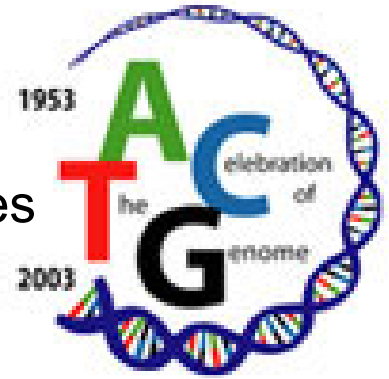
GENETIC DATA ANALYSIS

Genetic Data: Future of Personalized Healthcare

- To achieve personalization in Healthcare, there is a need for more advancements in the field of Genomics.
- The human genome is made up of DNA which consists of four different chemical building blocks (called bases and abbreviated A, T, C, and G).
- It contains 3 billion pairs of bases and the particular order of As, Ts, Cs, and Gs is extremely important. Size of a single human genome is about 3GB.
- Thanks to the Human Genome Project (1990-2003)
  - To determine the complete sequence of the DNA bases
  - The total cost was around $3 billion.



Genetic Data

- The whole genome sequencing data is currently being annotated and not many analytics have been applied so far since the data is relatively new.

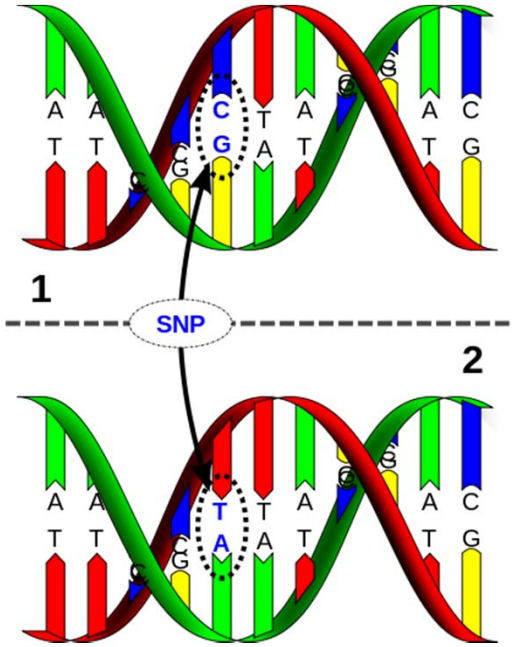
- Several publicly available genome repositories. http://aws.amazon.com/1000genomes/

- It costs around $5000 to get a complete genome. It is still in the research phase. Heavily used in the cancer biology.

- In this tutorial, we will focus on Genome-Wide Association Studies (GWAS).
  - It is more relevant to healthcare practice. Some clinical trials have already started using GWAS.
  - Most of the computing literature (in terms of analytics) is available for the GWAS. It is still in rudimentary stage for whole genome sequences.

Genome-Wide Association Studies (GWAS)

- Genome-wide association studies (GWAS) are used to identify common genetic factors that influence health and disease.
- These studies normally compare the DNA of two groups of participants: people with the disease (cases) and similar people without (controls). (One million Loci)
- Single nucleotide polymorphisms (SNPs) are DNA sequence variations that occur when a single nucleotide (A,T,C,or G) in the genome sequence differs between individuals.
- SNPs occur every 100 to 300 bases along the 3-billion-base human genome.



Important Computational Challenges in GWAS

- **Epistasis Modeling**
  - **GOAL:** To understand complex relationship between genotype and phenotype by identifying SNP-SNP interactionss.
  - **METHODS:** Exhaustive, Stochastic, and Heuristic.

- **High-dimensional SNP (Variable) Selection**
  - **GOAL:** To extract the SNPs that are significantly associated with the phenotype outcome. To obtain a set of reduced number of SNPs that are statistically significant to the genotype-phenotype relationship.
  - **METHODS:** Sparse Linear Methods and Random Forests.

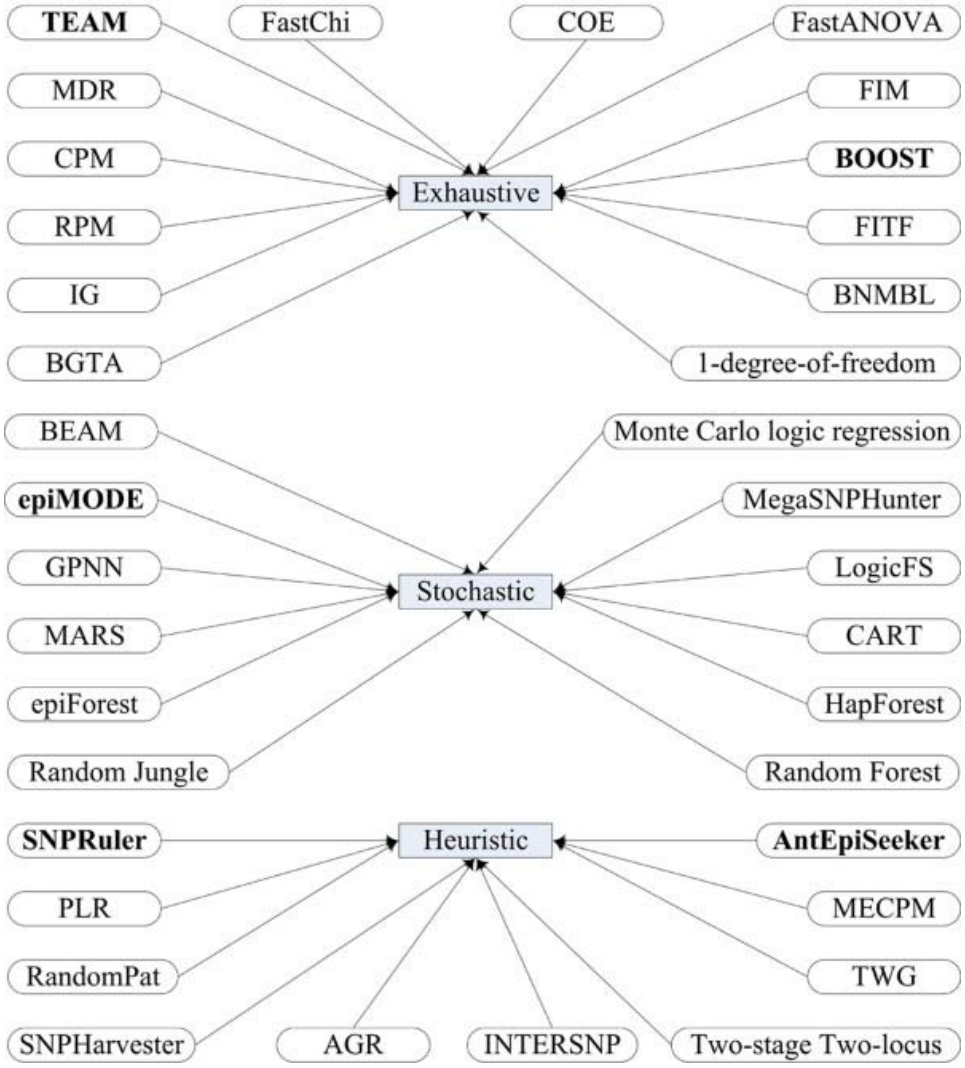
Epistasis Modeling

- For simple Mendelian diseases, single SNPs can explain phenotype very well.
- The complex relationship between genotype and phenotype is inadequately described by marginal effects of individual SNPs.
- Increasing empirical evidence suggests that interactions among loci contribute broadly to complex traits.

- The difficulty in the problem of detecting SNP pair interactions is the **heavy computational burden.**
  - To detect pairwise interactions from 300,000 SNPs genotyped in thousands of samples, a total of 4.5 X 10^{10} statistical tests are needed.
  - Since a huge number of possible combinations are tested, a large proportion of significant associations are expected to be false positives. Thus, reducing the number of false positives while retaining the significance power is another challenge.

Epistasis Detection Methods

- Exhaustive
  - Enumerates all *K*-locus interactions among SNPs.
  - Efficient implementations mostly aiming at reducing computations by eliminating unnecessary calculations.
- Non-Exhaustive
  - Stochastic: randomized search. Performance lowers when the # SNPs increase.
  - Heuristic: greedy methods that do not guarantee optimal solution.



Shang, Junliang, et al. "Performance analysis of novel methods for detecting epistasis." *BMC bioinformatics* 12.1 (2011): 475.

Variable Selection in SNP (High-Dimensional) Data

- In Genome-wide association study (GWAS), Single Nucleotide Polymorphism (SNPs) can be considered as features (usually in the range of hundreds of thousands).

- Not all the features are significantly associated with the phenotype outcome. A reduced number of features that are statistically significant might help us to better understand the genotype-phenotype relationship.

- A subset of features may produce predictive models with better accuracy. It removes the noisy, irrelevant and redundant features and finally, reduces the computational and memory usage complexity. Two popular methods are:
  - Sparse methods
  - Random Forests

Sparse Methods for SNP Data Analysis

- Successful identification of SNPs strongly predictive of disease promises a better understanding of the biological mechanisms underlying the disease.
- Scalability: The number of features is too high to be handled by traditional feature selection / ranking methods.

- Sparse linear methods have been used to fit the genotype data and obtain a selected set of SNPs.

- Minimizing the squared loss function (L) of N individuals and p variables (SNPs) is used for linear regression and is defined as

$$L(\beta_0, \beta) = \frac{1}{2}\sum_{i=1}^{N}(y_i - \beta_0 - x_i^T\beta)^2 + \lambda\sum_{j=1}^{p}\left|\beta_j\right|$$

where $x_i \in \mathbb{R}^p$ are inputs for the i^{th} sample, $y \in \mathbb{R}^N$ is the N vector of outputs, $\beta_0 \in \mathbb{R}$ is the intercept, $\beta \in \mathbb{R}^p$ is a p-vector of model weights, and λ is user penalty.

Packages for Lasso methods for SNP Data Analysis

- R package glmnet 1.7 with logistic loss (binomial family) implemented as a Fortran library
  - Link: http://www.jstatsoft.org/v33/i01/paper

- LIBLINEAR 1.8 [8][c], with ℓ_1-penalised squared hinge loss (model 5), implemented in C++
  - Link: http://www.csie.ntu.edu.tw/~cjlin/liblinear/

- SparSNP package http://bioinformatics.research.nicta.com.au/software/sparsnp

- HyperLasso , logistic regression with the double exponential (DE) prior (equivalent to lasso), implemented in C++. HyperLasso implements cyclical coordinate descent as well.
  - Software: http://www.ebi.ac.uk/projects/BARGEN/
  - Paper: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2464715/

Random Forests

- Feature importance measure is calculated by permuting each variable.

- Algorithm – High-level Description
  - Bootstrap sampling of the data.
  - Create trees with bootstrap samples.
  - For each tree
    - The classification error rate is calculated using out-of-bag (oob) samples.
    - For each variable in the tree, permute the variable values and compute the error rate. The increase in this value is a indication of the variable's importance.
  - Aggregate the error and importance measures from all trees to determine overall error rate and Variable Importance measures.

RESOURCES

Public Resources for Genetic (SNP) Data

- The Wellcome Trust Case Control Consortium (WTCCC) is a group of 50 research groups across the UK which was established in 2005.
- Data available at http://www.wtccc.org.uk/
- Seven different diseases: bipolar disorder (1868), coronary heart disease (1926), Crohn's disease (1748), hypertension (1952), rheumatoid arthritis (1860), type I diabetes (1963) or type II diabetes (1924).
- Around 3,000 healthy controls common for these disorders.
- The individuals were genotyped using Affymetrix chip and obtained approximately 500K SNPs.

- The database of Genotypes and Phenotypes (dbGaP) maintained by National Center of Biotechnology Information (NCBT) at NIH.
- Data available at http://www.ncbi.nlm.nih.gov/gap

Structured EHR Data Repositories

Dataset	Link	Description
Texas Hospital Inpatient Discharge	http://www.dshs.state.tx.us/thcic/hospitals/Inpatientpudf.shtm	Patient: hospital location, admission type/source, claims, admit day, age, icd9 codes + surgical codes
Framingham Health Care Data Set	http://www.framinghamheartstudy.org/share/index.html	Genetic dataset for cardiovascular disease
Medicare Basic Stand Alone Claim Public Use Files	http://resdac.advantagelabs.com/cms-data/files/bsa-puf	Inpatient, skilled nursing facility, outpatient, home health agency, hospice, carrier, durable medical equipment, prescription drug event, and chronic conditions on an aggregate level
VHA Medical SAS Datasets	http://www.virec.research.va.gov/MedSAS/Overview.htm	Patient care encounters primarily for Veterans: inpatient/outpatient data from VHA facilities
Nationwide Inpatient Sample	http://www.hcup-us.ahrq.gov/nisoverview.jsp	Discharge data from 1051 hospitals in 45 states with diagnosis, procedures, status, demographics, cost, length of stay
CA Patient Discharge Data	http://www.oshpd.ca.gov/HID/Products/PatDischargeData/PublicDataSet/index.html	Discharge data for licensed general acute hospital in CA with demographic, diagnostic and treatment information, disposition, total charges
MIMIC II Clinical Database	http://mimic.physionet.org/database.html	ICU data including demographics, diagnosis, clinical measurements, lab results, interventions, notes

Thanks to Prof. Joydeep Ghosh from UT Austin for providing this information

Publicly Available Medical Image Repositories

Image database Name	Modalities	No. Of patients	No. Of Images	Size Of Data	Notes/Applications	Download Link
Cancer Imaging Archive Database	CT DX CR	1010	244,527	241 GB	Lesion Detection and classification, Accelerated Diagnostic Image Decision, Quantitative image assessment of drug response	https://public.cancerimagingarchive.net/ncia/dataBasketDisplay.jsf
Digital Mammography database	DX	2620	9,428	211 GB	Research in Development of Computer Algorithm to aid in screening	http://marathon.csee.usf.edu/Mammography/Database.html
Public Lung Image Database	CT	119	28,227	28 GB	Identifying Lung Cancer by Screening Images	https://eddie.via.cornell.edu/crpf.html
Image CLEF Database	PET CT MRI US	unknown	306,549	316 GB	Modality Classification, Visual Image Annotation, Scientific Multimedia Data Management	http://www.imageclef.org/2013/medical
MS Lesion Segmentation	MRI	41	145	36 GB	Develop and Compare 3D MS Lesion Segmentation Techniques	http://www.ia.unc.edu/MSseg/download.php
ADNI Database	MRI PET	2851	67,871	16 GB	Define the progression of Alzheimer's disease	http://adni.loni.ucla.edu/data-samples/acscess-data/

Epidemiology Data

- The Surveillance Epidemiology and End Results Program (SEER) at NIH.
- Publishes cancer incidence and survival data from population-based cancer registries covering approximately 28% of the population of the US.
- Collected over the past 40 years (starting from January 1973 until now).
- Contains a total of 7.7M cases and >350,000 cases are added each year.
- Collect data on patient demographics, tumor site, tumor morphology and stage at diagnosis, first course of treatment, and follow-up for vital status.

Usage:

- Widely used for understanding disparities related to race, age, and gender.
- Can not be used for predictive analysis, but mostly used for studying trends.
- Medicare data for SEER patients is already linked.

SEER Database is available at http://seer.cancer.gov/
SEER-Medicare Linked Database available at http://healthservices.cancer.gov/seermedicare/

Public Health and Behavior Data Repositories

Dataset	Link	Description
Behavioral Risk Factor Surveillance System (BRFSS)	http://www.cdc.gov/brfss/technical_infodata/index.htm	Healthcare survey data: smoking, alcohol, lifestyle (diet, exercise), major diseases (diabetes, cancer), mental illness
Ohio Hospital Inpatient/Outpatient Data	http://publicapps.odh.ohio.gov/pwh/PWHMain.aspx?q=021813114232	Hospital: number of discharges, transfers, length of stay, admissions, transfers, number of patients with specific procedure codes
US Mortality Data	http://www.cdc.gov/nchs/data_access/cmf.htm	Mortality information on county-level
Human Mortality Database	http://www.mortality.org/	Birth, death, population size by country
Utah Public Health Database	http://ibis.health.utah.gov/query	Summary statistics for mortality, charges, discharges, length of stay on a county-level basis
Dartmouth Atlas of Health Care	http://www.dartmouthatlas.org/tools/downloads.aspx	Post discharge events, chronically ill care, surgical discharge rate

Thanks to Prof. Joydeep Ghosh from UT Austin for providing this information.

Conclusion

- Healthcare is a data-rich domain. As more and more data is being collected, there will be increasing demand for efficient data analytics.
- As the EHR data keeps growing at a rapid pace, more research on building new scalable predictive modeling platforms is required.
- An effective modeling platform for healthcare analytics research must integrate techniques from various disciplines such as information extraction, statistics, data mining, and visualization.
- Unraveling the “Big Data” related complexities can provide many insights about making the right decisions at the right time for the patients.
- Efficiently utilizing the colossal healthcare data repositories can yield some immediate returns in terms of patient outcomes and lowering care costs.

Acknowledgements

- Funding Sources
  - National Science foundation
  - National Institutes of Health
  - Susan G. Komen for the Cure
  - Delphinus Medical Technologies
  - IBM Research

Thank You

Questions and Comments

Feel free to email questions or suggestions to

jimeng@cs.cmu.edu

reddy@cs.wayne.edu