MODERN STATISTICS FOR BIOLOGY

Susan Holmes and Wolfgang Huber

Bios221/Stats366 Lecture 1 Fall 2019

Bios221/Stat366 Course Presentation - Lecture 1

Statistics, Biology

Susan Holmes, Statistics, Sequoia 102,
susan@stat.stanford.edu
Wolfgang Huber, Statistics, Sequoia 230,
whuber@stanford.edu

canvas:https://canvas.stanford.edu/courses/108798
TA: Nikos Ignatiadis ignat@stanford.edu.

Specificities of modern data

- Genetic data are discrete: Counts, transitions, states.
- Independence is not the norm, (dependent data).
- Contingency Tables, (chisquare or not).
- Large heterogeneous data sets.
- Need to interface statistics programs with database searches, ontologies, databases (glue).
- Non standard parameters we need to estimate: trees, graphs,....
- Complex Plotting procedures.
- Reproducibility of all research (diaries, write-ups, documentation).

General Principles do apply across technologies

- Generative - probability models (Poisson, binomial).
- Statistical methods (maximum likelihood).
- High quality graphics at three different levels.
- Data transformations.
- Removing unwanted variation.
- Experimental design.

Some Aims of the Course

Learn the useful Probabilistic Tools specific to genetic / protein / expression/ metabolic/ immunological or microbial data

Discrete random variables .

(Binomial, Multinomial, Poisson, Dirichlet,)

Monte Carlo Simulation.

(Bootstrap, nonparametric testing, power).

Filtering, de-noising, modeling, transforming.

(when to use log, other transforms).

Markov Chains

(transitions, dependencies).

Mixture models.

(latent variables, EM).

Expectation, conditional probability, variance.

(need to know basis).

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.
- Pattern, Trend Detection.
- Non parametric regression (smoothing).
- Machine Learning.

Learn to design your experiments and analyses

- Power computations, randomization, sensitivity, robustness.
- Reproducible research.

Learn to use R to do some statistical analyses of genomic/proteomic/count data

- R for sequence analyses and interfacing with databases (`Biostrings, DEseq2, Bayeseq, edgeR`).
- R/ Bioconductor suite of bioinformatics packages (`Biostrings, Biomart, `)
- R for phylogenetics (`ape, phangorn, distory, phyloseq`).
- R for high quality visualizations `ggplot2`.
- R for multivariate multi-table analyses `ade4, vegan, phyloseq`.
- R for input and normalization of data from modern technologies (`ShortRead, DEseq2, edgeR`).
- R for network and dynamic plotting `igraph, animation, statnet`.

Useful concepts and algorithms

- Mixture models (example of CpG islands, Cytof and FlowCytometry).
- Simulation and comparison with real data.
- Clustering methods/ EM algorithm.
- Normalization, Variance Stabilization, linear models.
- Methods for heterogeneous data, images, trees, networks.
- High dimensional data visualization.
- Combining data and information from databases (KEGG, GO, GeneBank, UCSC).
- Testing and Multiple Hypotheses Correction.
- Experimental Design.

Learn to read bioinformatics/computational 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

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

Due Date: October 11th.

Step 2: (final) Write a 15 page paper showing and interpreting your results.

Due date: November 17th.

Worlds of Variability

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 medical, genetic and ecological 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.
  Examples:
  Sometimes these tables represent transitions between ‘states’.

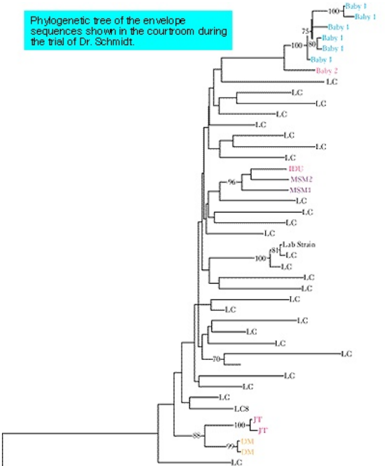
	followed by			
first	A	C	G	T
A	34	09	14	33
C	12	13	10	10
G	11	15	17	09
T	13	22	20	23

Phenotypic Data

eyes	Black	Brunette	Red	Blonde
Brown	68	20	15	5
Blue	119	84	54	29
Hazel	26	17	14	14
Green	7	94	10	16

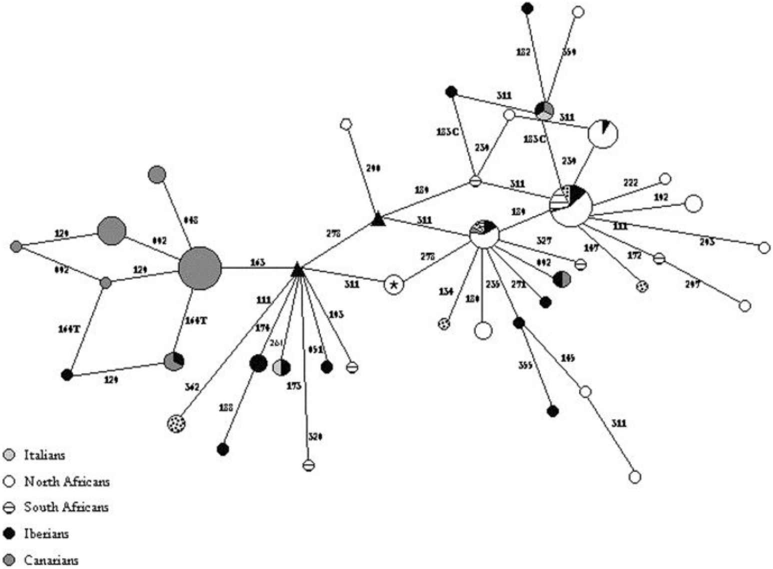
- Independence of observations or variables is not the norm, mostly the data show meaningful dependencies.
- Large data sets are much more common in molecular genetics than any other field of biology.
- We will need to interface statistical procedures with the large biological database (the glue can be languages such as `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),



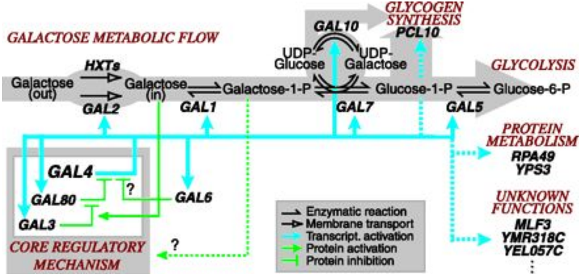
Phylogenetic tree of the envelope sequences shown in the courtroom during the trial of Dr. Schmidt.

graphs and networks

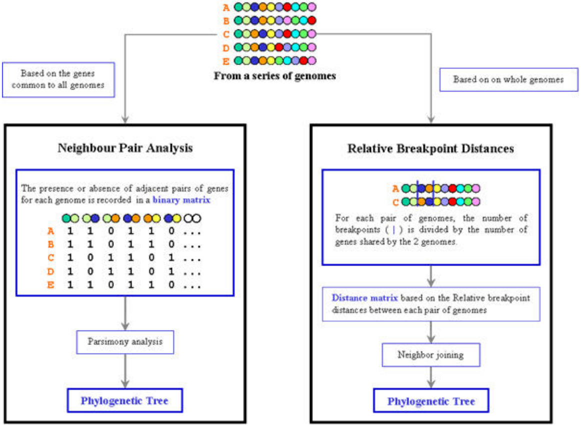


Genes work together and it is of importance to understand how they interact in gene transcription networks.

or metabolic networks and pathways.



and rankings (permutations).



From a series of genomes

Based on the genes common to all genomes

Based on on whole genomes

Neighbour Pair Analysis

The presence or absence of adjacent pairs of genes for each genome is recorded in a **binary matrix**

A	1	1	0	1	1	0	...
B	1	1	0	1	1	0	...
C	1	0	1	1	0	1	...
D	1	0	1	1	0	1	...
E	1	1	0	1	1	0	...

Parsimony analysis

Phylogenetic Tree

Relative Breakpoint Distances

For each pair of genomes, the number of breakpoints (|) is divided by the number of genes shared by the 2 genomes.

Distance matrix based on the Relative breakpoint distances between each pair of genomes

Neighbor joining

Phylogenetic Tree

Non standard parameters

Thus what one is actually estimating in genetics, immunology or microbiology 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 biology the parameters are much more complicated.

Challenges that we will focus on

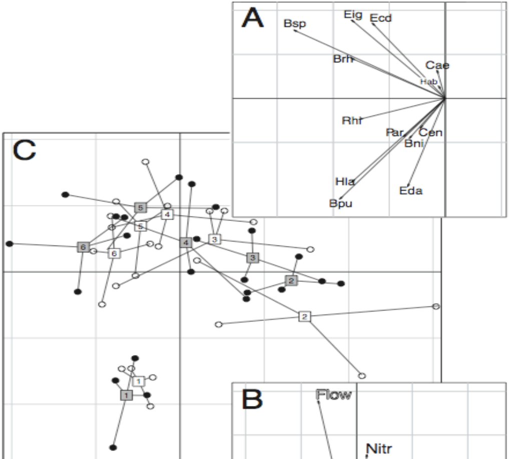
- Heterogeneity.
- Structured high-dimensionality.
- Graph or Tree integration.
- High quality graphics.
- Reproducibility.
- Validation and confirmatory analysis.

Heterogeneity of Data

- Status : response/ explanatory.
- Hidden (latent)/measured.
- Types :
  - Continuous
  - Binary, categorical
  - Graphs/ Trees
  - Images
  - Maps/ Spatial Information
  - Rankings
- Amounts of dependency: independent/time series/spatial.
- Different technologies used (16S-rRNA, illumina, Minion, Mass-cyto, Mass Spec, RNA-seq).

A Systems Approach leads to structured high dimensional data

We need a system that provides good data integration



Probabilistic tools that are useful in the analysis of DNA, protein or genetic 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 (we won't have time for alot).

We will not separate the theory necessary to our propos, and will remind the reader of the relevant definitions as we need them. (Expectations, variances and conditional probability are the foundations of much of what we will need, an introductory probability text such as Ross Pitmanor Grinstead and Snell are good places to look for more details).

What is Probability?

Allows one to go from a hypothetical model, usually parametric, to the probability of an event.
Probability is the theory of random variability, we use randomness to model uncertainty and noise and its consequences on what would be observed under certain probabilistic models.

What is Statistics?

Reasoning backwards from data to a potential explanation for what we see.

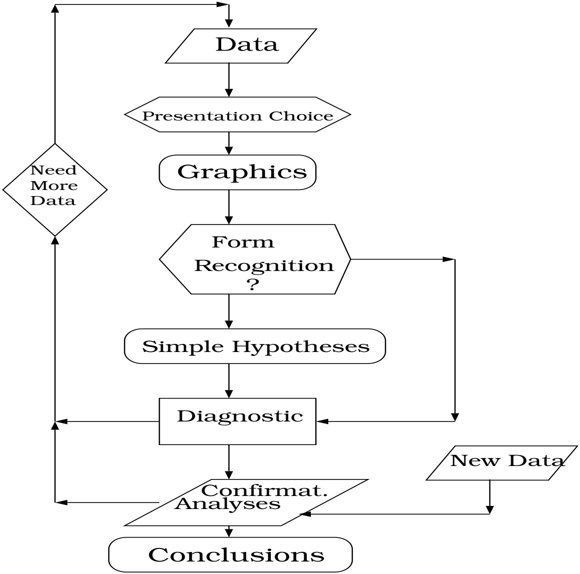
Statistics is a separate subject from probability, and although the latter serves as the mathematical basis for making inferences, statistics does not exist without data, and in the case of contemporary genetics, large data sets are the norm.

We need to know quite a few things about the probabilistic models that might have generated the data, in order to go back from the data and guess at the best possible model from which we think the data came.

Hardest part: curse of dimensionality.
Another source of complication we will encounter is the high dimensionality of the biological phenomena under hand; genes work together so their expression patterns can be highly correlated, so we cannot look at variables one at a time. The methods that enable the study of all the variability at once are multivariate techniques such as principal components analysis, correspondence analysis, multidimensional scaling and cluster analysis.
When one doesn't have a model,need to have a way of looking at the data. These methods are for data dimension reduction and visualization is another part of what we call exploratory data analyses.

Statistics is not only p-values

Statistics is often caricatured as a method for obtaining p-values; this is far from what statisticians spend most of their time doing. It is true that before the computer age, statisticians mostly used probability theory to make statements about their inferences.



Modern Statistical Analysis is Iterative

Web References

Class website: `http://bios221.stanford.edu`

- Install R
- Install RStudio
- Install Bioconductor
- Probability Distributions