Progress and lessons in the analysis of microarray data:

Wing Hung Wong
Departments of Statistics and Biostatistics
Harvard University

Outline

• Model-based approach: dealing with saturation and bad probes
  Xuemin Fang, Cheng Li
• Class prediction: Combined supervised and unsupervised learning
  Xuegong Zhang
• Incorporation of biological knowledge into computational algorithms: first steps
  Yan Cui
Gene expression can be measured by microarrays.

Affymetrix oligonucleotide arrays use multiple match-mismatch probe pairs to interrogate each gene.
one gene, 42 animals, raw data

Model for one gene

\[ X_{ij0} = \mu_j + \theta_i \phi_j + \epsilon_{ij0} \]
\[ X_{ij1} = \mu_j + \theta_i \phi_j + \theta_i \alpha_j + \epsilon_{ij1} \]

\( X_{ij0} \) and \( X_{ij1} \): observed MM and PM values for chip i, probe j,
\( \mu_j \): baseline hybridization level for probe pair j,
\( \theta_i \): expression index in chip i,
\( \phi_j, \alpha_j \): rate factors characterizing probe pair j

constraint: \( \mu_j \geq 0, \theta_i \geq 0, \phi_j \geq 0, \alpha_j \geq 0 \)
Corner misalignment discovered through outlier analysis:

Array 9 initially has an unusually large number of array and single outliers in the lower-left region. (B) The lower-left corner pixel position (white dot) appears to be off by about one feature and therefore leads to incorrect gridding and averaging of many features in the lower-left region. This is hard to detect by visual inspection of the original image. (C) After manually setting the correct corner pixel position, the array is salvaged.

Uses of the model (implemented in dChip Analyzer by Cheng Li)

- Automatic detection and handling of outlier observations, arrays or probes.
- Construction of standard errors and confidence intervals for expression indexes and fold changes.
- Assess cluster reliability, …
Affymetrix Latin Square Experiment
(Data Source: Steve Smeeken)

• One lot of 14 arrays were used for the study using spikes along with the background-- enriched RNA from DR153 (Ecoli) strain.

• Twenty assay spikes were labeled transcripts generated from the deletion area of DR153 chromosome, while their background RNA was isolated from DR153, subsequently enriched and direct-labeled.

• The twenty assay spikes were grouped into 14 sets and its concentration (nM) prepared in the mix according to the Latin-square scheme.

### Latin Square Scheme

| Mi   | x0 | b0 | x1 | c | e 
|------|----|----|----|---|---
| 1    | 0  | 4  | 2  | 1  | 3 
| 2    | 2  | 0  | 6  | 3  | 1
| 3    | 2  | 6  | 1  | 3  | 0 
| 4    | 1  | 3  | 0  | 6  | 2
| 5    | 6  | 1  | 3  | 0  | 2 
| 6    | 3  | 0  | 6  | 2  | 1
| 7    | 1  | 3  | 0  | 6  | 2 
| 8    | 6  | 1  | 3  | 0  | 2 
| 9    | 3  | 0  | 6  | 2  | 1
| 10   | 1  | 3  | 0  | 6  | 2 
| 11   | 6  | 1  | 3  | 0  | 2 
| 12   | 3  | 0  | 6  | 2  | 1
| 13   | 1  | 3  | 0  | 6  | 2 
| 14   | 6  | 1  | 3  | 0  | 2 

1 1024 512 256 128 64 32 32 16 16 8 8 4 4 2 2 1 0.5 0.5 0.25 0.25 0.125 0.125 0.0625 0.0625 0.03125 0.03125
2 512 256 128 64 32 16 16 8 8 4 4 2 2 1 0.5 0.5 0.25 0.25 0 0 0.0625 0.0625 0.03125 0.03125
3 256 128 64 32 16 8 8 4 4 2 2 1 0.5 0.5 0.25 0.25 0 0 0 0 0 0
4 128 64 32 16 8 8 4 4 2 2 1 0.5 0.5 0.25 0.25 0 0 0 0 0 0
5 64 32 16 8 8 4 4 2 2 1 0.5 0.5 0.25 0.25 0 0 0 0 0 0
6 32 16 8 8 4 4 2 2 1 0.5 0.5 0.25 0.25 0 0 0 0 0 0
7 16 8 8 4 4 2 2 1 0.5 0.5 0.25 0.25 0 0 0 0 0 0
8 8 4 4 2 2 1 0.5 0.5 0.25 0.25 0 0 0 0 0 0
9 4 2 2 1 0.5 0.25 0 0 0 0 0 0 0 0 0 0 0
10 2 1 0.5 0.25 0 0 0 0 0 0 0 0 0 0 0 0 0
11 1 0.5 0.25 0 0 0 0 0 0 0 0 0 0 0 0 0 0
12 0.5 0.25 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
13 0.25 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
14 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

6
Scanning Saturation

Probe 1-15: PM and MM data overlaid.

Probe 1-30: PM and MM data concatenated.

Probe 1-30: Red fitted curve overlaid.
Saturation

Histogram of PM value

QQ-norm Plot of PM value

Histograms of cel intensities of the 14 arrays
Handle Saturation

\[ X \sim \text{Normal}(\mu, \sigma^2) \]

Impute \( E(X \mid X > \text{observed value}) \)

Conditional Expectation

Given \( X > m \), what's \( E(X \mid X > m) \)?

\[
E(X \mid X > m) = \int_{-\infty}^{\inf} x f(x \mid x > m) \, dx
\]

\[
= \frac{\int_{m}^{\inf} x f(x) \, dx}{P(x > m)} = \frac{\int_{m}^{\inf} x f(x) \, dx}{1 - \Phi \left( \frac{m - \mu}{\sigma} \right)}
\]
Conditional Expectation (cont’d)

when $m > \mu$:

$$E(X \mid X > m) = \mu + \frac{\sigma e^{-(m-\mu)^2/2\sigma^2}}{\sqrt{2\pi} \left(1 - \Phi \left(\frac{m-\mu}{\sigma}\right)\right)}$$

when $m < \mu$:

$$E(X \mid X > m) = \mu + \frac{\sigma e^{-(m-\mu)^2/2\sigma^2} - \mu \Phi \left(\frac{m-\mu}{\sigma}\right)}{1 - \Phi \left(\frac{m-\mu}{\sigma}\right)}$$

After imputing saturated data points

Probe 1-15: PM and MM data overlaid.

Probe 1-30: PM and MM data concatenated

Probe 1-30: Red fitted curve overlaid.
Full model with imputation

Reduced (black) vs AD (orange) square root
Similar experiment using U95A human arrays (no deletions)

- Saturation is more subtle and hard to correct
- Instead, focus on how to model probe level data to get better linearity at the low intensity range.
- Finding 1: Extrapolation of probe sensitivity to low signal range produce better linearity
- Finding 2: Useful to treat MM as “unobserved” if it exceeds PM.

Results from Full model fitting:

- For each gene there’re 5 slides:
  1. Full model fitting for all arrays
  2-4. Full model fitting for the three windows
  5. Estimated theta vs absolute concentration
- In all the model fitting, we used constant background and omitted those MM values that are greater than their PM values.
- Number of arrays = 20
- Number of different concentration levels = 14
- Number of spike-in genes = 14
- Window size = 11
Gene 4 (9261)

- 38734_at
- M63603
- 5350
- phospholamban
- |circulation|physiological processes|cell motility|cell growth and/or maintenance|transport|transporter|intracellular|cell|
- [6]6q
- Cluster Incl. M63603: Human phospholamban mRNA, complete cds /cds=(181,339) /gb=M63603 /gi=189942 /ug=Hs.85050 /len=1635
seq(0, 31)
c(pm[i], mm[i])

0 5 10 15 20 25 30

200 400 600

c(

0 5 10 15 20 25 30

200 400 600 800

c(

0 5 10 15 20 25 30

200 400 600 800 1000 1200

c(

0 5 10 15 20 25 30

200 400 600 800

c(

0 5 10 15 20 25 30

2000 4000 6000

c(

0 5 10 15 20 25 30

2000 4000 6000 8000

c(

0 5 10 15 20 25 30

2000 4000 6000 8000 10000

c(

0 5 10 15 20 25 30

2000 6000 10000

c(

0 5 10 15 20 25 30

2000 6000 10000 14000

c(

Absolute concentration

theta

0 2 4 6 8 1 0 1 2 1 4

full model

THETA^(1/3)

0 2 4 6 8 1 0 1 2 1 4 6 8 10 12 14

full model with 3 windows
Gene 6 (2681)

- 36311_at
- U40370
- 5136
- phosphodiesterase 1A, calmodulin-dependent

Cluster Incl. U40370: Human 3, 5 cyclic nucleotide phosphodiesterase (HSPDE1A3A) mRNA, complete
cds/cds=(84, 1691)/gb=U40370/gi=1151108/ug=Hs.41717/len=2008

```
concentration = 4
seq(0, 31)
c(fit1$pm[i,], fit1$mm[i,])
0 5 10 15 20 25 30
500 1000 1500
```

```
concentration = 8
seq(0, 31)
c(fit1$pm[i,], fit1$mm[i,])
0 5 10 15 20 25 30
200 400 600 800 1000
```

```
concentration = 16
seq(0, 31)
c(fit1$pm[i,], fit1$mm[i,])
0 5 10 15 20 25 30
500 1000 1500
```

```
concentration = 32
seq(0, 31)
c(fit1$pm[i,], fit1$mm[i,])
0 5 10 15 20 25 30
500 1000 1500
```

```
concentration = 64
seq(0, 31)
c(fit1$pm[i,], fit1$mm[i,])
0 5 10 15 20 25 30
500 1000 1500
```

```
concentration = 128
seq(0, 31)
c(fit1$pm[i,], fit1$mm[i,])
0 5 10 15 20 25 30
500 1000 1500
```

```
concentration = 256
seq(0, 31)
c(fit1$pm[i,], fit1$mm[i,])
0 5 10 15 20 25 30
500 1000 1500
```

```
concentration = 512
seq(0, 31)
c(fit1$pm[i,], fit1$mm[i,])
0 5 10 15 20 25 30
500 1000 1500
```

```
concentration = 1024
seq(0, 31)
c(fit1$pm[i,], fit1$mm[i,])
0 5 10 15 20 25 30
500 1000 1500
```

```
concentration = 0
seq(0, 31)
c(fit1$pm[i,], fit1$mm[i,])
0 5 10 15 20 25 30
500 1000 1500
```

```
concentration = 0.25
seq(0, 31)
c(fit1$pm[i,], fit1$mm[i,])
0 5 10 15 20 25 30
500 1000 1500
```

```
concentration = 0.5
seq(0, 31)
c(fit1$pm[i,], fit1$mm[i,])
0 5 10 15 20 25 30
500 1000 1500
```

```
concentration = 1
seq(0, 31)
c(fit1$pm[i,], fit1$mm[i,])
0 5 10 15 20 25 30
500 1000 1500
```

```
concentration = 1
seq(0, 31)
c(fit1$pm[i,], fit1$mm[i,])
0 5 10 15 20 25 30
500 1000 1500
```

```
concentration = 2
seq(0, 31)
c(fit1$pm[i,], fit1$mm[i,])
0 5 10 15 20 25 30
500 1000 1500
```

```
concentration = 2
seq(0, 31)
c(fit1$pm[i,], fit1$mm[i,])
0 5 10 15 20 25 30
500 1000 1500
```

```
concentration = 2
seq(0, 31)
c(fit1$pm[i,], fit1$mm[i,])
0 5 10 15 20 25 30
500 1000 1500
```

```
concentration = 2
seq(0, 31)
c(fit1$pm[i,], fit1$mm[i,])
0 5 10 15 20 25 30
500 1000 1500
```
<table>
<thead>
<tr>
<th>Concentration</th>
<th>Graphs</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td><img src="#" alt="Graphs" /></td>
</tr>
<tr>
<td>0.25</td>
<td><img src="#" alt="Graphs" /></td>
</tr>
<tr>
<td>0.5</td>
<td><img src="#" alt="Graphs" /></td>
</tr>
<tr>
<td>1</td>
<td><img src="#" alt="Graphs" /></td>
</tr>
<tr>
<td>2</td>
<td><img src="#" alt="Graphs" /></td>
</tr>
<tr>
<td>4</td>
<td><img src="#" alt="Graphs" /></td>
</tr>
<tr>
<td>8</td>
<td><img src="#" alt="Graphs" /></td>
</tr>
<tr>
<td>16</td>
<td><img src="#" alt="Graphs" /></td>
</tr>
</tbody>
</table>
Absolute concentration

$\theta$

$\theta^{(1/3)}$

full model

full model with 3 windows
Class prediction from microarray data

Question:
Is metastasis partially predictable based on expression profiles?

Data:
Affymetrix U95A arrays on 89 cancer samples

Preliminary observations:

Lymph node positive (LN+) and negative (LN-) patients are evenly distributed in pathologically defined subgroups.

Known molecular markers are not predictive for lymph-node status -- not much better than random guessing.
Support Vector machine learning and gene selection

Start with \(d=12,558\) genes

**Step 1.** Build the SVM classifier with all the genes.

**Step 2.** Evaluate the contribution of each gene in this classifier.

**Step 3.** For certain pre-decided \(d\), select the top \(d\) genes from the rank list.

Repeat Step 1 and Step 2 with \(d\) genes.

Assess performance by Cross-Validation and Permutation test.

---

**Cross Validation** Errors on all 89 cases:

<table>
<thead>
<tr>
<th>#SelectedGenes</th>
<th>CV Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>12558</td>
<td>0.415730</td>
</tr>
<tr>
<td>1000</td>
<td>0.393258</td>
</tr>
<tr>
<td>500</td>
<td><strong>0.382022</strong></td>
</tr>
<tr>
<td>200</td>
<td>0.438202</td>
</tr>
<tr>
<td>100</td>
<td>0.460674</td>
</tr>
<tr>
<td>50</td>
<td>0.528090</td>
</tr>
<tr>
<td>30</td>
<td>0.449438</td>
</tr>
<tr>
<td>20</td>
<td>0.490909</td>
</tr>
</tbody>
</table>

1000 Permutation:

\[ P(\text{CV error} \leq 0.382) = 0.125 \]
Gene expression data (not metastasis) is used to cluster samples into two subgroups.

<table>
<thead>
<tr>
<th>#SelectedGenes</th>
<th>CV Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>12558</td>
<td>0.352941</td>
</tr>
<tr>
<td>1000</td>
<td>0.411765</td>
</tr>
<tr>
<td>500</td>
<td>0.323529</td>
</tr>
<tr>
<td>200</td>
<td>0.294118</td>
</tr>
<tr>
<td>100</td>
<td>0.323529</td>
</tr>
<tr>
<td>50</td>
<td>0.352941</td>
</tr>
<tr>
<td>30</td>
<td>0.352941</td>
</tr>
<tr>
<td>20</td>
<td>0.470588</td>
</tr>
</tbody>
</table>

Permutations: $P = 0.076$
Preliminary conclusions

- Gene expression data are able to improve prediction of metastasis
- Gene expression data identify two subgroups of patients that seem to follow different mechanism of metastasis.
- Combining supervised with unsupervised methods can be useful
An attempt for Integrative analysis of gene expression profile and gene function classification

Mouse Tumor Maintenance Experiment

- Doxycycline-inducible H-Ras mouse melanoma model null for INK4a
- The apoptotic phenotype peaks at ~48 hrs and tumor regression becomes noticeable at 72 hrs

Source: Lynda Chin, et al. Essential role for oncogenic Ras in tumor maintenance
cDNA Microarray Data

- The relative abundance of about 20,000 transcripts between RAS-off (doxycycline off) and RAS-expressing mouse tumor cells was measured at 0, 24, 36, 48 and 72 hours respectively after the withdrawal of doxycycline.

Ras Signaling Pathway

INK4a in cell cycle control

- Very large amount of informative data generated by each experiment
- There is already a great deal of knowledge about the biological system under investigation
- Basic Challenge: how to bring the biology into the computational analysis of experimental data
Indicator Table: Coding Gene Expression Profile

<table>
<thead>
<tr>
<th>Gene</th>
<th>Up/Exp1</th>
<th>Down/Exp1</th>
<th>Up/Exp2</th>
<th>Down/Exp2</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene 1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>Gene 2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>Gene 3</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>Gene 4</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>...</td>
</tr>
<tr>
<td>Gene 5</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>Gene 6</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Gene Ontology Types

- **Molecular Function**
  - the tasks performed by individual gene products; examples are *transcription factor* and *DNA helicase*

- **Biological Process**
  - broad biological goals, such as *mitosis* or *purine metabolism*, that are accomplished by ordered assemblies of molecular functions

- **Cellular Component**
  - subcellular structures, locations, and macromolecular complexes; examples include *nucleus*, *telomere*, and *origin recognition complex*

- As of June 30, 2001 GO contains 3021 process, 4018 function and 655 component terms.
Indicator Table: Coding GO Categories

<table>
<thead>
<tr>
<th>Func 1</th>
<th>Func 2</th>
<th>Func 3</th>
<th>Func 4</th>
<th>Func 5</th>
<th>...</th>
<th>...</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene 1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Gene 2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Gene 3</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Gene 4</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Gene 5</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Gene 6</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Analyzing Categorical Data: Looking gene expression data in the knowledge background of gene function classification

<table>
<thead>
<tr>
<th>Expression Categories</th>
<th>Function categories:</th>
</tr>
</thead>
<tbody>
<tr>
<td>up/Exp1</td>
<td>cell cycle</td>
</tr>
<tr>
<td>down/Exp1</td>
<td>apoptosis</td>
</tr>
<tr>
<td>up/Exp2</td>
<td>cell growth and maintenance</td>
</tr>
<tr>
<td>down/Exp2</td>
<td>signal transduction</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Indicator Table: Unified coding of expression categoris and functional categories

<table>
<thead>
<tr>
<th>Gene 1</th>
<th>Up/Exp1</th>
<th>Down/Exp1</th>
<th>Func 1</th>
<th>Func 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene 2</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Gene 3</td>
<td>0</td>
<td>1</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Gene 4</td>
<td>1</td>
<td>0</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Gene 5</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Gene 6</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Representing genes and categories in a graph

Produce informative graph by minimizing total squared length of the edges.
• An *Alternating Least Squares* (ALS) algorithm is employed for minimizing the loss function.
• This is similar to Homogeneity Analysis by means of alternating least squares, or Gifi analysis.

**Gifi Array Analyzer (Y. Cui)**

• A Java program implementing this method for analyzing and visualizing DNA microarray data
The Structure of *Gifi Array Analyzer*

- Load MicroArray Data
- Make Indicator Table
- Homogeneity Analysis
- Interactive Plot
  
  JDBC
  
  Download and Parse Databases files
  
  UniGene, LocusLink
  
  GeneOntology
  
  Run SQL Queries

Dox-inducible melanoma model:
Time course profiles during Dox withdrawal
Another Ink4-/- experiment
<table>
<thead>
<tr>
<th>Gene Ontology ID</th>
<th>Biological Process</th>
<th>Category Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>05010</td>
<td>Protein transport</td>
<td>130</td>
</tr>
<tr>
<td>05014</td>
<td>Cell movement</td>
<td>24</td>
</tr>
<tr>
<td>05031</td>
<td>Cell growth and maintenance</td>
<td>50</td>
</tr>
<tr>
<td>05032</td>
<td>Transcription factor</td>
<td>42</td>
</tr>
<tr>
<td>05033</td>
<td>Translation factor</td>
<td>41</td>
</tr>
<tr>
<td>05034</td>
<td>Metabolism</td>
<td>25</td>
</tr>
</tbody>
</table>

**Gene Ontology ID**

<table>
<thead>
<tr>
<th>Molecular Function</th>
<th>Category Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATP binding</td>
<td>274</td>
</tr>
<tr>
<td>GTP binding</td>
<td>211</td>
</tr>
<tr>
<td>Protein binding</td>
<td>214</td>
</tr>
<tr>
<td>Nucleic acid binding</td>
<td>42</td>
</tr>
</tbody>
</table>

**Gene Ontology ID**

<table>
<thead>
<tr>
<th>Cellular Component</th>
<th>Category Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Membranes</td>
<td>499</td>
</tr>
<tr>
<td>Membrane fraction</td>
<td>333</td>
</tr>
<tr>
<td>Mitochondrion</td>
<td>118</td>
</tr>
<tr>
<td>Nucleus</td>
<td>499</td>
</tr>
</tbody>
</table>

**Gene Ontology ID**

<table>
<thead>
<tr>
<th>Membrane fraction</th>
<th>Category Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Membrane fraction</td>
<td>333</td>
</tr>
<tr>
<td>Mitochondrion</td>
<td>118</td>
</tr>
</tbody>
</table>

**Gene Ontology ID**

<table>
<thead>
<tr>
<th>Nuclear envelope</th>
<th>Category Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nuclear envelope</td>
<td>499</td>
</tr>
<tr>
<td>Membrane fraction</td>
<td>333</td>
</tr>
<tr>
<td>Mitochondrion</td>
<td>118</td>
</tr>
</tbody>
</table>
Yeast MAPK Pathways
