Analysis and Computational Dissection of Molecular Signature Multiplicity

Alexander Statnikov

12/17/2008
Research Area

- Biomedical Informatics
- Clinical Informatics
- Imaging Informatics
- Public Health Informatics
- Educational Informatics
- Information Retrieval
- Bioinformatics

### Areas of Focus

<table>
<thead>
<tr>
<th>Predictive modeling and molecular diagnostics</th>
<th>Genetics and population analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pathway discovery and systems biology</td>
<td>Analysis of protein expression</td>
</tr>
<tr>
<td>Protein structure/function prediction</td>
<td>Analysis of gene expression</td>
</tr>
<tr>
<td>Gene structure/function prediction</td>
<td>High-throughput image analysis</td>
</tr>
<tr>
<td>Sequence analysis</td>
<td>Software, tools, web services</td>
</tr>
<tr>
<td>Genome annotation</td>
<td>…</td>
</tr>
</tbody>
</table>
Molecular Signature Multiplicity

• Different methods or samples from the same population lead to different but apparently maximally predictive signatures;

• Far-reaching implications for biological discovery and development of next generation patient diagnostics and personalized treatments:
  – Generation of biological hypotheses is very hard even when signatures are maximally predictive of the phenotype since thousands of completely different signatures are equally consistent with the data;
  – Produced signatures are not statistically generalizable to new cases, and thus not reliable enough for translation to clinical practice.
Molecular Signature Multiplicity

• Causes of this phenomenon are unknown; several contradictory conjectures exist in the field:
  – Signature multiplicity is due to small samples [Michiels et al., 2005]
  – Signature multiplicity leads to predictively non-reproducible signatures [Ein-Dor et al., 2006]; building reproducible signatures requires thousands of samples [Ioannidis, 2005]
  – Signature multiplicity is a by-product of the complex regulatory connectivity of genome [Dougherty and Brun, 2006]
  – Artifacts of data pre-processing, e.g. normalization [Gold et al., 2005; Qiu et al., 2005; Ploner et al., 2005]
Major Goals

1. Develop a Markov boundary characterization of molecular signature multiplicity phenomenon;
2. Design and study algorithms that can correctly identify the set of maximally predictive and non-redundant molecular signatures;
3. Conduct an empirical evaluation of the novel algorithms and compare to the existing state-of-the-art methods;
4. Test and refine previously stated hypotheses about the causes of signature multiplicity phenomenon.
Optimality Criteria of Signatures

Signatures that are focus of this project satisfy the following two optimality criteria:

- *maximally predictive of the phenotype* (they achieve best predictivity of the phenotype in the given dataset over all signatures based on different gene sets);
- *do not contain predictively redundant genes* (i.e., genes that can be removed from the signature without adversely affecting its predictivity).
Why Do We Need Algorithms to Extract As Many Optimal Signatures As Possible?

1. A deeper understanding of the signature multiplicity phenomenon and how it affects reproducibility of signatures;

2. Improving discovery of the underlying biological mechanisms by not missing genes that are implicated biologically in disease processes;

3. Catalyzing regulatory approval by establishing *in-silico equivalence* to previously validated signatures
Existing Algorithms for Multiple Signature Extraction: *Resampling-Based Methods*

1) Generate resampled datasets (e.g., by bootstrapping)

2) Apply a standard signature extraction algorithm (e.g., SVM-RFE)

- Based on assumption that multiplicity is strictly a small-sample phenomenon;
- An infinite number of resamplings is required to extract all optimal signatures;
- May stop producing multiple signatures in large sample sizes.
Existing Algorithms for Multiple Signature Extraction: *Iterative Removal*

- **Original data** (for all genes)
- **Reduced data** (excluding $X_1$ genes)
- **Reduced data** (excluding $X_1$ and $X_2$ genes)

Remove corresponding genes from the dataset

...until a signature has statistically significantly reduced predictivity

- Agnostic to what causes molecular signature multiplicity;
- Cannot discover signatures that have genes in common.
Existing Algorithms for Multiple Signature Extraction: *Stochastic Gene Selection*

**Genetic Algorithms (e.g., GA/KNN or GA/SVM)**

- Can output all signatures that are discoverable by a genetic algorithm when it is allowed to evolve an infinite number of generations.

**KIAMB**

- Stochastic Markov boundary method based on IAMB algorithm;
- In a specific class of distributions, every optimal signature will be output by this method with nonzero probability;
- Requires an infinite number of iterations to discover all optimal signatures; will discover same signature over and over again;
- Sample requirements are of exponential order to the number of genes in a signatures.
Existing Algorithms for Multiple Signature Extraction: *Brute-Force Exhaustive Search*

**LIKNON**

- Examines predictivity of all individual genes in the dataset, all pairs of genes, all triples of genes, and so on;
- It is infeasible when a signature has more than 2-3 genes;
- Agnostic to what causes signature multiplicity.

*In summary, no current algorithm provides a systematic and efficient approach for identification of the set of maximally predictive and non-redundant molecular signatures that exist in the underlying distribution.*
1. Markov Boundary Characterization of Molecular Signature Multiplicity
Key Definitions (1/2)

- **Definition of molecular signature:** A molecular signature is a mathematical/computational model (e.g., classifier or regression model) that predicts a phenotype of interest (e.g., diagnosis or response to treatment in human patients) given values of molecular variables (e.g., gene expression values).

- **Definition of maximally predictive molecular signature:** A maximally predictive molecular signature is a molecular signature that maximizes predictivity of the phenotype relative to all other signatures that can be constructed from the same dataset.

- **Definition of maximally predictive and non-redundant molecular signature:** A maximally predictive and non-redundant molecular signature based on variables \( \mathbf{X} \) is a maximally predictive signature such that any signature based on a proper subset of variables in \( \mathbf{X} \) is not maximally predictive.
• **Definition of Markov blanket:** A Markov blanket $\mathcal{M}$ of the response variable $T \in \mathbf{V}$ in the joint probability distribution $\mathbb{P}$ over variables $\mathbf{V}$ is a set of variables conditioned on which all other variables are independent of $T$, i.e. for every $X \in (\mathbf{V} \setminus \mathcal{M} \setminus \{T\})$, $T \perp X \, | \, \mathcal{M}$.

• **Definition of Market boundary (or non-redundant Markov blanket):** If $\mathcal{M}$ is a Markov blanket of $T$ and no proper subset of $\mathcal{M}$ satisfies the definition of Markov blanket of $T$, then $\mathcal{M}$ is called a *Markov boundary* (or *non-redundant Markov blanket*) of $T$. 
Theoretical Results

- Variable sets that participate in the maximally predictive signatures of $T$ are precisely the Markov blankets of $T$ and vice-versa;
- Similarly, variable sets that participate in the maximally predictive and non-redundant signatures of $T$ are precisely the Markov boundaries of $T$ and vice-versa;
- If a joint probability distribution $P$ over variables $V$ satisfies the intersection property*, then there exists a unique Markov boundary of $T$ [Pearl, 1988].

\[
X \perp Y \mid (Z \cup W) \quad \text{and} \quad X \perp W \mid (Z \cup Y) \implies X \perp (Y \cup W) \mid Z
\]
A Fundamental Reduction Used in This Project for the Analysis of Signatures

Signatures that have maximal predictivity of the phenotype relative to their genes.

- Since there is an infinite number of signatures with maximal predictivity, when I refer to a signature, I mean one of the predictively equivalent classifiers (e.g., $S_3$ or $S_4$ or $S_5$);
- Can study signature classes by reference only to their genes;
- This reduction is justified whenever the classifiers used can learn the minimum error decision function given sufficient sample.
Example of Markov Boundary Multiplicity

1. Many optimal signatures exist: e.g., \{A, C\} and \{B, C\} are maximally predictive and non-redundant signatures of \(T\). Furthermore, \{A, C\} and \{B, C\} remain maximally predictive even in infinite samples;
2. The network has very low connectivity;
3. Genes in optimal signatures do not have to be deterministically related: e.g., \(A\) and \(B\) are not deterministically related, yet convey individually the same information about \(T\);
4. If an algorithm selects only one optimal signature, then there is danger to miss biologically important causative genes;
5. The union of all optimal signatures includes all genes located in the local pathway around \(T\);
6. In this example the intersection of all optimal signatures contains only genes in the local pathway around \(T\).
2. A Novel Algorithm to Correctly Identify The Set Of Maximally Predictive and Non-Redundant Signatures
Generative algorithm TIE*

Inputs:
- dataset $\mathcal{D}$ (a sample of distribution $P$) for variables $\mathbf{V}$, including a response variable $T$;
- Markov boundary algorithm $\mathcal{X}$;
- strategy $\mathcal{Y}$ to generate subsets of variables that have to be removed to identify new Markov boundaries of $T$;
- criterion $\mathcal{Z}$ to verify Markov boundaries of $T$.

Output: all Markov boundaries (i.e., maximally predictive and non-redundant signatures) of $T$.

1. Use algorithm $\mathcal{X}$ to learn a Markov boundary $\mathbf{M}$ of $T$ from data $\mathcal{D}$ for variables $\mathbf{V}$ (i.e., in the original distribution)
2. Output $\mathbf{M}$
3. Repeat
4. Use strategy $\mathcal{Y}$ to generate a subset of variables $\mathbf{G}$ whose removal may lead to identification of a new Markov boundary of $T$
5. Use algorithm $\mathcal{X}$ to learn a Markov boundary $\mathbf{M}_{new}$ of $T$ from data $\mathcal{D}$ for variables $\mathbf{V} \setminus \mathbf{G}$ (i.e., in the embedded distribution)
6. If $\mathbf{M}_{new}$ is a Markov boundary of $T$ in the original distribution according to criterion $\mathcal{Z}$, output $\mathbf{M}_{new}$
7. Until all subsets $\mathbf{G}$ generated by strategy $\mathcal{Y}$ have been considered.
TIE* Algorithm for Gene Expression Data Analysis

An example of instantiated algorithm TIE* for gene expression data analysis

**Inputs:** dataset $D$ (a sample of distribution $P$) for variables $V$, including a response variable $T$.

**Output:** all Markov boundaries (i.e., maximally predictive and non-redundant signatures) of $T$.

1. Use algorithm HITON-PC to learn a Markov boundary $M$ of $T$ from data $D$ for variables $V$ (i.e., in the *original* distribution)
2. Output $M$
3. Repeat
4. Generate the smallest subset $G$ of the so far discovered Markov boundaries of $T$ such that (i) it was not considered in the previous iteration of the algorithm, and (ii) it does not properly include any subset that was generated in the previous iteration of the algorithm when $M_{new}$ was found not to be a Markov boundary of $T$
5. Use algorithm HITON-PC to learn a Markov boundary $M_{new}$ of $T$ from data $D$ for variables $V \setminus G$ (i.e., in the *embedded* distribution)
6. If the holdout validation estimate of predictivity of $T$ for the SVM classifier model induced from data $D$ using variables $M_{new}$ is statistically indistinguishable from the respective predictivity estimate for variables $M$, then $M_{new}$ is a Markov boundary of $T$ in the original distribution and it is output by the algorithm
7. Until no subset $G$ can be generated in line 4.
Trace of the TIE* Algorithm

\[ G = \{F\} \]
\[ M = \{A, B, F\} \]
\[ G = \{A\} \]
\[ M_{new} = \{A, B\} \]
\[ G = \{B\} \]
\[ M_{new} = \{C, B, F\} \]
\[ G = \{A, B\} \]
\[ M_{new} = \{A, D, E, F\} \]
\[ G = \{A, B\} \]
\[ M_{new} = \{C, D, E, F\} \]

Not a Markov boundary; Do not consider any \( G \) that is a superset of \{F\}
Markov boundary
Markov boundary
Markov boundary
Theoretical Results (1/2)

- **TIE\(^*\)** returns all and only Markov boundaries of \( T \) (i.e., maximally predictive and non-redundant signatures) if its input components \( X, Y, Z \) are admissible.

- **IAMB** is an admissible Markov boundary algorithm (input component \( X \)) under assumptions:
  - IAMB correctly outputs a Markov boundary if only the composition property holds.

- **HITON-PC** is an admissible Markov boundary algorithm (input component \( X \)) under assumptions:
  - HITON-PC correctly outputs a Markov boundary if the adjacency faithfulness assumption holds except for violations of the intersection axiom, global Markov condition holds, and there are no “spouses” in the Markov boundary.
Theoretical Results (2/2)

- Stated three strategies (*IncLex*, *IncMinAssoc*, and *IncMaxAssoc*) to generate subsets of variables that have to be removed from \( V \) to identify new Markov boundaries of \( T \) and proved their admissibility (input component \( \mathbb{Y} \))

- Stated two criteria (*Independence* and *Predictivity*) to verify Markov boundaries and proved their admissibility (input component \( \mathbb{Z} \))
A. Experiments with Artificial Simulated Data

Generative model is available, and the set of Markov boundaries (and thus the set of maximally predictive and non-redundant signatures) is known.

- Generate samples of systematically varied sizes;
- Compare to the gold standard;
- Test whether the TIE* algorithm behaves according to theoretical expectations and study its empirical properties;
- Obtain clues about behavior of TIE* and baseline comparison algorithms in experiments with real gene expression data.
Two artificial discrete networks were created:

- **TIED1** consists of 30 variables (including a response variable $T$) and contains 72 Markov boundaries of $T$;
- **TIED2** consists of 1,000 variables (including a response variable $T$) and contains the same 72 Markov boundaries of $T$ as TIED1.
Experiments

**Goal:** Compare TIE* to state-of-the-art algorithms (*Resampling-based methods*, KIAMB, and *Iterative Removal*) and examine sensitivity of the tested methods to high dimensionality.

**Findings:**

- TIE* correctly identifies the set of true Markov boundaries (maximally predictive and non-redundant signatures) in the datasets with 30 or 1,000 variables;
- Iterative Removal identifies only 1 signature;
- KIAMB fails to identify any true signature, and its output signatures have poor predictivity;
- Resampling-based methods either miss true signatures and/or output many redundant variables in the signatures.
Experiments with Linear Continuous Network \textit{LIND}

\textit{LIND} consists of 41 variables (including a response variable \(T\)) and contains 12 Markov boundaries of \(T\).
Goals:
1. Analyze behavior of TIE* as a function of sample size using data generated from a continuous network;
2. Compare criteria Independence and Predictivity for verification of Markov boundaries in the TIE* algorithm.

Findings:
• As sample size increases, the performance of both instantiations of TIE* generally improves and the algorithms discover the set of true Markov boundaries;
• \( \alpha \)-level in the criterion Predictivity significantly affects the number of Markov boundaries output by the TIE* algorithm;
• TIE* with criterion Predictivity typically leads to a larger number of output Markov boundaries and on average superior performance compared to criterion Independence.
Experiments with Discrete Network

**XORD**

XORD consists of 41 variables (including a response variable $T$) and contains 25 Markov boundaries of $T$. 
**Experiments**

**Goal:** Evaluate TIE* when the popular Markov boundary algorithms such as *IAMB* and *HITON-PC* are not applicable due to violations of their fundamental assumptions.

**Findings:**
- TIE* discovers the set of true Markov boundaries when the sample is $\geq 2,000$;
- There is $\sim 1$ false positive variable in each discovered Markov boundary for large sample sizes.
B. Experiments with Resimulated Microarray Gene Expression Data

- Resimulated data by design closely resembles real human lung cancer microarray gene expression data;
- The knowledge of a generative model allows to generate arbitrary large samples and study behavior of TIE* as a function of sample size;
- Unlike prior experiments with artificial simulated datasets, the set of maximally predictive and non-redundant signatures is not known \textit{a priori}. 
Goal: Examine whether the signature multiplicity phenomenon vanishes as the sample size grows.

Results:
Findings of Other Experiments

• TIE* is not sensitive to the choice of the initial signature discovered by the algorithm;
• Post-processing TIE* signatures with wrapping results in more signatures with smaller number of genes;
• Signatures output by tested non-TIE* methods are either redundant or have inferior predictivity compared to signatures output by TIE* techniques.
C. Experiments with Real Human Microarray Gene Expression Data

- **Independent-Dataset Experiments:** Using pairs of microarray datasets either from different laboratories or different platforms;
- **Single-Dataset Experiments:** Additional experiments with relatively large sample size microarray datasets;
- The primary goal of both experiments is to compare TIE* and baseline algorithms for multiple signature extraction in terms of maximal predictivity\# of induced signatures and reproducibility in independent data.
- **Operational definition of “maximal predictivity”:** Empirically best classification performance (AUC) achievable in each dataset over all tested methods consideration.
## Independent-Dataset Experiment: Datasets

<table>
<thead>
<tr>
<th>Task</th>
<th>Discovery dataset</th>
<th>Validation dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Lung Cancer Diagnosis</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lung tumors vs. normals (non-tumor lung samples)</td>
<td>Sample size: 203, Samples per class: lung tumors (186) normals (17) Number of genes: 12600 Microarray platform: Affymetrix U95A</td>
<td>Sample size: 96, Samples per class: lung tumors (86) normals (10) Number of genes: 7129 Microarray platform: Affymetrix HuGeneFL</td>
</tr>
<tr>
<td><strong>Lung Cancer Subtype</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Classification: adenocarcinoma vs. squamous cell carcinoma lung tumors</td>
<td>Sample size: 160, Samples per class: adenocarcinoma (139) squamous (21) Number of genes: 12600 Microarray platform: Affymetrix U95A</td>
<td>Sample size: 28, Samples per class: adenocarcinoma (14) squamous (14) Number of genes: 12533 Microarray platform: Affymetrix U95A</td>
</tr>
<tr>
<td><strong>Breast Cancer Subtype</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Classification: estrogen receptor positive (ER+) vs. ER- breast tumors; untreated patients</td>
<td>Sample size: 286, Samples per class: ER+ (209) ER- (77) Number of genes: 22283 Microarray platform: Affymetrix U133A</td>
<td>Sample size: 119, Samples per class: ER+ (85) ER- (34) Number of genes: 22283 Microarray platform: Affymetrix U133A</td>
</tr>
<tr>
<td><strong>Breast Cancer 5 Yr. Prognosis</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ER+ patients who developed distant metastases within 5 years (poor prognosis) vs. ones who did not (good prognosis)</td>
<td>Sample size: 204, Samples per class: poor prognosis (66) good prognosis (138) Number of genes: 22283 Microarray platform: Affymetrix U133A</td>
<td>Sample size: 72, Samples per class: poor prognosis (13) good prognosis (59) Number of genes: 22283 Microarray platform: Affymetrix U133A</td>
</tr>
<tr>
<td><strong>Glioma Subtype Classification</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>grade III vs. grade IV glioma tumors</td>
<td>Sample size: 100, Samples per class: grade III (24) grade IV (76) Number of genes: 22283 Microarray platform: Affymetrix U133A</td>
<td>Sample size: 85, Samples per class: grade III (26) grade IV (59) Number of genes: 22283 Microarray platform: Affymetrix U133A</td>
</tr>
<tr>
<td><strong>Leukemia 5 Yr. Prognosis</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>patients with disease-free survival &lt; 5 years (ones who had relapse or competing events within 5 years) vs. &gt; 5 years</td>
<td>Sample size: 164, Samples per class: survival &lt; 5 yr. (29) survival &gt; 5 yr. (135) Number of genes: 12625 Microarray platform: Affymetrix U95A</td>
<td>Sample size: 79, Samples per class: survival &lt; 5 yr. (18) survival &gt; 5 yr. (61) Number of genes: 22283 Microarray platform: Affymetrix U133A</td>
</tr>
</tbody>
</table>
## Detailed Results (1/3)

### Lung Cancer Diagnosis

<table>
<thead>
<tr>
<th>Method to induce multiple signatures</th>
<th>Number of signatures</th>
<th>Number of genes in a signature</th>
<th>Classification performance (AUC)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>mean</td>
<td>95% interval</td>
</tr>
<tr>
<td>TIE*</td>
<td>348/348/187</td>
<td>6</td>
<td>[4 - 8]</td>
</tr>
<tr>
<td>Resampling+SVM-RFE1</td>
<td>5000/2966/48</td>
<td>9</td>
<td>[1 - 43]</td>
</tr>
<tr>
<td>Resampling+SVM-RFE2</td>
<td>5000/341/61</td>
<td>1</td>
<td>[1 - 2]</td>
</tr>
<tr>
<td>Resampling+Univariate1</td>
<td>5000/2199/19</td>
<td>19</td>
<td>[1 - 62]</td>
</tr>
<tr>
<td>Resampling+Univariate2</td>
<td>5000/294/58</td>
<td>1</td>
<td>[1 - 2]</td>
</tr>
<tr>
<td>KIAMB1</td>
<td>985/985/985</td>
<td>41</td>
<td>[39 - 42]</td>
</tr>
<tr>
<td>KIAMB2</td>
<td>1489/1320/1246</td>
<td>48</td>
<td>[12 - 68]</td>
</tr>
<tr>
<td>KIAMB3</td>
<td>5000/271/157</td>
<td>9</td>
<td>[6 - 15]</td>
</tr>
<tr>
<td>Iterative Removal</td>
<td>51/51/51</td>
<td>7</td>
<td>[5 - 10]</td>
</tr>
</tbody>
</table>

### Lung Cancer Subtype Classification

<table>
<thead>
<tr>
<th>Method to induce multiple signatures</th>
<th>Number of signatures</th>
<th>Number of genes in a signature</th>
<th>Classification performance (AUC)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>mean</td>
<td>95% interval</td>
</tr>
<tr>
<td>TIE*</td>
<td>668/668/413</td>
<td>7</td>
<td>[5 - 8]</td>
</tr>
<tr>
<td>Resampling+SVM-RFE1</td>
<td>5000/4267/20</td>
<td>392</td>
<td>[1 - 5037]</td>
</tr>
<tr>
<td>Resampling+SVM-RFE2</td>
<td>5000/1206/107</td>
<td>2</td>
<td>[1 - 5]</td>
</tr>
<tr>
<td>Resampling+Univariate1</td>
<td>5000/4590/55</td>
<td>528</td>
<td>[1 - 8703]</td>
</tr>
<tr>
<td>Resampling+Univariate2</td>
<td>5000/917/81</td>
<td>3</td>
<td>[1 - 6]</td>
</tr>
<tr>
<td>KIAMB1</td>
<td>994/968/965</td>
<td>26</td>
<td>[24 - 26]</td>
</tr>
<tr>
<td>KIAMB2</td>
<td>1006/1005/1005</td>
<td>48</td>
<td>[47 - 50]</td>
</tr>
<tr>
<td>KIAMB3</td>
<td>3520/1364/1209</td>
<td>16</td>
<td>[8 - 31]</td>
</tr>
<tr>
<td>Iterative Removal</td>
<td>29/29/29</td>
<td>8</td>
<td>[5 - 12]</td>
</tr>
</tbody>
</table>
**Detailed Results (2/3)**

### Breast Cancer Subtype Classification

<table>
<thead>
<tr>
<th>Method to induce multiple signatures</th>
<th>Number of signatures</th>
<th>Number of genes in a signature</th>
<th>Classification performance (AUC)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean</td>
<td>95% interval</td>
<td>In discovery dataset</td>
</tr>
<tr>
<td></td>
<td>mean</td>
<td>95% interval</td>
<td>mean</td>
</tr>
<tr>
<td>TIE*</td>
<td>2776/2776/1602</td>
<td>17 [14 - 21]</td>
<td>0.847 [0.824 - 0.873]</td>
</tr>
<tr>
<td>Resampling+SVM-RFE1</td>
<td>5000/4601/22</td>
<td>1627 [1 - 10746]</td>
<td>0.845 [0.821 - 0.888]</td>
</tr>
<tr>
<td>Resampling+SVM-RFE2</td>
<td>5000/2033/65</td>
<td>18 [1 - 135]</td>
<td>0.858 [0.736 - 0.930]</td>
</tr>
<tr>
<td>Resampling+Univariate1</td>
<td>5000/4122/15</td>
<td>3560 [1 - 22283]</td>
<td>0.857 [0.826 - 0.920]</td>
</tr>
<tr>
<td>Resampling+Univariate2</td>
<td>5000/794/22</td>
<td>7 [1 - 18]</td>
<td>0.873 [0.754 - 0.930]</td>
</tr>
<tr>
<td>KIAMB1</td>
<td>983/970/960</td>
<td>31 [30 - 32]</td>
<td>0.8 [0.804 - 0.883]</td>
</tr>
<tr>
<td>KIAMB2</td>
<td>994/964/962</td>
<td>28 [27 - 29]</td>
<td>0.85 [0.802 - 0.884]</td>
</tr>
<tr>
<td>KIAMB3</td>
<td>943/570/493</td>
<td>14 [12 - 15]</td>
<td>0.856 [0.786 - 0.884]</td>
</tr>
<tr>
<td>Iterative Removal</td>
<td>34/34/34</td>
<td>19 [14 - 23]</td>
<td>0.833 [0.793 - 0.866]</td>
</tr>
</tbody>
</table>

### Breast Cancer 5 Yr. Prognosis

<table>
<thead>
<tr>
<th>Method to induce multiple signatures</th>
<th>Number of signatures</th>
<th>Number of genes in a signature</th>
<th>Classification performance (AUC)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean</td>
<td>95% interval</td>
<td>In discovery dataset</td>
</tr>
<tr>
<td></td>
<td>mean</td>
<td>95% interval</td>
<td>mean</td>
</tr>
<tr>
<td>TIE*</td>
<td>5342/5342/3321</td>
<td>84 [81 - 89]</td>
<td>0.671 [0.658 - 0.686]</td>
</tr>
<tr>
<td>Resampling+SVM-RFE1</td>
<td>5000/4755/42</td>
<td>4687 [2 - 22283]</td>
<td>0.684 [0.541 - 0.746]</td>
</tr>
<tr>
<td>Resampling+SVM-RFE2</td>
<td>5000/3407/350</td>
<td>56 [1 - 404]</td>
<td>0.586 [0.413 - 0.719]</td>
</tr>
<tr>
<td>Resampling+Univariate1</td>
<td>5000/4002/29</td>
<td>5791 [1 - 22283]</td>
<td>0.685 [0.573 - 0.741]</td>
</tr>
<tr>
<td>Resampling+Univariate2</td>
<td>5000/2573/139</td>
<td>44 [1 - 162]</td>
<td>0.62 [0.467 - 0.712]</td>
</tr>
<tr>
<td>KIAMB1</td>
<td>986/552/550</td>
<td>14 [14 - 14]</td>
<td>0.596 [0.507 - 0.693]</td>
</tr>
<tr>
<td>KIAMB2</td>
<td>988/969/955</td>
<td>28 [27 - 29]</td>
<td>0.595 [0.482 - 0.708]</td>
</tr>
<tr>
<td>KIAMB3</td>
<td>1182/916/889</td>
<td>23 [12 - 28]</td>
<td>0.596 [0.483 - 0.704]</td>
</tr>
<tr>
<td>Iterative Removal</td>
<td>31/31/31</td>
<td>28 [12 - 82]</td>
<td>0.69 [0.589 - 0.794]</td>
</tr>
</tbody>
</table>
Detailed Results (3/3)

**Glioma Subtype Classification**

<table>
<thead>
<tr>
<th>Method to induce multiple signatures</th>
<th>Number of signatures</th>
<th>Number of genes in a signature</th>
<th>Classification performance (AUC)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>mean 95% interval</td>
<td>In discovery dataset 95% interval</td>
</tr>
<tr>
<td>TIE*</td>
<td>5753/5753/4588</td>
<td>46 [45 - 53]</td>
<td>0.871 [0.860 - 0.885]</td>
</tr>
<tr>
<td>Resampling+SVM-RFE1</td>
<td>5000/4255/43</td>
<td>301 [2 - 3599]</td>
<td>0.808 [0.630 - 0.915]</td>
</tr>
<tr>
<td>Resampling+SVM-RFE2</td>
<td>5000/2055/126</td>
<td>3 [1 - 13]</td>
<td>0.694 [0.545 - 0.890]</td>
</tr>
<tr>
<td>Resampling+Univariate1</td>
<td>5000/4751/63</td>
<td>925 [2 - 17022]</td>
<td>0.84 [0.690 - 0.905]</td>
</tr>
<tr>
<td>Resampling+Univariate2</td>
<td>5000/1926/117</td>
<td>3 [1 - 15]</td>
<td>0.74 [0.495 - 0.900]</td>
</tr>
<tr>
<td>KIAMB1</td>
<td>973/658/654</td>
<td>15 [15 - 15]</td>
<td>0.765 [0.675 - 0.865]</td>
</tr>
<tr>
<td>KIAMB2</td>
<td>974/964/964</td>
<td>30 [29 - 30]</td>
<td>0.781 [0.685 - 0.880]</td>
</tr>
<tr>
<td>KIAMB3</td>
<td>1408/786/746</td>
<td>21 [6 - 30]</td>
<td>0.77 [0.685 - 0.865]</td>
</tr>
<tr>
<td>Iterative Removal</td>
<td>58/58/58</td>
<td>24 [15 - 44]</td>
<td>0.847 [0.744 - 0.921]</td>
</tr>
</tbody>
</table>

**Leukemia 5 Yr. Prognosis**

<table>
<thead>
<tr>
<th>Method to induce multiple signatures</th>
<th>Number of signatures</th>
<th>Number of genes in a signature</th>
<th>Classification performance (AUC)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>mean 95% interval</td>
<td>In discovery dataset 95% interval</td>
</tr>
<tr>
<td>TIE*</td>
<td>1804/1804/1561</td>
<td>22 [20 - 28]</td>
<td>0.714 [0.647 - 0.805]</td>
</tr>
<tr>
<td>Resampling+SVM-RFE1</td>
<td>5000/4643/158</td>
<td>1984 [1 - 8756]</td>
<td>0.631 [0.422 - 0.741]</td>
</tr>
<tr>
<td>Resampling+SVM-RFE2</td>
<td>5000/2537/570</td>
<td>15 [1 - 92]</td>
<td>0.543 [0.341 - 0.749]</td>
</tr>
<tr>
<td>Resampling+Univariate1</td>
<td>5000/3897/116</td>
<td>4024 [1 - 10507]</td>
<td>0.649 [0.431 - 0.756]</td>
</tr>
<tr>
<td>Resampling+Univariate2</td>
<td>5000/2516/465</td>
<td>48 [1 - 329]</td>
<td>0.539 [0.235 - 0.756]</td>
</tr>
<tr>
<td>KIAMB1</td>
<td>988/984/984</td>
<td>31 [29 - 31]</td>
<td>0.515 [0.351 - 0.681]</td>
</tr>
<tr>
<td>KIAMB2</td>
<td>1213/1131/1127</td>
<td>46 [13 - 56]</td>
<td>0.517 [0.341 - 0.687]</td>
</tr>
<tr>
<td>KIAMB3</td>
<td>4485/30/30</td>
<td>7 [6 - 10]</td>
<td>0.438 [0.348 - 0.632]</td>
</tr>
<tr>
<td>Iterative Removal</td>
<td>2/2/2</td>
<td>21 [19 - 23]</td>
<td>0.673 [0.630 - 0.716]</td>
</tr>
</tbody>
</table>
TIE* Signatures Have Maximal Predictivity

- TIE* achieves maximal predictivity in 5 out of 6 validation datasets;
- Non-TIE* methods achieve maximal predictivity in 0 to 2 datasets depending on the method;
- In the dataset where the predictivity of TIE* is statistically distinguishable from the empirically maximal one (*Lung Cancer Subtype Classification*), the magnitude of this difference is only 0.009 AUC on average over all discovered signatures.
TIE* Signatures Are Reproducible, Other Signatures May Be Overfitted

- TIE* has no overfitting on average over all signatures and datasets;
- Other methods achieve predictivity in the validation data that is lower than one in the discovery data (by 0.02-0.03 AUC), besides having inferior predictivity...
TIE* Signatures in Comparison with Other Signatures

Predictivity results for *Leukemia 5 Yr. Prognosis* task

Multiple signatures output by TIE* have maximal predictivity & low variance

Multiple signatures output by other methods do not achieve maximal predictivity and have high variance

Each dot in the plot corresponds to a signature (computational model) of the outcome: E.g., $\text{Outcome}(x) = \text{Sign}(w \cdot x + b)$, where $x, w \in \mathbb{R}^m, b \in \mathbb{R}$, $m$ is the number of genes in the signature.
## Single-Dataset Experiments: Datasets

<table>
<thead>
<tr>
<th>Task</th>
<th>Sample size</th>
<th>Samples per class</th>
<th>Number of genes</th>
<th>Microarray platform</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Lymphoma Subtype Classification I:</strong> Diffuse large-B-cell lymphoma (DLBCL) vs. Burkitt's lymphoma (BL) patients</td>
<td>303</td>
<td>DLBCL (258) BL (45)</td>
<td>2745</td>
<td>Human LymphDx 2.7k GeneChip</td>
</tr>
<tr>
<td><strong>Lymphoma Subtype Classification II:</strong> Diffuse large-B-cell lymphoma (DLBCL) vs. mediastinal large B-cell (MLBCL) patients</td>
<td>210</td>
<td>DLBCL (176) MLBCL (34)</td>
<td>32403 (44928)</td>
<td>Affymetrix U133A and U133B</td>
</tr>
<tr>
<td><strong>Breast Cancer Subtype Classification I:</strong> p53 mutant vs. wild-type breast tumors</td>
<td>251</td>
<td>p53 mutant (58) p53 wild-type (193)</td>
<td>22283</td>
<td>Affymetrix U133A</td>
</tr>
<tr>
<td><strong>Breast Cancer Subtype Classification II:</strong> estrogen receptor positive (ER+) vs. ER- breast tumors</td>
<td>247</td>
<td>ER+ (213) ER- (34)</td>
<td>22283</td>
<td>Affymetrix U133A</td>
</tr>
<tr>
<td><strong>Breast Cancer Subtype Classification III:</strong> progesterone receptor positive (PgR+) vs. PgR- breast tumors</td>
<td>251</td>
<td>PgR+ (190) PgR- (61)</td>
<td>22283</td>
<td>Affymetrix U133A</td>
</tr>
<tr>
<td><strong>Breast Cancer 5 Yr. Prognosis:</strong> ER+ patients who developed distant metastases within 5 years (poor prognosis) vs. ones who did not (good prognosis)</td>
<td>215</td>
<td>poor prognosis (51) good prognosis (164)</td>
<td>24496</td>
<td>Agilent Hu25K</td>
</tr>
<tr>
<td><strong>Bladder Cancer Stage Classification:</strong> stage Ta. vs. other stages (T1, T2, T3, T4) of bladder tumors</td>
<td>404</td>
<td>stage Ta (189) other stages (215)</td>
<td>1381 (3072)</td>
<td>MDL Human 3k</td>
</tr>
</tbody>
</table>

- Validation dataset → subset of 100 samples/patients
- Discovery dataset → all remaining samples/patients
- Repeat splits into discovery & validation datasets 10 times to minimize variance
Single-Dataset Experiments: Summary Results

- Results are similar to the ones from independent-dataset experiments;
- TIE* achieves maximal predictivity in 6 out of 7 validation datasets;
- Non-TIE* methods achieve maximal predictivity in 0 to 1 datasets depending on the method;
- In the dataset where TIE* has predictivity that is statistically distinguishable from the empirically maximal one (*Breast Cancer Subtype Classification II*), the magnitude of this difference is only <0.01 AUC on average over all discovered signatures.
4. Discussion and Interpretation of Results
Revisiting Previously Published Hypotheses about Signature Multiplicity

- Signature reproducibility neither precludes multiplicity nor requires sample sizes with thousands of subjects;
- Multiplicity of signatures does not require dense connectivity;
- Noisy measurements or normalization are not necessary conditions for signature multiplicity;
- Multiplicity can be produced by a combination of small sample size-related variance and intrinsic multiplicity in the underlying network;
- Multiple signatures output by TIE* are reproducible even though they are derived from small sample, noisy, and heavily-processed data.
A More Complete Picture is Emerging Regarding Causes of Multiplicity...

1. **Intrinsic information redundancy** in the underlying biological system;
2. **Variability in the output of gene selection and classifier algorithms** especially in small sample sizes;
3. **Small sample statistical indistinguishability of signatures** with different large sample predictivity and/or redundancy characteristics;
4. **Presence of hidden variables**;
5. **Correlated measurement noise**;
6. **RNA amplification techniques** that systematically distort measurements of transcript ratios;
7. **Cellular aggregation and sampling from mixtures of distributions** that affect inference of conditional independence relations;
8. **Normalization and other data pre-processing methods** that artificially increase correlations among genes;
9. **Engineered redundancy** in the assay technology platforms.
Directions for Future Research

• Use various types of real high-throughput molecular data (e.g., proteomics, metabolomics, SNP data, etc.) in addition to gene expression microarrays;

• Development of more computationally efficient multiple signature extraction algorithms for special distributions;

• Study instantiations of TIE* with sample-efficient Markov boundary methods that can detect “spouses”, e.g. HITON-MB and MMMB;

• Incorporate FDR in TIE*.
Conclusion

1. Developed a Markov boundary characterization of molecular signature multiplicity;
2. Designed a generative algorithm that can correctly identify the set of maximally predictive and non-redundant molecular signatures in principle independently of data distribution;
3. Conducted an empirical evaluation of the novel algorithm and compared it to existing state-of-the-art methods using artificial simulated, resimulated microarray gene expression, and real human microarray gene expression data;
4. Tested and refined several hypotheses about the causes of molecular signature multiplicity phenomenon.
Acknowledgements

Members of my Ph.D. committee:

◦ Constantin Aliferis (NYU)
◦ Gregory Cooper (Univ. of Pittsburgh)
◦ Douglas Hardin (Vanderbilt)
◦ Daniel Masys (Vanderbilt)
◦ Ioannis Tsamardinos (Univ. of Crete, Greece)

Other collaborators:

◦ Frank Harrell (Vanderbilt)
◦ Isabelle Guyon (ClopiNet)
Thank you for your attention!