Bioinformatics Homework 3

July 17, 2008

Due back, Tuesday, July 22, 2008

1. Please load the .mat file **`nequitansgenome.mat`** and load it in your workspace.
2. Save the m-files **`markovmatrix.m`**, **`montecarlomatrix.m`**, and **`gcdensity.m`** in your matlab working directory.
3. In your web browser, go to the website: http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=nuccore&id=38349555 the website describes the genome of the hyperthermophile *Nanoarchaeum equitans*
4. What is the size of the *N. equitans* genome? Apply command **`gcdensity`** to the entire genome to compute the GC-content. On the website choose **six** gene CDS regions, **six** tRNA regions, and **two** rRNA regions and note down their first and last positions. Next save those portions of the genome in different variables. For example, the command: **`gene_example = nequitansgenome(883:2691);`** will save the nucleotides from position 883 to 2691 in the variable **`gene_example`**. Avoid examples that have "joins" in them, but include some example with "complement" in them.
5. What does **`complement(2668..3189)`** (on the website) mean?
6. Next, apply the MATLAB command **`gcdensity`** to find the GC content of all 12 variables. What do you need to do (if anything) for the "complement" regions?
7. Use these numbers to compute and average GC content for the CDS regions and for the structural RNA regions.
8. How do your results compare with the results in the papers?
9. Next, read the attached text file: **`diary-MATLAB-lecture3.txt`** This is the record of what we did in class today. Try to understand the probabilities we generated.
10. Next, read the m-files **`markovmartix.m`** and **`montecarlomatrix.m`**
11. Please answer the questions in the commented out (in green) part of those m-files
12. Next, read the text file **`simulatingdna.txt`**
13. Familiarize yourself with the MATLAB command **`rand`**
14. Now generate the Monte Carlo matrix for the entire genome of *N. equitans.*

15. Use **`rand`** and the Monte Carlo matrix to simulate a sequence (i.e. generate an artificial sequence) for *N. equitans* which has between 20 and 30 nucleotides.
16. Compute the GC content of the simulated sequence. How does it compare with the original sequence?