BCBio 444: Bioinformatics Analysis

Karin S. Dorman

Department of Statistics
Department of Genetics, Development & Cell Biology
Program in Bioinformatics & Computational Biology
Iowa State University

August 29, 2017

Introductions

- Introductions: highest course in biology, math, statistics, computer science, major(s).
- syllabus

Introduction

Bioinformatics is like finding a needle in a haystack where every piece of hay looks like a needle. And the needle is cancer.

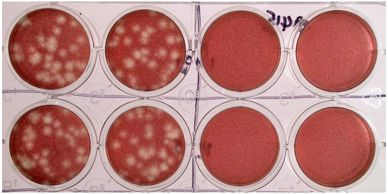
– darkhelmet41290 [reddit]

A Motivating Example

Much of lower-level bioinformatics and this course is about learning how to identify and use computational tools to answer standard questions, but it will not take long before you encounter data that looks different from standard types of data or biological questions unlike those that have known procedures to answer.

We will start with a motivating example that demonstrates how you can use the general-purpose tools of bioinformatics to put together your own methodology and answer for a example dataset.

Discovery: antiviral function



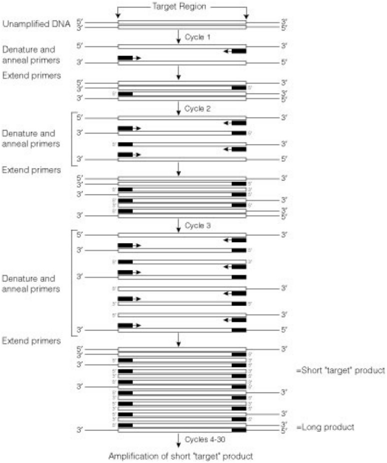
Suppose you have discovered a novel biological process that attacks and destroys some viruses. You have been able to grow a susceptible virus in two types of cells with respective to this novel process, one permissive and the other not. You *hypothesize* that the nonpermissive cells actively mutate the virus genome, rendering them nonfunctional. You suspect the mysterious function is specific, targeting and mutating one type of nucleotide base $\mathrm{N}_t \in \Omega = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$ in the virus to another, wrong nucleotide base $\mathrm{N}_m \in \Omega$ with $\mathrm{N}_m \neq \mathrm{N}_t$.

Your Goal

Your first goal is to determine what are N_t and N_m, in other words, what is the mutation that this novel biological process is inducing?

An Experiment

The susceptible virus in your study *integrates* its genome into host cells. You use this fact to design an experiment.



- Grow 1 plate each of (non)permissive cells.
- Add 10 moi of virus to each plate and incubate.
- Collect the cells and isolate the DNA.
- Amplify the virus genome using PCR (see right).
- Fragment and sequence.

Data in Fasta Format

Virus from nonpermissive cells:

```
>n.1
AAGGACCCTGTGCATAAAGTATATTATGACCCATCAAAAGACTTAATAGCAGAGATACA
GAAGCAAAGACAAGACCAATAGACATATCAGATTTATCAAGAACCATTTAAAAATCTGA
AAACAAGGAAATATGCAAGAAAAAAGTCTGCTCACAC...
>n.2
AAAAATAACATGGTAGAGCAGATGCATACAGATATAGTCAGTCTATAAGAACAAAGCCT
AAAGCCATGTGTAAAGTTAACCCCTCTCTGCGTTACTTTACATTGTAACAATGTCACAG
GGAACATCACAGAGAGAATCAGAGAAGAAAAAAAAAA...
...
```

Virus from permissive cells:

```
>p.1
GACCCTTATCCCGAACCCAAGGGAACCCGACAGGCCAGGAAGAATCGAAGAAGAAGGTG
GAGAGCAAGACAAAGAGAGATCCGTGCGATTAGTGAGCGGATTCTTAGCACTTGCCTGG
GACGACCTACGGAGCCTGTGCCTCTTCAGCTACCACC...
>p.2
ACCTAGTGTGAACAATGAGACACCAGGAATTAGATATCAGTACAATGTGCTTCCACAAG
GATGGAAAGGATCACCAGCAATATTCCAAAGTAGCATGACAAAAATCTTAGAACCTTTC
AGAAAGCAAAATCCAGAAATAACTATCTATCAATACA...
...
```

Questions

It could be that one or a few specific N_t nucleotides in the genome that are critical for virus function are being targeted for mutation. It is also possible that random N_t nucleotides are mutated until eventually virus function is disrupted.

- How can we use the data to distinguish these two hypotheses?
- How can we detect which, if any, nucleotide is targeted and how it is mutated?
- Why might the experiment not work?

Relevant data

I argue that the relevant data in the Fasta files for answering the previous slide's questions are the nucleotide counts.

Cell type	A	C	G	T
Nonpermissive cells	$n_{n\mathrm{A}}$	$n_{n\mathrm{C}}$	$n_{n\mathrm{G}}$	$n_{n\mathrm{T}}$
Permissive cells	$n_{p\mathrm{A}}$	$n_{p\mathrm{C}}$	$n_{p\mathrm{G}}$	$n_{p\mathrm{T}}$

- Will the counts $\boldsymbol{n}_{n\cdot} = (n_{n\mathrm{A}}, n_{n\mathrm{C}}, n_{n\mathrm{G}}, n_{n\mathrm{T}})$ and $\boldsymbol{n}_{p\cdot}$ be identical?
- Why will they vary?
- How can we determine which nucleotide, if any, is mutated and how it is mutated?
- When can we conclude that, yes, for example, the novel mechanism *does* mutate A to C?

Detecting signal

This is an example of detecting a signal. We can use two big ideas from statistics to help:

- Statistical hypothesis testing (*e.g.* z-test, t-test).
- Resampling to quantify variability.

First, let's review (or learn) basic probability & statistics...

Foundations of Probability

- **random experiment**: a repeatable process whose outcome cannot be predicted beforehand, but will be observed after the experiment is complete
- **outcome**: one possible output of a random experiment
- **sample space**: the set of possible outcomes of a random experiment
- **event**: a set of outcomes
- **probability**: given a random experiment, a measure of how likely an event is, in the range $[0, 1]$

In order to determine the probability of events, one must hypothesize a model. This is where the bioinformatics team needs to work together when developing new bioinformatics methods. Quantitative scientists propose models; biologists tear them down. Teamwork!

Examples – Probability

- *Experiment*: toss a coin; *outcome*: head (H), *sample space*: $\Omega = \{H, T\}$; *event*: $E = \{H\}$, *probability*: $P(\{H\}) = P(\{T\}) = 0.5$.
- *Experiment*: sequence a 15 bp fragment of mRNA; *outcome*: ACCGAGGTCTCTAAA; *sample space*:

$$\Omega = \underline{\hspace{6cm}};$$

event: $E = \{\text{YYYYYYYYYYYYYYY}\}$; *probability*:

$$P(\{\text{ACCGAGGTCTCTAAA}\}) = \underline{\hspace{6cm}}$$

$$P(\{\text{YYYYYYYYYYYYYYY}\}) = \underline{\hspace{6cm}}$$

Example – Nucleotide Counts/1

- **random experiment**: (1) collect data according to biological experiment, (2) count the nucleotides in the fasta files and store as count vectors $\boldsymbol{n}_p$ and $\boldsymbol{n}_n$.
- **outcome**: We will see an example on the next slides...
- **sample space**:

$$\Omega = \left\{ (\boldsymbol{n}_p, \boldsymbol{n}_n) : n_{hi} \in \{0, \mathbb{Z}^+\}, h \in \{p, n\}, i \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\} \right\}.$$

- **event**:

$$E = \left\{ (\boldsymbol{n}_p, \boldsymbol{n}_n) \in \Omega : n_{p\mathrm{A}} > n_{n\mathrm{A}} \right\}.$$

- **probability:**

What can we do for the probability?

Example – Nucleotide Counts/2

This is a silly model for the count data $\boldsymbol{n}_n$ and $\boldsymbol{n}_p$ as demonstrated in R code.

```
# generate two samples of 100 nucleotides by
# flipping a fair, 4-sided coin:
n.n.data <- rmultinom(n = 100, size = 1,
  prob = rep(0.25, 4))
n.p.data <- rmultinom(n = 100, size = 1,
  prob = rep(0.25, 4))
n.n <- rowSums(n.n.data)
n.p <- rowSums(n.p.data)
n.n
```

```
## [1] 29 25 19 27
```

```
n.p
```

```
## [1] 21 21 26 32
```

Example – Nucleotide Counts/3

Cell type	A	C	G	T
Nonpermissive cells	29	25	19	27
Permissive cells	21	21	26	32

- If you had to guess the target nucleotide N_t and mutated nucleotide N_m were from this data, what would you choose?
- In this case, the two rows of data are generated under identical conditions: there is no actual difference!
- So, how can we be sure a difference we see is real?

Review – Probability

- **Vocabulary.** random experiment, sample space, outcome, probability
- A model is necessary to compute probabilities.
- Most scientific experiments are random, at least there is measurement error: Outcomes contain noise.
- Random experiments are designed to answer questions or test hypotheses.
- Some noise can look like a meaningful pattern: *e.g.* it looked like G$\rightarrow$A mutation in the simulated count data.
- The triumvirate of bioinformatics:
    - Biological knowledge/cleverness will determine the right experiment & visible pattern to confirm the hypothesis;
    - Computers will help us extract the pattern;
    - Statistics (and computers) will help us *distinguish the pattern (signal) from the noise*.

Statistical Inference

Statistical inference is the process of deducing facts about the **population** based on a **simple random sample** (constituting **data**) from the population. There are two types of statistical inference:

- estimation
- hypothesis testing

Examples – Population/Sample

- Population: ISU students; Sample: this class (Is it a *random* sample? If I want to deduce the mathematical skills of ISU students, can I use you guys as the sample?)
- Population: mRNA in a cancer cell; Sample: a random set of mRNA from a random set of cancer cells from a random tumor
- What is the sample in the scientific experiment our biologist undertook? What is the population?

Statistic

A **statistic** is any function of a sample that requires nothing more than the sample to compute.

Example – Statistic

Which of these are statistics?

- Population: ISU students; Sample: this class. The average height of students in this room.
- Population: ISU students; Sample: this class. The number of inches your height differs from the mean ISU student height.
- Population: provirus sequences in permissive cells; Sample: our fasta file.

$$n_{pC}$$

- Population: provirus sequences in permissive *and* nonpermissive cells; Sample: our fasta file*s*.

$$n_{pC} - n_{nC}$$

Review – Statistical inference

Statistical inference is the process of deducing facts about the **population** based on a **simple random sample** (constituting **data**) from the population.

- estimation
- hypothesis testing

To perform statistical inference, we compute **statistics** on samples. Some statistics are useful for estimation: they are called **estimators**. Other statistics are useful for hypothesis testing: they are called **test statistics**.

Statistical hypothesis testing I

- Identify one, two or more hypotheses. A hypothesis is a model for reality. In statistical hypothesis testing it must be a really, really precise model. In fact, it must be capable of generating the random variables in your data sample.
- In *frequentist hypothesis testing*, we focus on one particular hypothesis called the **null hypothesis**, denoted by H_0. If we have many hypotheses, we would test each in turn.
- Then, we choose a **test statistic** that is *sensitive to the truth of* H_0, that *signals the validity of* H_0.

Example – z-test test statistic

The null hypothesis tested by the z-test is

$$H_0 : x_1, \ldots, x_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2), \text{ where } \mu = \mu_0.$$

The z-test uses the z test statistic, namely

$$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}},$$

where

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

is the **sample mean.**

Why does this **test statistic** signal the validity of H_0?

Example – Detecting C $\rightarrow$ A

Cell type	A	C	G	T
Nonpermissive cells	29	25	19	27
Permissive cells	21	21	26	32

$$T_1 = n_{pC} - n_{nC} + n_{nA} - n_{pA}$$

$$T_2 = \left(n_{pC} - n_{nC}\right)^2 + \left(n_{nA} - n_{pA}\right)^2$$

Hypothesis	T_1	T_2
G $\rightarrow$ A	15	113
T $\rightarrow$ G	13	89
G $\rightarrow$ C	11	65
$\vdots$	$\vdots$	$\vdots$
A $\rightarrow$ G	-15	113
$\vdots$	$\vdots$	$\vdots$

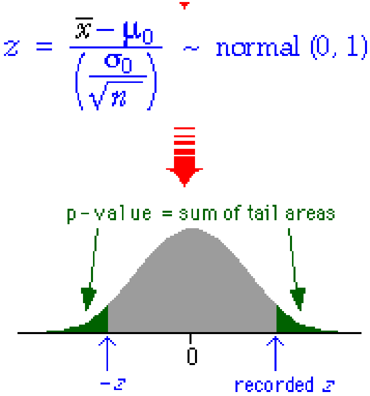
Statistical hypothesis testing II

- If we can derive the probability distribution of the *test statistic* under H_0, then we have all we need to draw an inference about the truthfullness of H_0. These are the methods you learn in basic statistics classes.
- In particular, we can compute the p-value, which is the probability of obtaining a test statistic T *as extreme or more extreme* than the observed test statistic t_0 when H_0 is true:

$$P(T \geq t_0 \mid H_0).$$

The above is an example of a **conditional probability** (click to remind yourself what this is).

z test p-value



This solution uses the **probability distribution** of the z **test statistic**, which you may have seen derived in a statistics class using mathematical theory. What if you don't know the probability distribution of your test statistic T?

Resampling

- The outcome of our random experiment has been distilled down to the test statistic T, which captures the signal in the data.
- The reason we need the probability distribution of the test statistic T is to understand how it varies because of the randomness of the experiment. We need to understand this variation to know if the signal exceeds the usual variation.
- If we could repeat the random experiment many times, then we could observe this variation directly.
- Remember your null model H_0 mimics the random experiment and is supposed to be able to create samples of data.
- Let's use it and repeatedly *simulate* the random experiment *in silico*.

Resampling – Monte Carlo Simulation

(General) Algorithm: Mimic the randomness/uncertainty of the random experiment using a computer.

- **Input**: the observed data $\boldsymbol{x} = (x_1, \ldots, x_n)$, a large number $B \in \mathbb{Z}$ for the number of times to repeat, and a model (constructed and confirmed with a biologist).
- Loop B times: at iteration i
  - Generate a **simulated** data set $\boldsymbol{x}^{(i)} = (x_1^{(i)}, \ldots, x_n^{(i)})$
  - Compute and store the test statistic: $T^{(i)} = T(\boldsymbol{x}^{(i)})$.
- Compute the *observed test statistic*: $t = T(\boldsymbol{x})$.
- **Output:** Compute the p-value as the proportion of simulation samples where $T^{(i)}$ is as or more extreme (shows more signal) than the observed test statistic t.

Example – Identifying N_t

Here is a demonstation in R to test the data simulated earlier with R. In particular, we are using proposed test statistic T_1.

```
t.bs <- NULL
for (i in 1:100) {
   # simulate data under model H0
   n.n.bs <- rowSums(rmultinom(n = 100,
            size = 1, prob = rep(0.25, 4)))
   n.p.bs <- rowSums(rmultinom(n = 100,
            size = 1, prob = rep(0.25, 4)))
   # compute simulated test statistic
   t.bs[i] <- max(n.p.bs - n.n.bs) +
            max(n.n.bs - n.p.bs)
}
# compute observed test statistic
t <- max(n.p - n.n) + max(n.n - n.p)
cat("The proportion as or more extreme is:",
   mean(t.bs > t))
```

```
## The proportion as or more extreme is: 0.34
```

Your task

- Count the nucleotides in the fasta files.
- Bootstrap the random experiment.
- Estimate the probability of simulation data as extreme or more extreme than the fasta files.

Advanced question: You will know from this that the mechanism in fact mutates nucleotides everywhere rather than at particularly sensitive locations in the genome. The next natural question is whether the novel mechanism specifically targets a preferred *motif*? A motif is a short nucleotide pattern. Answering this question is also within your grasp.

Important Concepts

- variation and noise in samples. Why do data in samples vary?
- random experiment, sample space, outcome, probability. Can you identify the these components of a "random experiment" in the elements of a scientific experiment?
- population, sample, statistical inference. Conclusions draw from statistical inference apply to the population or the sample?
- random variable. What are the random variable function's range and domain?
- test statistic. What is one thing a test statistic *must* accomplish?
- resampling. How can resampling help you assess whether a detected signal is significant?