

## Exploring Correlation

The accompanying Excel spreadsheet (*correl\_explore\_scenarios.xls*) illustrates the concept of the Pearson correlation coefficient as a measurement of similarity between gene expression patterns. Each of the four scenarios in the spreadsheet begins with log-transformed gene expression ratios of two genes, as measured in eight different samples. We will refer to the set of eight numbers for a particular gene as a “gene expression pattern,” or simply “pattern.” The correlation coefficient between the two gene expression patterns is calculated by Excel and displayed in the grey-shaded area to the right of the pattern data.

The first graph for each scenario (on the left hand side) plots the gene expression pattern for each of the two genes. One way to think about the correlation coefficient is as a measure of how well the two patterns “track” each other.

If the two patterns tend to go up and down together, from one sample to the next, then the patterns are highly positively correlated. The patterns in Scenario II have a fairly large positive correlation. The largest possible value for correlation is 1, and this occurs when the change from one sample to the next for one gene, divided by the change from one sample to the next for the other gene, is always the same number. In other words, the two gene expression patterns do not have to have to be the same order of magnitude to be highly correlated. For example, one gene may have values between  $-1$  and  $1$ , while the other gene has values between  $-100$  and  $100$ .

If the two patterns tend to be opposites of one another, i.e. one goes up while the other goes down, as you move from one sample to another, then the patterns are highly negatively correlated. The smallest possible value for correlation is  $-1$ .

The second graph for each scenario (on the right hand side) plots the log-transformed gene expression ratio for each sample as a point in the plane. The horizontal axis represents Gene 1, and the vertical axis represents Gene 2. The line of best fit (i.e. regression line) is shown on each graph of this type. If the line of best fit has a negative slope, the two patterns are negatively correlated; if the line has a positive slope, the two patterns are positively correlated. Note that the slope of the line does **not** measure the magnitude of the correlation. Rather, the magnitude of the correlation is determined by how close the points are to the line of best fit. If they are very close, the magnitude is large (near 1 or  $-1$ ). If they are scattered far from the line, the magnitude is near 0. The patterns in Scenario I have a correlation near 0.

The following exercises guide you through a brief exploration of the correlation coefficient. Answers are on the last page of this document.

1. In Scenario I, a single number can be changed for Gene 1 that results in dramatic changes in the correlation. Use the two graphs for the scenario to guide your experimentation of the following changes.
  - a. Change a single sample for Gene 1 that causes the correlation to jump up to approximately 0.68.
  - b. Change Scenario I, Gene 1, Sample 8, from 100 to  $-150$ . Note that the correlation jumps down to approximately  $-0.63$ . Try to explain this jump by seeing what changes in each of the two graphs.
2. To help answer the following, first notice that in Scenario II, the pattern for Gene 2 is evenly spaced between 10 and 80, changing in increments of 10.
  - a. Change the pattern for Gene 1 in Scenario II such that the correlation is exactly 1. You will need to change all but one or two of the values.
  - b. Change the pattern for Gene 1 in Scenario II such that the correlation is exactly  $-1$ . You will need to change all but one or two of the values.
3. Scenario III illustrates how sensitive the correlation can be to small changes. Here we examine a gene whose log ratio changes substantially across samples and a gene with essentially constant log ratio across samples.
  - a. Find a pair of samples for which Gene 2 can be changed from 7 to 6, resulting in a much larger positive correlation.
  - b. Return the two samples found in part (a) to their original values of 7, and find a new pair of samples for which Gene 2 can be changed from 7 to 6, resulting in a fairly large negative correlation.
4. Scenario IV shows that correlation is undefined if one of the patterns is constant across samples. As in the previous scenario, changing just one of the values for Gene 2 has a significant effect on the correlation.
  - a. Change the value for sample 1 from 4 to 3, and note the effect on correlation.
  - b. Change the value for sample 8 from 4 to 3, and note the effect on correlation.
  - c. Explain why one of these changes has a greater magnitude effect than the other.
  - d. Which single change from 4 to 3 would give the correlation nearest to 0? Why?

1. Possible answers:
  - a. Change Scenario I, Gene 1, Sample 4, from 100 to  $-150$ . The correlation jumps up to approximately 0.68.
  - b. Change Scenario I, Gene 1, Sample 8, from 100 to  $-150$ . The correlation jumps down to approximately  $-0.63$ .
2.
  - a. Change the pattern for Gene 1 to be evenly spaced and increasing, for example, increasing from  $-100$  to 110 in increments of 30. You can watch the correlation steadily approach 1 as you change the numbers for samples 1 through 8.
  - b. Change the pattern for Gene 1 to be evenly spaced and decreasing, for example, decreasing from 110 to  $-100$  in increments of 30.
3.
  - a. By changing only samples 2 and 6 for Gene 2 from 7 to 6, the correlation jumps to nearly 0.78.
  - b. By changing only samples 4 and 8 for Gene 2 from 7 to 6, the correlation falls to approximately  $-0.38$ .
4.
  - a. Changing sample 1 causes correlation to jump to 0.632.
  - b. Changing sample 8 causes correlation to jump to  $-0.577$ .
  - c. The first change has greater magnitude impact on the correlation because Sample 4, with Gene 1 value of 100, keeps the line from dropping too far on the right when Sample 8 is changed. The line tries to be close to all sample points. The closest point to Sample 1 on the left end is Sample 2, but it is further from Sample 1 than Sample 4 is from Sample 8, so the “pull” on the line when Sample 8 is changed is not as great.
  - d. Changing Sample 3, Gene 2, from 4 to 3 gives a correlation of 0.027. No other single change from 4 to 3 results in a correlation this close to 0. The reason this correlation is so near 0 is that this Gene 1 value (10) is closest to the average, so changing its Gene 2 value has little effect on the line of best fit.