Biomedical Big Data and Precision Medicine

Jie Yang

Department of Mathematics, Statistics, and Computer Science
University of Illinois at Chicago

October 8, 2015

1. Explosion of Biomedical Data
2. Types and Sources of Biomedical Data
   - Electronic clinical data
   - DNA sequencing data
   - DNA variation: SNP data
   - Gene expression data: DNA microarray and RNA-Seq
   - CNV and DNA methylation data
3. Precision Medicine
   - Precision medicine vs. personalized medicine
   - Precision medicine in genetic disease
   - Precision medicine: prevention
   - Biobanking and data sharing
4. Permanental Classification Approach

Explosion of Biomedical Data: Electronic Medical Records

- Hospitals and medical centers are collecting and maintaining comprehensive medical records electronically.

- For example, a comprehensive dataset extracted from the Advocate Health and Hospitals Corporation database contains the electronic medical records of 109,421 adult inpatients (162,466 encounters) discharged between March 1, 2011 and July 31, 2012.

- Electronic medical data may facilitate health study researchers to identify important traits associated with complex disease and evaluate the corresponding treatments or therapies.

Explosion of Biomedical Data: Genomic Data

- Genomic data of a large cohort of individuals are assembled and become available for health study researches.
- The eMERGE consortium funded by the National Human Genome Research Institute combines electronic medical records and genomic data from almost 200,000 individuals (Ashley, 2015).

- The combined data are extremely high-dimensional and also becoming bigger and bigger, especially the genomic part.
- The first-generation genomic data from genotyping chips targets genetic variants of about 20,000 genes, while the gnome sequencing technique (about 10-fold more expensive than chips) can provide 1000-fold more data (Ashley, 2015).

Electronic Clinical Data: Advocate Database

- Collected from eight Advocate Health Care hospitals.
- It includes 109,421 adult inpatients (162,466 index admissions) discharged from March 1, 2011 to July 31, 2012.
- There are 298 independent variables, including: (1) administration variables: age, gender, race, quarter of index admission, insurance, medical service group, language, employment status, marriage status, discharge disposition, against medical advice (AMA) at discharge, AMA history, Braden score, Charlson comorbidity index, number of emergency room, number of observation and number of inpatient in the last year and the last 6 months; (2) current and history condition, procedure and medication variables; (3) lab test results during the index admission.

Request Data from Database: An SQL Example

```
drop table IF EXISTS de ;
select distinct population_id, empi_id
into local temp table de ON COMMIT PRESERVE ROWS
from(---union:2810596
select distinct empi_id ,population_id from
  APP.Claim_detail
where service_from_date>=2014-01-01 and
  service_to_date<2015-04-01
union
select distinct empi_id ,population_id from
  APP.Claim_other
where stmt_from_date>=2014-01-01 and
  stmt_to_date<2015-04-01 )
a where population_id=4e2ef1a40340 ;
```

Genomic Data: DNA Sequencing Data

- A *genome* is the sum of all the DNA in an organism. There are four types of chemical bases in a genome: A, T, C, and G.
- The human genome consists of 23 chromosome pairs and has 3 billion pairs of bases (3.235×10^9 nucleotides).
- Shotgun sequencing: Developed in 1977. Break longer DNA sequences into random small segments and sequence them to obtain reads, then assemble short reads into a continuous sequence.
- High-throughput (or next-generation) sequencing: First developed in 2004. Parallelize the sequencing process and producing thousands or millions of sequences at once.

DNA Sequencing Data: An Example

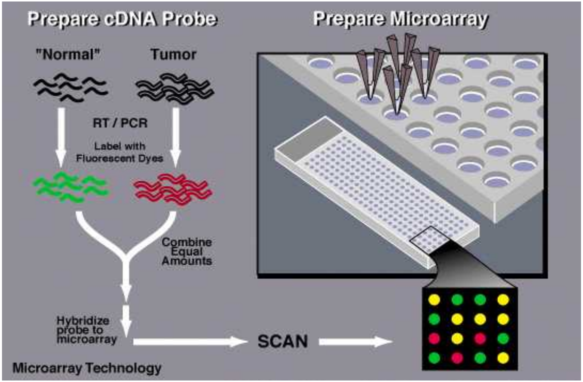
```
>gi|568815597|ref|NC_000001.11| Homo sapiens chromosome 1, GRCh38.p2
Primary Assembly
......
CCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCTAACCCAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCC
TAACCCTAACCCTAACCCTAACCCTAACCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACC
CTAACCCTAACCCCTAACCCTAACCCTAAACCCTAAACCCTAACCCTAACCCTAACCCTAACCCTAACCC
CAACCCCAACCCCAACCCCAACCCCAACCCCAACCCTAACCCCTAACCCTAACCCTAACCCTACCCTAAC
CCTAACCCTAACCCTAACCCTAACCCTAACCCCTAACCCCTAACCCTAACCCTAACCCTAACCCTAACCC
TAACCCTAACCCCTAACCCTAACCCTAACCCTAACCCTCGCGGTACCCTCAGCCGGCCCGCCCGCCCGGG
TCTGACCTGAGGAGAACTGTGCTCCGCCTTCAGAGTACCACCGAAATCTGTGCAGAGGACAACGCAGCTC
CGCCCTCGCGGTGCTCTCCGGGTCTGTGCTGAGGAGAACGCAACTCCGCCGTTGCAAAGGCGCGCCGCGC
CGGCGCAGGCGCAGAGAGGCGCGCCGCGCCGGCGCAGGCGCAGAGAGGCGCGCCGCGCCGGCGCAGGCGC
AGAGAGGCGCGCCGCGCCGGCGCAGGCGCAGAGAGGCGCGCCGCGCCGGCGCAGGCGCAGAGAGGCGCGC
CGCGCCGGCGCAGGCGCAGACACATGCTAGCGCGTCGGGGTGGAGGCGTGGCGCAGGCGCAGAGAGGCGC
GCCGCGCCGGCGCAGGCGCAGAGACACATGCTACCGCGTCCAGGGGTGGAGGCGTGGCGCAGGCGCAGAG
AGGCGCACCGCGCCGGCGCAGGCGCAGAGACACATGCTAGCGCGTCCAGGGGTGGAGGCGTGGCGCAGGC
GCAGAGACGCAAGCCTACGGGCGGGGGTTGGGGGGGCGTGTGTTGCAGGAGCAAAGTCGCACGGCGCCGG
GCTGGGGCGGGGGGAGGGTGGCGCCGTGCACGCGCAGAAACTCACGTCACGGTGGCGCGGCGCAGAGACG
.......
```

Single Nucleotide Polymorphism (SNP)

- A single nucleotide polymorphism (SNP, pronounced snip) is a DNA sequence variation occurring commonly within a population (e.g. 1%).
- For example, two sequenced DNA fragments from different individuals, AAGCCTA to AAGCTTA, contain a difference in a single nucleotide.
- SNPs occur in non-coding regions more frequently than in coding regions.
- As of June 8, 2015, dbSNP listed 149,735,377 SNPs in humans (`http://www.ncbi.nlm.nih.gov/SNP/`).
- A single SNP may cause a Mendelian disease. Examples include sickle-cell anemia, cystic fibrosis. For complex diseases, SNPs do not usually function individually.

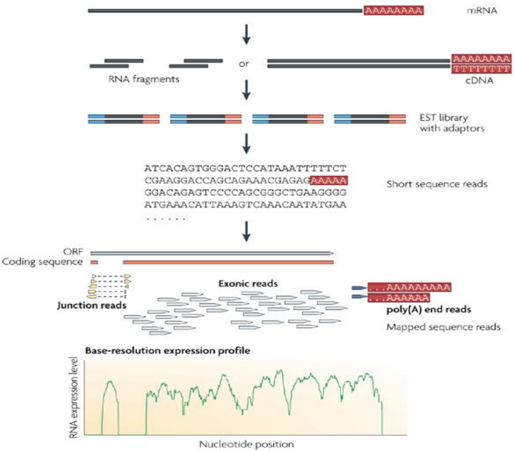
DNA Microarray Data

- A *DNA microarray* (also known as *DNA chip*) measures the expression levels of large numbers of genes simultaneously. For example, Arrayit's H25K (http://www.arrayit.com) contains 25,509 fully annotated human genes.



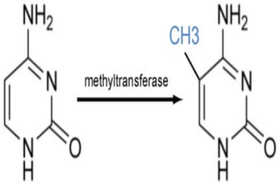
Alternative Technology: RNA-Seq

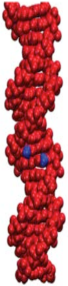
- *RNA-seq* (RNA sequencing) uses next-generation sequencing to reveal a snapshot of RNA presence, which can detect previously unidentified genes or transcripts.

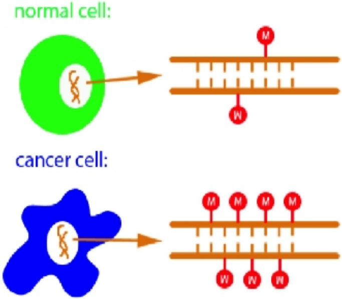


Other Types of Genomic Data

- Copy number variation (CNV): a structural variation, more than one copies of sections of DNA.
- DNA methylation: Methyl groups are added to DNA (see `http://www.ks.uiuc.edu/Research/methylation/`).





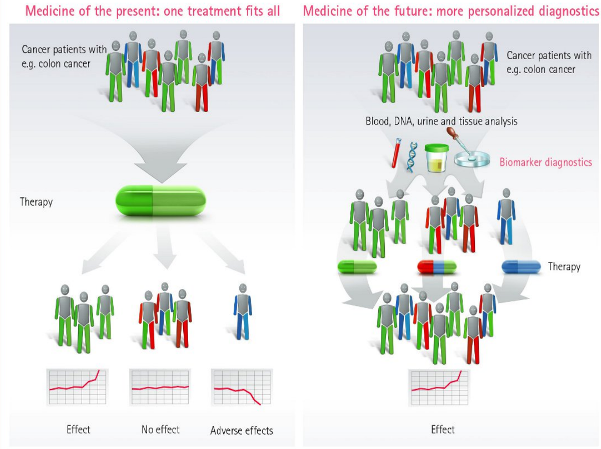


Precision Medicine Initiative

- *Precision medicine* refers to precisely classifying individuals into subpopulations according to a particular disease and precisely tailoring of medical treatments to subcategories of the disease (Committee on a Framework for Developing a New Taxonomy of Disease, 2011; Collins and Varmus, 2015; Ashley, 2015).
- Most medical treatments have been designed for the "average patient". Treatments can be very successful for some patients but not for others. Precision medicine takes into account individual differences in peoples genes, environments, and lifestyles for disease prevention and treatment.
- Launched with a $215 million investment in the President's 2016 Budget, the *Precision Medicine Initiative* will pioneer a new model of patient-powered research.

Precision Medicine vs. Personalized Medicine

- Precision Medicine classifies individuals into subpopulations.
- *Personalized Medicine* literally means the creation of drugs or medical devices that are unique to a patient.


Medicine of the present: one treatment fits all
Cancer patients with e.g. colon cancer
Therapy
Effect
No effect
Adverse effects
Medicine of the future: more personalized diagnostics
Cancer patients with e.g. colon cancer
Blood, DNA, urine and tissue analysis
Biomarker diagnostics
Therapy
Effect

Precision Medicine in Genetic Disease: An Example

- An example of precision medicine is the treatment of cystic fibrosis (inherited life-threatening disorder that damages the lungs and digestive system) with ivacaftor (Ramsey et al., 2011).
- The disease can be divided according to whether the defective channel reaches the cell surface or not.
- Ivacaftor (a potential treatment) increases the opening probability of the channel so that it is only effective in the subset of patients in whom the channel reaches the surface.
- The Cystic Fibrosis Foundation co-invested $150 million in the development of a particular drug targeting a precise subclass of patients.
- That initial investment increased in value to $3.3 billion by the time of the sale of the royalty rights in 2014 (Ashley, 2015).

Precision Medicine: Prevention

- A more sophisticated understanding of disease will almost certainly lead to more targeted and cost effective screening.

- One example is *familial hypercholesterolemia* (a genetic disorder characterized by high cholesterol levels), which carries a tier 1 recommendation for family screening from the Centers for Disease Control and Prevention.

- This genetic condition, which may be as common as 1:250 in the population, is associated with early myocardial infarction (heart attack).

- Cascade family screening with lipid panels and genetic testing has been shown to be highly cost-effective in identifying cases for potentially life-saving cholesterol-lowering therapy.

Biobanking and Data Sharing (Ashley, 2015)

- One major focus of the precision medicine initiative is the assembly of a large cohort of individuals willing to share their electronic medical record data and genomic data.
- In the first generation, genotyping chips targets the sequence of the approximately 20,000 genes. The next-generation sequencing technology is about 10-fold more expensive than chips, albeit for 1000-fold more data.
- The Million Veteran Program reports recruitment currently at more than 300,000 individuals, with thousands having been sequenced and hundreds of thousands having been genotyped.
- The eMERGE consortium combines electronic medical record data and genomic data from almost 200,000 individuals.

High-dimensional Classification and Clustering

- Achieving the goals of precision medicine requires sophisticated statistical and computational methods for high-dimensional classification and clustering. For a good review on commonly used classification and clustering methods, see, for example, Hastie et al. (2009).

- The medical records and genomic data of some patients may be collected under the same class label, say, lung cancer. How do we cluster the lung cancer patients into homogeneous subgroups so that the patients belonging to the same subgroup could be treated by a same therapy successfully? This is a *clustering* problem.

- If the clustering is successful, the next step would be classifying a new patient into one of the identified subgroups so that the patient could be treated by the most suitable therapy. This is a *classification* problem.

Permanental Classification Approach (Yang, Miescke, and McCullagh, 2012)

Permanental classification approach assumes that the observations of each cluster follow a *permanental process* (a Cox process with a special random intensity).

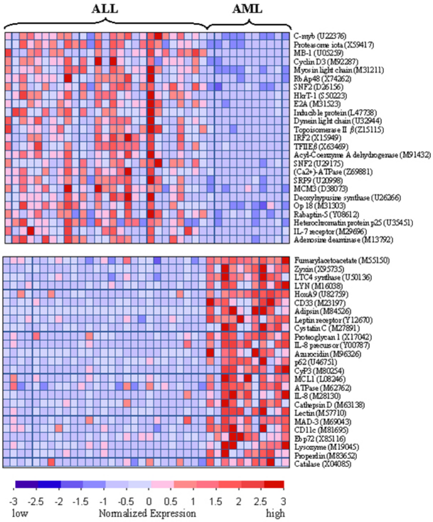
Given the observations $\mathbf{x} = \{x_1, \ldots, x_n\}$ with or without class labels $\mathbf{y} = \{y_1, \ldots, y_n\}$, the conditional distribution of the labels is

$$p_n(\mathbf{y} \mid \mathbf{x}) = \frac{\mathrm{per}_{\alpha_1}\left(K[\mathbf{x}^{(1)}]\right) \cdots \mathrm{per}_{\alpha_k}\left(K[\mathbf{x}^{(k)}]\right)}{\mathrm{per}_{\alpha_1+\cdots+\alpha_k}\left(K[\mathbf{x}]\right)}$$

Given a new unit u' with feature value x',

$$p_{n+1}(y(u') = r \mid \text{data}) \ \propto \ \frac{\mathrm{per}_{\alpha_r}\left(K[\mathbf{x}^{(r)} \cup \{x'\}]\right)}{\mathrm{per}_{\alpha_r}\left(K[\mathbf{x}^{(r)}]\right)}$$

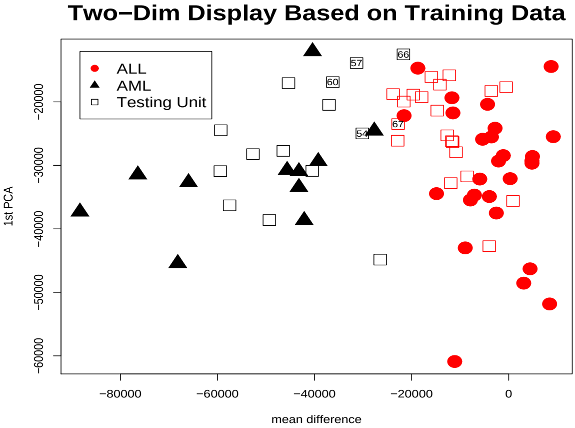
Leukemia Data (Golub et al., 1999)



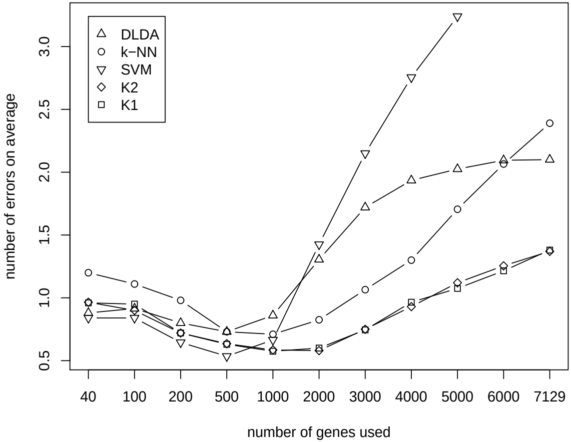
Two-Dimensional Display of Training Data

Training data: 38 samples (27 ALL, 11 AML)
Testing data: 34 samples (20 ALL, 14 AML)

Uncertain testing units (Golub et al., 1999)**:** No. 54, 57, 60, 66, 67



Number of Test Errors on Average (Dudoit et al. 2002)



Predicting Hypertension (Huang, Xu, and Yang, 2014)

Genome-wide association study (GWAS) data of 1043 individuals, three measurements for most participants, four time intervals (1981 to 1996, 1997 to 2000, 1998 to 2006 and 2009 to 2011), 65519 single-nucleotide polymorphisms (SNPs)

Prediction Errors of SVM and PC Using Common Variants

Number of SNPs	n=0	n=5	n=10	n=20	n=50	n=100	n=200
SVM (training)	0.230	0.029	0.018	0.013	0.018	0.005	0.001
SVM (testing)	0.242	0.146	0.135	0.127	0.126	0.121	0.125
PC (training)	0.223	0.103	0.097	0.040	0.033	0.034	0.032
PC (testing)	0.264	0.152	0.143	0.135	0.135	0.123	0.123

The best prediction error rates of SVM and PC are both close to 12%, which is much better than the error rate $\sim$ 22% based on logistic regression model.

Reference

- Ashley, E. (2015). The precision medicine initiative: A new national effort. *JAMA*, **313(21)**, 2119–2120.
- Wang, Gerstein, and Snyder (2009). RNA-Seq: a revolutionary tool for transcriptomics, *N*ature Reviews Genetics, **10**, 57–63.
- Collins, F. and H. Varmus (2015). A new initiative on precision medicine. *New England Journal of Medicine*, **372(9)**, 793–795.
- Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2nd ed.
- Huang, HH, Xu, T., and Yang, J. (2014). Comparing Logistic Regression, Support Vector Machines and Permanental Classification Methods in Predicting Hypertension, *BMC Proceedings*, **8**, Suppl.1:S96.
- Ramsey BW, Davies J, McElvaney NG, et al; VX08-770-102 Study Group (2011). A CFTR potentiator in patients with cystic fibrosis and the G551D mutation. *New England Journal of Medicine*, **365(18)**, 1663–1672.
- Yang, J., Miescke, K., and McCullagh, P. (2012). Classification based on a permanental process with cyclic approximation. *Biometrika*, 99, 775-786.