Statistical analysis for Biomarker identification

Raghunathan Srivatsan
IBAB, Bangalore
Biological systems and the processes driving them are intrinsically fluctuating (fuzzy) in nature.

Because of this, statistical analysis is employed to deal with the intrinsic variations in the data points from experiments.

This presentation will describe and demonstrate analysis of data for identifying biomarkers from three different types of experiments.

This way, we introduce few important aspects of bioinformatics data analysis.

( tip of the iceberg)
There are two main methods of statistics, both are used in biology.

1. **The frequentist method**:
   - Assumes prior distributions for the observed variables.
   - Relationships between data points and statistical parameters of the distributions are obtained by maximum likelihood estimates.
   
   [Used in data reduction, error analysis, hypothesis testing, curve fitting, estimation theory --- Binomial, Poisson, Gaussian distributions]

2. **Bayesian statistics**:
   - Assuming a prior probability for the event, use the data points to get an improved posterior probability. Computationally arrive at the distributions followed by the variables. [Bayes theorem is starting point]

   Hidden Markov models extensively used in gene sequencing analysis.
We will perform statistical analysis of data from three bioinformatics experiments to demonstrate various methods employed.

1. **Linkage Disequilibrium (LD) studies of SNP data**  
   (simple probability estimations)

2. **DNA expression microarray data**  
   (To identify significantly expressed genes under disease condition using student's t-test)

3. **Digital gene expression analysis of RNA-seq data**  
   (statistical test with negative binomial distribution)

Bayesian methods are not touched upon
Linkage Disequilibrium (LD)

LD is the non-random association of alleles at two or more loci, not necessarily on the same chromosome.

In a given population, we assume that the formation of haplotypes from alleles is a random phenomena.

Then, based on the observed frequency of alleles in a population, we can predict the rate of occurrence of certain combination of alleles or gene markers. [eg. SNP data from a population]

LD represents the occurrence of this observed combination in a population more than or less than the prediction based on random combination of alleles.

Amount of Linkage Disequilibrium (LD) is a measure of difference between the expected and observed allelic frequencies in the data.

Linkage Equilibrium ---> expected allelic frequencies

Linkage Disequilibrium ----> Deviation from the expectation
Theory

Take 2 loci A and B in a chromosome.
Let the two alleles be 1 and 2
This is a ’Two locus, two allele model’.
4 haplotypes are possible, each with a relative frequency in a population.

<table>
<thead>
<tr>
<th>Haplotype</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_1B_1$</td>
<td>$x_{11}$</td>
</tr>
<tr>
<td>$A_1B_2$</td>
<td>$x_{12}$</td>
</tr>
<tr>
<td>$A_2B_1$</td>
<td>$x_{21}$</td>
</tr>
<tr>
<td>$A_2B_2$</td>
<td>$x_{22}$</td>
</tr>
</tbody>
</table>

For each allele, the frequency can be worked out as follows:

<table>
<thead>
<tr>
<th>Allele</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_1$</td>
<td>$p_1 = x_{11} + x_{12}$</td>
</tr>
<tr>
<td>$A_2$</td>
<td>$p_2 = x_{21} + x_{22}$</td>
</tr>
<tr>
<td>$B_1$</td>
<td>$q_1 = x_{11} + x_{21}$</td>
</tr>
<tr>
<td>$B_2$</td>
<td>$q_2 = x_{21} + x_{22}$</td>
</tr>
</tbody>
</table>
- if alleles at the two loci are randomly associated with one another, then the frequencies of the four gametes are equal to the product of the frequencies of alleles it contains:

\[ x_{11} = p_1q_1 \]
\[ x_{12} = p_1q_2 \]
\[ x_{21} = p_2q_1 \]
\[ x_{22} = p_2q_2 \]

- in this situation, there is no linkage disequilibrium and gamete frequencies can be accurately followed using allele frequencies.
- if alleles at the two loci are not randomly associated, then there will be a deviation (D) in the expected frequencies:

\[ x_{11} = p_1q_1 + D \]
\[ x_{12} = p_1q_2 - D \]
\[ x_{21} = p_2q_1 - D \]
\[ x_{22} = p_2q_2 + D \]

- this parameter \textbf{D is the coefficient of linkage disequilibrium} first proposed by Lewontin and Kojima (1960).

- the most common expression of D is:

\[ D = x_{11}x_{22} - x_{12}x_{21} \]
If $D \neq 0$, two alleles are in Linkage Disequilibrium.

If $D = 0$, two alleles are in Linkage Equilibrium.

To remove the dependency of $D$ on the frequency of allele, $D$ is normalized:

$$D' = \frac{D}{D_{\text{max}}}$$

where,

$$D_{\text{max}} = \min(p_1q_1, p_2q_2) \quad \text{when } D < 0$$

$$D_{\text{max}} = \min(p_1q_2, p_2q_1) \quad \text{when } D > 0$$
\[ D' = 0 \] indicates that the two loci are independent of each other

\[ D' = 1 \] indicates a complete dependency between the two loci

Another way of describing LD is through correlation coefficient

\[ r = \frac{D}{\sqrt{p_1 p_2 q_1 q_2}} \]

When \( r^2 = 0 \), both the loci are in perfect equilibrium.

When \( r^2 = 1 \), both the loci are in perfect inequilibrium.

\[ D \] converges to zero with time at a rate depending on the recombination rate \( c \) of the loci, when no other force other than random transfers act.
Demo of Haploview
Microarray: a solid surface on which strands of polynucleotides have been attached in specific positions – called probes.

The probe consists of either

- CDNA printed on the surface or
- Shorter Oligonucleotides synthesized or deposited on the surface.

MRNA from samples is labelled with coloured dyes (Red, Green) and washed on the microarray with probes. These are called labeled targets.

Probes and samples are complementary.

Each probe on the array should bind to a quantity of labeled target.

The quantity of binding is proportional to the level of gene expression represented by that probe.
Illuminate the probe surface with laser light of same color as the fluorescent dye.

Measure the intensity of light over each probe on array.

After correcting experimental errors,

Intensity of fluorescence is proportional to the number of molecules of target bound to the surface.

Suppose,

control sample is treated with red dye

Treatment sample treated with green dye.

Depending upon the amount of control or treatment mRNA binding to the probe,

The spot reading can vary from red to green through varying yellow shades.

Expression levels are obtained for genomic region on the probe.
Different types of microarrays:
- Affymetrix GeneChip
- Illumina bead array
- Nimblegen array
- Two-color cDNA array
- Array CGH
- SNP microarrays
More........

Common theme:
- Representation of genes on slide
- Small portion of gene
- Larger sequence of gene

Different types of data processing, but common statistical methods of analysis.
Affymetrix – Steps of the Expression Assay

Image courtesy Affymetrix, www.affymetrix.com
Images

Full Array Image

Close-up of Array Image

Images courtesy Affymetrix, www.affymetrix.com
Disease studies with microarray data

In these studies, the expression levels of a given gene across many samples in normal condition are compared with disease condition.

<table>
<thead>
<tr>
<th>Region</th>
<th>Normal Condition</th>
<th>Disease Condition</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NS1   NS2 NS3 NS4</td>
<td>DS1   DS2 DS3 DS4</td>
</tr>
<tr>
<td>Region-1</td>
<td>n11   n12 n13 n14</td>
<td>d11   d12 d13 d14</td>
</tr>
<tr>
<td>Region-1</td>
<td>n21   n22 n23 n24</td>
<td>d21   d22 d23 d24</td>
</tr>
<tr>
<td>Region-3</td>
<td>n31   n32 n33 n34</td>
<td>d31   d32 d33 d34</td>
</tr>
<tr>
<td></td>
<td>...   ... .... ...</td>
<td>...   ... .... ...</td>
</tr>
</tbody>
</table>

Along each row, perform a statistical significance test between normal and disease data points to get the statistical significance of the observation.
Data Set: Gene expression profiling of disease Ewing's Sarcoma (Primary malignant bone tumor in children)

Microarray used: Affymetrix Human Genome HG95Av2 gene chip

Samples: 7 cases of primary tumor specimens, (treatment)

7 cases of metastases (control)

Analysis to identify differentially expressed genes to a good significance level of statistical test between control and treatment
Analysis Steps

Simpleaffy package in R/bioconductor was used.

- Probe summarization and Normalization between samples (MAS5 algorithm)
- Differential expression analysis (t-test)
- P-value correction with Bonferroni-Holmes method
- Annotation of probes to get gene names and other information
- Filtering based on statistical significance
- Study the short listed genes for biological relevance to the disease.
Statistical test

Two sample unpaired t-test

The t-statistic used is,

\[ t = \frac{\bar{X} - \bar{Y}}{S_w \sqrt{\frac{1}{n} + \frac{1}{m}}} \]

where,

\[ S_w = \sqrt{\frac{(n-1)S_x^2 + (m-1)S_Y^2}{n+m-2}} \]

With

\( \bar{X} \) is the mean of \( n \) control samples with variance \( S_X \)

\( \bar{Y} \) is the mean of \( m \) treatment samples with variance \( S_Y \)

For the computed t-value, get p-value.
(show the demonstration)
Relevant genes identified

**Tumor suppressor**: NEURL, FHIT, LLGL1, NF1

**Induction of apoptosis**: TNFRSF12, TGFB1, CASP10, TP63

**DNA repair**: IGHMBP2, XRCC2, ERCC2

( totally 24 genes in 7 functionalities, all *down regulated* in the poor prognosis signature patients).

RNA-seq Digital Gene Expression analysis

Study of RNA transcribed from a particular genome under investigation is called transcriptomics. Transcriptomes provide direct access to the gene regulation and protein information.

Many methods of RNA-sequencing exist, like:

- Sanger sequencing
- Serial Analysis of Gene Expression (SAGE)
- Massively Parallel Signature Sequencing (MPSS)

RNA-seq can be done on many platforms:

- Illumina : Genome analysing platform
- ABI : Solid sequencing
- Lifesciences : 454 sequencing
Basic methodology of RNA-seq

mRNA is extracted from a sample.

Many short regions of mRNA called **tags** of equal length are sequenced and recorded.

(Actually, cDNA corresponding to this mRNA is used for tags).

The length of the tag can vary between platforms and technology followed (36bp, 50bp, 100bp,...)

In **high-throughput sequencing**, about million tags or more are sequenced per sample. This is called a **library**.

In a library, each tag can occur multiple times.

The tag counts are representative of the abundance of corresponding mRNA molecule. This in turn represents either gene expression level or gene regulation level.

Called **Digital Gene Expression analysis**.
We can also map (align) millions of short reads to a reference genome. The number of overlapping mapped reads in a genomic region gives the abundance of corresponding mRNA molecule.
**Disease studies with RNA-seq data**

In these studies, the tag counts of a given genomic region across many samples in **normal condition** are compared with **disease condition**.

<table>
<thead>
<tr>
<th>Region</th>
<th>Normal Condition</th>
<th>Disease Condition</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NS1</td>
<td>NS2</td>
</tr>
<tr>
<td>Region-1</td>
<td>n11</td>
<td>n12</td>
</tr>
<tr>
<td>Region-1</td>
<td>n21</td>
<td>n22</td>
</tr>
<tr>
<td>Region-3</td>
<td>n31</td>
<td>n32</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Along each row, perform a statistical significance test between normal and disease data points to get the statistical significance of the observation.
Statistics used

Counts of each tag across multiple samples (libraries) follow a discrete statistical distribution.

Two main distributions have been widely used:

1. Poisson distribution
2. Negative binomial distribution

The negative binomial distribution has been widely studied and used in this type of analysis.

[Ref: Mark D Robinson et.al., Biostatistics (2008), 9, 2, pp. 321–332]
The tag count $y$ follows a negative binomial distribution,

$$y = NB(\mu, \phi)$$

$$f(y, \mu, \phi) = P(y) = \frac{\Gamma(y+\phi^{-1})}{\Gamma(\phi^{-1})\Gamma(y+1)}(\frac{1}{1+\mu \phi})^{\phi^{-1}}(\frac{\mu}{\phi-1+\mu})^{y}$$

Parameter phi is estimated from data. It accounts for the sample to sample variability.

A statistics is defined from phi and the data points which follows NB distribution under null hypothesis.

For observed statistic, p-value is computed to get significance.

Small number of samples can be analysed.

(Run the demo)