Medical Bioinformatics (CSCI295-L)

Genome-Wide Association Studies, Protein Folding and Immunogenomics

Sorin Istrail

November 2010

SYLLABUS

Contents

1 Short Introductions

Class 1

Topics covered:

1. The Course Sections
    (a) Genome-Wide Association Studies (GWAS)
    (b) Immunogenomics
    (c) Protein Folding
2. SNPs and the Human Genome: SNPs, haplotypes, mutation, recombination, genetic determinants of disease
3. SNP Genetic Variation: A look at the data
4. Fundamental Challenges: (1) Very limited success so far regarding the knowledge of the human genome sequence to the clinical impact ; and (2) Causality of disease mechanism may be unknowable

Class 2

- Topics covered:
    1. Introduction to Immunogenomics: proteomes, peptides, epitopes, T-cells, MHC, T-cell vaccines
    2. Introduction to Association Studies: haplotypes blocks, tagging SNPs: block definitions (e.g., 4-gametes block test), the Hudson-Kaplan 4-gametes block decomposition, minimum informative subset of tagging SNPs via Informativeness, graph theory modeling using the minimum set cover problem
- Class readings:
    - Guilt by Association
    - Viruses in the Sea

Class 3

- Topics covered:
    1. Introduction to Protein Folding: the comparative genomics landscape (1-dimensional = biomolecular sequence analysis, DNA, protein, regulatory regions, SNPs, haplotypes, genome assembly; 3-dimensional = protein structure; logic/chemistry ”dimension” = expression, networks), self-avidong walks, contacts, statistical mechanics, protein structure alignment, fold recognition, the Computational Protein Folding Competition (CASP= Critical Assessment of Structure Prediction), contact maps, contact map overlap structure alignment, fold recognition

2. Introduction to Genome-Wide Association Studies (GWAS): 17 caveats

- Class readings:
  - Drinking from the Fire Hose – Statistical Issues in GWAS

2 The Hardy-Weinberg Model

Class 4

- Topics covered:
  1. Population Genetics Models: population, diploid organism, random mating, non-overlapping generations, genotype-gamete-genotype, genotype frequency, gene frequency
  2. The Hardy-Weinberg model for one locus: Hardy-Weinberg Equilibrium (HWE), Fundamental Theorem 1 (HWE attained in one generation), Fundamental Theorem 2 (constancy of allele frequencies in every generation), HW frequencies satisfy the $Y^2 = XZ$ equation, the Hardy=Weinberg Theorem (one variable x only: $x = X + Y$)

Class 5

- Topics covered:
  1. The Hardy-Weinberg model for two loci: Linkage equilibrium (LE), recombination rate r (do not confuse with r^2 the measure of LD - different r notations), linkage disequilibrium (LD); double heterozygotes, recombinant and non-recombinant gametes, Lewontin's linkage disequilibrium parameter D, single locus HWE = no evolution (constancy of frequencies), two loci HW = potential for evolution based on parameters D and r.

3 Linkage Disequilibrium and GWAS

Class 6

- Topics covered:
  1. The first measure of linkage disequilibrium: D. The D equation, D' and the independence of allele frequencies
  2. Examples: Evaluation of HWE at each locus, and of LD between them.
  3. The second measure of linkage disequilibrium: r^2.
  4. Complete LD, Perfect LD, Useful LD.

- Class readings:
  - A haplotype map of the human genome

Class 7

- Topics covered:
  1. Properties of the LD measure r^2: its correlation coefficient definition, r^2 as a statistical test, description of the loss in efficiency when one marker is replaced by another marker, its relation to χ^2, drawbacks (e.g., measure is pairwise and not clear how to extend to many sites)
  2. Genome-wide association studies: an outline of the methodology (genome is huge so "interesting" pattern occur by chance; HW, Tagging SNPs, Haplotype Phasing are computational methods in active research not yet consensus on how to use and apply them to data; the problem of missing data; case and control samples have different parameters and hard to match and contrast statistically and algorithmically for the most discrepancy (e.g., different rates of missing data); LD is a non-quantitative phenomenon, there is no natural scale for it); tagging SNPs are effective only for capturing common variants, and tagging on one population only poorly performs in another population; population substructure can generate spurious phenotype associations; to phase of not to phase - associations of unphased genotypes is limited but phasing with an adhoc choice of method adds uncertainty as well.
  3. An introduction to HapMap
- Class readings:
  - Linkage disequilibrium and the mapping of complex human traits

4 Tagging SNPs and GWAS

Class 8

- Topics covered:
  1. The architecture of linkage disequilibrium based haplotype blocks across chromosome 6, 21, 22
- Class readings:
  - Blocks of limited haplotype diversity revealed by high-resolution scanning of the human chromosome 21

Class 9

- Topics covered:

  1. Tagging SNPs and the Minimum Informative Subset of Tagging SNPs
  2. A set of desiderata (axioms) for tagging SNPs selection:
     (a) To be extendable uniquely to multi markers
     (b) To be consistent with the LD measures
     (c) To be haplotype block free (no adhoc definition of block required)
     (d) To be hypothesis free (e.g., not specific to a certain disease or trait)
     (e) To be algorithmically sound (practical for genome-wide data0
     (f) To be statistically sound (no overfitting)
  3. The Informativeness measure satisfies all the above desiderata
  4. The modeling of Informativeness via the Minimum Set Cover Problem
  5. A dynamic programming algorithm for the minimally informative K SNPs problem.

- Class readings:

  - Efficiency and power in genetic association studies
  - Selecting a maximally informative set of SNPs for association analyses using LD
  - The structure of the Haplotype Blocks in the Human Genome

5 Haplotype Phasing and GWAS

Class 10

- Topics covered:

  1. The Haplotype Phasing Problem: The Clark Method and the Maximum Likelihood Method: genotypes, haplotypes, genotype explanations
  2. The Clark Method: greedy algorithm, Clark rule, three difficulties.

- Class readings:

  - Optimal haplotype block free selection of tagging SNPs for GWAS (and two supplemental documents)
  - Multiple Sclerosis GWAS: Risk alleles for multiple sclerosis identified by a genomewide study

Class 11

- Topics covered:

   1. The Expectation-Maximization Algorithm of Excoffier-Slatkin for computing haplotype frequencies and haplotype phase: intro to maximum likelihood, the EM algorithm background, computing haplotype frequencies and genotype explanations frequencies iteratively; the EM algorithm for the haplotype frequencies, and its application to haplotype phasing

- Class readings:

   - Inference of haplotypes from PCR-amplified samples of diploid populations
   - Maximum-Likelihood estimation of molecular haplotype frequencies in a diploid population

Class 12

- Topics covered:

   1. Guest Lecturer: Jonathan Yewdell (NIH): Topics in Immunogenomics

- Class readings:

   - Identifying cytotoxic T-cell epitopes from genomic and proteomic information

6 GWAS studies in detail: Multiple Sclerosis and Autism

Class 13

- Topics covered:

   1. The GWAS for Multiple Sclerosis

- Class readings:

   - Istrail, Yewdell et. al ”The immunopeptidome of human and their pathogens”

7 Wright-Fisher, Infinite Allels and Urn Models

Class 14

- Topics covered:

    1. The Wright-Fisher Model and the Infinite Allele Model
    2. Urn Models for population genetics: Polya Urn models
    3. The Wright-Fisher Model: Mutation, Random Genetic Drift, Selection
    4. Wright-Fisher and Markov Chains
    5. The Infinite Allele Model
    6. Stationarity of partitions
    7. Ewens' Sampling Lemma

- Class readings:

    – Genetic mapping of human disease (and supplementary material)

Class 15

- Topics covered:

    1. The Urn model for the Infinite Allele model is the Hoppe's Urn model
    2. Computing Q probabilities for the Hoppe's urn model
    3. The Clark Algorithm revisited: the analysis of the three difficulties based on the Infinite Allele model theory

8 Computational Complexity of the Haplotype Phasing Problem

Class 16

- Topics covered:

    1. Computational Complexity of the Haplotype Phasing Problem
        – The Parsimony Haplotype Phasing is NP-complete – The Hubbell Theorem 1.
        – The Global Maximum Likelihood Haplotype Phasing is NP-complete – The Hubell Theorem 2.
        – The Clark Maximal Resolution Haplotype Phasing is NP-complete.

- Class readings:

    – E. Hubbell's manuscript, 2000

9 Tests of Association

Class 17

- Topics covered:

  1. Tests of Association
     - 2×2 contingency tables, the hypergeometric distribution and the Fisher's Exact Test
     - Contingency tables of arbitrary size: The Chi-Square Test (the historical test) and the Kullback-Leibler relative entropy test (the right test)
  2. The Coalescent - introduction
     - Coalescent with mutation
     - Gene trees vs allele/haplotype trees vs haplotype networks

10 Complex Disease and Heritability

Class 18

- Topics covered:

  1. Complex Disease and Heritability
     - The "triangle" of the etiology of disease: single gene, polygenic, environment
     - Dichotomous vs continuous traits
     - Polygenic theory: Heritability and Thresholds
     - Polygenic susceptibility to disease
     - Regression, covariance, regression slope, correlation coefficient, Fisher's paper on the correlation between relatives
     - Heritability as a correlation coefficient

Class 19

- Topics covered:

  1. Guest Lecturer: Sam Broder (Celera, former Director of the National Cancer Institute): "Towards Evidence-Based Medicine: The Human Genome: 3 Billion Letters of Code, But Whos Counting?" followed by a Sweat Box Session on The Missing Heritability Puzzle – "The Genome as a Teacher: In This Class, Does the Teacher Grade on a Curve?"

Class 20

- Topics covered:
  1. Broad-sense and narrow sense heritability and the ”missing heritability” puzzle; common and rare variants
  2. The Coalescent with recombination; ancestral recombination graphs (ARG)
- Class readings:
  - Jon Mclellan and Mary-Claire King ”Genetic heterogeneity in human disease,” 2010
  - Eichler et al ”Missing heritability and strategies for finding the underlying causes of complex disease” 2010

11 The Coalescent and GWAS

Class 21

- Topics covered:
  1. The Miniciello-Durbin ARG reconstruction algorithm
  2. Population Substructure
- Class readings:
  - Miniciello-Durbin ”Mapping trait loci by use of inferred ancestral recombination graphs”

12 Statistical Hypothesis Testing in GWAS

Class 22

- Topics covered:
  1. Classical hypothesis testing – the five steps and GWAS
  2. Statistical power
  3. Multiple testing
  4. The Transmission/Disequilibrium test statistic (TDT)
  5. The Cochran-Mantel-Haenzel test statistic

13 GWAS and the Missing Heritability Puzzle

Class 23

- Topics covered:
  1. Disease Models: Common Disease Common Variant
  2. Rare alleles: Common Disease - many rare variants
  3. Genetic heterogeneity in human disease
  4. Missing heritability and strategies for finding underlying causes of complex disease

14 Protein Folding and Drug Design

Class 24

- Topics covered:
  1. Protein Folding in Lattice Models: Ken Dill's HP-Model
- Class readings:
  - F. Lam and S.Istrail "Combinatorial algorithms for protein folding in lattice models: a survey of mathematical results"

Class 25

- Topics covered:
  1. Protein Folding in Lattice Models: Folding algorithms

Class 26

- Topics covered:
  1. The Medicinal Chemist Compound Tinkering Problem: chemical graph theory

Class 27

- Topics covered:
  1. Concepts of drug-likeness, and the Lipinski rule of five

15 Immunogenomics

Class 28

- Topics covered:
  1. Drug resistance, codon-bias, RNA and HIV