Math Minute 6.1 Why Should You Log-Transform Microarray Data?

You have seen how red-to-green intensity ratios are computed from the intensity levels determined in the scanning process, and how they represent repression or induction of the genes. For example, a four-fold repression in gene expression results in a ratio of approximately 0.25. Similarly, a 16-fold repression results in a ratio of 0.0625, a 4-fold induction 4.0, and a 16-fold induction 16.0. But there is a much greater numerical difference between 4.0 and 16.0 than there is between 0.25 and 0.0625. Graph the expression patterns for genes C and N from Table 6.1 on a single set of axes, and notice how the induction stands out much more clearly than the repression, even though the repression is of the same magnitude. In addition to biasing our visual interpretation of gene expression patterns, the compression of ratios between 0 and 1 causes problems with mathematical techniques for analyzing and comparing gene expression patterns (Math Minute 6.2). Furthermore, to interpret a ratio less than 1 as a fold repression, you must take its reciprocal (e.g., $1/0.0625 = 16.0$), an operation that most people find difficult to do quickly and accurately in their heads.

To avoid these problems, investigators often use a **log transformation** of the ratio data. Log transformation is illustrated in the following example. Suppose the ratio is 0.0625. Take the base 2 logarithm of this number:

$$\log_2(0.0625) = \log_2(1/16) = \log_2(1) - \log_2(16) = -\log_2(16) = -4.$$

Because the $\log_2$ of 1/16 is the negative of the $\log_2$ of 16, a 16-fold induction and a 16-fold repression have the same magnitude (one positive and one negative) in the $\log_2$-transformed data. The magnitude of the number is the power of 2 needed to get the induction/repression number (e.g., $16 = 2^4$). The $\log_2$ transformations of the ratios in Table 6.1 are given in Table MM6.1.

Sometimes $\log_{10}$ is used instead of $\log_2$; in this case, the magnitude of the transformed data is the power of 10 needed to get the induction/repression number. Transforming the same ratios as above leads to the following:

$$\begin{array}{ll} \log_{10}(4) \approx 0.6 & \log_{10}(.25) \approx -0.6, \\ \log_{10}(16) \approx 1.2 & \log_{10}(0.0625) \approx -1.2. \end{array}$$

In $\log_2$ transformed data, a value of 2 corresponds to a ratio of 4; however, you would be surprised to see a value as large as 2 in $\log_{10}$ transformed data, since 2 corresponds to a ratio of 100. In general, a $\log_2$ transformation helps you easily identify doublings or halvings in ratios, while a $\log_{10}$ transformation helps you see order-of-magnitude changes. The key attribute of log-transformed expression data is that equally sized induction and repression receive equal treatment visually and mathematically.

Table MM6.1 Log_2 transformation of gene expression data in Table 6.1.

Name	0 hours	2 hours	4 hours	6 hours	8 hours	10 hours
gene C	0	3	3.58	4	3.58	3
gene D	0	1.58	2	2	1.58	1
gene E	0	2	3	3	3	3
gene F	0	0	0	−2	−2	−3.32
gene G	0	1	1.58	2	1.58	1
gene H	0	−1	−1.60	−2	−1.60	−1
gene I	0	2	3	2	0	−1
gene J	0	1	0	1	0	1
gene K	0	0	0	0	1.58	1.58
gene L	0	1	1.58	2	1.58	1
gene M	0	−1.60	−2	−2	−1.60	−1
gene N	0	−3	−3.59	−4	−3.59	−3

Math Minute 6.2 How Do You Measure Similarity between Expression Patterns?

In Table 6.1, the transcriptional responses of genes G and L are clearly similar, since they have the same ratios at each time point. But what about gene D? How similar is its response to that of genes G and L? A common way to measure the similarity between gene expression patterns like those in Table 6.1 is with the Pearson **correlation coefficient**, abbreviated with the letter *r*.

Correlation quantifies the extent to which the expression patterns of two genes go up together and down together over several time points or experimental conditions, even if the numbers are not of the same magnitude. A correlation coefficient of 1.0 between two genes means that their expression patterns track each other perfectly. A correlation coefficient of –1.0 between two genes means that their expression patterns track perfectly, but in opposition to one another (i.e., one goes up while the other goes down). A correlation coefficient near zero means that the expression patterns of the two genes do not track each other at all. You can experiment with the correlation guide web page and the Excel file correl_explore.xls to gain an intuitive understanding of correlation.

In the following example, we will work with the expression values given in Table 6.1 to further illustrate the disadvantages of analyzing raw (untransformed) expression ratios, as discussed in Math Minute 6.1. To find the correlation between genes D and L, denoted r_{DL}, first compute the sample mean and sample standard deviation of the expression values for each gene (i.e., each row):

$$\bar{X}_{\mathrm{D}} \approx 2.83 \quad \bar{X}_{\mathrm{L}} = 2.5 \qquad \sigma_{\mathrm{D}} \approx 1.067 \quad \sigma_{\mathrm{L}} \approx 0.957.$$

Subtract $\bar{X}_D$ from each value in the D row and divide each result by σ_{D}. The result is a row of **normalized** values in the D row:

$$\mathrm{D}_{\mathrm{norm}} = -1.715, 0.1593, 1.097, 1.097, 0.1593, -0.7779.$$

Do the same in the L row, this time subtracting $\bar{X}_{\mathrm{L}}$ and dividing by σ_{L}, to produce the following normalized row:

$$\mathrm{L}_{\mathrm{norm}} = -1.567, -0.5225, 0.5225, 1.567, 0.5225, -0.5225.$$

Now multiply the first number in $\mathrm{D}_{\mathrm{norm}}$ by the first number in $\mathrm{L}_{\mathrm{norm}}$, the second number in $\mathrm{D}_{\mathrm{norm}}$ by the second number in $\mathrm{L}_{\mathrm{norm}}$, and so on, keeping a running sum of these products. (You might recognize this operation as the dot product of the two vectors $\mathrm{D}_{\mathrm{norm}}$ and $\mathrm{L}_{\mathrm{norm}}$.) Finally, divide this sum (5.386) by the number of elements in each row (6) to get the correlation coefficient $r_{\mathrm{DL}} \approx 0.897$.

Table MM6.2 Correlation coefficient between each pair of genes, based on $\log_2$-transformed gene expression data in Table MM6.1.

	gene C	gene D	gene E	gene F	gene G	gene H	gene I	gene J	gene K	gene L	gene M	gene N
gene C	1	0.94	0.96	−0.40	0.95	−0.95	0.41	0.36	0.23	0.95	−0.94	−1
gene D	0.94	1	0.84	−0.10	0.94	−0.94	0.68	0.24	−0.07	0.94	−1	−0.94
gene E	0.96	0.84	1	−0.57	0.89	−0.89	0.21	0.30	0.43	0.89	−0.84	−0.96
gene F	−0.40	−0.10	−0.57	1	−0.35	0.35	0.60	−0.43	−0.79	−0.35	0.10	0.40
gene G	0.95	0.94	0.89	−0.35	1	−1	0.48	0.22	0.11	1	−0.94	−0.95
gene H	−0.95	−0.94	−0.89	0.35	−1	1	−0.48	−0.21	−0.11	−1	0.94	0.95
gene I	0.41	0.68	0.21	0.60	0.48	−0.48	1	0	−0.75	0.48	−0.68	−0.41
gene J	0.36	0.24	0.30	−0.43	0.22	−0.21	0	1	0	0.22	−0.24	−0.36
gene K	0.23	−0.07	0.43	−0.79	0.11	−0.11	−0.75	0	1	0.11	0.07	−0.23
gene L	0.95	0.94	0.89	−0.35	1	−1	0.48	0.22	0.11	1	−0.94	−0.95
gene M	−0.94	−1	−0.84	0.10	−0.94	0.94	−0.68	−0.24	0.07	−0.94	1	0.94
gene N	−1	−0.94	−0.96	0.40	−0.95	0.95	−0.41	−0.36	−0.23	−0.95	0.94	1

Notice that the expression ratios for genes H and M are the reciprocals of the ratios for genes L and D, respectively. In other words, gene H is repressed to exactly the same extent that gene L is induced, and gene M is repressed to the same extent as D is induced. We would thus expect r_{DL}, which compares the patterns of induction, to be the same as r_{HM}, which compares the patterns of repression. However, r_{HM} is 0.97, which is quite a bit larger than r_{DL}. This strange behavior occurs because correlation is sensitive to the relative magnitudes of the patterns, and can be prevented by first log-transforming the data (see Math Minute 6.1). The correlation coefficients of each pair of genes in Table MM6.1, computed with the log-transformed data, are shown in Table MM6.2. Note that r_{HM} is now the same as r_{DL}.

Because the correlation between genes D and L is close to 1, we conclude that gene D is highly similar to gene L (and thus also highly similar to G). This conclusion can lead to hypotheses about the function of gene D if genes G and L are well understood, but gene D is not. In Math Minute 6.3, we will see how correlations can be used to group highly similar genes so that these hypotheses can be made genome-wide.

MATH MINUTE DISCOVERY QUESTIONS

1. The diagonal entries of Table MM6.2 need not be calculated, because $r_{\mathrm{AA}} = 1$ for any gene A. Furthermore, the entries above (or below) the diagonal entries need not be calculated, because correlation is a symmetric function (i.e., $r_{\mathrm{AB}} = r_{\mathrm{BA}}$, for any two genes A and B). How many entries in Table MM6.2 must be calculated?
2. Find a simple formula that represents the number of entries in Table MM6.2 that must be calculated (i.e., your answer to question 1) as a function of the number of genes being compared.

Math Minute 6.3 How Do You Cluster Genes?

Figure 6.9 shows the results of clustering several thousand genes, just as Table 6.3 and Figure 6.8 illustrated the results of clustering the 12 genes in Table 6.2. The purpose of cluster analysis is to organize the genes into groups whose members' expression patterns are all similar to one another according to a particular similarity measure (e.g., Pearson correlation coefficient; see Math Minute 6.2). In Figure 6.8, you could arrange the 12 gene expression patterns by hand so that similar patterns were adjacent to one another as much as possible. However, computer algorithms are required to achieve the

MATH MINUTES
clustering
software

genome-wide clustering shown in Figure 6.9. There are many methods for clustering genes; the one used in our case studies is **hierarchical clustering**.

Hierarchical clustering works as follows. First, find the two most similar genes in the entire set of genes. Join these together into a cluster. Now join the next two most similar objects (an object can be a gene or a cluster), forming a new cluster. Add the new cluster to the list of available objects, and remove the two objects used to form the new cluster. Continue this process, joining objects in the order of their similarity to one another, until there is only one object on the list—a single cluster containing all genes.

To find the two most similar objects, we need a way to measure similarity when one or both objects being compared are clusters of genes. One way is to average the log-transformed expression patterns of the genes in a cluster, forming an average expression pattern that represents that cluster. The pattern is then treated as though it were a single gene, meaning that we must compute its correlation with the pattern of every other currently available object. Let's walk through the process to cluster the genes in Table 6.2, using the similarities in Table MM6.2. First, find the two most similar genes in the entire set of genes. Genes L and G are the most similar, because $r_{LG} = 1$. Join these together into a cluster, denoted [LG]. Cluster [LG] is added to the list of available objects, and the single genes L and G are removed from the list. Now join the next two most similar objects, using the procedure described earlier. (Note that in this case, the average of L and G is equal to both L and G, so we are saved the job of computing new correlations.) The most similar gene to the cluster [LG] is gene C, with $r_{CG} = r_{CL} = 0.95$. However, gene C and cluster [LG] are not the two most similar objects; rather, genes C and E are, with $r_{CE} = 0.96$. Thus, we join genes E and C to form cluster [EC].

At the next iteration, we need to know the correlation of each object with the average log-transformed expression patterns of genes E and C: 0, 2.5, 3.29, 3.5, 3.29, 3. The correlations of all available objects with this pattern representing [EC] are in Table MM6.3.

From Table MM6.3, we see that the object most similar to [EC] is cluster [LG], with a correlation of 0.93. Gene D is even more similar to [LG], since $r_{DG} = 0.94$. However, the two most similar objects now are genes N and H, with $r_{NH} = 0.95$. Therefore, we join genes N and H to form cluster [NH]. We have now completed 3 iterations of the hierarchical clustering algorithm. The entire clustering process for these 12 genes takes 11 iterations; the steps are summarized in Table MM6.4. Note that the final object created is the clustering of all 12 genes shown in Figure 6.8.

The hierarchical clustering process can also be summarized in a **dendrogram** similar to those discussed on pages 139–145. Figure MM6.1 shows the dendrogram for the hierarchical clustering detailed in Table MM6.4 and Figure 6.8. Notice that genes L and G are consolidated into a single node in the tree. The depth (from right to left) at which a node connects two objects represents the similarity between them. At any node that joins two branches, the top and bottom branches can be exchanged without changing the interpretation of the tree. Therefore, many different orderings of the leaves are consistent with the branching structure of a particular dendrogram. Dendrograms are used extensively in Chapter 7 to represent clusters.

Hierarchical clustering is the most popular method for finding trends in gene expression data, but there are several others. Another common method is the *k*-means cluster algorithm, which tries to find the best partition of the entire set of genes into precisely *k* groups. Several software programs for clustering gene expression data with hierarchical, *k*-means, and other methods are freely available for academic use. You can also experiment with clustering microarray data online. Each cluster algorithm may result in

Table MM6.3 Correlations between [EC] and all other objects.

D	F	H	I	J	K	M	N	[LG]
0.90	−0.48	−0.93	0.32	0.33	0.32	−0.90	−0.99	0.93

Table MM6.4 Summary of the hierarchical clustering algorithm applied to the 12 genes in Table 6.2.

	Two Most Similar Objects			
Iteration	Object 1	Object 2	Correlation	New Object
1	L	G	1.00	[LG]
2	E	C	0.96	[EC]
3	N	H	0.95	[NH]
4	M	[NH]	0.95	[MNH]
5	[LG]	D	0.94	[LGD]
6	[EC]	[LGD]	0.94	[ECLGD]
7	I	F	0.60	[IF]
8	J	[ECLGD]	0.29	[JECLGD]
9	K	[JECLGD]	0.19	[KJECLGD]
10	[KJECLGD]	[IF]	−0.12	[KJECLGDIF]
11	[MNH]	[KJECLGDIF]	−0.96	[MNHKJECLGDIF]

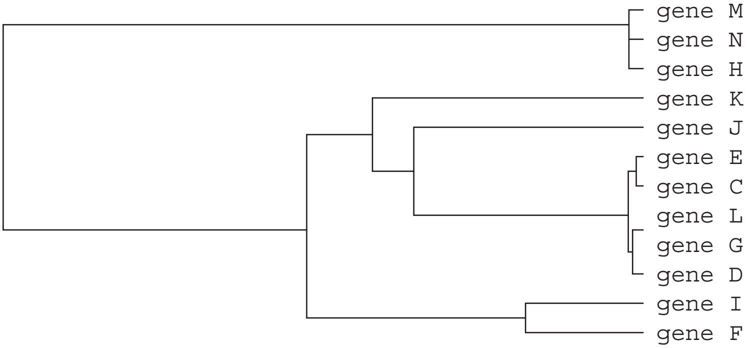

Figure MM6.1 Dendrogram of clustered genes from Table MM6.3 and Figure 6.8.

a different overall clustering of the data. Although there are mathematical methods for evaluating the extent to which clusters agree with the input similarity measurements, the last word in cluster evaluation belongs to the investigators who form and test hypotheses based on the clusters.

MATH MINUTE DISCOVERY QUESTIONS

1. Compute the correlations between cluster [NH] and all other objects, forming a table similar to Table MM6.3.
2. Explain why iteration 4 of the hierarchical clustering algorithm joins gene M with cluster [NH].
3. What new correlations must be computed in iteration 5 of the hierarchical clustering algorithm?
4. How many correlations must be computed to perform the first iteration of hierarchical clustering in the DeRisi diauxic shift data?