Mathematics, Genomics, and Cancer

Esfandiar Haghverdi

School of Informatics
IUB

April 6, 2009
Outline

- Introduction
- Class Comparison
- Class Discovery
- Class Prediction
- Example
- Biological states and state modulation
- Software Tools
- Research directions
Long history.

Very successful applications of mathematical modeling: from physiology of heart to cell motility.

Uses systems of differential equations to describe the underlying phenomena.

Not covered in this talk.

See the recent success story:
Can Mathematics Cure Leukemia?

Math & DNA Sequences

- More recent.

Very successful application of techniques from probability theory, statistics and discrete mathematics to solve problems associated with DNA sequencing. Not covered in this talk. See 2009 Einstein Lecture to be given by Michael S. Waterman: Reading DNA Sequences: 21st Century Technology with 18th Century Mathematics, on April 4 at the 2009 Spring Southeastern Section Meeting, at North Carolina State University.
More recent.

Very successful application of techniques from probability theory, statistics and discrete mathematics to solve problems associated with DNA sequencing.
More recent.

Very successful application of techniques from probability theory, statistics and discrete mathematics to solve problems associated with DNA sequencing.

Not covered in this talk.
More recent.

Very successful application of techniques from probability theory, statistics and discrete mathematics to solve problems associated with DNA sequencing.

Not covered in this talk.

See 2009 Einstein Lecture to be given by Michael S. Waterman: Reading DNA Sequences: 21st Century Technology with 18th Century Mathematics, on April 4 at the 2009 Spring Southeastern Section Meeting, at North Carolina State University.
Genetics in 2 minutes!

- DNA = a polynucleotide
- Nucleotide = sugar + phosphate group + base
- Base = a flat, single or double ring-shaped molecule: \(A, T, C, G, U\).
  DNA is a word on the alphabet \(\{A, C, G, T\}\).
- Double helix structure of DNA (Watson and Crick 1953):
  hydrogen bond between \(A\) and \(T\), and \(C\) and \(G\).
- RNA = a polynucleotide, single strand, formed from
  \(A, U, C, G\).
- Protein = a polypeptide, chain of amino acids held together with peptide bonds.
Central Dogma of Molecular Biology:

\[
\text{DNA} \xrightarrow{\text{transcription}} \text{mRNA} \xrightarrow{\text{translation}} \text{Protein}.
\]

Proteins are coded by codons, words of length 3 on \{A, U, C, G\} that specify the 20 known amino acids. For example: Glutamic Acid (glu) = GAG, Valine (val) = GUG.
Central Dogma of Molecular Biology:

\[ \text{DNA} \xrightarrow{\text{transcription}} \text{mRNA} \xrightarrow{\text{translation}} \text{Protein}. \]

Proteins are coded by codons, words of length 3 on \{A, U, C, G\} that specify the 20 known amino acids. For example: Glutamic Acid (glu) = GAG, Valine (val) = GUG.

Note the one letter difference: Sickle-cell anemia.
Central Dogma of Molecular Biology:

\[
\text{DNA} \xrightarrow{\text{transcription}} \text{mRNA} \xrightarrow{\text{translation}} \text{Protein}.
\]

Proteins are coded by codons, words of length 3 on \{A, U, C, G\} that specify the 20 known amino acids. For example: Glutamic Acid (glu) = GAG, Valine (val) = GUG.

Note the one letter difference: Sickle-cell anemia.

Gene = a locatable region of genomic sequence, corresponding to a unit of inheritance, and is associated with regulatory regions, transcribed regions and/or other functional sequence regions.
Some pictures
Loss/Gain of function and disease

Sickle Cell Anemia:
Autosomal recessive disorder
E6V in HBB causes interaction w/ F85 and L88
Formation of amyloid fibrils
Abnormally shaped red blood cells
Leads to sickle cell anemia
Manifestation of disease vastly different over patients


http://gingi.uchicago.edu/hbs2.html
Measuring mRNA

Microarrays are tools that make it possible to measure the level of mRNA for thousands of genes simultaneously. The assumption is that this is a good indication of actual \textit{in vivo} protein levels.
What are we trying to do?

- Extract meaningful information from gene expression data.
What are we trying to do?

- Extract meaningful information from gene expression data.
- In this talk, I present an overview which is by no means comprehensive.
What are we trying to do?

- Extract meaningful information from gene expression data.
- In this talk, I present an overview which is by no means comprehensive.
- There is no one-size-fits-all solution for the analysis and interpretation of genome wide data.
What are we trying to do?

- Extract meaningful information from gene expression data.
- In this talk, I present an overview which is by no means comprehensive.
- There is no one-size-fits-all solution for the analysis and interpretation of genome wide data.
- Many analysis options are available at all phases of analysis.
What are we trying to do?

- Extract meaningful information from gene expression data.
- In this talk, I present an overview which is by no means comprehensive.
- There is no one-size-fits-all solution for the analysis and interpretation of genome wide data.
- Many analysis options are available at all phases of analysis.
- An understanding of both biology and the computational methods is essential.
What are we trying to do?

- Extract meaningful information from gene expression data.
- In this talk, I present an overview which is by no means comprehensive.
- There is no one-size-fits-all solution for the analysis and interpretation of genome wide data.
- Many analysis options are available at all phases of analysis.
- An understanding of both biology and the computational methods is essential.
- Complex mathematical methods do not necessarily perform better than simpler ones.
What are we trying to do?

- Extract meaningful information from gene expression data.
- In this talk, I present an overview which is by no means comprehensive.
- There is no one-size-fits-all solution for the analysis and interpretation of genome wide data.
- Many analysis options are available at all phases of analysis.
- An understanding of both biology and the computational methods is essential.
- Complex mathematical methods do not necessarily perform better than simpler ones.
- Prepackaged analysis tools are not a good substitute for collaboration with mathematicians and computational/statistical scientists on complex problems.
Gene expression signatures: One can rationally distill a list of genes from an unbiased global scan of gene-expression changes observed across a carefully selected sample set.
Central Ideas

- **Gene expression signatures:** One can rationally distill a list of genes from an unbiased global scan of gene-expression changes observed across a carefully selected sample set.

- Biological states can be characterized by gene expression signatures.
Central Ideas

- **Gene expression signatures**: One can rationally distill a list of genes from an unbiased global scan of gene-expression changes observed across a carefully selected sample set.

- Biological states can be characterized by gene expression signatures.

- We can use gene expression signatures as surrogates for biological states.

GSEA: Gene Set Enrichment Analysis, A. Subramanian et al., *PNAS*, 102, 2005.

GSEA: Gene Set Enrichment Analysis, A. Subramanian et al., PNAS, 102, 2005.

Central Papers

- **GSEA:** Gene Set Enrichment Analysis, A. Subramanian et al., *PNAS*, **102**, 2005.
Goal: Identify differentially expressed genes.
Goal: Identify differentially expressed genes.

Methods:

- Calculate a test statistic ($t$-test, ANOVA $F$ statistic, non-parametric rank-based, ...)
- Determine the significance of the observed value for test statistic.
- Normality, equal variance, multiple testing, ...
Class Comparison

- **Goal:** Identify differentially expressed genes.

- **Methods:**
  - Calculate a test statistic ($t$-test, ANOVA $F$ statistic, non-parametric rank-based,...)
  - Determine the significance of the observed value for test statistic.
  - **Normality, equal variance, multiple testing,...**

- **Issues:**
  - Two or more experimental conditions
  - Conditions may be independent or related (time series)
  - Many different combinations of experimental variables
  - Replication, to estimate variability, to identify biologically reproducible changes
  - How to incorporate estimates of variation (model-based methods)
Goal: Identify meaningful patterns in the data, (aka. Unsupervised learning.)
Goal: Identify meaningful patterns in the data, (aka. Unsupervised learning.)

Methods:

- Dimensionality reduction methods
  - Most of the variation in data can be explained by a smaller number of transformed variables.
  - For example: SVD, PCA, MDS, ...

- Clustering
  - Data can be grouped into groups of similar points based on some similarity measure.

- Aggregation methods (e.g., HC)

- Partitioning or centroid methods (for example, $k$-means, SOM or Kohonen maps)

- Model-based methods (e.g., fitting into some mixture model)

- Optimization techniques (within class, between class)
Class Discovery

- **Goal**: Identify meaningful patterns in the data, (aka. Unsupervised learning.)

- **Methods**:
  - *Dimensionality reduction methods*: Most of the variation in data can be explained by a smaller number of transformed variables.
Class Discovery

- **Goal**: Identify meaningful patterns in the data, (aka. Unsupervised learning.)
- **Methods**:
  - *Dimensionality reduction methods*: Most of the variation in data can be explained by a smaller number of transformed variables.
    - For example: SVD, PCA, MDS, . . .
Goal: Identify meaningful patterns in the data, (aka. Unsupervised learning.)

Methods:

Dimensionality reduction methods: Most of the variation in data can be explained by a smaller number of transformed variables.

- For example: SVD, PCA, MDS, . . .

Clustering: Data can be grouped into groups of similar points based on some similarity measure.

- Aggregation methods (e.g., HC)
- Partitioning or centroid methods (for example, \( k \)-means, SOM or Kohonen maps)
- Model-based methods (e.g., fitting into some mixture model)
- Optimization techniques (within class, between class)
Issues:

- It is unbiased, no a priori assumption.
Issues:

- It is unbiased, no a priori assumption.
- The structure may not be of clinical or biological interest.
Issues:

- It is unbiased, no a priori assumption.
- The structure may not be of clinical or biological interest.
- Should be viewed as a first step in more detailed analysis.
Issues:

- It is unbiased, no a priori assumption.
- The structure may not be of clinical or biological interest.
- Should be viewed as a first step in more detailed analysis.
- There is no single best way to evaluate a clustering method or a cluster.
Class Discovery cont’d

**Issues:**

- It is unbiased, no a priori assumption.
- The structure may not be of clinical or biological interest.
- Should be viewed as a first step in more detailed analysis.
- There is no single best way to evaluate a clustering method or a cluster.
- How to evaluate clustering methods?
Issues:

- It is unbiased, no a priori assumption.
- The structure may not be of clinical or biological interest.
- Should be viewed as a first step in more detailed analysis.
- There is no single best way to evaluate a clustering method or a cluster.
- How to evaluate clustering methods?
- There is no single best clustering method for a data set.
Issues:

▶ It is unbiased, no a priori assumption.
▶ The structure may not be of clinical or biological interest.
▶ Should be viewed as a first step in more detailed analysis.
▶ There is no single best way to evaluate a clustering method or a cluster.
▶ How to evaluate clustering methods?
▶ There is no single best clustering method for a data set.
▶ Some desirable properties maybe: Stability (reliability), predictive power, reduction power, ...
Issues:

- It is unbiased, no a priori assumption.
- The structure may not be of clinical or biological interest.
- Should be viewed as a first step in more detailed analysis.
- There is no single best way to evaluate a clustering method or a cluster.
- How to evaluate clustering methods?
- There is no single best clustering method for a data set.
- Some desirable properties maybe: Stability (reliability), predictive power, reduction power, ...
- How to choose the number of clusters (Gordon, repeated sampling, gap statistic, ...)

Esfandiar Haghverdi  Mathematics, Genomics, and Cancer
Opportunities:

- Stochastic clustering (e.g., NMF)

Techniques from statistical physics (e.g., Deterministic Annealing)

Spectral methods (e.g., Diffusion maps on graphs)

Geometric methods (e.g., Diffusion maps on manifolds)

Information theoretic methods

Statistical theory of clustering (Cf. comparing clustering methods)
Opportunities:

- Stochastic clustering (e.g., NMF)
- Techniques from statistical physics (e.g., Deterministic Annealing)
Opportunities:

- Stochastic clustering (e.g., NMF)
- Techniques from statistical physics (e.g., Deterministic Annealing)
- Spectral methods (e.g., Diffusion maps on graphs)
Opportunities:
- Stochastic clustering (e.g., NMF)
- Techniques from statistical physics (e.g., Deterministic Annealing)
- Spectral methods (e.g., Diffusion maps on graphs)
- Geometric methods (e.g., Diffusion maps on manifolds)
Opportunities:

- Stochastic clustering (e.g., NMF)
- Techniques from statistical physics (e.g., Deterministic Annealing)
- Spectral methods (e.g., Diffusion maps on graphs)
- Geometric methods (e.g., Diffusion maps on manifolds)
- Information theoretic methods
Opportunities:

- Stochastic clustering (e.g., NMF)
- Techniques from statistical physics (e.g., Deterministic Annealing)
- Spectral methods (e.g., Diffusion maps on graphs)
- Geometric methods (e.g., Diffusion maps on manifolds)
- Information theoretic methods
- Statistical theory of clustering (Cf. comparing clustering methods)
Goal: Design an accurate classifier (predictor) under the guidance of a supervisor, (aka. Supervised learning problem.) E.g., predicting cancer (sub)types, clinical outcomes, etc.
Goal: Design an accurate classifier (predictor) under the guidance of a supervisor, (aka. Supervised learning problem.) E.g., predicting cancer (sub)types, clinical outcomes, etc.

Methods:
- Linear and quadratic discriminant analysis
- Weighted voting
- Shrunken centroids
- $k$-NN
- Neural nets
- SVM
- Decision tree classifiers
- Naive Bayes
- Bagging and boosting (combining classifiers)
Class Prediction, cont’d

*Issues:*

- features $\gg$ samples
Issues:

- features $\gg$ samples
- Overfitting (modeling the training data too exactly)
Issues:

- features $\gg$ samples
- Overfitting (modeling the training data too exactly)
- High level of noise
Issues:

- features $>>$ samples
- Overfitting (modeling the training data too exactly)
- High level of noise
- Which method to choose?
Issues:

- features >> samples
- Overfitting (modeling the training data too exactly)
- High level of noise
- Which method to choose?
  - Careful with comparisons
Issues:

- features >> samples
- Overfitting (modeling the training data too exactly)
- High level of noise
- Which method to choose?
  - Careful with comparisons
  - Some trends emerge (e.g., Diagonal LD does better than Fisher's LD, $k$-NN performs better after gene filtering, combined methods do better, simpler methods do better, ...)

Esfandiar Haghverdi  Mathematics, Genomics, and Cancer
Opportunities:

- Theory for method and classifier comparison
Opportunities:

▶ Theory for method and classifier comparison
▶ Combining knowledge from different methods
Opportunities:

- Theory for method and classifier comparison
- Combining knowledge from different methods
- Incorporating knowledge from complementary sources
Opportunities:

▶ Theory for method and classifier comparison
▶ Combining knowledge from different methods
▶ Incorporating knowledge from complementary sources
▶Boosting the power and rigor of analysis
Example (Golub et al. 1999)

ALL vs AML (The Biology of Cancer, R. Weinberg)
Normal kidney vs Renal cell carcinoma.
Sample: 38 bone marrow samples (27 ALL, 11 AML).
6817 genes
Example cont’d

- **Sample:** 38 bone marrow samples (27 ALL, 11 AML). 6817 genes

- **Test statistic:** $SNR = \frac{\mu_1 - \mu_2}{\sigma_1 + \sigma_2}$ to determine gene-class correlation

- **Significance test:** Permutation test (nhood analysis)

- **Signature:** 50 informative genes are selected

- **Prediction:** Given a new sample each informative gene casts a weighted vote, the votes are then summed to determine the winning class and to define the $0 < PS < 1$ (prediction strength) which needs to be above $0.3$ for each decisive vote.

- **Validity of the predictor:** Cross-validation and trying on test data.
Sample: 38 bone marrow samples (27 ALL, 11 AML). 6817 genes

Test statistic: \( SNR = \frac{\mu_1 - \mu_2}{\sigma_1 + \sigma_2} \) to determine gene-class correlation

Significance test: Permutation test (nhood analysis)
Sample: 38 bone marrow samples (27 ALL, 11 AML). 6817 genes

Test statistic: \( SNR = \frac{\mu_1 - \mu_2}{\sigma_1 + \sigma_2} \) to determine gene-class correlation

Significance test: Permutation test (nhood analysis)

Signature: 50 informative genes are selected
Sample: 38 bone marrow samples (27 ALL, 11 AML). 6817 genes

Test statistic: \( SNR = \frac{\mu_1 - \mu_2}{\sigma_1 + \sigma_2} \) to determine gene-class correlation

Significance test: Permutation test (nhood analysis)

Signature: 50 informative genes are selected

Prediction: Given a new sample each informative gene casts a weighted vote, the votes are then summed to determine the winning class and to define the \( 0 < PS < 1 \) (prediction strength) which needs to be above 0.3 for each decisive vote.
Example cont’d

- **Sample:** 38 bone marrow samples (27 ALL, 11 AML). 6817 genes
- **Test statistic:** $SNR = \frac{\mu_1 - \mu_2}{\sigma_1 + \sigma_2}$ to determine gene-class correlation
- **Significance test:** Permutation test (nhood analysis)
- **Signature:** 50 informative genes are selected
- **Prediction:** Given a new sample each informative gene casts a weighted vote, the votes are then summed to determine the winning class and to define the $0 < PS < 1$ (prediction strength) which needs to be above 0.3 for each decisive vote.
- **Validity of the predictor:** Cross-validation and trying on test data.
Suppose we have \( n \) samples.

- Gene vector: \( v(g) = (e_1, \cdots, e_n) \),
Suppose we have \( n \) samples.

- **Gene vector**: \( v(g) = (e_1, \ldots, e_n) \),
- **Class vector**: \( c = (c_1, \ldots, c_n) \): \( c_i = 1 \) if \( i \in AML \), and 0 if \( i \in ALL \).
Suppose we have $n$ samples.

- **Gene vector:** $v(g) = (e_1, \cdots, e_n)$,
- **Class vector:** $c = (c_1, \cdots, c_n)$: $c_i = 1$ if $i \in AML$, and 0 if $i \in ALL$.
- **Gene-Class Correlation:**
  
  $$P(g, c) = \frac{\mu_1(g) - \mu_2(g)}{\sigma_1(g) + \sigma_2(g)}$$
  for each gene $g$. 

- **Nhood:** $N_1(c, r) = \{g | P(g, c) = r\}$
- **Permutation test:** compare with $N_1(c^*, r)$, for 400 permutations.
- **Number of informative genes** is a free parameter, chosen to be 50: 25 top-most and 25 bottom-most.

**Robustness to noise**

**Ease of applicability.**
In symbols ...

Suppose we have \( n \) samples.

- **Gene vector:** \( v(g) = (e_1, \cdots, e_n) \),
- **Class vector:** \( c = (c_1, \cdots, c_n) \): \( c_i = 1 \) if \( i \in AML \), and 0 if \( i \in ALL \).
- **Gene-Class Correlation:** 
  \[
P(g, c) = \frac{\mu_1(g) - \mu_2(g)}{\sigma_1(g) + \sigma_2(g)}
\]
  for each gene \( g \).
- **Nhood:** \( N_1(c, r) = \{ g \mid P(g, c) = r \} \)
Suppose we have $n$ samples.

- **Gene vector**: $v(g) = (e_1, \ldots, e_n)$,
- **Class vector**: $c = (c_1, \ldots, c_n)$: $c_i = 1$ if $i \in AML$, and 0 if $i \in ALL$.
- **Gene-Class Correlation**: $P(g, c) = (\mu_1(g) - \mu_2(g))/(\sigma_1(g) + \sigma_2(g))$ for each gene $g$.
- **Nhood**: $N_1(c, r) = \{g \mid P(g, c) = r\}$
- **Permutation test**: compare with $N_1(c^*, r)$, for 400 permutations.
Suppose we have $n$ samples.

- **Gene vector:** $v(g) = (e_1, \cdots, e_n)$,
- **Class vector:** $c = (c_1, \cdots, c_n)$: $c_i = 1$ if $i \in AML$, and 0 if $i \in ALL$.
- **Gene-Class Correlation:**
  
  $$P(g, c) = (\mu_1(g) - \mu_2(g))/(\sigma_1(g) + \sigma_2(g))$$
  for each gene $g$.
- **Nhood:** $N_1(c, r) = \{g \mid P(g, c) = r\}$
- **Permutation test:** compare with $N_1(c^*, r)$, for 400 permutations.
- **Number of informative genes is a free parameter,** chosen to be 50: 25 top-most and 25 bottom-most.
Suppose we have $n$ samples.

- **Gene vector:** $v(g) = (e_1, \cdots, e_n)$,
- **Class vector:** $c = (c_1, \cdots, c_n)$: $c_i = 1$ if $i \in AML$, and 0 if $i \in ALL$.
- **Gene-Class Correlation:**
  $$P(g, c) = (\mu_1(g) - \mu_2(g))/(\sigma_1(g) + \sigma_2(g))$$ for each gene $g$.
- **Nhood:** $N_1(c, r) = \{g \mid P(g, c) = r\}$
- **Permutation test:** compare with $N_1(c^*, r)$, for 400 permutations.
- **Number of informative genes** is a free parameter, chosen to be 50: 25 top-most and 25 bottom-most.
  - Robustness to noise
Suppose we have $n$ samples.

- **Gene vector:** $\mathbf{v}(g) = (e_1, \cdots, e_n)$,
- **Class vector:** $\mathbf{c} = (c_1, \cdots, c_n)$: $c_i = 1$ if $i \in AML$, and 0 if $i \in ALL$.
- **Gene-Class Correlation:**
  $P(g, c) = (\mu_1(g) - \mu_2(g))/(\sigma_1(g) + \sigma_2(g))$ for each gene $g$.
- **Nhood:** $N_1(c, r) = \{g \mid P(g, c) = r\}$
- **Permutation test:** compare with $N_1(c^*, r)$, for 400 permutations.
- **Number of informative genes** is a free parameter, chosen to be 50: 25 top-most and 25 bottom-most.
  - Robustness to noise
  - Ease of applicability.
Predictor Design:

\[ v_g = a_g(x_g - b_g) \] where \( a_g = P(g, c), \)
\[ b_g = \left[ \mu_1(g) + \mu_2(g) \right]/2, \]
Predictor Design:

- \( v_g = a_g(x_g - b_g) \) where \( a_g = P(g, c) \),
  \( b_g = [\mu_1(g) + \mu_2(g)]/2 \),
- \( x_g \) normalized log expression level of gene \( g \) in sample \( x \)
In symbols cont’d

Predictor Design:

- $v_g = a_g(x_g - b_g)$ where $a_g = P(g, c)$,
  $b_g = [\mu_1(g) + \mu_2(g)]/2$,
- $x_g$ normalized log expression level of gene $g$ in sample $x$,
- $v_g \geq 0$ means $g$ votes for class 1 (AML) and $v_g < 0$ means $g$ votes for class 2 (ALL).
Predictor Design:

- \( v_g = a_g(x_g - b_g) \) where \( a_g = P(g, c) \),
  \( b_g = [\mu_1(g) + \mu_2(g)]/2 \),
- \( x_g \) normalized log expression level of gene \( g \) in sample \( x \)
- \( v_g \geq 0 \) means \( g \) votes for class 1 (AML) and \( v_g < 0 \) means \( g \) votes for class 2 (ALL).
- \( V_1 = \sum_{g \in IG} v_g \) for \( v_g \geq 0 \), and
Predictor Design:

- \( v_g = a_g(x_g - b_g) \) where \( a_g = P(g, c) \),
  \( b_g = \left[ \mu_1(g) + \mu_2(g) \right]/2 \),
- \( x_g \) normalized log expression level of gene \( g \) in sample \( x \)
- \( v_g \geq 0 \) means \( g \) votes for class 1 (AML) and \( v_g < 0 \) means \( g \) votes for class 2 (ALL).
- \( V_1 = \sum_{g \in IG} v_g \) for \( v_g \geq 0 \), and
- \( V_2 = \sum_{g \in IG} |v_g| \) for \( v_g < 0 \).
In symbols cont’d

Predictor Design:

- \( v_g = a_g(x_g - b_g) \) where \( a_g = P(g, c) \),
  \( b_g = \left[ \mu_1(g) + \mu_2(g) \right] / 2 \),
- \( x_g \) normalized log expression level of gene \( g \) in sample \( x \)
- \( v_g \geq 0 \) means \( g \) votes for class 1 (AML) and \( v_g < 0 \) means \( g \) votes for class 2 (ALL).
- \( V_1 = \sum_{g \in IG} v_g \) for \( v_g \geq 0 \), and
- \( V_2 = \sum_{g \in IG} |v_g| \) for \( v_g < 0 \).
- \( PS = (V_{win} - V_{lose})/(V_{win} + V_{lose}) \)
In symbols cont’d

Predictor Design:

- \( v_g = a_g(x_g - b_g) \) where \( a_g = P(g, c) \),
  \( b_g = [\mu_1(g) + \mu_2(g)]/2 \),

- \( x_g \) normalized log expression level of gene \( g \) in sample \( x \)

- \( v_g \geq 0 \) means \( g \) votes for class 1 (AML) and \( v_g < 0 \) means \( g \) votes for class 2 (ALL).

- \( V_1 = \sum_{g \in IG} v_g \) for \( v_g \geq 0 \), and

- \( V_2 = \sum_{g \in IG} |v_g| \) for \( v_g < 0 \).

- \( PS = (V_{win} - V_{lose})/(V_{win} + V_{lose}) \)

- \( V_1 > V_2 \) with \( PS > 0.3 \) means \( x \in AML \), if \( PS \leq 0.3 \) then uncertain.
Cross-validation: 36 were assigned classes (with 100%) accuracy and 2 were uncertain. Median $PS = 0.77$
Cross-validation: 36 were assigned classes (with 100%) accuracy and 2 were uncertain. Median $PS = 0.77$

Test data (34 samples): 24 bone marrow and 10 peripheral blood samples, 20 ALL, 14 AML. Result: 29 predicted with 100% accuracy and 5 uncertain. Median $PS = 0.73$. 
View each sample as a 6817-dimensional vector and cluster samples.
Example cont’d, clustering

- View each sample as a 6817-dimensional vector and cluster samples.
- 2-SOM on 38 samples:
  $A_1$: 24 ALL, 1 AML and $A_2$: 10 AML, 3 ALL.
4-SOM on the same samples:

- $B_1$: 10 AML, $B_2$: 8 T-ALL, 1 B-ALL
- $B_3$: 5 B-ALL, $B_4$: 13 B-ALL, 1 AML.
Example cont’d, clustering

- How to evaluate clusters to see if they represent true biological structure?
How to evaluate clusters to see if they represent true biological structure?

Idea: true structure implies more accurate predictor.
How to evaluate clusters to see if they represent true biological structure?

Idea: true structure implies more accurate predictor.

So design predictors based on clustering classes: leads to merging $B_3$ and $B_4$. 
Goal: Look at (Gene Set)-Class correlation instead of Gene-Class correlation.
Enrichment Analysis

- **Goal:** Look at (Gene Set)-Class correlation instead of Gene-Class correlation.

- **Motivation:**
  - **Mootha 03:** No single gene is significantly differentially expressed, yet sets of genes might express differentially.
  - **Subramanian 05:** 1. Robustness to different sites, and 2. Integrating biological knowledge.
Enrichment Analysis

- **Goal:** Look at (Gene Set)-Class correlation instead of Gene-Class correlation.

- **Motivation:**
  - **Mootha 03:** No single gene is significantly differentially expressed, yet sets of genes might express differentially.
  - **Subramanian 05:** 1. Robustness to different sites, and 2. Integrating biological knowledge.

- **Methods:**
  - GSEA (A. Subramanian, PNAS 2005). Lung adenocarcinoma with good/poor outcome. \(|S_B \cap S_M| = 12\) and \(|S_B \cap S_M \cap S_S| = 1\) whereas \(S_B\) in \(M\) was NES = 1.9, \(p < 0.001\) and \(S_M\) in \(B\) was NES=2.13, \(p < 0.001\).
Enrichment Analysis

- **Goal:** Look at (Gene Set)-Class correlation instead of Gene-Class correlation.

- **Motivation:**
  - **Mootha 03:** No single gene is significantly differentially expressed, yet sets of genes might express differentially.
  - **Subramanian 05:** 1. Robustness to different sites, and 2. Integrating biological knowledge.

- **Methods:**
  - GSEA (A. Subramanian, PNAS 2005). Lung adenocarcinoma with good/poor outcome. $|S_B \cap S_M| = 12$ and $|S_B \cap S_M \cap S_S| = 1$ whereas $S_B$ in $M$ was NES = 1.9, $p < 0.001$ and $S_M$ in $B$ was NES=2.13, $p < 0.001$.
  - Tibshirani and Efron
  - R. Gentleman (Bioconductor)
  - Module maps, a refinement of GSEA, gene set minimization (Segal et al. Nature Genetics, 04,05)
Tools

- Bioconductor (R)
Tools

- Bioconductor (R)
- BRB ArrayTools (NCI)
Tools

- Bioconductor (R)
- BRB ArrayTools (NCI)
- GenePattern, includes GeneCluster (Broad Institute)
- Connectivity Map (Broad Institute)
- Benchmark data sets
Tools

- Bioconductor (R)
- BRB ArrayTools (NCI)
- GenePattern, includes GeneCluster (Broad Institute)
- Connectivity Map (Broad Institute)
Tools

- Bioconductor (R)
- BRB ArrayTools (NCI)
- GenePattern, includes GeneCluster (Broad Institute)
- Connectivity Map (Broad Institute)
- Benchmark data sets
Some ideas ...

- Integrating biological knowledge into the mathematical models (e.g., (Conditional) Markov Random Fields, Regulatory networks, ...)
Some ideas ...

▶ Integrating biological knowledge into the mathematical models (e.g., (Conditional) Markov Random Fields, Regulatory networks, ...)
▶ Formalizing clustering algorithms
Some ideas ...

- Integrating biological knowledge into the mathematical models (e.g., (Conditional) Markov Random Fields, Regulatory networks, ...)
- Formalizing clustering algorithms
- Technology transfer from: Time-series analysis of financial data, VLDB, Theoretical Neuroscience,...
My interests

- Dimensionality reduction, especially manifold learning.
My interests

- Dimensionality reduction, especially manifold learning.
- Harmonic analysis (on graphs): eigenfunctions of graph Laplacian, diffusion wavelets.