Symbolic Discriminant Analysis for Mining Gene Expression Patterns

Jason H. Moore, Joel S. Parker, Lance W. Hahn

Program in Human Genetics,
Department of Molecular Physiology and Biophysics,
Vanderbilt University Medical School
Nashville, TN

Introduction

- Questions
  - Can we classify and/or predict biological and clinical endpoints using gene expression data? Which genes are important? What is the pattern or statistical relationship among the genes?
- Statistical Challenges
  - Modeling
    - What statistical method do you use? How do you select a statistical model?
  - Variable Selection
    - > 5,000 gene expression variables
    - How do you select a subset of variables?
    - 100 variables ~ $1.27 * 10^{30}$ subsets
- Objectives
  - Develop a computational or statistical methodology that is able to handle the model and variable selection challenges.
  - Use this methodology to identify patterns of gene expression that classify and predict clinical endpoints.

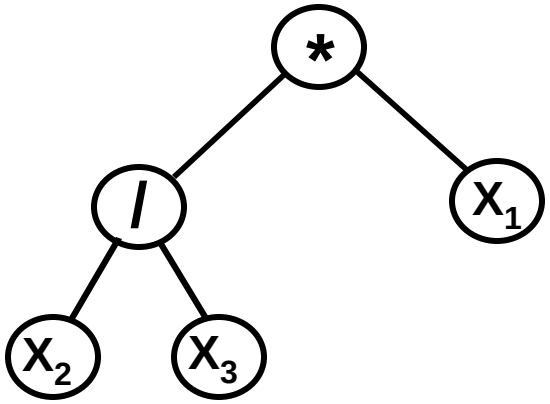
Symbolic Discriminant Analysis

- Supply list of gene expression variables
  - $X_1, X_2, \ldots X_{10,000}$
- Supply list of mathematical functions
  - +, -, *, /, abs, log, exp, sqrt
- Use variables and functions as building blocks

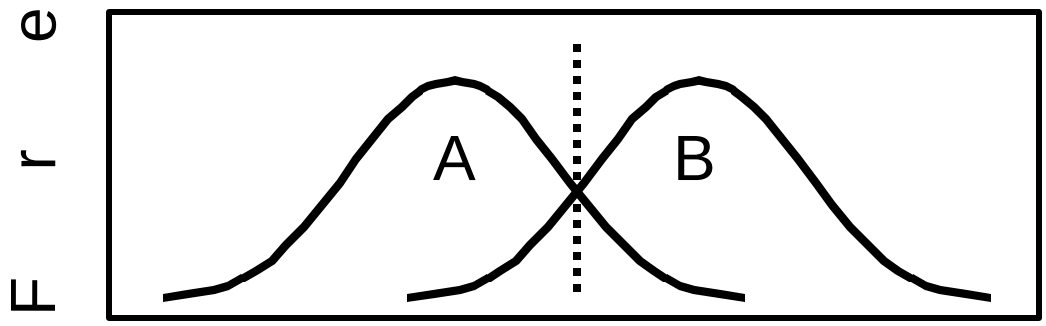
$$S_j = X_{1ij} * X_{2ij} / X_{3ij}$$

Symbolic Discriminant Analysis

- Supervised classification approach
- Use parallel genetic programming (GP) to build symbolic discriminant functions
- Misclassification rate is fitness function



$S = X_1 * X_2 / X_3$



Symbolic Discriminant Scores

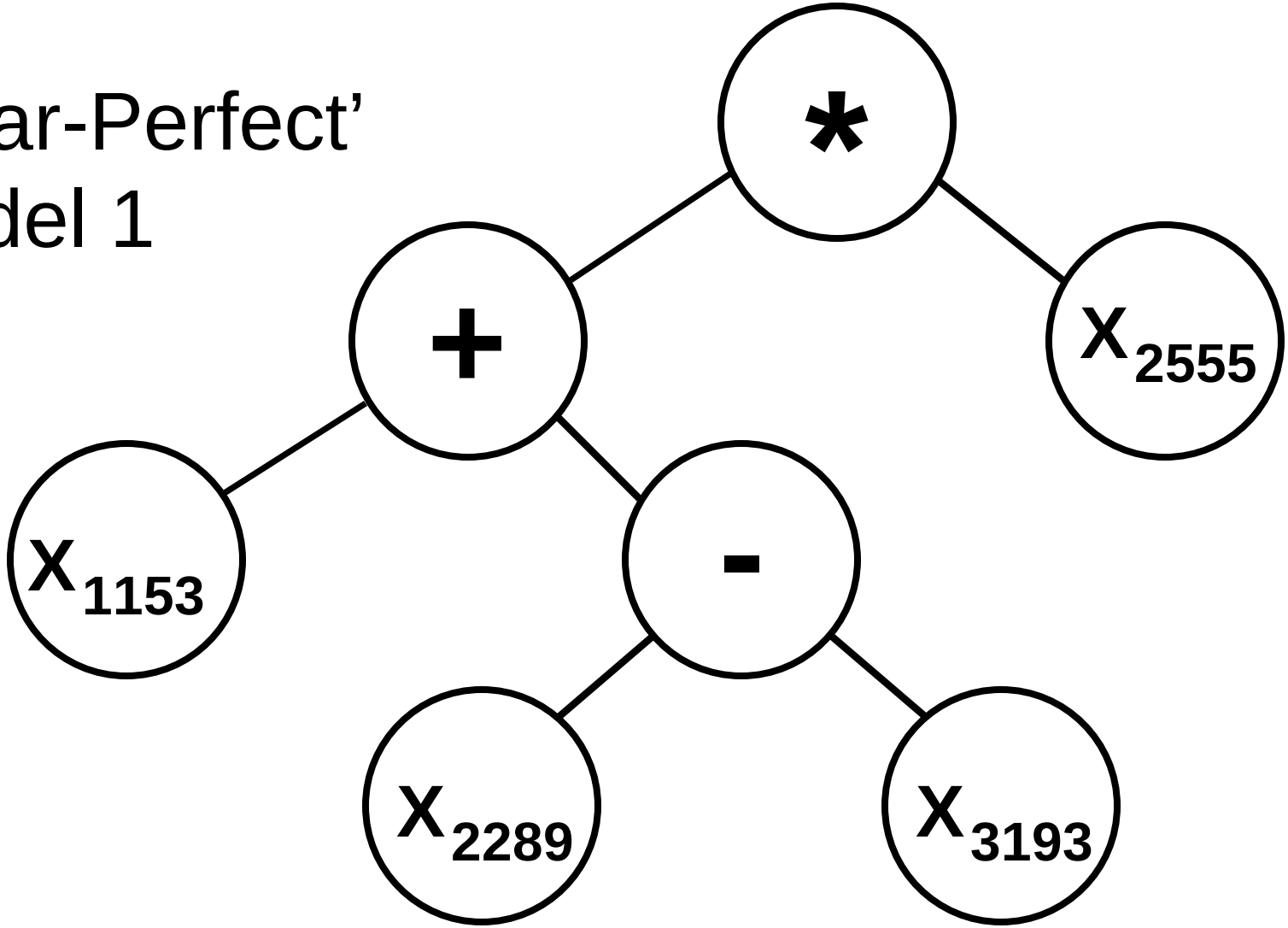
Application to Leukemia Data

- Leukemia Data (Golub et al. 1999)
  - Dataset 1 (n=38, training)
  - Dataset 2 (n=34, testing)
  - ~7100 expressed genes measured using Affymetrix oligonucleotide chips
- Cross Validation Strategy
  - Divide the training dataset into 38 equal parts.
  - Optimize SDA with each 37/38 of data.
  - Select SDA models that minimize the classification error and correctly predict the 1/38 of the data left out.
  - Estimate the prediction error using the testing dataset (n=34).
- Genetic Programming Settings
  - Population Size: 500
  - Iterations: 100
  - Populations: 4
  - Migration of best solutions every 25 iterations
  - Crossover probability: 0.6
  - Maximum depth: 6

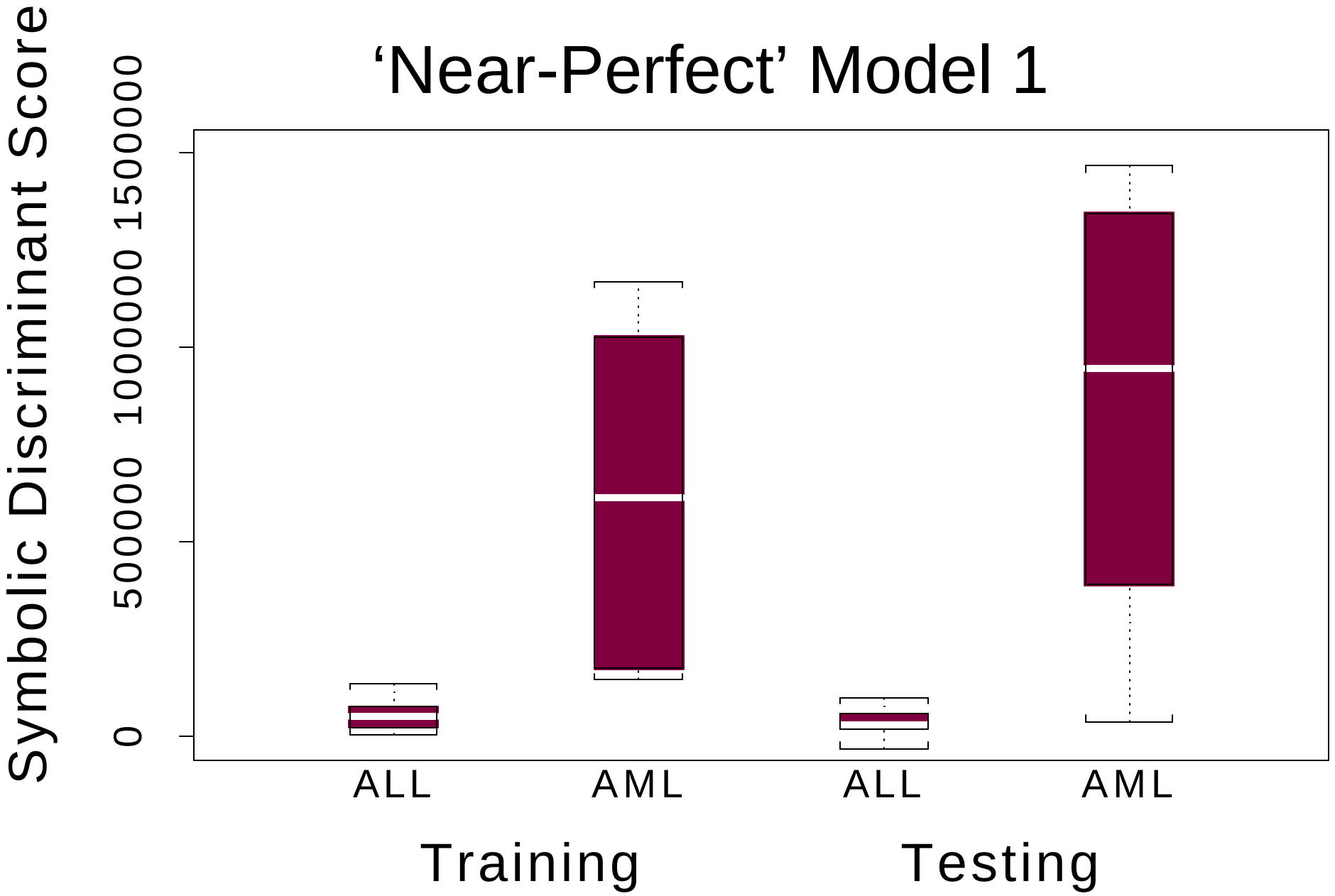
Results

- Identified 2 ‘near-perfect’ models
  - Classified 38/38 correctly
  - Predicted 33/34 correctly
- Identified 16 ‘very good’ models
  - Classified 38/38 correctly
  - Predicted 32/34 correctly
- Identified 36 ‘good’ models
  - Classified 38/38 correctly
  - Predicted 31/34 correctly

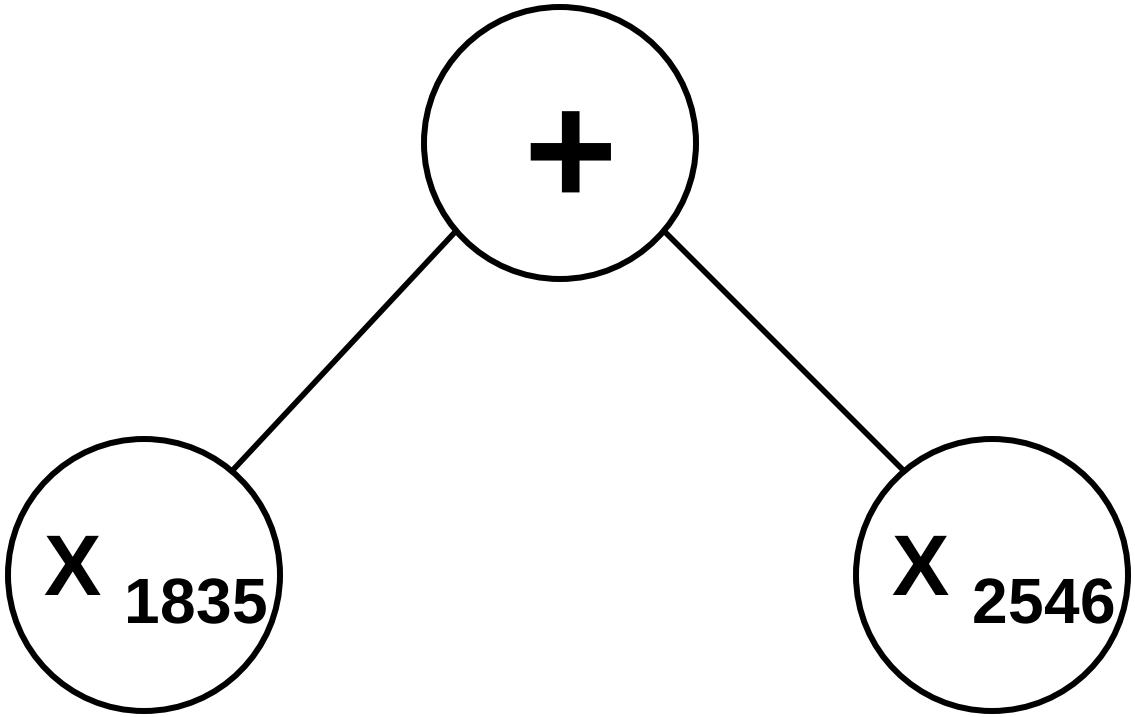
‘Near-Perfect’
Model 1



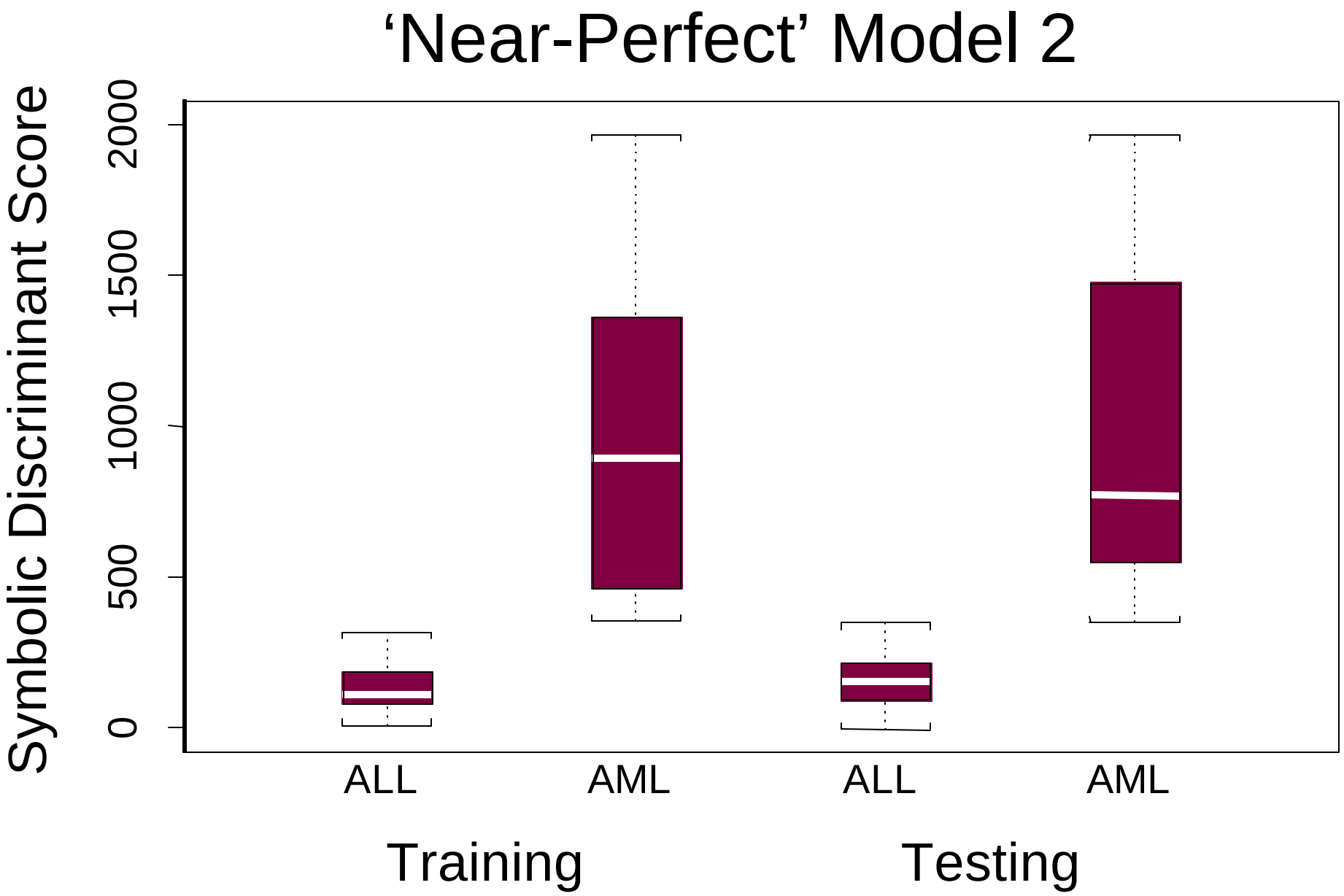
$$Y = X_{2555} * (X_{1153} + X_{2289} - X_{3193})$$


‘Near-Perfect’ Model 1
Symbolic Discriminant Score
0
5000000
10000000
15000000
ALL
AML
ALL
AML
Training
Testing

‘Near-Perfect’
Model 2



$$Y = X_{1835} + X_{2546}$$


‘Near-Perfect’ Model 2
Symbolic Discriminant Score
0
500
1000
1500
2000
ALL
AML
ALL
AML
Training
Testing

Which Genes Were Identified?

- Model 1: $Y = X_{2555} * (X_{1153} + X_{2289} - X_{3193})$
  - X_{2555}: Testis-specific cDNA on 17q
    - Cloned from a translocation, t(12;17), in a campomelic dysplasia patient.
  - X_{1153}: Erythroid beta-spectrin
    - Major component of red cell membrane, expressed during normal erythropoiesis
  - X_{2289}: Adipsin
    - Part of a gene cluster expressed during myeloid cell differentiation.
  - X_{3193}: Nucleoporin 98
    - Fuses with HOXA9 during an AML associated translocation, t(7;11)(p15;p15).
- Model 2: $Y = X_{1835} + X_{2546}$
  - X_{1835}: CD33
    - Differentiation antigen of AML progenitor cells.
  - X_{2546}: Rho E
    - Part of Rho family of signal transduction proteins
    - Lacks GTPase activity

Conclusions

- Symbolic discriminant analysis is a powerful alternative to traditional multivariate statistical methods.

- We anticipate this will be an important methodology to add to the repertoire approaches for mining gene expression patterns.