples nor a set of known invariant housekeeping genes. However, we did observe that the vast majority of genes assayed did not vary in expression very much among the samples. Under an assumption that a reasonably low percentage of the genes are truly differentially expressed across experiments, the median of all of the gene variances should provide a robust estimate of the experimental noise variance.

It is reasonable to assume an approximately normal distribution for the observed log ratios in genes that are *not* differentially expressed, that is, that are subject to experimental noise variation only. We can therefore base an objective method of identifying genes exhibiting statistically significant variation across experiments on the χ-square distribution. Specifically, if there are N experiments, σ^2 is the variance for a particular gene, and median(σ^2) is the median over all of the gene variances, then we treat the quantity $W = (N-1)\sigma^2/\text{median}(\sigma^2)$ as approximately χ-square with $N-1$ degrees of freedom. This is because the distribution nS^2/σ^2 is χ-squared with $n-1$ degrees of freedom when S^2 is the sample variance of n normal random variables identically distributed with mean m and variance σ^2 [(Hogg & Craig, 1978), p.175]. We set the probability that a gene is selected as significantly variable due to chance at 1%, and then any gene for which the quantity W exceeds the upper 1% χ-square percentage point (with $N-1$ degrees of freedom) is designated as showing significant variation. Using this measure, 1889 genes show variation that is significant at the 1% level over 92 experiments. We use only these genes in the multimodal calculations in the sections below.

Correlational Structure. Another important aspect of large scale gene expression data is that the values observed for different genes are correlated with each other in complex ways. Transcriptional control within a cell produces gene products to meet the current needs of the cell. For example, many metabolic needs are met by complex pathways of enzymatic reactions. The genes for enzymes that catalyze a set of reactions along a pathway are likely to be coregulated, since they all are involved in accomplishing the same biological function. Shared transcription regulatory mechanisms also suggest that the observed expression levels of different genes will be correlated with each other. The complexity and variability of specific and general expression controls suggest that the correlational structure of expression levels will be quite rich, that is, of high order and time varying.

The expression levels of a particular pair of genes may be correlated with each other at all times (pairwise correlation), or only when the expression of other genes are in particular states (higher order correlations). Calculating all possible high order correlations is computationally intractable for assays involving tens of thousands of genes or more, and would require large numbers of experiments in order to be adequately sensitive. However, we can make estimates about the extent of that correlation without making the complete calculation. Figure 1 shows a histogram of statistically significant pairwise correlations between the set of significantly varying genes. Every gene in the data set is statistically significantly correlated with at least one other gene. For comparison, Figure 1 also shows the amount of correlation at that level which would be expected by chance in such a large number of pairs, calculated by permuting the yeast data to form a data set with the same overall distributional characteristics but only random correlations. The comparison shows that there is substantially more pairwise correlation among genes than would be expected by chance.

Another way to estimate the degree of correlation among the genes is to use principle components analysis (PCA). If the genes were completely independent of each other, then no dimensionality reduction would be possible beyond that resulting from the fact that the data lie on a 92 dimensional subspace. If all of the genes were completely correlated, then a single component could be used to account for all of the observed variance. Doing PCA on the 2622 genes in the data set with no missing values over 92 experiments, it takes only 41 components to account for 90% of the variance, 56 components to account for 95% of the variance, and 77 components to account for 99% of the variance. This evidence strongly confirms the biological intuition that expression levels of various genes are correlated.

The rich correlational structure among the expression levels of many genes makes good distance calculations difficult. In order to use Mahalanobis distance appropriately in this situation, one must calculate the entire covariance matrix, not just the diagonal. This requirement is computationally demanding, and making reliable estimates of all these covariances requires a large amount of data. Euclidean and correlational distance measures, which both make assumptions about the independence of the dimensions, are distorted by the correlational structure of the genes. When calculating similarity between experiments, groups of genes that are highly correlated with each other will be weighed more strongly than genes that are not correlated with others. A simplified example of this phenomenon is shown in Figure 2. Previous efforts,

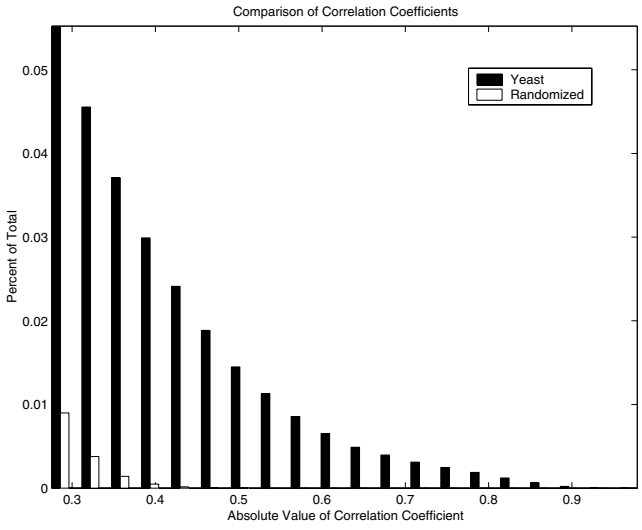

Fig. 1. Histograms of correlation coefficients of all pairwise correlations among genes that are statistically significant at the $p < 0.01$ level. The true yeast data (black bars) show substantially more significant pairwise correlation than a randomized control (white bars) generated from the same data.

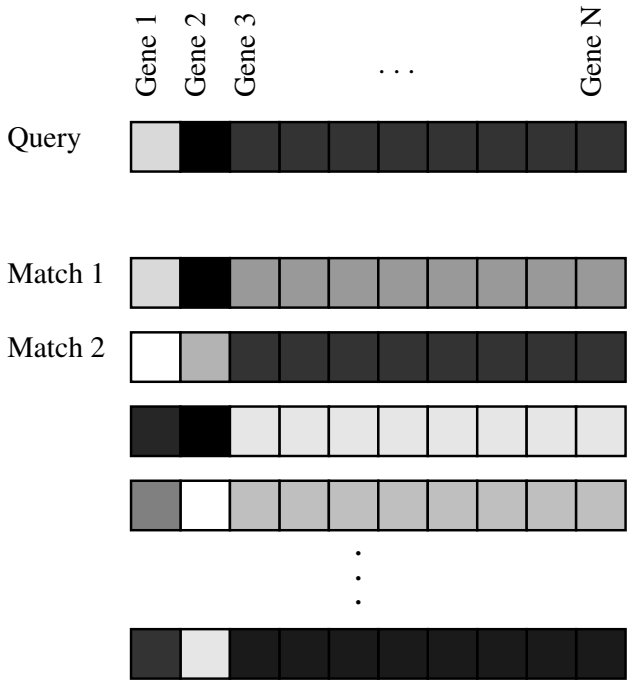

Fig. 2. An illustration of the influence of correlational structure on similarity. Each row represents an experiment in a database, and the greyscale indicates the expression level for all N gene columns. Note that genes 3 through N are all perfectly correlated. A similarity metric that gave all genes equal weight would rate the experiment labeled "Match2" as most similar to the query, since the value of the correlated genes are identical with the query. However, a metric that reduced the weight of the correlated genes to that of a single gene would rate the experiment labeled "Match 1" as more similar because of the identity of the values for genes 1 and 2 with the query.'

e.g. (Khan *et al.*, 1998) and (Raychaudhuri *et al.*, 2000) have used PCA to address related issues. One could transform the expression data using PCA, and then use pairwise correlation between the projected components. If the projection used captures most of the variance in the data, this is an approximation of Mahalanobis distance. In principle, however, there are several problems with this approach. First, computing the principle components or Mahalanobis distance requires that there be no data missing from any of the experiments; this is a rarely observed situation. Data values can be imputed, but there is no ideal imputation scheme for this application. Second, this approach only takes into account the pairwise correlations, ignoring higher order correlations. Such high order correlations are likely to be significant, since many biological phenomena depend on interactions of more than two biomolecules. Finally, this approach still assumes that the data have Gaussian structure; the validity of this assumption is addressed in the next section.

Distributional Form. Another important aspect of the data is its distribution. For example, Mahalanobis distance only makes sense when the data is normally distributed. Even Euclidean distance measures can be counterintuitive in spaces where the data is distributed in skewed or multimodal ways.

Consider the hypothetical distribution of the expression values of a single gene in Figure 3.

This gene is highly multimodal, with many observations

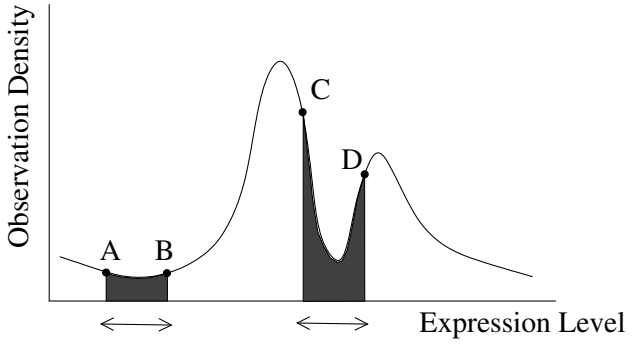

Fig. 3. A hypothetical multimodal density function for a single gene. Note that the Euclidean distance (indicated by the double arrows) between points A and B is the same as the distance between points C and D, but the pairs differ in the probability density (shaded area) between them.

near two particular values, and a smaller number of observations far from those modes. Consider the distance between points A and B, versus the distance between points C and D. Since expression values C and D occur in different modes, it is reasonable to assume that the biological activities associated with those expression levels are different. On the other hand, since the expression values of A and B are both in a rarely observed region, those values might be inferred to represent a biologically similar state, even if the Euclidean distance between A and B were larger than that between C and D. Our measure approximates the difference in the probability density between the two observations, making the distance between A and B smaller than the difference between C and D. (This does not directly assess whether the observations are in the same distributional mode.) It is important to note that the reason that A and B seem more similar than C and D cannot be captured in any summary measurement of the distribution (such as standard deviation). If the density function is significantly multimodal or skewed, capturing this intuition in a similarity metric requires use of the density function itself, or at least an estimate of it.

The empirical gene expression levels observed thus far suggest that many of these distributions are far from normal. Of the 1889 genes whose expression levels show a significant amount of variation in our data set, 181 (9%) are multimodal by the DIP test (Hartigan & Hartigan, 1985), and 727 (38%) fail the Kolmogorov-Smirnov test of normality at the 0.01 significance level; 679 (36%) of them are skewed. Several illustrative empirical density estimates are shown in Figure 4.

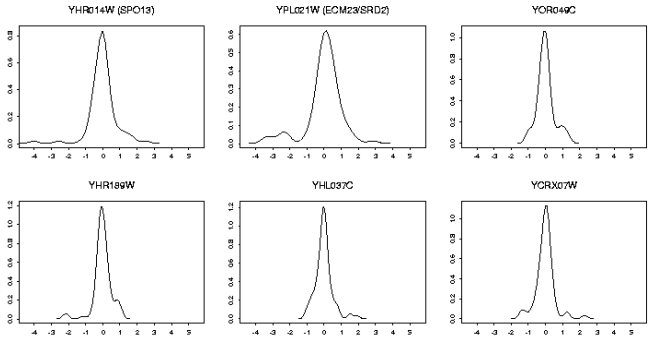

Fig. 4. Density plots of six representative multimodal gene expression log ratio distributions, from the yeast data set. The horizonta,l axis is the log ratio of the expression, and the vertical axis is the observed density; the changes in the expression level are statistically significant for each. YHR014W is a meiosis specific sporulation protein, YPL021W is homologous to Srd1p which is involved in processing pre-mRNA, YOR049C is similar to Rta1p which confers resistance to 7-amino cholesterol, YHR189W is a putative peptidyl-tRNA hydrolase, and YHL037C and YCRX07W are hypothetical proteins of unknown function.

METHODS

We desire a similarity metric over experiments that takes into account both the correlational structure among the genes and the complex distributional qualities of the individual gene density functions. The joint probability density function over all the genes contains the information we need to calculate such a similarity metric.

A Bayesian similarity metric

We view the problem as trying to answer the following question about two experiments: how can we decide if the experiments are two measurements of the same cellular state, versus measurements of two different cellular states?

This approach is somewhat akin to Bayesian methods in macromolecular sequence comparison, where one attempts to integrate over all possible sequences that could have been common ancestors of the two observed sequences. The main difference between comparing expression vectors and comparing sequences is that the model over possible expression states is more complex, so dynamic programming solutions do not apply.

Assume that we know the joint probability distribution function, $f(g_1, g_2, \ldots, g_n)$, which we will also write as $f(G)$. This is the distribution of the true expression levels (or log ratios) of the genes $g_1 \ldots g_n$ that the tissue can generate. Assume further that we also know the function for experimental error, that is a function $e(O, G)$ which defines the probability density of observing the values $o_1 \ldots o_n$ in an experiment given that the true expression levels were $g_1 \ldots g_n$.

Consider two experimental observations, X and Y. If X and Y are observations of the same cellular state, that is, they represent the same true expression level Z, then the likelihood of the observed data is

$$\int_Z (e(X, Z)e(Y, Z)f(Z))dZ. \tag{3}$$

Now consider two observations X and Y which are from independent cellular states; that is, where the expression levels are independent samples from the joint distribution. In that case, the likelihood of the observed data is

$$\int_Z (e(X, Z)f(Z))dZ \cdot \int_Z (e(Y, Z)f(Z))dZ. \tag{4}$$

The ratio between Eq. 3 and Eq 4

$$\frac{\int_Z (e(X, Z)e(Y, Z)f(Z))dZ}{\int_Z (e(X, Z)f(Z))dZ \cdot \int_Z (e(Y, Z)f(Z))dZ} \tag{5}$$

is the Bayes factor for distinguishing between the hypothesis that the two experimental observations are two instances of the same cellular state versus the hypothesis that they represent independent draws from the universe of possible gene expression states. When this score exceeds unity, then the odds are that the observations are instances of the same cellular state. Even when the odds of two observations being instances of the same state are quite small, the values take into account the correlational and distributional nature of the database, and can be used to generate rank orders. We therefore propose this ratio as a good similarity metric for expression array experiments.

The error function e represents the experimental error. We can reasonably assume that it is multivariate normal, and that the measurement errors for each gene are independent. This is largely an assumption of convenience. However, it is reasonable to expect symmetry in the errors, so the Gaussian model seems *prima face* reasonable. In contrast, we know from the above discussion that the distributional function f is clearly neither normal nor independent across the genes. Furthermore, since the number of genes in Z is very large (thousands at least), computing this integral directly is intractable.

An empirical estimate of the similarity metric

There are various plausible ways to estimate $f(Z)$. One could assume the genes were independent, and use empirical density estimation methods for each gene. Califano *et al.* (2000) propose a related approach for measuring the distance between individual genes. However, as the above arguments suggest, that independence assumption is unlikely to hold, and hence this is not an attractive method. A more reasonable approach is to use principle components analysis to transform the data so that the components are independent, and to work in that space. This would be appropriate if the correlational structure were all pairwise, and the distributions Gaussian - which is not the case here. A further possibility would be to approximate $f(Z)$ by a mixture of multivariate normals, e.g. using k-means clustering. Not only are there difficult modeling issues involved, such as selecting the number of components in the mixture, but such models are notoriously difficult to deconvolute even in one dimension, let alone in the current case of thousands of dimensions.

Our approach is to use the database of already observed experiments as an estimate of the true distribution function. Our estimator of $f(Z)$ is simply the probability mass function based on the empirical distribution function. It is a consistent estimator and has a long history of use in statistics. It is the estimator which is the basis of bootstrap methods. Although this assumption is somewhat questionable now, as the number of experiments in the database grows, using it as an estimate of the true distribution becomes more plausible.

Instead of integrating, we will sum over all of the entries in the database. Let N be the number of experiments in the database, let K be the number of genes assayed in the experiments, and let σ be an estimate of the average measurement error. Let D_{ij} be the value of the jth gene in the ith experiment in the database, and let X_j and Y_j be the jth gene in the two experiments that are being compared. We can define the similarity score for our Bayesian-based Gene Expression Search Tool (GEST) as

$$\frac{\frac{1}{N-2}\sum_{i=1}^{N-2}\prod_{j=1}^{K}\phi\left(\frac{X_j - D_{ij}}{\sigma}\right)\cdot\phi\left(\frac{Y_j - D_{ij}}{\sigma}\right)}{\frac{1}{N-2}\sum_{i=1}^{N-2}\prod_{j=1}^{K}\phi\left(\frac{X_j - D_{ij}}{\sigma}\right)\cdot\frac{1}{N-2}\sum_{i=1}^{N-2}\prod_{j=1}^{K}\phi\left(\frac{Y_j - D_{ij}}{\sigma}\right)} \tag{6}$$

where $\phi(n) = \frac{1}{\sqrt{2\pi}}\cdot e^{-\frac{n^2}{2}}$ and the summation over experiments disregards X and Y, thus the sum goes to $N-2$ rather than N. If they were included, they would dominate the metric, which would then be approximately proportional to the antilog of the Euclidean distance.

Computational issues

To make this scoring practical for database searches, several computational issues must be dealt with. First, all of the expression values in the database must be commensurate with the query. For example, if the expression values are expressed as ratios, the ratios must all be calculated relative to the same internal control. Also, the calculation assumes that all experiments have the same set of genes and there are no missing values. Standard techniques (e.g. using only the shared subset of genes, imputing missing values, etc.) can be used to bring real data into conformance with this latter assumption.

The computational complexity of this metric for comparing two experiments is clearly proportional to the product of the number of genes with the number of experiments in the database. Fortunately, the computational complexity of comparing a query experiment to every other experiment in the database is the same, since we can cache the comparison of each member of the database with each other and with the query. Calculating the individual scores is an additional linear factor in the number of experiments. Furthermore, if the database is relatively stable compared to the queries, we can precalculate the components of the equation that depend only on the database and save them.

RESULTS

To gain some practical sense of the biological significance of the use of different distance measures for searching databases of expression array results, we did an all-against-all comparison of the 92 yeast expression

experiments described above using three metrics: our Bayesian metric (GEST), Euclidean distance (ED) and correlational similarity (CS).

For the GEST metric, we used 0.25 as an estimate of the experimental error (σ). Of the 6024 genes appearing anywhere in the yeast data set, only 2622 appear in all 92 experiments with no missing values. However, an additional 2408 genes are present in at least 95% of the experiments. Since these values are likely to be missing at random, we can integrate over all possible values of the missing data. In this case, the integral simply results in 1, and removes the missing gene from equation 6. For the ED and CS metrics, we did the comparisons using the genes in common in each particular case. For the GEST metric, we did the same, subject to the additional constraint that each such gene must be present in 95% of all experiments.

The total running time for the GEST comparisons was 20901 seconds (a bit under 6 hours) on one processor of an SGI Origin 2000 server. The precalculation takes nearly all of that time, and once completed, a query can be searched against the full database in 235 seconds on the same machine. The calculation is straightforward to parallelize, and scales linearly with the number of experiments in the database.

Differences among similarity metrics

We compared every experiment in the yeast database to every other experiment using the Bayesian metric (GEST), Euclidean distance (ED) and correlational similarity (CS).

As a simple metric comparison, we looked at the best match (highest scoring non-identical experiment) for each of the 92 experiments in the database for all three metrics. For 16 experiments (17%), all three metrics agreed about the best match. In 8 experiments (9%) the GEST metric gave the same best match as the ED metric but that differed from the best match by the CS metric. The GEST best match was the same as the CS but not the ED in 4 experiments (4%). In 28 of the experiments (30%), the ED and the CS metrics agreed with each other, but disagreed with the GEST best match. In 32 experiments (35%), the metrics all gave different best matches.

We then tried a true statistical measure, comparing our results to the best objective standard we could find for the data set. As a "gold standard" for the true distance between pairs of cell cycle experiments, we made an estimate of the position in the cell cycle relative to the M/G1 to G1 boundary using the data from Spellman *et al.* (1998). Phases of the cell-cycle genes are delineated according to the color-coded overbar of Figure 1 in Spellman *et al.* (1998). Each time point can be assigned a number between 0 and 1 which represents the proportion of the cell-cycle completed, starting from the M/G1 to G1 boundary. For time series that span multiple cell cycles, the period of the cycle is adjusted to the length of the cycle in minutes as given in Spellman *et al.* (1998). For a data point at time t in a cell cycle of period T, with an M/G1 to G1 boundary at time t_0, the proportion assigned to the data point is given by $[(t - t_0) \bmod T]/T$. The underlying assumption here is that the closer two experiments are in the cell cycle, the more similar will be (the less distance between) their gene expression patterns.

Then, for each of the 73 of the 92 experiments assigned a cell cycle position, we compared the rank orderings of the matches produced by the sorted scores of the GEST and ED metrics and of the match rank ordering produced by the "true" distance given by our cell cycle distance "gold standard". This comparison was done via calculation of the appropriate Spearman and Kendall rank correlation coefficients. The basic reasoning: the closer a metric produced a match ordering to that produced by the "gold standard" of the cell cycle distance calculation, the better the metric. We tried this using all matches, and for the best N matches, for varying values of N (the idea being that the lower-scoring matches might be overwhelming the orderings with experimental noise). Unfortunately, the results were ambiguous across the set of 73 experiments. No statistically significant advantage could be detected for the Bayesian metric over Euclidean distance.

DISCUSSION

In the all-against-all comparison with the yeast database, we did not find robust statistically significant differences overall between our GEST metric and the ED and CS metrics. The problem here might lie with the data set. The scores in the rank orderings being compared were frequently very close. Comparing expression profiles for different mammalian tissue samples would provide a potentially more relevant basis for comparison. As the number of expression profiles in various databases grows larger and the estimate of the joint distribution of expression values becomes better, and as experimental measurements become more precise, the difference between the measures may grow to be significant.

Approximations for speedup

A key weakness of the approach is its computation time. Large databases are necessary for good estimates of the distributions of the expression values, but large databases mean large precalculation times. We have devised an shortcut that may be useful, but more testing of the current algorithm (itself an approximation) is needed before we move to a further approximation of the theoretical basis.

Identifying similar cellular states with respect to subsets of genes

If this method for searching a gene expression database is comparable to finding global sequence alignments, it

is also natural to consider analogs to local sequence alignments. Since proximity is not an issue for gene expression, the question becomes one of identifying significant subsets of genes that can be used to determine that a pair of experiments appear to be in similar states. We can adopt our metric to this task by just considering the subsets of genes that contribute positively to the score. However, there are open questions in attempting to determine the significance of scores calculated in that fashion. In the extreme, consider a pair of experiments that have only a single gene with a high score; although that one gene is a subset that makes the experiments appear to be similar, intuitively the subset is too small to be meaningful.

A useful extension to the method is to allow users to define a subset of genes of interest (e.g. those involved in cell cycle, or in apoptosis), and search for experiments that represent a similar cellular state with respect to those genes. It is also possible to allow users to select subsets of the experiments in a database to use for the density estimate. For example, one might be interested in finding experiments that are similar with respect to the joint density of expression of genes in a particular tissue, rather than over all possible cellular states.

Conclusions

We have demonstrated that gene expression data has a complex distributional form, involving multimodality, non-normality and complex correlational structure. These factors mitigate against doing database search using similarity metrics such as correlation coefficient or Euclidean distance. We introduced a similarity metric based on the Bayes ratio of the odds that a pair of experiments are two samples of the same cellular state versus being independent samples over the distribution of cellular states. That metric is impractical to calculate, so we provided a tractable approximation that uses a database of expression experiments as an estimate of the true joint distribution. We implemented this approximation in a Perl-based Gene Expression Search Tool (GEST). Using GEST, we compared this metric against Euclidean distance and correlational metrics on real data from yeast, and found differences among all of the metrics, but no statistically significant advantage for our new metric. More testing is required on better data sets (with a more accurate objective "gold standard") to find out how much of the theoretical advantage of the Bayesian approach described here can be implemented in practice.

REFERENCES

Abbott, A. (1999). Bioinformatics institute plans public database for gene expression data. *Nature*, **398**, 646.

Bassett, D., Eisen, M. & Boguski, M. (1999). Gene expression informatics – it's all in your mine. *Nature Genetics*, **21**, 51–55.

Califano, A., Stolovitzky, G. & Tu, Y. (2000). Analysis of gene expression microarrays for phenotype classification. In *Proceedings Eight International Conference on Intelligent Systems for Molecular Biology*. pp. 75–85.

Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P. O. & Herkowitz, I. (1998). The transcriptional program of sporulation in budding yeast. *Science*, **282**, 699–705.

DeRisi, J. L., Iyer, V. R. & Brown, P. O. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **278**, 680–686.

Eisen, M., Spellman, P., Brown, P. & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences USA*, **95**, 14863–14868.

Hartigan, J. & Hartigan, P. (1985). The DIP test of unimodality. *The Annals of Statistics*, **13**, 70–84.

Hogg, R. & Craig, A. (1978). Introduction to Mathematical Statistics. Macmillan Publishing Company.

Khan, J., Simon, R., Bittner, M., Chen, Y., Leighton, S., Pohida, T., Smith, P., Jiang, Y., Gooden, G., Trent, J. & Meltzer, P. (1998). Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays. *Cancer Research*, **58**, 5009–5013.

Raychaudhuri, S., Stuart, J. & Altman, R. (2000). Principal components analysis to summarize microarray experiments. In *Pacific Symposium on Biocomputing*. pp. 455–466.

Spellman, P., Sherlock, G., Zhang, M., Iyer, V., Anders, K., Eisen, M., Brown, P., Botstein, D. & Futcher, B. (1998). Comprehensive identification of the cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Molecular Biology of the Cell*, **9**, 3273–3297.

Wilson, D. & Martinez, T. (1997). Improved heterogenous distance functions. *Journal of Artificial Intelligence Research*, **6**, 1–34.