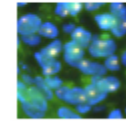




Genomic Markers. **Delivered.**



Biomarker Discovery in Genomic Data with Partial Clinical Annotation

Cole Harris, Noushin Ghaffari



Public microarray data

Status

- Data has been collected from a large number of samples
 - Breast cancer – 158 datasets in GEO repository

but

- Majority of data does not have associated clinical information to support clinically relevant biomarker discovery
 - Breast cancer prognosis – 22 datasets in GEO repository

Why?

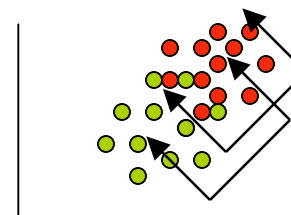
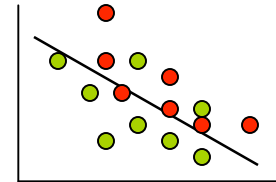
- Difficult and expensive to obtain
 - For prognostic marker discovery, long term follow-up is required.
 - For drug response marker discovery, patient response is required.

Concurrent mining approach

- *Our approach:*

- Map data across platforms to common gene set
- Within common genes, subsets of genes scored with objective function containing terms for:

- Accuracy in clinically annotated datasets
 - Crossvalidation, bootstrap
- Clustering in datasets lacking annotation
 - Inter-cluster vs. intra-cluster distance



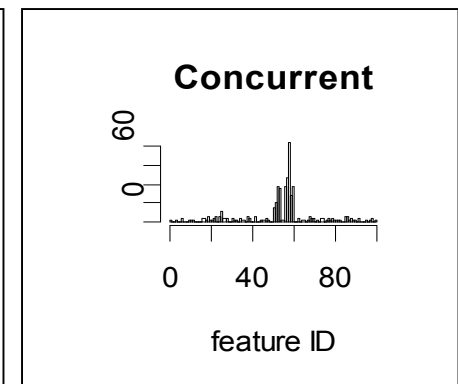
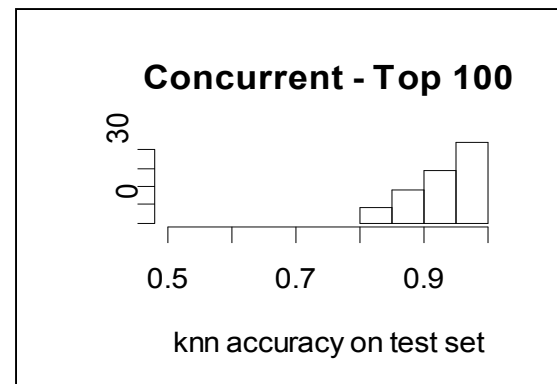
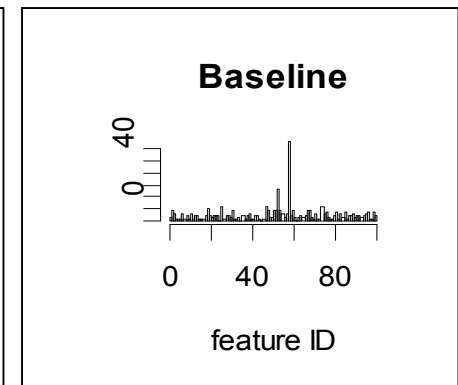
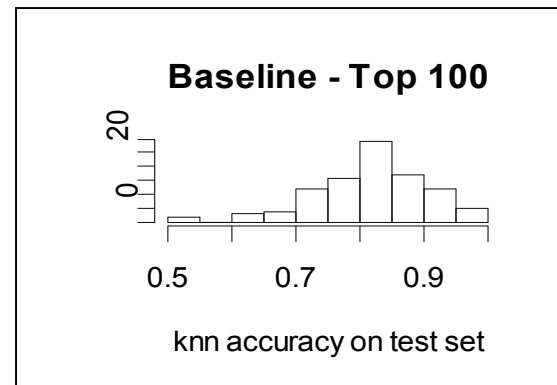
- *But still under development*

- Intrinsic assumptions violated?
 - under what circumstances?
- Optimal objective function?
 - simple approach?
 - information based approach?



Synthetic data example

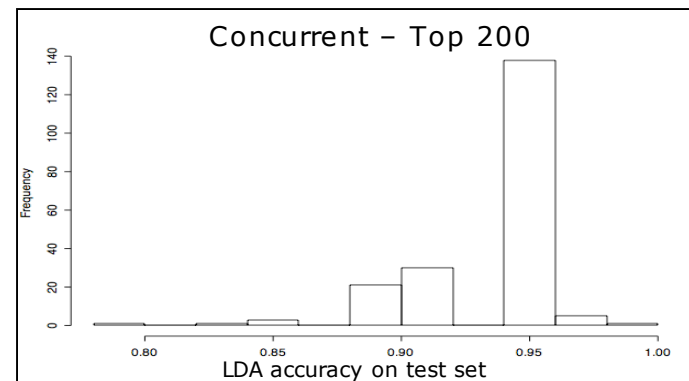
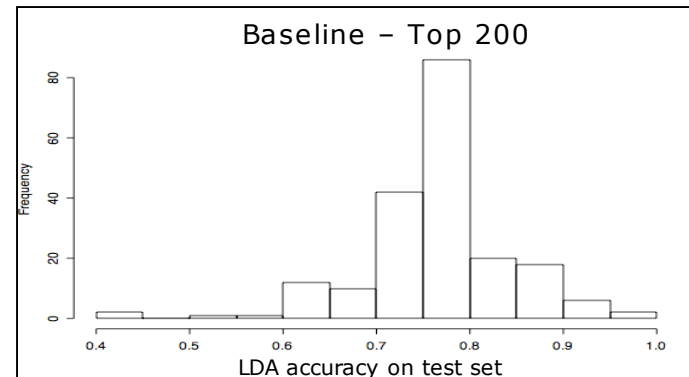
- **2 datasets:**
 - **Annotated:** 100 features across 40 samples (20X2 classes)
 - **Unannotated:** 100 features across 60 samples
 - **10 informative features**
 - IDs 51-60
- **20 annotated samples held out as test set** (10X2 classes)
- **GA search across 5-feature markers**
 - Baseline: LOOCV KNN (1nn) on annotated training data
 - Concurrent:
 - LOOCV KNN (1nn)
 - K-MEANS (nd=2)
 - distance between clusters/average cluster spread
 - error in expected cluster proportions





ALL/AML diagnosis

- **Data sources**
 - **Annotated:** Golub TR, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science. 1999 Oct 15;286(5439):531-7.
 - **Unannotated:** Armstrong SA, et al. MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. Nat Genet. 2002 Jan;30(1):41-7.
- **Annotated data**
 - Train - 38 samples (27 ALL, 11 AML)
 - Test - 34 samples (20 ALL, 14 AML)
 - 7129 genes
- **Unannotated data**
 - 52 samples (24 ALL, 28 AML)
 - 12,600 genes
- **6002 genes in common**
- **GA search across 3-gene markers**
 - Baseline: LDA on annotated training data
 - Concurrent:
 - LDA
 - K-MEANS (nd=2)
 - distance between clusters/average cluster spread
 - error in expected cluster proportions





Thank you for your attention

Questions?