6.S897/HST.956 Machine Learning for Healthcare

Lecture 20: Precision Medicine

Instructors: David Sontag, Peter Szolovits

1 Introduction to Disease Subtyping

Disease subtyping is the process of figuring elucidating the subtypes of a disease, you can do it by clustering instances of the disease by variety of features, including demographics. vital signs, medications, etc. This lecture will talk about using genetics to subtypes, as it is important for applications such as using the human genome project for prescribing therapies.

Data-driven disease subtyping is much more relevant now than it is before as we now have new capabilities to compile molecular data that were unimaginable before. More specifically, we have had multiple drivers of change:

1. The cost of recording the human genome has been reduced from millions to thousands of dollars
2. There has been increasing success of utilizing molecular information to improve diagnoses
3. There have been many advances in information technology, including electronic health records.
4. Public attitudes toward molecular data have shifted. The fear of horror stories happening with molecular data being collected and used is declining, and people are now seeing more the potential benefits of using them

2 Collecting Genome Data

When we do collect a ton of data, we need to think about how to integrate all of the high dimensional data we receive. Google maps havs a coordinate system that they can use to layer information about transportation, census, etc.. Thus, we want to try accomplishing a similar thing in healthcare. Each patient has individual data about them like genome, epigenome, symptoms, patient data, and we can use individuals to represent the latitude and longitude of the coordinate system.

Efforts have gone out to bring in large collections of data for this purpose. The NIH started project called All of Us, asking institutions around the USA to get people to volunteer genetic information and mental data. The collected data of million people was intended to be a representative sample of the US population. In contrast to some studies done only on european populations, this effort attempts to get a good sample of the diverse US population.

3 The Framingham study

The Framingham study was conducted in the 1940s. Every year or two years, surveyers went out and surveyed about 50K people about genetic data, habits, smoke, weight, height. They ended up collecting data over generations.The vision of the study was to build information commons integrating all types of information of which biomedical research can rest.

An interesting focus of the study was on taxonomy, inspired by Samuel Johnson's quote, "My diseases are an asthma and a dropsy and, what is less curable, seventy-five." For instance, dropsy is water sickness,

swelling, and edema. It is not a disease. but a symptom of bunch of diseases. The last time dropsy was listed as a cause of death is 1949, and since it is disappeared as disease from taxonomy. Researchers suggest that we also need to treat asthma like this, or more specifically, not as a disease, but symptom of causes, and specify its taxonomy

4 Precision Medicine Modality Space (PMMS)

Isaac Kohane gave the idea of a precision medicine modality space (PMMS), which is essentially a high dimensionality space. If we are lucky, we will find clusters in the data when represented in that space. When representing data in PMMS, we can analyze it and perform operations such as principal component analysis to figure out the lower dimensional structure of the data. Kohane says if patient falls within middle of cluster, then the patient is normal for the cluster. But if patient is on edge of cluster, then might be something wrong with the patient.

4.1 Example Scenario

For instance, a 13 year old boy is presented with abdominal pain, hourly diarrhea and blood per rectum. 10 years earlier, he was diagnosed with ulcerative colitis.

He essentially was a sick puppy, given that he developed ulcerative colitis at 3, and then all of a sudden 10 years later, breaks out with horrible abdominal pain. After unsuccessfully trying a bunch of treatment, doctors propose drastic measures, like removing his colon.

Scientists have proposed 3 main groups of ulcerative colitis patients:

1. Life long remission
2. Initially multi-year remission but refractory over decades
3. Initially have a remission but then no standard therapy works

If we had plotted is position in PMMS at age 3, he would have fallen in the middle of a cluster. Thus, perhaps we could have used data-driven methods to identify the patient as belonging to group 3 before his current crisis.

5 Challenges with Using Machine Learning

There are many challenges with utilizing machine learning methods to solve these problems. One of which is defining a good distance metric and a threshold for outliers. One must also find a good representation of time-varying data, define an optimal weighting of PMMS dimensions, and a method to find the most specific neighborhood for a patient. After taking a shallow dive into genetics, we will see how we deal with these challenges in data-driven methods for clustering using genetics.

6 History of Genetics

Biology is the science of exceptions. 25 years ago, I (Prof. Szolovits) was sitting next to Prof. Gerald Sussman during a lecture, and we were learning about the many ways in which a theory in biology does not apply. Prof. Sussman stated to me: Its only exceptions, not theories

In the past, people knew that children inherit traits from parents, and many notable studies in genetics followed. Mendel developed a theory about discrete factors of inheritance, called genes. His experiment with

pea plants lead to the theory of dominant vs recessive traits. Miescher discovered nuclein, a compound in cell nuclei, now called DNA (coined by Hershey and Chase). Watson/Crick/Franklin discovered that DNA is a double helix

We know that a gene is a DNA sequence a fundamental physical and functional unit of heredity. However, we still have a lot to learn, mainly what parts of DNA code which genes, or which parts code genes at all. Crick made a sequence hypothesis that specifies the pieces of nucleic acid by a sequence of bases.

6.1 Central Dogma

The central Dogma of genetics describes the flow of information form DNA to RNA to proteins. The current interpretation is that DNA is double stranded, and single stranded RNA are formed from binding to the DNA then breaking off, and then triplets of pieces (codon) create proteins.A few nobel prizes later, we learned that transcription process is regulated by promoter, repressor, enhancer regions on genome. Repressors prevents activators from binding or alters activator, promoter and enhancers do effectively the opposite. The promoter of thymidine kinase gene of herpes simplex virus, and the enhancer of the SV40 virus, are shown in Figure 1.

[Figure 6.19 & Figure 6.20 from *The Cell: A Molecular Approach* removed due to copyright restrictions.]

Figure 1: Promoter of thymidine kinase gene and Enhancer of the SV40 Virus.

In fact, of the DNA in your cells, only 1.5 percent are exons that code for proteins, as shown in Figure 2. Some people call the rest junk DNA, but evolutionarily the junk DNA would have been deleted over time. We now know a little more about what the junk DNA is comprised of, including introns and regulatory sequences Introns are spliced out sequences. There are regulatory sequences that are code for regulatory proteins (promoter, repressor, etc.).

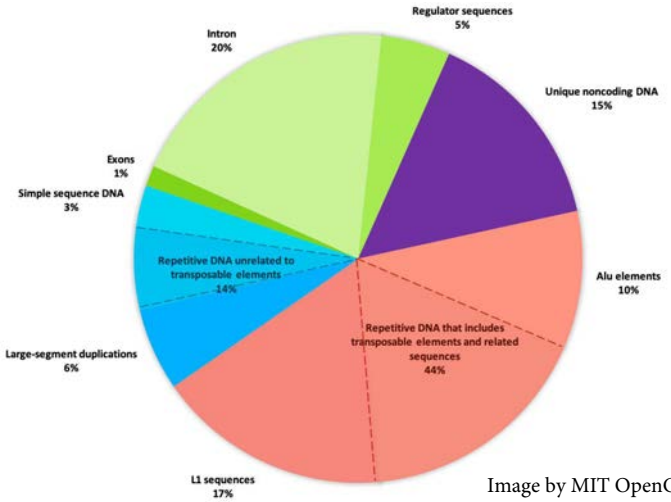

Image by MIT OpenCourseWare.

Figure 2: Distribution of components of DNA in cells

A new view states that a gene is any segment of DNA that is transcribed into RNA that has some function, and that view is becoming more and more accepted. In addition, new research has shown when DNA gets transcribed, it gets transcribed into pre mRNA, and then there is a process that splices out the introns before created mRNA. Then amino acids are formed from the mRNA.

6.2 Exceptions to the Central Dogma

Since biology is science of exceptions, there are also many exceptions to the Central Dogma. More specifically, there are specialized types of proteins and viruses that do not follow the general patterns

- Retroviruses: RNA that can turn into DNA via reverse transcriptase
- Prions are self-replicating proteins that are misfolded.
- DNA-modifying proteins repair DNA (crispr-CAs9)
- Retrotransposons modify DNA segments in Eukarya. It pops out pieces of DNA, which becomes inserted into another genome. It happens frequently in plants. For instance, wheat sometimes has a huge copy of DNA segments that initially didnt have.

To add in more complexity, there are other factors to consider.

- Various kinds of non-coding RNA participate in gene regulation
- RNA Interference (RNAi) latch on to RNA to prevent it to be translated into proteins.
- Once proteins are made, they are degraded differentially. Some proteins degrade much faster than others, so production rate not indicative of frequency
- Chromatin packages DNA into compact forms that are accessible to transcription only by histone modifications. We are still a little unsure how the Chromatin is unwrapped and accessed by proteins.

7 Cost of Genome Sequencing

Cost of Genome Sequencing The cost per genome has been declining for decades. For instance, Novogene will sequence all your exons for just hundreds of dollars.

A slightly more recent phenomenon is that people are saying we can sequence not only the DNA but also the RNA that came from the DNA. You can also buy a kit that will sample and for RNA sequences. Services exist to genotype cancers to see if perhaps which drugs best to treat cancer.

8 Early efforts to characterize Disease Subtypes Using Gene Expression Microarrays

The first class of disease subtyping papers came out around 2001, since then hundreds of thousands of similar kinds of analyses have been performed on different datasets. In one study by Alizadeh et. al. [1] on breast cancer cells, researchers extracted cancer tissue mRNA and then constructed DNA from that. They then amplified the DNA and marked it with red dye. They used collections of DNA fragments in each well of the microarray, and then took the amplified red-dye DNA and flowed it over the set of wells. Complimentary parts of DNA stuck to corresponding samples in the well.

Another approach involves marking normal tissue with green dye (i.e. or flourescent protein) and cancer tissue with red, and then one can measure the ratio of red to green at each spot in the microarray.

Researchers then clustered samples in gene expression space. They got a hierarchical clustering of genes and clustering of breast cancer biopsy specimens that represent genes in certain ways. It worked fairly well, and came out for a clustering with for example matched endothelial cell clusters with some tumor samples. It also clustered expression levels into 4 clusters, and in the figure below one can visually see how the clusters are much more similar to other samples in the cluster than samples in other clusters.

Researchers were able to determine 5 tumor subtypes based on clustering. When similar methods were done on lymphoma tissue, the results also were clusters that matched up with known groupings of lymphomas.

9 Genotypes and Phenotypes

To model relationships between genotypes and phenotypes, we need to have a list of all genotypes and phenotypes. Generating this list is doable for genotypes, but there are issues in defining phenotypes for this list. We could define phenotypes by disease, qualitative traits, quantitative traits, behavior, etc.

9.1 GWAS

Gene-wide association studies (GWAS) look for specific phenotype differences and then find what genetic differences correspond to those. People look for single-nucleotide polymorphisms, where a genome differs from population genome by 1 nucleotide. For example, in Huntingtons Disease, there are 30 repeats of a DNA sequence when the normal number of repeats is closer to 20.

The results from GWAS look like a Manhattan Plot, as shown Figure 3. All genes are laid out on sequence of chromosomes. For a specific phenotype, it plots difference in expression level based on presence or absence of gene. A high difference means that the there is some relation. More specifically, we can measure the ratio of odds of disease given genetic variant versus odds of disease given not having genetic variant. The odds ratio is measure of association between SNP (Single-Nucleotide Polymorphism) and phenotype. If the ratio is not one and statistically significant, it indicates an association

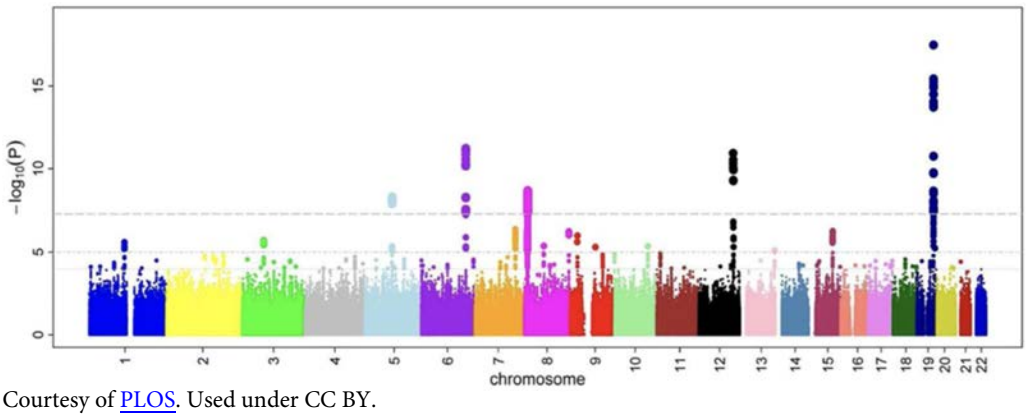

Figure 3: Manhattan plot representing the results from GWAS. The x-axis represents the specific gene, and the y-axis represents the phenotype difference between the gene being present and not being present.

GWAS typically identifies common variants with small effect sizes. Some studies result in odds ratio of 1.1/1.2, with although are statistically significant, are still really small. Highly penetrant medialain mutations are a kind of mutation not really elucidated by GWAS.

10 Current Companies

As an aside, there are certain companies now that ask for genetic makeup of children and parents/family, and then makes a recommendation at what the problem could be if there is a health problem with the child.

11 T2D Study

A study by Ulder et. al. [2] conducted a GWAS of Type-II diabetes. The goal was to identify soft clusters of genetic loci to suggest subtypes of T2D and possible mechanisms. Data was collected about 94 type 2 diabetes variants, including glycemic traits, BMI, height, etc. from multiple previous studies. They modeled the data through an association matrix X that had 47 traits by 94 variants. Traits were doubled, and one set was inverted where the z score was negative. This process was done since they used non-negative matrix factorization (NNMF) methods.

The researchers applied NNMF to factor the association matrix X into matrices W (47 by k) and H (k by 94). This method is an unsupervised learning method that attempts to find interesting patterns in data by dimensionality reduction. NNMF aims to minimize the L2 regularized loss $||X - WH||$ over the values of k, W, and H. Gibbs sampling and other stoastic tricks were used to measure P(X).

The researchers had data from about 17 thousand people of European ancestry, and the results of the study identified 5 subtypes of Type-II diabetes, with interpretations of Beta Cell, Proinsulin, Obesity, Lypodistrophy, and Liver/Lipid.

In the spider diagrams of the results are shown in Figure 4. In each of the 5 spider diagrams, in inner circle indicates a negative correlation, and the outer circle indicates a positive correlation. As shown by diagrams, factors are weighted differently in each of the 5 subtypes.

Figure 4: Spider Diagram of 5 Clusters.

12 PheWAS

Phenome-wide Association Study (PheWAS) is the complement of GWAS. It looks for phenotypic variations that correspond to specific genetic feature variations.

12.1 Example PheWAS study

Researchers Denny et. al. [3] from Vanderbilt then picked 5 genotypes (SNPs) , and then went thought tens of thousands of billing codes and by hand clustered them into 744 case groups (i.e. the phenotypes we are interested in). After plotting effect sizes, only multiple sclerosis really had a large effect size (statistically significant). The SNPs were selected because of its association with MS and Lupus, but the p-value with lupus was actually low. Thus, the study gave some unexpected results.

13 Expression Quantitative Trait Loci (eQTLs)

Expression Quantitative Trait Loci (eQTLs) eQTLs are genomic loci that explain variation in expression levels of mRNAs.We can take a look at gene expression levels and use those to define trait we are interested in. There exists differential expression in Different Populations. For example, Europeans and Africans have a 17

An interesting study in 2005 by Schadt et. al. [4] used bayesian methods to model complexities in genetic relationships. Using the trait locus (L), the rna expression level of that trait locus (R), and a complex trait (C), they modeled relationships between the three with many different bayesian graphs, and then picked the graph that yielded the highest likelihood of the data.

14 Gene Set Enrichment Analyses (GSEA)

Gene Set Enrichment Analysis (GSEA) is a method that determines whether a defined set of genes shows significant differences between two phenotypes. In these studies, even if genes pass hypothesis, there may not be a coherent understanding of relationships.

15 Deep Learning Applications

Deep Learning and sophisticated ML has not been used successfully yet for these applications. A lot of the studies you see coming up involves matrix factorization, clustering, bayesian networks, etc. Deep learning may make an impact in this field in the future.

16 Citations

1. Alizadeh AA, Ross DT, Perou CM, et al: Towards a novel classification of human malignan- cies based on gene expression patterns. J Pathol 195:41-52, 2001

2. Udler MS, Kim J, von Grotthuss M, Bons-Guarch S, Cole JB, Chiou J, et al. Type 2 diabetes genetic loci informed by multi-trait associations point to disease mechanisms and subtypes: A soft clustering analysis. PLoS medicine. 2018

3. Denny, J. C. et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. Bioinformatics 26, 12051210 (2010)

4. Schadt, E. E., Lamb, J., Yang, X., Zhu, J., Edwards, S., GuhaThakurta, D., et al. (2005). An integrative genomics approach to infer causal associations between gene 47 expression and disease. Nature Genetics, 37(7), 71071

6.S897 / HST.956 Machine Learning for Healthcare

Spring 2019