Value Added: Blending Math into a High School Genomics Lab

Laurie J. Heyer and A. Malcolm Campbell

Teacher Instructions and Student Activities Designed to Accompany DNA Chips: From Genes to Disease

Wet lab curriculum available at
http://www.bio.davidson.edu/projects/gcat/HSChips/HSchips.html

Mathematical Analysis of Gene Expression Ratios

Introduction

Your microarray experiment measures relative levels of mRNA gene expression in cancerous and healthy tissue of a single patient for six genes. The data from this experiment consists of six ratios. The numerator of each ratio is the gene expression level in cancerous tissue, and the denominator of each ratio is the gene expression level in healthy tissue. Now you will use the ratios from your microarray image and compare your data to ratios gathered from other microarrays.

Goals

Comparing your data to other data gathered from the same patient will help you understand the amount of experimental variation (sometimes called noise) in microarray experiments. Comparing similar data from several other patients will demonstrate how microarray experiments enable a better understanding of the genetic causes of particular cases of cancer. You will also see how to use microarray data to predict which patients will benefit most from chemotherapy and which patients will benefit least – the beginnings of personalized medicine

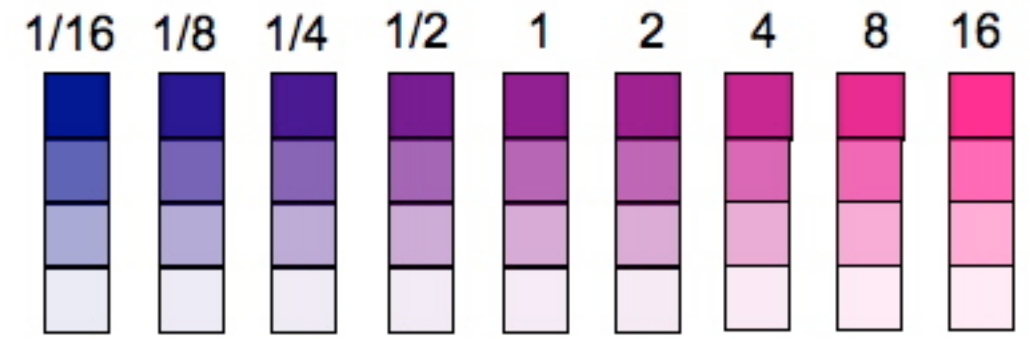
1. Determining ratios

By describing each spot on your microarray as different shades and intensities of pink, purple and blue, you have characterized the expression level of each gene in cancerous and healthy tissue. To effectively compare your results to those of other groups, you need a more quantitative measure than subjective phrases like "dark blue" or "pinkish purple." The goal of this activity is to turn colors into numbers that you can use in the remaining activities.

The scale shown below represents both the shade and the intensity of colors you might see in your microarray. The shade ranges from blue to pink as you go from left to right in the scale. The intensity ranges from deep to pale as you go from top to bottom in the scale. Match the colors in your microarray to those in the scale. Your colors may not match exactly. Estimate the ratios as best you can, **interpolating** between numbers as necessary. Record each ratio in the space provided.

Interpolating is the process of selecting a number in between two given numbers in a table. In the scale below, you may decide that your color is halfway between the colors for 2 and 4. You would therefore interpolate to find the number halfway between 2 and 4, resulting in a ratio of 3.

Color Ratio Scale



Teacher notes: Try to keep groups working independently on this step. They should not all get exactly the same ratios. In Section 3, it will be important that there is some variability in the ratios across different lab groups.

Gene expression ratios

Gene 1	Gene 2	Gene 3

Gene 4	Gene 5	Gene 6

Questions

1. What range of ratios could indicate that a gene was not expressed in cancerous tissue, but was expressed in healthy tissue?

   Answer: A ratio of 0, since the numerator would be 0. Noise and uncertainty might lead to a ratio near 0, but not exactly 0.

2. What range of ratios could indicate that a gene was not expressed in healthy tissue, but was expressed in cancerous tissue?

   Answer: Technically, the numerator would be greater than 0, but the denominator would be 0, resulting in an undefined ratio. Experimental noise might lead to a very large ratio if the denominator was judged to be a very small number, but not exactly 0. In the language of calculus, the limit of the ratio is infinity as the denominator approaches 0, but it is not precisely correct to say that the ratio "equals" infinity.

3. What range of ratios could indicate that a gene was equally expressed in both cancerous and healthy tissue?

   Answer: A ratio of 1, or very near 1, indicates equal expression in both conditions (equal numerator and denominator).

4. What range of ratios could indicate that a gene was not expressed in either cancerous or healthy tissue?

   Answer: If the gene is not expressed in either condition, the numerator and denominator are both 0, and the ratio is undefined. In calculus, if the numerator and denominator both approach 0, this is called an indeterminate form, and the limit of the ratio could be 0, a positive constant, or infinity. Likewise, the ratio might be judged to be near 0, or very large, or near 1, depending on noise in both the pink and blue channels.

5. Is it easier to determine the ratio when the expression levels are high (deep colors) or low (pale colors)? Explain your answer by relating the colors you see to the amount of mRNA produced.

   Answer: As explained in #4, the most difficult ratio to determine is when both expression levels are low. If at least one condition has a high expression level, the ratio can be determined much more easily.

6. Explain why there are more possible gene expression ratios than those shown in the color scale above.

   Answer: A ratio could be any non-negative real number. There are an infinite number of possibilities between each of the values shown in the color scale.

7. Explain how a single gene expression ratio (*e.g.*, 4) can correspond to many different levels of gene expression in the cancerous and healthy tissue samples.

   Answer: A ratio of 4 could result from a numerator of 40 and denominator of 10, or a numerator of 500 and a denominator of 125, or a numerator of 12436 and a denominator of 3109, or infinitely many other combinations of numerator and denominator.

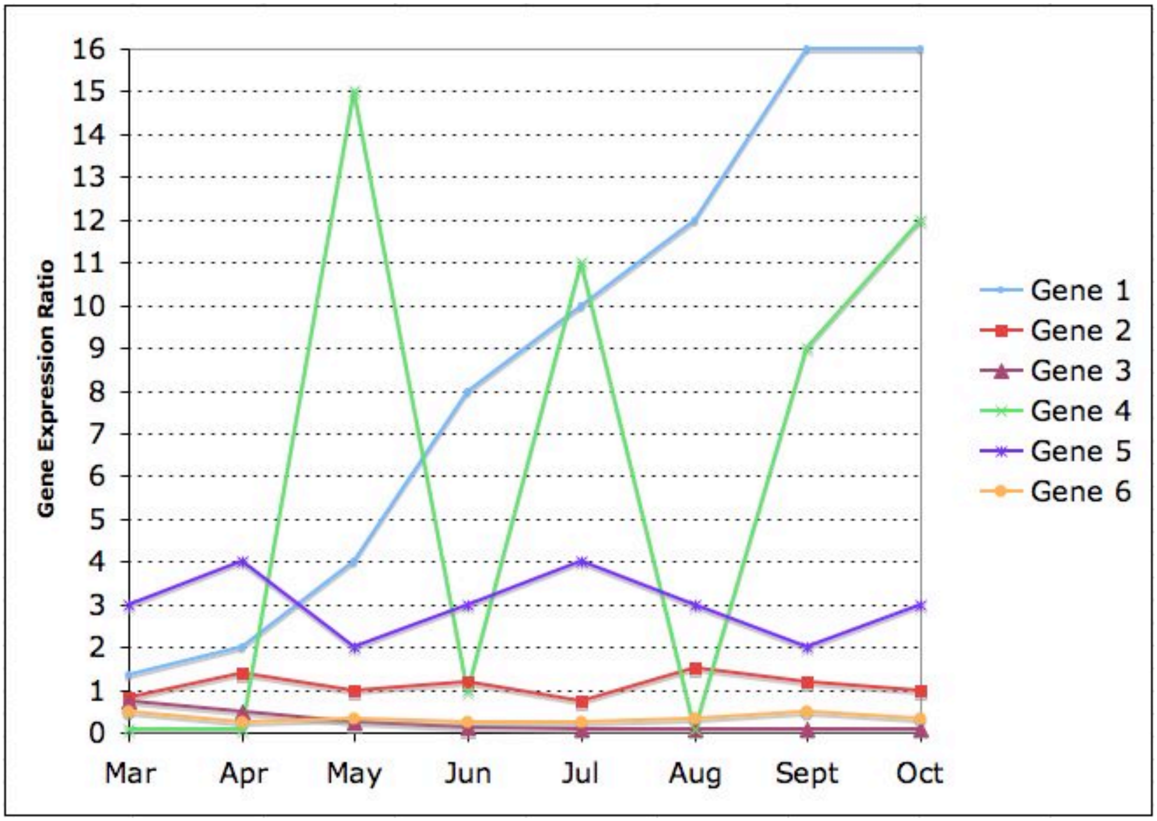
2. Transforming ratios

The goal of this activity is to see why it is useful to mathematically transform gene expression ratios, and to perform this transformation on the ratios from your microarray. Suppose the following gene expression ratios were measured in a single patient over a eight month time period, using one microarray per month, as the lung cancer progressed. Each microarray measures the expression ratios of six genes.

Table 1: Gene expression ratios during lung cancer progression

Gene	Mar	Apr	May	Jun	Jul	Aug	Sept	Oct
1	1.33	2	4	8	10	12	16	16
2	0.8	1.4	1	1.2	0.75	1.5	1.2	1
3	0.75	0.5	0.25	0.125	0.10	0.0833	0.0625	0.0625
4	0.1	0.08	15	1	11	0.07	9	12
5	3	4	2	3	4	3	2	3
6	0.5	0.25	0.33	0.25	0.25	0.33	0.5	0.33

Graph the ratio data in Table 1 on the axes provided below. Each gene should be graphed as a line so you should produce 6 different lines over the 8 month time period. Use different data point markers (*e.g.*, x, o, *), different line styles (*e.g.*, dashed, dotted) and/or different colors for each line, so you can easily tell them apart. Create a legend for your graph.



Questions

8. Look at the ratios for Gene 4, which range between 0.07 and 15, with no apparent pattern of increasing or decreasing. Explain how these values could result from very low gene expression levels (*i.e.*, pale colors) in both the cancerous and healthy tissues.

   Answer: The expression level of this gene is very low in both cancerous and healthy tissues. The true ratio could be much more consistent across the months of the study than our data indicate, but experimental noise leads us to a different value each time. See the answer to #4 for more discussion of this issue.

9. By looking at your graph, which genes would you be most interested in studying further to understand the progression of cancer? Support your choices with data.

   Answer: Looking at this graph, it appears that only Gene 1 is really changing much over the months of the study, so it would be interesting to study. Genes 3, 5 and 6 seem to have fairly consistent ratios, so while they may be indicating something about the presence of the cancer, or simply that the person is sick, they do not seem to be related to the progression of the cancer.

10. Convert the ratios for Gene 3 from decimals to fractions, and reduce the fractions. (Note that the 3 in 0.0833 is repeating.) Compare these fractions to the ratios for Gene 1. What pattern do you notice? Was this pattern easy or difficult to see in your graph? Does this change your answer to Question 9?

    Answer: $\boxed{\frac{3}{4}, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{12}, \frac{1}{16}, \frac{1}{16}}$ Notice that each fraction is the reciprocal of the ratios for Gene 1. This means that Gene 3 is being repressed to the same extent that Gene 1 is being induced, and the amount of repression is changing steadily over the months of the study. Now it appears that Gene 3 would be just as interesting to study as Gene 1 when trying to understand the progression of the cancer.

Compute the base 2 logarithm of each ratio in Table 1. Record your results in Table 2.

A **logarithm** is the power to which a base must be raised to produce the desired value. For example, $\log_2 8 = 3$ because 2 must be raised to the power of 3 to get 8, *i.e.*, $2^3 = 8$. Similarly, $\log_2 \tfrac{1}{16} = -4$, because $2^{-4} = \tfrac{1}{16}$. Using this exponent rule, you can compute (or at least estimate) base 2 logs in your head. However, it is also useful to know how to compute logs with a calculator.

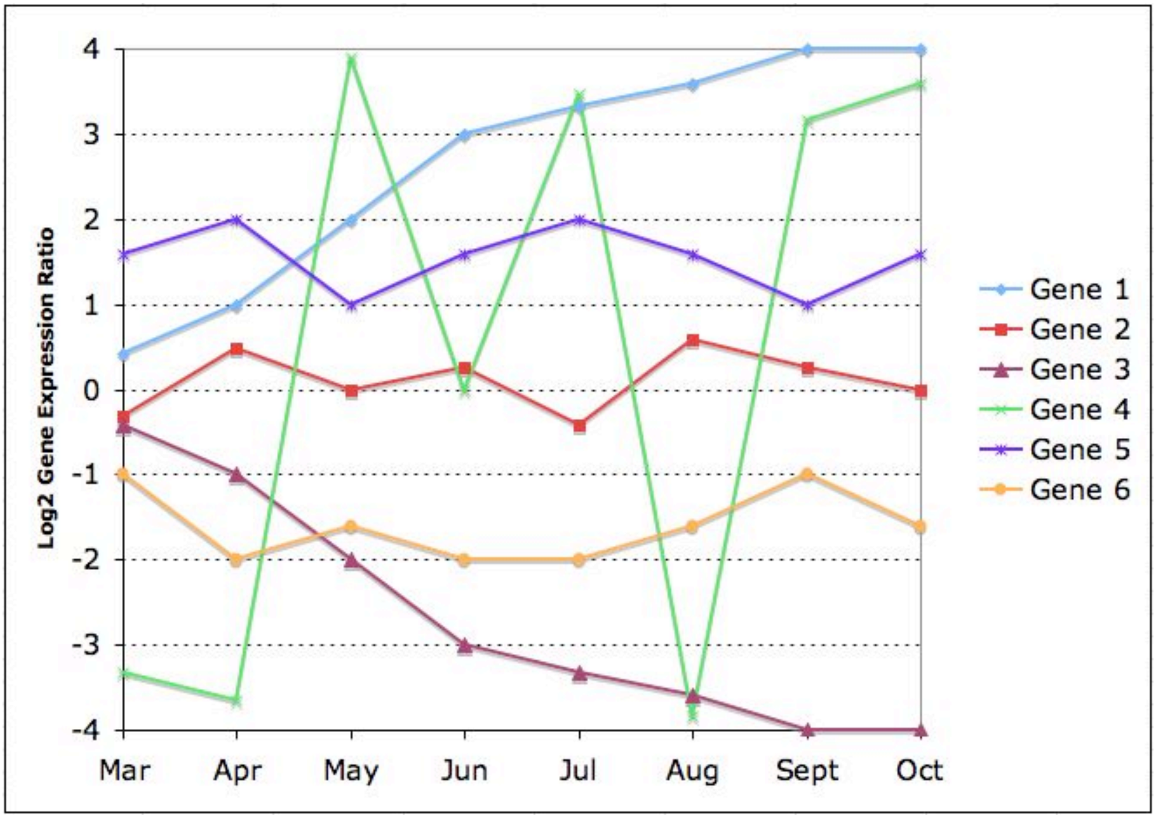
Log functions are built into most calculators and software such as Excel, but some calculators can only compute base 10 and natural (base *e*) logs. To compute a base 2 log with your calculator, you can use one of the following change of base formulas:

$$\log_2 x = \frac{\log_{10} x}{\log_{10} 2} \quad \text{or} \quad \log_2 x = \frac{\ln x}{\ln 2}$$

Table 2: Log$_2$ gene expression ratios during lung cancer progression

Gene	Mar	Apr	May	Jun	Jul	Aug	Sept	Oct
1	0.411	1.000	2.000	3.000	3.322	3.585	4.000	4.000
2	-0.322	0.485	0.000	0.263	-0.415	0.585	0.263	0.000
3	-0.415	-1.000	-2.000	-3.000	-3.322	-3.586	-4.000	-4.000
4	-3.322	-3.644	3.907	0.000	3.459	-3.837	3.170	3.585
5	1.585	2.000	1.000	1.585	2.000	1.585	1.000	1.585
6	-1.000	-2.000	-1.599	-2.000	-2.000	-1.599	-1.000	-1.599

Graph the log$_2$ gene expression ratios you recorded in Table 2 on the axes given below. As before, represent each gene with a single line, using the same color and symbol for each gene that you used in the previous graph.



11. By looking at the graph of $\log_2$ ratios, which genes would you select for further study of lung cancer progression? Support your choices with data. Reread your answers to Questions 9 and 10, and explain why the log-transformed ratios are more useful than ratios.

    Answer: The graph of log transformed ratios makes it clear that the expression ratios of both Gene 1 and Gene 3 are changing over time, and would be natural to select for further study of cancer progression. The reciprocal nature of the patterns is clearly illustrated in the log-transformed graph without having to find genes for which it would be appropriate to take reciprocals. This graph depicts repression of genes with the same magnitude as induction, so that genes that decrease with cancer progression are just as obvious as those that increase.

12. Explain why base 2 is useful when log transforming ratios. When would it be more useful to use base 10?

    Answer: Base 2 is useful for identifying doubling of halving of ratios. Every unit step on the vertical axis represents a doubling in the positive direction, or a halving in the negative direction. For example, the expression ratio of Gene 1 is doubled between April and May, and again between May and June. On the other hand, the expression ratio of Gene 3 in September is half what it was in June. Base 10 is useful for identifying changes in order of magnitude, and would be more helpful if the ratios were much larger, between 0.001 and 1000, for example. Note that $\mathrm{Log}_{10}(0.001) = -3$, and $\mathrm{Log}_{10}(1000) = 3$.

13. What mathematical problem could arise when you take logarithms of ratios? Suggest a way to handle this problem.

    Answer: If a ratio was determined to be exactly 0, rather than a small positive number, then the logarithm (in any base) is undefined. Excel will print #### for the answer to the log of 0. Numerically, it might be helpful to use a very large number, e.g. 99999 to represent this situation in the log-transformed data. Graphically, this causes a problem with scaling the axis to see the more reliable data. These undefined values could simply be left off of the graph, leaving a gap in the line representing that gene.

Log transformed ratios are not only superior to ratios when you are looking for interesting patterns in graphs, they are also better for every kind of mathematical analysis of gene expression ratios. Therefore, we will work exclusively with log-transformed ratios. Record the $\log_2$ ratios from your microarray in the space given below.

Log_2 gene expression ratios

Gene 1	Gene 2	Gene 3

Gene 4	Gene 5	Gene 6

3. Measuring variability

All laboratory experiments involve some degree of variation. Just as you do not get the exact same number of granules each time you measure a cup of sugar, you do not get the exact same volume of reagents each time you repeat an experiment. Even instruments that seem to measure things very precisely still have some variability. In this experiment, each lab group in your class is analyzing the same patient, but you will not all get exactly the same ratios. Part of this variability is due to slightly different amounts of reagents used by different groups. Part of the variability is due to different decisions made by the groups when they converted colors into ratios. Can you think of any other sources of variability in your results?

Mathematically, you can quantify measurement errors and other experimental variability using a quantity called **variance**. Often, investigators use the square root of the variance, called **standard deviation**, because its units are the same as the units of the original measurement. *The goal of this activity is to measure the variability of your microarray experiment using standard deviation.*

In the following table, record the $\log_2$ gene expression ratios obtained by each lab group in your class. Your teacher will number the lab groups so that everyone's table looks exactly the same. After you have recorded all the log-transformed ratios, compute the values in the last five rows as follows:

- N –the number of rows containing data (equal to the number of lab groups in your class, unless some groups failed to get readings for all six genes)
- Avg – average the values in the column. In statistics, this quantity is called $\bar{x}$, pronounced "x bar". If the ratio for group #1 is denoted by R_1, then the equation for average is $\bar{x} = \frac{\sum_{i=1}^{N} R_i}{N}$
- Sum of squares – square each value in the column, and add all the squared values, *i.e.*, compute $\sum_{i=1}^{N} R_i^2$
- Variance – square the column average, multiply the result by N, and subtract this value from the sum of squares; divide the entire result by N-1. The equation that represents what you just did is $\text{Var} = \frac{\sum_{i=1}^{N} R_i^2 - N \cdot \bar{x}^2}{N-1}$
- Std dev – take the square root of variance

Example: Suppose there are 5 lab groups in your class, and their expression ratios for Gene 1 are 8, 6, 7, 10 and 8. Then N=5, $\bar{x} = \frac{8+6+7+10+8}{5} = 7.8$, and the sum of squares is $8^2 + 6^2 + 7^2 + 10^2 + 8^2 = 313$. Therefore, the variance is $\frac{313-5\cdot(7.8)^2}{5-1} = \frac{8.8}{4} = 2.2$, and the standard deviation is $\sqrt{2.2} \approx 1.48$.

Teacher notes:

Be sure each lab group knows their number before beginning to fill out this table. It might be helpful to have a blank overhead of the table, and ask each group to fill in their values for everyone else to copy into their own tables. Average, variance and standard deviation are common descriptive statistics that can be calculated with Excel using the functions *average*, *var*, and *stdev*, respectively.

Lab Group	**Gene**					
	1	2	3	4	5	6
1						
2						
3						
4						
5						
6						
7						
8						
9						
10						
11						
12						
13						
14						
15						
16						
17						
18						
19						
20						
N						
Avg						
Sum of squares						
Variance						
Std dev						

Questions

14. Which gene has the most variable expression ratio? Why did the log ratios of this gene vary more from group to group than the log ratios of other genes?

    Answer: The gene with the largest standard deviation is the most highly variable. It may be Gene 4, since some groups will say the ratio is 0, some will say 1, and some may give other values. However, this variability is somewhat artificial, representing the different interpretations of an essentially clear spot. There will also be a numerical effect of using a very large number to represent the log of 0. Other likely candidates for the most variable gene are Gene 5 and Gene 6. In this experiment, variability is caused by low color intensity, which is difficult to assign to a ratio in the color scale. This same effect is seen in real microarray experiments: low intensities lead to more variability in ratios.

4. Clustering gene expression profiles

In Section 2, you gained hands-on experience with one useful application of microarray data: discovering genes that indicate the presence or progression of a disease such as lung cancer. You characterized genes by their pattern of gene expression over time, and found that Gene 1 is increasingly induced over time, and Gene 3 is increasingly repressed over time. In Section 3, you learned how to measure variability in expression levels across lab groups. Every time the experiment is repeated by the same or different lab groups, you should expect to get slightly different answers because of experimental errors.

However, you have only looked at gene expression ratios for a single patient so far. Is this patient representative of everyone with lung cancer? Are there sub-categories of lung cancer, or are all cases of lung cancer the same? What else could we learn about lung cancer by collecting data from more patients? In this section, you will learn how to compare gene expression patterns of different patients, and **cluster** (*i.e.,* group together) patients with similar patterns. If a group of patients have similar gene expression patterns across this set of six genes, and they also have similar clinical outcomes (for example, they both responded well to a particular type of chemotherapy), then we may be able to predict the clinical outcomes of future patients by measuring their gene expression levels in these six genes. This sort of personalized medical treatment is one of the major applications of the human genome project.

Table 3 gives gene expression ratios for six different genes in twelve different patients:

Table 3: Log$_2$ gene expression ratios in patients A-L

	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6	Avg	Std dev
Patient A	0	3	3.58	4	3.58	3	2.86	1.45
Patient B	0	1.58	2	2	1.58	1	1.36	0.76
Patient C	0	2	3	3	3	3	2.33	1.21
Patient D	0	0	0	-2	-2	-3.32	-1.22	1.42
Patient E	0	1	1.58	2	1.58	1	1.19	0.70
Patient F	0	-1	-1.60	-2	-1.60	-1	-1.20	0.70
Patient G	0	2	3	2	0	-1	1.00	1.55
Patient H	0	1	0	1	0	1	0.50	0.55
Patient I	0	0	0	0	1.58	1.58	0.53	0.82
Patient J	0	1	1.58	2	1.58	1	1.19	0.70
Patient K	0	-1.60	-2	-2	-1.60	-1	-1.37	0.76
Patient L	0	-3	-3.59	-4	-3.59	-3	-2.86	1.45

Questions

15. Fill in the blank cells in the final two columns of Table 3 for each of the twelve patients, using the statistical methods you learned in Section 3.

    Answer: See entries in Table 3 above.

Correlation is a way of comparing two lists of numbers (in mathematics, these lists are called **vectors**) to see how well the lists of numbers, or vectors, track one another. To compute the correlation between two patients, follow this three step process:

a. Compute the **sum product** of the corresponding vectors of log$_2$ gene expression ratios. To find the sum product, multiply the corresponding entries in each list, then sum all these products. For example, the sum product between Patient A and Patient B is (0)(0) + (3)(1.58) + (3.58)(2) + (4)(2) + (3.58)(1.58) + (3)(1) = 28.6068. In mathematics, the sum product is also called the **dot product** or the **inner product**.

b. Multiply the average of one vector by the average of the other vector and then multiply by *n*, the number of entries in each vector. In our example using Patients A and B, multiply 2.86 by 1.36, then by 6, to get 23.3795.

c. Subtract the result of Step b from the result of Step a, divide by the standard deviation of the first vector (found in the table you completed), divide by the standard deviation of the second vector, and finally, divide by *n* – 1. For example, for Patients A and B, subtract 23.3795 from 28.6068 to get 5.2273. Then divide by 1.45, divide by 0.76, and finally, divide by 5. The result is the **correlation coefficient** between Patient A and Patient B, 0.9442.

Questions:

16. Using the three-step process described above, compute correlation coefficients between every pair of patients. Enter your results in the blank cells of the following table. You do not need to compute values for the shaded cells, because the correlation is a symmetric relationship. For example, the correlation between patients A and B is the same as the correlation between patients B and A.

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

17. How many correlation coefficients must be computed in the above table (including the one given value)? How many correlation coefficients would need to be computed if there had been 20 patients? Find a general expression for the number of correlation coefficients that need to be computed if there are *n* patients.

Answer: There is 1 in the 2nd column, 2 in the 3rd column, and so on, up to 11 in the 12th column. So a total of $1+2+3+\cdots+11=\frac{11\times 12}{2}=66$ correlation coefficients must be computed. If there had been 20 patients, a total of $1+2+3+\cdots+19=\frac{19\times 20}{2}=190$ correlation coefficients would need to be computed. In general, the formula for counting the number of correlation coefficients when there are *n* patients is $\sum_{k=1}^{n-1} k=\frac{(n-1)n}{2}$, a formula that is found in most calculus textbooks in the section on Riemann Sums. Another way to think about this formula is the number of ways to choose 2 patients (a pair) from the *n* patients, which is the **binomial coefficient** $\binom{n}{2} = \frac{n(n-1)}{2}$. This formula is found in most statistics textbooks, and is also a fundamental concept in the mathematical subfield of combinatorics, also called discrete mathematics.

The clustering **algorithm** begins by finding the two patients that are most similar across their expression of the six genes. In this example, Patients J and E are the two most similar; they are actually identical! Join these two together into a single "average" patient by averaging their two expression vectors, and computed the correlation coefficient between this average patient and all other individual or average patients. (In this case, because J and E are identical, this step can be skipped.) Remove the individuals J and E from further consideration.

An **algorithm** is a step-by-step process that performs a computational task. Computer scientists often study algorithms to find ways to make them more efficient. Sorting, pattern matching, and clustering are examples of algorithms that are important in biological applications.

Continue this process, joining two patients or "average" patients together at each step, until all patients have been clustered. You may wish to draw your clustering results in a hierarchical tree, showing which two patients or "average" patients were joined at each step. You can represent the correlation coefficient between the patients that were joined by making the branches of the tree join at that height.

Teacher notes: The hierarchical clustering algorithm proceeds as follows:

First, find the two most similar patients in the entire set of patients. As described in the main text, patients E and J are the most similar since $r_{\mathrm{EJ}} = 1$. Join these together into a cluster, denoted [EJ]. Cluster [EJ] is added to the list of available objects, and the single patients E and J are removed from the list.

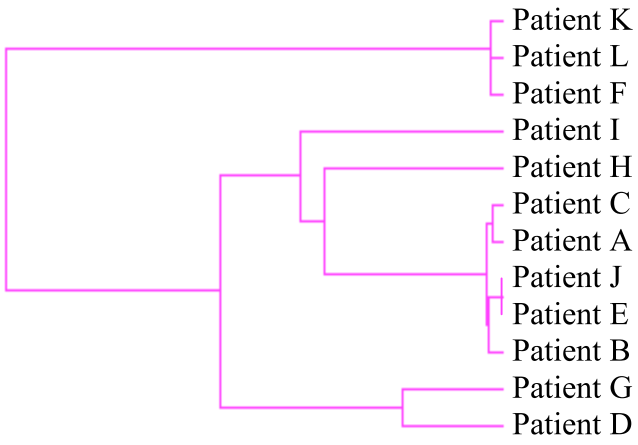
Now join the next two most similar objects, using the procedure described above. (Note that in this case, the average of E and J is equal to both E and J, so we are saved the job of computing new correlations.) The most similar patient to the cluster [EJ] is patient A, with $r_{\mathrm{AE}} = r_{\mathrm{AJ}} = 0.95$. However, patient C and cluster [EJ] are not the two most similar objects; rather patients A and C are, with $r_{\mathrm{AC}} = 0.96$. Thus we join patients A and C to form cluster [AC].

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

B	D	F	G	H	I	K	L	[EJ]
0.90	-0.48	-0.93	0.32	0.33	0.32	-0.90	-0.99	0.93

Now we see that the most similar object to [AC] is cluster [EJ], with a correlation of 0.93. Patient B is even more similar to [EJ], since $r_{BE} = 0.94$. But the two most similar objects now are patients F and L, with $r_{FL} = 0.95$. Therefore, we join patients F and L to form cluster [FL]. We have now completed 3 iterations of the hierarchical clustering algorithm. The entire clustering process for these 12 patients takes 11 iterations; the steps are summarized in the following table. Note that the final object created is the clustering of all 12 patients shown in the hierarchical tree, also called a **dendrogram**.

Iteration	Two most similar objects		Correlation	New Object
	Object 1	Object 2		
1	J	E	1.00	[EJ]
2	C	A	0.96	[AC]
3	L	F	0.95	[FL]
4	K	[FL]	0.95	[KFL]
5	[EJ]	B	0.94	[EJB]
6	[AC]	[EJB]	0.94	[ACEJB]
7	G	D	0.60	[DG]
8	H	[ACEJB]	0.29	[HACEJB]
9	I	[HACEJB]	0.19	[IHACEJB]
10	[IHACEJB]	[DG]	-0.12	[IHACEJBDG]
11	[KFL]	[IHACEJBDG]	-0.96	[KFLIHACEJBDG]



Questions

18. What group of five patients are highly similar to each other (at correlation greater than 0.9?

    Answer: ACEJB, joined at iteration 6, at correlation 0.94. This group is most easily seen in the dendrogram above.

19. What group of three patients are highly similar to each other (at correlation at least 0.95), but very dissimilar from the first group of five patients?

    Answer: KFL, joined at iteration 4, at correlation 0.95. Again, the answer to this question is easily seen in the tree.

20. If all lung cancers fell into two categories of gene expression patterns, what does this tell you about the single disease we call lung cancer? How might this information affect cancer patient diagnosis and treatment in the future?

    Answer: If there are two distinct gene expression profiles in lung cancers, perhaps the associated cancers are different enough to be considered different types of cancer. The causes, treatments and prognosis of the two subtypes of lung cancer may be very different. In the future, a gene expression profile might be part of the diagnosis to determine the optimal treatment.