CSCI2951-N: Advanced Algorithms in Computational Biology

http://www.cs.brown.edu/courses/csci2951-n/

Tuesdays and Thursdays 2:30-3:50 in the SWIG CIT 241

Prof. Sorin Istrail

- **Maximum Likelihood and Expectation-Maximization Algorithms**
  polynomial likelihood functions, Q functions, symmetries of likelihoods
  Biological Problem: Inferring haplotype frequencies in populations.
- **Set Cover Algorithms and Minimum Informative Subset**
  dominating sets, fixed parameter tractable algorithms, information theory
  Biological Problem: Tagging SNPs selection, LD.
- **Markov Chain Monte Carlo Algorithms**
  Metropolis algorithm, law of large numbers and sampling
  Biological Problem: Population Substructure
- **Knapsack Algorithms and Statistical Hypothesis Testing**
  the Neyman-Pearson lemma, multiple testing
  Biological Problem: Statistical Associations in GWAS
- **Graph Theory Algorithms**
  cycle basis of graphs, suffix-trees, graph coloring
  Biological Problem: Haplotype Reconstruction from next generation sequencing
- **Voting Theory Algorithms**
  social networks, Arrow paradox, von Neumann-Morgenstern utility theory
  Biological Problem: Protein Folding energy function inference

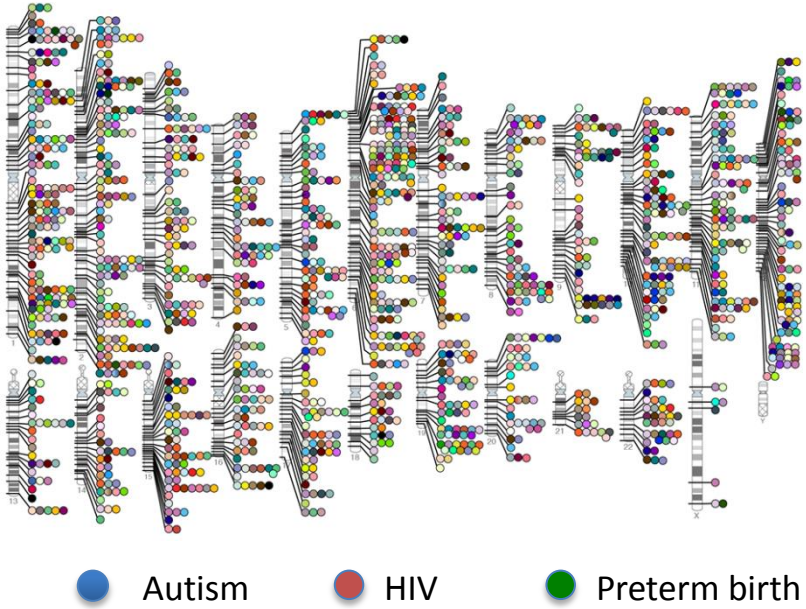
Algorithms

Set Cover, **Dominating Set**, **Knapsack** are classic NP-Complete problems

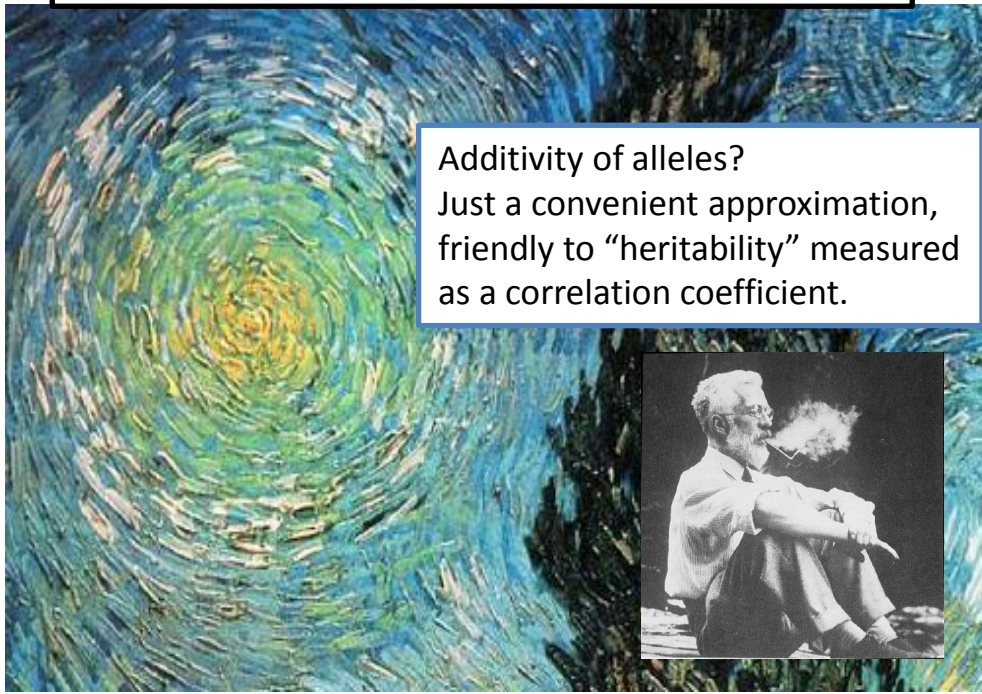
Haplotypes Reconstruction

Published Genome-Wide Associations through 2011
1,617 published GWA at $p \leq 5X10^{-8}$ for 249 traits

The Genome-Wide Association Studies (GWAS) Human Genome



The Missing Heritability Puzzle



Additivity of alleles?
Just a convenient approximation, friendly to "heritability" measured as a correlation coefficient.

Sir Ronald Fisher

CSCI2951-N: Advanced Algorithms in Computational Biology
http://www.cs.brown.edu/courses/csci2951-n/

What are the Genetic Determinants of Disease?

The Needles in the Haystack

The CDCV, a Rembrandt-like drawing metaphor, with few identical needles in a haystack, needs to be replaced now with a van Gogh-like drawing metaphor, with many needles each differently looking and private to areas in the haystack.

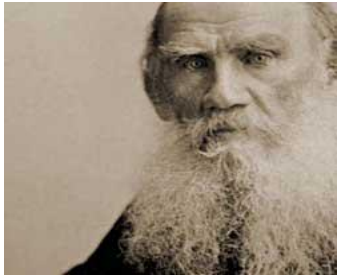
The Common Disease Common Variant (CDCV) Hypothesis is dead.

Long live the Common Disease Rare Variant Hypothesis!

Rembrandt van Rijn **A HAYSTACK NEAR A FARM** (1650)

Genetic Heterogeneity

"All happy families are alike; each unhappy family is unhappy in its own way."
Leo Tolstoy – *Anna Karenina*



Vincent van Gogh **NOON-REST FROM WORK** (1890)

Protein Folding

Social Choice Theory and the Thermodynamic Hypothesis

C. Anfinsen
Nobel Laureate
Chemistry (1972)

In the social network of amino acids in protein structures how do spatial pairwise preferences (individual values) can be aggregated to a universal energy function (social choice)?

Anfinsen's Hypothesis: There exist an universal energy function:
"These results suggest that the native molecule is the most stable configuration, ***thermodynamically speaking,*** and that the major force in the correct pairing of sulfhydryl groups in disulfide linkage is the concerted interaction of side-chain functional groups distributed along the primary sequence."

Arrow's General Impossibility Theorem:
It is impossible to formulate a social preference ordering that satisfies all of the following conditions:
Non-dictatorship: The preferences of an individual should not become the group ranking without considering the preferences of others.
Individual Sovereignty: each individual should be able to order the choices in any way and indicate ties
Unanimity: If every individual prefers one choice to another, then the group ranking should do the same
Freedom From Irrelevant Alternatives: If a choice is removed, then the others' order should not change
Uniqueness of Group Rank: The method should yield the same result whenever applied to a set of preferences. The group ranking should be transitive.

K. Arrow
Nobel Laureate
Economics (1952)