

## Statistical Analysis of High Content Screening Data

Rich Jacques

Department of Probability & Statistics  
University of Sheffield

22<sup>ND</sup> March 2006



## Outline

- Introduction
- Objectives
- Analyses
- Comparison of discriminant techniques
- Interim conclusions
- Further Work

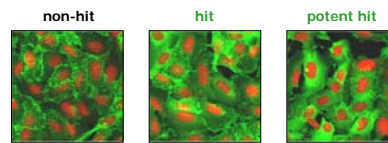
2

## Introduction

- In drug discovery at AstraZeneca the aim is to find compounds to fight a particular disease
- High Throughput Screens (HTS) allow a large number of compounds to be tested in a biological assay
- HTS produce high content images, one for each compound
- Advanced imaging algorithms used to produce a set of variables for each image

3

## Image Samples

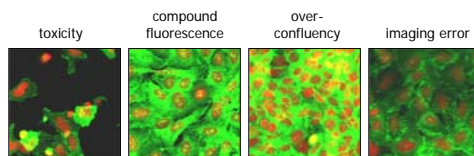


- Each compound classified by software as hit or non-hit
  - Only one variable used in this classification
- Images of the hits were checked by eye
  - False positives identified
- Images of non-hits were not checked
  - Not known how many non-hits were false negatives

4

## False Hits

- False positives are classified into nine different categories
- Example images from four of the nine false positive categories are shown below

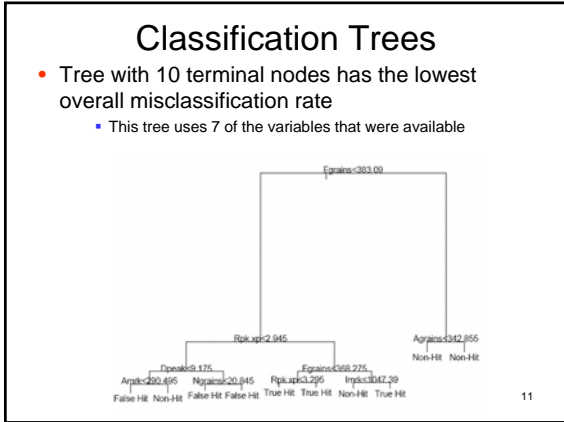
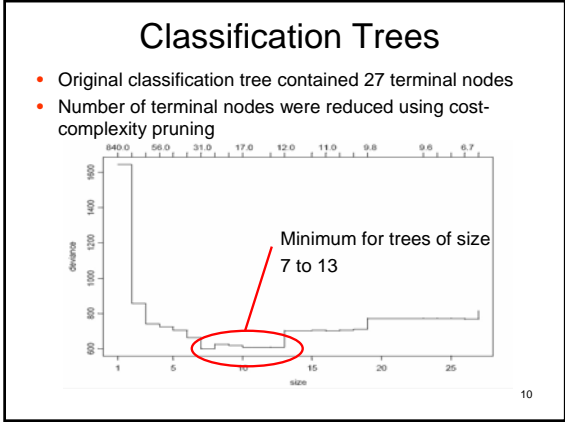
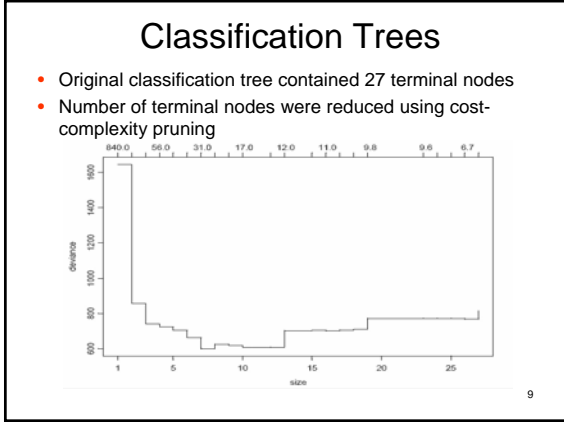
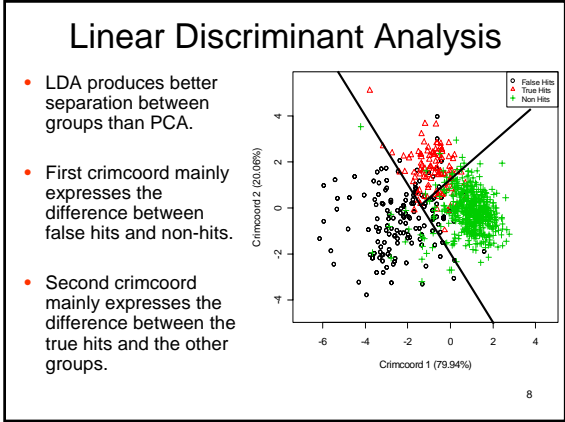
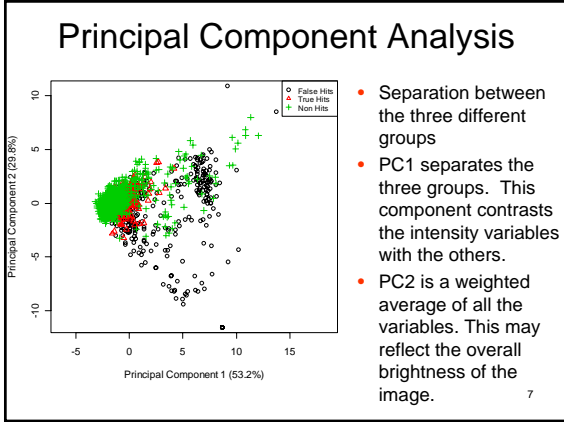


5

## Objectives

- Develop a multiparametric approach to enable a more refined compound selection
- Reduce the number of false positives in order to reduce the 'cost' of manual image inspection
- Avoid an increase in the number of false negatives since these could be developed into valuable new drugs

6



### Comparing Methods of Discrimination

Method	Data	Overall Misclassification Rate	% Misclassified		
			Non-Hits	False Hits	True Hits
LDA	Training	8.88%	3.30%	26.00%	36.04%
LDA	Test	9.37%	4.00%	29.70%	30.53%
QDA	Training	11.26%	9.00%	29.33%	7.21%
QDA	Test	12.30%	9.50%	24.85%	20.00%
C. Trees	Training	5.79%	1.90%	20.00%	21.62%
C. Trees	Test	7.94%	3.10%	23.03%	32.63%
N. Nets	Training	5.95%	3.90%	14.67%	12.61%
<b>N. Nets</b>	<b>Test</b>	<b>7.78%</b>	4.40%	23.64%	15.79%

12



### Comparing Methods of Discrimination

Method	Data	Overall Misclassification Rate	% Misclassified		
			Non-Hits	False Hits	True Hits
LDA	Training	8.88%	3.30%	26.00%	36.04%
LDA	Test	9.37%	4.00%	29.70%	30.53%
QDA	Training	11.26%	9.00%	29.33%	7.21%
QDA	Test	12.30%	9.50%	24.85%	20.00%
C. Trees	Training	5.79%	1.90%	20.00%	21.62%
C. Trees	Test	7.94%	<b>3.10%</b>	<b>23.03%</b>	32.63%
N. Nets	Training	5.95%	3.90%	14.67%	12.61%
N. Nets	Test	<b>7.78%</b>	4.40%	23.64%	<b>15.79%</b>

13

- ### Interim Conclusions
- Clear informative structure found in the data
    - Different types of hit can be distinguished
  - Neural networks have the lowest misclassification rate for test data
    - Neural networks harder to implement in HTS software than other discriminant techniques
  - Analysis identified five non-hits required to have their classification verified
    - Visual inspection identified that four should have originally been classified as true hits
- 14

- ### Future Work
- Classification using random forests
    - Is it possible to improve the misclassification from previous models?
  - Updating classification rules using unlabelled data
- 15

- Acknowledgements
    - Edward Ainscow
    - Advanced Science and Technology Laboratory, AstraZeneca Charnwood
  - References
    - Elizabeth Mills
    - Analysis of High-Content Cell-Biology Measurements
    - MSc Dissertation, University of Sheffield, 2004
- 

- 16