CM226 Fall 2021 Projects

Estimating (calling) variants from sequencing data

Possible projects

Many statistical models and more recently supervised machine learning

See: https://www.nature.com/articles/nbt.4235

Calling SNP genotypes versus insertions or deletions (Indels)

Using data from multiple individuals within the same population, multiple individuals from different populations

Genotype imputation

SNP genotype data is missing. Can we use patterns of correlation (LD) to impute missing data.

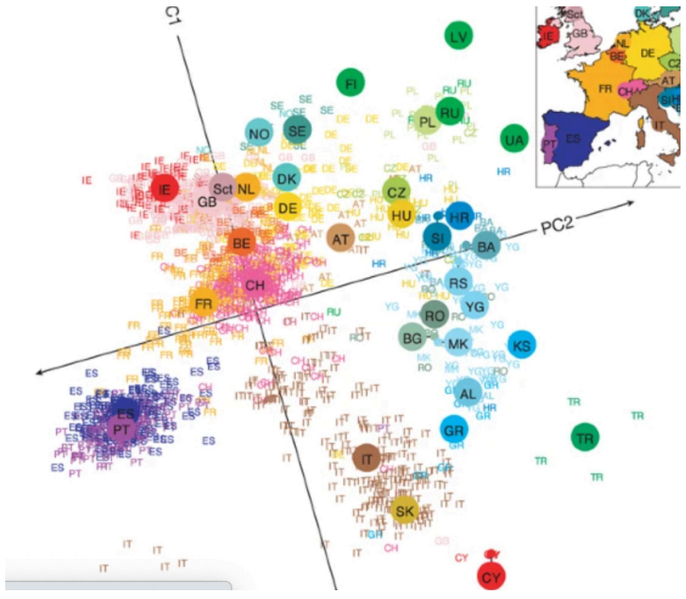
Many statistical models (HMMs, Matrix completion). What about deep learning ?

References: https://genome.cshlp.org/content/23/3/509.full.pdf

Data: 1000 Genomes project, Dataset 2 in google docs

Ancestry inference

Identify population of origin of a given individual based on their genome



https://www.nature.com/articles/ng.2285

Data: Dataset 2 in google doc, other public datasets

Local ancestry inference

Identify ancestry of admixed individuals (chromosome painting)

See 1000 Genomes project

Predict disease risk from genome sequence data

Given genome sequence (or genotype data) and disease status, learn a predictor that is accurate on new individual.

See Dataset 6 in google doc

Predict disease risk from genome sequence data

Given genome sequence (or genotype data) and disease status, learn a predictor that is accurate on new individual.

Variations: What if the individual is from a different population than the training data ?

What if individual data is not available but summary data is available ?

Genomic privacy

How many SNPs need to be made public to identify an individual in a database ?

How many SNPs with summary data need to be made public to predict an individual's phenotype ?

What if multiple phenotypes are available ?

Single-cell data

New technologies to assess gene and other cellular measurements in single cells.

- How to combine different datasets while avoiding confounding?
- How to fill in missing entries in this data ?
- How to predict biological states (disease or cell type)?
- How to perform better dimensionality reduction and visualization ?

See datasets 7 and 20

Microbiome/Metagenomics

We can now sequence microbes from different environments (including human tissues and body sites). Using microbial compositions to understand health vs disease.

Human microbiome project. See Dataset 1

Interpretable Machine Learning

A number of methods have been developed to interpret the predictions of black box deep learning algorithms.

https://arxiv.org/abs/1705.07874

It is unclear if this and/or methods solve the problem of interpretability. Critically study methods for interpretability.

COVID projects

Both viral genomic data and case data are available. Use these to understand transmission dynamics.

Datasets 3, 5 (SARS data), 8, 11.

For viral genomes, you need to apply to gisaid.org (takes 1-2 days to get approved).

Forecasting COVID-19 trajectory from case and death data

Goal: Answer “Will there be a surge?” or “When will social distancing end?”

Combine machine learning with epidemiological models.

Challenges: Need to incorporate sources of uncertainty and bias in the data.

https://www.medrxiv.org/content/10.1101/2020.05.30.20117796v2

Estimating prevalence of SARS-CoV-2 from testing data

Different types of tests (serology, PCR) have different error rates.

Given the error rates, how can we estimate the true prevalence of COVID-19?

Can we combine different tests to get improved estimates ?

https://covidtestingproject.org/about.html

Estimating SARS-CoV-2 spread from viral genome sequences

More than 10,000 viral strains sequenced.

Building evolutionary trees on this scale.

What mutations are key to the viral evolution (under selection and changing rapidly or highly conserved)?

https://www.gisaid.org

Predicting disease progression from image data

See datasets 9, 10, 12, 15, 18, 23, 24.

More generally, the MIMIC dataset has a number of clinical variables from a large cohort of patients. You will need to apply to get access.

Predicting DNA-DNA interaction

Gene regulation often occurs at a distance where a gene is regulated by genes located far away. Can we understand the factors that influence this interaction ?

Dataset 21, 19

Protein structure prediction

Predicting the 2D and 3D structure of proteins from its sequence is a major challenge in biology. Recently, deep learning methods have made major advances on this problem.

Dataset 16

Low-dimensional representation of gene expression data

Gene expression data is high-dimensional. Many methods exist to compress the data to a lower dimension for visualization and denoising. Which methods are most accurate and useful for downstream prediction?

Dataset 14.

Predicting the impact of cancer mutations

Identifying the clinical impact of mutations is a major open problem and is important in designing treaments.

Dataset 13

Others

Your own projects

DREAM competitions: http://dreamchallenges.org/

Additional projects (and open questions during the course)