Introduction to Bioinformatics
Problem Set 5: Statistics and Odds and Ends

1. Why did the search for FMRP in *Drosophila* work well when human FMRP protein was used as the query but failed so abysmally when the corresponding human gene was used as the query? Now's the time to find out.

1a. Hearken back to the tour *Search for FMRP in Drosophila* and consider item 14, the comparison of the human and *Drosophila* proteins. The comparison begins with a stretch of very high amino acid similarity, from the initial M (methionine) up to YK (tyrosine-lysine). How many amino acids are in that stretch?

1b. How many nucleotides are required at the beginning of the gene to encode those amino acids (M through YK)?

Let's get those nucleotides, from the human gene and from the *Drosophila* gene. If the amino acid sequences are extremely similar, then so should be the nucleotide sequences, no? The protein sequence found by Blast comes with a link to the corresponding gene, with the GenBank ID of LD09557. You determined during the tour that the GenBank sequence does not begin with the FMR gene but includes upstream sequence. You'll recall that the gene begins at nucleotide 423.

1c. In BioBIKE, DEFINE a variable (maybe `fly-seq`) as the beginning of the fly gene, using SEQUENCE-OF and specifying the FROM-GENBANK option. The argument should be the GenBank ID (in quotes). Also specify values for FROM and TO, so that you bring in only those nucleotides from the beginning of the gene to the end of the portion of the gene encoding Y and K. To check if you were successful, get the TRANSLATION-OF `fly-seq` and compare it to the amino sequence you considered in **1a**.

1d. Retrace your steps in the tour and find the GenBank ID for the human FMR gene and what nucleotide coordinate the gene starts with. DEFINE a variable (maybe `human-seq`) as in 1c, but using the appropriate GenBank ID.

1e. Align these two DNA sequences, using the ALIGNMENT-OF function. The argument should be a LIST consisting of two items: `fly-seq` and `human-seq`.

1f. Does the alignment look as good as the amino acid alignment? Why not? Do something to test your theory.

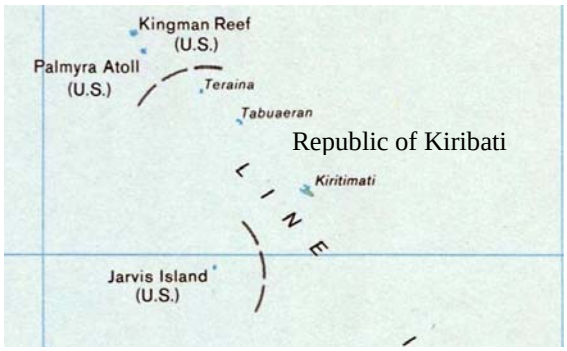
DEFINE-FUNCTION (See *Was Mendel Right? Part II* for examples in defining functions)

2. Define a function that accepts a number and returns its square.

3. Define a function that accepts a DNA sequence and returns a palindrome that begins with the given sequence. You'll want to know about the INVERSION-OF function in the STRING/SEQUENCE, String-production menu.

Statistics

4. Suppose that you are interested in the diversity of viruses in the world's coral reefs. You intend to sample viruses in various coral reefs and use their sequences to assess viral populations. You have chosen coral reefs adjoining two nearby south pacific islands (see figure at right), part of the Line Islands. One, Tabuaeran, is part of the Republic of Kiribati. The other, Palmyra, is a U.S. possession. Tabuaeran has a population of about 2500, while Palmyra is nearly uninhabited. It would be of interest to learn whether the coral reefs and the viruses that infect them differ in a nearly untouched reef as compared to one in which there has been significant human activity.*



Your plan is to compare viral sequences from each site. It would take about $40,000 to get good sequence coverage of a metagenomic sample, and you don't have that kind of money. Instead, you'll sequence clones of 200 viral sequences from each site and hope that fairly represents the total viral population.

4a. Before embarking on this project, let's step back and take a God-like view of the populations of viral sequences from the two sites. We'll do this by displaying the GC-fractions of metagenomes from Palmyra and Tabuaeran (nicknamed Tabu).

- Go to ViroBIKE (from the usual BioBIKE portal, click the ViroBIKE public site).
- Bring down the PLOT function (either from the Input-Output button or from the alphabetical list).
- This function requires a *data-list*, which can either be a list of values (to get one line on the graph) or a list of lists of values (to get multiple lines). We want to plot two sets of GC-fractions simultaneously, so select the *data-list* argument box and bring down the LIST function.
- Add another item so that LIST has two *item* boxes.
- Bring down GC-FRACTION-OF into the first *item* box.
- In the *entity* argument box of GC-FRACTION-OF, bring down CONTIGS-OF (from either the Genome button or the alphabetical list).
- In the *entity* box of CONTIGS-OF type `Palmyra` (<u>not</u> in quotes).
- Copy the GC-FRACTION-OF function with everything inside it and paste it into the other *item* box of the LIST function. Change `Palmyra` to `Tabu`.
- If you execute the function, you'll get a smush of many tens of thousands of points (or you'll time out trying). We need to organize the points by putting them into bins, counting how many contigs have GC-FRACTIONS between .30 and .31, how many

* Those hungry for more about coral reefs and microbial diversity can check out Dinsdale EA et al (2008). Microbial Ecology of Four Coral Atolls in the Northern Line Islands. PloS ONE 3(2):e1584.

between .31 and .32, etc. Mouse over the Options Icon of the PLOT function and select BIN-INTERVAL. Enter 0.01 into the *value* box.

- And for sake of esthetics, go back to the Options Icon and select X-LABEL. Type "GC Fraction" in the *value* box. Select Y-Label, and type "Frequency" in the *value* box.
- Now execute the function.

Does the solid line (Palmyra GC-fractions) coincide with the dashed line (Tabu GC-fractions? Which one is shifted to the right, i.e. to higher GC-fractions? What is the difference in the means of these two populations? 1.6% doesn't sound like a huge difference, but when you're looking at so many contigs (COUNT them if you don't believe me), it may well be significant.

4b. Back to real life. You don't have that picture of the viral sequence populations at the two sites – not enough money to get the sequences. Instead you're sequencing 200 viral fragments from each site. Let's simulate that situation by getting 200 samples each.

- DEFINE `palmyra-seqs` as the FIRST 200 CONTIGS-OF `palmyra`.
- Likewise, DEFINE `tabu-seqs` as the FIRST 200 CONTIGS-OF `tabu`.
- Redo the PLOT from **4a**, changing `palmyra` to `palmyra-seqs` and `tabu` to `tabu-seqs`. (Actually now that you have the sequences, CONTIGS-OF is no longer necessary. But it doesn't hurt).
- Oooh, spiky! Change the BIN-INTERVAL to .02 and try again.
- That's better, but is there any difference between the two? Calculate the difference in the mean GC-FRACTIONs as in **4a**, but with `palmyra-seqs` and `tabu-seqs`. ***Remember this number!***

There's still a difference, but it sure doesn't look as good on the graph!

4c. How can we test whether the difference in means is significant in this sample (as it is in the full population)? Suppose there is no real difference. Suppose that the two islands have the same virus, and when you sample at Palmyra and at Tabu, you're really sampling from the same pool. Let's simulate that situation.

- DEFINE `single-pool` as both `palmyra-seqs` and `tabu-seqs` JOINed together. Notice that the result (in the RESULT window) shows the Palmyra sequences in order. The Tabu sequences follow them (off screen).
- DEFINE `shuffled-pool` as `single-pool` SHUFFLEd. You'll find the SHUFFLE function under the List-Tables, List-Production. Notice that the result shows seemingly random sequences. It is actually `single-pool` in a random order.
- DEFINE `set1` as the FIRST 200 sequences of `shuffled-pool`.
- DEFINE `set2` as the LAST 200 sequences of `shuffled-pool`.

If Palmyra and Tabu have the same pool of viruses, then taking the first 200 of a mixed collection is no different from taking the last 200.

- DEFINE `mean1` as the MEAN of the GC-FRACTIONs-OF `set1`.

- DEFINE `mean2` in an analogous fashion.
- DEFINE `difference` as one mean SUBTRACTed from the other. Actually (since you just want to know if there is a significant difference between the Palmyra and Tabu means regardless of direction) surround the SUBTRACT function by ABS (absolute value).

Is the difference of the means GC-FRACTIONs of these two random subsets smaller than what you got with the real `palmyra-seqs` and `tabu-seqs`? Maybe yes, maybe no. What we really need to know is how frequently would random subsets of a single pool have a difference of means greater than what you observed with `palmyra-seqs` and `tabu-seqs`.

4d. Here's how we'll determine whether the observed difference in means is greater than what you'd expect by chance.

- Set up a PROGRAM (brought down from the Define button), consisting of the operations of **4c**. Excluding the definition of `single-pool`, there were six DEFINE operations. Use the More Icon to give PROGRAM six *form* boxes.
- Drag each of the operations of **4c** (except the definition of `single-pool`), in order, into the appropriate *form* box of PROGRAM.
- Execute the PROGRAM. It should give you the difference of means of two random subsets of `single-pool`, different from the difference you got in **4c**. Why is it different?
- Collapse the program (from the Action Icon at the upper left of PROGRAM.
- Name the collapsed PROGRAM box (from the Action Icon). Maybe something like difference between sets.

4e. Now that you know how to generate random subsets and determine the difference in their mean GC-fractions, you'll do it 100 times and ask how many of those times is the difference greater than the mean difference you observed with the Palmyra and Tabu samples.

- Bring down a FOR-EACH loop.
- This loop will have a Primary Control section, a Body section, and a Results section. Hide the other four sections.
- Open up the Body, and drag the PROGRAM box into the *body* form.
- From the Primary Control menu, choose *number* FROM *n1* TO *n2*.
- Add values to the *var*, *first value*, and *last value* boxes so that the variable `trial` goes from 1 to 100.
- You want to count the trial only if the difference between mean1 and mean2 is less than the difference of the means you actually observed in **4b**. Choose from the Results menu *when... count.*
- For *condition*, choose ORDER from the Flow-Logic button. Click the *comparison* icon and select >. Put `difference` (the variable you defined in your program) in the

first *any* box, and put the difference of means you calculated in **4b** in the second *any* box. Be sure the difference is a positive number.

- In the *value* box governed by COUNT, put `trial`. When `difference` is greater than the observed difference, you'll COUNT the trial.
- Execute the FOR-EACH loop. How many times out of 100 trials was the difference in means of the random subsets greater than the observed mean?
- Execute the loop again. Now how many times?
- (brought down from the Define button), consisting of the operations of **4c**. Excluding the definition of `single-pool`, there were six DEFINE operations. Use the More Icon to give PROGRAM six *form* boxes.
- What conclusion can you draw from this simulation?

That was a lot of work Surely there's an easier way to assess whether two sets of numbers are significantly different or instead are consistent with coming from a common pool.

4f. Yes indeed, there is an easier way! Actually, it's essentially the same thing you did but with a shortcut.

- Calculate the MEANs and STD-DEVs (standard deviation) of `palmyra-seqs` and `tabu-seqs.`
- Go to the t-test calculator (see the Links and Resources page of the course web site).
- Check the "Enter mean, SD, and N" radio button.
- Enter the mean, SD (standard deviation), and N (number of sequences) for group 1 (`palmyra-seqs`) and group 2 (`tabu-seqs`). Click `Calculate now.`
- Find the **P value**. How does it compare with the number of times the difference between the GC fractions of the random subsets exceeded the observed difference comparing `palmyra-seqs` and `tabu-seqs`?
- Note the t score and the degrees of freedom (df). The latter is always the sum of the N values minus 2.
- Go to the t-test probability calculator, also available from the Links and Resources page.
- Type in the t-value and the degrees of freedom. Then click Calculate. You should get the same P value. At the left, you'll see a graphical representation of what P value means. About 30% of the curve is colored red or blue. About 30% of the time the difference between random subsets exceeds the observed difference.

4g. What is a t-test? Suppose a t-test gives a P value of 0.05. How would you put this value in a meaningful English sentence?

5. In which of the following cases would a **chi-square test** be useful? In which would a **t-test** be useful?

a. Are the genes of *Anabaena* PCC 7120 longer, on average, than the genes of ss120?
b. Mendel counted 705 purple flowers and 224 white flowers. Is this reasonably close to a 3:1 ratio?
c. Your unidentified viral sequence has dinucleotide counts of {AA = 60, AC = 72, AG = 52, …}. Could the fragment reasonably have been derived from the virus Mx8?
d. Is expression of the gene encoding melanin induced by ultraviolet radiation? I've measured expression of the gene 12 times: 6 times with UV and 6 times without.

6. What is the average size of the genes of *Prochlorococcus marinus* ss120? Of *Anabaena* PCC 7120? Are the genes of ss120 significantly smaller than the genes of A7120? Answer the question by doing a t-test. How would you answer the same question with a simulation?

7. Are genes in *Synechococystis* PCC 6803 (S6803) that are annotated as "hypothetical" biased towards small genes?