CSE280a: Algorithms for genetics

Vineet Bafna

CSE280A
The scope/syllabus

- We will cover topics from Population Genetics.
- The focus will be on the use of **algorithms** and statistical methods for analyzing genetic data
  - Some background in algorithms (mathematical maturity) is helpful.
  - Background in basic probability and statistics.
  - Relevant biology will be discussed on a ‘need to know’ basis (We’ll cover some basics today).
The Diploid Genome Sequence of an Individual Human

When in Doubt, Spit It Out

A DATE WITH DNA K. C. Dustin and his wife, Debra Netschert, give saliva samples.

By ALLEN SALKIN
Published: September 12, 2008

it has sequenced the complete genome of an African man.
From an individual to a population

- It took a long time (10-15 yrs) to produce the draft sequence of the human genome.
- Soon (within 10-15 years), entire populations can have their DNA sequenced. Why do we care?
- Individual genomes vary by about 1 in 1000bp.
- These small variations account for significant phenotype differences.
  - Disease susceptibility.
  - Response to drugs
Population Genetics

- Individuals in a species (population) are phenotypically different.
- Often these differences are inherited (genetic). Understanding the genetic basis of these differences is a key challenge of biology!
- The analysis of these differences involves many interesting algorithmic questions.
- We will use these questions to illustrate algorithmic principles, and use algorithms to interpret genetic data.
Logistics

- Most communication is electronic.
  - Check the class website.
- All reading is available online.
- 5-6 assignments, (50%), 1 Final Exam (20%), and 1 research project (30%)
- Project Goal:
  - get started on research or complete an MS exam requirement.
  - the project may require development of a new tool analysis of genetic data or both.
Scope of genetics lectures

- Basic terminology
- Key principles
  - Sources of variation
  - HW equilibrium
  - Linkage
  - Coalescent theory
  - Recombination/Ancestral Recombination Graph
  - Haplotypes/Haplotype phasing
  - Population sub-structure
  - Demographics
  - Structural polymorphisms
  - Medical genetics basis: Association mapping/pedigree analysis
Why DNA?
Bodies to cells
Griffith’s experiment established a transforming principle

- Live S type bacteria could be isolated from the mix, and used to infect other cells. Until now, it was assumed that different types stayed fixed from generation to generation.
- Griffiths’ experiment suggested that bacteria could be transformed.
The Avery-MacLeod-McCarty experiment

• Avery et al. did the following:
  • Lysed dead S-cells to isolate components: sugar, proteins, DNA, and RNA.
  • Used enzymes to remove sugars, proteins, RNA, and DNA.
  • DNA was the only necessary ingredient for transformation.
DNA as a universal mechanism of heredity

- Proteins are the machines in the molecular factory of the cell.
  - Enzymatic reactions
  - Cell growth, division
  - Switching the production of other proteins on or off
  - Carrying a signal from one location to another
- Ditto for RNA. RNA also helps in the production of protein
- DNA is the inherited material, and must contain instructions for manufacturing all the proteins, as well as the location of the regulatory switch.
  - Variation in DNA → possible change in instructions → possible change in molecular machinery
The double helix

- Levene found that all DNA had repeated units of Phosphate linked to deoxyribose sugar, linked to a nitrogenous base.
- He assumed that it contained repeating units of all bases, but that could not be logically correct.
- His hypothesis was refuted by Chargaff who found that the different bases appear in different proportions, but A matches T and C matches G.
Watson and Crick’s structure

- Two antiparallel strands, each forming a helix
- Complementary nucleotides form hydrogen bonds satisfying Chargaff’s rules
- One strand contains all of the information. The other is complementary
DNA as an information carrier

- We can think of the strand as a long string of nucleotide bases (A,C,G,T)
- Specific substrings encode information about how to make proteins
- The two complementary strands allow DNA to replicate and cells to make copies of each other
Variation in DNA

- The DNA is inherited by the child from its parent.
- The copying is not identical, but might be mutated.
- If the mutation lies in a gene,....
  - Different proteins are produced
  - Different proteins are switched on or off
  - Different phenotype.
• Darwin suggested the following:
  - Organisms compete for finite resources.
  - Organisms with favorable mutations are more likely to reproduce, leading to fixation of favorable mutations (Natural selection).
  - Over time, the accumulation of many changes suitable to an environment leads to speciation.

• Kimura (1960s) observed
  - Most mutations are selectively neutral!
  - They drift in the population, eventually getting eliminated, or fixed by random chance.
This class

- We seek to separate the two types of mutations.
- Identifying mutations under selection is important for identifying the genetic basis of phenotypes.
- Neutral mutations help in reconstructing evolutionary scenarios and providing a baseline for parameter calculations.
(Comparative) Genomics vs. (Population) Genetics

- Mutations accumulate over time
- In looking at the DNA of different species
  - DNA has had a lot of time to mutate, and is not expected to be identical.
  - If the DNA is highly similar, that region is functionally important.
- In comparing DNA from a population
  - Not enough time to mutate, so DNA is expected to be identical.
  - The few differences that exist mediate phenotypic differences.
Population genetics

- By sampling DNA from a population, we can answer the following
  - What are the sources of variation?
  - As mutations arise, they are either neutral and subject to evolutionary drift, or they are (dis-)advantageous and under selective pressure. Can we tell?
  - If you had DNA from many sub-populations, Asian, European, African, can you separate them?
  - Why are some people more likely to get a disease than others? How is disease gene mapping done?
  - Phasing of chromosomes. How do we separate the maternal and paternal chromosomes
Terminology: allele

- **Allele**: A specific variant at a location
  - The notion of alleles predates the concept of gene, and DNA.
  - Initially, alleles referred to variants that described a measurable trait (round/wrinkled seed)
  - Now, an allele might be a nucleotide on a chromosome, with no measurable phenotype.
  - As we discuss source of variation, we will have different kinds of alleles.
**Locus**: The location of the **allele**
- A nucleotide position.
- A genetic marker
- A gene
- A chromosomal segment
Terminology

• **Genotype**: genetic makeup of (part of) an individual
• **Phenotype**: A measurable trait in an organism, often the consequence of a genetic variation
• Humans are **diploid**, they have 2 copies of each chromosome, and 2 alleles at each locus
  - They may have **heterozygosity/homozygosity** at a location
  - Other organisms (plants) have higher forms of **ploidy**.
  - Additionally, some sites might have 2 allelic forms, or even many allelic forms.
• **Haplotype**: genetic makeup of (part of) a single chromosome
What causes variation in a population?

- Mutations (may lead to SNPs)
- Recombinations
- Other crossover events (gene conversion)
- Structural Polymorphisms
- Small mutations that are sustained in a population are called SNPs
- The data is a matrix (rows are individuals, columns are loci).
SNP Matrix

- **Discard non-polymorphic sites**
- **Infinite sites assumption**: each site mutates at most once in history of the population.
- Often, 0 refers to ancestral, 1 to mutant allele

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>ATCGATAACTGGAGAAGCGAGGATTCTGAACCATTGG</td>
<td>000000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A2</td>
<td>ATCGATAACTGGAGAGCGAGGATTCTGAACCTATTGG</td>
<td>001001</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A3</td>
<td>ATCGATAACTGGAGAACCAGGAGGATTCTGAACCATTGG</td>
<td>110011</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A4</td>
<td>ATCGATAACTGGAGAACCAGGAGGATTCTGAACCATTGG</td>
<td>000111</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A5</td>
<td>ATCGATAACTGGAGAACCAGGAGGATTCTGAACCATTGG</td>
<td>111110</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A6</td>
<td>ATCGATAACTGGAGAACCAGGAGGATTCTGAACCATTGG</td>
<td>110111</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A7</td>
<td>ATCGATAACTGGAGAACCAGGAGGATTCTGAACCATTGG</td>
<td>011000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

CSE280

Vineet Bafna
Microsatellites/VNTRs/Short Tandem Repeats

- 2-13 nucleotides repeated \( \sim 10^2 \) times.
- Stutter errors cause variation during replication

Wiki: "STR-Slippage Dr.Peter Forster"
Short Tandem Repeats

- Polymorphism due to STRs

```
GCTAGATCATCATCATCATCATCATATTGCTAG
GCTAGATCATCATCATCATCATCATATTGCTAGTTA
GCTAGATCATCATCATCATCATCATCATATTGCTAG
GCTAGATCATCATCATCATCATCATATTGCTAGTTA
GCTAGATCATCATCATCATCATCATCATATTGCTAG
GCTAGATCATCATCATCATCATCATCATATTGCTAGTTA
```

- 4
- 3
- 5
- 3
- 5
STR can be used as a DNA fingerprint

- Consider a collection of regions with variable length repeats.
- Variable length repeats will lead to variable length DNA.
- The locations are far enough apart not to be linked.
- Vector of lengths is a fingerprint.

<table>
<thead>
<tr>
<th>individuals</th>
<th>loci</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
</tr>
</tbody>
</table>
Large scale structural changes (deletions/insertions/inversions) may occur in a population.
Copy Number variation
Certain diseases (cancers) are marked by an abundance of these events
Personalized genome sequencing

- These variants (of which 1,288,319 were novel) included
  - 3,213,401 single nucleotide polymorphisms (SNPs),
  - 53,823 block substitutions (2–206 bp),
  - 292,102 heterozygous insertion/deletion events (indels)(1–571 bp),
  - 559,473 homozygous indels (1–82,711 bp),
  - 90 inversions, as well as numerous segmental duplications and copy number variation regions.
- Non-SNP DNA variation accounts for 22% of all events identified in the donor, however they involve 74% of all variant bases. This suggests an important role for non-SNP genetic alterations in defining the diploid genome structure.
- Moreover, 44% of genes were heterozygous for one or more variants.
Human DNA

http://upload.wikimedia.org/wikipedia/commons/b/b2/Karyotype.png
Variation due to recombination

- Not all DNA recombines!
Human DNA

- Not all DNA recombines.
- mtDNA is inherited from the mother, and
- y-chromosome from the father

CSE280 Vineet Bafna
http://upload.wikimedia.org/wikipedia/commons/b/b2/Karyotype.png
Gene Conversion

- Gene Conversion versus single crossover
  - Hard to distinguish in a population
Quiz

- Allele
- Locus
- Recombination
- Mutation/Single nucleotide polymorphism
- STR (short tandem repeat)
  - How is DNA fingerprinting done
- Infinite sites assumption
A quick tour. How to identify the genetic basis of a phenotype via association
Abstraction of a causal mutation
A possible strategy is to collect cases (affected) and control individuals, and look for a mutation that consistently separates the two classes. Next, identify the gene.
Problem 1: many unrelated common mutations, around one every 1000bp
Case Control

Case 🙁

Control 😊
Problem 2: We may not sample the causal mutation.
How to hunt for disease genes

- We are guided by two simple facts governing these mutations
  1. Nearby mutations are correlated
  2. Distal mutations are not
Sample a population of individuals at variant locations across the genome. Typically, these variants are single nucleotide polymorphisms (SNPs).

Create a new bi-allelic variant corresponding to cases and controls, and test for correlations.

By our assumptions, only the proximal variants will be correlated.

Investigate genes near the correlated variants.
So, why should the proximal SNPs be correlated, and distal SNPs not?
Consider a fixed population (of chromosomes) evolving in time.
Each individual arises from a unique, randomly chosen parent from the previous generation.
(a) Genealogy of a chromosomal population
Infinite sites assumption: A mutation occurs at most once at a site.
The collection of acquired mutations in the extant population describe the SNPs
(b) Mutations: drift, fixation and elimination
(c) Removing extinct genealogies
Removing fixed mutations
The coalescent
• We drop the ancestral chromosomes, and place the mutations on the internal branches.
- A causal mutation creates a clade of affected descendants.
• Note that the tree (genealogy) is hidden.
• However, the underlying tree topology introduces a correlation between each pair of SNPs
(d) Causal, and correlated mutations
Recombination

Synapsis: Pairing of homologous chromosomes

Paternal  Maternal

Crossing over
Synapsis: Pairing of homologous chromosomes

a.

b.

c.
In our idealized model, we assume that each individual chromosome chooses two parental chromosomes from the previous generation.
Multiple recombination change the local genealogy
A bit of evolution

- Proximal SNPs are correlated, distal SNPs are not.
- The correlation (Linkage disequilibirium) decays rapidly after 20-50kb
Association mapping basics

- Test each polymorphic locus for correlation with case-control status.
- The correlation is measured using one of many statistical tests.
- Gene near a correlated locus is a candidate for mediating the case phenotype.
- Many factors confound the analysis:
  - Even the sources of variation are not well understood.
  - Understanding the confounding factors requires a knowledge of population genetics.
  - Getting around them requires a set of computational and statistical techniques.
Scope of genetics lectures

- Our focus will be on the computational algorithms used to probe genetic data
- Basic terminology
- Key principles
  - Sources of variation
  - HW equilibrium
  - Linkage (dis)-equilibrium
  - Coalescent theory
  - Recombination/Ancestral Recombination Graph
  - Haplotypes/Haplotype phasing
  - Population sub-structure
  - Structural polymorphisms
  - Medical genetics basis: Association mapping/pedigree analysis