Genomics
Genomics

• Statistical and Population Genomics:
  • Studying stochastic processes that shape DNA-sequence variation
  • Using DNA data to understand history, natural selection, and disease
Genomics example: DNA and geography

- **Matrix** $X$ ($m \times n$), contains data on $m$ mutations in $\{0,1\}$ for $n$ individuals

$$
X = \begin{bmatrix}
000000000000000000001 \\
10101001010001000101 \\
101010001010010000001 \\
000000000000000000001 \\
101010001010000000001 \\
000000000000100000000 \\
010001110101110011000 \\
010000000000000000110 \\
101010001010001000100 \\
101010001010011000101 \\
010110011011001100110 \\
10101001010001000101 \\
1010100010100100100100 \\
101010001010011000101 \\
010111011001101100110 \\
010111011001101100110 \\
010111011001101100110
\end{bmatrix}
$$

Mutation on chrom. 1, position 1021002

A genome
Genomics example: DNA and geography

- Matrix $X$ (mxn), contains data on $m$ mutations in $\{0,1\}$ for $n$ individuals
- Compute *eigen-decomposition* of $XX^T$

- This will give you *principal components* of $X$
- Individuals now described using some (e.g. 2) principal components

- Let’s do this with genomes from Europe...
Genomics example: DNA and geography

“Genes mirror geography within Europe”, Novembre et al. 2008 Nature
Genomics: models of ancestry

O = individual
Genomics example: fine-scale population structure

People of the British Isles (PoBI) data, Nature 2015

Genome of the Netherlands (GoNL) data, Nature Genetics, 2014
Genomics example: fine-scale population structure

https://ukancestrymap.github.io
Genomics example: identifying disease mutations

Genes involved in coronary artery disease (CAD), Nature Genetics, 2015
Genomics

• No need to know biology!
Relevant courses

Machine Learning:
- SB2a Foundations of Statistical Inference
- SB2b Statistical Machine Learning
- A12 Simulation and Statistical Programming
- SB1 Applied and Computational Statistics
- SC4 Advanced Topics in Statistical Machine Learning
- SC5 Advanced Simulation Methods
- SC6 Graphical Models
- SC7 Bayes Methods

Genomics:
- SB3a Applied Probability
- SC1 Stochastic Models in Mathematical Genetics
- SC2 Probability and Statistics for Network Analysis
- SC8 Topics in Computational Biology
Options in probability and related areas

James Martin

Thursday 15:00, Week 5, MT 2019
Simple random walk

Let \( p \in (0, 1) \). Let \( X_1, X_2, X_3, \ldots \) be i.i.d. with

\[
P(X_i = 1) = p, \quad P(X_i = -1) = 1 - p.
\]

Then \( \mathbb{E} X_i = \mu = 2p - 1, \text{Var} X_i = \sigma^2 = 4p(1 - p) \).

Let \( S_n = X_1 + X_2 + \cdots + X_n \) (position of random walk at step \( n \)).

Law of large numbers:

\[
\left\lfloor \frac{S_n}{n} \approx \mu \right\rfloor
\]

Central limit theorem:

\[
\left\lfloor \frac{S_n/n - \mu}{\sigma/\sqrt{n}} \approx \mathcal{N}(0, 1) \right\rfloor
\]
Simple random walk, \( n=100 \) \( p=0.75 \)
Simple random walk, n=10000 p=0.75
Simple random walk, n=10000 p=0.25
Simple random walk, $n=10000$ $p=0.5$
Simple random walk, n=10000 p=0.5
Percolation

Two-dimensional lattice $\mathbb{Z}^2$. Let $p \in (0, 1)$.

Let each edge of the lattice be open with probability $p$ and closed with probability $1 - p$, independently for different edges.

What does the connected component $C(v)$ of a given vertex $v \in \mathbb{Z}^2$ look like? (i.e. the set of vertices reachable from $v$ by following open edges)
Theorem

(i) If $p \leq 1/2$ then with probability 1 there is no infinite cluster;
(ii) if $p > 1/2$ then with probability 1 there is an infinite cluster (which is unique).
Critical behaviour

“Critical point” $p_c = 1/2$: fractal picture (“scale-invariant”).
Boundaries described by “Schramm-Loewner evolution” and related processes.
Predicted: “universal behaviour” – not dependent on particular lattice structure (but only known rigorously in certain very specific cases!)

Schramm, Lawler, Werner (Fields 2006), Smirnov (Fields 2010).
Critical exponents: size of largest cluster in box $[0, L]^2 \sim L^{91/48}$; $\mathbb{P}_{p_c}(|C(v)| = n) \sim n^{-96/91}$; $\mathbb{P}_{p_c}(v \leftrightarrow w) \sim |v - w|^{-5/24}$...
Part A:
- A8 Probability, A9 Statistics, A4 Integration
- A8 Graph Theory (8 lectures), A12 Simulation and Statistical Programming

Part B:
- SB3a Applied Probability
- B8.1 Probability, Measure and Martingales
- B8.2 Continuous Martingales and Stochastic Calculus
- B8.4 Information Theory, B8.3 Mathematical Models of Financial Derivatives, B5.1 Stochastic Modelling of Biological Processes, B8.5 Graph Theory, B4.1/4.2 Functional Analysis, ...

Part C:
- C8.1 Stochastic Differential Equations
- C8.2 Stochastic Analysis and PDEs
- SC9 Probability on Graphs and Lattices
- C8.4 Probabilistic Combinatorics
- C8.5 Introduction to Schramm-Loewner Evolution
- SC1 Stochastic Models in Mathematical Genetics
- SC2 Probability and Statistics for Network Analysis
- Many other options in Statistics, Analysis, Combinatorics, ....