High Dimensional Inference and Genomics

Sayan Mukherjee
Department of Statistical Science
Institute for Genome Sciences & Policy
Dept. of Computer Science
Duke University
Genomics

Genetics -- study of genes through variation.

Genomics -- study of the variation in genomes and their related products.

Types of genomic data
Sequence variation
Gene expression
Transcription factors
Epigenetics
High-dimensional data

Number of genes that express: ~30,000

Number of nucleotides: 3.2 billion

Number of SNPs: 1.42 million

Number of transcription factors: ~2,600 plus combinations

Number of epigenetic elements: ??????
Simulate network – forward problem

From Yalamanchili et al
Infer network – inverse problem

Given observations of genomic variables, different environments, and different phenotypes

Give us back
Prediction

Accuracy 85%

Accuracy 61%
Modeling variation

Wright 1921
Prior knowledge
Representation

Yeast: beer vs. bread, 2000 genes, gene network
Representation

Progression in prostate cancer, pathway networks
Uncertainty

Progression in prostate cancer, pathway networks
Computation

Number of variables: \( d = \text{Ginormous} \)

Size of model space: \( p = d^n \)

Search or sample models over this space:
  Monte Carlo from hell
  Variational methods will break

Information = uncertainty = computational complexity