Molecular Classification of Biological Phenotypes

Esfandiar Haghverdi

School of Informatics

November 13, 2008

Outline

- Introduction
- Class Comparison
- Class Discovery
- Class Prediction
- Example
- Biological states and state modulation
- Chemical Genomics
- Toxicogenomics
- Software Tools
- Ideas

Introduction

- Welcome to Genomics Data Mining (GDM) working group.

Introduction

- Welcome to Genomics Data Mining (GDM) working group.
- Meetings will be bi-weekly starting today.

Introduction

- Welcome to Genomics Data Mining (GDM) working group.
- Meetings will be bi-weekly starting today.
- Extract meaningful information about the system being studied from gene expression data.

Introduction

- Welcome to Genomics Data Mining (GDM) working group.
- Meetings will be bi-weekly starting today.
- Extract meaningful information about the system being studied from gene expression data.
- The work is motivated by recent progress in cancer genomics.

Introduction

- Welcome to Genomics Data Mining (GDM) working group.
- Meetings will be bi-weekly starting today.
- Extract meaningful information about the system being studied from gene expression data.
- The work is motivated by recent progress in cancer genomics.
- This overview is by no means comprehensive.

Introduction

- Welcome to Genomics Data Mining (GDM) working group.
- Meetings will be bi-weekly starting today.
- Extract meaningful information about the system being studied from gene expression data.
- The work is motivated by recent progress in cancer genomics.
- This overview is by no means comprehensive.
- There is no one-size-fits-all solution for the analysis and interpretation of genome wide data.

Introduction

- Welcome to Genomics Data Mining (GDM) working group.
- Meetings will be bi-weekly starting today.
- Extract meaningful information about the system being studied from gene expression data.
- The work is motivated by recent progress in cancer genomics.
- This overview 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.

Introduction

- Welcome to Genomics Data Mining (GDM) working group.
- Meetings will be bi-weekly starting today.
- Extract meaningful information about the system being studied from gene expression data.
- The work is motivated by recent progress in cancer genomics.
- This overview 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.

Introduction

- Welcome to Genomics Data Mining (GDM) working group.
- Meetings will be bi-weekly starting today.
- Extract meaningful information about the system being studied from gene expression data.
- The work is motivated by recent progress in cancer genomics.
- This overview 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.

Introduction

- Welcome to Genomics Data Mining (GDM) working group.
- Meetings will be bi-weekly starting today.
- Extract meaningful information about the system being studied from gene expression data.
- The work is motivated by recent progress in cancer genomics.
- This overview 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 computational/statistical scientists on complex problems.

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.

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.

Central Papers

- Gene expression signatures: Molecular Classification of Cancer, T.R. Golub et al., *Science* **286**, 531, 1999.

Central Papers

- Gene expression signatures: Molecular Classification of Cancer, T.R. Golub et al., *Science* **286**, 531, 1999.
- GSEA: Gene Set Enrichment Analysis, A. Subramanian et al., *PNAS*, **102**, 2005.

Central Papers

- Gene expression signatures: Molecular Classification of Cancer, T.R. Golub et al., *Science* **286**, 531, 1999.
- GSEA: Gene Set Enrichment Analysis, A. Subramanian et al., *PNAS*, **102**, 2005.
- GE-HTS: Gene Expression-based High-throughput Screening, K. Stegmaier et al., *Nature genetics* **36**, 257, 2004.

Central Papers

- Gene expression signatures: Molecular Classification of Cancer, T.R. Golub et al., *Science* **286**, 531, 1999.
- GSEA: Gene Set Enrichment Analysis, A. Subramanian et al., *PNAS*, **102**, 2005.
- GE-HTS: Gene Expression-based High-throughput Screening, K. Stegmaier et al., *Nature genetics* **36**, 257, 2004.
- Connectivity Map: The Connectivity Map, J. Lamb et al., *Science* **313**, 1929, 2006.

Class Comparison

- *Goal:* Identify differentially expressed genes.

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, FWER vs FDR

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, FWER vs FDR
- *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)

Class Comparison

Opportunities:

Time-series analysis:

- Regulatory pathway inference
- Yeast cell cycle (Fourier transform, . . .)
- Model organism (e.g., Drosophila, Daphnia) development
- Analysis of samples (cells) exposed to different doses of the same drug
- Analysis of expression patterns from related bacterial strains

Class Discovery

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

Class Discovery

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

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, ...

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, ...
  - *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 cont'd

Issues:

- It is unbiased, no a priori assumption.

Class Discovery cont'd

Issues:

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

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.

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.

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?

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?
- There is no single best clustering method for a data set.

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?
- There is no single best clustering method for a data set.
- Some desirable properties maybe: Stability (reliability), predictive power, reduction power, ...

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?
- 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, ...)

Class Discovery cont'd

Opportunities:

- Stochastic clustering (e.g., NMF)

Class Discovery cont'd

Opportunities:

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

Class Discovery cont'd

Opportunities:

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

Class Discovery cont'd

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)

Class Discovery cont'd

Opportunities:

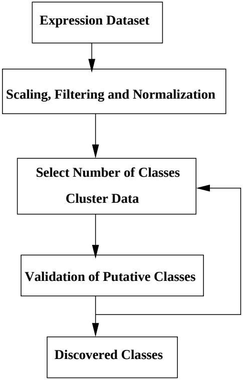
- Stochastic clustering (e.g., NMF)
- Techniques from statistical physics (e.g., Deterministic Annealing)
- Spectral methods (e.g., Diffusion maps on graphs)
- Geometric methods (e.g., Diffusion maps on manifolds)
- Information theoretic methods

Class Discovery cont'd

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)

Class Discovery Methodology



Class Prediction

- *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.

Class Prediction

- *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 $>>$ samples

Class Prediction, cont'd

Issues:

- features $>>$ samples
- Overfitting (modeling the training data too exactly)

Class Prediction, cont'd

Issues:

- features $>>$ samples
- Overfitting (modeling the training data too exactly)
- High level of noise

Class Prediction, cont'd

Issues:

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

Class Prediction, cont'd

Issues:

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

Class Prediction, cont'd

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, ...)

Class Prediction cont'd

Opportunities:

- Theory for method and classifier comparison

Class Prediction cont'd

Opportunities:

- Theory for method and classifier comparison
- Combining knowledge from different methods

Class Prediction cont'd

Opportunities:

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

Class Prediction cont'd

Opportunities:

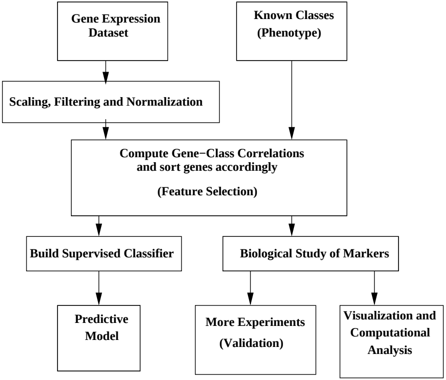
- Theory for method and classifier comparison
- Combining knowledge from different methods
- Incorporating knowledge from complementary sources
- Boosting the power and rigor of analysis

Class Prediction cont'd

Opportunities:

- Theory for method and classifier comparison
- Combining knowledge from different methods
- Incorporating knowledge from complementary sources
- Boosting the power and rigor of analysis
- Subpattern discovery, Califano et al. 99

Marker Selection Methodology



Gene Expression Dataset

Known Classes (Phenotype)

Scaling, Filtering and Normalization

Compute Gene−Class Correlations and sort genes accordingly

(Feature Selection)

Build Supervised Classifier

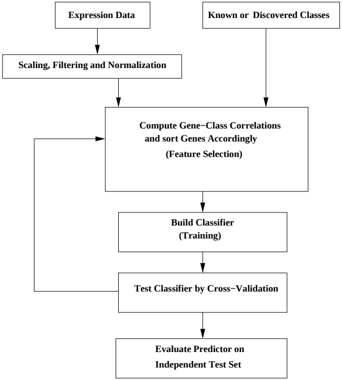
Biological Study of Markers

Predictive Model

More Experiments (Validation)

Visualization and Computational Analysis

Class Prediction Methodology



Expression Data

Known or Discovered Classes

Scaling, Filtering and Normalization

Compute Gene−Class Correlations
and sort Genes Accordingly
(Feature Selection)

Build Classifier
(Training)

Test Classifier by Cross−Validation

Evaluate Predictor on
Independent Test Set

Example (Golub et al. 1999)

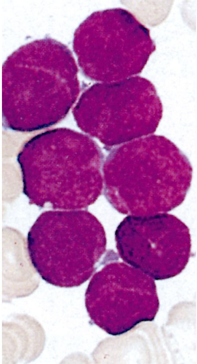


Figure 2-8a The Biology of Cancer (© Garland Science 2007)

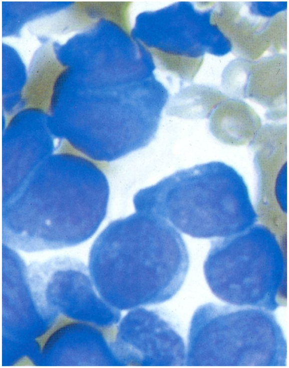
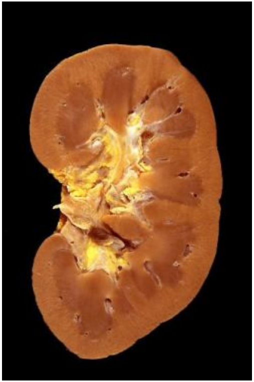
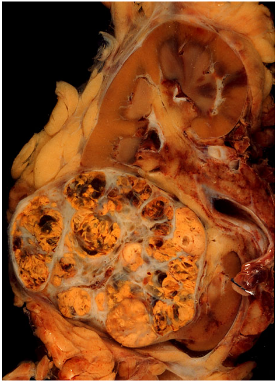


Figure 2-8b The Biology of Cancer (© Garland Science 2007)

ALL vs AML (The Biology of Cancer, R. Weinberg)

Compare to this!!





Normal kidney vs Renal cell carcinoma.

Example cont'd

- 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

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)

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

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.

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.

In symbols ...

Suppose we have n samples.

- Gene vector: $v(g) = (e_1, \cdots, e_n)$,

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 $i \in ALL$.

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: $P(g, c) = (\mu_1(g) - \mu_2(g))/(\sigma_1(g) + \sigma_2(g))$ for each gene g.

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: $P(g, c) = (\mu_1(g) - \mu_2(g))/(\sigma_1(g) + \sigma_2(g))$ for each gene g.
- Nhood: $N_1(c, r) = \{g \mid P(g, c) = r\}$

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: $P(g, c) = (\mu_1(g) - \mu_2(g))/(\sigma_1(g) + \sigma_2(g))$ 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.

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: $P(g, c) = (\mu_1(g) - \mu_2(g))/(\sigma_1(g) + \sigma_2(g))$ 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.

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: $P(g, c) = (\mu_1(g) - \mu_2(g))/(\sigma_1(g) + \sigma_2(g))$ 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

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: $P(g, c) = (\mu_1(g) - \mu_2(g))/(\sigma_1(g) + \sigma_2(g))$ 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.

In symbols cont'd

Predictor Design:

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

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

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).

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

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 $v_g < 0$.

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 $v_g < 0$.
- $PS = (V_{win} - V_{lose})/(V_{win} + V_{lose})$

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.

In symbols, cont'd

- Cross-validation: 36 were assigned classes (with 100%) accuracy and 2 were uncertain. Median $PS = 0.77$

In symbols, cont'd

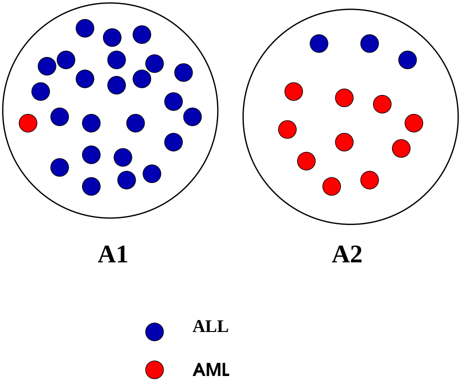
- Cross-validation: 36 were assigned classes (with 100%) accuracy and 2 were uncertain. Median $PS = 0.77$
- Test data (34 samples): 24 bone marrow and 10 peripheral blood samples, 20 ALL, 14 AML.
  Result: 29 predicted with 100% accuracy and 5 uncertain.
  Median $PS = 0.73$.

Example cont'd, clustering

- View each sample as a 6817-dimensional vector and cluster samples.

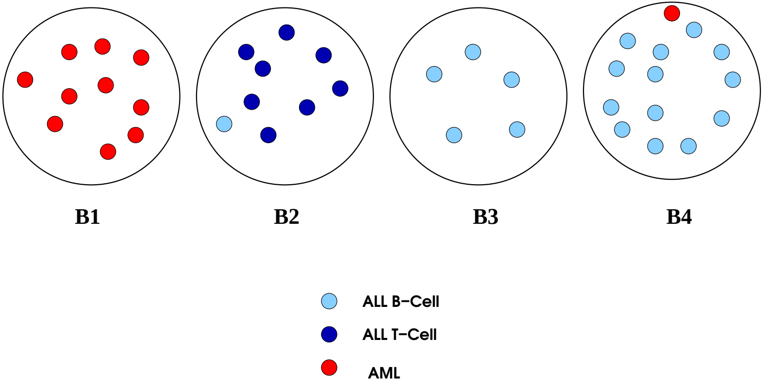
Example cont'd, clustering

- 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

- 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?

Example cont'd, clustering

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

Example cont'd, clustering

- 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.

Enrichment Analysis

- *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)

EA cont'd

Opportunities:

- Using BLAST theory to enhance the predictive power.

EA cont'd

Opportunities:

- Using BLAST theory to enhance the predictive power.
- Random walks on networks.

Chemical Genomics

- Generating large collections of small molecules and using them to modulate cellular states.

Chemical Genomics

- Generating large collections of small molecules and using them to modulate cellular states.
- One approach is to screen different compounds that induce state modulations, using signatures for the states.

Chemical Genomics

- Generating large collections of small molecules and using them to modulate cellular states.
- One approach is to screen different compounds that induce state modulations, using signatures for the states.
- GE-HTS (Stegmaier et al., Nature Genetics, 2004) 1,739 compounds are screened for AML/neutrophil and AML/monocyte terminal cell differentiation.

Chemical Genomics

- Generating large collections of small molecules and using them to modulate cellular states.
- One approach is to screen different compounds that induce state modulations, using signatures for the states.
- GE-HTS (Stegmaier et al., Nature Genetics, 2004) 1,739 compounds are screened for AML/neutrophil and AML/monocyte terminal cell differentiation.
- *Connectivity Map* (J. Lamb et al., Science 06):

  disease – gene – drug.

CG, cont'd

Issues:

- Necessary modifications when dealing with more heterogeneous situations, e.g., BC vs Leukemia.

CG, cont'd

Issues:

- Necessary modifications when dealing with more heterogeneous situations, e.g., BC vs Leukemia.
- Choice of cell type

CG, cont'd

Issues:

- Necessary modifications when dealing with more heterogeneous situations, e.g., BC vs Leukemia.
- Choice of cell type
- Measurement time

CG, cont'd

Issues:

- Necessary modifications when dealing with more heterogeneous situations, e.g., BC vs Leukemia.
- Choice of cell type
- Measurement time
- Concentration and treatment duration

CG, cont'd

Issues:

- Necessary modifications when dealing with more heterogeneous situations, e.g., BC vs Leukemia.
- Choice of cell type
- Measurement time
- Concentration and treatment duration
- Analytical methods for detecting relevant signals

Toxicogenomics

- Identification of potential human and environmental toxicants, and their putative mechanisms of action, through the use of genomics resources.

Toxicogenomics

- Identification of potential human and environmental toxicants, and their putative mechanisms of action, through the use of genomics resources.
- Gene expression is altered during toxicity.

Toxicogenomics

- Identification of potential human and environmental toxicants, and their putative mechanisms of action, through the use of genomics resources.
- Gene expression is altered during toxicity.
- Challenge: Given a set of experimental conditions, define the characteristic and specific pattern of gene expression elicited by a given toxicant. (Toxicant signatures!)

Toxicogenomics

- Identification of potential human and environmental toxicants, and their putative mechanisms of action, through the use of genomics resources.
- Gene expression is altered during toxicity.
- Challenge: Given a set of experimental conditions, define the characteristic and specific pattern of gene expression elicited by a given toxicant. (Toxicant signatures!)
- Given a model organism:
  Known toxicants → signatures → database → determine the action mechanism of an unknown toxicant.

Toxicogenomics

- Identification of potential human and environmental toxicants, and their putative mechanisms of action, through the use of genomics resources.
- Gene expression is altered during toxicity.
- Challenge: Given a set of experimental conditions, define the characteristic and specific pattern of gene expression elicited by a given toxicant. (Toxicant signatures!)
- Given a model organism:
  Known toxicants → signatures → database → determine the action mechanism of an unknown toxicant.
- Connectivity Map: gene – toxicant

TG, cont'd

Issues:

- Proper definition of signature similarity

TG, cont'd

Issues:

- Proper definition of signature similarity
- Model system selection

TG, cont'd

Issues:

- Proper definition of signature similarity
- Model system selection
- Dose selection

TG, cont'd

Issues:

- Proper definition of signature similarity
- Model system selection
- Dose selection
- Measurement time (Time series analysis?)

TG, cont'd

Issues:

- Proper definition of signature similarity
- Model system selection
- Dose selection
- Measurement time (Time series analysis?)
- Other factors: age, diet, etc

Tools

- Bioconductor

Tools

- Bioconductor
- BRB ArrayTools (NCI, Richard Simon)

Tools

- Bioconductor
- BRB ArrayTools (NCI, Richard Simon)
- GenePattern, includes GeneCluster (Broad Institute)

Tools

- Bioconductor
- BRB ArrayTools (NCI, Richard Simon)
- GenePattern, includes GeneCluster (Broad Institute)
- Connectivity Map (Broad Institute)

Tools

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

Tools

- Bioconductor
- BRB ArrayTools (NCI, Richard Simon)
- GenePattern, includes GeneCluster (Broad Institute)
- Connectivity Map (Broad Institute)
- Benchmark data sets
- Local implementations with interface to the tools above

Some ideas ...

- Technical improvements at different levels

Some ideas ...

- Technical improvements at different levels
- Integrating biological knowledge into the mathematical models (after Random Markov Fields)

Some ideas ...

- Technical improvements at different levels
- Integrating biological knowledge into the mathematical models (after Random Markov Fields)
- New mathematical tools

Some ideas ...

- Technical improvements at different levels
- Integrating biological knowledge into the mathematical models (after Random Markov Fields)
- New mathematical tools
- Using signatures to infer signalling pathways

Some ideas ...

- Technical improvements at different levels
- Integrating biological knowledge into the mathematical models (after Random Markov Fields)
- New mathematical tools
- Using signatures to infer signalling pathways
- Using signatures to improve clustering algorithms

Some ideas ...

- Technical improvements at different levels
- Integrating biological knowledge into the mathematical models (after Random Markov Fields)
- New mathematical tools
- Using signatures to infer signalling pathways
- Using signatures to improve clustering algorithms
- Technology transfer from: Time-series analysis of financial data, VLDB, Theoretical Neuroscience

Ideas, cont'd

- New biologically relevant and important questions:

Ideas, cont'd

- New biologically relevant and important questions:
- Daphnia based toxicogenomics

Ideas, cont'd

- New biologically relevant and important questions:
- Daphnia based toxicogenomics
- Daphnia based ecogenomics

Ideas, cont'd

- New biologically relevant and important questions:
- Daphnia based toxicogenomics
- Daphnia based ecogenomics
- Signatures in developmental stages of model organisms

Ideas, cont'd

- New biologically relevant and important questions:
- Daphnia based toxicogenomics
- Daphnia based ecogenomics
- Signatures in developmental stages of model organisms
- Time series analysis of environmental effects

Ideas, cont'd

- New biologically relevant and important questions:
- Daphnia based toxicogenomics
- Daphnia based ecogenomics
- Signatures in developmental stages of model organisms
- Time series analysis of environmental effects
- Other genomic signatures: DNA methylation patterns, microRNA profiles, metabolite profiles, ...