STAT877: Statistical Methods for Molecular Biology

Christina Kendziorski
with Cecile Ane, Karl Broman, Mark Craven, Colin Dewey, Sunduz Keles, Bret Larget, Michael Newton, Sushmita Roy, Sijian Wang and Yingqi Zhao

http://www.biostat.wisc.edu/~kendzior/stat877.html
Disclaimer: there is a lot of material crammed onto the next five slides, and a few others spread throughout. I wanted you to have a record of this information in writing – please never make slides like this for a scientific presentation!

And see Brad Efron’s 13 rules for giving a really bad talk.
What this class is

- The course will provide a statistical perspective on some current biological problems, with an introduction to statistical analysis in genomics, phylogenetics, gene regulation, gene expression, sequencing, linkage analysis, and related areas.

- Within each of the areas covered, we will discuss motivation, study design and associated limitations, data collection, data features, models, assumptions, fitting algorithms, diagnostics, results, interpretation, and open questions.

- Statistical concepts will include: stochastic modeling, hierarchical modeling, likelihood methods, Bayesian methods, multivariate analysis methods, model selection, high-dimensional parameters, experimental design strategies, and methods for multiple testing.

- Biological concepts will include: measurement of DNA, RNA, and protein; genomic data resources; the relationship between genotype and phenotype in various settings, and phylogenies.
What this class isn’t

- Overview of basic statistics for molecular biologists (hypothesis testing, p-values, linear models, regression, etc.)

- We assume
  - Some knowledge of very basic biology and genetics (Genetics 466 is recommended, but not required). 466: Mendelian genetics, mapping, molecular genetics, genetic engineering, cytogenetics, quantitative genetics, and population genetics.
  - That you will prepare for class and discuss any perceived deficiencies in these areas with CK as soon as possible.
Topics covered

Jan 21, 23: Intro; data collection and resources (CKendziorski)
Jan 28, 30, Feb.4: QTL mapping (KBroman)
Feb 6,11,13,18: Phylogenies (BLarget and CAne)
Feb 20,25: Alignment (MCraven and CDewey)
Feb 27, Mar 4, 6: Chip-Seq (SKeles)
Mar 11,13: RNA-seq (CDewey Mar.11, CKendziorski on Mar.13)
Mar 18,20: Spring Break
Mar 25,27: GSEA, FDR (MNewton)
Apr 1,3,8: Classification and clustering (SWang)
Apr 10: Biomarker development (CKendziorski)
Apr 15,17: Network reconstruction (SRoy)
Apr 22,24: Dynamic treatment regimes and personalized medicine (YZhao)
Apr 29, May1: Student presentations
May 6,8: Student presentations
Evaluation

- Quiz on basic biology - January 28 (5%)
- Homework assignments (50%; calculated from top 5).
- Course project paper (outline 5%; draft 5%; final 15%)
- Course project presentation (20%)
Course projects

- Individual or groups (2-3) to be decided by January 28 (interests/background collected at end of class today).

- Project outlines due week of February 10

- Project drafts due week of March 17

- Outlines must include: clear statement of problem(s) being addressed, deliverables, specification of who is primarily responsible for what (if you are working in a group).

- Drafts must include: clear statement of problem(s) being addressed, why doing so is interesting/important, background specifying what has been done in this area and why what you’re doing is necessary and likely to be better than what is currently available, sketch on details of approach.

- Final papers must include what’s in the draft along with details of the approach, results from simulations and/or case-study data, and a discussion of results.

- Final presentation: 20 minutes (10 minute oral presentation; 10 minute oral exam).
Central Dogma of Molecular Biology
Central Dogma (extended by CK for this class)

DNA → mRNA → Protein → Traits: could be binary or continuous (quantitative).

- Intermediate phenotypes
- Phenotypes

- Brief background on where/how DNA is stored.
- What does each of the structures above look like? How are they made?
- How are they measured? [Later]
- What types of statistical questions commonly arise? [Later]
Genes contain instructions for making proteins.

Proteins act alone or in complexes to perform many cellular functions.

U.S. DEPARTMENT OF ENERGY
Much of the information in the next third of this lecture is taken from Wikipedia
Cells

- Cells are the smallest unit of life that can replicate independently, and are often called the "building blocks of life".

- There are two types of cells, eukaryotic, which contain a nucleus, and prokaryotic, which do not. Prokaryotic cells are usually single-celled organisms, while eukaryotic cells are usually part of multicellular organisms.

- The vast majority of DNA in a human cell is contained in the cell nucleus (the nuclear genome). In humans, the nuclear genome is divided into 46 linear DNA molecules called chromosomes, including 22 homologous chromosome pairs and a pair of sex chromosomes.

- There is also some DNA in the mitochondria. The mitochondrial genome is a circular DNA molecule distinct from the nuclear DNA. Although the mitochondrial DNA is very small compared to nuclear chromosomes, it codes for 13 proteins involved in mitochondrial energy production and specific tRNAs.

From Wikipedia
Cells

- A cell’s nucleus contains its chromosomes, and is the place where almost all DNA replication and RNA synthesis (transcription) occur.

- The nucleus is spherical and separated from the cytoplasm by a double membrane called the nuclear envelope. The nuclear envelope isolates and protects a cell's DNA from various molecules that could accidentally damage its structure or interfere with its processing.

- During processing of protein, DNA is transcribed, or copied into a special RNA, called messenger RNA (mRNA); mRNA is then transported out of the nucleus, where it is translated into a specific protein molecule.
A Cell
And another version
Recall: The vast majority of DNA in a human cell is contained in the cell nucleus (the nuclear genome). And most of the time when people say human genome they are referring to the nuclear genome.

The human genome is distributed along 23 pairs of chromosomes: 22 autosomal pairs and the sex chromosome pair (XX for females and XY for males).

In each pair, one chromosome is maternally inherited, the other paternally inherited.

On average, a single human chromosome consists of DNA molecules that are almost 5 centimeters long (calculation a bit later).
Human Chromosomes

HUMAN CHROMOSOMES

- Centromere
- Telomere
- Chromatid
Number of chromosomes - comparison

<table>
<thead>
<tr>
<th>Organism</th>
<th>Number of chromosomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>pea plant</td>
<td>14</td>
</tr>
<tr>
<td>sun flower</td>
<td>34</td>
</tr>
<tr>
<td>cat</td>
<td>38</td>
</tr>
<tr>
<td>puffer fish</td>
<td>42</td>
</tr>
<tr>
<td>human</td>
<td>46</td>
</tr>
<tr>
<td>dog</td>
<td>78</td>
</tr>
</tbody>
</table>

From http://www.genomenewsnetwork.org/resources/whats_a_genome/Chp1_2_1.shtml
Unrelated to complexity

<table>
<thead>
<tr>
<th>Organism</th>
<th>Genome size (base pairs)</th>
<th>Chromosome number (n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amoeba dubia</td>
<td>670,000,000,000</td>
<td>Several hundred</td>
</tr>
<tr>
<td>Trumpet lily (Lilium longiflorum)</td>
<td>90,000,000,000</td>
<td>12</td>
</tr>
<tr>
<td>Mouse (Mus musculus)</td>
<td>3,454,200,000</td>
<td>20</td>
</tr>
<tr>
<td>Human (Homo sapiens)</td>
<td>3,200,000,000</td>
<td>23</td>
</tr>
<tr>
<td>Carp (Cyprinus carpio)</td>
<td>1,700,000,000</td>
<td>49</td>
</tr>
<tr>
<td>Chicken (Gallus gallus)</td>
<td>1,200,000,000</td>
<td>39</td>
</tr>
<tr>
<td>Housefly (Musca domestica)</td>
<td>900,000,000</td>
<td>6</td>
</tr>
<tr>
<td>Tomato (Lycopersicon esculentum)</td>
<td>655,000,000</td>
<td>12</td>
</tr>
</tbody>
</table>

What’s in a number? Genome size and chromosome number seem unrelated to complexity. The figures above are for haploid genomes — most cells are diploid (2n), carrying two copies of each chromosome.

From http://universe-review.ca/F11-monocell13.htm
Central Dogma (extended by CK for this class)

- DNA ➔ mRNA ➔ Protein ➔ Traits: could be binary or continuous (quantitative).

  intermediate phenotypes

  phenotypes

- Brief background on where/how DNA is stored.
- What does each of the structures above look like?
- How are they measured? [Later]
- What types of statistical questions commonly arise?
**DNA**

- A deoxyribonucleic acid or DNA molecule is a double-stranded polymer composed of four basic molecular units called nucleotides.
- Nucleotides: adenine (A), guanine (G), cytosine (C), and thymine (T).
- Each nucleotide comprises: one of these bases, a phosphate group, and a deoxyribose sugar.

![Diagram of a nucleotide showing the phosphate group, deoxyribose sugar, and nitrogenous base.](image)

*Fig. 3.1. A nucleotide consists of a pentose sugar, a phosphate group and a nitrogenous base.*
DNA

- Base pairing: A with T; G with C
- Chains are held together by hydrogen bonds between the nitrogen bases

Figure: http://www.rpi.edu/dept/chem-eng/Biotech-Environ/Projects00/rdna/rdna.html
DNA

- Base pairing: A with T; G with C
- Chains are held together by hydrogen bonds between the nitrogen bases

Figure: http://www.angelfire.com/sc3/toxchick/biochemistry/biochemistry15.html
DNA

- Polynucleotide chains are directional molecules, with slightly different structures marking the two ends of the chains, the so-called 3’ (pronounced 3 prime) end and the 5’ end.
- The 5’ end is the leading end and the 3’ end is the tail end of a nucleic acid. This terminology is due to the numbering of carbon atoms in the sugar ring (reference?).
- The 5’ end carries a phosphate group and the 3’ end carries a sugar group.
- The two complementary strands of DNA are antiparallel (i.e., the 5’ end to 3’ end directions for each strand are opposite).
- Nucleic Acids can only be synthesized in vivo in the 5’ to 3’ direction and so, traditionally, DNA and RNA sequences are written from 5’ to 3’.
DNA

- The 5’ end is the leading end and the 3’ end is the tail end of a nucleic acid. This terminology is due to the numbering of carbon atoms in the sugar ring (?).
Figure https://www.boundless.com/chemistry/polymers/nucleic-acids/dna-replication-and-protein-synthesis-22cb2ab5-fa3b-4ec7-8141-774ee03c80c9/
Cool fact about DNA

- Nucleotide: building blocks of DNA (measured in base pairs (bp)).

- The haploid human genome contains approximately 3 billion base pairs of DNA packaged into 23 chromosomes. That makes a total of 6 billion base pairs of DNA per cell.

- Each base pair is around 0.34 nanometers long (a nanometer is one-billionth of a meter). Each diploid cell therefore contains about 2 meters of DNA \([(0.34 \times 10^{-9}) \times (6 \times 10^9)]\).

- Moreover, it is estimated that the human body contains about 50 trillion cells—which works out to 100 trillion meters of DNA per human.

- Considering the fact that the Sun is 150 billion meters from Earth, this means that each of us has enough DNA to go from here to the Sun and back more than 300 times, or around Earth's equator 2.5 million times! How is this possible?

Inside the nucleus, DNA wraps around histones
Central Dogma (extended by CK for this class)

DNA → mRNA → Protein → Traits: could be binary or continuous (quantitative).

- Brief background on where/how DNA is stored.
- What does each of the structures above look like?
- How are they measured? [Later]
- What types of statistical questions commonly arise?
Evolving definition of a gene

1860s-1900s: a discrete unit of heredity (Mendel)

1910s: a distinct locus (Morgan)

1940s: the blueprint for a protein (Beadle & Tatum)

1960s: a transcribed code (Watson & Crick)

Genome era: a locatable region of genomic sequence, corresponding to a unit of inheritance, which is associated with regulatory regions, transcribed regions and/or other functional sequence regions

Gerstein et al. 2007: The gene is a union of genomic sequences encoding a coherent set of potentially overlapping functional products
Locating genes on chromosomes

- The approximate location of a gene is often specified by the chromosome number (1, 2, …, 22, X, Y), the arm (p, q), and the position on the arm (regions, bands within regions, sub-bands within bands, …).
Typical eukaryotic protein coding gene

- Promoters and enhancers determine what portions of the DNA will be transcribed into the precursor mRNA (pre-mRNA).
- The pre-mRNA is then spliced into messenger RNA (mRNA) which is later translated into protein.
Transcription in a bit more detail

- RNA polymerase binds to promoter DNA and moves a transcription bubble, like the slider of a zipper, which splits the double helix DNA molecule into two strands.
- RNA polymerase adds matching RNA nucleotides that are paired with complementary DNA nucleotides of the template DNA strand.
- RNA sugar-phosphate backbone forms with assistance from RNA polymerase to form an RNA strand.
- Hydrogen bonds of the untwisted RNA + DNA helix break, freeing the newly synthesized RNA strand.
- Most of the time, the RNA is further processed (with the addition of a 3'UTR poly-A tail and a 5'UTR cap) and exits to the cytoplasm through the nuclear pore complex.
- The stretch of DNA transcribed into an RNA molecule is called a transcription unit and encodes at least one gene. If the gene transcribed encodes a protein, the result of transcription is messenger RNA (mRNA), which will then be used to create that protein via the process of translation. Alternatively, the transcribed gene may encode for either non-coding RNA genes (such as microRNA, lincRNA, etc.) or ribosomal RNA (rRNA) or transfer RNA (tRNA),
Transcription in a bit less detail

Initiate

Elongate

Terminate
Transcription of RNA from DNA

- The bottom strand of the DNA molecule above is the template for RNA synthesis.
- RNA polymerase makes a copy of the DNA sequence but substitutes uridine (U) in place of thymine (T).

The bottom strand of the DNA duplex is used as the template to synthesize RNA. However, the sequence of bases in the RNA is the same as in the top strand of the DNA, with U in place of T.
mRNA structure

The structure of a typical human protein coding mRNA including the untranslated regions (UTRs)

orientation 5’ to 3’

UTR = untranslated region: mRNA stability
mRNA localization
translational efficiency
Translation: mRNA to protein via ribosome & tRNA

- Ribosome attaches to mRNA
- tRNAs carry specific amino acids into the ribosome
- Specific amino acids are added to the chain according to mRNA codons.
- Moves along mRNA building the protein (sequence of amino acids)
Translation: mRNA to protein via ribosome & tRNA
Cartoon summary of transcription and translation

Image adapted from: National Human Genome Research Institute.

Figure http://www.genomebc.ca/education/articles/transcription-and-translation-overview/
Amino acids are encoded by triples of mRNA nucleotides called **codons**

### The Genetic Code

<table>
<thead>
<tr>
<th></th>
<th>U</th>
<th>C</th>
<th>A</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td>UUU</td>
<td>UUC</td>
<td>UAU</td>
<td>UGU</td>
</tr>
<tr>
<td></td>
<td>Phenylalanine</td>
<td>Serine</td>
<td>Stop</td>
<td>Cysteine</td>
</tr>
<tr>
<td></td>
<td>Leucine</td>
<td></td>
<td></td>
<td>Stop</td>
</tr>
<tr>
<td></td>
<td>Leucine</td>
<td>Proline</td>
<td></td>
<td>Tryptophan</td>
</tr>
<tr>
<td>C</td>
<td>CUU</td>
<td>CCC</td>
<td>CAU</td>
<td>CGU</td>
</tr>
<tr>
<td></td>
<td>Leucine</td>
<td>Proline</td>
<td>Histidine</td>
<td>Arginine</td>
</tr>
<tr>
<td></td>
<td>Leucine</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>AUG</td>
<td>GAC</td>
<td>GAA</td>
<td>GGU</td>
</tr>
<tr>
<td></td>
<td>Methionine</td>
<td>Asparagine</td>
<td>Serine</td>
<td>Glycine</td>
</tr>
<tr>
<td></td>
<td>Serine</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>GUU</td>
<td>GCC</td>
<td>GUA</td>
<td>GUU</td>
</tr>
<tr>
<td></td>
<td>Valine</td>
<td>Alanine</td>
<td>Aspartic acid</td>
<td>Glycine</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: The table lists the amino acids corresponding to each codon.
Met
Thr
Glu
Leu
Arg
Ser
stop
Evolving definition of a gene

Mark B. Gerstein et al. Genome Res. 2007; 17: 669-681
Survey

- Name
- Department
- Background – undergraduate degree, current program, RA?
- Interests?
- Currently have a research project?
- Would you like to work alone or in a group? This request will be met.
Common Expression Data Structure

- \( g = 1, 2, \ldots, m \) genes (transcripts)
- \( k = 1, 2, \ldots, K \) conditions
- \( j = 1, 2, \ldots, j_k \) replicates
- \( t = 1, 2, \ldots, T \) time points

Can \( n \) be decreased if the number of subjects is increased?

\[ \mu_{g_1} = \mu_{g_2} \forall g, t ? \]