BNFO 301 – Introduction to Bioinformatics
Problem Set 5 – Microarray Analysis and Statistics

1. Hihara et al [(2001) Plant Cell 13:793-806] examined the expression of genes of the cyanobacterium *Synechocystis* PCC 6803 in response to high light intensity. A summary of their microarray data is available within BioBIKE as the data table "Hihara2001". Four conditions were examined: 15 min, 60 min, 6 hrs, and 15 hrs of high light intensity, all compared to continuous low light. The conditions are labeled 1 through four, respectively.

1a. Find the relative intensity (60 min high light vs low light) of several of the genes you looked at on the Stanford MicroArray Database. You can do this using the following syntax (the gene SLL1461 is used as an example):

```
(RATIO-OF SLL1461 IN "Hihara2001" COLUMN 2)
```

1b. Does the data from the SMD correspond with the data of Hihara et al?

1c. Find the genes that are most highly expressed in 60 min high light intensity relative to low light intensity. Do this in two ways (try both). First write a loop that considers each gene and collects the name of the gene along with the number returned by RATIO-OF. Second, use RATIO-OF to act on all genes of S6803 at once. Connect the numbers to the genes in the following way:

```
(INTERLEAVE (GENES-OF S6803)
                     (RATIO-OF …))
```

Then sort the genes as follows (using an old command that will soon be replaced, sorry).

```
(SORT * 'GREATER-THAN :KEY 'SECOND)
```

(Using * in this way presumes that the list is the previous result. If it isn't, then use (RESULT n) where n is the line number).

1d. Find the genes that are LEAST highly expressed in 60 min high light intensity relative to low light intensity.

1e. Revisit Question 1b.

2. A reasonable expectation is that expression of genes involved in photosynthesis are affected by high light intensity. Let's see if that is or is not the case.

2a. Find the genes related to photosynthesis. You can do this by accessing the Gene Ontology (GO) category "photosynthesis". Do this in the following way:

```
GO.photosynthesis
```

Click on the resulting link and examine what Gene Ontology considers information pertinent to photosynthesis. You'll notice a provocative category labeled GO.Related-genes. Click on a few of the listed genes and decide if indeed these are genes related to photosynthesis. You can access all of the genes in this category in the standard way, using bracket notation:

```
GO.photosynthesis[GO.Related-genes]
```

Define a variable containing all the genes related to photosynthesis, according to GO.

2b. Find the expression ratios of the photosynthesis-related-genes at 60 min of high light intensity. What generality, if any, do you find?

2c. Let's test that generality. Find out what fraction of all *Synechocystis* genes have expression ratios less than one after 60 minutes of high light intensity. How many photosynthetic-related-genes have expression ratios less than one under the same conditions?

2d. Use a Chi-square test on the results of 2c. ***What do those results mean?***

2e. Do a large number of simulated experiments to answer the same question addressed by a Chi-square test.

3. Consider Chi-Square.

3a. Define a function that calculates chi-squared scores, given two input arguments: a list of integers and a list of expected frequencies. Here are two lists you can use to test your function, which you can think of as (1) the observed nucleotide counts (A, C, G, and T) for a gene 100-nt in length, and (2) the expected frequencies of the four nucleotides:

```
(LIST 28 22 28 22) (LIST .25 .25 .25 .25)
```

You should get back the result 1.44. **Provide the code for your function.**

Hint: You will need to loop through both lists simultaneously, accessing the observed counts for A and the expected frequency for A, then the observed counts for C and the expected frequency for C, etc. To do this, you might draw inspiration from the following model:

```
(FOR-EACH animal IN (LIST "cat" "dog" "pig")
     FOR adjective IN (LIST "feline" "canine" "porcine")
     DO (DISPLAY-LINE animal *tab* adjective))
```

3b. How do you ***interpret*** the 1.44 result from my example? You may use the figure to the right in your answer, and you'll also need a specific number that you can get off the web (see Resources & Links on the course web site). Run the example lists through your function again, but this time multiplying each number in the first list by 10. ***Now*** how do you interpret the results? **Provide perhaps two carefully wrought sentences, each of which uses words like "probability" and "population".**