Chapter 9

A Probabilistic Method for Determining Gene Relationships from Expression Data

9.1 Introduction

One of the ultimate goals in biology is to understand the function of all genes and the structure of the interaction networks among them. Aside from its scientific value, a complete and detailed understanding would have a profound impact on the field of medicine. For example, it would become possible to design accurate diagnostics for nearly any condition, and it would be possible to accurately predict the effectiveness and side-effects of drugs or genetic treatments [103].

Microarrays (discussed in Chapters 7 and 8) and SAGE (serial analysis of gene expression) have proven to be powerful tools in the pursuit of this goal, providing genome-wide high-throughput measurements of cellular mRNA levels ("gene expression levels") as a readout of the state of the underlying genetic network. Experiments are often designed to compare the state of the genetic network under two or more conditions. For example, one can observe how the pattern of expression of genes involved in development changes over time, or one can observe the difference in network state between healthy tissue and cancerous tissue. Alternatively, one can monitor the induced changes in expression due to a perturbation such as a structural network change (e.g., knockout or overexpressed gene) or a temporary change induced by a drug, toxin, pathogen, hormone, or other factor.

By observing patterns of induced changes, it is possible to make some hypotheses about the underlying genetic network structure. For example, genes that show similar changes in expression under a variety of conditions are likely to be closely related, perhaps performing a similar function or belonging to the same regulatory pathway.[1] In a growing body of knowledge in the literature and in numerous online databases, such hypotheses are being pieced together into a unified picture that will ultimately describe the whole underlying system. Already, the individual hypotheses have suggested useful biological experiments and have helped to identify new candidate drug targets and diagnostic markers for further exploration.

With the vast amounts of data being generated by microarray and SAGE experiments, the field of bioinformatics has proliferated. Many statistical techniques have emerged to deduce relationships among genes from this wealth of data, in order to assign functions to previously unknown genes and to piece together the network of gene regulation. In this chapter, I first briefly review several such techniques and argue the advantages of non-metric techniques, such as Guilt by Association (GBA), in particular. Though this method was developed by Walker *et al.* [282] to infer the relatedness of genes based on cDNA library data, we have extended it so that expression ratio data can also be analyzed. I present a detailed description of our modifications as well as our implementation of the modified algorithm in computer software. The software uses several tricks to permit the calculations to be performed in a reasonable amount of time. Our computed estimates of gene relationships (p-values) are available in an online database for further investigation.

9.2 Analyzing expression data

In a typical microarray study,[2] messenger RNA (mRNA) is extracted from a sample, transcribed into complementary DNA (cDNA), and labeled with a radioactive or fluorescent marker. A microarray contains thousands of spatially-identified tethered DNA “probes”. When the sample is hybridized to

[1] It should be noted that a lack of correlation between expression patterns does not necessarily indicate that genes are unrelated. It may simply mean that the set of experimental conditions was not sufficiently broad to cause changes in the relevant pathway.

[2] Though I use the term “microarray” for concreteness, the discussion of experimental principles and data analysis is valid for a wide variety of high-throughput technologies that measure gene expression.

the array, these probes capture complementary cDNA molecules from the sample, and the intensity (radioactivity or fluorescence) measured at each array position can thus be read out to determine the concentration of the corresponding cDNA species in the sample. This concentration corresponds to the original mRNA level or “gene expression level”.

To cancel out many uncertain sources of noise and variation, experiments are often carried out in a differential fashion and one determines an *expression ratio* for each gene, where the expression level in one sample is divided by the level in a reference sample. A differential experiment may be performed by labeling each of the two samples with a different fluorophore and hybridizing them to the same microarray, or by measuring absolute levels on two different microarrays. The resulting expression ratios are associated with the experimental differences between the two samples. A ratio greater than 1 indicates that the gene was up-regulated (expressed at a higher level) compared to the reference, and a ratio less than 1 indicates that the gene was down-regulated (expressed at a lower level) with respect to the reference.

While an individual differential experiment can provide meaningful information, studies typically compare many samples (prepared under different experimental conditions) to the same reference. The term “condition” is used in the broadest possible sense. Some examples of experiments that have been reported include: (i) comparing samples taken at different times or stages of development to a baseline sample; (ii) comparing samples from different tissue types or different types of cancers to a pool of cDNA from all samples; or (iii) comparing samples exposed to certain nutrients, drugs, toxins, or pathogens to an untreated sample. The result of performing such a series of experiments is a set of expression ratios for each gene, called an expression “vector”, “profile”, or “pattern”. We are interested here in the analysis of such vectors.

Numerous methods have emerged for sifting through the vast quantities of published gene expression data. They are based on the simple idea that genes showing similar patterns of expression across many experiments are likely to be related in function or to play a part in the same biological pathway.

One important distinction among analysis methods is how the similarity of expression patterns is comapred. “Metric” methods use a distance measure—such as Euclidean distance or Pearson correlation distance—that satisfies the axioms of a metric. Distance metrics satisfy certain mathematical properties, including the triangle inequality, which states that:

$$d_{AC} \leq d_{AB} + d_{BC}, \tag{9.1}$$

where d_{ij} is the distance between the expression vectors of two genes i and j. A common criticism of methods using distance metrics is that certain biological relationships cannot be accurately described due to this constraint. For example, if genes X and Y have unrelated biological functions yet are both regulated to some degree by a common transcriptional regulator Z, then one would expect the proper description to be $d_{XZ}, d_{YZ} \ll d_{XY}$, which is not compatible with the triangle inequality. Other shortcomings of distance metrics have been cited as well. For example, methods based on Euclidean distance cannot handle expression vectors with missing data points (due to the inability to properly orient an incomplete vector in expression space), nor can they properly handle important relationships such as negative correlations between genes [33]. Furthermore, distance metrics tend to be highly sensitive to measurement errors in expression data (which is inherently very noisy), and many metrics introduce biases such as assigning more significance to larger ratio values in the expression vector [32]. “Non-metric” measures of gene relatedness, such as probabilities [282] or mutual information [32], do not suffer these drawbacks.

A second important distinction among analysis methods is whether the final description of the relationships consists of gene clusters or gene networks (e.g., Bayesian [83, 131] or Relevance [32] networks). Clustering is the process of finding groups (clusters) of genes with the most closely related expression vectors. A wide variety of methods have been used [129], including k-means, Gaussian mixture models, fuzzy c-means, self-organizing maps [262], and hierarchical clustering into dendrograms [70]. Since the emergence of high-throughput platforms for measuring gene expression, clustering has been the predominant method of analysis and has led to a great many important biological discoveries. However, clustering has many shortcomings. First, many methods do not permit genes to be members of multiple clusters, therefore preventing an accurate description of genes involved in multiple pathways or under the control of multiple regulatory factors. Some methods have difficulty describing other biologically relevant situations, such as negatively correlated genes, or genes exhibiting non-linear relationships. Another drawback is that some clustering algorithms require seemingly arbitrary quantities such as the final number of clusters or other parameters to be known in advance. Lastly, clustering methods use global correlation measures and attempt to place *all* genes into clusters, even though it is unlikely that the weakest relations are believable.

Network approaches tend to extract prominent relationships from the observed data rather than trying to fit all of it. This can be especially useful in situations where gene relationships are only apparent under a small subset of experimental conditions and would be masked in global comparisons. In the Relevance network approach of Butte and Kohane [32], a non-metric quantity called mutual information is computed for each pair of genes to indicate the probability that they are related, and then a threshold probability level is imposed, effectively converting all the probabilities into binary values: related or unrelated. Butte and Kohane permuted their data to determine the maximum mutual information that could be obtained by random chance and used this as a cutoff. The result is several disjoint networks of genes, each containing links representing only the most believable gene relationships; improbable links are simply discarded. The threshold value affects the size of these networks and the number of connections between them. In practice, one must tune the threshold probability to achieve the desired trade-off between false positive and false negative error rates. Bayesian approaches attempt to determine the most probable genetic regulatory network structure given the available data. Again, rather than just grouping genes that behave similarly, a Bayesian network precisely identifies specific links between individual genes. Bayesian networks have the advantage of being able to naturally incorporate different measurement models into the analysis (e.g., noisy or stochastic expression data rather than fixed ratios) [32] and even to combine different types of data into the analysis [83, 131]. For example, clinical and protein interaction data could be combined with expression ratio data to increase the accuracy of predictions.

To avoid the many pitfalls associated with clustering using distance metrics, the methods described in the next sections use non-metric probability calculations. These probabilities can be used to identify candidate drug targets or diagnostic markers by finding genes closely related to known targets, or can be used to construct relevance networks.

9.3 Guilt by Association

Our non-metric analysis method described in the next section is based closely on the Guilt by Association (GBA) algorithm introduced by Walker *et al.* [282], in which pairs of related genes are identified based on their "associations" in tissue libraries. A tissue library (also known as a cDNA library) is essentially a collection of the mRNA content of a sample that has been reverse transcribed into cDNA. Transcripts are identified and counted by a method such as sequencing or SAGE. GBA measures the association between genes with a non-metric probability function to avoid the disadvantages of using distance metrics.

Library data is attractive for analysis because it is often more quantitatively accurate than microarray data, especially for transcripts expressed at very low levels [213, 124]. In addition, transcript counting is far less noisy [282, 280] and can more accurately detect even very slight differences in expression levels between samples [213]; in microarrays, differences smaller than a factor of 2 are often considered insignificant and are thus ignored. Furthermore, library data is more "portable" than expression ratio data because it consists of absolute measurements of transcript abundance. SAGE measurements from different experiments can be directly compared, whereas the differences between microarray formats, reference samples, and normalization strategies make direct comparison of microarray experiments difficult [213].

Like microarray data, many cDNA libraries have been published online [13]. Data from several different libraries can be combined to construct expression vectors for each gene. If a pair of genes has similar expression vectors across a set of libraries then the genes are likely to be related.

The first step in GBA analysis is to discretize the library expression data. Walker *et al.* used binary values: 1 if the gene was present in a given library (regardless of the number of copies); and

0 if the gene was absent. Table 9.1 illustrates library data that has been discretized in this manner. Discretization is intended to simplify the analysis, as well as to reduce the impact of any quantitative differences between different libraries (e.g., if they are normalized or subtracted), and to remove the magnitudes of expression to allow the detection of relationships between genes that are not linear and monotonic.

cDNA Library:	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	...
Gene A	1	0	0	1	1	0	1	1	0	1	...
Gene B	1	1	0	1	1	0	1	0	0	1	...
Gene C	0	0	1	1	0	0	0	1	1	0	...
⋮											

Table 9.1: **Example of discretized expression of genes in cDNA libraries**. Each row of the table represents the expression of a particular transcript (gene or EST) in many different libraries (represented by columns). A value of "0" indicates that the transcript was not detected in the library, while a value of "1" indicates that at least one copy was found.

For each pair of genes, Walker *et al.* compute the probability (p-value) that the observed pattern of co-expression could have arisen by random chance. To do so, the discretized expression data is organized into a 2×2 contingency table (Table 9.2). This table summarizes the pattern of *co-expression* of the two genes. Entries represent the number of libraries in which gene A and gene B are both present (n_{11}), the number of libraries in which the genes are both absent (n_{00}), and the number of libraries in which one gene is present while the other is absent (n_{10} and n_{01}). For each row and column, "margin totals" are computed. For example, the row totals r_1 and r_0 represent the total number of libraries in which gene A was present or absent, respectively. The total of these margin totals is the total number of libraries, N. It is from this table that the p-value is computed. One makes the "null hypothesis" that the genes are independent and computes the probability that the observed pattern of co-expression could occur randomly, *assuming fixed margin totals*, then tests the validity of this hypothesis. A low p-value implies that the null hypothesis should be rejected and that the genes are likely to be related.

	Gene B = 1	**Gene B = 0**	**Total**
Gene A = 1	$n_{11} = 5$	$n_{10} = 1$	$r_1 = 6$
Gene A = 0	$n_{01} = 1$	$n_{00} = 3$	$r_0 = 4$
Total	$c_1 = 6$	$c_0 = 4$	$N = 10$

Table 9.2: **Example of co-expression pattern of genes in cDNA libraries**. This 2×2 contingency table summarizes for genes A and B the co-expression data from Table 9.1. Each of the values in the table represents a certain number of cDNA libraries. n_{11} is the number of libraries in which gene A is present and gene B is also present; n_{10} is the number of libraries in which gene A is present but gene B is absent, etc. The last row and column are "margin totals". The row total $r_1 = n_{11} + n_{10}$ is the total number of libraries in which gene A was found, and $r_0 = n_{01} + n_{00}$ is the total number of libraries from which gene A was absent. Column totals represent analogous quantities for gene B. The total of the margin totals is simply the total number of libraries, $N = r_1 + r_0 = c_1 + c_0$. The data in the contingency table are used to compute the likelihood that the pair of genes is related, by methods described in the text.

One method for evaluating the null hypothesis is to perform a chi-squared test. First, an *expected* count is computed for each table cell (row i, column j),

$$E_{ij} = \frac{r_i \times c_j}{N}. \tag{9.2}$$

The chi-squared statistic is then computed:

$$\chi^2 = \sum_{i,j} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}, \tag{9.3}$$

where O_{ij} is the *observed* count in each cell (from the original contingency table), and the sum is computed over all four cells. Observing that there is just one degree of freedom (e.g., n_{11}) when margin totals are fixed, one then computes the probability of the χ^2 statistic. A low probability indicates a large deviation between expected and observed counts and hence the relation between the genes is non-random.

The validity of the chi-squared test depends on having a sufficient sample size. However, for many pairs of genes—particularly those present in very few cDNA libraries—the minimum requirements for validity are not met. One commonly used validity requirement is that the total sample size must be greater than 40 or that all tables cells must have expected values of at least 5 if the total sample size is in the range 20–40. A thorough analysis of validity conditions was reported by Tejedor and Andrés [266].

When the conditions are not met, the p-value calculation must instead be performed by an alternate method such as Fisher's exact test. This involves computing the probability of observing the actual data or more "extreme" data with the same margin totals. One enumerates all possible tables with the same margin totals, computes the probability of each table, then sums all probabilities that are less than or equal to that of the observed table. The probability of one particular table (i.e., one particular co-expression pattern) is:

$$P_{2\times2} = \frac{\binom{N}{n_{11}}\binom{N-n_{11}}{n_{10}}\binom{N-n_{11}-n_{10}}{n_{01}}\binom{N-n_{11}-n_{10}-n_{01}}{n_{00}}}{\binom{N}{r_1}\binom{N-r_1}{r_0}\binom{N}{c_1}\binom{N-c_1}{c_0}} = \frac{r_1!r_0!c_1!c_0!}{N!n_{11}!n_{10}!n_{01}!n_{00}!}. \tag{9.4}$$

This equation can be interpreted as the number of possible arrangements of data with the observed numbers of correlations (n_{11}, n_{10}, etc.) preserved, divided by the total number of possible arrangements of data with only the restriction that the observed margin totals are preserved.

For 2×2 contingency tables, it is straightforward to enumerate all other possible tables to determine which are more extreme because there is only one degree of freedom. One needs only to vary one of the cells over all possible values and compute the other cells using the margin totals, omitting any tables containing negative-valued cells. An example is shown in Figure 9.1 for the contingency table of Table 9.2.

The lower the p-value (whether computed by a chi-squared test or Fisher's exact test), the less likely is the null hypothesis, and the more likely it is that the pair of genes is related. Walker *et al.* used Guilt by Association with Fisher's exact test to determine which genes were most closely associated with known genes involved in prostate cancer [282], Parkinson's disease and schizophrenia [281], and the cell cycle [280]. In the first two studies, a set of 522 human cDNA libraries was used; in the third, 1176 libraries. It was observed that the computed p-values for known gene relations were lower (therefore more significant) than correlation coefficients [281]. Furthermore, several known relations were not detected by correlation methods but were detected by GBA [282, 280].

It should be noted that the magnitudes of the p-values are only approximate because several assumptions made by the calculations do not hold in general. For example, the cDNA libraries

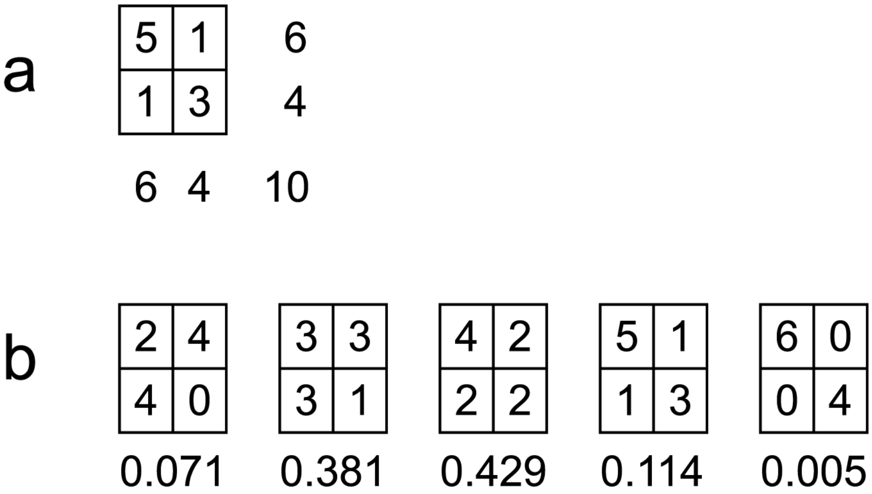


Figure 9.1: **Example of performing Fisher's exact test**. (a) The contingency table from Table 9.2 with the margin totals shown. (b) All possible tables with the same margin totals. The value underneath each table is the probability, $P_{2\times2}$, of that particular table from Equation 9.4. The overall p-value for the observed data is then the sum of all probabilities less than or equal to 0.114: $P = 0.114 + 0.071 + 0.005 = 0.190$.

are not all independent—many belong to sets of experiments with only small differences between each sample. However, even with appropriate corrections, the most reliable relations remain significant [282].

Though only cDNA library data were analyzed by Walker *et al.*, presumably this method can also be applied to data from single channel oligonucleotide experiments such as those using Affymetrix GeneChips or arrays using radioactive labels.

9.4 Extension of GBA to expression ratio data

Building on earlier work by Brody and Quake (`http://thebigone.stanford.edu/yeast/`), we extended the Guilt by Association method to use differential expression *ratio* data rather than library data. The motivation for this was two-fold: (i) there is far more expression ratio data that is publicly available; and (ii) ratios can represent richer, more-complex relationships between genes than presence or absence in cDNA libraries. Furthermore, by combining data from multiple microarray studies it may be possible to uncover gene relationships that are not clearly apparent in any individual study. Our algorithm involves discretization of expression data prior to analysis and a non-metric measure of the relation between gene pairs, for the same reasons put forth by Walker *et al.* [282, 281, 280].

Brody and Quake analyzed yeast expression data from one set of microarray experiments by a simplified approach; here we further develop the algorithm and analyze human expression data combined from many different microarray studies.

9.4.1 Algorithm

To perform a modified Guilt by Association analysis, we first discretized expression ratios to three distinct values, $+$, $-$, or 0, representing up-regulated, down-regulated, or unchanged expression, respectively. Discretization helps to address the problem of high variability in ratio data and also helps to identify complex non-linear correlations by ignoring the magnitude of expression changes. A hypothetical set of discretized data is shown in Table 9.3. We performed the discretization simply by imposing a fixed "noise threshold". Ratios that exceeded the upper threshold, $T_+ = 1.414$, were designated up-regulated; ratios that were smaller than the lower threshold, $T_- = 0.707$, were designated down-regulated; and all others were designated as unchanged. Our simple method has several shortcomings—for example, it ignores genes that undergo only small expression changes in response to perturbations. However, several more sophisticated approaches have been reported in the literature for determining whether expression changes are statistically significant. For example, one can apply statistical analysis of variance (ANOVA) [146, 66, 144], or one can design experiments to include redundant measurements and additional controls [145, 107]. At the time our work was performed, most published datasets were not properly analyzed and did not supply all the needed information (such as raw scanner data) for us to be able to correct the analysis. Furthermore, many datasets did not include sufficient experimental replicates and controls to eliminate many sources of noise and variation. The field has progressed tremendously in the past 5 years, and several recent studies now include estimates of statistical significance with each expression ratio.

Analogous to the 2×2 case, the pattern of co-expression of the discretized data for each pair of genes (Table 9.3) can be summarized in a 3×3 contingency table (see Table 9.4). This table lists the

Experiment:	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	...
Gene A	+	0	0	−	+	0	+	+	+	0	−	−	0	+	+	...
Gene B	+	+	0	−	+	−	+	0	X	X	+	0	+	−	−	...
Gene C	X	+	−	+	0	0	0	+	−	0	0	−	0	0	+	...
⋮																

Table 9.3: **Example of discretized gene expression ratios in expression datasets**. Each row corresponds to a particular transcript (gene or EST). For each experiment (column), the ratio is expressed as + (up-regulated), − (down-regulated), or 0 (unchanged). Note that sometimes the data for a particular gene is missing from an experiment due to a defect in the array or other problem (indicated by an "X"). When comparing two genes, only experiments in which both genes have a valid data point are included in the calculation.

number of experiments in which the genes are both up-regulated (n_{++}), both down-regulated (n_{--}), etc. Margin totals are computed for the table and then a p-value is computed based on a chi-squared test or Fisher's exact test. With fixed margin totals, there are four degrees of freedom in a 3×3 table. The chi-squared test is valid if no more than 20% of the expected frequences are less than 5 and none is less than 1. In many cases, this condition is not met, even in datasets consisting of dozens of individual experiments. For example, highly correlated genes have a high value for n_{++} but low values for all other table cells.

	Gene B = +	**Gene B = 0**	**Gene B = −**	**Total**
Gene A = +	$n_{++} = 3$	$n_{+0} = 1$	$n_{+-} = 2$	$r_+ = 6$
Gene A = 0	$n_{0+} = 2$	$n_{00} = 1$	$n_{0-} = 1$	$r_0 = 4$
Gene A = −	$n_{-+} = 1$	$n_{-0} = 1$	$n_{--} = 1$	$r_- = 3$
Total	$c_+ = 6$	$c_0 = 3$	$c_- = 4$	$N = 13$

Table 9.4: **Example of co-expression pattern of genes from discretized ratio data**. The pattern of co-expression of genes A and B in Table 9.3 are summarized in this contingency table. n_{++} is the number of microarray experiments in which both genes were up-regulated, n_{+0} is the number of experiments in which gene A was up-regulated while gene B was unchanged, etc. Margin totals are computed for each row (r_+, r_0, and r_-) and column (c_+, c_0, and c_-). From the table, the likelihood of the null hypothesis (that the genes are independent) is computed using a chi-squared or Fisher's exact test.

Using generalizations of Fisher's exact test to 3×3 tables, a p-value can be computed as follows. The probability of a single table is given by:

$$P_{3\times3} = \frac{r_+!r_0!r_-!c_+!c_0!c_-!}{N!n_{++}!n_{+0}!n_{+-}!n_{0+}!n_{00}!n_{0-}!n_{-+}!n_{-0}!n_{--}!}, \tag{9.5}$$

and the overall p-value is obtained by summing probabilities over all possible tables that have the same margin totals as the observed data but are more "extreme". In 3×3 tables with fixed margin totals, there are four degrees of freedom so it is much more difficult to find all the tables than in the 2×2 case. When dealing with datasets consisting of on the order of 100 experiments, we observed that calculations for some pairs of genes involved the evaluation of tens of thousands of possible tables. Brute force methods can easily identify all tables, but more efficient algorithms have also been developed [190, 191, 46].

9.4.2 Implementation

We wrote computer software in perl and C++ to create and maintain a database of gene expression ratio data (in original and discretized form) along with p-values between each gene pair computed by the GBA method. The database is available online at `http://thebigone.stanford.edu/pvalue/`. Raw microarray data was obtained from the Stanford Microarray Database (`http://genome-www.stanford.edu/microarray/`). We downloaded data from several available experiments relating to human tissue and cell lines, as indicated in Table 9.5.

Reference	**Description**	**Number of Genes**	**Number of Experiments**
[127]	Fibroblast response to serum	8600	19
[24]	Peripheral blood mononuclear cell response to bacterial infection	7600	182
[211]	Breast tumours	8100	84
[232]	Clustering of genes based on tumour type in cancer cell lines	8000	68
[4]	Distinguishing types of B-cell lymphoma by expression differences	17900 (cDNA clones)	133
[210]	Expression patterns in mammary epithelial cells and breast cancers	5000	33
[294]	Identification of cell-cycle associated genes in cancer cell lines	16300, 29600	90
[61]	Response of macrophages to a bacterial transcription factor required for virulence	22600 (cDNA transcripts)	53
[59]	Temporal expression profile of prostate cancer cell line after treatment with synthetic androgen	18000	30

Table 9.5: **Human microarray datasets used in GBA analysis**.

Most microarray publications list probes used on the arrays by their GenBank Accession Number [19]. We consulted the UniGene database (build 150) [214] to determine the gene (or EST) represented by each probe. Storing the raw data based on sequence identifiers rather than genes allowed us to easily keep up to date[3] with UniGene as genes were added and corrections were made with each new "build". Using the UniGene database also allowed us to aggregate data from all probes representing the same gene. For each unique gene (UniGene "cluster"), a single expression ratio for each microarray experiment was determined by taking the median of the ratios of all constituent probes. Expression ratios for all genes/clusters were then discretized and stored. The latest version of our database contains expression data for approximately 35000 unique clusters.

Though combining measurements in this manner makes theoretical sense, we noticed in several cases that probes corresponding to the same UniGene cluster had very different expression patterns and probably should not be combined. These cases may indicate errors in the UniGene database or may represent misidentification of probes in array experiments. It is expected that UniGene errors will eventually be resolved with future updates so no effort was made at this time to detect or correct these questionable cluster assignments. However, in the meantime, we did build a second database of discretized expression ratios based on individual sequences rather than clusters. This database contained approximately 80000 entries. For clarity in the subsequent discussion, only the first database is described.

For each pair of genes in the database, a p-value was computed based on the discretized ratio data and stored. Storing p-values is necessary if one wishes to search the database for the most probable gene relationships, for example. Due to the large number of gene pairs, we chose not to store *all* p-values but rather only those less than a certain threshold (10^{-2}). Not only did this reduce the data storage requirements, but it also dramatically improved the speed at which results could be returned when querying the database. The threshold was selected somewhat arbitrarily, but a later analysis (see Figure 9.2) revealed it to be an acceptable choice because only p-values much lower than this are thought to represent significant relationships in our database.

[3] Each update requires that sequences be reassigned to clusters, expression ratios be re-aggregated and re-discretized, and p-values be re-computed for all gene pairs. To improve efficiency, one could detect which sequences and clusters had been affected by the update and re-compute only those.

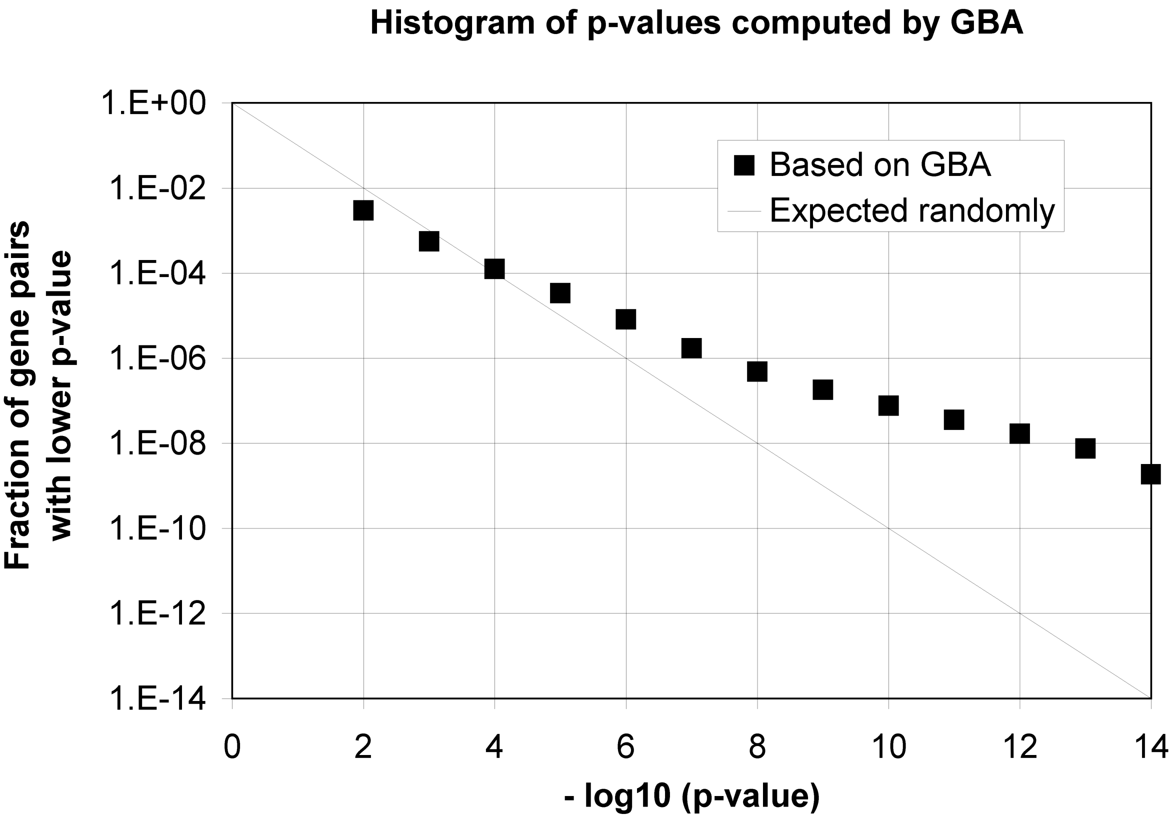


Figure 9.2: **Analysis to determine p-value representing the threshold of significance**. Due to violations of certain assumptions of the GBA algorithm, the magnitudes of the computed p-values are not reliable. To determine at what p-value gene pairs can be considered to have a significant relation, we performed a simple graphical analysis. A histogram was generated, indicating for each p-value the fraction of gene pairs having that p-value or lower. The continuous line shows the fraction that would be expected by random chance (i.e., $1/p$), and the square markers indicate the fractions tabulated from our database of p-values (combined from 9 datasets using the p_{min} algorithm). One can observe that the GBA data is distinctly non-random. At a p-value of 10^{-4}, the lines cross. For lower p-values, there are more gene pairs in the GBA data than expected by random chance, suggesting that 10^{-4} represents the threshold of significance. This is only an approximation and must be tuned to achieve the desired trade-off between false positives and false negatives.

P-value calculations were implemented in C++ code, using logarithms of the relevant equations to improve computational accuracy. Three different p-values were calculated for every pair of genes in our dataset: (i) the individual table probability $P_{3\times3}$, (ii) the chi-squared probability, and (iii) the probability computed by Fisher's exact test.

Despite the apparent complexity of Equation 9.5, $P_{3\times3}$ can be computed very inexpensively by pre-computing a table of $\log(n!)$ for $n = 0..N$ once at the start of the run. (The time needed for this pre-computation is amortized over all genes pairs.) Compared with Brody and Quake's original code, this simple modification reduced the execution time by 20%. The implementation of the chi-squared calculation [216] is also relatively inexpensive.

On the other hand, Fisher's exact test is a very expensive calculation, due to the large number of tables that may exist for a given set of margin totals. Though brute force methods can be used to find all tables and compute their probabilities, more efficient algorithms have been published [190, 191, 46]. We implemented this computation by calling an external FORTRAN 77 subroutine published by Mehta and Patel [191]. I modified the code slightly to avoid duplicating some calculations when calling the subroutine billions of times. Calculations of the Fisher's exact test p-value for some gene pairs took many seconds (on an 800 MHz AMD Athlon computer), so it was not practical to complete a full run for all gene pairs (0.6 billion pairs in one database, 3.2 billion pairs in the other). However, it was not necessary to perform the full calculation most of the time—in many cases the other calculations provide an excellent approximation.

For very small p-values, the single-table probability agrees very well with Fisher's exact test. This is not surprising—when the p-value is low, there are very few, if any, additional contingency tables that are more extreme than the observed data; therefore, there are few terms in the Fisher summation. For increasing p-values, the values rapidly diverge. It may be possible to derive a threshold p-value below which the single-table value can safely be used as an approximation to the Fisher value within some specified tolerance.

Though not accurate, the single-table p-value also has some utility for high p-values. Because it is always an *under*estimate of the Fisher's exact test p-value, it can be used as a quick screen to avoid unnecessary and expensive calculations. If the single-table p-value is greater than the database cutoff value, then we know that the gene pair will not be stored in the database because the Fisher's exact test p-value will be even higher. Thus the full Fisher's exact test computation can be skipped.

We also found that the agreement between Fisher's exact test and the chi-squared test was quite good in most cases, for both high and low p-values. The only exceptions were cases where the requirements for validity of the chi-squared test were not satisfied. For example, highly correlated genes frequently had low values in many cells of the contingency table. This suggests that the chi-squared method can be used to compute most p-values to a good approximation, except in cases where Fisher's exact test must be used due to violation of the validity conditions.

It should be noted that in Brody and Quake's analysis of yeast expression data, all p-values are based on the single-table value. Thus, it is expected that the results are only accurate for the lowest (most significant) p-values. Fortunately, these are the ones that are generally the most interesting.

In addition to the software mentioned above, additional programs were written in the PHP scripting language to provide a web-based interface to the database. The database can be browsed for gene pairs having the smallest p-values (i.e., most probable relationships) or for all genes having a probable relationship with a particular gene (identified by UniGene cluster, sequence accession number, gene name, or gene description). For each pair of genes, the p-value is given along with links to view the raw or discretized data on which the calculation was based. One other quantity that is shown is the "dot product", computed by multiplying integer representations of the discretized ratios ($+1$, 0, -1) for the two genes in each experiment and summing over all experiments. Gene pairs that are highly correlated will have a large positive dot product, pairs that are highly anti-correlated will have a large negative dot product, and those related in more complex ways will have an intermediate value.

9.4.3 Combining datasets

Since many published microarray studies explore only a small range of experimental conditions, a large part of the cell's genetic network is not interrogated, and the relationships between many pairs of genes remain hidden. By combining expression measurements from multiple studies, however, one can compare expression vectors across a much broader range of conditions, and relationships are more likely to be revealed, if they exist. As more and more studies are published, the effectiveness of combining them will improve.

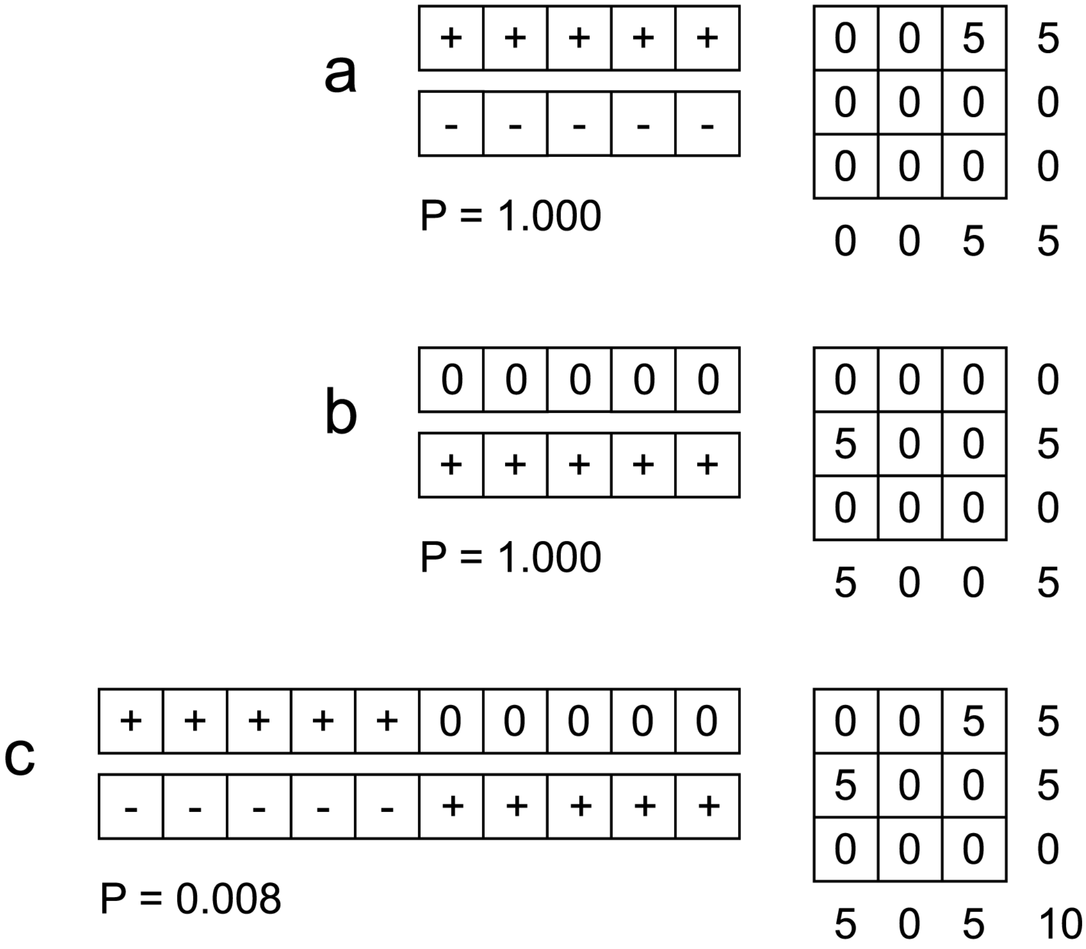
We pursued two approaches for combining sets of experiments ("datasets") from multiple studies. In the first, we simply combined all datasets into very long expression vectors for each gene, such that each vector contained ratios from all experimental conditions in all studies. However, we observed that low p-values were being computed for many pairs of genes thought to be unrelated.

This problem arises from the details of how microarray experiments are performed. In the ideal case, an experiment would compare samples consisting of a single cell type to a reference consisting of an identical cell type. The observed differences in the samples would reflect real expression changes resulting directly from the experimental conditions. However, many studies use mixed cell types, either inadvertently because micro-dissection was not used to isolate individual cells during sample preparation, or because samples were intentionally pooled. Pooling is often performed to ensure that the reference sample contains molecules representing all cDNA sequences to avoid the problem of dividing by zero in ratio calculations.

Comparing different cell types in a microarray experiment results in systematic biases in the expression ratios for the whole set of experiments. For example, when one cell type is compared to a pool of cell types, expression ratios reflect biases such as the fundamental differences in expression levels between cells of different types in addition to any real expression changes due to the experimental conditions. Such biases can lead to the appearance of false correlations (see Figure 9.3). It could also be argued that biases such as expression differences due to differences in cell type represent meaningful information that should be included in the analysis; however, in practice, it is not possible to differentiate meaningful biases from the many possible meaningless ones.

The majority of public datasets available at the time of this analysis had obvious biases such as different cell types, though a few had no obvious biases [127, 294].[4] Instead of restricting our analysis

[4] It is likely that even these apparently bias-free studies have some sources of hidden bias such as differences between dyes or detectors in the two fluorescence channels [18].



a

+	+	+	+	+
-	-	-	-	-

P = 1.000

0	0	5	5
0	0	0	0
0	0	0	0
0	0	5	5

b

0	0	0	0	0
+	+	+	+	+

P = 1.000

0	0	0	0
5	0	0	5
0	0	0	0
5	0	0	5

c

+	+	+	+	+	0	0	0	0	0
-	-	-	-	-	+	+	+	+	+

P = 0.008

0	0	5	5
5	0	0	5
0	0	0	0
5	0	5	10

Figure 9.3: **Example of bias problem when combining expression datasets**. Biases present in microarray experiments, due to the use of pools of cell types in the reference sample, for example, lead to problems when performing p-value calculations on combined datasets. (a) A hypothetical dataset where gene X (upper row) appears up-regulated in all five hypothetical experiments due to a bias, and gene Y (lower row) appears down-regulated in all experiments. The contingency table for this particular dataset is shown at the right. Because all measurements are identical, the fact that gene X and gene Y show a correlated pattern of expression is not considered significant. The p-value is very high: $P = 1.000$. Thus, a systematic bias does not create a problem (i.e., false positive) if the dataset is analyzed individually. (b) Another hypothetical dataset where gene X appears unchanged in all experiments, and gene Y appears up-regulated in all experiments due to a bias. Again, the contingency table and (insignificant) p-value are shown. (c) If these two datasets are combined into long expression vectors containing all experiments from both sets, the two groupings of biased measurements now falsely appear highly significant ($P = 0.008$). The low p-value arises because it is unlikely that independent genes would be observed to be co-expressed with $n_{+-} = 5$ and $n_{0+} = 5$ in 10 experiments. Thus, combining datasets with different biases into long expression vectors leads to very misleading p-values.

to this small fraction of studies, however, we combined expression measurements by an alternate approach.

For each gene pair, we computed p-values separately for each dataset, then combined them into a single p-value. There are several ways of accomplishing this. The simplest method is simply to take the minimum of all the p-values:

$$p_{min} = \min(p_1, p_2, ..., p_k), \tag{9.6}$$

where k is the number of datasets and p_i is the p-value computed from the ith dataset. The rationale behind this approach is that gene relationships will be revealed in some sets of experiments but not in others. If at least one set shows a significant relation, then it is sufficient to assume the genes are related. Other methods for combining p-values have been reported as well [306, 292]. For example, one can compute a "Fisher statistic",

$$S = -2\sum_{i=1}^{k} \ln p_i, \tag{9.7}$$

and then compute an overall p-value by interpretting this statistic as a χ^2 value with $2k$ degrees of freedom. One problem with this method is that it requires *all* p-values to be stored in the database—a cutoff cannot be used. It is not clear what is the best method to combine values: the first method is vulnerable to outliers with low p-value, while the second may allow a significant result in one dataset to be "washed out" among many datasets that show no significant result. It is also unclear whether all individual p-values should be weighted equally or whether p-values derived from datasets consisting of more experiments should be given more weight.

We observed that among the 9 datasets used in our analysis, only a few p-values were available to be combined for each gene pair. Some of the missing values are due to our use of a p-value cutoff when building the database. However, the p_{min} algorithm is unaffected because the missing values can safely be assumed to be greater than any of the included values. Other missing values reflect the fact that not all microarrays contain the same set of genes, and thus the comparison of expression patterns for a given gene pair may not be possible on some arrays. By including many more datasets in the analysis, one can increase the average number of individual p-values that are combined for each gene pair.

9.4.4 Results

In our database, many of the lowest p-values corresponded to expected relationships such as genes that code for different modules of the same protein complex (such as major histocompatibility complex (MHC) proteins and immunoglobulins) or genes that are in close proximity on the chromosome. A large fraction of highly significant relations involved at least one unknown gene. These pairs most likely represent identical genes. Indeed, a more recent version of the UniGene database (build 186) shows that many of the pairs we identified initially have now been merged into the same UniGene cluster. This suggests the method could be used to assign putative functions to unknown ESTs used as microarray probes.

Among the lowest p-values are also pairs of genes representing different enzymes in the same metabolic pathway, as well as many pairs of genes involved in the cell cycle, including many of the same relations found by Walker [280]. Our database did not contain very many of the relationships pertaining to prostate cancer, Parkison's disease, and schizophrena as found by Walker *et al.* [282, 281]. Presumably this is simply due to the fact that the microarray experiments included in our analysis did not include all of the relevant genes and that the experimental conditions in these studies were not designed to perturb the relevant pathways.

In addition to verifying several of the most significant relations, we found that many sets of significant gene pairs picked randomly from the database correspond to suspected biological relations. Since it is very tedious to perform literature and database searches for each gene in a pair to determine whether the relation makes biological sense, we instead compared our results to suspected groups of significant genes published in an extensive study by Segal *et al.* [239]. As part of that study, more than 450 biologically significant "modules" of related genes were identified. We found that most pairs of genes with low p-value picked randomly from our database consist of genes belonging to the same module.

In spite of these comparisons, there is an overall lack of sources of "correct answers" against which to compare generated hypotheses [33]. Eventually, improved annotations and more complete databases will allow algorithms such as ours to be fully evaluated in terms of the accuracy and completeness of the set of predicted relationships.

9.5 Related Work

Earlier sections in this chapter have described the relation of our work to the GBA method of Walker *et al.* [282, 281, 280] who analyzed profiles of expression in cDNA libraries.

Our work also has many characteristics in common with the work of Butte and Kohane [32], in which vectors of yeast expression ratios were discretized into n subranges and compared based on their mutual information. Mutual information is a non-metric measure of the shared information between two vectors. The higher the mutual information, the less likely the vectors are randomly related to one another, and the more likely there exists a biological relationship between the genes. The authors' analysis revealed many relationships that could be validated in the literature, including pairs of identical genes, genes in the same pathway, and genes with similar functions. Butte and Kohane selected $n = 10$ in their analysis, which is significantly higher than our value of $n = 3$. While higher values of n utilize more information from the expression ratios, they increase the susceptibility to noise. Our $n = 3$ approach has the advantage of being compatible with statistical approaches that determine whether a gene is significantly up- or down-regulated. This is particularly helpful in the case of genes that do not exhibit wide swings in expression levels and for which small changes in expression can often be very significant. Such small expression changes are typically ignored by methods that look only at magnitudes of expression ratios.

Bowers *et al.* [26] recently reported an interesting analysis of protein "phylogenetic profiles" (as opposed to gene expression profiles) that bears some similarity to our work. For each protein, an N-dimensional profile is constructed, with ones and zeros representing whether the protein (or a close homolog) is present or not present in each of N organisms. Rather than analyzing pairs of profiles, the authors investigated protein triplets. They identified pairs of profiles, a and b, that individually were not good predictors of a third profile, c, but whose logically combined profiles described c well. Comparisons were based on a non-metric measure related to the entropy of the individual and joint profiles. Each triplet could be classified as one of eight possible "logic relationships" and could be combined together to infer the structure of the protein interaction network. The analysis of triplets can detect relations that might go unnoticed if only examining pairwise relationships. The authors suggest that the underlying principles of their analysis could be applied to other sets of genomic data including expression profiles. Perhaps the work of Butte and Kohane [32] or our modified Guilt by Association algorithm could serve as a starting point.

Other areas of related work include Bayesian networks [83, 131] and Boolean networks [165]. These approaches can model additional dimensions of relationships between genes, including temporal (causation) and spatial effects, ultimately leading to a more accurate and complete picture the genetic network in humans and other organisms. However, the data for performing such analyses remains scarce. In the meantime, methods such as those described above for predicting gene relationships based on expression data will continue to be immensely useful in deducing the functions of unknown genes and discovering new candidate drug targets and diagnostic markers.

9.6 Future Directions

At the time we created this database, it was difficult to draw meaningful conclusions, beyond the simple verification of some known biological relationships, due to the relatively small number and narrow range of published human microarray studies available, and due to the many errors and omissions in the UniGene database. Furthermore, without the raw image data it was not possible to determine for low expression ratios whether the degree of up- or down-regulation should be considered significant. It is likely that many low ratios were misclassified by our simple threshold approach.

In recent years, UniGene has been updated many times and hundreds of new human microarray studies (consisting of many thousands of individual experiments) have been published, reflecting a much more complete set of interrogation conditions. In addition, a far greater set of known relations now exists in databases and in the literature for assessing the accuracy of the results. Updating the database to include hundreds of new microarray studies and to employ a more sophisticated discretization algorithm would consume considerable computing resources but could ultimately produce a valuable data mining tool. Researchers at Peking University have implemented a public database, *GBA server* [296], that accumulates EST library data and p-value calculations based on the GBA methods of Walker *et al.* [282]. A similar implementation for the modified GBA method that we have presented here could potentially serve as a valuable resource for the online bioinformatics community. An attractive feature of GBA databases is that their effectiveness and reliability increases with time as more and more data is integrated.

An additional worthwhile direction of future work concerns the user interface. In addition to presenting the output as a list of genes with significant relations, it would be useful to explore the use of web-based graphical tools such as TouchGraph [240] to display interactive relevance networks of the relations having p-values below some threshold. A simple analysis (Figure 9.2) suggests a p-value cutoff of 10^{-4} or lower for our current database. We found 67193 pairs of genes with a p-value below this threshold (out of about 0.5 billion possible pairs). It is not unreasonable to graphically navigate a set of data with this size.

Acknowledgment

I am grateful to Jim Brody for introducing me to the field of bioinformatics, and for helping me to get up to speed with web-based scripting languages and tools for analyzing microarray data. I also thank Rebecca Shafee for helpful discussions about combining expression datasets and for her significant contributions in terms of parsing and importing datasets as well as exploring and verifying the predicted gene relationships. It was a pleasure to work with people with such enthusiasm for science.