Linear Separability and Classifiability in Gene Expression Datasets

Benny Chor  Giora Unger

School of Computer Science
Tel-Aviv University

May 35, 2004
High Throughput BioTechnologies

- Sequencing
- 2D gels
- DNA micro-arrays (chips)
- Mass spec
High Throughput BioTechnologies

Computational challenge:
High Throughput BioTechnologies

- Computational challenge:
  - Analyze the huge amounts of data.
High Throughput BioTechnologies

- Computational challenge:
  - Analyze the huge amounts of data.
  - Overcome "curse of dimensionality" and noise.
High Throughput BioTechnologies

- Computational challenge:
  - Analyze the huge amounts of data.
  - Overcome "curse of dimensionality” and noise.
  - Discover meaningful biological signals.
DNA MicroArrays

In a single experiment, measure expression levels of \( n \) genes (\( n \approx 2,000 - 22,000 \)).
DNA MicroArrays

- In a single experiment, measure expression levels of $n$ genes ($n \approx 2,000 - 22,000$).
- Perform $m$ experiments (on different conditions/tissues/patients) ($m \approx 50 - 200$).
DNA MicroArrays

- In a single experiment, measure expression levels of $n$ genes ($n \approx 2,000 - 22,000$).
- Perform $m$ experiments (on different conditions/tissues/patients) ($m \approx 50 - 200$).
- Major goals:
DNA MicroArrays

- In a single experiment, measure expression levels of $n$ genes ($n \approx 2,000 - 22,000$).
- Perform $m$ experiments (on different conditions/tissues/patients) ($m \approx 50 - 200$).
- Major goals:
  - Understand mechanisms controlling gene expression in cells/tissues.
DNA MicroArrays

- In a single experiment, measure expression levels of $n$ genes ($n \approx 2,000 - 22,000$).
- Perform $m$ experiments (on different conditions/tissues/patients) ($m \approx 50 - 200$).
- Major goals:
  - Understand mechanisms controlling gene expression in cells/tissues.
  - Apply knowledge in clinical (medical) scenarios.
DNA MicroArrays

- In a single experiment, measure expression levels of \( n \) genes (\( n \approx 2,000 - 22,000 \)).
- Perform \( m \) experiments (on different conditions/tissues/patients) (\( m \approx 50 - 200 \)).
- Major goals:
  - Understand mechanisms controlling gene expression in cells/tissues.
  - Apply knowledge in clinical (medical) scenarios.
- In many clinical (medical) experiments, tissues are taken from two different populations (e.g. type A cancer vs. type B cancer).
Tools for Analyzing DNA MicroArrays

Clustering
Tools for Analyzing DNA MicroArrays

Clustering
Tools for Analyzing DNA MicroArrays

- Clustering

- Generalizations and Variants
Tools for Analyzing DNA MicroArrays

- Clustering

- Generalizations and Variants
  - Biclustering
Tools for Analyzing DNA MicroArrays

- Clustering

- Generalizations and Variants
  - Biclustering
  - Order Preserving Submatrices
Tools for Analyzing DNA MicroArrays

- Clustering

- Generalizations and Variants
  - Biclustering
  - Order Preserving Submatrices
  - Principal Component Analysis
Tools for Analyzing DNA MicroArrays

- Clustering

- Generalizations and Variants
  - Biclustering
  - Order Preserving Submatrices
  - Principal Component Analysis
  - Plaid Models
Tools for Analyzing DNA MicroArrays

- Clustering

- Generalizations and Variants
  - Biclustering
  - Order Preserving Submatrices
  - Principal Component Analysis
  - Plaid Models
This Work

- Look for **signals** based on **pairs of genes**
This Work

- Look for signals based on pairs of genes
- Interested in separating tissues of type A from tissues of type B.
This Work

- Look for signals based on pairs of genes
- Interested in separating tissues of type A from tissues of type B.
- Separation is interpreted geometrically (linear separation)
Pairs of Genes

Consider expression levels of a pair of genes \((g_1, g_2)\).
Pairs of Genes

Consider expression levels of a pair of genes \((g_1, g_2)\).
Pairs of Genes

Consider expression levels of a pair of genes \((g_1, g_2)\).

- This is a plot of points in 2D plane.
Pairs of Genes

Consider expression levels of a pair of genes \((g_1, g_2)\).

This is a plot of points in 2D plane.
Each point represents one experiment.
Pairs of Genes

Consider expression levels of a pair of genes \((g_1, g_2)\).

This is a plot of points in 2D plane.

- Each point represents one experiment.
- Color of points indicate their class (A or B).
Linear Separation

Expression levels of a pair of genes \((g_1, g_2)\), again.
Linear Separation

Expression levels of a pair of genes \((g_1, g_2)\), again.
Linear Separation

Expression levels of a different pair of genes \((g_1, g_2)\).
Linear Separation

Expression levels of a different pair of genes \((g_1, g_2)\).
Linear Separation II

Genes $g_1, g_2$ linearly separate class A from class B.
Linear Separation II

Genes $g_1, g_2$ linearly separate class A from class B.
Probability of Separation

Suppose you’re told that a pair of genes, \((g_1, g_2)\), separates class \(A\) from \(B\). Are you surprised?
Probability of Separation

- Suppose you’re told that a pair of genes, \((g_1, g_2)\), separates class A from B. Are you surprised?
- Well, the pair either does separate or it does not.
Probability of Separation

Suppose you’re told that a pair of genes, \((g_1, g_2)\), separates class A from B. Are you surprised?

Well, the pair either does separate or it does not. Need some context to discuss surprise.
Probability of Separation

- Suppose you’re told that a pair of genes, \((g_1, g_2)\), separates class A from B. Are you surprised?
- Well, the pair either does separate or it does not.
- Need some context to discuss surprise.
- *E.g.* how many pairs are separating.
Expected Number of Separating Pairs

*Claim:* If *labels* are assigned *at random*, chances of separation are *slim.*
Expected Number of Separating Pairs

- **Claim**: If *labels* are assigned *at random*, chances of separation are *slim*.
- Two examples, both with $m_1 = 3, m_2 = 6$. 

Expected Number of Separating Pairs

- **Claim:** If labels are assigned at random, chances of separation are **slim**.
- Two examples, both with $m_1 = 3, m_2 = 6$.

  . . . . . . . . . . . .
Expected Number of Separating Pairs

- **Claim**: If labels are assigned at random, chances of separation are slim.
- Two examples, both with $m_1 = 3, m_2 = 6$. 

  .   .   .   .   .   .   .   .   .   .   .   .
Expected Number of Separating Pairs

Claim: If labels are assigned at random, chances of separation are slim.

Two examples, both with \( m_1 = 3, m_2 = 6 \).
Expected Number of Separating Pairs

Claim: If labels are assigned at random, chances of separation are slim.
Expected Number of Separating Pairs

Claim: If labels are assigned at random, chances of separation are slim.

\[
\text{Prob(separation)} \leq \frac{\binom{m_1 + m_2}{2}}{\binom{m_1 + m_2}{m_1}}, \text{ where } m_1 \text{ is the number of type A tissues, and } m_2 \text{ of type B tissues (planar } k\text{-sets problem).}
\]
Expected Number of Separating Pairs

Claim: If labels are assigned at random, chances of separation are slim.

\[ \text{Prob(separation)} \leq \frac{\binom{m_1+m_2}{2}}{\binom{m_1}{m_1}} \]

where \( m_1 \) is the number of type A tissues, and \( m_2 \) of type B tissues (planar \( k \)-sets problem).
Expected Number of Separating Pairs

Claim: If labels are assigned at random, chances of separation are slim.

\[ \text{Prob(separation)} \leq \frac{\binom{m_1+m_2}{2}}{\binom{m_1+m_2}{m_1}} \], where \( m_1 \) is the number of type A tissues, and \( m_2 \) of type B tissues (planar \( k \)-sets problem).

\[ \implies \text{expected number of separating pairs is low:} \]
Expected Number of Separating Pairs

- **Claim**: If labels are assigned at random, chances of separation are slim.

  \[
  \text{Prob(separation)} \leq \frac{\binom{m_1+m_2}{2}}{\binom{m_1}{m_1}}
  \]
  where \(m_1\) is the number of type A tissues, and \(m_2\) of type B tissues (planar \(k\)-sets problem).

- \(\implies\) expected number of separating pairs is low:
  For \(n\) genes, the expected number of separating pairs is
  \[
  \leq \binom{n}{2} \frac{\binom{m_1+m_2}{2}}{\binom{m_1}{m_1}}
  \]
# Our 10 Datasets

<table>
<thead>
<tr>
<th>Dataset Name</th>
<th>Cancer Type</th>
<th># Genes</th>
<th>( m = m_1 + m_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gordon 2002</td>
<td>Lung</td>
<td>12533</td>
<td>70 = 20 + 50</td>
</tr>
<tr>
<td>Small Beer 2002</td>
<td>Lung</td>
<td>4966</td>
<td>58 = 50 + 8</td>
</tr>
<tr>
<td>Bhattacharjee 2001</td>
<td>Lung</td>
<td>4392</td>
<td>50 = 40 + 10</td>
</tr>
<tr>
<td>Squamous 2001</td>
<td>Lung</td>
<td>4295</td>
<td>30 = 15 + 15</td>
</tr>
<tr>
<td>Beer 2002</td>
<td>Lung</td>
<td>7129</td>
<td>48 = 40 + 8</td>
</tr>
<tr>
<td>Golub 1999</td>
<td>Leukemia</td>
<td>7129</td>
<td>38 = 27 + 11</td>
</tr>
<tr>
<td>Khan 2001</td>
<td>SRBCT</td>
<td>2308</td>
<td>26 = 13 + 13</td>
</tr>
<tr>
<td>Alon 1999</td>
<td>Colon</td>
<td>2000</td>
<td>28 = 15 + 13</td>
</tr>
<tr>
<td>Adeno Beer 2002</td>
<td>Lung</td>
<td>7129</td>
<td>25 = 15 + 10</td>
</tr>
<tr>
<td>van’t Veer 2002</td>
<td>Breast</td>
<td>21906</td>
<td>28 = 14 + 14</td>
</tr>
</tbody>
</table>
## Expected Upper Bounds for 10 Datasets

<table>
<thead>
<tr>
<th>Dataset Name</th>
<th>Cancer Type</th>
<th># Genes</th>
<th>( m = m_1 + m_2 )</th>
<th>Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gordon 2002</td>
<td>Lung</td>
<td>12533</td>
<td>70 = 20+50</td>
<td>(10^{-6})</td>
</tr>
<tr>
<td>Small Beer 2002</td>
<td>Lung</td>
<td>4966</td>
<td>58 =50+8</td>
<td>11</td>
</tr>
<tr>
<td>Bhattacharjee 2001</td>
<td>Lung</td>
<td>4392</td>
<td>50 =40+10</td>
<td>1</td>
</tr>
<tr>
<td>Squamous 2001</td>
<td>Lung</td>
<td>4295</td>
<td>30 =15+15</td>
<td>26</td>
</tr>
<tr>
<td>Beer 2002</td>
<td>Lung</td>
<td>7129</td>
<td>48 =40+8</td>
<td>76</td>
</tr>
<tr>
<td>Golub 1999</td>
<td>Leukemia</td>
<td>7129</td>
<td>38 =27+11</td>
<td>15</td>
</tr>
<tr>
<td>Khan 2001</td>
<td>SRBCT</td>
<td>2308</td>
<td>26 = 13+13</td>
<td>83</td>
</tr>
<tr>
<td>Alon 1999</td>
<td>Colon</td>
<td>2000</td>
<td>28 = 15+13</td>
<td>20</td>
</tr>
<tr>
<td>Adeno Beer 2002</td>
<td>Lung</td>
<td>7129</td>
<td>25 =15+10</td>
<td>2332</td>
</tr>
<tr>
<td>van’t Veer 2002</td>
<td>Breast</td>
<td>21906</td>
<td>28 = 14+14</td>
<td>2261</td>
</tr>
</tbody>
</table>
What Next?

- Wanna find all separating pairs of genes in these (and other) datasets.
What Next?

- Wanna find all separating pairs of genes in these (and other) datasets.
- Compare # to upper bound.
What Next?

- Wanna find all separating pairs of genes in these (and other) datasets.
- Compare # to upper bound.
- If # exceptionally high, possible evidence of interesting interactions between members of pairs, relevant to condition.
What Next?

- Wanna find all separating pairs of genes in these (and other) datasets.
- Compare # to upper bound.
- If # exceptionally high, possible evidence of interesting interactions between members of pairs, relevant to condition.
- Such pairs may warrant additional exploration in lab.
Houston, We Have a Problem
Houston, We Have a Problem

A small size dataset (e.g. Beer, Golub) has 7,000 genes ⇒ 24,500,000 pairs of genes.
Houston, We Have a Problem

- A small size dataset (e.g. Beer, Golub) has 7,000 genes ⇒ 24,500,000 pairs of genes.
- A medium size dataset (e.g. Bhattacharjee, Gordon) has 12,000 genes ⇒ 72,000,000 pairs of genes.
Houston, We Have a Problem

- A **small size** dataset (e.g. Beer, Golub) has 7,000 genes ⇒ **24,500,000 pairs** of genes.
- A **medium size** dataset (e.g. Bhattacharjee, Gordon) has 12,000 genes ⇒ **72,000,000 pairs** of genes.
- A **large size** dataset (e.g. van’t Veer) has 20,000 genes ⇒ **200,000,000 pairs** of genes.
Houston, We Have a Problem

- A **small size** dataset (e.g. Beer, Golub) has 7,000 genes ⇒ 24,500,000 pairs of genes.
- A **medium size** dataset (e.g. Bhattacharjee, Gordon) has 12,000 genes ⇒ 72,000,000 pairs of genes.
- A **large size** dataset (e.g. van’t Veer) has 20,000 genes ⇒ 200,000,000 pairs of genes.

Computationally intensive algorithms (e.g. SVM) cannot check that many pairs of genes in a reasonable amount of time.
Algorithm Incremental Approach

With actual datasets, the vast majority of pairs are non-separating.
Algorithm Incremental Approach

- With actual datasets, the vast majority of pairs are non-separating.
- If could filter out non-separating pairs efficiently, be able to explore the few separating pairs.
**Algorithm Incremental Approach**

- With actual datasets, the vast majority of pairs are **non-separating**.

- If could **filter out** non-separating pairs efficiently, be able to explore the **few separating pairs**.

- For a given pair, a **small number** of points often suffices to give evidence of non-separability.
Algorithm Incremental Approach

- With actual datasets, the vast majority of pairs are non-separating.

- If could filter out non-separating pairs efficiently, be able to explore the few separating pairs.

- For a given pair, a small number of points often suffices to give evidence of non-separability.
Algorithm **Incremental** Approach

- With actual datasets, the vast majority of pairs are **non-separating**.
- If could **filter out** non-separating pairs efficiently, be able to explore the **few separating pairs**.
- For a given pair, a **small number** of points often suffices to give evidence of non-separability.
Algorithm Running Times

Identifying all separating pairs of genes for a whole dataset, on a standard PC (Pentium 4, 2.0GHz, 1 GB RAM):

- 31 seconds for a small size dataset (Golub – 7129 genes).
Algorithm Running Times

Identifying all separating pairs of genes for a whole dataset, on a standard PC (Pentium 4, 2.0GHz, 1 GB RAM):

- 31 seconds for a small size dataset (Golub – 7129 genes).
- 2 minutes for a medium size dataset (Gordon – 12533 genes).
Algorithm Running Times

Identifying all separating pairs of genes for a whole dataset, on a standard PC (Pentium 4, 2.0GHz, 1 GB RAM):

- 31 seconds for a small size dataset (Golub – 7129 genes).
- 2 minutes for a medium size dataset (Gordon – 12533 genes).
- 3 minutes for a large size dataset (van’t Veer – 21906 genes).
And the Winners Are . . .

<table>
<thead>
<tr>
<th>Dataset</th>
<th># Genes</th>
<th>$m = m_1 + m_2$</th>
<th>Exp. Upper Bound</th>
<th># Separating Pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gordon 2002</td>
<td>12533</td>
<td>$70 = 20+50$</td>
<td>$10^{-6}$</td>
<td>17796</td>
</tr>
<tr>
<td>Small Beer 2002</td>
<td>4966</td>
<td>$58 =50+8$</td>
<td>11</td>
<td>120336</td>
</tr>
<tr>
<td>Bhattacharjee 2001</td>
<td>4392</td>
<td>$50 =40+10$</td>
<td>1</td>
<td>87918</td>
</tr>
<tr>
<td>Squamous 2001</td>
<td>4295</td>
<td>$30 =15+15$</td>
<td>26</td>
<td>$1.7 \cdot 10^6$</td>
</tr>
<tr>
<td>Beer 2002</td>
<td>7129</td>
<td>$48 =40+8$</td>
<td>76</td>
<td>226862</td>
</tr>
<tr>
<td>Golub 1999</td>
<td>7129</td>
<td>$38 =27+11$</td>
<td>15</td>
<td>19641</td>
</tr>
<tr>
<td>Khan 2001</td>
<td>2308</td>
<td>$26 = 13+13$</td>
<td>83</td>
<td>20056</td>
</tr>
</tbody>
</table>
And the Winners Are . . .

<table>
<thead>
<tr>
<th>Dataset</th>
<th># Genes</th>
<th>$m = m_1 + m_2$</th>
<th>Exp. Upper Bound</th>
<th># Separating Pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gordon 2002</td>
<td>12533</td>
<td>70 = 20+50</td>
<td>$10^{-6}$</td>
<td>17796</td>
</tr>
<tr>
<td>Small Beer 2002</td>
<td>4966</td>
<td>58 =50+8</td>
<td>11</td>
<td>120336</td>
</tr>
<tr>
<td>Bhattacharjee 2001</td>
<td>4392</td>
<td>50 =40+10</td>
<td>1</td>
<td>87918</td>
</tr>
<tr>
<td>Squamous 2001</td>
<td>4295</td>
<td>30 =15+15</td>
<td>26</td>
<td>$1.7 \cdot 10^6$</td>
</tr>
<tr>
<td>Beer 2002</td>
<td>7129</td>
<td>48 =40+8</td>
<td>76</td>
<td>226862</td>
</tr>
<tr>
<td>Golub 1999</td>
<td>7129</td>
<td>38 =27+11</td>
<td>15</td>
<td>19641</td>
</tr>
<tr>
<td>Khan 2001</td>
<td>2308</td>
<td>26 = 13+13</td>
<td>83</td>
<td>20056</td>
</tr>
</tbody>
</table>

Found seven datasets where actual # separating pairs much larger than expected upper bound (by factors ranging from 250 to $10^9$).
And the Winners Are . . .

<table>
<thead>
<tr>
<th>Dataset</th>
<th># Genes</th>
<th>$m = m_1 + m_2$</th>
<th>Exp. Upper Bound</th>
<th># Separating Pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gordon 2002</td>
<td>12533</td>
<td>70 = 20+50</td>
<td>$10^{-6}$</td>
<td>17796</td>
</tr>
<tr>
<td>Small Beer 2002</td>
<td>4966</td>
<td>58 =50+8</td>
<td>11</td>
<td>120336</td>
</tr>
<tr>
<td>Bhattacharjee 2001</td>
<td>4392</td>
<td>50 =40+10</td>
<td>1</td>
<td>87918</td>
</tr>
<tr>
<td>Squamous 2001</td>
<td>4295</td>
<td>30 =15+15</td>
<td>26</td>
<td>$1.7 \cdot 10^6$</td>
</tr>
<tr>
<td>Beer 2002</td>
<td>7129</td>
<td>48 =40+8</td>
<td>76</td>
<td>226862</td>
</tr>
<tr>
<td>Golub 1999</td>
<td>7129</td>
<td>38 =27+11</td>
<td>15</td>
<td>19641</td>
</tr>
<tr>
<td>Khan 2001</td>
<td>2308</td>
<td>26 = 13+13</td>
<td>83</td>
<td>20056</td>
</tr>
</tbody>
</table>

- Found seven datasets where actual # separating pairs much larger than expected upper bound (by factors ranging from $250$ to $10^9$).
- This is what I call surprise!
Hey, Aren’t You Cheating?!

Used to have 10 datasets. You’ve showed us only 7.
Hey, Aren’t You Cheating?!

Used to have 10 datasets. You’ve showed us only 7.

<table>
<thead>
<tr>
<th>Dataset</th>
<th># Genes</th>
<th>$m = m_1 + m_2$</th>
<th>Exp. Upper Bound</th>
<th># Separating Pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gordon 2002</td>
<td>12533</td>
<td>70 = 20+50</td>
<td>$10^{-6}$</td>
<td>17796</td>
</tr>
<tr>
<td>Small Beer 2002</td>
<td>4966</td>
<td>58 =50+8</td>
<td>11</td>
<td>120336</td>
</tr>
<tr>
<td>Bhattacharjee 2001</td>
<td>4392</td>
<td>50 =40+10</td>
<td>1</td>
<td>87918</td>
</tr>
<tr>
<td>Squamous 2001</td>
<td>4295</td>
<td>30 =15+15</td>
<td>26</td>
<td>$1.7 \cdot 10^6$</td>
</tr>
<tr>
<td>Beer 2002</td>
<td>7129</td>
<td>48 =40+8</td>
<td>76</td>
<td>226862</td>
</tr>
<tr>
<td>Golub 1999</td>
<td>7129</td>
<td>38 =27+11</td>
<td>15</td>
<td>19641</td>
</tr>
<tr>
<td>Khan 2001</td>
<td>2308</td>
<td>26 = 13+13</td>
<td>83</td>
<td>20056</td>
</tr>
<tr>
<td>Alon 1999</td>
<td>2000</td>
<td>28 = 15+13</td>
<td>20</td>
<td>602</td>
</tr>
<tr>
<td>Adeno Beer 2002</td>
<td>7129</td>
<td>25 =15+10</td>
<td>2332</td>
<td>1944</td>
</tr>
<tr>
<td>van’t Veer 2002</td>
<td>21906</td>
<td>28 = 14+14</td>
<td>2261</td>
<td>737</td>
</tr>
</tbody>
</table>
Hey, Aren’t You Cheating?!

Used to have 10 datasets. You’ve showed us only 7.

<table>
<thead>
<tr>
<th>Dataset</th>
<th># Genes</th>
<th>$m = m_1 + m_2$</th>
<th>Exp. Upper Bound</th>
<th># Separating Pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gordon 2002</td>
<td>12533</td>
<td>$70 = 20+50$</td>
<td>$10^{-6}$</td>
<td>17796</td>
</tr>
<tr>
<td>Small Beer 2002</td>
<td>4966</td>
<td>$58 =50+8$</td>
<td>11</td>
<td>120336</td>
</tr>
<tr>
<td>Bhattacharjee 2001</td>
<td>4392</td>
<td>$50 =40+10$</td>
<td>1</td>
<td>87918</td>
</tr>
<tr>
<td>Squamous 2001</td>
<td>4295</td>
<td>$30 =15+15$</td>
<td>26</td>
<td>$1.7 \cdot 10^6$</td>
</tr>
<tr>
<td>Beer 2002</td>
<td>7129</td>
<td>$48 =40+8$</td>
<td>76</td>
<td>226862</td>
</tr>
<tr>
<td>Golub 1999</td>
<td>7129</td>
<td>$38 =27+11$</td>
<td>15</td>
<td>19641</td>
</tr>
<tr>
<td>Khan 2001</td>
<td>2308</td>
<td>$26 = 13+13$</td>
<td>83</td>
<td>20056</td>
</tr>
<tr>
<td>Alon 1999</td>
<td>2000</td>
<td>$28 = 15+13$</td>
<td>20</td>
<td>602</td>
</tr>
<tr>
<td>Adeno Beer 2002</td>
<td>7129</td>
<td>$25 =15+10$</td>
<td>2332</td>
<td>1944</td>
</tr>
<tr>
<td>van’t Veer 2002</td>
<td>21906</td>
<td>$28 = 14+14$</td>
<td>2261</td>
<td>737</td>
</tr>
</tbody>
</table>

- All datasets are equal, but some datasets are more equal than others.
## Separability and Classifiability

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Exp. Upper Bound</th>
<th># Separating Pairs</th>
<th>SVM Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gordon 2002</td>
<td>$10^{-6}$</td>
<td>17796</td>
<td>0.07%</td>
</tr>
<tr>
<td>Small Beer 2002</td>
<td>11</td>
<td>120336</td>
<td>0%</td>
</tr>
<tr>
<td>Bhattacharjee 2001</td>
<td>1</td>
<td>87918</td>
<td>1.54%</td>
</tr>
<tr>
<td>Squamous 2001</td>
<td>26</td>
<td>$1.7 \cdot 10^6$</td>
<td>0%</td>
</tr>
<tr>
<td>Beer 2002</td>
<td>76</td>
<td>226862</td>
<td>0%</td>
</tr>
<tr>
<td>Golub 1999</td>
<td>15</td>
<td>19641</td>
<td>2.9%</td>
</tr>
<tr>
<td>Khan 2001</td>
<td>83</td>
<td>20056</td>
<td>2.4%</td>
</tr>
<tr>
<td>Alon 1999</td>
<td>20</td>
<td>602</td>
<td>16%</td>
</tr>
<tr>
<td>Adeno Beer 2002</td>
<td>2332</td>
<td>1944</td>
<td>22%</td>
</tr>
<tr>
<td>van’t Veer 2002</td>
<td>2261</td>
<td>737</td>
<td>17%</td>
</tr>
</tbody>
</table>
## Separability and Classifiability

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Exp. Upper Bound</th>
<th># Separating Pairs</th>
<th>SVM Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gordon 2002</td>
<td>$10^{-6}$</td>
<td>17796</td>
<td>0.07%</td>
</tr>
<tr>
<td>Small Beer 2002</td>
<td>11</td>
<td>120336</td>
<td>0%</td>
</tr>
<tr>
<td>Bhattacharjee 2001</td>
<td>1</td>
<td>87918</td>
<td>1.54%</td>
</tr>
<tr>
<td>Squamous 2001</td>
<td>26</td>
<td>$1.7 \cdot 10^6$</td>
<td>0%</td>
</tr>
<tr>
<td>Beer 2002</td>
<td>76</td>
<td>226862</td>
<td>0%</td>
</tr>
<tr>
<td>Golub 1999</td>
<td>15</td>
<td>19641</td>
<td>2.9%</td>
</tr>
<tr>
<td>Khan 2001</td>
<td>83</td>
<td>20056</td>
<td>2.4%</td>
</tr>
<tr>
<td>Alon 1999</td>
<td>20</td>
<td>602</td>
<td>16%</td>
</tr>
<tr>
<td>Adeno Beer 2002</td>
<td>2332</td>
<td>1944</td>
<td>22%</td>
</tr>
<tr>
<td>van’t Veer 2002</td>
<td>2261</td>
<td>737</td>
<td>17%</td>
</tr>
</tbody>
</table>

An interesting relation:
## Separability and Classifiability

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Exp. Upper Bound</th>
<th># Separating Pairs</th>
<th>SVM Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gordon 2002</td>
<td>$10^{-6}$</td>
<td>17796</td>
<td>0.07%</td>
</tr>
<tr>
<td>Small Beer 2002</td>
<td>11</td>
<td>120336</td>
<td>0%</td>
</tr>
<tr>
<td>Bhattacharjee 2001</td>
<td>1</td>
<td>87918</td>
<td>1.54%</td>
</tr>
<tr>
<td>Squamous 2001</td>
<td>26</td>
<td>$1.7 \cdot 10^6$</td>
<td>0%</td>
</tr>
<tr>
<td>Beer 2002</td>
<td>76</td>
<td>226862</td>
<td>0%</td>
</tr>
<tr>
<td>Golub 1999</td>
<td>15</td>
<td>19641</td>
<td>2.9%</td>
</tr>
<tr>
<td>Khan 2001</td>
<td>83</td>
<td>20056</td>
<td>2.4%</td>
</tr>
<tr>
<td>Alon 1999</td>
<td>20</td>
<td>602</td>
<td>16%</td>
</tr>
<tr>
<td>Adeno Beer 2002</td>
<td>2332</td>
<td>1944</td>
<td>22%</td>
</tr>
<tr>
<td>van’t Veer 2002</td>
<td>2261</td>
<td>737</td>
<td>17%</td>
</tr>
</tbody>
</table>

An interesting relation:

High separability $\Rightarrow$ Low classification error
Separability and Classifiability

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Exp. Upper Bound</th>
<th># Separating Pairs</th>
<th>SVM Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gordon 2002</td>
<td>$10^{-6}$</td>
<td>17796</td>
<td>0.07%</td>
</tr>
<tr>
<td>Small Beer 2002</td>
<td>11</td>
<td>120336</td>
<td>0%</td>
</tr>
<tr>
<td>Bhattacharjee 2001</td>
<td>1</td>
<td>87918</td>
<td>1.54%</td>
</tr>
<tr>
<td>Squamous 2001</td>
<td>26</td>
<td>$1.7 \cdot 10^6$</td>
<td>0%</td>
</tr>
<tr>
<td>Beer 2002</td>
<td>76</td>
<td>226862</td>
<td>0%</td>
</tr>
<tr>
<td>Golub 1999</td>
<td>15</td>
<td>19641</td>
<td>2.9%</td>
</tr>
<tr>
<td>Khan 2001</td>
<td>83</td>
<td>20056</td>
<td>2.4%</td>
</tr>
<tr>
<td>Alon 1999</td>
<td>20</td>
<td>602</td>
<td>16%</td>
</tr>
<tr>
<td>Adeno Beer 2002</td>
<td>2332</td>
<td>1944</td>
<td>22%</td>
</tr>
<tr>
<td>van’t Veer 2002</td>
<td>2261</td>
<td>737</td>
<td>17%</td>
</tr>
</tbody>
</table>

An interesting relation:

high separability $\Rightarrow$ low classification error

low separability $\Rightarrow$ high classification error
Conclusions and Open Problems

- Developed efficient, incremental algorithm to identify all pairs of separating genes.
Conclusions and Open Problems

- Developed efficient, incremental algorithm to identify all pairs of separating genes.
- Found seven datasets that are highly separable.
Conclusions and Open Problems

- Developed efficient, incremental algorithm to identify all pairs of separating genes.
- Found seven datasets that are highly separable.
- An interesting relation between linear separation and classification error.
Conclusions and Open Problems

- Developed efficient, incremental algorithm to identify all pairs of separating genes.
- Found seven datasets that are highly separable.
- An interesting relation between linear separation and classification error.
- Can linear separation improve classification?
Conclusions and Open Problems

- Developed efficient, incremental algorithm to identify all pairs of separating genes.
- Found seven datasets that are highly separable.
- An interesting relation between linear separation and classification error.
- Can linear separation improve classification?
  Apparently not...
Conclusions and Open Problems

- Developed efficient, incremental algorithm to identify all pairs of separating genes.
- Found seven datasets that are highly separable.
- An interesting relation between linear separation and classification error.
- Can linear separation improve classification? Apparently not...
- Todo:
Conclusions and Open Problems

- Developed efficient, incremental algorithm to identify all pairs of separating genes.
- Found seven datasets that are highly separable.
- An interesting relation between linear separation and classification error.
- Can linear separation improve classification? Apparently not...
- Todo:
  - Explore the biological mechanisms underlying separability in specific datasets.
Conclusions and Open Problems

- Developed efficient, incremental algorithm to identify all pairs of separating genes.
- Found seven datasets that are highly separable.
- An interesting relation between linear separation and classification error.
- Can linear separation improve classification? Apparently not...
- Todo:
  - Explore the biological mechanisms underlying separability in specific datasets.
  - Go from pairs to triplets, quadruples, ...