Computational Biology through Statistics and the Web

Susan Holmes
Statistics Department, Stanford
and INRA- Biométrie, Montpellier, France
susan@stat.stanford.edu
http://www-stat.stanford.edu/~susan/

Funded in part by a grant from NSF-DMS
Do we care about inferences for phylogenetic trees?

Cetacees: recognising what is being sold as Whale meat in Japan?
PCR sequencing: only in hotel room

Phylogenetic Identification of Whale and Dolphin Products

<table>
<thead>
<tr>
<th></th>
<th>KOREA 1994-07</th>
<th>JAPAN 1993-94</th>
<th>1998-00</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pygmy Right Bowhead Right</td>
<td>54</td>
<td>31</td>
<td>29</td>
</tr>
<tr>
<td>N. Minke</td>
<td>2</td>
<td>134</td>
<td>57</td>
</tr>
</tbody>
</table>
| S. Minke | 2 | 4 |-
| Bryde’s | 2 | 1 | 4 |
| Pygmy Bryde’s | 2 | | |
| Sei | | | |
| Humpback | 21 | 1 | |
| Fin | | | |
| Blue | 2 | | |
| Gray | 1 | 1 | |
| Pygmy Sperm Sperm | 1 | | |
| Other Beaked whales | 1 | | |
| Baird’s | 19 | 8 | |
| Beaked Cuvier’s | 1 | 3 | |
| Porpoises | 3 | 2 | |
| Killer whale | 1 | 1 | |
| Dolphin | 10 | 33 | 15 |
| Total | 81 | 257* | 117 |

1999 = 2
DNA sequences

Where do they come from? Sequence Yourself (PBS)

Data Bases: TIGR

Exploring our molecular selves: Molecular Selves

Human Genome: Human Genome Browser from UCSC Gateway and examples

Other genomes: The Institute for Genomic Research
Human immunodeficiency virus: Phylogeny and the origin of HIV-1
The origin of human immunodeficiency virus type 1 (HIV-1) is controversial.
Phylogeny has showed that viruses obtained from the Democratic Republic of Congo in Africa have a quantitatively different phylogenetic tree structure from those sampled in other parts of the world.

**Quest for the origin of AIDS**

This indicates that the structure of HIV-1 phylogenies is the result of epidemiological processes acting within human populations alone, and is not due to multiple cross-species transmission initiated by oral polio vaccination.

**Serial Passage** Conversely, phylogenetic analysis of HIV-1 sequences indicates that group M originated before the vaccination campaign, supporting a model of ‘natural transfer’ from chimpanzees to humans. If this timescale
is correct, then the OPV theory remains a viable hypothesis of HIV-1 origins only if the subtypes of group M differentiated in chimpanzees before their transmission to humans.
Korber and colleagues extrapolated the timing of the origin of HIV-1 group M back to a single viral ancestor in 1931, give or take about 12 years for 95% confidence limits.

Because this calendar of events obviously pre-dated the OPV trials, in the revised version of his book, Hooper suggested that group M first began to diverge in chimpanzees, and that there were then several independent transfers of virus to humans via OPV. In that case, several OPV batches should bear evidence of their production in chimpanzee tissue, yet no such evidence has been found.
The OPV batch that Hooper considered to be under most suspicion, however, was CHAT 10A-11.

An original vial of the batch was found at Britain’s National Institute for Biological Standards and Control, and the new tests show that it was prepared from rhesus-macaque cells.
Clues de decoding

Database: LANL HIV database

Finding matches
BLAST

Software for aligning sequences:
Clustal-W

Software for Building Trees:
Phylip.

Glossary: NiHGP glossary
How sure are we of the answers?

Phylogenetic Trees and Variability.

- Aggregating/Combining trees,
- Stability of sets of trees,
- Comparisons of sets of trees of several kinds.
- Explanation of one set of trees by another.
- Combining trees with other data.
- Confidence Statements for trees.

Specificities of this particular statistics class
• Genetic data is discrete: Counts, transitions, states.

• Independence is not the norm, (dependent data).

• Contingency Tables, (chisquare or not).

• Large data sets.

• Need to interface statistics programs with database searches (glue).

• Non standard parameters we need to estimate: trees, graphs,....
Some Aims of the Course

Learn the useful Probabilistic Tools specific to genetic/protein/expression data

Discrete random variables.
(Binomial, Multinomial, Poisson, Dirichlet,)
Monte Carlo Simulation.
Random walks, Gibbs sampling, Markov Chains.
Hidden Markov Models.
Expectation, conditional probability, variance.
Learn the statistical tools for analyzing large data sets

- Extreme values. (maximum, minimum)
- Multivariate analyses (PCA, SVD, CA, DA, Clustering).
- Maximum Likelihood Estimation, Bayesian methods, EM, variance stabilization.
- Multiple Testing.
- Non parametric regression (smoothing).
Learn to use R to do some statistical analyses of genomic/proteomic data

- HMM models for CpG islands.
- Simulating evolution, comparison with real data.
- Bootstrapping trees, confidence statements.
- Motif finding with the Gibbs sampler.
- Clustering methods.
- High dimensional data visualization.
- Microarray analyses (Bioconductor) (exploratory and discriminant).
Learn to read bioinformatics/statistical genetics papers
You will learn how to read and more important reproduce the results from the papers we read.
Learn to write up a statistical analyses

Fulfils the writing in the major requirement for the Mathematical and Computational Sciences major.

Step 1: (midterm) Find a real problem, and make a proposal of a simulation study or statistical analyses of a large data set.

Step 2: (final) Write a 10-15 page paper showing and interpreting your results.
Biology cannot be easily summarized into simple principles because it is a world of complex variation. It is variability that has enabled evolution, and it is variability that ensures the robustness of complex biological systems, it’s the rule rather than the exception in biological systems.

Statistics and probability provide many tools for decomposing the signals in genetic data.
Particularities of genomic data

- Genetic sequence data are often discrete: either binary, or categorical (A,C,G,T), most of the data comes in the form of counts, or frequency tables, we call these contingency tables.

Example:
Chargaff showed that the proportions of A,C,G,T were

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>T</th>
<th>G</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>30.3</td>
<td>30.3</td>
<td>19.5</td>
<td>19.9</td>
</tr>
<tr>
<td>Mycobacterium Tuberculosis</td>
<td>15.1</td>
<td>14.6</td>
<td>34.9</td>
<td>35.4</td>
</tr>
<tr>
<td>Sea Urchin</td>
<td>32.8</td>
<td>32.1</td>
<td>17.7</td>
<td>18.4</td>
</tr>
</tbody>
</table>

These proportions are not consistent with an equal distribution of the nucleotides in the 4 possible
Example of a simple test: Chi-square test for goodness of fit:

Null hypothesis: \( p_A = p_C = p_G = p_T = \frac{1}{4} \),

Statistic (measures distance to \( H_0 \)) = \[ \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \]
With 6 degrees of freedom, we have a significant Chisquare at the 0.01 level at 16.8, but here the statistic is equal to 29, well above the 1 % significance level.

```r
> charg
[1,]  30.3  30.3  19.5  19.9
[2,]  15.1  14.6  34.9  35.4
[3,]  32.8  32.1  17.7  18.4
> sum((charg-25)^2/25)
[1] 29.3152
> pchisq(29.3,6)
[1] 0.9999466
```
or Phenotypic Data

<table>
<thead>
<tr>
<th>eyes</th>
<th>Black</th>
<th>Brunette</th>
<th>Red</th>
<th>Blonde</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brown</td>
<td>68</td>
<td>20</td>
<td>15</td>
<td>5</td>
</tr>
<tr>
<td>Blue</td>
<td>119</td>
<td>84</td>
<td>54</td>
<td>29</td>
</tr>
<tr>
<td>Hazel</td>
<td>26</td>
<td>17</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td>Green</td>
<td>7</td>
<td>94</td>
<td>10</td>
<td>16</td>
</tr>
</tbody>
</table>

- Independence of observations or variables is not the norm, mostly the data show meaningful dependencies.
- Large data sets are much more common in molecular genetics than any other field of biology.
- We will need to interface statistical procedures with the large genetic database searches (the glue can be languages such as perl, R, python).
The parameters that we will be interested in are non standard, not just real vectors, they are trees (family trees of genes and of species are called phylogenetic trees and are of importance in the study of molecular evolution),
Phylogenetic tree of the envelope sequences shown in the courtroom during the trial of Dr. Schmidt.
graphs and networks

(genomes work together and it is of importance to understand how they interact in gene transcription networks:
or metabolic networks and pathways)
and rankings (permutations).
Non standard parameters

Thus what one is actually estimating in genetics is also very different from classical statistics. In classical statistics, we estimate what we don’t know, so primarily often denote it by a greek letter called a parameter. The most often the parameter is a real number. We might have an estimate on its own or we might get a lower or higher estimate which constitutes a confidence interval. We will see that in genetics the parameters are much more complicated.
Multinomial Distribution

\[ P(x_1, x_2, \ldots, x_m | p_1, \ldots, p_m) = \frac{n!}{\prod x_i!} \prod p_i^{x_i} \]

\[ = \binom{n}{x_1, x_2, \ldots, x_m} p_1^{x_1} p_2^{x_2} \cdots p_m^{x_m} \]
Example: Different codons can encode the same amino acid. Mycobacterium tuberculosis H37Rv [gbbct]: 3873 CDS’s (CDS=Coding Sequences) (1321373 codons)

<table>
<thead>
<tr>
<th>Triplet</th>
<th>Frequency: per thousand</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>UUU</td>
<td>6.2( 8159)</td>
<td></td>
</tr>
<tr>
<td>UUC</td>
<td>23.3(30825)</td>
<td></td>
</tr>
<tr>
<td>UUA</td>
<td>1.6( 2140)</td>
<td></td>
</tr>
<tr>
<td>UUG</td>
<td>17.9(23684)</td>
<td></td>
</tr>
<tr>
<td>CUU</td>
<td>5.4( 7188)</td>
<td></td>
</tr>
<tr>
<td>CUC</td>
<td>17.3(22811)</td>
<td></td>
</tr>
<tr>
<td>CUA</td>
<td>4.8( 6278)</td>
<td></td>
</tr>
<tr>
<td>CUG</td>
<td>50.5(66756)</td>
<td></td>
</tr>
<tr>
<td>AUU</td>
<td>6.5( 8551)</td>
<td></td>
</tr>
<tr>
<td>AUC</td>
<td>33.9(44767)</td>
<td></td>
</tr>
<tr>
<td>ACU</td>
<td>3.7( 4837)</td>
<td></td>
</tr>
<tr>
<td>ACC</td>
<td>35.2(46519)</td>
<td></td>
</tr>
<tr>
<td>AAA</td>
<td>5.3( 7003)</td>
<td></td>
</tr>
<tr>
<td>AAC</td>
<td>19.9(26348)</td>
<td></td>
</tr>
<tr>
<td>AAG</td>
<td>1.9( 2474)</td>
<td></td>
</tr>
<tr>
<td>AGA</td>
<td>1.9( 2474)</td>
<td></td>
</tr>
<tr>
<td>AGC</td>
<td>1.9( 2474)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>AUA</td>
<td>2.2 (2893)</td>
<td>ACA</td>
</tr>
<tr>
<td>AUG</td>
<td>18.4 (24348)</td>
<td>ACG</td>
</tr>
<tr>
<td>GUU</td>
<td>8.0 (10578)</td>
<td>GCU</td>
</tr>
<tr>
<td>GUC</td>
<td>32.7 (43214)</td>
<td>GCC</td>
</tr>
<tr>
<td>GUA</td>
<td>4.7 (6274)</td>
<td>GCA</td>
</tr>
<tr>
<td>GUG</td>
<td>40.1 (52998)</td>
<td>GCG</td>
</tr>
</tbody>
</table>
Proline is $CC^*$ (regular expression), there are 4 alternative spellings:

<table>
<thead>
<tr>
<th>Codon</th>
<th>o/oo</th>
<th>count</th>
<th>$p_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCU</td>
<td>3.4</td>
<td>(4457) 0.059</td>
<td></td>
</tr>
<tr>
<td>CCC</td>
<td>17.0</td>
<td>(22503) 0.294</td>
<td></td>
</tr>
<tr>
<td>CCA</td>
<td>6.1</td>
<td>(8085) 0.106</td>
<td></td>
</tr>
<tr>
<td>CCG</td>
<td>31.4</td>
<td>(41507) 0.542</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>57.9</td>
<td>76552 1.001</td>
<td></td>
</tr>
</tbody>
</table>
Looking for epitopes. Epitope: *Specific potion of a macromolecular antigen to which an antibody binds. In the case of a protein antigen recognized by a T-cell, the epitope or determinant is the peptide portion that binds to an Major Histocompatibility Complex (MHC) molecule for recognition by the T cell receptor (TCR)*

From: *Web glossary for Antibody*

A type of protein made by certain white blood cells in response to a foreign substance (antigen). Each antibody can bind to only a specific antigen. The purpose of this binding is to help destroy the antigen. Antibodies can work in several ways, depending on the nature of the
antigen. Some antibodies destroy antigens directly. Others make it easier for white blood cells to destroy the antigen.

Suppose we have a chemical test for an allergic that gives hit score of either 0 or 1.

In a background noisy environment the 1’s occur with a small probability say \( p_0 = \frac{1}{100} \).

If the protein sequence tested provokes an allergic reaction, (we will call it an epitope), the probability of a 1 will be close to the overall proportion of people who have been sensitized to this epitope in the population, maybe \( p = \frac{10}{100} \).
We test 150 protein sequences (150 windows moved along a sequence) on 50 people, for each we have the number of hits out of 50:
Histogram of Epitope Hits along the Protein

Simplest Setup: Poisson iid background rate is $50/100 = .5 = \lambda$. The maximum we see here is: 9, what are the chances that for a Poisson background with parameter $\lambda = 0.5$, we observe a value as big as 9.
Here we will use **extreme value theory**: the distribution of the order statistic \( x(n) = \max\{x_1, x_2, x_3, \ldots, x_{150}\} \).

What are the chances that the maximum is as big as 9?

\[
P(x(n) \geq 9) = 1 - P(x(n) < 9) = 1 - \prod_{i=1}^{n} P(x_i < 9)
\]

(supposing independence)

\[
\prod_{i=1}^{n} P(x_i < 9) = \left( \sum_{k=0}^{8} \frac{e^{-\lambda} \lambda^k}{k!} \right)^n = \left( 1 - \sum_{k=9}^{\infty} \frac{e^{-\lambda} \lambda^k}{k!} \right)^n
\]

(supposing identically dist. Poisson) Call

\[
\epsilon = \sum_{k=9}^{\infty} \frac{e^{-\lambda} \lambda^k}{k!} = P(x_i \geq 9).
\]
How small is $\epsilon$?

R

> sum(dpois(0:5,0.5))
[1] 0.9999858

> ppois(5,0.5)
[1] 0.9999858

> sum(dpois(0:8,0.5))
[1] 1

$$(1 - \epsilon)^n = \exp(n \log(1 - \epsilon)) = \exp(-n\epsilon) = e^{-nP(x_i \geq 9)}$$
How do we conclude that the spike that we see corresponds to an epitope?

```r
> epsilon=1-ppois(8,0.5)
> epsilon
[1] 3.43549e-09
> 150*epsilon
[1] 5.153236e-07
> exp(-150*epsilon)
[1] 0.9999995
> 1- exp(-150*epsilon)
[1] 5.153234e-07
```

\[
P(x_{(n)} \geq 9) = 1 - e^{-nP(x_i \geq 9)} = 1 - 0.9999995 = 5.10^{-7}
\]

This is a very small probability.
Probabilistic tools that are useful in the analysis of DNA, protein or gene expression data: binomials, Beta, multinomials, Dirichlet, Poisson, chisquare distributions. We will see how useful Monte Carlo simulations are for complex situations where analytical solutions are unavailable. Even more sophisticated are the current use of Monte Carlo Markov chains and random walks for computation of posterior distributions in modern Bayesian settings and the use of hidden Markov models for sequence analysis.
Web References

- Probability Distributions
- Probability by Surprise
- Statistics of Sequence Similarity Scores
- Genetics Glossary for the layperson
Modeling the Genes Grammar

Hidden Markov Models

Estimating the most likely states: Viterbi Algorithm:
Viterbi movie
Viterbi Animation: Steps of the Viterbi Method
Computational Biology Resources

Stat 166: Undergraduate Computational Biology

Animations for Biology

Making Sense of DNA and Protein Sequences: an Interactive NCBI Mini-Course Minicourse

Transmembrane Helix: TMH Finder

Genscan: Genscan: Gene Annotation Program NY Times Article About This
Monte Carlo Methods

Buffon’s Needle: Buffon’s Needle Applet
Four unnormalized data sets: R, G, Rb, Gb

Imported Data from SMD
Mean intensities.
Channel 1: Green, gDNA.
Channel 2: Red, mRNA.
Example: T-Cells and Cancer
A little immunology

T-lymphocyte cells (T-cells) originally derive from stem cells of the bone marrow. At around the time of birth, lymphocytes derived in this way leave the marrow and pass to the thymus gland in the chest, where they multiply.
The lymphocytes are processed by the thymus gland, so that between them they carry the genetic information necessary to react with a multitude of possible antigens.
The human genome is presently estimated to contain as few as 25 thousand genes (we should soon know the number exactly).

The number of T-cell receptors for antigen (TCRs) that we make is estimated at $25 \times 10^6$.

**Antigens** are macromolecules that elicit an immune response in the body. Antigens can be proteins, polysaccharides, conjugates of polysaccharides and proteins.

T-cell diversity is attained by a complex 5-stage genetic rearrangement that occurs at random in the developing T-cells in the thymus.
T-cells are tested for their ability to recognise and bond to antigens in the thymus.

It is thought that this occurs by positive and negative selection.

The postulation is that positive selection eliminates
T-cells that do not bond tightly enough to antigen type molecules produced in the thymus and negative selection eliminates T-cells that bond too tightly to self type molecules found in the thymus.

This produces a mature T-cell that is effective against antigens but is also self tolerant. It is known that approximately 95% of all developing T-cells die in the thymus.

The T-cells, each processed to 'recognise' and interact with a specific antigen, circulate permanently between the blood and lymphatic systems.

On recognition of the antigen by a helper T-cell, one or a group of lymphocytes takes up residence in secondary lymphoid tissue (e.g. lymph glands, spleen, bone
marrow) and divides to form two types of cells, memory cells, which are lymphocytes processed in the same way as themselves, and killer cells, which interact with the antigen.

T-cell receptor sites sit on the surface of T cells and provide the specificity in antigen binding. They have two components that give each different type a unique surface morphology, and it is this that allows bonding to antigens.
This enhanced electron microscope picture shows a Killer T-cell (top right) about to attack a larger cancer cell.
T cell populations in the periphery

T-Cells are responsible for destroying infected or Cancerous Cells, and for coordinating all Acquired Immune Responses. For this reason, T-Cell Immunity is generally called Cellular Immunity.

Two sub-types of T-Cells are responsible

- The Cytotoxic T-Cell (CD8+)
- The Helper T-Cell (CD4+)

Note: CD’s (CD = Cluster of Designation) are proteins on the surface of these white blood cells that help them communicate with other cells in our bodies to eliminate
microorganisms. Without these cell surface proteins the immune system would not be able to communicate properly and would not be able to eliminate foreign invaders from our bodies. They are a very important part of our immune system.
Well known CD genes

For those who spend their lifes perusing the gene lists like I do, another source of proteins:

Surface Markers of T and B Cells

<table>
<thead>
<tr>
<th>Lymphocyte</th>
<th>Human</th>
<th>Mouse</th>
</tr>
</thead>
<tbody>
<tr>
<td>T Cells immature</td>
<td>CD5, CD1, CD2,</td>
<td>Thy-1, Ly-1, 2, 3, and T1</td>
</tr>
<tr>
<td>Mature Subsets</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TH and TD Cells</td>
<td>CD4, CD3, TCR</td>
<td>Ty-1, Ly-1, L3T4, Qa1</td>
</tr>
<tr>
<td>TS and TC Cell</td>
<td>Cd8, cd3, TCR</td>
<td>Thy-1, Ly-2,3</td>
</tr>
<tr>
<td>Activated T cells</td>
<td>HLA-Dr, Cd25 (tac)</td>
<td>Thy-1, Ly-2,3</td>
</tr>
<tr>
<td>B Cells</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B Lymphocytes</td>
<td>CD19-22, C1R, mlg</td>
<td>lyb-2,3, 5, mlg PC1</td>
</tr>
<tr>
<td>Plasma Cells</td>
<td>?</td>
<td></td>
</tr>
<tr>
<td>Null Cells</td>
<td></td>
<td></td>
</tr>
<tr>
<td>K, NK, NC, LAK</td>
<td>Display neither T nor B markers</td>
<td>NK1, CD11b, CD16</td>
</tr>
</tbody>
</table>

Each T-Cell has a unique surface molecule much like an ImmunoGlobulin’s, called a T-Cell Receptor (TCR).
Unlike Ig’s that can recognize any molecule, the TCR is restricted to recognize only short Amino Acid Chains, displayed on the surface of cells, in conjunction with a molecule, called the Major Histocompatibility Complex (MHC).

Almost all the body’s cells are constantly producing MHC molecules and attaching small internal proteins to
them, for expression on their surface. T-Cells probe the surface of all cells for MHC complexes.

Mature CD8+ Cells move through the body, searching for cells that possess complexes, to which the TCR will bind and proceed to destroy those cells.

Memory CD8+ Cells function like Memory B-Cells, they persist and will multiply and mature if they are re-exposed to the same MHC complex.

CD8 positive (CD8+) T cells are able to detect and destroy cells of our body that have changed and are no longer normal. If cells in our body are infected with a virus oftentimes the virus will cause the cell to place viral proteins on the surface of the cell. The CD8 cells can detect these changes however they usually need CD4
positive T cells help to destroy these infected cells.

αβ T cells:

- make up 90-95% of peripheral T cells
- express CD4 or CD8, i.e., are MHC class II or I restricted
- consist of naive, effector or memory T cells
- naive T cells express high levels of L-selectin, a homing receptor which allows extravasation into lymph nodes via high endothelial cells of postcapillary venules and recirculation. Naive T cells also express CD45RA. In the absence of antigenic stimulation naive T cells live 5-7 weeks. If a naive T cell recognizes specific antigen complexed with MHC of an antigen presenting cell or
target cell it will become activated and initiate a primary response.

- activated T cells differentiate within 48 hours into effector cells (T helper 1 and 2, T cytotoxic, T delayed-type hypersensitivity) 5-7 days post-activation. Effector T cells live days to a few weeks and express high levels of adhesion molecules.

- memory T cells generated during a primary immune response persist for many years in a resting state but continue to express many of the same cell-surface molecules as effector T cells. Memory T cells express CD45RO but not L-selectin. Instead of recirculating through LN, memory T cells recirculate to the tissue in which they were originally stimulated. Since memory
T cells have less stringent activation requirements than naive T cells, they respond rapidly to a second antigenic challenge.
Biological Questions

- Do cancer patients show differential expression in any genes expressed in T-cells?
- Are there any differences between naive effector and memory T-cells?
- What are the steps involved in T-cell differentiation?
70 agilent arrays, 13,000 genes in 5 batches. Much batch to batch effect, no replication.

- five patients with melanoma, five healthy subjects.

- Stimulated and unstimulated T-cells of 7 different types (naive effector and memory, and stimulated naive, effector and memory).

- Stratagene Human Reference.