Quantifying the discrimination power of various conditions in the Yeast data set

A. Jagota[1], M. Masso, W. W. vanOsdol[2]

1 University of California, Santa Cruz

2 Alza Corporation, Mountain View, California

Question and Motivation

- Many datasets of micro-array gene expression data contain expression patterns of genes over a set of *conditions* $C_1,...,C_k$ (alpha, elu, etc).

- These data sets are also *labeled* in that each gene is annotated with the broad *functional class* it belongs to (DNA replication, cell cycle, etc).

- *Do genes in different functional classes respond differently to different conditions?*

- If so, this might be exploitable to build
  - Better functional class predictors (by selectively using temporal patterns of particular conditions).
  - Better clustering methods (by treating the expression pattern of a gene not as a single vector but rather as a set of time-series, one time-series per condition).

Main Results

- This poster proposes a simple and intuitive measure of the discrimination power of a condition on a labeled data set.

- This measure may be used to rank different conditions in terms of their ability to predict the various functional classes.

- Applying this measure to a subset of the CAMDA data set revealed that the ELU condition had the poorest predictive accuracy on the chosen subset.

- A CART classifier was applied to this same data set to predict functional classes from the temporal patterns of individual conditions, one by one. The CART analysis revealed that the ELU condition was the poorest predictor, which agrees with the discrimination power result.

The Discrimination Power of a Condition

- Let $\mathbf{D}_i$ denote a data set of (temporal) patterns of the expressions of a set of genes $g_1,...,g_n$ for a specific condition C_i.
- Let $\mathbf{D}_i$ be *labeled*, specifically each pattern $\mathbf{d}_j$ (for gene g_j) has a class label, one of $1,...,k$, (for functional class of gene g_j).
- Let $\mathbf{D}_i^c$ denote the subset of $\mathbf{D}_i$ of those patterns whose functional class is c.
- Our measure of the discrimination power of condition C_i on data set $\mathbf{D}_i$ is:

$$\mathrm{dp}(C_i) = \underbrace{\frac{1}{|\mathbf{D}_i|}\sum_{c=1}^{k}\sum_{\mathbf{d}\in\mathbf{D}_i^c}\rho(\mathbf{d},\mu_i^c)}_{\text{average intra-class tightness}} - \underbrace{\frac{2}{k(k-1)}\sum_{\{c,c'\}\subseteq\{1,...,k\}}\rho(\mu_i^c,\mu_i^{c'})}_{\text{average inter-class separation}} \quad (0.1)$$

- Here μ_i^c is the mean vector of patterns in $\mathbf{D}_i^c$, and ρ is the usual correlation coefficient (Eisen et al 1998).
- The first term in (0.1) measures the *average intra-class tightness* and the second term measures the *average inter-class separation*.
- In the first term, the average is taken over all patterns in $\mathbf{D}_i$ and in the second term the average is taken over all pairs of classes.
- This averaging ensures that the contributions of the two terms are of the same scale.

Discrimination Power Results

Table 1: Discrimination power of four conditions, *alpha*, *cdc15*, *cdc28*, and *elu* on a subset of the CAMDA data set that contained expression patterns of 157 genes from the nine most populated functional classes. The nine functional classes with their populations were: *DNA repair (12), DNA replication (27), cell cycle (27), cell wall biogenesis (15), chromatin structure (16), cytoskeleton (17), mating (13), transcription (11)*, and *transport (19).*

9-class problem	Alpha	Cdc15	Cdc28	Elu
Average tightness	0.46	0.45	0.46	0.57
Average separation	-0.18	-0.13	-0.26	-0.59
Discrimination Power	0.28	0.32	0.2	-0.02

Discrimination Power Results

Table 2: Discrimination power of the same four conditions on a two-class problem: data set comprised of DNA replication genes (27) and cell cycle genes (27).

DNA replication vs cell cycle	Alpha	Cdc15	Cdc28	Elu
Average tightness	0.438	0.32	0.458	0.673
Average separation	-0.519	-0.583	-0.504	-0.66
Discrimination Power	-0.081	-0.263	-0.046	0.013

Discrimination Power Results

Table 3: Discrimination power of the same four conditions on another two-class problem: data set comprised of DNA replication genes (27) and Transport genes (27).

DNA replication vs Transport	Alpha	Cdc15	Cdc28	Elu
Average tightness	0.47	0.53	0.54	0.55
Average separation	-0.39	0.41	0.21	-0.27
Discrimination Power	0.08	0.94	0.75	0.28

CART Analysis Results

Analog of Table 1: Prediction accuracy of individual conditions on 9-class problem. ELU portion in agreement with **Table 1**.

ALPHA	CDC15	CDC28	ELU
35%	41%	40%	25%

Analog of Table 2: DNA replication versus cell cycle. Don't agree with **Table 2**.

ALPHA	CDC15	CDC28	ELU
92%	80%	76%	69%

Analog of Table 3: DNA replication versus Transport. CDC15, CDC28 agree well with **Table 3**.

ALPHA	CDC15	CDC28	ELU
84.7%	93.2%	93.3%	82.6%

(Function, Condition) Tightnesses

Table 4

	Alpha	Cdc15	Cdc28	Elu
Cell cycle	0.3	0.15	0.36	0.58
Cell wall biogenesis	0.33	0.4	0.43	0.51
Chromatin structure	0.7	0.66	0.63	0.85
Cytoskeleton	0.44	0.48	0.58	0.56
DNA repair	0.79	0.6	0.8	0.75
DNA replication	0.58	0.49	0.55	0.77
Mating	0.49	0.63	0.4	0.46
Transcription	0.27	0.24	0.15	0.24
Transport	0.32	0.59	0.52	0.26

9-class CART Analysis Paired With Tightnesses

Table 5: In each cell, the first entry is from **Table 4**. The second entry is from the CART 9-class analysis, specifically the prediction accuracy of CART on the particular (class,condition) pair. Rows (or row slices) in which tightness and accuracy seem correlated. Rows (or row slices) which buck this trend

	Alpha	Cdc15	Cdc28	Elu
Cell cycle	0.3, 62.5%	0.15, 43.75%	0.36, 62.5%	0.58, 78%
Cell wall biogenesis	0.33, 0%	0.4, 0%	0.43, 0%	0.51, 0%
Chromatin structure	0.7, 0%	0.66, 46.6%	0.63, 56.2%	0.85, 0%
Cytoskeleton	0.44, 0%	0.48, 0%	0.58, 0%	0.56, 0%
DNA repair	0.79, 0%	0.6, 0%	0.8, 0%	0.75, 0%
DNA replication	0.58, 92.5%	0.49, 61.5%	0.55, 74%	0.77, 63%
Mating	0.49, 68.7%	0.63, 76.5%	0.4, 53%	0.46, 0%
Transcription	0.27, 0%	0.24, 0%	0.15, 0%	0.24, 0%
Transport	0.32, 0%	0.59, 89%	0.52, 50%	0.26, 0%

Discussion and Future Work

- The discrimination power and the CART analyses reveal that different conditions have differing abilities to predict functional classes.
- Both the discrimination power and the CART analysis suggests that ELU is the poorest discriminator among the four conditions.
- Our immediate future interest is in building a classifier that exploits the differing discrimination power of different conditions. This may take the form of a decision tree method.
- We are also interested in exploiting these ideas in cluster analysis, in particular in developing a (dis)similarity measure that treats different conditions differently.