

I519 Lab 1 Fall 2008
School of Informatics
Indiana University

Prof. M.M. Dalkilic
Associate Instructor Chao Ji

September 4, 2008

Introduction

- Get your lab accounts
- I will be candid and tell you the size of the class is barely manageable. Therefore, Chao will be expecting easily graded projects. I will be most almost intransigent in overturning any of Chao decisions on your homework. If you have questions about presentation, ask him first—but make it simple for him to grade. I have given him explicit instructions not to spend time trying to get programs to run, looking for code, trying to navigate directories. All of this will be turned in on course. Chao will create drop boxes that allow depositing of materials during a particular interval. **This homework is due at the beginning of laboratory.**
- You are free to discuss abstractly the problems—no other kind of assistance is acceptable

R

This section has to do with R code. There are tasks, scripts, and questions that require written answers. The scripts in this homework are trivial—but in the future, we will be exacting in how we ask you to name and place them. *All* the written material should be compiled into an easily readable document.

The file `basecodes.txt` contains single letter codes for nucleic acids. The file I've created is shown below:

```
Code Meaning Complement Notes
A A T Adenine
C C G Cytosine
G G C Guanine
T T A Thymine
M A|C ? aMino
R A|G ? puRine
W A|T ? Weak interactions 2H bonds
S C|G ? Strong interactions 3H bonds
Y C|T ? pYrimidine
K G|T ? Keto
V A|C|G ? none
H A|C|T ? none
D A|G|T ? none
B C|G|T ? none
N A|C|T|G ? any base
```

The first column is the code. The second column is the nucleic acid that is symbolized. Some symbols, like M, denote multiple nucleic acids. The third column is only partially completed and is supposed to contain the complement. I've completed the first four rows.

1. Edit the file replacing the question mark with its complement. The file should be named `encoded.txt`.
2. Write an R script that reads in the file. Write a **simple** script to replace the question mark with to the variable `encode`. Keep the header.
3. Execute the `attach` and names commands on `encodes`.
4. Do a summary on `encodes`
5. Give a command that selects all the columns of the first four rows.
6. On page 11 of the R book, the author doesn't explicitly state the trees are the same species—this would be unlikely. His method of randomization (using a lottery) is not correct given the trees are different. How should this problem be properly randomized? Give two other elements that might have to be taken into account, but many not necessarily be able to randomly modelled. What would you suggest?
7. Let V be the set of vultures and L a binary predicate $L(x, y)$ when x is located in y . Here are the sentences he formally states: the text page 3:

- $\exists x \in V \wedge L(x, \text{local park})$
- $\forall x \in V \Rightarrow \neg L(x, \text{local park})$

The $\exists x$ means “there's at least one x ”, while $\forall x$ means “all x exist.” When he discusses that the first one is not refutable because the observer did not see a vulture, how is this connected the formal sentence I wrote? When he said the second sentence was refutable, how is this connected to the second sentence I wrote? This is a concept stolen (borrowed) from logic. A sentence is *satisfiable* if it can be made true by a mapping of variables to truth values (environment). If one possesses a procedure that can *only* determine if a sentence is satisfiable, then how can one determine if the sentence is *always* true? How does this relate to the author's statement “absence of evidence is not evidence of absence?”

Language of your choice sans Java

You may use “scripting” languages as well, except javascript. The program should be in zipped file call I5191. When unzipped, the top level *must* include a README file. The programs should be in their own directories I5191D1 and I5191D2. In each directory the program should be invoked by simply typing I519P1 and I519P2. If there are *insignificant* elements to make the program run, this should be clearly state in both the README and in both the directors. The analytics should be submitted together with the R script typed as Assignment 1 I519 Fall 08.

Genome is composed of sequences of complexes composed of a sequence of phosphate bound to a sugar bound to a nitrogenous base. These groups are called nucleotides—for ease of *representation* we denoted each of the four nucleotides by A, T, C, and G named adenosine, thymosine, cytosine, and guanosine. This group forms an alphabet of sorts that we'll denote by Σ_D . I will pose a problem to you that has an analytic solution, but can be recovered through replication and randomization (where have you seen that before)?

- Suppose the probability of picking any $x \in \Sigma_D$ is $\frac{1}{4}$ —they all equally likely to be picked. Now, how long do expect the stretch of DNA be, given your choice from Σ_D —given you continually pick from Σ_D until all nucleotides are seen.
- Suppose $P = \langle 2.2.3.3 \rangle$ for Σ_D . Again, how long do expect the stretch of DNA to be?

- Consider two independent events x, y associated with probabilities $p, 1-p$. You'd like to find on average (the expected value) of how many times you must pick from these two to get, say, x . The r.v. pdf would be $P[X = k] = (1-p)^{k-1}p$ where X is the trials until success.

$$E[X] = \sum_{k=0}^{\infty} k(1-p)^{k-1}p \quad (1)$$

$$\text{let } q = 1-p \quad (2)$$

$$= p \sum_{k=0}^{\infty} \frac{d}{dq}(d^k) \quad (3)$$

$$= p \frac{d}{dq}(\sum_0^{\infty} q^k) \quad (4)$$

$$= \frac{p}{(1-q)^2} \quad (5)$$

$$= \frac{1}{p} \quad (6)$$

I've left a little bit of algebra out—(THIS IS A HINT BTW)

1 Mount

There does not exist an ANSI standard for format of biological data. Do you think this has been helpful or hurtful in the progress in the longrun—how about now? The relational model, based on predicate calculus, cannot model complex relationships. Yet, much work has been done (40 yrs. worth) to optimize certain functions. The current formats are called **unstructured** because no agreeable format exists—groups have been pushing XML. Simply look around your work and home environment—how many XML databases are present—take some time looking at the specification of XML. Does this help the average scientist?