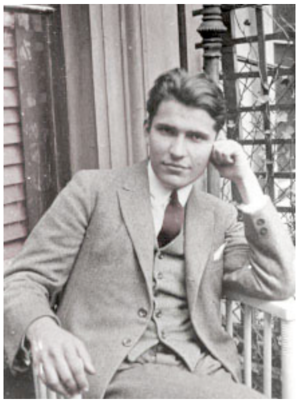
DNA

Chargaff’s Numbers



Erwin Chargaff, ca.1930

A biochemist at Columbia University, Erwin Chargaff discovered the base-pairing regularities or “complementarity relationships” of nucleic acids that provided one of the key steps in developing a structural model for DNA. Chargaff’s discovery (1950) that in large sequences such as entire genomes, the “number of As always equals the number of Ts and that the number of Cs always equals the number of Gs was a pivotal clue for Watson and Crick’s discovery of the double helix structure of DNA (1953). *You can read more about Chargaff here:*

`http://en.wikipedia.org/wiki/Erwin_Chargaff`

Write a Perl program to read a file filled with DNA and print a neat summary of Chargaff’s numbers. For this program, “Chargaff’s numbers” are defined as:

(0) The **number** of Adenine (**A**) nucleotides in the file of DNA.
(1) The **number** of Cytosine (**C**) nucleotides in the file of DNA.
(2) The **number** of Guanine (**G**) nucleotides in the file of DNA.
(3) The **number** of Thymine (**T**) nucleotides in the file of DNA.
(4) The **total number** of nucleotides in the file.

(5) The **total number** of nucleotides in the file that are ***not*** A, C, G, or T.

(6) The **proportion** of Adenine (**A**) nucleotides in the file of DNA.
(7) The **proportion** of Cytosine (**C**) nucleotides in the file of DNA.
(8) The **proportion** of Guanine (**G**) nucleotides in the file of DNA.
(9) The **proportion** of Thymine (**T**) nucleotides in the file of DNA.

(10) The **percentage** of Adenine (**A**) nucleotides in the file of DNA.
(11) The **percentage** of Cytosine (**C**) nucleotides in the file of DNA.
(12) The **percentage** of Guanine (**G**) nucleotides in the file of DNA.
(13) The **percentage** of Thymine (**T**) nucleotides in the file of DNA.

(14) Print the **likelihood** of finding the specific 8-mer motif: **`ACGTACGT`**

Note: We are assuming that each nucleotide is independent, thus you can use the multiplication rule to determine the likelihood of a motif. When calculating the probability of finding this 8-mer, you *must* use the log method outlined on pages (bottom of 85 and top of 86).

Superior Effort:

(15) Finally, print the length in centimeters (cm) of a strand of DNA at least 100,000 bases long). To start, note that there are 4 million bases (4 x 10^6bp) in a millimeter of DNA. Your Perl program must compute each of the conversions, e.g., from millimeters to cm.

NOTE: Print the proportions and percents with `printf`; use a format of `%5.2f` Experiment with values to use in the `printf` for log-likelihood. See the code for hints when you want to print in scientific notation for the length of DNA calculation, e.g., `%5.1e`. **Hint: Read up on `printf` (pages 86-89).**

How to get started:

(0) Fetch the Chargaff starter folder (a "Starter Kit") by visiting our moodle site (or alternatively from off-campus: http://cs.wheatoncollege.edu/mleblanc/dna/)

(1) Download our "starter kit" to help you begin this program. Click on the "**Chargaff's Numbers (zip)**"link.

(2) Save the .zip file on your Desktop. Un-zip the file. You should have a folder that contains the files:

a) README_spec.doc -- this file
b) Chargaff.pl -- a starter Perl program (we've done some of work for you!)
c) test1.fna -- a very small FASTA format file to test your math at first
d) test2.fna -- another small FASTA format file to test your math
e) Nanoarchaeum_equitans_Kin4_M.fna
-- a FASTA format file containing the entire genome of this microbe

(3) This program is **Due on Thursday, September 24, 2009**.

(4) Like all of your programs, you should take extra care to:

(a) write coherent documentation at the top of your program (in **pod** format)
(b) use good indentation and liberal use of blank lines to separate sections of your program; use our starting program for help …
(c) use lots of # comments to document sections of your code, including lines to separate sections, e.g., #--------------------------
(d) use *very* good variable names
(e) test your program on multiple test files (NOTE: you can make up your own files, e.g., `test3.fna`, so you can test a certain A,C,G,T distribution). Certainly you should confirm that your results are correct using the given files "`test1/2.fna`", as well as test the given file "Nanoarchaeum".

(f) It is "ok" to ask a colleague what they are getting for answers; however, it is *not* ok to ask them how they used Perl to get those answers!