Instructor: Sael Lee

CS549 Spring – Computational Biology

LECTURE 11: BIOMARKER DISCOVERY

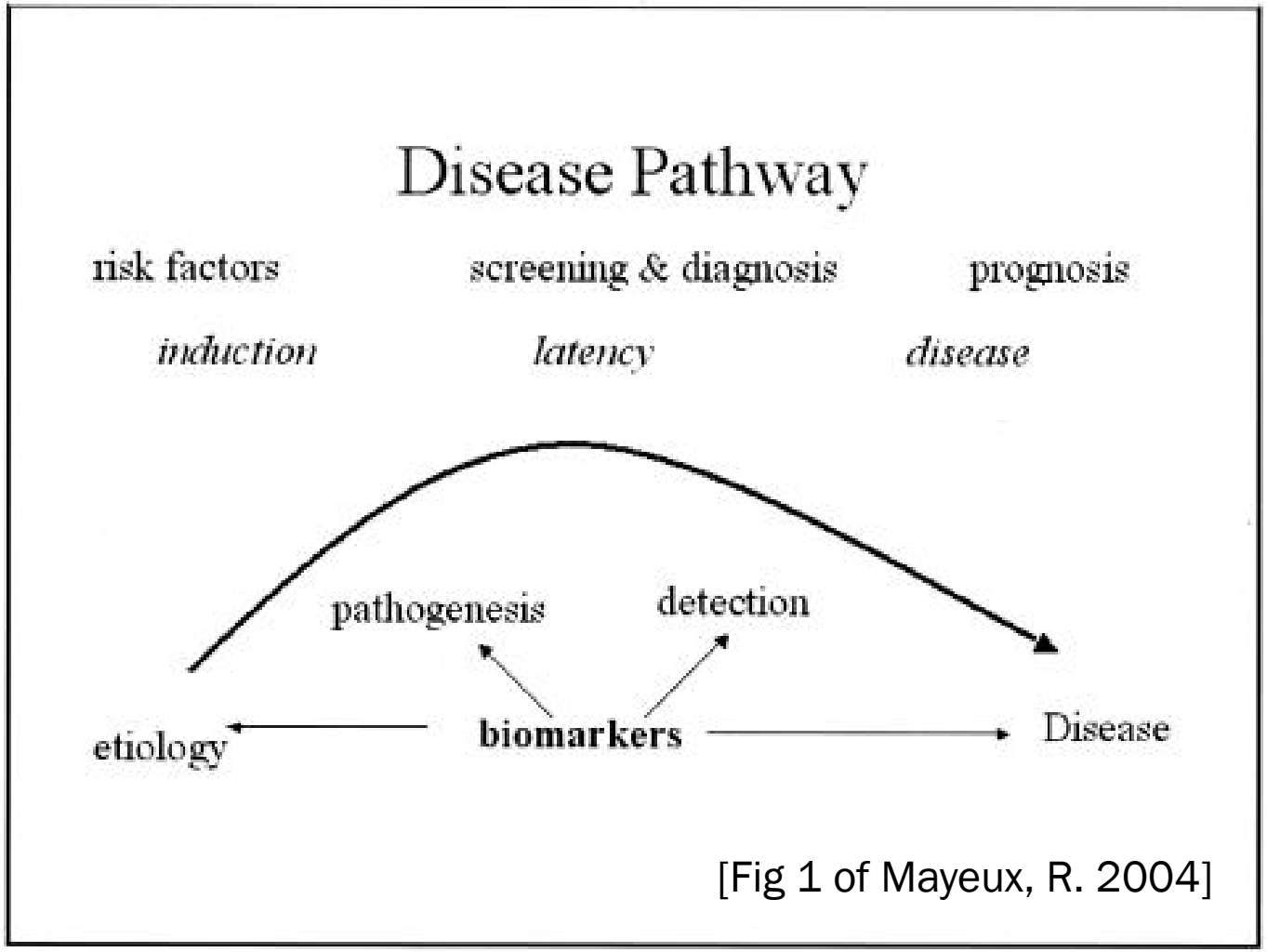
Resources: Steven Skiena's CSE 549 lecture 15-18 slides

WHAT IS A BIOMARKER?

- **Biomarker**, or biological marker, is any type of indicator of biological state.
  - "cellular, biochemical or molecular alterations that are measurable in biological media such as human tissues, cells, or fluids." - [B S Hulka (1990) New York: Oxford University Press]
- It objectively measures the states of biology in medicine, cell biology, geology, ecotoxicology, etc.
- The most popular uses are in medicine to measure states in:
  - Normal biological process
  - Pathogenic process
  - Pharmacological responds to therapeutics

http://www.news-medical.net/health/Biomarker-What-is-a-Biomarker.aspx

DISEASE PATHWAY AND POTENTIAL IMPACT OF BIOMARKER



[Fig 1 of Mayeux, R. 2004]

Mayeux, R. (2004). Biomarkers: potential uses and limitations. *NeuroRx : the journal of the American Society for Experimental NeuroTherapeutics*, *1*(2), 182–8.

EXAMPLE OF BIOMARKERS IN CLINICAL USAGE

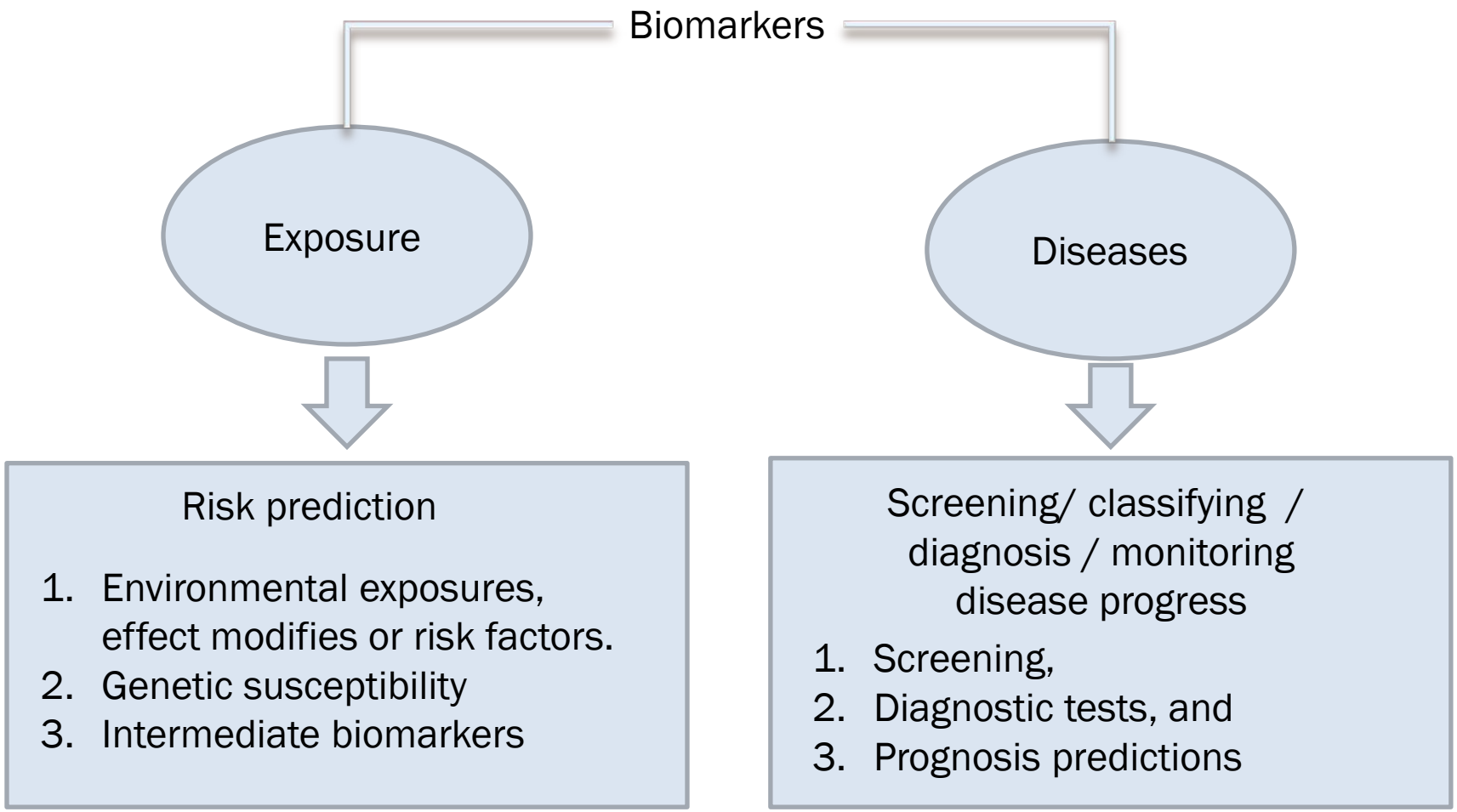
- Diagnosis and management of
  - cardiovascular disease,
  - infections,
  - immunological and genetic disorders,
  - cancer
  - nervous system disorders
  - absorption and metabolism of exposures (drug / other treatments / toxic materials)
  - Diseases risk prediction

CAPABILITIES OF BIOMARKERS [TABLE 1 OF MAYEUX, R. 2004]

- Delineation of events between exposure and disease
- Establishment of dose-response
- Identification of early events in the natural history
- Identification of mechanisms by which exposure and disease are related
- Reduction in misclassification of exposures or risk factors and disease
- Establishment of variability and effect modification
- Enhanced individual and group risk assessments

Mayeux, R. (2004). Biomarkers: potential uses and limitations. *NeuroRx : the journal of the American Society for Experimental NeuroTherapeutics*, *1*(2), 182–8.

TYPES OF BIOMARKERS



POSSIBLE SHORT COMES OF BIOMARKERS

1. Difference in amount of an external exposure
2. Difference in the way a putative toxin is metabolized
3. Personal difference / Group difference / measurement error

1. Content validity
   - degree to which a biomarker reflects the study
2. Construct validity
   - relevant characteristics of the disease or trait
3. Criterion validity
   1. sensitivity,
   2. specificity, and
   3. predictive power

DATA USED FOR BIOMARKER DISCOVERY

- Bio-specimens used:
  - Blood, brain, cerebrospinal fluid, spinal fluid, muscle, nerve, skin, and other body fluids
  - In both the healthy and diseased state
- **DNA, RNA, or protein**
  - **EX> Microarray chips, Genome sequences,**
- Cytogenetic markers
  - ex> chromosome structure
- Tissue markers
  - Microscope level visible differences
- Behavior markers
- Measure toxicants in body fluids & tissues
- Death of marker animals
  - Ex> environmental conditions.

BIOMARKER FOR CANCER TREATMENT: GENENTECH DESCRIPTION

Youtube

Understanding Biomarkers -Genentech scientist Jeff Settleman

- Cancer research – understanding cancer mechanism
- Risk of developing cancer
- Medication/treatment decision

MICROARRAY TECHNOLOGY

- Experimental techniques exist to enable biologists to measure the expression of a single gene under certain conditions. (in the past)
- Microarray technology enables one to do such experiments on a vastly larger scale. (currently)

- RNA-seq data: Becoming more popular

FOCUSING ON GENE EXPRESSION

- Certain technologies have been developed where different compounds are anchored to tiny *beads*, so reacting beads can be labeled, isolated, and identified.
- But the best solution is to attach distinct compounds to different regions of a solid substrate so you know *where* they are.

WHAT DOES MICROARRAY MEASURE

- Analysis of post translational modifications in genes
  - ex.> methylation states.
- Sequencing variants of a *known* genome
  - detecting single nucleotide polymorphisms (SNPs)
- Identifying a specific strain of virus
  - (e.g. the Affymetrix HIV-1 array).
- Measuring differential expression of all genes in tumor and normal cells,
  - to determine which genes may cause/cure cancer

- Identify which treatment a specific tumor should respond best to.
  - Paired treatment
- Measuring differential expression of all genes in different tissue types,
  - to determine what makes one cell type different than another.
- Measuring differential expression of all genes in different time
  - Circadian rhythm
- Measuring copy number variants from chromosomal anomalies or cancer.
- Obtaining individual's genotype / SNP data, e.g. 23andMe

DNA MICROARRAY

cDNA microarray YouTube 1. – Gabriel Mckinsey

DNA Microarray YouTube 2.

- Single stranded DNA/RNA molecules are anchored by one end to the plate/substrate.
  - These molecules will seek to hybridize with complementary strands floating in solution.
- The target molecules are fluorescently labeled,
  - so that the spots on the *chip/array* where hybridization occurs can be identified.
- The strength of the detected signal somewhat reflects the amount of stuff which binds to it,
  - and thus the amount of the target in solution.
- Such *quantitative* expression data is not very reliable, however.

http://www.3d-gene.com/en/about/abo_001.html

THERE ARE MANY POSSIBLE SOURCE OF ERROR

- Accuracy and fluctuations in scanning the fluorescent signals
- The strength of the bond formed between two single stranded DNA/RNA molecules is a function of
  - (1) the length of the bonded molecules,
  - (2) the base composition of the molecules, since A/T and C/G bond with different energies,
  - (3) the number and location of base mismatches, since end mismatches cause less trouble.
- Efficiency of hybridization of labeled cDNA to each slides
  - ***Cross hybridization*** is a source of many false positive errors,
    - a closely related DNA sequence binds at the probe in the absence of the desired target.
  - Heat breaks these bonds, so the ***stringency* of hybridization** can be effected by changing the temperature and other conditions.
  - ***Self hybridization*** occurs when probe molecules fold and hybridize with themselves, thus rendering them less effective at hybridizing with the target.
    - This occurs particularly in *self palindromic* probes.
- Variations within and between oligonucleotide spots,
- Efficiency of dye incorporation: Image Processing Issues

COMPLEXITY IN ANALYSIS OF MICROARRAY DATA

- Underlying biological processes being investigated are often not understood and are almost certainly complex
- Measures the steady-state level of an unstable molecule , mRNA
  - Depends on the rate of transcription and degradation of the mRNA.

CLASSIFICATION AND CLUSTERING PROBLEM

- Finding Biomarkers using microarray data becomes **feature selection** (gene selection) problem in **classification** (supervised learning) and **clustering** (unsupervised learning)

FEATURE SELECTION AND BIOMARKER DISCOVERY

- Feature selection challenge specific to microarray data:
  - Large feature (gene) and small number of data (samples)
  - Reproducibility is low
    - need stable feature selection method.
- Cause of instability
  - Algorithm design without considering stability
  - The existence of multiple sets of true markers
  - Small number of samples in high dimensional data

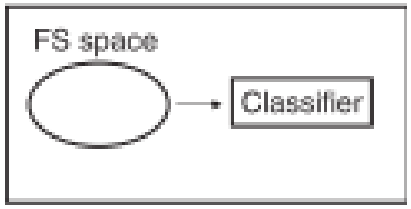
FEATURE SELECTION

- Selected features can be singular or form groups.
  - Singular: early onset genetic diseases
  - **Group feature: complex diseases**
    - cancer, diabetes, etc

- Incorporation of prior-knowledge in to feature selection.
  - **Best to incorporate all we know** esp. since variable samples are always small
    - Interaction between genes

TYPES OF FEATURE SELECTION METHOD

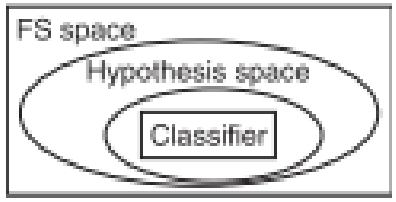
Filtering Methods



relevance of features is evaluated by looking only at the intrinsic properties of the data

* Often feature relevance score is used to evaluate each feature (gene)

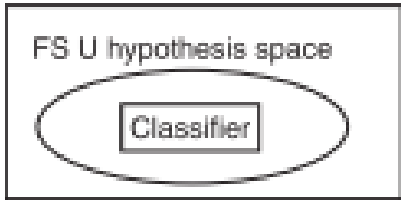
Wrapper Methods



model hypothesis search is embed within the feature subset search

-> various subsets of features are generated and evaluated

Embedded Method



optimal feature subset search is built into the classifier construction

-> a search in the combined space of feature subsets and hypotheses

Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, *23*(19), 2507–17.