Modeling Patterns in Single-Nucleotide Polymorphism Data for Predicting Cancer Susceptibility: A Case Study in Multiple Myeloma

Michael J. Waddell

Laboratory of Dr. C. David Page
Departments of Computer Science and Biostatistics and Medical Informatics
University of Wisconsin
Background

- Genetics technology has grown dramatically in the past 2 decades.
  - PCR technology developed (1985)
  - BLAST algorithm developed (1990)
  - DNA chips made commercially available (1996)
  - SNP consortium founded (1999)
  - Human Genome Project working draft published (2001)

- This growth has enabled identification of genes responsible for predisposition to some inherited human disorders.
Gregor Mendel (1822–1884)

- However, these successes have been primarily limited to simple Mendelian disorders.
- Mendelian (or monogenic) disorders are rare.
- The vast majority of disorders are likely polygenic.
Finding genes

- There are two main approaches for finding genes responsible for disorders:
  1. Linkage analysis and positional cloning
  2. Direct analysis of candidate genes (via association studies)
Finding genes

- There are two main approaches for finding genes responsible for disorders:
  1. Linkage analysis and positional cloning
     - Requires studies of families with known pedigrees.
     - Genetic markers (RFLPs, microsatellite loci, SNPs, etc.) closest to the gene(s) of interest will be strongly correlated with the disease pattern in the family.
     - Results may not be generalizable.
  2. Direct analysis of candidate genes (via association studies)
Finding genes

- There are two main approaches for finding genes responsible for disorders:
  1. Linkage analysis and positional cloning
  2. Direct analysis of candidate genes (via association studies)
Finding genes

- There are two main approaches for finding genes responsible for disorders:
  1. Linkage analysis and positional cloning
  2. Direct analysis of candidate genes (via association studies)
     - Also uses genetic markers
     - Does not require families
     - Lots of false positives (esp. with non-monogenic disorders)
       - More features than datapoints
       - Low prior probability of association
       - Can be mislead by population stratification (ie: ethnicity, gender, etc.)
Finding genes

• There are two main approaches for finding genes responsible for disorders:
  1. Linkage analysis and positional cloning
  2. Direct analysis of candidate genes (via association studies)

• These standard approaches have successfully identified genes causing:
  ♠ Breast cancer (BRCA-1 & -2)
  ♠ Colon cancer (FAP & HNPCC)
  ♠ Diabetes (MODY-1, -2 & -3)
  ♠ etc.
Finding genes

- There are two main approaches for finding genes responsible for disorders:
  1. Linkage analysis and positional cloning
  2. Direct analysis of candidate genes (via association studies)

- These standard approaches have successfully identified genes causing:
  - Breast cancer (BRCA-1 & -2)
  - Colon cancer (FAP & HNPCC)
  - Diabetes (MODY-1, -2 & -3)
  - etc.

  (All of which are Mendelian or near-Mendelian.)
Finding genes

- There are two main approaches for finding genes responsible for disorders:
  1. Linkage analysis and positional cloning
  2. Direct analysis of candidate genes (via association studies)
- These standard approaches fail when attempting to identify a set of genes, each of modest effect, whose combined effects cause a particular trait (aka: a polygenic trait).
Genetic Heterogeneity

- Genetic heterogeneity: distinct mutations cause the same, indistinguishable, phenotype.
- Types of genetic heterogeneity:
  1. Distinct loci that interact to cause the phenotype
  2. Distinct loci that are capable of independently causing the phenotype
- Standard techniques can deal with small numbers of independent loci.
- They fail with large numbers of independent loci or interactions among loci.
Machine Learning to the Rescue

1. Divide the patients into two groups: healthy and disease.

2. Machine learning or statistical modeling algorithms can then be used to construct a model of the disorder.

3. This model can be validated using cross-validation.

4. If the model is sufficiently accurate, it can be studied to gain insight into the disease.

⭐ Algorithms have been developed that deal well with interactions and redundant features.
SNPs

- Part of the genetic variation among individuals is the **cumulative** effect of variations at a number of single-base locations within the genome.
- These locations are known as **SNPs** (Single Nucleotide Polymorphisms).
- A “**SNP pattern**” consists of the DNA bases present at a large number of SNP positions.
- SNPs can be used in linkage analysis or association studies to identify **markers** for genes associated with a disorder.
**Benefits of Using SNPs**

- A person’s SNP pattern is highly unlikely to change over time or as a result of disease.
- SNP data can be collected from any tissue in the body (not just from diseased tissue).
- This allows a larger number of samples to be obtained (especially controls) since faster and less invasive procedures are used.
3 Challenges of Using SNPs

1. There are now over one million SNPs known but measuring them all is typically cost-prohibitive.

   - SNP data contain measurements for only a small fraction of known SNPs (typically a few thousand).

   - If prior knowledge is available, focus the SNPs collected to particular region(s) of the genome.

   - Otherwise, choose SNPs to give good overall coverage of the genome. (Subsequent studies can use these results as prior knowledge.)
3 Challenges of Using SNPs

2. SNP data commonly contain missing values.
   - This can adversely affect many algorithms used for classification tasks.
   - When choosing an algorithm to use, this must be taken into consideration in order to choose an appropriate one.
3 Challenges of Using SNPs

3. SNP data are "unphased."

Patient 1:

\[
\begin{array}{ccc}
C & G & T \\
\uparrow & \uparrow & \uparrow \\
T & A & T \\
\uparrow & \uparrow & \uparrow \\
\end{array}
\]

Patient 2:

\[
\begin{array}{ccc}
C & A & T \\
\uparrow & \uparrow & \uparrow \\
T & G & T \\
\uparrow & \uparrow & \uparrow \\
\end{array}
\]

<table>
<thead>
<tr>
<th>SNP 1</th>
<th>SNP 2</th>
<th>SNP 3</th>
<th>…</th>
<th>class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient 1</td>
<td>C</td>
<td>T</td>
<td>A</td>
<td>G</td>
</tr>
<tr>
<td>Patient 2</td>
<td>C</td>
<td>T</td>
<td>A</td>
<td>G</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
</tbody>
</table>
3 Challenges of Using SNPs

3. SNP data are “unphased.”
   - There are 2 main ways of dealing with SNP data:
     1. Perform haplotyping to determine the phasing of the SNP data.
        - Algorithms for haplotyping are not guaranteed to be correct.
        - These algorithms generally require data on related individuals (pedigrees).
     2. Work with the data in its unphased form.
Multiple Myeloma

- Multiple Myeloma (MM) is a uniformly fatal malignancy of the plasma cells.

- MM occurs with relatively high frequency in older adults (0.035% of the US population aged 70+).

- MM occurs with much lower frequency in younger adults (0.002% of the US population aged 30–54).

- We hypothesize that those diagnosed with MM at a young age have a genetic predisposition to the disease.
Data Set

- “Unphased” SNP data for 80 patients (based on 3000 SNPs)
  - 40 “old” patients: diagnosed with MM after age 70
  - 40 “young” patients: diagnosed with MM before age 40
- The SNPs were selected to give good overall coverage of the human genome — not based on prior knowledge of MM.
Support Vector Machines

Figure 1: Linear support vector machines (SVMs) maximize the “margin” between the bounding planes.
Support Vector Machines

- Non-linear SVMs also exist, but their output is harder to gain insights from.
- With a linear SVM, features (SNPs) with larger coefficients in the linear separator are more important.
- SVMs (whether linear or not), require features to be numerical and continuous.
- SNP data is not numerical or continuous.
**SVMs Require Continuous Features**

- In unphased SNP data, each feature takes on one of 3 discrete values:
  1. Homozygous for the dominant SNP
  2. Homozygous for the recessive SNP
  3. Heterozygous

- Therefore, we convert the discrete values to the values: -1, 0, 1
  (Where 0 represents heterozygous and the two homozygous cases are arbitrarily mapped to ±1.)

- “Phased” SNP data would have a 4th possible value.
SVMs Require Continuous Features

Figure 2: This choice of mapping is made because it allows the SVM to divide between the presence and absence of either SNP.
SVMs Require Continuous Features

- This mapping allows the SVM to divide between the presence and absence of either SNP.
- It does not allow the SVM to divide between the presence and absence of homozygosity. (But this is unlikely to be biologically relevant.)
- An alternative mapping that would allow both of these divisions would use 2 features per SNP. (However, this would double the number of features and machine learning algorithms perform better with fewer features.)
"Curse of Dimensionality"

- The “Curse of Dimensionality” — having many more features than examples — is a major problem in machine learning.

- We employ 3 methods to deal with this:
  1. Use leave-one-out cross-validation to assess the accuracy of the model since it is robust to high-dimensional data.
  2. Use SVMs because they are more robust than some other algorithms on high-dimensional data.
  3. Use feature selection to reduce the number of features given to the SVM.
Feature Selection

- Feature selection is commonly used to improve performance of modeling algorithms.
- It is common practice, though incorrect, to perform feature selection once over the whole data set.
- This practice leads to information “leaking” from the validation set into the training set.
- To avoid this common pitfall, the top 10% of SNPs were chosen by information gain on each fold of cross-validation prior to model construction.
Results

<table>
<thead>
<tr>
<th>Actual</th>
<th>Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Old</td>
<td>Old</td>
</tr>
<tr>
<td>Young</td>
<td>Young</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 3:** Our approach yields an accuracy estimate of 71% by leave-one-out cross validation. This is **significantly** better than random guessing (p < 0.01).
Conclusions

- Our accuracy of 71% is not as high as we have obtained using microarray data.
- However, this prediction is based only on SNP data (which are not affected by disease progression like microarray data).
- Also, our SNP coverage was relatively sparse (only 3000 SNPs were used).
- Thus, we conclude that SNP data do provide predictive ability for cancer susceptibility.
Interpreting SVM Results

- Ideally, the SVM model would be based on only a few SNPs.
- An SVM only “uses” those SNPs with non-zero coefficients.
- Our SVM produced a model that uses over 150 SNPs.
- 48 SNPs had non-zero coefficients on every cross-validation fold.
## SNPs With Non-Zero Coefficients

<table>
<thead>
<tr>
<th>SNP</th>
<th>Chrom.</th>
<th>Contig Accession</th>
<th>Contig Position</th>
<th>Chrom. Position</th>
<th>Hit Orientation</th>
</tr>
</thead>
<tbody>
<tr>
<td>TSC0022568</td>
<td>20</td>
<td>NT_011387.8</td>
<td>23917072</td>
<td>23925072</td>
<td>minus strand</td>
</tr>
<tr>
<td>974272</td>
<td>8</td>
<td>NT_008046.11</td>
<td>6451653</td>
<td>97542762</td>
<td>minus strand</td>
</tr>
<tr>
<td>930891</td>
<td>5</td>
<td>NT_023132.10</td>
<td>2310237</td>
<td>171083062</td>
<td>plus strand</td>
</tr>
<tr>
<td>930311</td>
<td>2</td>
<td>NT_037537.1</td>
<td>1516282</td>
<td>172969092</td>
<td>plus strand</td>
</tr>
<tr>
<td>921897</td>
<td>5</td>
<td>NT_006431.11</td>
<td>4641524</td>
<td>61626387</td>
<td>plus strand</td>
</tr>
<tr>
<td>912797</td>
<td>1</td>
<td>NT_004636.13</td>
<td>1545773</td>
<td>66344816</td>
<td>plus strand</td>
</tr>
<tr>
<td>910069</td>
<td>6</td>
<td>NT_025741.10</td>
<td>16628671</td>
<td>139745590</td>
<td>plus strand</td>
</tr>
<tr>
<td>880307</td>
<td>3</td>
<td>NT_022509.9</td>
<td>2184897</td>
<td>179348004</td>
<td>minus strand</td>
</tr>
<tr>
<td>869380</td>
<td>1</td>
<td>NT_004612.13</td>
<td>1632279</td>
<td>210865199</td>
<td>plus strand</td>
</tr>
<tr>
<td>759002</td>
<td>7</td>
<td>NT_033968.2</td>
<td>1421567</td>
<td>51360645</td>
<td>plus strand</td>
</tr>
<tr>
<td>755614</td>
<td>3</td>
<td>NT_037588.0</td>
<td>31998</td>
<td>unplaced</td>
<td>minus strand</td>
</tr>
<tr>
<td>737453</td>
<td>2</td>
<td>NT_005403.11</td>
<td>16204291</td>
<td>214479968</td>
<td>plus strand</td>
</tr>
<tr>
<td>726828</td>
<td>15</td>
<td>NT_003268.0</td>
<td>126217</td>
<td>unplaced</td>
<td>minus strand</td>
</tr>
<tr>
<td>726828</td>
<td>3</td>
<td>NT_005684.9</td>
<td>71977</td>
<td>77618043</td>
<td>plus strand</td>
</tr>
<tr>
<td>717480</td>
<td>12</td>
<td>NT_035247.0</td>
<td>106169</td>
<td>unplaced</td>
<td>plus strand</td>
</tr>
<tr>
<td>717480</td>
<td>13</td>
<td>NT_009952.11</td>
<td>17735730</td>
<td>99066060</td>
<td>plus strand</td>
</tr>
<tr>
<td>712269</td>
<td>17</td>
<td>NT_030843.4</td>
<td>770495</td>
<td>19404437</td>
<td>minus strand</td>
</tr>
<tr>
<td>707860</td>
<td>6</td>
<td>NT_007592.11</td>
<td>8361187</td>
<td>17560475</td>
<td>plus strand</td>
</tr>
<tr>
<td>679206</td>
<td>10</td>
<td>NT_030059.8</td>
<td>8262073</td>
<td>102231090</td>
<td>plus strand</td>
</tr>
<tr>
<td>598985</td>
<td>8</td>
<td>NT_034927.0</td>
<td>61452</td>
<td>unplaced</td>
<td>minus strand</td>
</tr>
<tr>
<td>598985</td>
<td>8</td>
<td>NT_023744.11</td>
<td>1619368</td>
<td>2954228</td>
<td>plus strand</td>
</tr>
</tbody>
</table>
## SNPs With Non-Zero Coefficients

<table>
<thead>
<tr>
<th>SNP</th>
<th>Chrom.</th>
<th>Contig Accession</th>
<th>Contig Position</th>
<th>Chrom. Position</th>
<th>Hit Orientation</th>
</tr>
</thead>
<tbody>
<tr>
<td>257769</td>
<td>5</td>
<td>NT_005676.11</td>
<td>12192025</td>
<td>16122034</td>
<td>plus strand</td>
</tr>
<tr>
<td>220872</td>
<td>11</td>
<td>NT_033899.3</td>
<td>4804715</td>
<td>116778426</td>
<td>plus strand</td>
</tr>
<tr>
<td>1992291</td>
<td>2</td>
<td>NT_005204.11</td>
<td>6319962</td>
<td>27596897</td>
<td>plus strand</td>
</tr>
<tr>
<td>1990272</td>
<td>4</td>
<td>NT_006342.13</td>
<td>974650</td>
<td>10545064</td>
<td>plus strand</td>
</tr>
<tr>
<td>1989229</td>
<td>17</td>
<td>NT_037978.1</td>
<td>98150</td>
<td></td>
<td>plus strand</td>
</tr>
<tr>
<td>1989229</td>
<td>17</td>
<td>NT_010748.10</td>
<td>88877</td>
<td>45477210</td>
<td>minus strand</td>
</tr>
<tr>
<td>1882055</td>
<td>7</td>
<td>NT_007819.11</td>
<td>35393849</td>
<td>35742551</td>
<td>minus strand</td>
</tr>
<tr>
<td>1863239</td>
<td>7</td>
<td>NT_007819.11</td>
<td>36884313</td>
<td>37233015</td>
<td>minus strand</td>
</tr>
<tr>
<td>1862336</td>
<td>19</td>
<td>NT_011109.13</td>
<td>9443320</td>
<td>47679528</td>
<td>plus strand</td>
</tr>
<tr>
<td>1848280</td>
<td>2</td>
<td>NT_005265.11</td>
<td>8167756</td>
<td>188270833</td>
<td>minus strand</td>
</tr>
<tr>
<td>1544035</td>
<td>18</td>
<td>NT_025028.11</td>
<td>4990909</td>
<td>68576178</td>
<td>plus strand</td>
</tr>
<tr>
<td>1540801</td>
<td>3</td>
<td>NT_022517.13</td>
<td>1185663</td>
<td>19978681</td>
<td>plus strand</td>
</tr>
<tr>
<td>1519822</td>
<td>8</td>
<td>NT_023679.13</td>
<td>2113240</td>
<td>50269239</td>
<td>minus strand</td>
</tr>
<tr>
<td>1510332</td>
<td>3</td>
<td>NT_022463.12</td>
<td>879640</td>
<td>162748670</td>
<td>plus strand</td>
</tr>
<tr>
<td>1458904</td>
<td>8</td>
<td>NT_008251.11</td>
<td>403353</td>
<td>34579397</td>
<td>plus strand</td>
</tr>
<tr>
<td>1411426</td>
<td>9</td>
<td>NT_035014.2</td>
<td>1022925</td>
<td>127413028</td>
<td>minus strand</td>
</tr>
<tr>
<td>1396614</td>
<td>5</td>
<td>NT_006713.11</td>
<td>6482883</td>
<td>77351542</td>
<td>minus strand</td>
</tr>
<tr>
<td>1371374</td>
<td>5</td>
<td>NT_016755.11</td>
<td>2785127</td>
<td>124857651</td>
<td>plus strand</td>
</tr>
<tr>
<td>1366199</td>
<td>5</td>
<td>NT_030685.4</td>
<td>553283</td>
<td>115729193</td>
<td>plus strand</td>
</tr>
<tr>
<td>1335286</td>
<td>13</td>
<td>NT_024524.11</td>
<td>27334658</td>
<td>74268506</td>
<td>plus strand</td>
</tr>
<tr>
<td>1332234</td>
<td>9</td>
<td>NT_023974.13</td>
<td>3907284</td>
<td>31122790</td>
<td>plus strand</td>
</tr>
</tbody>
</table>
### SNPs With Non-Zero Coefficients

<table>
<thead>
<tr>
<th>SNP</th>
<th>Chrom.</th>
<th>Contig Accession</th>
<th>Contig Position</th>
<th>Chrom. Position</th>
<th>Hit Orientation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1275054</td>
<td>11</td>
<td>NT_033899.3</td>
<td>13114602</td>
<td>125088313</td>
<td>minus strand</td>
</tr>
<tr>
<td>124756</td>
<td>7</td>
<td>NT_007933.10</td>
<td>52899826</td>
<td>126200901</td>
<td>plus strand</td>
</tr>
<tr>
<td>116016</td>
<td>14</td>
<td>NT_026437.9</td>
<td>34866829</td>
<td>65373186</td>
<td>minus strand</td>
</tr>
<tr>
<td>1028484</td>
<td>6</td>
<td>NT_007540.11</td>
<td>2825315</td>
<td>64980073</td>
<td>plus strand</td>
</tr>
<tr>
<td>1024892</td>
<td>2</td>
<td>NT_022184.10</td>
<td>1857674</td>
<td>71585093</td>
<td>minus strand</td>
</tr>
<tr>
<td>1023617</td>
<td>5</td>
<td>NT_023132.10</td>
<td>4669166</td>
<td>173441991</td>
<td>plus strand</td>
</tr>
<tr>
<td>1019791</td>
<td>5</td>
<td>NT_034779.2</td>
<td>4954101</td>
<td>155150577</td>
<td>minus strand</td>
</tr>
<tr>
<td>1017935</td>
<td>4</td>
<td>NT_022782.10</td>
<td>2048357</td>
<td>42629852</td>
<td>minus strand</td>
</tr>
<tr>
<td>1014025</td>
<td>17</td>
<td>NT_010799.11</td>
<td>881757</td>
<td>28044553</td>
<td>plus strand</td>
</tr>
<tr>
<td>1002624</td>
<td>14</td>
<td>NT_037845.1</td>
<td>13318680</td>
<td>26562678</td>
<td>plus strand</td>
</tr>
</tbody>
</table>
**Receiver Operator Characteristic (ROC) Curve**

Figure 4: The ROC curve shows that linear SVMs perform **significantly** better than random guessing (dotted line). It also shows the accuracy if we tuned the SVM model to **bound** the false positive rate (since MM is rare). The point (5%, 42.5%) is noted in **green**. The point without tuning (35%, 77.5%) is noted in **blue**.
Comparisons with Other Algorithms

- After finishing analysis of the linear SVM results, we decided to try some other algorithms on this problem:
  - non-linear (Gaussian and polynomial) SVMs
  - decision trees
  - naïve Bayesian networks
  - ensembles of voting decision stumps
## Comparisons with Other Algorithms

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>“Old” Accuracy</th>
<th>“Young” Accuracy</th>
<th>Overall Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear SVMs</td>
<td>65.0%</td>
<td>77.5%</td>
<td>71.2%</td>
</tr>
<tr>
<td>Polynomial SVMs</td>
<td>50.0%</td>
<td>70.0%</td>
<td>60.0%</td>
</tr>
<tr>
<td>Gaussian SVMs</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Decision Trees</td>
<td>45.0%</td>
<td>52.5%</td>
<td>48.8%</td>
</tr>
<tr>
<td>Boosted Decision Trees</td>
<td>50.0%</td>
<td>50.0%</td>
<td>50.0%</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>42.5%</td>
<td>45.0%</td>
<td>43.8%</td>
</tr>
<tr>
<td>Ensemble of Voters</td>
<td>10.0%</td>
<td>40.0%</td>
<td>25.0%</td>
</tr>
</tbody>
</table>
Comparisons with Other Algorithms

- We see from this comparison, that our choice of using linear SVMs for this task was good.
- However, this comparison raises a number of questions:
  1. Why did polynomial SVMs do worse than linear?
  2. Why did Gaussian SVMs get 0% accuracy?
  3. Why did all other algorithms do the same or worse than random guessing?
Comparisons with Other Algorithms

- We see from this comparison, that our choice of using linear SVMs for this task was good.
- However, this comparison raises a number of questions:
  1. Why did polynomial SVMs do worse than linear?
     - Polynomial SVMs *can* separate between the absence and presence of homozygosity (which is not biologically relevant) they were likely led astray by irrelevant correlations.
  2. Why did Gaussian SVMs get 0% accuracy?
  3. Why did all other algorithms do the same or worse than random guessing?
Comparisons with Other Algorithms

- We see from this comparison, that our choice of using linear SVMs for this task was good.
- However, this comparison raises a number of questions:
  1. Why did polynomial SVMs do worse than linear?
  2. Why did Gaussian SVMs get 0% accuracy?
     - Because of the very large number of features compared to the number of patients, it is possible that Gaussian SVMs fit the training data so tightly that it simply memorized that data and was not able to generalize it at all.
     - Other ideas?
  3. Why did all other algorithms do the same or worse than random guessing?
Comparisons with Other Algorithms

• We see from this comparison, that our choice of using linear SVMs for this task was good.

• However, this comparison raises a number of questions:

  1. Why did polynomial SVMs do worse than linear?
  2. Why did Gaussian SVMs get 0% accuracy?
  3. Why did all other algorithms do the same or worse than random guessing?
     ♠ Naïve Bayes and EOVs assume feature independence which is strongly violated in this domain.
     ♠ Decision trees are not only able to separate presence and absence of homozygosity, but they are not robust with high-dimensional data.
Future Work

- Expand the study to include more patients in order to further validate these results.
- Use a denser coverage of the genome than the 3000 SNPs used.
- Use a more focused coverage of the genome: focused on those regions found in this study to be significant for MM predisposition.
- Further tune the SVM algorithm to use a smaller set of features for classification to gain better insight into those regions/genes that are important.
Future Work

- Obtain control data points — SNP data on individuals, at both “young” and “old” ages, without MM — to validate that SNP patterns do not change with age.

- Compare the SNPs found to be important with the genes found using related gene microarray studies.
Acknowledgments

• David Page
  University of Wisconsin

• Fenghuang Zhan, Bart Barlogie and John Shaughnessy, Jr.
  University of Arkansas for Medical Sciences

• Computation and Informatics in Biology and Medicine (CIBM)
  Training Program

• National Library of Medicine
  Grant No. 5T15LM007359

• SVM\textsuperscript{light} (http://svmlight.joachims.org)

• C5.0 (http://www.rulequest.com)

• EOV (http://www.cs.wisc.edu/~mwaddell/eov.html)