Statistical Modelling and Methods in Bioinformatics

Loyola University Chicago
Department of Microbiology and Immunology
Timothy E. O’Brien
10th March 2017
Why Replicate Our Studies?

Figure 9.1 Different levels of replication in a microarray experiment.
“In this experiment, the independent cell cultures are biological replicates. The replicates at RNA samples and arrays are technical replicates which are similar to the repeated measurements. They are less useful for identifying significantly expressed genes between the two treatments. However, technical replicates are essential in experiments designed for evaluating the technology and in identifying the sources of variation. The variability between the duplicated arrays estimates the variability of the procedure after RNA extraction and the variability between the duplicated RNA samples estimates the variability from both RNA extraction and the array hybridization.” (Lee, pp.203-4)
Power: Statistical Modelling May Help
### The Marginal Table:

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
<th>% diseased</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>53</td>
<td>430</td>
<td>483</td>
<td>11.0%</td>
</tr>
<tr>
<td>F</td>
<td>15</td>
<td>176</td>
<td>191</td>
<td>7.9%</td>
</tr>
</tbody>
</table>

### The Stratified Tables for the Same Data:

<table>
<thead>
<tr>
<th>Younger patients</th>
<th>Disease status</th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
<th>% diseased</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>53</td>
<td>414</td>
<td>467</td>
<td>11.3%</td>
</tr>
<tr>
<td></td>
<td>F</td>
<td>11</td>
<td>37</td>
<td>48</td>
<td>22.9%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Older patients</th>
<th>Disease status</th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
<th>% diseased</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>0</td>
<td>16</td>
<td>16</td>
<td>0.0%</td>
</tr>
<tr>
<td></td>
<td>F</td>
<td>4</td>
<td>139</td>
<td>143</td>
<td>2.8%</td>
</tr>
</tbody>
</table>
DNA/Protein Sequence Alignment Methods

• **Traditional Methods:**
  • Global: Needleman/Wunsch method
  • Local: Smith/Waterman algorithm
  • End-Space Free

• **Modern Methods using Hidden Markov Models**
Needleman/Wunsch Global Algorithm:

|   | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
| A | 5 | -2 | -1 | -2 | -1 | -1 | 0 | -2 | -1 | -2 | -1 | -1 | -3 | -1 | 1 | 0 | -3 | -3 | -2 | 0 |
| R | -2 | 7 | -1 | -2 | -4 | 1 | 0 | -3 | 0 | -4 | -3 | 3 | -2 | -3 | -3 | -1 | -1 | -3 | -3 | -2 | -3 |
| N | -1 | -1 | 7 | 2 | -2 | 0 | 0 | 0 | 1 | -3 | -4 | 0 | -2 | -4 | -2 | 1 | 0 | -4 | -2 | -3 | -3 |
| D | -2 | -2 | 2 | 8 | -4 | 0 | 2 | -1 | -1 | -4 | -4 | -1 | -1 | -4 | -5 | -1 | 0 | -1 | -5 | -3 | -4 |
| C | -1 | -4 | -2 | -4 | 13 | -3 | -3 | -3 | -3 | -2 | -2 | -3 | -2 | -4 | -1 | -1 | -1 | -5 | -3 | -1 | -3 |
| Q | -1 | 1 | 0 | 0 | -3 | 7 | 2 | -2 | 1 | -3 | -2 | 2 | 0 | -4 | -1 | 0 | -1 | -1 | -1 | -3 | -3 |
| E | -1 | 0 | 0 | 2 | -3 | 2 | 6 | 3 | 0 | -4 | -3 | 1 | 2 | 3 | 2 | 0 | -1 | 1 | -1 | -3 | -2 | -3 |
| G | 0 | -3 | 0 | -1 | -3 | -2 | -3 | 8 | -2 | -4 | -4 | -2 | -3 | -4 | -2 | 1 | -2 | -3 | -1 | -1 | -3 | -2 | -3 |
| H | -2 | 0 | 0 | 0 | -3 | 2 | 6 | 3 | 0 | -4 | -3 | 1 | 2 | 1 | 2 | 0 | -1 | 1 | -2 | -1 | -1 | -3 | -2 |
| I | -1 | 4 | -3 | -4 | -2 | -3 | 4 | -4 | -4 | 5 | 2 | -3 | 2 | 0 | -3 | -3 | 1 | -3 | -1 | 1 | 4 | 2 | -3 |
| L | -2 | 3 | 4 | -4 | -2 | -3 | 4 | -3 | 2 | 5 | -3 | 3 | 1 | -4 | -3 | 1 | -2 | 1 | 1 | 0 | -1 | -2 | -3 |
| K | -1 | 3 | 0 | 0 | 3 | -2 | 1 | 2 | 0 | -3 | -3 | 6 | -2 | -4 | -1 | 0 | 0 | 1 | -3 | -2 | -3 | -2 | -3 |
| M | -1 | 2 | -2 | -4 | -2 | 0 | -2 | -3 | 1 | 2 | 3 | 2 | 7 | 0 | -3 | 2 | 1 | 1 | 0 | 1 | 0 | -3 | -2 |
| F | -3 | -3 | -4 | -5 | -2 | 4 | 3 | -4 | -1 | 0 | 1 | 4 | 0 | 8 | -4 | -3 | 2 | 1 | 4 | 1 | -4 | -3 | -3 |
| P | -1 | 3 | -2 | 1 | 4 | -1 | 2 | -2 | -3 | 4 | -4 | 1 | 3 | -4 | 10 | -1 | 1 | -4 | -3 | -3 | 1 | 5 | 2 | -4 | -2 |
| S | 1 | 1 | 1 | 0 | -1 | 0 | -1 | 0 | 1 | -3 | -3 | 0 | -2 | -3 | 1 | 5 | 2 | -4 | -2 | -2 | 1 | 4 | 1 | -4 | -3 |
| T | 0 | 1 | 0 | -1 | 1 | 1 | -2 | -2 | -3 | 2 | -3 | -1 | -1 | -1 | -2 | 1 | 2 | 5 | 3 | 2 | 0 | 1 | -2 | -2 | -3 |
| W | -3 | -3 | -4 | -5 | -5 | -1 | 3 | 3 | -3 | -3 | -3 | -3 | -2 | -3 | -1 | 1 | -4 | -3 | 15 | 2 | 3 | 1 | 5 | 4 | 2 | -3 |
| Y | -2 | 1 | -2 | 3 | -3 | 2 | 1 | -2 | 1 | -2 | 2 | 0 | 3 | -2 | 2 | 1 | 8 | 1 | 2 | 0 | 1 | -2 | -2 | -3 |
| V | 0 | -3 | -3 | -4 | -1 | -3 | -3 | -4 | -4 | 4 | 1 | -3 | 1 | -3 | 2 | 0 | -3 | -3 | 1 | 5 | 2 | 3 | 5 | 2 | -3 |

**Figure 2.2** The BLOSUM50 substitution matrix. The log-odds values have been scaled and rounded to the nearest integer for purposes of computational efficiency. Entries on the main diagonal for identical residue pairs are highlighted in bold.
Figure 2.5 Above, the global dynamic programming matrix for our example sequences, with arrows indicating traceback pointers; values on the optimal alignment path are shown in bold. Below, a corresponding optimal alignment, which has total score 1.
Smith/Waterman Local Alignment Method:

Figure 2.6 Above, the local dynamic programming matrix for the example sequences. Below, the optimal local alignment, with score 28.
End-Space Free Alignment Algorithm:

**Figure 2.8** Above, the overlap dynamic programming matrix for the example sequences. Below, the optimal overlap alignment, with score 25.
Using Hidden Markov Models (HMM) for Sequence Alignment:

- Uses the EM (Expectation/Maximization) Algorithm
- http://www.nature.com/nbt/journal/v22/n10/pdf/nbt1004-1315.pdf
A casino has two dice:

- **Fair die:**
  \[ P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = 1/6 \]
- **Loaded die:**
  \[ P(1) = P(2) = P(3) = P(4) = P(5) = 1/10; \quad P(6) = 1/2 \]
- **Dealer switches between dice as:**
  - \( \text{Prob}(\text{Fair} \rightarrow \text{Loaded}) = 0.01 \)
  - \( \text{Prob}(\text{Loaded} \rightarrow \text{Fair}) = 0.2 \)
  - Transitions between dice obey a Markov process

**Game:**

1. You bet $1
2. You roll (always with a fair die)
3. Casino player rolls (maybe with fair die, maybe with loaded die)
4. Highest number wins $2
An HMM for the occasionally dishonest casino

**Diagram:**

- **Transition Probabilities:**
  - From **F** to **F**: 0.99
  - From **F** to **L**: 0.01
  - From **L** to **F**: 0.2
  - From **L** to **L**: 0.8

**Observation Probabilities:**

- **F**:
  - P(1|F) = 1/6
  - P(2|F) = 1/6
  - P(3|F) = 1/6
  - P(4|F) = 1/6
  - P(5|F) = 1/6
  - P(6|F) = 1/6

- **L**:
  - P(1|L) = 1/10
  - P(2|L) = 1/10
  - P(3|L) = 1/10
  - P(4|L) = 1/10
  - P(5|L) = 1/10
  - P(6|L) = 1/2
Question # 1 – Evaluation

GIVEN

A sequence of rolls by the casino player

124552646214614613613666166466163661636616361...

QUESTION

How likely is this sequence, given our model of how the casino works?

This is the EVALUATION problem in HMMs
Question # 2 – Decoding

GIVEN

A sequence of rolls by the casino player

1245526462146146136136661664661636616366163...

QUESTION

What portion of the sequence was generated with the fair die, and what portion with the loaded die?

This is the DECODING question in HMMs
Question # 3 – Learning

GIVEN
A sequence of rolls by the casino player

124552646214614613613666166466163661636616361651...

QUESTION

How “loaded” is the loaded die? How “fair” is the fair die? How often does the casino player change from fair to loaded, and back?

This is the LEARNING question in HMMs
What is a hidden Markov model?

Statistical models called hidden Markov models are a recurring theme in computational biology. What are hidden Markov models, and why are they so useful for so many different problems?

Often, biological sequence analysis is just a matter of putting the right label on each residue. In gene identification, we want to associate residues in a query sequence with homologous residues in a target database sequence. We can always write an ad hoc program for any given problem, but the same frustrating issues will always recur. One is that we want to incorporate heterogeneous sources of information. A genefinder, for instance, ought to combine splice-site consensus, codon bias, coding-length preferences and open reading frame analysis into one scoring system. How should these parameters be set? How should different kinds of information be weighted? A second issue is to interpret results probabilistically. Finding a best scoring answer is one thing, but what does the score mean, and how confident are we that the best scoring answer is correct? A third issue is extensibility. The moment we perfect our ad hoc genefinder, we wish we had also modeled translational initiation consensus, alternative splicing and a polyadenylation signal. Too often, piling more reality onto a fragile ad hoc program makes it collapse under its own weight.

Hidden Markov models (HMMs) are a formal foundation for making probabilistic models of linear sequence 'labeling' problems. They provide a conceptual toolkit for building complex models just by drawing an intuitive picture. They are at the heart of a diverse range of programs, including gene-finding, profile searches, multiple sequence alignment and regulatory site identification. HMMs are the Legoos of computational sequence analysis.

A toy HMM: 5' splice site recognition

As a simple example, imagine the following caricature of a 5' splice-site recognition problem. Assume we are given a DNA sequence that begins in an exon, contains one 5' splice site and ends in an intron. The problem is to identify where the switch from exon to intron occurred—where the 5' splice site (5'SS) is.

For us to guess intelligently, the sequences of exons, splice sites and introns must have different statistical properties. Let's imagine two simple differences: say that exons have a uniform base composition on average (25% each base), introns are AT-rich (say, 40% each for AT, 10% each for CG), and the 5'SS consensus nucleotide is almost always a G (say, 95% G and 5% A).

Starting from this information, we can draw an HMM (Fig. 1). The HMM invokes three states, one for each of the three labels we might assign to a nucleotide: E (exon), 5' (5'SS) and I (intron). Each state has its own emission probabilities (shown above the state), which model the base composition of exons, introns and the consensus G at the 5'SS. Each state also has transition probabilities (arrows), the probabilities of moving from this state to a new state. The transition probabilities describe the linear order in which we expect the states to occur: one or more Es, one 5', one or more Is.

So, what's hidden?

decoding:

Figure 1. A toy HMM for 5' splice site recognition. See text for explanation.

- Log P values for each state:
  - E: -41.22
  - 5': -43.00
  - I: -42.58

- Priors for each state:
  - E: 45%
  - 5': 28%
  - I: 27%
Thanks for your attention and questions