CSCI2950-C: Topics in Computational Biology
Lecture 2, Part 2
Finding the Mutations that Cause Cancer

Ben Raphael
September 19, 2011
http://cs.brown.edu/courses/csci2950-c/

Course Topics

1. “Resequencing” algorithms
   (also RNA-Seq, ChIP-Seq, *-Seq)
2. De novo assembly algorithms

---

Genome
Millions -billions nucleotides

Next-generation DNA sequencing

10-100’s million reads
Reads: 30-1000 nucleotides
Origins of Cancer

Tumor growth driven by “...a particular, incorrect chromosome combination which is the cause of the abnormal growth characteristics passed on to daughter cells...” (Boveri 1914)

DNA Replication and Mutation

- Single nucleotide variant
- Copy number variants
- Structural variants

DNA Replication → Mitosis → Translocation

Leukemia
Breast Cancer
Genome Sequencing and Comparison

Comparative genomics
Differences between species?

Personal genomics
Genetic basis for traits of individuals?

Cancer genomics
Which somatic mutations lead to cancer?

Somatic Mutations and Cancer

Clonal Theory (Nowell 1976)

Passenger mutations

Founder cell

Driver mutation

“typical tumor”:
~10 driver mutations
100’s – 1000’s of passenger mutations

Sequence genome
Challenges in Cancer Genomics

1. **Measurement** of all somatic mutations.

   - **Human genome:** ~3 billion letters
   - **Next-generation DNA sequencing**
     - Reads of 30-1000 nucleotides
   - **Algorithms**
   - **Genome and Mutations**

   10-100’s million *reads*
   Read: 30-1000 letters

2. **Distinguish** functional *(driver)* mutations from background *(passenger)* mutations.
Functional Mutations

Next-generation DNA sequencing

Recurrent mutations/mutated genes $\rightarrow$ functional

Recurrent Mutations

Cancer Gene Identification Problem

Given: Mutation matrix $A$

$$A_{ij} = \begin{cases} 
1, & \text{Patient } i \text{ has mutation in gene } j \\
0, & \text{otherwise.}
\end{cases}$$

Find: Genes mutated in more patients than expected by chance.
Cancer gene hunting

316 Ovarian Cancer Patients. *(Nature, June 30, 2011)*

A Game of Chance

Outcomes = \{H, T\}
Pr[H] = q
Pr[T] = 1 - q

Coin flipping game
H: I win.
T: You win.
q = probability I win = \(\frac{1}{2}\), fair coin.
Let’s Play!

Outcomes = \{H, T\}

\begin{align*}
\Pr[ H ] &= q = \frac{1}{2} \\
\Pr[ T ] &= 1 - q = \frac{1}{2} \\
\end{align*}

Flip coin 100 times. H, T, T, H, T, H, T, H, ..., H, T.

80 H.

I win 80/100 times.

Am I cheating???

Cheating???

Outcomes = \{H, T\}

\begin{align*}
\Pr[ H ] &= \frac{1}{2} \\
\Pr[ T ] &= \frac{1}{2} \\
\end{align*}

What is \( \Pr[ 80 \text{ H on 100 coin flips}] \)?

Outcome = sequences of HTHTTHHH...TTHT

Outcome (H/T) of each flip independent of other flips.

\( \Pr[\text{sequence S }] = \frac{1}{2}^{\# \text{ H in } S} \frac{1}{2}^{\# \text{ T in } S} \).

\[
\begin{align*}
\Pr[ 80 \text{ H }] &= \binom{100}{80} \left( \frac{1}{2} \right)^{80} \left( \frac{1}{2} \right)^{20} \\
&= \frac{100!}{80!20!} \\
&= \text{is number of ways to choose 80 positions for H in sequence}
\end{align*}
\]
Cheating???

Outcomes = \{H, T\}
Pr[ H ] = q = \frac{1}{2}
Pr[ T ] = q = \frac{1}{2}

X = number of \(H\) on 100 flips of a fair coin.
X takes values in \([0, 100]\), according to a probability function.

\[
Pr[X = k] = \binom{100}{k} q^k (1 - q)^{100-k}
\]

Binomial Distribution

\[
Pr \left[ X = 80 \right] \approx 4 \times 10^{-10}
\]

\[
Pr \left[ X = 50 \right] \approx 0.08
\]

Should you Accuse???

Intuition: If \(r\) is too large, then I am probably cheating.

Accuse of cheating if \(Pr[X \geq r] \leq \alpha\).

\(\alpha\) is “probability of false accusation”.
\(\alpha = 0.05? \ 0.01?\)

\[
Pr \left[ X \geq 80 \right] \approx 5 \times 10^{-10}
\]

\[
Pr \left[ X \geq 65 \right] \approx 0.002
\]
Hypothesis Testing

Observe $r$ H’s.

$X$ = number of $H$ on 100 flips of fair coin.

Null hypothesis $H_0 : \Pr[H] = q = \frac{1}{2}$.

Alternative hypothesis: $q > \frac{1}{2}$

Let $p = \Pr[X \geq r]$ under $H_0$.

If $p \leq \alpha$, then $H_0$ unlikely. **Reject $H_0$.**

$p$ is the **$p$-value** of the test.

Recurrent Mutations

Cancer Gene Identification Problem

**Given:** Mutation matrix $A$

$$A_{ij} = \begin{cases} 
1, & \text{if patient } i \text{ has mutation in gene } j \\
0, & \text{otherwise.}
\end{cases}$$

**Find:** Genes mutated in more patients *than expected by chance*. 
Recurrent Mutations

Single-gene Test

Given: $r$ patients have mutation in gene $g$.

Passenger mutation probability $q$

Compute: Is $p = \Pr(X \geq r) \leq \alpha$?

$X$ is number of “H” on $m$ flips where $\Pr[H] = q$.

Cancer gene hunting

Cancer gene hunting


Surprise Happens

*p*-values have uniform probability
Bonferroni

Chose fixed threshold $\alpha$ to reject $q = \frac{1}{2}$ (null hypothesis)

**Error**: reject $q = \frac{1}{2}$ when coin is fair

$$\Pr[\text{error}] = \Pr[\text{reject } q = \frac{1}{2} \text{ when coin is fair}] = \alpha.$$

More Coins = More Errors

**Multiple Hypothesis Testing Problem**

$$\Pr[\text{no errors}] = (1 - \alpha)^{50}$$
$$\Pr[\geq 1 \text{ error}] = 1 - (1 - \alpha)^{50} \approx 0.4 \text{ when } \alpha = 0.01$$
$$\Pr[\geq 1 \text{ error}] = \Pr[\sum_i \text{error for coin } i] \leq \sum_i \Pr[\text{error for coin } i] = 50\alpha.$$

**Bonferroni correction**: To have $\Pr[\geq 1 \text{ error}] \leq \alpha$

Set $\alpha^{\text{(single coin)}} = \alpha/50.$
Cancer Gene Hunting

Ovarian Cancer

~20000 Genes
316 Patients

Mutation Matrix

Recurrent Mutations

91 glioblastoma multiforme samples [TCGA (2008)]

What genes are altered/mutated?

genes

patients

Mutation matrix

<table>
<thead>
<tr>
<th>Genes</th>
<th>Patients</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Mutation matrix

- mut = mutated
- not mut = not mutated

<table>
<thead>
<tr>
<th>Genes</th>
<th>Sample mutations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Graph showing distribution of mutations across samples.
Summary

Somatic mutations: Find recurrently mutated genes.

- Binomial Distribution
- Hypothesis Testing
- Multiple Hypothesis Correction